From Searching to Researching

Three stages of digital scholarship

Finding: we make catalogs and enable searching so that scholars can find the objects they wish to study

Reading: the reading or examination is done remotely

Analyzing: algorithms find the result that we want

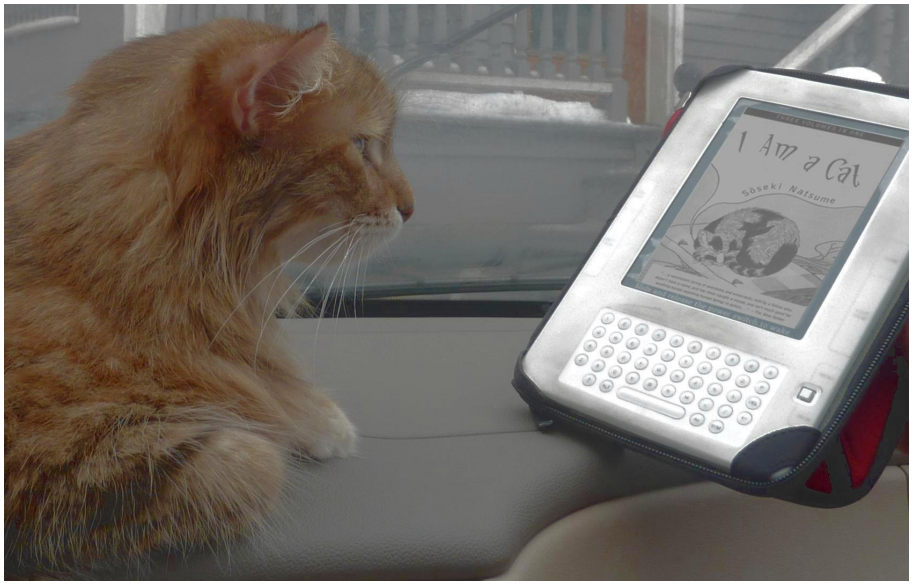
Searching

Starting with Vannevar Bush, we had the idea of machines finding materials for us to read. In the 1950s and 1960s we developed digital methods for text searching.

At first, after searching people would then read on paper. Even in 1990 my colleagues interviewed chemists who said they liked not only the format and content of paper journals, but even the feel and smell of the publication.

Reading

Today, reading online is normal. In 2011 Amazon started reporting more Kindle sales than paper sales for books. Many scholarly resources are online-only. The chemical journals mentioned on the last slide are discontinuing paper editions. Even some art galleries are now selling on-line only.



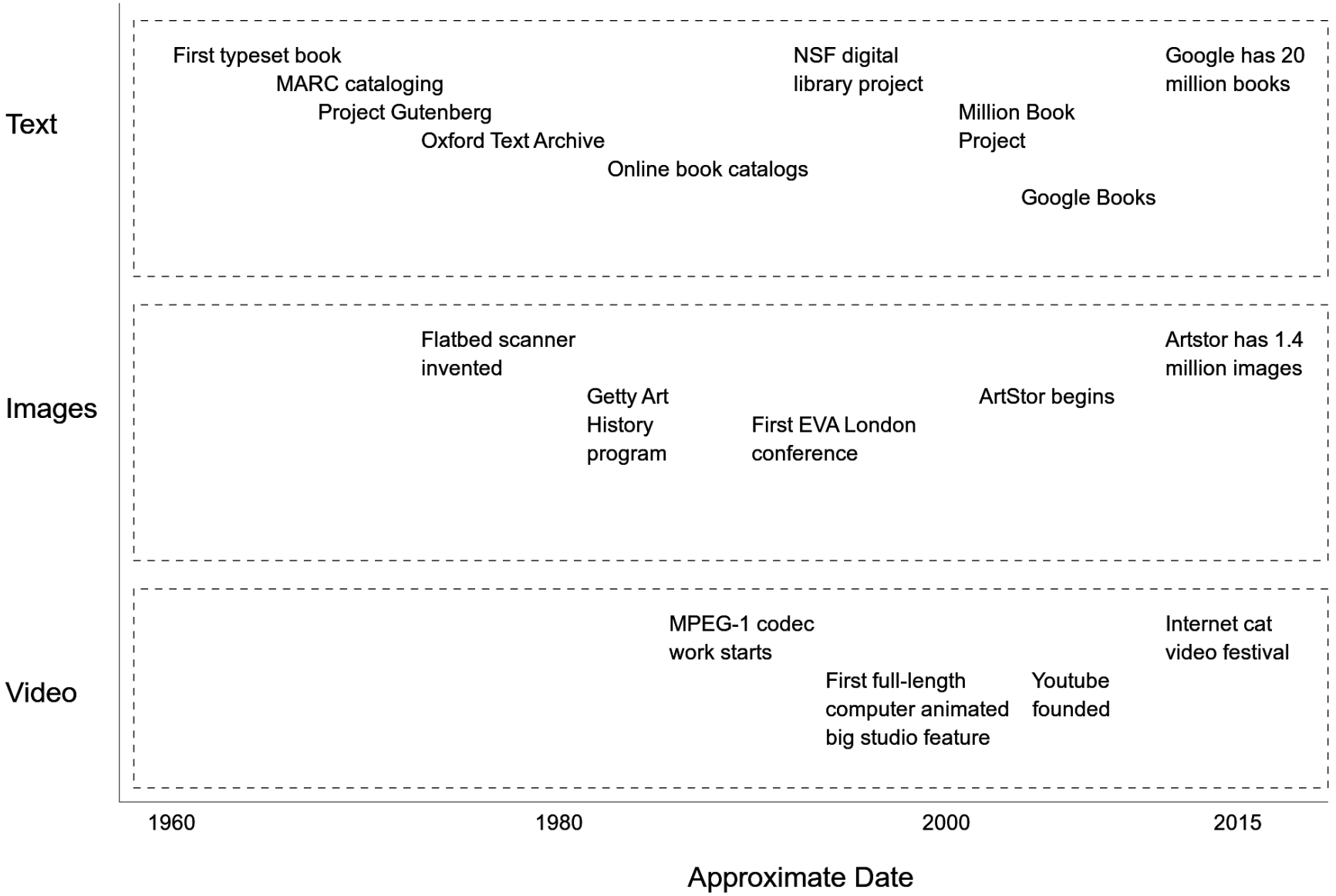
There are signs that the growth of e-reading has leveled off, although we're not at "peak Kindle." Some reports suggest people remember better what they read on paper.

Analysis

Today, algorithms assist scholars in reaching conclusions from materials. Google Ngram searches show word use over time, and Google Trends shows both time and area of use. Formal encoding of text properties helps such research. Authorship studies were among the earliest uses of digital texts.

Citations and twitter are now used in faculty evaluation. Soon, “likes” will count and Nora the piano-playing cat will get tenure.

Collections in different media



Scanning progress

The Internet Archive “Scribe” scanner. We are now so good at scanning books that the average 19th century U.S. book has been scanned half a dozen times.

Reading online

Both Amazon sales figures and observing any train or plane confirms that reading has shifted from paper to screens.

Will this mean that people read only “snippets”?

This is not a new complaint. Plato (in the *Phaedrus*) complained that writing was an excuse for not memorizing works. In 1987 I heard a lecturer complain that tourists took photographs instead of drawing sketches.

There are more books published today than ever before; this means that there will be fewer readers per book. But each reading may be of better selected material.

Authorship studies

The first applications of "stylometrics" were authorship studies, with Mosteller's work on the *Federalist Papers* leading the way. Mosteller and Wallace worked by hand, but nowadays machines do similar counts of the individual preferences of authors for syntactic structures, sentence lengths or individual words. Other studies trace influence and describe style.

In many cases, these are the same goals, but different methods.

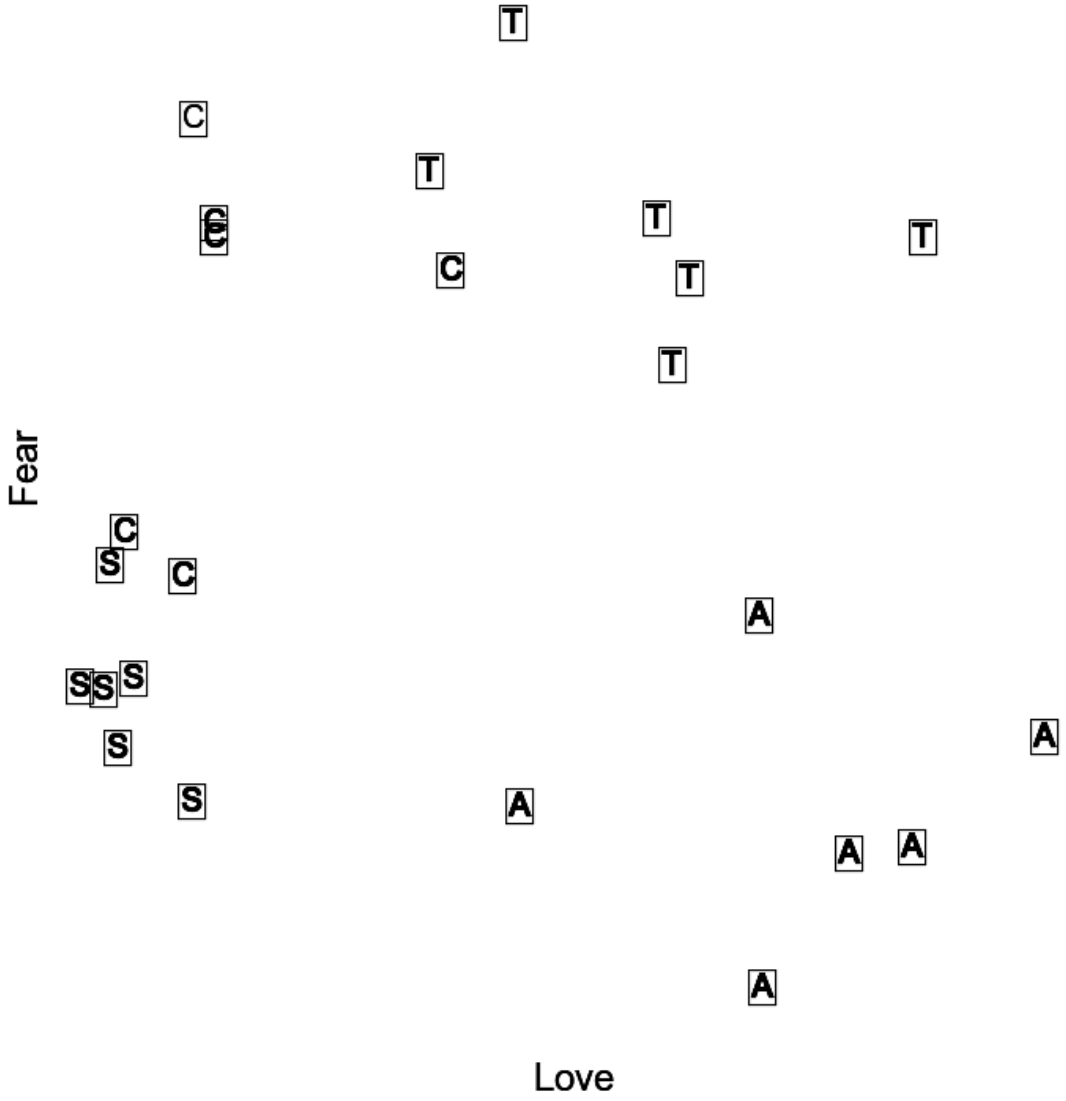
Sentiment analysis is used not only in stylometrics but for marketing and intelligence work.

Uses of *love* words vs. *fear* words

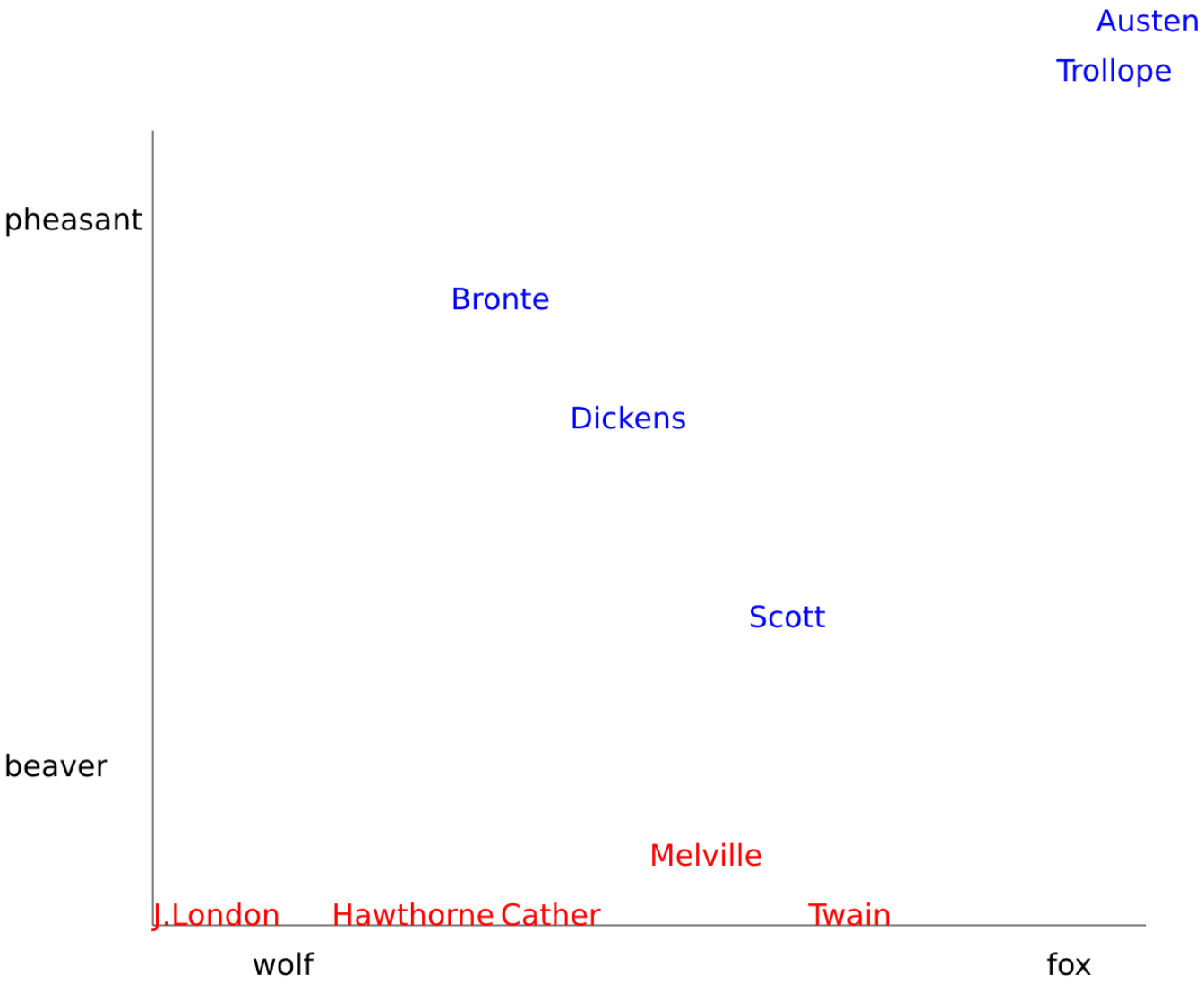
Counting words from the "love" category in a 1913 Roget thesaurus compared with "fear" words. Authors:

A: Jane Austen
C: Willkie Collins
S: Sir Walter Scott
T: Anthony Trollope

(6 novels each)



Stylistic cues to location



Counting *wolf* vs. *fox* and *pheasant* vs. *beaver* in various authors

Tracking “ideas” through text

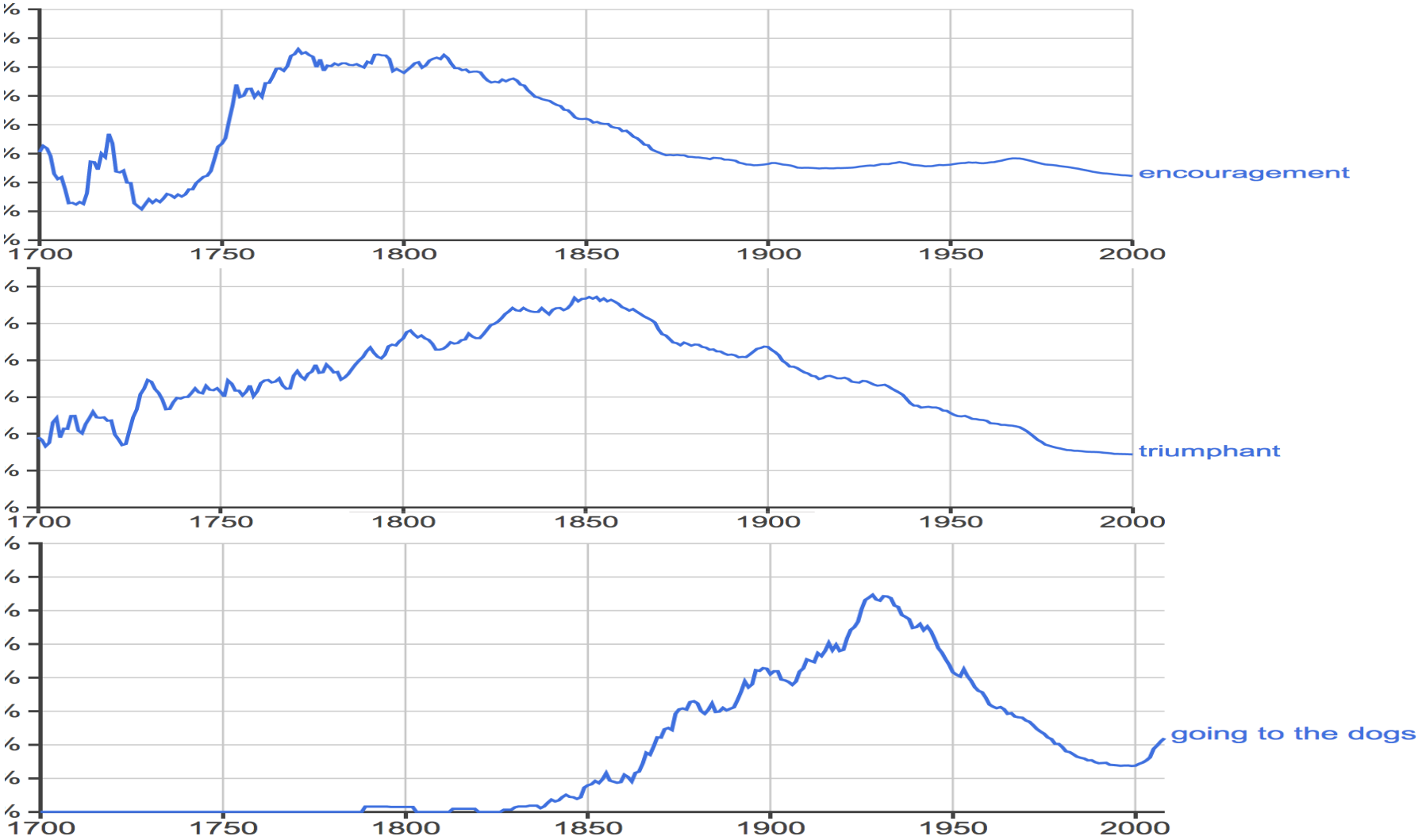
Schilit and Kolak, in “Exploring a Digital Library through Key Ideas,” tracked quotations and references through Google Books.

The question of relating words to concepts faces problems of both synonymous words and ambiguous words.

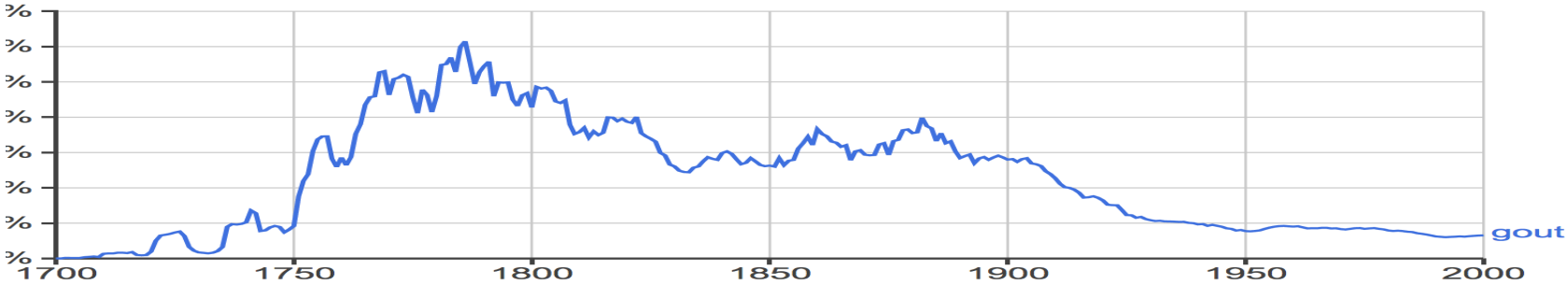
Unfortunately, it’s still difficult for people outside Google to do some of this work.

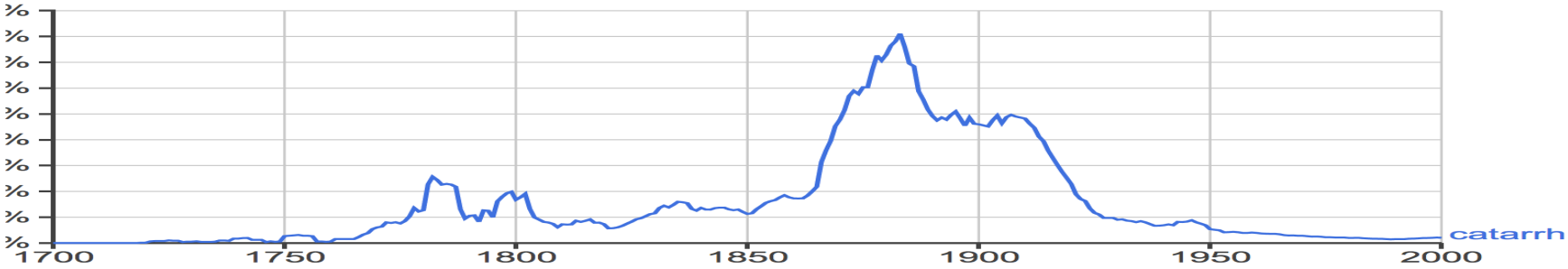
Intellectual history in Ngrams

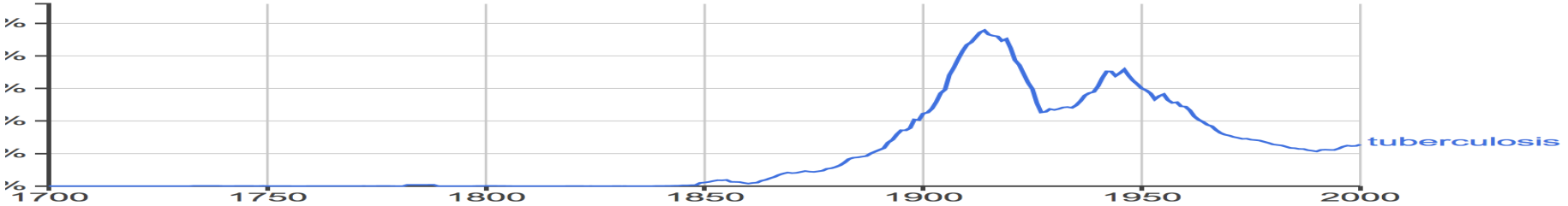
The 18th century things was optimistic, the 19th century thought things were perfect, and the 20th century was a downer.

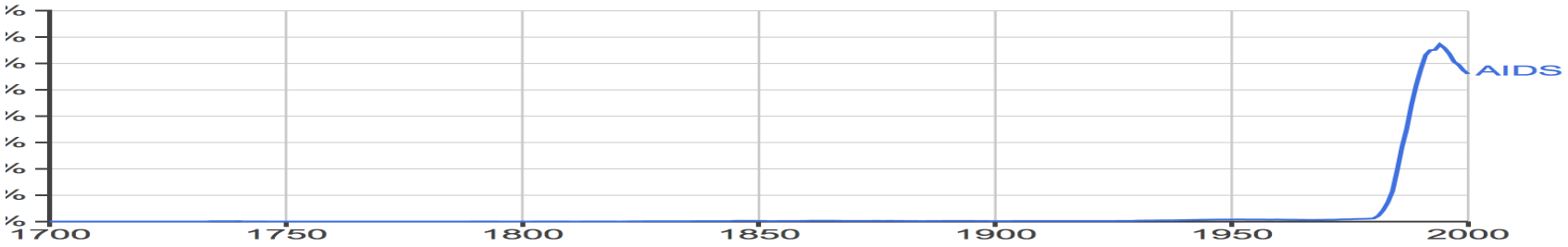


Medical changes in Ngrams

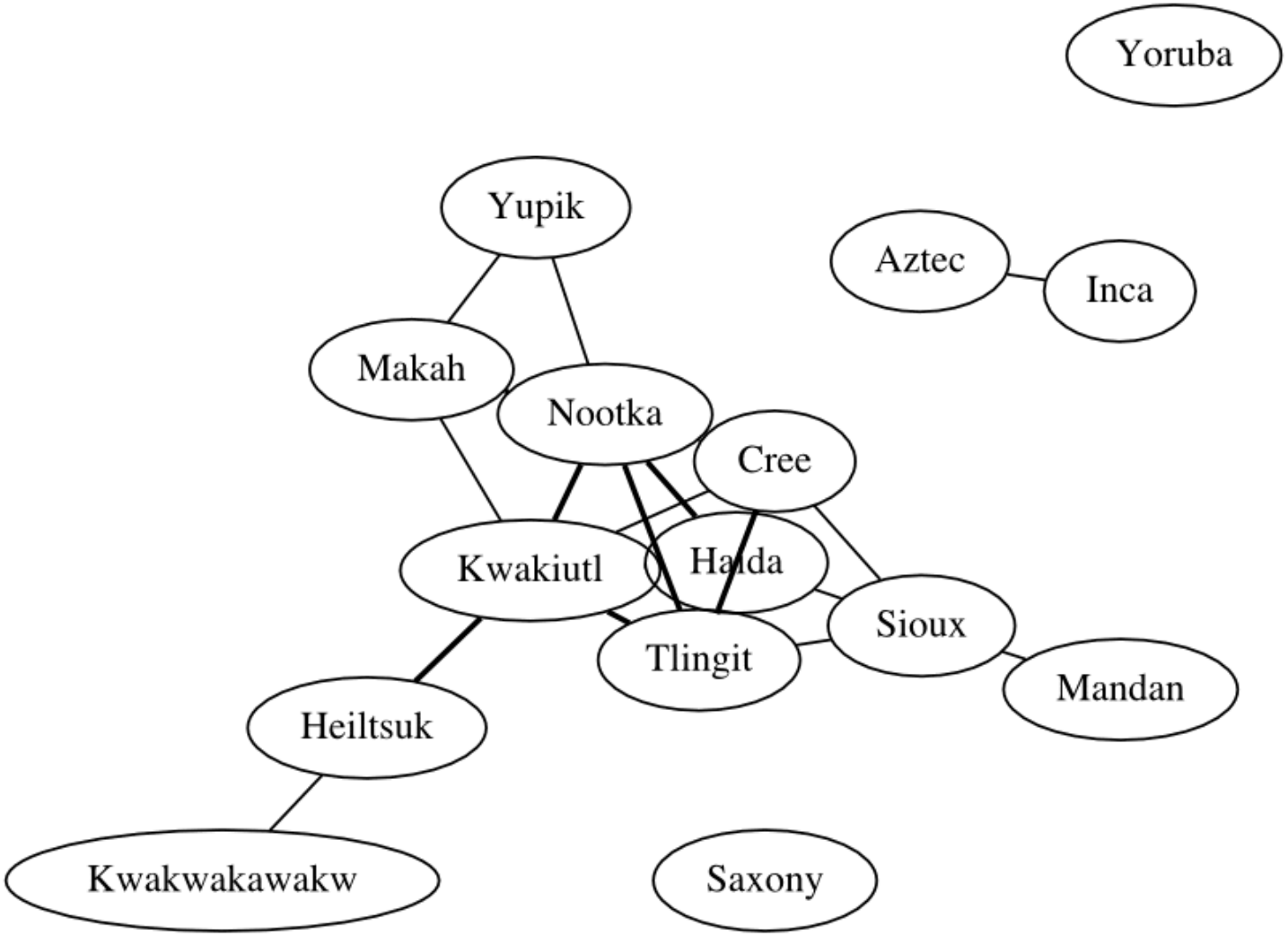








Links found in catalog text



Overlaps of words in descriptions of objects in the British Museum used to cluster the words.

Sense disambiguation

```
coal
 1 a piece of glowing carbon or charred wood : ember
 2 charcoal
 3 a black or brownish black solid combustible substance
   formed by the partial decomposition of vegetable
   matter without free access of air and under the
   influence of moisture and often increased pressure
   and temperature that is widely used as a natural fuel
   pieces or a quantity of the fuel broken up for burning
coal
 1 to burn to charcoal : char
 2 to supply with coal
 0 to take in coal
```

```
ash
 1 any of a genus (Fraxinus) of trees of the olive family
   with pinnate leaves, thin furrowed bark, and gray
   branchlets
 2 the tough elastic wood of an ash
ash
 1 the solid residue left when combustible material is
   thoroughly burned or is oxidized by chemical means
   fine particles of mineral matter from a volcanic vent
 2 ruins
 3 the remains of the dead human body after cremation or
   disintegration
 4 something that symbolizes grief, repentance, or
   humiliation
 5 deathly pallor
ash
 0 to convert into ash
```

Using dictionary definitions to determine that the meaning of *ash* in the phrase *coal ash* should not be the tree. Illustration from 1986 paper.

Not really used. (a) We now have much larger text collections to do similar tasks; (b) there is no standard dictionary or thesaurus to label the senses; (c) people work harder on recall tools than precision tools. Wordnet is perhaps the most accepted standard for senses, but it's limited compared with a big dictionary.

Drifting meanings over time

Senses of **Train** In Different Periods

Date	Work	Sense								
		N1	N2	N3	N4	N5	V1	V2	V3	Other
c1600	Shakespeare	0	50%	0	6%	0	44%	0	0	
1611	*King James Bible*	0	50%	0	0	0	50%	0	0	
1719	*Robinson Crusoe*	0	0	0	0	100%	0	0	0	
17??	ESTC*	0	4%	21%	0	0	53%	18%	0	4%
1760	*Tristram Shandy*	0	38%	62%	0	0	0	0	0	
1813	*Pride & Prejudice*	0	0	100%	0	0	0	0		
1851	*Moby Dick*	33%	0	0	0	0	17%	0	0	50%
c1900	Sherlock Holmes	46%	4%	22%	0	0	28%	0	0	
c1960	Brown corpus	23%	3%	0	2%	0	72%	0	1%	–
1988	AP wire (Feb 2)	54%	0	0	0	0	43%	0	2%	

* ESTC = Eighteenth Century Short Title Catalog.

Summary: N1 *railway train*, N2 *wagon train*, N3 *train of thought*, N4 *dress*, N5 *gunpowder*, V1 *teach*, V2 *prune*, V3 *aim*.

Synonymy is the reverse problem

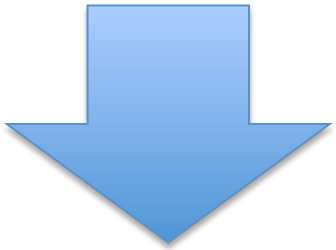
	What do you do to a gun?				
Source	*fire*	*empty*	*discharge*	*operate*	*shoot*
Robinson Crusoe	16				
Tristram Shandy			1		
Moby Dick			2		
Sherlock Holmes	2	2			
Brown corpus	4			1	1

Sense disambiguation gone bad

She taught 20 courses based on seven different preparations.

Google Translate,
English➔Russian

Она учила 20 курсов на основе семи различных препаратов.



Google Translate,
Russian➔English

She taught 20 courses on the basis of seven different drugs.

Next: Pictures

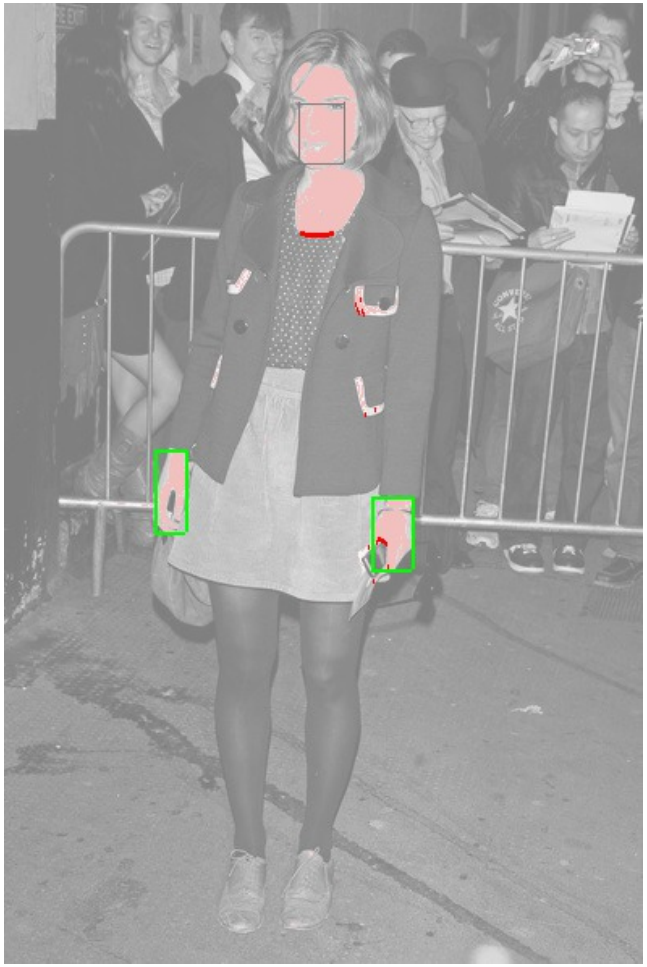
All of the text problems: feature extraction, retrieval, summarization, … are harder for images.

But they are still important, and getting more so, as everybody is taking pictures all the time.

Image processing is specialized
- we have general purpose text processing, but not images
- consider software for faces, CAD, maps, photos, …

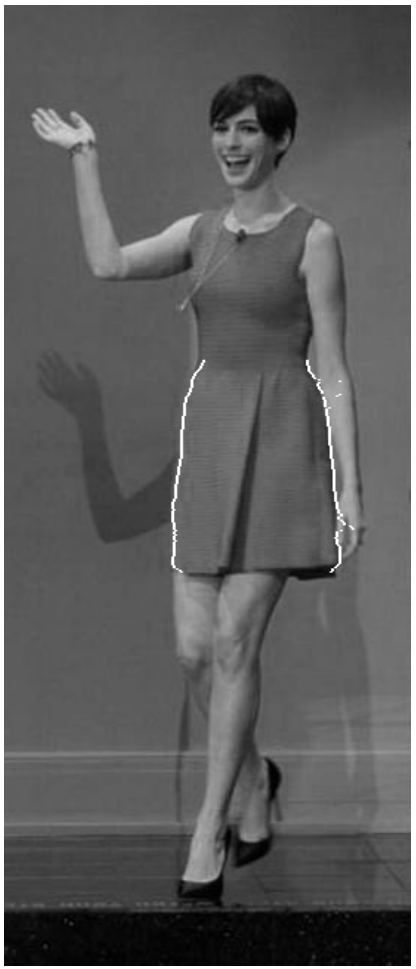
Here are a few examples of costume analysis.

Sometimes color analysis works



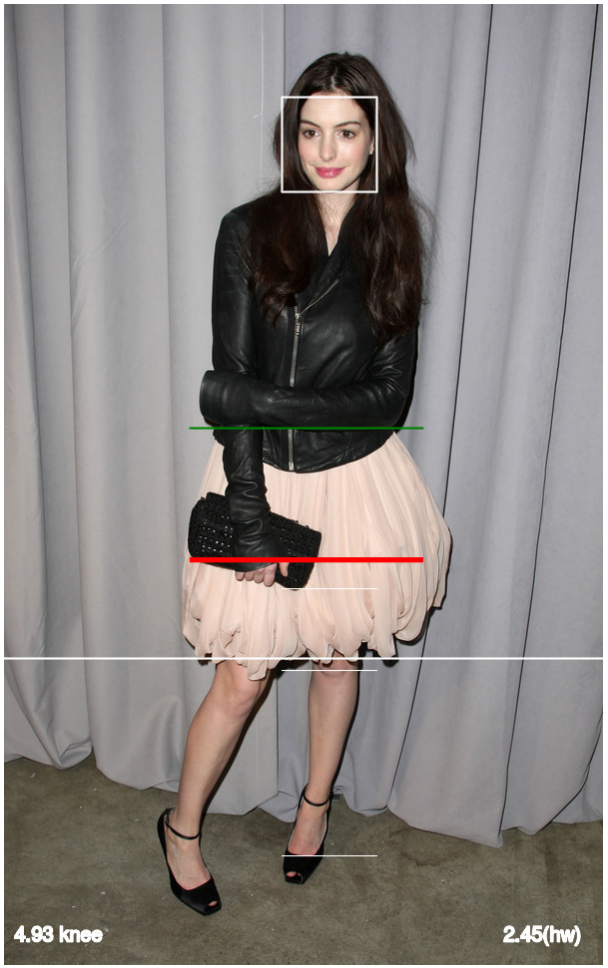
Looking for neckline.

Even simpler



Looking for skirt width

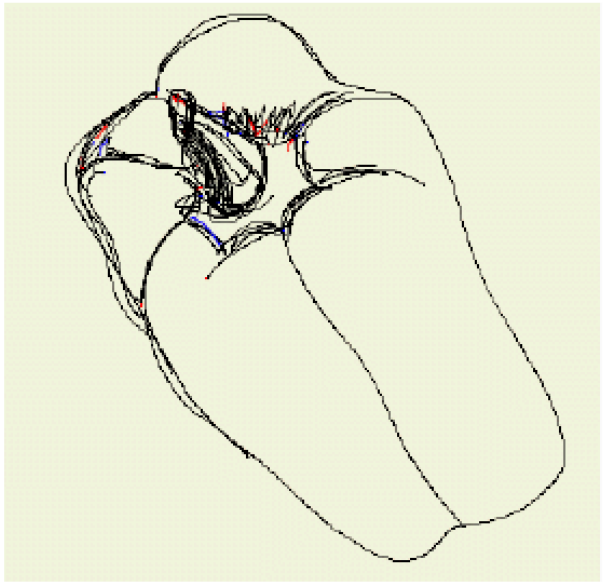
Works even better on hemlines

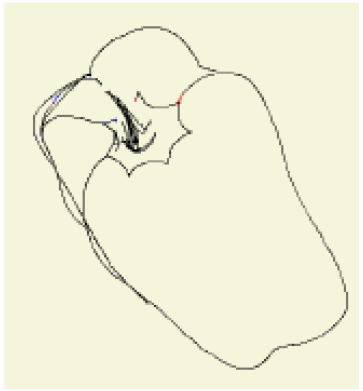


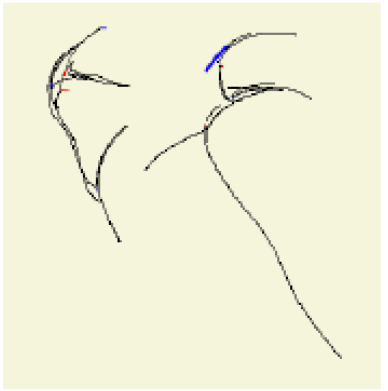
But the costume historian who suggested this works in a time period when there were no short skirts.

Summarizing an image

Look at the order of strokes, drawing on a tablet.



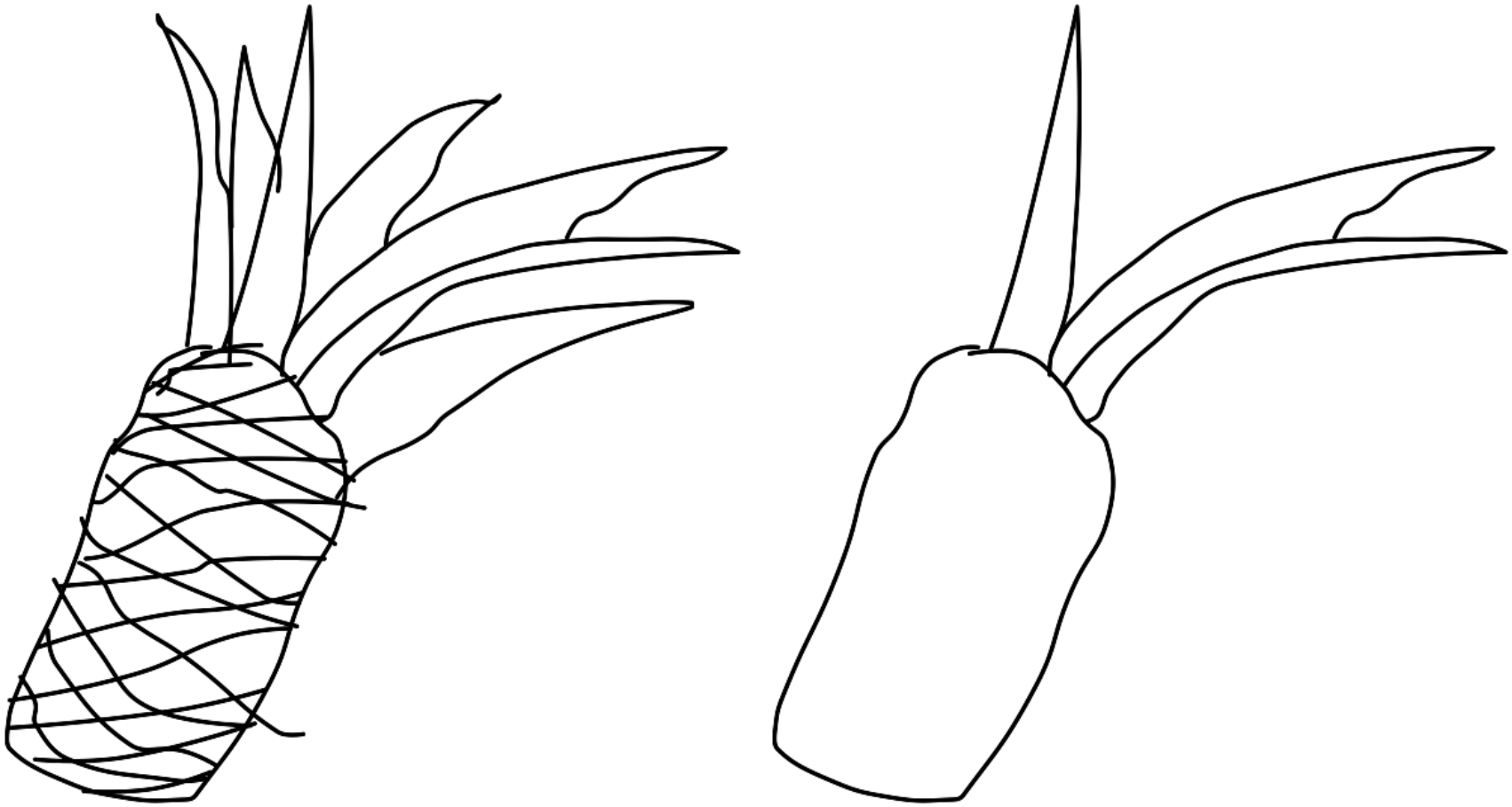




was drawn as shown (top left first, bottom right last)

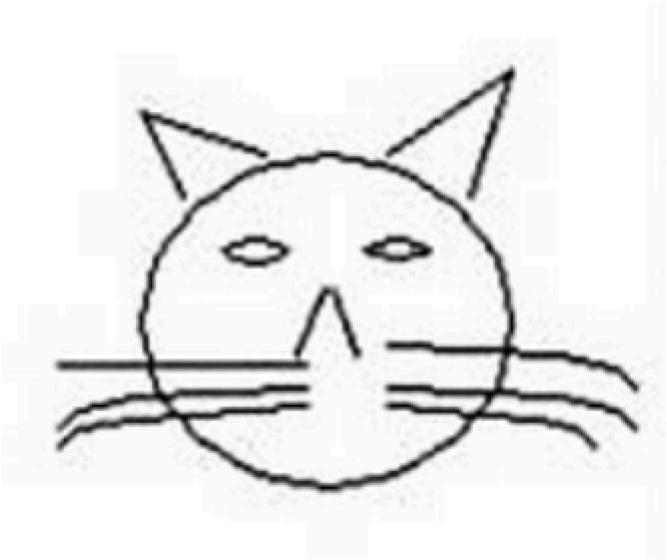
Drawing by Annamarie Klose

Again first strokes



From the SIGGRAPH sketch database

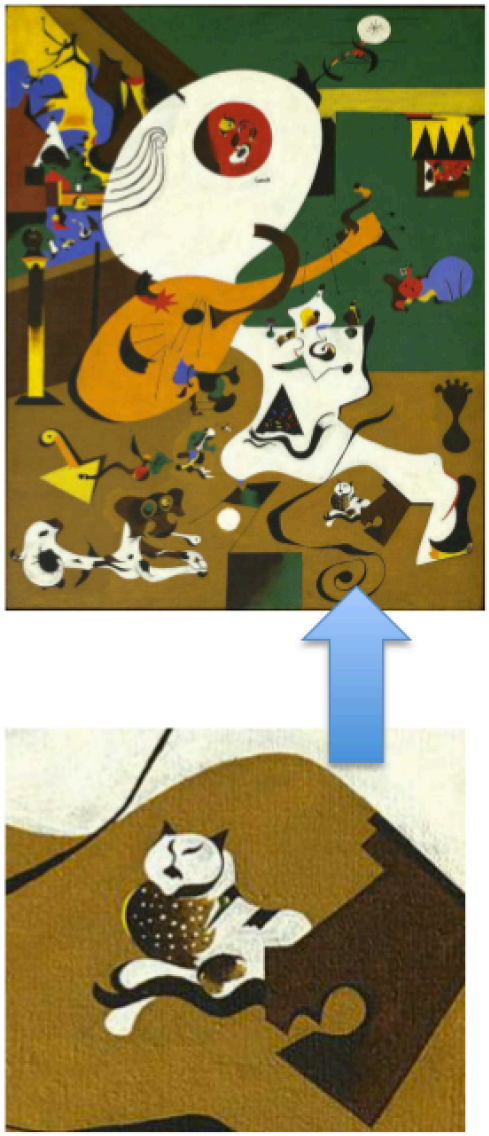
Impossible problem 1



Just for fun, this is edge detection on the cat picture – cats are just too fuzzy. Google has a good cat recognizer, but producing the left image is a different problem.

Impossible problem 2

Henrik Sorgh (1661) on left, Joan Miró (1928) on right.



Conclusions

Searching is now the province of machines and digital text: Google answers over three billion queries a day.

Reading is increasingly online, with scholarly publishers going all-electronic and Amazon stomping on the paper publishers of general books.

Researching is moving to algorithms, with authorship studies, stylistic analysis, network analysis and crowdsourcing (Goodreads) supplementing, but not yet replacing, traditional criticism.

Inspiration

The phrase “big data beats better algorithms” is encouraging fairly trivial analysis; is that what we really want?

Paul has led us to bigger data, better algorithms, and better problem understanding. We need all of those.