

History in the Age of Abundance:

Skills, Tools, and Methods
for the 21st-Century
Historian

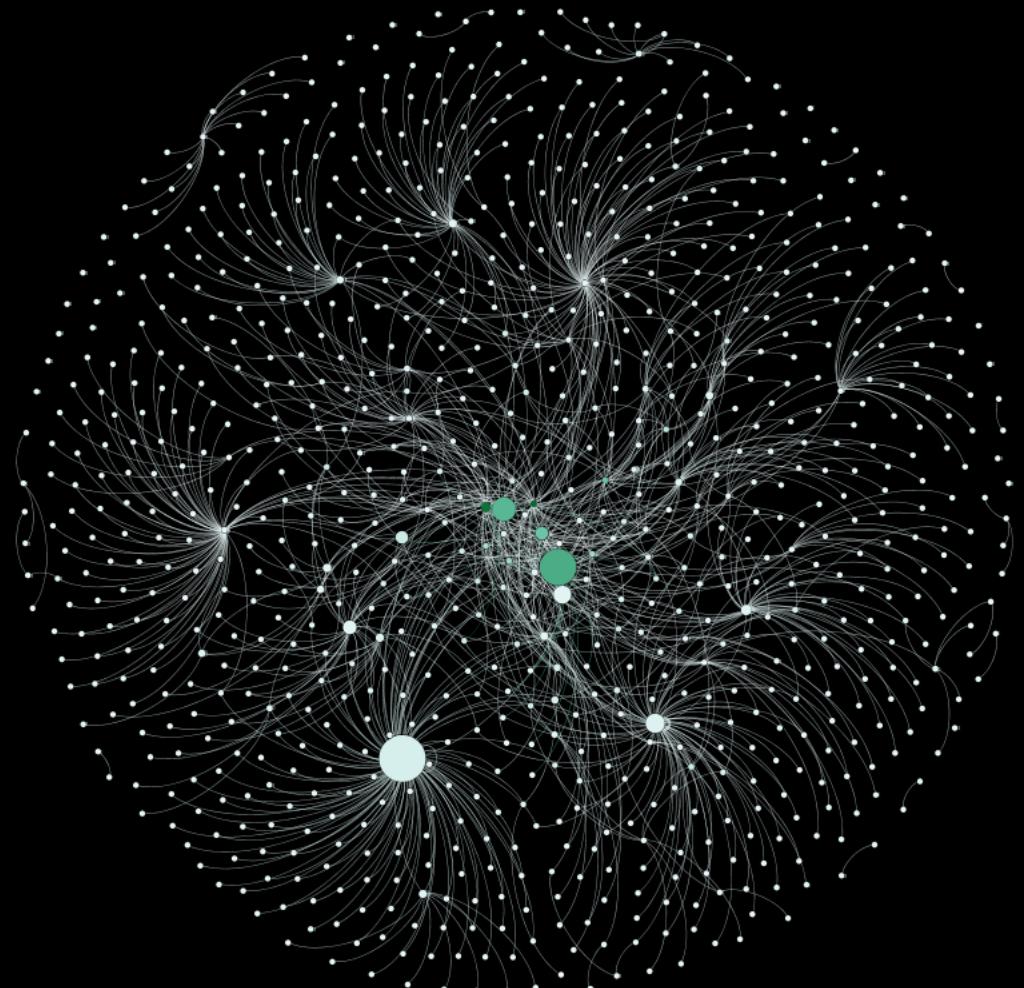
Ian Milligan

Associate Professor

Department of History



UNIVERSITY OF
WATERLOO

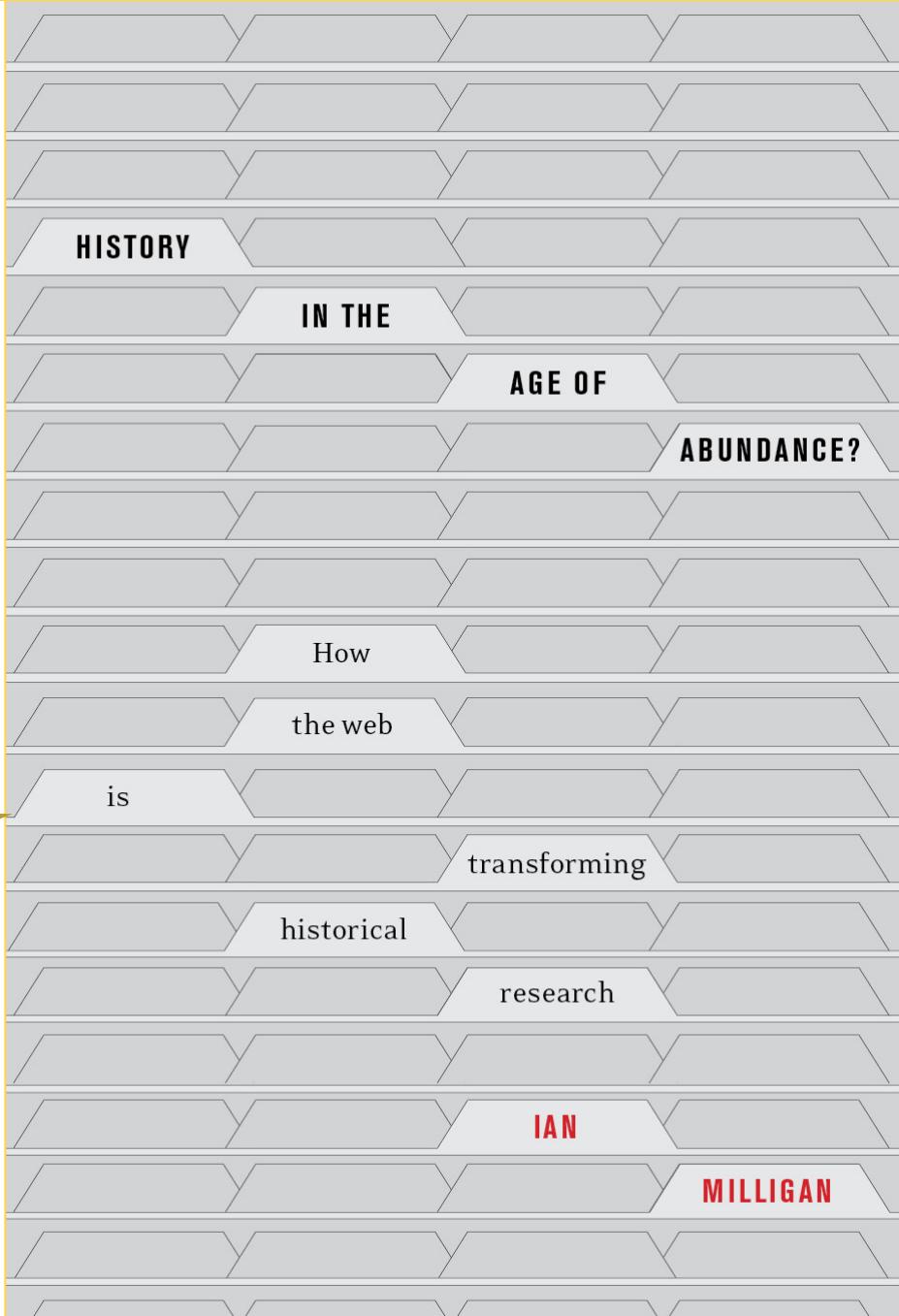


SLIDES PDF AVAILABLE AT

<https://www.ianmilligan.ca/talk/york-hist/>

Shameless Plug!

- ***History in the Age of Abundance? How the Web is Transforming Historical Research.*** McGill-Queen's University Press, 2019.
- \$28 on Amazon, or there's a copy at YUL or a few at Toronto Public Library



PLAN FOR THE TALK

- The Problem of TMI
- The Broad Solution to this Problem
- The Current Status of this Problem
- What to do through three personas
 - Computational Social Scientist/Humanist
 - Digital Social Scientist/Humanist
 - Social Scientist/Humanist
- Conclusions around history in the 21st century
- Questions

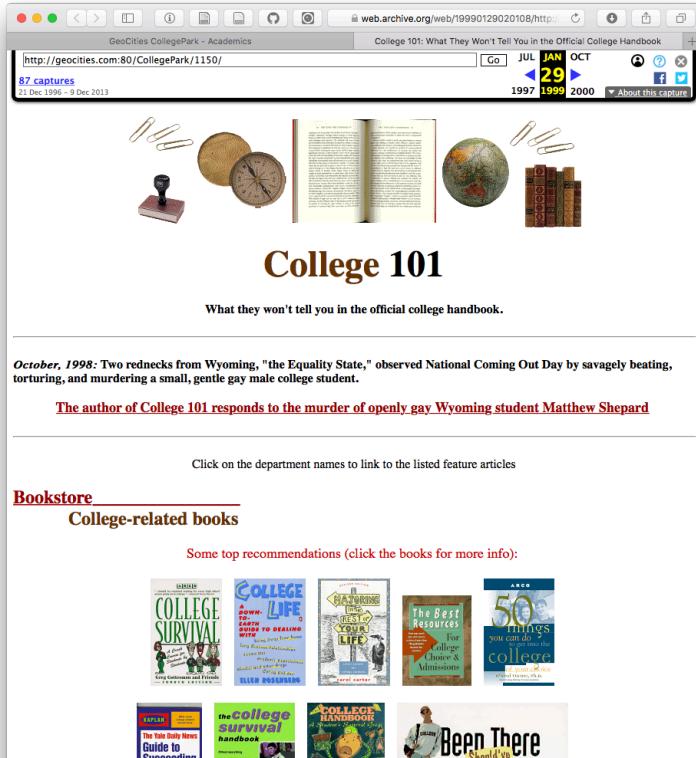


WHAT'S THE PROBLEM OF HISTORY IN THE “AGE OF ABUNDANCE”?

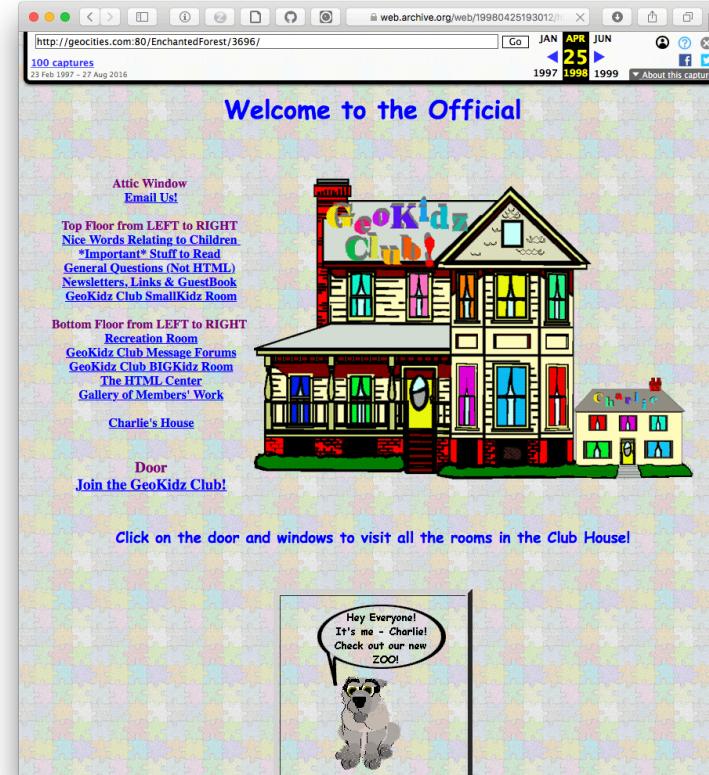
THE PROBLEM



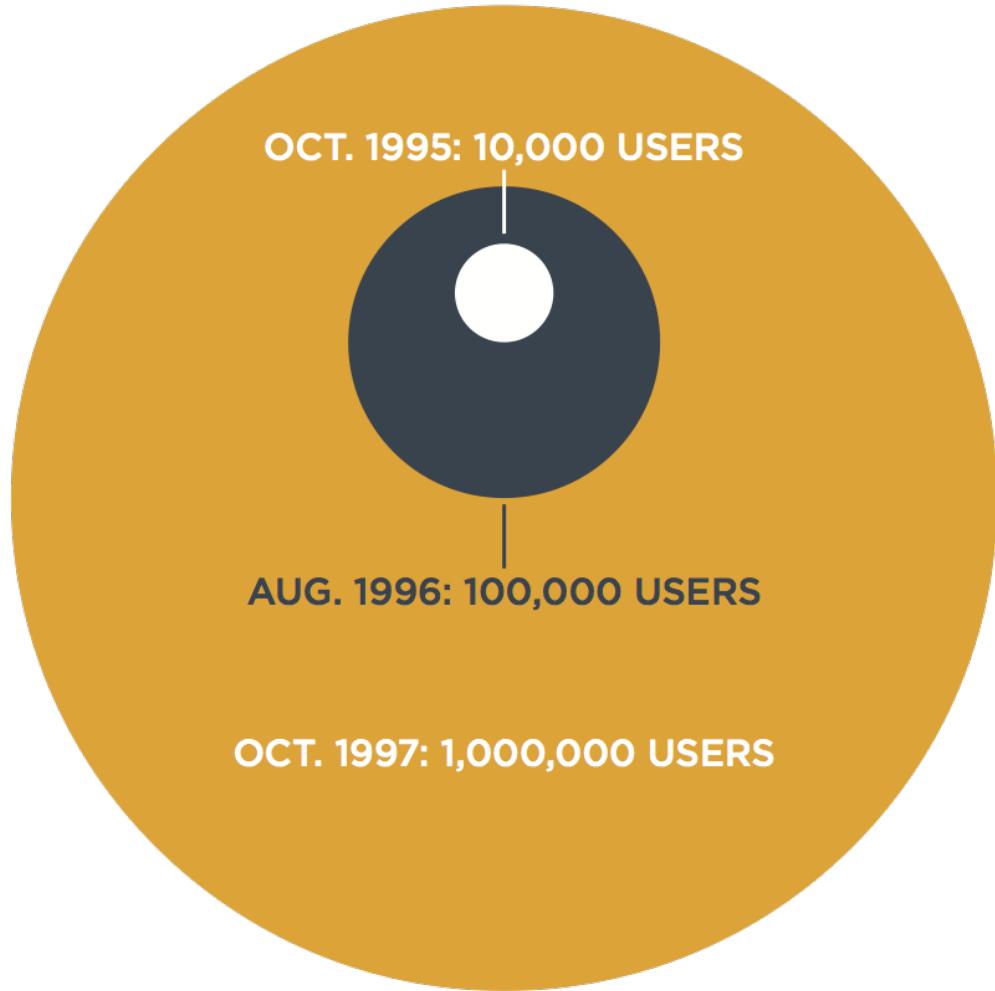
THE PROBLEM



GeoCities.com



THE PROBLEM



In other words...

**WEB ARCHIVES REPRESENT A NEW MEANS
OF KNOWLEDGE DISTRIBUTION AND
ACQUISITION**

A PERSONAL EXAMPLE OF OUR CHANGING HISTORICAL RECORD

rec.games.miniatures.warhammer >
Civil War in Warhammer 40,000
7 posts by 7 authors 8+1

★ Nyarlathotep Who can say, but it certainly sounds amusing. Later days, Nyar. 12/7/95

Peter Milligan 12/7/95 

★ - hide quoted text -
I was recently playing a game of Warhammer 40,000 when the opponents had a disagreement and decided that they would have a civil war. The two armies in question were Space Wolf, Blood Angels and Imperial Guard. The Space Wolf player opened up with his Cyclone and using 12 missiles, wiped out the 1500 point Imperial army, and the survivors were picked off by the Blood Angel's thud gun.

Although I was delighted by this, it made for a short game, as the only model I got to use was my Leman Russ, who picked off the survivors after the Blood angels picked off the Space Wolves.

Do you think Civil War is an acceptable rule?

★ Bub Well If you're in another universe why not. But in terms of 40K no :) -- Bub 7262...@compuserve.com 12/8/95

★ Joel E Slovacek once in a while, so long as the people fighting the war understand that they will lose (since they have 12/8/95

★ Chad Lubrecht What is this civil war rule? I seem to have missed the original posting, sorry! 12/8/95

A PERSONAL EXAMPLE OF OUR CHANGING HISTORICAL RECORD

- **First**, Don't Read the Post;
- **Secondly**, my discomfort is perhaps in and of itself significant;
- **Third**, the musings of an eleven-year old child are now available..
- **Fourth** – in many countries today, such musings would be subject to legal deposit.

rec.games.miniatures.warhammer >
Civil War in Warhammer 40,000
7 posts by 7 authors 8+1 12/7/95

★ Nyarlathotep Who can say, but it certainly sounds amusing. Later days, Nyar. 12/7/95

Peter Milligan 12/7/95

- hide quoted text -
I was recently playing a game of Warhammer 40,000 when the opponents had a disagreement and decided that they would have a civil war. The two armies in question were Space Wolf, Blood Angels and Imperial Guard. The Space Wolf player opened up with his Cyclone and using 12 missiles, wiped out the 1500 point Imperial army, and the survivors were picked off by the Blood Angel's thud gun.
Although I was delighted by this, it made for a short game, as the only model I got to use was my Leman Russ, who picked off the survivors after the Blood angels picked off the Space Wolves.
Do you think Civil War is an acceptable rule?

Bub Well If you're in another universe why not. But in terms of 40K no :) – Bub 7262...@compuserve.com 12/8/95

Joel E Slovacek once in a while, so long as the people fighting the war understand that they will lose (since they have 12/8/95

Chad Lubrecht What is this civil war rule? I seem to have missed the original posting, sorry! 12/8/95

 UNIVERSITY OF
WATERLOO

**THIS IS NOT A NICHE CONCERN FOR ANY
POST-1990 TOPIC**

WHAT OPPORTUNITIES DO THEY PRESENT?

- The way that we preserve our culture is changing;
 - **Scale:** Internet Archive has 635 billion+ URLs; 40PB of unique data (and non-Internet Archive collectors probably have about the same again).
 - **Scope:** Data that never before would have been collected is now being collected about people who aren't traditionally in the historical record.
- Any researcher tackling post-1996 topics will realistically need to understand the vast arrays of text, image, etc. that comprise our modern cultural record.
- The **Wayback Machine** isn't enough; will need to explore data at scale.

**WHO HERE HAS USED THE INTERNET
ARCHIVE?**

THE WAYBACK MACHINE IN ACTION

The screenshot shows the Internet Archive Wayback Machine homepage. At the top, there's a navigation bar with icons for file operations (New, Open, Save, Print, Copy, Paste, Find, Help), a search bar containing "archive.org/web/", and download/upload buttons. Below the bar is a dark header with the Internet Archive logo (a classical building icon) and a user profile for "ianmillig...". The main content area features the Wayback Machine logo ("INTERNET ARCHIVE WayBack Machine") and a search bar with the placeholder "http://". A "BROWSE HISTORY" button is also present. Below this, a message encourages users to "Find the Wayback Machine useful?" followed by a "DONATE" button. A row of five thumbnail images shows different web pages from the archive. In the bottom right corner, there's a "Tools" section with links to the "Wayback Machine Availability API" and "WordPress Broken Link Checker". The bottom left corner contains the text "HISTORY IN THE AGE OF ABUN".

archive.org/web/

ABOUT CONTACT BLOG PROJECTS HELP DONATE JOBS VOLUNTEER PEOPLE

INTERNET ARCHIVE
WayBack Machine

Explore more than 362 billion [web pages](#) saved over time

http://

BROWSE HISTORY

Find the Wayback Machine useful? [DONATE](#)

Tools

Wayback Machine Availability API
Build your own tools.

WordPress Broken Link Checker
Banish broken links from your blog.

404 Handler for Webmasters

HISTORY IN THE AGE OF ABUN



UNIVERSITY OF
WATERLOO

DOWNSIDES OF THE WAYBACK MACHINE

- **Wayback Machine** is great if you know what you're looking for;
 - Ever-improving keyword search functionality
 - Represents a great stride in accessibility more generally
- But it isn't great for more detailed research queries:
 - You may want to do complicated queries (i.e. websites that say X and link to Y);
 - You may want to do exploratory text mining;
 - You may want to work with images en masse;
 - Etc.

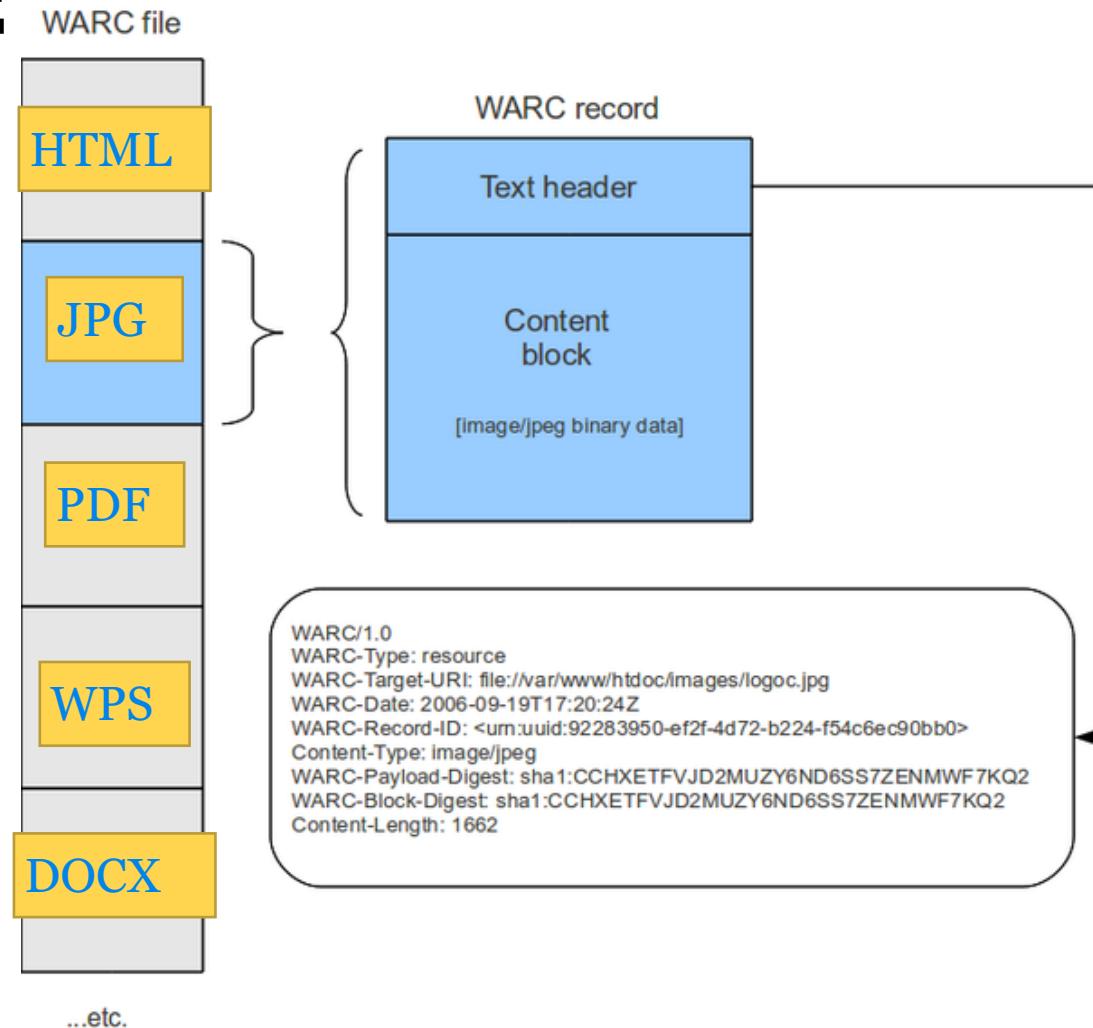


**SO INSTEAD, WE'LL NEED TO WORK WITH
DATA AT SCALE.**

What does this entail?

WORKING WITH DATA AT SCALE

- The **WARC** (ISO 28500:2009) file
- Pictured at right
- Hard to use and a bit idiosyncratic, with a smaller user base, so the first step is to usually transform the data into something that's a bit more common!

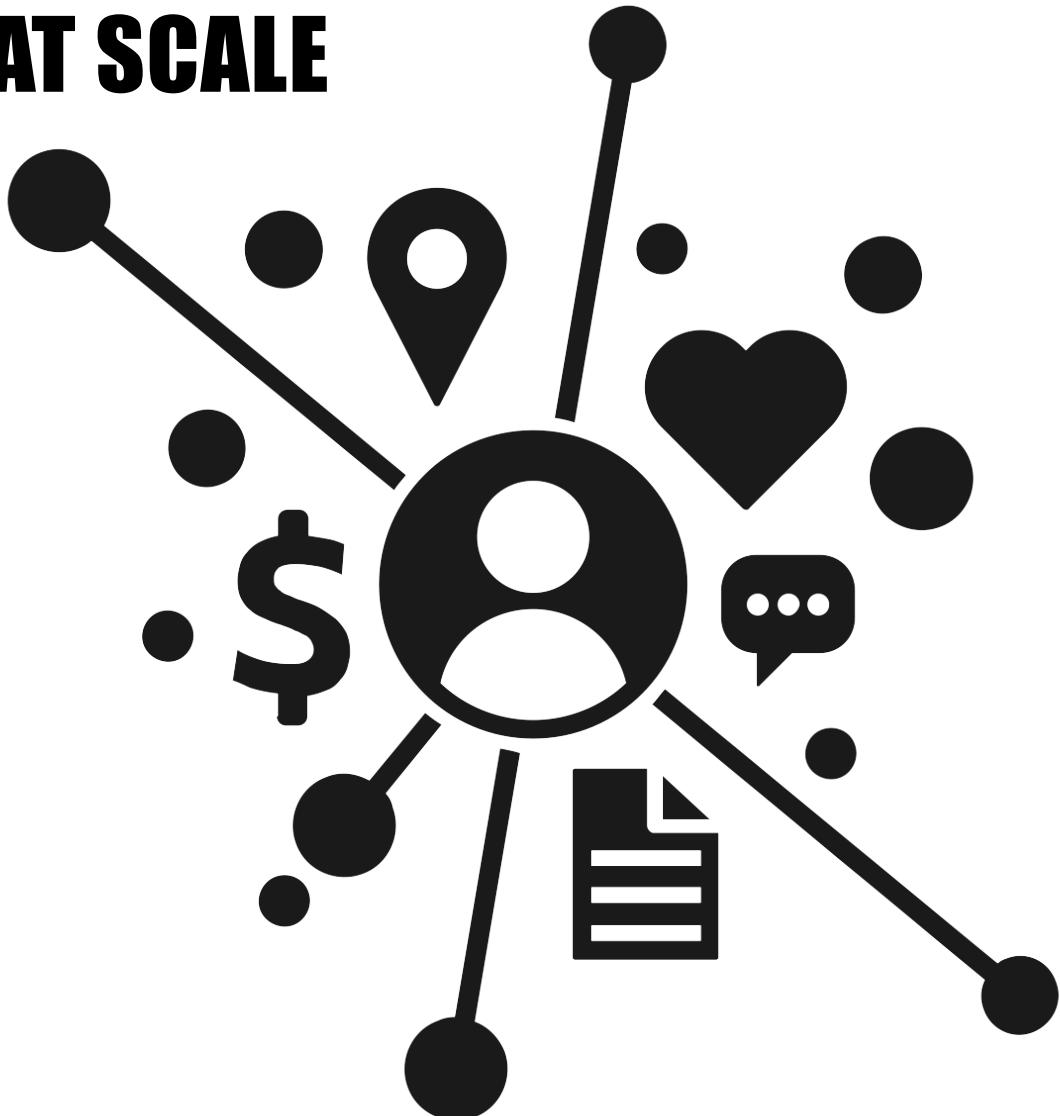


**YOUR COMPREHENSIVE EXAMS PREPPED
YOU FOR WORKING WITH DATA, RIGHT?**

Uh oh... what kind of skills are you going to need?

SKILLSET ONE: WORKING WITH DATA AT SCALE

- An understanding of:
 - The Basics of **Natural Language Processing** (NLP)
 - **Basic Statistical Knowledge** to work with Quantitative Data (frequency of websites, terms, normalizing numbers, etc.)
 - Flexible **Data Science** skills (or StackOverflow skills..)
- In other words, being equipped with the skills and capacities to analyze text/data at scale.



SKILLSET TWO: UNDERSTANDING HOW DATA IS CONSTRUCTED

- They also need to have a solid understanding of:
 - **How and why the data was collected**, i.e. selection criteria;
 - What data **wasn't collected**;
 - How the **software used to create the dataset** has changed over time;
- How to **clean or normalize data** when necessary (i.e. as a URL changes from <http://www.ndp.ca> to <https://www.ndp.ca> to <http://ndp.ca> in the crawl, recognizing that those are probably the same website! You see that, the computer doesn't unless you teach it to).



OH YEAH...

**PLUS THE NORMAL SKILLS OF THE
HISTORIAN TOO..**

SO LET'S TAKE STOCK

- **Historians will need to understand and study the Web** in order to come to grips of history after the mid-1990s – not just for history of the Web, of course, but for the history of our society and culture as reflected on the Web
- **Existing tools like the Wayback Machine aren't enough** to tackle this problem
- **Historians will need new skills** for working with and understanding data, plus their traditional competencies



ARE HISTORIANS READY?

(Leading question alert!)

NO...

(Sad trombone)

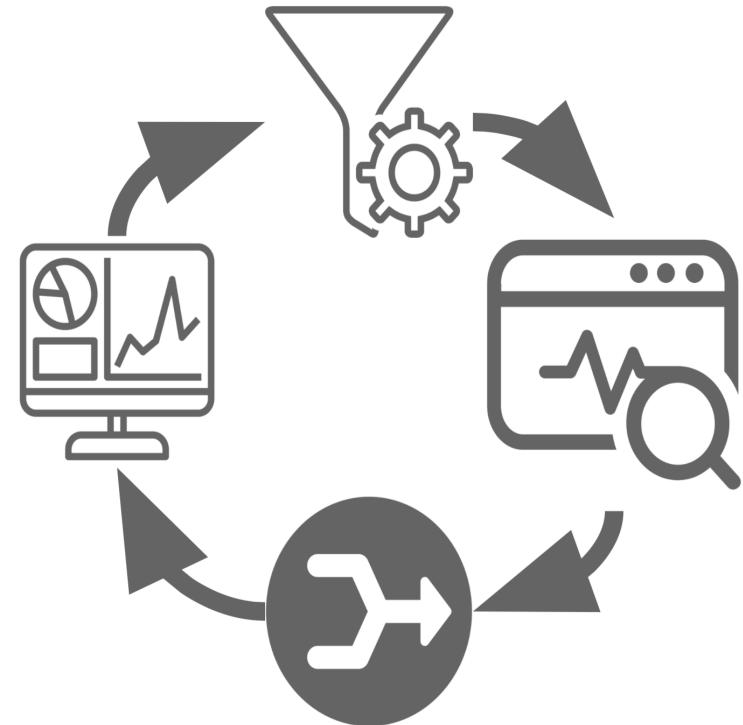
FIRST OF ALL, “WE” CAN’T DO IT ALONE

- I’ll return to this point, but when I use “we” it’s not the **royal we**; it’s the Archives Unleashed “we”
 - Nick Ruest (York University Libraries)
 - Jimmy Lin (Waterloo Computer Science)
 - + undergrads in CS; grads in CS and History
- I’ll return to some smiling mugshots in a few



THE CURRENT STATUS OF WORKING WITH WEB ARCHIVES AT SCALE?

- We have developed something called the **Archives Unleashed Toolkit**, based on working with scholars at seven datathons + numerous workshops
- It allows scholar to use the **FAAV** cycle to explore it:
 - **Filter**
 - **Analyze**
 - **Aggregate**
 - **Visualize**
- Allows for complicated research queries using Apache Spark



PROBLEM SOLVED, RIGHT?

Check out our cutting-edge interface....

THE ARCHIVES UNLEASHED TOOLKIT

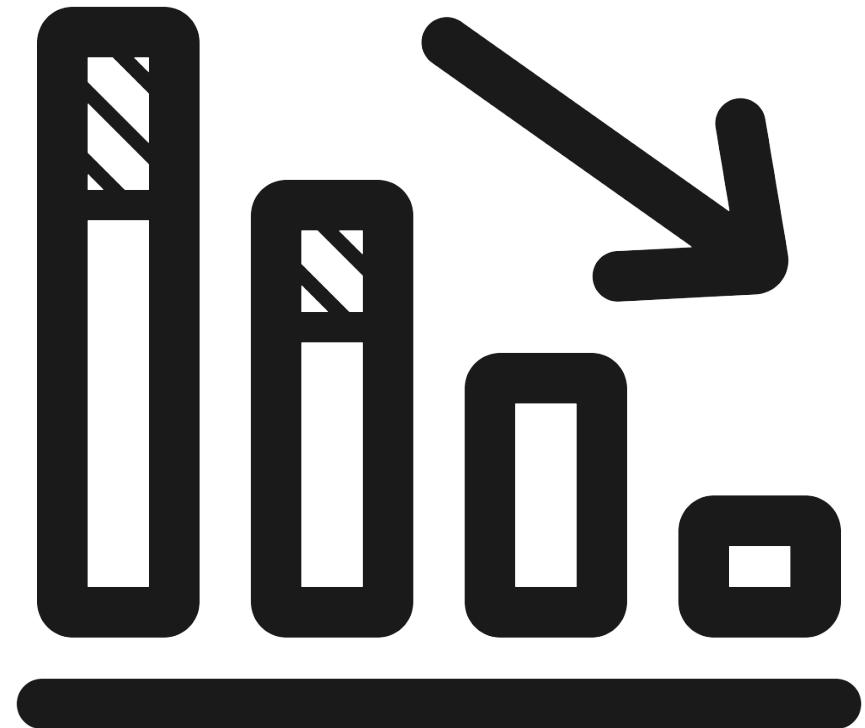
- It's easy to use, as long as you:
 - Know how to use the command line;
 - How to access a server;
 - How to use the Spark Shell;
 - How to code, at least somewhat, in Scala;
 - And, have a lot of patience for open-source documentation!

(OK, it's not easy to use..)

```
import io.archivesunleashed._ import  
io.archivesunleashed.matchbox._  
RecordLoader.loadArchives("example.a  
rc.gz", sc) .keepValidPages()  
.keepDate(List("200804"),  
ExtractDate.DateComponent.YYYYMM)  
.map(r => (r.getCrawlDate,  
r.getDomain, r.getUrl,  
RemoveHTML(r.getContentString)))  
.saveAsTextFile("plain-text-date-  
filtered-200804/")
```

... AND HISTORIANS HAVE BEEN TURNING AWAY FROM QUANT TO QUAL

- Studies of our introductory historiography textbooks show this diminishing.
 - John Tosh, *Pursuit of History*
 - 1st, 2nd edition: “History by Numbers” (entire chapter)
 - By 5th edition, no quantitative history at all.
- **“Nevertheless, it is curious that at a time when both the use of and the breadth of humanities data is growing, quantitative skills ... seem to no longer form a core component of our undergraduate history programmes.”** (James Baker, <https://blogs.bl.uk/digital-scholarship/2014/04/digital-history-and-the-death-of-quant.html>)



WHERE TOOLS END AND USERS BEGIN

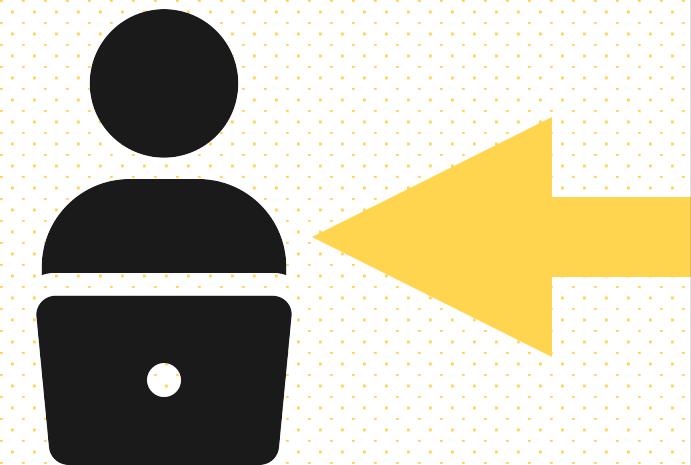


WHERE TOOLS END AND USERS BEGIN



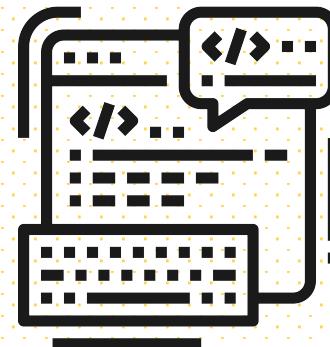
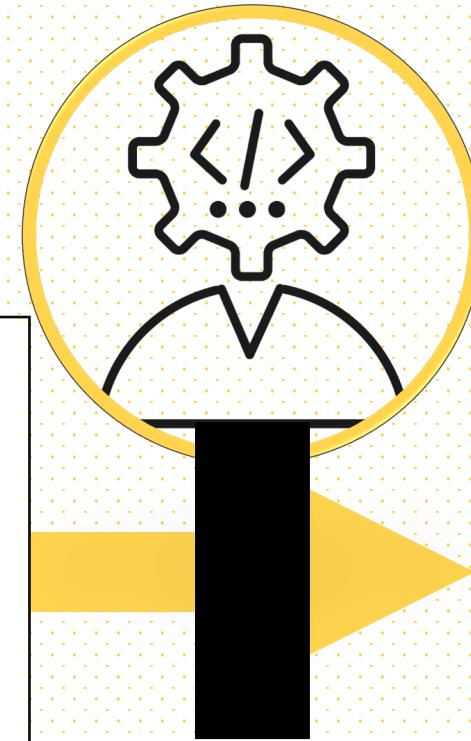
A now-dated cultural reference, thanks for killing the magic,
Game of Thrones Season 8.

WHERE TOOLS END AND USERS BEGIN



OPTION ONE: Historians become programmers, capable of advanced web archive analysis.

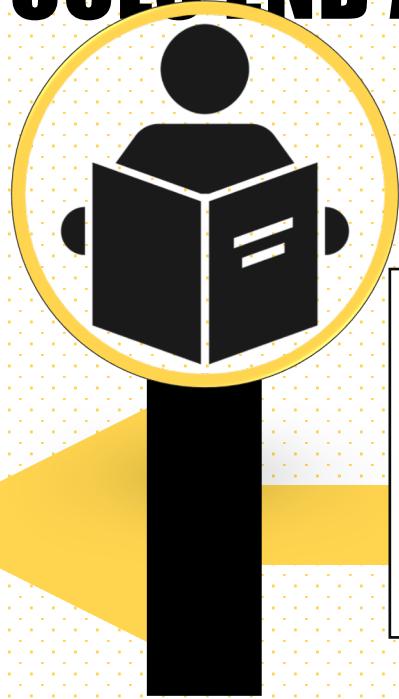
- Writing code;
- Contributing to open-source projects;
- Maybe need a bit of support but in general are self-sufficient in a computational environment.



Advanced
Web Archive Analysis

Historian with
Research Question

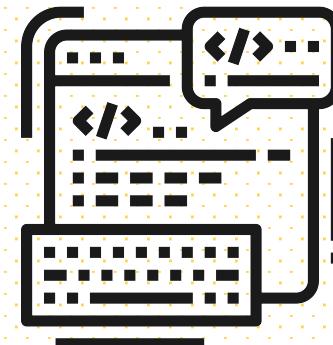
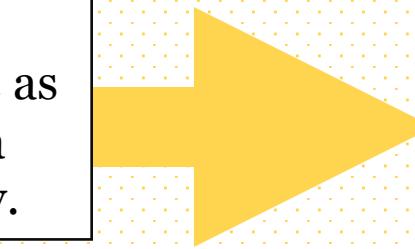
WHERE TOOLS END AND USERS BEGIN



Historian with
Research Question

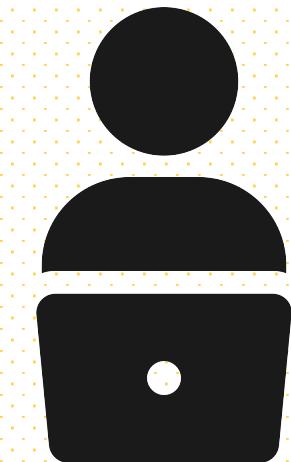
OPTION TWO: Historians do not develop technical skills, continuing solely as subject-matter experts.

- An approach that sees using a web archive as similar to using PROQUEST or JSTOR – a bit of work but nothing out of the ordinary.

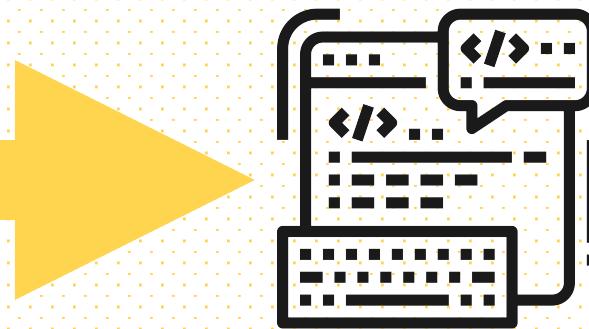


Advanced
Web Archive Analysis

WHERE TOOLS END AND USERS BEGIN



**Historian with
Research Question**



Advanced

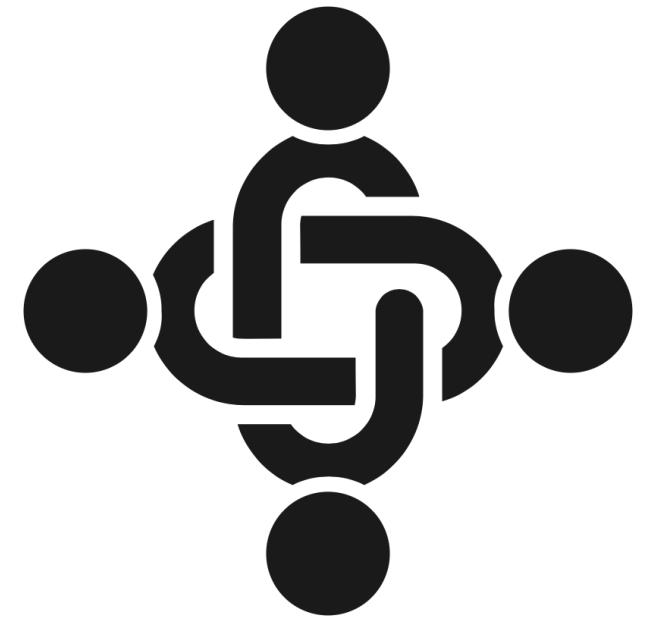
Web Archive Analysis

OPTION THREE: The middle ground

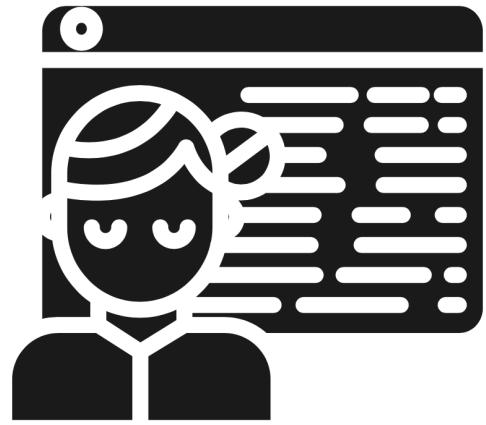
- Some computational skills, but platforms designed to resemble conventional research processes as much as possible
- Positioned in a world where you can draw on standard library research support (i.e. not web archiving particularly, but “working with text”)
- No command lines!

WHERE TOOLS END AND USERS BEGIN

- We can draw on a model of successful **interdisciplinary cooperation** to see how this can be effected
- If a computer scientist and a historian work together, it isn't just "historian: now you become a computer scientist;" nor is it "computer scientist: now you become a historian." Compromise is worked out in terms of:
 - Workflow (Platforms? Collaboration?)
 - Publication venues
 - The shape of the work
- It really is often meeting half way, and being conscious

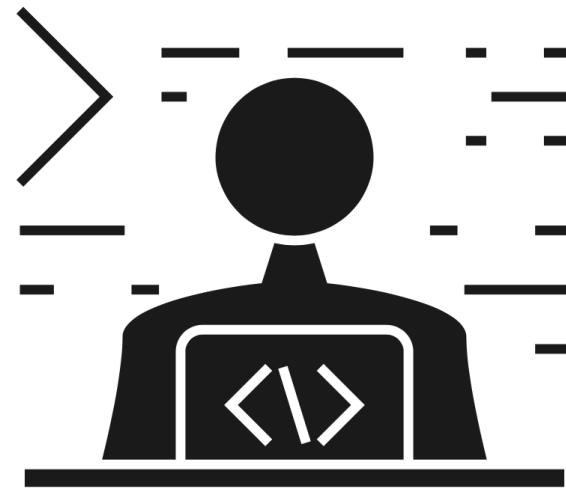


THREE PERSONAS



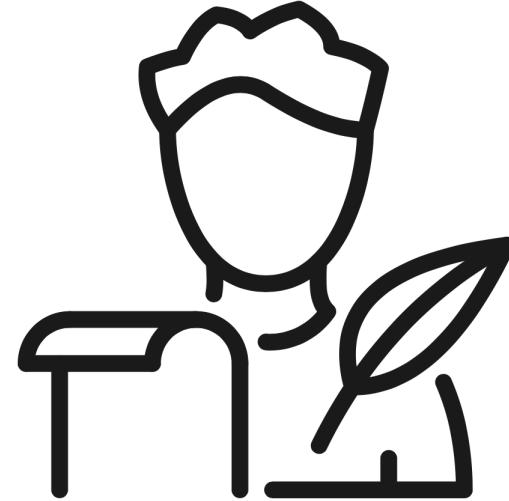
Person A

Computational
Humanist



Person B

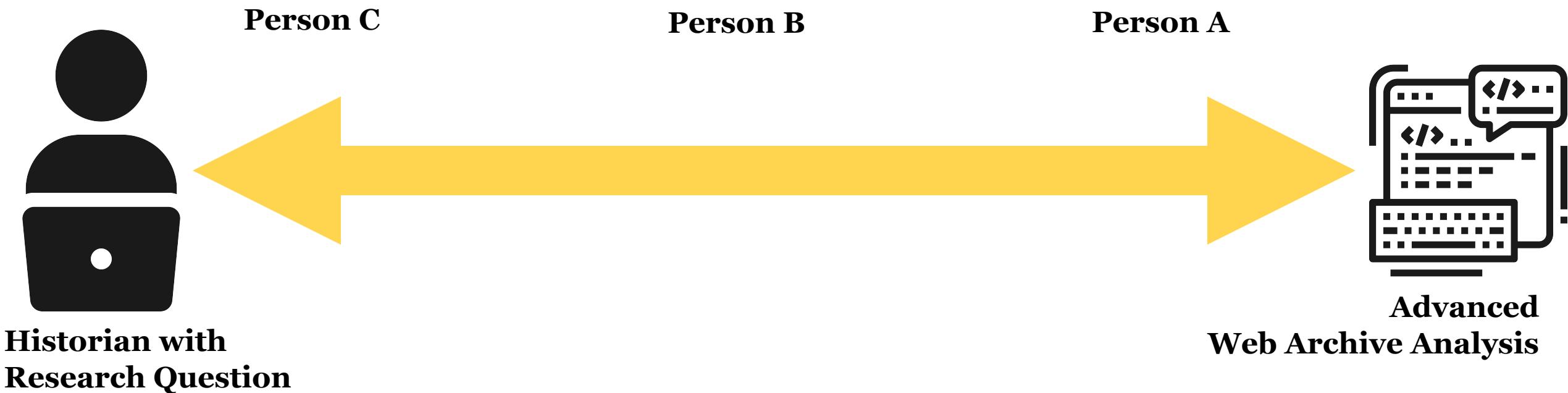
Digital Humanist



Person C

Conventional
Historian

IN OTHER WORDS...



**HOW CAN WE SUPPORT THIS
COMPUTATIONAL TURN FOR ALL THREE
PERSONAS?**

Enter our project...

Archives Unleashed



FIRST, WE NEED AN INTERDISCIPLINARY TEAM



Ian Milligan

Historian



Nick Ruest (YORK!)

Library/Archives



Jimmy Lin

Computer Science

WE CAN'T GO IT ALONE....!

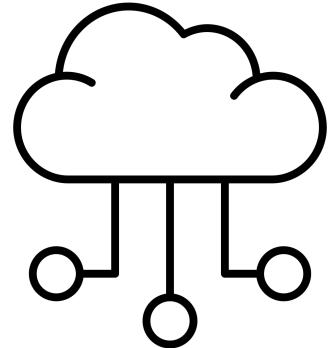
(So why would we expect our users to go it alone??)

AND A MISSION

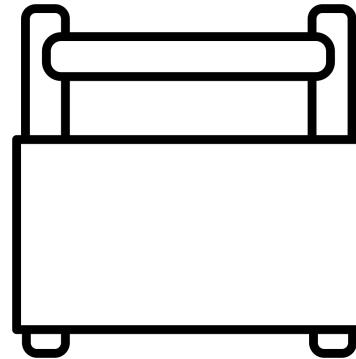
Archives Unleashed aims to make petabytes of historical internet content accessible to scholars and others interested in researching the recent past.

ARCHIVES UNLEASHED PROJECT

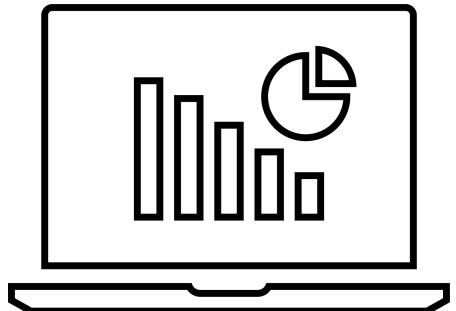
AND A SET OF TOOLS...



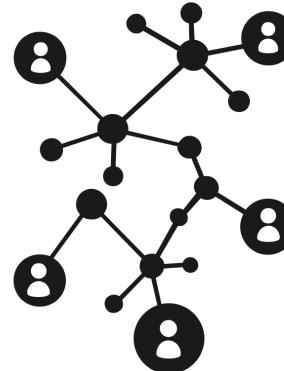
Cloud



Toolkit



Notebooks



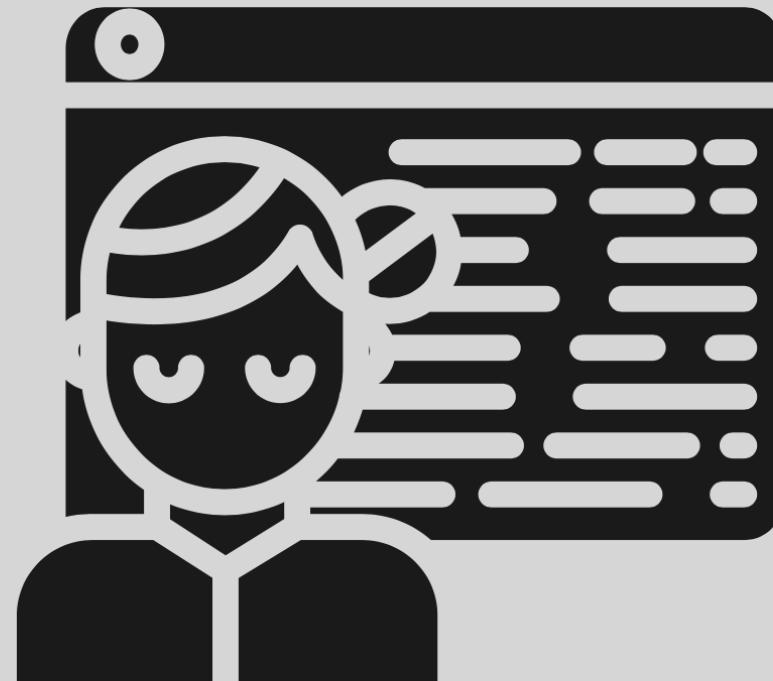
Datathons

PERSONA A

The sort of scholar we would have called a
“computational humanist” or
“computational social scientist.”

Comfortable installing packages,
understanding dependencies, fluent on
the command line, and can Stack
Overflow like a master.

THE COMPUTATIONAL HISTORIAN



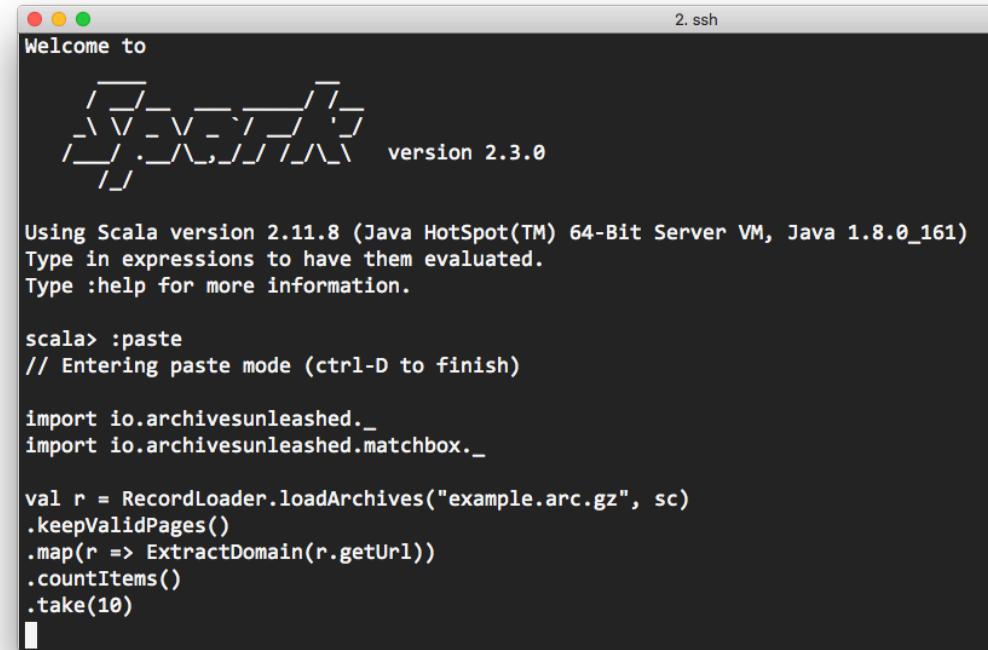
COMPUTATIONAL HUMANITIES

- Needs to have some sort of toolkit to translate WARC files into something they can work with
- Can interpret tested yet still dense documentation
- Can troubleshoot error messages (i.e. knows what to Google – this is a hard-earned skill)
- Can take sample scripts, execute them to see outcome, and then change them to run on their own data with own questions
- Can use the command line!

```
import io.archivesunleashed._ import  
io.archivesunleashed.matchbox._  
RecordLoader.loadArchives("example.arc.g  
z", sc) .keepValidPages()  
.keepDate(List("200804"),  
ExtractDate.DateComponent.YYYYMM) .map(r  
=> (r.getCrawlDate, r.getDomain,  
r.getUrl,  
RemoveHTML(r.getContentString)))  
.saveAsTextFile("plain-text-date-  
filtered-200804/")
```

COMPUTATIONAL HUMANITIES

- The **Archives Unleashed Toolkit** was designed around this persona.
- Allows a user to take WARCs and:
 - Determine elemental statistics about a collection;
 - Extract particular images, domains, URLs, pages with keywords, etc.
 - Do sophisticated Apache Spark-powered network analysis; and
 - Write custom Scala scripts to do almost anything you imagine with our set of custom web archiving User Defined Functions



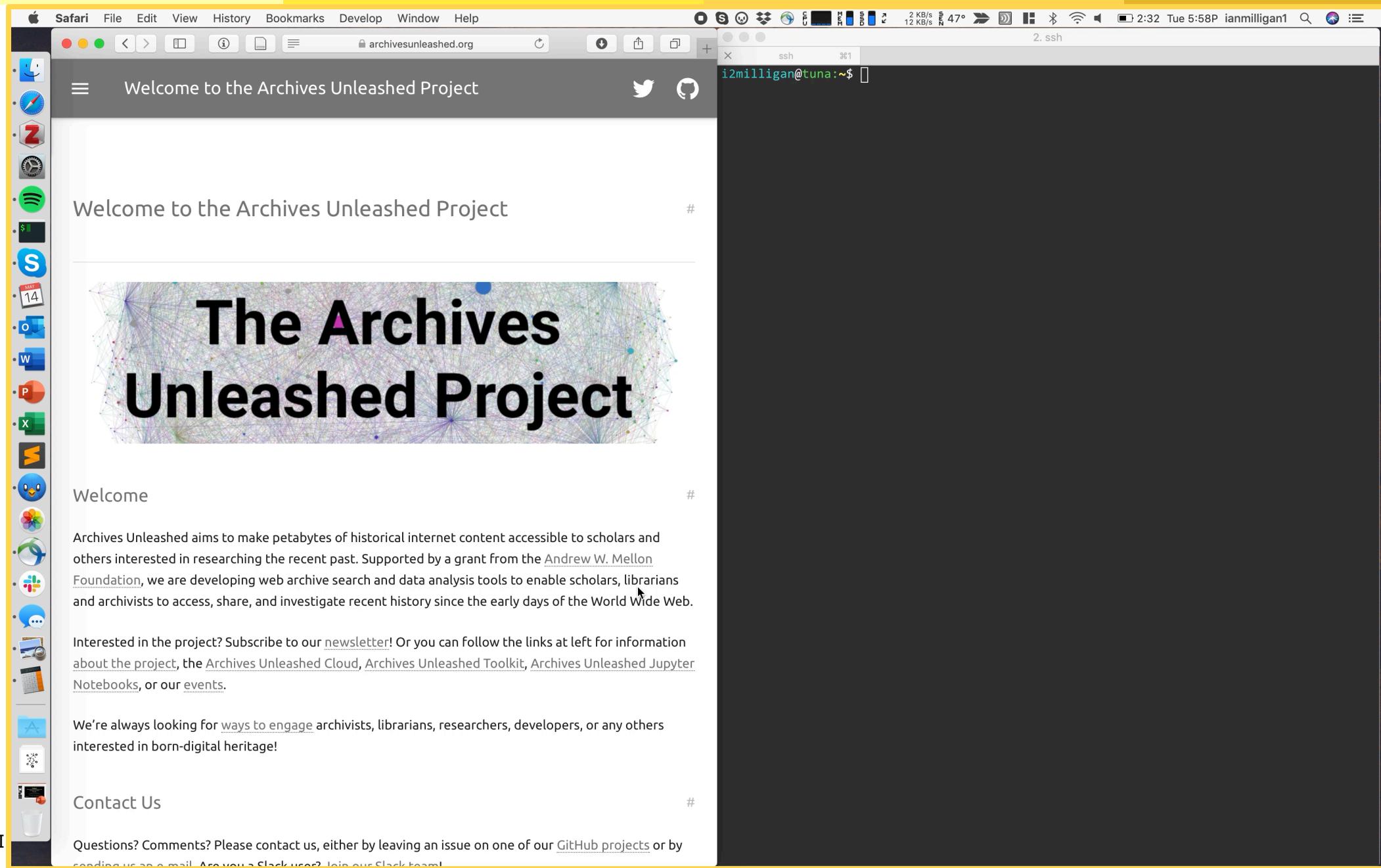
The screenshot shows a terminal window titled "Welcome to" with a decorative logo consisting of various symbols like brackets and slashes. The title bar also shows "2. ssh". The Scala shell version is listed as "version 2.3.0". Below the title, it says "Using Scala version 2.11.8 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_161)". It provides instructions: "Type in expressions to have them evaluated." and "Type :help for more information.". A Scala command is being typed in the shell:

```
scala> :paste
// Entering paste mode (ctrl-D to finish)

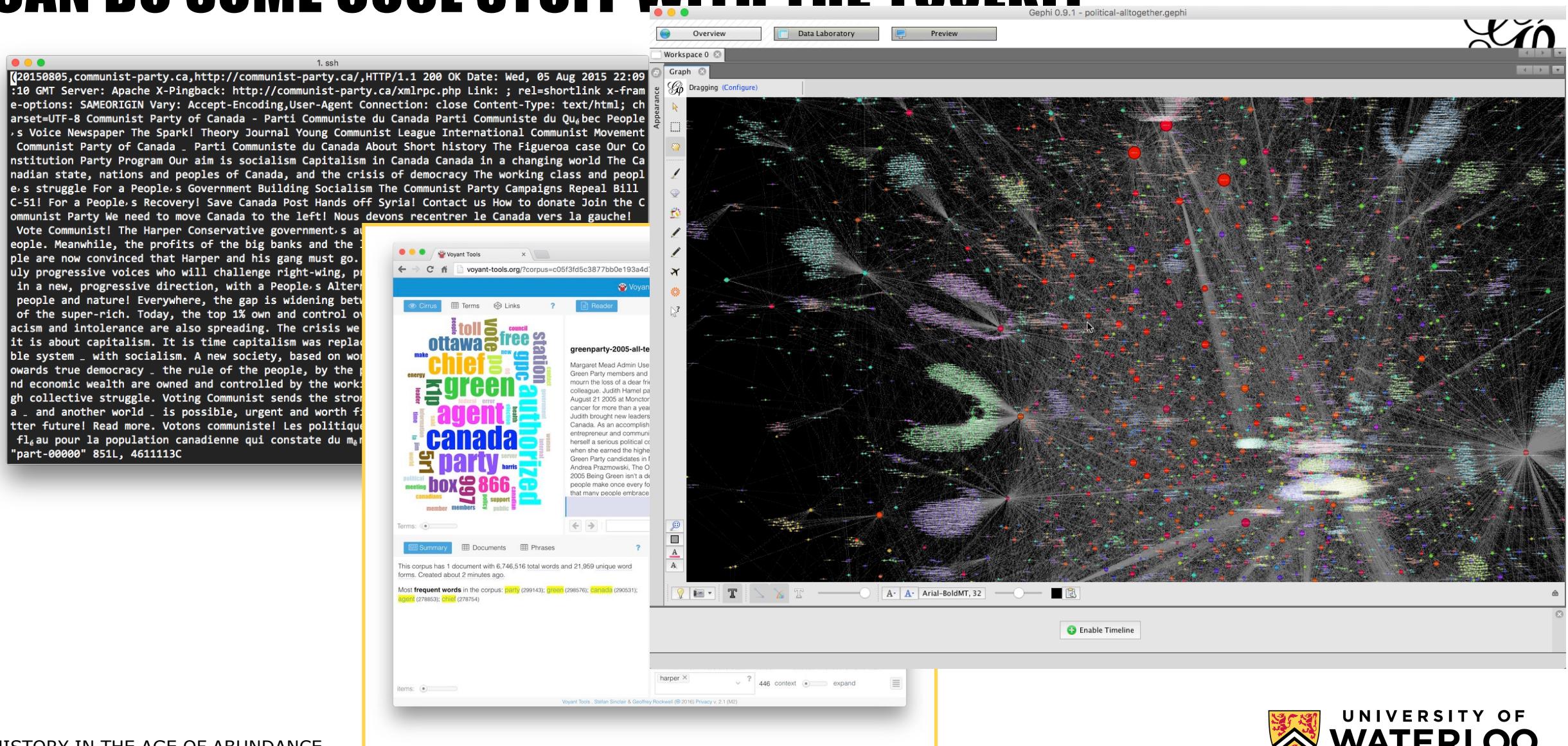
import io.archivesunleashed._
import io.archivesunleashed.matchbox._

val r = RecordLoader.loadArchives("example.arc.gz", sc)
  .keepValidPages()
  .map(r => ExtractDomain(r.getUrl()))
  .countItems()
  .take(10)
```

LET'S SEE THIS IN ACTION...



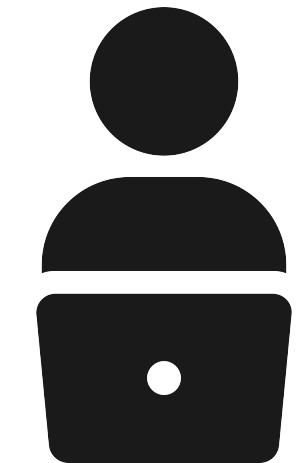
CAN DO SOME COOL STUFF WITH THE TOOLKIT



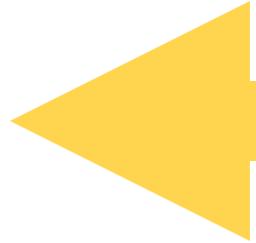


**WE MADE THE MISTAKE OF ASSUMING
THESE WERE OUR MAIN USERS...**

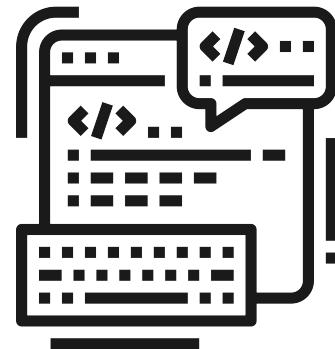
IN OTHER WORDS...



Historian with
Research Question



Person C



Advanced
Web Archive Analysis

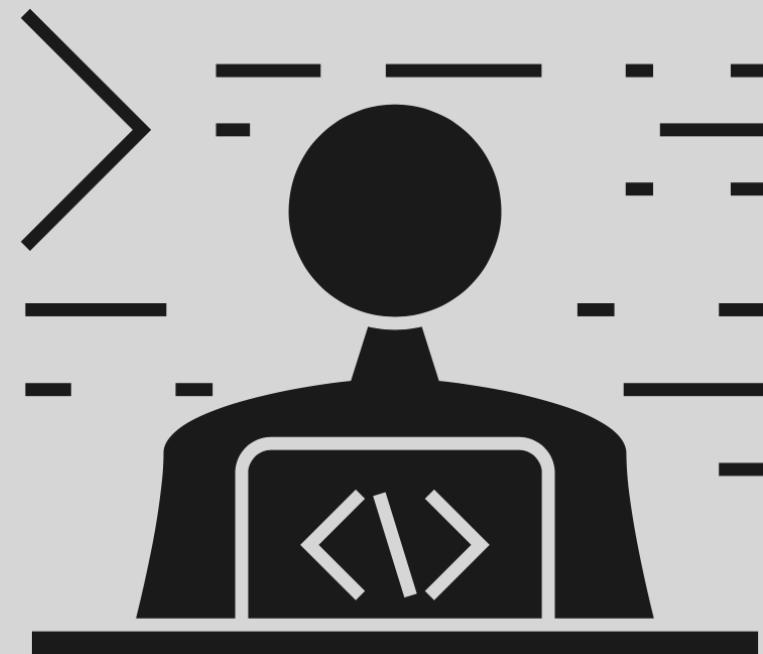
WE DIDN'T GO FAR ENOUGH... EXPECTED THEM TO COME TO US!

**ENTER THE ANDREW W. MELLON
FOUNDATION AND OUR GOAL OF MAKING
AN ACCESSIBLE TOOL**

PERSONA B

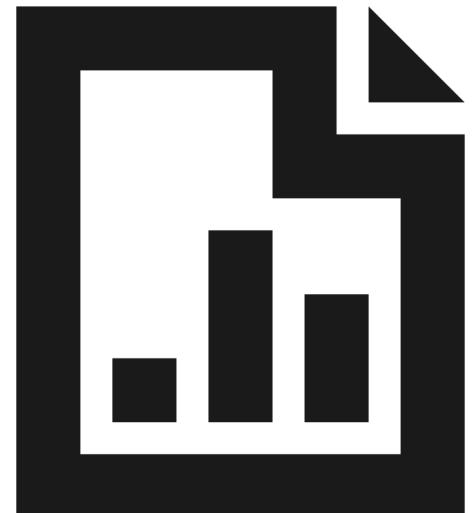
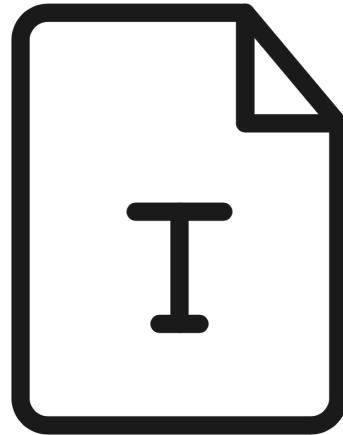
The sort of scholar we would have called a “digital humanist.” They’re comfortable with computers, can use some off-the-shelf tools like Voyant or Gephi to work with text/networks, and can use tutorials like those of the *Programming Historian* to learn some basic Python or R.

THE DIGITAL HISTORIAN



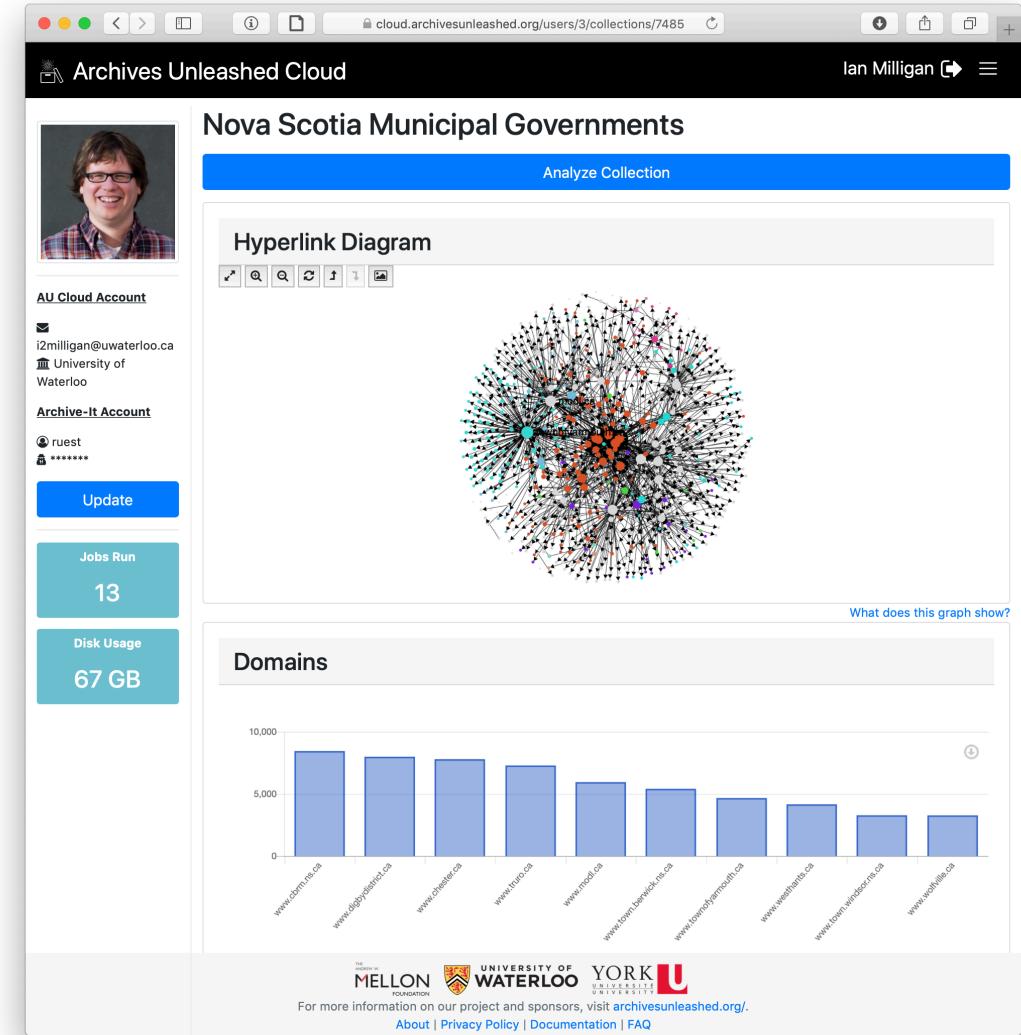
DIGITAL HUMANIST

- Needs to have some sort of toolkit to translate WARC files into something they can work with
- Can think critically about data
- Can use off-the-shelf tools to work with text/data, i.e. Voyant Tools or Gephi or other things based out of the *Programming Historian*
- In general doesn't want to use the command line or write custom programs



COMPUTATIONAL HUMANITIES

- The **Archives Unleashed Cloud** was designed around this persona.
- Allows a user to take WARCs and:
 - Use a modern UI to sync their collections from the provider;
 - Run basic analyses in the browser to find major sites of interest;
 - Download derivative file formats that can integrate with standard workflows.
- In other words: let's get the WARC out of the equation and *translate* it into a standard file format.



LET'S SEE THIS IN ACTION....

THE CLOUD IN ACTION

The screenshot shows a web browser window for the Archives Unleashed Cloud at cloud.archivesunleashed.org. The title bar reads "cloud.archivesunleashed.org". The main content area features a large, complex network graph on a black background. The graph consists of numerous small white dots (nodes) connected by thin white lines (edges), forming a dense, organic cluster. Three larger, semi-transparent teal dots are highlighted within this cluster, representing specific data points of interest. At the top of the page, there is a navigation bar with icons for back, forward, search, and file operations. To the right of the bar, the user's name "Ian Milligan" is displayed with a profile icon and a three-dot menu. A yellow banner at the top of the content area contains the text "New user? You currently need Archive-It credentials to use our service. Please read our FAQ." with a close button ("X"). The bottom of the page includes logos for Mellon Foundation, University of Waterloo, and York University, along with links to the project's About, Privacy Policy, Documentation, and FAQ pages.

HISTORY IN THE AGE OF

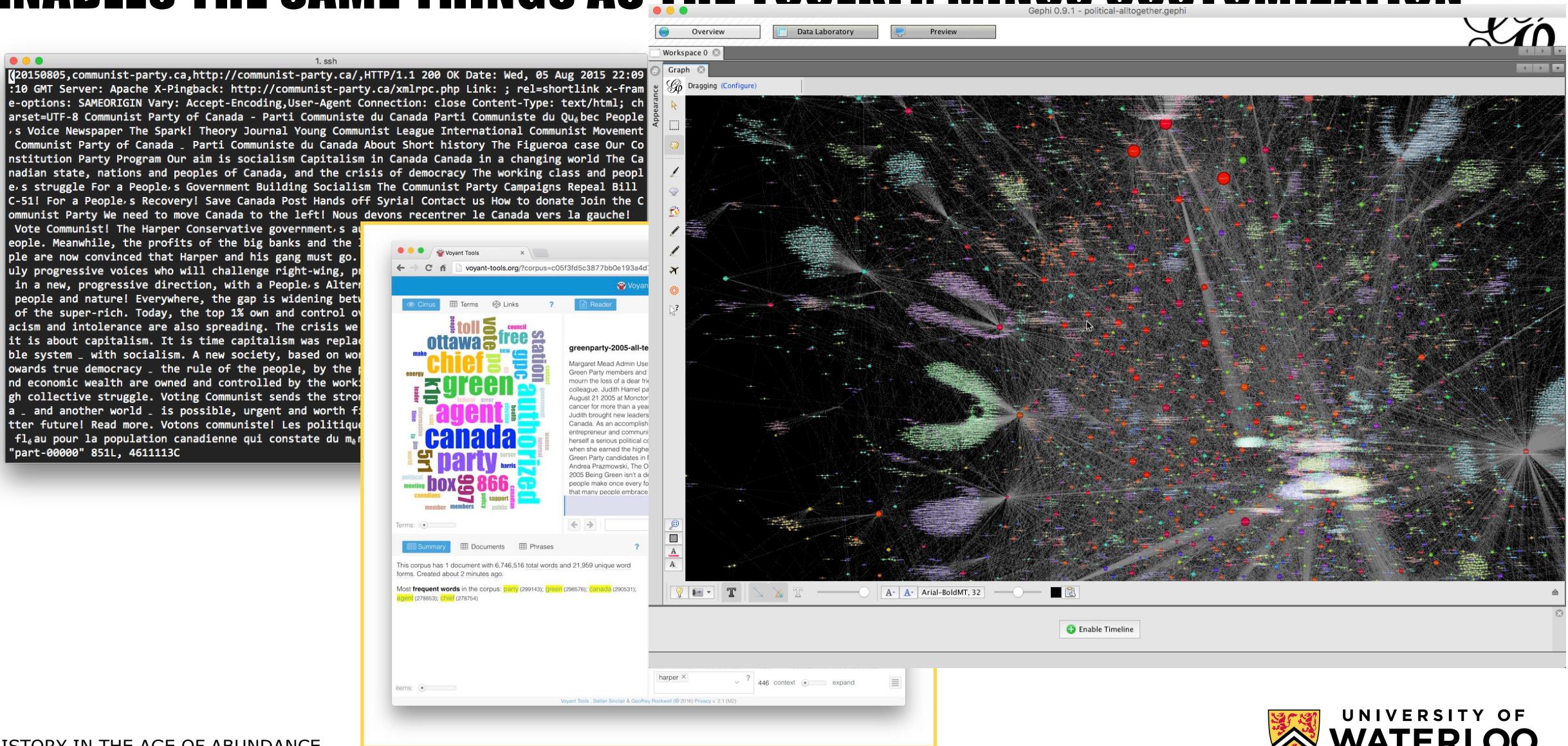
MELLON FOUNDATION UNIVERSITY OF WATERLOO YORK UNIVERSITY

For more information on our project and sponsors, visit archivesunleashed.org/.

About | Privacy Policy | Documentation | FAQ

UNIVERSITY OF WATERLOO

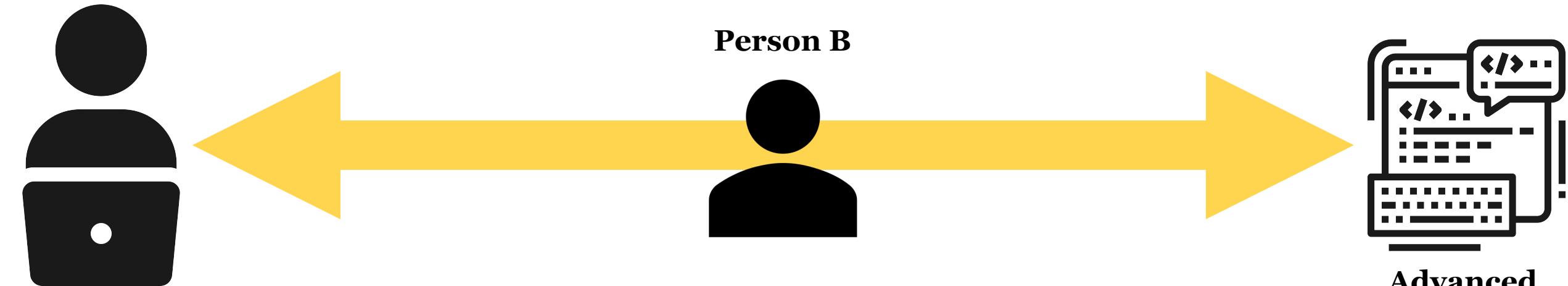
ENABLES THE SAME THINGS AS THE TOOLKIT. MINUS CUSTOMIZATION





**WE'RE MOVING TOWARDS OUR USERS...
BUT THEY STILL REQUIRE SOMEWHAT
SPECIALIZED SKILLS TO INTERPRET DATA.**

IN OTHER WORDS...



**Historian with
Research Question**

Person B

**Advanced
Web Archive Analysis**

WE'RE GETTING THERE.. THEY COME PART WAY, WE COME PART WAY.

PERSONA C

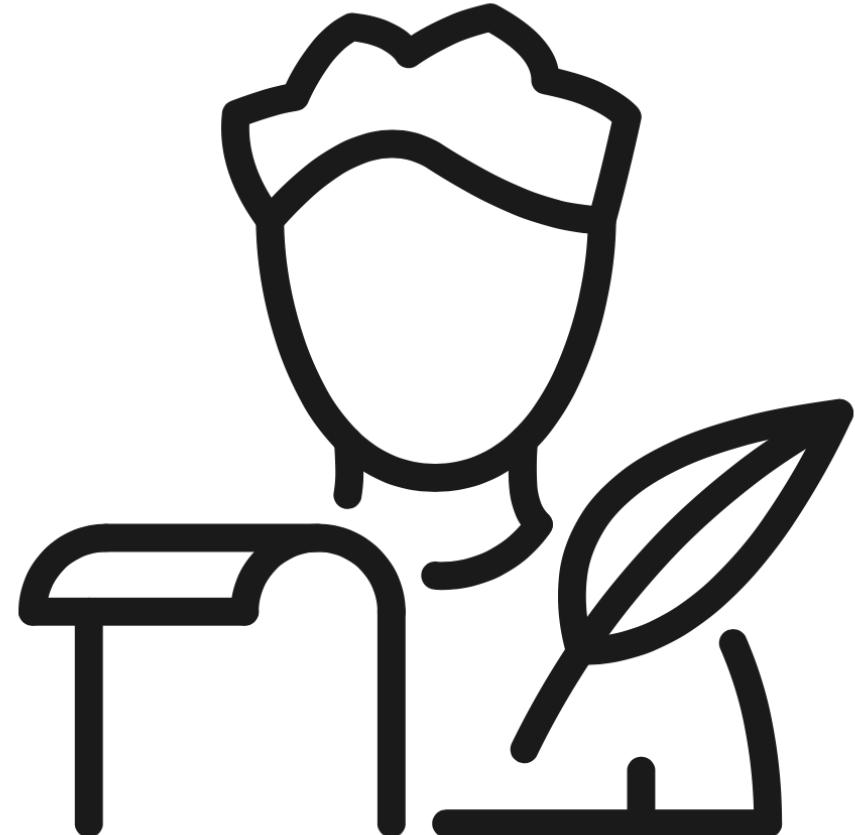
The sort of scholar who might use computers – like for Word, some light Excel – but otherwise generally just wants to do historical research without earning new technical skills.

THE CONVENTIONAL HISTORIAN



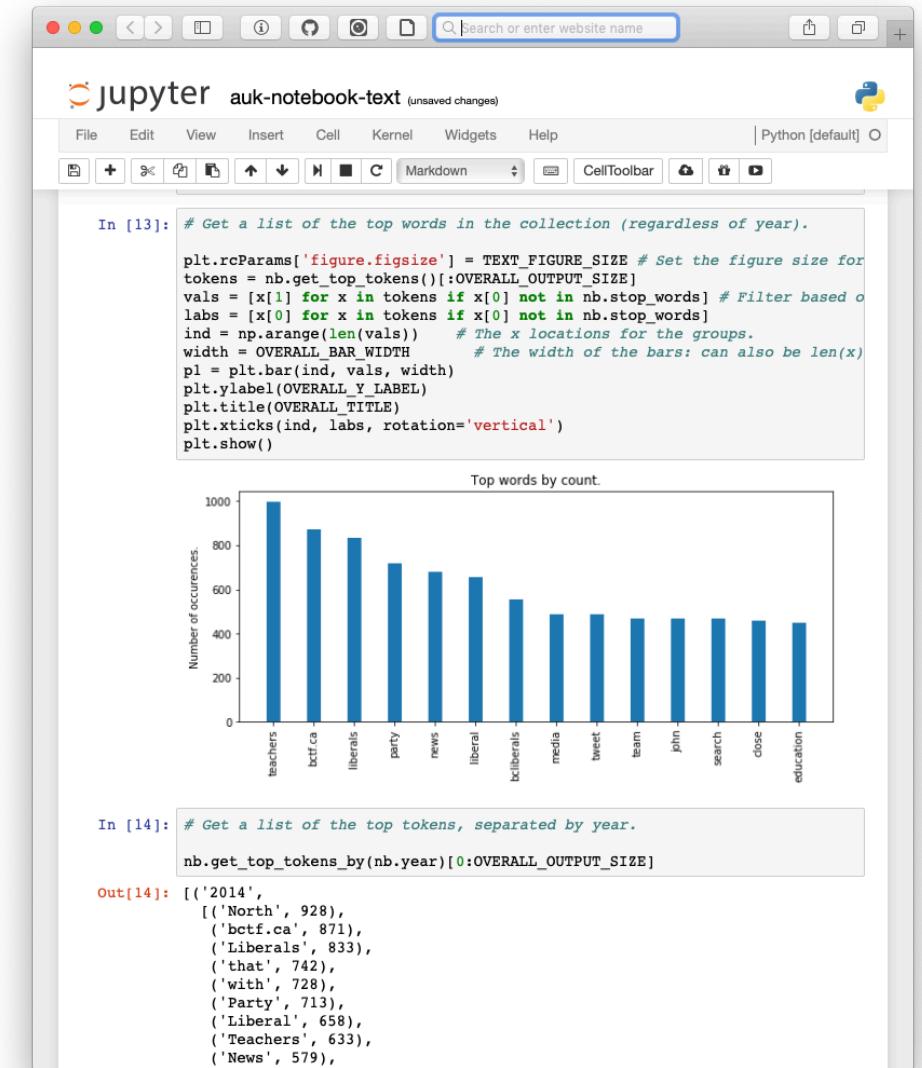
CONVENTIONAL HISTORIAN

- This is the toughest group to serve – if in the Toolkit model we weren't going out far enough to serve people in a reasonable area; this person might not be coming out far enough to meet us...
- But we will try.



CONVENTIONAL HISTORIAN

- **The Archives Unleashed Jupyter Notebooks** were designed with this in mind.
 - Uses web browser (which we can all use);
 - Slightly confusing layout of pressing play buttons and executing code, but thanks to Markdown integration we can mark it up.
 - Extensively playtested (and playtesting) with PhD students and MA students who have varying levels of computational comfort.
- **Exploring one-click hosting with Google Colab/Google Drive**



STILL NOT THE EASIEST TO USE..

But we're trying.

LET'S SEE THEM IN ACTION!

THE NOTEBOOKS IN ACTION

The screenshot shows a Jupyter Notebook interface running on a Mac OS X system. The title bar reads "localhost:8888/notebooks/auk-notebook-text.ipynb". The notebook has a "Python [default]" kernel selected. The main content area displays the "Welcome to the Text Analysis Notebook" page, which includes a small icon of a box with a starburst, a heading, a paragraph of text, and several links. Below this, there are sections titled "How Jupyter Notebooks Work" and "Setup". The "Setup" section contains a heading "The Collection ID" and a code cell at the bottom. The code cell contains the following Python code:

```
In [ ]: COLLECTION_ID = '4867' # Change to switch collections.  
AUK_PATH = 'data/' # Change value to full path to your data, including trailing slash.
```

IN OTHER WORDS...



WE'RE GETTING THERE... BUT THEY STILL NEED TO COME OUT A BIT..!

**FINALLY, WE BUILD CAPACITY TO HELP
PEOPLE COME OUT TO US THROUGH A
SERIES OF DATATHONS.**

DATATHONS



DATATHONS

- Helping to lower barriers
- Bringing people together
- Establishing a true community of practice to work with web archives



THE TL;DR

Historians in the future will need to understand the Web.

We need to make sure they're ready.

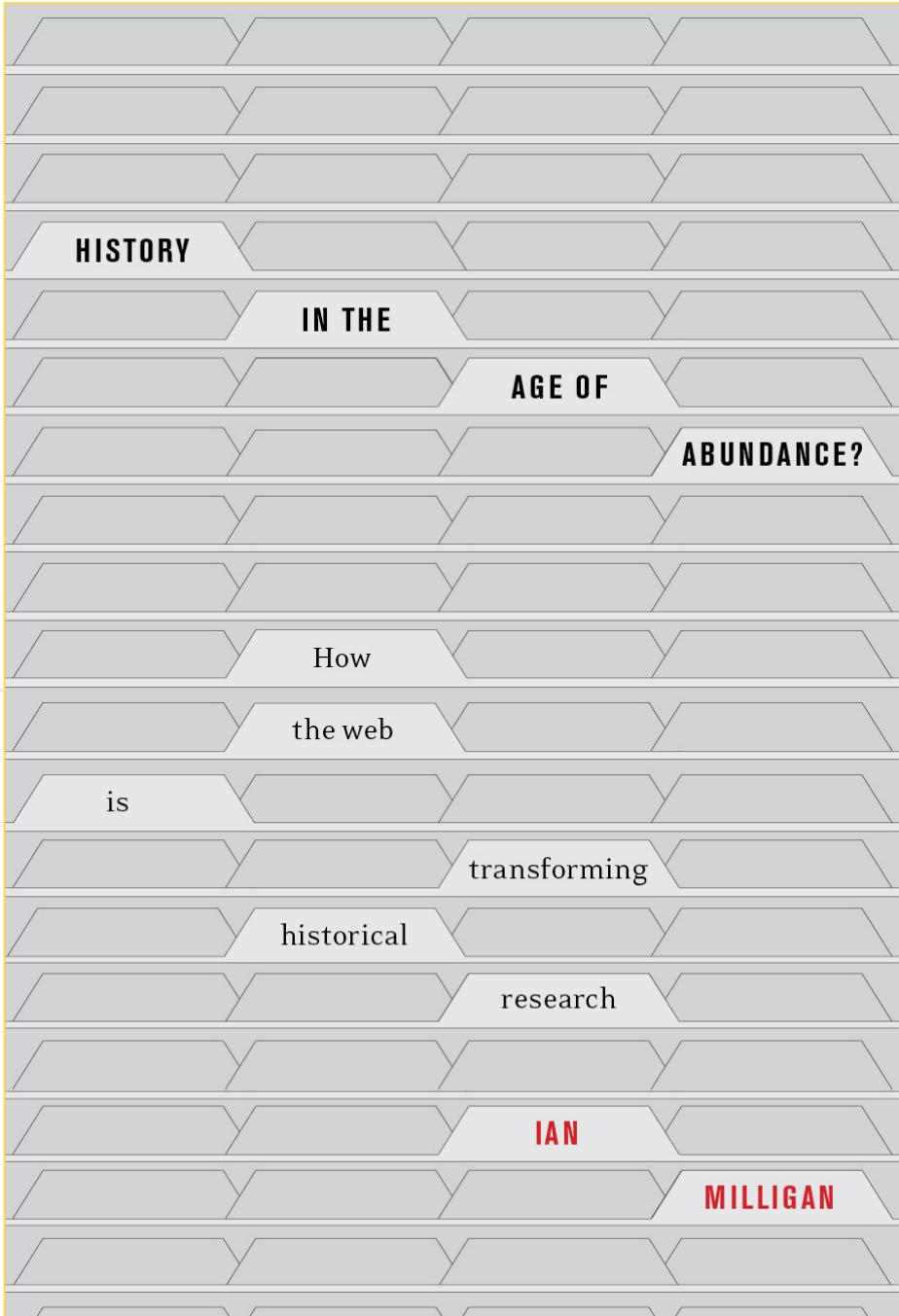
CONCLUDE

- Part of this is new, usable tools... (i.e. we come out towards them);
- Part of this is new cultures in the humanities and social sciences (i.e. they come out towards US)
- But overall, we all need to work together in light of the recognition that you can't study the 1990s without web archives – and we'll be studying the 1990s soon.



WANT TO LEARN MORE?

- You can ask me questions.
- Or you can funnel like a few royalty cents to me (or yeah, just pick up a copy at York University Libraries or Toronto Public Library, I can manage).
- ***History in the Age of Abundance? How the Web is Transforming Historical Research.*** McGill-Queen's University Press, 2019.



THANKS TO OUR FUNDERS!

THE
ANDREW W.
MELLON
FOUNDATION



compute | calcul
canada | canada



Social Sciences and Humanities
Research Council of Canada

Conseil de recherches en
sciences humaines du Canada

Canada