Big Data: Big Deal?
New Challenges for Scholars and Librarians

John Unsworth

New England Technical Services Librarians
Annual Conference

College of Holy Cross, Worcester, MA

May 3rd

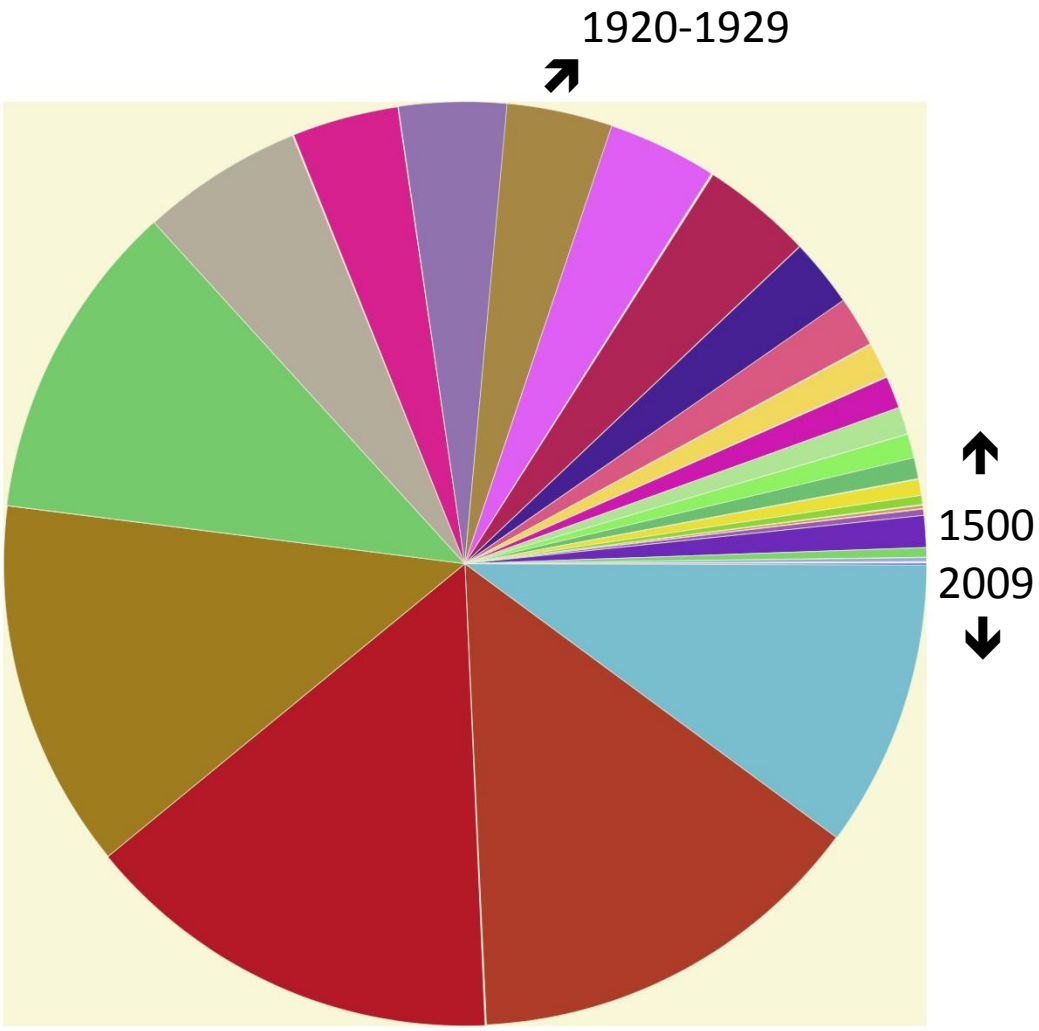
Brief history of digitized books

- 1971: Project Gutenberg
- 1987: Perseus Project
- 1990: Library of Congress's American Memory project
- 1994: NSF Digital Libraries Initiative
- 1995: Making of America, JSTOR
- 2001: Million Books project
- 2004: Google Books
- 2005: Open Content Alliance
- 2008: HathiTrust
- 2010: DPLA

How much?

Currently Digitized in HathiTrust (as of May 2, 2012):

- 10,210,206 total volumes
- 5,422,297 book titles
- 269,166 serial titles
- 3,573,572,100 pages
- 458 terabytes
- 121 miles
- 8,296 tons
- 2,901,693 volumes (~28% of total) in the public domain



Date range	Count	Percent
2000-2009 [3]	556,333	10.11
1990-1999 [4]	779,942	14.18
1980-1989 [5]	813,644	14.79
1970-1979 [6]	712,039	12.94
1960-1969 [7]	618,282	11.24
1950-1959 [8]	312,386	5.68
1940-1949 [9]	205,535	3.74
1930-1939 [10]	209,438	3.81
1920-1929 [11]	205,109	3.73
1910-1919 [12]	210,668	3.83
1900-1909 [13]	213,965	3.89
1890-1899 [14]	132,167	2.40
1880-1889 [15]	98,442	1.79
1870-1879 [16]	68,452	1.24
1860-1869 [17]	62,904	1.14
1850-1859 [18]	52,927	0.96
1840-1849 [19]	47,025	0.85
1830-1839 [20]	40,795	0.74
1820-1829 [21]	31,544	0.57
1810-1819 [22]	18,089	0.33
1800-1899 [23]	6,560	0.12
1800-1809 [24]	13,244	0.24
1700-1799 [25]	61,638	1.12
1600-1699 [26]	18,996	0.35
1500-1599 [27]	8,027	0.15
Pre-1500 [28]	3,767	0.07

Copyright

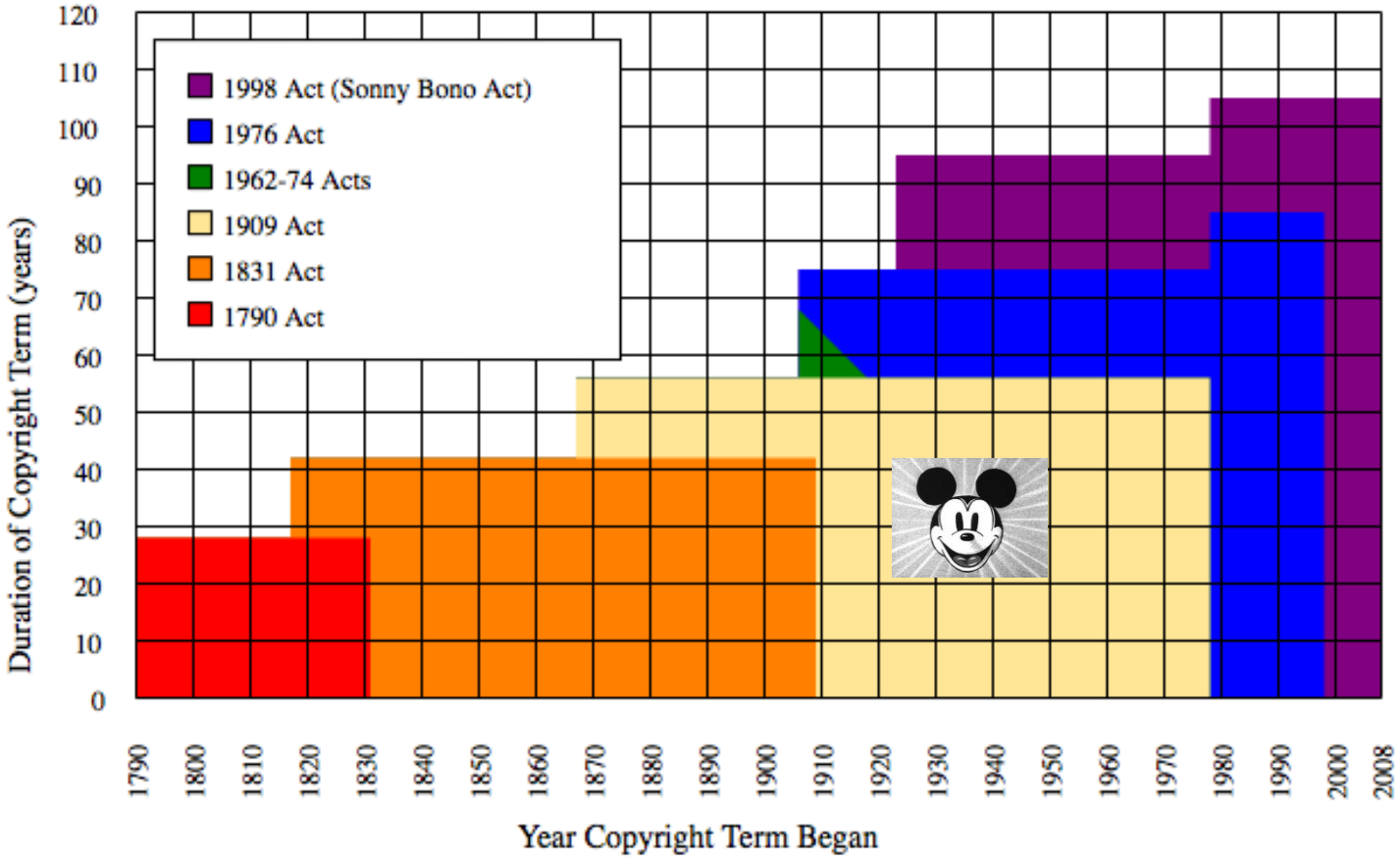
- 1989: US joins the Berne Convention
- 1996: World Intellectual Property Organization Copyright Treaty
- 1998: Digital Millennium Copyright Act
- 2003: Eldred v. Ashcroft
- 2005: Authors Guild v. Google
- 2008: Kahle v. Gonzalez
- 2011: Google Books settlement rejected; Authors Guild et al. v. HathiTrust et al.
- 2012: Authors Guild v. Google goes to court

Mickey v. Old Possum

In which we learn:

- How Mickey Mouse explains the current legal status of an exemplary work of 20th-century literature, namely T.S. Eliot's *The Waste Land*.
- Why, if Mickey had existed long before *The Waste Land* was written, it probably wouldn't have been publishable.
- What the imperative to save Mickey does to starve new forms of scholarship.

Copyright Creep



http://en.wikipedia.org/wiki/Copyright_Term_Extension_Act

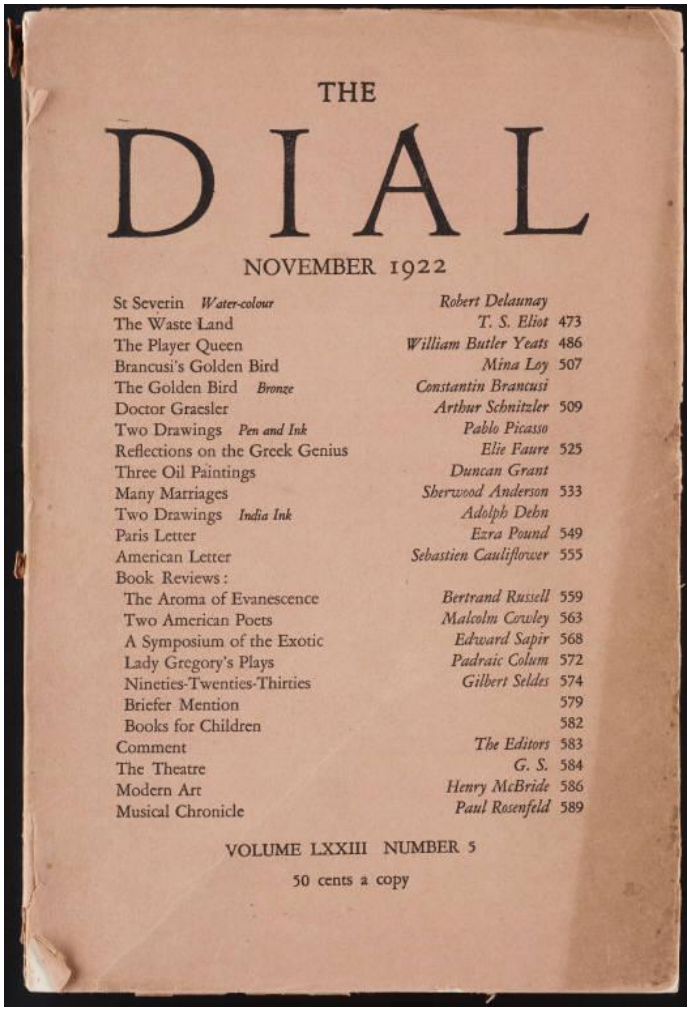
Steamboat Willie

"*Steamboat Willie* has been close to entering the public domain in the United States several times. Each time, copyright protection in the United States has been extended. It could have entered public domain in 4 different years; first in 1956, renewed to 1984, then to 2003 by the Copyright Act of 1976, and finally to the current public domain date of 2023 by the Copyright Term Extension Act (also known pejoratively as the Mickey Mouse Protection Act)[3] of 1998. The U.S. copyright on *Steamboat Willie* will be in effect through 2023 unless there is another change of the law."

http://en.wikipedia.org/wiki/Steamboat_Willie

The Waste Land

T.S. Eliot, by Wyndham Lewis, 1938



THE

DIAL

NOVEMBER 1922

VOLUME LXXIII NUMBER 5

50 cents a copy

Original publication of the poem: 1922, in The Dial (an American literary magazine

Copyright and The Waste Land

- “The copyright was registered in the United States sometime in 1922.
- The copyright gave 28 years of protection plus any additional time to cause it to expire after midnight on the last day of the year. Thus it was protected up to and throughout 1950 (1922 + 28).
- In 1950 the copyright could be renewed for 28 more years meaning that it would enter the public domain **in the United States** after the end of 1978 (1950 + 28).
- In the United States, the Copyright Act of 1976 extended the renewal from 28 years to 47 years giving *The Waste Land* protection for 19 more years or throughout 1997 (1950 + 28 + 19).”

http://world.std.com/~raparker/exploring/thewasteland/excopy.html

Copyright and The Waste Land

- “On January 1, 1998, *The Waste Land* went into public domain **in the United States**.
- On October 27, 1998 U.S. public law 105-298 extended renewal of copyrighted items (that were still under protection) by 20 years.
- *The Waste Land* was, however, already in the public domain **in the United States** and thus remains in that state.
- If *The Waste Land* [had been] written in 1923 it would be protected for 95 years (28 + 28 + 19 + 20) plus the remainder of the last calendar year meaning that it would go into the public domain (in the US) January 1, 2019.”

http://world.std.com/~raparker/exploring/thewasteland/excopy.html

And in England…

- *"The Waste Land* is still under copyright restrictions in the United Kingdom and most likely in the countries of the European Union, the Commonwealth of Nations and other countries. Copies of T.S. Eliot's poems, plays, essays and other of his works that are placed on computers for public access through the internet may be infringing on copyrights held by Faber and Faber, Mrs. T.S. Eliot and others."

Copyright information about the Waste Land comes from R.A. Parker, "Exploring the Waste Land," a hobbyist site at http://www.std.com/~raparker/exploring/thewasteland/excopy.html

The Waste Land

T.S. ELIOT

Poem
The full published text of
The Waste Land (1922)

Performance
A specially filmed performance of
the entire poem by Fiona Shaw

Manuscript
A facsimile of Eliot's original manuscript
with hand-written edits by Ezra Pound

Perspectives
Commentary on the poem and on
Eliot from a range of interesting people

Readings
Hear the poem spoken aloud by
different voices including Eliot himself

Gallery
A selection of photographs and images
related to the poem

Tips
How to get the best from this
electronic edition of The Waste Land

Notes
Annotations and references
explaining the text of the poem

CREDITS SETTINGS

TOUCH PRESS ff

Necessary Characteristics of Cyberinfrastructure

- It will be accessible as a public good.
- It will be sustainable.
- It will provide interoperability.
- It will facilitate collaboration.
- It will support experimentation.

--ACLS Report on Cyberinfrastructure for the Humanities and Social Sciences (2006)

HATHI TRUST

A Shared Digital Repository

HathiTrust Research Center

Goals of the HTRC

- Maintain repository of text mining algorithms, retrieval tools, derived data sets, and indices available for human and programmatic discovery.
- Be a user-driven resource, with an active advisory board, and a community model that allows users to share tools and results.
- Support interoperability across collections and institutions, through use of inCommon SAML identity.
- See also: http://www.ideals.illinois.edu/handle/2142/29936 -- a report prepared by the Illinois Center for Informatics Research in Science and Scholarship, on the experience of Google Digital Humanties grant recipients.

RESEARCH CENTER

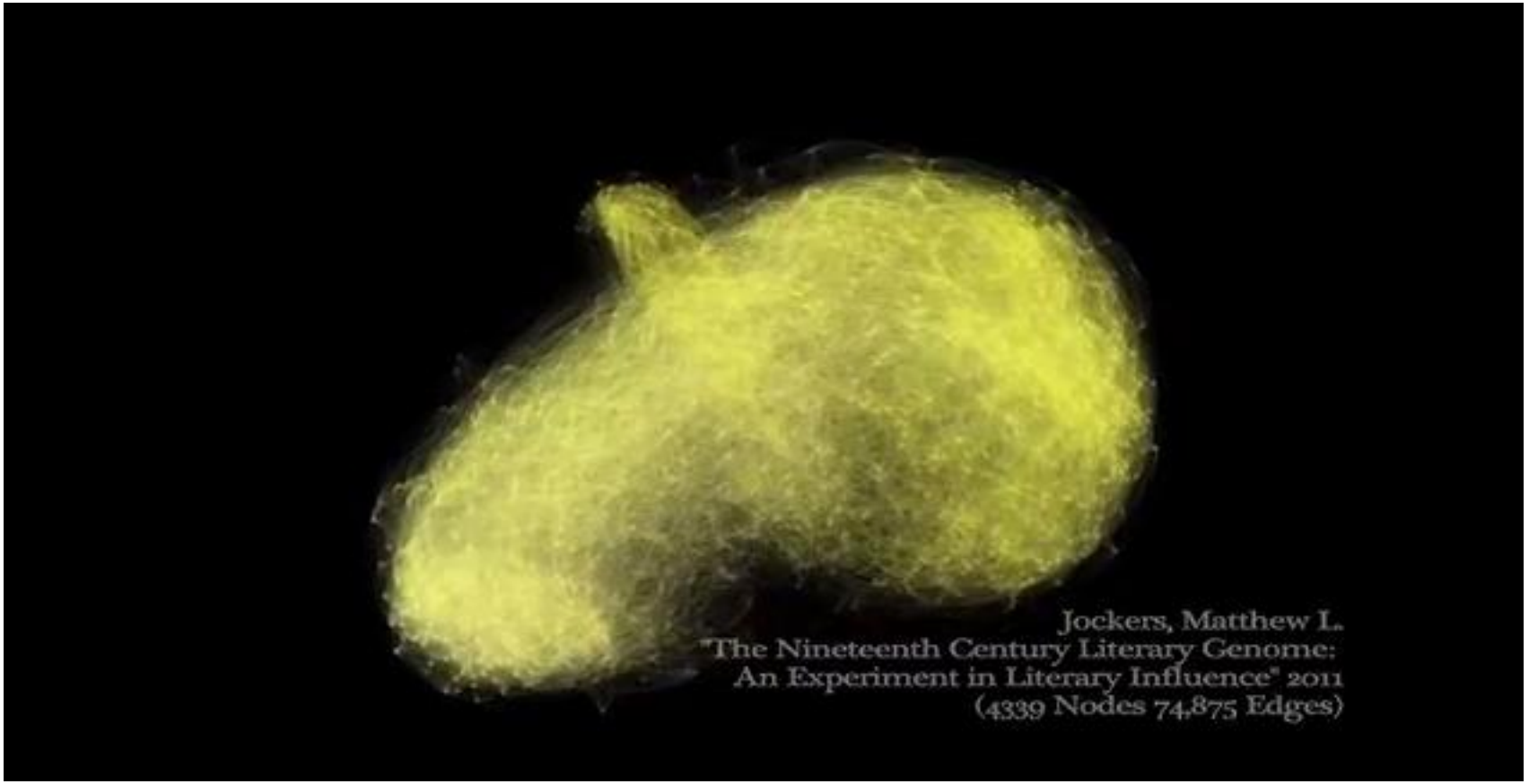
Non-consumptive Research

"Research in which computational analysis is performed on one or more Books, but not research in which a researcher reads or displays substantial portions of a Book to understand the intellectual content presented within the Book."

Non-consumptive Research

- One of HTRC's unique challenges is support for non-consumptive research.
- This will entail bringing algorithms to data, and exporting results, and/or providing people with secure computational environments in which they can work with copyrighted materials without exporting them.
- Why is this worth doing? Because it enables a new art of information that can be used to make new kinds of arguments (and possibly to settle some old ones).

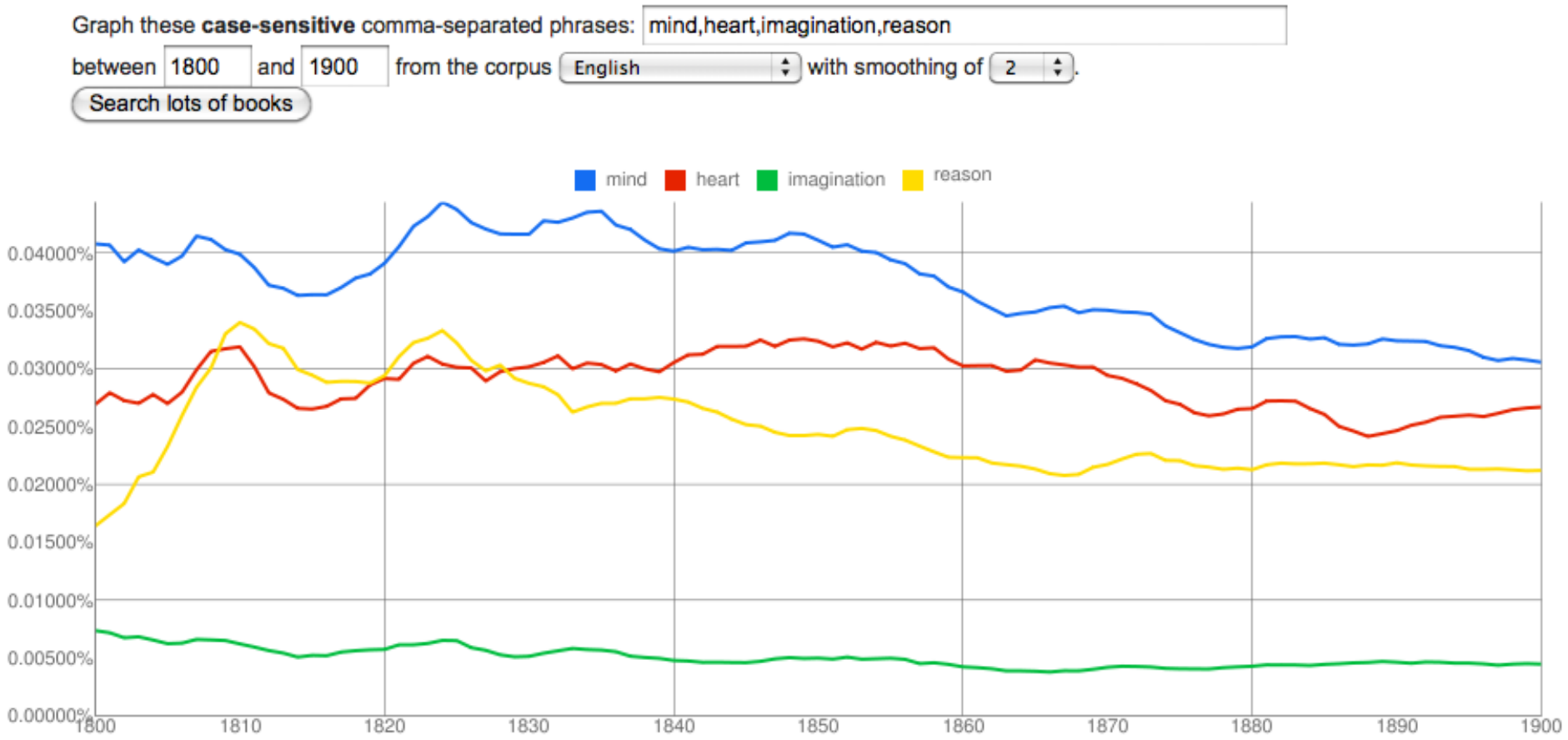
Matt Jockers, “The Nineteenth-Century Literary Genome” via Digital Humanities Specialist (aka Elijah Meeks) http://dhs.stanford.edu



Arguing with Data

- Data enables arguments based on quantitative and/or empirical data
- Data still requires interpretation, and you can still make better and worse interpretations, and more or less compelling arguments
- In addition to new kinds of arguments, you can make new kinds of mistakes, especially mistakes based on incomplete data or on an incomplete understanding of data

Mistakes based on incomplete data

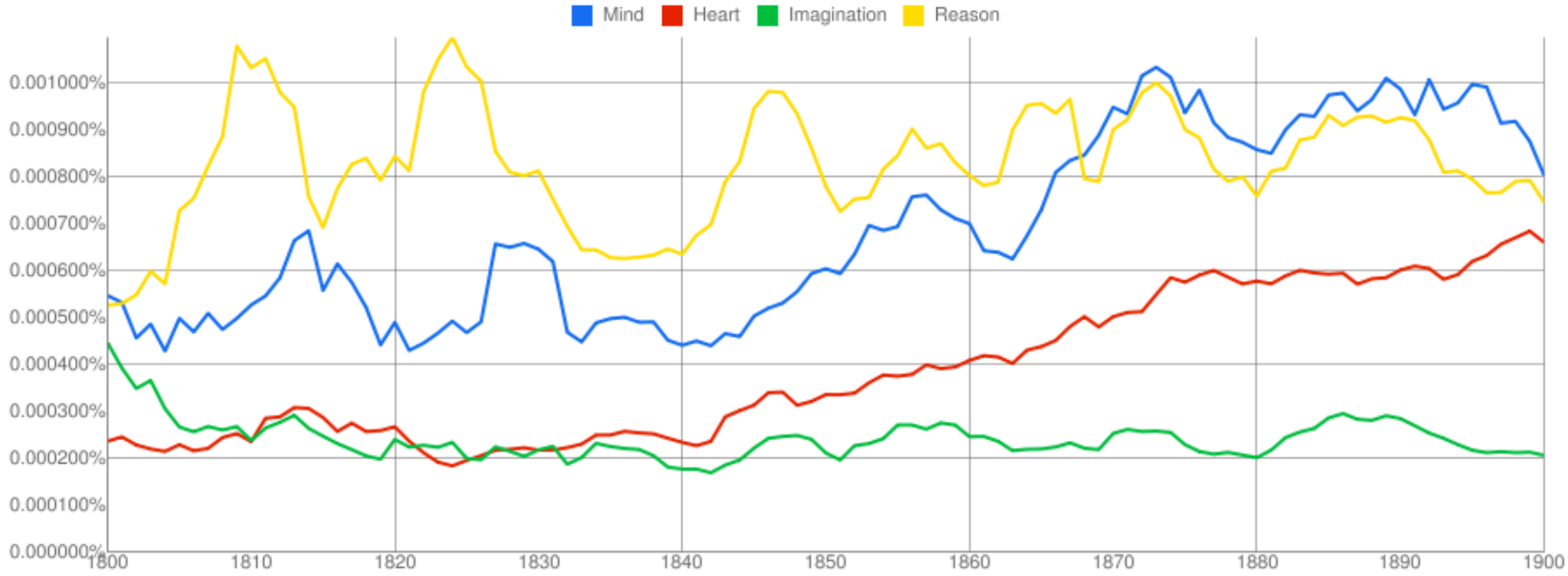


Mistakes based on incomplete data

Graph these **case-sensitive** comma-separated phrases: Mind,Heart,Imagination,Reason

between 1800 and 1900 from the corpus English with smoothing of 2.

Search lots of books



New kinds of arguments

http://tedunderwood.wordpress.com/

Ted Underwood is exploring the changing etymological basis of diction in English, over a 200-year period, especially the shift from words derived from German, to words derived from Latin, and back again.

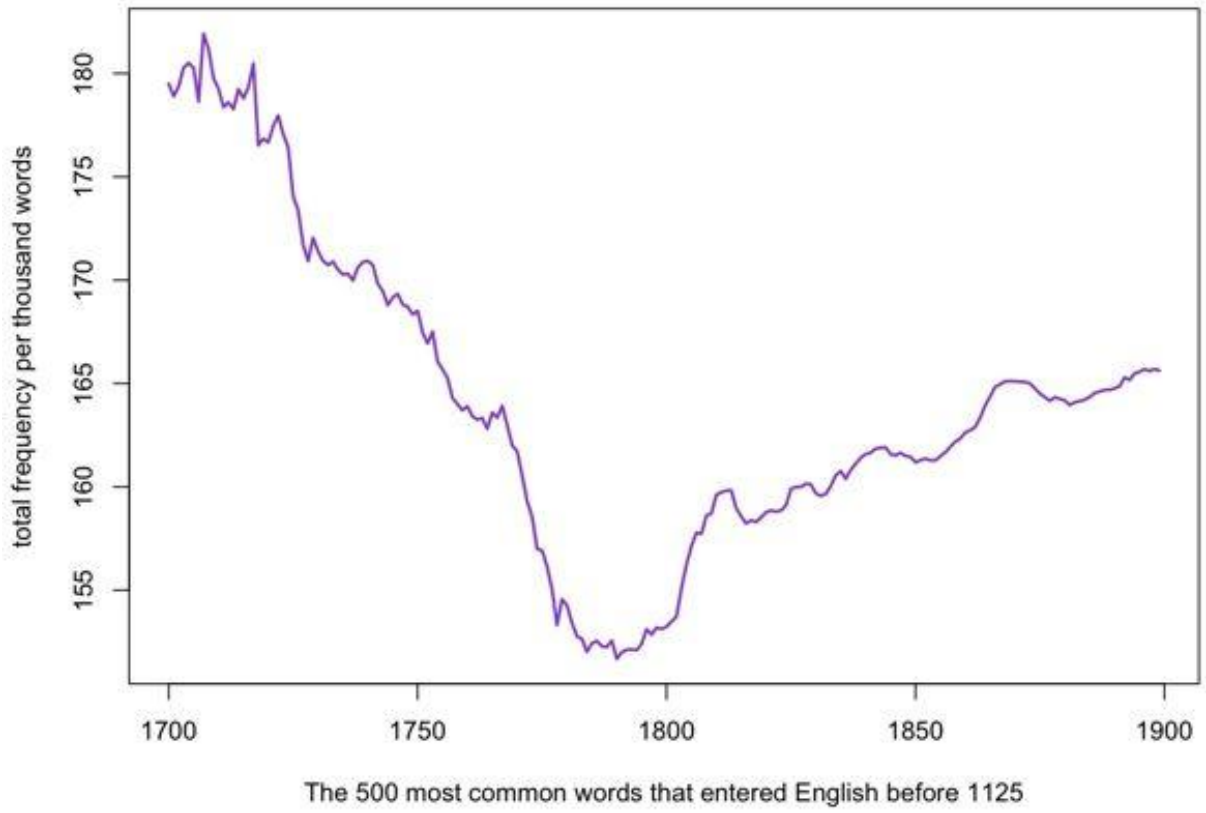
Etymology and Style

Ted Underwood, 2011

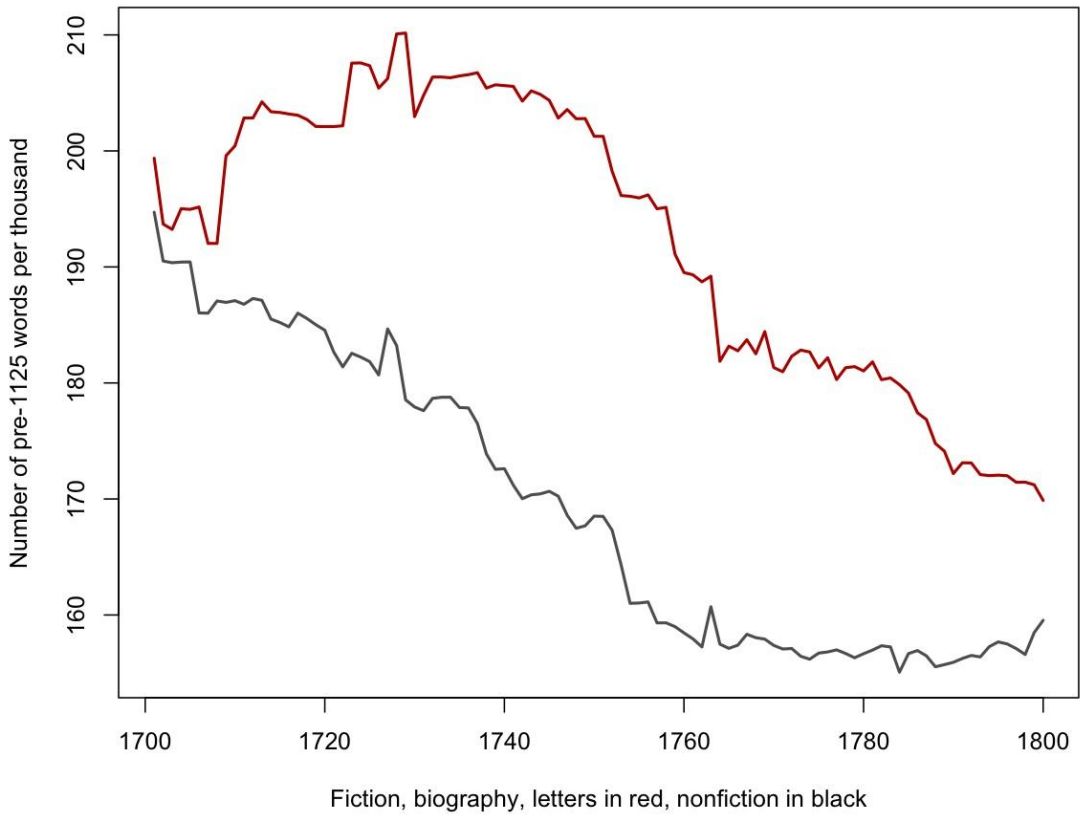
- English professors have a long, lively history of drawing specious conclusions from the "Latinate" or "Germanic" character of a particular writer's style.
- There is nevertheless good evidence that older words do predominate in informal, and especially spoken English. [Laly Bar-Ilan and Ruth A. Berman, "Developing register differentiation: the Latinate-Germanic divide in English," *Linguistics* 45 (2007): 1-35.]
- Can we use this fact to trace broad changes of register in the history of written English?

The fundamental distinction is not Latinate/Germanic, but date of entry. French was the written language for 200 years; words that entered English before that point had to be used in the spoken language to survive. This includes “Latinate” words like “street” and “wall.”

http://bit.ly/h8cJem



To understand the significance of the result, it needs to be broken down by genre. Initial results suggest that fiction and nonfiction prose both become more formal (less like speech) in the 18c. Drama and poetry change little, although older, less formal, “speechlike” words always predominate in drama.



Datum = Something Given

So, Ted's investigation concerns historical trends: as such, it is reasonable to think that it might be interesting to extend beyond 1900.

Can we do that? Only if we are given the data.