1851 to the present

A history of computational methods in the humanities

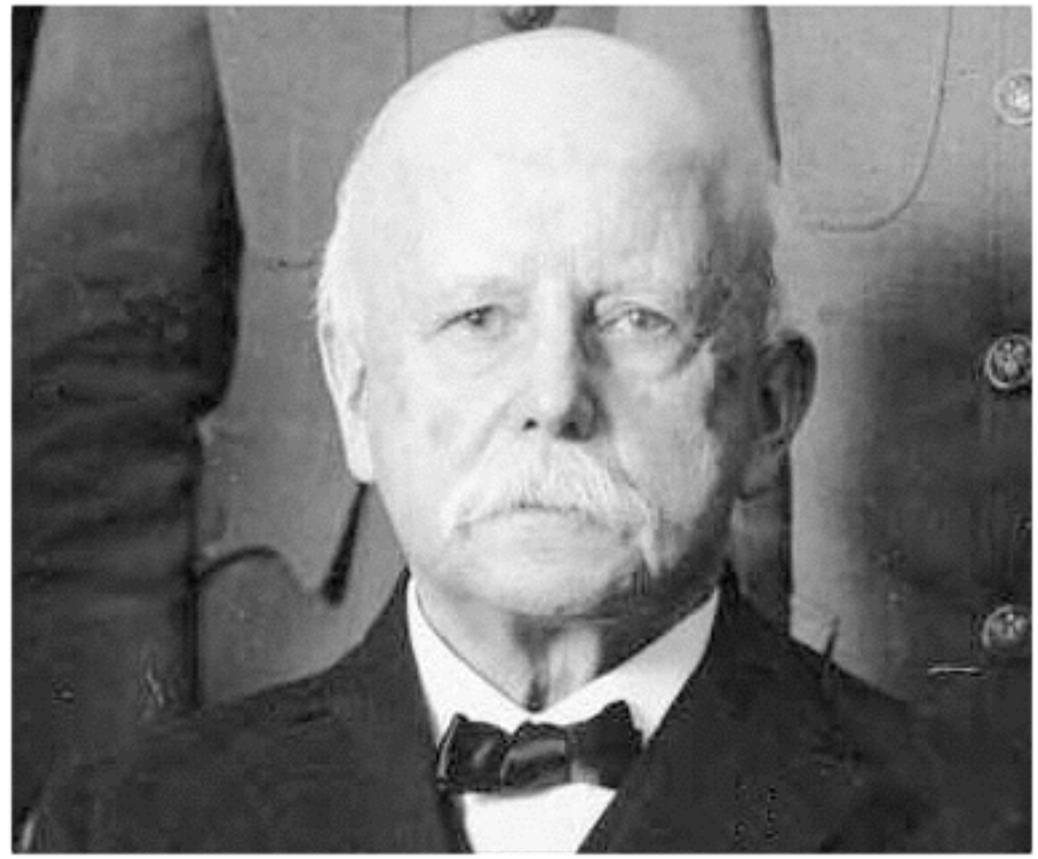
John Unsworth
Brandeis University
@
University of Miami
October 23rd, 2014

Augustus De Morgan

image from Wiki Commons

Women operating Burroughs adding machines

image from The Babbage Institute: http://purl.umn.edu/62887



Thomas Corwin Mendenhall

image from Wiki Commons

Not yet "humanities computing" or "digital humanities"

Stylometry

An arcane solution to a rare problem.

Agustus De Morgan

- Born 1806 in India. Educated at Oxford and Cambridge.
- Mathematician, logician, actuary, arithmetician
- Influenced by Boole
- Early Mathematics Faculty member at University College London
- A fan of the self-taught Indian mathematician Ramchundra
- Teacher of Ada Lovelace

Thomas Corwin Mendenhall

- Born 1841 in Ohio
- Primary school principal, Normal School graduate, high school teacher, Professor of physics and mechanics at Ohio State University, Visiting Professor (1878) at Tokyo Imperial University, Professor in the Signal Corps, Superintendent of the US Coast and Geodetic Survey, President of Worcester Polytechnic Institute, President of the American Association for the Advancement of Science.
- Measured gravity, systematically observed lightning, devised railroad signaling for weather information, calculated solar wavelengths, used statistical analysis to detect style in writing.
- Influenced by De Morgan

"A Memoir of Augustus de Morgan"

- Published by his wife, Sophia Elizabeth Morgan, in 1882
- Includes this letter to a friend, the Reverend W. Heald, from 1851

Count a large number of words in Herodotus—say all the first book—and count all the letters; divide the second numbers by the first, giving the average number of letters to a word *in that book.*

Do the same with the second book. I should expect a very close approximation. If Book I. gave 5·624 letters per word, it would not surprise me if Book II. gave 5·619. I judge by other things.

But I should not wonder if the same result applied to two books of Thucydides gave, say 5·713 and 5·728. That is to say, I should expect the slight differences between one writer and another to be well maintained against each other, and very well agreeing with themselves. If this fact were established there, if St. Paul's Epistles which begin with Παυλος gave 5·428 and the Hebrews gave 5·516, for instance, I should feel quite sure that the *Greek* of the Hebrews (passing no verdict on whether Paul wrote in Hebrew and another translated) was not from the pen of Paul.

If scholars knew the law of averages as well as mathematicians, it would be easy to raise a few hundred pounds to try this experiment on a grand scale. I would have Greek, Latin, and English tried, and I should expect to find that one man writing on two different subjects agrees more nearly with himself than two different men writing on the same subject. Some of these days spurious writings will be detected by this test. Mind, I told you so. With kind regards to all your family, I remain, dear Heald,

Yours sincerely,

A. De Morgan.

Count a large number of words in Herodotus—say all the first book—and count all the letters; divide the second numbers by the first, giving the average number of letters to a word *in that book.*

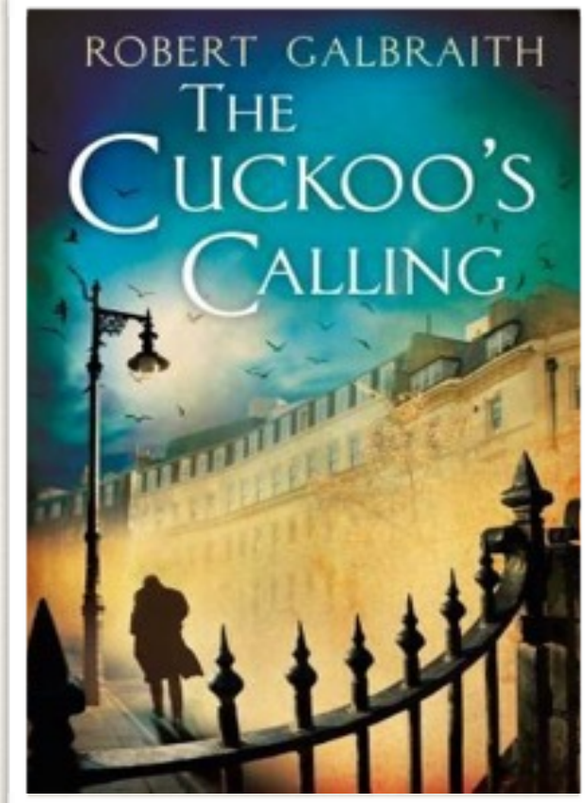
Do the same with the second book. I should expect a very close approximation. If Book I. gave 5·624 letters per word, it would not surprise me if Book II. gave 5·619. I judge by other things.

But I should not wonder if the same result applied to two books of Thucydides gave, say 5·713 and 5·728. That is to say, I should expect the slight differences between one writer and another to be well maintained against each other, and very well agreeing with themselves. If this fact were established there, if St. Paul's Epistles which begin with Παυλος gave 5·428 and the Hebrews gave 5·516, for instance, I should feel quite sure that the *Greek* of the Hebrews (passing no verdict on whether Paul wrote in Hebrew and another translated) was not from the pen of Paul.

If scholars knew the law of averages as well as mathematicians, it would be easy to raise a few hundred pounds to try this experiment on a grand scale. I would have Greek, Latin, and English tried, and I should expect to find that one man writing on two different subjects agrees more nearly with himself than two different men writing on the same subject. Some of these days spurious writings will be detected by this test. Mind, I told you so. With kind regards to all your family, I remain, dear Heald,

Yours sincerely,

A. De Morgan.



JK Rowling, identified by Patrick Juola et al. as "Robert Galbraith," based on stylometry.

"The Characteristic Curves of Composition"

- Published by TC Mendenhall in *Science* (11 March 1887: Vol. ns-9 no. 214S pp. 237-246) five years after *A Memoir of Augustus De Morgan* appeared

DOI:10.1126/science.ns-9.214S.237

- It begins: "Agustus DeMorgan somewhere remarks (I think it is in his 'Budget of paradoxes') that some time somebody will institute a comparison among writers in regard to the average length of words used in composition, and that it may be found possible to identify the author of a book, a poem, or a play, in this way."

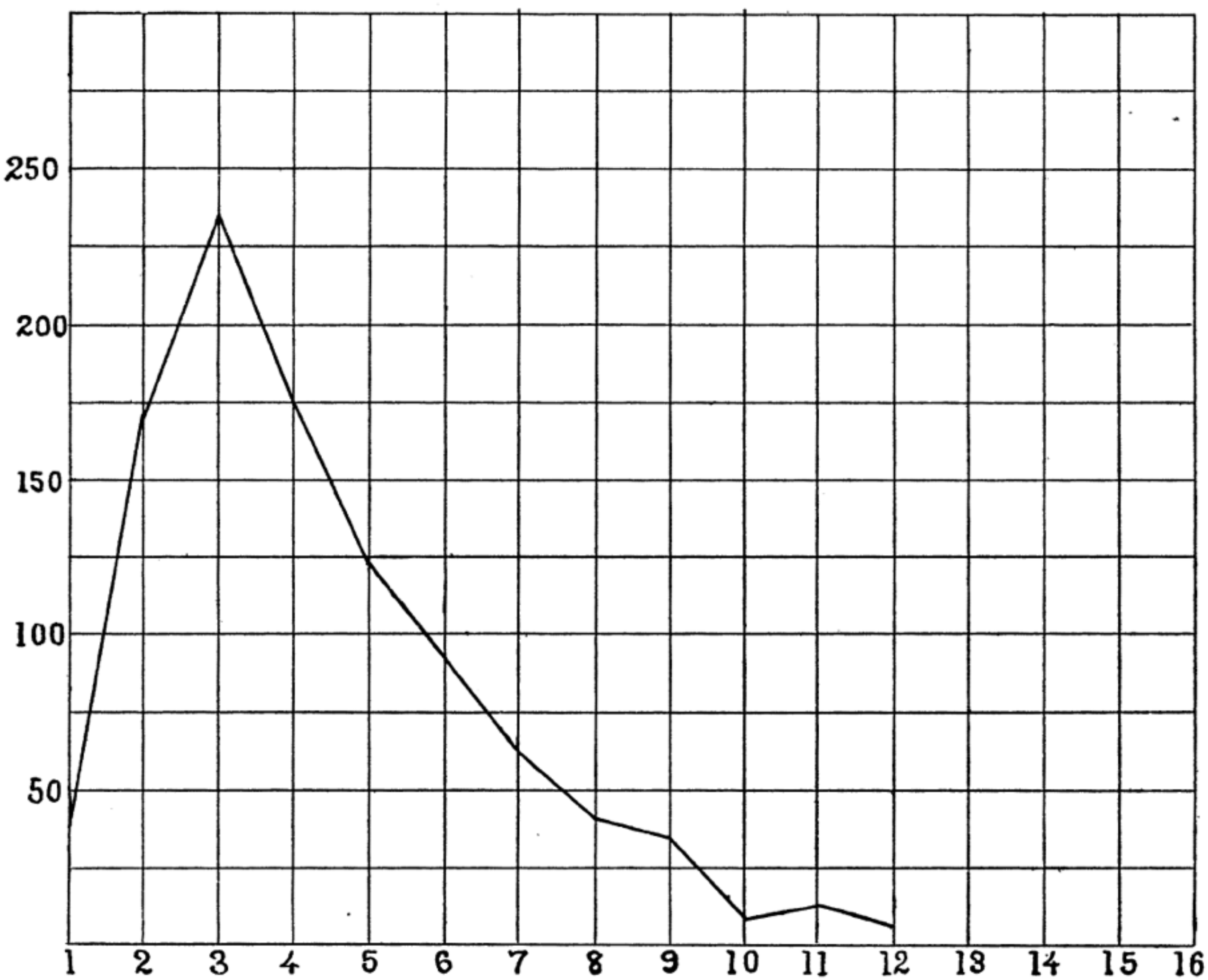


FIG. 1. — FIRST ONE THOUSAND WORDS IN 'OLIVER TWIST.'

Style as spectrum

- “The nature of the process is extremely simple, but it may be useful to point out its similarity to a well-known method of material analysis, the consideration of which actually first suggested to the writer its literary analogue.”

- “By the use of the spectroscope, a beam of non-homogenous light is analyzed, and its components assorted according to their wave-length…. So certain and uniform are the results of this analysis, that the appearance of a particular spectrum is indisputable evidence of the presence of the element to which it belongs.”

"A Mechanical Solution of a Literary Problem"

- Published by Mendenhall 14 years later, in *Popular Science Monthly* (1901, 60: 97-105)
- Describes an experiment in which "two ladies computed the number of words of two letters, three, and so on in Shakespeare, Marlowe, Bacon, and many other authors in an attempt to determine who wrote Shakespeare."
- The counting was carried out by "Mrs. Richard Mitchell and Miss Amy C. Whitman, of Worcester, Massachusetts," with a particular focus on the number of times the author used words of different lengths.


8 8 8 8 8 8 8
7 7 7 7 7 7 7
6 6 6 6 6 6 6
5 5 5 5 5 5 5
4 4 4 4 4 4 4
3 3 3 3 3 3 3
2 2 2 2 2 2 2

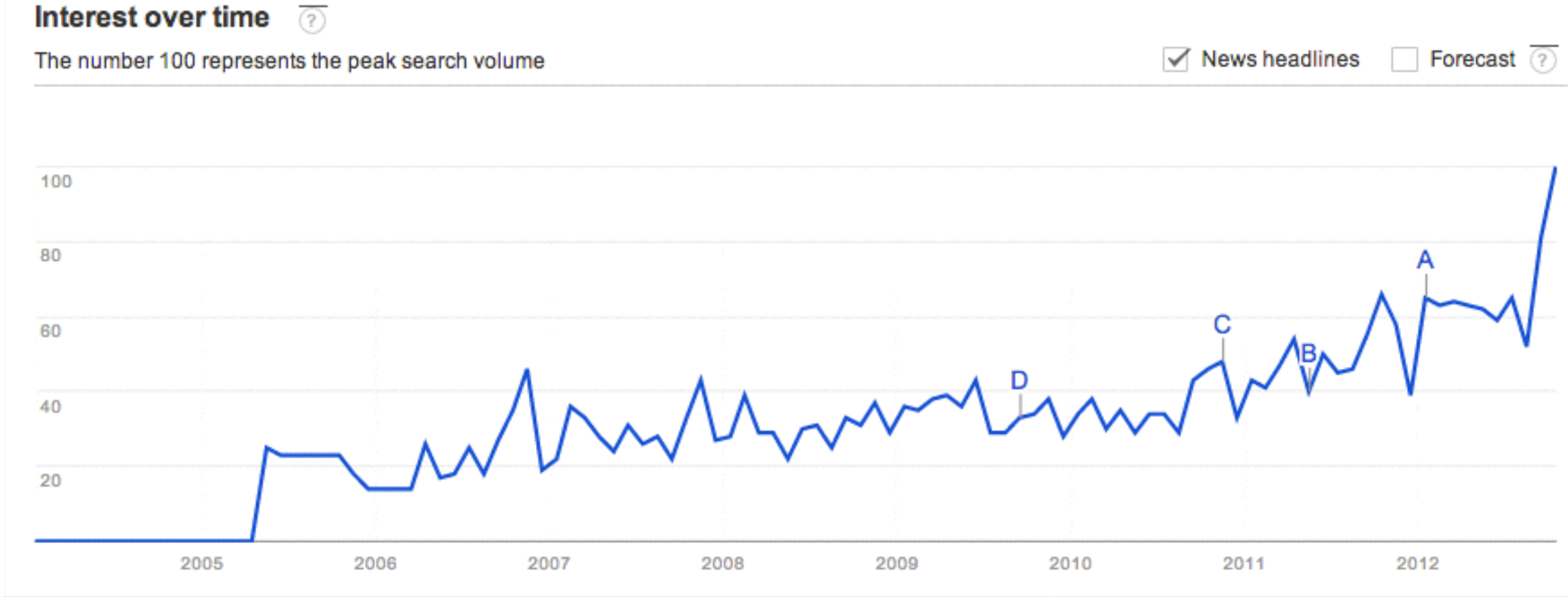
A simple counting machine

- "The operation of counting was greatly facilitated by the construction of a simple counting machine by which a registration of a word of any given number of letters was made by touching a button marked with that number."
- "One of the counters, with book in hand, called off 'five,' 'two,' 'three,' etc., as rapidly as possible … the other registering, as called, by pressing the proper buttons."

"After some preliminary work the counting of Shakespeare was seriously begun, and the result from the start with the first group of a thousand words was a decided surprise. Two things appeared from the beginning: Shakespeare's vocabulary consisted of words whose average length was a trifle below four letters, less than that of any writer of English before studied; and his word of greatest frequency was the four-letter word, a thing never met with before."

–TC Mendenhall, "A Mechanical Solution of a Literary Problem" Popular Science, 1901

For a contemporary discussion, see David Hoover, "Quantitative Analysis and Literary Studies," *A Companion to Digital Literary Studies,* ed. Susan Schreibman and Ray Siemens. Oxford: Blackwell, 2008.
http://goo.gl/lyvgZM



a term introduced in 2004

How did we get from Stylometry to Digital Humanities?

In Several Chapters

Textual Computing

- Computer Centers, late 1940s through the present (Tuebingen, Oxford, NCSA)
- Scholarly Societies and Journals, mid-1960s through the present (ACH, ALLC)
- Standards efforts, late 1980s to present (TEI)
- Library Digitization & Digital Humanities Centers, 1990s to present (esp. research libraries)
- Big Data, Google Books, the HathiTrust

Before going further, it needs to be said:

DH is no longer just about text and numbers

- Image analysis (including spectral)
- Maker Labs
- Music information retrieval
- Embodied computing
- Embedded computing
- Multimedia
- Retro computing
- Gaming
- Design
- Community Informatics

Historically, though, that's where it began:

Father Busa

- In his doctoral dissertation, published in 1949, Roberto Busa concentrated on the concept of presence in the works of Thomas Aquinas. He wrote out by hand 10,000 3x5" cards each containing a sentence with the word "in" or a word connected with "in."
- In the same year, he visited Thomas Watson and enlisted IBM in helping him construct a punch-card index of every word in the works of St. Thomas Aquinas (11M words)

Textual Computing

...from the perspective of miniaturization:

"I began, in 1949, with only electro-countable machines with punched cards. My goal was to have a file of 13 million of these cards, one for each word, with a context of 12 lines stamped on the back. The file would have been 90 meters long, 1.20 m in height, 1 m in depth, and would have weighed 500 tonnes."

Textual Informatics

- “I call the first current “documentary," in memory of the American Documentation Society, and of the Deutsche Gesellschaft für Dokumentation in the 1950s. It includes databanks, the Internet, and the World Wide Web, which today are the infrastructures of telecommunications and are in continuous ferment.
- The second current I call "editorial." This is represented by CDs and their successors, including the multimedia ones, a new form of reproduction of a book, with audio-visual additions.. . .
- I call the third current "hermeneutic" or interpretative, that informatics most associated with linguistic analysis”

Deus ex Machina

Miniaturization and Informatics: Room for Data

"In His mercy, around 1955, God led men to invent magnetic tapes. The first were the steel ones by Remington, closely followed by the plastic ones of IBM. Until 1980, I was working on 1,800 tapes, each one 2,400 feet long, and their combined length was 1,500 km, the distance from Paris to Lisbon, or from Milan to Palermo."

— *Busa, "Foreword," Blackwell Companion to Digital Humanities*

The medium of representation

"At this time much attention was paid to the limitations of the technology. Data to be analysed was either texts or numbers. It was input laboriously by hand either on punch cards with each card holding up to eighty characters or one line of text (uppercase letters only), or on paper tape where lower case letters were perhaps possible but which could not be read in any way at all by a human being. . . . Character-set representation was soon recognized as a substantial problem. . . . Various methods were devised to represent upper and lower case letters on punched cards, most often by inserting an asterisk or similar character before a true upper case letter. Accents and other non-standard characters had to be treated in a similar way and non-roman alphabets were represented entirely in transliteration."

Representation and discipline

"Most large-scale datasets were stored on magnetic tape, which can only be processed serially. It took about four minutes for a full-size tape to wind from one end to the other and so software was designed to minimize the amount of tape movement. Random access to data such as happens on a disk was not possible. Data had therefore to be stored in a serial fashion. This was not so problematic for textual data, but for historical material it could mean the simplification of data, which represented several aspects of one object (forming several tables in relational database technology), into a single linear stream. This in itself was enough to deter historians from embarking on computer-based projects."

—*Susan Hockey, "A History of Humanities Computing," Companion to Digital Humanities, ed. Susan Schreibman, Ray Siemens, John Unsworth. Oxford: Blackwell, 2004.*
http://www.digitalhumanities.org/companion/

EDSAC II at Cambridge, 1960

The 60s & 70s

Organizing a Field:

Centers, Associations,
Journals, Projects

From Projects to Centers

Most early DH centers grew out of specific projects—Wisbey's Center for Literary and Linguistic Computing at Cambridge and Wilhelm Ott's metrical analysis with Fortran in the 1960s, Bob Kraft's Computer Assisted Tools for Septuagint Studies begun in the early 1970s at Penn, or the Perseus Project begun at Harvard in 1985. Often these projects were driven by the vision and dedication of a single individual.

Starting a Center (1985)

"If a college has two or three faculty committed to humanities computing, for whatever reasons, it has what's needed to get started. From that point on, centres of quite different characters take root. Several models operate successfully throughout North America. They develop according to the professional goals of those faculty and so any one cannot easily be taken as "the best way" to found a centre. To administrators … the argument that computing humanists will better enable their institutions to meet society's needs will be almost universally admitted. This is especially true now that the novelty of seeing humanities faculty using computers has been exhausted and it is no longer "innovative" (in a national or an international community) to set up humanities computing centres."

— Ian Lancashire, http://www.digitalhumanities.org/humanist/Archives/Virginia/v02/0338.html

1962-1978

- 1962: Association for Machine Translation and Computational Linguistics (AMTCL) founded; became the Association for Computational Linguistics in 1968.
- 1963: Roy Wisbey founded the Centre for Literary and Linguistic Computing in Cambridge to support his work with Early Middle High German Texts.
- 1966: Computers and the Humanities founded
- 1968: David Packard's Concordance to Livy.
- 1970: The first instance of what later became the Association for Literary and Linguistic Computing/Association for Computers and the Humanities (ALLC/ACH) joint annual conference held at the University of Cambridge
- 1972?: Bob Kraft at Penn establishes Computer Assisted Tools for Septuagint Studies
- 1973: Founding of The Association for Literary and Linguistic Computing
- 1978: Founding of the Association for Computers and The Humanities; TUSTEP gets its name

The 80s and 90s

The Golden Age of Projects

IBM personal computer, 1980s

1980s and 90s

- 1985: Perseus Project begun at Harvard
- 1986: Literary and Linguistic Computing founded
- 1986: SGML specification released
- 1987: Text-Encoding Initiative, Humanist begun
- 1991: Electronic Beowulf Project
- 1992: H-Net founded
- 1993: Mosaic released, IATH founded at Virginia, STG founded at Brown, University of Michigan Digital Library formed
- 1994: First edition of the TEI guidelines; Center for History and New Media founded
- 1996: First draft of XML spec released
- 1996: Brewster Kahle founds the Internet Archive (bless him)

IBM Server, 1990s

Model	IBM POWERServer 560
CPU	POWER 50MHz
Ram	192 MB
Disk	9100 MB (IBM DDRS-39130) 4560 MB (IBM DDRS-34560) 13660 MB Total disk space.
Network	10 Mbit Ethernet 100 Mbit SAS-FDDI
OS	AIX 4.3
Other	2.3 GB Exabyte 8200 Tapedrive
Manufactured	From: 1992-03-27 To: 1993-12-21
Dimensions	Width: 360mm Depth: 675mm Height: 610mm
Weight	36.7kg to 53.1kg
Events	*2005-07-01: maswan* Ersatt av hatchepsut under våren.

The hand-held computer, circa 2007

And today, the iPhone 6 has a processor that is *28 times faster* than that IBM 560, with *5 times as much RAM* and *9.3 times the storage* capacity, at ***one quarter of a percent*** of the weight.

Big Data, Big Deals, Big Risks

2000 onwards

Digital Humanities at Scale

2001-2011

- 2001: TEI incorporated as a non-profit membership organization
- 2002: The Million Book Project launched at Carnegie Mellon
- 2003: HASTAC founded
- 2004: Google Books announced at the Frankfurt Book Fair
- 2004: Blackwell Companion to Digital Humanities
- 2006: ACLS report on Cyberinfrastructure for Humanities and Social Sciences
- 2006: NEH Office of Digital Humanities opens
- 2007: The Million Book Project wraps up
- 2007: The MONK Project (text-mining workbench with 150M words of literary text in English, 1600-1923)
- 2008: Planning begins for the HathiTrust Research Center, established in 2011

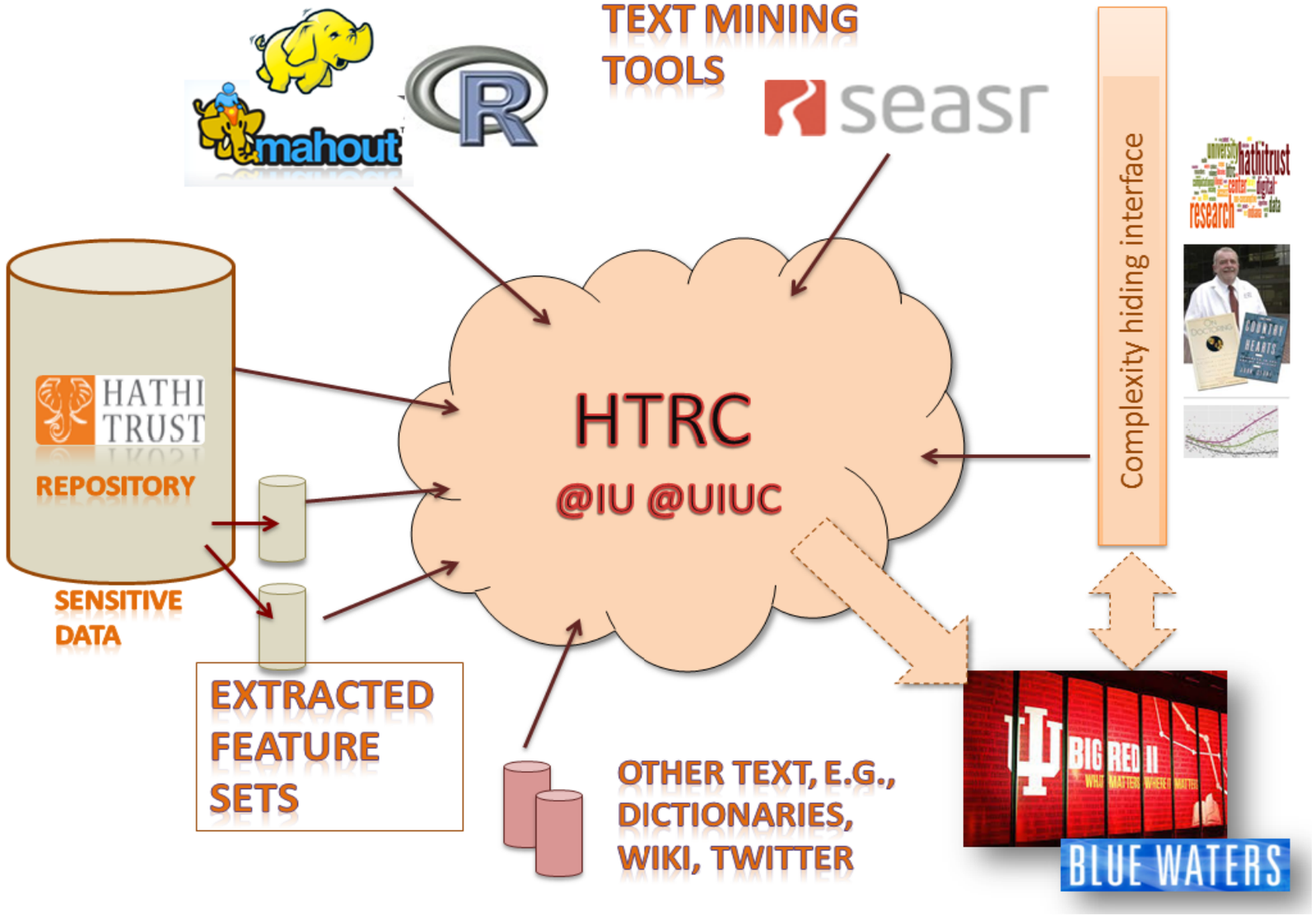
Words Jane Austen Avoids

How Much Data?

- 12,398,938 total volumes
- 6,375,777 book titles
- 325,433 serial titles
- 4,339,628,300 pages
- 556 terabytes
- 147 miles
- 10,074 tons
- 4,533,763 volumes(~37% of total) in the public domain

Actual elephant at Leith Public Library in Scotland

HTRC v2.0

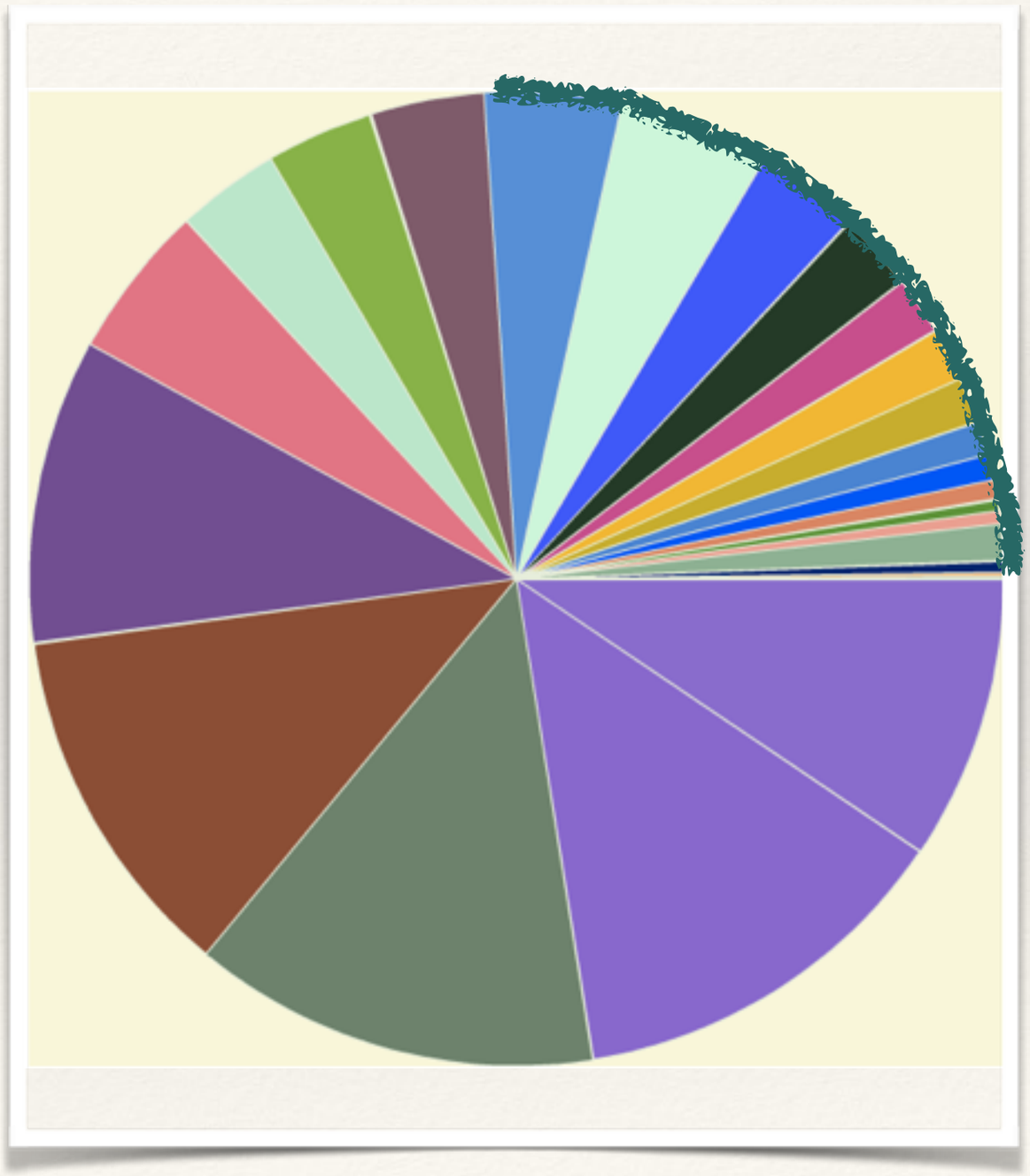

mahout
TEXT MINING TOOLS
seasr
HATHI TRUST
REPOSITORY
SENSITIVE DATA
HTRC
@IU @UIUC
EXTRACTED FEATURE SETS
OTHER TEXT, E.G., DICTIONARIES, WIKI, TWITTER
Complexity hiding interface
BIG RED II
BLUE WATERS

HathiTrust Research Center

- The HathiTrust Research Center (HTRC) is a collaborative research center launched jointly by Indiana University and the University of Illinois at Urbana-Champaign, along with the HathiTrust Digital Library to help meet the technical challenges that researchers face when dealing with massive amounts of digital text.
- The HTRC is focused on developing cutting-edge software tools, services, and cyberinfrastructure to enable advanced computational access to the growing digital record of human knowledge. Leveraging data storage and computational infrastructure at Indiana University and the University of Illinois at Urbana-Champaign, the HTRC is provisioning a secure computational and data environment for scholars to perform research using the HathiTrust corpus.
- The center is breaking new ground in the areas of text mining and non-consumptive research that will allow scholars to fully utilize content of the HathiTrust Library while preventing intellectual property misuse within the confines of current U.S. copyright law.

Early Work at Scale

- Initial work is being done with the out-of-copyright component of the HT collection, particularly that produced by libraries rather than by Google.



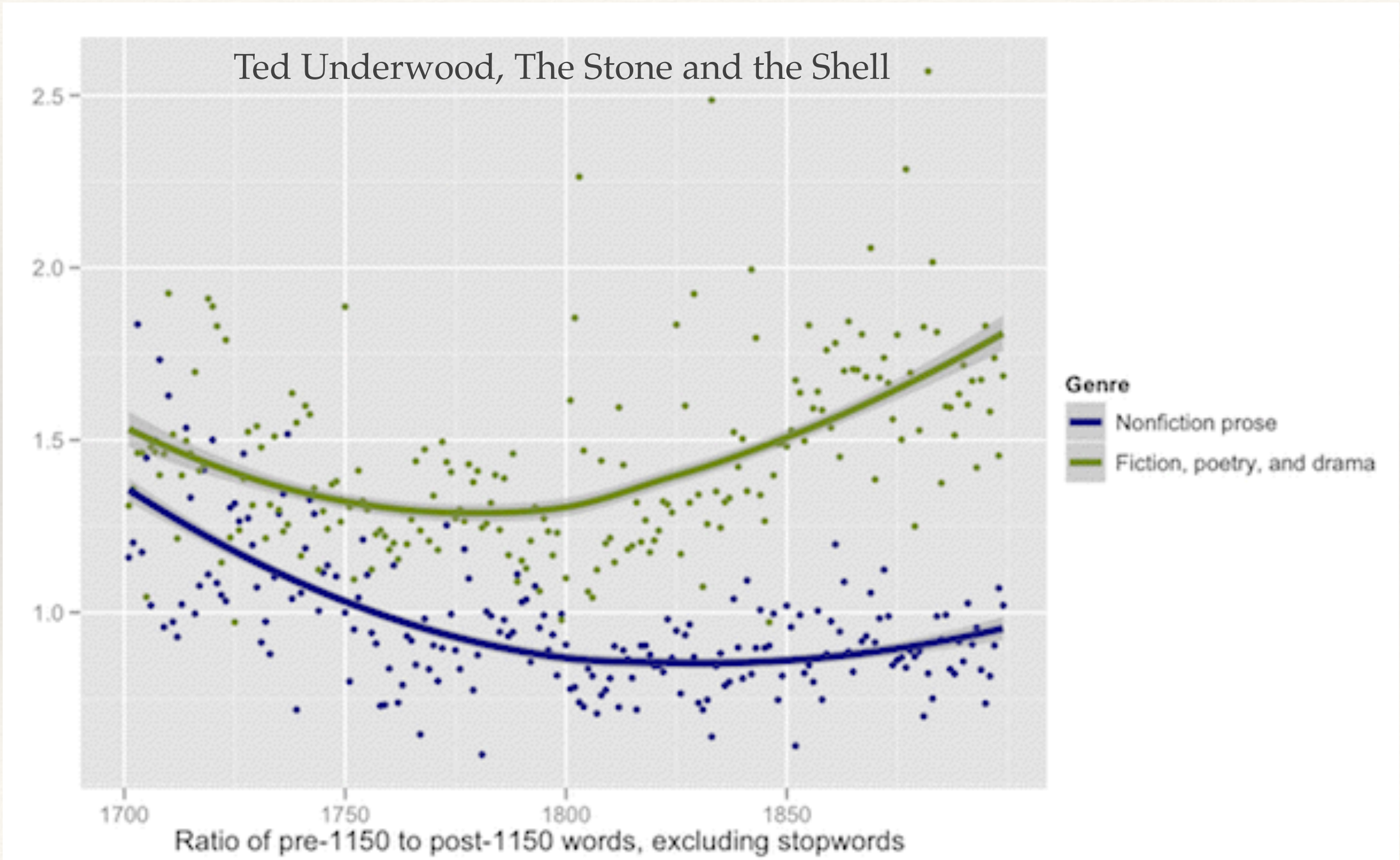
Images from Internet Archive Books

- Kalev Leetaru, GSLIS alumnus, figured out how to locate and extract images from all of the books in the Internet Archive, and used a variety of metadata, plus nearby text, to tag them and make them searchable on Flickr

Workset Creation

- What materials do you have that pertain to Japan? How many volumes are in Japanese?
- How would we gather up all the works that deal with Francis Bacon? How about those contemporaries with whom he worked?
- What musical scores are in the corpus? What works contain music notation?
- Which works have back of book indexes that I might analyze?
- How would I gather works by 16th-century women? By 19th-century men?
- Which works are fiction? Which are non-fiction? Which are commentaries? Essays? Poetry? Prose?
- How would I gather together all the images of Victorian England?
- How would I gather works similar to those that I currently I have in hand? Can I define different kinds of similarity?

Correlation of Etymology and Genre in English



Computer Science > Computation and Language

Mapping Mutable Genres in Structurally Complex Volumes

Ted Underwood, Michael L. Black, Loretta Auvil, Boris Capitanu

(Submitted on 12 Sep 2013 (v1), last revised 18 Sep 2013 (this version, v2))

To mine large digital libraries in humanistically meaningful ways, scholars need to divide them by genre. This is a task that classification algorithms are well suited to assist, but they need adjustment to address the specific challenges of this domain. Digital libraries pose two problems of scale not usually found in the article datasets used to test these algorithms. 1) Because libraries span several centuries, the genres being identified may change gradually across the time axis. 2) Because volumes are much longer than articles, they tend to be internally heterogeneous, and the classification task needs to begin with segmentation. We describe a multi-layered solution that trains hidden Markov models to segment volumes, and uses ensembles of overlapping classifiers to address historical change. We test this approach on a collection of 469,200 volumes drawn from HathiTrust Digital Library. To demonstrate the humanistic value of these methods, we extract 32,209 volumes of fiction from the digital library, and trace the changing proportions of first- and third-person narration in the corpus. We note that narrative points of view seem to have strong associations with particular themes and genres.

Comments: Preprint accepted for the 2013 IEEE International Conference on Big Data. Revised to include corroborating evidence from a smaller workset

Subjects: **Computation and Language (cs.CL)**; Digital Libraries (cs.DL)

Cite as: **arXiv:1309.3323 [cs.CL]**

(or **arXiv:1309.3323v2 [cs.CL]** for this version)

Submission history

From: Ted Underwood [view email]

[v1] Thu, 12 Sep 2013 22:27:59 GMT (174kb,D)

[v2] Wed, 18 Sep 2013 17:37:27 GMT (152kb,D)

Additional Obscure Readings

- *Analytics of literature, a manual for the objective study of English prose and poetry,* by L. A. Sherman. 1893.
- Conrad Mascol, aka William B. Smith, 1888; C. Mascol, "Curves of Pauline and Pseudo-Pauline Style I," *Unitarian Review* 30 (November 1888): 452–60.
- Conrad Mascol, "Curves of Pauline and Pseudo-Pauline Style II," *Unitarian Review* 30 (December 1888: 539–46).
- Tasman, Paul. 1958. “Indexing the Dead Sea Scrolls by Electronic Data Processing.” New York: IBM Corporation, 12; cited in *Formatting The Word of God*, An Exhibition at Bridwell Library Perkins School of Theology, Southern Methodist University, October 1998 through January 1999. Ed. Valerie R. Hotchkiss and Charles C. Ryrie. Bridwell Library: Dallas, Texas, 1998.