Introduction to Machine Learning in the Digital Humanities

Paul Barrett
Nathan Taback
Welcome to DHSI 2018!

Thanks for joining the DHSI community!

In this booklet, you will find essential course materials prefaced by some useful information about getting settled initially at UVic, finding your way around, getting logged in to our network (after you’ve registered the day before our courses begin), and so on.

Given our community’s focus on things computational, it will be a surprise to no one that we might expect additional information online for some of the classes - your instructors will let you know - or that the most current version of all DHSI-related information may be found on our website at dhsi.org.

To access the DHSI wifi network, simply go into your wireless settings and connect to the “DHSI” network and enter the password “dhsi2018”.

And please don’t hesitate to be in touch with us at institut@uvic.ca or via Twitter at @AlyssaA_DHSI or @DHInstitute if we can be of any help ....
The 2018 schedule is just about ready! A very few things to confirm, add, etc, but this is the place to be to find out what is happening when / where ...

Psst: Some Suggested Outings

If you're here a day or two before we begin, or staying a day or two afterwards, here are a few ideas of things you might consider doing ....

▼ Suggested Outing 1, Botanical Beach (self-organised; car needed)
A self-guided visit to the wet, wild west coast tidal shelf (and historically-significant former research site) at Botanical Beach; we recommend departing early (around 8.00 am) to catch low tide for a better view of the wonderful undersea life! Consider bringing a packed lunch to nibble-on while looking at the crashing waves when there, and then have an afternoon drink enjoying the view from the deck of the Port Renfrew Hotel.

▼ Suggested Outing 2, Butchart Gardens (self-organised)
A shorter journey to the resplendently beautiful Butchart Gardens and, if you like, followed by (ahem) a few minutes at the nearby Church and State Winery, in the Saanich Peninsula. About an hour there by public bus from UVic, or 30 minutes by car.

▼ Suggested Outing 3, Saltspring Island (self-organised; a full day, car/bus + ferry combo)
Why not take a day to explore and celebrate the funky, laid back, Canadian gulf island lifestyle on Saltspring Island. Ferry departs regularly from the Schwartz Bay ferry terminal, which is about one hour by bus / 30 minutes by car from UVic. You may decide to stay on forever ....

▼ Suggested Outing 4, Paddling Victoria's Inner Harbour (self-organised)
A shorter time, seeing Victoria's beautiful city centre from the waterways that initially inspired its foundation. A great choice if the day is sunny and warm. Canoes, kayaks, and paddle boards are readily rented from Ocean River Adventures and conveniently launched from right behind the store. Very chill.

And more!
Self-organised High Tea at the Empress Hotel, scooter rentals, visit to the Royal BC Museum, darts at Christies Carriage House, a hangry breakfast at a local diner, whale watching, kayaking, brew pub sampling (at Spinnaker's, Swans, Moon Under Water, and beyond!), paddle-boarding, a tour of used bookstores, and more have also been suggested!

Sunday, 3 June 2018 [DHSI Registration + Suggested Outings]

9:00 to 4:00
▼ Early Class Meeting: 4. [Foundations] DH For Department Chairs and Deans (Hickman 120, Classroom)
Further details are available from instructors in mid May to those registered in the class. Registration materials will be available in the classroom.

3:00 to 5:00
DHSI Registration (MacLaurin Building, Room A100)
After registration, many will wander to Cadboro Bay and the pub at Smuggler's Cove OR the other direction to Shelbourne Plaza and Maude Hunter's Pub OR even into the city for a nice meal.

Monday, 4 June 2018
Your hosts for the week are Alyssa Arbuckle, Ray Siemens, and Dan Sondheim.

7:45 to 8:15
Last-minute Registration (MacLaurin Building, Room A100)

8:30 to 10:00
Welcome, Orientation, and Instructor Overview (MacLaurin A144)
10:15 to Noon

Lunch break / Unconference Coordination Session (MacLaurin A144) (Grab a sandwich and come on down!)
Undergraduate Meet-up, Brown-Bag (details via email)

Classes in Session

12:15 to 1:15

Institute Panel: Perspectives on DH (or, #myDHis ...)
Chair: Alyssa Arbuckle (U Victoria) (MacLaurin A144)

- Milena Radzikowska (Mt Royal C): "Release the Kraken: Story-Driven Prototyping for the Digital Humanities."
  Abstract: I have spent the last 15 years of my career designing text analysis tools for use by humanities scholars. In this brief presentation, I propose to share a concept-based approach to interface design for DH.

- Emily Murphy (U Victoria): "#MyDHis Edgy."
  Abstract: I will build upon—or, possibly, perform a misprision of—a tweet by Polina Vinogradova; "#myDHis messy, dusty, edgy, and radically inclusive!" Vinogradova evokes the mess and dust of the archives, the edges that connect nodes of a network, and the political impetus to think of cultural history and community together. I argue that these aspects of DH have a renewed importance as we head into a moment of feminist historiography.

- Margaret Konkol (Old Dominion U): "Prototyping Mina Loy’s Alphabet with a 3D Printer."
  Abstract: This talk discusses the interpretive and methodological implications of using 3D printing technologies to prototype the archival diagrams of a proposed but never constructed plastic segmental alphabet letter kit—a game designed by modernist poet Mina Loy for F.A.O Schwarz. Although intended as a toy for young children, "The Alphabet that Builds Itself," as a work of "object typography" articulates a theory of language as kinetic, geometric, recombinant, and open to mutation. Alphabetic segments extend into the x, y, and z coordinates in exponential iterations and conjoin with magnets. Combining elements of contemporaneous typefaces like Futura and Gill Sans, which represented modernity’s functional ideals and democratic principles of simplicity, these recombinant letters represent, as this talk argues, Loy’s unpublished modernist poem, an articulation of Loy’s concept of language as a physical fact in which substance, not just form, is semantic.

4:10 to 5:00

Lee Zickel (Case Western Reserve U): "Comfortably Trepid."
Abstract: #myDHs found outside the well-established, DH-friendly institutions, at an institution that is devoted predominantly to Medicine and Engineering. I, and with increasing frequency other DH practitioners and instructors, am not positioned in a DH Lab or Humanities Center, but in ITS. Part teacher, part technologist, part translator, I will briefly discuss my work supporting humanists and social scientists, particularly those who are new to or less comfortable with computational methodologies.

Dorothy Kim (Vassar C): "#MyDHis Antifascist."
Abstract: I've spent a lot of time in the last 12 months thinking about fascism, digital humanities, its long histories, and what it means to do DH work that centers social justice particularly in this global rise of late fascism. I will speak briefly about DH's history, including the medieval history related to Busa but how that history really connects to data systems that created the Holocaust and also participated in the Cold War nuclear military complex.
Randa El Khatib (U Victoria): "Learning from the Iterative Process."
Abstract: #MyDHis Iterative. In addition to the improvements that come with iterative projects, the iterative process itself is a fruitful area for scholarly inquiry. Within this iterative context, the various teams that I work with and I have been reflecting on and rethinking central DH practices, such as what it means to collaborate, prototype, remix, and implement DH values in our work. In this talk, I will present the various lessons learnt along the way.

Sarah Melton (Boston C): "#MyDHis...People."
Abstract: Taking seriously Miriam Posner’s exhortation to “commit to DH people, not DH projects,” I invite us to reflect on how people are the core of DH. In this brief talk, I will explore the intersections between DH, labor, and infrastructure.

5:00 to 6:00 Opening Reception (University Club)
We are grateful to Gale Cengage for its sponsorship.

Tuesday, 5 June 2018

9:00 to Noon
Classes in Session

12:15 to 1:15
Lunch break / Unconference
"Mystery" Lunches
vable DHSI Lunchtime Workshop Session (click for workshop details and free registration for DHSI participants)
- 73. Introduction to ORCID (Digital Scholarship Commons, Classroom).

1:30 to 4:00
Classes in Session

DHSI Colloquium Lightning Talk Session 1 (MacLaurin A144)
Chair: James O’Sullivan
- New Modes of DH and Archival Skills Acquisition in a Graduate Public History Course. Paulina Rousseau (Ryerson U)
- Walking a Transect: Exploring a Soundscape. John Barber (Washington State U)
- Centering the Edge Case: Designing Services for Humanities Data Research. Grace Afsari-Mamagani (New York U)
- Orwellian Vocabulary and the 21st-Century Politics. Ilgin Kizilgunesler (U Manitoba)
- Making Open Data from a Gray Archive. Sara Palmer (Emory U)

6:00 to 8:00
DHSI Newcomer's Beer-B-Q (Felicitas, Student Union Building)

Wednesday, 6 June 2018

9:00 to Noon
Classes in Session

Lunch break / Unconference
"Mystery" Lunches
Brown Bag Lecture: Alexandra Branzan Albu (U Victoria): "Visual Recognition of Symbolic and Natural Patterns" (Digital Scholarship Commons, 3rd Floor McPherson Library)

Abstract: Image-based object recognition is a visual pattern recognition problem; one may characterize visual patterns as either symbolic or natural. Symbolic patterns evolved for human communication; they include but are not limited to text, forms, tables, graphics, engineering drawings etc. Symbolic patterns vary widely in terms of size, style, language, alphabet and fonts; however, literate humans can easily compensate for this variability and instantly recognize most symbolic patterns. On the other hand, natural patterns characterize images of physical structures; they often lack the intrinsic discriminability and structure of symbolic patterns, and vary widely in terms of pose, perspective, and lighting.

This lecture will explore similarities and differences in approaches designed for recognizing visual and symbolic patterns, and will address the following questions via examples.
- What are the distinctive characteristics of natural patterns? What dimensions of variability can we infer?
- What are the distinctive characteristics of symbolic patterns? What dimensions of variability can we infer?

Alexandra Branzan Albu is an Associate Professor with the Department of Electrical and Computer Engineering and cross-listed with Computer Science. Her research interests are related to image analysis, computer vision, and visual computing. She is actively pursuing outreach activities dedicated to increasing the women's presence in electrical engineering and computer science.

1:30 to 4:00
Classes in Session
Thursday, 7 June 2018

9:00 to Noon

Classes in Session

12:15 to 1:15

UVIC Library/ETCL lunchtime talk: “A Humanities Application of 3D printing and Machine Translation in the ChessBard and Loss Sets” by Dr. Aaron Tucker

Digital Scholarship Commons, 3rd floor, Mearns Centre for Learning / McPherson Library

Bring your lunch and come on up!.

[Instructor lunch meeting]

1:30 to 4:00

Classes in Session

4:15 to 5:15

DHSI Colloquium Lightning Talk Session 3 (MacLaurin A144)

Chair: James O'Sullivan

• Documenting Deportation: A Collaborative Digital Collection. Paulina Rousseau (Ryerson U)
• Unleashing the Power of Texts as Networks: Visualizing the Scholastic Commentaries and Texts Archive. Jeffrey Witt (Loyola U Maryland) and Drew Winget (Stanford U)
• #haunteDH: Punching holes in the International Busa Machine Narrative. Arun Jacob (McMaster U)
• Text in World: Computational Analysis of Trauma in Genocide Narratives. Nanditha Narayananamoorthy (U York) and Krish Perumal (U Toronto)

7:30 to 9:30

(Groovy?) Movie Night (MacLaurin A144)

Friday, 8 June 2018 [DHSI; DLFxDHSI Opening]

9:00 to Noon

DHSI Classes in Session

12:15 to 1:15

DHSI Lunch Reception / Course E-Exhibits (MacLaurin A100)

1:00 to 2:00

DLFxDHSI Registration (MacLaurin A100)

1:30 to 1:50

[DHSI] Remarks, A Week in Review (MacLaurin A144)

2:00 to 3:00

Joint Institute Lecture (DHSI and DLFxDHSI):
Bethany Nowviskie (CLIR DLF and U Virginia): "Reconstitute the World: Machine-reading Archives of Mass Extinction"

Chair: Lisa Goddard (U Victoria)

(MacLaurin A144)

Abstract: The basic constitution of our digital collections becomes vastly more important in the face of two understandings: first, that archives of modernity are archives of the sixth great mass extinction of life on our planet; and next, that we no longer steward cultural heritage for human readers alone. In the same way that we people are shaped by what we read, hear, and see, the machine readers that follow us into and perhaps beyond the Anthropocene have begun to learn from "unsupervised" encounters with our digital libraries. What will we preserve for the living generations and artificial intelligences that will come? What do we neglect, or even choose to extinguish? And from an elegiac archive, a library of endings, can we create forward-looking, speculative collections—collections from which to deep-dream new futures? The most extra/ordinary power we possess is the power to make poetry from records of the past. Could it be called on, one day, to reconstitute the world?
Joint Reception: DHSI and DLFxDHSI (University Club)

DLFxDHSI Poster/Demo Session

• DHSI Colloquium Poster/Demo Session
  • Media as a Colonial Artifact in Menzies’ Journal. Paula Johanson (U Victoria)
  • Camp Edit: the Institute for the Editing of Historical Documents. Nikolaus Wasmoen (Association for Documentary Editing, U Buffalo), Jennifer Stiertzer (Association for Documentary Editing, U Virginia), and Cathy Moran Hajo (Association for Documentary Editing, Ramapo C)
  • A Digital Archaeology of Life in Cleveland’s Depression-Era Slums. Charlie Harper (Case Western Reserve U) and Jared Bendis (Case Western Reserve U)
  • Feminist Pest Control: controlling and not controlling nonhuman pests. Lindsay Garcia (C of William and Mary)
  • Legends of the Buddhist Saints. Jonathan S. Walters (Whitman C) and Dana Johnson (Freelance Web Developer)
  • Accessibility in Digital Environments Via TEI-Encoded Uncontracted Braille. Gia Alexander (Texas A&M U)
  • Translation3point0: Why Literary Translation Data Matters. Katie King (U Washington)
  • PoéticaSonora: A Digital Audio Repository Prototype for Latin American Sound Art and Poetry. Aurelio Meza (Concordia U)
  • Beauty and the Book: Pre-Raphaelite Artistic Practice Contained. Josie Greenhill (U Victoria)
  • Poetic Procedures/Digital Deformances. Corey Sparks (California State U, Chico)
  • Miranda, the Folger Shakespeare Library’s new Digital Asset Platform. Meaghan Brown (Folger Shakespeare Library)
  • Living Song Project. Quinn Patrick Ankrum (U Cincinnati) and Elizabeth Avery (U Oklahoma)
  • Digital Frankenstein Variorum. Rikk Mulligan (Carnegie Mellon U)

8:30 to 9:00 DLFxDHSI Registration (MacLaurin A100)

9:00 to 5:30 DLFxDHSI UnConference Sessions

• DHSI All Day Workshop Session (click for workshop details and free registration for DHSI participants)

9:00 to 4:00

• 53. Building Your Academic Digital Identity (MacLaurin D105, Classroom)

Saturday, 9 June 2018 [DLFxDHSI + DHSI Conference and Colloquium]

9:00 to 5:00

DHSI Colloquium Day Conference (MacLaurin A144)

Welcome

People I: Documenting Online Lives. Chair: Molly Nebiolo (University of New York)

  - Examining Gendered Harassment Online and in Silicon Valley. Andrea Flores (Utica College)
  - This is Just to Say I Have <X> the <Y> in your <Z>: Modernist Memes in an Era of Public Apology. Shawna Ross (Texas A&M University)

Break

People II: Documenting Lives Online. Chair: Dheepa Sundaram (College of Wooster)

  - Youtube Yoga and Ritual on Demand: The Virtual Economics of Hindu Soteriology. Dheepa Sundaram (College of Wooster)
  - The Resemblage Project: Creativity and Digital Health Humanities in Canada. Andrea Charise (University of Toronto) and Stefan Krecsy (University of Toronto)

Lunch

Projects I: Building and Analyzing. Chair: Yannis Rammos (New York University)

  - Building the ARTECHNE Database: New directions in Digital Art History. Marieke Hendriksen (Old Dominion University)
  - The Ineffective Inquisition: The Holy Office’s Sphere of Influence in Early Modern New Spain. Kira Homo (Pennsylvania State University)

Break

Projects II: Mapping and Visualizing. Chair: Innocent Opara (Qumet Institute)

  - Mapping Sarah Sophia Bank’s Numismatic Collection. Erica Hayes (North Carolina State University) and Kacie Wills (University of California, Riverside)
  - Text Mining and Visualizing 18th Century American Correspondence. Ashley Sanders Garcia (University of California, Los Angeles)

Break

Practices: Digital Scholarship on Campus and in the Classroom. Chair: Alyssa Arhuckle (University of Victoria)
Concluding Remarks

Sunday, 10 June 2018 [SINM + DHSI Registration, Workshops]

8:30 to 9:00
Symposium on Indigenous New Media Registration (MacLaurin A100)

9:00 to 5:00
DHSI Registration (MacLaurin A100)

9:00 to 4:00
▼ SINM Sessions
- 63. Symposium on Indigenous New Media: Reading Group (Hickman 105, Classroom)
- 72. Symposium on Indigenous New Media: Indigitization (Hickman 120, Classroom)
  Full details here

9:00 to 4:00
▼ DHSI All Day Workshop Sessions (click for workshop details and free registration for DHSI participants)
- 53. Building Your Academic Digital Identity (MacLaurin D105, Classroom)
- 54. An Introduction to the Archaeology of 1980s Computing (MacLaurin D114, Classroom)

9:00 to Noon
▼ DHSI AM Workshop Sessions (click for workshop details and free registration for DHSI participants)
- 55. Regular Expressions (MacLaurin D111, Classroom)
- 56. 3D Visualization for the Humanities (MacLaurin D010, Classroom)
- 58. DH Fieldwork Methods (MacLaurin D016, Classroom)
- 60. Pedagogy of the Digitally Oppressed: Inculcating De-/Anti-/Post-Colonial Digital Humanities (MacLaurin D107, Classroom)
- 61. Introduction to #GraphPoem. Digital Tools for Poetry Computational Analysis and Graph Theory Apps in Poetry (MacLaurin D101, Classroom)
- 62. Creating a CV for Digital Humanities Makers (MacLaurin D115, Classroom)

1:00 to 4:00
▼ DHSI PM Workshop Sessions (click for workshop details and free registration for DHSI participants)
- 64. Agent-Based Modelling in the Humanities (MacLaurin D111, Classroom)
- 65. Unleash Linux on MacOS (MacLaurin D010, Classroom)
- 66. DHSI Knits: History of Textiles and Technology (MacLaurin D016, Classroom)
- 67. Crowdsourcing as a Tool for Research and Public Engagement (MacLaurin D109, Classroom)
- 69. Web Annotation as Critical Humanities Practice (MacLaurin D103, Classroom)
- 70. Dynamic Ontologies for the Humanities (MacLaurin D107, Classroom)
- 71. Social Media Research in the Humanities (MacLaurin D101, Classroom)

4:10 to 5:00
▼ Joint Institute Lecture (DHSI and SINM):
  David Gaertner (U British Columbia): "A Landless Territory?: CyberPowWow and the Politics of Indigenous New Media."
  Chair: Deanna Reder (Simon Fraser U)
  (MacLaurin A144)
  Abstract: Following the 1997 launch of Skawennati’s (Mohawk) CyberPowWow, digital space has become a vital new territory for the resurgence of Indigenous storytelling and cultural practice: "We have signed a new treaty," Cree artist Archer Pechawis wrote of this period, "and it is good. We have the right to hunt, fish, dance and make art at www.CyberPowWow.net, .org and .com for as long as the grass grows and the rivers flow." This talk will critically explore the theoretical, cultural, political-economic, and gendered dynamics underwriting the histories and futures of Indigenous new media. Particular attention will be given in examining the ways in which new media and digital storytelling connect to and support key issues in the field of Indigenous studies, such as sovereignty, self-determination, decolonization, and land rights.

After the day, many will wander to Cadboro Bay and the pub at Smuggler's Cove OR the other direction to Shelbourne Plaza and Maude Hunter's Pub OR even into the city for a bite to eat.
Your hosts for the week are Ray Siemens and Dan Sondheim.

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>7:45 to 8:15</td>
<td>DHSI Last-minute Registration (MacLaurin A100)</td>
</tr>
<tr>
<td>8:30 to 10:00</td>
<td>DHSI Welcome, Orientation, and Instructor Overview (MacLaurin A144)</td>
</tr>
<tr>
<td>9:00 to 4:00</td>
<td>SINM Sessions</td>
</tr>
</tbody>
</table>

- DHSI Classes in Session (click for details and locations)
  - 29. [Foundations] Models for DH at Liberal Arts Colleges (& 4 yr Institutions) (MacLaurin D109, Classroom)
  - 32. Stylometry with R: Computer-Assisted Analysis of Literary Texts (Clearihue A102, Lab)
  - 33. Digital Storytelling (MacLaurin D111, Classroom)
  - 34. Text Mapping as Modelling (Clearihue D131, Classroom)
  - 35. Geographical Information Systems in the Digital Humanities (Clearihue A105, Lab)
  - 36. Open Access and Open Social Scholarship (MacLaurin D114, Classroom)
  - 37. Introduction to Machine Learning in the Digital Humanities (Cornett A229, Classroom)
  - 38. Queer Digital Humanities: Intersections, Interrogations, Iterations (MacLaurin D110, Classroom)
  - 41. Using Fedora Commons / Islandora (Human and Social Development A160, Lab)
  - 42. Documenting Born Digital Creative and Scholarly Works for Access and Preservation (MacLaurin D115, Classroom)
  - 43. Games for Digital Humanists (MacLaurin D016, Classroom & Human and Social Development A170, Lab)
  - 44. XPath for Document Archeology and Project Management (Cornett A128, Classroom)
  - 46. Surveillance and the Digital Humanities (MacLaurin D103, Classroom)
  - 47. Text Analysis with Python and the Natural Language ToolKit (Clearihue A103, Lab)
  - 48. Information Security for Digital Researchers (Clearihue D130, Classroom)
  - 49. Wrangling Big Data for DH (Human and Social Development A150, Lab)
  - 50. Accessibility & Digital Environments (MacLaurin D101, Classroom)
  - 51. Critical Pedagogy and Digital Praxis in the Humanities (MacLaurin D105, Classroom)
  - 52. Drupal for Digital Humanities Projects (MacLaurin D107, Classroom)

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:15 to Noon</td>
<td>DHSI Classes in Session</td>
</tr>
</tbody>
</table>

- Joint Institute Lecture (DHSI and SINM):
  - Jordan Abel (Simon Fraser U): “Indigeneity, Conceptualism, and the Borders of DH.”
  - Chair: Michelle Brown (U Hawaii) (MacLaurin A144)

Abstract: This talk brings together digital humanities discourses in computational textual analysis and Indigenous Literary Studies to analyze a corpus comprised of every book of Indigenous poetry published in Canada, extending from Pauline Johnson's 1895 book The White Wampum to Marilyn Dumont's 2015 book The Pemmican Eaters. While the main goal of this research project initially centered on the topic modeling of a corpus of Indigenous poetry, the project also addresses the systemic barriers that have prevented such work gaining traction, and likewise attempts to address the specific challenges that Indigenous writing (and in particular Indigenous poetry) present to current Digital Humanities methodologies.

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>5:00 to 6:00</td>
<td>Joint Reception: DHSI and SINM (University Club)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 to Noon</td>
<td>DHSI Classes in Session</td>
</tr>
</tbody>
</table>

Tuesday, 12 June 2018

- Joint Institute Lecture (DHSI and SINM):
  - Jordan Abel (Simon Fraser U): “Indigeneity, Conceptualism, and the Borders of DH.”
  - Chair: Michelle Brown (U Hawaii) (MacLaurin A144)

Abstract: This talk brings together digital humanities discourses in computational textual analysis and Indigenous Literary Studies to analyze a corpus comprised of every book of Indigenous poetry published in Canada, extending from Pauline Johnson's 1895 book The White Wampum to Marilyn Dumont's 2015 book The Pemmican Eaters. While the main goal of this research project initially centered on the topic modeling of a corpus of Indigenous poetry, the project also addresses the systemic barriers that have prevented such work gaining traction, and likewise attempts to address the specific challenges that Indigenous writing (and in particular Indigenous poetry) present to current Digital Humanities methodologies.

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>12:15 to 1:15</td>
<td>Lunch break / Unconference Coordination Session (MacLaurin A144)</td>
</tr>
</tbody>
</table>

DHSI Undergraduate Meet-up, Brown-Bag (details via email)

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:30 to 4:00</td>
<td>DHSI Classes in Session</td>
</tr>
</tbody>
</table>

- Joint Institute Lecture (DHSI and SINM):
  - Jordan Abel (Simon Fraser U): “Indigeneity, Conceptualism, and the Borders of DH.”
  - Chair: Michelle Brown (U Hawaii) (MacLaurin A144)

Abstract: This talk brings together digital humanities discourses in computational textual analysis and Indigenous Literary Studies to analyze a corpus comprised of every book of Indigenous poetry published in Canada, extending from Pauline Johnson's 1895 book The White Wampum to Marilyn Dumont's 2015 book The Pemmican Eaters. While the main goal of this research project initially centered on the topic modeling of a corpus of Indigenous poetry, the project also addresses the systemic barriers that have prevented such work gaining traction, and likewise attempts to address the specific challenges that Indigenous writing (and in particular Indigenous poetry) present to current Digital Humanities methodologies.

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
</table>
| 4:10 to 5:00 | Joint Institute Lecture (DHSI and SINM):
  - Jordan Abel (Simon Fraser U): “Indigeneity, Conceptualism, and the Borders of DH.”
  - Chair: Michelle Brown (U Hawaii) (MacLaurin A144)

Abstract: This talk brings together digital humanities discourses in computational textual analysis and Indigenous Literary Studies to analyze a corpus comprised of every book of Indigenous poetry published in Canada, extending from Pauline Johnson's 1895 book The White Wampum to Marilyn Dumont's 2015 book The Pemmican Eaters. While the main goal of this research project initially centered on the topic modeling of a corpus of Indigenous poetry, the project also addresses the systemic barriers that have prevented such work gaining traction, and likewise attempts to address the specific challenges that Indigenous writing (and in particular Indigenous poetry) present to current Digital Humanities methodologies.

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>5:00 to 6:00</td>
<td>Joint Reception: DHSI and SINM (University Club)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 to Noon</td>
<td>Classes in Session</td>
</tr>
</tbody>
</table>

- Joint Institute Lecture (DHSI and SINM):
  - Jordan Abel (Simon Fraser U): “Indigeneity, Conceptualism, and the Borders of DH.”
  - Chair: Michelle Brown (U Hawaii) (MacLaurin A144)

Abstract: This talk brings together digital humanities discourses in computational textual analysis and Indigenous Literary Studies to analyze a corpus comprised of every book of Indigenous poetry published in Canada, extending from Pauline Johnson's 1895 book The White Wampum to Marilyn Dumont's 2015 book The Pemmican Eaters. While the main goal of this research project initially centered on the topic modeling of a corpus of Indigenous poetry, the project also addresses the systemic barriers that have prevented such work gaining traction, and likewise attempts to address the specific challenges that Indigenous writing (and in particular Indigenous poetry) present to current Digital Humanities methodologies.

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>12:15 to 1:15</td>
<td>Lunch break / Unconference Coordination Session (MacLaurin A144)</td>
</tr>
</tbody>
</table>

DHSI Undergraduate Meet-up, Brown-Bag (details via email)
Wednesday, 13 June 2018

9:00 to Noon Classes in Session

12:15 to 1:15 Lunch break / Unconference
"Mystery" Lunches

1:30 to 4:00 Classes in Session

4:15 to 5:15 DHSI Colloquium Lightning Talk Session 4 (MacLaurin A144)
Chair: Lindsey Seatter

- Mapping Indigenous and Chicana/o Environmental Imaginaries using GIS. Stevie Ruiz (California State U, Northridge), Quetzalli Enrique (California State U, Northridge), Enrique Ramirez (California State U, Northridge), and Tomas Figueroa (California State U, Northridge)
- "Doing DH with Graphic Narratives. John Barber (Washington State U)
- "But is it any good?: A quantitative approach to the popularity of digital fanfiction. Suzanne Black (U Edinburgh)
- The American Prison Writing Archive (APWA). Doran Larson (Hamilton C), Janet Simons (Digital Humanities Initiative, Hamilton C), and William Rasenberger (Hamilton C)

6:00 to 8:00 DHSI Newcomer's Beer-B-Q (Felicitas, Student Union Building)

Thursday, 14 June 2018

9:00 to Noon Classes in Session

12:15 to 1:15 Lunch break / Unconference
"Mystery" Lunches

1:30 to 4:00 Classes in Session

4:15 to 5:15 DHSI Colloquium Lightning Talk Session 5 (MacLaurin A144)
Chair: Lindsey Seatter

- Faraway, so close: Has the political environment really changed in Ecuador?. Luis Meneses (Electronic Textual Cultures Lab, U Victoria)
- Re-mixing Melville's Reading: Text Analysis of Marginalia with R and XSLT. Christopher Ohge (U London, School of Advanced Study) and Steven Olsen-Smith (Boise State U)
- Developing Interactive and Open-Source OER: Inquiry-Based Music Theory. Evan Williamson (U Idaho)
- Spatial Humanities and the Web of Everywhere. Ken Cooper (SUNY Geneseo)

6:00 to 7:00 "Half Way There (yet again)!" [An Informal, Self-Organized Birds of a Feather Get-Together] (Felicitas, Student Union Building)
Bring your DHSI nametag and enjoy your first tipple on us!

Friday, 15 June 2018

9:00 to Noon Classes in Session

12:15 to 1:15 Lunch Reception / Course E-Exhibits (MacLaurin A100)
<table>
<thead>
<tr>
<th>Time</th>
<th>Event Description</th>
</tr>
</thead>
</table>
| 1:30 to 2:30 | Institute Lecture: William Bowen (U Toronto Scarborough): “Discovery, Collaboration and Dissemination: Lessons Learned and Plans for the Future” (MacLaurin A144)  
Abstract: Much has changed and continues to change in digital humanities since the formal establishment of Iter in the Fall of 1997. However, the mandate of the not-for-profit partnership to support “the advancement of learning in the study and teaching of Middle Ages and Renaissance (400–1700) through the development and distribution of online resources” continues to have relevance. This presentation explores the striking challenges faced by Iter and presents our current thinking on the realization of this mandate for the future through a platform with a focus on facilitating the discovery of the academic resources necessary to our work; creating an environment for collaboration, sharing and developing projects; and on enabling the distribution and publication of our scholarship. |
| 2:40 to 3:00 | Awards and Bursaries Recognition  
Closing, DHSI in Review (MacLaurin A144) |

**Contact info:**

institut@uvic.ca  
P: 250-472-5401  
F: 250-472-5681
Introduction to Machine Learning in Digital Humanities
Digital Humanities Summer Institute
University of Victoria
June 11-15, 2018

Instructors: Paul Barrett (paul.barrett@utoronto.ca) and Nathan Taback (nathan.taback@utoronto.ca)

The learning objective of this course is to become familiar with machine learning (ML) techniques used in the digital humanities (DH).

By the end of the course students will be able to:

- Describe how ML algorithms are used in the digital humanities.
- Describe and use standard tools in supervised ML including linear and logistic regression and trees.
- Critique specific applications in DH that make use of supervised ML techniques.
- Describe and use standard tools in unsupervised ML including cluster analysis.
- Critique specific applications in DH that make use of unsupervised ML techniques.
- Describe and use basic topic models to discover hidden topics/themes in documents.

Software for Machine Learning

We will be using R, an open source statistical language, to implement the ML algorithms that will be discussed in class. We are using the cloud version of R. Before beginning the class, please register an R Studio Cloud account at the following link:

https://rstudio.cloud
Daily Schedule

Morning – 9:00 -12:00 (except 10:15 on day 1), afternoon – 1:30 - 4:00.

Students should bring a laptop to class with R installed.

Afternoon case studies will involve small group work on a case study and class presentations.

Day 1 (Morning)

- Introduction to digital humanities – Paul come up with a few key highlights from some of the debates.
- Introduction to machine learning
- Introduction to GitHub
- Introduction to R using R Studio (students should install R and R Studio before the first class)

Day 1 (Afternoon)

Bob Dylan: “Gotta Serve Somebody” – Data Visualization
- Raise questions to the class
- N-grams and tf-idf
- Get them to work on a new text. Create a new notebook

Day 2 (Morning)

- Sentiment analysis on twitter using tidytext
- Interact with class to show the change from unlimited to unlimited & get them to score the tweets themselves
- #DHSI2017 – is the conference going well?
- Different lexicons for sentiment analysis

- Introduction to Linear Regression, Logistic Regression, Prediction Accuracy (same as Linear but Binary instead of Quantitative),
- Predicting email SPAM using regression
- Classification trees (is Cat? Yes / No)

- tf-idf / n grams

Day 2 (Afternoon)
Supervised ML using R: logistic regression
Students do Shakespeare and Jonson supervised learning example
Plan B if Shakespeare doesn’t work: Email version with modified code

Day 3 (Morning)
- Introduction to unsupervised machine learning techniques used in DH.
- Clustering: HG Wells
- Topic Modeling

Day 3 (Afternoon)
- Example of topic modeling – what to use? CanLit?
- Probably something from Gutenberg. Get them to download, topic model, say something compelling about the material.
- Take out French language stuff?
- Editorial decisions – how do they affect this?

Day 4 (Morning)
- Intro: Did Wells influence the Brontë sisters?
- Student teams work on data set with questions that we provide.
- A few options: Shakespeare, CanLit, Reuters, Jane Austen, Twitter: Choose a hash tag and figure out what the topics are.
- OR we carry over what we haven’t gotten into

Day 4 (Afternoon)
More work time

Day 5 (Morning)
- Putting it all together.
Artificial Intelligence — The Revolution Hasn’t Happened Yet

Artificial Intelligence (AI) is the mantra of the current era. The phrase is intoned by technologists, academicians, journalists and venture capitalists alike. As with many phrases that cross over from technical academic fields into general circulation, there is significant misunderstanding accompanying the use of the phrase. But this is not the classical case of the public not understanding the scientists—here the scientists are often as befuddled as the public. The idea that our era is somehow seeing the emergence of an intelligence in silicon that rivals our own entertains all of us—enthralling us and frightening us in equal measure. And, unfortunately, it distracts us.

There is a different narrative that one can tell about the current era. Consider the following story, which involves humans, computers, data and life-or-death decisions, but where the focus is something other than
intelligence-in-silicon fantasies. When my spouse was pregnant 14 years ago, we had an ultrasound. There was a geneticist in the room, and she pointed out some white spots around the heart of the fetus. “Those are markers for Down syndrome,” she noted, “and your risk has now gone up to 1 in 20.” She further let us know that we could learn whether the fetus in fact had the genetic modification underlying Down syndrome via an amniocentesis. But amniocentesis was risky—the risk of killing the fetus during the procedure was roughly 1 in 300. Being a statistician, I determined to find out where these numbers were coming from. To cut a long story short, I discovered that a statistical analysis had been done a decade previously in the UK, where these white spots, which reflect calcium buildup, were indeed established as a predictor of Down syndrome. But I also noticed that the imaging machine used in our test had a few hundred more pixels per square inch than the machine used in the UK study. I went back to tell the geneticist that I believed that the white spots were likely false positives—that they were literally “white noise.” She said “Ah, that explains why we started seeing an uptick in Down syndrome diagnoses a few years ago; it’s when the new machine arrived.”

We didn’t do the amniocentesis, and a healthy girl was born a few months later. But the episode troubled me, particularly after a back-of-the-envelope calculation convinced me that many thousands of people had gotten that diagnosis that same day worldwide, that many of them had opted for amniocentesis, and that a number of babies had died needlessly. And this happened day after day until it somehow got fixed. The problem that this episode revealed wasn’t about my individual medical care; it was about a medical system that measured variables and outcomes in various places and times, conducted statistical analyses, and made use of the results in other places and times. The problem had to do not just with data analysis per se, but with what database researchers call “provenance”—broadly, where did data arise, what inferences were drawn from the data, and how relevant are those inferences to the present situation? While a trained human might be able to work all of this out on a case-by-case basis, the issue was that of designing a planetary-scale medical system that could do this without the need for such detailed human oversight.
I'm also a computer scientist, and it occurred to me that the principles needed to build planetary-scale inference-and-decision-making systems of this kind, blending computer science with statistics, and taking into account human utilities, were nowhere to be found in my education. And it occurred to me that the development of such principles—which will be needed not only in the medical domain but also in domains such as commerce, transportation and education—were at least as important as those of building AI systems that can dazzle us with their game-playing or sensorimotor skills.

Whether or not we come to understand “intelligence” any time soon, we do have a major challenge on our hands in bringing together computers and humans in ways that enhance human life. While this challenge is viewed by some as subservient to the creation of “artificial intelligence,” it can also be viewed more prosaically—but with no less reverence—as the creation of a new branch of engineering. Much like civil engineering and chemical engineering in decades past, this new discipline aims to corral the power of a few key ideas, bringing new resources and capabilities to people, and doing so safely. Whereas civil engineering and chemical engineering were built on physics and chemistry, this new engineering discipline will be built on ideas that the preceding century gave substance to—ideas such as “information,” “algorithm,” “data,” “uncertainty,” “computing,” “inference,” and “optimization.” Moreover, since much of the focus of the new discipline will be on data from and about humans, its development will require perspectives from the social sciences and humanities.

While the building blocks have begun to emerge, the principles for putting these blocks together have not yet emerged, and so the blocks are currently being put together in ad-hoc ways.

*Thus, just as humans built buildings and bridges before there was civil engineering, humans are proceeding with the building of societal-scale, inference-and-decision-making systems that involve machines, humans and the environment. Just as early buildings and bridges sometimes fell to the ground—in unforeseen ways and with tragic consequences—many of our early societal-scale inference-and-decision-making systems are already exposing serious conceptual flaws.*
And, unfortunately, we are not very good at anticipating what the next emerging serious flaw will be. What we’re missing is an engineering discipline with its principles of analysis and design.

The current public dialog about these issues too often uses “AI” as an intellectual wildcard, one that makes it difficult to reason about the scope and consequences of emerging technology. Let us begin by considering more carefully what “AI” has been used to refer to, both recently and historically.

Most of what is being called “AI” today, particularly in the public sphere, is what has been called “Machine Learning” (ML) for the past several decades. ML is an algorithmic field that blends ideas from statistics, computer science and many other disciplines (see below) to design algorithms that process data, make predictions and help make decisions. In terms of impact on the real world, ML is the real thing, and not just recently. Indeed, that ML would grow into massive industrial relevance was already clear in the early 1990s, and by the turn of the century forward-looking companies such as Amazon were already using ML throughout their business, solving mission-critical back-end problems in fraud detection and logistics-chain prediction, and building innovative consumer-facing services such as recommendation systems. As datasets and computing resources grew rapidly over the ensuing two decades, it became clear that ML would soon power not only Amazon but essentially any company in which decisions could be tied to large-scale data. New business models would emerge. The phrase “Data Science” began to be used to refer to this phenomenon, reflecting the need of ML algorithms experts to partner with database and distributed-systems experts to build scalable, robust ML systems, and reflecting the larger social and environmental scope of the resulting systems.

This confluence of ideas and technology trends has been rebranded as “AI” over the past few years. This rebranding is worthy of some scrutiny.

Historically, the phrase “AI” was coined in the late 1950’s to refer to the heady aspiration of realizing in software and hardware an entity possessing human-level intelligence. We will use the phrase “human-imitative AI” to refer to this aspiration, emphasizing the notion that the artificially intelligent entity should seem to be one of us, if not physically
at least mentally (whatever that might mean). This was largely an academic enterprise. While related academic fields such as operations research, statistics, pattern recognition, information theory and control theory already existed, and were often inspired by human intelligence (and animal intelligence), these fields were arguably focused on “low-level” signals and decisions. The ability of, say, a squirrel to perceive the three-dimensional structure of the forest it lives in, and to leap among its branches, was inspirational to these fields. “AI” was meant to focus on something different—the “high-level” or “cognitive” capability of humans to “reason” and to “think.” Sixty years hence, however, high-level reasoning and thought remain elusive. The developments which are now being called “AI” arose mostly in the engineering fields associated with low-level pattern recognition and movement control, and in the field of statistics—the discipline focused on finding patterns in data and on making well-founded predictions, tests of hypotheses and decisions.

Indeed, the famous “backpropagation” algorithm that was rediscovered by David Rumelhart in the early 1980s, and which is now viewed as being at the core of the so-called “AI revolution,” first arose in the field of control theory in the 1950s and 1960s. One of its early applications was to optimize the thrusts of the Apollo spaceships as they headed towards the moon.

Since the 1960s much progress has been made, but it has arguably not come about from the pursuit of human-imitative AI. Rather, as in the case of the Apollo spaceships, these ideas have often been hidden behind the scenes, and have been the handiwork of researchers focused on specific engineering challenges. Although not visible to the general public, research and systems-building in areas such as document retrieval, text classification, fraud detection, recommendation systems, personalized search, social network analysis, planning, diagnostics and A/B testing have been a major success—these are the advances that have powered companies such as Google, Netflix, Facebook and Amazon.

One could simply agree to refer to all of this as “AI,” and indeed that is what appears to have happened. Such labeling may come as a surprise to optimization or statistics researchers, who wake up to find themselves...
suddenly referred to as “AI researchers.” But labeling of researchers aside, the bigger problem is that the use of this single, ill-defined acronym prevents a clear understanding of the range of intellectual and commercial issues at play.

The past two decades have seen major progress—in industry and academia—in a complementary aspiration to human-imitative AI that is often referred to as “Intelligence Augmentation” (IA). Here computation and data are used to create services that augment human intelligence and creativity. A search engine can be viewed as an example of IA (it augments human memory and factual knowledge), as can natural language translation (it augments the ability of a human to communicate). Computing-based generation of sounds and images serves as a palette and creativity enhancer for artists. While services of this kind could conceivably involve high-level reasoning and thought, currently they don’t—they mostly perform various kinds of string-matching and numerical operations that capture patterns that humans can make use of.

Hoping that the reader will tolerate one last acronym, let us conceive broadly of a discipline of “Intelligent Infrastructure” (II), whereby a web of computation, data and physical entities exists that makes human environments more supportive, interesting and safe. Such infrastructure is beginning to make its appearance in domains such as transportation, medicine, commerce and finance, with vast implications for individual humans and societies. This emergence sometimes arises in conversations about an “Internet of Things,” but that effort generally refers to the mere problem of getting “things” onto the Internet—not to the far grander set of challenges associated with these “things” capable of analyzing those data streams to discover facts about the world, and interacting with humans and other “things” at a far higher level of abstraction than mere bits.

For example, returning to my personal anecdote, we might imagine living our lives in a “societal-scale medical system” that sets up data flows, and data-analysis flows, between doctors and devices positioned in and around human bodies, thereby able to aid human intelligence in making diagnoses and providing care. The system would incorporate information from cells in the body, DNA, blood tests, environment,
population genetics and the vast scientific literature on drugs and treatments. It would not just focus on a single patient and a doctor, but on relationships among all humans—just as current medical testing allows experiments done on one set of humans (or animals) to be brought to bear in the care of other humans. It would help maintain notions of relevance, provenance and reliability, in the way that the current banking system focuses on such challenges in the domain of finance and payment. And, while one can foresee many problems arising such a system—including privacy issues, liability issues, security issues, etc—these problems should properly be viewed as challenges, not showstoppers.

We now come to a critical issue: Is working on classical human-imitative AI the best or only way to focus on these larger challenges? Some of the most heralded recent success stories of ML have in fact been in areas associated with human-imitative AI—areas such as computer vision, speech recognition, game-playing and robotics. So perhaps we should simply await further progress in domains such as these. There are two points to make here. First, although one would not know it from reading the newspapers, success in human-imitative AI has in fact been limited—we are very far from realizing human-imitative AI aspirations. Unfortunately the thrill (and fear) of making even limited progress on human-imitative AI gives rise to levels of over-exuberance and media attention that is not present in other areas of engineering.

Second, and more importantly, success in these domains is neither sufficient nor necessary to solve important IA and II problems. On the sufficiency side, consider self-driving cars. For such technology to be realized, a range of engineering problems will need to be solved that may have little relationship to human competencies (or human lack-of-competencies). The overall transportation system (an II system) will likely more closely resemble the current air-traffic control system than the current collection of loosely-coupled, forward-facing, inattentive human drivers. It will be vastly more complex than the current air-traffic control system, specifically in its use of massive amounts of data and adaptive statistical modeling to inform fine-grained decisions. It is those challenges that need to be in the forefront, and in such an effort a focus on human-imitative AI may be a distraction.
As for the necessity argument, it is sometimes argued that the human-imitative AI aspiration subsumes IA and II aspirations, because a human-imitative AI system would not only be able to solve the classical problems of AI (as embodied, e.g., in the Turing test), but it would also be our best bet for solving IA and II problems. Such an argument has little historical precedent. Did civil engineering develop by envisaging the creation of an artificial carpenter or bricklayer? Should chemical engineering have been framed in terms of creating an artificial chemist? Even more polemically: if our goal was to build chemical factories, should we have first created an artificial chemist who would have then worked out how to build a chemical factory?

A related argument is that human intelligence is the only kind of intelligence that we know, and that we should aim to mimic it as a first step. But humans are in fact not very good at some kinds of reasoning—we have our lapses, biases and limitations. Moreover, critically, we did not evolve to perform the kinds of large-scale decision-making that modern II systems must face, nor to cope with the kinds of uncertainty that arise in II contexts. One could argue that an AI system would not only imitate human intelligence, but also “correct” it, and would also scale to arbitrarily large problems. But we are now in the realm of science fiction—such speculative arguments, while entertaining in the setting of fiction, should not be our principal strategy going forward in the face of the critical IA and II problems that are beginning to emerge. We need to solve IA and II problems on their own merits, not as a mere corollary to an human-imitative AI agenda.

It is not hard to pinpoint algorithmic and infrastructure challenges in II systems that are not central themes in human-imitative AI research. II systems require the ability to manage distributed repositories of knowledge that are rapidly changing and are likely to be globally incoherent. Such systems must cope with cloud-edge interactions in making timely, distributed decisions and they must deal with long-tail phenomena whereby there is lots of data on some individuals and little data on most individuals. They must address the difficulties of sharing data across administrative and competitive boundaries. Finally, and of particular importance, II systems must bring economic ideas such as incentives and pricing into the realm of the statistical and computational infrastructures that link humans to each other and to valued goods. Such
II systems can be viewed as not merely providing a service, but as creating *markets*. There are domains such as music, literature and journalism that are crying out for the emergence of such markets, where data analysis links producers and consumers. And this must all be done within the context of evolving societal, ethical and legal norms.

Of course, classical human-imitative AI problems remain of great interest as well. However, the current focus on doing AI research via the gathering of data, the deployment of “deep learning” infrastructure, and the demonstration of systems that mimic certain narrowly-defined human skills—with little in the way of emerging explanatory principles—tends to deflect attention from major open problems in classical AI. These problems include the need to bring meaning and reasoning into systems that perform natural language processing, the need to infer and represent causality, the need to develop computationally-tractable representations of uncertainty and the need to develop systems that formulate and pursue long-term goals. These are classical goals in human-imitative AI, but in the current hubbub over the “AI revolution,” it is easy to forget that they are not yet solved.

IA will also remain quite essential, because for the foreseeable future, computers will not be able to match humans in their ability to reason abstractly about real-world situations. We will need well-thought-out interactions of humans and computers to solve our most pressing problems. And we will want computers to trigger new levels of human creativity, not replace human creativity (whatever that might mean).

It was John McCarthy (while a professor at Dartmouth, and soon to take a position at MIT) who coined the term “AI,” apparently to distinguish his budding research agenda from that of Norbert Wiener (then an older professor at MIT). Wiener had coined “cybernetics” to refer to his own vision of intelligent systems—a vision that was closely tied to operations research, statistics, pattern recognition, information theory and control theory. McCarthy, on the other hand, emphasized the ties to logic. In an interesting reversal, it is Wiener’s intellectual agenda that has come to dominate in the current era, under the banner of McCarthy’s terminology. (This state of affairs is surely, however, only temporary; the
pendulum swings more in AI than in most fields.)

But we need to move beyond the particular historical perspectives of McCarthy and Wiener.

*We need to realize that the current public dialog on AI—which focuses on a narrow subset of industry and a narrow subset of academia—risks blinding us to the challenges and opportunities that are presented by the full scope of AI, IA and II.*

This scope is less about the realization of science-fiction dreams or nightmares of super-human machines, and more about the need for humans to understand and shape technology as it becomes ever more present and influential in their daily lives. Moreover, in this understanding and shaping there is a need for a diverse set of voices from all walks of life, not merely a dialog among the technologically attuned. Focusing narrowly on human-imitative AI prevents an appropriately wide range of voices from being heard.

While industry will continue to drive many developments, academia will also continue to play an essential role, not only in providing some of the most innovative technical ideas, but also in bringing researchers from the computational and statistical disciplines together with researchers from other disciplines whose contributions and perspectives are sorely needed—notably the social sciences, the cognitive sciences and the humanities.

On the other hand, while the humanities and the sciences are essential as we go forward, we should also not pretend that we are talking about something other than an engineering effort of unprecedented scale and scope—society is aiming to build new kinds of artifacts. These artifacts should be built to work as claimed. We do not want to build systems that help us with medical treatments, transportation options and commercial opportunities to find out after the fact that these systems don’t really work—that they make errors that take their toll in terms of human lives and happiness. In this regard, as I have emphasized, there is an engineering discipline yet to emerge for the data-focused and learning-
focused fields. As exciting as these latter fields appear to be, they cannot yet be viewed as constituting an engineering discipline.

Moreover, we should embrace the fact that what we are witnessing is the creation of a new branch of engineering. The term “engineering” is often invoked in a narrow sense—in academia and beyond—with overtones of cold, affectless machinery, and negative connotations of loss of control by humans. But an engineering discipline can be what we want it to be.

*In the current era, we have a real opportunity to conceive of something historically new—a human-centric engineering discipline.*

I will resist giving this emerging discipline a name, but if the acronym “AI” continues to be used as placeholder nomenclature going forward, let’s be aware of the very real limitations of this placeholder. Let’s broaden our scope, tone down the hype and recognize the serious challenges ahead.

Michael I. Jordan

...
the computational, inferential, cognitive and biological sciences, first as a graduate student at UCSD and then as a faculty member at MIT and Berkeley. One of his recent roles is as a Faculty Partner and Co-Founder at AI@The House—a venture fund and accelerator in Berkeley. This fund aims to support not only AI activities, but also IA and II activities, and to do so in the context of a university environment that includes not only the engineering disciplines, but also the perspectives of the social sciences, the cognitive sciences and the humanities.
ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid’s blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

Just as the 18-year-old girls were realizing they were too big for the tiny conveyances — which belonged to a 6-year-old boy — a woman came running after them saying, “That’s my kid’s stuff.” Borden and her friend immediately dropped the bike and scooter and walked away.
But it was too late — a neighbor who witnessed the heist had already called the police. Borden and her friend were arrested and charged with burglary and petty theft for the items, which were valued at a total of $80.

Compare their crime with a similar one: The previous summer, 41-year-old Vernon Prater was picked up for shoplifting $86.35 worth of tools from a nearby Home Depot store.

Prater was the more seasoned criminal. He had already been convicted of armed robbery and attempted armed robbery, for which he served five years in prison, in addition to another armed robbery charge. Borden had a record, too, but it was for misdemeanors committed when she was a juvenile.

Yet something odd happened when Borden and Prater were booked into jail: A computer program spat out a score predicting the likelihood of each committing a future crime. Borden — who is black — was rated a high risk. Prater — who is white — was rated a low risk.

Two years later, we know the computer algorithm got it exactly backward. Borden has not been charged with any new crimes. Prater is serving an eight-year prison term for subsequently breaking into a warehouse and stealing thousands of dollars’ worth of electronics.

Scores like this — known as risk assessments — are increasingly common in courtrooms across the nation. They are used to inform decisions about who can be set free at every stage of the criminal justice system, from assigning bond amounts — as is the case in Fort
Lauderdale — to even more fundamental decisions about defendants’ freedom. In Arizona, Colorado, Delaware, Kentucky, Louisiana, Oklahoma, Virginia, Washington and Wisconsin, the results of such assessments are given to judges during criminal sentencing.

Rating a defendant’s risk of future crime is often done in conjunction with an evaluation of a defendant’s rehabilitation needs. The Justice Department’s National Institute of Corrections now encourages the use of such combined assessments at every stage of the criminal justice process. And a landmark sentencing reform bill currently pending in Congress would mandate the use of such assessments in federal prisons.

In 2014, then U.S. Attorney General Eric Holder warned that the risk scores might be injecting bias into the courts. He called for the U.S. Sentencing Commission to study their use. “Although these measures were crafted with the best of intentions, I am concerned that they inadvertently undermine our efforts to ensure individualized and equal justice,” he said, adding, “they may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society.”

The sentencing commission did not, however, launch a study of risk scores. So ProPublica did, as part of a larger examination of the powerful, largely hidden effect of algorithms in American life.

We obtained the risk scores assigned to more than 7,000 people arrested in Broward County, Florida, in 2013 and 2014 and checked to see how many were charged with new crimes over the next two years, the same benchmark used by the creators of the algorithm.

The score proved remarkably unreliable in forecasting violent crime: Only 20 percent of the people predicted to commit violent crimes actually went on to do so.
When a full range of crimes were taken into account — including misdemeanors such as driving with an expired license — the algorithm was somewhat more accurate than a coin flip. Of those deemed likely to re-offend, 61 percent were arrested for any subsequent crimes within two years.

We also turned up significant racial disparities, just as Holder feared. In forecasting who would re-offend, the algorithm made mistakes with black and white defendants at roughly the same rate but in very different ways.

- The formula was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants.
- White defendants were mislabeled as low risk more often than black defendants.

Could this disparity be explained by defendants’ prior crimes or the type of crimes they were arrested for? No. We ran a statistical test that isolated the effect of race from criminal history and recidivism, as well as from defendants’ age and gender. Black defendants were still 77 percent more likely to be pegged as at higher risk of committing a future violent crime and 45 percent more likely to be predicted to commit a future crime of any kind. (Read our analysis.)

The algorithm used to create the Florida risk scores is a product of a for-profit company, Northpointe. The company disputes our analysis.

In a letter, it criticized ProPublica’s methodology and defended the accuracy of its test: “Northpointe does not agree that the results of your analysis, or the claims being made based upon that analysis, are correct or that they accurately reflect the outcomes from the application of the model.”

Northpointe’s software is among the most widely used assessment tools in the country. The company does not publicly disclose the calculations used to arrive at defendants’ risk scores, so it is not possible for either defendants or the public to see what might be driving the disparity. (On Sunday, Northpointe gave ProPublica the basics of its future-crime formula — which includes factors such as education levels, and whether a defendant has a job. It did not share the specific calculations, which it said are proprietary.)
Northpointe’s core product is a set of scores derived from 137 questions that are either answered by defendants or pulled from criminal records. Race is not one of the questions. The survey asks defendants such things as: “Was one of your parents ever sent to jail or prison?” “How many of your friends/acquaintances are taking drugs illegally?” and “How often did you get in fights while at school?” The questionnaire also asks people to agree or disagree with statements such as “A hungry person has a right to steal” and “If people make me angry or lose my temper, I can be dangerous.”

The appeal of risk scores is obvious: The United States locks up far more people than any other country, a disproportionate number of them black. For more than two centuries, the key decisions in the legal process, from pretrial release to sentencing to parole, have been in the hands of human beings guided by their instincts and personal biases.

If computers could accurately predict which defendants were likely to commit new crimes, the criminal justice system could be fairer and more selective about who is incarcerated and for how long. The trick, of course, is to make sure the computer gets it right. If it’s wrong in one direction, a dangerous criminal could go free. If it’s wrong in another direction, it could result in someone unfairly receiving a harsher sentence or waiting longer for parole than is appropriate.

The first time Paul Zilly heard of his score — and realized how much was riding on it — was during his sentencing hearing on Feb. 15, 2013, in court in Barron County, Wisconsin. Zilly had been convicted of stealing a push lawnmower and some tools. The prosecutor recommended a year in county jail and follow-up supervision that could help Zilly with “staying on the right path.” His lawyer agreed to a plea deal.

But Judge James Babler had seen Zilly’s scores. Northpointe’s software had rated Zilly as a high risk for future violent crime and a medium risk for general recidivism. “When I look at the risk assessment,” Babler said in court, “it is about as bad as it could be.”

Then Babler overturned the plea deal that had been agreed on by the prosecution and defense and imposed two years in state prison and three years of supervision.
before deciding whether they should be released. Race, nationality and skin color were often used in making such predictions until about the 1970s, when it became politically unacceptable, according to a survey of risk assessment tools by Columbia University law professor Bernard Harcourt.

In the 1980s, as a crime wave engulfed the nation, lawmakers made it much harder for judges and parole boards to exercise discretion in making such decisions. States and the federal government began instituting mandatory sentences and, in some cases, abolished parole, making it less important to evaluate individual offenders.

But as states struggle to pay for swelling prison and jail populations, forecasting criminal risk has made a comeback.

Dozens of risk assessments are being used across the nation — some created by for-profit companies such as Northpointe and others by nonprofit organizations. (One tool being used in states including Kentucky and Arizona, called the Public Safety Assessment, was developed by the Laura and John Arnold Foundation, which also is a funder of ProPublica.)

There have been few independent studies of these criminal risk assessments. In 2013, researchers Sarah Desmarais and Jay Singh examined 19 different risk methodologies used in the United States and found that “in most cases, validity had only been examined in one or two studies” and that “frequently, those investigations were completed by the same people who developed the instrument.”

Their analysis of the research through 2012 found that the tools “were moderate at best in terms of predictive validity,” Desmarais said in an interview. And she could not find any

Two Drug Possession Arrests

Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.
substantial set of studies conducted in the United States that examined whether risk scores were racially biased. “The data do not exist,” she said.

Since then, there have been some attempts to explore racial disparities in risk scores. One 2016 study examined the validity of a risk assessment tool, not Northpointe’s, used to make probation decisions for about 35,000 federal convicts. The researchers, Jennifer Skeem at University of California, Berkeley, and Christopher T. Lowenkamp from the Administrative Office of the U.S. Courts, found that blacks did get a higher average score but concluded the differences were not attributable to bias.

The increasing use of risk scores is controversial and has garnered media coverage, including articles by the Associated Press, and the Marshall Project and FiveThirtyEight last year.

Most modern risk tools were originally designed to provide judges with insight into the types of treatment that an individual might need — from drug treatment to mental health counseling.

“What it tells the judge is that if I put you on probation, I’m going to need to give you a lot of services or you’re probably going to fail,” said Edward Latessa, a University of Cincinnati professor who is the author of a risk assessment tool that is used in Ohio and several other states.

But being judged ineligible for alternative treatment — particularly during a sentencing hearing — can translate into incarceration. Defendants rarely have an opportunity to challenge their assessments. The results are usually shared with the defendant's attorney, but the calculations that transformed the underlying data into a score are rarely revealed.

“Risk assessments should be impermissible unless both parties get to see all the data that go into them,” said Christopher Slobogin, director of the criminal justice program at Vanderbilt Law School. “It should be an open, full-court adversarial proceeding.”

Black Defendants’ Risk Scores
These charts show that scores for white defendants were skewed toward lower-risk categories. Scores for black defendants were not. (Source: ProPublica analysis of data from Broward County, Fla.)

Proponents of risk scores argue they can be used to reduce the rate of incarceration. In
2002, Virginia became one of the first states to begin using a risk assessment tool in the sentencing of nonviolent felony offenders statewide. In 2014, Virginia judges using the tool sent nearly half of those defendants to alternatives to prison, according to a state sentencing commission report. Since 2005, the state’s prison population growth has slowed to 5 percent from a rate of 31 percent the previous decade.

In some jurisdictions, such as Napa County, California, the probation department uses risk assessments to suggest to the judge an appropriate probation or treatment plan for individuals being sentenced. Napa County Superior Court Judge Mark Boessenecker said he finds the recommendations helpful. “We have a dearth of good treatment programs, so filling a slot in a program with someone who doesn’t need it is foolish,” he said.

However, Boessenecker, who trains other judges around the state in evidence-based sentencing, cautions his colleagues that the score doesn’t necessarily reveal whether a person is dangerous or if they should go to prison.

“A guy who has molested a small child every day for a year could still come out as a low risk because he probably has a job,” Boessenecker said. “Meanwhile, a drunk guy will look high risk because he’s homeless. These risk factors don’t tell you whether the guy ought to go to prison or not; the risk factors tell you more about what the probation conditions ought to be.”

Sometimes, the scores make little sense even to defendants.

James Rivelli, a 54-year old Hollywood, Florida, man, was arrested two years ago for shoplifting seven boxes of Crest Whitestrips from a CVS drugstore. Despite a criminal record that included aggravated assault, multiple thefts and felony drug trafficking, the Northpointe algorithm classified him as being at a low risk of reoffending.

“I am surprised it is so low,” Rivelli said when told by a reporter he had been rated a 3 out of a possible 10. “I spent five years in state prison in Massachusetts. But I guess they don’t count that here in Broward County.” In fact, criminal records from across the nation are supposed to be included in risk assessments.

Less than a year later, he was charged with two felony counts for shoplifting about $1,000 worth of tools from Home Depot. He said his crimes were fueled by drug addiction and that
NORTHPOINTE WAS FOUNDED in 1989 by Tim Brennan, then a professor of statistics at the University of Colorado, and Dave Wells, who was running a corrections program in Traverse City, Michigan.

Wells had built a prisoner classification system for his jail. “It was a beautiful piece of work,” Brennan said in an interview conducted before ProPublica had completed its analysis. Brennan and Wells shared a love for what Brennan called “quantitative taxonomy” — the measurement of personality traits such as intelligence, extroversion and introversion. The two decided to build a risk assessment score for the corrections industry.

Brennan wanted to improve on a leading risk assessment score, the LSI, or Level of Service Inventory, which had been developed in Canada. “I found a fair amount of weakness in the LSI,” Brennan said. He wanted a tool that addressed the major theories about the causes of
crime.

Brennan and Wells named their product the Correctional Offender Management Profiling for Alternative Sanctions, or COMPAS. It assesses not just risk but also nearly two dozen so-called “criminogenic needs” that relate to the major theories of criminality, including “criminal personality,” “social isolation,” “substance abuse” and “residence/stability.” Defendants are ranked low, medium or high risk in each category.

Two DUI Arrests

As often happens with risk assessment tools, many jurisdictions have adopted Northpointe’s software before rigorously testing whether it works. New York State, for instance, started using the tool to assess people on probation in a pilot project in 2001 and rolled it out to the rest of the state's probation departments — except New York City — by 2010. The state didn’t publish a comprehensive statistical evaluation of the tool until 2012. The study of more than 16,000 probationers found the tool was 71 percent accurate, but it did not evaluate racial differences.

A spokeswoman for the New York state division of criminal justice services said the study did not examine race because it only sought to test whether the tool had been properly calibrated to fit New York’s probation population. She also said judges in nearly all New York counties are given defendants’ Northpointe assessments during sentencing.

In 2009, Brennan and two colleagues published a validation study that found that Northpointe’s risk of recidivism score had an accuracy rate of 68 percent in a sample of 2,328 people. Their study also found that the score was slightly less predictive for black men than white men — 67 percent versus 69 percent. It did not examine racial disparities beyond that, including whether some groups were more likely to be wrongly labeled higher risk.
Brennan said it is difficult to construct a score that doesn’t include items that can be correlated with race — such as poverty, joblessness and social marginalization. “If those are omitted from your risk assessment, accuracy goes down,” he said.

In 2011, Brennan and Wells sold Northpointe to Toronto-based conglomerate Constellation Software for an undisclosed sum.

Wisconsin has been among the most eager and expansive users of Northpointe’s risk assessment tool in sentencing decisions. In 2012, the Wisconsin Department of Corrections launched the use of the software throughout the state. It is used at each step in the prison system, from sentencing to parole.

In a 2012 presentation, corrections official Jared Hoy described the system as a “giant correctional pinball machine” in which correctional officers could use the scores at every “decision point.”

Wisconsin has not yet completed a statistical validation study of the tool and has not said when one might be released. State corrections officials declined repeated requests to comment for this article.

Some Wisconsin counties use other risk assessment tools at arrest to determine if a defendant is too risky for pretrial release. Once a defendant is convicted of a felony anywhere in the state, the Department of Corrections attaches Northpointe’s assessment to the confidential presentence report given to judges, according to Hoy’s presentation.

In theory, judges are not supposed to give longer sentences to defendants with higher risk scores. Rather, they are supposed to use the tests primarily to determine which defendants are eligible for probation or treatment programs.

### Prediction Fails Differently for Black Defendants

<table>
<thead>
<tr>
<th></th>
<th>WHITE</th>
<th>AFRICAN AMERICAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labeled Higher Risk, But Didn’t Re-Offend</td>
<td>23.5%</td>
<td>44.9%</td>
</tr>
<tr>
<td>Labeled Lower Risk, Yet Did Re-Offend</td>
<td>47.7%</td>
<td>28.0%</td>
</tr>
</tbody>
</table>
Overall, Northpointe’s assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

But judges have cited scores in their sentencing decisions. In August 2013, Judge Scott Horne in La Crosse County, Wisconsin, declared that defendant Eric Loomis had been “identified, through the COMPAS assessment, as an individual who is at high risk to the community.” The judge then imposed a sentence of eight years and six months in prison.

Loomis, who was charged with driving a stolen vehicle and fleeing from police, is challenging the use of the score at sentencing as a violation of his due process rights. The state has defended Horne's use of the score with the argument that judges can consider the score in addition to other factors. It has also stopped including scores in presentencing reports until the state Supreme Court decides the case.

“The risk score alone should not determine the sentence of an offender,” Wisconsin Assistant Attorney General Christine Remington said last month during state Supreme Court arguments in the Loomis case. “We don’t want courts to say, this person in front of me is a 10 on COMPAS as far as risk, and therefore I’m going to give him the maximum sentence.”

That is almost exactly what happened to Zilly, the 48-year-old construction worker sent to prison for stealing a push lawnmower and some tools he intended to sell for parts. Zilly has long struggled with a meth habit. In 2012, he had been working toward recovery with the help of a Christian pastor when he relapsed and committed the thefts.

After Zilly was scored as a high risk for violent recidivism and sent to prison, a public defender appealed the sentence and called the score’s creator, Brennan, as a witness.

Brennan testified that he didn't design his software to be used in sentencing. “I wanted to stay away from the courts,” Brennan said, explaining that his focus was on reducing crime rather than punishment. “But as time went on I started realizing that so many decisions are made, you know, in the courts. So I gradually softened on whether this could be used in the courts or not.”
Still, Brennan testified, “I don’t like the idea myself of COMPAS being the sole evidence that a decision would be based upon.”

After Brennan’s testimony, Judge Babler reduced Zilly's sentence, from two years in prison to 18 months. “Had I not had the COMPAS, I believe it would likely be that I would have given one year, six months,” the judge said at an appeals hearing on Nov. 14, 2013.

Zilly said the score didn’t take into account all the changes he was making in his life — his conversion to Christianity, his struggle to quit using drugs and his efforts to be more available for his son. “Not that I’m innocent, but I just believe people do change.”

FLORIDA’S BROWARD COUNTY, where Brisha Borden stole the Huffy bike and was scored
as high risk, does not use risk assessments in sentencing. “We don’t think the [risk assessment] factors have any bearing on a sentence,” said David Scharf, executive director of community programs for the Broward County Sheriff’s Office in Fort Lauderdale.

Broward County has, however, adopted the score in pretrial hearings, in the hope of addressing jail overcrowding. A court-appointed monitor has overseen Broward County’s jails since 1994 as a result of the settlement of a lawsuit brought by inmates in the 1970s. Even now, years later, the Broward County jail system is often more than 85 percent full, Scharf said.

In 2008, the sheriff’s office decided that instead of building another jail, it would begin using Northpointe’s risk scores to help identify which defendants were low risk enough to be released on bail pending trial. Since then, nearly everyone arrested in Broward has been scored soon after being booked. (People charged with murder and other capital crimes are not scored because they are not eligible for pretrial release.)

The scores are provided to the judges who decide which defendants can be released from jail. “My feeling is that if they don’t need them to be in jail, let’s get them out of there,” Scharf said.

Scharf said the county chose Northpointe’s software over other tools because it was easy to use and produced “simple yet effective charts and graphs for judicial review.” He said the system costs about $22,000 a year.

In 2010, researchers at Florida State University examined the use of Northpointe’s system in Broward County over a 12-month period and concluded that its predictive accuracy was “equivalent” in assessing defendants of different races. Like others, they did not examine whether different races were classified differently as

Two Shoplifting Arrests

After Rivelli stole from a CVS and was caught with heroin in his car, he was rated a low risk. He later shoplifted $1,000 worth of tools from a Home Depot.
low or high risk.

Scharf said the county would review ProPublica's findings. “We’ll really look at them up close,” he said.

Broward County Judge John Hurley, who oversees most of the pretrial release hearings, said the scores were helpful when he was a new judge, but now that he has experience he prefers to rely on his own judgment. “I haven’t relied on COMPAS in a couple years,” he said.

Hurley said he relies on factors including a person’s prior criminal record, the type of crime committed, ties to the community, and their history of failing to appear at court proceedings.

ProPublica’s analysis reveals that higher Northpointe scores are slightly correlated with longer pretrial incarceration in Broward County. But there are many reasons that could be true other than judges being swayed by the scores — people with higher risk scores may also be poorer and have difficulty paying bond, for example.

Most crimes are presented to the judge with a recommended bond amount, but he or she can adjust the amount. Hurley said he often releases first-time or low-level offenders without any bond at all.

However, in the case of Borden and her friend Sade Jones, the teenage girls who stole a kid’s bike and scooter, Hurley raised the bond amount for each girl from the recommended $0 to $1,000 each.

Hurley said he has no recollection of the case and cannot recall if the scores influenced his decision.
The girls spent two nights in jail before being released on bond.

“We literally sat there and cried” the whole time they were in jail, Jones recalled. The girls were kept in the same cell. Otherwise, Jones said, “I would have gone crazy.” Borden declined repeated requests to comment for this article.

Jones, who had never been arrested before, was rated a medium risk. She completed probation and got the felony burglary charge reduced to misdemeanor trespassing, but she has still struggled to find work.

“I went to McDonald's and a dollar store, and they all said no because of my background,” she said. “It’s all kind of difficult and unnecessary.”
Due process requires that in criminal proceedings the accused be able to confront and cross examine the witnesses against him. How does one cross examine a computer algorithm that
judges are required to use? I hope the constitutional challenges come soon.

JL ➔ FastClock • 2 years ago

Depends on what you mean by "required to use". All sentencing risk assessment being used right now are advisory - meaning that they do not commit a judge to a course of action based on the score. A similar challenge was brought up in VA and failed because judges have a right to ignore the risk assessments scores completely.

uegidock ➔ JL • 2 years ago

And it's only used during sentencing right? It's one thing to challenge your accuser, but once you're found guilty it seems like the rules change.

Lars Carlson • 2 years ago

Isn't the very concept of 'predictive law enforcement' both entirely unconstitutional and utterly immoral? Who is allowing this? What kind of Orwellian nightmare is 'law enforcement' foisting on us? Here's some 'predictive law-enforcement' for you: investigate every single cop on the force and every single taxpayer making more than $500k a year. You'll find crime there with no problem, no problem at all.

poor_richard2 ➔ Lars Carlson • 2 years ago

The "for-profit" prison industry generally pressures Republican legislatures to write these laws. Its far more profitable than slave plantations, and removes Democratic voters en masse from voter registration rolls.

hogwash_bill ➔ poor_richard2 • 2 years ago

Front-running Democrats are in their pockets too. Hillary received $133,000 from for-profit prisons, the same amount as Marco Rubio. Only Bernie held the moral high ground.

Jm ➔ poor_richard2 • 2 years ago

Not only are Democrats complicit in this with Republicans, but Bill Clinton himself largely opened the floodgates for the modern prison industry. All of them should be fed to sharks in the ocean, at least then some good would come from their remaining existence.
from their remaining existence.

DPsombra • poor_richard2 • 2 years ago

I've pretty much always likened how we treat our prisoners to the "new slavery." But, it's cool! The Kardashians are on the air, so we don't have to think about these things!

poor_richard2 • 2 years ago

I've pretty much always likened how we treat our prisoners to the "new slavery." But, it's cool! The Kardashians are on the air, so we don't have to think about these things!

poor_richard2 • 2 years ago

Hillary sucks billionaire donor along every day for money.

poor_richard2 • 2 years ago

Hillary sucks billionaire donor along every day for money.

This comment is awaiting moderation. Show comment.

Jm • richstacy • 2 years ago

Left and right are BOTH wrong. You have been brainwashed and are fighting against yourself. WAKE UP.

poor_richard2 • 2 years ago

Hillary sucks billionaire donor along every day for money.

poor_richard2 • 2 years ago

Hillary sucks billionaire donor along every day for money.

This comment is awaiting moderation. Show comment.

Larry Strong • Lars Carlson • 2 years ago

Welcome to the Minority Report. And yes, the police are often the biggest criminals of all.

Larry Strong • Lars Carlson • 2 years ago

Welcome to the Minority Report. And yes, the police are often the biggest criminals of all.

Nixak*77* • Larry Strong • 2 years ago

My Brief Assessment of this 'Unbiased' 'Minority 'pre-crime' Report'-

This so-called 'unbiased' [Oh really?] Assessment report, as a computer based 'predictor' of the likelihood that an offender may offend again, asks 137 questions- which IMO could / should be cut-down to just 1/2 or even 1/3 of that. A number of questions are thinly disguised racially biased, & many are quite redundant [IE: it asks about gang affiliation at-least 3 or 4 different Xs]. In fact many or most [or even all] of some entire sections are questionable at best: IE: Under Family Criminality: Asks if you were raised by both parents or NOT, or if you're adopted or raised in foster care [obviously this is biased vs Black people], then if your parents &/or even siblings were ever arrested, & then asks again if they were ever convicted [= redundant].

Then there's the sections on Work / Hire-ability, & Education- obviously since
Blacks' unemployment rate has for decades been 2Xs that of whites, this is biased vs Blacks. Then it even asks even if you do work, does your [lack of] Education & Skills set only allow you to at-best get minimum wage jobs- Huhh WTF!! Damn! Talk about a Black guy can't win for losing!!

Then there's a section asking if you've got financial problems. So just how

MikeCody  Lars Carlson  2 years ago
Every decision made concerning bail, probation, or parole is an example of predictive law enforcement. In theory, these tools reduce the ability of the decision makers to decide based on the fact that the defendant is the County Commissioner's nephew. Like any computer based scoring system, the algorithms need tuning over time as statistical analysis finds flaws.

jakee308  Lars Carlson  2 years ago
Well they're not actually predicting crime as much as they're predicting whether someone is likely to commit a crime again.

And there are studies that show some people will continue to commit crimes no matter how much or how often they get encarcerated.

But yes, this is the first step on a path to predict who will be a criminal and who will not. They've got some pilot programs going in Great Britain doing just that. The Brits actually like to lock people up more than the US but they don't lock them up for as long so it doesn't show up as much. Plus they're a smaller country so the raw numbers are lower.

Nigel Tolley  jakee308  2 years ago
Not sure where you get that idea.
The UK, like most places, has realised that prison is the last resort. You go to prison and your life is affected forever - you get hooked on drugs, you lose your house, it destroys the futures of your kids.
Hence prison is a last resort.
Also, magistrates can only go up to 3 years (might still be 2) in prison, maximum. They have to pass the case up the tree to a professional judge in more serious cases.
No, it isn't perfect. The UK government passes bad laws, same as any other. But a lot fewer die with random violence.

England also has quadruple the violent crime rate of America 2,000 per 100,000 vs. 466. and also leads Europe. It is NOT the model of an effective criminal justice system.

It may not be perfect, but consider that I know precisely one person who has been murdered, and the murder rate here is 1.0 per 100,000, vs 3.8 per 100,000 in the USA.

As for reported violence? If you never see violence, you remember it. If it is everyday, you forget the majority.

But enough statistics, have an "anecdata".

I work a "high risk" job. One I feel pretty sure would’ve got me into a lot of violence in the USA - indeed, I know that warrants are largely served by SWAT in the US now. In the course of ~10 years doing warrant entries, breaking in to thousands of homes and businesses a year, without anything more than a bit of paper, a Warrant Officer (unarmed, not even a stab vest) & an engineer (for the gas or electric meter), I have experienced violence. We've had a few "close calls" - a guy with guns, a few with axes, etc., many drug farms (nearly all grow houses), & yes, I've been assaulted. I've been assaulted twice.

Twice. In total, over thousands of jobs where *we literally bang on your door until you answer*, and if you don't, I pick the locks and we let ourselves in. And if you are inside, we tell you we have a warrant and refuse to leave.

We call the police if we feel we need to, I think 3 times in one week is the most, & we sometimes do 8 jobs a day.
So no. Actual "what an American would call violence" violent crime is tiny over here. And our murder rate is a quarter yours.

Barry • 2 years ago
Wow. Unbelievable that an inaccurate computer algorithm attempting to predict future crime should be allowed to even be seen by a judge. This is unconstitutional.

Criminals should be punished for what they did, not what they might do.

Nixak*77* • 2 years ago
'Failed' Polygraphs are NOT allowed as evidence against a defendant, for this very reason. Though They supposedly have a 80% 'success' rate of predicting if someone's Lying- That means they have a 20% FAILURE Rate [false positives / false negatives or undetermined] = Well below the legal standard of ' Guilt Beyond a Reasonable Doubt'!

As such this so-called 'unbiased' 'Minority {pre-crime} Report' which is even MORE UNRELIABLE than polygraphs are, should have also NEVER been allowed in court-effectively used as 'evidence' against defendants!!!

Christopher Perrien • 2 years ago
Good point, though recidivism I am sure plays into most judges' judgements.

Barry • 2 years ago
That is perfectly OK. Turning that decision making over to a computer that proves inaccurate is not OK. The criteria for applying past crimes to punishment should be simple and transparent.

disagree • 2 years ago
I'll accept that only if they get the algorithm to 98-99%. And even then I think a jury should still be required to sign off on it.

Kal E • 2 years ago
If you are tasked with determining whether or not someone is allowed to be released under bail, or forced to wait in jail until the case comes up, and you are asked to
consider risk of flight (i.e., skipping town) or re-offending, what would you take into consideration? As you go through that thought process, imagine programming a machine to do the same and take the same variables into deciding- that’s it. That’s the boogey man that most people reacting with worry don’t understand. It’s actually not that scary or complicated.

What is scary is that the results of this program have been shown to be inaccurate and racially biased (even after controlling for different rates of crimes between certain races).

Even scarier is when 10,000 judges across the country make decisions where no one can see their "algorithm" and bias- and we just let them continue to perpetuate injustice. I prefer an algorithm that everyone can see, study, and work to fix. It’s easier to fix and test the algorithm than to train and hope judges don’t bring bias into decision-making.

OK, it may generally make sense to penalize people for their own past bad behavior. But should we increase someone's penalty because his dad went to prison when he was a child?

It depends. First, we would need to establish, through rigorous research, that there is causation between those two variables. Assuming that there is, and that adding that variable to our model increases its predictive strength, then we can consider it on ethical grounds and decide if it should be added to the model. Regardless, it should be a matter a public policy and entirely open (no proprietary secret models); not a willy-nilly free market for predictive models with every city/county/state deciding what to do on their own, with no federal oversight/controls- which is what we have now.
I have been working as a statistician for 10 years, which is relevant when I say that predictive models are always wrong. They are even more wrong when it comes to predicting human behavior.

Surprisingly, this model seems to be doing pretty well on an overall level. 70% for a consumer model would be amazing. The problem is, this isn't a consumer model, this is a model that is impacting the lives of individuals. And it is wrong, analysts know it is going to be wrong some percentage of the time.

I have some general ideas how the strong racial bias could have happened from a modeling perspective. Often there are hidden relationships in the data that are difficult to see, which is why the concept of "Correlation does not imply Causation" is so critical to all modeling endeavors.

For instance the percentage of the black population with family members in prison is higher than the non black population. It is likely that race SHOULD have been considered in order to control for the pre existing racial bias in our current system. There are many other possible reasons for the model to behave this way.

But the reasons behind this racial bias are not the point. This is a model that the modelers should never have built, in my opinion. Or at the very least it should never have been deployed this way. Handing a simplistic analysis like this off to non analysts without a very strong sense of it's error and overall limitations is just a set up for failure.

Some times the question is not "Can we predict this?" But "Should we?"

If you actually look at the questionnaire on which this {non}'Predictive' {mis}'Analysis' is based, you'll see it's quite redundant as well as biased [IE: asks about gang affiliation at-least 3Xs - 4Xs]. It has whole sections where IMO much/most if not ALL the questions are of dubious relevance [IE: one section asks mainly about boredom]. IMO the sections asking about family ties [were / are you raised in a 2 Parent home], employment & education status, & if one lives in a hi-crime area- are all [blatantly] Biased against Black people [= 'clever' ways of 'determining' if someone's most likely Black without actually asking the question outright].

Thus IMO this questionnaire is at-least 2Xs - 3Xs too long- it's that redundant, biased &/or dubiously relevant [see my above comment where I go into more detail about this].
And the alternative is what? Release every person unless we are 100% certain they will re-offend? Lock everyone up? Lean heavily on flawed human intuition, which is often less accurate and more biased than measures like this?

Build a system that is designed to reduce recidivism, (Especially through drug rehab and occupational programs) and also use punishments that fit the crime committed?

And in the real world we still need to exercise some sort of judgement about who we release, under what conditions, etc. Which side do we err on?

Clearly no system is perfect, but some are better than others. Even if you are willing to accept (much) higher crime rates in your community to reduce incarceration, the reality is that many people are not willing to make these choices. You cannot simply will them away. Given this reality it seems to me that there are good welfare enhancing arguments for accepting a model like this, even knowing that they are imperfect (both type I and type II errors), because they are the least bad choice that we have. You know what they say about making the perfect the enemy of the good.....

I personally think that making one person's life harder because a flawed computer model said they are a certain way, is always wrong. These aren't just statistics, they are actual people.

I certainly think criminals should still be punished and incarcerated, but I think the entire system as a whole needs further reform. That is not my area of expertise.

FYI: This 'Minority {pre-crime} Report' seems to be used by judges to justify sentencing people to extra-time [or NOT], NOT by parole review
boards &/or judges to decide if those already doing time may / should be eligible for early release.

Random Critical Analysis  Nixak*77*  • 2 years ago

I know but the issues are very similar. The algorithm isn't doing anything that judges and other humans haven't been doing for ages using their own heuristics and the like (keep in mind that this is only being used as a guide)

Incidentally, propublica's claims of bias here are unfounded. I know two people that independently analyzed it and I saw their work-- no sign of bias (if properly analyzed) and the prediction actually looks quite good in the sample propublica provided (obviously quite predictive, linear trend between risk and probability of offending, no sign of b/w bias, etc). I doubt the median judge would outperform on his own.

Nixak*77*  • 2 years ago

FYI: I looked at this 137 questionnaire myself & IMO it has multiple instances of NOT only [NOT so hidden] biases, redundancies & even dubiously relevant questions, but that's so even for some of its entire sections!! As such IMO this questionnaire is at-least 2Xs to 3Xs too long [See my 1st comment(s) above].

Random Critical Analysis  Nixak*77*  • 2 years ago

I haven't spent much time looking at the questionnaire, but the actual risk scores generated by the algorithm show no signs of bias against blacks. In fact, at any given recidivism risk score blacks are at least as likely to reoffend (often more likely though these diffs aren't usually significant). If the questionnaire is as "biased" as you claim, how do you explain that?

Nixak*77*  • 2 years ago

This article has given several detailed cases of comparatively similar cases for Black vs white 'offenders', which clearly shows some degree of racial bias. I don't know how many actual cases this Propublica investigative team checked into [I think they gave links to their full assessments of some of them].
assessments of cases, as they did the actual questionnaire], but I looked
at this questionnaire myself, & I stand by my assessment which I go over
more fully in my 1st comment(s) above.

This comment is awaiting moderation. Show comment.

Nixak*77* ➔ Random Critical Analysis • 2 years ago

Bottom Line: After reading this questionnaire if I were the lawyer for a
Black defendant, I’d advise my Black Client to refuse to answer this
questionnaire the same way lawyers usually advise their clients against
testifying against / incriminating themselves in court & against taking
lie-detector tests- which at an 80% success rate is apparently
significantly more 'reliable' than this biased 'Minority {pre-crime}
Report'.

In fact IMO the NAACP & ACLU should challenge the validity of using
this questionnaire in court, using the illegality of using failed lie-
detector tests as 'evidence' vs defendants in court as a precedence, along
w this questionnaire's apparent racial biases.
IMO it's amazing you think this test that's just 62% - 63% accurate [per
you] is somehow 'good' enough- when almost every official test one
takes, a 62% score is a 'D minus'!! Especially considering that failed lie-
detector tests [w an 80% accuracy rate] can NOT be uses against
defendants in court, even though an 80% score = a 'B minus'.

I take it that you are neither a statistician nor a lawyer.

Well if you must know- I know enough Math to understand basic stats
concepts, & know enough about basic law to understand the legality /
constitutionality of this biased 'Minority {pre-crime} Report'
questionnaire should be contested, & 'defendants' should have the legal
right to refuse to take it.
Because of racist enforcement of laws AND the creation of laws that target black populations - going all the way back to the 19th Century "Black codes" - Black people have been criminalized and therefore overwhelmingly overpopulated in the baseline figures used to make statistical analyses of crime. This then creates a feedback loop where Black people are increasingly targeted, arrested, and imprisoned. In other words, it becomes a self-fulfilling prophesy that enables racist whites to blame Black people and hide their racism behind statistics.

Black people are 6 times as likely as whites to be murdered. That figure is based on a simple count of murder victims and is not skewed by any "racial bias" either in the statistical process or by the law courts. The trouble with you leftist is that you cannot even acknowledge the real problems which are destroying black America, because anyone who even so much as mentions this stuff is immediately branded a racist.

It is the lack of sheer economic opportunities available for all Americans that is not being addressed by those who just remind us the Dire implications of So Called Black on Black crimes. Criminal Activision at All Levels are merely symptoms of A consistent Lack of Economic Development and Capital investment in Human Development which is whofully, or Absent in Major Urban Centers which Accidentally are homes to large minority populationd. Law Enforcement and Prisons are only one Aspect of the Solution but we will be fooling ourselves if we don't address the fundamental lack of opportunities, in vestments and sheer attitudes towards minority populations across the country.

And almost all of those murders are committed by young black men. Black on black violent crime is at epidemic levels. And the Director of the FBI noted only yesterday that the murder rate has gone up 60% in many large American cites, probably due to the fact that the police are now afraid to do their jobs.

Interestingly enough, if ProPublica have done their stats homework correctly here (and that is an if - it's not easy), then what they've found is evidence of this. If the profiling
software doesn't look explicitly at race, but returns predictions that are disfavorable to blacks - well, another way of saying the same thing is that their test shows blacks do better than their circumstances suggest. So if they included race as an explicit factor in their model, not only would it be more predictive, it would be better for blacks as well...

There are some big ifs about that, though. Frankly, I have my doubts. Also, even if this happy result holds up for this particular variable (race), I think that we should oppose it on general principle. You should not be judged for those of your characteristics you obviously have no power over.

The correct approach is to make the models so that they are actively blinded to the protected characteristics - so that if you look at the variables, your guess about the subject's race or gender would be no better than random. This is possible in this day and age. Northpointe's techniques are simplistic by today's standards (and probably their own time's).

And yes, predictive accuracy might suffer - but that’s the price we pay for equality. In the long term, I think we are also better off with it.
OPINION

Artificial Intelligence’s White Guy Problem

By Kate Crawford
June 25, 2016

ACCORDING to some prominent voices in the tech world, artificial intelligence presents a looming existential threat to humanity: Warnings by luminaries like Elon Musk and Nick Bostrom about “the singularity” — when machines become smarter than humans — have attracted millions of dollars and spawned a multitude of conferences.

But this hand-wringing is a distraction from the very real problems with artificial intelligence today, which may already be exacerbating inequality in the workplace, at home and in our legal and judicial systems. Sexism, racism and other forms of discrimination are being built into the machine-learning algorithms that underlie the technology behind many “intelligent” systems that shape how we are categorized and advertised to.

Take a small example from last year: Users discovered that Google’s photo app, which applies automatic labels to pictures in digital photo albums, was classifying images of black people as gorillas. Google apologized; it was unintentional.
But similar errors have emerged in Nikon’s camera software, which misread images of Asian people as blinking, and in Hewlett-Packard’s web camera software, which had difficulty recognizing people with dark skin tones.

This is fundamentally a data problem. Algorithms learn by being fed certain images, often chosen by engineers, and the system builds a model of the world based on those images. If a system is trained on photos of people who are overwhelmingly white, it will have a harder time recognizing nonwhite faces.

A very serious example was revealed in an investigation published last month by ProPublica. It found that widely used software that assessed the risk of recidivism in criminals was twice as likely to mistakenly flag black defendants as being at a higher risk of committing future crimes. It was also twice as likely to incorrectly flag white defendants as low risk.

The reason those predictions are so skewed is still unknown, because the company responsible for these algorithms keeps its formulas secret — it’s proprietary information. Judges do rely on machine-driven risk assessments in different ways — some may even discount them entirely — but there is little they can do to understand the logic behind them.

Police departments across the United States are also deploying data-driven risk-assessment tools in “predictive policing” crime prevention efforts. In many cities, including New York, Los Angeles, Chicago and Miami, software analyses of large sets of historical crime data are used to forecast where crime hot spots are most likely to emerge; the police are then directed to those areas.

At the very least, this software risks perpetuating an already vicious cycle, in which the police increase their presence in the same places they are already policing (or overpolicing), thus ensuring that more arrests come from those areas. In the United
States, this could result in more surveillance in traditionally poorer, nonwhite neighborhoods, while wealthy, whiter neighborhoods are scrutinized even less. Predictive programs are only as good as the data they are trained on, and that data has a complex history.

Histories of discrimination can live on in digital platforms, and if they go unquestioned, they become part of the logic of everyday algorithmic systems. Another scandal emerged recently when it was revealed that Amazon's same-day delivery service was unavailable for ZIP codes in predominantly black neighborhoods. The areas overlooked were remarkably similar to those affected by mortgage redlining in the mid-20th century. Amazon promised to redress the gaps, but it reminds us how systemic inequality can haunt machine intelligence.

And then there's gender discrimination. Last July, computer scientists at Carnegie Mellon University found that women were less likely than men to be shown ads on Google for highly paid jobs. The complexity of how search engines show ads to internet users makes it hard to say why this happened — whether the advertisers preferred showing the ads to men, or the outcome was an unintended consequence of the algorithms involved.

Regardless, algorithmic flaws aren’t easily discoverable: How would a woman know to apply for a job she never saw advertised? How might a black community learn that it were being overpoliced by software?

We need to be vigilant about how we design and train these machine-learning systems, or we will see ingrained forms of bias built into the artificial intelligence of the future.
Like all technologies before it, artificial intelligence will reflect the values of its creators. So inclusivity matters — from who designs it to who sits on the company boards and which ethical perspectives are included. Otherwise, we risk constructing machine intelligence that mirrors a narrow and privileged vision of society, with its old, familiar biases and stereotypes.

If we look at how systems can be discriminatory now, we will be much better placed to design fairer artificial intelligence. But that requires far more accountability from the tech community. Governments and public institutions can do their part as well: As they invest in predictive technologies, they need to commit to fairness and due process.

While machine-learning technology can offer unexpected insights and new forms of convenience, we must address the current implications for communities that have less power, for those who aren’t dominant in elite Silicon Valley circles.

Currently the loudest voices debating the potential dangers of superintelligence are affluent white men, and, perhaps for them, the biggest threat is the rise of an artificially intelligent apex predator.

But for those who already face marginalization or bias, the threats are here.

Kate Crawford is a principal researcher at Microsoft and co-chairwoman of a White House symposium on society and A.I.

Follow The New York Times Opinion section on Facebook and Twitter (@NYTOpinion), and sign up for the Opinion Today newsletter.

A version of this article appears in print on June 26, 2016, on Page SR11 of the New York edition with the headline: A.I.'s White Guy Problem
Ten simple rules for responsible big data research

Matthew Zook1*, Solon Barocas2, danah boyd2,3, Kate Crawford2,4, Emily Keller3, Seeta Peña Gangadharan5, Alyssa Goodman6, Rachelle Hollander7, Barbara A. Koenig8, Jacob Metcalf9, Arvind Narayanan10, Alondra Nelson11, Frank Pasquale12

1 Department of Geography, University of Kentucky, Lexington, Kentucky, United States of America, 2 Microsoft Research, New York, New York, United States of America, 3 Data & Society, New York, New York, United States of America, 4 Information Law Institute, New York University, New York, New York, United States of America, 5 Department of Media and Communications, London School of Economics, London, United Kingdom, 6 Harvard-Smithsonian Center for Astrophysics, Harvard University, Cambridge, Massachusetts, United States of America, 7 Center for Engineering Ethics and Society, National Academy of Engineering, Washington, DC, United States of America, 8 Institute for Health Aging, University of California-San Francisco, San Francisco, California, United States of America, 9 Ethical Resolve, Santa Cruz, California, United States of America, 10 Department of Computer Science, Princeton University, Princeton, New Jersey, United States of America, 11 Department of Sociology, Columbia University, New York, New York, United States of America, 12 Carey School of Law, University of Maryland, Baltimore, Maryland, United States of America

* zook@uky.edu

Introduction

The use of big data research methods has grown tremendously over the past five years in both academia and industry. As the size and complexity of available datasets has grown, so too have the ethical questions raised by big data research. These questions become increasingly urgent as data and research agendas move well beyond those typical of the computational and natural sciences, to more directly address sensitive aspects of human behavior, interaction, and health. The tools of big data research are increasingly woven into our daily lives, including mining digital medical records for scientific and economic insights, mapping relationships via social media, capturing individuals’ speech and action via sensors, tracking movement across space, shaping police and security policy via “predictive policing,” and much more.

The beneficial possibilities for big data in science and industry are tempered by new challenges facing researchers that often lie outside their training and comfort zone. Social scientists now grapple with data structures and cloud computing, while computer scientists must contend with human subject protocols and institutional review boards (IRBs). While the connection between individual datum and actual human beings can appear quite abstract, the scope, scale, and complexity of many forms of big data creates a rich ecosystem in which human participants and their communities are deeply embedded and susceptible to harm. This complexity challenges any normative set of rules and makes devising universal guidelines difficult.

Nevertheless, the need for direction in responsible big data research is evident, and this article provides a set of “ten simple rules” for addressing the complex ethical issues that will inevitably arise. Modeled on PLOS Computational Biology’s ongoing collection of rules, the recommendations we outline involve more nuance than the words “simple” and “rules” suggest. This nuance is inevitably tied to our paper’s starting premise: all big data research on social, medical, psychological, and economic phenomena engages with human subjects, and researchers have the ethical responsibility to minimize potential harm.
The variety in data sources, research topics, and methodological approaches in big data belies a one-size-fits-all checklist; as a result, these rules are less specific than some might hope. Rather, we exhort researchers to recognize the human participants and complex systems contained within their data and make grappling with ethical questions part of their standard workflow. Towards this end, we structure the first five rules around how to reduce the chance of harm resulting from big data research practices; the second five rules focus on ways researchers can contribute to building best practices that fit their disciplinary and methodological approaches. At the core of these rules, we challenge big data researchers who consider their data disentangled from the ability to harm to reexamine their assumptions. The examples in this paper show how often even seemingly innocuous and anonymized data have produced unanticipated ethical questions and detrimental impacts.

This paper is a result of a two-year National Science Foundation (NSF)-funded project that established the Council for Big Data, Ethics, and Society, a group of 20 scholars from a wide range of social, natural, and computational sciences (http://bdes.datasociety.net/). The Council was charged with providing guidance to the NSF on how to best encourage ethical practices in scientific and engineering research, utilizing big data research methods and infrastructures [1].

1. Acknowledge that data are people and can do harm

One of the most fundamental rules of responsible big data research is the steadfast recognition that most data represent or impact people. Simply starting with the assumption that all data are people until proven otherwise places the difficulty of disassociating data from specific individuals front and center. This logic is readily evident for "risky" datasets, e.g., social media with inflammatory language, but even seemingly benign data can contain sensitive and private information, e.g., it is possible to extract data on the exact heart rates of people from YouTube videos [2]. Even data that seemingly have nothing to do with people might impact individuals’ lives in unexpected ways, e.g., oceanographic data that change the risk profiles of communities’ and properties’ values or Exchangeable Image Format (EXIF) records from photos that contain location coordinates and reveal the photographer’s movement or even home location.

Harm can also result when seemingly innocuous datasets about population-wide effects are used to shape the lives of individuals or stigmatize groups, often without procedural recourse [3,4]. For example, social network maps for services such as Twitter can determine credit-worthiness [5], opaque recidivism scores can shape criminal justice decisions in a racially disparate manner [6], and categorization based on zip codes resulted in less access to Amazon Prime same-day delivery service for African-Americans in United States cities [7]. These high-profile cases show that apparently neutral data can yield discriminatory outcomes, thereby compounding social inequities.

Other cases show that “public” datasets are easily adapted for highly invasive research by incorporating other data, such as Hague et al.’s [8] use of property records and geographic profiling techniques to allegedly identify the pseudonymous artist Banksy [9]. In particular, data ungoverned by substantive consent practices, whether social media or the residual DNA we continually leave behind us, may seem public but can cause unintentional breaches of privacy and other harms [9,10].

Start with the assumption that data are people (until proven otherwise), and use it to guide your analysis. No one gets an automatic pass on ethics.

2. Recognize that privacy is more than a binary value

Breaches of privacy are key means by which big data research can do harm, and it is important to recognize that privacy is contextual [11] and situational [12], not reducible to a simple
public/private binary. Just because something has been shared publicly does not mean any subsequent use would be unproblematic. Looking at a single Instagram photo by an individual has different ethical implications than looking at someone’s full history of all social media posts. Privacy depends on the nature of the data, the context in which they were created and obtained, and the expectations and norms of those who are affected. Understand that your attitude towards acceptable use and privacy may not correspond with those whose data you are using, as privacy preferences differ across and within societies.

For example, Tene and Polonetsky [13] explore how pushing past social norms, particularly in novel situations created by new technologies, is perceived by individuals as “creepy” even when they do not violate data protection regulations or privacy laws. Social media apps that utilize users’ locations to push information, corporate tracking of individuals’ social media and private communications to gain customer intelligence, and marketing based on search patterns have been perceived by some to be “creepy” or even outright breaches of privacy. Likewise, distributing health records is a necessary part of receiving health care, but this same sharing brings new ethical concerns when it goes beyond providers to marketers.

Privacy also goes beyond single individuals and extends to groups [10]. This is particularly resonant for communities who have been on the receiving end of discriminatory data-driven policies historically, such as the practice of redlining [14, 15]. Other examples include community maps—made to identify problematic properties or an assertion of land rights—being reused by others to identify opportunities for redevelopment or exploitation [16]. Thus, reusing a seemingly public dataset could run counter to the original privacy intents of those who created it and raise questions about whether it represents responsible big data research.

Situate and contextualize your data to anticipate privacy breaches and minimize harm. The availability or perceived publicness of data does not guarantee lack of harm, nor does it mean that data creators consent to researchers using their data.

3. Guard against the reidentification of your data

It is problematic to assume that data cannot be reidentified. There are numerous examples of researchers with good intentions and seemingly good methods failing to anonymize data sufficiently to prevent the later identification of specific individuals [17]; in other cases, these efforts were extremely superficial [18, 19]. When datasets thought to be anonymized are combined with other variables, it may result in unexpected reidentification, much like a chemical reaction resulting from the addition of a final ingredient.

While the identificatory power of birthdate, gender, and zip code is well known [20], there are a number of other parameters—particularly the metadata associated with digital activity—that may be as or even more useful for identifying individuals [21]. Surprising to many, unlabeled network graphs—such as location and movement, DNA profiles, call records from mobile phone data, and even high-resolution satellite images of the earth—can be used to reidentify people [22]. More important than specifying the variables that allow for reidentification, however, is the realization that it is difficult to recognize these vulnerable points a priori [23]. Factors discounted today as irrelevant or inherently harmless—such as battery usage—may very well prove to be a significant vector of personal identification tomorrow [24]. For example, the addition of spatial location can turn social media posts into a means of identifying home location [25], and Google’s reverse image search can connect previously separate personal activities—such as dating and professional profiles—in unanticipated ways [26]. Even data about groups—“aggregate statistics”—can have serious implications if they reveal that certain communities, for example, suffer from stigmatized diseases or social behavior much more than others [27].
Identify possible vectors of reidentification in your data. Work to minimize them in your published results to the greatest extent possible.

4. Practice ethical data sharing

For some projects, sharing data is an expectation of the human participants involved and thus a key part of ethical research. For example, in rare genetic disease research, biological samples are shared in the hope of finding cures, making dissemination a condition of participation. In other projects, questions of the larger public good—an admittedly difficult to define category—provide compelling arguments for sharing data, e.g., the NIH-sponsored database of Genotypes and Phenotypes (dbGaP), which makes deidentified genomic data widely available to researchers, democratizing access, or the justice claim made by the Institute of Medicine about the value of mandating that individual-level data from clinical trials be shared among researchers [28]. Asking participants for broad, as opposed to narrowly structured consent for downstream data management makes it easier to share data. Careful research design and guidance from IRBs can help clarify consent processes. However, we caution that even when broad consent was obtained upfront, researchers should consider the best interests of the human participant, proactively considering the likelihood of privacy breaches and reidentification issues. This is of particular concern for human DNA data, which is uniquely identifiable.

These types of projects, however—in which rules of use and sharing are well governed by informed consent and right of withdrawal—are increasingly the exception rather than the rule for big data. In our digital society, we are followed by data clouds composed of the trace elements of daily life—credit card transactions, medical test results, closed-circuit television (CCTV) images and video, smart phone apps, etc.—collected under mandatory terms of service rather than responsible research design overseen by university compliance officers. While we might wish to have the standards of informed consent and right of withdrawal, these informal big data sources are gathered by agents other than the researcher—private software companies, state agencies, and telecommunications firms. These data are only accessible to researchers after their creation, making it impossible to gain informed consent a priori, and contacting the human participants retroactively for permission is often forbidden by the owner of the data or is impossible to do at scale.

Of course, researchers within software companies and state institutions collecting these data have a special responsibility to address the terms under which data are collected; but that does not exempt the end-user of shared data. In short, the burden of ethical use (see Rules 1 to 3) and sharing is placed on the researcher, since the terms of service under which the human subjects’ data were produced can often be extremely broad with little protection for breaches of privacy. In these circumstances, researchers must balance the requirements from funding agencies to share data [29] with their responsibilities to the human beings behind the data they acquired. A researcher needs to inform funding agencies about possible ethical concerns before the research begins and guard against reidentification before sharing.

Share data as specified in research protocols, but proactively address concerns of potential harm from informally collected big data.

5. Consider the strengths and limitations of your data; big does not automatically mean better

In order to do both accurate and responsible big data research, it is important to ground datasets in their proper context including conflicts of interests. Context also affects every stage of research: from data acquisition, to cleaning, to interpretation of findings, and dissemination of the results. During the step of data acquisition, it is crucial to understand both the source of
the data and the rules and regulations with which they were gathered. This is especially important in cases of research conducted in relatively loose regulatory environments, in which use of answers to research questions may conflict with the expectations of those who provided the data. One possible approach might be the ethical norms employed to track the provenance of artifacts, often in cooperation and collaboration with the communities from which they come (e.g., archaeologists working in indigenous communities to determine the disposition of material culture). In a similar manner, computer scientists use data lineage techniques to track the evolution of a dataset and often to trace bugs in the data.

Being mindful of the data’s context provides the foundation for clarifying when your data and analysis are working and when they are not. While it is tempting to interpret findings based on big data as a clear outcome, a key step within scientific research is clearly articulating what data or an indicator represent and what they do not. Are your findings as clear-cut if your interpretation of a social media posting switches from a recording of fact to the performance of a social identity? Given the messy, almost organic nature of many datasets derived from social actions, it is fundamental that researchers be sensitive to the potential multiple meanings of data.

For example, is a Facebook post or an Instagram photo best interpreted as an approval/disapproval of a phenomenon, a simple observation, or an effort to improve status within a friend network? While any of these interpretations are potentially valid, the lack of context makes it even more difficult to justify the choice of one understanding over another. Reflecting on the potential multiple meanings of data fosters greater clarity in research hypotheses and also makes researchers aware of the other potential uses of their data. Again, the act of interpretation is a human process, and because the judgments of those (re)using your data may differ from your own, it is essential to clarify both the strengths and shortcomings of the data.

For example, is a Facebook post or an Instagram photo best interpreted as an approval/disapproval of a phenomenon, a simple observation, or an effort to improve status within a friend network? While any of these interpretations are potentially valid, the lack of context makes it even more difficult to justify the choice of one understanding over another. Reflecting on the potential multiple meanings of data fosters greater clarity in research hypotheses and also makes researchers aware of the other potential uses of their data. Again, the act of interpretation is a human process, and because the judgments of those (re)using your data may differ from your own, it is essential to clarify both the strengths and shortcomings of the data.

**Document the provenance and evolution of your data. Do not overstate clarity; acknowledge messiness and multiple meanings.**

**6. Debate the tough, ethical choices**

Research involving human participants at federally funded institutions is governed by IRBs charged with preventing harm through well-established procedures and are familiar to many researchers. IRBs, however, are not the sole arbiter of ethics; many ethical issues involving big data are outside of their governance mandate. Precisely because big data researchers often encounter situations that are foreign to or outside of the mandate of IRBs, we emphasize the importance of debating the issues within groups of peers.

Rather than a bug, the lack of clear-cut solutions and governance protocols should be more appropriately understood as a feature that researchers should embrace within their own work. Discussion and debate of ethical issues is an essential part of professional development—both within and between disciplines—as it can establish a mature community of responsible practitioners. Bringing these debates into coursework and training can produce peer reviewers who are particularly well placed to raise these ethical questions and spur recognition of the need for these conversations.

A precondition of any formal ethics rules or regulations is the capacity to have such open-ended debates. As digital social scientist and ethicist Annette Markham [30] writes, “we can make [data ethics] an easier topic to broach by addressing ethics as being about choices we make at critical junctures; choices that will invariably have impact.” Given the nature of big data, bringing technical, scientific, social, and humanistic researchers together on projects enables this debate to emerge as a strength because, if done well, it provides the means to understand the ethical issues from a range of perspectives and disrupt the silos of disciplines.
There are a number of good models for interdisciplinary ethics research, such as the trainings offered by the Science and Justice research center at the University of California, Santa Cruz [32] and Values in Design curricula [33]. Research ethics consultation services, available at some universities as a result of the Clinical and Translational Science Award (CTSA) program of the National Institutes of Health (NIH), can also be resources for researchers [34].

Some of the better-known “big data” ethical cases—i.e., the Facebook emotional contagion study [35]—provide extremely productive venues for cross-disciplinary discussions. Why might one set of scholars see this as a relatively benign approach while other groups see significant ethical shortcomings? Where do researchers differ in drawing the line between responsible and irresponsible research and why? Understanding the different ways people discuss these challenges and processes provides an important check for researchers, especially if they come from disciplines not focused on human subject concerns.

Moreover, the high visibility surrounding these events means that (for better or worse) they represent the “public” view of big data research, and becoming an active member of this conversation ensures that researchers can give voice to their insights rather than simply being at the receiving end of policy decisions. In an effort to help these debates along, the Council for Big Data, Ethics, and Society has produced a number of case studies focused specifically on big data research and a white paper with recommendations to start these important conversations (http://bdes.datasociety.net/output/).

Engage your colleagues and students about ethical practice for big data research.

7. Develop a code of conduct for your organization, research community, or industry

The process of debating tough choices inserts ethics directly into the workflow of research, making “faking ethics” as unacceptable as faking data or results. Internalizing these debates, rather than treating them as an afterthought or a problem to outsource, is key for successful research, particularly when using trace data produced by people. This is relevant for all research including those within industry who have privileged access to the data streams of digital daily life. Public attention to the ethical use of these data should not be avoided; after all, these datasets are based on an infrastructure that billions of people are using to live their lives, and there is a compelling public interest that research is done responsibly.

One of the best ways to cement this in daily practice is to develop codes of conduct for use in your organization or research community and for inclusion in formal education and ongoing training. The codes can provide guidance in peer review of publications and in funding consideration. In practice, a highly visible case of unethical research brings problems to an entire field, not just to those directly involved. Moreover, designing codes of conduct makes researchers more successful. Issues that might otherwise be ignored until they blow up—e.g., Are we abiding by the terms of service or users’ expectations? Does the general public consider our research “creepy”? [13]—can be addressed thoughtfully rather than in a scramble for damage control. This is particularly relevant to public-facing private businesses interested in avoiding potentially unfavorable attention.

An additional and longer-term advantage of developing codes of conduct is that it is clear that change is coming to big data research. The NSF funded the Council for Big Data, Ethics, and Society as a means of getting in front of a developing issue and pending regulatory changes within federal rules for the protection of human subjects that are currently under review [1]. Actively developing rules for responsible big data research within a research community is a key way researchers can join this ongoing process.
Establish appropriate codes of ethical conduct within your community. Make industry researchers and representatives of affected communities active contributors to this process.

8. Design your data and systems for auditability

Although codes of conduct will vary depending on the topic and research community, a particularly important element is designing data and systems for auditability. Responsible internal auditing processes flow easily into audit systems and also keep track of factors that might contribute to problematic outcomes. Developing automated testing processes for assessing problematic outcomes and mechanisms for auditing other’s work during review processes can help strengthen research as a whole. The goal of auditability is to clearly document when decisions are made and, if necessary, backtrack to an earlier dataset and address the issue at the root (e.g., if strategies for anonymizing data are compromised).

Designing for auditability also brings direct benefits to researchers by providing a mechanism for double-checking work and forcing oneself to be explicit about decisions, increasing understandability and replicability. For example, many types of social media and other trace data are unstructured, and answers to even basic questions such as network ties, location, and randomness depend on the steps taken to collect and collate data. Systems of auditability clarify how different datasets (and the subsequent analysis) differ from each other, aiding understanding and creating better research.

Plan for and welcome audits of your big data practices.

9. Engage with the broader consequences of data and analysis practices

It is also important for responsible big data researchers to think beyond the traditional metrics of success in business and the academy. For example, the energy demands for digital daily life, a key source of big data for social science research, are significant in this era of climate change [36]. How might big data research lessen the environmental impact of data analytics work? For example, should researchers take the lead in asking cloud storage providers and data processing centers to shift to sustainable and renewable energy sources? As important and publicly visible users of the cloud, big data researchers collectively represent an interest group that could rally behind such a call for change.

The pursuit of citations, reputation, or money is a key incentive for pushing research forward, but it can also result in unintended and undesirable outcomes. In contrast, we might ask to what extent is a research project focused on enhancing the public good or the underserved of society? Are questions about equity or promoting other public values being addressed in one’s data streams, or is a big data focus rendering them invisible or irrelevant to your analysis [37]? How can increasingly vulnerable yet fundamentally important public resources—such as state-mandated cancer registries—be protected? How might research aid or inhibit different business and political actors? While all big data research need not take up social and cultural questions, a fundamental aim of research goes beyond understanding the world to considering ways to improve it.

Recognize that doing big data research has societal-wide effects.

10. Know when to break these rules

The final (and counterintuitive) rule is the charge to recognize when it is appropriate to stray from these rules. For example, in times of natural disaster or a public health emergency, it may be important to temporarily put aside questions of individual privacy in order to serve a larger
public good. Likewise, the use of genetic or other biological data collected without informed consent might be vital in managing an emerging disease epidemic.

Moreover, be sure to review the regulatory expectations and legal demands associated with protection of privacy within your dataset. After all, this is an exceedingly slippery slope, so before following this rule (to break others), be cautious that the “emergency” is not simply a convenient justification. The best way to ensure this is to build experience in engaging in the tough debates (Rule 6), constructing codes of conduct (Rule 7), and developing systems for auditing (Rule 8). The more mature the community of researchers is about their processes, checks, and balances, the better equipped it is to assess when breaking the rules is acceptable. It may very well be that you do not come to a final clear set of practices. After all, just as privacy is not binary (Rule 2), neither is responsible research. Ethics is often about finding a good or better, but not perfect, answer, and it is important to ask (and try to answer) the challenging questions. Only through this engagement can a culture of responsible big data research emerge.

Understand that responsible big data research depends on more than meeting checklists.

Conclusion

The goal of this set of ten rules is to help researchers do better work and ultimately become more successful while avoiding larger complications, including public mistrust. To achieve this, however, scholars must shift from a mindset that is rigorous when focused on techniques and methodology and naïve when it comes to ethics. Statements to the effect that “Data is [sic] already public” [38] are unjustified simplifications of much more complex data ecosystems embedded in even more complex and contingent social practices. Data are people, and to maintain a rigorously naïve definition to the contrary [18] will end up harming research efforts in the long run as pushback comes from the people whose actions and utterances are subject to analysis.

In short, responsible big data research is not about preventing research but making sure that the work is sound, accurate, and maximizes the good while minimizing harm. The problems and choices researchers face are real, complex, and challenging and so too must be our response. We must treat big data research with the respect that it deserves and recognize that unethical research undermines the production of knowledge. Fantastic opportunities to better understand society and our world exist, but with these opportunities also come the responsibility to consider the ethics of our choices in the everyday practices and actions of our research. The Council for Big Data, Ethics, and Society (http://bdes.datasociety.net/) provides an initial set of case studies, papers, and even ten simple rules for guiding this process; it is now incumbent on you to use and improve these in your research.

Acknowledgments

This article also benefitted from the input of Geoff Bowker and Helen Nissenbaum.

References


5. Danyllo WA, Alisson VB, Alexandre ND, Moacir LM, Jansepetrus BP, Oliveira RF. Identifying relevant users and groups in the context of credit analysis based on data from Twitter. InCloud and Green Computing (CGC), 2013 Third International Conference on 2013 Sep 30 (pp. 587–592). IEEE.


Topic Modeling for Humanists: A Guided Tour

It’s that time again! Somebody else posted a really clear and enlightening description of topic modeling on the internet. This time it was Allen Riddell, and it’s so good that it inspired me to write this post about topic modeling that includes no actual new information, but combines a lot of old information in a way that will hopefully be useful. If there’s anything I’ve missed, by all means let me know and I’ll update accordingly.

Introducing Topic Modeling

Topic models represent a class of computer programs that automagically extracts topics from texts. What a topic actually is will be revealed shortly, but the crux of the matter is that if I feed the computer, say, the last few speeches of President Barack Obama, it’ll come back telling me that the president mainly talks about the economy, jobs, the Middle East, the upcoming election, and so forth. It’s a fairly clever and exceptionally versatile little algorithm that can be customized to all sorts of applications, and a tool that many digital humanists would do well to have in their toolbox.

From the outset it’s worth clarifying some vocabulary, and mentioning what
topic models can and cannot do. “LDA” and “Topic Model” are often thrown around synonymously, but LDA is actually a special case of topic modeling in general produced by David Blei and friends in 2002. It was not the first topic modeling tool, but is by far the most popular, and has enjoyed copious extensions and revisions in the years since. The myriad variations of topic modeling have resulted in an alphabet soup of names that might be confusing or overwhelming to the uninitiated; ignore them for now. They all pretty much work the same way.

When you run your text through a standard topic modeling tool, what comes out the other end first is several lists of words. Each of these lists is supposed to be a “topic.” Using the example from before of presidential addresses, the list might look like:

1. Job Jobs Loss Unemployment Growth
2. Economy Sector Economics Stock Banks
3. Afghanistan War Troops Middle-East Taliban Terror
4. Election Romney Upcoming President
5. ... etc.

The computer gets a bunch of texts and spits out several lists of words, and we are meant to think those lists represent the relevant “topics” of a corpus. The algorithm is constrained by the words used in the text; if Freudian psychoanalysis is your thing, and you feed the algorithm a transcription of your dream of bear-fights and big caves, the algorithm will tell you nothing about your father and your mother; it’ll only tell you things about bears and caves. It’s all text and no subtext. Ultimately, LDA is an attempt to inject semantic meaning into vocabulary; it’s a bridge, and often a helpful one. Many dangers face those who use this bridge without fully understanding it, which is exactly what the rest of this post will help you avoid.
Learning About Topic Modeling

The pathways to topic modeling are many and more, and those with different backgrounds and different expertise will start at different places. This guide is for those who’ve started out in traditional humanities disciplines and have little background in programming or statistics, although the path
becomes more strenuous as we get closer Blei’s original paper on LDA (as that is our goal.) I will try to point to relevant training assistance where appropriate. A lot of the following posts repeat information, but there are often little gems in each which make them all worth reading.

**No Experience Necessary**

The following posts, read in order, should be completely understandable to pretty much everyone.

**The Fable**

Perhaps the most interesting place to start is the stylized account of topic modeling by Matt Jockers, who weaves a tale of authors sitting around the LDA buffet, taking from it topics with which to write their novels. According to Jockers, the story begins in a quaint town, . . .

somewhere in New England perhaps. The town is a writer’s retreat, a place they come in the summer months to seek inspiration. Melville is there, Hemingway, Joyce, and Jane Austen just fresh from across the pond. In this mythical town there is spot popular among the inhabitants; it is a little place called the “LDA Buffet.” Sooner or later all the writers go there to find themes for their novels. . .

The blog post is a fun read, and gets at the general idea behind the process of a topic model without delving into any of the math involved. Start here if you are a humanist who’s never had the chance to interact with topic models.

**A Short Overview**

Clay Templeton over at MITH wrote a short, less-stylized overview of topic modeling which does a good job discussing the trio of issues currently of im-
In this post I map out a basic genealogy of topic modeling in the humanities, from the highly cited paper that first articulated Latent Dirichlet Allocation (LDA) to recent work at MITH.

Templeton’s piece is concise, to the point, and offers good examples of topic models used for applications you’ll actually care about. It won’t tell you any more about the process of topic modeling than Jockers’ article did, but it’ll get you further into the world of topic modeling as it is applied in the humanities.

**An Example: The American Political Science Review**

Now that you know the basics of what a topic model actually is, perhaps the best thing is to look at an actual example to ground these abstract concepts. David Blei’s team shoved all of the journal articles from *The American Political Science Review* into a topic model, resulting in a list of 20 topics that represent the content of that journal. Click around on the page; when you click one of the topics, it sends you to a page listing many of the words in that topic, and many of the documents associated with it. When you click on one of the document titles, you’ll get a list of topics related to that document, as well as a list of other documents that share similar topics.

This page is indicative of the sort of output topic modeling will yield on a corpus. It is a simple and powerful tool, but notice that none of the automated topics have labels associated with them. The model requires us to make meaning out of them, they require interpretation, and without fully understanding the underlying algorithm, one cannot hope to properly interpret the results.
Written by yours truly, this next description of topic modeling begins to get into the formal process the computer goes through to create the topic model, rather than simply the conceptual process behind it. The blog post begins with a discussion of the predecessors to LDA in an attempt to show a simplified version of how LDA works, and then uses those examples to show what LDA does differently. There’s no math or programming, but the post does attempt to bring up relevant vocabulary and define them in terms familiar to those without programming experiencing.

With this matrix, LSA uses singular value decomposition to figure out how each word is related to every other word. Basically, the more often words are used together within a document, the more related they are to one another. It’s worth noting that a “document” is defined somewhat flexibly. For example, we can call every paragraph in a book its own “document,” and run LSA over the individual paragraphs.

Only the first half of this post is relevant to our topic modeling guided tour. The second half, a section on topic modeling and network analysis, discusses various extended uses that are best left for later.

Computational Process

Ted Underwood provides the next step in understanding what the computer goes through when topic modeling a text.

... it’s a long step up from those posts to the computer-science articles that explain “Latent Dirichlet Allocation” mathematically. My goal in this post is to provide a bridge between those two levels of difficulty.

Computer scientists make LDA seem complicated because they care about proving that their algorithms work. And the proof is indeed
brain-squashingly hard. But the practice of topic modeling makes good sense on its own, without proof, and does not require you to spend even a second thinking about “Dirichlet distributions.” When the math is approached in a practical way, I think humanists will find it easy, intuitive, and empowering. This post focuses on LDA as shorthand for a broader family of “probabilistic” techniques. I’m going to ask how they work, what they’re for, and what their limits are.

His is the first post that talks in any detail about the iterative process going into algorithms like LDA, as well as some of the assumptions those algorithms make. He also shows the first formula appearing in this guided tour, although those uncomfortable with formulas need not fret. The formula is not essential to understanding the post, but for those curious, later posts will explicate it. And really, Underwood does a great job of explaining a bit about it there.

Be sure to read to the very end of the post. It discusses some of the important limitations of topic modeling, and trepidations that humanists would be wise to heed. He also recommends reading Blei’s recent article on Probabilistic Topic Models, which will be coming up shortly in this tour.

**Computational Process From Another Angle**

It may not matter whether you read this or the last article by Underwood first; they’re both first passes to what the computer goes through to generate topics, and they explain the process in slightly different ways. The highlight of Edwin Chen’s blog post is his section on “Learning,” followed a section expanding that concept.

*And for each topic t, compute two things: 1) \( p(\text{topic } t \mid \text{ document } d) \) = the proportion of words in document \( d \) that are currently assigned to topic \( t \), and 2) \( p(\text{word } w \mid \text{ topic } t) \) = the proportion of assignments to topic \( t \) over all documents that come from this word \( w \). Reassign \( w \) a*
new topic, where we choose topic $t$ with probability $p(\text{topic } t \mid \text{document } d) \times p(\text{word } w \mid \text{topic } t)$ (according to our generative model, this is essentially the probability that topic $t$ generated word $w$, so it makes sense that we resample the current word’s topic with this probability).

This post both explains the meaning of these statistical notations, and tries to actually step the reader through the process using a metaphor as an example, a bit like Jockers’ post from earlier but more closely resembling what the computer is going through. It’s also worth reading through the comments on this post if there are parts that are difficult to understand.

This ends the list of articles and posts that require pretty much no prior knowledge. Reading all of these should give you a great overview of topic modeling, but you should by no means stop here. The following section requires a very little bit of familiarity with statistical notation, most of which can be found at this Wikipedia article on Bayesian Statistics.

**Some Experience Required**

Not much experience! You can even probably ignore most of the formulae in these posts and still get quite a great deal out of them. Still, you’ll get the most out of the following articles if you can read signs related to probability and summation, both of which are fairly easy to look up on Wikipedia. The dirty little secret of most papers that include statistics is that you don’t actually need to understand all of the formulae to get the gist of the article. If you want to fully understand everything below, however, I’d highly suggest taking an introductory course or reading a textbook on Bayesian statistics. I second Allen Riddell in suggesting Hoff’s *A First Course in Bayesian Statistical Methods* (2009), Kruschke’s *Doing Bayesian Data Analysis* (2010), or Lee’s *Bayesian Statistics: An Introduction* (2004). My own favorite is Kruschke’s; there are puppies on the cover.
Return to Blei

David Blei co-wrote the original LDA article, and his descriptions are always informative. He recently published a great introduction to probabilistic topic models for those not terribly familiar with it, and although it has a few formulae, it is the fullest computational description of the algorithm, gives a brief overview of Bayesian statistics, and provides a great framework with which to read the following posts in this series. Of particular interest are the sections on “LDA and Probabilistic Models” and “Posterior Computation for LDA.”

LDA and other topic models are part of the larger field of probabilistic modeling. In generative probabilistic modeling, we treat our data as arising from a generative process that includes hidden variables. This generative process defines a joint probability distribution over both the observed and hidden random variables. We perform data analysis by using that joint distribution to compute the conditional distribution of the hidden variables given the observed variables. This conditional distribution is also called the posterior distribution.

Really, read this first. Even if you don’t understand all of it, it will make the following reads easier to understand.

Back to Basics

The post that inspired this one, by Allen Riddell, explains the mixture of unigrams model rather than the LDA model, which allows Riddell to back up and explain some important concepts. The intended audience of the post is those with an introductory background in Bayesian statistics but it offers a lot even to those who do not have that. Of particular interest is the concrete example he uses, articles from German Studies journals, and how he actually walks you through the updating procedure of the algorithm as it infers topic and document distributions.
The second move swaps the position of our ignorance. Now we guess which documents are associated with which topics, making the assumption that we know both the makeup of each topic distribution and the overall prevalence of topics in the corpus. If we continue with our example from the previous paragraph, in which we had guessed that “literary” was more strongly associated with topic two than topic one, we would likely guess that the seventh article, with ten occurrences of the word “literary”, is probably associated with topic two rather than topic one (of course we will consider all the words, not just “literary”). This would change our topic assignment vector to \( z = (1,1,1,1,1,2,1,1,2,1,2,2,2,2,2,2,2,2) \). We take each article in turn and guess a new topic assignment (in many cases it will keep its existing assignment).

The last section, discussing the choice of number of topics, is not essential reading but is really useful for those who want to delve further.

Some Necessary Concepts in Text Mining

Both a case study and a helpful description, David Mimno’s recent article on Computational Historiography from ACM Transactions on Computational Logic goes through a hundred years of Classics journals to learn something about the field (very similar Riddell’s article on German Studies). While the article should be read as a good example of topic modeling in the wild, of specific interest to this guide is his “Methods” section, which includes an important discussion about preparing text for this sort of analysis. In order for computational methods to be applied to text collections, it is first necessary to represent text in a way that is understandable to the computer. The fundamental unit of text is the word, which we here define as a sequence of (unicode) letter characters. It is important to distinguish two uses of word: a word type is a distinct sequence of characters, equivalent to a dictionary headword or lemma; while a
What follows is a description of the primitive objects of a text analysis, and how to deal with variations in words, spelling, various languages, and so forth. Mimno also discusses smoothed distributions and word distance, both important concepts when dealing with these sorts of analyses.

Further Reading

By now, those who managed to get through all of this can probably understand most of the original LDA paper by Blei, Ng, and Jordan (most of it will be review!), but there’s a lot more out there than that original article. Mimno has a wonderful bibliography of topic modeling articles, and they’re tagged by topic to make finding the right one for a particular application that much easier.

Applications: How To Actually Do This Yourself

David Blei’s website on topic modeling has a list of available software, as does a section of Mimno’s Bibliography. Unfortunately, almost everything in those lists requires some knowledge of programming, and as yet I know of no really simple implementation of topic modeling. There are a few implementations for humanists that are supposed to be released soon, but to my knowledge, at the time of this writing the simplest tool to run your text through is called MALLET.

MALLET is a tool that does require a bit of comfort with the command-line, though it’s really just the same four commands or so over and over again. It’s a fairly simply software to run once you’ve gotten the hang of it, but that
first part of the learning curve could be a bit more like a learning cliff.

On their website, MALLET has a link called “Tutorial” – don’t click it. Instead, after downloading and installing the software, follow the directions on the “Importing Data” page. Then, follow the directions on the “Topic Modeling” page. If you’re a Windows user, Shawn Graham, Ian Milligan, and I wrote a tutorial on how to get it running when you run into a problem (and if this is your first time, you will), and it also includes directions for Macs. Unfortunately, a more detailed tutorial is beyond the scope of this tour, but between these links you’ve got a good chance of getting your first topic model up and running.

Examples in the DH World

There are a lot of examples of topic modeling out there, and here are some that I feel are representative of the various uses it can be put to. I’ve already mentioned David Mimno’s computational historiography of classics journals, as well as Allen Riddell’s similar study of German Studies publications. Both papers are good examples of using topic modeling as a meta-analysis of a discipline. Turning the gaze towards our collective navels, Matt Jockers used LDA to find what’s hot in the Digital Humanities, and Elijah Meeks has a great process piece looking at topics in definitions of digital humanities and humanities computing.

Lisa Rhody has an interesting exploratory topical analysis of poetry, and Rob Nelson as well discusses (briefly) making an argument via topic modeling applied to poetry, which he expands in this New York Times blog post. Continuing in the literary vein, Ted Underwood talks a bit about the relationship of words to topics, as well as a curious find linking topic models and family relations.

One of the great and oft-cited examples of topic modeling in the humanities is Rob Nelson’s Mining the Dispatch, which looks at the changing discus-
sion during the American Civil War through an analysis of primary texts. Just as Nelson looks at changing topics in the news over time, so too does Newman and Block in an analysis of eighteenth century newspapers, as well as Yang, Torget, and Mihalcea in a more general look at topic modeling and newspapers. In another application using primary texts, Cameron Blevins uses MALLET to run an in-depth analysis of an eighteenth century diary.

Future Directions

This is not actually another section of the post. This is your conscience telling you to go try topic modeling for yourself.

---

Aditi Muralidharan

July 26, 2012

Scott! You write an excellent introduction to the subject. I’m linking to it from my blog, and pointing everyone who wants to get started in topic modeling to it from now.

Reply to Aditi

Elijah Meeks

July 26, 2012

---
Aditi is right, this is really good, Scott, and required reading for anyone in DH.

Reply to Elijah

---

**Rosvita Rauch**  
August 1, 2012

Yes, thank you! This is a very useful overview. Rosvita

Reply to Rosvita

---

**Dario**  
October 18, 2013

“as yet I know of no really simple implementation of topic modeling.” There is one software implementing topic modeling entirely on a GUI. However it is a commercial software, and for an educational license you need to get more than 600$ out of your pockets. The software is T-Lab (www.tlab.it/en), and among several other things it implements topic modeling using LDA and Gibbs.

Reply to Dario

---

**Lakshmi Murthy**  
December 17, 2013

This is an excellent introduction to Topic Modeling. I am trying to use Topic Modeling for my project at NYU Stern and this article is a great starting point. Thank you.

Reply to Lakshmi
There’s also a nice implementation called “Gensim: topic modeling for humans”. It has lucid tutorials and uses the Python language (IMO the easiest language for tech novices).

I found a GUI that sits on top of MALLET that has worked well for me: https://code.google.com/p/topic-modeling-tool/

Fantastic introduction, thank you – I have lots of new tabs open with the articles you summarized in, ready to be enjoyed and hopefully understood! I must also say, the formatting of this post was very nice indeed.
These four papers were presented at the 2013 MLA Convention in Boston, at a roundtable called “The Dark Side of Digital Humanities.” Held in a large packed room, the session provoked a great deal of often-heated commentary—in the Twitter feed during and after the roundtable, in the discussion following the presentations, and in several blog posts and articles in the days following the convention. To get a sense of the aim of the roundtable, here is a selection from the roundtable proposal made to the MLA selection committee:

The same neoliberal logic that informs the ongoing destruction of the mainstream humanities has encouraged foundations, corporations, and university administrations to devote new resources to the digital humanities. Indeed it is largely due to the apparently instrumental or utilitarian value of the digital humanities that university administrators, foundation officers, and government agencies are so eager to fund DH projects, create DH undergraduate and graduate programs, and hire DH faculty. And because there is no sign that these funding streams are going to dry up any time soon, and no sign on the horizon of an increase in funding for the “crisis humanities,” there is great potential for increased tension between the “haves” of digital humanities and the “have-nots” of mainstream humanities.

As a result of this tension, DH finds itself faced with a choice between what this roundtable playfully refers to as the “dark side” and “the light side” of the force. From the rise of for-profit universities to the push to develop online “content modules” branded with the names of established universities, it is clear that the 21st century university is fundamentally networked, nearly impossible to envisage without the objects and methodological practices of the computational sciences. What are the relations, then, between DH as a strict tool- and interface-based practice and the institutional logics of the new neoliberal networked universities? What can we make, further, of the links between the claims made on behalf of both online learning initiatives and the new tools for digital humanities research: that they each have a radical, open, democratic aspect that is linked to mass literacy
movements, making scholarly materials widely available to populations that had not previously had such access? What are the relations between new reading techniques (text mining, distant reading) and new modes of content delivery? Is it even possible to have “distant reading” without somehow also contributing to the project of distant education? Part of the work of this panel will be to envisage a model of digital humanities that is not rooted in technocratic rationality or neoliberal economic calculus but rather that emerges from as well as informs traditionary practices of humanist inquiry.

Our interest in this roundtable is on the impact of digital humanities on research and teaching in the humanities in higher education—the question of how digital humanities will impact the future of the humanities in general. Composed of entry-level, mid-career, and senior scholars with a history of curricular, scholarly, and hands-on engagement with digital media, this roundtable will pose several questions and challenges to the digital humanities. Taking neoliberalism as the economic framework within which we are reluctantly operating, we want to explore alternative paths on which digital humanists might travel to ameliorate, rather than exacerbate, some of the internecine divisions that this economic crisis has precipitated and intensified.

In order to preserve the flavor of the roundtable itself, the panelists have chosen to present their contributions in virtually unrevised form. Fuller versions, with appropriate scholarly apparatus, can be found in a special issue of Differences (vol. 25, no. 1) 2014.

Wendy Hui Kyong Chun
This talk was given on January 4, 2013 at the Modern Language Association (MLA) convention. It focuses on a paradox surrounding DH: the disparity between the hype surrounding DH and the material work conditions surrounding much DH (adjunct/soft money positions, the constant drive to raise funds, the lack of scholarly recognition of DH work for promotions). In it, I call for us to work together—across the various fields and divisions—to create a university that is fair and just for all (teachers, students, researchers). I also call for us to find value in what is often discarded as “useless” in order to take on the really hard problems that face us.

We have been asked to be provocative, so I will use my eight minutes to provoke: to agitate and perhaps aggravate, excite, and perhaps incite. I want to propose that the dark side of the digital humanities is its bright side, its alleged promise: its alleged promise to save the humanities by making them
and their graduates relevant, by giving their graduates technical skills that will allow them to thrive in a difficult and precarious job market. Speaking partly as a former engineer, this promise strikes me as bull: knowing GIS or basic statistics or basic scripting (or even server-side scripting) is not going to make English majors competitive with engineers or CS geeks trained here or increasingly abroad (**straight up programming jobs are becoming increasingly less lucrative**).1

But let me be clear. My critique is not directed at DH per se. DH projects have extended and renewed the humanities and revealed that the kind of critical thinking (close textual analysis) that the humanities have always been engaged in is and has always been central to crafting technology and society. DH projects such as "Feminist Dialogues in Technology," a distributed online cooperative course that will be taught in fifteen universities across the globe—courses that use technology not simply to disseminate but also to rethink and regenerate cooperatively education at a global scale—these projects are central. As well, the humanities should play a big role in "big data" not simply because we are good at pattern recognition (because we can read narratives embedded in data), but also and more importantly because we can see what big data ignores. We can see the ways in which so many big data projects, by restricting themselves to certain databases and terms, shine a flashlight under a streetlamp.

I also want to stress that my sympathetic critique is not aimed at the humanities, but at the general euphoria surrounding technology and education. That is, it takes aim at the larger project of rewriting political and pedagogical problems into technological ones, into problems that technology can fix. This rewriting ranges from the idea that MOOCs, rather than a serious public commitment to education, can solve the problem of the spiraling cost of education (MOOCs that enroll, but do not graduate; MOOCs that miss the point of what we do, for when lectures work, they work because they create communities, because they are, to use Benedict Anderson’s phrase, "extraordinary mass ceremonies") to the blind embrace of technical skills. To put it as plainly as possible: there are a lot of unemployed engineers out there, from forty-something assembly programmers in Silicon Valley to young kids graduating from community colleges with CS degrees and no jobs. Also, there is a huge gap between industrial skills and university training. Every good engineer has to be re-taught how to program; every film graduate re-taught how to make films.

My main argument is this: the vapid embrace of the digital is a form of what Lauren Berlant has called "cruel optimism." Berlant argues, "[A] relation of cruel optimism exists when something you desire is actually an obstacle to your flourishing" (1). She emphasizes that optimistic relations are not inherently cruel, but become so when "the object that draws your attachment actively impedes the aim that brought you to it initially." Crucially, this attachment is doubly cruel "insofar as
the very pleasures of being inside a relation have become sustaining regardless of the content of the relation, such that a person or world finds itself bound to a situation of profound threat that is, at the same time, profoundly confirming” (2).

So, the blind embrace of DH (**think here of “The Old Order Changeth***) allows us to believe that this time (once again) graduate students will get jobs. It allows us to believe that the problem facing our students and our profession is a lack of technical savvy rather than an economic system that undermines the future of our students.

As Lauren Berlant points out, the hardest thing about cruel optimism is that, even as it destroys us in the long term, it sustains us in the short term. DH allows us to tread water: to survive, if not thrive (**think here of the ways in which so many DH projects and jobs depend on soft money and the ways in which DH projects are often—and very unfairly—not counted towards tenure or promotion***). It allows us to sustain ourselves and to justify our existence in an academy that is increasingly a sinking ship.

The humanities are sinking—if they are—not because of their earlier embrace of theory or multiculturalism, but because they have capitulated to a bureaucratic technocratic logic. They have conceded to a logic, an enframing (**to use Heidegger’s term***) that has made publishing a question of quantity rather than quality, so that we spew forth MPUs or minimum publishable units. A logic, an enframing that can make teaching a burden rather than a mission, so that professors and students are increasingly at odds. A logic, an enframing that has divided the profession and made us our own worst enemies so that those who have jobs for life, deny jobs to others—others who have often accomplished more than they (than we)—have.

The academy is a sinking ship—if it is—because it sinks our students into debt, and this debt, generated by this optimistic belief that a university degree automatically guarantees a job, is what both sustains and kills us. This residual belief/hope stems from another time when most of us could not go to university—another time when young adults with degrees received good jobs, not necessarily because of what they learned, but because of the society in which they lived.

Now, if the bright side of the digital humanities is the dark side, let me suggest that the dark side—what is now considered to be the dark side—may be where we need to be. The dark side, after all, is the side of passion. The dark side, or what has been made dark, is what all that bright talk has been turning away from (critical theory, critical race studies—all that fabulous work that #transformDH is doing).

This dark side also entails taking on our fears and biases to create deeper collaborations with the sciences and engineering. It entails forging joint (frictional and sometimes fractious) coalitions to take on problems such as education, global change, etc. It means realizing that the humanities do not have a lock on creative or critical thinking and realizing that
research in the sciences can be as useless as research in the humanities—and that this is a good thing. It is called basic research.

It also entails realizing that what is most interesting about the digital in general is perhaps not what has been touted as its promise, but rather what is been discarded or decried as its trash (**think here of all those failed DH tools, which have still opened up new directions**). It entails realizing that what is most interesting is what has been discarded or decried as inhuman: rampant publicity, anonymity, the ways in which the Internet vexes the relationship between public and private, the ways it compromises our autonomy and involves us with others and other machines in ways we do not entirely know and control (**think here of the constant and promiscuous exchange of information that drives the Internet, something that is usually hidden from us**).

As Natalia Cecire has argued DH is best when it takes on the humanities, as well as the digital. Maybe, just maybe, by taking on the inhumanities, we will transform the digital as well.

Richard Grusin

The proposal I submitted for the 2013 roundtable opened with the following questions: “Is it only an accident that the emergence of digital humanities has coincided with the intensification of the economic crisis in the humanities in higher education? Or is there a connection between these two developments?” I began with these questions to help make sense of a feeling that has bothered me since MLA 2011—the incommensurate affective moods between panels on “digital” humanities and those on what might be understood as “crisis” humanities. This mood did not appear suddenly in 2011 but has been emerging, largely unspoken or ignored, at least since the financial meltdown of 2008. Nor has it gone away, as demonstrated by the current MOOC bubble, which generates digital utopian arguments about the remaking of higher education while intensifying the sense of precarity that has come to replace the security of tenure as the predominant affective mood of the academy. (See Figure 38.1.)

The first convention held on the new January schedule, MLA11 had been premediated as something of a new start for the Modern Language Association. This sense of a new beginning was accompanied in Los Angeles by a sense of loss evident in panels devoted to the crisis in the humanities that had been produced by radical funding cuts in public support for education in Europe, Australia, and the United States. These cuts, and the concomitant transformation of the professoriate, have been under way for several decades now (particularly in the United States), but in the recessionary aftermath of the financial crisis of 2008, they reached a level unimaginable to most academics. Panels on the immediacy of the crisis in the humanities were accompanied by wide-
spread historical critique of the devastating effects of the neoliber.

The urgency of this new “critical university studies” was especially palpable in California, where the UC and CSU systems have only intensified their corporatism under continued funding cutbacks from the state.

Yet MLA11 was not all doom and gloom. The sessions I attended on the digital humanities were marked by an affectivity of vitality and growth, of optimism and new beginnings. A comparatively prosperous IT funding climate created a set of issues and concerns for DH scholars very different from the economic crisis so palpable elsewhere. Packed panels on the future of digital humanities or the role of social media in fostering public intellectuals were filled with laughter, hope, and a sense of empowerment coming partly from the growing investment of human and economic capital in digital humanities projects by university administrators and partly from the financial resources available to DH teachers, scholars, and developers from corporate, nonprofit, and governmental foundations. DH panels, too, addressed challenges produced by the changing climate in the humanities. Of most concern among DHers was the difficulty in getting departmental and university tenure committees to provide appropriate credit to digital work that does not end up as refereed articles or scholarly monographs and the lack of professional recognition for technical labor, which was too often performed by nontenure track members of the academic precariat. For the purposes of this roundtable I would characterize the problem of reforming criteria for tenure and promotion a “first world problem” and note instead the way in which the institutional structure of digital humanities threatens to intensify (both within DH itself and among the humanities more broadly) the proliferation of temporary, insecure labor that is rampant not only in the academy but throughout twenty-first-century capitalism.
Paradoxically, the key to this dual intensification of academic precarity is the very act that digital humanists often use to distinguish themselves from the traditional humanities: “making things.” At MLA11, DH panels devoted a good deal of energy to boundary drawing, which often depended on the distinction between making or producing things and critiquing them. In the panel on “The History and Future of Digital Humanities,” for example, I learned that I was not a digital humanist because I did not code (“Keeping a blog does not make you a digital humanist”) or because I did not “make things” (tell that to anyone who has labored for an hour or more over a single sentence). In the aftermath of MLA11, this invidious distinction between making things and merely critiquing them has come to be one of the generally accepted differences that marks DH off from the humanities in general. One could see the distinction at play in the brief Twitter exchange between HASTAC co-founder Cathy Davidson and Vectors founding editor Tara McPherson (see Figure 38.2). To McPherson’s boundary-drawing “I worry that much of theory/cult studies tends toward critique as an end in itself,” Davidson quickly replies: “Could not agree more. Critique hard. New ideas much harder. Making stuff work really, really hard!”

Put most starkly, academics on the left (which is pretty
much everyone doing theory and cultural studies) blame the crisis in the humanities on the corporatization of the academy and the neoliberal insistence that the value of higher education must be understood instrumentally in economic terms. Thus the shrinking of the tenured and tenure-track professoriate, which has resulted in the sharp growth of temporary and part-time labor in the academy, has been justified by university administrators and state legislatures in terms of bottom-line economics and the need for higher education to train students for jobs not to read literature or study culture. Consciously or not, McPherson and Davidson echo the instrumentalism of neoliberal administrators and politicians in devaluing critique (or by extension any other humanistic inquiry that does not make things) for being an end in itself as opposed to the more valuable and useful acts “of making stuff work.” But perhaps even more interestingly, as movements such as #transformDH have been articulating, it is the distinction between making things and doing more traditional scholarly work that perpetuates a class system within DH that generates an almost unbridgeable divide between those on the tenure-track, those in what have come to be called “alt-ac” positions, and those in even more precarious and temporary positions.

Sadly this pattern continues to reproduce itself in the current explosion of MOOC mania in print and online media, where much of the burgeoning interest in MOOCs has come from liberal administrators caught up in the convergence of neoliberal calculus and digital utopianism. At the same time that the market logic of neoliberalism has been used to decimate the mainstream humanities from within and without, this same logic has encouraged foundations, corporations, and university administrations to devote new resources to the digital humanities and to the development of MOOCs and other online forms of “content delivery.” If it is largely due to their instrumental or utilitarian value that university administrators, foundation officers, and government agencies are eager to fund DH projects, create DH undergraduate and graduate programs, and hire DH faculty, it is also the case that this neoliberal instrumentalism reproduces within the academy (both in traditional humanities and in digital humanities alike) the precaritization of labor that marks the dark side of information capitalism in the twenty-first century.
My remarks at the “Dark Side of the Digital Humanities” MLA roundtable on January 4, 2013, represent some preliminary thoughts and questions about games that I explore in greater detail in two essays that appeared in boundary 2 and Differences. My decision to include digital games in this conversation was not an attempt to claim the absolute centrality of games for the digital humanities. Additionally, my topic selection did not carry with it a necessary insistence upon a conflation between the “digital humanities” and “new media studies.” Since 2013, and for the foreseeable future, these disciplinary categories, and the boundaries between them, are porous. They continue to be debated and renegotiated by
For the purpose of the broad and inclusive conversation that Richard Grusin organized for MLA, I decided to work within a broad rubric of “Comparative Media Studies,” especially as it has been developed by N. Katherine Hayles in 2012 in *How We Think*. This inclusive category encourages conversations among scholars working in areas that include the materiality of print and digital productions (John Cayley, Matthew Kirschenbaum, and Jerome McGann); critical code studies (Wendy Chun, Matthew Fuller, and Lev Manovich); platform studies (Ian Bogost and Nick Montfort); technologically mediated forms of social interaction (Jodi Dean and Geert Lovink), information networks (Tiziana Terranova and Eugene Thacker) and electronic literature and digital art forms (N. Katherine Hayles, Henry Jenkins, Mark Marino, and Stephanie Strickland); the philosophical dimensions of digital media (Alexander Galloway, Richard Grusin, Mark Hansen, Friedrich Kittler, and McKenzie Wark); the cultural implications of digital technologies (Lisa Nakamura, Tara McPherson, and Rita Raley); the educational affordances of digital technologies (Cathy Davidson, Nichole Pinkard, and Katie Salen); and so on. This category also allows us to discuss a number of projects that include data mining, social network analysis, digital editions of print works, historical simulations, electronic literature, digital art, game design, and much more.

During our MLA roundtable, I was interested in producing a provocation and, briefly, introducing what is likely to remain one major problem of and for the digital humanities: the problem of games and gamification. The text that follows is meant as a starting point for a continued exchange. Perhaps, like the beginning of a game, it can be conceived as an invitation to play.

In recent years, games have touched practically every aspect of contemporary life. This certainly has something to do with a colossal video game industry that saw about $25 billion of revenue in 2011 in the United States alone with approximately 183 million American “active gamers” (that is, people who claim to play digital games an average of thirteen hours a week). Mobile gaming revenues rose to $1.2 billion in 2012 from $462 million just five years earlier. Even with some stagnation in U.S. console sales, global digital game markets have also seen significant growth.

The expanding centrality of games, however, has also in many ways exceeded the realm of “gamers” through what is often called “gamification.” Gamification, a term that derives from behavioral economics, refers to the use of game mechanics in traditionally nongame activities. This buzzword emerged only in the twenty-first century but has already found its way into writing on business, marketing, psychology, and design. We have seen the structure and logic of games creep into consumerism, crowdsourcing, and social media applications. For example, the *Chore Wars* website, whose celebratory tagline claims that “finally, you can claim experience
points for housework,” converts undesirable chores into a game complete with superheroic role-playing and points that spur competition among housemates. Nike+ shoes use sensors to transform a tedious running routine into a daily contest by tracking statistics, assigning achievement points, and allowing users to interface with cute avatars. TaskRabbit provides an online space for outsourcing minor jobs such as grocery delivery to other users while motivating contributors through a leaderboard and a statistics tracker that resembles a video game progress bar. Phylo, a game released by Jérôme Waldispühl’s team at McGill University, invites players to help researchers with a common problem in comparative genomics—Multiple Sequence Alignments—by participating in pattern recognition challenges. All of these sites and apps (of which there are many others) suggest that life in the early twenty-first century is becoming permeated by games. Especially throughout the overdeveloped world in which digital media, smartphones, and high-speed Internet access have achieved a ubiquitous status for many people, games have become an exemplary cultural form that serves as a prominent metaphor of success.

Gamification is increasingly becoming a problem of and, in some ways, a problem for the digital humanities. This is especially noticeable in the realm of education. Over the last two years, we have seen numerous instances of game-based learning, including how-to guides (Education Gamification Survival Kit) and charter schools with gameplay curricula (Katie Salen’s Quest to Learn and ChicagoQuest schools). Another ongoing initiative that has received a great deal of attention is the MacArthur Foundation’s “Badges for Lifelong Learning” that began as a Digital Media and Learning competition. Subsequently, the badges concept was adopted by organizations such as the Digital Youth Network: a Chicago-based “digital literacy program that creates opportunities for youth to engage in learning environments that span both in-school and out-of-school contexts.” The Digital Youth Network awards badges to youth who develop skills in technology, new media art, and social media participation. The gamelike impulse to collect badges serves as motivation for continued learning and produces a “visual portfolio of competencies” for participating youth and mentors.

Adopters of gamification across different fields, including education, regularly proclaim it to be an unparalleled organizational technique. One leading proponent, Jane McGonigal, suggests that “reality is broken” and can only be saved through games that turn “a real problem into a voluntary obstacle” and activate “genuine interest, curiosity, motivation, effort, and optimism” among their players (Reality Is Broken, 311). Alongside beaming support for gamification as a cutting-edge panacea, however, there has been some resistance to this concept and its widespread application. Curiously, much of the criticism has come from game designers. Gamification has been condemned in these circles for adopting only the
least artistic aspects of contemporary digital games—namely, their repetitive grinding and achievement-oriented operant conditioning. In a brief, polemical position paper published in *The Atlantic*, Ian Bogost contends that, above all, gamification is, in a philosophical sense, “bullshit.” Drawing from moral philosopher Harry Frankfurt, he explains that “bullshit is used to conceal, to impress or to coerce.” Gamification, for Bogost, engages in precisely this form of obfuscation insofar as it “takes games—a mysterious, magical, powerful medium that has captured the attention of millions of people—and makes them accessible in the context of contemporary business.” Condemning the rhetorical deceptiveness of the term, Bogost suggests the alternative term “exploitationware,” which decouples “gamification” from “games” (“Gamification is Bullshit”).

As one starting point to this roundtable discussion, I hope this brief introduction to what we might call the problematic of gamification will suffice. As teachers, researchers, and university administrators, we are bound to see many more instances of gamification in the coming years. Digital games will remain a major topic of both the digital humanities and new media studies. So they are worth discussing. My own visceral reaction to the phenomenon has often been one of skepticism—or at least critical reflectiveness. Game-based badges or experience points motivate people to perform repetitive tasks but not necessarily to engage closely with texts or to undertake projects at a more complex level. At the same time, I am also a game designer and a scholar of digital games. In 2011, I co-founded an organization called Game Changer Chicago (GCC) with Melissa Gilliam, a professor of Obstetrics, Gynecology, and Pediatrics and Chief of Family Planning at the University of Chicago. GCC uses digital storytelling and game-oriented methods to teach disadvantaged youth on the South Side of Chicago about sexual and reproductive health.5 We have focused on topics that include teen pregnancy, sexual violence, and socioeconomic health disparities. At GCC, our team produces interactive graphic novels, card games, and Alternate Reality Games projects with youth and for other youth to play. Through this new media production work and the research associated with it, I have found that when games are well designed, they entail many benefits. Such games offer players interactive contexts for thinking through and experimenting with complex problems in a hands-on fashion. Digital games enable multiple learning styles and engage players at several levels simultaneously through text, graphics, animation, audio, algorithms, haptic feedback, and different forms of interactivity. They spur decision making, enable role-playing, encourage play and discussion, and do many other things that exceed the addictiveness of point accumulation and victory that characterizes gamification.

So, then, despite the use of gamification for questionable ends (e.g., slot machines in Las Vegas), games are not, for me, a categorical evil but rather a rich problematic through which
we might think, feel, and process our historical present. For this reason, I include games under Richard Grusin’s heading of the ‘Dark Side of the Digital Humanities.’ I finish with three sets of questions that seek to navigate that darkness—a darkness that is, at different moments, terrifying and thrilling:

1. How should we think about games at a historical moment when gamification is arguably not merely a local phenomenon (for instance, in business, marketing, or education) but increasingly the form that economic and social reality takes in our world? Does it make sense to “game” an educational system that is founded on inequalities in a world that already uses games as a dominant metaphor and method?

2. Do the benefits of “badges” and other techniques of gamification outweigh their potential to operate as a reductive form of behaviorism? What are the benefits and limitations of incorporating badges into our pedagogy? Can we imagine (as many educators, theorists, and organizers are already attempting to do) badges that move beyond the superficial level of short-term behavioral modification? Can we instead create an infrastructure that builds a desire for lifelong learning and material skills into narratives, journeys, and games that youth (especially those youth coming from flailing or failing school systems) find compelling?

3. How might we imagine what are called “serious games” or “countergames” as complicating gamification? I am not necessarily advocating for either of these terms. However, along with scholar-designers such as Ian Bogost, Mary Flanagan, and Tracy Fullerton, I am committed to creating games that do not simply condition behaviors but encourage more complex forms of thought, speculation, practice, and action. For example, in 2012, along with my co-directors Katherine Hayles and Patrick LeMieux, I created an Alternate Reality Game called *Speculation* that explored the greed-driven culture of Wall Street investment banks and the 2008 economic crisis through a number of mini-games, collaborative narratives, and online forums.6 This game experimented with a design that was more speculative (in a number of senses) than didactic. This final question, then, is one that I ask myself on a weekly basis. Within a period of gamification, how might we think, play, and act critically through games?

Rita Raley

For “The Dark Side of Digital Humanities” (tweeted at #s07), we were charged with producing eight-minute statements designed to provoke a wide-ranging discussion of the unsaid,
understated, or under-theorized economic and political is-

tues that are associated with, attend upon, or otherwise follow

from the digital humanities as an institutional entity. In our

respective prefatory statements we noted that we had been

asked to provoke, but stimulate is closer to the thinking be-

hind the roundtable. The formulation of the title of the round-

table was itself a provocation, however, and an exemplary in-

stance of “behavioral priming,” to borrow a phrase from N.

Katherine Hayles’s paper delivered the following day. One

imagines that even the addition of a question mark in the pro-

gram copy might have produced a different affective response

in the audience, among which there still seems to be a fair bit

of indignation, at least insofar as one can glean the mood

from Twitter and blog postings. That the indignant audience

should now include many who were not even at the confer-

cence, much less at the session, can only confirm Teresa Bren-

nan’s thesis on the “transmission of affect”—it was not simply

biochemical response but also suggestion that produced the

(contagious) affects of #s07.7

The upset seems in part to derive from a misunderstanding

about our critical object: though our roundtable referred in

passing to existing projects, collectives, and games that we

take to be affirmative and inspiring, the “digital humanities”

under analysis was a discursive construction and, I should

add, clearly noted as such throughout. That audience mem-

bers should have professed not to recognize themselves in

our presentations is thus to my mind all to the good, even if it

somewhat misses the mark. Indeed I would say that human-

ists above all else need continually to work to perceive and

negotiate the institutional imaginary of informational tech-

nology so as not to fall into the trap of unconsciously adopt-

ing its optics. (My own cynicism about that institutional imag-

inary deepens with every administrative inquiry: I teach and

write about digital media, so clearly I should want to partici-

pate in working groups and pilot programs for online

education.)

Our topic is the dark side of the digital humanities. Not

quite the evil side, as Matthew Fuller and Andrew Goffey term

it, but, one hopes, not entirely unrelated. Evil media studies

pursues “practices of trickery, deception, and manipulation”—

one might even say tactics here—practices or tactics that en-

deavor “to escape [both] the order of critique” with all of its

melancholic negativity, as well as “the postulates of represen-

tation,” with their moralizing insistence on substance, es-

sence, truth. The dark side might on the face of it seem to

suggest precisely that “order of critique,” but our objective to-

day is not to diagnose so as to circumscribe and pronounce

upon the truth of things—not to uniformly fix what is after all

a diverse set of techniques and activities within a singular

frame and to seek out the hidden ideological core buried deep

within it; not then to bring to light “the” dark side of “the” digi-

tal humanities. But it is to suggest that there are critical blind

spots and assumptions that ought to be discussed before we
triumphantly embrace the notion that the digital humanities is the only game in town worth playing or, even, the only conference sessions worth attending, not simply the “next big thing” but the only thing. If, as sometimes seems to be the case, the digital humanities is the hill on which the humanities has chosen to stake its last claim for relevance, to fight its last battle for recognition, then we would do well to examine the field and identify not just the exploits but perhaps also the lines of escape.

This is not new thinking of course, and indeed the cultural politics of the digital humanities—its lacunae, protocols, and technocratic function—are central research problems for many of my colleagues in the Transcriptions Center at UC Santa Barbara. For example, two of our graduate students, Amanda Phillips and Anne Cong-Huyen, have been active in a #transformDH initiative that explores the intersections of the digital humanities and race, gender, and sexuality. And at the MLA convention in 2011, Alan Liu succinctly formulated the as-yet unanswered question that continues to serve as a critical challenge for all of us today: How, he asked, do “the digital humanities advance, channel, or resist the great postindustrial, neoliberal, corporatist, and globalist flows of information-cum-capital”? To answer the question of how the digital humanities “advance and channel” such flows, one simply needs to track monetary circulation and study the attendant promotional materials. In our current mercantile knowledge regime, with its rational calculus of academic value—seats occupied, publications counted, funds procured—the digital humanities are particularly well positioned to answer administrative and public demands to make knowledge useful: after all, research based on quantification is itself readily available to quantification. Cynically, in an institutional context in which a corporate administrative class is already mystified by humanities research that it cannot assess in terms of the amorphous metrics of “excellence” and “innovation,” one might say that the digital humanities are also particularly well positioned to exploit the expectation that we should be affectively awed by instrumentation (“oh my god, this lab, this application, is so cool”). In the “new world of brain-currency” shaped by engineers and economists, as Richard Hoggart once described it, it is the digital humanists who serve as cashiers, no longer ordinary schoolmasters peddling language as symbolic capital but academic service staff providing skills-based training—visual literacies, communicative competence, technological proficiency, data management—reinstantiating in the process the very categorical distinctions between theory and practice that DIY and maker culture have long sought to challenge.

Advancing and channeling the great flows of information-cum-capital requires a certain elasticity, more specifically, the capacity to become more agile so as to achieve operability and move to market more quickly. Agility is more easily attained without the practical and financial burdens of in-
frastructure; if networking, storage, and computing are automated, if they are virtualized, redundancy is eliminated and companies (universities, labs, centers) are left with legacy hardware that can only be repurposed as art and furniture. Why invest in servers, then, if Amazon, Microsoft, and Google can offer IT as a service? Contemporary doxa holds that treating infrastructure and platform as services makes it possible to free up resources for innovation and experimentation, for the symbolic work claimed as the particular province of the human: architecture and design. But accepting IT as a service also means accepting terms of use, and if the digital humanities has had very little to say about protocols of finance and governance, it has arguably had even less to say about the very protocols that govern our everyday use of university Gmail accounts (or indeed the whole of Google Education). As many have suggested but fewer have done, we ought to be marshaling the full critical, philosophical, and rhetorical resources at our disposal in order to think about all of the criteria that structure our communicative acts, from RFC standards and interface design, to privacy policies and terms of service.

The lesson one would like to think that the UC Office of the President had to learn, with its attempt to modernize its logo, is that interfaces and corporations alike have short life spans. Perhaps we too have to be jolted out of the cycle of innovating for the next grant cycle so that we might collaboratively speculate on a less-instrumental future for the humanities as a whole, one that brings into play the affordances of digital media but does so with a measured skepticism that might serve as a buffer against the irrational exuberance that too often characterizes the framing of our projects, initiatives, and entrepreneurial efforts.

To conclude, here are the questions I offer for discussion.

(1) Daniel Bell argued in *The Intellectual and the University* that the principle task of humanitas was to defend against the “increasingly powerful armory of intellectual techniques” (game theory, cybernetics, simulation) at the disposal of technocracy (Bell, 4–6). How are we now to regard the embrace of these very techniques, particularly when the actual work is outsourced to technical staff or when putatively interdisciplinary collaborations between humanists and computer scientists rely on a textbook division of labor? How, moreover, are we to regard the schism between high-end tool development as research and undergraduate pedagogy that maintains traditional disciplinary structures?

(2) What are the connections between the production of the aesthetic as techne in digital humanities research and contemporary courseware initiatives, and in what sense is each oriented toward technocratic knowledge production? What are the relations between new reading techniques (text mining, distant
(3) It is universally acknowledged that the digital humanities have made important contributions to traditional scholarship in literary studies, in particular introducing provocative questions about scale, multimodal scholarship, and changing reading and writing practices. Still one might ask why and how it is that it has come to function as the solution to every crisis of disciplinary legitimacy and every methodological impasse. For example, the project of symptomatic reading is said to be exhausted, thus necessitating the turn toward surface reading, of which "digital modes" of reading serve as the preferred instance (Best and Marcus). But we might also ask if is there a sense in which our institutions have been caught flat-footed by the forces of disruptive innovation and by the disaggregation of higher education: university education conceived as piecework is apportioned to tutors and lecturers; tutoring centers develop on the model of the call center; online study groups develop and gradually morph into online universities such as P2P.12 Can we then understand the exuberance that surrounds the digital humanities to be less of an attempt to shape a future than a salvific attempt to develop a sustainable organizational model for our profession that would include evaluative criteria and pedagogical practices particular to our current sociotechnological milieu? Are we still playing catch-up, and is the enthusiastic, transmedial promotion cover for our belatedness?13

(Administrator: you can have any faculty position you like, as long as it is digital.)

Notes

1 ***The sections in asterisks were either points implied in my visuals or in my 2013 MLA talk, which I have elaborated on in this written version. For an almost word-for-word transcription, see Alexis Lothian’s excellent notes: http://www.queergeektheory.org/2013/01/mla13-the-dark-side-of-digital-humanities/.

2. This chapter appears essentially in its original form, with minor revisions and additions that gesture toward related and future work. For the essays I mention in the text, see Jagoda, "Gamification," and Jagoda, "Gaming the Humanities."

3. For updated numbers, see, for instance, Newzoo, ‘2015 Global Games

5. Game Changer Chicago, http://gamechanger.uchicago.edu/.


7. Steven Pile succinctly outlines the spatial transfer of affect. See Pile, “Distant Feelings.”


12. “P2PU helps you navigate the wealth of open education materials that are out there, creates small groups of motivated learners, and supports the design and facilitation of courses.” See https://www.p2pu.org/en/.

13. Strenuous individual efforts aside, such as Katherine Hayles’s showcasing media studies at the MLA during her tenure as chair of the Division on Literary Criticism (Washington, D.C., 2005), it is, I hope, not controversial to suggest that the MLA as an organization was slow to make structural adjustments that would reflect the profound transformations in our medial environments and practices and that, from one angle, it is possible to read the exuberant embrace of social media platforms such as Twitter as compensatory.

Bibliography

Bell, Daniel. The Intellectual and the University. New York: City College, 1966.


Jagoda, Patrick. “Gamification and Other Forms of Play.” boundary 240, no. 2 (Summer 2013): 113–44.


API and Atom Feed

Data about this text is available via read-only JSON API endpoints: sentences, annotations, comments, and index keywords.

Comments posted on this text can be followed by subscribing to this text's Atom Feed.
I am on the side of the makers. I believe that the humanities can be a place not just to think about things, but to do things. Doing, when done right, can expand the scope of our critical activity, prepare our students for work in the world, and finally—and this despite the protestations of some—enact meaningful change in our communities (Fish). I write, then, being inspired by research at institutions such as the Critical Making Lab at University of Toronto, Concept Lab at UC Irvine, and metaLab at Harvard, along with many similar research centers that routinely engage with material culture as a matter of scholarly practice. In my courses as well, students create models, curate exhibitions, file patents, convene conferences, write grant applications, send letters to the Senate, draw, build, and code. However, the academy also presents some unique challenges to critical making of that sort, particularly when it comes to sustainable tool development. As tool makers, we should heed the lessons of the numerous forgotten projects that did not find an audience or failed to make an impact. For every line of code actively running Pandoc, NLTK, or Zotero, there are hundreds that lie fallow in disuse. Yet even in failure, this codebase can teach us something about the relationship between tools and methods.¹

In reflecting on my own failed projects, I have come to believe that with some notable exceptions, the university is an unfit place to develop “big” software. We are much better poised to remain agile, to tinker, and to experiment. The digital humanities (DH) can be understood as part of a wider “computational turn” affecting all major disciplines: see computational biology, computational linguistics, computational social science, computational chemistry, and so on. Computation in the humanities supplements the traditional research toolkit of a historian, a literary scholar, and a philosopher.² In this chapter, however, I would like to bring into question a specific mode of tool making, practiced within the digital humanities and without, of the kind that confuses tools with methods. The tools I have in mind prevent or—more perniciously—tacitly discourage critical engagement with methodology. To discern the problem with tools more clearly, imagine a group of astronomers using a telescope that reveals to them wondrous star constellations. Yet our hypothetical scientists cannot tell if these stars actually exist or whether they are merely an artifact of a faulty telescope. This has always been the tool-wielder’s dilemma. Contemporary research instrumentation in our field, from natural language processing to network analysis, involves complex mechanisms. Their inner workings often lie beyond the full comprehension of the
casual user. To use such tools well, we must, in some real sense, understand them better than the tool makers. At the very least, we should know them well enough to comprehend their biases and limitations.

The best kind of tools are therefore the ones that we make ourselves. After spending days wrangling a particularly messy corpus, I might write a script that automates data cleanup. My code may strip out extraneous HTML markup, for example. I could then release the script as a software library to help others who face the same task. With time, I might add a graphical user interface (GUI) or even build a website that makes using my scripts that much easier. Such small acts accelerate the research capabilities of the field as a whole. I would do nothing to discourage analogously altruistic sharing. But let us be sure that in using tools we also do not forget to master them from the inside out. What if my code implicitly mangles important metadata; or worse, what if it alters primary sources in an unexpected and tendentious ways? Let the tool makers make such biases explicit to the public.

Methods Within
Some tools encourage intellectual laziness by obscuring methodology. More often, it is not the tool but rather a mode of lazy thinking that is at fault. For example: the nltk.cluster module bundled in Python's Natural Language Toolkit (NLTK) framework (Bird, Klein, and Loper) contains an implementation of something called "k-means clustering," an unsupervised method of finding groups of similar documents within a large collection. The "unsupervised" part means that we are looking for hidden structure without making any assumptions about the documents at the outset (Na, Xumin, and Yohng). The documents may be grouped by the preponderance of personal pronouns or perhaps by sentence length. We do not know what elements the algorithm will identify, only that it will make piles "typical" of our corpus. The tricky part comes in estimating the number of expected document clusters (that is the $k$ variable). In a corpus of nineteenth-century novels, for example, one may expect a dozen or so clusters, which could perhaps correspond to novelistic genres. When clustering a large database of diplomatic communiques, one would reasonably expect more fine-grained "piles" of documents, which could have something to do with regional differences or with major political events. In either case, the algorithm will blindly return some groupings of distinctly related documents. But whatever the results of clustering, they are difficult to interpret in terms of meaningful literary-historical categories like "genre" or "period." Some of our piles will correspond to genres and periods, while others will seem meaningless. The algorithm produces nonhierarchical results—that is, the output is not ordered according to value or significance. As the algorithm is also nondeterministic, meaning that it will perform differently each time it is run, the group-
ings will also vary with each iteration. To complicate matters, NLTK implements other clustering algorithms, like expectation–maximization (E-M) and group average agglomerative clustering (GAAC). These methods will likely chance upon yet other hidden relations between documents and other ways of organizing the material into piles. The algorithm will always return a result, according to some set of formal commonalities. But what these results mean and why they matter is open to interpretation. To make the clusters meaningful requires a deep understanding of the underlying logic.

NLTK facilitates such discovery by distributing detailed documentation along with the code. The documentation does more than just describe the code: it reveals implicit assumptions, citing external sources throughout. In experimenting with NLTK, I was able to get some output from the clustering methods in a matter of days. It took me months to understand what they could mean and how they could be applicable to my research. Just applying the tool or even “learning to code” alone was therefore insufficient for making sense of the results. What could help me, then, and what is only now beginning to surface in DH literature is a critical conversation about methodology.

Unlike some other tools of its kind, NLTK is particularly good at revealing its methods. Its codebase is open to inspection; it is easy to read; and it contains much commentary along with links to related research. The NLTK project began in 2001, at the University of Pennsylvania, in a collaboration between a linguist and his student (Loper and Bird). Research based on the module started appearing in print several years later, around 2004. NLTK reached version 1.0 eight years after its inception, in 2009. In the intervening time, immense care must have went into the critical apparatus that ships with the tool. And I suspect that at this late stage of the project, more hours have gone into the writing of its documentation than into the crafting of its code. As of 2015, the NLTK GitHub page lists no fewer than 130 contributors.

Reflecting on the history of NLTK gives us a glimpse into the realities of responsible academic making. Not every project will need to go through such a long development cycle or include such detailed documentation. But even my own small collection of data cleaning scripts would need substantial work to reach the level of polish required for empowered use of the kind NLTK enables. Note also that NLTK itself is only a “wrapper” around a set of statistical methods for the analysis of natural language. That layer of encapsulation already poses a number of problems for the researcher. Using NLTK responsibly demands a degree of statistical literacy along with programming experience. The cited methodology often contains a mixture of code and mathematical formula. Yet higher-level encapsulations of NLTK, like a web-based topic modeler, for example, would further remove the user from that implicit logic. Each level of abstraction in the movement from statistical methods, to Python code, to graphical
user interface introduces its own set of assumptions, compromises, and complications. Any “ease of use” gained in simplifying the instrument comes at the expense of added and hidden complexity.

Figure 9.1. Layers of encapsulation.

Hidden complexity puts the wielder of the tool in danger of resembling a hapless astronomer. To avoid receiving wondrous pictures from broken telescopes, in the way of actual astronomers, we must learn to disassemble our instruments and to gain access to their innermost meaning-making apparatus. Any attempt to further repackage or to simplify the tool can only add another layer of obfuscation.

It follows, then, that without a critical discussion about implicit methods, out-of-the-box tool use is best treated with a measure of suspicion. The makers of out-of-the-box tools should similarly weigh the altruistic desire to make research easier against the potential side effects that come with increased complexity. The tool can only serve as a vehicle for
methodology. Insight resides in the logic within. When exposed, methodology becomes subject to debate and improvement. Tools proliferate and decline in quality relative to the researcher’s experience. If tomorrow’s scholars move from Python to Haskell, the effort of learning the underlying algorithms is what will transfer with the language. Methodology is what remains long after the tools pass into obsolescence.

Unplanned Obsolescence
In addition to methodological concerns, tool making also involves pragmatic considerations about sustainability. Software is cheap and fun to build by contrast to the expense and drudgery of its maintenance. “Ninety percent of coding is debugging. The other 10 percent is writing bugs.”4 The aphorism comes naturally to program managers and software engineers who have gone through the full software product development cycle. In the excitement of building new tools, it is however easy to underestimate the challenges of long-term application maintenance. Academic attention spans are naturally cyclical: articles are published, interest wanes, funding dries up, students graduate. Scholars start anew each year and each semester. Software support requires the continuity of care and much more of it as a codebase matures. Standards change, dependencies break, platforms decay, users have questions. The case for the humanities as a laboratory for innovation is strong, but I doubt that many are prepared to make “critical customer support” a part of their research agenda.

Software development requires immense resources, as digital humanists from George Mason and the University of Virginia will tell you. Smaller teams should think twice before investing time and money into tool development. Not every method needs to be packaged into a tool. Some projects would be better off contributing to existing efforts or using their resources to encourage methodological literacy. In fact, if you build it, they might not come at all. Start-ups know that beyond the initial excitement of a product launch, the challenge of any new application lies in the acquisition and the retention of users, no matter how “disruptive” or “innovative” the technology.

A few years ago, I spent some time working with a talented French developer to design a collaborative translation platform. Despite his skills and dedication to the project, the tool did not gain significant traction among language teachers, translators, or students. I learned then that no amount of innovative engineering or beautiful web design could guarantee participation. Neither of us had the time nor the resources to advocate for the service. Advocacy would require arranging for training, outreach, fundraising, and support: services we could not provide in addition to our professional obligations. It was however tempting to think that social and institutional change could ride on the coat tails of software alone. If we build it right, the two of us thought, we could transform the
practice of translation in the classroom. Yet we failed to consider the difficulty of implementing that vision into practice. We built the tool but not the community around it. The classroom environment resisted change, and for a good reason. Upon reflection, we saw that language teaching was grounded in proven, if sometimes imperfect, practices. Our platform development should have considered the strengths of that tradition and not just its weaknesses. Before rushing to innovate, we could have started with smaller classroom experiments to test our intuitions. We could have arranged for interviews, focus groups, and pilot studies. To give you a sense of our miscalculation, consider Duolingo, a similar (and earlier) effort led by researchers from Carnegie Mellon University, which amassed more than four million dollars of investment from the National Science Foundation and Union Square Ventures before bringing their service to the public. In retrospect, it was hubris to attempt platform building without similar commitments.

Consider also the following in the case of our hypothetical “wrapper” around NLTK—the one that would simplify the use of natural language processing for the nontechnical audience. Every contemporary Mac and Linux laptop machine comes prepackaged with powerful command-line tools for text manipulation: software utilities like wc, sort, and uniq. When chained together, these simple instruments are used to count and sort words in a document or to generate a term-frequency distribution useful for formal text analysis. They are free, simple to learn, versatile, and require no additional installation. They come with their own textbook, accessible from the terminal. Yet most of my students, even at the intermediate level, remain unaware of such tools already at hand. Many were not exposed to the basics of file paths, networking, or operating systems. How can one better facilitate the practice of computational text analysis without closing the digital literacy gap that separates mere users from empowered tinkerers and tool makers? A proposal to implement yet another tool duplicating the functionality of ubiquitous native utilities gives me pause. We must first reflect on the reasons as to why there was no adoption in the first place. That is not to say that existing word-frequency tools cannot be refined in some way. But, any new project that hopes to innovate would have to at least match the power of the existing instrumentation and then improve on it in some palpable way. And even then, our hypothetical project would face the same barriers to literacy and adoption as the original toolkit. These would have to be addressed before writing a single line of code.

Furthermore, whatever adoption the new alternative might achieve risks fracturing the existing user base, already limited to a small number of practitioners. By analogy, a new publishing platform that hopes to uniformly “disrupt” academic publishing is far more likely to enter an already fragmented market rife with good alternatives that are struggling to survive. The fragmentation prevents any one of them from gaining criti-
Fig. 4.1 Instrumental efficacy alone therefore cannot address the lack of adoption. For example, legacy platforms like Microsoft Word or clunky journal management systems (used behind the scenes for peer review) do not account for the range of "planned obsolescence" problems in academic publishing that Kathleen Fitzpatrick identified in her recent book on the subject. The tool comprises but a small part of a much larger publishing ecosystem. It can act as a wedge that initiates change, but not without a larger communal effort to address the way we read, write, and do research. The world does not suffer from a lack of better text editors, for example. Rather, the adoption of powerful free and open source software is stymied by insufficient training, institutional momentum, and the lack of intellectual buy-in. Rather than fracturing the community, by creating another text editor for example, we would often do better to join forces: to congeal our efforts around common standards and best practices. Unfortunately for us, funding agencies favor promises of bold innovation where it would be more prudent to invest into organic growth. The effort to shift the habitus of a community, as Pierre Bourdieu would describe it, involves a delicate balance between disruption and continuance. Much can be learned from the success of the open-source and free culture movements in this regard (Weber). Take, for example, the story of Wikipedia and MediaWiki. MediaWiki, the software platform powering Wikipedia, was neither the first nor the most technically sophisticated wiki software package. But in the hands of Wikipedians, MediaWiki became a tool capable of transforming the contemporary information landscape. Despite some of its problems, Wikipedia struck the right balance between traditional forms of knowledge-making such as the encyclopedia and innovative editorial structures such as commons-based peer production. Wikipedia the community inspires me more than MediaWiki the tool. In the Wikipedia world, the platform is secondary to community development.

The care of academic research communities, of the kind that encourages empowered tool use, happens in departments and through professional organizations. Programs like the Digital Humanities Summer Institute answer the need for training necessary to do research in our rapidly developing field. However, more resources are needed to initiate methodological and not just instrumental innovation. Few humanities-based alternatives exist to institutional structures in other fields like the Society for Political Methodology and the International Association of Legal Methodology; journals like Sociological Methods & Research, Journal of Mixed Methods Research, and International Journal of Qualitative Methods; prizes and funding opportunities like the Political Methodology Career Achievement and Emerging Scholars Awards, or the Program for Promoting Methodological Innovation in Humanities and Social Sciences administered by the Japan Society for the Promotion of Science. To sharpen our tools we must similarly prioritize methodological development. Only
then can we build platforms that answer to the values of humanistic critical inquiry.

A shared concern with data and computation has brought a number of disciplines closer together. Biologists, linguists, economists, and sociologists increasingly integrate their methodologies, as evidenced by a vigorous cross-disciplinary publishing record. DH is primed to join that conversation, but only if its methods develop without abridgment. Tools are great when they save time, but not when they shield us from the complexity of thought. Working as a digital humanist or a new media scholar means taking on extra responsibilities: to do well by history when writing history, to do good science when doing science, and to engineer things that last when making things.

Notes

1. William Pannapacker has written eloquently on the topic in the Chronicle of Higher Education. See “Pannapacker from MLA: The Success of ‘Failure.’”

2. I do not mean to imply that DH can be reduced to computation. See Ramsay and Rockwell, “Developing Things,” and also Elliott, MacDougall, and Turkel, “New Old Things.”

3. Astronomers also use k-means clustering to identify star constellations. See also MacQueen, “Some Methods for Classification and Analysis of Multivariate Observations.”

4. The quote is commonly attributed to Bram Cohen, the creator of BitTorrent, posted on Twitter in 2011. There are however numerous earlier instances of the exact quote, itself a variation of Sturgeon’s Law coined by Theodore Sturgeon (the American science fiction writer) in a 1957 article for Venture magazine and cited as such in the Oxford English Dictionary.

5. If you are behind one of these machines now, search for your terminal application using Spotlight and type man wc in the prompt (q to exit). For mere examples, see https://github.com/xpmethod/dhnotes/blob/master/command-line/109-text.md.

6. For more on the influence of Wikipedia, see Collier and Bear; and Callahan and Herring. It is a point made by Benjamin Mako Hill in his Almost Wikipedia. Another good summary comes from Garber, “The Contribution Conundrum.”

Bibliography


Elliott, D., R. MacDougall, and W. J. Turkel. “New Old Things: Fabrication,


**API and Atom Feed**

Data about this text is available via read-only JSON API endpoints: sentences, annotations, comments, and index keywords.

Comments posted on this text can be followed by subscribing to this text's Atom Feed.
Where Is Methodology in Digital Humanities?

TANYA E. CLEMENT

For the very idea of “practices” has a satisfyingly concrete ring to it, if no longer through that classical (and singular) opposition to “theory,” as something that (in the plural) installs us immediately in the interstices of effective social power, in its minutest details.

—Evan Watkins, Work Time: English Departments and the Circulation of Cultural Value

In a 2009 Digital Humanities Quarterly piece, Christine Borgman asks, “Where are the social studies of digital humanities?” Suggesting that ethnographic and other social studies of scientific information work have significantly shaped how scholars have come to understand scholarly cyberinfrastructure in the sciences, she argues that the practices of digital humanists should be similarly studied. And while most digital humanists do not employ the qualitative methods of data gathering to which Borgman refers, such as survey research, observations, and interviews, perhaps the absence of these methods indicates that DH it is still not clear where such methods might fit within the epistemological landscape of the humanities. After all, social scientists have long studied—and often directly impacted—scholarly information system development in the sciences using these familiar methods. By contrast, digital humanists are using methods that are largely new to the humanities, and perhaps for this reason, we are less adept at expressing how these forms of study map to our theoretical concerns.

How we validate and share knowledge within and between epistemological frameworks, whether it is the humanities or the social sciences, has much to do with how we articulate the link between our methods and our theories. Situating seemingly positivist social science methods within a humanist framework is about more than the interpretive methods we might employ. Likewise, collaborating with social scientists or impacting social science studies, which are also often shaped by cultural studies, critical race theory, feminist inquiry, or postcolonialism, requires more than staking a claim to constructivist theories. In ethnographic studies such as those Borgman cites, an articulation of methodology helps the researcher describe a systematic approach to fieldwork and data analysis methods, one that ultimately facilitates a deeper engagement with theory. This chapter therefore aims to distill a range of methodological perspectives employed in the study of information systems in the social sciences and digital humanities, with the goal of suggesting bridges not only between the social sciences and digital humanities but between
I begin with a brief review of studies on information work in the social sciences and the digital humanities in order to situate these at once disparate and interconnected discourse communities. Then, by focusing on the methodological perspectives that underpin crucial projects in each field, I show how each understands the links between their methods and their theories. This attention allows us to understand the theoretical implications of adopting methods that are more familiar to the social sciences in DH as well as to identify the theoretical implications of digital humanities methods for the humanities writ large.

Establishing Social Science Methodology in Information Work

Information systems and the people who work with them have long been studied by social scientists. Michael Buckland, among the most well-known theorists of information, identifies three meanings of the term: information as knowledge, information as thing, and information as process. For Buckland, information-as-knowledge is intangible since it is based on personal, subjective, and conceptual understandings. Information-as-thing has materiality, however, since to communicate these understandings they have “to be expressed, described, or represented in some physical way, as a signal, text, or communication” within a system (2). The systems that these information workers typically work with include computers and networking technologies, whether they are standalone or embedded in a larger system, but information-as-thing allows the social scientist to isolate information objects associated with specific forms of information work from the systems that engage those objects (whether that object is digital or not) as culturally informed processes—for example, the sculptures and exhibits of a museum or gallery; the books, documents, and taxonomies of a library; the DNA or microscope of a scientific lab; the code, bits, and bytes of computer programmers; the numbers and graphs of data analysts, and so on. An investigation of how workers interact with and through these information objects and systems through social-scientific means often yields insights about how information is understood in those fields.

At the same time, much social science information work scholarship is undergirded by a desire for a better understanding of information-as-process—what Foucault would call “an archaeology of knowledge”—or the systems of power and influence that shape information systems and therefore knowledge production, identity construction, and intersubjectivity. Social science researchers work with disempowered communities to better understand systems of power and resistance in the modern metropolis (Burawoy); study the relations of employment and the role of the worker in the constitution of a worker’s identity (Orr); and study interpretive flexibility and
human agency in information technology development and use in large, multinational software consulting firms (Orlikowski). Embedding themselves, sometimes as workers in these communities, these researchers attempt to foreground their own research practices in their studies of others. Of particular interest to digital humanists, Christopher Kelty’s *Two Bits: The Cultural Significance of Free Software* and Matthew Hull’s *Government of Paper: The Materiality of Bureaucracy in Urban Pakistan* are ethnographic studies in information work of which the objects of study are text and technology; their subjects concern the creation, dissemination, and authorization of knowledge; their goals are to explore “recursive publics” and the political economies of paper, respectively. Other social scientists, such as those to whom Borgman refers, also consider how the universal claims of science are localized as a result of unspoken, ontological, epistemological, and practical diversities in the day-to-day worlds of scientific labs (Knorr-Cetina; Sommerlund; Voskuhl). These studies share a theoretical kinship with humanistic studies concerning information and knowledge production, even as they employ divergent methods.

Qualitative social scientific methods in information and knowledge work typically focus on the direct observation of practices rather than on reviews of theories or findings (Geertz “Thick Description’). Social scientists employ such methods through standardized procedures, which are considered essential for conducting a qualitative study that yields valid and rigorous scholarship. Howard Becker’s seminal 1996 essay, “The Epistemology of Qualitative Research,” maintains that scholars who employ quantitative methods justify their results by proving that their data is reliable (i.e., repeatable), but those who rely on qualitative methods are more concerned with showing that their data is accurate and precise (or based on close observation) and broad (based on a wide range of variables). In other words, researchers who engage in qualitative methods are especially conscientious about their methods producing accurate data—as close to objective as one can get—so that their results can be considered valid by their peers, who typically find quantitative methods more exact (Becker).

In such studies, an articulation of methodology helps researchers reinforce the systematic nature of their chosen approach. As I will argue, this is an act that ultimately facilitates a deeper engagement with theory. Consider that social scientists view technique as a particular and situated way of applying a method or systematic approach, and methodology as the reciprocal relationship between method and theory (Katz; Burawoy, 271). Data produced outside of a theoretical framework is considered merely “sociological aestheticism” (Geertz, “Thick Description’), simply description (Katz), or at worst, “haphazard” and “fortuitous” (Snow, Morill, and Anderson, 184). In the dominant model of social science scholarship, therefore, the researcher includes an explicit statement about which theories he or she is engaging within any initial ac-
count of his or her methodological conditions (Snow, Morill, and Anderson, 194). This theory is described in the context of the study, including how the theory was formed and how the particular setting of the study compares to that formation, as well as any co-related historical factors such as how the theory has been used in the past or how it is discussed in current scholarship (195). This form of explicit engagement with theory is crucial in social science epistemologies that warrant a degree of scientific objectivity.

Establishing Digital Humanities Methodology in Information Work

Digital humanists also operate on information-as-thing (the word, the line of poetry, the page, the piece of code) as well as understand information as information-as-process—the continual state of becoming informed through understandings that are constantly shifting. The study of information work in digital humanities and of digital humanists as information workers has been primarily concerned with the cultural contexts that influence these shifting understandings—what we might also describe as knowledge production—especially as these contexts relate to the academy. Johanna Drucker, Kathleen Fitzpatrick, Julia Flanders, John Unsworth, and many others have written about the processes of scholarly knowledge production in changing publication and employment practices, for example. Anne Balsamo, N. Katherine Hayles, Alan Liu, and Patrik Svensson have each published extended studies on the future of researchers and educators in the humanities and arts—a future in which, as a result of a postindustrial value system, literature and the arts are increasingly undervalued. These examples demonstrate the overarching concern of digital humanists with the cultural contexts of academe and how the changing nature of information work will alter the future of the humanities.

What is needed to connect this body of work to the discourse of information work in the social sciences is a more precise articulation of how digital humanists themselves function as information workers. In fact, Liu provides a taxonomy of the knowledge worker that is particularly helpful in understanding digital humanists as information workers:

Knowledge workers =

Academic intellectuals +
(technical + professional + managerial) intelligentsia +
trailing edge clerical workers (Liu, Laws of Cool, 392)

Broadly speaking, digital humanists are typically academic intellectuals and “intelligentsia,” and their information work might include research, writing, and publication, project conceptualization, teaching, and service to the institution and the field at large. Doing this work within the context of information technologies can include (among a variety of activities), algorithm development and implementation; coding and encoding; e-mailing; data curation, management, and analysis;
generating and analyzing statistics; social networking; meeting virtually; note-taking and publication; user experience and user testing; as well as database, software, web, and visualization development and interpretation. Noting that his definition of the knowledge worker is a class-based concept, Liu reminds us that information work is a process of knowledge production that is embedded in culture. As such, this list is not exhaustive; it does not include the social infrastructure development work that supports these activities with technology such as curriculum development, fund-raising, networking, staffing, and general project management, but it makes the point well: the digital humanist can be defined as a knowledge worker and an information worker.

And yet there is a prevailing notion in the social sciences that humanists are not trained to study themselves as information workers. For instance, the Information Work Research Group in the School of Information at the University of Texas at Austin, of which I am a part, trains graduate students in social scientific methods—that is, direct, in situ observations and semistructured interviews—out of the sense that such methods provide an “essential means for understanding what information professionals actually do at work, why they do it, and how they do it,” digital humanists included. Borgman attributes the differences between humanists’ “fluid” methods and more exacting social scientific methods to the fact that humanities scholars tend to learn methods that are specific to content areas and through independent research rather than as a “common substrate of research methods courses and practices that span the social sciences” (Borgman, Big Data, Little Data, No Data, 164). The implication is that a level of scientific objectivity is lacking in the humanist’s more subjective approach to studying knowledge production and information work.

Indeed, some digital humanists openly discount the need for the objective stance that is central to social science research. For example, Liu calls his study a “census, a propaedeutic, an introduction” in lieu of “a more scientific study,” and likens his reading of “cool” websites to the act of “thrusting one’s hand into the water to see if it is cold, hot, or lukewarm” (Laws of Cool, 183; 233). Similarly, in her survey of social science studies on issues of attention in the digital age, Hayles notes that “few scholars in the humanities have the time—or the expertise—to backtrack through cited studies and evaluate them for correctness and replicability” (How We Think, 68). Arguing that “perhaps our most valuable yardstick for evaluating these results . . . is our own experience,” Hayles simultaneously promotes a de-siloing of knowledge work, maintaining that “the scientific research is valuable and should not be ignored” (68). To be sure, these examples leverage experiential knowledge with other interpretive data-gathering methods, such as archival work and close reading. But in their nods to the value of social science research, Liu, Hayles, and other digital humanists implicitly deny their own
need for a systematic repeatable method and at times openly disavow a desire to be "scientific" at all.6

Certainly, digital humanists are deeply concerned with employing accurate and variable as well as systematic methods in their studies. John Unsworth, for instance, calls methods “primitives” and lists activities such as discovering, annotating, comparing, illustrating, referring, representing, and sampling as among the methods employed in humanities knowledge work. The European Union’s Digital Research Infrastructure for the Arts and Humanities initiative has gone so far as to create a Taxonomy of Digital Research Activities in the Humanities (TaDiRAH), which includes three broad categories: research activities, research objects, and research techniques. Research techniques include a long list of topic areas about which we are used to seeing debates in DH, including concordance-building, crowdsourcing, encoding, gamification, topic modeling, and versioning. In this case, research activities, which includes capture, creation, enrichment, analysis, interpretation, storage, dissemination, and meta-activities (such as assessing, community building, project management, and teaching and learning), in effect extend Unsworth’s list of primitives to add methods typically employed in a digital context.7 There is indeed a broad range of topics and methods in which digital humanists engage as we pursue our inquiries into information-as-process.

It is telling, then, that most critiques of DH—both from the social sciences and the humanities—do not point to a lack of accuracy, variability, or other limitation of method. Rather, most critiques of DH point to a decoupling of method from the theoretical perspectives that would ordinarily help situate the kind of intellectual effort being engaged (Drucker; Hall; Liu, “Where Is Cultural Criticism in the Digital Humanities”; McPherson, “Why Are the Digital Humanities So White”?). To frame the issue another way, consider how reductive it would seem to describe the mere presence of the techniques and methods as doing digital humanities. It would be like saying that doing ethnography simply entails establishing relationships, watching people, transcribing interviews, and keeping a diary (Geertz, “Thick Description”). Digital humanities research must include enough detail or “evidence” to form accurate and convincing accounts, and accounts are much more accurate when they reflect as broad a spectrum of perspectives as possible. These imperatives can be achieved—to a varying degree—through many types of methods, but situating digital humanities within a humanist epistemological framework must also entail an explicit articulation of our methodological perspectives, or how our techniques are tied to theory. Digital humanities scholarship that does not engage with theory risks being perceived as unconcerned with interpretive, situated, and subjective knowledge production, and therefore displaced from the epistemic culture of the humanities. Articulating our methodology, moreover, gives us an opportunity to explain why we do what we do, which in turn allows us to argue for
the specific contributions of our findings to ourselves, to other humanists, to those possible collaborators in other disciplines who rely on methodology as a signpost, and to the world.

Methodological Perspectives in DH Information Work

In contrast to social science scholarship on information work, digital humanities studies of information work often lack methodological discussions—even while methodological perspectives, as I term them, are always at play. Methodological perspectives are akin to what Sandra Harding has called “methodological features” in feminist social science inquiry—the empirical and theoretical resources, intellectual rationales, and relations between subject and object of study—that, once identified, help a researcher to pinpoint how they are applying “the general structure of scientific theory to research on women and gender” (Harding, *Feminism and Methodology*, 9). In the context of DH, using the term methodological perspectives rather than methodological features underscores the fact that even so-called features are subjective and influenced by one's situated epistemic culture. In digital humanities, we reflect our methodological perspectives when we choose to study certain texts (or certain things as texts) or when we discuss why methods are best employed through certain techniques—why deformance in Adobe Photoshop can advance our thinking about the hermeneutics of visual art (Samuels and McGann), for example; or how algorithmic thinking with word frequencies might advance our thinking about how gender roles play out in *King Lear* (Ramsay, “Reconceiving Text Analysis”); or how surface reading with social network analysis allows us to articulate the archival silences in Thomas Jefferson’s archive (Klein). These and other examples show us methodological perspectives (a “thinking through” of the link between techniques, methods, and theories) already present in literary scholarship in digital humanities. More often than not, however, these ties are implicit—the means to the end is either foregrounded in exclusion of a productive critique, or the research strategy is subsumed by a finding that the researcher argues has been “discovered” rather than constructed.

But methodological perspectives in the humanities and social sciences can overlap, which means that digital humanists must be even more diligent about articulating how our perspectives are situated within a humanist epistemological framework. That is, new methods in digital humanities such as statistical analysis, visualization, or ethnography do not exist in a vacuum. If digital humanists choose to employ methods that are more common to the social sciences, we must understand the relevant articulation work that surrounds similar kinds of methods in the social sciences. Or consider the reverse: theories in feminist inquiry, postcolonial studies, and activism have never been solely humanistic research perspec-
tives; scholars from outside the humanities also school themselves in those histories before applying their methods. In other words, theories that engage self-reflexivity as a methodological perspective are essential to scholarship in the humanities. But reflexive awareness is also at the foundation of methodological perspectives in qualitative social science research—research that has been deeply influenced by constructivist paradigms. There is a general understanding in the social sciences—the same that should be reflected in digital humanities work—that methodological perspectives, as driven by the historical, present, and perceived future context of a project, shape and are shaped by practical and theoretical concerns.

A reflexive understanding of knowledge production and information-as-process is a significant research perspective that translates across information studies in the social sciences and digital humanities. Examining similar methodological perspectives in information work research in the social sciences—namely, authority creation, hermeneutics, and becoming answerable—informs how we might apply humanist theories to research on information-as-process in the digital humanities. Comparing these perspectives from the viewpoint of social science and digital humanities studies not only allows us to develop a better understanding of methodological perspectives in general, but also offers specific examples of how these similar perspectives play out in different epistemic cultures. Though shared across disciplines, these perspectives ultimately reflect unique epistemes through their employment of technique and method in individual studies. Learning to express these differences is vital not only for digital humanists who seek to situate their work in conversation with social science research, but also for those who seek to situate themselves in common as humanists.

Notes

1. Scholars in science and technology studies (STS) also rely on organization and management scholarship in workplace studies in which researchers study workers including consultants, doctors, or engineers in the context of particular organizations (Garcia et al.) as well as the information work practices of traditional information institutions including libraries, archives, and museums that collect, preserve, interpret, and disseminate various kinds of information (Marty; LeMaistre et al.).

2. This kind of fieldwork is exemplified in the work of scholars such as Daniela Rosner, who considers “object bias” in Computer-Supported Cooperative Work (CSCW) studies. From a reflexive position, Rosner focuses on the social and functional role of artifacts as objects that are spatial and temporal flows with emergent compositional elements and constituent surfaces.

3. Other examples might include Latour and Woolgar’s ethnography of the scientific laboratory life (1986); Bowker and Star’s ethnography of infrastructure and metadata (2000); Suchman’s ethnography of technology design and production and consumption and use (2006); and Geiger and Ribes’s trace ethnography of log data (2011).

This definition is from a grant proposal narrative that is not publicly available.

This stance is unlike social scientists such as Donna Haraway and Sandra Harding, who use a feminist epistemology to critique objectivist stances and to claim that valid scientific methods require that we claim our situatedness on what we are able to see (Harding, *Whose Science? Whose Knowledge?*, 106; Haraway, 583).

Other good examples include Hayles’s coverage of different modes of reading (57–68) and her work close reading telegraph manuals alongside electronic and paper-based postmodern literature, Liu’s survey across business and management literature (*Laws of Cool*, 76–175) and his close reading of ‘cool’ websites from the 1990s (*Laws of Cool*, 176–285), and the Stanford Literary Lab Pamphlets, which focus on texts that range hundreds of years and techniques that include topic modeling, sentiment analysis, and visualization techniques.

That epistemology has bearing on the methodological discussion is not lost on Harding. In *Feminism and Methodology*, she notes that her recommendation for employing these three methodological features is meant to counteract “traditional” theories of knowledge. “Traditional” methodological features in her treatment of the topic, however, are those that are common to social science investigations. I am consciously situating this conversation squarely in the humanities episteme by choosing the alternative term “methodological perspective.”

Harpham defines the objective of humanities scholarship as ‘self-understanding’ (23) and Gilman calls this reflexivity “the self-conscious awareness of the methodological approaches that one uses” (384). Menand also identifies reflexive processes as the defining feature of humanistic knowledge by arguing that in “developing tools for understanding ourselves” and “everything in the world of values”—what Latour calls ‘states of affairs’ (232)—humanists are instantiating “the fact of situatedness” that ultimately leads to a necessary skepticism to objectivity and positivism (Menand, 15). For Stanley Fish likewise, this skepticism means that humanists understand the constructed nature of “the cultural systems within which we live and move and have our beings” as “the given” or “normative” (377). This investment in the constructed nature of knowledge production in all fields means methodological perspectives in the humanities must reflect an understanding that knowledge is “knowing, observer dependent, emergent, and process-driven rather than entity-defined” (Drucker, 87).

In “the Chicago School,” originating at the University of Chicago in the 1940s and 1950s, ethnography was considered an empirical, scientific study and ethnographic writings typically began to include methodology sections (and books on methodology and epistemology) that shied away from more subjective (historical research) or objective (statistical techniques) methods. Valid data was achieved through in situ observation that provided insight into the participant’s interpretations of events. Such insights were often the result of gaining the confidence of and having and showing empathy for subjects. As in literary study, ethnographers made a postmodern, postcolonialist turn in the 1980s and 1990s that is reflected in books that challenged the efficacy (and ethics) of ethnography as an objective or impartial look at the world and insisted, instead, on the potential of fieldwork that produced subjective, contingent, and situated knowledges (Adler).

Becker’s response is to assure his audience that qualitative methods, while not objective in the same way as quantitative methods, are still valid
in the context of older and new methodological traditions. Harpham's remarks in "Beneath and beyond the 'Crisis in the Humanities'" are a response to what McGann calls a general "malaise" that has had humanities scholarship and education in a "holy mess for some time" (McGann, 72) and are also meant to assure his audience that humanist inquiries are still valid—in that they are still relevant—within the context of global economies and advanced information technologies.

For a social science researcher like Anne Beaulieu, that has meant new means for studying the information work of women's studies scholars through online forums and listservs ("Mediating Ethnography"). Through a methodological perspective concerned with authority creation and gendered information work, Beaulieu engages feminist theories that are concerned with making previously unseen work observable and therefore discoverable as new subjects for study.

#transformDH, http://transformdh.org/

Postcolonial Digital Humanities, http://dhpoco.org/

One concern, as Drucker points out, is that the "cultural authority" of visualization, publishing, and design technologies are still claimed by fields in the sciences and technology (85).

In direct response to topic modeling and visualization techniques, Liu has called this "the meaning problem in the Digital Humanities" based on projects that perform "tabula rasa interpretation" or "the initiation of interpretation through the hypothesis- free discovery of phenomena" ("The Meaning of the Digital Humanities," 414).

Chun and Balsamo are excellent examples of this kind of methodological perspective at work in digital humanities. Blanchette gives an excellent overview of critical making in particular from the social science perspective.

Posner has made this call in her keynote at the Keystone Digital Humanities Conference, which she has published on her blog. In The Laws of Cool, Liu warns that in the age of "millennial knowledge" or "knowledge that is antihistorical (anti-obsolescent) . . . . The centrality of the challenge to academic knowledge thus stands starkly revealed: [corporate] knowledge work is not just indifferent to humanistic knowledge, it opposes it on principle" (6), while in her book, Hayles writes, "If the Traditional Humanities are at risk of becoming marginal to the main business of the contemporary academy and society, the Digital Humanities are at risk of becoming a trade practice held captive by the interest of corporate capitalism" (53–54). Finally, Drucker is concerned with the ramifications of DH's wholesale adoption of visualization techniques that "come entirely from realms outside the humanities—management, social sciences, natural sciences, business, economics, military surveillance, entertainment, gaming, and other fields," arguing that humanistic principles are at best peripheral to the methodological premises of these other disciplines, but more often than not that these other methodologies are "at odds with—even hostile to—humanistic values and thought" (86). In the context of these concerns, Liu situates the work of literature and art as committing necessary "disruptions" in the flow of informationalism (Laws of Cool, 427). "The creative arts," Liu writes, "as cultural criticism must be the history not of things created . . . but of things destroyed in the name of creation" (8).

The ACH (Association for Computers and the Humanities) has gone to great lengths to help support open access and fair use in DH by joining the DH community in filing two amicus briefs in lawsuits related to digitization of in-copyright and orphaned works in the Google Books and HathiTrust corpora. Spearheaded by Matthew Jockers, Matthew Sag, and Jason Schultz on behalf of the DH community, the briefs describe "how DH scholars employ innovative data-mining techniques in ways consistent with fair use, and how scholarship could be held back if this
kind of research is not well supported by the courts.” As of October 11, 2012, based in part on the evidence from these briefs, the United States district court ruled favorably for continued fair use in digital research (“Brief of Digital Humanities and Law Scholars”).

20. Haraway has warned that the science community considers a history of knowledge production as “histories of the technologies” (587) rather than a history of the communities that use these technologies. As a reflection of Haraway’s work as well as the work of others such as Harding and Suchman, ethnographic studies have considered technologies, instead, as ways of life and ways of social orders.

21. While some of these examples come from DH scholarship, these authors, as professors and graduate students in iSchools or professionals in libraries are scholars whose work has been heavily influenced by both the humanities and information studies. The work done by Feinberg, Carter, and Bullard describes a process in which they purposefully use selection, description, organization, and arrangement to explicate resource collections as forms of rhetorical expression.

22. Guided by feminist inquiry, Radway seeks to understand how romance novels impact gender relations. Accordingly, Radway’s methods include using in situ observations and interviews to study these dynamics since these methods allow her to shift her focus away from the reified literary text and the implied reader and toward actual, situated women reading the text. Watkins identifies the practices that must (and do) promote disruptions in the formation of human capital in which graduates from English departments inevitably engage.

Bibliography


Knorr-Cetina, K. Epistemic Cultures: How the Sciences Make Knowledge.


API and Atom Feed

Data about this text is available via read-only JSON API endpoints: sentences, annotations, comments, and index keywords.

Comments posted on this text can be followed by subscribing to this text’s Atom Feed.
Do Digital Humanists Need to Understand Algorithms?

BENJAMIN M. SCHMIDT

Algorithms and Transforms

Ian Bogost recently published an essay arguing that fetishizing algorithms can pollute our ability to accurately describe the world we live in. “Concepts like ‘algorithm,’” he writes, “have become sloppy shorthands, slang terms for the act of mistaking multipart complex systems for simple, singular ones” (Bogost). Even critics of computational culture succumb to the temptation to describe algorithms as though they operate with a single incontrovertible beauty, he argues; this leaves them with a “distorted, theological view of computational action” that ignores human agency.

As one of the few sites in the humanities where algorithms are created and deployed, the digital humanities are ideally positioned to help humanists better understand the operations of algorithms rather than blindly venerate or condemn them. But too often, we deliberately occlude understanding and meaning in favor of an instrumental approach that simply treats algorithms as tools whose efficacy can be judged intuitively. The underlying complexity of computers makes some degree of ignorance unavoidable. Past a certain point, humanists certainly do not need to understand the algorithms that produce results they use; given the complexity of modern software, it is unlikely that they could.

But although there are elements to software we can safely ignore, some basic standards of understanding remain necessary to practicing humanities data analysis as a scholarly activity and not merely a technical one. While some algorithms are indeed byzantine procedures without much coherence or purpose, others are laden with assumptions that we are perfectly well equipped to understand. What an algorithm does is distinct from, and more important to understand, than how it does it. I want to argue here that a fully realized field of humanities data analysis can do better than to test the validity of algorithms from the outside; instead, it will explore the implications of the assumptions underlying the processes described in software. Put simply: digital humanists do not need to understand algorithms at all. They do need, however, to understand the transformations that algorithms attempt to bring about. If we do so, our practice will be more effective and more likely to be truly original.

The core of this argument lies in a distinction between algorithms and transformations. An algorithm is a set of precisely specifiable steps that produce an output. “Algorithms” are central objects of study in computer science; the primary...
intellectual questions about an algorithm involve the resources necessary for those steps to run (particularly in terms of time and memory). “Transformations,” on the other hand, are the reconfigurations that an algorithm might effect. The term is less strongly linked to computer science: its strongest disciplinary ties are to mathematics (for example, in geometry, to describe the operations that can be taken on a shape) and linguistics (where it forms the heart of Noam Chomsky’s theory of “transformational grammar”).

Computationally, algorithms create transformations. Intellectually, however, people design algorithms in order to automatically perform a given transformation. That is to say: a transformation expresses a coherent goal that can be understood independently of the algorithm that produces it. Perhaps the simplest example is the transformation of sorting. “Sortedness” is a general property that any person can understand independently of the operations that produce it. The uses that one can make of alphabetical sorting in humanities research—such as producing a concordance to a text or arranging an index of names—are independent of the particular algorithm used to sort. There are, in fact, a multitude of particular algorithms that enable computers to sort a list. Certain canonical sorting algorithms, such as quicksort, are fundamental to the pedagogy in computer science. (The canonical collection and explanation of sorting algorithms is the first half of Knuth’s canonical computer science text.) It would be ludicrous to suggest humanists need to understand an algorithm like quicksort to use a sorted list. But we do need to understand sortedness itself in order to make use of the distinctive properties of a sorted list.

The alternative to understanding the meaning of transformations is to use algorithms instrumentally; to hope, for example, that an algorithm like Latent Dirichlet Allocation will approximate existing objects like “topics,” “discourses,” or “themes” and explore the fissures where it fails to do so. (See, for example, Rhody; Goldstone and Underwood; Schmidt, “Words Alone.”) This instrumental approach to software, however, promises us little in the way of understanding; in hoping that algorithms will approximate existing meanings, it in many ways precludes them from creating new ones. The signal criticism of large-scale textual analysis by traditional humanists is that it tells scholars nothing they did not know before. This critique is frequently misguided; but it does touch on a frustrating failure, which is that distant reading as commonly practiced frequently fails to offer any new ways of understanding texts.

Far more interesting, if less immediately useful, will be to marry large-scale analysis to what Stephen Ramsay calls “algorithmic criticism”: the process of using algorithmic transformations as ways to open texts for new readings (Ramsay). This is true even when, as in some of the algorithms Ramsay describes, the transformation is inherently meaningless. But transformations that embody a purpose themselves can help
us to create new versions of text that offer fresh or useful perspectives. Seeking out and describing how those transformations function is a type of work we can do more to recognize and promote.

The Fourier Transform and Literary Time

A debate between Annie Swafford and Matt Jockers over Jockers's “Syuzhet” package for exploring the shape of plots through sentiment analysis offers a useful case study of how further exploring a transformation's purpose can enrich our vocabulary for describing texts. Although Swafford's initial critique raised several issues with the package, the bulk of her continuing conversation with Jockers centered on the appropriateness of his use of a low-pass filter from signal processing as a "smoothing function." Jockers argued it provided an excellent way to "filter out the extremes in the sentiment trajectories." Swafford, on the other hand, argued that it was often dominated by "ringing artifacts" which, in practice, means the curves produced place almost all their emphasis "at the lowest point only and consider rises or falls on either side irrelevant" (Jockers, "Revealing Sentiment"; Swafford "Problems"; Swafford, "Why Syuzhet Doesn’t Work").

The Swafford and Jockers debate hinged over not just an algorithm, but a concretely defined transformation. The discrete Fourier transform undergirds the low-pass filters that Jockers uses to analyze plot. The thought that the Fourier transform might make sense as a formation for plot is an intriguing one; it is also, as Swafford argues, quite likely wrong. The ringing artifacts that Swafford describes are effects of a larger issue: the basic understanding of time embodied in the transformation itself.

The purpose of the Fourier transform is to represent cyclical events as frequencies by breaking complex signals into their component parts. Some of the most basic elements of human experience—most notably, light and sound—physically exist as repeating waves. The Fourier transform offers an easy way to describe these infinitely long waves as a short series of frequencies, constantly repeating. The pure musical note "A," for example, is a constant pulsation at 440 cycles per second; as actually produced by a clarinet, it has (among other components) a large number of regular "overtones," less powerful component notes that occur at a higher frequency and enrich the sound beyond a simple tone. A filter like the one Jockers uses strips away these regularities; it is typically used in processes like MP3 compression to strip out notes too high for the human ear to hear. When applied even more aggressively to such a clarinet tone, it would remove the higher frequencies, preserving the note "A" but attenuating the distinctive tone of the instrument.

The idea that plots might be represented in the frequency domain is fascinating, but makes some highly questionable assumptions. Perhaps the most striking assumption is that
plots, like sound or light, are composed of endlessly repeating signals. A low-pass filter like the one Jockers employs ignores any elements that seem to be regularly repeating in the text and instead focuses on the longest-term motions; those that take place over periods of time greater than a quarter or a third the length of the text. The process is analogous to predicting the continuing sound of the clarinet based on a sound clip of the note “A” just 1/440th of a second long, a single beat of the base frequency. This, remarkably, is feasible for the musical note, but only because the tone repeats endlessly. The default smoothing in the Syuzhet package assumes that books do the same; among other things, this means the smoothed versions assume the start of every book has an emotional valence that continues the trajectory of its final sentence. (I have explained this at slightly greater length in Schmidt, “Commodius Vici.”)

For some plots, including Jockers’s primary example, Portrait of the Artist as a Young Man, this assumption is not noticeably false. But for other plots, it causes great problems. Figure 48.1 shows the plot of Portrait and four other novels, with text taken from Project Gutenberg. William Dean Howell’s The Rise of Silas Lapham is a story of ruination; Ragged Dick, by Horatio Alger, is the archetypal “Rags to Riches” novel of the nineteenth century; Madame Bovary is a classically tragic tale of decline. Three different smoothing functions are shown: a weighted moving average, among the simplest possible functions; a loess moving average, which is one of the most basic and least assumption-laden algorithms used in exploratory data analysis; and the low-pass filter included with Syuzhet.4

The problems with the Fourier transform here are obvious. A periodic function forces Madame Bovary to be “as well off” after her death as before her infidelity. The less assumption-laden methods, on the other hand, allow her fate to collapse at the end and for Ragged Dick’s trajectory to move upward instead of ending on the downslope. Andrew Piper suggests that it may be quite difficult to answer the question, “How do we know when a curve is ‘wrong’?” (Piper, “Validation”). But in this case, the wrongness is actually quite apparent; only the attempt to close the circle can justify the downturn in Ragged Dick’s fate at the end of the novel.

What sort of evidence is this? By Jockers’s account the Bovary example is simply a negative “validation” of the method, by which I believe he means a sort of empirical falsification of the claim that this is the best method in all cases (Jockers, “Requiem”). Swafford’s posts imply similarly that case-by-case validation and falsification are the gold standard. In her words, the package (and perhaps the digital humanities as a whole) need “more peer review and rigorous testing—designed to confirm or refute hypotheses” (Swafford, “Continuing”).

Seen in these terms, the algorithm is a process whose operations are fundamentally opaque; we can poke or prod to see if it matches our hopes, but we can never truly know it. But when the algorithm is a means of realizing a meaningful
transformation, as in the case of the Fourier transform, we can do better than this kind of quality assurance testing; we can interpretively know in advance where a transformation will fail. I did not choose *Madame Bovary* at random to see if it looked good enough; instead, the implications of the smoothing method made it obvious that the tragedy, in general, was a type of novel that this conception of sentiment that Syuzhet’s smoothing could not comprehend. I will admit, with some trepidation, that I have never actually read either *Madame Bovary* or *Ragged Dick*, but each is the archetype of a plot wholly incompatible with low-pass filter smoothing. Any other novel that ends in death and despair or extraordinary good fortune would fail in the same way.

Figure 48.1. Four plot trajectories.

These problems carry through to Jockers’s set of fundamental plots: all begin and end at exactly the same sentiment. But the obvious problems with this assumption were not noted in the first two months of the package’s existence (which surely included far more intensive scrutiny than any peer-review process might have). One particularly interesting reason that these failings were not immediately obvious is that line charts, like Figure 48.1, do not fully embody the assumptions of the Fourier transform. The statistical graphics we use to represent results can *themselves* be thought of as meaningful transformations into a new domain of analysis. And in this case, the geometries and coordinate systems we use to chart plots are
themselves emblazoned with a particular model. Such line charts assume that time is linear and infinite. In general, this is far and away the easiest and most accurate way to represent time on paper. It is not, though, true to the frequency domain that the Fourier transform takes for granted. If the Fourier transform is the right way to look at plots, we should be plotting in polar coordinates, which wrap around to their beginning. I have replotted the same data in Figure 48.2, with percentage represented as an angle starting from 12:00 on a clock face and the sentiment defined not by height but by distance from the center.

Figure 48.2. Four plot trajectories plotted in polar coordinates.

Here, the assumptions of the Fourier transform are much more clear. For all of the novels here, time forms a closed loop; the ending points distort themselves to line up with the beginning, and vice versa. The other algorithms, on the other hand, allow great gaps: the Madame Bovary arc circles inward as if descending down a drain, and Ragged Dick propels outward into orbit.

These circular plots are more than falsifications. Fully embracing the underlying assumptions of the transform in this way does not only highlight problems with the model; it suggests a new perspective for thinking about plots. This view highlights the gap between the beginning and end as a central feature of the novel; in doing so, it challenges us to think
of the time that plots occupy as something other than straightforwardly linear.

This is a conversation worth having, in part because it reminds us to question our other assumptions about plots and time. The infinite time that the Cartesian plot implies is, in some ways, just as false as the radial one. Many smoothing methods (including the one I would like to see used in Syuzhet, loess regression), can easily extrapolate past the beginning and end of the plot. That this is possible shows that they are, in some ways, equally unsuitable for the task at hand.

The heart of the distinction between fabula and syuzhet, in fact, is that there is no way to speak about “before the beginning” of a novel, or what words Shakespeare might have written if he had spent a few more hours working past the end of Hamlet. Any model that implies such phrases exist is obviously incorrect.

But even when arguably false, these transformations may yet be productive of new understandings and forms of analysis. While this cyclical return is manifestly inappropriate to the novel, it has significant implications for the study of plot more generally. By asking what sorts of plots of the frequency domain might be useful for, we can abstractly identify whole domains where new applications may be more appropriate.

For example: the ideal form of the three-camera situation comedy is written so that episodes can air in any arbitrary order in syndication. That is to say, along some dimensions they should be cyclical. For sitcom episodes, cyclicality is a useful framework to keep in mind. The cleanness of the fit of sentiment, theme, or other attributes may be an incredibly useful tool both to understand how commercial implications intertwine with authorial independence, or for understanding the transformation of a genre over time. Techniques of signal processing could be invaluable in identifying, for example, when and where networks allow writers to spin out multi-episode plot lines.7

Though the bulk of the Swafford and Jockers conversation centered on the issue of smoothing, many digital humanists seem to have found a second critique Swafford offered far more interesting. She argued that the sentiment analysis algorithms provided by Jockers’s package, most of which were based on dictionaries of words with assigned sentiment scores, produced results that frequently violated “common sense.” While the first issue seems blandly technical, the second offers a platform for digital humanists to talk through how we might better understand the black boxes of algorithms we run. What does it mean for an algorithm to accord to common sense? For it to be useful, does it need to be right 100 percent of the time? 95 percent? 50.1 percent? If the digital humanities are to be a field that appropriates tools created by others, these are precisely the questions it needs to practice answering.

To phrase the question this way, though, is once again to consider the algorithm itself as unknowable. Just as with the
Fourier transform, it is better to ask consciously what the transformation of sentiment analysis does. Rather than thinking of the sentiment analysis portion of Syuzhet as a set of word lists to be tested against anonymous human subjects, for example, we should be thinking about the best way to implement the underlying algorithms behind sentiment analysis—logistic regression, perhaps—to distinguish between things other than the binary of “positive” and “negative.” Jockers’s inspiration, Kurt Vonnegut, for example, believed that the central binary of plot was fortune and misfortune, not happiness and sadness; while sentiment analysis provides a useful shortcut, any large-scale platforms might do better to create a classifier that actually distinguishes within that desired binary itself. Andrew Piper’s work on plot structure involves internal comparisons within the novel itself (Piper, “Novel Devotions”). Work like this can help us to better understand plot by placing it into conversation with itself and by finding useful new applications for transformations from other fields.

Doing so means that digital humanists can help to dispel the myths of algorithmic domination that Bogost unpacks, rather than participating in their creation. When historians applied psychoanalysis to historical subjects, we did not suggest they “collaborate” with psychoanalysts and then test their statements against the historical record to see how much they held true; instead, historians themselves worked to deploy concepts that were seen as themselves meaningful. It is good and useful for humanists to be able to push and prod at algorithmic black boxes when the underlying algorithms are inaccessible or overly complex. But when they are reduced to doing so, the first job of digital humanists should be to understand the goals and agendas of the transformations and systems that algorithms serve so that we can be creative users of new ideas, rather than users of tools the purposes of which we decline to know.

Notes


2. http://www.matthewjockers.net/2015/02/02/syuzhet

3. It may be worth emphasizing that a low-pass filter removes all elements above a certain frequency; it does not reduce to its top five or ten frequencies, which is a different, equally sensible compression scheme.

4. For all three filters, I have used a span approximating a third of the novel. The loess span is one-third; the moving average uses a third of the novel at a time; and the cutoff for the low-pass filter is three. To avoid jagged breaks at outlying points, I use a sine-shaped kernel to weight the moving average so that each point weights far-away points for its average less than the point itself.


This does not necessarily mean that Fourier transform is the best way to think of plots as radial. Trying to pour plot time into the bottle of periodic functions, as we are seeing, produces extremely odd results. As Scott Enderle points out, even if a function is completely and obviously cyclical, it may not be regular enough for the Fourier transform to accurately translate it to the frequency domain (Enderle).

Bibliography


API and Atom Feed

Data about this text is available via read-only JSON API endpoints: 

sentences, annotations, comments, and index keywords.

Comments posted on this text can be followed by subscribing to this text's Atom Feed.
Putting the Human Back into the Digital Humanities: Feminism, Generosity, and Mess

ELIZABETH LOSH, JACQUELINE WERNIMONT, LAURA WEXLER, HONG-AN WU

In her 2015 talk on women’s history in the digital world, Claire Potter observes “that like digital humanities, histories of media are so intertwined with histories of gender, race, and class as to require feminism” (“Putting the Humanities in Action: Why We Are All Digital Humanists, and Why That Needs to Be a Feminist Project”). Her talk took place during a spring when news of police violence against unarmed black people in places ranging from Ferguson, Missouri, to Staten Island, New York, was so regular as to operate as a kind of violent cultural heartbeat documented by #BlackLivesMatter. It also came at the beginning of a summer in which several major digital humanities events were troubled by talks and panels that seemed to belie any sense of greater intersectional sensitivity within the field. For example, a 2015 keynote address by David Hoover prompted a robust discussion on social media about the future inclusivity of “big tent” digital humanities (“Out-of-the-Box Text Analysis for the Digital Humanities”). In a peculiar turn, we saw greater popular awareness about the structural racism and sexism in the United States emerge at the same moment that it felt as if a progressive, interdisciplinary academic field had failed to make good on those same insights.

As members of FemTechNet who are interrogating norms around technology, we agree wholeheartedly with Potter that feminist digital humanities has a significant role to play in emerging academic and social efforts to ask “bigger questions and locate bigger answers” without reverting to dreams of “ludicrous racial, gender, and class harmonies.” FemTechNet and ally organizations such as the FemBot Collective are attending to fundamental design challenges posed by existential conditions of difference, discord, bad actors, and mess. As our manifesto asserts, “FemTechNet understands that technologies are complex systems with divergent values and cultural assumptions. We work to expand critical literacies about the social and political implications of these systems.”

The digital is a medium as well as a method, hence a proper concern of media studies critique. As Tara McPherson observes in a 2009 essay, far too often the more conservative digital humanities tradition—which is grounded in humanities computing, the print canon, and text-encoding initiatives—leads scholars to ignore “the epistemological, phenomenological, ethical, and cultural dimensions of the visually intense and media-rich worlds we inhabit” (McPherson, “Introduction: Media Studies and the Digital Humanities,” 119) and ultimately fails to take “the human back into the digital realm.”
McPherson also suggests that media studies would benefit from more rigorous study of algorithmic interaction and topics such as visualization or information management. Thus digital humanities is presented with an rich opportunity to lead academic change in gender/women’s studies, media studies, and elsewhere—not just at the technical level, but at theoretical and social levels as well—but it needs to be an intersectional feminist digital humanities in order to do so.

McPherson was one of the original members of FemTechNet, an international feminist collective of hundreds of scholars, students, artists, and activists who study technology and computation, which is becoming even more widely known now for its interventions in digital humanities work. FemTechNet has answered the call for the digital humanities to incorporate contemporary media studies even further by highlighting work by noted feminist theorists in the fields of STS (science and technology studies) and HCI (human-computer interaction) in the Feminist Digital Humanities course offered annually at the Digital Humanities Summer Institute at the University of Victoria. Such research emphasizes the situated, material, embodied, affective, and labor-intensive character of engagements with computational media.

The larger FemTechNet umbrella organization operates as a nonhierarchical collective that supports and advances shared objectives and methods, while both recognizing that local specificities will shape implementations and honoring domain expertise of collective members. We mandate no particular ideological or methodological approach, we share copiously, and we respect the situated labor of our colleagues. Currently underway in FemTechNet, production of the Ethnic Studies and Critical Race Pedagogy Workbook, spearheaded by Genevieve Carpio, Anne Cong-Huyen, Christofer Rodelo, Veronica Paredes, and Lisa Nakamura, among others, is a significant addition to this capacity.

Addressing biases toward imagined technocratic rationality in the digital humanities is not the special domain of FemTechNet; many feminists outside of the collective have offered important correctives to the field. For example, Julia Flanders has commented on low-status and low-wage labor in DH, Amy Earhart has written about abandoned and obsolescent projects and broken links, and Bethany Nowviskie has interrogated masculinist hubris in blog postings that range from humorously mocking phallic obsessions with size and tools in the digital humanities to wistfully meditating on the complicity of DH with the trajectory toward extinction in the anthropocene. In the context of this volume, it is noteworthy that several pieces within the Debates in DH series have usefully challenged theoretical and practical assumptions within the field from a feminist perspective. While we are using “Digital Humanities” as a heuristic in order to think through DH as a disciplinary field, we are sympathetic with Jamie
“Skye” Bianco’s intervention that “this DH” is not “one,” but many different digital humanities (Bianco).

While there is a great deal of excellent feminist, queer, and antiracist work within DH discourses, there remains significant room for development. Scott Weingart, Jeana Jorgensen, and Nickoal Eichmann have been working on analyses of the annual ADHO conference, which remains stubbornly male dominated (only one-third of papers are presented by women). Weingart observes that this major professional conference is also topical skewed toward masculinized methods, such as “stylometrics, programming and software, standards, image processing, network analysis, etc.” Thus DH—when approached as an object of distant reading—remains a field oriented toward instrumental engagements with digital technologies rather than negotiations in critical communities of practice. There is clearly more that feminist scholars can do to make the field more theoretically sophisticated and institutionally disruptive.

FemTechNet has embraced the recent turn in media studies toward analyzing media archeology, the apparatus, interface design, infrastructure, embodied and affective labor, and mess. “Mess” serves as a theoretical intervention in popular notions of digital media as neat, clean, and hyper-rational and serves as a powerful reminder that “the practice of any technology in the world is never quite as simple, straightforward, or idealized as it is imagined to be” (Dourish and Bell, 4). Beyond simply troubling the neat veneer of computing culture, Paul Dourish and Genevieve Bell note that attending to the messiness of digital technologies is also a way of recognizing that “technologies are contested . . . they are different among the different groups, places, contexts, and circuits” in which they are employed (5).

Recognizing Bad Actors as a Design Problem
To be critical of the social structures that are manifested and enforced by computational means, it is necessary to evaluate carefully and with discretion the type of datasets that the digital humanities employ. The unchallenged acceptance of datasets, like the uncritical inclusion of the newest computational media, further reinforces the idea that DH best proceeds along a technology-driven model of neoliberal development that dominates the global economy and creates further bias against those using legacy systems (such as feature phones or low-bandwidth networks), which is an issue in urban and rural America as well as in the developing world.

Too narrow a vision for the digital humanities obscures important conflicts among users, including contentious issues about how structures of power and privilege can be reproduced in computational systems and the need for flexible tactics around negotiation that recognize differences among stakeholders. For example, Wikipedia prides itself on the transparency and egalitarianism of its organizational dynam-
ics. It emphasizes a strong community ethos around collaborative procedures that include “civility” among its core five pillars. Clay Shirky even defines Wikipedia as “a process, not a product” that “assumes that new errors will be introduced less frequently than existing ones will be corrected” (Here Comes Everybody, 119). Yet “civility” can also be employed as a repressive device and deliberations among editors can still become messy.

Wikipedia’s pose of maintaining a “neutral point of view” can be itself problematic for feminists who do not wish to be “neutral” but rather to address its systematic bias against representing women, feminism, invisible social actors, lost histories, and the logics of reproduction rather than production. Even Wikipedia’s attempts to address gender imbalance can have, and have had, regrettable outcomes, despite numerous projects and initiatives to include more women among its notable figures. In 2013, controversy erupted after the New York Times reported on the fact that prominent authors’ names were being moved from the “American Novelists” category to the “American Women Novelists” category, thereby undercutting their centrality among all novelists or literature generally.

Senior Wikipedia editor Adrienne Wadewitz—who often used the term “digital liberal arts” rather than “digital humanities”—worked tirelessly to improve the quality and coverage of Wikipedia as an online encyclopedia and repository of images and video. She also aspired to improve a gender gap in participation on a site in which over 90 percent of the editors identify as male. As Wadewitz asserted, “The point of doing feminist outreach is you need to find not only women but also feminists. Right now only 10 percent of editors are women, but just because we recruit more women doesn’t mean we recruit more feminists.” In considering who gets “written out of history” she encouraged active questioning of “the structures of knowledge” rather than training editors to “replicate the structures of the past” (Losh, “How to Use Wikipedia as a Teaching Tool: Adriianne Wadewitz”). Wadewitz worked closely with FemTechNet from the time of the collective’s founding in 2012 until her death in 2014. Because she was so effective, FemTechNet made its Wikipedia strategy central to its curriculum and later joined those organizing Global Women Wikipedia Write-Ins, Art+Feminism Edit-A-Thons, and other ally events. Unfortunately, it can be difficult to apportion academic credit to a collaboratively authored resource that is perpetually susceptible to change.

As digital humanities initiatives aim to replicate the successes of user-friendly interoperable Web 2.0 interfaces and to capitalize on robust community contribution practices on social media platforms, it is important to acknowledge the potential negative unintended consequences of appropriating these platform design choices for user-generated content. Mark Nunes has argued that Wikipedia has become “the poster child for institutional anxiety over Web 2.0 knowledge communities” (Error Glitch, 168) because of edits that reflect
acts of bad faith, including vandalism that intentionally introduces errors. From its inception, FemTechNet has had a Wikipedia Committee devoted to effecting change within a significant knowledge repository, and we believe that Wikipedia has the opportunity to be an important site for feminist knowledge dissemination and the democratization of digital humanities projects. It is also a community where we can intervene in the everyday and sometimes extraordinary sexism, heteronormativism, and racism of Web 2.0.

Bad actors can compromise the safety and security of many types of digital humanities users, if we approach the digital humanities from a "big tent" perspective, including many self-identified online feminists. Safety is particularly at risk when harassment, ridicule, or abuse escalates into coordinated attack efforts that use anonymous accounts to cloak the identities of hostile participants. For example, one FemTechNet student working on a project theorizing feminist code found her online reputation targeted by mockery in 2013 on Reddit and GitHub. Barbed comments about her person seemed to undermine her security and privacy as a student.

FemTechNet instructors teaching about the development of independent video games—particularly feminist, queer, or trans titles—often felt forced to abandon social media in favor of more private walled gardens for discussions, and developers like Zoe Quinn or Mattie Brice received death threats, rape threats, and "doxxing" intended to destroy their credit ratings and encourage further harassment. After the 4chan forum site banned this conduct, opponents of feminist "social justice warriors" turned to 8chan as a staging ground. Although these incidents have become news items relatively recently, awareness of how rape culture may be manifested in online interactions in cyberspace dates back at least to work done on MOOs, MUDs, and other spaces for text chat in the 1980s and 1990s.4 There is a long history as well of such hostility to feminist journalists, public intellectuals, and prominent online bloggers speaking for underrepresented minorities in general. FemTechNet has seen that we are in need of a robust and strategic intersectional analysis in imagining digital humanities interfaces that might also serve as channels for social and public exchanges.

**Building Architectures for Safety and Risk**

Academic communities are no longer able to behave like gated communities; hybrid experiences in which online and face-to-face interactions converge and also obliterate distinctions between "gown" and "town." In response to the systematic exclusion seen in Wikipedia, Reddit, and the various social media venues where GamerGate and other "raids" have played out, FemTechNet proposed a year-long program to address antifeminist violence online. Threats against women and/or feminist public intellectuals reached a fever pitch with specific, detailed threats of sexual violence and assassination
directed at prominent feminist bloggers and YouTube hosts. While these women face harassment on a daily basis, threats were being leveraged in contexts, like at the state university in Utah, where law enforcement officials refused to prohibit guns in a room where Feminist Frequency creator Anita Sarkeesian was to speak.5

In contexts such as these, feminists are fundamentally at risk and engendering safety, online or “in real life,” seems difficult at best. In order to help address these disturbing situations and to ensure that feminist voices will not be de facto silenced, FemTechNet proposed a year-long content production and curation project to the DML Trust Challenge, supported by the MacArthur Foundation. We secured funding late in spring of 2015 and are currently at work creating a living digital space that houses critical and supportive information on digital security, documenting harassment, local support networks, and identity protection online.

Using the Scalar platform (scalar.usc.edu), which was initially designed for more conventional multimodal digital humanities projects, this Trust Challenge project is creating a digital collection that accepts contributions from many kinds of stakeholders in order to better keep up with the speed of digital culture, in which new forms of risk and harassment emerge with frightening speed. As Whitney Phillips has observed, so-called trolls are in a “simultaneously symbiotic and exploitative relationship to mainstream culture, particularly in the context of corporate media” (This Is Why We Can’t Have Nice Things, 21), meaning that—much like the twenty-four-hour news cycle—accelerated multichannel media acclimatization has driven and will continue to drive rapid change in harassment methods and modes.

This work continues the FemTechNet traditions of recognizing and using distributed expertise and fostering networks of collaboration that capitalize on the human resources of everyday cyberspaces, drawing on at least thirty-five domain experts to broaden the scope of our work. It also entails a willingness to hold sometimes competing paradigms and goals together in a single project, despite the existence of tensions around risk, privilege, expertise, ownership, and appropriation. This is to say, even as we work to ensure both “brave” and “safe” spaces online, we recognize a certain ambiguity in that notions of “trust” and “safety” are also central to a largely corporate, utopian narrative about the nature and function of digital communities and technologies. But we continue to hope that our underlying values, which are not those of corporations, will leave their mark and that this presents an important example of work by digital humanists that is simultaneously engaged in addressing so-called practical needs and transforming theoretical and ideological paradigms.

Similarly, we continue to engage with the rising public awareness of the dangers of engaging law enforcement engendered by digital discourse around police violence against unarmed black people in Ferguson, Baltimore, New York City,
and elsewhere. Activities to create digital collections around related acts of witnessing and bearing witness might seem to be out of the purview of a digital humanities dictated by the priorities of academic institutions, but members of FemTechNet such as Beth Coleman, Alondra Nelson, Jessie Daniels, and Kelli Moore seek to have them understood as central rather than peripheral to digital humanities work.6

The limitations and sometimes outright danger of seeking legal recourse has been long known to those who have experienced sexual assault and gender or race-based violence. Just as activists and theorists of "terrestrial" or "in real life" (IRL) violence have debated the value and utility of legal frameworks dependent on security and safety, we find ourselves similarly engaged in understanding how to best support women, queer/trans folks, and people of color online within a cultural system that can still be fundamentally oppressive and exploitative. Further, we are cognizant that while threats of violence to self and family are always damaging, we have colleagues both in the United States and abroad who have been murdered and/or sexually assaulted in efforts to silence their work in terrestrial and digital contexts,7 and we are also collaborating with FemTechNet partners working on the issue of street harassment and sexual violence, such as Jasmeen Patheja of Bangalore's Blank Noise. For the most vulnerable, to be visible online is to be visible as an all-too-human target.

Seeking Alternatives to Niceness

In contexts where women and feminists are fighting to be heard and to live freely both online and off, discourses of access and civility with DH can seem appealing. Invocations of inclusivity and acrimony-free spaces offer a utopian vision of a discipline where discord and dissensus are unnecessary. While pieces like Tom Scheinfeldt's "Why Digital Humanities Is 'Nice'" claim that DH is concerned with method rather than theory and therefore is naturally less contentious in its interpersonal relations, many might also hear echoes of a question that McPherson poses in another one of her essays: "Why Are the Digital Humanities So White?" In other words, asserting an absence of conflict around power relations can undermine claims for diversity, equity, and inclusion. What role can a genuinely messy, heterogeneous, and contentious pluralism play in this version of digital humanities if niceness is enshrined as a core value?

In its very etymology, the word "nice" points to its own negative undercurrents, as a term that has evolved from meaning "timid" to "fussy, fastidious" to "dainty, delicate" to "precise, careful" to "agreeable, delightful" to "kind, thoughtful." None of these should be attributes to which the humanities aspires. Of course, "niceness" might seem to be an even more compliant, feminized, and passive stance in academia than the highly problematic notion of "civility" invoked in cases like those of Steven Salaita,8 Saida Grundy,9 and other faculty
members persecuted by their institutions for unpopular opinions expressed in tweets from private accounts (Koh, "Nice- ness, Building, and Opening the Genealogy of the Digital Humanities"). Nonetheless, both niceness and civility do have their defenders among digital humanists seeking community and desiring a disciplinary home without domestic tension.

Certainly DH feminism is not without its own internal conflicts, just as feminist movements have always struggled to negotiate participation amid intersectional identities and to overcome the unconditional acceptance of default positions and interpretations of white feminism as the norm. By arguing that we need to defend digital spaces that are both safe and brave, we do not want to occupy the position of the “tone police” who exclude challengers to norms. As Bonnie Stewart has recently noted, Internet shaming can include important positive modes of “calling out” injustice, as well as negative modes of trolling.

Fortunately, there are now a number of useful touchstones in the field. Lisa Spiro offered “This Is Why We Fight: Defining the Values of Digital Humanities” as a counterpoint to the “who’s in? who’s out?” debates around defining DH as a field (“This Is Why We Fight: Defining the Values of Digital Humanities”). In that piece she calls for a “core values statement” as a way to “communicate its identity to itself and the general public, guide its priorities, and perhaps heal its divisions.” In that same collection, “This Digital Humanities Which Is Not One” (Bianco) invoked Luce Irigaray’s powerful “This Sex Which Is Not One” in order to disrupt the notion, implicit in Spiro’s call, that there is “a” digital humanities discipline, rather than seeing it as multiple.

So what do we have left if we shouldn’t settle for just being “nice” or “civil” or “respectful,” and we do not want to flatten a rich field into a homogenous discipline? In Designing Culture, FemTechNet cofounder Anne Balsamo lists the principle of “intellectual generosity” first among feminist virtues that include “confidence,” “humility,” “flexibility,” and “integrity.” Balsamo observes that intellectual generosity includes “the sincere acknowledgment of the work of others” and fosters “intellectual risk-taking and courageous acts of creativity” (Balsamo, 163). We would add that as of this writing, #BlackLivesMatter continues to underline the urgency of feminist antiracism as a first principle.

In conclusion, we advocate for a repositioning of digital humanities by putting the “genres of the human,” to use Sylvia Wynter’s important term, back at the center of these inquiries and by scrutinizing how gender, embodiment, and affect are often relegated to the periphery. It is vital to attend to how corpora composed of supposedly neutral and transparent databases and tools may obscure the many ways that objects of study are positioned in relationship to human—and race, classed, and gendered—constructs of discovery, revelation, display, exhibition, desire, curation, witnessing, and bearing witness. These acts of searching and finding are not neutral
facts of scholarship because they may also compromise trust, privacy, dignity, and consent and must be pursued in a spirit that is mindful of the presence and potential activities of bad actors. It is only through acknowledging and addressing how both traditional and computational media are constructed, consumed, and utilized by humans as political social actors with intersectional positionalities that digital humanities can raise the crucial questions of gender, race, nationality, class, power, and representation. We urge our colleagues in the material, mediated, and messy digital humanities to join us in embracing an ethos of generosity that supports collaboration and inclusion in the field.

Notes


4. MUD: a multiuser dungeon/dimension/domain. MUDs are real-time multiplayer game worlds that often text-based. MOO: an object-oriented MUD. Both are characterized by being network accessible, multiuser, and in the case of MOO, programmable, interactive systems well suited to the construction of text-based adventure games, conferencing systems, and other collaborative software.


6. See, for example, the #fergusonsyllabus discussion as inaugurated by Marcia Chatelain, including her discussion of it in The Atlantic (http://www.theatlantic.com/education/archive/2014/08/how-to-teach-kids-about-whats-happening-in-ferguson/379049/). See also Chad William’s #charlestonsyllabus (http://aaaihs.org/resources/charlestonsyllabus/) and Jacqueline Wernimont’s “Build a Better DH Syllabus” (https://jwernimont.wordpress.com/2015/02/17/build-a-better-dh-syllabus/).

7. For example, Sabeen Mahmud, a human rights and free speech activist who organized Pakistan’s first hackathon in April 2013, was murdered April 24, 2015.

8. Salaita was fired from the University of Illinois shortly after taking a new position there; the case revolved around a set of tweets that left administrators and donors uncomfortable. The Center for Constitutional Rights is suing on Salaita’s behalf; see http://www.ccrjustice.org/Salaita.


**API and Atom Feed**

Data about this text is available via read-only JSON API endpoints: sentences, annotations, comments, and index keywords.

Comments posted on this text can be followed by subscribing to this text’s Atom Feed.
A comparative study of machine learning methods for authorship attribution

Matthew L. Jockers
Department of English, Stanford University, Stanford, CA 94305, USA

Daniela M. Witten
Department of Statistics, Stanford University, Stanford, CA 94305, USA

Abstract

We compare and benchmark the performance of five classification methods, four of which are taken from the machine learning literature, in a classic authorship attribution problem involving the Federalist Papers. Cross-validation results are reported for each method, and each method is further employed in classifying the disputed papers and the few papers that are generally understood to be coauthored. These tests are performed using two separate feature sets: a “raw” feature set containing all words and word bigrams that are common to all of the authors, and a second “pre-processed” feature set derived by reducing the raw feature set to include only words meeting a minimum relative frequency threshold. Each of the methods tested performed well, but nearest shrunken centroids and regularized discriminant analysis had the best overall performances with 0/70 cross-validation errors.

1 Introduction

In statistical or quantitative authorship attribution, a work of unknown or disputed authorship is classified to a known author based on a training set of works of known authorship. Unlike typical document classification, however, in authorship attribution one does not desire to classify documents based on document content. Instead, one wishes to perform classification based upon author signal, or “style.” While there are several case specific issues (such as register, genre, subject, time period, sample length, etc.) that must be considered when testing for authorship, the two most important factors from a machine learning perspective involve the choice of features and the selection of an appropriate and effective classification technique. With regard to the choice of features, there is a growing consensus that analysis of high frequency words (mostly function, or closed class, words) and/or n-grams provides the most consistently reliable results in authorship attribution problems (Burrows 2002; Diederich et al., 2000; Griewe 2007; Hoover 2003a,b; Koppel et al., 2007; Martindale and McKenzie 1995; Uzuner and Katz 2005; Zhao and Zobel 2005; Yu 2008). There is, however, no similar consensus in terms of a best classifier. This lack of consensus may be attributed to the fact that, with a very few notable exceptions, the available methodologies have not been properly or fully compared.
Much of the existing attribution research has focused upon testing the efficacy of single methods by conducting experiments within a corpus of certain, known provenance (Argamon 2008; Burrows 2002, 2007; Hoover 2001, 2004a,b; Tweedie et al., 1996). Few studies in authorship attribution have been designed to test the relative merits of one method against another. Some initial benchmarking research includes Yu (2008), Jockers et al. (2008), and a more extensive analysis by Zhao and Zobel (2005). In Yu (2008) the goal was to compare methods in a more general text classification problem and not specifically in an authorship attribution setting. Yu compares naive Bayes and support vector machines in classifying “kinds of emotion” such as eroticism in Emily Dickinson and sentimentalism in early American novels. In Jockers et al. (2008), two methods (Delta and nearest shrunken centroids) are applied to an authorship attribution problem and the cross-validation results of the two methods are discussed; however, the study was not expressly designed for comparative benchmarking. In Zhao and Zobel (2005), an experiment is constructed for the purpose of benchmarking several machine learning methods. The Zhao and Zobel study presents a limited, though carefully controlled, experiment designed for comparing five machine-learning methods. Using a feature set of “365 function words,” the authors evaluate classification techniques that have been previously employed in authorship attribution problems. Of the five methods tested, the researchers conclude that Bayesian networks are generally the most effective and that decision trees perform poorly by comparison. The authors also note that when limited positive training data are available, nearest neighbor methods perform well.

The test corpus utilized by Zhao and Zobel consists of 200,000+ Associated Press newswire articles written by 2,380 authors. These articles are relatively short. The corpus contains many articles by single individual authors, with as many as 800 documents by a single writer in one extreme case. Zhao and Zobel also adjust the size of the document pool in order to assess the effect of sample size on classification performance. The corpus employed here, while useful, is not typical of many authorship problems. Despite a carefully constructed and compelling experiment, one wonders how applicable the approaches recommended would be in more classic authorship attribution problems involving longer, often more “literary” texts such as Shakespeare’s plays, The Federalist Papers of Madison, Hamilton and Jay, or even the anonymous political expose Primary Colors. Moreover, the techniques Zhao and Zobel test do not include some of the most recent methodologies from the machine learning literature. So while Zhao and Zobel’s work marks a significant methodology, it must be likewise seen as a beginning point rather than a definitive or exhaustive analysis.

Though not a benchmarking study per se, it is worth noting another experiment from 2004 in which Juola et al. (2006) orchestrated a comparative testing of authorship attribution methods in the form of an authorship attribution contest. The challenge involved twelve participants (or participant teams) working with a controlled corpus that Juola had carefully compiled. The multilingual corpus is divided into thirteen different “problem sets” with varying degrees of complexity. Participants were invited to apply methods of their choice to the thirteen diverse problem sets. After the deadline for submissions, Juola compiled the attribution results and ranked the participants in terms of their accuracy, thus providing an empirical evaluation of approaches. The highest scoring researchers (see Koppel and Schler 2004) “scored an average success rate of 71% ... using Support Vector Machine with a linear kernel function” (Juola et al., 2006).2

One difficulty with the Juola experiment is that several of the thirteen problem sets are quite small (allowing researchers limited ability to train a classifier on multiple samples of known authorship) and, thus, the corpus does not facilitate thorough cross-validation testing within the known texts. While a corpus with a very small sample size may be more realistic than the massive corpus utilized by Zhao and Zobel, the small corpus size hampered the ability of competitors to properly tune their algorithms. Most classification methods involve one or more tuning parameters (such as the number of features used in the analysis) that must be chosen based on the specific problem at hand.
Given the small sample size, competitors were most likely forced to select tuning parameter values without proper cross-validation. For this reason, the results for at least some of the problems within the Juola experiment are likely to contain an element of randomness that could have been avoided if a larger corpus had been available. Nonetheless, the empirical testing performed in this experiment is certainly worthwhile, and the results of the contest are valuable and should inspire additional analysis.

In this research, we offer a natural extension of the work begun by Zhao and Zobel. Our objective is three-fold: to expose the authorship attribution community to classification methods that have not been previously applied to authorship problems, to compare the relative performance of these methods, and to apply these methods to a classic authorship attribution problem. Additionally, we compare classification results obtained using two possible feature sets. The first feature set is not pre-processed: we let the classification algorithms determine which features to use based on cross-validation of internal tuning parameters. The second feature set is pre-processed in order to filter out context-sensitive features and limit the available feature set to features of a certain frequency that are common to all texts in the corpus. Our classification methods and feature selection methodologies are described in further detail below.

For our test corpus, we decided against using Zhao and Zobel’s corpus of newswire articles on the grounds that it is both unfamiliar to authorship researchers and not typical of authorship problems. We decided against employing the Juola problem set corpus due the small sample size and relative obscurity of at least some of the problems. The Federalist Papers corpus was selected for this research on the grounds that it met the two primary criteria of being both familiar to authorship researchers and of adequate size to afford thorough testing. As pointed out by a reviewer, the Federalist corpus is not the only suitable problem set for a benchmarking analysis. However, in addition to being one of the most widely used corpora for authorship attribution testing and investigation, the Federalist Papers is a “real” authorship corpus, with an ample set of works of known authorship and a smaller subset of disputed texts. It has the advantage of being well understood. As early as 1997, Richard S. Forsyth had noted that the Federalist Paper problem “is possibly the best candidate for an accepted benchmark in stylometry” (Forsyth 1997). The Federalist collection has the advantage of being homogeneous and the “closed set” of potential authors tightly constrained. Moreover, significant prior research on the Federalist Papers provides opportunity for comparison with other approaches.

2 Classification Methods

We compare five classification methods in this analysis: Delta (Argamon 2008; Burrows 2002; Hoover 2004a,b), k-nearest neighbors (KNN), the support vector machine (SVM), nearest shrunken centroids (NSC; Tibshirani et al., 2003), and regularized discriminant analysis (RDA; Guo et al., 2007). Of these methods, only one (Delta) is specifically designed for authorship attribution. An overview of the other methods can be found in Hastie et al. (2009). They are general-purpose classification methods from the machine learning literature. Most of these methods involve one or more tuning parameters, which are chosen via cross-validation on the training data. Delta involves a tuning parameter that determines the number of features used in the classification. KNN’s tuning parameter is the number of nearest neighbors to be used in classification of a test observation. Many versions of SVMs exist in the literature. We used a SVM with a linear kernel and one-against-one classification, as implemented in the “e1071” library of the statistical software language R. The SVM has a single tuning parameter, which determines the cost of violating the constraints. NSC has one tuning parameter, which controls the number of features used, and RDA has two tuning parameters, one of which controls the number of features used. It is worth noting that KNN and SVM result in classifiers that use all of the features present in the data, whereas
NSC, RDA and Delta perform built-in feature selection.

3 Text Acquisition, Preparation, and Tokenization

The text of the Federalist Papers was acquired through Project Gutenberg and compared against the versions available online at the Avalon Project of Yale’s Law School (*The Federalist Papers* 2009). In some cases formatting corrections to the Gutenberg texts were made and in all cases the boilerplate Gutenberg text was removed before analysis. To allow for word and bigram tokenization using the scripts developed by the authors, the corrected text was first marked up into XML. “Div” elements separated each paper and its author-related metadata: its title and the status of the paper’s authorship (e.g. Madison, Hamilton, Jay, Coauthored, Disputed). Using scripts developed for this project, the XML text was lowercased and tokenized in order to produce raw counts and relative frequencies for each word and word bigram within each text sample. The decision to use word features alone was based on the growing consensus that analysis of high frequency words (mostly function, or closed-class words) and/or n-grams provide the most consistently reliable results in authorship attribution problems (see, for example, Burrows 2002; Diederich *et al.*, 2000; Grieve 2007; Hoover 2003a,b; Koppel *et al.*, 2007; Martindale and McKenzie 1995; Uzuner and Katz 2005; Zhao and Zobel 2005; Yu 2008). We formatted the resulting data as a matrix of dimension 85×69,969 (number of texts by number of features). Further analysis and handling of the data was conducted in the open-source R statistical software package (http://cran.r-project.org).

4 Data Pre-processing

We performed all analyses on two different versions of the data, which we will refer to as “raw features” and “pre-processed features”. Despite its name, we did impose one restriction on the features contained in the raw data set. We required that every feature occur at least once in the texts of Jay, Madison, and Hamilton. That is, a feature that occurs in Hamilton and Madison texts but never in Jay was excluded from the analysis. This was done both for computational convenience and in order to avoid context-specific features that might skew the attribution by subject over style. The resulting matrix was of dimension 85×2,907.

The pre-processed feature set was composed of the subset of the raw features that appear across the corpus with a mean relative frequency of at least 0.05%. That is, the pre-processed feature set consists of features that are used by each author and also occur with sufficiently high overall relative frequency. This pre-processing results in the exclusion of features that are likely to be context sensitive (the effect of which is to avoid classifying texts based on a shared subject rather than upon a shared style). The pre-processed data matrix was of dimension 85×298.

With the exception of Delta, all methods were performed on the 85×2,907 or 85×298 matrix of feature frequencies, where the count for a given feature in a given text was converted to a frequency by dividing it by the total number of frequency counts occurring in that text. Delta was performed as specified in Argamon (2008) and Burrows (2002, 2003).

5 Exploratory Data Analysis

To visually explore the data before performing classification, we used principal components analysis (PCA) on the raw feature set. PCA (see Hastie *et al.*, 2009) is a method for projecting data on to a low-dimensional subspace that is frequently used in authorship attribution research. It is not possible to visualize the eighty-five texts directly, since they lie in 2,907-dimensional space. Instead, we compute principal components (PCs), which are directions in the 2,907-dimensional space along which much of the variance of the data occurs. The first PC is the dimension that explains the greatest possible part of the variation in the data, the second PC...
explains the next greatest possible part of the variation, and so on.

In Fig. 1, we plot the projections of the eighty-five texts on to the first, second, and third PCs. We see that the first PC shows separation between Jay and the other authors, and the second PC shows separation between the coauthored texts and the others. However, the first PC does not distinguish between Hamilton, Madison, and the disputed samples. The second and third PCs suggest some separation between Hamilton and Madison, but the disputed samples are located between Hamilton and Madison in this projection. There is a suggestion that some of the disputed samples may lie closer to Madison than to Hamilton in the plot of the second and third PCs. Though principal components analysis reveals some degree of separation between the authors, it does not provide a tool for predicting the authorship of a new text.

6 Results

We will refer to the Hamilton, Madison, and Jay samples as the training data, and the disputed and coauthored texts will be called the test data. For each method, we performed ten-fold cross-validation on the training data in order to estimate its accuracy and select tuning parameter values. Then the methods were fit on the full training data set (using the tuning parameter values that resulted in smallest cross-validation errors) and tested on the test data set. The resulting output constitutes our predictions for the disputed and coauthored texts. Each method was performed on two feature sets: the

<table>
<thead>
<tr>
<th>Method</th>
<th>Number of CV errors</th>
<th>Number of features used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delta pp</td>
<td>3/70</td>
<td>75</td>
</tr>
<tr>
<td>Delta Raw</td>
<td>3/70</td>
<td>400</td>
</tr>
<tr>
<td>KNN pp</td>
<td>3/70</td>
<td>298</td>
</tr>
<tr>
<td>KNN Raw</td>
<td>2/70</td>
<td>2,907</td>
</tr>
<tr>
<td>NSC pp</td>
<td>0/70</td>
<td>199</td>
</tr>
<tr>
<td>NSC Raw</td>
<td>0/70</td>
<td>718</td>
</tr>
<tr>
<td>RDA pp</td>
<td>1/70</td>
<td>243</td>
</tr>
<tr>
<td>RDA Raw</td>
<td>0/70</td>
<td>312</td>
</tr>
<tr>
<td>SVM pp</td>
<td>4/70</td>
<td>298</td>
</tr>
<tr>
<td>SVM Raw</td>
<td>10/70</td>
<td>2,907</td>
</tr>
</tbody>
</table>

pp, pre-processed.
raw feature set, and the pre-processed feature set, as described earlier.

The number of cross-validation errors resulting from each method is given in Table 1. These results suggest that in this experiment all of the methods performed quite well. NSC and RDA show very few cross-validation errors, and SVM performs the least effectively, with ten cross-validation errors reported with the raw feature set and four with the pre-processed data. The number of features used by each method is shown in Table 1. The NSC, RDA, and Delta classifiers derive added interpretability from the fact that they use only a subset of the words and bigrams; this subset can be examined in order to determine what types of words and bigrams differ between the authors. The smallest cross-validation error for KNN on the raw data was achieved using \( K = 6 \) neighbors.

The attributions for the disputed and coauthored papers are found in Table 2. The attribution data displayed in Table 2 is in general agreement with prior attribution work suggesting that most of the disputed papers are likely written by Madison. There is a hint that a few may be from Hamilton.

Since NSC was the best overall classifier in these tests, we provide a few additional observations regarding the NSC results. Table 3 shows the fifty most important features used by NSC. A positive score indicates that the given author used the word or word bigram more than average, and a negative score indicates underuse.

The NSC output includes probabilities indicating the classifier’s certainty of each sample text belonging to a given candidate author, within the closed set of candidate authors. The probabilities are listed in Table 4.

The results in Table 4 suggest a surprisingly high degree of certainty that Madison authored the texts in question. In particular, NSC is confident that Madison authored even the papers that are known to have been coauthored. In interpreting these probabilities, it is important to keep in mind that NSC (like any classifier) is quite sensitive to the parameters used to fit the model, such as the feature set and specific choice of training samples used. The extremely high probabilities in Table 4 may be due in part to the use of context-specific words by the classifier. That is, if the Madison training texts and the test texts address a particular topic that is not addressed by the Hamilton or Jay training texts, then the NSC classifier might use these words as very strong evidence that the test texts were written by Madison. Therefore, one should interpret the probabilities in Table 4 as the probability of each test text being written by a given author under the NSC model. The sensitivity of an individual classifier argues for the application of multiple classifiers for a single problem, as we have done in this study.\(^\text{10}\)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No. 18</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Hamilton</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
</tr>
<tr>
<td>No. 19</td>
<td>Madison</td>
<td>Madison</td>
<td>Hamilton</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Hamilton</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
</tr>
<tr>
<td>No. 20</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Hamilton</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
</tr>
<tr>
<td>No. 49</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Hamilton</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
</tr>
<tr>
<td>No. 50</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Hamilton</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
</tr>
<tr>
<td>No. 51</td>
<td>Madison</td>
<td>Madison</td>
<td>Hamilton</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Hamilton</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
</tr>
<tr>
<td>No. 52</td>
<td>Madison</td>
<td>Madison</td>
<td>Hamilton</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Hamilton</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
</tr>
<tr>
<td>No. 53</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Hamilton</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
</tr>
<tr>
<td>No. 54</td>
<td>Madison</td>
<td>Hamilton</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Hamilton</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
</tr>
<tr>
<td>No. 55</td>
<td>Madison</td>
<td>Hamilton</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Hamilton</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
</tr>
<tr>
<td>No. 56</td>
<td>Madison</td>
<td>Madison</td>
<td>Hamilton</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Hamilton</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
</tr>
<tr>
<td>No. 57</td>
<td>Madison</td>
<td>Hamilton</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Hamilton</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
</tr>
<tr>
<td>No. 58</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
</tr>
<tr>
<td>No. 62</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
</tr>
<tr>
<td>No. 63</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
<td>Madison</td>
</tr>
</tbody>
</table>

Papers eighteen to twenty are thought to be coauthored by Hamilton and Madison, and the remaining are of unknown authorship.

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Attributions by method for disputed and coauthored papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. 18</td>
<td>Madison</td>
</tr>
<tr>
<td>No. 19</td>
<td>Madison</td>
</tr>
<tr>
<td>No. 20</td>
<td>Madison</td>
</tr>
<tr>
<td>No. 49</td>
<td>Madison</td>
</tr>
<tr>
<td>No. 50</td>
<td>Madison</td>
</tr>
<tr>
<td>No. 51</td>
<td>Madison</td>
</tr>
<tr>
<td>No. 52</td>
<td>Madison</td>
</tr>
<tr>
<td>No. 53</td>
<td>Madison</td>
</tr>
<tr>
<td>No. 54</td>
<td>Madison</td>
</tr>
<tr>
<td>No. 55</td>
<td>Madison</td>
</tr>
<tr>
<td>No. 56</td>
<td>Madison</td>
</tr>
<tr>
<td>No. 57</td>
<td>Madison</td>
</tr>
<tr>
<td>No. 58</td>
<td>Madison</td>
</tr>
<tr>
<td>No. 62</td>
<td>Madison</td>
</tr>
<tr>
<td>No. 63</td>
<td>Madison</td>
</tr>
</tbody>
</table>
A reviewer asked how our results would change if we did not initially restrict the “raw” data set to consist of words that occur in all three of the authors’ training texts. To answer this question, all analyses were repeated on the data set of dimension $85^2 C_2^{69, 969}$ consisting of all features that occur in any of the texts of known authorship. Restricting this data to the features that occur with a mean relative frequency of at least 0.05% (the “pre-processed” data) resulted in a data set of dimension $85^2 C_2^{158}$. (The pre-processed data set is smaller when the raw data set contains more features, since this reduces the relative frequency of each feature.) The results on these new raw and pre-processed data sets were substantively similar to those reported earlier. In particular, NSC assigned all texts to Madison using both the raw (3/70 CV errors) and the pre-processed (1/70) data sets.

### 7 Conclusions

Machine learning methods that are not specific to authorship attribution perform very well on this problem and may perform well on other authorship attribution problems as well.\textsuperscript{11} While SVM has been employed rather extensively on authorship problems (e.g. Fung 2003; Hirst and Feiguina 2007; Juola \textit{et al.}, 2006), NSC and RDA have not, and

---

**Table 3** Fifty most important features for NSC classifier based on raw data

<table>
<thead>
<tr>
<th>Feature</th>
<th>Hamilton Score</th>
<th>Madison Score</th>
<th>Jay Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0.2412</td>
<td>−0.105</td>
<td>−1.7906</td>
</tr>
<tr>
<td>america</td>
<td>−0.1449</td>
<td>0</td>
<td>1.4893</td>
</tr>
<tr>
<td>an</td>
<td>0.2422</td>
<td>−0.2594</td>
<td>−1.3678</td>
</tr>
<tr>
<td>and</td>
<td>−0.3347</td>
<td>0.297</td>
<td>2.2061</td>
</tr>
<tr>
<td>arms and</td>
<td>−0.0289</td>
<td>0</td>
<td>0.6999</td>
</tr>
<tr>
<td>be more</td>
<td>−0.0283</td>
<td>0</td>
<td>0.8047</td>
</tr>
<tr>
<td>been</td>
<td>0</td>
<td>0.0773</td>
<td>−0.7353</td>
</tr>
<tr>
<td>by</td>
<td>−0.2041</td>
<td>0.805</td>
<td>0</td>
</tr>
<tr>
<td>by the</td>
<td>−0.1438</td>
<td>0.8094</td>
<td>−0.2656</td>
</tr>
<tr>
<td>confederacies</td>
<td>−0.009</td>
<td>0</td>
<td>1.1143</td>
</tr>
<tr>
<td>four</td>
<td>0</td>
<td>−0.0512</td>
<td>0.6787</td>
</tr>
<tr>
<td>given</td>
<td>−0.0156</td>
<td>0</td>
<td>0.8156</td>
</tr>
<tr>
<td>has been</td>
<td>0.023</td>
<td>0</td>
<td>−0.7837</td>
</tr>
<tr>
<td>importance</td>
<td>−0.0301</td>
<td>0</td>
<td>0.7586</td>
</tr>
<tr>
<td>in the</td>
<td>0.0859</td>
<td>0</td>
<td>−0.6783</td>
</tr>
<tr>
<td>in</td>
<td>0.2102</td>
<td>−0.2745</td>
<td>−0.9989</td>
</tr>
<tr>
<td>independent</td>
<td>0</td>
<td>0</td>
<td>0.7267</td>
</tr>
<tr>
<td>is</td>
<td>0</td>
<td>0.0859</td>
<td>−0.9019</td>
</tr>
<tr>
<td>national government</td>
<td>0</td>
<td>−0.1011</td>
<td>0.8511</td>
</tr>
<tr>
<td>nations</td>
<td>−0.024</td>
<td>−0.0017</td>
<td>1.287</td>
</tr>
<tr>
<td>of</td>
<td>0.3769</td>
<td>−0.1734</td>
<td>−2.9831</td>
</tr>
<tr>
<td>of a</td>
<td>0.2567</td>
<td>−0.3661</td>
<td>−1.2168</td>
</tr>
<tr>
<td>of america</td>
<td>−0.1082</td>
<td>0</td>
<td>1.102</td>
</tr>
<tr>
<td>of the</td>
<td>0.1998</td>
<td>0</td>
<td>−2.0792</td>
</tr>
<tr>
<td>on</td>
<td>−0.3511</td>
<td>1.31</td>
<td>0</td>
</tr>
<tr>
<td>on the</td>
<td>−0.2469</td>
<td>1.0483</td>
<td>0</td>
</tr>
<tr>
<td>one</td>
<td>−0.063</td>
<td>0</td>
<td>0.6553</td>
</tr>
<tr>
<td>only to</td>
<td>−0.0737</td>
<td>0</td>
<td>0.7166</td>
</tr>
<tr>
<td>others</td>
<td>−0.0558</td>
<td>0</td>
<td>0.69</td>
</tr>
<tr>
<td>powers</td>
<td>−0.1778</td>
<td>0.8355</td>
<td>0</td>
</tr>
<tr>
<td>soon</td>
<td>−0.0077</td>
<td>0</td>
<td>0.7338</td>
</tr>
<tr>
<td>that they</td>
<td>−0.0644</td>
<td>0</td>
<td>1.0944</td>
</tr>
<tr>
<td>the</td>
<td>0.1427</td>
<td>0.5186</td>
<td>−3.9449</td>
</tr>
<tr>
<td>them</td>
<td>−0.0568</td>
<td>0</td>
<td>0.736</td>
</tr>
<tr>
<td>there</td>
<td>0.3111</td>
<td>−0.791</td>
<td>−0.5824</td>
</tr>
<tr>
<td>they</td>
<td>−0.0204</td>
<td>0</td>
<td>0.8312</td>
</tr>
<tr>
<td>they should</td>
<td>−0.0283</td>
<td>0</td>
<td>0.7041</td>
</tr>
<tr>
<td>this</td>
<td>0.2419</td>
<td>−0.2932</td>
<td>−1.2704</td>
</tr>
<tr>
<td>three</td>
<td>−0.0182</td>
<td>0</td>
<td>0.8145</td>
</tr>
<tr>
<td>to</td>
<td>0.4428</td>
<td>−1.0046</td>
<td>−1.3278</td>
</tr>
<tr>
<td>to be</td>
<td>0.0591</td>
<td>0</td>
<td>−0.7669</td>
</tr>
<tr>
<td>to the</td>
<td>0.1271</td>
<td>0</td>
<td>−1.2119</td>
</tr>
<tr>
<td>treaties</td>
<td>−0.0486</td>
<td>0</td>
<td>1.0861</td>
</tr>
<tr>
<td>treaties and</td>
<td>−0.0634</td>
<td>0</td>
<td>0.9153</td>
</tr>
<tr>
<td>upon</td>
<td>0.5181</td>
<td>−1.3328</td>
<td>−1.1763</td>
</tr>
<tr>
<td>useful</td>
<td>−0.0103</td>
<td>0</td>
<td>0.6865</td>
</tr>
<tr>
<td>well</td>
<td>−0.0465</td>
<td>0</td>
<td>0.982</td>
</tr>
<tr>
<td>well as</td>
<td>−0.0576</td>
<td>0</td>
<td>0.7236</td>
</tr>
<tr>
<td>which</td>
<td>0.0504</td>
<td>0</td>
<td>−1.1303</td>
</tr>
<tr>
<td>wise</td>
<td>0</td>
<td>−0.0307</td>
<td>1.0514</td>
</tr>
</tbody>
</table>

**Table 4** NSC probabilities

<table>
<thead>
<tr>
<th>Paper</th>
<th>Hamilton</th>
<th>Madison</th>
<th>Jay</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. 18</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>No. 19</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>No. 20</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>No. 49</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>No. 50</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>No. 51</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>No. 52</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>No. 53</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>No. 54</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>No. 55</td>
<td>0.0038</td>
<td>0.9962</td>
<td>0</td>
</tr>
<tr>
<td>No. 56</td>
<td>1e−04</td>
<td>0.9999</td>
<td>0</td>
</tr>
<tr>
<td>No. 57</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>No. 58</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>No. 62</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>No. 63</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
both performed particularly well in these experiments. Authorship attribution researchers would likely benefit from greater exposure to and further testing of these approaches. NSC has the advantage over RDA of having only one tuning parameter, as well as greater interpretability. With the exception of our prior work with NSC (Jockers et al., 2008), neither NSC nor RDA has been employed in authorship attribution problems. These are two examples of high-dimensional classification methods that result from regularization of classical methods intended for low-dimensional settings. We believe that these methods and other penalized classification approaches show promise for the problem of authorship attribution, which is characterized by high dimensionality and low sample size.

References


Notes

1 See also http://www.mathcs.duq.edu/~juola/authorship_contest.html.
2 For more on the use of SVM, see also Koppel et al. (2006, 2007).
3 And, the individual samples were frequently too short to allow for text segmentation.
4 Several other studies (Dumais et al., 1998; Joachims 1998; Yang and Pedersen 1997) evaluating text classification algorithms use texts drawn from similar resources, i.e. news articles and web documents.
6 Among others, see in particular (Bosch and Smith, 1998; Fung, 2003; Holmes and Forsyth, 1995; Khmelev and Tweedie, 2001; Martindale and McKenzie, 1995; Mosteller and Wallace, 1964; Tweedie et al., 1996).
7 The second RDA parameter controls regularization of the covariance matrix.
8 See http://avalon.law.yale.edu/subject_menus/fed.asp.
9 Hoover (2008) provides a useful overview of PCA in authorship research.
10 Luyckx and Daelemans (2008) argue that many authorship studies overestimate the performance of their systems by limiting the pool of potential candidates to too small a set of potential authors. The performance observed here should be viewed in the context of this particular problem and not necessarily extended to other authorship problems involving hundreds (Luyckx and Daelemans, 2008) or thousands (Koppel et al., 2006) of potential authors.
11 For example, we found (Jockers et al., 2008) that NSC performed substantially better than Delta.
A rumor prevails that literary scholars should and do neglect using digital applications that aid interpretation because most of these tools seem too objective or deterministic—digital tools seem to take the “human” (e.g., the significance of gender, race, class, religion, sexuality, and history) out of literary study. The thinking is that twentieth- and twenty-first-century literary (and cultural) theory, which tends to value the literary texts and aspects of them that resist simple evaluative resolutions, is incompatible with digital methodologies, which are supposedly geared toward simplifications and fast solutions. A consequence of this thinking is the perception that a literary scholar’s research questions are not readily transferable to modes of research, dissemination, design, preservation, and communication that rely on algorithms, software, and the Internet. Happily, this perspective is changing: in “The State of the Digital Humanities: A Report and a Critique,” Alan Liu asserts that text analysis, visualization, and data mining represent paradigmatic shifts in the work of the humanities that force scholars to reflect on the relation between information and new media and technology and that require them “to investigate underlying database, data-flow, cross-platform data architecture” (14). Liu poses a challenge to digital humanities scholars to show how these methodologies—which Franco Moretti and others have called “distant reading”—compare with and contribute to more traditional close reading practices (27). This essay works in answer to this challenge by presenting several computer-assisted modes of scholarship that depend on differential (close and distant, subjective and objective) reading practices, technologies of self-reflection and collaboration, and the value of plausibility, all of which have always been crucial to literary inquiry.

One main thrust of the argument that literary study and digital methodologies are incompatible is that digital methodologies function outside the contexts that are meaningful to literary study. On the contrary, much of the rigorous literary scholarship that depends on digital methodologies is deeply entrenched in current traditions of humanist inquiry, as I demonstrate in this essay. The first part introduces what Marjorie Perloff calls “differential reading,” which positions close and distant reading practices as both subjective and objective methodologies. The second part discusses textual analysis and visualization and the extent to which differential reading requires the technologies of self-reflection or self-consciousness. In labeling self-reflection a technology, I take my cue from Richard Poirier, who claims, “All literature is to some extent aware of itself as a technology” (113), and from Martha Nell Smith, who defines technology as “the means by which we accomplish various ends—the tools and devices on which our critical suppositions rely” (“Computing” 836); both invite us to understand self-consciousness, access, and collaboration in humanist inquiry as technologies. Technologies of access and collaboration are the main focus of the third part, in which I look at the practice of data mining as a methodology squarely situated within traditionally humanist technologies. Finally, I look at plausibility in differential reading practices, present in both close or traditional and distant or new methodologies.

Differential Reading and the “Double Discipline” of Digital Humanities

There are at least two instances in which the computer is used as a tool for interpreting the work of Gertrude Stein. In his foreword to the 1995 Dalkey edition of Stein’s The Making of Americans (1925), William Gass mentions his computational tool, “a magnifying glass which [he] can draw down out of its shy place in the corner,” to “examine the layout of the page” by “enlarg[ing] and mak[ing] comprehensible some chosen bit” (vii). An earlier mention of computer usage occurs in Carolynn Van Dyke’s 1993 article on Stein’s novel Lucy Church Amicably (1930), “‘Bits of Information and Tender Feeling’: Gertrude Stein and Computer-Generated Prose.” Van Dyke creates “computer-generated texts similar to Stein’s work,” because, she hypothesizes, they “may help to illuminate [Stein’s] writing not only because they furnish a novel basis for comparison but also, more particularly, because their principles of
composition can be described with some certainty” (169–70). What is significant about Gass’s and Van Dyke’s uses is not their differences but their similarities. Gass is using his tool to magnify one sentence of Stein’s text because he believes that it will augment his ability to examine the rest of the novel. Doing so is “convenient” and generalizable: he maintains that “almost any sentence would yield the same results” (vii). Van Dyke, in generating syntactic, semantic, and pragmatic patterns and comparing these with the style in Stein’s novel, seeks “to explore both the nature of Stein’s art and certain wider questions about linguistic and literary meaning” (170); that is, Van Dyke is also in pursuit of useful generalizations. Her digital tool is as much a magnifying glass as Gass’s: each is a tool that helps the scholar get a better look at a small part of the text to learn something about the workings of the whole. In general, computer-assisted methodologies such as text analysis, visualizations, and data mining are just such tools, but they often provide the view the magnifying glass gives the user when he or she turns it upside down. These methodologies defamiliarize texts, making them unrecognizable in a way (putting them at a distance) that helps scholars identify features they might not otherwise have seen, make hypotheses, generate research questions, and figure out prevalent patterns and how to read them.

Both close and distant reading practices can facilitate interpretation through subjective and objective means. A useful term for thinking about how new methodologies engage similar modes of interpretation with what Geoffrey Harpham calls the “double discipline” (32) of subjective and objective practices is what Marjorie Perloff calls “differential reading.” Perloff’s call for interpretive practices that engage differential reading is specifically geared toward reinvigorating or reconceptualizing the practice of literary study in the double discipline of subjective objectivity (xxv–xxxiv). To this end, she identifies the following four approaches to literature:

- As rhetoric or practical criticism: “the examination of diction and syntax, rhythm and repetition, and the various figures of speech” (6)
- As philosophy or the “potential expression of truth and knowledge” (7)
- As art or a unique aesthetic construct—a form of discourse inherently other, of which the objective is the “pleasure of representation” and the “pleasure of recognition,” or the pleasure “of taking in impersonations, fictions, and language creations of others and recognizing their justice” (17)
- As “cultural production”—“for its political role, its exposure of the state of a given society” (9)

Perloff argues that each of the classifications above must be adopted in conversation with the others; that is, notions of justice become contingent not only on the historicity of text (on its political role and its exposure of the state of this society positioned in this moment) but also on its rhetorical practices (the argument embodied in its structure and themes) and ultimately on its potential for expression and for giving pleasure in terms of representation and recognition. Perloff’s emphasis on a dialectic of knowledge (“connaissance”) comprising the “pleasure of representation” and the “pleasure of recognition” (“reconnaissance”) (17) is echoed by Stanley Fish, who calls theoretical work “the pleasure of making visible the work of so many hitherto invisible hands” (377), and Harpham, who argues that the “the truth of the past exists largely in textual traces” that “can always be reassembled from a different point of view, with different emphases, presumptions, and priorities” (24). This process of assembly, Harpham continues, in addition to being pleasurable, “is accompanied, especially for the scholar, by a distinct sense of power.” Perloff sees this empowerment as engendered by a broadened perspective: “the wider one’s reading in a specified area, the greater the pleasure of a given text and the greater the ability to make connections between texts” (16).

Reading differentially—that is, closely and at a distance, subjectively and objectively—with new methodologies necessitates Liu’s mandated close attention to databases, data flows, and data architectures, including attention to the human element behind them. For example, a scholarly edition that incorporates archival materials and is published and accessed within the technological infrastructure (the digital repository) of an academic archive or library is a database with an interface that reflects, as Derrida reminds us, a structure of “archivization” (17). The interface poses as the objective gatekeeper to archives of seemingly agnostic content, but the content and its functionality generally reflect the situated and subjective practices of a particular institutional setting. In their introduction to the special issue Toward a Poetics of the Archive, Paul Voss and Marta Werner look dreamily toward
Textual Analysis, Visualization, and the Technologies of Self-Reflective Analysis

Using linguistic text analysis to chart subtle linguistic variations has some precedent in literary studies. In John Burrows's classic 1987 text *Computation into Criticism*, for instance, his goal is an attempt "to prevent broad resemblances from obscuring subtle differences" by exploring how counting what had previously been considered "insignificant" words (e.g., "of," "the," "in") can point to "larger" arguments about style in Jane Austen's oeuvre (179–80). Burrows writes, "From no other evidence than statistical analysis of the relative frequencies of the very common words, it is possible to differentiate sharply and appropriately among the idiolects of Jane Austen's characters and even to trace the ways in which an idiolect can develop in the course of a novel" (4). The multivariate statistical techniques that Burrows used are still cited as standard protocol in a study performed almost ten years later by Wayne McKenna and Alexis Antonia.3 McKenna and Antonia examine James Joyce's "Nausicaa" episode from *Ulysses* by focusing on the modals "could" and "would," the causal conjunctives "so" and "because," and prepositional phrases beginning with "like." When mapped and plotted against characters in other episodes, the authors argue, these words reflect the extent to which the character Gerty MacDowell does not understand her social world or have any power in it. Likewise, Stephen Ramsay has used *StageGraph* to cluster Shakespearean plays on the basis of low-level structural elements such as the length of acts, scene changes, and character movements ("In Praise of Pattern"). The clusters map remarkably well to genre classifications previously discussed by critics, a finding that allows Ramsay to explore the idea that abstract genres such as tragedy or history have a direct correlation to these "low hanging" structural elements. By identifying quantifiable pieces of a text using word frequencies and locations, these scholars have generated computer-assisted close readings of the structures of texts that correspond to, contradict, or otherwise provide interesting insight into what has been assumed about the texts on an abstract level.

There are several freely available tools that can help scholars generate hypotheses about a digital text. For example, I used *HyperPo* (now evolved into *Voyant*) to compare word frequencies in *The Making of Americans* with word frequencies in a small sample of other texts on *Project Gutenberg*. Visualizing these results using a "stack graph for categories" visualization available from *Many Eyes* immediately shows one way the text is different from other well-known texts of comparable size.4 The results, as seen in figure 1, make it clear that *The Making of Americans* has the largest number of words, or tokens (see the label "Total"), the least number of unique words or types ("Unique"), and
the largest average word frequency (“Avg. Frequency”). This text also has the largest standard deviation (“STDV Frequency”), because each repeated word is repeated more times than is expected in a bell curve based on the previous three statistics. The graph also makes visible the fact that the low number of unique words (or words that are used at least once in each text) and the high average frequency of words in *The Making of Americans* are almost exactly the inverse of the numbers for Joyce’s *Finnegans Wake*. The distant view, of course, does not tell us anything about the nature of Joyce’s unique words or Stein’s repetition. For such insight we must turn to the texts, where we can see that Joyce includes words that are “word-sounds” such as “tauftauf thuartpeartrick” and “bababadalgh-araghtakaminarronkonbronntronntuonnnthuntuonntrovrrhounawnskawntoohooohordenenturnuk!,” linked words such as “upturnpikepointandplace” and “devlinsfirst,” and many other experimental words that are only used once. On reading *The Making of Americans*, we can quickly see its experimentation—almost any page includes passages in which words are repeated over and over. By using these tools to chart word frequencies from *The Making of Americans* and to compare these with other texts, we have fast proof that the repetition in the text is highly unusual. We can see that anomalies between and among texts are primarily structural. The next step requires differential reading to discover the function of these anomalies: like flipping the magnifying glass, we can use digital methodologies to move in closer and to draw our attention back out as we seek to understand the relation of patterns we see at a distance and those we see up close.

![Figure 1. Table of word frequencies from texts comparable in size or composition date to Stein’s *The Making of Americans*. The novels are arranged in the order that they were published, starting with the earliest at the bottom.](image)

Noting trends in word frequencies, however, provides us with a simplified view of the text. The computer’s ability to sort and illustrate quantified data helps identify patterns, but understanding why a pattern occurs and determining whether it is one that offers insight into a text requires technologies of self-reflective inquiry. Harpham sums up humanistic study as “[t]he scholarly study of documents and artifacts produced by human beings in the past [that] enables us to see the world from different points of view so that we may better understand ourselves” (23): the next step for the computing humanist, then, is to find or create computer-assisted research practices that are self-reflective or self-conscious, and many scholars are doing just that. John F. Sowa writes that “to be useful, a computer program must represent information about things in the world,” much of which can be interpreted in wildly different ways since “computerized information passes through many levels of representations of representations” (186); Ben Shneiderman considers the effect of visualizations, which provide a window into research results but have an inherent limitation of space that results in “occlusion of data, disorientation, and
misinterpretation”; and D. Seulley and Bradley Pasanek preface their discussion of data mining with the observation that these methodologies “will always be subject to experimenter bias” (413). Many digital humanists have been trained in literary study and use this training to consider new modes of inquiry in discussions concerning materiality and the digital text (Clement, “Digital Regiving”; Kirschenbaum; McGann, Radiant Textuality; Smith, “Importance”); Lacanian psycholinguistics and the computer screen’s “flickering signifier” (Hayles); and the French Oulipo movement as an influence in exploratory, computational analysis studies (Ramsay, “Reconceiving”; Rockwell; Sinclair). Further, situated and self-reflective reading practices have directly affected how scholars think about the process of using these digital methodologies. John Unsworth touts “the importance of failure” within digital methods (“Documenting”). Willard McCarty’s notion of a “via negative,” or a “negative way,” to knowledge involves an iterative, trial-and-error process (5; 39–41). This short list of examples reflects the awareness within digital humanities that “situated knowledges” are entwined in the “silences, absences, and distortions in dominant paradigms” that compose the layers of representations of representations that literature and digital tools employ (Haraway; Hawkesworth 8).

Using tools to facilitate differential reading practices can facilitate self-reflective or self-conscious critical practices by helping us interpret the patterns we see in texts and in how others have read texts. Vocabulary Management Profiles (VMP), for example, was developed to serve as a measuring stick that marks changes among passages or between points in a text where an author describes, tells, or analyzes a story. VMP 2.2 enables users to chart patterns that map to points in a text that indicate shifts in style, thus permitting differential reading. The software’s method of analysis assumes that new episodes, new settings, and new characters are signaled by an increase in new vocabulary while the description or analysis of these activities usually involves more repetition of already used words. The tool’s analytic procedure entails determining the average frequency ratio (between 0.0 and 1.0) for each word across a text (Youmans, “How to Generate VMP 2.2s”). In the visualizations this tool generates (see fig. 2 and fig. 3), the y-axis represents the frequency ratio while the x-axis represents the location of each word across the text. This relation is represented by a line graph that maps narrative style changes. Peaks on the charts signal new vocabulary words and thus new episodes, settings, or characters. Valleys signal repetition and thus “a continuation of the episode, description, or characterization” (Youmans, “Vocabulary-Management Profile”). Figure 2 tells us more about the behavior of repetitive patterns in The Making of Americans that we see in figure 1: the repetition in the text increases and decreases in certain spots of the narrative. In essence, these VMP 2.2 visualizations show points in the text at which the most dynamic changes in repetition occur.

Figure 2. A VMP 2.2 visualization of narrative versus description in The Making of Americans.

The VMP 2.2 visualizations afford me a new perspective from which I can read passages that scholars have identified
as significant in The Making of Americans. The hump at point B in figure 3 and the zoomed-in version of the same pattern in figure 4 indicate the “Hodder episode.” Most critics date the composition of The Making of Americans, published by Contact in Paris in 1925, to have begun between 1901 (Katz) or 1903 (Wald) and concluded in 1911; during this time, Stein was also working on Q.E.D. (1903), Fernhurst (1904–05), and the character sketches published as Three Lives (1906). The Hodder episode is considered significant because this short narrative mirrors an incident in Stein’s circle between her acquaintances Mary Gwinn and Alfred Hodder and is also fictionalized in Q.E.D., a novel it is argued Stein was unable to publish because it portrays an affair between two women. In the episode, one of the women, Cora Dounor, is having an affair with Martha Hersland’s husband, Phillip Redfern (Alfred Hodder’s supposed fictional counterpart). The episode also marks “the longest sustained narrative in the plot” (Wald 286). Priscilla Wald sees this episode as “a key for the overall project,” as well as a key to reading the more traditional (diegetic) narratives within the text (286). Leon Katz, meanwhile, argues that the Hodder episode marks both the clumsy insertion of earlier drafts and the end of the novel’s “best writing in conventional idiom” with its “stunningly effective incantatory prose rhythms that lend color and great weight to the quality of her observation”—a style that Katz argues stems from Stein’s work in Three Lives and that would eventually become the experimental writing that emerges as Stein’s “aesthetic ideal,” which seeks “absolute consistency” between “overt subject matter and form” (224).

This analysis provides for different readings. The data peak at point B, which corresponds to the Hodder episode in the text, shows that VMP 2.2 can mark the same episode as critics, on the basis of changes in the underlying textual patterns, or form, of the text. That critics’ findings are mirrored in the VMP 2.2 visualization confirms the worth of the tool’s analysis. Yet the overlap also demonstrates how digital methodologies can help us expand our understanding of texts. That is, the VMP 2.2 data for The Making of Americans show what critics have found, that Stein is both telling a story (diegesis) and explaining that story (exegesis) at the same moment in the text, and also elucidate patterns critics have disregarded. We can see from the visualization of the words in the Hodder episode in figure 4, for example, that the narrative style changes before the narrative has moved to the next topic: the words that correspond to the downward trend on the graph are in the paragraph that appears toward the conclusion of the Hodder episode, after Martha Hersland has read a letter from her husband’s desk and discovered that he is having an affair. The paragraph (and thus the episode) concludes with, “She read it to the end, she had her evidence.” This sentence ends at point A in figure 4. Though the story of Redfern’s infidelity ends at point A, the next paragraph shifts into the narrator’s first-person ruminations about the subjective meanings of categories about words (“the meaning of the words they are using . . . later have not any meaning”) in relation to Redfern:
and some then have a little shame in them when they are copying an old piece of writing where they were using words that sometime had real meaning for them and now have not any real meaning in them . . . now I commence again with words that have meaning, a little perhaps I had forgotten when it came to copying the meaning in some of the words I have just been writing. Now to begin again with what I know of the being in Phillip Redfern, now to begin again a description of Phillip Redfern and always now I will be using words having in my feeling, thinking, imagining very real meaning . . . (441)

These later mentions of Redfern (as seen in fig. 4) map to Stein’s change to a more repetitive style and the narrator’s attempt to describe Redfern through diagrammic typing or repetition with variation. And so “redfern” and thus the subject of Redfern disappears at point B in figure 4, well after the change in style at point A, indicating that the two changes (subject and style), contrary to what critics have contended, are not necessarily concurrent. A clear pattern is emerging as a result of these changes from exegesis to diegesis and back; less clear is the extent to which pattern changes are a comment on the overt subject of the text (the affair), as Wald and Katz contend, or on identity construction—a more subtle subject—that seems to emerge from a distant reading of the text (Clement, “‘Thing’”). Ultimately, these analytics and visualizations help us generate new knowledge by facilitating new readings of the text and by affording a self-reflective stance for comparisons, a perspective from which we can begin to ask why we as close readers have found some patterns and yet left others undiscovered.

**Data Mining, Visualizations, and the Technologies of Collaboration**
I developed a simple hypothesis from these initial explorations of word frequencies: arguments scholars make about *The Making of Americans* are based on limited knowledge of the text’s underlying structure because the underlying patterns are difficult to discern with close reading. Data-mining procedures proved to be productive in initially illuminating complex structural patterns that helped me discern those underlying patterns. There are three main steps that comprise predictive data-mining analyses. The overall goal is to examine a large collection of documents such as the three thousand paragraphs constituting *The Making of Americans*; the first step is to determine decision criteria for classification. The decision criteria could include features of the text to be analyzed such as $n$-grams ($n$ number of characters or words of text), parts of speech, or phonetic sounds and certain behaviors or relations such as words or phrases that are repeated or words or phrases that are collocated (found in proximity to one another). The second step is to use the data-mining algorithm to analyze and map the behavior of or patterns created by the decision criteria in a subset of the document collection or corpora. The third step is to use the data-mining algorithm to apply that mapping to new documents to find similar patterns or behaviors.

In our case study for the MONK project, we used a frequent-pattern-analysis algorithm to extract features from Stein’s *The Making of Americans* for data mining. MONK (*Metadata Offer New Knowledge*) is a collaborative project funded by the Andrew W. Mellon Foundation that includes departments of computing, design, library science, and English at several universities in the United States and Canada. The goal of MONK was to develop data-mining and visualization applications that would help scholars leverage their access to large-scale text collections. The *Data to Knowledge* (*D2K*) data-mining environment we used to generate our decision criteria identified thousands of co-occurring, repetitive patterns in *The Making of Americans*. Establishing these patterns is a function of moving a window over trigrams (a three-word series), one word at a time, until each of the text’s three thousand paragraphs has been analyzed for co-occurring trigrams. Figure 5 shows how this analysis works on one sentence, in which the first two trigram sequences (A and B) are shown with the last trigram (C) and the resulting set of trigrams for the whole sentence. Looking at $n$-grams allows for an element of “fuzzy matching” that is useful when considering repetition with variation because it facilitates searching for like patterns that are not exact duplicates. For example, one result of executing the *D2K* frequent-pattern-analysis algorithm on trigrams from *The Making of Americans* is a subset of four trigrams (“a description of,” “now a description,” “this is now,” and “is now a”) that co-occur in three different paragraphs on pages 290 and 291. These four trigrams appear in the following three sentences in those paragraphs: “This is now a description of such feeling,” “This is now a description of my feeling,” “This is now a description of all of them.” Executing the frequent-pattern-analysis algorithm on longer $n$-grams produces matches of greater length. An analysis executed on 36-grams produces a subset of co-occurring patterns that enables us to find two multiparagraph sections that present an unusual pattern in the text: these sections share the same 495 words on page 444 (in ch. 4) and on page 480 (in ch. 5). The midsection between the two pages (p. 462) is the exact center of the novel, which means these pages form a bridge over the middle of the book. This startling discovery confirms the idea that the repetition in the text is not completely random. Making the same discovery would have been difficult through close reading, since the text is replete with many shorter repetitions, and impossible through more straightforward string searches without preknowledge of its existence.
Knowing that this bridge existed led me to further hypothesize that the relation of the patterns in the first part of the text to patterns in the second half of the text was purposeful and significant to the text’s experimentation. A combination of close and distant reading, of subjective and objective reading, would be required to expand my understanding of the text. Executing the algorithm on the text generated thousands of patterns (since each slight variation in a repetition generated a new pattern) and thus a long list of results that was impossible to read. To address this difficulty, we developed the interface FeatureLens, which allowed me to sort the results in different ways and view them in the context of the text. Being able to see where patterns map back to the text became increasingly important as I performed close readings to investigate the relation of the patterns in the two parts by reading at a distance. The intricate and complex patterns I discovered confirmed my hypothesis. From this perspective, in which I could read *The Making of Americans* differentially, I made the argument that the discourse about identity formation continues to develop through the compositional progress of the text even as the narrative progression within the text unwraps and ceases (Clement, “Digital Regiving”).

Close reading and distant reading in data-mining projects are facilitated most by “technologies of collaboration” (Smith, “Computing” 845). In general, using data mining and visualizations as methodologies for exploring literary texts means using cutting-edge tools that are not available to all scholars. As a result, data-mining projects in particular call for the “technologies of collaboration” or “the work of many hands” (Unsworth, “Creating”). MONK, for example, is built on two previous collaborative research projects, Nora and WordHoard, and depends on SEASR (Software Environment for the Advancement of Scholarly Research) to provide tools such as D2K, which was developed by the Automated Learning Group at the National Center for Supercomputing Applications. Further, the online MONK tool contains a selection of texts spanning from the sixteenth century through the late nineteenth century that collaborators have transcribed and encoded in TEI-Analytics (TEI-A), a TEI (Text Encoding Initiative) markup created for analytics, through Abbot, a tool created by Martin Mueller, Steve Ramsay, and Brian L. Pytlik Zillig. I helped develop the FeatureLens application in collaboration with a team of researchers at the Human-Computer Interaction Lab at the University of Maryland, College Park. The data that I used to visualize the patterns of repetition were created in coordination with Mueller, who helped develop the algorithm used to parse repetition in the text. This algorithm was developed in coordination with Craig Berry, who wrote the software to study repetitive patterns in the Iliad and Odyssey as well as those in the poems of Hesiod and the Homeric Hymns. In the digital humanities community, most projects are replete with collaborators, and resources are continually shared, reused, and remixed; yet even in such a context, data-mining methodologies stand out as being particularly dependent on
The extent to which the development of data-mining projects relies on collaborative practices is evident in a recent report published by the Council on Library and Information Resources (CLIR), *One Culture: Computationally Intensive Research in the Humanities and Social Sciences: A Report on the Experiences of First Respondents to the Digging into Data Challenge* (Williford and Henry). In this report, the authors summarize, contextualize, and analyze eight projects that were the first recipients of Digging into Data Challenge grants, awarded in 2009 and 2011 and funded by the National Science Foundation, the National Endowment for the Humanities, the Social Sciences and Humanities Research Council in Canada, and the Joint Information Systems Committee in the United Kingdom. The grants supported a diverse range of projects with teams of three to thirty-four people and resources that included images, text, and recordings of letters, trial records, and speech. Each project also included a significant investment in the human labor required to preprocess and clean the texts that would be used for the analyses as well as the labor required for testing and fine-tuning algorithms for new and varied data. Student workers are often unseen collaborators, but this report details their significant contributions. These adaptive and iterative processes and the management of so many people and resources also required significant collaborative work from project management staff.

Technologies of collaboration tend to bring to the fore the significant role subjective practices play in data-mining projects. It quickly became clear to the CLIR investigators, for example, that their initial questions set up a binary between “old” (i.e., human) versus “new” (i.e., computational) practices that did not correspond well to the varied experiences of the researchers. Originally, the CLIR investigators had asked:

1. Why do you as a scholar need a computer to do your work?; and
2. What kinds of new research can be done when computer algorithms are applied to large data corpora? (9–10)

The CLIR investigators noted that the eight projects “reflect more complex, iterative interactions between human- and machine-mediated methods than are implied by our second question. Rather than being a combination of fixed, clearly defined entities—the researcher’s question, the algorithm, and the corpus—the projects are structures built with continually moving parts” (10). Finally, after some initial work with the project participants, the writers realized that “there was never clear separation between past and present, traditional and digital, or other bounded concepts” and that “[m]any of the researchers interviewed for this study assiduously avoided making such distinctions” (10).

Observations about the data-mining process made by collaborators also support the supposition that it is the combination of data mining and visualization (distant reading) with the ability to read and contextualize one’s results (close reading) that generated many of the most productive studies and that is needed to address the remaining gaps and difficulties. Much as I did in looking at *The Making of Americans* with VMP 2.2, the team working on Structural Analysis of Large Amounts of Music Information depended on a human-generated “ground truth”—here, student-created metadata—by which the computational analysis was measured. Students worked first at labeling various structural features of recordings for which the team already had metadata, to ensure that the students were accurate; then, the students applied their knowledge to a larger set of recordings, to which machine learning was later applied and tweaked until it could produce results similar to the students’. The metadata represents rigorous close reading practices and analysis, subjective practices that dictate in what way the machine learning must be tweaked (Williford and Henry, “Case Studies”). In another project, researchers analyzing images of quilts from the *Digging into Image Data to Answer Authorship Related Questions* identified the need for “new tools to facilitate interdisciplinary collaboration and iteration on results of computational analysis of data sets, particularly in a manner that can incorporate the participation of citizen scholars” as an important next step in their work. The struggle faced by the team on the project *Digging into the Enlightenment: Mapping the Republic of Letters* seems particularly important. The project was somewhat hobbled by “incomplete data,” including missing dates and names on many of the letters; the team concluded that “[h]umanistic inquiry . . . is freeform, fluid, and exploratory; not easily translatable into a computationally reproducible set of actions.” This observation led the CLIR investigators to conclude that there was a
Valuing the Plausible in Digital Humanities Inquiries

Data mining is a computational process that is at once exciting and expensive in terms of time, labor, and processing power. The technologies of collaboration that make this work possible also provide opportunities to analyze the situated practices that make up the data architecture behind that work. The milestones in my own research, for example, would have been inaccessible without the entire extended MONK team and a complicated network of grant-funded and institutional support. Data-mining projects, like all projects, are shaped by issues that are philosophical, but the dependence these projects have on expensive resources means they are also colored by broad brushstrokes of technological, practical, financial, and political factors that, as Liu contends, we must learn to read and interrogate. The work can be exclusionary, since individual academics often do not have the resources needed to develop scholarly projects that incorporate digital methodologies and since these limited resources—not the evolving philosophies about the value of digital analysis versus human analysis—remain the largest obstacles to the wider adoption of data mining as a means for producing scholarship in literary studies.

Mary Hawkesworth describes this notion of plausibility in science as the result of disbanding Wilfrid Sellars’s “myth of the given”:

Once the “myth of the given” has been abandoned and once the belief that the absence of one invariant empirical test for the truth of a theory implies the absence of all criteria for evaluative judgment has been repudiated, then it is possible to recognize the rational grounds for assessing the merits of alternative theoretical interpretations... the stimuli that trigger interpretation limit the class of plausible characterizations without dictating one absolute description. (48–49; emphasis added)

One can consider Hawkesworth’s notion of value in the context of digital methodologies as a measurement of the extent to which these new methodologies lead one to question one’s interpretations. In other words, digital methodologies offer not an “invariant empirical test for the truth of a theory” but only plausibly sound interpretations. We have seen that values mediate every aspect of using digital tools, from selecting objects of study and methods of analysis to formulating and validating the knowledge produced. Digital projects are founded on so many small steps of human interaction and input that they inevitably reflect—as do all modes of literary inquiry—gendered, racial, economic, social, historicized, and politicized contexts. When we use simpler approaches—when we use digital methodologies to count large sets of data or to magnify small sets of data—however, it is easier for us to describe how we use this new information to make new readings of our cultural artifacts. The more complex projects require that we articulate the more complex interplay between subjective and objective practices, but literary inquiry...
The notion of plausibility with computational practices will become more productive as we learn to consider the technologies and the layering of representations of representations that make up the digital methodologies we use to look at literary texts. Consider, for example, this description of typical data-mining procedures for developers who generally “have only a superficial understanding” of what is usually numerical data:

They accept what they are given by the domain experts and do not have a deep understanding of the measurements or their relationship with each other. Results are analyzed primarily by empirical analysis. When something goes awry, we may have difficulty in attributing this to problems with the collection process or the specification of the features. (Weiss, Indurkhya, Zhang, and Damerau 51)

The authors maintain, however, that text-mining procedures are easier for developers than more quantitative data mining. “For text mining,” they write, “we are much closer to understanding the data, and we all have some expertise. The document is text. We can read and comprehend it, and we analyze a result by going directly to the documents of interest” (51–52; emphasis added). But what if the text is The Making of Americans or Finnegans Wake or any text that emphasizes the multiple, meaning-making properties of a literary text? How often is there a common, agreed-upon meaning of the words in a literary text? What if the language is code? Sowa notes how programming languages incorporate the “vagueness, uncertainty, randomness, and ignorance” that constitutes any human language at work (352). We must consider also the well-documented limitations surrounding practices of text encoding and textual knowledge representation. Julia Flanders views subjectivity as one aspect of humanist inquiry that standardized encoding practices tend to disregard because of precedents set by institutional-level projects such as those produced by libraries and the commercial industry. She offers an alternative vision that includes “understand[ing] XML as a way of expressing perspectival understandings of the text: not as a way of capturing what is timeless and essential,” but rather as a way of reflecting “shifting vantage points from which the text appears to us, the shifting relationships that constrain our understanding of it, the adaptability and strategic positioning of our own readerly motivations” (ch. 2, para. 60). Incorporating a sense of the subjectivity of seemingly objective practices like data mining corresponds to imagining alternative uses for any structured environment that incorporates systems of representation. Literary study and differential reading practices that consistently present us with the same messiness and doubt that make the complexities of concepts like race, gender, class, and culture most immediately relevant also remind us that the products of our technologies, our devices, always fall short of our perceived notions of the real or the authentic.

Ultimately, the rule of plausibility dictates that differently situated eyes panning multiple directions (or realities) not only are more powerful than a small magnifying glass but also serve different purposes and research agendas. This many-eyed perspective might be like “eye vision,” which involves shooting a dynamic event, such as a soccer game, from multiple cameras placed at different angles. A computer combines the video streams from these cameras, and the resulting images duplicate a multidimensional viewpoint. That we are aware it is a virtual reality keeps us mindful of the processes we use to produce it, but the experience of this encompassing vantage point allows for a feeling of justice or authenticity that is based on plausible complexities, not simplified and immutable truths. While computers cannot necessarily do what humanists also cannot do—such as solve literary conundrums—computational practices do allow scholars to experiment with texts in ways that were formerly prohibitive in print culture. Sometimes the view facilitated by digital tools generates the same data human beings (or humanists) could generate by hand, but more quickly—an important advantage when so many literary texts go unread and, essentially, undervalued. At other times, these vantage points are remarkably different from that which has been afforded within print culture and provide us with a new perspective on texts that continue to compel and surprise us by being so provocative and complex—so human.

Notes
1. On the history of difficulty in literary study, see Diepeveen; Poirier.

2. Perloff writes, for example, “if theories of poetry-as-rhetoric regard James Joyce and Ezra Pound as key modernists, the theory of poetry-as-philosophy would (and has) put Samuel Beckett or Paul Celan at that center" (7).

3. Some examples of Burrows’s techniques include principal component analysis, used to reduce a data set to its most useful dimensions of variance, and probability distribution tests such as Student’s t-test and the Mann-Whitney test.

4. This is another opportunity to note the subjective nature of how texts can be chosen for comparison. The texts considered for this study were limited by availability primarily because many of the texts published at the same time as The Making of Americans are under copyright and not freely available as full texts. All the texts seen in figure 1 were freely available from Project Gutenberg and were chosen because they are seminal texts of varied lengths from the nineteenth and early twentieth centuries.

5. Youmans found a strong correlation between paragraph boundaries and valleys in VMP 1 visualizations that were constructed with thirty-five-word moving intervals (“New Tool”). In contrast, Youmans used a fifty-five-word interval to investigate the correlation between VMP 1 charts and the boundaries between numbered sections in two short stories by William Faulkner (“Vocabulary Management Profile”).

6. This ratio is based on the type ratio divided by the token ratio for each word, where type equals the single occurrence of each distinct word (or form of lexeme) and tokens are the total number of words. The formula for computing a visualization in VMP 2.2 counts new vocabulary as 1.0 and repeated words as a ratio greater than 0.0 based on how recently the word occurred in the text. “Recently” is determined by “(Number of Current Word minus Number of Previous Occurrence minus 1)/(Total Tokens in the Text minus 1)” (Youmans, “How to Generate VMP 2.2s”). The creators maintain that this procedure mimics a “second reading” of a text because the first thirty-five-word window begins with the seventeenth word from the end of a text and wraps through to the eighteenth word of the beginning of the text. The next ratio is computed for the sixteenth word from the end of the text through to the nineteenth word of the beginning of the text and so on throughout the text, thereby creating a moving window that generates an average of ratios for each word.

7. These collections include the Early American Fiction Collection, Documenting the American South, Nineteenth-Century Fiction, and Wright American Fiction.

8. D2K was developed by the Automated Learning Group at the National Center for Supercomputing Applications.

9. N-grams are sequences of n-length items and are usually used as a basis for analysis in natural language processing and genetic sequence analysis. For the algorithm used in the MONK project, see Pei, Han, and Mao.

10. The list of participants and more information about the project are available at the MONK Workbench Home Test.

11. At the time, Berry was working with Northwestern University’s Academic Technologies, a unit within the Northwestern University Library. They ultimately created the Chicago Homer, which produces the same kind of data from the repetitive patterns in the Iliad and Odyssey as well as in the poems of Hesiod and the Homeric Hymns.

12. Other exclusions are also possible: the CLIR report notes that none of the principal investigators in the first round were women, whereas in the second round, funded in December 2011, “nine of the fourteen funded projects have a woman as a principal investigator” (Williford and Henry, “Case Studies” [introd., n1]).

13. In championing computational modeling, McCarty admits to the historicized scholar’s situatedness; in “raising the level of complexity in the questions we can ask of our sources [we do] better justice to them . . . justice is justice as we now conceive of it, not by an ahistorical, absolute measure” (128).

Works Cited


DOI: 10.1632/lsda.2013.8

Comments

1 Comment on paragraph 13

1. Tanya Clement March 26, 2015 at 12:05 am

Very nice catch! You are absolutely right—an error in the description of the visualization. They are not in the order of publishing date at all. They are ordered in accordance to how they can make evident this: “The graph also makes visible the fact that the low number of unique words (or words that are used at least once in each text) and the high average frequency of words in The Making of Americans are almost exactly the inverse of the numbers for Joyce’s Finnegans Wake.” In other words, the other texts have very little to show in terms of these differences. I believe, after having made this visualization almost ten years ago, that the visualization is positioning the texts in such a way to maximize the space for showing the percentage ranges (on the y-axis) for the indicators on the x-axis. Thanks, Paul!

1 Comment on paragraph 14

1. P March 24, 2015 at 2:36 pm

Maybe a slight problem with the figure or caption when it says the novels are arranged by publication dates with the earliest at the bottom. Datewise, they’re mixed up.

9 Pingbacks and trackbacks

1. Experimenting with Books | Digital Humanities @ USF May 31, 2013 at 7:01 pm

[…] about distant reading abound. In “Text Analysis, Data Mining, and Visualizations in Literary Scholarship,” Tanya Clement writes: “One main thrust of the argument that literary study and digital […]

2. My #hybrid Life | solo June 7, 2013 at 12:30 am

[…] I thought about that for a bit before I formulated a plan for a poem written by me with the help of ... . [...]

3. CCR 733: Rhetoric, Composition, and Digital Humanities (sp14) | Collin Gifford Brooke August 10, 2013 at 9:54 pm

[...] Clement, “Text Analysis, Data Mining, and Visualizations in Literary Scholarship.” LSDA Johanna Drucker, “Graphesis: Visual knowledge production and [...]”

4. Annotated Bibliography: Digital Humanities Methods | page tectonics February 11, 2014 at 8:16 pm


5. Working DH Literature Review | page tectonics April 1, 2014 at 12:13 pm

[...] Tanya. “Text Analysis, Data Mining, and Visualizations in Literary Scholarship.” Literary Studies in the Digital Age. MLA Commons. [...]

6. The Weekly Create – September 18, 2014 – How to Read the Human Brain (Kinda, Sorta, Not Reeealllyyy…) | compositionandrhetoricstudies October 20, 2014 at 5:13 pm


8. Ett ödjmukt försök i digital humaniora | Elin goes digital December 19, 2014 at 12:35 pm


[...] digital methodologies, which are supposedly geared toward simplifications and fast solutions”.2 Återigen måste jag åberopa det humanistiska ansvaret och den kritiska blicken som vi kan [...]

This site is part of the MLA network on Humanities Commons. Explore other sites on this network or register to build your own.
Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

1. INTRODUCTION

Statistics starts with data. Think of the data as being generated by a black box in which a vector of input variables \( x \) (independent variables) go in one side, and on the other side the response variables \( y \) come out. Inside the black box, nature functions to associate the predictor variables with the response variables, so the picture is like this:

\[
\begin{array}{c}
y \\
\downarrow \\
\text{nature}
\end{array} \quad \begin{array}{c}
x \end{array}
\]

There are two goals in analyzing the data:

**Prediction.** To be able to predict what the responses are going to be to future input variables;

**Information.** To extract some information about how nature is associating the response variables to the input variables.

There are two different approaches toward these goals:

**The Data Modeling Culture**

The analysis in this culture starts with assuming a stochastic data model for the inside of the black box. For example, a common data model is that data are generated by independent draws from

response variables = \( f(\text{predictor variables}, \text{random noise, parameters}) \)

The values of the parameters are estimated from the data and the model then used for information and/or prediction. Thus the black box is filled in like this:

\[
\begin{array}{c}
\downarrow \\
\text{linear regression} \\
\text{logistic regression} \\
\text{Cox model} \\
\end{array} \quad \begin{array}{c}
y \\
\uparrow \\
\text{x}
\end{array}
\]

**Model validation.** Yes–no using goodness-of-fit tests and residual examination.

**Estimated culture population.** 98% of all statisticians.

**The Algorithmic Modeling Culture**

The analysis in this culture considers the inside of the box complex and unknown. Their approach is to find a function \( f(x) \)—an algorithm that operates on \( x \) to predict the responses \( y \). Their black box looks like this:

\[
\begin{array}{c}
\downarrow \\
\text{unknown} \\
\uparrow \\
\text{x}
\end{array}
\]

**Model validation.** Measured by predictive accuracy.

**Estimated culture population.** 2% of statisticians, many in other fields.

In this paper I will argue that the focus in the statistical community on data models has:

- Led to irrelevant theory and questionable scientific conclusions;
• Kept statisticians from using more suitable algorithmic models;
• Prevented statisticians from working on exciting new problems;

I will also review some of the interesting new developments in algorithmic modeling in machine learning and look at applications to three data sets.

2. ROAD MAP

It may be revealing to understand how I became a member of the small second culture. After a seven-year stint as an academic probabilist, I resigned and went into full-time free-lance consulting. After thirteen years of consulting I joined the Berkeley Statistics Department in 1980 and have been there since. My experiences as a consultant formed my views about algorithmic modeling. Section 3 describes two of the projects I worked on. These are given to show how my views grew from such problems.

When I returned to the university and began reading statistical journals, the research was distant from what I had done as a consultant. All articles begin and end with data models. My observations about published theoretical research in statistics are in Section 4.

Data modeling has given the statistics field many successes in analyzing data and getting information about the mechanisms producing the data. But there is also misuse leading to questionable conclusions about the underlying mechanism. This is reviewed in Section 5. Following that is a discussion (Section 6) of how the commitment to data modeling has prevented statisticians from entering new scientific and commercial fields where the data being gathered is not suitable for analysis by data models.

In the past fifteen years, the growth in algorithmic modeling applications and methodology has been rapid. It has occurred largely outside statistics in a new community—often called machine learning—that is mostly young computer scientists (Section 7). The advances, particularly over the last five years, have been startling. Three of the most important changes in perception to be learned from these advances are described in Sections 8, 9, and 10, and are associated with the following names:

*Rashomon*: the multiplicity of good models;
*Occam*: the conflict between simplicity and accuracy;
*Bellman*: dimensionality—curse or blessing?

Section 11 is titled “Information from a Black Box” and is important in showing that an algorithmic model can produce more and more reliable information about the structure of the relationship between inputs and outputs than data models. This is illustrated using two medical data sets and a genetic data set. A glossary at the end of the paper explains terms that not all statisticians may be familiar with.

3. PROJECTS IN CONSULTING

As a consultant I designed and helped supervise surveys for the Environmental Protection Agency (EPA) and the state and federal court systems. Controlled experiments were designed for the EPA, and I analyzed traffic data for the U.S. Department of Transportation and the California Transportation Department. Most of all, I worked on a diverse set of prediction projects. Here are some examples:

Predicting next-day ozone levels.
Using mass spectra to identify halogen-containing compounds.
Predicting the class of a ship from high altitude radar returns.
Using sonar returns to predict the class of a submarine.
Identity of hand-sent Morse Code.
Toxicity of chemicals.
On-line prediction of the cause of a freeway traffic breakdown.
Speech recognition
The sources of delay in criminal trials in state court systems.

To understand the nature of these problems and the approaches taken to solve them, I give a fuller description of the first two on the list.

3.1 The Ozone Project

In the mid- to late 1960s ozone levels became a serious health problem in the Los Angeles Basin. Three different alert levels were established. At the highest, all government workers were directed not to drive to work, children were kept off playgrounds and outdoor exercise was discouraged.

The major source of ozone at that time was automobile tailpipe emissions. These rose into the low atmosphere and were trapped there by an inversion layer. A complex chemical reaction, aided by sunlight, cooked away and produced ozone two to three hours after the morning commute hours. The alert warnings were issued in the morning, but would be more effective if they could be issued 12 hours in advance. In the mid-1970s, the EPA funded a large effort to see if ozone levels could be accurately predicted 12 hours in advance.

Commuting patterns in the Los Angeles Basin are regular, with the total variation in any given
daylight hour varying only a few percent from one weekday to another. With the total amount of emissions about constant, the resulting ozone levels depend on the meteorology of the preceding days. A large data base was assembled consisting of lower and upper air measurements at U.S. weather stations as far away as Oregon and Arizona, together with hourly readings of surface temperature, humidity, and wind speed at the dozens of air pollution stations in the Basin and nearby areas.

Altogether, there were daily and hourly readings of over 450 meteorological variables for a period of seven years, with corresponding hourly values of ozone and other pollutants in the Basin. Let $\mathbf{x}$ be the predictor vector of meteorological variables on the $n$th day. There are more than 450 variables in $\mathbf{x}$ since information several days back is included. Let $y$ be the ozone level on the $(n+1)$st day. Then the problem was to construct a function $f(\mathbf{x})$ such that for any future day and future predictor variables $\mathbf{x}$ for that day, $f(\mathbf{x})$ is an accurate predictor of the next day’s ozone level $y$.

To estimate predictive accuracy, the first five years of data were used as the training set. The last two years were set aside as a test set. The algorithmic modeling methods available in the pre-1980s decades seem primitive now. In this project large linear regressions were run, followed by variable selection. Quadratic terms in, and interactions among, the retained variables were added and variable selection used again to prune the equations. In the end, the project was a failure—the false alarm rate of the final predictor was too high. I have regrets that this project can’t be revisited with the tools available today.

### 3.2 The Chlorine Project

The EPA samples thousands of compounds a year and tries to determine their potential toxicity. In the mid-1970s, the standard procedure was to measure the mass spectra of the compound and try to determine its chemical structure from its mass spectra.

Measuring the mass spectra is fast and cheap. But the determination of chemical structure from the mass spectra requires a painstaking examination by a trained chemist. The cost and availability of enough chemists to analyze all of the mass spectra produced daunted the EPA. Many toxic compounds contain halogens. So the EPA funded a project to determine if the presence of chlorine in a compound could be reliably predicted from its mass spectra.

Mass spectra are produced by bombarding the compound with ions in the presence of a magnetic field. The molecules of the compound split and the lighter fragments are bent more by the magnetic field than the heavier. Then the fragments hit an absorbing strip, with the position of the fragment on the strip determined by the molecular weight of the fragment. The intensity of the exposure at that position measures the frequency of the fragment. The resultant mass spectra has numbers reflecting frequencies of fragments from molecular weight 1 up to the molecular weight of the original compound. The peaks correspond to frequent fragments and there are many zeroes. The available data base consisted of the known chemical structure and mass spectra of 30,000 compounds.

The mass spectrum predictor vector $\mathbf{x}$ is of variable dimensionality. Molecular weight in the data base varied from 30 to over 10,000. The variable to be predicted is

$$y = \begin{cases} 1: & \text{contains chlorine,} \\ 2: & \text{does not contain chlorine.} \end{cases}$$

The problem is to construct a function $f(\mathbf{x})$ that is an accurate predictor of $y$ where $\mathbf{x}$ is the mass spectrum of the compound.

To measure predictive accuracy the data set was randomly divided into a 25,000 member training set and a 5,000 member test set. Linear discriminant analysis was tried, then quadratic discriminant analysis. These were difficult to adapt to the variable dimensionality. By this time I was thinking about decision trees. The hallmarks of chlorine in mass spectra were researched. This domain knowledge was incorporated into the decision tree algorithm by the design of the set of 1,500 yes–no questions that could be applied to a mass spectra of any dimensionality. The result was a decision tree that gave 95% accuracy on both chlorines and nonchlorines (see Breiman, Friedman, Olshen and Stone, 1984).

### 3.3 Perceptions on Statistical Analysis

As I left consulting to go back to the university, these were the perceptions I had about working with data to find answers to problems:

- (a) Focus on finding a good solution—that’s what consultants get paid for.
- (b) Live with the data before you plunge into modeling.
- (c) Search for a model that gives a good solution, either algorithmic or data.
- (d) Predictive accuracy on test sets is the criterion for how good the model is.
- (e) Computers are an indispensable partner.
4. RETURN TO THE UNIVERSITY

I had one tip about what research in the university was like. A friend of mine, a prominent statistician from the Berkeley Statistics Department, visited me in Los Angeles in the late 1970s. After I described the decision tree method to him, his first question was, “What’s the model for the data?”

4.1 Statistical Research

Upon my return, I started reading the Annals of Statistics, the flagship journal of theoretical statistics, and was bemused. Every article started with

Assume that the data are generated by the following model: . . .

followed by mathematics exploring inference, hypothesis testing and asymptotics. There is a wide spectrum of opinion regarding the usefulness of the theory published in the Annals of Statistics to the field of statistics as a science that deals with data. I am at the very low end of the spectrum. Still, there have been some gems that have combined nice theory and significant applications. An example is wavelet theory. Even in applications, data models are universal. For instance, in the Journal of the American Statistical Association (JASA), virtually every article contains a statement of the form:

Assume that the data are generated by the following model: . . .

I am deeply troubled by the current and past use of data models in applications, where quantitative conclusions are drawn and perhaps policy decisions made.

5. THE USE OF DATA MODELS

Statisticians in applied research consider data modeling as the template for statistical analysis: Faced with an applied problem, think of a data model. This enterprise has at its heart the belief that a statistician, by imagination and by looking at the data, can invent a reasonably good parametric class of models for a complex mechanism devised by nature. Then parameters are estimated and conclusions are drawn. But when a model is fit to data to draw quantitative conclusions:

• The conclusions are about the model’s mechanism, and not about nature’s mechanism.

It follows that:

• If the model is a poor emulation of nature, the conclusions may be wrong.

These truisms have often been ignored in the enthusiasm for fitting data models. A few decades ago, the commitment to data models was such that even simple precautions such as residual analysis or goodness-of-fit tests were not used. The belief in the infallibility of data models was almost religious. It is a strange phenomenon—once a model is made, then it becomes truth and the conclusions from it are infallible.

5.1 An Example

I illustrate with a famous (also infamous) example: assume the data is generated by independent draws from the model

\[ y = b_0 + \sum_{m=1}^{M} b_m x_m + \epsilon, \]

where the coefficients \( \{b_m\} \) are to be estimated, \( \epsilon \) is \( N(0, \sigma^2) \) and \( \sigma^2 \) is to be estimated. Given that the data is generated this way, elegant tests of hypotheses, confidence intervals, distributions of the residual sum-of-squares and asymptotics can be derived. This made the model attractive in terms of the mathematics involved. This theory was used both by academic statisticians and others to derive significance levels for coefficients on the basis of model (R), with little consideration as to whether the data on hand could have been generated by a linear model. Hundreds, perhaps thousands of articles were published claiming proof of something or other because the coefficient was significant at the 5% level.

Goodness-of-fit was demonstrated mostly by giving the value of the multiple correlation coefficient \( R^2 \) which was often closer to zero than one and which could be over inflated by the use of too many parameters. Besides computing \( R^2 \), nothing else was done to see if the observational data could have been generated by model (R). For instance, a study was done several decades ago by a well-known member of a university statistics department to assess whether there was gender discrimination in the salaries of the faculty. All personnel files were examined and a data base set up which consisted of salary as the response variable and 25 other variables which characterized academic performance; that is, papers published, quality of journals published in, teaching record, evaluations, etc. Gender appears as a binary predictor variable.

A linear regression was carried out on the data and the gender coefficient was significant at the 5% level. That this was strong evidence of sex discrimination was accepted as gospel. The design of the study raises issues that enter before the consideration of a model—Can the data gathered
answer the question posed? Is inference justified when your sample is the entire population? Should a data model be used? The deficiencies in analysis occurred because the focus was on the model and not on the problem.

The linear regression model led to many erroneous conclusions that appeared in journal articles waving the 5% significance level without knowing whether the model fit the data. Nowadays, I think most statisticians will agree that this is a suspect way to arrive at conclusions. At the time, there were few objections from the statistical profession about the fairy-tale aspect of the procedure. But, hidden in an elementary textbook, Mosteller and Tukey (1977) discuss many of the fallacies possible in regression and write “The whole area of guided regression is fraught with intellectual, statistical, computational, and subject matter difficulties.”

Even currently, there are only rare published critiques of the uncritical use of data models. One of the few is David Freedman, who examines the use of regression models (1994); the use of path models (1987) and data modeling (1991, 1995). The analysis in these papers is incisive.

5.2 Problems in Current Data Modeling

Current applied practice is to check the data model fit using goodness-of-fit tests and residual analysis. At one point, some years ago, I set up a simulated regression problem in seven dimensions with a controlled amount of nonlinearity. Standard tests of goodness-of-fit did not reject linearity until the nonlinearity was extreme. Recent theory supports this conclusion. Work by Bickel, Ritov and Stoker (2001) shows that goodness-of-fit tests have very little power unless the direction of the alternative is precisely specified. The implication is that omnibus goodness-of-fit tests, which test in many directions simultaneously, have little power, and will not reject until the lack of fit is extreme.

Furthermore, if the model is tinkered with on the basis of the data, that is, if variables are deleted or nonlinear combinations of the variables added, then goodness-of-fit tests are not applicable. Residual analysis is similarly unreliable. In a discussion after a presentation of residual analysis in a seminar at Berkeley in 1993, William Cleveland, one of the fathers of residual analysis, admitted that it could not uncover lack of fit in more than four to five dimensions. The papers I have read on using residual analysis to check lack of fit are confined to data sets with two or three variables.

With higher dimensions, the interactions between the variables can produce passable residual plots for a variety of models. A residual plot is a goodness-of-fit test, and lacks power in more than a few dimensions. An acceptable residual plot does not imply that the model is a good fit to the data.

There are a variety of ways of analyzing residuals. For instance, Landwher, Preibon and Shoemaker (1984, with discussion) gives a detailed analysis of fitting a logistic model to a three-variable data set using various residual plots. But each of the four discussants present other methods for the analysis. One is left with an unsettled sense about the arbitrariness of residual analysis.

Misleading conclusions may follow from data models that pass goodness-of-fit tests and residual checks. But published applications to data often show little care in checking model fit using these methods or any other. For instance, many of the current application articles in JASA that fit data models have very little discussion of how well their model fits the data. The question of how well the model fits the data is of secondary importance compared to the construction of an ingenious stochastic model.

5.3 The Multiplicity of Data Models

One goal of statistics is to extract information from the data about the underlying mechanism producing the data. The greatest plus of data modeling is that it produces a simple and understandable picture of the relationship between the input variables and responses. For instance, logistic regression in classification is frequently used because it produces a linear combination of the variables with weights that give an indication of the variable importance. The end result is a simple picture of how the prediction variables affect the response variable plus confidence intervals for the weights. Suppose two statisticians, each one with a different approach to data modeling, fit a model to the same data set. Assume also that each one applies standard goodness-of-fit tests, looks at residuals, etc., and is convinced that their model fits the data. Yet the two models give different pictures of nature’s mechanism and lead to different conclusions.

McCullah and Nelder (1989) write “Data will often present with almost equal emphasis on several possible models, and it is important that the statistician recognize and accept this.” Well said, but different models, all of them equally good, may give different pictures of the relation between the predictor and response variables. The question of which one most accurately reflects the data is difficult to resolve. One reason for this multiplicity is that goodness-of-fit tests and other methods for checking fit give a yes–no answer. With the lack of
power of these tests with data having more than a small number of dimensions, there will be a large number of models whose fit is acceptable. There is no way, among the yes–no methods for gauging fit, of determining which is the better model. A few statisticians know this. Mountain and Hsiao (1989) write, “It is difficult to formulate a comprehensive model capable of encompassing all rival models. Furthermore, with the use of finite samples, there are dubious implications with regard to the validity and power of various encompassing tests that rely on asymptotic theory.”

Data models in current use may have more damaging results than the publications in the social sciences based on a linear regression analysis. Just as the 5% level of significance became a de facto standard for publication, the Cox model for the analysis of survival times and logistic regression for survive–nonsurvive data have become the de facto standard for publication in medical journals. That different survival models, equally well fitting, could give different conclusions is not an issue.

5.4 Predictive Accuracy

The most obvious way to see how well the model box emulates nature’s box is this: put a case \( x \) down nature’s box getting an output \( y \). Similarly, put the same case \( x \) down the model box getting an output \( y’ \). The closeness of \( y \) and \( y’ \) is a measure of how good the emulation is. For a data model, this translates as: fit the parameters in your model by using the data, then, using the model, predict the data and see how good the prediction is.

Prediction is rarely perfect. There are usually many unmeasured variables whose effect is referred to as “noise.” But the extent to which the model box emulates nature’s box is a measure of how well our model can reproduce the natural phenomenon producing the data.

McCullagh and Nelder (1989) in their book on generalized linear models also think the answer is obvious. They write, “At first sight it might seem as though a good model is one that fits the data very well; that is, one that makes \( \hat{\mu} \) (the model predicted value) very close to \( y \) (the response value).” Then they go on to note that the extent of the agreement is biased by the number of parameters used in the model and so is not a satisfactory measure. They are, of course, right. If the model has too many parameters, then it may overfit the data and give a biased estimate of accuracy. But there are ways to remove the bias. To get a more unbiased estimate of predictive accuracy, cross-validation can be used, as advocated in an important early work by Stone (1974). If the data set is larger, put aside a test set.

Mosteller and Tukey (1977) were early advocates of cross-validation. They write, “Cross-validation is a natural route to the indication of the quality of any data-derived quantity. . . . We plan to cross-validate carefully wherever we can.”

Judging by the infrequency of estimates of predictive accuracy in JASA, this measure of model fit that seems natural to me (and to Mosteller and Tukey) is not natural to others. More publication of predictive accuracy estimates would establish standards for comparison of models, a practice that is common in machine learning.

6. THE LIMITATIONS OF DATA MODELS

With the insistence on data models, multivariate analysis tools in statistics are frozen at discriminant analysis and logistic regression in classification and multiple linear regression in regression. Nobody really believes that multivariate data is multivariate normal, but that data model occupies a large number of pages in every graduate textbook on multivariate statistical analysis.

With data gathered from uncontrolled observations on complex systems involving unknown physical, chemical, or biological mechanisms, the a priori assumption that nature would generate the data through a parametric model selected by the statistician can result in questionable conclusions that cannot be substantiated by appeal to goodness-of-fit tests and residual analysis. Usually, simple parametric models imposed on data generated by complex systems, for example, medical data, financial data, result in a loss of accuracy and information as compared to algorithmic models (see Section 11).

There is an old saying “If all a man has is a hammer, then every problem looks like a nail.” The trouble for statisticians is that recently some of the problems have stopped looking like nails. I conjecture that the result of hitting this wall is that more complicated data models are appearing in current published applications. Bayesian methods combined with Markov Chain Monte Carlo are cropping up all over. This may signify that as data becomes more complex, the data models become more cumbersome and are losing the advantage of presenting a simple and clear picture of nature’s mechanism.

Approaching problems by looking for a data model imposes an a priori straight jacket that restricts the ability of statisticians to deal with a wide range of statistical problems. The best available solution to a data problem might be a data model; then again it might be an algorithmic model. The data and the problem guide the solution. To solve a wider range of data problems, a larger set of tools is needed.
Perhaps the damaging consequence of the insistence on data models is that statisticians have ruled themselves out of some of the most interesting and challenging statistical problems that have arisen out of the rapidly increasing ability of computers to store and manipulate data. These problems are increasingly present in many fields, both scientific and commercial, and solutions are being found by nonstatisticians.

7. ALGORITHMIC MODELING

Under other names, algorithmic modeling has been used by industrial statisticians for decades. See, for instance, the delightful book “Fitting Equations to Data” (Daniel and Wood, 1971). It has been used by psychometricians and social scientists. Reading a preprint of Gifi’s book (1990) many years ago uncovered a kindred spirit. It has made small inroads into the analysis of medical data starting with Richard Olshen’s work in the early 1980s. For further work, see Zhang and Singer (1999). Jerome Friedman and Grace Wahba have done pioneering work on the development of algorithmic methods. But the list of statisticians in the algorithmic modeling business is short, and applications to data are seldom seen in the journals. The development of algorithmic methods was taken up by a community outside statistics.

7.1 A New Research Community

In the mid-1980s two powerful new algorithms for fitting data became available: neural nets and decision trees. A new research community using these tools sprang up. Their goal was predictive accuracy. The community consisted of young computer scientists, physicists and engineers plus a few aging statisticians. They began using the new tools in working on complex prediction problems where it was obvious that data models were not applicable: speech recognition, image recognition, nonlinear time series prediction, handwriting recognition, prediction in financial markets.

Their interests range over many fields that were once considered happy hunting grounds for statisticians and have turned out thousands of interesting research papers related to applications and methodology. A large majority of the papers analyze real data. The criterion for any model is what is the predictive accuracy. An idea of the range of research of this group can be got by looking at the Proceedings of the Neural Information Processing Systems Conference (their main yearly meeting) or at the Machine Learning Journal.

7.2 Theory in Algorithmic Modeling

Data models are rarely used in this community. The approach is that nature produces data in a black box whose insides are complex, mysterious, and, at least, partly unknowable. What is observed is a set of x’s that go in and a subsequent set of y’s that come out. The problem is to find an algorithm \( f(x) \) such that for future x in a test set, \( f(x) \) will be a good predictor of y.

The theory in this field shifts focus from data models to the properties of algorithms. It characterizes their “strength” as predictors, convergence if they are iterative, and what gives them good predictive accuracy. The one assumption made in the theory is that the data is drawn i.i.d. from an unknown multivariate distribution.

There is isolated work in statistics where the focus is on the theory of the algorithms. Grace Wahba’s research on smoothing spline algorithms and their applications to data (using cross-validation) is built on theory involving reproducing kernels in Hilbert Space (1990). The final chapter of the CART book (Breiman et al., 1984) contains a proof of the asymptotic convergence of the CART algorithm to the Bayes risk by letting the trees grow as the sample size increases. There are others, but the relative frequency is small.

Theory resulted in a major advance in machine learning. Vladimir Vapnik constructed informative bounds on the generalization error (infinite test set error) of classification algorithms which depend on the “capacity” of the algorithm. These theoretical bounds led to support vector machines (see Vapnik, 1995, 1998) which have proved to be more accurate predictors in classification and regression then neural nets, and are the subject of heated current research (see Section 10).

My last paper “Some infinity theory for tree ensembles” (Breiman, 2000) uses a function space analysis to try and understand the workings of tree ensemble methods. One section has the heading, “My kingdom for some good theory.” There is an effective method for forming ensembles known as “boosting,” but there isn’t any finite sample size theory that tells us why it works so well.

7.3 Recent Lessons

The advances in methodology and increases in predictive accuracy since the mid-1980s that have occurred in the research of machine learning has been phenomenal. There have been particularly exciting developments in the last five years. What has been learned? The three lessons that seem most
important to one:

Rashomon: the multiplicity of good models;
Occam: the conflict between simplicity and accuracy;
Bellman: dimensionality—curse or blessing.

8. RASHOMON AND THE MULTIPLICITY OF GOOD MODELS

Rashomon is a wonderful Japanese movie in which four people, from different vantage points, witness an incident in which one person dies and another is supposedly raped. When they come to testify in court, they all report the same facts, but their stories of what happened are very different.

What I call the Rashomon Effect is that there is often a multitude of different descriptions \( f(\mathbf{x}) \) in a class of functions giving about the same minimum error rate. The most easily understood example is subset selection in linear regression. Suppose there are 30 variables and we want to find the best five variable linear regressions. There are about 140,000 five-variable subsets in competition. Usually we pick the one with the lowest residual sum-of-squares (RSS), or, if there is a test set, the lowest test error. But there may be (and generally are) many five-variable equations that have RSS within 1.0% of the lowest RSS (see Breiman, 1996a). The same is true if test set error is being measured.

So here are three possible pictures with RSS or test set error within 1.0% of each other:

**Picture 1**

\[
y = 2.1 + 3.8x_3 - 0.6x_8 + 83.2x_{12} - 2.1x_{17} + 3.2x_{27},
\]

**Picture 2**

\[
y = -8.9 + 4.6x_5 + 0.01x_8 + 12.0x_{15} + 17.5x_{21} + 0.2x_{22},
\]

**Picture 3**

\[
y = -76.7 + 9.3x_2 + 22.0x_7 - 13.2x_8 + 3.4x_{11} + 7.2x_{28}.
\]

Which one is better? The problem is that each one tells a different story about which variables are important.

The Rashomon Effect also occurs with decision trees and neural nets. In my experiments with trees, if the training set is perturbed only slightly, say by removing a random 2–3% of the data, I can get a tree quite different from the original but with almost the same test set error. I once ran a small neural net 100 times on simple three-dimensional data reselecting the initial weights to be small and random on each run. I found 32 distinct minima, each of which gave a different picture, and having about equal test set error.

This effect is closely connected to what I call instability (Breiman, 1996a) that occurs when there are many different models crowded together that have about the same training or test set error. Then a slight perturbation of the data or in the model construction will cause a skip from one model to another. The two models are close to each other in terms of error, but can be distant in terms of the form of the model.

If, in logistic regression or the Cox model, the common practice of deleting the less important covariates is carried out, then the model becomes unstable—there are too many competing models. Say you are deleting from 15 variables to 4 variables. Perturb the data slightly and you will very possibly get a different four-variable model and a different conclusion about which variables are important. To improve accuracy by weeding out less important covariates you run into the multiplicity problem. The picture of which covariates are important can vary significantly between two models having about the same deviance.

Aggregating over a large set of competing models can reduce the nonuniqueness while improving accuracy. Arena et al. (2000) bagged (see Glossary) logistic regression models on a database of toxic and nontoxic chemicals where the number of covariates in each model was reduced from 15 to 4 by standard best subset selection. On a test set, the bagged model was significantly more accurate than the single model with four covariates. It is also more stable. This is one possible fix. The multiplicity problem and its effect on conclusions drawn from models needs serious attention.

9. OCCAM AND SIMPLICITY VS. ACCURACY

Occam’s Razor, long admired, is usually interpreted to mean that simpler is better. Unfortunately, in prediction, accuracy and simplicity (interpretability) are in conflict. For instance, linear regression gives a fairly interpretable picture of the \( y, \mathbf{x} \) relation. But its accuracy is usually less than that of the less interpretable neural nets. An example closer to my work involves trees.

On interpretability, trees rate an A+. A project I worked on in the late 1970s was the analysis of delay in criminal cases in state court systems. The Constitution gives the accused the right to a speedy trial. The Center for the State Courts was concerned
STATISTICAL MODELING: THE TWO CULTURES

Table 1
Data set descriptions

<table>
<thead>
<tr>
<th>Data set</th>
<th>Training Sample size</th>
<th>Test Sample size</th>
<th>Variables</th>
<th>Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer</td>
<td>699</td>
<td>—</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>351</td>
<td>—</td>
<td>34</td>
<td>2</td>
</tr>
<tr>
<td>Diabetes</td>
<td>768</td>
<td>—</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>Glass</td>
<td>214</td>
<td>—</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>Soybean</td>
<td>683</td>
<td>—</td>
<td>35</td>
<td>19</td>
</tr>
<tr>
<td>Letters</td>
<td>15,000</td>
<td>5000</td>
<td>16</td>
<td>26</td>
</tr>
<tr>
<td>Satellite</td>
<td>4,435</td>
<td>2000</td>
<td>36</td>
<td>6</td>
</tr>
<tr>
<td>Shuttle</td>
<td>43,500</td>
<td>14,500</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>DNA</td>
<td>2,000</td>
<td>1,186</td>
<td>60</td>
<td>3</td>
</tr>
<tr>
<td>Digit</td>
<td>7,291</td>
<td>2,007</td>
<td>256</td>
<td>10</td>
</tr>
</tbody>
</table>

that in many states, the trials were anything but speedy. It funded a study of the causes of the delay. I visited many states and decided to do the analysis in Colorado, which had an excellent computerized court data system. A wealth of information was extracted and processed.

The dependent variable for each criminal case was the time from arraignment to the time of sentencing. All of the other information in the trial history were the predictor variables. A large decision tree was grown, and I showed it on an overhead and explained it to the assembled Colorado judges. One of the splits was on District N which had a larger delay time than the other districts. I refrained from commenting on this. But as I walked out I heard one judge say to another, “I knew those guys in District N were dragging their feet.”

While trees rate an A+ on interpretability, they are good, but not great, predictors. Give them, say, a B on prediction.

9.1 Growing Forests for Prediction

Instead of a single tree predictor, grow a forest of trees on the same data—say 50 or 100. If we are classifying, put the new \( x \) down each tree in the forest and get a vote for the predicted class. Let the forest prediction be the class that gets the most votes. There has been a lot of work in the last five years on ways to grow the forest. All of the well-known methods grow the forest by perturbing the training set, growing a tree on the perturbed training set, perturbing the training set again, growing another tree, etc. Some familiar methods are bagging (Breiman, 1996b), boosting (Freund and Schapire, 1996), arcing (Breiman, 1998), and additive logistic regression (Friedman, Hastie and Tibshirani, 1998).

My preferred method to date is random forests. In this approach successive decision trees are grown by introducing a random element into their construction. For example, suppose there are 20 predictor variables. At each node choose several of the 20 at random to use to split the node. Or use a random combination of a random selection of a few variables. This idea appears in Ho (1998), in Amit and Geman (1997) and is developed in Breiman (1999).

9.2 Forests Compared to Trees

We compare the performance of single trees (CART) to random forests on a number of small and large data sets, mostly from the UCI repository (ftp.ics.uci.edu/pub/MachineLearningDatabases). A summary of the data sets is given in Table 1.

Table 2 compares the test set error of a single tree to that of the forest. For the five smaller data sets above the line, the test set error was estimated by leaving out a random 10% of the data, then running CART and the forest on the other 90%. The left-out 10% was run down the tree and the forest and the error on this 10% computed for both. This was repeated 100 times and the errors averaged. The larger data sets below the line came with a separate test set. People who have been in the classification field for a while find these increases in accuracy startling. Some errors are halved. Others are reduced by one-third. In regression, where the

Table 2
Test set misclassification error (%)

<table>
<thead>
<tr>
<th>Data set</th>
<th>Forest</th>
<th>Single tree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast cancer</td>
<td>2.9</td>
<td>5.9</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>5.5</td>
<td>11.2</td>
</tr>
<tr>
<td>Diabetes</td>
<td>24.2</td>
<td>25.3</td>
</tr>
<tr>
<td>Glass</td>
<td>22.0</td>
<td>30.4</td>
</tr>
<tr>
<td>Soybean</td>
<td>5.7</td>
<td>8.6</td>
</tr>
<tr>
<td>Letters</td>
<td>3.4</td>
<td>12.4</td>
</tr>
<tr>
<td>Satellite</td>
<td>8.6</td>
<td>14.8</td>
</tr>
<tr>
<td>Shuttle ( \times 10^3 )</td>
<td>7.0</td>
<td>62.0</td>
</tr>
<tr>
<td>DNA</td>
<td>3.9</td>
<td>6.2</td>
</tr>
<tr>
<td>Digit</td>
<td>6.2</td>
<td>17.1</td>
</tr>
</tbody>
</table>
forest prediction is the average over the individual tree predictions, the decreases in mean-squared test set error are similar.

9.3 Random Forests are A + Predictors

The Statlog Project (Mitchie, Spiegelhalter and Taylor, 1994) compared 18 different classifiers. Included were neural nets, CART, linear and quadratic discriminant analysis, nearest neighbor, etc. The first four data sets below the line in Table 1 were the only ones used in the Statlog Project that came with separate test sets. In terms of rank of accuracy on these four data sets, the forest comes in 1, 1, 1, 1 for an average rank of 1.0. The next best classifier had an average rank of 7.3.

The fifth data set below the line consists of $16 \times 16$ pixel gray scale depictions of handwritten ZIP Code numerals. It has been extensively used by AT&T Bell Labs to test a variety of prediction methods. A neural net handcrafted to the data got a test set error of 5.1% vs. 6.2% for a standard run of random forest.

9.4 The Occam Dilemma

So forests are A+ predictors. But their mechanism for producing a prediction is difficult to understand. Trying to delve into the tangled web that generated a plurality vote from 100 trees is a Herculean task. So on interpretability, they rate an F. Which brings us to the Occam dilemma:

- Accuracy generally requires more complex prediction methods. Simple and interpretable functions do not make the most accurate predictors.

Using complex predictors may be unpleasant, but the soundest path is to go for predictive accuracy first, then try to understand why. In fact, Section 10 points out that from a goal-oriented statistical viewpoint, there is no Occam’s dilemma. (For more on Occam’s Razor see Domingos, 1998, 1999.)

10. BELLMAN AND THE CURSE OF DIMENSIONALITY

The title of this section refers to Richard Bellman’s famous phrase, “the curse of dimensionality.” For decades, the first step in prediction methodology was to avoid the curse. If there were too many prediction variables, the recipe was to find a few features (functions of the predictor variables) that “contain most of the information” and then use these features to replace the original variables. In procedures common in statistics such as regression, logistic regression and survival models the advised practice is to use variable deletion to reduce the dimensionality. The published advice was that high dimensionality is dangerous. For instance, a well-regarded book on pattern recognition (Meisel, 1972) states “the features... must be relatively few in number.” But recent work has shown that dimensionality can be a blessing.

10.1 Digging It Out in Small Pieces

Reducing dimensionality reduces the amount of information available for prediction. The more predictor variables, the more information. There is also information in various combinations of the predictor variables. Let’s try going in the opposite direction:

- Instead of reducing dimensionality, increase it by adding many functions of the predictor variables.

There may now be thousands of features. Each potentially contains a small amount of information. The problem is how to extract and put together these little pieces of information. There are two outstanding examples of work in this direction, The Shape Recognition Forest (Y. Amit and D. Geman, 1997) and Support Vector Machines (V. Vapnik, 1995, 1998).

10.2 The Shape Recognition Forest

In 1992, the National Institute of Standards and Technology (NIST) set up a competition for machine algorithms to read handwritten numerals. They put together a large set of pixel pictures of handwritten numbers (223,000) written by over 2,000 individuals. The competition attracted wide interest, and diverse approaches were tried.

The Amit–Geman approach defined many thousands of small geometric features in a hierarchical assembly. Shallow trees are grown, such that at each node, 100 features are chosen at random from the appropriate level of the hierarchy; and the optimal split of the node based on the selected features is found.

When a pixel picture of a number is dropped down a single tree, the terminal node it lands in gives probability estimates $p_0, \ldots, p_9$ that it represents numbers 0, 1, ..., 9. Over 1,000 trees are grown, the probabilities averaged over this forest, and the predicted number is assigned to the largest averaged probability.

Using a 100,000 example training set and a 50,000 test set, the Amit–Geman method gives a test set error of 0.7%—close to the limits of human error.

10.3 Support Vector Machines

Suppose there is two-class data having prediction vectors in $M$-dimensional Euclidean space. The prediction vectors for class #1 are $\{x(1)\}$ and those for
class #2 are \( \{\mathbf{x}(2)\} \). If these two sets of vectors can be separated by a hyperplane then there is an optimal separating hyperplane. “Optimal” is defined as meaning that the distance of the hyperplane to any prediction vector is maximal (see below).

The set of vectors in \( \{\mathbf{x}(1)\} \) and in \( \{\mathbf{x}(2)\} \) that achieve the minimum distance to the optimal separating hyperplane are called the support vectors. Their coordinates determine the equation of the hyperplane. Vapnik (1995) showed that if a separating hyperplane exists, then the optimal separating hyperplane has low generalization error (see Glossary).

In two-class data, separability by a hyperplane does not often occur. However, let us increase the dimensionality by adding as additional predictor variables all quadratic monomials in the original predictor variables; that is, all terms of the form \( x_{m1}x_{m2} \). A hyperplane in the original variables plus quadratic monomials in the original variables is a more complex creature. The possibility of separation is greater. If no separation occurs, add cubic monomials as input features. If there are originally 30 predictor variables, then there are about 40,000 features if monomials up to the fourth degree are added.

The higher the dimensionality of the set of features, the more likely it is that separation occurs. In the ZIP Code data set, separation occurs with fourth degree monomials added. The test set error is 4.1%. Using a large subset of the NIST data base as a training set, separation also occurred after adding up to fourth degree monomials and gave a test set error rate of 1.1%.

Separation can always be had by raising the dimensionality high enough. But if the separating hyperplane becomes too complex, the generalization error becomes large. An elegant theorem (Vapnik, 1995) gives this bound for the expected generalization error:

\[
\text{Ex}(\text{GE}) \leq \text{Ex}(\text{number of support vectors})/(N - 1),
\]

where \( N \) is the sample size and the expectation is over all training sets of size \( N \) drawn from the same underlying distribution as the original training set.

The number of support vectors increases with the dimensionality of the feature space. If this number becomes too large, the separating hyperplane will not give low generalization error. If separation cannot be realized with a relatively small number of support vectors, there is another version of support vector machines that defines optimality by adding a penalty term for the vectors on the wrong side of the hyperplane.

Some ingenious algorithms make finding the optimal separating hyperplane computationally feasible. These devices reduce the search to a solution of a quadratic programming problem with linear inequality constraints that are of the order of the number \( N \) of cases, independent of the dimension of the feature space. Methods tailored to this particular problem produce speed-ups of an order of magnitude over standard methods for solving quadratic programming problems.

Support vector machines can also be used to provide accurate predictions in other areas (e.g., regression). It is an exciting idea that gives excellent performance and is beginning to supplant the use of neural nets. A readable introduction is in Cristianini and Shawe-Taylor (2000).

11. INFORMATION FROM A BLACK BOX

The dilemma posed in the last section is that the models that best emulate nature in terms of predictive accuracy are also the most complex and inscrutable. But this dilemma can be resolved by realizing the wrong question is being asked. Nature forms the outputs \( \mathbf{y} \) from the inputs \( \mathbf{x} \) by means of a black box with complex and unknown interior.

\[
\begin{array}{c}
\text{nature} \\
\downarrow \\
\mathbf{x} \\
\end{array}
\]

Current accurate prediction methods are also complex black boxes.

\[
\begin{array}{c}
\text{neural nets} \\
\downarrow \\
\text{support vectors} \\
\downarrow \\
\mathbf{x} \\
\end{array}
\]

So we are facing two black boxes, where ours seems only slightly less inscrutable than nature's. In data generated by medical experiments, ensembles of predictors can give cross-validated error rates significantly lower than logistic regression. My biostatistician friends tell me, “Doctors can interpret logistic regression.” There is no way they can interpret a black box containing fifty trees hooked together. In a choice between accuracy and interpretability, they'll go for interpretability.

Framing the question as the choice between accuracy and interpretability is an incorrect interpretation of what the goal of a statistical analysis is.
The point of a model is to get useful information about the relation between the response and predictor variables. Interpretability is a way of getting information. But a model does not have to be simple to provide reliable information about the relation between predictor and response variables; neither does it have to be a data model.

- The goal is not interpretability, but accurate information.

The following three examples illustrate this point. The first shows that random forests applied to a medical data set can give more reliable information about covariate strengths than logistic regression. The second shows that it can give interesting information that could not be revealed by a logistic regression. The third is an application to a microarray data where it is difficult to conceive of a data model that would uncover similar information.

**11.1 Example I: Variable Importance in a Survival Data Set**

The data set contains survival or nonsurvival of 155 hepatitis patients with 19 covariates. It is available at ftp.ics.uci.edu/pub/MachineLearning-Databases and was contributed by Gail Gong. The description is in a file called hepatitis.names. The data set has been previously analyzed by Diaconis and Efron (1983), and Cestnik, Konenenko and Bratko (1987). The lowest reported error rate to date, 17%, is in the latter paper.

Diaconis and Efron refer to work by Peter Gregory of the Stanford Medical School who analyzed this data and concluded that the important variables were numbers 6, 12, 14, 19 and reports an estimated 20% predictive accuracy. The variables were reduced in two stages—the first was by informal data analysis. The second refers to a more formal (unspecified) statistical procedure which I assume was logistic regression.

Efron and Diaconis drew 500 bootstrap samples from the original data set and used a similar procedure to isolate the important variables in each bootstrapped data set. The authors comment, “Of the four variables originally selected not one was selected in more than 60 percent of the samples. Hence the variables identified in the original analysis cannot be taken too seriously.” We will come back to this conclusion later.

**Logistic Regression**

The predictive error rate for logistic regression on the hepatitis data set is 17.4%. This was evaluated by doing 100 runs, each time leaving out a randomly selected 10% of the data as a test set, and then averaging over the test set errors.

Usually, the initial evaluation of which variables are important is based on examining the absolute values of the coefficients of the variables in the logistic regression divided by their standard deviations. Figure 1 is a plot of these values.

The conclusion from looking at the standardized coefficients is that variables 7 and 11 are the most important covariates. When logistic regression is run using only these two variables, the cross-validated error rate rises to 22.9%. Another way to find important variables is to run a best subsets search which, for any value $k$, finds the subset of $k$ variables having lowest deviance.

This procedure raises the problems of instability and multiplicity of models (see Section 7.1). There are about 4,000 subsets containing four variables. Of these, there are almost certainly a substantial number that have deviance close to the minimum and give different pictures of what the underlying mechanism is.
Random Forests

The random forests predictive error rate, evaluated by averaging errors over 100 runs, each time leaving out 10% of the data as a test set, is 12.3%—almost a 30% reduction from the logistic regression error.

Random forests consists of a large number of randomly constructed trees, each voting for a class. Similar to bagging (Breiman, 1996), a bootstrap sample of the training set is used to construct each tree. A random selection of the input variables is searched to find the best split for each node.

To measure the importance of the mth variable, the values of the mth variable are randomly permuted in all of the cases left out in the current bootstrap sample. Then these cases are run down the current tree and their classification noted. At the end of a run consisting of growing many trees, the percent increase in misclassification rate due to noising up each variable is computed. This is the measure of variable importance that is shown in Figure 1.

Random forests singles out two variables, the 12th and the 17th, as being important. As a verification both variables were run in random forests, individually and together. The test set error rates over 100 replications were 14.3% each. Running both together did no better. We conclude that virtually all of the predictive capability is provided by a single variable, either 12 or 17.

To explore the interaction between 12 and 17 a bit further, at the end of a random forest run using all variables, the output includes the estimated value of the probability of each class vs. the case number. This information is used to get plots of the variable values (normalized to mean zero and standard deviation one) vs. the probability of death. The variable values are smoothed using a weighted linear regression smoother. The results are in Figure 3 for variables 12 and 17.
The graphs of the variable values vs. class death probability are almost linear and similar. The two variables turn out to be highly correlated. Thinking that this might have affected the logistic regression results, it was run again with one or the other of these two variables deleted. There was little change.

Out of curiosity, I evaluated variable importance in logistic regression in the same way that I did in random forests, by permuting variable values in the 10\% test set and computing how much that increased the test set error. Not much help—variables 12 and 17 were not among the 3 variables ranked as most important. In partial verification of the importance of 12 and 17, I tried them separately as single variables in logistic regression. Variable 12 gave a 15.7\% error rate, variable 17 came in at 19.3\%.

To go back to the original Diaconis–Efron analysis, the problem is clear. Variables 12 and 17 are surrogates for each other. If one of them appears important in a model built on a bootstrap sample, the other does not. So each one’s frequency of occurrence is automatically less than 50\%. The paper lists the variables selected in ten of the samples. Either 12 or 17 appear in seven of the ten.

### 11.2 Example II Clustering in Medical Data

The Bupa liver data set is a two-class biomedical data set also available at ftp.ics.uci.edu/pub/MachineLearningDatabases. The covariates are:

1. mcv mean corpuscular volume
2. alkphos alkaline phosphotase
3. sgpt alamine aminotransferase
4. sgot aspartate aminotransferase
5. gammagt gamma-glutamyl transpeptidase
6. drinks half-pint equivalents of alcoholic beverage drunk per day

The first five attributes are the results of blood tests thought to be related to liver functioning. The 345 patients are classified into two classes by the severity of their liver malfunctioning. Class two is severe malfunctioning. In a random forests run,
the misclassification error rate is 28%. The variable importance given by random forests is in Figure 4.

Blood tests 3 and 5 are the most important, followed by test 4. Random forests also outputs an intrinsic similarity measure which can be used to cluster. When this was applied, two clusters were discovered in class two. The average of each variable is computed and plotted in each of these clusters in Figure 5.

An interesting facet emerges. The class two subjects consist of two distinct groups: those that have high scores on blood tests 3, 4, and 5 and those that have low scores on those tests.

11.3 Example III: Microarray Data

Random forests was run on a microarray lymphoma data set with three classes, sample size of 81 and 4,682 variables (genes) without any variable selection [for more information about this data set, see Dudoit, Fridlyand and Speed, (2000)]. The error rate was low. What was also interesting from a scientific viewpoint was an estimate of the importance of each of the 4,682 gene expressions.

The graph in Figure 6 was produced by a run of random forests. This result is consistent with assessments of variable importance made using other algorithmic methods, but appears to have sharper detail.

11.4 Remarks about the Examples

The examples show that much information is available from an algorithmic model. Friedman (1999) derives similar variable information from a different way of constructing a forest. The similarity is that they are both built as ways to give low predictive error.

There are 32 deaths and 123 survivors in the hepatitis data set. Calling everyone a survivor gives a baseline error rate of 20.6%. Logistic regression lowers this to 17.4%. It is not extracting much useful information from the data, which may explain its inability to find the important variables. Its weakness might have been unknown and the variable importances accepted at face value if its predictive accuracy was not evaluated.

Random forests is also capable of discovering important aspects of the data that standard data models cannot uncover. The potentially interesting clustering of class two patients in Example II is an illustration. The standard procedure when fitting data models such as logistic regression is to delete variables; to quote from Diaconis and Efron (1983) again, “…statistical experience suggests that it is unwise to fit a model that depends on 19 variables with only 155 data points available.” Newer methods in machine learning thrive on variables—the more the better. For instance, random forests does not overfit. It gives excellent accuracy on the lymphoma data set of Example III which has over 4,600 variables, with no variable deletion and is capable of extracting variable importance information from the data.
These examples illustrate the following points:

• Higher predictive accuracy is associated with more reliable information about the underlying data mechanism. Weak predictive accuracy can lead to questionable conclusions.

• Algorithmic models can give better predictive accuracy than data models, and provide better information about the underlying mechanism.

12. FINAL REMARKS

The goals in statistics are to use data to predict and to get information about the underlying data mechanism. Nowhere is it written on a stone tablet what kind of model should be used to solve problems involving data. To make my position clear, I am not against data models per se. In some situations they are the most appropriate way to solve the problem. But the emphasis needs to be on the problem and the data.

Unfortunately, our field has a vested interest in data models, come hell or high water. For instance, see Dempster’s (1998) paper on modeling. His position on the 1990 Census adjustment controversy is particularly interesting. He admits that he doesn’t know much about the data or the details, but argues that the problem can be solved by a strong dose of modeling. That more modeling can make error-ridden data accurate seems highly unlikely to me.

Terrabytes of data are pouring into computers from many sources, both scientific, and commercial, and there is a need to analyze and understand the data. For instance, data is being generated at an awesome rate by telescopes and radio telescopes scanning the skies. Images containing millions of stellar objects are stored on tape or disk. Astronomers need automated ways to scan their data to find certain types of stellar objects or novel objects. This is a fascinating enterprise, and I doubt if data models are applicable. Yet I would enter this in my ledger as a statistical problem.

The analysis of genetic data is one of the most challenging and interesting statistical problems around. Microarray data, like that analyzed in Section 11.3 can lead to significant advances in understanding genetic effects. But the analysis of variable importance in Section 11.3 would be difficult to do accurately using a stochastic data model.

Problems such as stellar recognition or analysis of gene expression data could be high adventure for statisticians. But it requires that they focus on solving the problem instead of asking what data model they can create. The best solution could be an algorithmic model, or maybe a data model, or maybe a combination. But the trick to being a scientist is to be open to using a wide variety of tools.

The roots of statistics, as in science, lie in working with data and checking theory against data. I hope in this century our field will return to its roots. There are signs that this hope is not illusory. Over the last ten years, there has been a noticeable move toward statistical work on real world problems and reaching out by statisticians toward collaborative work with other disciplines. I believe this trend will continue and, in fact, has to continue if we are to survive as an energetic and creative field.

GLOSSARY

Since some of the terms used in this paper may not be familiar to all statisticians, I append some definitions.

Infinite test set error. Assume a loss function $L(y, \hat{y})$ that is a measure of the error when $y$ is the true response and $\hat{y}$ the predicted response. In classification, the usual loss is 1 if $y \neq \hat{y}$ and zero if $y = \hat{y}$. In regression, the usual loss is $(y - \hat{y})^2$. Given a set of data (training set) consisting of $\{(y_n, x_n) \mid n = 1, 2, \ldots, N\}$, use it to construct a predictor function $\phi(x)$ of $y$. Assume that the training set is i.i.d drawn from the distribution of the random vector $Y, X$. The infinite test set error is $E(L(Y, \phi(X)))$. This is called the generalization error in machine learning.

The generalization error is estimated either by setting aside a part of the data as a test set or by cross-validation.

Predictive accuracy. This refers to the size of the estimated generalization error. Good predictive accuracy means low estimated error.

Trees and nodes. This terminology refers to decision trees as described in the Breiman et al book (1984).

Dropping an $x$ down a tree. When a vector of predictor variables is “dropped” down a tree, at each intermediate node it has instructions whether to go left or right depending on the coordinates of $x$. It stops at a terminal node and is assigned the prediction given by that node.

Bagging. An acronym for “bootstrap aggregating.” Start with an algorithm such that given any training set, the algorithm produces a prediction function $\phi(x)$. The algorithm can be a decision tree construction, logistic regression with variable deletion, etc. Take a bootstrap sample from the training set and use this bootstrap training set to construct the predictor $\phi_1(x)$. Take another bootstrap sample and using this second training set construct the predictor $\phi_2(x)$. Continue this way for $K$ steps. In regression, average all of the $\{\phi_k(x)\}$ to get the
bagged predictor at \( x \). In classification, that class which has the plurality vote of the \( \{ \phi_k(x) \} \) is the bagged predictor. Bagging has been shown effective in variance reduction (Breiman, 1996b).

**Boosting.** This is a more complex way of forming an ensemble of predictors in classification than bagging (Freund and Schapire, 1996). It uses no randomization but proceeds by altering the weights on the training set. Its performance in terms of low prediction error is excellent (for details see Breiman, 1998).

**ACKNOWLEDGMENTS**

Many of my ideas about data modeling were formed in three decades of conversations with my old friend and collaborator, Jerome Friedman. Conversations with Richard Olshen about the Cox model and its use in biostatistics helped me to understand the background. I am also indebted to William Meisel, who headed some of the prediction projects I consulted on and helped me make the transition from probability theory to algorithms, and to Charles Stone for illuminating conversations about the nature of statistics and science. I’m grateful also for the comments of the editor, Leon Gleser, which prompted a major rewrite of the first draft of this manuscript and resulted in a different and better paper.

**REFERENCES**


D. R. Cox

Professor Breiman’s interesting paper gives both a clear statement of the broad approach underlying some of his influential and widely admired contributions and outlines some striking applications and developments. He has combined this with a critique of what, for want of a better term, I will call mainstream statistical thinking, based in part on a caricature. Like all good caricatures, it contains enough truth and exposes enough weaknesses to be thought-provoking.

There is not enough space to comment on all the many points explicitly or implicitly raised in the paper. There follow some remarks about a few main issues.

One of the attractions of our subject is the astonishingly wide range of applications as judged not only in terms of substantive field but also in terms of objectives, quality and quantity of data and so on. Thus any unqualified statement that “in applications...” has to be treated sceptically. One of our failings has, I believe, been, in a wish to stress generality, not to set out more clearly the distinctions between different kinds of application and the consequences for the strategy of statistical analysis. Of course we have distinctions between decision-making and inference, between tests and estimation, and between estimation and prediction and these are useful but, I think, are, except perhaps the first, too phrased in terms of the technology rather than the spirit of statistical analysis. I entirely agree with Professor Breiman that it would be an impoverished and extremely unhistorical view of the subject to exclude the kind of work he describes simply because it has no explicit probabilistic base.

Professor Breiman takes data as his starting point. I would prefer to start with an issue, a question or a scientific hypothesis, although I would be surprised if this were a real source of disagreement. These issues may evolve, or even change radically, as analysis proceeds. Data looking for a question are not unknown and raise puzzles but are, I believe, atypical in most contexts. Next, even if we ignore design aspects and start with data, key points concern the precise meaning of the data, the possible biases arising from the method of ascertainment, the possible presence of major distorting measurement errors and the nature of processes underlying missing and incomplete data and data that evolve in time in a way involving complex interdependencies. For some of these, at least, it is hard to see how to proceed without some notion of probabilistic modeling.

Next Professor Breiman emphasizes prediction as the objective, success at prediction being the criterion of success, as contrasted with issues of interpretation or understanding. Prediction is indeed important from several perspectives. The success of a theory is best judged from its ability to predict in new contexts, although one cannot dismiss as totally useless theories such as the rational action theory (RAT), in political science, which, as I understand it, gives excellent explanations of the past but which has failed to predict the real political world. In a clinical trial context it can be argued that an objective is to predict the consequences of treatment allocation to future patients, and so on.

If the prediction is localized to situations directly similar to those applying to the data there is then an interesting and challenging dilemma. Is it preferable to proceed with a directly empirical black-box approach, as favored by Professor Breiman, or is it better to try to take account of some underlying explanatory process? The answer must depend on the context but I certainly accept, although it goes somewhat against the grain to do so, that there are situations where a directly empirical approach is better. Short term economic forecasting and real-time flood forecasting are probably further exemplars. Key issues are then the stability of the predictor as practical prediction proceeds, the need from time to time for recalibration and so on.

However, much prediction is not like this. Often the prediction is under quite different conditions from the data; what is the likely progress of the incidence of the epidemic of v-CJD in the United Kingdom, what would be the effect on annual incidence of the epidemic of v-CJD in the United States of reducing by 10% the medical use of X-rays, etc.? That is, it may be desired to predict the consequences of something only indirectly addressed by the data available for analysis. As we move toward such more ambitious tasks, prediction, always hazardous, without some understanding of underlying process and linking with other sources of information, becomes more
and more tentative. Formulation of the goals of analysis solely in terms of direct prediction over the data set seems then increasingly unhelpful.

This is quite apart from matters where the direct objective is understanding of and tests of subject-matter hypotheses about underlying process, the nature of pathways of dependence and so on.

What is the central strategy of mainstream statistical analysis? This can most certainly not be discerned from the pages of Bernoulli, The Annals of Statistics or the Scandinavian Journal of Statistics nor from Biometrika and the Journal of Royal Statistical Society, Series B or even from the application pages of Journal of the American Statistical Association or Applied Statistics, estimable though all these journals are. Of course as we move along the list, there is an increase from zero to 100% in the papers containing analyses of “real” data. But the papers do so nearly always to illustrate technique rather than to explain the process of analysis and interpretation as such. This is entirely legitimate, but is completely different from live analysis of current data to obtain subject-matter conclusions or to help solve specific practical issues. Put differently, if an important conclusion is reached involving statistical analysis it will be reported in a subject-matter journal or in a written or verbal report to colleagues, government or business. When that happens, statistical details are typically and correctly not stressed. Thus the real procedures of statistical analysis can be judged only by looking in detail at specific cases, and access to these is not always easy. Failure to discuss enough the principles involved is a major criticism of the current state of theory.

I think tentatively that the following quite commonly applies. Formal models are useful and often almost, if not quite, essential for incisive thinking. Descriptively appealing and transparent methods with a firm model base are the ideal. Notions of significance tests, confidence intervals, posterior intervals and all the formal apparatus of inference are valuable tools to be used as guides, but not in a mechanical way; they indicate the uncertainty that would apply under somewhat idealized, may be very idealized, conditions and as such are often lower bounds to real uncertainty. Analyses and model development are at least partly exploratory. Automatic methods of model selection (and of variable selection in regression-like problems) are to be shunned or, if use is absolutely unavoidable, are to be examined carefully for their effect on the final conclusions. Unfocused tests of model adequacy are rarely helpful.

By contrast, Professor Breiman equates mainstream applied statistics to a relatively mechanical process involving somehow or other choosing a model, often a default model of standard form, and applying standard methods of analysis and goodness-of-fit procedures. Thus for survival data choose a priori the proportional hazards model. (Note, incidentally, that in the paper, often quoted but probably rarely read, that introduced this approach there was a comparison of several of the many different models that might be suitable for this kind of data.) It is true that many of the analyses done by nonstatisticians or by statisticians under severe time constraints are more or less like those Professor Breiman describes. The issue then is not whether they could ideally be improved, but whether they capture enough of the essence of the information in the data, together with some reasonable indication of precision as a guard against under or overinterpretation. Would more refined analysis, possibly with better predictive power and better fit, produce subject-matter gains? There can be no general answer to this, but one suspects that quite often the limitations of conclusions lie more in weakness of data quality and study design than in ineffective analysis.

There are two broad lines of development active at the moment arising out of mainstream statistical ideas. The first is the invention of models strongly tied to subject-matter considerations, representing underlying dependencies, and their analysis, perhaps by Markov chain Monte Carlo methods. In fields where subject-matter considerations are largely qualitative, we see a development based on Markov graphs and their generalizations. These methods in effect assume, subject in principle to empirical test, more and more about the phenomena under study. By contrast, there is an emphasis on assuming less and less via, for example, kernel estimates of regression functions, generalized additive models and so on. There is a need to be clearer about the circumstances favoring these two broad approaches, synthesizing them where possible.

My own interest tends to be in the former style of work. From this perspective Cox and Wermuth (1996, page 15) listed a number of requirements of a statistical model. These are to establish a link with background knowledge and to set up a connection with previous work, to give some pointer toward a generating process, to have primary parameters with individual clear subject-matter interpretations, to specify haphazard aspects well enough to lead to meaningful assessment of precision and, finally, that the fit should be adequate. From this perspective, fit, which is broadly related to predictive success, is not the primary basis for model choice and formal methods of model choice that take no account
of the broader objectives are suspect in the present context. In a sense these are efforts to establish data descriptions that are potentially causal, recognizing that causality, in the sense that a natural scientist would use the term, can rarely be established from one type of study and is at best somewhat tentative.

Professor Breiman takes a rather defeatist attitude toward attempts to formulate underlying processes; is this not to reject the base of much scientific progress? The interesting illustrations given by Beveridge (1952), where hypothesized processes in various biological contexts led to important progress, even though the hypotheses turned out in the end to be quite false, illustrate the subtlety of the matter. Especially in the social sciences, representations of underlying process have to be viewed with particular caution, but this does not make them fruitless.

The absolutely crucial issue in serious mainstream statistics is the choice of a model that will translate key subject-matter questions into a form for analysis and interpretation. If a simple standard model is adequate to answer the subject-matter question, this is fine: there are severe hidden penalties for overelaboration. The statistical literature, however, concentrates on how to do the analysis, an important and indeed fascinating question, but a secondary step. Better a rough answer to the right question than an exact answer to the wrong question, an aphorism, due perhaps to Lord Kelvin, that I heard as an undergraduate in applied mathematics.

I have stayed away from the detail of the paper but will comment on just one point, the interesting theorem of Vapnik about complete separation. This confirms folklore experience with empirical logistic regression that, with a largeish number of explanatory variables, complete separation is quite likely to occur. It is interesting that in mainstream thinking this is, I think, regarded as insecure in that complete separation is thought to be a priori unlikely and the estimated separating plane unstable. Presumably bootstrap and cross-validation ideas may give here a quite misleading illusion of stability. Of course if the complete separator is subtle and stable Professor Breiman’s methods will emerge triumphant and ultimately it is an empirical question in each application as to what happens.

It will be clear that while I disagree with the main thrust of Professor Breiman’s paper I found it stimulating and interesting.

Comment

Brad Efron

At first glance Leo Breiman’s stimulating paper looks like an argument against parsimony and scientific insight, and in favor of black boxes with lots of knobs to twiddle. At second glance it still looks that way, but the paper is stimulating, and Leo has some important points to hammer home. At the risk of distortion I will try to restate one of those points, the most interesting one in my opinion, using less confrontational and more historical language.

From the point of view of statistical development the twentieth century might be labeled “100 years of unbiasedness.” Following Fisher’s lead, most of our current statistical theory and practice revolves around unbiased or nearly unbiased estimates (particularly MLEs), and tests based on such estimates. The power of this theory has made statistics the dominant interpretational methodology in dozens of fields, but, as we say in California these days, it is power purchased at a price: the theory requires a modestly high ratio of signal to noise, sample size to number of unknown parameters, to have much hope of success. “Good experimental design” amounts to enforcing favorable conditions for unbiased estimation and testing, so that the statistician won’t find himself or herself facing 100 data points and 50 parameters.

Now it is the twenty-first century when, as the paper reminds us, we are being asked to face problems that never heard of good experimental design. Sample sizes have swollen alarmingly while goals grow less distinct (“find interesting data structure”). New algorithms have arisen to deal with new problems, a healthy sign it seems to me even if the innovators aren’t all professional statisticians. There are enough physicists to handle the physics case load, but there are fewer statisticians and more statistics problems, and we need all the help we can get. An
The attractive feature of Leo’s paper is his openness to new ideas whatever their source.

The new algorithms often appear in the form of black boxes with enormous numbers of adjustable parameters (“knobs to twiddle”), sometimes more knobs than data points. These algorithms can be quite successful as Leo points out, sometimes more so than their classical counterparts. However, unless the bias-variance trade-off has been suspended to encourage new statistical industries, their success must hinge on some form of biased estimation. The bias may be introduced directly as with the “regularization” of overparameterized linear models, more subtly as in the pruning of overgrown regression trees, or surreptitiously as with support vector machines, but it has to be lurking somewhere inside the theory.

Of course the trouble with biased estimation is that we have so little theory to fall back upon. Fisher’s information bound, which tells us how well a (nearly) unbiased estimator can possibly perform, is of no help at all in dealing with heavily biased methodology. Numerical experimentation by itself, unguided by theory, is prone to faddish wandering:

**Rule 1.** New methods always look better than old ones. Neural nets are better than logistic regression, support vector machines are better than neural nets, etc. In fact it is very difficult to run an honest simulation experiment, and easy to inadvertently cheat by choosing favorable examples, or by not putting as much effort into optimizing the dull old standard as the exciting new challenger.

**Rule 2.** Complicated methods are harder to criticize than simple ones. By now it is easy to check the efficiency of a logistic regression, but it is no small matter to analyze the limitations of a support vector machine. One of the best things statisticians do, and something that doesn’t happen outside our profession, is clarify the inferential basis of a proposed new methodology, a nice recent example being Friedman, Hastie, and Tibshirani’s analysis of “boosting,” (2000). The past half-century has seen the clarification process succeed successfully at work on nonparametrics, robustness and survival analysis. There has even been some success with biased estimation in the form of Stein shrinkage and empirical Bayes, but I believe the hardest part of this work remains to be done. Papers like Leo’s are a call for more analysis and theory, not less.

Prediction is certainly an interesting subject but Leo’s paper overstates both its role and our profession’s lack of interest in it.

- The “prediction culture,” at least around Stanford, is a lot bigger than 2%, though its constituency changes and most of us wouldn’t welcome being typecast.
- Estimation and testing are a form of prediction: “In our sample of 20 patients drug A outperformed drug B; would this still be true if we went on to test all possible patients?”
- Prediction by itself is only occasionally sufficient. The post office is happy with any method that predicts correct addresses from hand-written scrawls. Peter Gregory undertook his study for prediction purposes, but also to better understand the medical basis of hepatitis. Most statistical surveys have the identification of causal factors as their ultimate goal.

The hepatitis data was first analyzed by Gail Gong in her 1982 Ph.D. thesis, which concerned prediction problems and bootstrap methods for improving on cross-validation. (Cross-validation itself is an uncertain methodology that deserves further critical scrutiny; see, for example, Efron and Tibshirani, 1996). The *Scientific American* discussion is quite brief, a more thorough description appearing in Efron and Gong (1983). Variables 12 or 17 (13 or 18 in Efron and Gong’s numbering) appeared as “important” in 60% of the bootstrap simulations, which might be compared with the 59% for variable 19, the most for any single explanator.

In what sense are variable 12 or 17 or 19 “important” or “not important”? This is the kind of interesting inferential question raised by prediction methodology. Tibshirani and I made a stab at an answer in our 1998 *annals* paper. I believe that the current interest in statistical prediction will eventually invigorate traditional inference, not eliminate it.

A third front seems to have been opened in the long-running frequentist-Bayesian wars by the advocates of algorithmic prediction, who don’t really believe in any inferential school. Leo’s paper is at its best when presenting the successes of algorithmic modeling, which comes across as a positive development for both statistical practice and theoretical innovation. This isn’t an argument against traditional data modeling any more than splines are an argument against polynomials. The whole point of science is to open up black boxes, understand their insides, and build better boxes for the purposes of mankind. Leo himself is a notably successful scientist, so we can hope that the present paper was written more as an advocacy device than as the confessions of a born-again black boxist.
INTRODUCTION

Professor Breiman’s paper is an important one for statisticians to read. He and Statistical Science should be applauded for making this kind of material available to a large audience. His conclusions are consistent with how statistics is often practiced in business. This discussion will consist of an anecdotal recital of my encounters with the algorithmic modeling culture. Along the way, areas of mild disagreement with Professor Breiman are discussed. I also include a few proposals for research topics in algorithmic modeling.

CASE STUDY OF AN ALGORITHMIC MODELING CULTURE

Although I spent most of my career in management at Bell Labs and Bellcore, the last seven years have been with the research group at Fair, Isaac. This company provides all kinds of decision support solutions to several industries, and is very well known for credit scoring. Credit scoring is a great example of the problem discussed by Professor Breiman. The input variables, \( \mathbf{x} \), might come from company databases or credit bureaus. The output variable, \( y \), is some indicator of credit worthiness.

Credit scoring has been a profitable business for Fair, Isaac since the 1960s, so it is instructive to look at the Fair, Isaac analytic approach to see how it fits into the two cultures described by Professor Breiman. The Fair, Isaac approach was developed by engineers and operations research people and was driven by the needs of the clients and the quality of the data. The influences of the statistical community were mostly from the nonparametric side—things like jackknife and bootstrap.

Consider an example of behavior scoring, which is used in credit card account management. For pedagogical reasons, I consider a simplified version (in the real world, things get more complicated) of monthly behavior scoring. The input variables, \( \mathbf{x} \), in this simplified version, are the monthly bills and payments over the last 12 months. So the dimension of \( \mathbf{x} \) is 24. The output variable is binary and is the indicator of no severe delinquency over the next 6 months. The goal is to estimate the function, \( f(\mathbf{x}) = \log(\Pr(y = 1|\mathbf{x})/\Pr(y = 0|\mathbf{x})) \). Professor Breiman argues that some kind of simple logistic regression from the data modeling culture is not the way to solve this problem. I agree. Let’s take a look at how the engineers at Fair, Isaac solved this problem—way back in the 1960s and 1970s.

The general form used for \( f(\mathbf{x}) \) was called a segmented scorecard. The process for developing a segmented scorecard was clearly an algorithmic modeling process.

The first step was to transform \( \mathbf{x} \) into many interpretable variables called prediction characteristics. This was done in stages. The first stage was to compute several time series derived from the original two. An example is the time series of months delinquent—a nonlinear function. The second stage was to define characteristics as operators on the time series. For example, the number of times in the last six months that the customer was more than two months delinquent. This process can lead to thousands of characteristics. A subset of these characteristics passes a screen for further analysis.

The next step was to segment the population based on the screened characteristics. The segmentation was done somewhat informally. But when I looked at the process carefully, the segments turned out to be the leaves of a shallow-to-medium tree. And the tree was built sequentially using mostly binary splits based on the best splitting characteristics—defined in a reasonable way. The algorithm was manual, but similar in concept to the CART algorithm, with a different purity index.

Next, a separate function, \( f(\mathbf{x}) \), was developed for each segment. The function used was called a scorecard. Each characteristic was chopped up into discrete intervals or sets called attributes. A scorecard was a linear function of the attribute indicator (dummy) variables derived from the characteristics. The coefficients of the dummy variables were called score weights.

This construction amounted to an explosion of dimensionality. They started with 24 predictors. These were transformed into hundreds of characteristics and pared down to about 100 characteristics. Each characteristic was discretized into about 10 attributes, and there were about 10 segments. This makes \( 100 \times 10 \times 10 = 10,000 \) features. Yes indeed, dimensionality is a blessing.
What Fair, Isaac calls a scorecard is now elsewhere called a generalized additive model (GAM) with bin smoothing. However, a simple GAM would not do. Client demand, legal considerations and robustness over time led to the concept of score engineering. For example, the score had to be monotonically decreasing in certain delinquency characteristics. Prior judgment also played a role in the design of scorecards. For some characteristics, the score weights were shrunk toward zero in order to moderate the influence of these characteristics. For other characteristics, the score weights were expanded in order to increase the influence of these characteristics. These adjustments were not done willy-nilly. They were done to overcome known weaknesses in the data.

So how did these Fair, Isaac pioneers fit these complicated GAM models back in the 1960s and 1970s? Logistic regression was not generally available. And besides, even today’s commercial GAM software will not handle complex constraints. What they did was to maximize (subject to constraints) a measure called divergence, which measures how well the score, \( S \), separates the two populations with different values of \( y \). The formal definition of divergence is:

\[
2(E[S|y = 1] - E[S|y = 0])^2/(V[S|y = 1] + V[S|y = 0]).
\]

This constrained fitting was done with a heuristic nonlinear programming algorithm. A linear transformation was used to convert to a log odds scale.

Characteristic selection was done by analyzing the change in divergence after adding (removing) each candidate characteristic to (from) the current best model. The analysis was done informally to achieve good performance on the test sample. There were no formal tests of fit and no tests of score weight statistical significance. What counted was performance on the test sample, which was a surrogate for the future real world.

These early Fair, Isaac engineers were ahead of their time and charter members of the algorithmic modeling culture. The score formula was linear in an exploded dimension. A complex algorithm was used to fit the model. There was no claim that the final score formula was correct, only that it worked well on the test sample. This approach grew naturally out of the demands of the business and the quality of the data. The overarching goal was to develop tools that would help clients make better decisions through data. What emerged was a very accurate and palatable algorithmic modeling solution, which belies Breiman’s statement: “The algorithmic modeling methods available in the pre-1980s decades seem primitive now.” At a recent ASA meeting, I heard talks on treed regression, which looked like segmented scorecards to me.

After a few years with Fair, Isaac, I developed a talk entitled, “Credit Scoring—A Parallel Universe of Prediction and Classification.” The theme was that Fair, Isaac developed in parallel many of the concepts used in modern algorithmic modeling.

Certain aspects of the data modeling culture crept into the Fair, Isaac approach. The use of divergence was justified by assuming that the score distributions were approximately normal. So rather than making assumptions about the distribution of the inputs, they made assumptions about the distribution of the output. This assumption of normality was supported by a central limit theorem, which said that sums of many random variables are approximately normal—even when the component random variables are dependent and multiples of dummy random variables.

Modern algorithmic classification theory has shown that excellent classifiers have one thing in common, they all have large margin. Margin, \( M \), is a random variable that measures the comfort level with which classifications are made. When the correct classification is made, the margin is positive; it is negative otherwise. Since margin is a random variable, the precise definition of large margin is tricky. It does not mean that \( E[M] \) is large. When I put my data modeling hat on, I surmised that large margin means that \( E[M]/\sqrt{V(M)} \) is large. Lo and behold, with this definition, large margin means large divergence.

Since the good old days at Fair, Isaac, there have been many improvements in the algorithmic modeling approaches. We now use genetic algorithms to screen very large structured sets of prediction characteristics. Our segmentation algorithms have been automated to yield even more predictive systems. Our palatable GAM modeling tool now handles smooth splines, as well as splines mixed with step functions, with all kinds of constraint capability. Maximizing divergence is still a favorite, but we also maximize constrained GLM likelihood functions. We also are experimenting with computationally intensive algorithms that will optimize any objective function that makes sense in the business environment. All of these improvements are squarely in the culture of algorithmic modeling.

**OVERFITTING THE TEST SAMPLE**

Professor Breiman emphasizes the importance of performance on the test sample. However, this can be overdone. The test sample is supposed to represent the population to be encountered in the future. But in reality, it is usually a random sample of the
current population. High performance on the test sample does not guarantee high performance on future samples, things do change. There are practices that can be followed to protect against change.

One can monitor the performance of the models over time and develop new models when there has been sufficient degradation of performance. For some of Fair, Isaac’s core products, the redevelopment cycle is about 18–24 months. Fair, Isaac also does “score engineering” in an attempt to make the models more robust over time. This includes damping the influence of individual characteristics, using monotone constraints and minimizing the size of the models subject to performance constraints on the current test sample. This score engineering amounts to moving from very nonparametric (no score engineering) to more semiparametric (lots of score engineering).

**SPIN-OFFS FROM THE DATA MODELING CULTURE**

In Section 6 of Professor Breiman’s paper, he says that “multivariate analysis tools in statistics are frozen at discriminant analysis and logistic regression in classification . . . .” This is not necessarily all that bad. These tools can carry you very far as long as you ignore all of the textbook advice on how to use them. To illustrate, I use the saga of the Fat Scorecard.

Early in my research days at Fair, Isaac, I was searching for an improvement over segmented scorecards. The idea was to develop first a very good global scorecard and then to develop small adjustments for a number of overlapping segments. To develop the global scorecard, I decided to use logistic regression applied to the attribute dummy variables. There were 36 characteristics available for fitting. A typical scorecard has about 15 characteristics. My variable selection was structured so that an entire characteristic was either in or out of the model. What I discovered surprised me. All models fit with anywhere from 27 to 36 characteristics had the same performance on the test sample. This is what Professor Breiman calls “Rashomon and the multiplicity of good models.” To keep the model as small as possible, I chose the one with 27 characteristics. This model had 162 score weights (logistic regression coefficients), whose P-values ranged from 0.0001 to 0.984, with only one less than 0.05; i.e., statistically significant. The confidence intervals for the 162 score weights were useless. To get this great scorecard, I had to ignore the conventional wisdom on how to use logistic regression.

So far, all I had was the scorecard GAM. So clearly I was missing all of those interactions that just had to be in the model. To model the interactions, I tried developing small adjustments on various overlapping segments. No matter how hard I tried, nothing improved the test sample performance over the global scorecard. I started calling it the Fat Scorecard.

Earlier, on this same data set, another Fair, Isaac researcher had developed a neural network with 2,000 connection weights. The Fat Scorecard slightly outperformed the neural network on the test sample. I cannot claim that this would work for every data set. But for this data set, I had developed an excellent algorithmic model with a simple data modeling tool.

Why did the simple additive model work so well? One idea is that some of the characteristics in the model are acting as surrogates for certain interaction terms that are not explicitly in the model. Another reason is that the scorecard is really a sophisticated neural net. The inputs are the original inputs. Associated with each characteristic is a hidden node. The summation functions coming into the hidden nodes are the transformations defining the characteristics. The transfer functions of the hidden nodes are the step functions (compiled from the score weights)—all derived from the data. The final output is a linear function of the outputs of the hidden nodes. The result is highly nonlinear and interactive, when looked at as a function of the original inputs.

The Fat Scorecard study had an ingredient that is rare. We not only had the traditional test sample, but had three other test samples, taken one, two, and three years later. In this case, the Fat Scorecard outperformed the more traditional thinner scorecard for all four test samples. So the feared overfitting to the traditional test sample never materialized. To get a better handle on this you need an understanding of how the relationships between variables evolve over time.

I recently encountered another connection between algorithmic modeling and data modeling. In classical multivariate discriminant analysis, one assumes that the prediction variables have a multivariate normal distribution. But for a scorecard, the prediction variables are hundreds of attribute dummy variables, which are very nonnormal. However, if you apply the discriminant analysis algorithm to the attribute dummy variables, you can get a great algorithmic model, even though the assumptions of discriminant analysis are severely violated.
A SOLUTION TO THE OCCAM DILEMMA

I think that there is a solution to the Occam dilemma without resorting to goal-oriented arguments. Clients really do insist on interpretable functions, \( f(x) \). Segmented palatable scorecards are very interpretable by the customer and are very accurate. Professor Breiman himself gave single trees an A+ on interpretability. The shallow-to-medium tree in a segmented scorecard rates an A++. The palatable scorecards in the leaves of the trees are built from interpretable (possibly complex) characteristics. Sometimes we can't implement them until the lawyers and regulators approve. And that requires super interpretability. Our more sophisticated products have 10 to 20 segments and up to 100 characteristics (not all in every segment). These models are very accurate and very interpretable.

I coined a phrase called the “Ping-Pong theorem.” This theorem says that if we revealed to Professor Breiman the performance of our best model and gave him our data, then he could develop an algorithmic model using random forests, which would outperform our model. But if he revealed to us the performance of his model, then we could develop a segmented scorecard, which would outperform his model. We might need more characteristics, attributes and segments, but our experience in this kind of contest is on our side.

However, all the competing models in this game of Ping-Pong would surely be algorithmic models. But some of them could be interpretable.

THE ALGORITHM TUNING DILEMMA

As far as I can tell, all approaches to algorithmic model building contain tuning parameters, either explicit or implicit. For example, we use penalized objective functions for fitting and marginal contribution thresholds for characteristic selection. With experience, analysts learn how to set these tuning parameters in order to get excellent test sample or cross-validation results. However, in industry and academia, there is sometimes a little tinkering, which involves peeking at the test sample. The result is some bias in the test sample or cross-validation results. This is the same kind of tinkering that upsets test of fit pureness. This is a challenge for the algorithmic modeling approach. How do you optimize your results and get an unbiased estimate of the generalization error?

GENERALIZING THE GENERALIZATION ERROR

In most commercial applications of algorithmic modeling, the function, \( f(x) \), is used as a surrogate for the decision process, and misclassification error is used as a surrogate for profit. However, I see a mismatch between the algorithms used to develop the models and the business measurement of the model’s value. For example, at Fair, Isaac, we frequently maximize divergence. But when we argue the model’s value to the clients, we don’t necessarily brag about the great divergence. We try to use measures that the client can relate to. The ROC curve is one favorite, but it may not tell the whole story. Sometimes, we develop simulations of the client’s business operation to show how the model will improve their situation. For example, in a transaction fraud control process, some measures of interest are false positive rate, speed of detection and dollars saved when 0.5% of the transactions are flagged as possible frauds. The 0.5% reflects the number of transactions that can be processed by the current fraud management staff. Perhaps what the client really wants is a score that will maximize the dollars saved in their fraud control system. The score that maximizes test set divergence or minimizes test set misclassifications does not do it. The challenge for algorithmic modeling is to find an algorithm that maximizes the generalization dollars saved, not generalization error.

We have made some progress in this area using ideas from support vector machines and boosting. By manipulating the observation weights used in standard algorithms, we can improve the test set performance on any objective of interest. But the price we pay is computational intensity.

MEASURING IMPORTANCE—IS IT REALLY POSSIBLE?

I like Professor Breiman’s idea for measuring the importance of variables in black box models. A Fair, Isaac spin on this idea would be to build accurate models for which no variable is much more important than other variables. There is always a chance that a variable and its relationships will change in the future. After that, you still want the model to work. So don’t make any variable dominant.

I think that there is still an issue with measuring importance. Consider a set of inputs and an algorithm that yields a black box, for which \( x_1 \) is important. From the “Ping Pong theorem” there exists a set of input variables, excluding \( x_1 \) and an algorithm that will yield an equally accurate black box. For this black box, \( x_1 \) is unimportant.
IN SUMMARY

Algorithmic modeling is a very important area of statistics. It has evolved naturally in environments with lots of data and lots of decisions. But you can do it without suffering the Occam dilemma; for example, use medium trees with interpretable GAMs in the leaves. They are very accurate and interpretable. And you can do it with data modeling tools as long as you (i) ignore most textbook advice, (ii) embrace the blessing of dimensionality, (iii) use constraints in the fitting optimizations (iv) use regularization, and (v) validate the results.

Comment

Emanuel Parzen

1. BREIMAN DESERVES OUR APPRECIATION

I strongly support the view that statisticians must face the crisis of the difficulties in their practice of regression. Breiman alerts us to systematic blunders (leading to wrong conclusions) that have been committed applying current statistical practice of data modeling. In the spirit of “statistician, avoid doing harm” I propose that the first goal of statistical ethics should be to guarantee to our clients that any mistakes in our analysis are unlike any mistakes that statisticians have made before.

The two goals in analyzing data which Leo calls prediction and information I prefer to describe as “management” and “science.” Management seeks profit, practical answers (predictions) useful for decision making in the short run. Science seeks truth, fundamental knowledge about nature which provides understanding and control in the long run. As a historical note, Student’s $t$-test has many scientific applications but was invented by Student as a management tool to make Guinness beer better (bitter?).

Breiman does an excellent job of presenting the case that the practice of statistical science, using only the conventional data modeling culture, needs reform. He deserves much thanks for alerting us to the algorithmic modeling culture. Breiman warns us that “if the model is a poor emulation of nature, the conclusions may be wrong.” This situation, which I call “the right answer to the wrong question,” is called by statisticians “the error of the third kind.” Engineers at M.I.T. define “suboptimization” as “elegantly solving the wrong problem.”

Breiman presents the potential benefits of algorithmic models (better predictive accuracy than data models, and consequently better information about the underlying mechanism and avoiding questionable conclusions which results from weak predictive accuracy) and support vector machines (which provide almost perfect separation and discrimination between two classes by increasing the dimension of the feature set). He convinces me that the methods of algorithmic modeling are important contributions to the tool kit of statisticians.

If the profession of statistics is to remain healthy, and not limit its research opportunities, statisticians must learn about the cultures in which Breiman works, but also about many other cultures of statistics.

2. HYPOTHESES TO TEST TO AVOID BLUNDERS OF STATISTICAL MODELING

Breiman deserves our appreciation for pointing out generic deviations from standard assumptions (which I call bivariate dependence and two-sample conditional clustering) for which we should routinely check. “Test null hypothesis” can be a useful algorithmic concept if we use tests that diagnose in a model-free way the directions of deviation from the null hypothesis model.

Bivariate dependence (correlation) may exist between features [independent (input) variables] in a regression causing them to be proxies for each other and our models to be unstable with different forms of regression models being equally well fitting. We need tools to routinely test the hypothesis of statistical independence of the distributions of independent (input) variables.

Two sample conditional clustering arises in the distributions of independent (input) variables to discriminate between two classes, which we call the conditional distribution of input variables $X$ given each class. Class I may have only one mode (cluster) at low values of $X$ while class II has two modes.

Emanuel Parzen is Distinguished Professor, Department of Statistics, Texas A&M University, 415 C Block Building, College Station, Texas 77843 (e-mail: eparzen@stat.tamu.edu).
(clusters) at low and high values of $X$. We would like to conclude that high values of $X$ are observed only for members of class II but low values of $X$ occur for members of both classes. The hypothesis we propose for testing is equality of the pooled distribution of both samples and the conditional distribution of sample I, which is equivalent to $P$[class $I | X] = P$[class $I$]. For successful discrimination one seeks to increase the number (dimension) of inputs (features) $X$ to make $P$[class $I | X]$ close to 1 or 0.

3. STATISTICAL MODELING, MANY CULTURES, STATISTICAL METHODS MINING

Breiman speaks of two cultures of statistics; I believe statistics has many cultures. At specialized workshops (on maximum entropy methods or robust methods or Bayesian methods or . . .) a main topic of conversation is “Why don’t all statisticians think like us?”

I have my own eclectic philosophy of statistical modeling to which I would like to attract serious attention. I call it “statistical methods mining” which seeks to provide a framework to synthesize and apply the past half-century of methodological progress in computationally intensive methods for statistical modeling, including EDA (exploratory data analysis), FDA (functional data analysis), density estimation, Model DA (model selection criteria data analysis), Bayesian priors on function space, continuous parameter regression analysis and reproducing kernels, fast algorithms, Kalman filtering, complexity, information, quantile data analysis, nonparametric regression, conditional quantiles.

I believe “data mining” is a special case of “data modeling.” We should teach in our introductory courses that one meaning of statistics is “statistical data modeling done in a systematic way” by an iterated series of stages which can be abbreviated SIEVE (specify problem and general form of models, identify tentatively numbers of parameters and specialized models, estimate parameters, validate goodness-of-fit of estimated models, estimate final model nonparametrically or algorithmically). MacKay and Oldford (2000) brilliantly present the statistical method as a series of stages PPDAC (problem, plan, data, analysis, conclusions).

4. QUANTILE CULTURE, ALGORITHMIC MODELS

A culture of statistical data modeling based on quantile functions, initiated in Parzen (1979), has been my main research interest since 1976. In my discussion to Stone (1977) I outlined a novel approach to estimation of conditional quantile functions which I only recently fully implemented. I would like to extend the concept of algorithmic statistical models in two ways: (1) to mean data fitting by representations which use approximation theory and numerical analysis; (2) to use the notation of probability to describe empirical distributions of samples (data sets) which are not assumed to be generated by a random mechanism.

My quantile culture has not yet become widely applied because “you cannot give away a good idea, you have to sell it” (by integrating it in computer programs usable by applied statisticians and thus promote statistical methods mining).

A quantile function $Q(u)$, $0 \leq u \leq 1$, is the inverse $F^{-1}(u)$ of a distribution function $F(x)$, $-\infty < x < \infty$. Its rigorous definition is $Q(u) = \inf(x: F(x) \geq u)$. When $F$ is continuous with density $f$, $P(Q(u)) = u$, $q(u) = Q(u) = 1/f(Q(u))$. We use the notation $Q$ for a true unknown quantile function, $Q$ for a raw estimator from a sample, and $Q$ for a smooth estimaor of the true $Q$.

Concepts defined for $Q(u)$ can be defined also for other versions of quantile functions. Quantile functions can “compress data” by a five-number summary, values of $Q(u)$ at $u = 0.05, 0.25, 0.75, 0.1, 0.9$ (or 0.05, 0.95). Measures of location and scale are $QM = 0.5(Q(0.25) + Q(0.75))$, $QD = 2(Q(0.75) - Q(0.25))$. To use quantile functions to identify distributions fitting data we propose the quantile—quantile function $Q/Q(u) = (Q(u) - QM)/QD$. Five-number summary of distribution becomes $QM, QD, Q/Q(0.5)$ skewness, $Q/Q(0.1)$ left-tail, $Q/Q(0.9)$ right-tail. Elegance of $Q/Q(u)$ is its universal values at $u = 0.25, 0.75$. Values $|Q/Q(u)| > 1$ are outliers as defined by Tukey EDA.

For the fundamental problem of comparison of two distributions $F$ and $G$ we define the comparison distribution $D(u; F, G)$ and comparison density $d(u; F, G) = D(u; F, G)$. For $F, G$ continuous, define $D(u; F, G) = G(F^{-1}(u))$, $d(u; F, G) = g(F^{-1}(u))/f(F^{-1}(u))$ assuming $f(x) = 0$ implies $g(x) = 0$, written $G \ll F$. For $F, G$ discrete with probability mass functions $p_F$ and $p_G$ define (assuming $G \ll F$) $d(u; F, G) = p_G(F^{-1}(u))/p_F(F^{-1}(u))$.

Our applications of comparison distributions often assume $F$ to be an unconditional distribution and $G$ a conditional distribution. To analyze bivariate data $(X, Y)$ a fundamental tool is dependence density $d(t, u) = d(u; F_Y, F_Y|X=q_Y(t))$. When $X, Y$ is jointly continuous,

$$d(t, u) = f_X(Q_X(t), Q_Y(u))/f_X(Q_X(t))f_Y(Q_Y(u)).$$
The statistical independence hypothesis \( F_{XY} = F_X F_Y \) is equivalent to \( d(t, u) = 1, \) \( t, u. \) A fundamental formula for estimation of conditional quantile functions is
\[
Q_{Y|X=x}(u) = Q_Y(D^{-1}(u; F_Y, F_{Y|X=x})) = Q_Y(s), u = D(s; F_Y, F_{Y|X=x}).
\]

To compare the distributions of two univariate samples, let \( Y \) denote the continuous response variable and \( X \) be binary 0, 1 denoting the population from which \( Y \) is observed. The comparison density is defined (note \( F_Y \) is the pooled distribution function)
\[
d_1(u) = d(u; F_Y, F_{Y|X=1}) = P[X = 1|Y = Q_Y(u)]/P[X = 1].
\]

5. QUANTILE IDEAS FOR HIGH DIMENSIONAL DATA ANALYSIS

By high dimensional data we mean multivariate data \((Y_1, \ldots, Y_m)\). We form approximate high dimensional comparison densities \( d(u_1, \ldots, u_m) \) to test statistical independence of the variables and, when we have two samples, \( d_1(u_1, \ldots, u_m) \) to test equality of sample \( I \) with pooled sample. All our distributions are empirical distributions but we use notation for true distributions in our formulas. Note that
\[
\int_0^1 du_1 \ldots \int_0^1 du_m d(u_1, \ldots, u_m) = 1.
\]

A decile quantile bin \( B(k_1, \ldots, k_m) \) is defined to be the set of observations \((Y_1, \ldots, Y_m)\) satisfying, for \( j = 1, \ldots, m, Q_{Y_j}((k_j - 1)/10) < Y_j \leq Q_{Y_j}(k_j/10) \). Instead of deciles \( k/10 \) we could use \( k/M \) for another base \( M. \)

To test the hypothesis that \( Y_1, \ldots, Y_m \) are statistically independent we form for all \( k_j = 1, \ldots, 10, \)
\[
d(k_1, \ldots, k_m) = P[Bin(k_1, \ldots, k_m)]/P[Bin(k_1, \ldots, k_m)|\text{independence}].
\]

To test equality of distribution of a sample from population \( I \) and the pooled sample we form
\[
d_1(k_1, \ldots, k_m) = P[Bin(k_1, \ldots, k_m)|\text{population } I]/P[Bin(k_1, \ldots, k_m)|\text{pooled sample}]
\]

for all \((k_1, \ldots, k_m)\) such that the denominator is positive and otherwise defined arbitrarily. One can show (letting \( X \) denote the population observed)
\[
d_1(k_1, \ldots, k_m) = P[X = I|\text{observation from Bin}(k_1, \ldots, k_m)]/P[X = I].
\]

To test the null hypotheses in ways that detect directions of deviations from the null hypothesis our recommended first step is quantile data analysis of the values \( d(k_1, \ldots, k_m) \) and \( d_1(k_1, \ldots, k_m). \)

I appreciate this opportunity to bring to the attention of researchers on high dimensional data analysis the potential of quantile methods. My conclusion is that statistical science has many cultures and statisticians will be more successful when they emulate Leo Breiman and apply as many cultures as possible (which I call statistical methods mining). Additional references are on my web site at stat.tamu.edu.

Rejoinder

Leo Breiman

I thank the discussants. I’m fortunate to have comments from a group of experienced and creative statisticians—even more so in that their comments are diverse. Manny Parzen and Bruce Hoadley are more or less in agreement, Brad Efron has serious reservations and D. R. Cox is in downright disagreement.

I address Professor Cox’s comments first, since our disagreement is crucial.

D. R. COX

Professor Cox is a worthy and thoughtful adversary. We walk down part of the trail together and then sharply diverge. To begin, I quote: “Professor Breiman takes data as his starting point. I would prefer to start with an issue, a question, or a scientific hypothesis,...” I agree, but would expand the starting list to include the prediction of future events. I have never worked on a project that has started with “Here is a lot of data; let’s look at it and see if we can get some ideas about how to use it.” The data has been put together and analyzed starting with an objective.

C1 Data Models Can Be Useful

Professor Cox is committed to the use of data models. I readily acknowledge that there are situations
where a simple data model may be useful and appropriate; for instance, if the science of the mechanism producing the data is well enough known to determine the model apart from estimating parameters. There are also situations of great complexity posing important issues and questions in which there is not enough data to resolve the questions to the accuracy desired. Simple models can then be useful in giving qualitative understanding, suggesting future research areas and the kind of additional data that needs to be gathered.

At times, there is not enough data on which to base predictions; but policy decisions need to be made. In this case, constructing a model using whatever data exists, combined with scientific common sense and subject-matter knowledge, is a reasonable path. Professor Cox points to examples when he writes:

"Often the prediction is under quite different conditions from the data; what is the likely progress of the incidence of the epidemic of v-CJD in the United Kingdom, what would be the effect on annual incidence of cancer in the United States reducing by 10% the medical use of X-rays, etc.? That is, it may be desired to predict the consequences of something only indirectly addressed by the data available for analysis... prediction, always hazardous, without some understanding of the underlying process and linking with other sources of information, becomes more and more tentative."

I agree.

**C2 Data Models Only**

From here on we part company. Professor Cox's discussion consists of a justification of the use of data models to the exclusion of other approaches. For instance, although he admits, "... I certainly accept, although it goes somewhat against the grain to do so, that there are situations where a directly empirical approach is better..." the two examples he gives of such situations are short-term economic forecasts and real-time flood forecasts—among the less interesting of all of the many current successful algorithmic applications. In his view, the only use for algorithmic models is short-term forecasting; there are no comments on the rich information about the data and covariates available from random forests or in the many fields, such as pattern recognition, where algorithmic modeling is fundamental.

He advocates construction of stochastic data models that summarize the understanding of the phenomena under study. The methodology in the Cox and Wermuth book (1996) attempts to push understanding further by finding casual orderings in the covariate effects. The sixth chapter of this book contains illustrations of this approach on four data sets.

The first is a small data set of 68 patients with seven covariates from a pilot study at the University of Mainz to identify psychological and socioeconomic factors possibly important for glucose control in diabetes patients. This is a regression-type problem with the response variable measured by GHb (glycosylated haemoglobin). The model fitting is done by a number of linear regressions and validated by the checking of various residual plots. The only other reference to model validation is the statement, "$R^2 = 0.34$, reasonably large by the standards usual for this field of study." Predictive accuracy is not computed, either for this example or for the three other examples.

My comments on the questionable use of data models apply to this analysis. Incidentally, I tried to get one of the data sets used in the chapter to conduct an alternative analysis, but it was not possible to get it before my rejoinder was due. It would have been interesting to contrast our two approaches.

**C3 Approach to Statistical Problems**

Basing my critique on a small illustration in a book is not fair to Professor Cox. To be fairer, I quote his words about the nature of a statistical analysis:

> "Formal models are useful and often almost, if not quite, essential for incisive thinking. Descriptively appealing and transparent methods with a firm model base are the ideal. Notions of significance tests, confidence intervals, posterior intervals, and all the formal apparatus of inference are valuable tools to be used as guides, but not in a mechanical way; they indicate the uncertainty that would apply under somewhat idealized, maybe very idealized, conditions and as such are often lower bounds to real uncertainty. Analyses and model development are at least partly exploratory. Automatic methods of model selection (and of variable selection in regression-like problems) are to be shunned or, if use is absolutely unavoidable, are to be examined carefully for their effect on the final conclusions. Unfocused tests of model adequacy are rarely helpful."
Given the right kind of data: relatively small sample size and a handful of covariates, I have no doubt that his experience and ingenuity in the craft of model construction would result in an illuminating model. But data characteristics are rapidly changing. In many of the most interesting current problems, the idea of starting with a formal model is not tenable.

C4 Changes in Problems

My impression from Professor Cox’s comments is that he believes every statistical problem can be best solved by constructing a data model. I believe that statisticians need to be more pragmatic. Given a statistical problem, find a good solution, whether it is a data model, an algorithmic model or (although it is somewhat against my grain), a Bayesian data model or a completely different approach.

My work on the 1990 Census Adjustment (Breiman, 1994) involved a painstaking analysis of the sources of error in the data. This was done by a long study of thousands of pages of evaluation documents. This seemed the most appropriate way of answering the question of the accuracy of the adjustment estimates.

The conclusion that the adjustment estimates were largely the result of bad data has never been effectively contested and is supported by the results of the Year 2000 Census Adjustment effort. The accuracy of the adjustment estimates was, arguably, the most important statistical issue of the last decade, and could not be resolved by any amount of statistical modeling.

A primary reason why we cannot rely on data models alone is the rapid change in the nature of statistical problems. The realm of applications of statistics has expanded more in the last twenty-five years than in any comparable period in the history of statistics.

In an astronomy and statistics workshop this year, a speaker remarked that in twenty-five years we have gone from being a small sample-size science to a very large sample-size science. Astronomical data bases now contain data on two billion objects comprising over 100 terabytes and the rate of new information is accelerating.

A recent biostatistics workshop emphasized the analysis of genetic data. An exciting breakthrough is the use of microarrays to locate regions of gene activity. Here the sample size is small, but the number of variables ranges in the thousands. The questions are which specific genes contribute to the occurrence of various types of diseases.

Questions about the areas of thinking in the brain are being studied using functional MRI. The data gathered in each run consists of a sequence of 150,000 pixel images. Gigabytes of satellite information are being used in projects to predict and understand short- and long-term environmental and weather changes.

Underlying this rapid change is the rapid evolution of the computer, a device for gathering, storing and manipulation of incredible amounts of data, together with technological advances incorporating computing, such as satellites and MRI machines.

The problems are exhilarating. The methods used in statistics for small sample sizes and a small number of variables are not applicable. John Rice, in his summary talk at the astronomy and statistics workshop said, “Statisticians have to become opportunistic.” That is, faced with a problem, they must find a reasonable solution by whatever method works. One surprising aspect of both workshops was how opportunistic statisticians faced with genetic and astronomical data had become. Algorithmic methods abounded.

C5 Mainstream Procedures and Tools

Professor Cox views my critique of the use of data models as based in part on a caricature. Regarding my references to articles in journals such as JASA, he states that they are not typical of mainstream statistical analysis, but are used to illustrate technique rather than explain the process of analysis. His concept of mainstream statistical analysis is summarized in the quote given in my Section C3. It is the kind of thoughtful and careful analysis that he prefers and is capable of.

Following this summary is the statement:

By contrast, Professor Breiman equates mainstream applied statistics to a relatively mechanical process involving somehow or other choosing a model, often a default model of standard form, and applying standard methods of analysis and goodness-of-fit procedures.

The disagreement is definitional—what is “mainstream”? In terms of numbers my definition of mainstream prevails, I guess, at a ratio of at least 100 to 1. Simply count the number of people doing their statistical analysis using canned packages, or count the number of SAS licenses.

In the academic world, we often overlook the fact that we are a small slice of all statisticians and an even smaller slice of all those doing analyses of data. There are many statisticians and nonstatisticians in diverse fields using data to reach conclusions and depending on tools supplied to them.
by SAS, SPSS, etc. Their conclusions are important and are sometimes published in medical or other subject-matter journals. They do not have the statistical expertise, computer skills, or time needed to construct more appropriate tools. I was faced with this problem as a consultant when confined to using the BMDP linear regression, stepwise linear regression, and discriminant analysis programs. My concept of decision trees arose when I was faced with nonstandard data that could not be treated by these standard methods.

When I rejoined the university after my consulting years, one of my hopes was to provide better general purpose tools for the analysis of data. The first step in this direction was the publication of the CART book (Breiman et al., 1984). CART and other similar decision tree methods are used in thousands of applications yearly in many fields. It has proved robust and reliable. There are others that are more recent; random forests is the latest. A preliminary version of random forests is free source with \texttt{/77} code, \texttt{S+} and \texttt{R} interfaces available at www.stat.berkeley.edu/users/breiman.

A nearly completed second version will also be put on the web site and translated into Java by the Weka group. My collaborator, Adele Cutler, and I will continue to upgrade, add new features, graphics, and a good interface.

My philosophy about the field of academic statistics is that we have a responsibility to provide the many people working in applications outside of academia with useful, reliable, and accurate analysis tools. Two excellent examples are wavelets and decision trees. More are needed.

**BRAD EFRON**

Brad seems to be a bit puzzled about how to react to my article. I'll start with what appears to be his biggest reservation.

**E1 From Simple to Complex Models**

Brad is concerned about the use of complex models without simple interpretability in their structure, even though these models may be the most accurate predictors possible. But the evolution of science is from simple to complex.

The equations of general relativity are considerably more complex and difficult to understand than Newton's equations. The quantum mechanical equations for a system of molecules are extraordinarily difficult to interpret. Physicists accept these complex models as the facts of life, and do their best to extract usable information from them.

There is no consideration given to trying to understand cosmology on the basis of Newton's equations or nuclear reactions in terms of hard ball models for atoms. The scientific approach is to use these complex models as the best possible descriptions of the physical world and try to get usable information out of them.

There are many engineering and scientific applications where simpler models, such as Newton's laws, are certainly sufficient—say, in structural design. Even here, for larger structures, the model is complex and the analysis difficult. In scientific fields outside statistics, answering questions is done by extracting information from increasingly complex and accurate models.

The approach I suggest is similar. In genetics, astronomy and many other current areas statistics is needed to answer questions, construct the most accurate possible model, however complex, and then extract usable information from it.

Random forests is in use at some major drug companies whose statisticians were impressed by its ability to determine gene expression (variable importance) in microarray data. They were not concerned about its complexity or black-box appearance.

**E2 Prediction**

Leo's paper overstates both its [prediction's] role, and our profession's lack of interest in it... Most statistical surveys have the identification of causal factors as their ultimate role.

My point was that it is difficult to tell, using goodness-of-fit tests and residual analysis, how well a model fits the data. An estimate of its test set accuracy is a preferable assessment. If, for instance, a model gives predictive accuracy only slightly better than the "all survived" or other baseline estimates, we can't put much faith in its reliability in the identification of causal factors.

I agree that often "... statistical surveys have the identification of casual factors as their ultimate role." I would add that the more predictively accurate the model is, the more faith can be put into the variables that it fingers as important.

**E3 Variable Importance**

A significant and often overlooked point raised by Brad is what meaning can one give to statements that "variable \(X\) is important or not important." This has puzzled me on and off for quite a while. In fact, variable importance has always been defined operationally. In regression the "important"
variables are defined by doing “best subsets” or variable deletion.

Another approach used in linear methods such as logistic regression and survival models is to compare the size of the slope estimate for a variable to its estimated standard error. The larger the ratio, the more “important” the variable. Both of these definitions can lead to erroneous conclusions.

My definition of variable importance is based on prediction. A variable might be considered important if deleting it seriously affects prediction accuracy. This brings up the problem that if two variables are highly correlated, deleting one or the other of them will not affect prediction accuracy. Deleting both of them may degrade accuracy considerably. The definition used in random forests spots both variables.

“Importance” does not yet have a satisfactory theoretical definition (I haven’t been able to locate the article Brad references but I’ll keep looking). It depends on the dependencies between the output variable and the input variables, and on the dependencies between the input variables. The problem begs for research.

E4 Other Reservations

Sample sizes have swollen alarmingly while goals grow less distinct (“find interesting data structure”).

I have not noticed any increasing fuzziness in goals, only that they have gotten more diverse. In the last two workshops I attended (genetics and astronomy) the goals in using the data were clearly laid out. “Searching for structure” is rarely seen even though data may be in the terabyte range.

The new algorithms often appear in the form of black boxes with enormous numbers of adjustable parameters (“knobs to twiddle”).

This is a perplexing statement and perhaps I don’t understand what Brad means. Random forests has only one adjustable parameter that needs to be set for a run, is insensitive to the value of this parameter over a wide range, and has a quick and simple way for determining a good value. Support vector machines depend on the settings of 1–2 parameters. Other algorithmic models are similarly sparse in the number of knobs that have to be twiddled.

New methods always look better than old ones. . . . Complicated models are harder to criticize than simple ones.

In 1992 I went to my first NIPS conference. At that time, the exciting algorithmic methodology was neural nets. My attitude was grim skepticism. Neural nets had been given too much hype, just as AI had been given and failed expectations. I came away a believer. Neural nets delivered on the bottom line! In talk after talk, in problem after problem, neural nets were being used to solve difficult prediction problems with test set accuracies better than anything I had seen up to that time.

My attitude toward new and/or complicated methods is pragmatic. Prove that you’ve got a better mousetrap and I’ll buy it. But the proof had better be concrete and convincing.

Brad questions where the bias and variance have gone. It is surprising when, trained in classical bias-variance terms and convinced of the curse of dimensionality, one encounters methods that can handle thousands of variables with little loss of accuracy. It is not voodoo statistics; there is some simple theory that illuminates the behavior of random forests (Breiman, 1999). I agree that more theoretical work is needed to increase our understanding.

Brad is an innovative and flexible thinker who has contributed much to our field. He is opportunistic in problem solving and may, perhaps not overtly, already have algorithmic modeling in his bag of tools.

BRUCE HOADLEY

I thank Bruce Hoadley for his description of the algorithmic procedures developed at Fair, Isaac since the 1960s. They sound like people I would enjoy working with. Bruce makes two points of mild contention. One is the following:

High performance (predictive accuracy) on the test sample does not guarantee high performance on future samples; things do change.

I agree—algorithmic models accurate in one context must be modified to stay accurate in others. This does not necessarily imply that the way the model is constructed needs to be altered, but that data gathered in the new context should be used in the construction.

His other point of contention is that the Fair, Isaac algorithm retains interpretability, so that it is possible to have both accuracy and interpretability. For clients who like to know what’s going on, that’s a sellable item. But developments in algorithmic modeling indicate that the Fair, Isaac algorithm is an exception.

A computer scientist working in the machine learning area joined a large money management
company some years ago and set up a group to do portfolio management using stock predictions given by large neural nets. When we visited, I asked how he explained the neural nets to clients. “Simple,” he said; “We fit binary trees to the inputs and outputs of the neural nets and show the trees to the clients. Keeps them happy!” In both stock prediction and credit rating, the priority is accuracy. Interpretability is a secondary goal that can be finessed.

**MANNY PARZEN**

Manny Parzen opines that there are not two but many modeling cultures. This is not an issue I want to fiercely contest. I like my division because it is pretty clear cut—are you modeling the inside of the box or not? For instance, I would include Bayesians in the data modeling culture. I will keep my eye on the quantile culture to see what develops.

Most of all, I appreciate Manny’s openness to the issues raised in my paper. With the rapid changes in the scope of statistical problems, more open and concrete discussion of what works and what doesn’t should be welcomed.

**WHERE ARE WE HEADING?**

Many of the best statisticians I have talked to over the past years have serious concerns about the viability of statistics as a field. Oddly, we are in a period where there has never been such a wealth of new statistical problems and sources of data. The danger is that if we define the boundaries of our field in terms of familiar tools and familiar problems, we will fail to grasp the new opportunities.

**ADDITIONAL REFERENCES**


For all the talk about how computers allow for new levels of scale in humanities research, new debates over institutional structures, and new claims to scientific rigor, it is easy to lose sight of the radical difference between the way human beings and computer programs “read” texts. Topic modeling, one of the most touted methods of finding patterns in large corpora, relies on a procedure that has little resemblance to anything a human being could do. Each text is converted into a matrix of word frequencies, transforming it into an entirely numerical dataset. The computer is directed to create a set of probability tables populated with random numbers, and then it gradually refines them by computing the same pair of mathematical functions hundreds or thousands of times in a row. After a few billion or perhaps even a few trillion multiplications, additions, and other algebraic operations, it sutures words back onto this numerical structure and presents them in a conveniently sorted form. This output, like the paper spit out by a fortune-telling machine, is supposed to tell us the “themes” of the texts being analyzed. While some of the earliest computational text-analysis projects, like Father Roberto Busa’s famous collaboration with IBM on the Index Thomisticus, began by attempting to automate procedures that scholars had already been doing for centuries, topic modeling takes us well beyond the mechanical imitation of human action (Hockey). When we incorporate text-mining software into our scholarly work, machines are altering our interpretive acts in altogether unprecedented ways.

Yet, as Alan Liu has argued, there has been relatively little interchange between the scholars who are applying these computational methods to literary history and those in fields like media studies who critically examine the history and culture from which this computational technology emerged (“Where Is Cultural Criticism in the Digital Humanities?”). Many scholars of technology, including Lisa Gitelman, Wendy Hui Kyong Chun, Tara McPherson, and David Golumbia, have argued that the seemingly abstract structures of computation can serve ideological ends; but scholars who apply text mining to literary and cultural history have largely skirted the question of how the technologies they use might be influenced by the military and commercial contexts from which they emerged (Gitelman, Paper Knowledge; Chun, Control and Freedom; McPherson, “Why Are the Digital Humanities So White?”; Golumbia, Cultural Logic of Computation). As a way of gesturing toward a fuller understanding of the cultural context surrounding text-mining methods, I will give a brief
account of the origins of a popular technique for topic modeling, Latent Dirichlet Allocation (LDA), and attempt to situate text mining in a broader history of thinking about language. I identify a congruity between text mining and the language standardization efforts that began in the seventeenth and eighteenth centuries, when authors such as John Locke called for the stabilization of vocabularies and devalued “literary” dimensions of language such as metaphor, wordplay, and innuendo as impediments to communication. I argue that, when applied to the study of literary and cultural texts, statistical text-mining methods tend to reinforce conceptions of language and meaning that are, at best, overly dependent on the “literal” definitions of words and, at worst, complicit in the marginalization of nonstandard linguistic conventions and modes of expression.

While text-mining methods could potentially give us an ideologically skewed picture of literary and cultural history, a shift toward a media studies perspective could enable scholars to engage with these linguistic technologies in a way that keeps their alienness in sight, foregrounding their biases and blind spots and emphasizing the historical contingency of the ways in which computers “read” texts. What makes text mining interesting, in this view, is not its potential to “revolutionize” the methodology of the humanities, as Matthew Jockers claims, but the basic fact of its growing influence in the twenty-first century, given the widespread adoption of statistical methods in applications like search engines, spellcheckers, autocomplete features, and computer vision systems. Thinking of text-mining programs as objects of cultural criticism could open up an interchange between digital scholarship and the critical study of computers that is productive in both directions. The work of media theorists who study the ideological structures of technology could help us better understand the effects that computerization could have on our scholarly practice, both in explicitly digital work and in more traditional forms of scholarship that employ technologies like databases and search engines. On the other side, experimenting with techniques such as topic modeling in a critical frame could support a more robust analysis of the cultural authority that makes these technologies seem natural at the present moment, baring the ideological assumptions that underlie the quantification of language, and creating, perhaps, a renewed sense of the strangeness of the idea that words can be understood through the manipulation of numbers.

Models of Language

“Topic modeling” does not refer to any single method, but rather to a number of distinct technologies that attempt to determine the “topics” of texts automatically. The implementation most commonly used in the humanities is a program called MALLET, developed by Andrew McCallum and others and based on an algorithm developed by David Blei (Blei, Ng,
Provided with a collection of text files, MALLET can produce "topics" that look, in the output of the program, like this:

- Passions, passion, pleasure, person, love, pride, object, hatred, humility
- Men, interest, natural, society, property, actions, justice, human, moral
- Reason, nature, give, principles, general, observe, relations, common, subject
- Idea, ideas, objects, existence, mind, perceptions, impressions, form, time
- Object, imagination, relation, effect, present, mind, idea, experience, force

This model was trained using the text of David Hume's *Treatise of Human Nature*, divided into Hume's relatively short sections. Each line represents a "topic"—a cluster of words that tend to appear together in the same section. MALLET presents these topics as lists of words (e.g., "passions passion pleasure . . ."), starting with the word most strongly affiliated with the topic and proceeding downward. The program associates each text with one or more of these topics, which constitute a guess as to what that text is "about." There is no certainty to this process; the topics are produced by an approximate method and so the results are slightly different every time the program is run. The meaning of the results is further complicated by the fact that the "topics" in the output do not necessarily correspond to anything for which a simple description might exist. In many cases, they seem to be based more on sets of discursive conventions than on what we normally think of as "topics," and the results often include one or more topics that are totally inscrutable. Interpretation emerges as a key issue, especially given that the method depends on a complex set of assumptions that are colored by the institutional situation from which topic modeling emerged.

The idea of using a computer to automatically identify "topics" is in large part a product of the desire to exploit the increasingly large amount of text that was being distributed electronically in the late twentieth century. While the earliest attempts at automated "topic detection" go back to the 1960s, the field expanded greatly starting around 1990. (For an example of a very early attempt, see Borko and Bernick.) Many of the efforts from the 1990s dealt primarily with text from newswires and were designed for applications in finance and national security. The major accomplishments of this period include a software package known as SCISOR (System for Conceptual Information Summarization, Organization, and Retrieval), developed in the early 1990s, and the DARPA-funded Topic Detection and Tracking initiative, which ran from 1996 to 1997 (Jacobs and Rau; Allan et al.). The primary goal of the DARPA initiative, which drew participants from Carnegie Mellon University and the University of Massachusetts, was to come up with a way of automatically detecting the occurrence of major world events, such as volcanic eruptions and
political elections, through the text analysis of news feeds. The methods developed for this project mostly worked in a different way from the topic-modeling software that is now most familiar in the humanities. Instead of dealing with static collections of texts, they were designed to process continuous text feeds that changed from one topic to another at irregular intervals. One of the primary functions of the software was to determine when these transitions took place.

The topic-modeling techniques most commonly used in DH emerged around the same time as these projects, but they came from an area of research that was more oriented toward static collections than continuous news feeds. One of the most influential methods to emerge from this area is Latent Semantic Indexing (LSI), which was introduced in 1990 by Deerwester et al. (“Indexing by Latent Semantic Analysis”). Unlike the methods designed for the “segmentation” of newswire text, LSI and the other methods it inspired are meant to work with collections of discrete documents. The most common LSI-derived methods of topic modeling also depend on the “bag of words” assumption. Under this assumption, the computer takes no account of sentence divisions, syntax, or even the relative positions of words, considering only how frequently each word type appears in each document. Using these frequencies, LSI identifies clusters of associated words (“topics”) and links them to particular documents, something that can serve two major purposes. First, as the name Latent Semantic Indexing suggests, it can be used as a subject index, helping users find documents that are relevant to particular topics; and second, it produces a “reduced description” of the corpus that can be used to visualize patterns in a large amount of text.

In the original version of LSI, the topics are computed through a more-or-less arbitrary procedure that was empirically found to produce reasonable results for the test dataset, a collection of information science abstracts (Deerwester et al., 19; Blei, Ng, and Jordan, 994). In 1999, Thomas Hofmann developed a new variant of LSI based on a probabilistic model, a mathematical construct that offers a sort of rationale for the method (Hofmann, “Probabilistic Latent Semantic Indexing”). Texts, the model asserts, are composed of mixtures of certain “topics,” each of which has an associated vocabulary; a text about, for instance, fishing and economics is most likely to contain words that are strongly associated with these topics. Hofmann’s procedure can be used to determine the “topic” definitions that best fit a given collection of text based on this model. In 2003, David Blei, Andrew Y. Ng, and Michael I. Jordan introduced a further modification of the method called Latent Dirichlet Allocation (LDA), which is the variant used by MALLET and remains the most popular form of topic modeling among humanists (Blei, Ng, and Jordan; McCallum). LDA uses a similar model to Hofmann’s, but it adds a mathematical function called a Dirichlet distribution to determine the probabilities of certain topics occurring together. Adding this
function makes the model fully generative, which means, in a computer-science context, that it offers a complete mathematical description of the process by which the input texts were (hypothetically) generated, including a way of determining the probabilities of specific outcomes. (For a general introduction to generative modeling in relation to other forms of machine learning, see Jebara.) The generative model underlying LDA is something like this: first, the writer picks a “mixture” of topics to write about; then the writer constructs the text word-by-word by first randomly choosing a topic from the mixture and then picking a word based on the probability table for that topic. This generative model allows a computer to perform two complementary operations: a topic-modeling program can “learn” what words are associated with what topics based on a corpus of text; then, it can use this model to infer the likely topics of other texts.

It should be apparent from my admittedly rough description that this form of statistical modeling carries a heavy weight of epistemological baggage. Prominent among the disciplinary norms that govern the legitimacy of evidence and methodology in machine learning is the idea that the performance of the tools should be judged against a “gold standard” that defines the correct output—a practice that assumes the desired result to be both fixed ahead of time and accessible through some means outside of the method itself (Juckett). The institutional formation from which text-mining software emerged has also influenced the sorts of language for which it is designed. As I noted previously, the original version of LSI was initially tested with information science abstracts. The paper that introduces LDA draws its examples from AP and Reuters news articles, while Blei’s later work has included models based on articles from the journal Science, the Yale Law Journal, and the New York Times (Blei, Ng, and Jordan; Blei, “Probabilistic Topic Models”; Blei, “Topic Modeling and Digital Humanities”). The sorts of text on which these methods are generally tested have a number of commonalities. They are primarily written in a standard dialect and orthography; they tend to privilege the informational over the aesthetic dimensions of language; and they primarily consist of prose. Many of the examples used in testing these methods are also, it is worth noting, the sorts of text that the military-industrial apparatus would have a clear interest in mining. The language commonly used in articles about topic modeling—articles by Hofmann and Blei describe users going on “quests” for information in collections of texts that bear “hidden” or “latent” meanings—is suggestive of the ultimate purpose of the technology (Hofmann; Blei, “Probabilistic Topic Models”). By automatically determining what large numbers of documents are “about,” the software can help operators find and “extract” what they need from texts that are assumed to be repositories of information.

While LDA has proved to work reasonably well when applied to texts that are outside of its original purview, including
nineteenth-century novels, literary criticism, and early eighteenth-century essays, it is reasonable to ask whether the results it produces are affected by the assumptions that went into the development of the software. One scholar who has considered the biases of topic modeling while employing it in humanistic research, Lisa Marie Rhody, argues that topic models of poetry must be read in a different way from those based on scientific journals (“Topic Modeling and Figurative Language”). A reason, she suggests, is that poetry characteristically uses a relatively large amount of figurative language and produces meaning in a much wider variety of ways than do “non-figurative” texts. An attempt to topic model poetry thus encounters particular interpretive difficulties, but it also “illustrates how figurative language resists thematic topic assignments and by doing so, effectively increases the attractiveness of topic modeling as a methodological tool for literary analysis of poetic texts.” In the topic model that she produced based on the Revising Ekphrasis collection of poetry, Rhody finds that some of the most interesting “topics” correspond less to what poems are “about” than to particular poetic traditions. She argues, in particular, that a topic with the top words “death life heart dead long world blood earth man soul men face day pain die” corresponds to the language of elegiac poetry, and she uses it to highlight elegiac qualities in poems by African American poets that are not explicitly about death.

While Rhody frames her argument in terms of a “caricatured” view that hyper-emphasizes the figurative nature of poetry, the distinction between figurative and non-figurative language is slightly misleading as an explanation of why topic modeling works particularly well with scientific texts. Many phrases occur repeatedly in scientific abstracts that are arguably figurative: ideas being “underlined” and “highlighted,” “first steps” toward solutions, “root causes” of problems. From the perspective of topic modeling, what is important is not that words be used literally, but that the vocabularies of texts correlate with their topics in a uniform fashion. Scientific language fits this requirement particularly well in part because it is produced within a system of overlapping subdisciplines that have distinctive lexicons. The slippery question of whether well-worn phrases like “root cause” are figurative is beside the point; what is important to a topic-modeling program is that this formula is repeated commonly in engineering abstracts but is relatively rare in physics, which makes it a potential distinguishing factor between the two disciplines. In addition to having to work with highly specialized technical vocabularies, scientific writers are encouraged to stick with established usages rather than inventing novel expressions for things that have already been described, which creates highly repetitive patterns in word usage that facilitate the detection of topics. Because of the tight control of the vocabulary to be used within each specialization, the process of scientific writing hews very close to the generative model by which LDA assumes texts were written—much closer, I venture, than the
process of writing poetry, although Rhody's example shows that some poetic traditions do have distinctive vocabularies that topic modeling can detect.

The assumption that word choice follows uniformly from the "topic" of a text—whatever we take the "topics" to represent—presumes a sort of linguistic standardization that is historically bound up with structures of authority. As John Guillory has noted, while we now tend to see a plain and direct style as the default for most forms of writing, rhetoricians in the early-modern period placed a greater value on copia, a style that involves a profusion of different ways of saying the same thing ("The Memo and Modernity"). The idea that words should have fixed meanings was largely a product of the latter half of the seventeenth century, when authors like John Locke, Thomas Sprat, and John Wilkins began to see the fluidity of language as an obstruction to clear thought. The compilation of dictionaries for European languages, which began in earnest around that time, created a newly sharp division between standard and nonstandard uses of words, largely based on which usages were "authorized" by their inclusion in the works of eminent writers. It also sharpened the divisions between languages; as Benedict Anderson argues, the "lexicographic revolution" led Europeans to see languages as the property of particular groups, creating imagined communities of German, French, and English speakers where previously there had been a profusion of local dialects (84). This shift in the way Europeans thought about language enabled the creation of highly uniform styles like the ones now used in scientific writing and gave a greater stability to the word usage of many other forms of writing, reducing both variation for variation's sake and many regionalisms. The reason topic modeling does particularly well at identifying themes in technical and informational genres like news articles, abstracts, and encyclopedia entries is that they actively strive to follow this sort of standard, sticking for the most part to usages that have the force of authority behind them.

While topic modeling's affinity for uniform language can be accounted for if the collection of texts is fairly homogeneous, as in Blei's collection of newspaper articles and Rhody's collection of poetry, it becomes a more difficult problem when the method is applied to a corpus that includes texts of various types with varying relations to linguistic standards. When trained on collections that include mostly standardized text but some text that follows other conventions—as is the case with many collections of nineteenth-century fiction—topic models tend to relegate the nonstandard words to a small number of topics while excluding them from the rest of the model. For example, Matthew Wilkins's topic model of the Wright American Fiction corpus includes these two topics: "uv wuz ez hev wich hed sed sez ther ef" and "dat master slave negro massa slaves white black dis dey" (Wilkins, 100-topic model of the Wright American Fiction corpus). The language of the mock-rustic characters of humorists like David
Ross Locke, Charles Farrar Browne, and George William Bagby gets its own topic, while the spellings that some novelists used to represent African American speech are mixed together with words having to do with slavery. Topic modeling is fairly good at distinguishing languages, something that could potentially be useful, but this tendency to separate linguistic conventions could easily become problematic if we are not extremely careful in how we interpret the results. To the extent that it is relegated to its own topic, orthographically distinctive text is prevented from influencing the other topics in the model. If we are to use the other topics as a way of tracking themes or "discourses" in the collection, we are effectively excluding the words of characters who are presented in caricature from affecting our results, repeating the structure of authority that enables their speech to be coded as nonstandard.

LDA is not just tuned to work best with standardized (and, one might say, hegemonic) forms of language; it also structures its results in a way that encourages interpretation in terms of the standardized meanings of words. In chapter 45 in this volume, Tanya Clement discusses a property common to many text-mining techniques, a dependence on the assumption that "the Word" is a stable and inherently meaningful unit of language ("The Ground Truth of DH Text Mining"). The tendency of text-mining programs to accentuate the stability of words results, in part, from the way in which statistical methods tend to smooth out individual discrepancies so as to emphasize the overall patterns in a dataset. This smoothing is not an accident, but a necessary result of the need to avoid what statisticians call overfitting (Dietterich). A model that exactly accounts for every nuance of a dataset tends to be too complex to be useful—to take an image from Jorge Luis Borges, it is like a map that is as large as the territory it represents—and thus, some cases that deviate from general trends have to be ignored (Borges, "On Exactitude in Science"). Though the practicalities of modeling require the smoothing out of differences, this process is an ideologically loaded way of dealing with language, and the much-vaunted comprehensibility of topic models depends on it. Each "word" in the output of a topic-modeling program stands for many instances of that word in the input, each one with a unique syntactic context that the model largely ignores. An interpretation of these aggregate-words can easily slide into the assumption that all of these instances can be encompassed by a single meaning.

An example of this smoothing-out of instabilities in word meaning occurs in Matthew Jockers’s book Macroanalysis. After introducing the idea of topic modeling, Jockers presents two topics from a model of the Stanford Literary Lab’s corpus of novels. Jockers proceeds to interpret the appearance of the word stream in the list of top words for one of these topics, alongside indian, indians, chief, savages, warriors, men, party, etc.:
Here Jockers seems to be doing something familiar to literary critics: determining the meaning of a word based on context. But the “particular use” of stream to which Jockers is referring is neither a word type (the word stream considered in the abstract) nor a word token (a particular instance of the word in a text)—it is an entry in a probability table that was generated through an approximate optimization method. This unit does not correspond to a single “use” of a word in any usual sense, but rather derives from patterns among many different instances of the word in the corpus. Although some of these instances might refer to a body of flowing water, there is no guarantee that they all use the word in the same sense—there are, for instance, at least a few dozen references to a “stream of settlers” in nineteenth-century texts that discuss conflicts between Europeans and Native Americans, and if these are present in Jockers’s corpus they would likely be included in the topic he discusses. In attempting to determine what the stream in the topic model “refers to,” Jockers interprets this abstract composite as if it were the same type of thing as a word token in a literary text, a move that presupposes the stability of the word’s signification in the parts of the corpus covered by the topic. This sort of interpretation-in-aggregate is not necessarily illegitimate if we recognize it for what it is, but Jockers’s application of simple and familiar terms of interpretation to a topic model belies the very complex and potentially problematic set of assumptions that underlie what he is doing in this passage.

The bias toward standardized forms of language is present not only in topic modeling, but in many other text-mining methods that depend on statistical analysis of words. The affinity of these methods for particular forms of language becomes readily apparent in an exchange between Jockers and Annie Swafford about Jockers’s Syuzhet program (Jockers, “Revealing Sentiment and Plot Arcs with the Syuzhet Package”). This package uses sentiment analysis software to guess the emotional valence of each sentence of a novel and plots an “arc” that is derived from these results. In a blog post, Swafford points out a number of problems with this method, among them the inability of sentiment analysis to account for the nuances of literary language (“Problems with the Syuzhet Package”). Responding to the latter problem, Jockers admits his frustration: “Things like irony, metaphor, and dark humor are the monsters under the bed that keep me up at night” (“Some thoughts . . .”). The difficulty of accounting for these aspects of language in projects like Syuzhet seems to stem from the fact that all of the sentiment analysis methods pres-
Currently available are designed to suit the language of, in Swafford’s words, “a tweet or product review” (“Continuing the Syuzhet Discussion”). In other words, Jockers’s analysis depends on a tool designed to suit the contemporary descendants of the Enlightenment project of rationalizing language and standardizing the meanings of words. Although some of the problems that Swafford points out are specific to the software that Syuzhet uses, many other text-mining techniques share the tendency to work best with texts that straightforwardly follow standard usages while treating the existence of “non-literal” language, when they deal with it at all, as a problem to be solved. If we are to adopt text-mining tools in humanistic research, we will need to take account of the assumptions they make about language and how those assumptions could serve ideological interests.

Alien Reading

Although these observations suggest that there are good reasons for scholars to be wary of the adoption of text-mining software in the humanities, it would be a mistake simply to dismiss it as irrelevant to our concerns. In his 2014 article, “Theorizing Research Practices We Forgot to Theorize Twenty Years Ago,” Ted Underwood forcefully points out that literary scholars outside of the digital humanities have already been using text-mining software on a regular basis for decades in the form of databases and search engines, but we have done little to theorize the role that these technologies play in scholarly practice (64). Many of the databases scholars commonly use already depend on generative models and other text-mining techniques for correcting scanning errors and accounting for spelling variants. In the present day, it is virtually impossible for scholars to avoid text-mining software altogether, even if many of us only encounter it indirectly through platforms like Google or JSTOR. If, as scholars, we are to engage with these technologies on our own terms, then we will have to find a way of making their roles in humanistic research a matter of active concern. Experimenting with text-mining programs in English departments could serve as a safeguard against the possibility that we unknowingly absorb these tools into our practice without reflecting on the assumptions about language and knowledge that underlie them and considering the effects they could have on our work.

The unreflective computerization that Underwood points out presents a particular problem in the present moment because of a current trend toward user interfaces that cover up the complexity of what goes on inside the machine. As Lori Emerson argues in Reading Writing Interfaces, the naturalistic interfaces of modern computers make their operations seem much simpler and more familiar than they really are, encouraging a passive, consumer-like orientation towards the computer rather than a deep understanding of it (1–19). The interfaces of tablet computers especially make heavy use of ele-
ments that mimic the behavior of physical objects, appealing to very familiar intuitions about how objects behave. This makes the devices easy to use up to a point, but it gives the typical user little insight into how they work. Search engines can similarly be much more complex than their user interfaces suggest, employing sophisticated algorithms for cleaning up and indexing texts, identifying synonyms, and determining the “relevance” of results that depend on strong assumptions about language and that could potentially introduce biases into research. While many text-mining programs present their results using familiar terms like *word*, *topic*, and *similarity*, the mathematical structures underneath are often fundamentally different from the ways in which human beings ordinarily understand these concepts. The apparent simplicity of interfaces like the search box allows us to use these technologies in our scholarship without confronting the complexity of what they do and the ways in which their designs might conflict with our precepts as scholars.

While Underwood responds to this problem with an embrace of statistical modeling, it is also possible to employ text-mining programs without accepting the thinking behind them, pushing back against naturalistic user interface design by drawing attention to aspects of the software that conflict with a humanistic view of interpretation. This approach would involve encountering text mining as an alien form of reading—alien both in the fact that it emerged from a discipline with very different concerns from our own and the fact that it is performed by a machine, the sort of nonhuman agent that Ian Bogost has sought to understand with his idea of *alien phenomenology* (Bogost). Rhody’s work with topic modeling is one example of a project that employs text-mining technologies while keeping in mind the ways in which their assumptions might clash with the concerns of humanists. I would like to suggest an approach that goes further into a critique of the technology itself, engaging with text-mining tools as embodied, historically situated cultural productions that are potentially problematic. Understanding the extent to which our use of digital tools can reinforce hegemonic views of language requires a sort of scholarship that takes up a critical, perhaps even antagonistic attitude toward computerized modes of processing language. One thing that we, as humanists, can do to further this goal is to experiment with text-mining programs in a context that enables us to brush them against the grain, analyzing their assumptions and showing how they are positioned in the wider intellectual and cultural scene of the twenty-first century—writing, as it were, the *Tristram Shandy* to information science’s *Essay concerning Human Understanding*.

This statement could perhaps be accused of encouraging the sort of navel-gazing focus on methodology that, as Cameron Blevins argues in chapter 26 of this volume, has characterized recent work in digital history; but we need not consider text-mining practices in isolation from the rest of the
world ("Digital History’s Perpetual Future Tense"). Sterne’s *Tristram Shandy* is much more than just a satire of Locke’s call for the stabilization of words; it situates this impulse among many other aspects of the life of the eighteenth-century English middle class and its relationship to the intellectual culture of its past. In a long view, text-mining software is a part of the same history that literary critics study, a twenty-first-century expression of a standardizing impulse that has had a productive (if sometimes hostile) interchange with imaginative literature for centuries and that bears complex relationships to older practices of industrial management, library organization, and philology. These histories are relevant to many of the questions that more traditional forms of literary scholarship ask, bound up as they are with age-old practices of reading and writing like excerpting, cataloging, and the creation of grammars. A critical engagement with text-mining software can also help us understand those aspects of computational methods that are genuinely new, especially the use of statistical methods. Experimenting with text-mining software can highlight the strangeness of computational technology in comparison to what has come before—a strangeness that, to use a commonplace from the field of media studies, those of us who live on the cusp of its emergence may be much better poised to see than future generations.

Engaging with text mining as an alien form of reading requires that we resist attempts to present computational results in forms that readily appeal to our assumptions and intuitions about language. The ease with which we can identify the “words” in the output of MALLET with our usual notion of the word makes it too easy to overlook the radical difference between how these units function in the program and the ways in which words can work in a human mind. While we cannot expect everyone who uses text-mining software to attempt a complete understanding of what is going on inside the computer, we should at least make an effort to appreciate the extent to which the tools we use are unknown to us, especially given the possibility that what happens inside could serve ideological ends. One can get a vivid sense of the gap between machine reading and our intuitive conceptions of language by examining the entries to Darius Kazemi’s *National Novel Generating Month* an annual contest that challenges people to write a computer program that generates a 50,000 word novel (*NaNoGenMo 2014*). Most of the results are essentially unreadable, serving more as comments on process and algorithm than as ways of producing something that really resembles a novel, and they often ultimately direct our attention back on the role of computation itself in the generative process. For instance, Sean Connor’s entry produces a randomized novel by piecing together sequences of words and punctuation marks from L. Frank Baum’s fourteen Oz novels. This is one paragraph of the output:

“Same with me, please,” interrupted the girl Ruler for
The program that generated this text is based on a Markov chain model, the same sort of generative model that is commonly used in regularizing texts for the purposes of search engines, among many other applications. Although we cannot draw any major conclusions about how the technology works by reading specific examples of output, the practice of generating text using statistical models that were primarily designed for the processing of existing texts can be useful simply as a reminder of the fact that these models only loosely correspond to the way human languages work. The statistical methods that exist at present diverge in many ways from our ordinary expectations about what a text should look like, something that becomes much easier to see when one employs them for writing rather than for reading.

Text-generation programs have been employed for a number of purposes in the humanities, not all of which are specifically critical of the technology underlying them. Stephen Ramsay's Reading Machines proposes an Oulipan approach to literary criticism in which computerized transformations serve to enable richer and more complex interpretations of texts (15–17). Others, such as Mark Sample, have connected text generation to Lisa Samuels and Jerome J. McGann's idea of "deformance," a practice that creates modified versions of texts as a way of exploring their autopoietic capabilities (Sample, "Notes towards a Deformed Humanities;" Samuels and McGann, 'Deformance and Interpretation'). For instance, Sample's Twitter account "This is Just to Say" (@JustToSayBot11) produces randomly generated parodies of William Carlos Williams's poem of the same name, replacing the sweet and cold plums with something different every time. But it is also possible to think of text generation more as an interrogation of the technology itself than as a way of encountering literary texts. One way of understanding how a text generator could serve as a critique of technology is through Sean Sturm and Stephen Turner's idea of digital caricature. Drawing on the work of the philosopher Vilém Flusser, Sturm and Turner suggest that we think of computation as "a caricature of thinking," a diminished imitation of mental operations that can potentially be viewed as a joke (para. 30–31). Finding humor in the "drop-down menu-isation" that computers impose on design, they argue, involves understanding it not just in terms of symbolic logic, but also from the perspective of "a region of primitively evolved drives" that computers lack (para. 33). The failures of methods like the Markov chain model to produce convincing imitations of novels can serve as caricatures in just the sense that Sturm and Turner discuss, eliciting laughter because they reveal the machine's incongruity with the social world in which we expect writing
to take place. While computer scientists will undoubtedly de-
velop better software that can create more convincing imita-
tions of human writing, employing these programs as jokes
allows us to revel in their present limitations, taking the op-
portunity they provide to show how the mechanisms under-
neath the software differ from human intelligence. Given the
increasing inescapability of digitally inflected modes of
thought, Sturm and Turner suggest, the best way to under-
stand what it means to be human today is to laugh at
computers.

But while digital caricature can serve a useful purpose by
provoking an awareness of the difference between human
and machine reading, it cannot substitute for a historical per-
spective on these technologies. The absurd text created by
novel generators can give us a visceral sense of how comput-
tational models differ from our intuitive understandings of
language, but it can only get us so far in understanding how
those models relate to ideology. For this we need to supple-
ment our experimentation with text-mining methods with
research that situates them historically—both in the short
term, looking at the institutional contexts from which they
emerged, and in the long term, looking at how they relate to
the histories of linguistic thought, philosophy, communica-
tion, and labor organization. This is an area where scholars of
literature and intellectual history could have a particularly pro-
ductive interchange with media theorists who critically study
contemporary technology. Text-mining systems are playing
increasingly large roles in our lives, our teaching, and our
scholarship, and digital humanists, especially those who are
versed in both statistical modeling and literary theory, are
uniquely positioned to examine the linguistic ideologies that
underlie them. Placing text mining in dialogue with the past
could be useful not just for theorizing the implications of new
scholarly tools like search engines, but also for interpreting
historical texts in ways that are of particular relevance to the
present shift from print to digital reading. To do this, we need
a different form of scholarship from the one that applies a
computer science methodology to the study of literary history.
A media-studies approach would engage with programs like
MALLET as cultural artifacts from the twenty-first century,
products of a mechanization of language that is in some ways
similar to views that have been put forth in the past, and that is
in some ways new.

Notes

1. Jockers models nineteenth-century novels in Macroanalysis, 118–153;
Goldstone and Underwood apply topic modeling to literary criticism in
“The Quiet Transformations of Literary Studies,” Collin Jennings and I
created a topic model for Joseph Addison and Richard Steele’s The
Spectator, available to view online at

3. I draw this conclusion from a search of the Thomson Reuters Web of Science database.

4. For an anecdote about a PhD student in biology being excoriated for writing “like a poet,” see Ruben.

5. See Locke, An Essay concerning Human Understanding, especially 437–65; Sprat, History of the Royal Society of London; and John Wilkins, An Essay towards a Real Character.


7. The HathiTrust database returns 329 results for “stream of settlers” together with “Indians,” constituting at least thirty distinct books.

8. On one connection between computers and industrialism, see McPherson. On accounting as a precedent for hypertext, see Duguid. On the relationship between computational linguistic techniques and philology, see Lennon.


Bibliography


—. "Revealing Sentiment and Plot Arcs with the Syuzhet Package." *Matthew L. Jockers* (blog), February 2, 2015. [http://www.matthewjockers.net/2015/02/02/syuzhet](http://www.matthewjockers.net/2015/02/02/syuzhet).


THE COMPUTATIONAL TURN:
THINKING ABOUT THE DIGITAL HUMANITIES

David M. Berry

Introduction

Few dispute that digital technology is fundamentally changing the way in which we engage in the research process. Indeed, it is becoming more and more evident that research is increasingly being mediated through digital technology. Many argue that this mediation is slowly beginning to change what it means to undertake research, affecting both the epistemologies and ontologies that underlie a research programme. Of course, this development is variable depending on disciplines and research agendas, with some more reliant on digital technology than others, but it is rare to find an academic today who has had no access to digital technology as part of their research activity. Library catalogues are now probably the minimum way in which an academic can access books and research articles without the use of a computer, but, with card indexes dying a slow and certain death (Baker, 1996, 2001), there remain few outputs for the non-digital scholar to undertake research in the modern university. Email, Google searches and bibliographic databases are become increasingly crucial, as more of the world libraries are scanned and placed online. Whilst some decry the loss of the skills and techniques of older research traditions, others have warmly embraced what has come to be called the digital humanities (Schreibman et al., 2008; Schnapp & Presner, 2009; Presner, 2010; Hayles, 2011).

The digital humanities try to take account of the plasticity of digital forms and the way in which they point toward a new way of working with representation and mediation, what might be called the digital ‘folding’ of reality, whereby one is able to approach culture in a radically new way. To mediate an object, a digital or computational device requires that this object be translated into the digital code that it can understand. This minimal transformation is effected
through the input mechanism of a socio-technical device within which a model or image is stabilised and attended to. It is then internally transformed, depending on a number of interventions, processes or filters, and eventually displayed as a final calculation, usually in a visual form. This results in real-world situations where computation is event-driven and divided into discrete processes to undertake a particular user task. The key point is that without the possibility of discrete encoding there is no object for the computational device to process. However, in cutting up the world in this manner, information about the world necessarily has to be discarded in order to store a representation within the computer. In other words, a computer requires that everything is transformed from the continuous flow of our everyday reality into a grid of numbers that can be stored as a representation of reality which can then be manipulated using algorithms. These subtractive methods of understanding reality (episteme) produce new knowledges and methods for the control of reality (techne). They do so through a digital mediation, which the digital humanities are starting to take seriously as their problematic.

The digital humanities themselves have had a rather interesting history. Starting out as ‘computing in the humanities’, or ‘humanities computing’, in the early days they were often seen as a technical support to the work of the ‘real’ humanities scholars, who would drive the projects. This involved the application of the computer to the disciplines of the humanities, something that has been described as treating the ‘machine’s efficiency as a servant’ rather than ‘its participant enabling of criticism’ (McCarty, 2009). As Hayles explains, changing to the term “Digital Humanities” was meant to signal that the field had emerged from the low-prestige status of a support service into a genuinely intellectual endeavour with its own professional practices, rigorous standards, and exciting theoretical explorations’ (Hayles, 2011). Ironically, as the projects became bigger and more complex, and as it developed computational techniques as an intrinsic part of the research process, technically proficient researchers increasingly saw the computational as part and parcel of what it meant to do research in the humanities itself. That is, computational technology has become the very condition of possibility required in order to think about many of the questions raised in the humanities today. For example, as Schnapp and Presner explain in the Digital Humanities Manifesto 2.0,

The first wave of digital humanities work was quantitative, mobilizing the search and retrieval
The question of quite how the digital humanities undertake their research, and whether the notions of first and second wave digital humanities captures the current state of different working practices and methods in the digital humanities, remains contested. Yet these can be useful analytical concepts for thinking through the changes in the digital humanities. We might, however, observe the following: first-wave digital humanities involved the building of infrastructure in the studying of humanities texts through digital repositories, text markup, etc., whereas second-wave digital humanities expands the notional limits of the archive to include digital works, and so bring to bear the humanities’ own methodological toolkits to look at ‘born-
digital’ materials, such as electronic literature (e-lit), interactive fiction (IF), web-based artefacts, and so forth.

I would like to explore here a tentative path for a third wave of the digital humanities, concentrated around the underlying computationality of the forms held within a computational medium. That is, I propose to look at the digital component of the digital humanities in the light of its medium specificity, as a way of thinking about how medial changes produce epistemic changes. This approach draws from recent work in software studies and critical code studies, but it also thinks about the questions raised by platform studies, namely the specifics of general computability made available by specific platforms (Fuller, 2008; Manovich, 2008; Montfort & Bogost, 2009; Berry, 2011). I also want to suggest that neither first nor second-wave digital humanities really problematized what Lakatos (1980) would have called the ‘hard-core’ of the humanities, the unspoken assumptions and ontological foundations which support the ‘normal’ research that humanities scholars undertake on an everyday basis. Indeed, we could say that third-wave digital humanities points the way in which digital technology highlights the anomalies generated in a humanities research project and which leads to the questioning of the assumptions implicit in such research, e.g. close reading, canon formation, periodization, liberal humanism, etc. We are, as Presner argues, ‘at the beginning of a shift in standards governing permissible problems, concepts, and explanations, and also in the midst of a transformation of the institutional and conceptual conditions of possibility for the generation, transmission, accessibility, and preservation of knowledge’ (2010: 10).

To look into this issue, I want to start with an examination of the complex field of understanding culture through digital technology. Indeed, I argue that to understand the contemporary born-digital culture and the everyday practices that populate it – the focus of a digital humanities second wave – we need a corresponding focus on the computer code that is entangled with all aspects of our lives, including reflexivity about how much code is infiltrating the academy itself. As Mathew Fuller argues, ‘in a sense, all intellectual work is now “software study”, in that software provides its media and its context... [yet] there are very few places where the specific nature, the materiality, of software is studied except as a matter of engineering’ (2006). We also need to bring to the fore the ‘structure of feeling’ that computer code facilitates and the way in which people use software in their research thinking and everyday
practices. This includes the increase in the acceptance and use of software in the production, consumption and critique of culture.

Thus, there is an undeniable cultural dimension to computation and the medial affordances of software. This connection again points to the importance of engaging with and understanding code: indeed, computer code can serve as an index of digital culture (imagine digital humanities mapping different programming languages to the cultural possibilities and practices that it affords, e.g. HTML to cybertulture, AJAX to social media). This means that we can ask the question: what is culture after it has been ‘softwarized’? (Manovich, 2008:41). Understanding digital humanities is in some sense then understanding code, and this can be a resourceful way of understanding cultural production more generally: for example, just as digital typesetting transformed the print newspaper industry, eBook and eInk technologies are likely to do so again. We thus need to take computation as the key issue that is underlying these changes across mediums, industries and economies.

Knowing knowledge

In trying to understand the digital humanities our first step might be to problematize computationality, so that we are able to think critically about how knowledge in the 21st century is transformed into information through computational techniques, particularly within software. It is interesting that at a time when the idea of the university is itself under serious rethinking and renegotiation, digital technologies are transforming our ability to use and understand information outside of these traditional knowledge structures. This is connected to wider challenges to the traditional narratives that served as unifying ideas for the university and, with their decline, has led to difficulty in justifying and legitimating the postmodern university vis-à-vis government funding.

Historically, the role of the university has been closely associated with the production of knowledge. For example, in 1798 Immanuel Kant outlined an argument for the nature of the university titled The Conflict of the Faculties. He argued that all of the university’s activities should be organised by a single regulatory idea, that of the concept of reason. As Bill Readings (1996) stated:

\[
\text{Reason on the one hand, provide[d] the ratio for all the disciplines; it [was] their organizing}
\]
principle. On the other hand, reason [had] its own faculty, which Kant names[d] "philosophy" but which we would now be more likely to call the 'humanities'. (Readings, 1996: 15)

Kant argued that reason and the state, knowledge and power, could be unified in the university by the production of individuals who would capable of rational thought and republican politics – students trained for the civil service and society. Kant was concerned with the question of regulative public reason, that is, with how to ensure stable, governed and governable regimes which can rule free people, in contrast to tradition represented by monarchy, the Church or a Leviathan. This required universities, as regulated knowledge-producing organisations, to be guided and overseen by the faculty of philosophy, which could ensure that the university remained rational. This was part of a response to the rise of print culture, growing literacy and the kinds of destabilising effects that this brought. Thus, without resorting to dogmatic doctrinal force or violence, one could have a form of perpetual peace by the application of one's reason.2

This was followed by the development of the modern university in the 19th century, instituted by the German Idealists, such as Schiller and Humboldt, who argued that there should be a more explicitly political role to the structure given by Kant. They argued for the replacement of reason with culture, as they believed that culture could serve as a 'unifying function for the university' (Readings, 1996: 15). For the German Idealists like Humboldt, culture was the sum of all knowledge that is studied, as well as the cultivation and development of one's character as a result of that study. Indeed, Humboldt proposed the founding of a new university, the University of Berlin, as a mediator between national culture and the nation-state. Under the project of 'culture', the university would be required to undertake both research and teaching, i.e., the production and dissemination of knowledge respectively. The modern idea of a university therefore allowed it to become the preeminent institution that unified ethnic tradition and statist rationality by the production of an educated cultured individual. The German Idealists proposed

that the way to reintegrate the multiplicity of known facts into a unified cultural science is through Bildung, the ennoblement of character...
The university produces not servants but subjects.

www.culturemachine.net • 6
That is the point of the pedagogy of Bildung, which teaches knowledge acquisition as a process rather than the acquisition of knowledge as a product. (Reading, 1996: 65-67)

This notion was given a literary turn by the English, in particular John Henry Newman and Matthew Arnold, who argued that literature, not culture or philosophy, should be the central discipline in the university, and also in national culture more generally. Literature therefore became institutionalised within the university ‘in explicitly national terms and [through] an organic vision of the possibility of a unified national culture’ (Readings, 1996: 16). This became regulated through the notion of a literary canon, which was taught to students to produce literary subjects as national subjects.

Readings argues that in the postmodern university we now see the breakdown of these ideals, associated particularly with the rise of the notion of the ‘university of excellence’ -- which for him is a concept of the university that has no content, no referent. What I would like to suggest is that today, we are beginning to see instead the cultural importance of the digital as the unifying idea of the university. Initially this has tended to be associated with notions such as information literacy and digital literacy, betraying their debt to the previous literary conception of the university, albeit understood through vocational training and employment. However, I want to propose that, rather than learning a practice for the digital, which tends to be conceptualised in terms of ICT skills and competences (see for example the European Computer Driving License), we should be thinking about what reading and writing actually should mean in a computational age. This is to argue for critical understanding of the literature of the digital, and through that develop a shared digital culture through a form of digital Bildung. Here I am not calling for a return to the humanities of the past, to use a phrase of Fuller (2010), ‘for some humans’, but rather to a liberal arts that is ‘for all humans’. To use the distinction introduced by Hofstadter (1963), this is to call for the development of a digital intellect -- as opposed to a digital intelligence. Hofstadter writes:

Intellect... is the critical, creative, and contemplative side of mind. Whereas intelligence seeks to grasp, manipulate, re-order, adjust, intellect examines, ponders, wonders, theorizes, criticizes, imagines. Intelligence will seize the immediate meaning in a situation and evaluate it.
Intellect evaluates evaluations, and looks for the meanings of situations as a whole... Intellect [is] a unique manifestation of human dignity. (Hofstadter, 1963: 25)

The digital assemblages that are now being built not only promise great change at the level of the individual human actor. They provide destabilising amounts of knowledge and information that lack the regulating force of philosophy -- which, Kant argued, ensures that institutions remain rational. Technology enables access to the databanks of human knowledge from anywhere, disregarding and bypassing the traditional gatekeepers of knowledge in the state, the universities and the market. There no longer seems to be the professor who tells you what you should be looking up and the 'three arguments in favour of it' and the 'three arguments against it'. This introduces not only a moment of societal disorientation, with individuals and institutions flooded with information, but also offers a computational solution to this state of events in the form of computational rationalities--something that Turing (1950) described as super-critical modes of thought. Both of these forces are underpinned at a deep structural level by the conditions of possibility suggested by computer code.

As mentioned previously, computer code enables new communicative processes, and with the increasing social dimension of networked media the possibility of new and exciting forms of collaborative thinking arises. This is not the collective intelligence discussed by Levy (1999); rather, it is the promise of a collective intellect. The situation is reminiscent of the medieval notion of the universitas, but recast in a digital form, as a society or association of actors who can think critically together, mediated through technology. It further raises the question of what new modes of collective knowledge software can enable or constitute. Can software and code take us beyond the individualising trends of blogs, comments, twitter feeds, and so forth, and make possible something truly collaborative -- something like the super-critical thinking that is generative of ideas, modes of thought, theories and new practices? There is certainly something interesting about real-time stream forms of digital memory in that they are not afforded towards the past, as history, but neither are they directed towards a form of futurity. Instead we might say they seem to nowmediate? newmediate? lifemediate? Jetztzeitmediate? (Benjamin, 1992: 252-3)? In other words, they gather together the newness of a particular group of streams, a kind of collective writing, that has the potential
to be immensely creative. These are possible rich areas for research for a third-wave digital humanities that seeks to understand these potentially new forms of literature and the medium that supports them.

For the research and teaching disciplines within the university, the digital shift could represent the beginnings of a moment of ‘revolutionary science’, in the Kuhnian sense of a shift in the ontology of the positive sciences and the emergence of a constellation of new ‘normal science’ (Kuhn 1996). This would mean that the disciplines would, ontologically, have a very similar Lakatosian computational ‘hard core’ (Lakatos, 1980). This has much wider consequences for the notion of the unification of knowledge and the idea of the university (Readings, 1996). Computer science could play a foundational role with respect to the other sciences, supporting and directing their development, even issuing ‘lucid directives for their inquiry’. Perhaps we are beginning to see reading and writing computer code as part of the pedagogy required to create a new subject produced by the university, a computational or data-centric subject. This is, of course, not to advocate that the existing methods and practices of computer science become hegemonic, rather that a humanistic understanding of technology could be developed, which also involves an urgent inquiry into what is human about the computational humanities or social sciences. In a related manner, Fuller (Fuller, S., 2006) has called for a ‘new sociological imagination’, pointing to the historical project of the social sciences that have been committed to ‘all and only humans’, because they ‘take all human beings to be of equal epistemic interest and moral concern’ (Fuller, 2010: 242). By drawing attention to ‘humanity’s ontological precariousness’ (244), Fuller rightly identifies that the project of humanity requires urgent thought, and, we might add, even more so in relation to the challenge of a computationality that threatens our understanding of what is required to be identified as human at all.

If software and code become the condition of possibility for unifying the multiple knowledges now produced in the university, then the ability to think oneself, taught by rote learning of methods, calculation, equations, readings, canons, processes, etc., might become less important. Although there might be less need for an individual ability to perform these mental feats or, perhaps, even recall the entire canon ourselves due to its size and scope, using technical devices, in conjunction with collaborative methods of working and studying, would enable a cognitively supported method
instead. The internalisation of particular practices that have been instilled for hundreds of years in children and students would need to be rethought, and in doing so the commonality of thinking qua thinking produced by this pedagogy would also change. Instead, reasoning could shift to a more conceptual or communicative method of reasoning, for example, by bringing together comparative and communicative analysis from different disciplinary perspectives, and by knowing how to use technology to achieve a usable result – a rolling process of reflexive thinking and collaborative rethinking.

Relying on technology in a more radically decentred way, depending on technical devices to fill in the blanks in our minds and to connect knowledge in new ways, would change our understanding of knowledge, wisdom and intelligence itself. It would be a radical decentring in some ways, as the Humboldtian subject filled with culture and a certain notion of rationality would no longer exist; rather, the computational subject would know where to recall culture as and when it was needed in conjunction with computationally available others, a just-in-time cultural subject, perhaps, to feed into a certain form of connected computationally supported thinking through and visualised presentation. Rather than a method of thinking with eyes and hand, we would have a method of thinking with eyes and screen.

This doesn’t have to be dehumanising. Latour and others have rightly identified the domestication of the human mind that took place with pen and paper (Latour, 1986). This is because computers, like pen and paper, help to stabilise meaning by cascading and visualising encoded knowledge that allows it to be continually ‘drawn, written, [and] recoded’ (Latour, 1986: 16). Computational techniques could give us greater powers of thinking, larger reach for our imaginations, and, possibly, allow us to reconnect to political notions of equality and redistribution based on the potential of computation to give to each according to their need and to each according to their ability. This is the point made forcefully by Fuller (2010: 262), who argues that we should look critically at the potential for inequality which is created when new technologies are introduced into society. This is not merely a problem of a ‘digital divide’, but a more fundamental one of how we classify those that are more ‘human’ than others, when access to computation and information increasingly has to pass through the market.
Towards a digital humanities?

The importance of understanding computational approaches is increasingly reflected across a number of disciplines, including the arts, humanities and social sciences, which use technologies to shift the critical ground of their concepts and theories – something that can be termed a computational turn. This is shown in the increasing interest in the digital humanities (Schreibman et al., 2008) and computational social science (Lazer et al., 2009), as evidenced, for example, by the growth in journals, conferences, books and research funding. In the digital humanities ‘critical inquiry involves the application of algorithmically facilitated search, retrieval, and critical process that... originat[es] in humanities-based work'; therefore ‘exemplary tasks traditionally associated with humanities computing hold the digital representation of archival materials on a par with analysis or critical inquiry, as well as theories of analysis or critical inquiry originating in the study of those materials’ (Schreibman et al., 2008: xxv). In social sciences, Lazer et al. argue that ‘computational social science is emerging that leverages the capacity to collect and analyze data with an unprecedented breadth and depth and scale’ (2009).

Latour speculates that there is a trend in these informational cascades, which is certainly reflected in the ongoing digitalisation of arts, humanities and social science projects that tends towards ‘the direction of the greater merging of figures, numbers and letters, merging greatly facilitated by their homogenous treatment as binary units in and by computers’ (Latour, 1986: 16). The financial considerations are also new with these computational disciplines, as they require more money and organisation than the old individual scholar of lore did. Not only are the start-up costs correspondingly greater, usually needed to pay for the researchers, computer programmers, computer technology, software, digitisation costs, etc., but there are real questions about sustainability of digital projects, such as: ‘Who will pay to maintain the digital resources?’ ‘Will the user forums, and user contributions, continue to be monitored and moderated if we can’t afford a staff member to do so? Will the wiki get locked down at the close of funding or will we leave it to its own devices, becoming an online-free-for all?’ (Terras, 2010). It also raises a lot of new ethical questions for social scientists and humanists to grapple with. As argued in Nature,

For a certain sort of social scientist, the traffic patterns of millions of e-mails look like manna
from heaven. Such data sets allow them to map formal and informal networks and pecking orders, to see how interactions affect an organization’s function, and to watch these elements evolve over time. They are emblematic of the vast amounts of structured information opening up new ways to study communities and societies. Such research could provide much-needed insight into some of the most pressing issues of our day, from the functioning of religious fundamentalism to the way behaviour influences epidemics... But for such research to flourish, it must engender that which it seeks to describe... Any data on human subjects inevitably raise privacy issues, and the real risks of abuse of such data are difficult to quantify, (Nature, 2007)

For Latour, ‘sociology has been obsessed by the goal of becoming a quantitative science. Yet it has never been able to reach this goal because of what it has defined as being quantifiable within the social domain...’. Thus, he adds, ‘[i]t is indeed striking that at this very moment, the fast expanding fields of “data visualisation”, “computational social science” or “biological networks” are tracing, before our eyes, just the sort of data’ that sociologists such as Gabriel Tarde, at the turn of the 20th century, could merely speculate about (Latour, 2010: 116).

Further, it is not merely the quantification of research which was traditionally qualitative that is offered with these approaches. Rather, as Unsworth argues, we should think of these computational ‘tools as offering provocations, surfacing evidence, suggesting patterns and structures, or adumbrating trends’ (Unsworth, quoted in Clement et al., 2008). For example, the methods of ‘cultural analytics’ make it possible, through the use of quantitative computational techniques, to understand and follow large-scale cultural, social and political processes for research projects – that is, it offers massive amounts of literary or visual data analysis (see Manovich and Douglas, 2009). This is a distinction that Moretti (2007) referred to as distant versus close readings of texts. As he points out, the traditional humanities focuses on a ‘minimal fraction of the literary field’,

A canon of two hundred novels, for instance, sounds very large for nineteenth-century Britain
(and is much larger than the current one), but is still less than one per cent of the novels that were actually published: twenty thousand, thirty, more, no one really knows -- and close reading won’t help here, a novel a day every day of the year would take a century or so... And it’s not even a matter of time, but of method: a field this large cannot be understood by stitching together separate bits of knowledge about individual cases, because it isn’t a sum of individual cases: it’s a collective system, that should be grasped as such, as a whole, (Moretti, 2007: 3-4)

It is difficult for the traditional arts, humanities and social sciences to completely ignore the large-scale digitalisation effort going on around them, particularly when large quantities of research money are available to create archives, tools and methods in the digital humanities and computational social sciences. However, less understood is the way in which the digital archives being created are deeply computational in structure and content, because the computational logic is entangled with the digital representations of physical objects, texts and ‘born digital’ artefacts. Computational techniques are not merely an instrument wielded by traditional methods; rather they have profound effects on all aspects of the disciplines. Not only do they introduce new methods, which tend to focus on the identification of novel patterns in the data as against the principle of narrative and understanding, they also allow the modularisation and recombination of disciplines within the university itself.

Computational approaches facilitate disciplinary hybridity that leads to a post-disciplinary university -- which can be deeply unsettling to traditional academic knowledge. Software allows for new ways of reading and writing. For example, this is what Tanya Clement says on the distant reading of Gertrude Stein’s *The Making of Americans,*

*The Making of Americans* was criticized by [those] like Malcolm Cowley who said Stein’s ‘experiments in grammar’ made this novel ‘one of the hardest books to read from beginning to end that has ever been published’.... The highly repetitive nature of the text, comprising almost 900 pages and 3174 paragraphs with only approximately 5,000 unique words, makes
keeping tracks of lists of repetitive elements unmanageable and ultimately incomprehensible... [However] text mining allowed me to use statistical methods to chart repetition across thousands of paragraphs... facilitated my ability to read the results by allowing me to sort those results in different ways and view them within the context of the text. As a result, by visualizing clustered patterns across the text’s 900 pages of repetitions... [th]is discovery provides a new key for reading the text as a circular text with two corresponding halves, which substantiates and extends the critical perspective that Making is neither inchoate nor chaotic, but a highly systematic and controlled text. This perspective will change how scholars read and teach The Making of Americans. (Clement, quoted in Clement, Steger, Unsworth, and Uszkalo, 2008)

I wouldn’t want to overplay the distinction between pattern and narrative as differing modes of analysis. Indeed, patterns implicitly require narrative in order to be understood, and it can be argued that code itself consists of a narrative form that allows databases, collections and archives to function at all. Nonetheless, pattern and narrative are useful analytic terms that enable us to see the way in which the computational turn is changing the nature of knowledge in the university and, with it, the kind of computational subject that the university is beginning to produce. As Bruce Sterling argues,

‘Humanistic heavy iron’: it’s taken a long time for the humanities to get into super computing, and into massive database management. They are really starting to get there now. You are going to get into a situation where even English professors are able to study every word ever written about, or for, or because of, Charles Dickens or Elizabeth Barrett Browning. That’s just a different way to approach the literary corpus. I think there is a lot of potential there. (Sterling, 2010)

Indeed, there is a cultural dimension to this process and, as we become more used to computational visualisations, we will expect to see them and use them with confidence and fluency. The computational subject is a key requirement for a data-centric age,
certainly when we begin to look at case studies that demonstrate how important a computational comportment can be in order to perform certain forms of public and private activities in a world that is increasingly pervaded by computational devices. In short, Bildung is still a key idea in the digital university, not as a subject trained in a vocational fashion to perform instrumental labour, nor as a subject skilled in a national literary culture, but rather as a subject which can unify the information that society is now producing at increasing rates, and which understands new methods and practices of critical reading (code, data visualisation, patterns, narrative) and is open to new methods of pedagogy to facilitate it. Indeed, Presner (2010) argues that the digital humanities must be engaged with the broad horizon of possibilities for building upon excellence in the humanities while also transforming our research culture, our curriculum, our departmental and disciplinary structures, our tenure and promotion standards, and, most of all, the media and format of our scholarly publications. (Presner, 2010: 6)

This is a subject that is highly computationally communicative, and that is also able to access, process and visualise information and results quickly and effectively. At all levels of society, people will increasingly have to turn data and information into usable computational forms in order to understand it at all. For example, one could imagine a form of computational journalism that enables the public sphere function of the media to make sense of the large amount of data which governments, amongst others, are generating, perhaps through increasing use of ‘charticles’, or journalistic articles that combine text, image, video, computational applications and interactivity (Stickney, 2008). This is a form of ‘networked’ journalism that ‘becomes a non-linear, multi-dimensional process’ (Beckett, 2008: 65). Additionally, for people in everyday life who need the skills that enable them to negotiate an increasingly computational field – one need only think of the amount of data in regard to managing personal money, music, film, text, news, email, pensions, etc. – there will be calls for new skills of financial and technical literacy, or, more generally, a computational literacy or computational pedagogy that the digital humanities could contribute to.
Humanity and the humanities

As the advantages of the computational approach to research (and teaching) become persuasive to the positive sciences, whether history, biology, literature or any other discipline, the ontological notion of the entities they study begins to be transformed. These disciplines thus become focused on the *computationality* of the entities in their work. Here, following Heidegger, I want to argue that there remains a location for the possibility of philosophy to explicitly question the ontological understanding of what the computational is in regard to these positive sciences. Computationality might then be understood as an ontotheology, creating a new ontological ‘epoch’ as a new historical constellation of intelligibility. The digital humanists could therefore orient themselves to questions raised when computationality is itself problematized in this way (see Liu 2011).

With the notion of ontotheology, Heidegger is following Kant’s argument that intelligibility is a process of filtering and organising a complex overwhelming world by the use of ‘categories’, Kant’s ‘discursivity thesis’. Heidegger historicizes Kant’s cognitive categories by arguing that there is ‘succession of changing historical ontotheologies that make up the “core” of the metaphysical tradition. These ontotheologies establish “the truth concerning entities as such and as a whole”, in other words, they tell us both what and how entities are – establishing both their essence and their existence’ (Thomson, 2009: 149-150). Metaphysics, grasped ontotheologically, ‘temporarily secures the intelligible order’ by understanding it ‘ontologically’, from the inside out, and ‘theologically’, from the outside in, which allows the formation of an epoch, a ‘historical constellation of intelligibility which is unified around its ontotheological understanding of the being of entities’ (Thomson, 2009: 150). As Thomson argues:

*The positive sciences all study classes of entities... Heidegger... [therefore] refers to the positive sciences as ‘ontic sciences’. Philosophy, on the other hand, studies the being of those classes of entities, making philosophy an ‘ontological science’ or, more grandly, a ‘science of being’* (Thomson 2003: 529).

Philosophy as a field of inquiry, one might argue, should have its ‘eye on the whole’, and it is this focus on ‘the landscape as a whole’ which
distinguishes the philosophical enterprise and which can be extremely useful in trying to understand these ontotheological developments (Sellars, 1962: 36). If code and software are to become objects of research for the humanities and social sciences, including philosophy, we will need to grasp both the ontic and ontological dimensions of computer code. Broadly speaking, then, this paper suggests that we take a philosophical approach to the subject of computer code, paying attention to the wider aspects of code and software, and connecting them to the materiality of this growing digital world. With this in mind, the question of code becomes central to understanding in the digital humanities, and serves as a condition of possibility for the many computational forms that mediate out experience of contemporary culture and society.

Endotes

1 HTML is the HyperText Markup Language used to encode webpages. AJAX is shorthand for Asynchronous JavaScript and XML, which is a collection of client side technologies that enable an interactive and audio-visual dynamic web.

2 I am indebted to Alan Finlayson for his comments on this section.

3 For example in The Idea of a University (Newman, 1996) and Culture and Anarchy (Arnold, 2009).

4 See http://www.bcs.org/server.php?show=nav.5829

5 What Heidegger calls ‘the Danger’ (die Gefahr) is the idea that a particular ontotheology should become permanent, particularly the ontotheology associated with technology and enframing (see Heidegger 1993).

6 See Thomson (2003: 531) for a discussion of how Heidegger understood this to be the role of philosophy.

7 Kirschenbaum argues:

I believe such trends will eventually affect the minutiae of academic policy. The English
department where I teach, like most which offer the doctorate, requires students to demonstrate proficiency in at least one foreign language. Should a graduate student be allowed to substitute demonstrated proficiency in a computer-programming language instead? Such questions have recently arisen in my department and elsewhere; in my own case, almost a decade ago, I was granted permission to use the computer language Perl in lieu of proficiency in the second of two languages that my department required for the Ph.D. I successfully made the case that given my interest in the digital humanities, this was far more practical than revisiting my high-school Spanish. (Kirschenbaum 2009, emphasis added)

8 This does not preclude other more revolutionary human-computer interfaces that are under development, including haptic interfaces, eye control interfaces, or even brain-wave controlled software interfaces.

9 See [http://www.thecomputationalturn.com/](http://www.thecomputationalturn.com/)

10 See the open digital humanities translation of Plato’s *Protagoras* for a good example of a wiki-based project, [http://openprotagoras.wikidot.com/](http://openprotagoras.wikidot.com/)

11 Here I don’t have the space to explore the possibilities of a transformation of the distinction between research and teaching by digital technologies, themselves a result of the Humboldtian notion of the university. We might consider that a new hybridized form of research-teaching or teaching-research might emerge, driven, in part, by the possibility of new knowledges being created and discovered within the teaching process itself. This would mean that the old distinctions of research as creative, and teaching as dissemination would have to change too.

References


http://cnx.org/content/m34246/1.6/?format=pdf


http://www.humanitiesblast.com/manifesto/Manifesto_V2.pdf


http://www.wired.com/beyond_the_beyond/2010/02/atemporality-for-the-creative-artist/


