Yale

Mining Large Datasets for the Humanities

IFLA 2014 LYON

Peter Leonard, PhD

Librarian for Digital Humanities Research
Yale University Library

How can libraries support
humanities scholars
in making sense of
large digitized collections
of cultural material?

Three Aspects:

How can libraries support **humanities scholars** in **making sense** of **large digitized collections of cultural material**?

First Part:

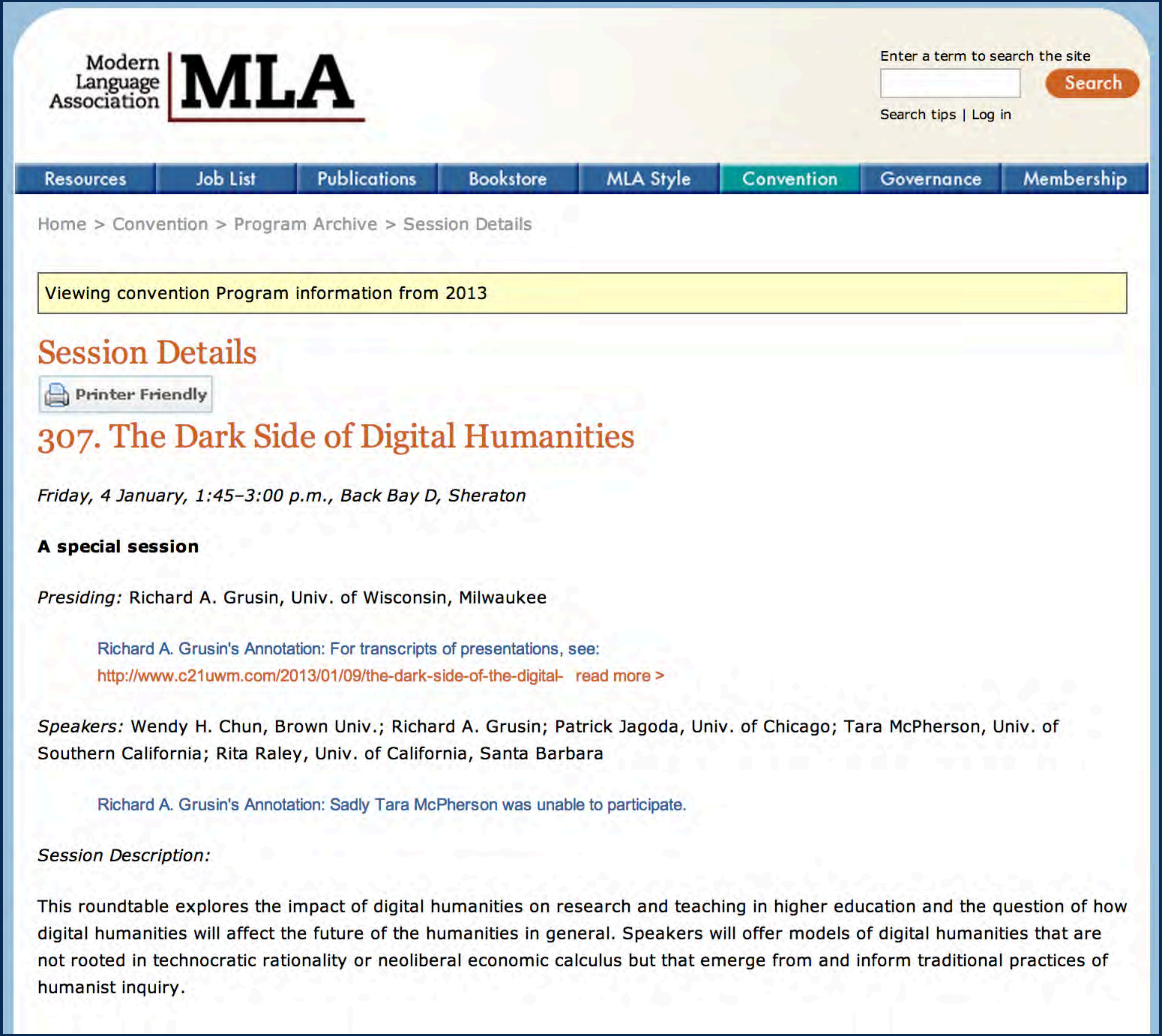
How can libraries support
humanities scholars
in **making sense** of
large digitized collections
of cultural material?

Humanities Scholars’ Challenges:

- Not lack of interest
- Lack of quantitative training
- Lack of “laboratory” model & teamwork

Library Opportunities:

- Collaboration between Subject Librarians, Data Librarians, & Scholars
- Library as neutral ground for STEM & Humanists



Viewing convention Program information from 2013

Session Details

307. The Dark Side of Digital Humanities

Friday, 4 January, 1:45–3:00 p.m., Back Bay D, Sheraton

A special session

Presiding: Richard A. Grusin, Univ. of Wisconsin, Milwaukee

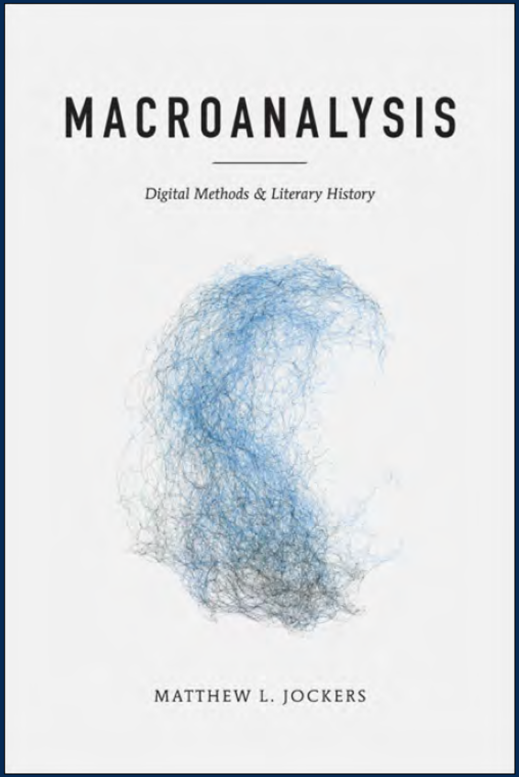
Richard A. Grusin's Annotation: For transcripts of presentations, see: http://www.c21uwm.com/2013/01/09/the-dark-side-of-the-digital- read more >

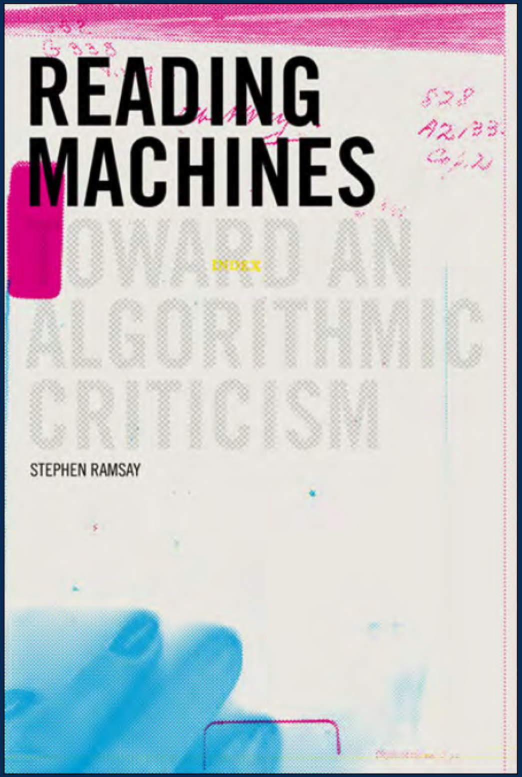
Speakers: Wendy H. Chun, Brown Univ.; Richard A. Grusin; Patrick Jagoda, Univ. of Chicago; Tara McPherson, Univ. of Southern California; Rita Raley, Univ. of California, Santa Barbara

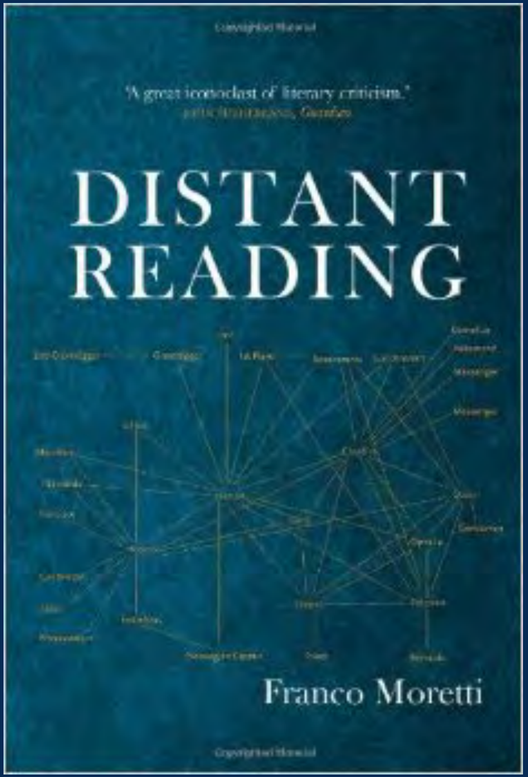
Richard A. Grusin's Annotation: Sadly Tara McPherson was unable to participate.

Session Description:

This roundtable explores the impact of digital humanities on research and teaching in higher education and the question of how digital humanities will affect the future of the humanities in general. Speakers will offer models of digital humanities that are not rooted in technocratic rationality or neoliberal economic calculus but that emerge from and inform traditional practices of humanist inquiry.


MACROANALYSIS
Digital Methods & Literary History
MATTHEW L. JOCKERS


READING MACHINES
TOWARD AN ALGORITHMIC CRITICISM
INDEX
STEPHEN RAMSAY


Copyrighted Material
'A great iconoclast of literary criticism.'
Guardian
DISTANT READING
Franco Moretti
Copyrighted Material

How can libraries support **humanities scholars** in **making sense** of **large digitized collections of cultural material?**

Some answers…

- Big opportunity for libraries in this area
- Possibilities & limits of collaboration
- Keep up on disciplinary debates & literature

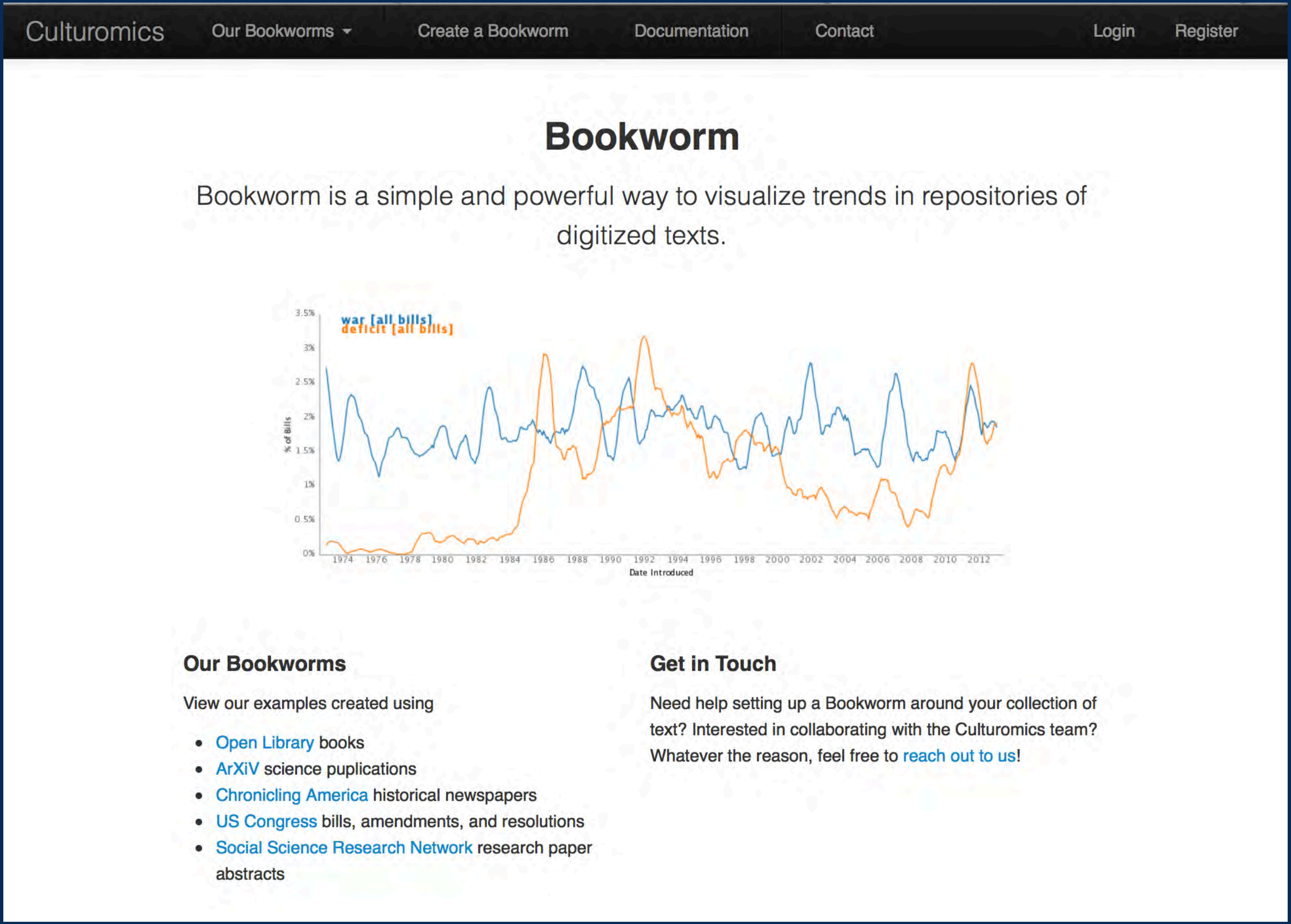
Second Part:

How can libraries support
humanities scholars
in **making sense** of
large digitized collections
of cultural material?

Making Sense

- Looking for something you think is there

- Letting the data organize itself



Bookworm

Bookworm is a simple and powerful way to visualize trends in repositories of digitized texts.

Our Bookworms

View our examples created using

- Open Library books
- ArXiV science puplications
- Chronicling America historical newspapers
- US Congress bills, amendments, and resolutions
- Social Science Research Network research paper abstracts

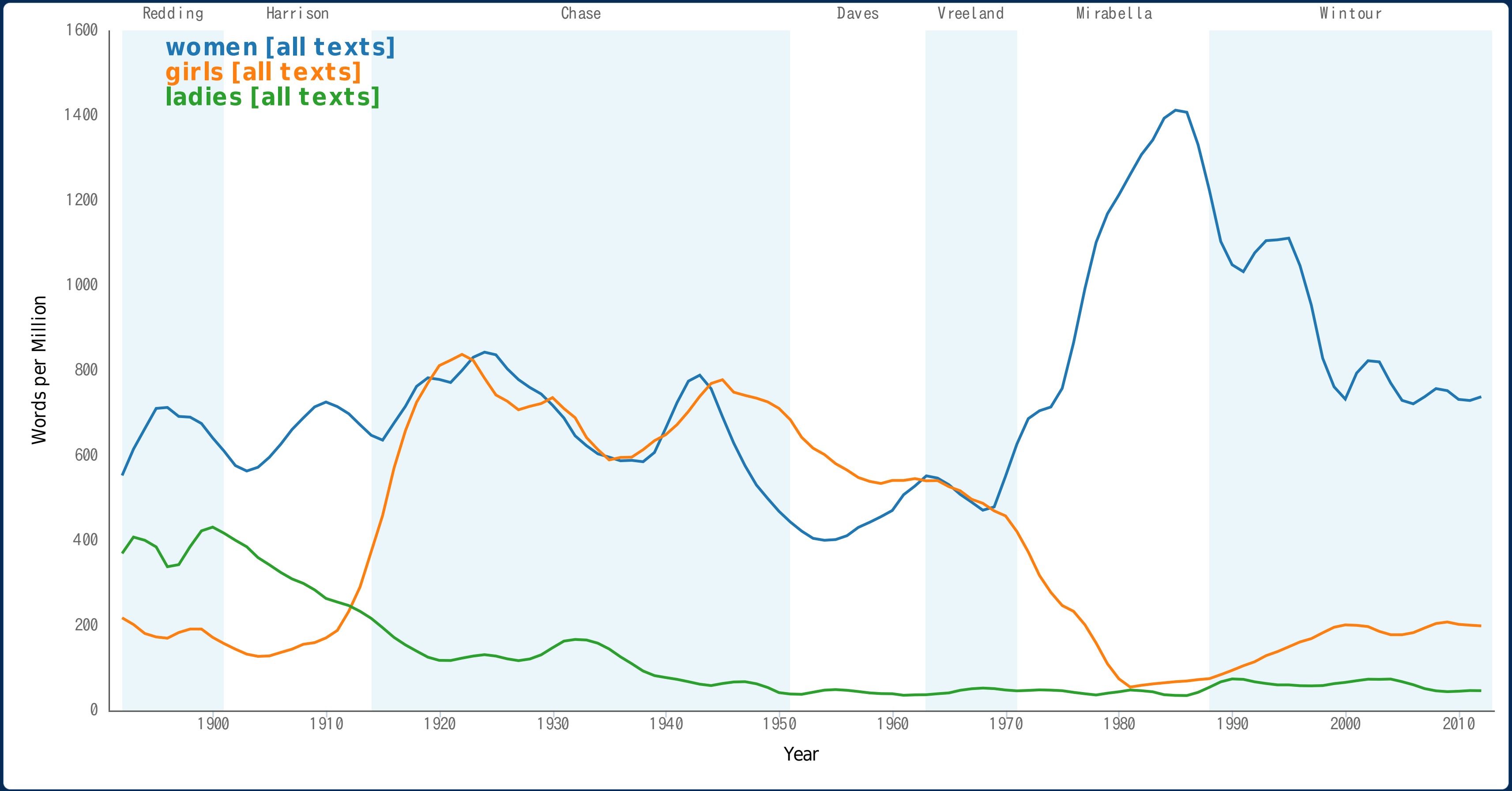
Get in Touch

Need help setting up a Bookworm around your collection of text? Interested in collaborating with the Culturomics team? Whatever the reason, feel free to reach out to us!

UE
MARCH
1969
VOGUE
148
OCT. - DEC.
1966
VOGUE
APRIL - JUNE
1966
VOGUE
APRIL - JUNE
1972
VOGUE
OCT. - DEC.
1961
VOGUE
OCT. 15 - DE
1968

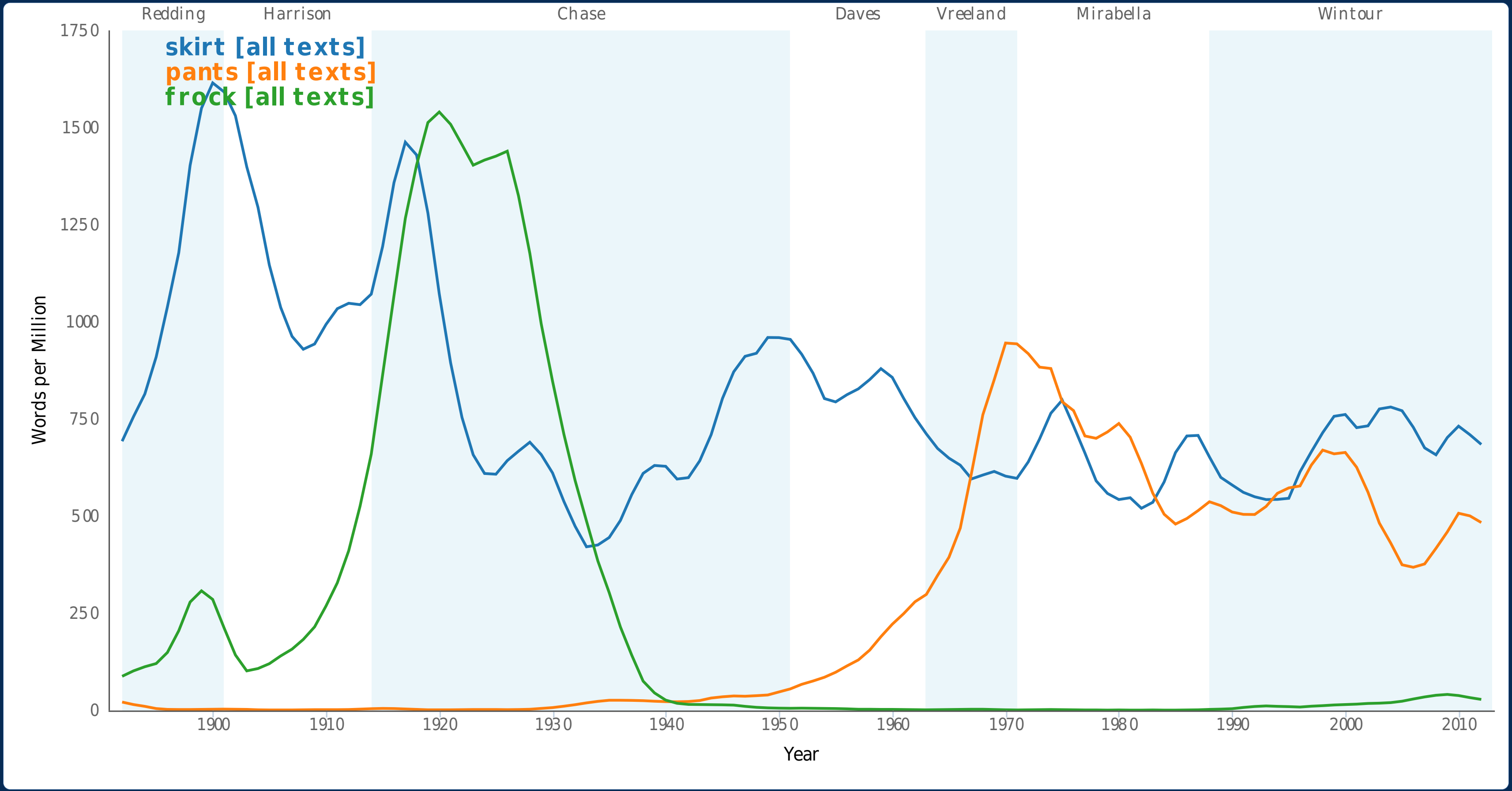
Looking for something you think is there:

“women” vs. “girls” vs. “ladies”



Looking for something you think is there:

“skirt” vs. “pants” vs. “frock”



Special issue of Poetics: "**Topic Models and the Cultural Sciences**"

Editors' Introduction:

"Topic models: What they are and why they matter."
John Mohr (Sociology, UCSB) and Petko Bogdanov (Computer Science, UCSB)

In this short essay we provide a brief, non-technical introduction to the text mining methodology known as topic modeling. We start with the most basic question—what is a topic model? We review the theory behind the method and then focus in on the concept of a 'topic.' Here we

e have understood and interpreted the
comment briefly on some of the
hen proceed to the second question
/e answer by describing some of the
olars in the social and cultural sciences
ng the eight research papers collected for
ection 1 contains two papers that more
opic models to address social scientific
dels to analyze the way that public
atives of national security are
ays to apply topic models to analyze

OFFICE OF **DIGITAL HUMANITIES**

RESOURCE

Topic Modeling for Humanities Research

AUGUST 13, 2014 | BY NEH STAFF

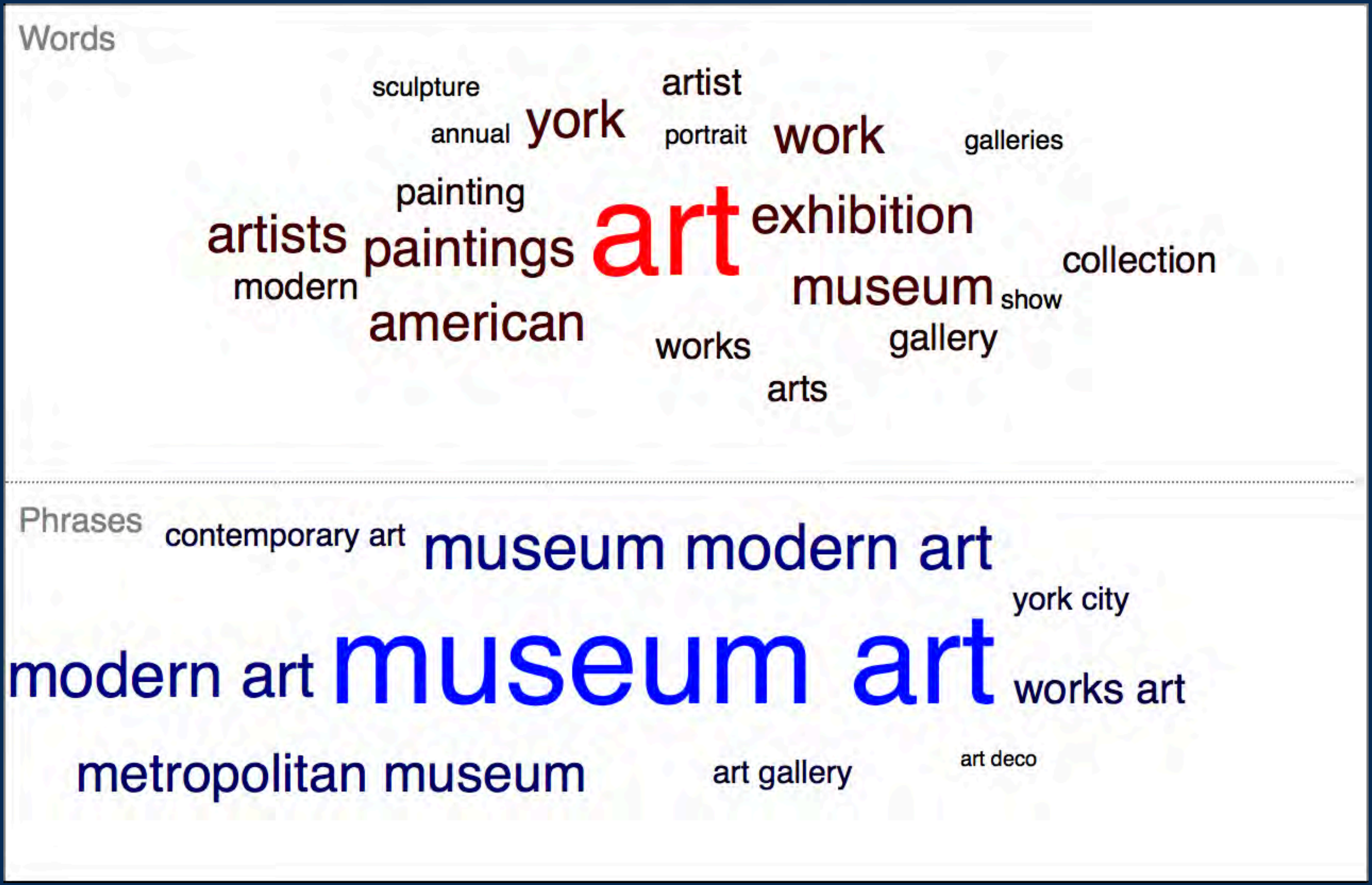
The Maryland Institute for Technology in the Humanities (MITH) will host the Topic Modeling for Humanities Research workshop on November 3, 2012. The workshop will facilitate a unique opportunity for cross-fertilization, information exchange, and collaboration between and among humanities scholars and researchers in natural language processing on the subject of topic modeling applications and methods. The workshop will be organized into three primary areas: 1) an overview of how topic modeling is currently being used in the humanities; 2) an inventory of extensions of the LDA model that have particular relevance for humanities research questions; and 3) a discussion of software implementations, toolkits, and interfaces. The workshop was funded by a Digital Humanities Start-Up Grant (HD-51627-12).

Paper #1: "Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of Government Arts Funding in the U.S."
Paul DiMaggio (Sociology, Princeton University), Manish Nag (Sociology, Princeton University), and David Blei (Computer Science, Princeton University).

Topic modeling provides a valuable method for identifying the linguistic contexts that surround social institutions or policy domains. This paper uses Latent Dirichlet Allocation (LDA) to analyze how one such policy domain, government assistance to artists and arts organizations, was framed in almost 8,000 articles. These comprised all articles that referred to government support for the arts in the U.S. published in five U.S. newspapers between 1986 and 1997, a period during which such assistance, once noncontroversial, became a focus of contention. We illustrate the strengths of topic modeling as a means of analyzing large text corpora, discuss the proper choice

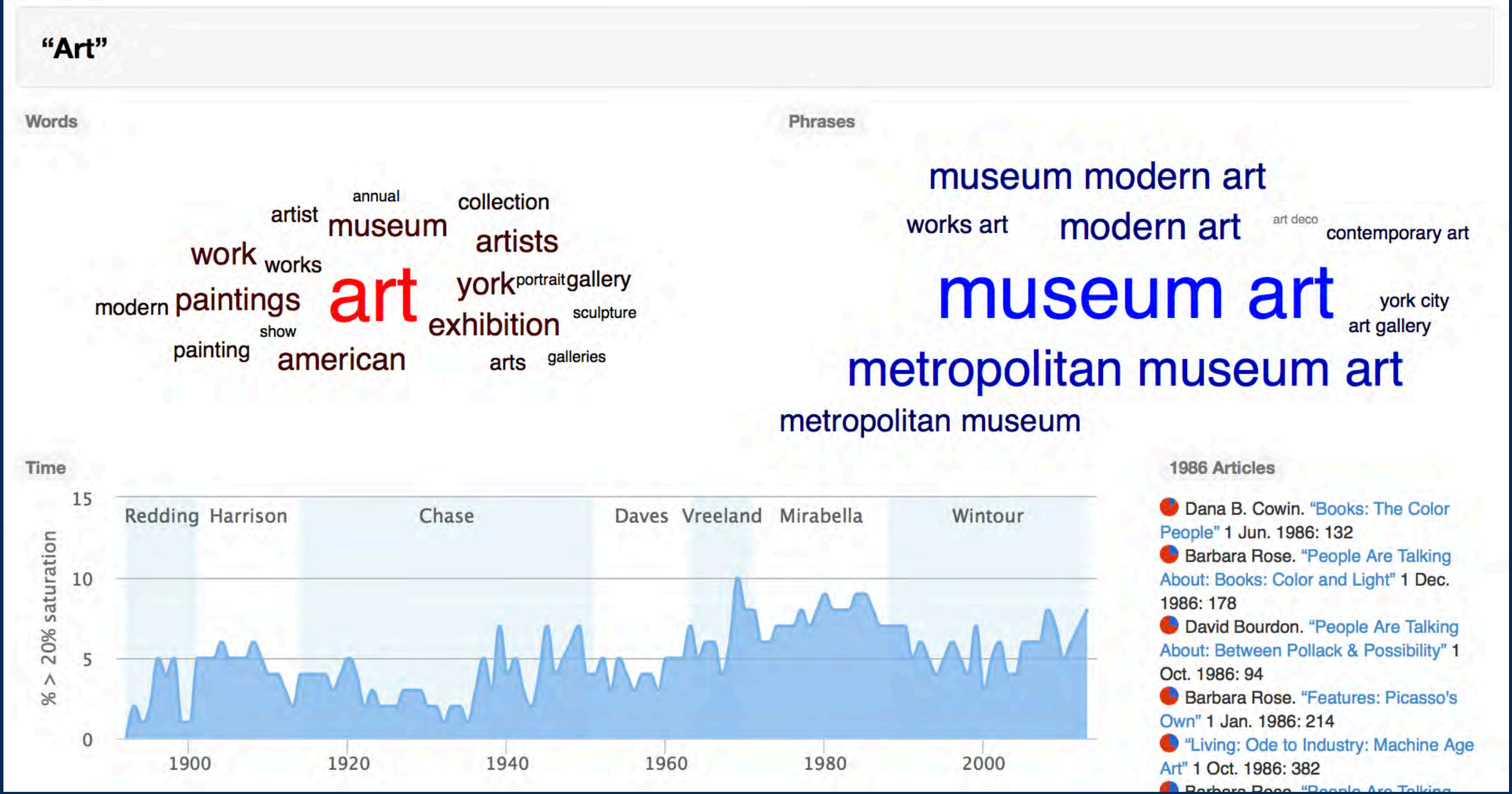
Letting the data organize itself:

“Topic 1”



Letting the data organize itself:

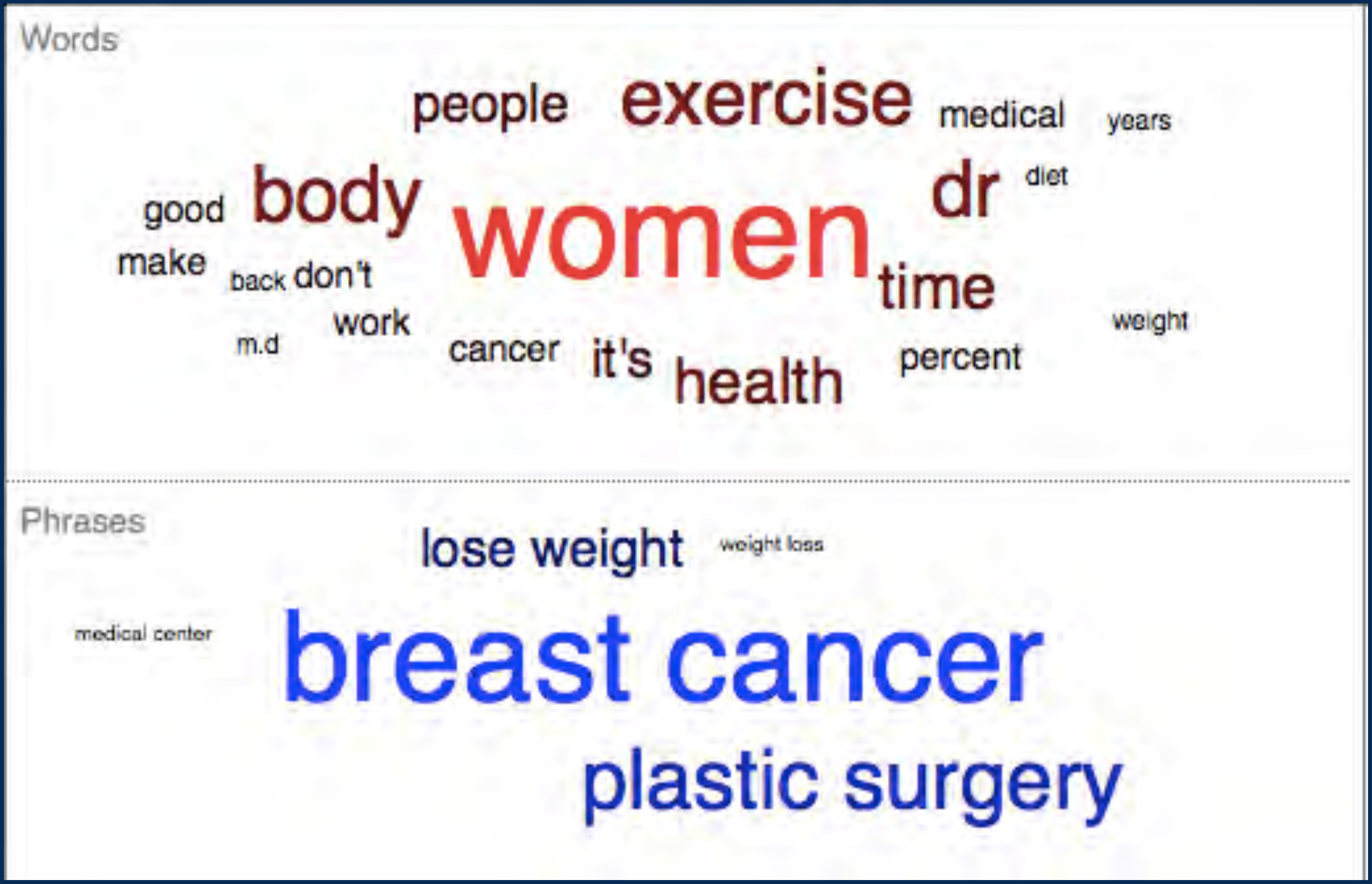
“Topic 1” (Art & Museums)



1986 Articles

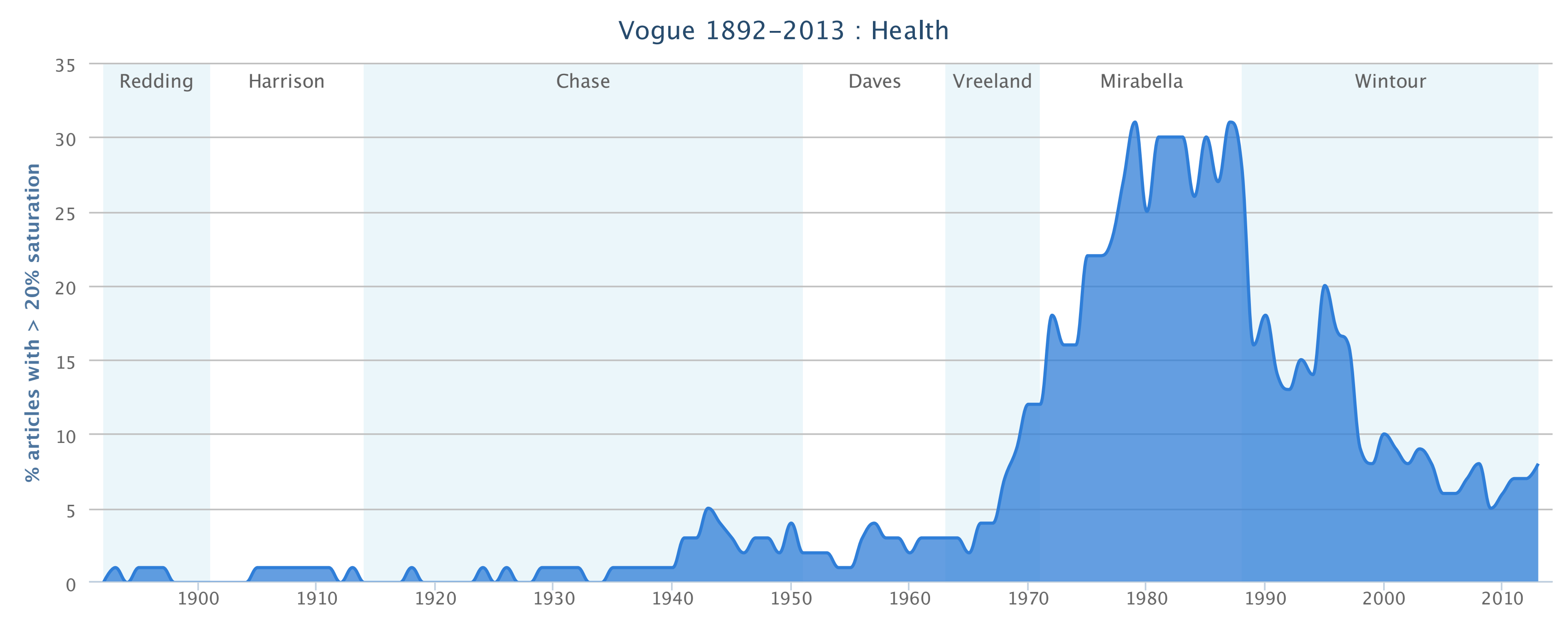
- Dana B. Cowin. “Books: The Color People” 1 Jun. 1986: 132
- Barbara Rose. “People Are Talking About: Books: Color and Light” 1 Dec. 1986: 178
- David Bourdon. “People Are Talking About: Between Pollack & Possibility” 1 Oct. 1986: 94
- Barbara Rose. “Features: Picasso's Own” 1 Jan. 1986: 214
- “Living: Ode to Industry: Machine Age Art” 1 Oct. 1986: 382
- Barbara Rose. “People Are Talking

Letting the data organize itself:
“Topic 14”



Letting the data organize itself:

“Topic 14” (Women’s Health)



Peak Women’s Health (1970s-80s)

99% “Q & A: The Pill” 1 Dec. 1987: 361

98% Jane Ogle. “Facts on Fat: Obesity—a Heavy Health-risk Factor” 1 Aug. 1979: 249

96% Melva Weber. “Inner Info: Contraception” 1 Aug. 1978: 210

95% Charles Kuntzleman. “What Is the Best Way That You Can Shape Up for Active Sports?” 1 Aug. 1979: 82

95% Jane Ogle. “Why Crash Diets Don't Work” 1 Aug. 1979: 248

91% Melva Weber. “Latest in the IUD Dust-Up…” 1 Mar. 1975: 88

89% Ellen Switzer. “Your Blood Pressure” 1 May. 1973: 152

Proto Women’s Health (1910s, 1930s, etc)

66% “Correct Breathing as a Figure Builder” 13 May. 1909: 894

50% “How to Reduce Weight Judiciously” 15 Jun. 1910: 10

44% “Health Laws for Rheumatics” 15 Mar. 1911: 100

43% “Mechanical Massage” 18 Jul. 1907: 84

29% “Teaching Poise to Children” 11 Sep. 1909: 342

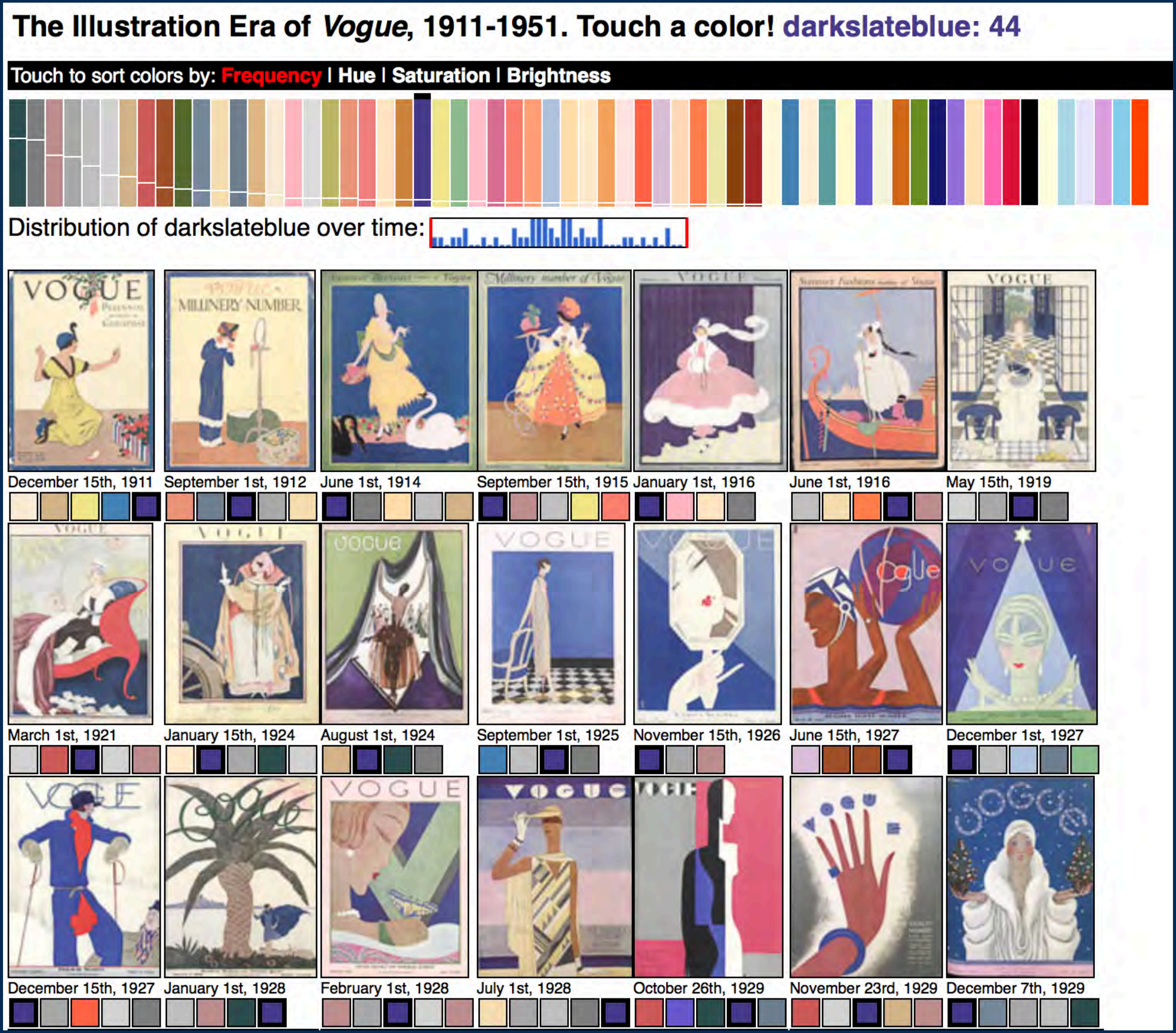
26% “Tuberculosis: A Preventable & Curable Disease” 12 Aug. 1909: 18

26% “Good Form for These Ruthless New Dresses” 15 Apr. 1931: 93

How can libraries support **humanities scholars** in **making sense** of **large digitized collections of cultural material?**

Some Answers…

- Be open to new methods from outside the Humanities
- Balance subject expertise with algorithmic discovery



The Illustration Era of *Vogue*, 1911-1951. Touch a color! darkslateblue: 44

Touch to sort colors by: Frequency | Hue | Saturation | Brightness

Distribution of darkslateblue over time:

December 15th, 1911

September 1st, 1912

June 1st, 1914

September 15th, 1915

January 1st, 1916

June 1st, 1916

May 15th, 1919

March 1st, 1921

January 15th, 1924

August 1st, 1924

September 1st, 1925

November 15th, 1926

June 15th, 1927

December 1st, 1927

December 15th, 1927

January 1st, 1928

February 1st, 1928

July 1st, 1928

October 26th, 1929

November 23rd, 1929

December 7th, 1929

Third Part:

How can libraries support **humanities scholars** in **making sense** of **large digitized collections of cultural material**?

Digital Cultural Material

- HathiTrust
- JSTOR
- Internet Archive

Digital Cultural Material

- HathiTrust **Research Center**
- JSTOR **Data for Research Program**
- Internet Archive **Bulk Download Tool**

Large digitized collections of cultural material

Vendor-digitized cultural material

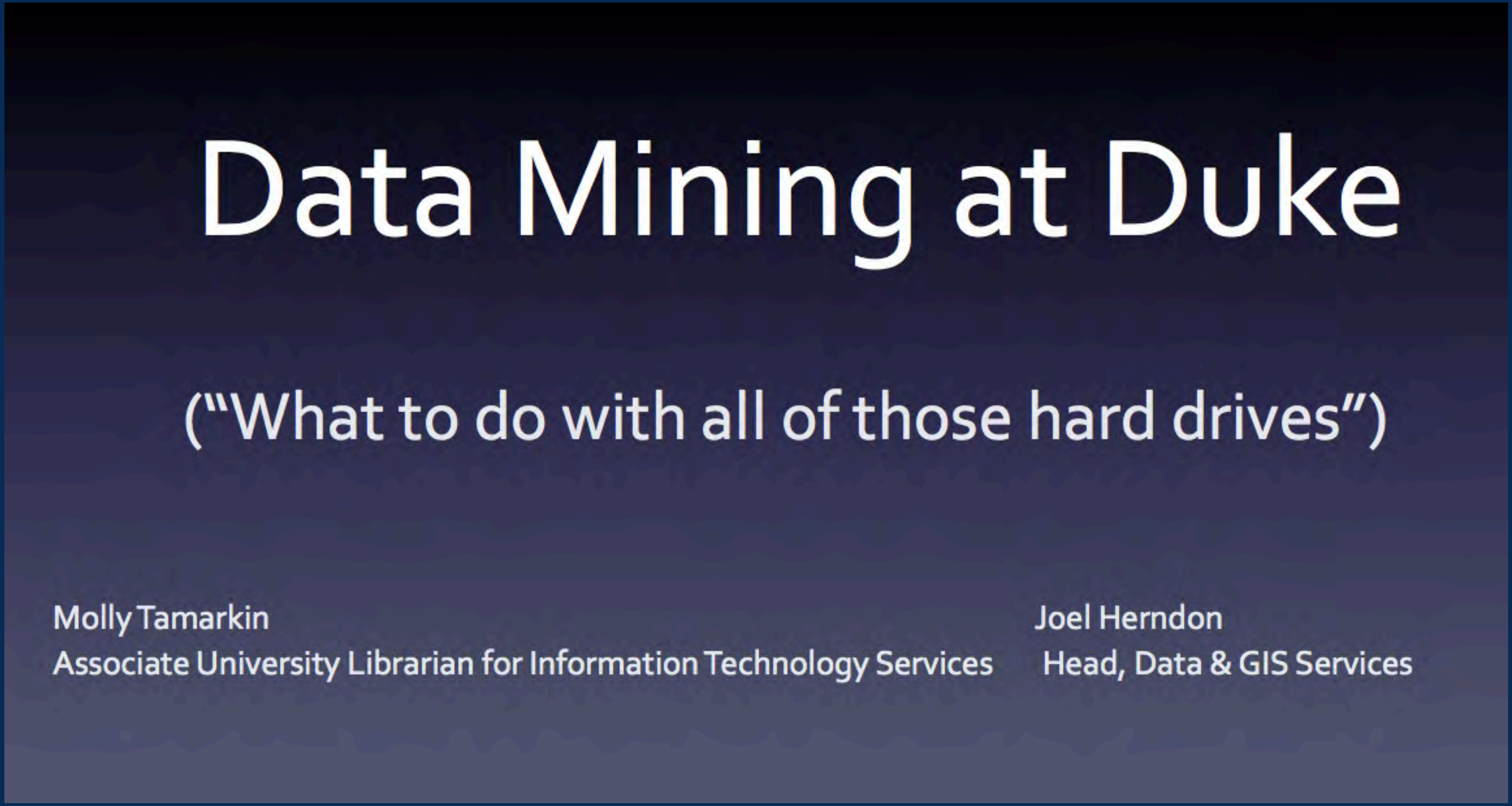
Challenges

- Copyright restrictions
- License restrictions
- Lack of awareness

Opportunities

- Pre-digitized
- Often item-level descriptions
- Sometimes local copies present due to *perpetual access licenses*

CNI 2012:



Data Mining at Duke

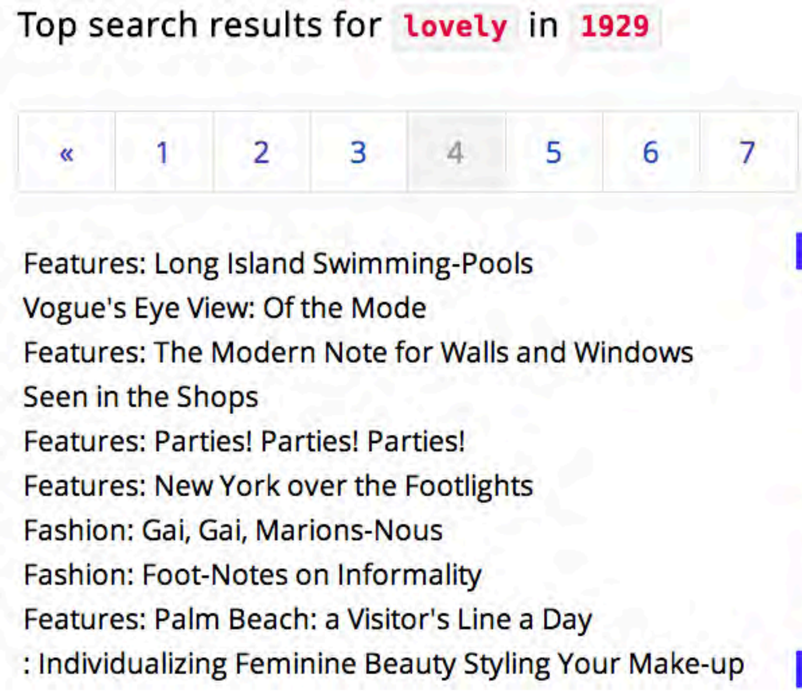
("What to do with all of those hard drives")

Molly Tamarkin
Associate University Librarian for Information Technology Services

Joel Herndon
Head, Data & GIS Services

Two aspects of data mining

- Analysis
- As TDM simply employs computers to "read" material and extract facts one already has the right as a human to read and extract facts from, it is difficult to see how the technical copying by a computer can be used to justify copyright and database laws regulating this activity. (IFLA 2013)
- Presentation



Top search results for `lovely` in `1929`

Features: Long Island Swimming-Pools
Vogue's Eye View: Of the Mode
Features: The Modern Note for Walls and Windows
Seen in the Shops
Features: Parties! Parties! Parties!
Features: New York over the Footlights
Fashion: Gai, Gai, Marions-Nous
Fashion: Foot-Notes on Informality
Features: Palm Beach: a Visitor's Line a Day
: Individualizing Feminine Beauty Styling Your Make-up

}

`http://search.proguest.com/vogue/...`

Two aspects of data mining

- Analysis
  - As TDM simply employs computers to “read” material and extract facts one already has the right as a human to read and extract facts from, it is difficult to see how the technical copying by a computer can be used to justify copyright and database laws regulating this activity. (IFLA 2013)
- Presentation
  - Researchers must be able to share the results of text and data mining, as long as these results are not substitutable for the original copyright work. (IFLA 2013)

How can libraries support **humanities scholars** in **making sense** of **large digitized collections of cultural material**?

Some Answers…

- Amazing humanities data may be hiding in your basement
- Separate analysis from display
- Advocate for full access for *analysis*
- Partner with vendor for *display*

Opportunities for Libraries (and Librarians) in Humanities Data Mining

- Extend scholarly support further into the research lifecycle

- Ensure better use of licensed electronic resources

Special Thanks:

Daniel Dollar Director of Collection Development

Michael Dula Chief Technology Officer

Joan Emmet Licensing & Copyright Librarian

Lindsay King Public Services Librarian, Haas Arts Library

Julie Linden Assistant Director of Collection Development

Alan Solomon Head, Humanities Collections and Research Education