Processing Humanities Multimedia

Garth Evans
Josh Romphf
Welcome to DHSI 2019!

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.

Do check in there first if you need anything that’s not in this coursepak.

To access the DHSI wifi network, simply go into your wireless settings and connect to the “DHSI” network and enter the password “dhsi2019”.

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 ....
DHSI Wi-Fi

Network name: DHSI
Passkey: dhsi2019
The 2019 schedule is just taking shape nicely! A very few things to confirm, add, etc, still 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 Penninsula. 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!

9:00 to 4:00

Early Class Meeting: 4. [Foundations] DH For Department Chairs and Deans (David Strong Building C124, 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, 3 June 2019

Your hosts for the week are Alyssa Arbuckle, Ray Siemens, and Jannaya Friggstad Jensen.

7:45 to 8:15

Last-minute Registration (MacLaurin Building, Room A100)
Welcome, Orientation, and Instructor Overview (MacLaurin A144)
- Welcome to the Territory
- Welcome to DHSI: Ray Siemens, Alyssa Arbuckle
- Welcome from UVic: Jonathan Bengtson (University Librarian), Alexandra D'Arcy (Associate Dean Research, Humanities)

Classes in Session (click for details and locations)
- 1. [Foundations] Digitisation Fundamentals and their Application (Clearihue A103, Lab)
- 2. [Foundations] Introduction to Computation for Literary Criticism (Clearihue A102, Lab)
- 4. [Foundations] DH For Department Chairs and Deans (David Strong Building C124, Classroom)
- 5. [Foundations] Developing a Digital Project (With Omeka) (Clearihue A031, Lab)
- 9. Out-of-the-Box Text Analysis for the Digital Humanities (Human and Social Development A160, Lab)
- 10. Sound and Digital Humanities (Cornett A120, Classroom)
- 11. Critical Pedagogy and Digital Praxis in the Humanities (Clearihue D132, Classroom)
- 12. Digital Humanities for Japanese Culture: Resources and Methods (McPherson Library A003, Classroom)
- 14. Retro Machines & Media (McPherson Library 129, Classroom)
- 15. Geographical Information Systems in the Digital Humanities (Clearihue A105, Lab)
- 16. Introduction to IIIF: Sharing, Consuming, and Annotating the World’s Images (Cornett A121, Classroom)
- 17. Web APIs with Python (Human and Social Development A170, Lab)
- 18. Ethical Data Visualization: Taming Treacherous Data (Cornett A128, Classroom)
- 19. Linked Open Data and the Semantic Web (Cornett A132, Classroom)
- 20. Palpability and Wearable Computing (McPherson Library A025, Classroom)
- 21. The Frontend: Modern JavaScript & CSS Development (Clearihue A030, Lab)
- 25. Information Security for Digital Researchers (David Strong Building C114, Classroom)

Lunch break / Unconference Coordination Session (MacLaurin A144)
(Grab a sandwich and come on down!)
Discussion topics, scheduling, and room assignments from among all DHSI rooms will be handled at this meeting.

Classes in Session

Institute Lecture: Jacqueline Wernimont (Dartmouth C): "Sex and Numbers: Pleasure, Reproduction, and Digital Biopower"
Chair: Anne Cong-Huyen (U Michigan) (MacLaurin A144)

Abstract: Drawing from Numbered Lives (MIT 2018), this talk will consider a long history of sex-number entanglement in Anglo-American Cultures. Drawing on historical and contemporary objects and practices, Wernimont will ask "in what ways do theories of biopower, critical gender and critical race studies, and media studies" suggest that we can understand this set of entanglements and their impacts. NB: While relevant, this talk will not include discussions of sexual trauma or violence. It will include frank discussion of sex acts and various ways of translating sexual behavior into numbers.

Opening Reception (University Club)

Tuesday, 4 June 2019

Classes in Session

Lunch break / Unconference
*Mystery* Lunches

Classes in Session

DHSI Conference and Colloquium Lightning Talk Session 1 (MacLaurin A144)
Wednesday, 5 June 2019

9:00 to Noon
Classes in Session

Lunch break / Unconference

"Mystery" Lunches

Presentation: An Introduction to Scholarly Publishing with Manifold (MacLaurin A144)
Lunch included for those who [register here]

This presentation introduces Manifold Scholarship, a Mellon-funded digital publishing platform developed by the CUNY Graduate Center, The University of Minnesota Press, and Cast Iron Coding. Manifold allows you to create beautiful, dynamic open access projects that can include text, images, video, embedded resources, and social annotation. We will provide an overview of Manifold and demonstrate how faculty, students and staff in the digital humanities can use Manifold to publish open access scholarly works, conduct and participate in peer review, and create custom edited versions of public domain course texts and OER.

1:30 to 4:00
Classes in Session

DHSI Conference and Colloquium Lightning Talk Session 2 (MacLaurin A144)
Chair: Kim O'Donnell (Simon Fraser U)

• Catherine Ryu (Michigan State U), "Tone Perfect: Developing a Multimodal Audio Database for Mandarin Chinese as an Open Source"
• Kenzie Burchell (U Toronto Scarborough), "Making Responsible Reporting Practices Visible: Comparing newswire coverage of humanitarian crises in Syria"
• Jessica Linzel (Brock U), "The Shopkeeper Aristocracy: Mapping Trade Networks in Colonial Niagara"
• Kirsten Painter (U Washington), "From Bogatyr’s to Bread: Digitization & Online Exhibition of Rare Russian Children’s Books at the U Washington"
• John Barber (Washington State U), "A Mighty Span"

4:15 to 5:15
DHSI Conference and Colloquium Lightning Talk Session 3 (MacLaurin A144)
Chair: Kim O'Donnell (Simon Fraser U)

• Colleen Kolba (U South Florida), "What Comics can Teach our Students about Multimodal Literacy"
• Trish Baer (ETCL; U Victoria), "Preserving Digital Legacies: Archived Websites and Digital Discoverability"
• Suchismita Dutta (U Miami), "The Importance of Archival Transcription for Genre Building"
• Jeffrey Lawler (California State U, Long Beach), "Twining our way through the Past: Video Game Authoring as History Pedagogy"

Wednesday, 5 June 2019

6:00 to 8:00
DHSI Newcomer’s Gathering (Grad House Restaurant, Graduate Student Centre)
Come down, buy meal and a beverage, and make some new friends!

Thursday, 6 June 2019

9:00 to Noon
Classes in Session

12:15 to 1:15
"Mystery" Lunches
[Instructor lunch meeting]

1:30 to 4:00
Classes in Session

DHSI Conference and Colloquium Lightning Talk Session 3 (MacLaurin A144)
Chair: Kim O'Donnell (Simon Fraser U)

• Colleen Kolba (U South Florida), "What Comics can Teach our Students about Multimodal Literacy"
• Trish Baer (ETCL; U Victoria), "Preserving Digital Legacies: Archived Websites and Digital Discoverability"
• Suchismita Dutta (U Miami), "The Importance of Archival Transcription for Genre Building"
• Jeffrey Lawler (California State U, Long Beach), "Twining our way through the Past: Video Game Authoring as History Pedagogy"
Friday, 7 June 2019 [DHSI; ADHO Pedagogy SIG Conference Opening]

9:00 to Noon
Classes in Session

12:15 to 1:15
Lunch Reception / Course E-Exhibits (MacLaurin A100)

1:30 to 1:50
Remarks, A Week in Review (MacLaurin A144)

2:00 to 3:00
Joint Institute Lecture (DHSI and ADHO Pedagogy SIG Conference):
Matt Gold (CUNY Graduate Center and Association for Computers and the Humanities): “Thinking Through DH: Proposals for Digital Humanities Pedagogy”
Chair: Diane Jakacki (Bucknell U)
(MacLaurin A144)

Abstract: How do we teach digital humanities, and how should DH be taught? What, indeed, should we teach when we teach DH? This talk will present a proposal for grounding digital humanities pedagogical practice in the research interests of our students and the epistemological foundations of our methods rather than through an approach grounded more central in data and methods.

3:30 to 5:00
Joint Reception: DHSI and ADHO Pedagogy SIG Conference (University Club)
E-Poetry Event (Chris Tanasescu)
Watch this space for details, including how to participate!

DHSI Conference and Colloquium Poster/Demo Session
- Pia Russel (U Victoria); Emily Stremel (U Victoria), “British Columbia’s Historical Textbooks Digital Library”
- Cody Hennesy (U Minnesota); Rachael Samberg (U California, Berkeley); Stacy Reardon (U California, Berkeley), “Finding the Haystack: Literacies for Accessing and Using Text as Data”
- Paula Johanson (ETCL; Independent Scholar), “Proving Seahorses and Juan de Fuca’s Travels in The Curve of Time”
- Tara Baillargeon (Marquette U); Elizabeth Wawrzyniak (Marquette U), “FellowsHub: J. R. R. Tolkien Fanzine Portal”
- Caterina Agostini (Rutgers U), “Art at the Time of Syphilis: A First-Person Medical Narrative in Benvenuto Cellini’s Vita”
- Lauren Elle DeGaine (ETCL; U Victoria), “Women at the Front: A Digital Exhibit of Victorian Frontpiece Illustrations”
- Adam Griggs (Mercer U); Kathryn Wright (Mercer U); Christian Pham (Mercer U); Gail Morton (Mercer U); Stephanie Miranda (Mercer U), “Digitizing Middle Georgia’s History of Slavery”

Saturday, 8 June 2019 [Conference, Colloquium, and Workshop Sessions]

8:00 to 9:00
Conference / Workshop Registration (MacLaurin A100)
The day’s events are included with your DHSI registration. If you’re not registered in DHSI, you’re very welcome to join us by registering here as a Conference / Colloquium / Workshop participant. We’ll have a nametag waiting for you!

Coffee, Tea, &c? Looking for some morning coffee or tea, or a small nibble? Options and hours of operation for weekend campus catering are available here. Mystic Market usually opens around 10.00.

9:00 to 4:00
DHSI Conference and Colloquium Sessions
ADHO Pedagogy SIG Conference Sessions
Right2Left Workshop Sessions

9:00 to 4:00
All Day DHSI Workshop Session (click for workshop details and free registration for DHSI participants)
- 55. Introduction to Machine Learning in the Digital Humanities [8-9 June; All day, each day] (David Strong Building C124, Classroom)

9:00 to 9:10
Informal Greetings, Room Set-up (Lobby, outside Hickman 105)

Session 1
DHSI Colloquium and Conference (Hickman 105)
Digital Humanities & Literature, Chair: Kim O’Donnell (Simon Fraser U)
- Youngmin Kim (Dongguk U), “Transdiscursivity in the Convergence of Digital Humanities and World Literature”
- Caroline Winter (U Victoria), “Digitizing Adam Smith’s Literary Library”
- Kaitlyn Fralick (U Victoria); Kailey Fukushima (U Victoria); Sarah Karlson (U Victoria), “Victorian Poetry
9:10 to 10:30  
**ADHO Pedagogy SIG Conference** (*Hickman 110*)  
Chair: Katherine Faull (Bucknell U)  
- Aaron Tucker and Nada Savicevic (Ryerson U), "Write Here, Right Now: An Open Source eTextbook for the Flipped Classroom"  
- Heather McAlpine (U Fraser Valley), "Digital Meters: Using Text Encoding to Teach Literature in the Undergraduate Classroom"  
- Tiina H. Airaksinen (U Helsinki), "Digital Humanities in Cultural Studies: Creating a MOOC course for University Students and A-Level Students"  

**Right2Left Workshop** (*Hickman 116*)  
Keynote - Nathan P. Gibson (Ludwig Maximilians U, München): "Thinking in -JTR: Reorienting the Directional Assumptions of Global Digital Scholarship"

10:30 to 10:40  
**Break**

10:40 to Noon  
**Session 2**  
**DHSI Colloquium and Conference** (*Hickman 105*)  
Digital Humanities & Society, Chair: Eleanor Reed (Hastings C)  
- Joel Zapata (Southern Methodist U), "Uncovering the Southern Plains’ Mexican American Civil Rights Movement"  
- Ayo Oseisanwo (U Ibadan), "Online Newspaper Construction of Agitation for the Sovereign State of Biafra in Nigeria"  
- Joseph Jones (U British Columbia), "Testbed for an Approach to Distant Reading: Fictions That Represent Vietnam War Resisters in Canada"  
- Brendan Mackie (U California, Berkeley), "Visualizing Long-Term Cultural Change: An Example From The Birth of Civil Society"

**ADHO Pedagogy SIG Conference** (*Hickman 110*)  
Chair: Laura Estill (St Francis Xavier U)  
- Jane Jackson (Chinese U of Hong Kong), "Interrogating digital spaces for intercultural meaning-making"  
- Ryan Ikeda (UC Berkeley), "Disrupting Digital Literacy: Situating Electronic Literature Among Public Education Initiatives"  
- Christopher Church, Katherine Hepworth (U Nevada, Reno), "We’re STEAMed! A call for balancing technical instruction and disciplinary content in the digital humanities"  
- Chelsea Milbourne (Cal Poly, San Luis Obispo), "Finding the Right Fit between Technology and Class Content: Reflections on Including Web Development in a Digital Storytelling Course"

**Right2Left Workshop** (*Hickman 116*)  
- Edward “Eddie” Surman (Claremont Graduate U), "Qualitative Digital Text Analysis and #Right2Left Languages: A Demonstration of Atlas.ti using the Hebrew Bible"

Noon to 1:10  
**Lunch** (We recommend *Mystic Market* on weekends!)

1:10 to 2:30  
**Session 3**  
**DHSI Colloquium and Conference** (*Hickman 105*)  
Digital Humanities & Community, Chair: Claire Carlin (U Victoria)  
- Pia Russel (U Victoria); Emily Stremel (U Victoria), "Mentorship and disability: Supporting disabled employees in digital humanities"  
- Amy Lueck (Santa Clara U), "Virtually Emplacing Indigenous Memory"  
- Md. Shehabul Alam (National U Bangladesh), "Integrating Library Service with Union Information and Service Center: A Joint Initiative towards Digital Bangladesh"  
- Veronica Gomez (Instituto de Humanidades y Ciencias Sociales (HuCSo) - UNL-CONICET), "Latin American E-literature and Location: The Nation Revisited in Electronic Literature Organization (ELO)"  

**ADHO Pedagogy SIG Conference** (*Hickman 110*)  
Chair: Chris Tănăsescu (UC Louvain)  
- Laura Estill (St Francis Xavier U), "One Assignment, Three Ways: Assessing DH Projects in a Literature Course"  
- Felix Bayode Oke, Stella N. Kpolugbo (Anchor U Lagos), "The Multimodal Technique as a Pedagogical Tool in Pelu Awofeso’s White Lagos: A Definitive and Visual Guide to the Eyo Festival"  
- Shu Wan (U Iowa), "A digital "historical gaze" of Chinese students in Iowa, 1911-1930"  
- Francesca Giannetti (Rutgers U, New Brunswick), "So near while apart: Correspondence Editions as Critical Library Pedagogy and Digital Humanities Methodology"

**Right2Left Workshop** (*Hickman 116*)  
- Najla Jarkas (American U Beirut) and David Joseph Wrisley (NYU Abu Dhabi), "RTL Software Localization and Digital Humanities: the Case Study of Translating Voyant Tools into Arabic"
2:30 to 2:40
Break

Session 4

DHSI Colloquium and Conference (Hickman 105)
Digital Humanities & Media, Chair: Caroline Winter (U Victoria)
- Ashleigh Cassmere-Stanfield (U Chicago), “Sonifying Hamlet and Reading the Room”

ADHO Pedagogy SIG Conference (Hickman 110)
Chair: Aaron Tucker (Ryerson U)
Youngmín Kim (Dongguk U), “Teaching Digital Humanities and World Literature in Class”
Alice Fleerackers, Juan Pablo Alperin, Esteban Morales, Remi Kalir (Simon Fraser U, U Colorado Denver), “Online annotations in the classroom: How, why, and what do students learn from annotating course material?”
Andie Silva (York C and Graduate Center, CUNY), “Keeping it Local: Undergraduate DH as Feminist Practice”

Right2Left Workshop (Hickman 116)
- Joanna Byszuk (Institute of Polish Language, Polish Academy of Sciences, Warsaw/Computational Stylistics Group) and Alexey Khismatulin (Institute of Oriental Manuscripts, Russian Academy of Sciences, Saint Petersburg), ”Attribution of Authorship for Medieval Persian Quasidas with Stylometry”
- Ilan Benattar (New York U), ”#Right2Left Biblical Translations in Jewish Textual History: Case Studies in Judeo-Arabic and Judeo-Spanish”

Sunday, 9 June 2019 [Workshop Sessions]

8:00 to 5:00
The day’s events are included with your DHSI registration. If you’re not registered in DHSI, you’re very welcome to join us by registering here as a Conference / Colloquium / Workshop participant. We’ll have a nametag waiting for you!

Coffee, Tea, &c?
Looking for some morning coffee or tea, or a small nibble? Options and hours of operation for weekend campus catering are available here. Mystic Market usually opens around 10.00.

9:00 to 4:00
All Day Workshop Sessions (click for workshop details and free registration for DHSI participants)
- 55. Introduction to Machine Learning in the Digital Humanities [8-9 June; All day, each day] (David Strong Building C124, Classroom)
- 56. Pedagogy of the Digitally Oppressed: Anti-Colonial DH Methods and Praxis [9 June; All Day] (Hickman 116, Classroom)
- 57. Natural Language Processing and Network Coding Apps for Text & Textual Corpus Analysis in the Humanities [9 June; All Day] (David Strong Building C114, Classroom)

9:00 to Noon
AM Workshop Sessions (click for workshop details and free registration for DHSI participants)
- 59. 3D Visualization for the Humanities [9 June; AM] (Cornett A229, Classroom)
- 60. It’s All Relational: AbTeC’s Indigenous Video Game Workshops as Storytelling Praxis [9 June; AM] (Cornett A121, Classroom)
- 61. Spatial DH: De-Colonizing Cultural Territories Online [9 June; AM] (Clearihue D130, Classroom)
- 63. Creating a CV for Digital Humanities Makers [9 June; AM] (David Strong Building C108, Classroom)

Noon to 1:00
Lunch (We recommend Mystic Market on weekends!)

1:00 to 4:00
PM Workshop Sessions (click for workshop details and free registration for DHSI participants)
- 65. Indigenous Futures in the Classroom and Beyond [9 June; PM] (Cornett A121, Classroom)
- 66. DHSI Knits: History of Textiles and Technology [9 June; PM] (Fine Arts 109, Classroom)
- 68. Linked Open Datafication for Humanities Scholars [9 June; PM] (McPherson Library A003, Classroom)
- 69. Stylo - WYSIWYM Text Editor for Humanities Scholars [9 June; PM] (McPherson Library A025, Classroom)

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.

Monday, 10 June 2019
Your hosts for the week are Ray Siemens and Jannaya Friggstad Jensen.

**7:45 to 8:15**
DHSI Last-minute Registration (MacLaurin A100)

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

**10:15 to Noon**
Classes in Session (click for details and locations)
- 29. [Foundations] Understanding The Predigital Book: Technologies of Inscription (McPherson Library A003, Classroom)
- 30. [Foundations] Databases for Digital Humanists (McPherson Library 210, Classroom)
- 33. Digital Storytelling (Cornett A120, Classroom)
- 34. Text Encoding as Modelling (Clearihue D131, Classroom)
- 35. Stylometry with R: Computer-Assisted Analysis of Literary Texts (Clearihue A102, Lab)
- 36. Open Access and Open Social Scholarship (Clearihue D130, Classroom)
- 37. Digital Games as Tools for Scholarly Research, Communication and Pedagogy (Cornett A229, Classroom)

**12:15 to 1:15**
Lunch break / Unconference Coordination Session (MacLaurin A144)
(Grab a sandwich and come on down!)

**1:30 to 4:00**
Classes in Session

**4:10 to 5:00**
Institute Lecture: Angel David Nieves (San Diego State U): “3D Mapping and Forensic Traces of Testimony: Documenting Apartheid-Era Crimes Through the Digital Humanities”
Chair: Constante Crompton (U Ottawa)
(MacLaurin A144)

Abstract: In 1989 the killing of a queer, 14-year-old youth in Winnie Mandela’s house named Stompie Seipei (an event that few in South Africa are willing to recall, let alone discuss, in any detail) – is perhaps one of the most glaring examples where the queer and activist community was suppressed or erased from anti-apartheid/liberation histories. Digital humanities may actually help both reconstruct and recover a history that is still very early in the telling, despite what is commonly believed about the liberation struggle and the contributions of queer activists in the dismantling of apartheid. Perhaps it could explain why a youth such as Seipei was killed – or at the very least, provide a more complex and messy narrative that permits one to know more how the history of queer anti-apartheid activists was suppressed. This talk outlines a methodology for "messy thinking and writing" in the digital humanities that – through a queer and feminist intersectional framework – permits a more complex layering of oral histories and 3D historical reconstructions.

**5:00 to 6:00**
Reception (University Club)

---

**Tuesday, 11 June 2019**

**9:00 to Noon**
Classes in Session

**12:15 to 1:15**
Lunch break / Unconference

"Mystery" Lunches
<table>
<thead>
<tr>
<th>Time</th>
<th>Event Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:30 to 4:00</td>
<td>Classes in Session</td>
</tr>
<tr>
<td>6:00 to 8:00</td>
<td>DHSI Newcomer's Gathering (Grad House Restaurant, Graduate Student Centre) &lt;br&gt;Come down, buy meal and a beverage, and make some new friends!</td>
</tr>
<tr>
<td>9:00 to Noon</td>
<td>Classes in Session</td>
</tr>
<tr>
<td>12:15 to 1:15</td>
<td>Lunch break / Unconference</td>
</tr>
<tr>
<td>12:15 to 1:15</td>
<td>&quot;Mystery&quot; Lunches</td>
</tr>
<tr>
<td>12:15 to 1:15</td>
<td>Presentation: An Introduction Jupyter Notebooks for Researchers (MacLaurin A144) &lt;br&gt;This presentation introduces Jupyter Notebooks for researchers, via a partnership between Compute Canada and the Pacific Institute for the Mathematical Sciences (PIMS) including a large number of Canadian institutions. Read more here. Presenting is James Colliander, PIMS Director and team.</td>
</tr>
<tr>
<td>1:30 to 4:00</td>
<td>Classes in Session</td>
</tr>
<tr>
<td>6:00 to 7:00</td>
<td>&quot;Half Way There (yet again)!&quot; [An Informal, Self-Organized Birds of a Feather Get-Together] (Felicitas, Student Union Building) &lt;br&gt;Bring your DHSI nametag and enjoy your first tipple on us! [A great opportunity for an interest group meet-up]</td>
</tr>
<tr>
<td>9:00 to Noon</td>
<td>Classes in Session</td>
</tr>
<tr>
<td>12:15 to 1:15</td>
<td>Lunch break / Unconference</td>
</tr>
<tr>
<td>12:15 to 1:15</td>
<td>&quot;Mystery&quot; Lunches</td>
</tr>
<tr>
<td>12:15 to 1:15</td>
<td>[Instructor lunch meeting]</td>
</tr>
<tr>
<td>1:30 to 4:00</td>
<td>Classes in Session</td>
</tr>
</tbody>
</table>
| 4:10 to 5:00 | Institute Lecture: Karina van Dalen-Oskam (Huygens Institute and U Amsterdam; Alliance of Digital Humanities Organizations): “The Riddle of Literary Quality: Some Answers” <br>Chair: Aaron Mauro (Penn State, Behrend C) (MacLaurin A144) <br>Abstract: What is literature, and can you measure it? That is the key question of the project The Riddle of Literary Quality. “The Riddle” is a research project of the Huygens Institute for the History of the Netherlands (Amsterdam) in collaboration with the Fryeke Akademy (Leeuwarden) and the Institute for Logic, Language and Computation (University of Amsterdam). The Riddle combines computational analysis of writing style with the results of a large online survey of readers, completed by almost 14,000 participants. In my talk, I will go into
some of the main results of the project.

Friday, 14 June 2019

9:00 to Noon
Classes in Session

12:15 to 1:15
Lunch Reception / Course E-Exhibits (MacLaurin A100)

1:30 to 2:00
Closing, DHSI in Review (MacLaurin A144)

Contact info:
institut@uvic.ca  P: 250-472-5401  F: 250-472-5681
Processing Humanities Multimedia - DHSI 2018

Instructors:

Garth Evans (garth.evans@ubc.ca)
Josh Romphf (jromphf@library.rochester.edu)

Description:

From YouTube, to image repositories, to podcasts, to scraping media from web services like eBay, Reddit, and 4Chan the wealth of information available to humanities scholars that falls outside the realm of “traditional sources” is staggering and will continue to increase for the foreseeable future. Traditional scholarly approaches will still have their place among these new media objects but will frequently need to be used in conjunction with methods for handling large volumes of new media. But what are these methods and when/how are they used? This course answers these questions by starting from a basic introduction to media types and their potential research value and then leading the hands-on process for building a pipeline for processing each, from collecting the material through to processing it and finally storing it. Exact sources of the media to be used are still being considered but still images, sound files, and video will all feature prominently. No previous experience working with media files of any type is required but would certainly be an asset.

Introduction:

Welcome everyone! Throughout the week we’ll be experimenting with a variety of media processing technologies. The course will primarily be workshop based, with plenty of time for independent projects and troubleshooting. While we’ll be covering a lot of technologies, one thing we won’t be doing is shooting any video or recording any audio (at least not in the traditional “production” sense). Rather, we’ll be manipulating existing media files, so we highly encourage you to bring some of your own to work with, but we’ll also supply several examples. Please refer to the table below for the week’s structure. We’re intentionally keeping this pretty loose; so if there’s some subject matter you’d like to touch on, please let us know. Given the nature of media processing, many of the tools we’ll be using require “low level” access to the underlying hardware of our machines. Similarly, most of these tools need to be compiled for native processor architectures. What this means is that we may be spending a bit of time installing
some of these tools. Since our hope isn’t for us to be in “compiler hell,” we’ll rely on Compute Canada’s vast resources in order to do a lot of this work remotely. For those that would like to work locally, we’ll also supply some build scripts and instructions well in advance of our class. Finally, we think it’s important to stress that no programming or multimedia experience is required for this course. With a bit of curiosity and patience, we can all have a great time.

Agenda:

<table>
<thead>
<tr>
<th></th>
<th>Monday</th>
<th>Tuesday</th>
<th>Wednesday</th>
<th>Thursday</th>
<th>Friday</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Morning</strong></td>
<td>Introduction to the course / DHSI / any housekeeping</td>
<td>Introduction to Multimedia Formats: Image, Audio, and Video</td>
<td>Working with Media Collections: Data Retrieval (Internet Archive Case Study)</td>
<td>Brainstorm and Work on Multimedia Projects</td>
<td>Work on Multimedia Projects / Closing Remarks</td>
</tr>
<tr>
<td><strong>Afternoon</strong></td>
<td>From Analog to Digital: A Brief History of Recorded Image and Sound</td>
<td>Digging into Multimedia Objects: Command Line Processing</td>
<td>What’s Next: Introduction to Extended Reality</td>
<td>Work on Multimedia Projects</td>
<td>DHSI Project Show and Tell</td>
</tr>
</tbody>
</table>
Processing Humanities Multimedia – DHSI 2018

Instructors:

Garth Evans (garth.evans@ubc.ca)
Josh Romphf (jromphf@library.rochester.edu)

Description:

From YouTube, to image repositories, to podcasts, to scraping media from web services like eBay, Reddit, and 4Chan the wealth of information available to humanities scholars that falls outside the realm of “traditional sources” is staggering and will continue to increase for the foreseeable future. Traditional scholarly approaches will still have their place among these new media objects but will frequently need to be used in conjunction with methods for handling large volumes of new media. But what are these methods and when/how are they used? This course answers these questions by starting from a basic introduction to media types and their potential research value and then leading the hands-on process for building a pipeline for processing each, from collecting the material through to processing it and finally storing it. Exact sources of the media to be used are still being considered but still images, sound files, and video will all feature prominently. No previous experience working with media files of any type is required but would certainly be an asset.

Introduction:

Welcome everyone! Throughout the week we’ll be experimenting with a variety of media processing technologies. The course will primarily be workshop based, with plenty of time for independent projects and troubleshooting. While we’ll be covering a lot of technologies, one thing we won’t be doing is shooting any video or recording any audio (at least not in the traditional “production” sense). Rather, we’ll be manipulating existing media files, so we highly encourage you to bring some of your own to work with, but we’ll also supply several examples. Please refer to the table below for the week’s structure. We’re intentionally keeping this pretty loose; so if there’s some subject matter you’d like to touch on, please let us know. Given the nature of media processing, many of the tools we’ll be using require “low level” access to the underlying hardware of our machines. Similarly, most of these tools need to be compiled for native processor architectures. What this means is that we may be spending a bit of time installing
some of these tools. Since our hope isn’t for us to be in “compiler hell,” we’ll rely on Compute Canada’s vast resources in order to do a lot of this work remotely. For those that would like to work locally, we’ll also supply some build scripts and instructions well in advance of our class. Finally, we think it’s important to stress that no programming or multimedia experience is required for this course. With a bit of curiosity and patience, we can all have a great time.

**Agenda:**

<table>
<thead>
<tr>
<th></th>
<th>Monday</th>
<th>Tuesday</th>
<th>Wednesday</th>
<th>Thursday</th>
<th>Friday</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Afternoon</strong></td>
<td>From Analog to Digital: A Brief History of Recorded Image and Sound</td>
<td>Digging into Multimedia Objects: Command Line Processing</td>
<td>Working with Media Collections: Data Retrieval (Internet Archive Case Study)</td>
<td>Work on Multimedia Projects</td>
<td>DHSI Project Show and Tell</td>
</tr>
</tbody>
</table>
Digitization of cultural heritage over last 20 years has opened up very interesting possibilities for the study of our cultural past using computational “big data” methods. Today, as over two billion people create global “digital culture” by sharing their photos, video, links, writing posts, comments, ratings, etc., we can also use the same methods to study this universe of contemporary digital culture.

In this chapter I will discuss a number of issues regarding the “shape” of the digital visual collections we have, from the point of view of researchers who use computational methods. They are working today in many fields including computer science, computational sociology, digital art history, digital humanities, digital heritage and Cultural Analytics – which is the term I introduced in 2007 to refer to all of this research, and also to a particular research program of our own lab that has focused on exploring large visual collections.

Regardless of what analytical methods are used in this research, the analysis has to start with some concrete existing data. The “shapes” of existing digital collections may enable some research directions and make others more difficult. So what is the data universe created by digitization, what does it make possible, and also impossible?

The Islands and The Ocean

Before born-digital content, media creators first used physical and later electronic media (video and audio). Starting in the middle 1990s, gradually more and more of this content has being digitized. We can call such content born-analog.

The very first project to digitize cultural texts and make them freely available was Project Gutenberg that started in 1970. Today the largest sites for digitized content include Europeana (over 53 milion “artworks, artefacts, books, videos, and sounds from across
Europe” as of 2016), Digital Public Library of America (over 13 million items as of 2016), HathiTrust (13 million volumes as of 2015), Digital Collections at the Library of Congress and Internet Archive. The latter contains digital collections of various types of media ranging from largest collection of historical software to 10.7 billion historical texts (as of 12/2016).

The sites typically offer a number of useful ways to navigate these massive collections. For example, the Digital Public Library of America (DPLA) supports direct search, view by Timeline, Map view, and Thematic Exhibitions. Both DPLA and Europeana also encourage and help developers create experimental interfaces and apps that expand how their artifacts can be viewed and used. But in terms of using them for Cultural Analytics research, they do have one limitation. While the works in these and other collections can always be viewed online, not all works can be downloaded (or downloaded in mass using an API), because of the restrictions imposed by owners of the works.

The site which in my view is most interesting in this genre is Google Arts & Culture. It has fewer works but the most fluid interface. This site grew from the earlier Google Art Project that worked with many museums to scan artworks and then presented them online in a “virtual museum” interface. Today it offers virtual tours of many museums, millions of digitized artworks and photographs from the past, contemporary art. Media projects and photo stories are also created. The interfaces include zoom, timeline, search by color, thematic exhibitions, and also categories (artists, mediums, art movements, partners, names of objects, and places). When I was exploring the website (July 2016), it was offering 3,000 thematic exhibitions on all kinds of cultural topics. When we started our own Cultural Analytics Lab (culturalanalytics.info) in 2007, it was a bet. While contemporary culture was already well represented on the web, the kinds of large-scale online digital collections with multiple navigation functions and API like Europeana or DPLA did not yet exist. But I assumed that within the next few years, millions of digital images of historical art, photography and other media would become available. However, it was not clear at that time how inclusive they would become.

In the article I wrote about Cultural Analytics in March 2009, I described my experience of trying to use the existing digital image collections available at that time. I was interested in the following question: What did people paint around the world in 1930 – aside from a number of modernist “isms” that encompassed at best 150 artists (working in Paris, Amsterdam, Berlin and a few other cities) who are now included in the Western art historical canon? I was not thinking of “paintings in tens of thousands of small museums in small

---

5 Manovich 2009.
cities,” rather of paintings of nationally “important” artists that have entered in art history canons in their countries.

I did a search on artstor.org – a leading commercial service for digital images of art used in most art history classes in U.S. and also in other countries. In 2009 it already contained close to one million digital images of art, architecture and design. These images came from many important USA museums, art collections, and university libraries. To collect the images of artworks that are outside of the usual Western art historical canon on Artstor, we excluded Western Europe and North America from the search. This left the rest of the world: Eastern Europe, South-East Asia, East Asia, West Asia, Oceania, Central America, South America, and Africa. Not a small area! But when we searched Artstor for paintings done in these parts of the world in 1930, we only found a few dozen images. So, while there were very large numbers of images of paintings of canonical artists from Europe and USA painted in the same year, there were only a few images for a whole continent like East Asia.

This highly uneven distribution of digitized cultural artifacts is not due to Artstor’s choices. Artstor does not digitize images itself. Instead, it makes images available that have been submitted by museums and other cultural institutions. The results of our search reflects what participating museums collect and what they think should be digitized first. In other words, a number of major US collections and a slide library of a major research university (where by 2007 the proportion of Asian students was 45%) together contained only a few dozen paintings created outside of the West in 1930 which were digitized. In contrast, searching for Picasso returned around 700 images. Describing this example, I wrote in this 2009 article:

If this example is any indication, digital art repositories may be amplifying the already existed biases and filters of modern cultural canons. Instead of transforming the “top forty” into “the long tail,” digitization can be producing the opposite effect.

What remains outside of the digitized collections is all the rest: provincial nineteen century newspapers sitting in some library somewhere; millions of paintings in tens of thousands of small museums in small cities around the world; millions of thousands of specialized magazines in all kinds of fields and areas which no longer even exist; millions of home movies and photographs… This creates a problem for Cultural Analytics, which has a potential to map everything that remains outside the canon – and to begin writing a more inclusive cultural history without “great names.” We want to understand not only the exceptional but also the typical; not only the few “cultural sentences spoken by a few “great men” but the patterns in all cultural sentences spoken by everybody else; what is outside a few great museums rather than what is inside and what has already been discussed extensively and too many times.

---

6 The very first large institutional collection that formed the core of Artstor was the slide library of the University of California, San Diego (UCSD) – the same university where I had been teaching since 1996. The library had over 200,000 slides, and they were all digitized and included in Artstor. In 2009, this was the largest single collection in Artstor. The slides were either directly created by art history faculty teaching in Visual Art Department, or by art library staff following lists of images faculty provided. This collection is very interesting because it reflects the biases of art history as it was taught over a few decades when color slides were the main media for teaching and studying art.
I worried that what has been digitized, is only an “island,” and that a massive cultural “ocean” remains inaccessible for quantitative analysis. Luckily, such amplification of biases and focus on only “what is important” did not happen. Exploring the online libraries of digitized cultural artifacts seven years later, I am amazed by their richness and variety. The reason is that Europeana, DPLA, Library of Congress, NYPL, Internet Archive or Google Arts & Cultures do not just offer us images of high art like art museums. Instead, they are extensions of traditional libraries. And the libraries in modern times have an important function besides offering readers books and periodicals – they are places to which numerous people and organizations donate their archives. As these archives started to be digitized, an amazingly rich and varied historical cultural landscape started to emerge online.

For example, here are three examples among hundreds of digital image collections from the New York Public Library (NYPL):

“Photographs of The Catskill Water Supply System in Process of Construction.” 55 albumen print photographs created between 1906 and 1915.7

“Buttolph Collection of Menus” – A collection of Miss Frank E. Buttolph (1850–1924), a somewhat mysterious and passionate figure, whose mission in life was to collect menus donated to NYPL in 1899, 18,964 digitized items.8

“Catalog of the Chiroptera, by G. E. Dobson” – 31 digitized prints from a 1878 book.9

And here are examples listed in the blog post from europeana.eu referred to as “highlights of the new datasets ingested in the last months”:

Almost 100 objects (drawings, paintings, photographs) from Telegraph Museum in UK.
Over 3 000 photographs, XIX and XX century, mostly buildings from Culture Centre in Helsingborg.
Collection of 620 botanical drawings by Georg Schweinfurth from Botanic Garden and Botanical Museum Berlin-Dahlem.10

Comparing these collections with those of the digital image offerings of the largest art museums, we find that they are complete opposites of each other. Although modern art museums’ collections like that of libraries also developed through both their purchasing programs and donations, what was donated to them – or what museums chose to accept – was quite different. Libraries ended up housing millions of all kinds of heterogeneous items, few of them financially valuable. In contrast, modern art museums have traditionally focused on what has been recognized as very valuable. Indeed, original European “museums” included estates of very rich people, parts of royal palaces, or treasures of cathedrals and churches. For example, Vatican Museums originated in 1506 when Pope Julius II purchased

the ancient sculpture of Laocoön and his Sons and placed it on public display. (I should note that digitized collections of design and crafts museums such as Victoria and Albert in London or Cooper-Hewitt in New York are closer to that of libraries – their holdings are more varied and also organized in more categories than those of art museums.)

Libraries vs. Museums

However, there is also another aspect in museum’s history. Some of the original European museums contained not art but “curiosities.” One such famous museum is The Kunstkamera that was established in St. Petersburg in 1716 by Peter the Great to present “natural and human curiosities and rarities.” Another is the British Museum that opened in London in 1759, that initially showed a private collection of the physician and scientist Sir Hans Sloane.

Art history since the 20th century has created a highly controlled system that divides our visual heritage into “art” and everything else, and organizes the former by artists (their national origin, time period, and medium and style). The digital collections of art museums today also look ordered and systematic.

We are used to their ordered classifications. In comparison, the meta-collections of digitized visual artifacts by Europeana, DPLA and others may remind us of the cabinets of curiosities. Instead of a military-like “parades” of art history played in physical museums or on their sites, we find “trivia” and “ephemera.” (The latter word comes from Greek and New Latin where it referred to insects or flowers that were alive sometimes for less than a day.)

Browsing through page after page describing endless collections that often contain a few dozen or even only a few items – like the ones in the examples above – I often feel uncanny. In this view, the past looks un-periodic and un-systematized. Endless “deposits” of human material cultures have remained inside libraries, have then been digitized and are now connected by common metadata standards, web protocols, Javascript code, APIs and other computer machinery.

Labyrinth, kaleidoscope, Kunstkamera, Memex’ hypertext, random access memory, relational database – none of these models describe my experience of navigating digital cultural collections. For instance, consider Europeana with its 53 million items. The idea behind this massive multi-year project was to connect digitized artifacts from thousands of European museums and regional archives. So, rather than having to search all their individual sites, you can use the Europeana platform as a single point of access. The platform provides a common interface to all of the objects but it does not store them. They are stored at individual museums and archives. European Film Gateway, one of Europeana’s projects, does the same for dozens of European film archives.

Technically and conceptually, this works brilliantly. But experientially, the result has some unintended consequences. Instead of creating a kind of “united Europe” – a single pan-European space for cultural heritage – Europeana may be fragmenting it. As I browse through endless separate collections or individual items from these collections that fit my search terms, countries, geographic relations, and time periods are dissolved. Instead of a
“European” continent, it feels that I am looking at random survived files of many alien civilizations that got all mixed together.

This feeling is created by both very heterogeneous topics, and by equally heterogeneous styles. Photographs created in all kinds of techniques, engravings, etching, newspaper illustrations, covers of cigarette cases, early hand-colored photos, paintings … images are in rectangular formats, round frames, part of a text page, drawn in a corner of a hand written letter… texts typed, types set, hand written, printed on early dot matrix printers, carefully drawn with a brush … every possible subject and form of visual communication is here. (If Instagram platform during 2010–2015 can be thought as the extreme example of visual constraints, with all image being the same size and format and belonging to one medium, a digital historical collection is the other extreme).

But this heterogeneity, richness and variety is actually a good thing. It makes us aware of how rigid and limited our concepts of an “image” are today – a few clearly separated mediums, rectangular formats, and also separation between images and texts. So, while the abundance of communication “species” in digital libraries is on first sight disorientating – and it is certainly a challenge for large scale analysis using Computer Vision systems initially developed for contemporary photos – in the long run it is best for us. It forces us to face the human visual culture as it really exists historically – thousands of variations and their combinations, rather a net set of a small number of categories.

Cultural Sampling

The “islands” of digitized historical contents are constantly growing. But will they ever be big enough to let us understand the “ocean” – i.e., construct a sufficiently detailed map of the human visual history of the last few centuries? Richness and variety do not mean comprehensiveness. In other words: while digitization and organization of digitized items by Europeana, DPLA, and other projects continues, the most basic question for any quantitative study of cultural history remains unaddressed. This question is, how can we compile representative samples that systematically cover everything created in a particular period, geographic area and media – or in many such periods and areas together?11

Anthropologists do use sampling methods in their research when they excavate sites or study groups of people (such as in urban anthropology that looks at contemporary cities). But there is a basic large question which is more difficult to address: Since the kinds and quantities of artifacts that remained from various ancient civilizations vary significantly, do they together add to a representative sample? (Of course, as excavations of sites and analysis of new artifacts continue, this sample is being continuously refined.)

Since I am a historian of modern visual culture and media of the last 200 years, I am confident that for this period we do not have any comprehensive sample of visual culture in this period before the arrival of social media. So, while the “islands” are increasing

11 For an overview of different sampling methods, see Cook 2011 and Chambers / Skinner 2003.
in size and number, reconstructing the whole ocean maybe may become very difficult.
I am using the term “sample” in the sense it is used in statistics: a smaller subset of the
larger data. Constructing proper samples and determining the validity of predictions based
on these samples is a one of the main areas of statistics. In all social sciences including
sociology, demographics, psychology, and political science these questions are particularly
crucial, since these disciplines often use small human groups for surveys or observation.
Construction of proper samples is also crucial for marketing research, human-computer
interaction research and all other applied fields where researchers want to find people’s
attitude about existing products, interest in new products and new features, their lifestyle
aspirations, etc. And while the arrival of big social media data in the second part of the
2000s has changed the situation significantly, because now businesses can follow online
millions of individuals tracking what pages they visit, what they click on, which ads they look
at, and what they purchase, small groups continued to be widely used. (You can ask people
who agreed to participate all kinds questions, or place them in situations and see what they
chose – something which is not always possible online.)

We do not have systematic samples of modern visual and media culture. Instead we
have numerous separate collections and archives that are being digitized. Therefore, the
kind of question I asked in 2009 –What did people painted around the world in 1930? – is
still unanswerable. And for many other questions, the situation is even worse. Consider for
example the history of photography. While working on a book about Instagram aesthetics
in the context of modern design, art and photography, I had a pretty big sample of Insta-
gram: 16 million photos shared in 17 global cities between 2012 and 2016.\textsuperscript{12} It is important
to note that these are not photos with particular tags. Instead, they are all geo-coded
photos shared in larger city areas during a particular period. According to a few of computer
science publications that analyzed large samples of Instagram posts in 2014, during that
time Instagram users shared locations for 20\% of their photos.\textsuperscript{13} This means that our data-
sets also represent approximately 20\% of all Instagram photos shared in a given area and
period. From a sampling point of view, these are very good samples. Not only are they quite
substantial but we also know what part of a “population” is represented. (“Population” in
statistics is a technical term that refers to the whole data that for practical reasons is not
accessible to us. Instead, we can use small samples from which we can probabilistically infer
characteristics of the whole data.)

I certainly did not expect to find anything like these samples for vernacular photo-
graphy in the 20\textsuperscript{th} century. But I assumed that after all digitization work of the last twenty
years, I can easily find samples of at least few thousand digitized photographs for particular
decades, and maybe even for particular countries. It turns out that nothing like this existed.

What has been digitized and made available online are various collections of vernacular
photography from particular private collections. They added certain photos to their collec-

\textsuperscript{12} Manovich 2016.
\textsuperscript{13} Manikonda et al. 2014.
tions because each photo was interesting to them for some reason. Museum exhibitions of vernacular photography that I consulted were similarly “non-objective” – they were assembled by curators who had particular curatorial ideas. I also found some user groups on Flickr with “found photographs” contributed by group members. Every collection I consulted was the result of individual or groups’ taste and ideas of what should be included. Often people were only interested in more “artistic” and “avant-garde” examples of vernacular photography, rather than the typical.

To my knowledge, nobody has ever thought to create a representative sample that would contain characteristics of the field of vernacular photography as a whole in particular historical periods, types of cameras and printing, and so on (for example, photos made with Kodak Brownie cameras of 1900, or first portable 35-mm Leicas in 1925, or prints using Kodacolor after 1942, or Polaroid prints after 1972.) So now that we have learned from computer science studies of massive social media samples that we can look at any culture as a statistical population asking about distributions, averages, variance, clusters, and so on, we want similar historical samples. But they do not exist.

For example, the National Gallery of Art in Washington presented an exhibition in 2010 called The Art of American Snapshot, 1888–1978: From the collection of Robert E. Jackson. According to the curators, “Organized chronologically, the exhibition focuses on the changes in culture and technology that enabled and determined the look of snapshots. It examines the influence of popular imagery, as well as the use of recurring poses, viewpoints, framing, camera tricks, and subject matter, noting how they shift over time.”

The online exhibition catalog shows that curators did an excellent job of capturing a number of aspects of vernacular photography and its evolution. However, since the exhibition only had 200 photographs for a 90-year period, that meant that the historical map exhibition constructs was very “low resolution” (to use the spatial metaphor) and also not complete. If we want to understand differences in snapshot photography between different countries, or find gradual changes in style or subjects that are not related only to the introduction of new photography technologies, or see if there may be some regional or demographic differences, we cannot accomplish this with 200 photos.

For a comparison, consider the Gallup U.S. Daily poll. For this poll, Gallup interviews (over the phone) 500 people across U.S. every day. For a country of 300 million people, this looks like a tiny sample. But because Gallup selects people at random and conducts these interviews every day, it accumulates 15,000 responses per month, and 175,000 per year. We also learn that “Gallup also weights its final samples to match the U.S. population

according to gender, age, race, Hispanic ethnicity, education, region, population density, and phone status. This weighting is done using data from a number of other surveys. For example, to weight by population density, Gallop uses U.S. Census reports. This systematic approach to sampling and analysis of the results is typical of all natural and social sciences, public administration, demographics, public polls, marketing research, and countless other areas. In fact, the only area where it is absent is humanities.

The question humanists have been asking is about canon, and how to make canons in their field more representative. There is a parallel here with the kind of weighting Gallup and other organizations that collect demographic data do. However, sometimes in the attempts to compensate for a lack of representation of older canons, the new canons are “weighted” more towards groups that were previously not represented. So as a result, we once again get something completely driven by ideologies, rather than a balanced sample.

A “balanced cultural sample” can be defined in multiple ways, all equally informative and complementary to each other. For example, we can include a proportion of all works produced in particular media, period, and place. Or we can focus instead not on what has been produced, but what audiences actually read, watched, or listened to. We may decide to select only works that achieved certain recognition (which would be equivalent of likes and favorites in contemporary social media), or disregard this information. But whatever we do, we need a systematic procedure, not simply a taste judgment. Statistics has developed a sophisticated theory of sampling which includes many methods, and since these methods are used today in all sciences, they should be adopted for analysis of historical cultural artifacts as well – if we are interested in understanding them as a kind of ecological or geological system, where all participants and artifacts are important – as opposed to only a set of “masterpieces.”

The idea of creating systematic and representative samples of culture is interesting by itself, because it leads to all kinds of follow up questions. And since our textbooks, museums, cultural portals, classes, and documentaries always represent human arts and cultures using only selected examples, the questions about cultural sampling are important in general, even if we are not conducting quantitative analysis. They relate to how we understand, represent and teach human cultural history – and also how we think about our cultural present, with its new scale of numbers of participants, their cultural interactions and experiences.

For example, imagine a hypothetical scenario where we can include any painting created in France in the 19th century in our sample. Now imagine that we want to create a representative sample, so we randomly select X number of paintings. Such a sample will include several academic salon paintings, realistic paintings, portraits and so on. And it would miss the 19th century art which we now recognize as most important – works by Impressionists and Post-impressionists. Why? It has been estimated that 13 key French Impressionists artists together created 13,000 paintings and pastels during their lifetimes. But this is a
very small number in comparison to all paintings created by artists living in France during the whole 19th century. So, a random sample would likely miss them all.

This is exactly the same problem, which accompanies a great deal of quantitative social media research in Computer Science. In many articles, authors explain how they carefully construct a random sample drawn from all users of Pinterest, Instagram or Twitter. Using such samples, they then develop statistical models that account for some characteristics of the behavior and posts of these users. This research is very interesting and important. But using a single global sample of a network with hundreds of millions of people from most countries in the world sharing billions of daily text posts, images and video has serious limitations. We can only see the “typical.” So we miss all kinds of regional variations, and presence and activity of endless users who don’t have the typical behaviors and posts. In other words, if any of these networks have their own “Impressionists,” they are not visible in the analysis that uses single random samples.

Sometimes, the sampling procedures used end up only including particular types of users. For example, in the paper “Analyzing User Activities, Demographics, Social Network Structure and User-Generated Content on Instagram” (2014), the researchers state: “To the best of our knowledge, we believe this is the first paper to conduct an extensive and deep analysis of Instagram’s social network, user activities, demographics, and the content posted by users on Instagram.”18 This is how they describe the method they used to create a user sample for their study:

First, we retrieved the unique IDs of users who had pictures that appeared on Instagram’s public timeline by using Instagam API, which displays a subset of Instagram media that was most popular at the moment. This process resulted in a sample of unique users. However, after careful examination of each user in this sample, we found that these users were mostly celebrities (which explains why their posts were so popular). To avoid the sampling bias, for each user in this sample, we crawled the IDs of both their followers and friends, and later merged two lists to form one unified seed user list which contained 1 million unique users.

The final dataset has 5,659,795 images for 369,828 users (the rest had private accounts). Out of these images, 1,064,041 have geo-locations. But how well are these users representing the Instagram universe? Most people follow other people as opposed to celebrities. People who do follow celebrities and their friends are likely only one type of Instagram user. Additionally, given that the number of Instagram users in every country differs, with the biggest countries also often having larger number of users, such a “random” sample likely better represents some countries than others.

These considerations do not invalidate the results in this and all other papers that use a single large sample from massive global social networks. Their findings are valid. They just may not apply to every type of user or type of post on such networks. (Note that we are not talking about individual users but groupings, each with their own characteristics. In other words, these are like 19th century Impressionists who had common characteristics.)

18 Manikonda et al. 2014.
We also need to recall here perhaps the most fundamental “Achilles’ heel” of statistics. “The goal of statistics is to represent the facts in the most condensed way” (1833). But we pay a big price for such compression. The measures used in descriptive statistics summarize some population (i.e., a set of items) but they may not correspond to any concrete members of this population. For an example, let’s take a series of numbers: 1,1,2,3,2,9,9,10,11,11,11. The average (called “mean” in statistics) of this series is 6.36. But we don’t have any actual numbers close to this mean! No. 4, 5, or 6. Instead, we have two “clusters”: 1 to 3, and another one from 9 to 11. (This is called a bimodal distribution.)

In other words, the standard statistical measures of a large population can easily miss the presence of various groupings in this population. So, if we represent some “cultural population” – be it 19th century paintings or 20th century cinema, Instagram today, or global music videos – with a single random sample, we can miss all kinds of groupings (1960s New Wave or 1920s Soviet Montage school in cinema history; contemporary music videos from India, Korea, Vietnam, Thailand or Kazakhstan which have their own differences despite overall similarity; and so on.) And the characteristics which we will find may describe the “average” which never existed in reality. That is, it may not correspond to any actual group. And rather than capturing the presence of multiple distinct groups, it can hide them from view.

In fact, I would like to claim that in human societies and cultures there are no “averages.” Certainly, we can follow Adolphe Quetelet who in the early 1830s was the first to start to measure the physical characteristics of humans such as height and weight and found that their distributions followed “normal” curves.19 If we perform such measurements today, we will find similar distributions. And, in a sample of a million people, certainly many would have the exact height specified by the mean. In the same way, if we for example measure the length of tens of thousands of modern novels, we will find that some do have exactly the same length as the average novel.

But such results only hold if we limit the study of cultural artifacts, interactions, and experiences to one characteristic at a time. If we look at several selfies sampled from Instagram, we can calculate the average degree of smile, size of a face in a photo, and its position. And if the sample size is big enough, some actual selfies will have exactly the same numbers as the averages. But just as a face of every person is unique, like their fingerprints, their photos are also unique. So if we multiply the number of characteristics, eventually we will not find any real selfie that matches the sample averages on all of them. The same applies to any other type of cultural expression, past or present.

There is one field that does think about cultural sampling and it is using statistical methods to create and analyze these samples. This field is the sociology of culture. The most well-known book in this field remains famous Distinction: A Social Critique of the Judgement of Taste by French sociologist Pierre Bourdieu. Published in 1979, it has been recognized as one of the ten most important books of sociology in the 20th century. Bourdieu offered

19 Tyler 1872.
powerful intellectual ideas and theories that connected people’s cultural tastes and their socio-economic statuses. These theories were grounded in the statistical analysis of two large surveys of tastes of the French public conducted in the 1960s. Bourdieu collaborated with French “data scientists” (to use contemporary term) who developed a new analytical and visualization methods to represent relations between many elements, and he used this method in all of his later studies including Distinction.

Today sociologists of culture continue to use surveys of groups of people, but they also use samples from cultural publications. One example of the former is a study where the researchers “asked 1544 German-speaking research participants to list adjectives that they use to label aesthetic dimensions of literature in general and of individual literary forms and genres in particular (novels, short stories, poems, plays, comedies).” The example of the later is a study called “Institutional Recognition in the Transnational Literary Field, 1955–2005.” It uses “a sample of articles from 1955, 1975, 1995 and 2005 in French, German, Dutch and US elite papers (N=2,419).” Here is another example: an analysis of fashion discourse during 1949–2010 that uses 1301 fashion reviews from The New York Times and The International Herald Tribune. Although such samples are rather small in comparisons to social media scale, they are sufficient to answer particular questions the researchers asked in these studies.

When I first thought of cultural analytics in 2005, I imagined being able to construct detailed world-wide maps of particular fields – such as painting, cinema, graphic design or music video – for long historical periods. But as I realized that digitization efforts are not creating systematic samples such maps would require, I had to abandon these ideas for the time being. So instead, I focused on a different type of sampling that I could do given what has been digitized – by type of media. Starting in 2008, in our lab, we have worked on over 40 datasets that cover almost every major type of visual media today. We analyzed comics and Manga series, video games, feature films, documentaries, motion graphics, music video, political video ads, print magazines, historical photographs, born-digital photographs and other images, and interactive virtual worlds. We also deliberately included dataset that lie at the extremes of a high – low and professional – non-professional dimensions: from paintings of van Gogh, Mondrian and Rothko to 10 million Instagram photos shared in New York City by 5 million people. And we have also deliberately balanced Western and non-Western cultural sources. The latter include Japanese video games, music videos from across Korea, Instagram photos shared in seventeen global cities that cover four continents. We published analysis using Instagram photos shared in Tel Aviv, Israel during Fallen Soldiers and Victims of Terrorism Remembrance Day, and another analysis of Instagram photos shared during February 2014 Maidan revolution in Kiev, Ukraine.

20 Knoopa et al. 2016.
21 Verboord et al. 2015.
22 Van de Peer 2014.
In fact, the advantage of using social media data is that it is not “canonical” or “national.” Popular networks such as Facebook, Instagram, and others are used in every country except the few where they are/were blocked for periods of time (In the case of Facebook, Bangladesh, China, Iran, North Korea, Syria\textsuperscript{23}). As of May 2016, the messaging app WhatsApp that started in China was used in 109 countries, with one billion users sending 42 billion messages daily.\textsuperscript{24} And by the same time, 80% of Instagram 500M active users were outside U.S.\textsuperscript{25}

For example, when we were creating our Instagram samples datasets between 2012 and 2016, Instagram API allowed anyone to download all geo-tagged photos shared within a particular rectangular area defined by its latitude and longitude. Each area could be 5km x 5km in size, and collecting from a number of areas was not more complicated. So it was equally easy to download images from parts of Manhattan, or Moscow, or Bangkok, or Kiev, and so on. (To download all geotagged images shared during five months in Manhattan, we combined a number of areas to enclose the island in a large rectangle, and then filtered out the data outside of Manhattan boundaries).

This means that in practice, comparing many areas from around the world is as easy as comparing nearby areas from the same city – as long people share sufficient amounts of social media in these global areas. The global perspective is “built in” in social media. This of course also applies to the standard formats, constraints and affordances particular networks and apps provide for their users. Everyone who used Twitter between 2007 and 2017 had to fit their messages into the same 140 characters. Everyone who was using Instagram between 2010 and 2015 had to submit to its square image format and the same size: 640 x 640 (or 612 x 612). Everyone has access to exactly the same functions (adding hashtags, optional geo-tagging, etc.) and the same UI. This by itself raises an important question: does social media software lead to less diversity in user-generated content? This was one of the key questions for me during my eight years of research.

Data Representation

However, like every other type of data about society, social media data has its own limitations, and they are not insignificant. I will briefly discuss five issues which are all about representation – what gets represented (and available for research) and what is absent. While the use of social networks and the web continues to grow around the world, billions of people do not use them. Here is a concrete example from our own research of how this situation limits what we can “see” using their data. In 2014, Twitter agreed to provide selected researchers with access to any part of their data if they used it in new interesting ways. Thirteen hundred labs from around the world applied, and we were one of six labs.

\textsuperscript{23} Kirkland 2014.
\textsuperscript{24} Smith 2016.
\textsuperscript{25} Facebook 2016.
that won. I asked Twitter to give us all tweets with geo-located images shared with them. Twitter added images functionality in 2011, and we were given access to all tweets with geo-located images shared worldwide between 2011 and 2014. When we plotted locations of a random sample of 100 million tweets from this data approximately half of the populated Earth surface had no coverage.

The second issue has to do with demographics of users who do use social networks. In “developed” countries and global megacities, people from all demographic groups use the networks. In a country like the USA, there is no significant differences in social network use between women and men, or different races, or people with different level of education – but there are still big differences between age groups. This is also true globally – although the differences are getting smaller with time. A report on social media use among people who were online in 34 countries in first quarter in 2016 found that 92% of those who are in 45–54 age group have social media accounts; for people in 55–65 age group the figure is 82%.26

In many developing countries, the proportions of people using social networks among those using the web are higher than in developed countries. At first, this looks like good news because it could mean that we get data on cultural activities of larger proportion of populations in these countries. However, the reality is different. As the report explains, “As many as 98% of Internet users in countries like Malaysia, Brazil, Indonesia and Vietnam are on at least one network. In part, that’s a result of their lower Internet penetration levels, which means online adults in these regions are more likely than their counterparts in Europe or North America to come from young, urban and relatively affluent segments.27

The third issue is uneven spatial distribution of social networks activity and content even in big urban areas where we see very high use – until we zoom in. The amount of sharing and participation can vary dramatically between city areas, as we show in the Inequaligram project. We collected and analyzed 7,442,454 public geo-tagged Instagram images shared in Manhattan over five months. The inequality we found between the more populated and less populated parts of Manhattan was staggering. We found that the ratio between a square km area with most images and the area with least images was 250,000:1. According to our analysis, 50% of all images shared by local residents are within only 21% of Manhattan area. For visitors, this difference is almost twice as big: 50% of their images were shared in only 12% of the Manhattan area. In summary, even for such a densely populated urban area as Manhattan, its Instagram collective image only reflects part of it and not all.

The forth issue is what content people share, what comments they make, and what they are willing to say online. Social networks are not a mirror of society. Just as people in other areas of their lives play roles, follow norms, present particular identities and behave in ways expected from them (by “mainstream,” or their particular “subcultures,” or “tribes), they do this online. And because their posts and comments can be seen by all other network

26 GWI 2016.
27 Ibid.
users (unless they make posts or the whole account private), appear in Google search, and are saved by the networks, shared with marketers, etc., they are likely to be extra-careful. And, just as with professional cultural products, some of user-generated content is driven by conventions, stereotypes and models people see around them. For example, we find endless photos in “table top” genre on Instagram created by regular users, overwhelming proportions of selfies smile (see our selfiecity.net and selfiecity.net/London for more details), and travel photos follow their own conventions. All this means that the “culture” we can analyze using social media is its own universe, and not a simple sample of people’s cultural activities, taste and opinions outside the networks.

Finally, the fifth issue is access to social media data. In the middle part of 2000, all large social networks created APIs that allow people to freely download large data samples containing user posts and all public information about them visible online – date and time a post was shared, location (if user shared this information), username, tags, comments, and numbers of likes and re-shares. In the case of visual networks such as Instagram and Flickr, image and video along with their user descriptions and all other information was also available for downloads. Flickr launched its API in 2004, and Facebook and Twitter in 2006.28

While these APIs were intended for developers building apps that use data from the platforms, and for users to share contents between networks and also their blogs, computer science researchers, data visualization artists, and other creative technologists realized that they can also freely access this data, and numerous studies and projects were created. Hundreds of thousands of computer and social scientists and students used these APIs to download data, analyze it and publish papers.

However, there have always been limits on how much data can be downloaded. For example, during the period we were actively downloading Instagram data (2012–2016), it had a limit of 3000 images per hour, and only images from the last few days were available. Nevertheless, we were able to assemble 16 million Instagram photos shared in 17 global cities in different periods between 2012 and 2016. But given that in 2016 people are sharing 80 millions of images on Instagram per day, what we were able to assemble was a tiny portion.

However, because of the concerns with privacy and unauthorized use of posts, some of the biggest networks gradually limited or closed API access to bulk user data. Facebook limited the use of its API on April 30, 2015, and Instagram stopped allowing bulk downloads on June 1, 2016. At this moment (end of 2016), Twitter is still accessible, along with some networks popular in particular geographic areas such as Russian VK.

In summary, we know that social media and the web are not used by everyone; the proportions and demographics of those who use social media varies from place to place; and what people publish and share constitutes its own cultural reality as opposed to being a transparent window into the realities outside. We should always keep these limitations in mind. At the same time, using the web and social media data and contemporary

28 Lane 2012.
technologies for tracking and analyzing it questions the very idea of representation. This concerns the very foundation of modern research methods based on sampling.

These methods assume that for practical reasons we cannot have access to the complete “population” (i.e., full data). We can only access and analyze one or more samples of the population. Accordingly, modern statistics is divided into two areas. Inferential statistics is a set of methods for estimating characteristics of the population based on its sample(s). Descriptive statistics only describes the properties of whatever data we have, and it does not assume that this data came from a larger population.

However, when we analyze web and social media content and interactions, we often can have full data. Certainly, the companies that run social networks, media sharing sites or publishing platforms can record all interactions happening on their platforms. This is true for Facebook, YouTube, Twitter, Pinterest, Spotify, Amazon, Scribd, Shutterstock, Behance, academia.edu, and other social media and publication services. This does not mean that a company will be analyzing all their data, or keeping it forever, or even have its own researchers work on it – because companies don’t want to sued, have bad publicity or get in trouble with governments. So the data is anonymized, sampled when needed, and only particular parts of the data are made available to internal researchers depending on what lab they work for. However, the largest companies certainly take advantage of having massive data about user interactions on their platforms, using it to train systems that recommend other users to follow or other videos to watch and decide which posts from friends to show, select trending topics etc. Big data is also driving the main source of income for big social media companies – i.e. automatic advertising systems such as Google AdWords and Facebook Ads.

Although academic researchers do not have direct access to complete data from these companies, it is possible to use their APIs to download complete data that satisfies particular criteria, such as all activity on a particular platform within a particular time period. Many papers use such datasets. In our own work, we also followed this approach. We were using Instagram API to download all publically shared geo-coded images shared in a particular geographic area over a period of time. In fact, every Instagram dataset we used was generated in this way. For example, to create a dataset of 7,442,454 public Instagram images shared in Manhattan over five months, we used a single Mac to run our custom download program 24/7 during this whole period. As far as we know, the images we downloaded are all images people shared within this area and time with geo-location (which constitutes approximately 20% of everything shared).

Why may we want to use complete cultural data? If we are only interested in extracting general patterns, characteristics, and types — for example, the 10 most common types of images on Instagram — we certainly do not need all of the data. But such summarization and aggregation common to the use of statistical methods in 19th and 20th century is only one way to use cultural data. As I explained above, using small samples from diverse cultural “population” (such as trillions of Instagram images) may only reveal the “typical” and “most popular” and miss “regional variations” and “presence and activity of endless users who do not have the typical behaviors and posts.” Therefore, ideally Cultural Analytics should try
to obtain and analyze complete data generated by some cultural process (be it career of a single photographer or all photos shared on Instagram).

Rather than only treating culture as “data points” that together create patterns that we want to discover, disregarding the individual points afterwards, Cultural Analytics should pay equal attention to both patterns and individual artifacts, experiences and interactions. As creators and audience members, we engage and enjoy concrete artifacts and experiences, and not “patterns.” A particularly successful artifact is often described as “unique” – i.e. it cannot be reduced to already existing patterns. As aesthetic subjects, we search and enjoy such uniqueness. One of the goals of Cultural Analytics is to help us find truly unique artifacts in the infinite universes of media now being created. And even if other artifacts are not unique in most ways, they may still have something unique in other ways, which can get lost if we reduce them to patterns. For instance, every human face is unique, and therefore even the most conventionally-driven photo of this face will be special for us. (In this aspect, Cultural Analytics should combine special perspective of sciences and of humanities – the former’s concern with general laws and regularities, and the latter’s concern with unique cultural objects.)

To conclude, I would like to note one techno-cultural development of the last 20 years that connects many issues I have discussed – the rise of search as a new dominant mode for interacting with information. This development is just one of many consequences of the dramatic and rapid expansion of information and content being produced which we have experienced since the middle of the 1990s. To serve the search results, Google, Bing, Baidu, Yandex, and other search engines analyze many different types of data – including both metadata of particular web pages (so-called “meta elements”) and their content. For example, according to Google, its search engine algorithm uses more than 200 input types.29

However, Google, Yandex or Bing do not reveal the measurements of web pages they analyze – they only serve their conclusions, i.e. which sites best fits the search string user entered determined by their propriety algorithms that combine these measures. In contrast, the goal of Cultural Analytics is to enable what we may call “deep cultural search” – give users the open-source tools so they themselves can analyze any type of cultural content in detail and use the results of this analysis in new ways.

References

Lev Manovich


Knoopa, Christine A., Valentin Wagnera, Thomas Jacobsen, Winfried Menninghaus: “Mapping the aesthetic space of literature ‘from below’.” In: Poetics 56 (June), pp. 35–49.


Abstract  Contemporary and future historians need to grapple with and confront the challenges posed by web archives. These large collections of material, accessed either through the Internet Archive’s Wayback Machine or through other computational methods, represent both a challenge and an opportunity to historians. Through these collections, we have the potential to access the voices of millions of non-elite individuals (recognizing of course the cleavages in both Web access as well as method of access). To put this in perspective, the Old Bailey Online currently describes its monumental holdings of 197,745 trials between 1674 and 1913 as the “largest body of texts detailing the lives of non-elite people ever published.” GeoCities.com, a platform for everyday web publishing in the mid-to-late 1990s and early 2000s, amounted to over thirty-eight million individual webpages. Historians will have access, in some form, to millions of pages: written by everyday people of various classes, genders, ethnicities, and ages. While the Web was not a perfect democracy by any means – it was and is unevenly accessed across each of those categories – this still represents a massive collection of non-elite speech.

Yet a figure like thirty-eight million webpages is both a blessing and a curse. We cannot read every website, and must instead rely upon discovery tools to find the information that we need. Yet these tools largely do not exist for web archives, or are in a very early state of development: what will they look like? What information do historians want to access? We cannot simply map over web tools optimized for discovering current information through online searches or metadata analysis. We need to find information that mattered at the time, to diverse and very large communities. Furthermore, web pages cannot be viewed in isolation, outside of the networks that they inhabited. In theory, amongst corpuses of millions of pages, researchers can find whatever they want
The Promise and Pitfalls of Web Archives

to confirm. The trick is situating it into a larger social and cultural context: is it representative? Unique?

In this paper, “Lost in the Infinite Archive,” I explore what the future of digital methods for historians will be when they need to explore web archives. Historical research of periods beginning in the mid-1990s will need to use web archives, and right now we are not ready. This article draws on first-hand research with the Internet Archive and Archive-It web archiving teams. It draws upon three exhaustive datasets: the large Web ARChive (WARC) files that make up Wide Web Scrapes of the Web; the metadata-intensive WAT files that provide networked contextual information; and the lifted-straight-from-the-web guerilla archives generated by groups like Archive Team. Through these case studies, we can see – hands-on – what richness and potentials lie in these new cultural records, and what approaches we may need to adopt. It helps underscore the need to have humanists involved at this early, crucial stage.

Keywords: archive; world wide web; historical studies; webscraping; digital history

The Web is having a dramatic impact on how we research and understand the recent past. Historians, who have long laboured under conditions of source scarcity – we wish we had more information about the past, but it was not recorded or preserved – are now confronted with primary sources on a scale that defies both conventional methodologies and standard computational methods.1 Web archives offer profound promise. Take a comparative example. The Old Bailey Online describes its holdings of 197,745 trials between 1674 and 1913 as the ‘largest body of texts detailing the lives of non-elite people ever published’.2 The web archive of GeoCities, a platform for web publishing that operated from the mid-1990s to the early 2000s, amounts to over 38 million pages. Eventually, historians will have access to billions of such sources written by people of various classes, genders, ethnicities, and ages. While the World Wide Web is not a perfect democracy, by any means and any of the categories listed above, it still represents a massive shift. As a result, web archives exemplify this conundrum and represent challenge as well as opportunity.

What information do we want to access? How was the information collected? How do national boundaries intersect with the realm of the Internet? What are the implications of working with such large archives, collected without the informed consent or even knowledge of the overwhelming majority of contributors? These are pressing concerns. For the most part, historians cannot write histories of the 1990s unless they use web archives: with them, military historians will have access to the voices of rank-and-file soldiers on discussion boards; political historians, to blogs, the cut and thrust of websites, electoral commentary and
beyond; and of course, social and cultural historians, to the voices of the people on a scale never before possible.

The stakes are high. If we do not come to grasps with web archives, the histories that we write will be fundamentally flawed. Imagine a history of the late 1990s or early 2000s that draws primarily on print newspapers, ignoring the revolution in communications technology that fundamentally affected how people share, interact, and leave historical traces behind. Yet even as we use web archives, we need to be cognizant of their functionalities, strengths, and weaknesses: we need to begin to theorize and educate ourselves about them, just as historians have been cognizant of analog archives since the cultural turn. As new discovery methods for finding information in web archives begin to appear, historians need to be ready to participate; otherwise we might not know why one particular response is number one, versus number one million.

The sheer amount of social, cultural, and political information generated and presented almost every day within the web archive since the Internet Archive began collecting in 1996 represents a complex data set that will fundamentally reshape the historical profession. We need to be ready.

ON COMPLEX DATA SETS: THREE DIFFERENT EXAMPLES

This is not an abstract concern: the history of the 1990s will be written soon. While there is no common rule for when a topic becomes ‘history,’ it took less than 30 years after the tumultuous year of 1968 for a varied, developed, and contentious North American historiography to appear on the topic of life in the 1960s.3 Carrying out ‘recent histories,’ be they of the 1970s or of events only a few years ago, brings with them a host of methodological issues from a lack of historiography, historical participants who can ‘talk back,’ and issues of copyright and privacy.4 The year 2021 will mark the 30th anniversary of the creation of the first publicly accessible website. Just as media, government, and business radically transformed their practices in the 1990s, historians must do so as well to analyze this information. ‘New media’ is not that new anymore.

Historians run very real risks if they are not prepared. Currently, the main way to access the archived Web is through the Wayback Machine, most notably associated with the Internet Archive. The Internet Archive emerged out of a concern around a ‘digital dark age’ in the mid-1990s, where rapid technological evolution led to fears around whether our heritage was being preserved. Responding to this, Internet entrepreneur Brewster Kahle founded the Internet Archive in June 1996, which began to rapidly grow their web archive collection. They did so by sending ‘web crawlers,’ automated software programs, out into the Web to download webpages that they found. This crawling process meant that depending on how the Web developed and the limits placed on a crawler, the crawler could indefinitely collect – generating an infinite archive.5
While the Internet Archive was collecting data from 1996 onwards, the next step was to make it accessible to researchers. In 2001, they launched the still-dominant form of interacting with web archives: the Wayback Machine. You can try it yourself at http://archive.org/web. It is limited. A user needs to know the exact Uniform Resource Locator (URL) that they are looking for: a website like http://www.geocities.com/enchantedforest/1008/index.html, for example. The page is then retrieved from the web archive and displayed. If you know the URL of the page you are interested in, and only want to read a few, the Wayback Machine works by generating facsimiles of those pages. They are not perfect, as they may not collect embedded images, or might grab them at slightly different times (to avoid overloading any single server, the crawler might download the text of a website and then the image a few hours or even days later; this can lead to the storing of websites that never existed in the first place). Beyond technical issues, it is difficult to find documents with the Wayback Machine unless you know the URL that you want to view.

This latter shortcoming disqualifies it as a serious research tool unless it is paired with a search engine of some kind. Historians are used to full-text search interfaces. However, imagine conducting research through date-ordered keyword search results, carried out on billions of sites. It would produce an outcome similar to the current methods by which historians search digitized newspapers. In the absence of contextual information about the results found, they can be useless. It is possible to find almost anything you want within 38 million web pages. I can find evidence on any matter of topics that advances one particular argument or interpretation. Without the contextual information provided by the archive itself, we can be misled.

Three case studies can help us better understand the questions, possibilities, and challenges facing historians as we enter this archival territory. The first is the Wide Web Scrape, a compilation of billions of objects collected by the Internet Archive between 9 March and 23 December 2011. Next, I explore work that I have been doing with a collection of political websites created between 2005 and 2015. Finally, I explore the GeoCities end-of-life torrent, to get at the heart of ethical challenges.

Together, these studies suggest a path forward for historians. Those of us who use web archives do not need to become programmers, but do need to become aware of basic Web concepts: an understanding of what metadata is, how the Web works, what a hyperlink is, and basic definitional concepts such as URLs. Beyond this, however, is the crucial dimension of algorithmic awareness. When we query archives, we need to know why some results are coming to the top and others at the bottom. If we turn our research over to black boxes, the results that come from them can reaffirm biases: websites belonging to the powerful, for example, rather than the marginalized voices we might want to explore and
consider. The decisions that we as historians make now will have profound effects as tools begin to be developed to access web archives.

**DATA IS BIGGER THAN THE NATION: THE WIDE WEB SCRAPE**

As a data set, the Wide Web Scrape is exhaustive, transcending national borders. The 2,713,676,341 item captures—websites, images, PDFs, Microsoft Word documents, and so forth—are stored across 85,570 WebARChive (WARC) files. The WARC file format, which is certified by the International Standards Organization, preserves web-archived information in a concatenated form. Generated by the Internet Archive, these files also serve as a good introduction to the geographic challenges of web archives: historians tend towards geographic boundaries, but these archives can transcend them. WARC files are an abundant resource, but that abundance is double edged.

As a Canadian historian looking for a relatively circumscribed corpus, I decided to focus on the Canadian Web, or websphere, as best I could. The ‘Canadian Web’, is however, intrinsically a misnomer. The Web does not work within national boundaries. It is a global network, transcending traditional geopolitical barriers (local fissures still appear, as seen in ‘this video is not available in your country’ messages). The Internet Archive exploits the Web’s borderless nature in their global crawling of material in a way national domain crawls by national institutions cannot. From Denmark to Britain, researchers collecting and studying national webspheres have taken different approaches. Some, such as the Danish NetLab, have confined their studies to national top-level domains (.dk). Others, such as the British Library’s born-digital legal deposit scheme, use algorithms and human intervention to find British sites outside of the .uk domain.

What does the data collected along the lines of a national websphere—a top-level domain such as .ca—look like? While all archival records are only as useful as the discovery tools that accompany them—a misfiled box in a conventional archive might as well not exist—the size of these collections elude traditional curation. From the holdings of the Wide Web Scrape, we examined the CDX files (akin to archival finding aids which contain information about the records found within archival boxes), and which can be measured in gigabytes rather than terabytes. They contain millions of lines of text like:

```
```
From this, we can learn a few things: in this case, we learn that the record is justlabour.yorku.ca, collected on 14 July 2011 at 7:37 GMT. It redirected (HTML code 302) to the table of contents for volume 16. If you visit justlabour.yorku.ca today, you’ll be redirected to a more recent issue. CDX files help us find specific records. Accordingly, I used them to download a sample of 622,365.ca URLs.

Working with this data set was an interesting window into the choices historians need to make when they work with large data sets from the Web. Derived data—plain text, named entities (discussed later), extracted links, hyperlinks with anchor text—can be useful. Yet at every stage they present historians with questions. Some extracted hyperlinks will be relative—that is, /destination.html rather than http://www.history.ca/destination.html. Should they be reclassified if we want to make a chart of all the hyperlinks connecting different websites, and at what stage? To create plain text files, we use the warcbase platform. Working with this data set was an interesting window into the choices historians need to make when they work with large data sets from the Web. Derived data—plain text, named entities (discussed later), extracted links, hyperlinks with anchor text—can be useful. Yet at every stage they present historians with questions. Some extracted hyperlinks will be relative—that is, /destination.html rather than http://www.history.ca/destination.html. Should they be reclassified if we want to make a chart of all the hyperlinks connecting different websites, and at what stage? To create plain text files, we use the warcbase platform. I was able to run textual analysis, extract location data, postal codes, and names of people, and explore the topics people were discussing. This method had the downside, however, of removing images, backgrounds, and layouts, meaning that text is taken out of context. While the size of the data sets under discussion mitigates this to some extent, we are still profoundly altering sources.

There were three promising ways to query this data, each of which sheds light on various web archival challenges: keywords, named entity recognition (which finds entities like locations and names within text), and hyperlink structures. To search a large body of material with keywords, the Apache Solr search engine is ideal. It can index material and respond to queries from a number of front-ends that can run locally on a computer. The United Kingdom’s Web Archive, for example, uses a custom front-end Solr portal that provides full-text search access to their collections. One view prompts you to enter a query, and to then subsequently see the relative frequency of that term rise and fall over time (how often was the word ‘nationalize’ used in 2006, for example, compared to 2012). With specific queries, this search approach works well. Yet on a broad scale, when looking for cultural trends, more context is necessary.

The most promising keyword approach to my data set was clustering, which takes a set of documents and groups them. If a web collection contained websites about cats, dogs, and pigs, the algorithm might cluster the cat sites together. Conversely, it might find another characteristic—the ages of the authors, perhaps—and cluster them that way. There are several different algorithms to choose from, although in my experience the Lingo clustering algorithm provides the best results (See Fig. 1).

The free Carrot2 front end (http://project.carrot2.org/), which interfaces easily with a Solr database, is the most useful. From a query for ‘children’, we see that this sample of 622,365 websites contains pages relating to child health, health.
centres, service providers, public health, educational services, and consumer products such as Tylenol. Clicking on the graphical representation brings the user to a list of documents, and another click brings up an individual document. The image on the right is the graphical representation of overlapping clusters, such as the simplified Figure 2.

If a dot is connected to two clusters, it belongs to both. These connections can provide a rough sense of how representative things are: there are many websites about breastfeeding, for example, but not many about Christian early childhood education institutions. More importantly, it is possible to isolate a corpus to study. Used jointly, the Solr database and Carrot2 front end help transcend the Wayback Machine’s limitations.

The main drawback with this approach is the need to know what you are looking for. Extracting commonly mentioned locations can be fruitful, as in Figure 3.

Extracted using a combination of Stanford Named Entity Recognition (NER), Google Maps API, and verification by student research assistants, this
process found location names—for example, ‘Toronto’ or ‘Johannesburg’—and geolocated them by assigning coordinates. While longitudinal data will be more useful, allowing us to see how various locations changed over time, at this point we can see the attention paid towards Canadian trading partners and the complete absence of attention towards sub-Saharan Africa. Within Canada, Québec is overrepresented vis-à-vis the province of Ontario.

Web-wide scrapes represent the dream of social history: a massive documentary record of the lives of everyday people, their personal websites, small businesses, labour unions, community groups, and so forth. Yet the value of this information is balanced by the sheer size and complexity of these data sets. Web-wide scrapes are one extreme of what we can do with web archives: exploring a massive record of human activity, collected on a previously unimaginable scale.

**Figure 3.** Countries (other than Canada) mentioned in .ca top-level domain sample (left); Canadian provinces mentioned (right).

**ARCHIVE-IT POLITICAL COLLECTIONS: AN IDEAL SIZE?**

Web-wide scrapes are time consuming and expensive to work with. Recognizing this, web archivists have begun to move towards more accessible data sets that bridge the gap between the lightweight CDX file and the heavy-duty WARC file (both of which we have seen in the preceding section). In this section, I argue that while our first inclination, as with the Wide Web Scrape, might be to go right to the content, more fruitful historical information can be found within the metadata.

Archive-It, a web archiving subscription service provided by the Internet Archive for universities and other institutions, recently piloted their research services portal. It provides access to Web Archive Transformation, or WAT, files: a happy medium between CDXs and WARCs. These provide rich metadata: everything that a CDX has, plus metatext about the website, the title, and the links and anchor text from each site. They are essentially the WARC's sans content, making them much smaller.
Beginning a decade ago, the University of Toronto Library (UTL) has put together thematic web collections with Archive-It. One of their major collections is about Canadian political parties and political interest groups, collected quarterly since 2005. Canada has seen pivotal changes within its political sphere over the last ten years, between 2005 and 2015: an arguable militarization of Canadian society, the transition from the ‘natural governing party’ of the centrist Liberal Party of Canada to the Conservative Party of Canada (and back in late 2015), as well as major policy changes on foreign policy, science policy, and climate change. Given these critical shifts, it is surprising on one level that UTL’s collection was not used more—the collection, for example, has never been cited before we began to work with it. On another level, however, it is unsurprising: the current portal to work with the collection at https://archive-it.org/collections/227 has only a very basic search function. It was only by reaching out to librarians at UTL and the Internet Archive that I was able to get the files and begin to explore what we could actually do with them. Ultimately, it became clear that metadata was just as—and in many cases more—useful than the content itself (we ended up providing access to the content through http://webarchives.ca, an implementation of the British Library’s Shine frontend).

By using either the Internet Archive’s web analysis workshop or warcbase, a web archiving platform, we can extract links the WAT files in this collection by domain. The results look similar to the example in Table 1.

In this case, we can see that among the sub-sites that make up the Conservative Party’s website there are ten links to websites within the liberal.ca domain, and vice versa. This sort of data needs to be used with caution, however: one strategic, high-profile link to a website might have more impact than lots of smaller links. For example, a single link on the front page of a political party’s website has far greater impact than hundreds of links contained in the footers of biographical statements. We call this the ‘weight’ because it dictates how much emphasis should be put on the lines that connect various nodes.

This data can be useful on a large scale. Consider Figure 4, which visualizes the external links stemming from and between the websites of Canada’s three main political parties. Each line, or edge, represents a hyperlink between domains (or nodes).
Above, we can see which pages only link to the left-leaning New Democratic Party (NDP or ndp.ca), those that link only to the centrist Liberals (liberal.ca) in the top, and those that only connect to and from the right-wing Conservative Party at right. In the middle are the websites that either link to all three parties or to just two of the three (to the left and right of the Liberal node, respectively). Even from this graph we can see that while many groups link to only the Liberals and the NDP, or to the Liberals and the Conservatives, few link just to the NDP and the Conservatives.

By taking quarterly slices of the data, we can also use metadata to identify the broad contours of a narrative as in Figure 5.

We can see that several entities link to all three parties, such as the environmentalist davidssuzuki.org or the Assembly of First Nations (afn.ca), and we can also see how all of the organizations linked to each other. The Liberal Party was then in power and was under attack by both the opposition parties. In particular, the left-leaning NDP linked hundreds of times to their ideologically close cousins, the centrist Liberals, as part of their electoral attacks, ignoring the
right-leaning Conservative Party in the process. Link metadata illuminates more than a close reading of an individual website would.

We can also find sections of this collection that link far more to themselves than to other parts. These divisions lend themselves well to specific extraction. Consulting the UTL’s entire collection via WARC files may be too difficult, but link analysis can tell us what to download. One experiment proved interesting. I took the two main political parties, the Liberals and Conservatives, over the period of study and (relying solely on links) found the communities that grew out of their party websites. The results were interesting: liberal.ca was in the same community as interest groups such as the National Association of Women and Law and media organizations such as *Maclean’s* magazine and the Canadian Broadcasting Corporation. Most interestingly, the left-wing New Democratic Party of Canada appeared in the same community. For the Conservatives, they were grouped with many cabinet ministers’ pages, but also with groups such as Consumers First, which fought for price parity between Canada and America.

By extracting some of these pages and topic modeling the results, we can confirm existing narratives and raise new questions. Topic modeling finds
topics’ in text. For example, imagine that I am writing about women in a male-dominated labour movement. When I write about the women, I use words like ‘femininity’, ‘equity’, ‘differential’, and ‘women’. Men: masculinity’, ‘wildcat’, or ‘foremen’. In this thought experiment, imagine I am drawing these words from buckets full of slips of paper. Topic modeling reverses that process, putting those words back into the bucket and telling me what is in it. It is a quick way to get a sense of what might be happening in a large body of text.\textsuperscript{18}

Taking the link community that appeared around political parties, we were able to find topics most closely connected to them. In December 2014, the Liberals were highlighting cuts to social programs, issues of mental health, municipal issues, housing, and their new leader, Justin Trudeau (now, as of October 2015, the new Prime Minister of Canada). The Conservatives: Ukraine, the economy, family and senior issues, and the high-profile stimulus-based Economic Action Plan. For 2006, the results were surprising. The Liberals: community questions, electoral topics (given the federal election), universities, human rights, childcare support, and northern issues. The Conservatives: some education and governance topics, but notably, several relating to Canada’s aboriginal population. While the Liberals had advanced a comprehensive piece of legislation designed to improve the conditions of Canada’s aboriginal population, Conservative interest in the topic was surprising: perhaps it reflects the Conservative opposition to it? As one commenter on an earlier draft suggested, it may represent the influence of key advisors, one of whom was a leading Conservative scholar of native-newcomer relations. Questions are raised, suggesting promise in marrying content and metadata in such a manner.

A PLACE OF THEIR OWN: EXPLORING THE ETHICAL MINEFIELD OF GEOCITIES

In general, the sheer scale of distantly reading millions of websites or exploring the public record of political parties has kept us in the previous cases removed from everyday individuals. As the Web became mainstream in the mid-to-late 1990s, GeoCities played a critical role. For the first time, users could create their own web pages without learning HTML or FTP. On sites like GeoCities, they could become part of part of virtual communities, held together by volunteers, neighbourhood watches, web rings, and guestbooks. Even though in 1999 GeoCities was perhaps the third most popular website in existence, Yahoo! deleted it in 2009. Dedicated teams of Internet archivists, such as Archive Team (http://archiveteam.org), created the web archive that we can use today. It is large: at its peak, GeoCities had around 38 million pages.

GeoCities began in late 1994 as a service predicated on geospatial metaphors and giving voices to those who ‘had not had an equal voice in society’.\textsuperscript{19} Users could easily create new sites within an existing GeoCities community,
such as the Enchanted Forest for children or Area 51 for science fiction fans. They received an ‘address’ based on their neighbourhood: www.geocities.com/EnchantedForest/1005/index.html. In an era when the Web was understood as a new ‘frontier’, this claim to an actual address resonated. User numbers skyrocketed, from 1,400 in July 1995 to 100,000 by August 1996 and a million by October 1997.

I have been exploring the question of how community was created and enacted there. A significant minority of users threw themselves into the site. When a user arrived to create their site, they had to choose where to live: a small ‘cottage’ in the Enchanted Forest, perhaps, or a ‘tent’ in Pentagon. Reminders exhorted them to fit into the site’s theme, reach out to neighbours, and crucially – in a move reminiscent of the American 1862 *Homestead Act* – ‘move in’ and improve their property within a week. Some users became community leaders, welcoming new arrivals and teaching them the ropes. An awards economy boomed, with users creating their own awards and giving them to other sites. They visited each other’s guestbooks. Messages are disproportionately from GeoCities users rather than visitors from outside. This community structure persisted until 1999, when Yahoo! bought GeoCities and turned it into a conventional web host.

Like in the previous section, we can explore neighbourhoods with topic modelling. We can see topics in the Enchanted Forest about parties, friends, soldiers and children’s characters such as Pingu. In Heartland, topics relating to family, church, and genealogy appear, and in the LGBT-focused WestHollywood, the focus is on gender, transgender issues, and fighting against hate crimes. Over time, the topics discussed in some neighbourhoods changed. Pentagon moved beyond being a hub for deployed and constantly moving service people towards serving as a forum for political discussions and military history. Heartland came to advance a vision of family focused on Christianity and genealogy. These findings demonstrate that neighbourhoods both shaped and were shaped by user contributions.

How did this come to be? By extracting links, we can begin to find the central nodes that dozens or even hundreds of other websites linked to, as well as the web of connections that held everybody together. This gives us a few hundred websites per neighbourhood to investigate: the community leaders who received kudos from their members, sites that accumulated awards, those with active guestbooks. These factors produced many hyperlinks, both in and out, making these sites critical nodes.

Websites like GeoCities raise ethical questions. Unlike in our previous case studies, which dealt with institutional websites, in GeoCities we are dealing with largely personal websites from over a decade ago. The majority of these people almost certainly did not create these sites with a future historian in mind, nor are they likely to be aware that their sites live on within the Internet Archive or the Archive Team torrent. They did not give consent to the archiving
of their sites, nor did they have access to a robots.txt file that could have changed access parameters (see http://archive.org/about/exclude.php). Indeed, unless they remember their URL, users cannot see if their sites were archived in order to pursue their removal from the archive. Traditional archival collections often have restrictions: donor requests, privacy legislation, or the protection of personal information on medical, financial, or other grounds. While historians have ethical responsibilities at all times, in many cases the onus of making a collection available and accessible lies with institutions. Oral historians, on the other hand, operate outside traditional institutions, instead working in the personal spaces of their interviewees. Institutional review boards, committees that oversee how human subjects are used in research within most North American contexts, govern their work. While none of the above is simple, it is well-travelled ground. Where do web archives fall between these poles?

Strictly speaking, as we generally treat websites as ‘publications’, it is legal to quote from tweets, blogs, websites, and so forth. Legal does not equal ethical, though. As Aaron Bady notes, ‘The act of linking or quoting someone who does not regard their twitter as public is only ethically fine if we regard the law as trumping the ethics of consent.’23 We need to consider user privacy expectations, which is at the heart of the distinction between a political candidate’s site and a GeoCities homestead. This is not to treat users as dupes but to recognize that somebody posting a website in an obscure corner of GeoCities might have an expectation of privacy: many of these sites would not have been discovered by regular users but are easily discovered by web crawlers methodically crawling a community structure.

We can find guidance from web scholars. danah boyd, a web scholar, notes that students with open Facebook profiles regarded a teacher visiting their page as a breach of privacy, social norms, and etiquette.24 The Association of Internet Researchers provides guidance that has researchers consider the public or private nature of the website and the differences between dealing with sources en masse versus individually.25 Stine Lomberg has emphasized the importance of distance but also, when exploring content, of considering user expectations of privacy.26

Historians need to consider these factors when deciding how to appropriately use this material. Some GeoCities cases bring these questions into perspective. Memorial sites, by people who lost children or other loved ones, are both private and intimate but also have well-travelled guestbooks, often by people who lost loved ones of their own. Other searches bring up pages about suicide or depression. These can only be found thanks to today’s modern discovery tools. If a 15-year old wrote to the government with a rant, privacy legislation would excise her or his name; if you find the rant in GeoCities, the name—or their pseudonym (which can sometimes be connected to real names) – would be there. These are resources that would never make it into a traditional archive.
We have power because we can access the blogs, ruminations, and personal moments of literally millions of people that would never before have been accessed—but we need to use this power responsibly. With the Wayback Machine, the lack of full-text search provides some privacy, but as we undertake more computational inquiries historians can uncover things forgotten since their creation. My own take on this question is twofold, drawing on earlier literature: we need to consider the scale at play. Mining a few thousand sites and dealing with—and writing about—people in aggregate presents few privacy concerns, whereas zooming in on a handful of websites and closely reading them does. A website many other sites connect to, a proud prominent view counter in the corner (or other equivalent markers of popularity that have supplanted this now dated approach), a well-travelled guestbook, signals a website of an owner who wanted to be read and encountered, and who conceived of themselves as part of a broader Web of documents. A smaller website addressed to an internal audience, written by a teenager and full of revealing messages and pictures, is a different thing altogether.

GeoCities represents a new kind of primary source: the largely non-commercialized, unfettered thoughts of millions of everyday people in the mid-to-late 1990s, left for historians today. We can learn invaluable things, from the forms online community took on the Web to the opinions and thoughts on a host of social, political, or cultural issues or topics.

CONCLUSIONS

These three disparate web archiving case studies all demonstrate the critical questions that lie at the heart of these new complex data sets. The technical challenges are clear: not enough processing power or computer memory, the need to find access to a computing cluster, and the variety of file formats and types that underlie them. Rather than a narrow-lens pedagogical approach that stresses say the WARC file, historians who want to use these sources—arguably a necessity when undertaking topics in the 1990s and beyond—need to have a flexible understanding of software and standards.

While this article has focused on the research process, further issues will emerge when scholars attempt to publish this type of work. Images, already a sticking point with many publishers, are borrowed, altered, shared, throughout the Web: can one publish a notable image found in a 1996-era web archive if this has no contactable author or even real name? How can we share our research data with each other if we need to worry about digital rights? How do we balance global copyright regimes with the local contexts of journals and academics? At the least, pedagogical training in copyright is needed, as well as advocacy around orphan works and strengthening fair dealing/use.
Despite these challenges and cautions, which need to be heeded as we move forward, I want to return to the original promise articulated at the beginning of this paper. Each of these case studies, from the Wide Web Scrape to the political movements archive to GeoCities, presents promise. They provide more voices from a more diverse body of people, furthering the goals of social historians to write their histories from the bottom up, to move our stories away from the elites and dominant players of society to the everyday. Web archives are not going to have a slight impact on the practice of history: they are going to force a profound shift. We will have more sources than ever before, by people who never could have conceivably reached large audiences or had their words recorded. We should be optimistic, but we need to be prepared.

END NOTES

3 For examples from the Canadian context, see C. Levitt, Children of privilege: student revolt in the sixties: a study of student movements in Canada, the United States, and West Germany (Toronto, 1984) or D. Owram, Born at the right time: a history of the baby boom generation (Toronto, 1997).

Warbase is documented at https://github.com/lintool/warbase/wiki.


F. Turner, From counterculture to cybertulture: Stewart Brand, the Whole Earth network, and the rise of digital utopianism (Chicago, 2008).


d. boyd, It’s complicated: the social lives of networked teens (New Haven, 2014), 58.


Automation of Sight: 
From Photography to Computer Vision

Lev Manovich

An earlier version of this paper was presented at a conference "Photography and the Photographic: Theories, Practices, Histories," University of California at Riverside, April 1994.

Prologue

Nothing perhaps symbolizes mechanization as dramatically as the first assembly lines installed by Henry Ford in U.S. factories in 1913. It seemed that mechanical modernity was at its peak. Yet, in the same year the Spanish inventor Leonardo Torres y Quevedo had already advocated the industrial use of programmed machines.\(^1\) He pointed out that although automatons existed before, they were never used to perform useful work:

> The ancient automatons...imitate the appearance and movement of living beings, but this has not much practical interest, and what is wanted is a class of apparatus which leaves out the mere visible gestures of man and attempts to accomplish the results which a living person obtains, thus replacing a man by a machine.\(^2\)

With mechanization, work is performed by a human but his or her physical labor is augmented by a machine. Automation takes mechanization one step further: the machine is programmed to replace the functions of human organs of observation, effort, and decision.

Mass automation was made possible by the development of digital computers during World War II and thus became synonymous with computerization. The term "automation" was coined in 1947; and in 1949 Ford began the construction of the first automated factories.

Barely a decade later, automation of imaging and of vision were well under way. By the early 1960s, construction of static and moving two-dimensional and perspectival images, correction of artifacts in photographs, the identification of objects from their images, and many other visual tasks were already handled by computers. A number of new disciplines were emerging as well -- computer image processing, computer vision, computer graphics, computer-aided design.
What these new disciplines had all in common is that they employed perspectival images. In other words, automation of imaging and vision was first of all directed at perspectival sight.

The reasons for this are two-fold. On the one hand, by the time digital computers became available, modern society was already heavily invested in lens-based methods of image gathering (photography, film, television) which all produced perspectival images. Therefore, it is not surprising that it would want first of all to automate various uses of such images in order to obtain a new return from its investment. On the other hand, the automation of perspectival sight has already begun well before this century with the development of perspective machines, descriptive and perspective geometry and, of course, photography. Computers certainly proved to be very fast perspectival machines, but they were hardly the first.

**Perspective, Perspectival Machines, Photography**

From the moment of adaptation of perspective, artists and draftsmen have attempted to aid the laborious manual process of creating perspectival images. Between the sixteenth and the nineteenth century various "perspectival machines" were constructed. They were used to construct particularly challenging perspectival images, to illustrate the principles of perspective, to help students learn how to draw in perspective, to impress artists' clients, or to serve as intellectual toys. Already in the first decades of the sixteenth century, Dürer described a number of such machines. One device is a net in the form of a rectangular grid, stretched between the artist and the subject. Another uses a string representing a line of sight. The string is fixed on one end, while the other end is moved successively to key points on the subject. The point where the string crosses the projection plane, defined by a wooden frame, is recorded by two crossed strings. For each position, a hinged board attached to the frame is moved and the point of intersection is marked on its surface. It is hard to claim
that such a device, which gave rise to many variations, made the creation of perspectival images more efficient, however the images it helped to produce had reassuring mechanical precision. Other major types of perspectival machines that appeared subsequently included the perspectograph, pantograph, physionotrace, and optigraph.

Why manually move the string imitating the ray of light from point to point? Along with perspectival machines a whole range of optical apparatuses was in use, particularly for depicting landscapes and in conducting topographic surveys. They included versions of camera obscura from large tents to smaller, easily transportable boxes. After 1800, the artist's ammunition was strengthened by camera lucida, patented in 1806. Camera lucida utilized a prism with two reflecting surfaces at 135°. The draftsman carefully positioned his eye to see both the image and the drawing surface below and traced the outline of the image with a pencil.

Optical apparatuses came closer than previous perspectival devices to the automation of perspectival imaging. However, the images produced by camera obscura or camera lucida were only ephemeral and considerable effort was still required to fix these images. A draftsman had to meticulously trace the image to transform it into the permanent form of a drawing.

With photography, this time-consuming process was finally eliminated. The process of imaging physical reality, the creation of perspectival representations of real objects was now automated. Not surprisingly, photography was immediately employed in a variety of fields, from aerial photographic surveillance to criminal detection. Whenever the real had to be captured, identified, classified, stored, photography was put to work.

Photography automated one use of perspectival representation -- but not others. According to Bruno Latour, the greatest advantage of perspective over other kinds of representations is that it establishes a "four-lane freeway" between physical reality and its representation. We can combine real and imagined objects in a single geometric model and go back and forth between reality and the model. By the twentieth century, the creation of a geometric model of both existing and imagined reality still remained a time consuming manual process, requiring the techniques of perspectival and analytical
Manovich -- 5

geometry, pencil, ruler, and eraser. Similarly, if one wanted to visualize the model in perspective, hours of drafting were required. And to view the model from another angle, one had to start all over again. The automation of geometrical modeling and display had to wait the arrival of digital computers.

3-D Computer Graphics: Automation of Perspectival Imaging

Digital computers were developed towards the end of World War II. The automation of the process of constructing of perspectival images of both existent and non-existent objects and scenes followed quite soon. By the early 1960s Boeing designers already relied on 3-D computer graphics for the simulation of landings on the runway and of pilot movement in the cockpit.

By automating perspectival imaging, digital computers completed the process which began in the Renaissance. This automation became possible because perspectival drawing has always been a step-by-step procedure, an algorithm involving a series of steps required to project coordinates of points in 3-D space onto a plane. Before computers the steps of the algorithm were executed by human draftsmen and artists. With a computer, these steps can be executed automatically and, therefore, much more efficiently.

The details of the actual perspective-generating algorithm which could be executed by a computer were published in 1963 by Lawrence G. Roberts, then a graduate student at MIT. The perspective-generating algorithm constructs perspectival images in a manner quite similar to traditional perspectival techniques. In fact, Roberts had to refer to German textbooks on perspectival geometry from the early 1800s to get the mathematics of perspective. The algorithm reduces reality to solid objects, and the objects are further reduced to planes defined by straight lines. The coordinates of the endpoint of each line are stored in a computer. Also stored are the parameters of a virtual camera -- the coordinates of a point of view, the direction of sight, and the position of a
projection plane. Given this information, the algorithm generates a perspectival image of an object, point by point.

The subsequent development of computer graphics can be seen as the struggle to automate other operations involved in producing perspectival stills and moving images. The computerization of perspectival construction made possible the automatic generation of a perspectival image of a geometric model as seen from an arbitrary point of view -- a picture of a virtual world recorded by a virtual camera. But, just like with the early perspectival machines described by Dürer, early computer graphics systems did not really save much time over traditional methods. To produce a film of a simulated landing, Boeing had to supplement computer technology with manual labor. As in traditional animation, twenty-four plots were required for each second of film. These plots were computer-generated and consisted of simple lines. Each plot was then hand-colored by an artist. Finished plots were filmed, again manually, on an animation stand.\textsuperscript{11} Gradually, throughout the 1970s and the 1980s, the coloring stage was automated as well. Many algorithms were developed to add the full set of depth cues to a synthetic image -- hidden line and hidden surface removal, shading, texture, atmospheric perspective, shadows, reflections, and so on.\textsuperscript{12}

At the same time, to achieve interactive perspectival display, special hardware was built. Each step in the process of 3-D image synthesis was delegated to a special electronic circuit: a clipping divider, a matrix multiplier, a vector generator. Later on, such circuits became specialized computer chips, connected together to achieve real-time, high resolution, photorealistic 3-D graphics. Silicon Graphics Inc., one of the major manufacturers of computer graphics hardware, labeled such a system "geometry engine."

The term appropriately symbolizes the second stage of the automation of perspectival imaging. At the first stage, the photographic camera, with perspective physically built into its lens, automated the process of creating perspectival images of existing objects. Now, with the perspectival algorithm and other necessary geometric operations embedded in silicon, it become possible to display and interactively manipulate models of non-existent objects as well.
Computer Vision: Automation of Sight

In his papers, published between 1963 and 1965, Roberts formalized the mathematics necessary for generating and modifying perspective views of geometric models on the computer. This, writes William J. Mitchell, was "an event as momentous, in its way, as Brunelleschi's perspective demonstration." However, Roberts developed techniques of 3-D computer graphics having in mind not the automation of perspectival imaging but another, much more daring goal -- "to have the machine recognize and understand photographs of three dimensional objects." Thus, the two fields were born simultaneously: 3-D computer graphics and computer vision, automation of imaging and of sight.

The field of computer vision can be seen as the culmination of at least two centuries-long histories. The first is the history of mechanical devices designed to aid human perception, such as Renaissance perspectival machines. This history reaches its final stage with computer vision, which aims to replace human sight altogether. The second is the history of automata, whose construction was especially popular in the seventeenth and eighteenth centuries. Yet, despite similarity in appearance, there is a fundamental difference between Enlightenment automata which imitated human's or animal's bodily functions and the modern day robots equipped with computer vision systems, artificial legs, arms, etc. As noted by Leonardo Torres, old automata, while successfully copying the appearance and movement of living beings, had no economic value. Indeed, such voice synthesis machines as Wolgang von Kempelen's 1778 device which directly imitated the functioning of the oral cavity or Abbé Mical's Têtes Parlantes (1783) operated by a technician hiding offstage and pressing a key on a keyboard were used only for entertainment. When in 1913 Torres called for automata that would "accomplish the results which a living person obtains, thus replacing a man by a machine" he was expressing a fundamentally new idea of using automata for productive labor. A few years later, the brother of the Czech writer Karel Capek coined the word robot from the Czech word robota, which
means "forced labor." Capek’s play *R.U.R.* (1921) and Fritz Lang’s *Metropolis* (1927) clearly demonstrate this new association of automata with physical industrial labor.

Therefore, it would be erroneous to conclude that, with computer vision, twentieth century technology simply added the sense of sight to eighteenth century mechanical statues. But even to see computer vision as the continuation of Torres’, Capek’s or Lang’s ideas about industrial automation which replaces manual labor would not be fully accurate. The idea of computer vision became possible and the economic means to realize this idea became available only with the shift from industrial to post-industrial society after World War II. The attention turned from the automation of the body to the automation of the mind, from physical to mental labor. This new concern with the automation of mental functions such as vision, hearing, reasoning, problem solving is exemplified by the very names of the two new fields that emerged during the 1950s and 1960s -- artificial intelligence and cognitive psychology. The latter gradually replacing behaviorism, the dominant psychology of the "Fordism" era. The emergence of the field of computer vision is a part of this cognitive revolution, a revolution which was financed by the military escalation of the Cold War. This connection is solidified in the very term "artificial intelligence" which may refer simultaneously to two meanings of "intelligence": reason, the ability to learn or understand, and information concerning an enemy or a possible enemy or an area. Artificial intelligence: artificial reason to analyze collected information, collected intelligence.

In the 1950s, faced with the enormous task of gathering and analyzing written, photographic, and radar information about the enemy, the CIA and the NSA (National Security Agency) began to fund the first artificial intelligence projects. One of the earliest projects was a Program for Mechanical Translation, initiated in the early 1950s in the attempt to automate the monitoring of Soviet communications and media. The work on mechanical translation was probably the major cause of many subsequent developments in modern linguistics, its move towards formalization; it can be discerned in Noam Chomsky’s early theory
which, by postulating the existence of language universals in the domain of grammar, implied that translation between arbitrary human languages could be automated. The same work on mechanical translation was also one of the catalysts in the development of the field of pattern recognition, the precursor to computer vision. Pattern recognition is concerned with automatically detecting and identifying predetermined patterns in the flow of information. A typical example is character recognition, the first stage in the process of automating translation. Pattern recognition was also used in the U.S. for the monitoring of Soviet radio and telephone communication. Instead of listening to every transmission, an operator would be alerted if computer picked up certain words in the conversation.

As a "logistics of perception" came to dominate modern warfare and surveillance and as the space race began, image processing became another major new field of research. Image processing comprises techniques to improve images for human or computer interpretation. In 1964, the space program for the first time used image processing to correct distortions in the pictures of the Moon introduced by a on-board television camera of Ranger 7. In 1961, the National Photographic Interpretation Center (NPIC) was created to produce photoanalysis for the rest of the U.S. intelligence community and, as Manual De Landa points out, by the end of the next decade computers "were routinely used to correct for distortions made by satellite's imaging sensors and by atmospheric effects, sharpen out-of-focus images, bring multicolored single images out of several pictures taken in different spectral bands, extract particular features while diminishing or eliminating their backgrounds altogether..." De Landa also notes that computer analysis of photographic imagery became the only way to deal with the pure volume of intelligence being gathered: "It became apparent during the 1970s that there is no hope of keeping up with the millions of images that poured into NPIC...by simply looking at them the way they had been looked at in World War II. The computers therefore also had to be taught to compare new imagery of a given scene with old imagery, ignoring what had not changed and calling the interpreter's attention to what had."
The techniques of image processing, which can automatically increase an image's contrast, remove the effects of blur, extract edges, record differences between two images, and so on, greatly eased the job of human photoanalysts. And the combining of image processing with pattern recognition made it possible in some cases to delegate the analysis of photographs to a computer. For instance, the technique of pattern matching used to recognize printed characters can also be used to recognize objects in a satellite photograph. In both cases, the image is treated as consisting of two-dimensional forms. The contours of the forms are extracted from the image are then compared to templates stored in computer memory. If a contour found in the image matches a particular template, the computer signals that a corresponding object is present in a photograph.

A general purpose computer vision program has to be able to recognize not just two-dimensional but three-dimensional objects in a scene taken from an arbitrary angle. Only then it can be used to recognize an enemy's tank, to guide an automatic missile towards its target or to control a robotic arm on the factory floor. The problem with using simple template matching is that "a two-dimensional representation of a two-dimensional object is substantially like the object, but a two-dimensional representation of a three-dimensional object introduces a perspective projection that makes the representation ambiguous with respect to the object." While pattern recognition was working for images of two-dimensional objects, such as letters or chromosomes, a different approach was required to "see" in 3-D.

Roberts' 1965 paper "Machine Recognition of Three-dimensional Solids" is considered to be the first attempt at solving the general task of automatically recognizing three-dimensional objects. His program was designed to understand the artificial world composed solely of polyhedral blocks -- a reduction of reality to geometry that would have pleased Cézanne. Using image processing techniques, a photograph of a scene was first converted into a line drawing. Next, the techniques of 3-D computer graphics were used:
Roberts' program had access to three-dimensional models of objects: a cube, a rectangular solid, a wedge, and a hexagonal prism. They were represented by coordinates \((x, y, z)\) of their vertices. A program recognized these objects in a line drawing of the scene. A candidate model was selected on the basis of simple features such as a number of vertices. Then the selected model was rotated, scaled, projected, and matched with the input line drawing. If the match was good, the object was recognized, as were its position and size. Roberts' program could handle even a composite object made of multiple primitive shapes; it subtracted parts of a line drawing from the drawing as they were recognized, and the remaining portions were analyzed further.\(^25\)

Was this enough to completely automate human vision? This depends upon how we define vision. The chapter on computer vision in The Handbook of Artificial Intelligence (1982) opens with the following definition: "Vision is the information-processing task of understanding a scene from its projected images."\(^26\) But what does "understanding a scene" mean? With computer vision research financed by the military-industrial complex, the definition of understanding becomes highly pragmatic. In the best tradition of the pragmatism of James and Pierce, cognition is equated with action. The computer can be said to "understand" a scene if it can act on it -- move objects, assemble details, destroy targets. Thus, in the field of computer vision "understanding a scene" implies two goals. First, it means the identification of various objects represented in an image. Second, it means reconstruction of three-dimensional space from the image. A robot, for instance, need not only recognize particular objects, but it has to construct a representation of the surrounding environment to plan its movements. Similarly, a missile not only has to identify a target but also to determine the position of this target in three-dimensional space.

It can be seen that Roberts' program simultaneously fulfilled both goals. His program exemplified the approach taken by most computer vision researchers in the following two decades. A typical program first reconstructs the
three-dimensional scene from the input image and then matches the reconstructed objects to the models stored in memory. If the match is good, the program can be said to recognize the object, while simultaneously recording its position.

A computer vision program thus acts like a blind person who "sees" objects (i.e., identifies them) by reading their shapes through touch. As for a blind person, understanding the world and the recognition of shapes are locked together; they cannot be accomplished independently of one another.

In summary, early computer vision was limited to recognition of two-dimensional forms. Later, researchers began to tackle the task of recognizing 3-D objects which involves reconstruction of shapes from their perspectival representations (a photograph or a video image). From this point on, the subsequent history of computer vision research can be seen as a struggle against perspective inherent to the photographic optics.

The Retreat of Perspective

With the emergence of the field of computer vision, perspectival sight reaches its apotheosis and at the same time begins its retreat. At first computer vision researchers believed that they could invert the perspective and reconstruct the represented scene from a single perspectival image. Eventually, they realized that it is often easier to bypass perspectival images altogether and use other means as a source of three-dimensional information.

Latour points out that with the invention of perspective it became possible to represent absent things and plan our movement through space by working on representations. To quote him again, "one cannot smell or hear or touch Sakhalin island, but you can look at the map and determine at which bearing you will see the land when you send the next fleet." Roberts' program extended these abilities even further. Now the computer could acquire full knowledge of the three-dimensional world from a single perspectival image! And because the
program determined the exact position and orientation of objects in a scene, it became possible to see the reconstructed scene from another viewpoint. It also became possible to predict how the scene would look from an arbitrary viewpoint. Finally, it also became possible to guide automatically the movement of a robot through the scene.

Roberts' program worked using only a single photograph -- but only because it was presented with a highly artificial scene and because it "knew" what it could expect to see. Roberts limited the world which his program could recognize to simple polyhedral blocks. The shapes of possible blocks were stored in a computer. Others simplified the task even further by painting all objects in a scene the same color.

However, given an arbitrary scene, composed from arbitrary surfaces of arbitrary color and lighted in an arbitrary way, it is very difficult to reconstruct the scene correctly from a single perspectival image. The image is "underdetermined." First, numerous spatial layouts can give rise to the same two-dimensional image. Second, "the appearance of an object is influenced by its surface material, the atmospheric conditions, the angle of the light source, the ambient light, the camera angle and characteristics, and so on," and all of these different factors are collapsed together in the image. Third, perspective, as any other type of projection, does not preserve many geometric properties of a scene. Parallel lines turn into convergent lines; all angles change; equal lines appear unequal. All this makes it very difficult for a computer to determine which lines belong to a single object.

Thus, perspective, which until now served as a model of visual automation, becomes the drawback which needs to be overcome. Perspective, this first step towards the rationalization of sight (Ivins) has eventually become a limit to its total rationalization -- the development of computer vision.

The realization of the ambiguities inherent in a perspectival image itself came after years of vision research. In trying to compensate for these ambiguities, laboratories began to scrutinize the formal structure of a perspectival image with a degree of attention unprecedented in the history of perspective. For
instance, in 1968 Adolpho Guzman classified the types of junctions that appear in line representations after he realized that a junction type can be used to deduce whether regions of either side of a junction line were part of the same object. In 1986 David Lowe presented a method to calculate the probability that a particular regularity in an image (for instance, parallel lines) reflects the physical layout of the scene rather than being an accident due to a particular viewpoint. All other sources of depth information such as shading, shadows or texture gradients were also systematically studied and described mathematically.

Despite these advances, a single perspectival image remained too ambiguous a source of information for practical computer vision systems. An alternative has been to use more than one image at a time. Computer stereo systems employ two cameras which, like human eyes, are positioned a distance apart. If the common feature can be identified in both images, then the position of an object can be simply determined through geometric calculations. Other systems use a series of continuous images recorded by a video camera.

But why struggle with the ambiguity of perspectival images at all? Instead of inferring three-dimensional structure from a two-dimensional representation, it is possible to measure depth directly by employing various remote sensing technologies. In addition to video cameras, modern vision systems also utilize a whole range of different range finders such as lasers or ultrasound. Range finders are devices which can directly produce a three-dimensional map of an object. The same basic principle employed in radar is used: the time required for an electro-magnetic wave to reach an object and reflect back is proportional to the distance to the object. But while radar reduces an object to a single point and in fact is blind to close-by objects, a range finder operates at small distances. By systematically scanning the surface of an object, it directly produces a "depth map," a record of an object's shape which can be then matched to geometric models stored in computer memory thus bypassing the perspectival image altogether.

Thus, perspective occupies a special role in the history of computer imaging. A first algorithm .... Yet, while giving rise to new technologies of "geometric vision," perspective also becomes a limit to the final automation of sight -- recognition of objects by a computer.
Perspective, this first step towards the "rationalization of sight" (Ivins) has eventually become a limit to its total rationalization -- the development of computer vision. The perspective algorithm, a foundation of both computer graphics and computer vision, is used to generate perspectival views given a geometric model and to deduce the model given a perspectival view. Yet, while giving rise to new technologies of "geometric vision," perspective also becomes a limit to the final automation of sight -- recognition of objects by a computer. Finally, it is displaced from its privileged role, becoming just one among other techniques of space mapping and visualization.

Epilogue

The Renaissance's adaptation of perspective represented the first step in the automation of sight. While other cultures used sophisticated methods of space mapping, the importance of perspective lies not in its representational superiority but in its algorithmic character. This algorithmic character enabled the gradual development of visual languages of perspective and descriptive geometry and, in parallel, of perspectival machines and technologies, from a simple net described by Dürer to photography and radar. And when digital computers made possible mass automation in general, automation of perspectival vision and imaging followed soon.

The use of computers allowed to extend perspective, utilizing to the extreme its inherent qualities such as the algorithmic character and the reciprocal relationship it establishes between reality and representation. The perspective algorithm, a foundation of both computer graphics and computer vision, is used to generate perspectival views given a geometric model and to deduce the model given a perspectival view. Yet, while giving rise to new technologies of "geometric vision," perspective also becomes a limit to the final automation of sight -- recognition of objects by a computer. Finally, it is
displaced from its privileged role, becoming just one among other techniques of space mapping and visualization.


2 Qtd. in ibid., 67.

3 For a survey of perspectival instruments, see Martin Kemp, The Science of Art (New Haven: Yale University Press, 1990), 167-220.

4 Ibid., 171-172.

5 Ibid., 200.


11 This mixture of automated and pre-industrial labor is characteristic of the early uses of computers for the production of images. In 1955 the psychologist Attneave was the first to construct an image which was to become one of the icons of the age of digital visuality -- random squares pattern. A pattern consisted of a grid made from small squares colored black or white. A computer generated table of random numbers has been used to determine the colors of the square -- odd number for one color, even number for another. Using this procedure, two research assistants manually filled in 19,600 squares of the pattern. Paul Vitz and Arnold B. Glimcher, Modern Art and Modern Science (New York: Praeger Publishers, 1984), 234. Later, many artists, such as Harold Cohen, used computers to generate line drawings which they then colored by hand, transferred to canvas to serve as a foundation for painting, etc.


13 Mitchell, The Reconfigured Eye, 118.
14 "Retrospectives II: The Early Years in Computer Graphics at MIT, Lincoln Lab, and Harvard," 57.

Remko Scha, "Virtual Voices," MediaMatic 7, no. 1 (1992): 33. Scha describes two fundamental approaches taken by the developers of voice imitating machines: the genetic method which imitates the physiological processes that generate speech sounds in the human body and the gennematic method which is based on the analysis and reconstruction of speech sounds themselves without considering the way in which the human body produces them. While the field of computer vision, and other fields of artificial intelligence, first clearly followed gennematic method, in the 1980s, with the
growing popularity of neural networks, there was a shift towards the genetic method --
direct imitation of the physiology of the visual system. In a number of laboratories,
scientists begin to build artificial eyes which move, focus, and analyze information
exactly like human eyes.

16 Eames and Eames, A Computer Perspective, 100.

17 Manuel De Landa, “Policing the Spectrum,” in War in the Age of Intelligent Machines

18 Ibid., 214.

19 The first paper on image processing was published in 1955.

L.S.G. Kovasznay, and H.M. Joseph,


22 Within the field of computer vision, a scene is defined as a collection of three-dimensional objects depicted in an input picture. David McArthur, "Computer Vision and Perceptual Psychology," *Psychological Bulletin* 92, no. 2 (1982): 284.


26 Ibid., 127.


29 Ibid., 128.

30 Ibid., 131.
The Past, Present, and Future of Virtual and Augmented Reality Research: A Network and Cluster Analysis of the Literature

Pietro Cipresso1,2*, Irene Alice Chicchi Giglioli3, Mariano Alcañiz Raya3 and Giuseppe Riva1,2

1 Applied Technology for Neuro-Psychology Lab, Istituto Auxologico Italiano, Milan, Italy, 2 Department of Psychology, Catholic University of the Sacred Heart, Milan, Italy, 3 Instituto de Investigación e Innovación en Bioingeniería, Universitat Politècnica de València, Valencia, Spain

The recent appearance of low cost virtual reality (VR) technologies – like the Oculus Rift, the HTC Vive and the Sony PlayStation VR – and Mixed Reality Interfaces (MRITF) – like the Hololens – is attracting the attention of users and researchers suggesting it may be the next largest stepping stone in technological innovation. However, the history of VR technology is longer than it may seem: the concept of VR was formulated in the 1960s and the first commercial VR tools appeared in the late 1980s. For this reason, during the last 20 years, 100s of researchers explored the processes, effects, and applications of this technology producing 1000s of scientific papers. What is the outcome of this significant research work? This paper wants to provide an answer to this question by exploring, using advanced scientometric techniques, the existing research corpus in the field. We collected all the existent articles about VR in the Web of Science Core Collection scientific database, and the resultant dataset contained 21,667 records for VR and 9,944 for augmented reality (AR). The bibliographic record contained various fields, such as author, title, abstract, country, and all the references (needed for the citation analysis). The network and cluster analysis of the literature showed a composite panorama characterized by changes and evolutions over the time. Indeed, whether until 5 years ago, the main publication media on VR concerned both conference proceeding and journals, more recently journals constitute the main medium of communication. Similarly, if at first computer science was the leading research field, nowadays clinical areas have increased, as well as the number of countries involved in VR research. The present work discusses the evolution and changes over the time of the use of VR in the main areas of application with an emphasis on the future expected VR’s capacities, increases and challenges. We conclude considering the disruptive contribution that VR/AR/MRITF will be able to get in scientific fields, as well in human communication and interaction, as already happened with the advent of mobile phones by increasing the use and the development of scientific applications (e.g., in clinical areas) and by modifying the social communication and interaction among people.

Keywords: virtual reality, augmented reality, quantitative psychology, measurement, psychometrics, scientometrics, computational psychometrics, mathematical psychology
INTRODUCTION

In the last 5 years, virtual reality (VR) and augmented reality (AR) have attracted the interest of investors and the general public, especially after Mark Zuckerberg bought Oculus for two billion dollars (Luckerson, 2014; Castelvecchi, 2016). Currently, many other companies, such as Sony, Samsung, HTC, and Google are making huge investments in VR and AR (Korolov, 2014; Ebert, 2015; Castelvecchi, 2016). However, if VR has been used in research for more than 25 years, and now there are 1000s of papers and many researchers in the field, comprising a strong, interdisciplinary community, AR has a more recent application history (Burdea and Coiffet, 2003; Kim, 2005; Bohil et al., 2011; Cipresso and Serino, 2014; Wexelblat, 2014). The study of VR was initiated in the computer graphics field and has been extended to several disciplines (Sutherland, 1965, 1968; Mazuryk and Gervautz, 1996; Choi et al., 2015). Currently, videogames supported by VR tools are more popular than the past, and they represent valuables, work-related tools for neuroscientists, psychologists, biologists, and other researchers as well. Indeed, for example, one of the main research purposes lies from navigation studies that include complex experiments that could be done in a laboratory by using VR, whereas, without VR, the researchers would have to go directly into the field, possibly with limited use of intervention. The importance of navigation studies for the functional understanding of human memory in dementia has been a topic of significant interest for a long time, and, in 2014, the Nobel Prize in “Physiology or Medicine” was awarded to John M. O’Keefe, May-Britt Moser, and Edvard I. Moser for their discoveries of nerve cells in the brain that enable a sense of place and navigation. Journals and magazines have extended this knowledge by writing about “the brain GPS,” which gives a clear idea of the mechanism. A huge number of studies have been conducted in clinical settings by using VR (Bohil et al., 2011; Serino et al., 2014), and Nobel Prize winner, Edvard I. Moser commented about the use of VR (Minderer et al., 2016), highlighting its importance for research and clinical practice. Moreover, the availability of free tools for VR experimental and computational use has made it easy to access any field (Riva et al., 2011; Cipresso, 2015; Brown and Green, 2016; Cipresso et al., 2016).

Augmented reality is a more recent technology than VR and shows an interdisciplinary application framework, in which, nowadays, education and learning seem to be the most field of research. Indeed, AR allows supporting learning, for example increasing-on content understanding and memory preservation, as well as on learning motivation. However, if VR benefits from clear and more definite fields of application and research areas, AR is still emerging in the scientific scenarios.

In this article, we present a systematic and computational analysis of the emerging interdisciplinary VR and AR fields in terms of various co-citation networks in order to explore the evolution of the intellectual structure of this knowledge domain over time.

Virtual Reality Concepts and Features

The concept of VR could be traced at the mid of 1960 when Ivan Sutherland in a pivotal manuscript attempted to describe VR as a window through which a user perceives the virtual world as if looked, felt, sounded real and in which the user could act realistically (Sutherland, 1965).

Since that time and in accordance with the application area, several definitions have been formulated: for example, Fuchs and Bishop (1992) defined VR as “real-time interactive graphics with 3D models, combined with a display technology that gives the user the immersion in the model world and direct manipulation” (Fuchs and Bishop, 1992); Gigante (1993) described VR as “The illusion of participation in a synthetic environment rather than external observation of such an environment. VR relies on a 3D, stereoscopic head-tracker displays, hand/body tracking and binaural sound. VR is an immersive, multi-sensory experience” (Gigante, 1993); and “Virtual reality refers to immersive, interactive, multi-sensory, viewer-centered, 3D computer generated environments and the combination of technologies required building environments” (Cruz-Neira, 1993).

As we can notice, these definitions, although different, highlight three common features of VR systems: immersion, perception to be present in an environment, and interaction with that environment (Biocca, 1997; Lombard and Ditton, 1997; Loomis et al., 1999; Heeter, 2000; Biocca et al., 2001; Bailenson et al., 2006; Skalski and Tamborini, 2007; Andersen and Thorpe, 2009; Slater, 2009; Sundar et al., 2010). Specifically, immersion concerns the amount of senses stimulated, interactions, and the reality's similarity of the stimuli used to simulate environments. This feature can depend on the properties of the technological system used to isolate user from reality (Slater, 2009).

Higher or lower degrees of immersion can depend by three types of VR systems provided to the user:

- **Non-immersive systems** are the simplest and cheapest type of VR applications that use desktops to reproduce images of the world.
- **Immersive systems** provide a complete simulated experience due to the support of several sensory outputs devices such as head mounted displays (HMDs) for enhancing the stereoscopic view of the environment through the movement of the user’s head, as well as audio and haptic devices.
- **Semi-immersive systems** such as Fish Tank VR are between the two above. They provide a stereo image of a three dimensional (3D) scene viewed on a monitor using a perspective projection coupled to the head position of the observer (Ware et al., 1993). Higher technological immersive systems have showed a closest experience to reality, giving to the user the illusion of technological non-mediation and feeling him or her of “being in” or present in the virtual environment (Lombard and Ditton, 1997). Furthermore, higher immersive systems, than the other two systems, can give the possibility to add several sensory outputs allowing that the interaction and actions were...
Finally, the user's VR experience could be disclosed by measuring presence, realism, and reality's levels. Presence is a complex psychological feeling of "being there" in VR that involves the sensation and perception of physical presence, as well as the possibility to interact and react as if the user was in the real world (Heeter, 1992). Similarly, the realism's level corresponds to the degree of expectation that the user has about the stimuli and experience (Baños et al., 2000, 2009). If the presented stimuli are similar to reality, VR user's expectation will be congruent with reality expectation, enhancing VR experience. In the same way, higher is the degree of reality in interaction with the virtual stimuli, higher would be the level of realism of the user's behaviors (Baños et al., 2000, 2009).

From Virtual to Augmented Reality

Looking chronologically on VR and AR developments, we can trace the first 3D immersive simulator in 1962, when Morton Heilig created Sensorama, a simulated experience of a motorcycle running through Brooklyn characterized by several sensory impressions, such as audio, olfactory, and haptic stimuli, including also wind to provide a realist experience (Heilig, 1962). In the same years, Ivan Sutherland developed The Ultimate Display that, more than sound, smell, and haptic feedback, included interactive graphics that Sensorama didn't provide. Furthermore, Philco developed the first HMD that together with The Sword of Damocles of Sutherland was able to update the virtual images by tracking user's head position and orientation (Sutherland, 1965). In the 70s, the University of North Carolina realized GROPE, the first system of force-feedback and Myron Krueger created VIDEPLACE an Artificial Reality in which the users' body figures were captured by cameras and projected on a screen (Krueger et al., 1985). In this way two or more users could interact in the 2D-virtual space. In 1982, the US Air Force created the first flight simulator [Visually Coupled Airbone System Simulator (VCASS)] in which the pilot through an HMD could control the pathway and the targets. Generally, the 80's were the years in which the first commercial devices began to emerge: for example, in 1985 the VPL company commercialized the DataGlove, glove sensors' equipped able to measure the flexion of fingers, orientation and position, and identify hand gestures. Another example is the Eyphone, created in 1988 by the VPL Company, an HMD system for completely immersing the user in a virtual world. At the end of 80's, Fake Space Labs created a Binocular-Omni-Orientational Monitor (BOOM), a complex system composed by a stereoscopic-displaying device, providing a moving and broad virtual environment, and a mechanical arm tracking. Furthermore, BOOM offered a more stable image and giving more quickly responses to movements than the HMD devices. Thanks to BOOM and DataGlove, the NASA Ames Research Center developed the Virtual Wind Tunnel in order to research and manipulate airflow in a virtual airplane or space ship. In 1992, the Electronic Visualization Laboratory of the University of Illinois created the CAVE Automatic Virtual Environment, an immersive VR system composed by projectors directed on three or more walls of a room.

More recently, many videogames companies have improved the development and quality of VR devices, like Oculus Rift, or HTC Vive that provide a wider field of view and lower latency. In addition, the actual HMD's devices can be now combined with other tracker system as eye-tracking systems (FOVE), and motion and orientation sensors (e.g., Razer Hydra, Oculus Touch, or HTC Vive).

Simultaneously, at the beginning of 90', the Boing Corporation created the first prototype of AR system for showing to employees how set up a wiring tool (Carmigniani et al., 2011). At the same time, Rosenberg and Feiner developed an AR fixture for maintenance assistance, showing that the operator performance enhanced by added virtual information on the fixture to repair (Rosenberg, 1993). In 1993 Loomis and colleagues produced an AR GPS-based system for helping the blind in the assisted navigation through adding spatial audio information (Loomis et al., 1998). Always in the 1993 Julie Martin developed "Dancing in Cyberspace," an AR theater in which actors interacted with virtual object in real time (Cathy, 2011). Few years later, Feiner et al. (1997) developed the first Mobile AR System (MARS) able to add virtual information about touristic buildings (Feiner et al., 1997). Since then, several applications have been developed: in Thomas et al. (2000), created ARQuake, a mobile AR video game; in 2008 was created Wikitude that through the mobile camera, internet, and GPS could add information about the user's environments (Perry, 2008). In 2009 others AR applications, like AR Toolkit and SiteLens have been developed in order to add virtual information to the physical user's surroundings. In 2011, Total Immersion developed D'Vive, and AR system for designing projects (Maurugeon, 2011). Finally, in 2013 and 2015, Google developed Google Glass and Google HoloLens, and their usability have begun to test in several field of application.

Virtual Reality Technologies

Technologically, the devices used in the virtual environments play an important role in the creation of successful virtual experiences. According to the literature, can be distinguished input and output devices (Burdea et al., 1996; Burdea and Coiffet, 2003). Input devices are the ones that allow the user to communicate with the virtual environment, which can range from a simple joystick or keyboard to a glove allowing capturing finger movements or a tracker able to capture postures. More in detail, keyboard, mouse, trackball, and joystick represent the desktop input devices easy to use, which allow the user to launch continuous and discrete commands or movements to the environment. Other input devices can be represented by tracking devices as bend-sensing gloves that capture hand movements, postures and gestures, or pinch gloves that detect the fingers movements, and trackers able to follow the user's movements in the physical world and translate them in the virtual environment.

On the contrary, the output devices allow the user to see, hear, smell, or touch everything that happens in the virtual environment. As mentioned above, among the visual devices can be found a wide range of possibilities, from the simplest or least
immersive (monitor of a computer) to the most immersive one such as VR glasses or helmets or HMD or CAVE systems.

Furthermore, auditory, speakers, as well as haptic output devices are able to stimulate body senses providing a more real virtual experience. For example, haptic devices can stimulate the touch feeling and force models in the user.

Virtual Reality Applications
Since its appearance, VR has been used in different fields, as for gaming (Zyda, 2005; Meldrum et al., 2012), military training (Alexander et al., 2017), architectural design (Song et al., 2017), education (Englund et al., 2017), learning and social skills training (Schmidt et al., 2017), simulations of surgical procedures (Gallagher et al., 2005), assistance to the elderly or psychological treatments are other fields in which VR is bursting strongly (Freeman et al., 2017; Neri et al., 2017).

A recent and extensive review of Slater and Sanchez-Vives (2016) reported the main VR application evidences, including weakness and advantages, in several research areas, such as science, education, physical training, as well as medical phenomena, behavioral, and could be used in other fields, like travel, meetings, collaboration, industry, news, and entertainment. Furthermore, another review published this year by Freeman et al. (2017) focused on VR in mental health, showing the efficacy of VR in assessing and treating different psychological disorders as anxiety, schizophrenia, depression, and eating disorders.

There are many possibilities that allow the use of VR as a stimulus, replacing real stimuli, recreating experiences, which in the real world would be impossible, with a high realism. This is why VR is widely used in research on new ways of applying psychological treatments or training, for example, to problems arising from phobias (agoraphobia, phobia to fly, etc.) (Botella et al., 2017). Or, simply, it is used like improvement of the traditional systems of motor rehabilitation (Llorens et al., 2014; Borrego et al., 2016), developing games that ameliorate the tasks. More in detail, in psychological treatment, Virtual Reality Exposure Therapy (VRET) has showed its efficacy, allowing to patients to gradually face fear stimuli or stressed situations in a safe environment where the psychological and physiological reactions can be controlled by the therapist (Botella et al., 2017).

Augmented Reality Technologies
Technologically, the AR systems, however various, present three common components, such as a geospatial datum for the virtual object, like a visual marker, a surface to project virtual elements to the user, and an adequate processing power for graphics, animation, and merging of images, like a pc and a monitor (Carmigniani et al., 2011). To run, an AR system must also include a camera able to track the user movement for merging the virtual objects, and a visual display, like glasses through that the user can see the virtual objects overlaying to the physical world. To date, two-display systems exist, a video see-through (VST) and an optical see-though (OST) AR systems (Botella et al., 2005; Juan et al., 2005, 2007). The first one, disclosures virtual objects to the user by capturing the real objects/scenes with a camera and overlaying virtual objects, projecting them on a video or a monitor, while the second one, merges the virtual object on a transparent surface, like glasses, through the user see the added elements. The main difference between the two systems is the latency: an OST system could require more time to display the virtual objects than a VST system, generating a time lag between user's action and performance and the detection of them by the system.

Augmented Reality Applications
Although AR is a more recent technology than VR, it has been investigated and used in several research areas such as architecture (Lin and Hsu, 2017), maintenance (Schwald and De Laval, 2003), entertainment (Ozbek et al., 2004), education (Ninca rean et al., 2013; Bacca et al., 2014; Akçayır and Akçayır, 2017), medicine (De Buck et al., 2005), and psychological treatments (Juan et al., 2005; Botella et al., 2005, 2010; Breitón-López et al., 2010; Wrzesien et al., 2011a,b, 2013; see the review Chicchi Giglioli et al., 2015). More in detail, in education several AR applications have been developed in the last few years showing the positive effects of this technology in supporting learning, such as an increased-on content understanding and memory preservation, as well as on learning motivation (Radu, 2012, 2014). For example, Ibáñez et al. (2014) developed a AR application on electromagnetism concepts' learning, in which students could use AR batteries, magnets, cables on real superfficies, and the system gave a real-time feedback to students about the correctness of the performance, improving in this way the academic success and motivation (Di Serio et al., 2013). Deeply, AR system allows the possibility to learn visualizing and acting on composite phenomena that traditionally students study theoretically, without the possibility to see and test in real world (Chien et al., 2010; Chen et al., 2011).
TABLE 1 | Category statistics from the WoS for the entire period and the last 5 years.

<table>
<thead>
<tr>
<th>%</th>
<th>Frequency</th>
<th>Subject category (for all the period)</th>
<th>%</th>
<th>Frequency</th>
<th>Subject category (for the last 5 years)</th>
</tr>
</thead>
</table>

As well in psychological health, the number of research about AR is increasing, showing its efficacy above all in the treatment of psychological disorder (see the reviews Baus and Bouchard, 2014; Chiuchi and Girolami et al., 2015). For example, in the treatment of anxiety disorders, like phobias, AR exposure therapy (ARET) showed its efficacy in one-session treatment, maintaining the positive impact in a follow-up at 1 or 3 month after. As VRET, ARET provides a safety and an ecological environment where any kind of stimulus is possible, allowing to keep control over the situation experienced by the patients, gradually generating situations of fear or stress. Indeed, in situations of fear, like the phobias for small animals, AR applications allow, in accordance with the patient’s anxiety, to gradually expose patient to fear animals, adding new animals during the session or enlarging their or increasing the speed. The various studies showed that AR is able, at the beginning of the session, to activate patient’s anxiety, for reducing after 1 h of exposition. After the session, patients even more than to better manage animal’s fear and anxiety, were able to approach, interact, and kill real feared animals.

MATERIALS AND METHODS

Data Collection
The input data for the analyses were retrieved from the scientific database Web of Science Core Collection (Falagas et al., 2008) and the search terms used were “Virtual Reality” and “Augmented Reality” regarding papers published during the whole timespan covered.

Web of science core collection is composed of: Citation Indexes, Science Citation Index Expanded (SCI-EXPANDED) –1970-present, Social Sciences Citation Index (SSCI) –1970-present, Arts and Humanities Citation Index (A&HCI) –1975-present, Conference Proceedings Citation Index- Science (CPCI-S) –1990-present, Conference Proceedings Citation Index- Social Science & Humanities (CPCI-SSH) –1990-present, Book Citation Index– Science (BKCI-S) –2009-present, Book Citation Index– Social Sciences & Humanities (BKCI-SSH) –2009-present, Emerging Sources Citation Index (ESCI) –2015-present, Chemical Indexes, Current Chemical Reactions (CCR-EXPANDED) –2009-present (Includes Institut National de la Propriete Industrielle structure data back to 1840), Index Chemicus (IC) –2009-present.

The resultant dataset contained a total of 21,667 records for VR and 9,944 records for AR. The bibliographic record contained various fields, such as author, title, abstract, and all of the references (needed for the citation analysis). The research tool to visualize the networks was Cite space v.4.0.R5 SE (32 bit) (Chen, 2006) under Java Runtime v.8 update 91 (build 1.8.0_91-b15). Statistical analyses were conducted using Stata MP-Parallel Edition, Release 14.0, StataCorp LP. Additional information can be found in Supplementary Data Sheet 1.

The betweenness centrality of a node in a network measures the extent to which the node is part of paths that connect an arbitrary pair of nodes in the network (Freeman, 1977; Brandes, 2001; Chen, 2006).

Structural metrics include betweenness centrality, modularity, and silhouette. Temporal and hybrid metrics include citation burstness and novelty. All the algorithms are detailed (Chen et al., 2010).

RESULTS
The analysis of the literature on VR shows a complex panorama. At first sight, according to the document-type statistics from the Web of Science (WoS), proceedings papers were used extensively as outcomes of research, comprising almost 48% of the total (10,392 proceedings), with a similar number of articles on the subject amounting to about 47% of the total of 10,199 articles. However, if we consider only the last 5 years (7,755 articles representing about 36% of the total), the situation changes with about 57% for articles (4,445) and about 33% for proceedings (2,578). Thus, it is clear that VR field has changed in areas other than at the technological level.

About the subject category, nodes and edges are computed as co-occurring subject categories from the Web of Science “Category” field in all the articles.

According to the subject category statistics from the WoS, computer science is the leading category, followed by engineering, and, together, they account for 15,341 articles, which make up about 71% of the total production. However, if we
FIGURE 1 | Category from the WoS: network for the last 5 years.

FIGURE 2 | Country network (node dimension represents centrality).
consider just the last 5 years, these categories reach only about 55%, with a total of 4,284 articles (Table 1 and Figure 1).

The evidence is very interesting since it highlights that VR is doing very well as new technology with huge interest in hardware and software components. However, with respect to the past, we are witnessing increasing numbers of applications, especially in the medical area. In particular, note its inclusion in the top 10 list of rehabilitation and clinical neurology categories (about 10% of the total production in the last 5 years). It also is interesting that neuroscience and neurology, considered together, have shown an increase from about 12% to about 18.6% over the last 5 years. However, historic areas, such as automation and control systems, imaging science and photographic technology, and robotics, which had accounted for about 14.5% of the total articles ever produced were not even in the top 10 for the last 5 years, with each one accounting for less than 4%.

About the countries, nodes and edges are computed as networks of co-authors countries. Multiple occurrence of a country in the same paper are counted once.

The countries that were very involved in VR research have published for about 47% of the total (10,200 articles altogether). Of the 10,200 articles, the United States, China, England, and Germany published 4921, 2384, 1497, and 1398, respectively. The situation remains the same if we look at the articles published over the last 5 years. However, VR contributions also came from all over the globe, with Japan, Canada, Italy, France, Spain, South Korea, and Netherlands taking positions of prominence, as shown in Figure 2.

Network analysis was conducted to calculate and to represent the centrality index (Freeman, 1977; Brandes, 2001), i.e., the dimension of the node in Figure 2. The top-ranked country, with a centrality index of 0.26, was the United States (2011), and England was second, with a centrality index of 0.25. The third, fourth, and fifth countries were Germany, Italy, and Australia, with centrality indices of 0.15, 0.15, and 0.14, respectively.

About the Institutions, nodes and edges are computed as networks of co-authors Institutions (Figure 3).

The top-level institutions in VR were in the United States, where three universities were ranked as the top three in the
world for published articles; these universities were the University of Illinois (159), the University of South California (147), and the University of Washington (146). The United States also had the eighth-ranked university, which was Iowa State University (116). The second country in the ranking was Canada, with the University of Toronto, which was ranked fifth with 125 articles and McGill University, ranked 10th with 103 articles.

Other countries in the top-ten list were Netherlands, with the Delft University of Technology ranked fourth with 129 articles; Italy, with IRCCS Istituto Auxologico Italiano, ranked sixth (with the same number of publication of the institution ranked fifth) with 125 published articles; England, which was ranked seventh with 125 articles from the University of London’s Imperial College of Science, Technology, and Medicine; and China with 104 publications, with the Chinese Academy of Science, ranked ninth. Italy’s Istituto Auxologico Italiano, which was ranked fifth, was the only non-university institution ranked in the top-10 list for VR research (Figure 3).

About the Journals, nodes, and edges are computed as journal co-citation networks among each journals in the corresponding field. The top-ranked Journals for citations in VR are Presence: Teleoperators & Virtual Environments with 2689 citations and CyberPsychology & Behavior (Cyberpsychol BEHAV) with 1884 citations; however, looking at the last 5 years, the former had increased the citations, but the latter had a far more significant increase, from about 70% to about 90%, i.e., an increase from 1029 to 1147.

Following the top two journals, IEEE Computer Graphics and Applications (IEEE Comput Graph) and Advanced Health Telematics and Telemedicine (St HEAL T) were both left out of the top-10 list based on the last 5 years. The data for the last 5 years also resulted in the inclusion of Experimental Brain Research (Exp BRAIN RES) (625 citations), Archives of Physical Medicine and Rehabilitation (Arch PHYS MED REHAB) (622 citations), and Plos ONE (619 citations) in the top-10 list of three journals, which highlighted the categories of rehabilitation and clinical neurology and neuroscience and neurology. Journal co-citation analysis is reported in Figure 4, which clearly shows four distinct clusters.

Network analysis was conducted to calculate and to represent the centrality index, i.e., the dimensions of the nodes in Figure 4. The top-ranked item by centrality was Cyberpsychol BEHAV, with a centrality index of 0.29. The second-ranked item was Arch PHYS MED REHAB, with a centrality index of 0.23. The third was Behaviour Research and Therapy (Behav RES THER), with a centrality index of 0.15. The fourth was BRAIN, with a centrality index of 0.09. The fifth was CyberPsychology & Behavior, with a centrality index of 0.08. The sixth was Presence: Teleoperators & Virtual Environments, with a centrality index of 0.07. The seventh was Arch Phys Med Rehab, with a centrality index of 0.06. The eighth was Plos ONE, with a centrality index of 0.05. The ninth was Experimental Brain Research, with a centrality index of 0.04. The tenth was IEEE Computer Graphics and Applications, with a centrality index of 0.03.
index of 0.14. The fifth was Exp BRAIN RES, with a centrality index of 0.11.

**Who’s Who in VR Research**

Authors are the heart and brain of research, and their roles in a field are to define the past, present, and future of disciplines and to make significant breakthroughs to make new ideas arise (Figure 5).

Virtual reality research is very young and changing with time, but the top-10 authors in this field have made fundamentally significant contributions as pioneers in VR and taking it beyond a mere technological development. The purpose of the following highlights is not to rank researchers; rather, the purpose is to identify the most active researchers in order to understand where the field is going and how they plan for it to get there.

The top-ranked author is Riva G, with 180 publications. The second-ranked author is Rizzo A, with 101 publications. The third is Darzi A, with 97 publications. The forth is Aggarwal R, with 94 publications. The six authors following these three are Slater M, Alcaniz M, Botella C, Wiederhold BK, Kim SI, and Gutierrez-Maldonado J with 90, 90, 85, 75, 59, and 54 publications, respectively (Figure 6).

Considering the last 5 years, the situation remains similar, with three new entries in the top-10 list, i.e., Muhlberger A, Cipresso P, and Ahmed K ranked 7th, 8th, and 10th, respectively.

The authors’ publications number network shows the most active authors in VR research. Another relevant analysis for our focus on VR research is to identify the most cited authors in the field.

For this purpose, the authors’ co-citation analysis highlights the authors in term of their impact on the literature considering
the entire time span of the field (White and Griffith, 1981; González-Teruel et al., 2015; Bu et al., 2016). The idea is to focus on the dynamic nature of the community of authors who contribute to the research.

Normally, authors with higher numbers of citations tend to be the scholars who drive the fundamental research and who make the most meaningful impacts on the evolution and development of the field. In the following, we identified the most-cited pioneers in the field of VR Research.

The top-ranked author by citation count is Gallagher (2001), with 694 citations. Second is Seymour (2004), with 668 citations. Third is Slater (1999), with 649 citations. Fourth is Grantcharov (2003), with 563 citations. Fifth is Riva (1999), with 546 citations. Sixth is Aggarwal (2006), with 505 citations. Seventh is Satava (1994), with 477 citations. Eighth is Witmer (2002), with 454

**Citation Network and Cluster Analyses for VR**

Another analysis that can be used is the analysis of document co-citation, which allows us to focus on the highly-cited documents that generally are also the most influential in the domain (Small, 1973; González-Teruel et al., 2015; Orosz et al., 2016).


The network of document co-citations is visually complex (Figure 7) because it includes 1000s of articles and the links among them. However, this analysis is very important because it can be used to identify the possible conglomeration of knowledge in the area, and this is essential for a deep understanding of the area. Thus, for this purpose, a cluster analysis was conducted (Chen et al., 2010; González-Teruel et al., 2015; Klavans and Boyack, 2015). Figure 8 shows the clusters, which are identified with the two algorithms in Table 2.

The identified clusters highlight clear parts of the literature of VR research, making clear and visible the interdisciplinary nature of this field. However, the dynamics to identify the past, present, and future of VR research cannot be clear yet. We analysed the relationships between these clusters and the temporal dimensions of each article. The results are synthesized in Figure 9. It is clear that cluster #0 (laparoscopic skill), cluster #2 (gaming and rehabilitation), cluster #4 (therapy), and cluster #14 (surgery) are the most popular areas of VR research. (See Figure 9 and Table 2 to identify the clusters.) From Figure 9, it also is possible to identify the first phase of laparoscopic skill (cluster #6) and therapy (cluster #7). More generally, it is possible to identify four historical phases (colors: blue, green, yellow, and red) from the past VR research to the current research.

We were able to identify the top 486 references that had the most citations by using burst citations algorithm. Citation burst is an indicator of a most active area of research. Citation burst is a detection of a burst event, which can last for multiple years as well as a single year. A citation burst provides evidence that a particular publication is associated with a surge of citations. The

---

**FIGURE 7** | Network of document co-citations: the dimensions of the nodes represent centrality, the dimensions of the characters represent the rank of the article rank, and the numbers represent the strengths of the links. It is possible to identify four historical phases (colors: blue, green, yellow, and red) from the past VR research to the current research.

Citation Network and Cluster Analyses for AR
Looking at Augmented Reality scenario, the top ranked item by citation counts is Azuma (1997) in Cluster #0, with citation counts of 231. The second one is Azuma et al. (2001) in Cluster #0, with citation counts of 220. The third is Van Krevelen (2010) in Cluster #5, with citation counts of 207. The 4th is Lowe (2004) in Cluster #1, with citation counts of 157. The 5th is Wu (2013) in Cluster #4, with citation counts of 144. The 6th is Dunleavy (2009) in Cluster #4, with citation counts of 122. The 7th is Zhou (2008) in Cluster #5, with citation counts of 118. The 8th is Bay (2008) in Cluster #1, with citation counts of 117. The 9th is Newcombe (2011) in Cluster #1, with citation counts of 109. The 10th is Carmigniani et al. (2011) in Cluster #5, with citation counts of 104.

The network of document co-citations is visually complex (Figure 10) because it includes 1000s of articles and the links among them. However, this analysis is very important because it can be used to identify the possible conglomerate of knowledge in the area, and this is essential for a deep understanding of the area. Thus, for this purpose, a cluster analysis was conducted (Chen et al., 2010; González-Teruel et al., 2015; Klavans and Boyack, 2015). Figure 11 shows the clusters, which are identified with the two algorithms in Table 3.

The identified clusters highlight clear parts of the literature of AR research, making clear and visible the interdisciplinary nature of this field. However, the dynamics to identify the past, present, and future of AR research cannot be clear yet. We analysed the relationships between these clusters and the temporal dimensions of each article. The results are synthesized in Figure 12. It is clear that cluster #1 (tracking), cluster #4 (education), and cluster #5 (virtual city environment) are the current areas of AR research. (See Figure 12 and Table 3 to identify the clusters.) It is possible...
TABLE 2 | Cluster ID and silhouettes as identified with two algorithms (Chen et al., 2010).

<table>
<thead>
<tr>
<th>ID</th>
<th>Size</th>
<th>Silho-uette</th>
<th>Mean (Citee Year)</th>
<th>Label (TFIDF, tf*idf weighting algorithm)</th>
<th>Label (LLR, log-likelihood ratio, p-level)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>84</td>
<td>0.812</td>
<td>2005</td>
<td>(25.82) laparoscopic skill; (25.01) proficiency; (24.5) basic laparoscopic skill; (24.14) trainer; (23.79) establishing validity</td>
<td>Training (143.21, 1.0E-4); performance (73.38, 1.0E-4); laparoscopic skill (72.93, 1.0E-4)</td>
</tr>
<tr>
<td>1</td>
<td>77</td>
<td>0.758</td>
<td>1992</td>
<td>(17.76) ergonomic; (17.66) reality; (16.83) virtual reality; (16.04) virtual environment; (15.76) assembly</td>
<td>Ergonomic (54.1, 1.0E-4); virtual reality interface (34.63, 1.0E-4); developing virtual environment (34.48, 1.0E-4)</td>
</tr>
<tr>
<td>2</td>
<td>62</td>
<td>0.992</td>
<td>2007</td>
<td>(24.5) gaming; (24.5) wi; (24.47) stroke; (23.07) rehabilitation; (22.38) cerebral palsy</td>
<td>Stroke (82.9, 1.0E-4); children (75.13, 1.0E-4); stroke rehabilitation (57.95, 1.0E-4)</td>
</tr>
<tr>
<td>3</td>
<td>61</td>
<td>0.758</td>
<td>1994</td>
<td>(15) reality; (14.66) virtual reality; (14.25) surgery; (14.1) telemedical information society; (13.73) chemistry</td>
<td>Telemedical information society (34.85, 1.0E-4); gaining insight (23.21, 1.0E-4); next decade (18.32, 1.0E-4)</td>
</tr>
<tr>
<td>4</td>
<td>56</td>
<td>0.934</td>
<td>2008</td>
<td>(25.4) therapy; (23.55) exposure therapy; (22.41) disorder; (21.60) virtual reality exposure therapy; (20.99) post-traumatic stress</td>
<td>Treatment (109.92, 1.0E-4); post-traumatic stress disorder (78.95, 1.0E-4); virtual reality exposure therapy (66.15, 1.0E-4)</td>
</tr>
<tr>
<td>5</td>
<td>49</td>
<td>0.885</td>
<td>1992</td>
<td>(16.03) reality; (15.31) virtual reality; (15.01) autistic children; (12.79) child; (12.79) children</td>
<td>Autistic children (29.81, 1.0E-4); possibilities (23.84, 1.0E-4); communication (22.08, 1.0E-4)</td>
</tr>
<tr>
<td>6</td>
<td>41</td>
<td>0.855</td>
<td>1998</td>
<td>(17.6) laparoscopic skill; (16.95) direct observation; (16.95) measuring operative performance; (16.95) videotape; (16.15) measuring</td>
<td>Laparoscopic skills training (52.73, 1.0E-4); measuring operative performance (40.97, 1.0E-4); videotape (40.97, 1.0E-4)</td>
</tr>
<tr>
<td>7</td>
<td>41</td>
<td>0.946</td>
<td>1998</td>
<td>(20.71) therapy; (18.76) exposure therapy; (17.85) exposure; (17.35) anxiety; (17.2) virtual reality exposure therapy</td>
<td>Virtual reality exposure therapy (32.01, 1.0E-4); spider phobia (27.67, 1.0E-4); ptsd vietnam veteran (22.12, 1.0E-4)</td>
</tr>
<tr>
<td>8</td>
<td>38</td>
<td>1</td>
<td>1989</td>
<td>(30.67) Japanese institutional mechanism; (30.67) systems perspective; (20.88) mechanism; (19.25) perspective; (17.97) system</td>
<td>Japanese institutional mechanism (615.45, 1.0E-4); systems perspective (615.45, 1.0E-4); virtual reality (16.28, 1.0E-4)</td>
</tr>
<tr>
<td>9</td>
<td>21</td>
<td>1</td>
<td>1987</td>
<td>(23.27) routine use; (23.27) current application; (23.27) behavioral-assessment; (23.27) obstacle; (23.27) future possibilities</td>
<td>Future possibilities (168.77, 1.0E-4); routine use (168.77, 1.0E-4); current application (168.77, 1.0E-4)</td>
</tr>
<tr>
<td>10</td>
<td>18</td>
<td>0.934</td>
<td>1991</td>
<td>(12.45) reality; (12.26) virtual-reality; (9.73) medicine; (9.07) virtual reality; (5.71) technology</td>
<td>Virtual-reality (88.95, 1.0E-4); medicine (34.87, 1.0E-4); pretty interface (9.63, 0.005)</td>
</tr>
<tr>
<td>11</td>
<td>16</td>
<td>0.937</td>
<td>1990</td>
<td>(13.37) tutorial; (12.45) reality; (11.98) virtual reality; (11.12) virtual reality technology; (10.78) technology</td>
<td>Tutorial (51.15, 1.0E-4); virtual reality technology (44.66, 1.0E-4); space (16.78, 1.0E-4)</td>
</tr>
<tr>
<td>12</td>
<td>12</td>
<td>1</td>
<td>1988</td>
<td>(20.05) special effect; (20.05) cyberspace; (13.65) space; (11.38) effect; (10.73) reality</td>
<td>Special effect (128.6, 1.0E-4); cyberspace (128.6, 1.0E-4); virtual reality (27.79, 1.0E-4)</td>
</tr>
<tr>
<td>13</td>
<td>8</td>
<td>0.995</td>
<td>1997</td>
<td>(14.88) neural substrate; (14.88) human spatial navigation; (14.88) cognitive map; (11.58) navigation; (10.64) cognitive</td>
<td>Neural substrate (72.6, 1.0E-4); human spatial navigation (66.58, 1.0E-4); cognitive map (66.58, 1.0E-4)</td>
</tr>
<tr>
<td>14</td>
<td>6</td>
<td>0.993</td>
<td>2008</td>
<td>(12.06) neurosurgery; (9.74) computer technology; (9.74) surgical application; (9.43) surgery; (8.55) teaching</td>
<td>Neurosurgery (28.72, 1.0E-4); computer technology (18.1, 1.0E-4); surgical application (18.1, 1.0E-4)</td>
</tr>
</tbody>
</table>

to identify four historical phases (colors: blue, green, yellow, and red) from the past AR research to the current research.

We were able to identify the top 394 references that had the most citations by using burst citations algorithm. Citation burst is an indicator of a most active area of research. Citation burst is a detection of a burst event, which can last for multiple years as well as a single year. A citation burst provides evidence that a particular publication is associated with a surge of citations. The burst detection was based on Kleinberg’s algorithm (Kleinberg, 2002, 2003). The top ranked document by bursts is Azuma (1997) in Cluster #0, with bursts of 101.64. The second one is Azuma et al. (2001) in Cluster #0, with bursts of 84.23. The third is Lowe (2004) in Cluster #1, with bursts of 64.07. The 4th is Van Krevelen (2010) in Cluster #5, with bursts of 50.99. The 5th is Wu (2013) in Cluster #4, with bursts of 47.23. The 6th is Hartley (2000) in Cluster #0, with bursts of 37.71. The 7th is Dunleavy (2009) in Cluster #4, with bursts of 33.22. The 8th is Kato (1999) in Cluster #8, with bursts of 32.16. The 9th is Newcombe (2011) in Cluster #0, with bursts of 30.72. The 10th is Feiner (1993) in Cluster #0, with bursts of 29.72.

**DISCUSSION**

Our findings have profound implications for two reasons. At first the present work highlighted the evolution and development of VR and AR research and provided a clear perspective based on solid data and computational analyses. Secondly our findings on VR made it profoundly clear that the clinical dimension is one of the most investigated ever and seems to
Figure 9 clarifies the past, present, and future of VR research. The outset of VR research brought a clearly-identifiable development in interfaces for children and medicine, routine use and behavioral-assessment, special effects, systems perspectives, and tutorials. This pioneering era evolved in the period that we can identify as the development era, because it was the period in which VR was used in experiments associated with new technological impulses. Not surprisingly, this was exactly concomitant with the new economy era in which significant investments were made in information technology, and it also was the era of the so-called ‘dot-com bubble’ in the late 1990s.

The confluence of pioneering techniques into ergonomic studies within this development era was used to develop the first effective clinical systems for surgery, telemedicine, human spatial navigation, and the first phase of the development of therapy and laparoscopic skills. With the new millennium, VR research switched strongly toward what we can call the clinical-VR era, with its strong emphasis on rehabilitation, neurosurgery, and a new phase of therapy and laparoscopic skills. The number of applications and articles that have been published in the last 5 years are in line with the new technological development that we are experiencing at the hardware level, for example, with so many new, HMDs, and at the software level with an increasing number of independent programmers and VR communities.
Finally, Figure 12 identifies clusters of the literature of AR research, making clear and visible the interdisciplinary nature of this field. The dynamics to identify the past, present, and future of AR research cannot be clear yet, but analyzing the relationships between these clusters and the temporal dimensions of each article tracking, education, and virtual city environment are the current areas of AR research. AR is a new technology that is showing its efficacy in different research fields, and providing a novel way to gather behavioral data and support learning, training, and clinical treatments.

Looking at scientific literature conducted in the last few years, it might appear that most developments in VR and AR studies have focused on clinical aspects. However, the reality is more complex; thus, this perception should be clarified. Although researchers publish studies on the use of VR in clinical settings, each study depends on the technologies available. Industrial development in VR and AR changed a lot in the last 10 years. In the past, the development involved mainly hardware solutions while nowadays, the main efforts pertain to the software when developing virtual solutions. Hardware became a commodity that is often available at low cost. On the other hand, software needs to be customized each time, per each experiment, and this requires huge efforts in term of development. Researchers in AR and VR today need to be able to adapt software in their labs.

Virtual reality and AR developments in this new clinical era rely on computer science and vice versa. The future of VR and AR is becoming more technological than before, and each day, new solutions and products are coming to the market. Both from software and hardware perspectives, the future of AR and VR depends on huge innovations in all fields. The gap between the past and the future of AR and VR research is about the “realism” that was the key aspect in the past versus the “interaction” that is the key aspect now. First 30 years of VR and AR consisted of a continuous research on better resolution and improved perception. Now, researchers already achieved a great resolution and need to focus on making the VR as realistic as possible, which is not simple. In fact, a real experience implies a realistic interaction and not just great resolution. Interactions can be improved in infinite ways through new developments at hardware and software levels.

Interaction in AR and VR is going to be “embodied,” with implication for neuroscientists that are thinking about new solutions to be implemented into the current systems (Blanke...
et al., 2015; Riva, 2018; Riva et al., 2018). For example, the use of hands with contactless device (i.e., without gloves) makes the interaction in virtual environments more natural. The Leap Motion device allows one to use hands in VR without the use of gloves or markers. This simple and low-cost device allows the VR users to interact with virtual objects and related environments in a naturalistic way. When technology is able to be transparent, users can experience increased sense of being in the virtual environments (the so-called sense of presence).

Other forms of interactions are possible and have been developing continuously. For example, tactile and haptic device able to provide a continuous feedback to the users, intensifying their experience also by adding components, such as the feeling of touch and the physical weight of virtual objects, by using force feedback. Another technology available at low cost that facilitates interaction is the motion tracking system, such as

---

**FIGURE 11** | Document co-citation network by cluster: the dimensions of the nodes represent centrality, the dimensions of the characters represent the rank of the article rank and the red writing reports the name of the cluster with a short description that was produced with the mutual information algorithm; the clusters are identified with colored polygons.

---

**TABLE 4** | Cluster ID and silhouettes as identified with two algorithms (Chen et al., 2010).

<table>
<thead>
<tr>
<th>ID</th>
<th>Size</th>
<th>Silhouette</th>
<th>Mean (Citee Year)</th>
<th>Label (TFIDF, tf*idf weighting algorithm)</th>
<th>Label (LLR, log-likelihood ratio, p-level)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>122</td>
<td>0.669</td>
<td>1999</td>
<td>(18.41) internet</td>
<td>Internet (39.96, 1.0E-4)</td>
</tr>
<tr>
<td>1</td>
<td>66</td>
<td>0.806</td>
<td>2007</td>
<td>(16.67) tracking</td>
<td>Mobile phone (47.52, 1.0E-4)</td>
</tr>
<tr>
<td>2</td>
<td>65</td>
<td>0.827</td>
<td>1994</td>
<td>(17.48) natural environment</td>
<td>Natural feature tracking (57.72, 1.0E-4)</td>
</tr>
<tr>
<td>3</td>
<td>56</td>
<td>0.89</td>
<td>2004</td>
<td>(17.33) liver</td>
<td>Laparoscopic surgery (30.43, 1.0E-4)</td>
</tr>
<tr>
<td>4</td>
<td>50</td>
<td>0.943</td>
<td>2011</td>
<td>(19.32) education</td>
<td>Education (64.26, 1.0E-4)</td>
</tr>
<tr>
<td>5</td>
<td>48</td>
<td>0.86</td>
<td>2007</td>
<td>(15.96) virtual city environment</td>
<td>Virtual city environment (32.68, 1.0E-4)</td>
</tr>
<tr>
<td>6</td>
<td>20</td>
<td>0.997</td>
<td>1989</td>
<td>(21.63) knowledge-based augmented reality</td>
<td>Knowledge-based augmented reality (250.67, 1.0E-4)</td>
</tr>
<tr>
<td>7</td>
<td>19</td>
<td>0.926</td>
<td>1992</td>
<td>(19.32) hand-eye calibration</td>
<td>Hand-eye calibration (104.98, 1.0E-4)</td>
</tr>
</tbody>
</table>

---

1https://www.leapmotion.com/
Microsoft Kinect, for example. Such technology allows one to track the users’ bodies, allowing them to interact with the virtual environments using body movements, gestures, and interactions. Most HMDs use an embedded system to track HMD position and rotation as well as controllers that are generally placed into the user's hands. This tracking allows a great degree of interaction and improves the overall virtual experience.

A final emerging approach is the use of digital technologies to simulate not only the external world but also the internal bodily signals (Azevedo et al., 2017; Riva et al., 2017): interoception, proprioception and vestibular input. For example, Riva et al. (2017) recently introduced the concept of “sonoception” (www.sonoception.com), a novel non-invasive technological paradigm based on wearable acoustic and vibrotactile transducers able to alter internal bodily signals. This approach allowed the development of an interoceptive stimulator that is both able to assess interoceptive time perception in clinical patients (Di Lernia et al., 2018b) and to enhance heart rate variability (the short-term vagally mediated component—rMSSD) through the modulation of the subjects’ parasympathetic system (Di Lernia et al., 2018a).

In this scenario, it is clear that the future of VR and AR research is not just in clinical applications, although the implications for the patients are huge. The continuous development of VR and AR technologies is the result of research in computer science, engineering, and allied sciences. The reasons for which from our analyses emerged a “clinical era” are threefold. First, all clinical research on VR and AR includes also technological developments, and new technological discoveries are being published in clinical or technological journals but with clinical samples as main subject. As noted in our research, main journals that publish numerous articles on technological developments tested with both healthy and patients include Presence: Teleoperators & Virtual Environments, Cyberpsychology & Behavior (Cyberpsychol BEHAV), and IEEE Computer Graphics and Applications (IEEE Comput Graph). It is clear that researchers in psychology, neuroscience, medicine, and behavioral sciences in general have been investigating whether the technological developments of VR and AR are effective for users, indicating that clinical behavioral research has been incorporating large parts of computer science and engineering. A second aspect to consider is the industrial development. In fact, once a new technology is envisioned and created it goes for a patent application. Once the patent is sent for registration the new technology may be made available for the market, and eventually for journal submission and publication. Moreover, most VR and AR research that proposes the development of a technology moves directly from the presenting prototype to receiving the patent and introducing it to the market without...
publishing the findings in scientific paper. Hence, it is clear that if a new technology has been developed for industrial market or consumer, but not for clinical purpose, the research conducted to develop such technology may never be published in a scientific paper. Although our manuscript considered published researches, we have to acknowledge the existence of several researches that have not been published at all. The third reason for which our analyses highlighted a “clinical era” is that several articles on VR and AR have been considered within the Web of Knowledge database, that is our source of references. In this article, we referred to “research” as the one in the database considered. Of course, this is a limitation of our study, since there are several other databases that are of big value in the scientific community, such as IEEE Xplore Digital Library, ACM Digital Library, and many others. Generally, the most important articles in journals published in these databases are also included in the Web of Knowledge database; hence, we are convinced that our study considered the top-level publications in computer science or engineering. Accordingly, we believe that this limitation can be overcome by considering the large number of articles referenced in our research.

Considering all these aspects, it is clear that clinical applications, behavioral aspects, and technological developments in VR and AR research are parts of a more complex situation compared to the old platforms used before the huge diffusion of HMD and solutions. We think that this work might provide a clearer vision for stakeholders, providing evidence of the current research frontiers and the challenges that are expected in the future, highlighting all the connections and implications of the research in several fields, such as clinical, behavioral, industrial, entertainment, educational, and many others.

AUTHOR CONTRIBUTIONS

PC and GR conceived the idea. PC made data extraction and the computational analyses and wrote the first draft of the article. IG revised the introduction adding important information for the article. PC, IG, MR, and GR revised the article and approved the last version of the article after important input to the article rationale.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2018.02086/full#supplementary-material

REFERENCES


Sullivan, D’Fusion Supports iPhone4S and 3DMax 2012.


Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer GC declared a shared affiliation, with no collaboration, with the authors PC and GR to the handling Editor at the time of the review.

Copyright © 2018 Cipresso, Giglioli, Raya and Riva. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.