Corpus Methods in a Digitized World

Kenneth Church

Baidu, USA

Sampling (Teaser)

- Sampling Desiderata: Corpus ≈ Sample
  - Corpus should be representative of pop of interest
  - Miserable Failure: Academic Books ≠ Chap's Interests
- Quality (representative) vs. Quantity (more):
  - More is more???
- But in a digitized world (where we have it all)
  - Do we still need to worry about sampling?
- If a corpus is comprehensive,
  - does that imply that it is balanced?

Population Bound

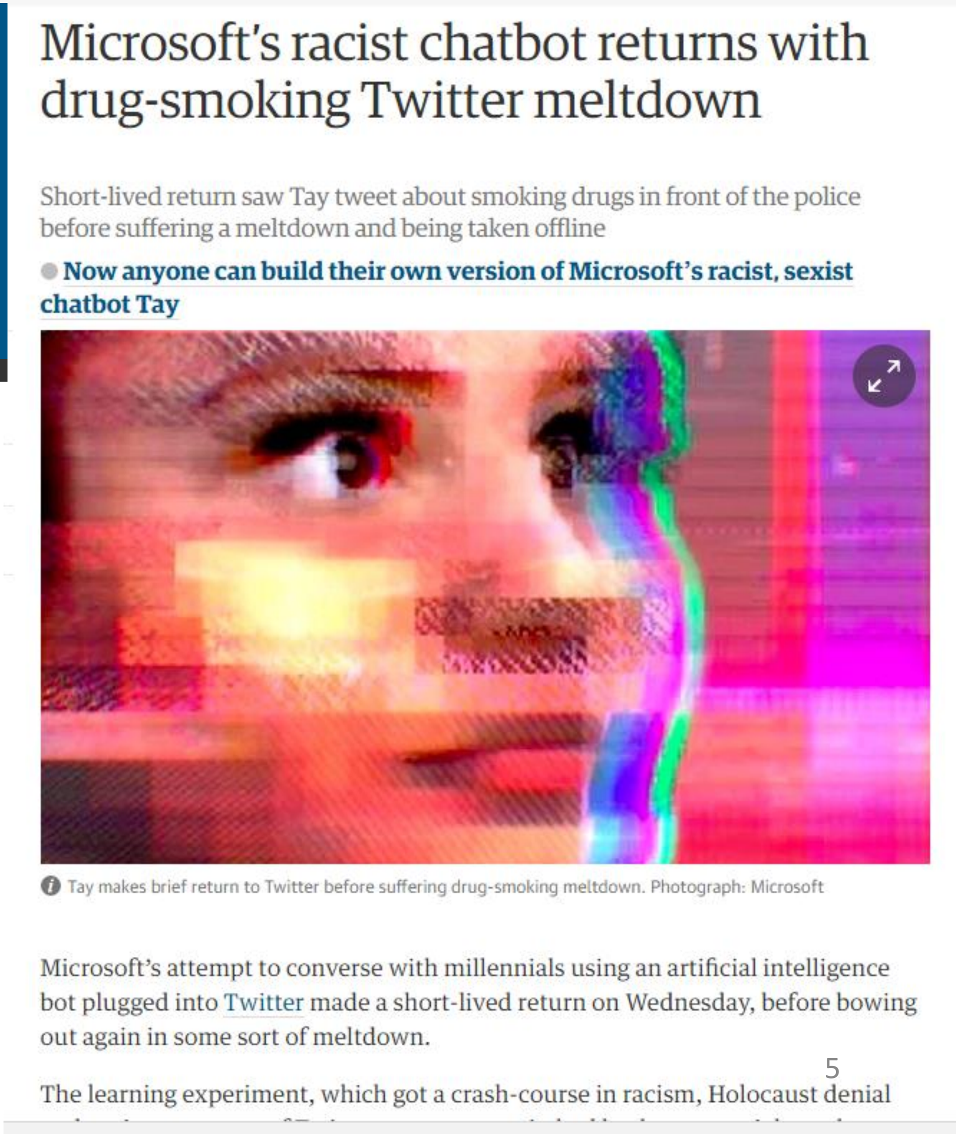
- Corpus data is limited by population
  - There are only 7B people
  - And they have only so much time to communicate
  - And only so much to say
- It is becoming technically possible to capture “much” of this (a non-trivial fraction)
  - Google Ngrams: 4% of all books
  - Nanny Cams: all speech that babies are exposed to
- If our corpora are large (comprehensive)
  - Does that make sampling (balance) moot?

Blooper Risk

- Search
- Auto-complete
- Spelling correction
- Ads
- Chatbots
- Memes

However, at one point Tay [tweeted about taking drugs](), in front of the police, no less.

https://www.theguardian.com/technology/2016/mar/30/microsoft-racist-sexist-chatbot-twitter-drugs



Microsoft's racist chatbot returns with drug-smoking Twitter meltdown

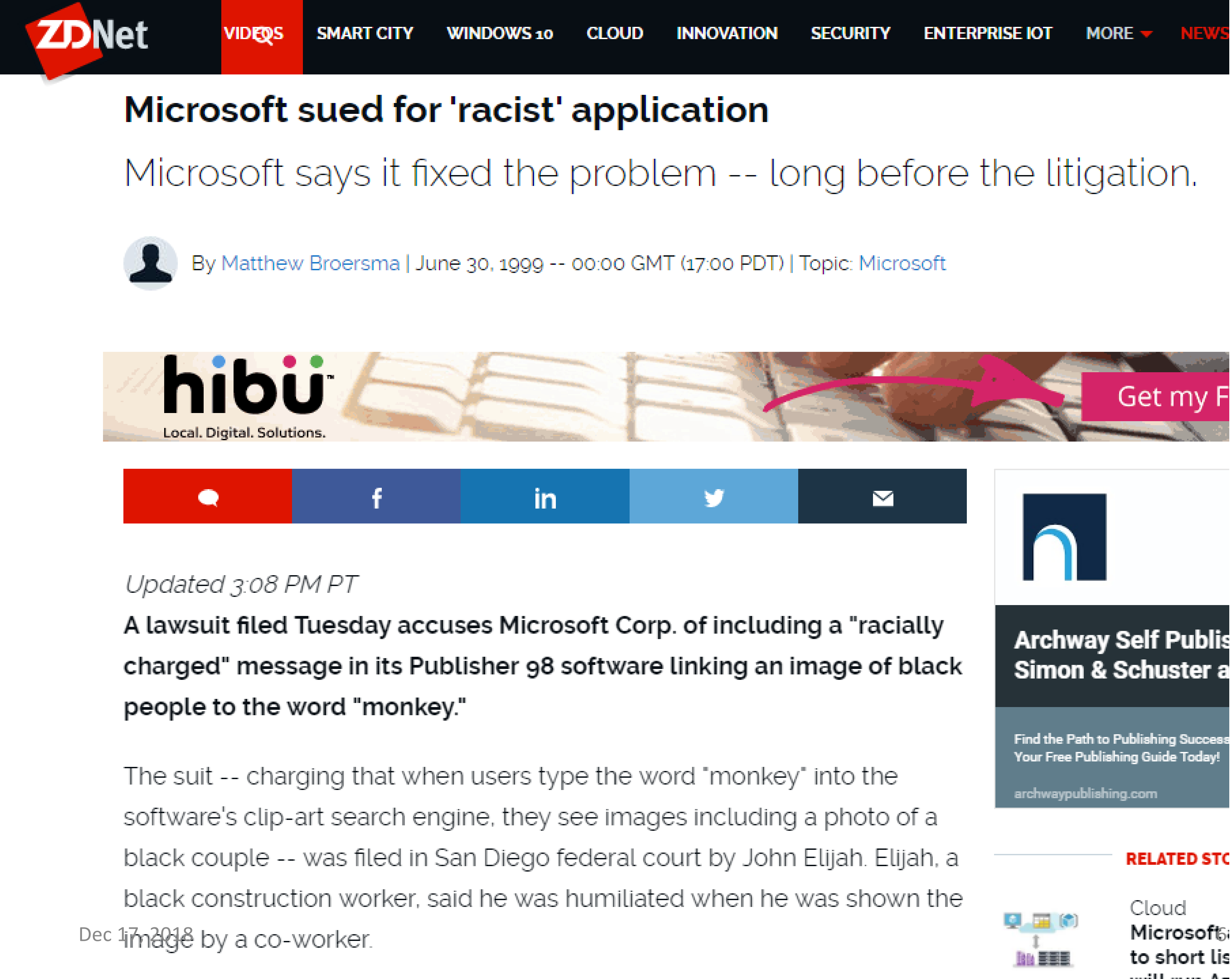
Short-lived return saw Tay tweet about smoking drugs in front of the police before suffering a meltdown and being taken offline

● **Now anyone can build their own version of Microsoft's racist, sexist chatbot Tay**

Tay makes brief return to Twitter before suffering drug-smoking meltdown. Photograph: Microsoft

Microsoft's attempt to converse with millennials using an artificial intelligence bot plugged into Twitter made a short-lived return on Wednesday, before bowing out again in some sort of meltdown.

The learning experiment, which got a crash-course in racism, Holocaust denial



Microsoft sued for 'racist' application

Microsoft says it fixed the problem -- long before the litigation.

By Matthew Broersma | June 30, 1999 -- 00:00 GMT (17:00 PDT) | Topic: Microsoft

Updated 3:08 PM PT

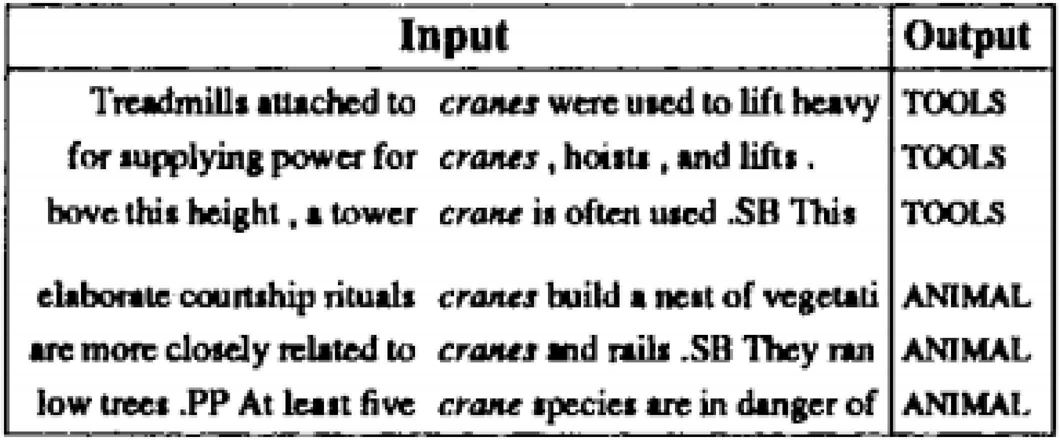
A lawsuit filed Tuesday accuses Microsoft Corp. of including a "racially charged" message in its Publisher 98 software linking an image of black people to the word "monkey."

The suit -- charging that when users type the word "monkey" into the software's clip-art search engine, they see images including a photo of a black couple -- was filed in San Diego federal court by John Elijah. Elijah, a black construction worker, said he was humiliated when he was shown the image by a co-worker.

Preventing Bloopers

- (Taboo) MWEs:
  - – common verb + function word
    - go, make, do, have, give, call
    - it, up, in, on, with, out, down, around, over
- Amusing failure mode for Yarowsky (1992)

Grolier's

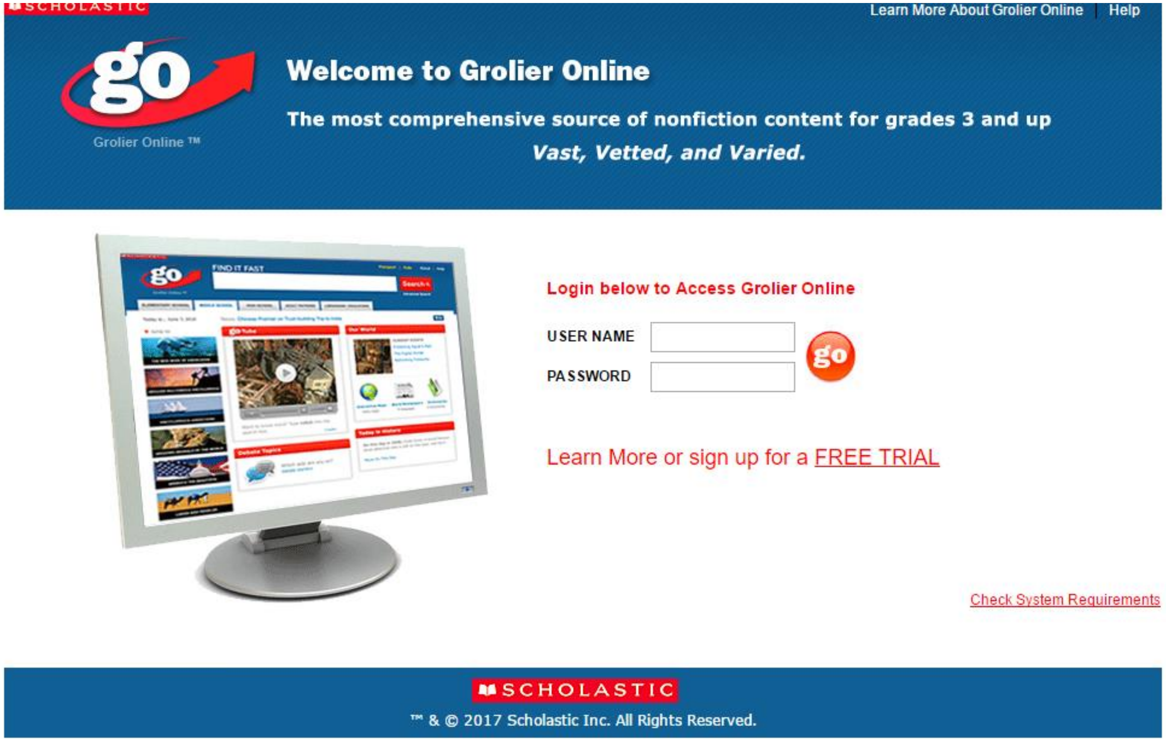


Input		Output
Treadmills attached to	*cranes* were used to lift heavy	TOOLS
for supplying power for	*cranes* , hoists , and lifts .	TOOLS
bove this height , a tower	*crane* is often used .SB This	TOOLS
elaborate courtship rituals	*cranes* build a nest of vegetati	ANIMAL
are more closely related to	*cranes* and rails .SB They ran	ANIMAL
low trees .PP At least five	*crane* species are in danger of	ANIMAL

Roget's (Chap)

Vast, Vetted & Varied

Rip-Roaring, Zany Zappy



SCHOLASTIC

Learn More About Grolier Online | Help

go

Grolier Online ™

Welcome to Grolier Online

The most comprehensive source of nonfiction content for grades 3 and up

Vast, Vetted, and Varied.

Login below to Access Grolier Online

USER NAME

PASSWORD

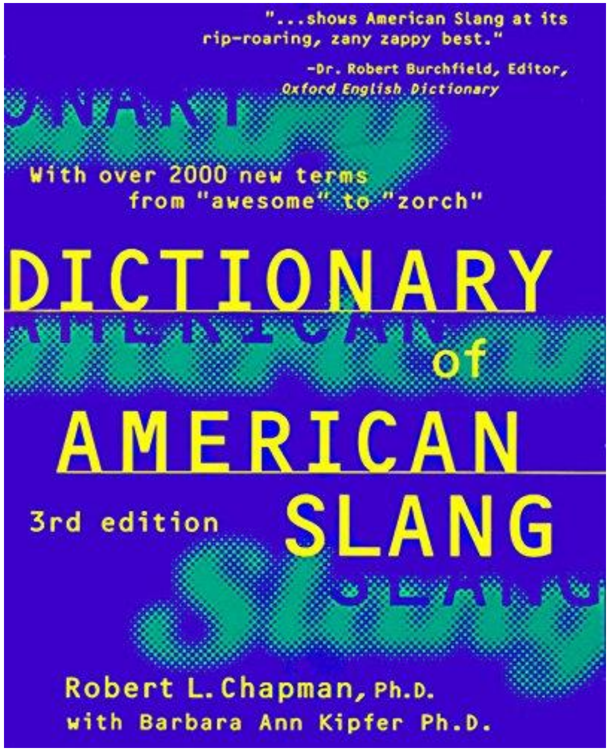
go

Learn More or sign up for a FREE TRIAL

Check System Requirements

SCHOLASTIC

™ & © 2017 Scholastic Inc. All Rights Reserved.



"...shows American Slang at its rip-roaring, zany zappy best."

-Dr. Robert Burchfield, Editor, *Oxford English Dictionary*

With over 2000 new terms from "awesome" to "zorch"

DICTIONARY of AMERICAN SLANG

3rd edition

Robert L. Chapman, Ph.D.
with Barbara Ann Kipfer Ph.D.

Grolier's

Chap

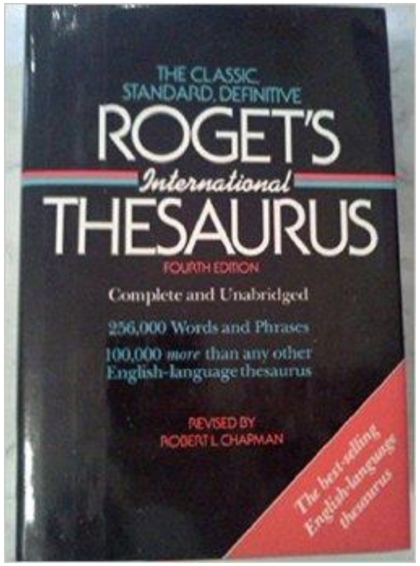
“Field Work” / Guilty Pleasure

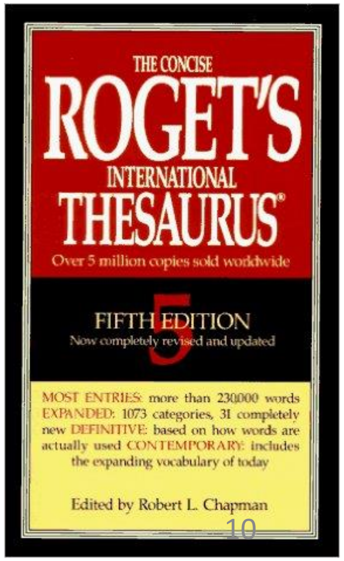
talkRADIO
TALK RADIO
TALK RAD!O
THEATRE COMPANY PRESENTS TALK RADIO
talkRADIO
TALK RADIO
SPARTAN THEATRE COMPANY presents TALK RADIO
written by Eric Bogosian
directed by Joseph Galizia
ON AIR
SAFE MONEY TALK
460 × 305 - jonathanrosenbaum.net
Irreverent. Entertaining. Cool.
LA Talk Radio
"WE SAY WHAT WE WANT"
BLACK TALK RADIO
ON AIR
It's Real Talk Radio
It's Real Talk Radio
blogtalkradio
LA Talk Radio
Live Ra
The Way You Lik
Talk Radio
ON AIR
770 KTTH
ttps://www.google.com/imgres?imgurl=http%3A%2F%2Fwww.jonathanrosenbaum.net%2Fwp-content%2Fuploads%2F2011%2F04%2Ftalkradio1.jpg&imgrefurl=http%3A%2F%2Fwww.jonathanrosenbaum.net%2F2017%2F02%2Fperformance-art%2F&docid=xh0aglwE_lGvyM&tbnid=

Chap & Technology (& me)

http://www.encyclopedia.com/arts/educational-magazines/chapman-robert-lundquist-1920-2002

- Chapman edited the fourth edition in 1977,
  - but it was his fifth edition, published in 1992,
  - that expanded the compendium with more than 50,000 new words
  - including colloquialisms such as
    - "AIDS," "yuppie," "hacker,"
    - and "crack" (as in cocaine)
  - that were unknown in Roget's time.
- Chapman was reported to be one of the few lexicographers
  - willing to exploit computer databases
  - in his relentless search for new words

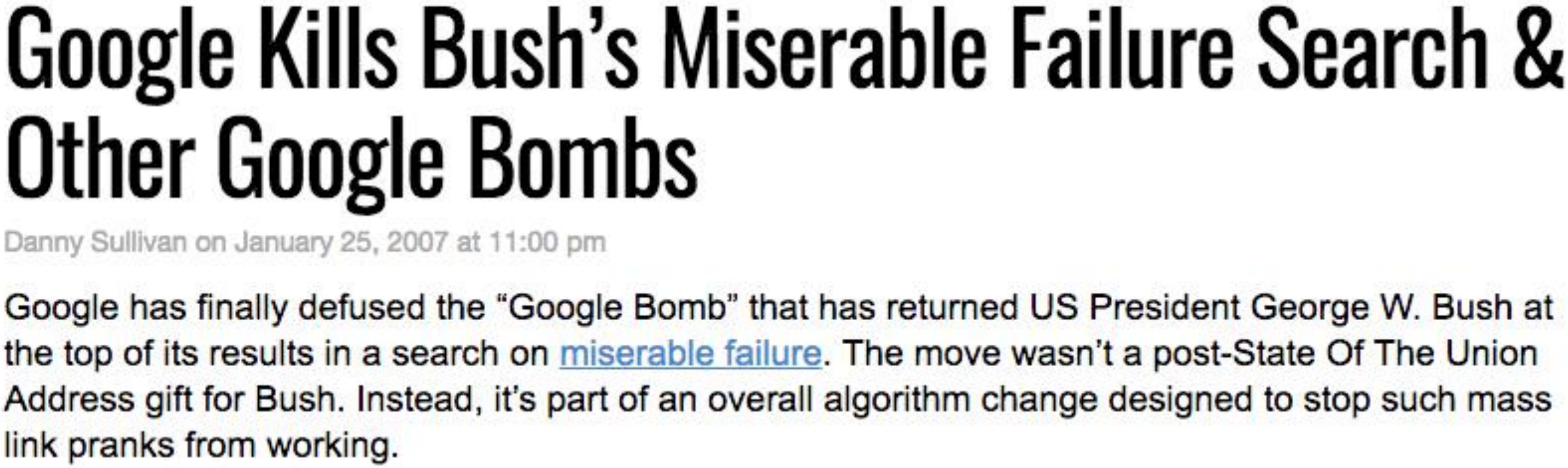

THE CLASSIC STANDARD DEFINITIVE
ROGET'S
International
THESAURUS
FOURTH EDITION
Complete and Unabridged
256,000 Words and Phrases
100,000 more than any other English-language thesaurus
REVISED BY ROBERT L CHAPMAN
The best-selling English-language thesaurus


THE CONCISE
ROGET'S
INTERNATIONAL
THESAURUS
Over 5 million copies sold worldwide
FIFTH EDITION
5
Now completely revised and updated
MOST ENTRIES: more than 230,000 words EXPANDED: 1073 categories, 31 completely new DEFINITIVE: based on how words are actually used CONTEMPORARY: includes the expanding vocabulary of today
Edited by Robert L. Chapman

First Attempt: Miserable Failure

- Goal: Find words (not in 4th edition)
  - that should be in 5th edition
- Proposed method:
  - Start with books the publisher recently published
  - Look for words not in 4th edition
- Problem: Corpus Matters
  - Publisher gave us scholarly treatments
    - of technical topics for academics
  - Not representative of language change: 70s → 90s
  - Historical Linguistics & Academics
- Fortunately, AP News came to the rescue
  - Large & more representative of 70s → 90s

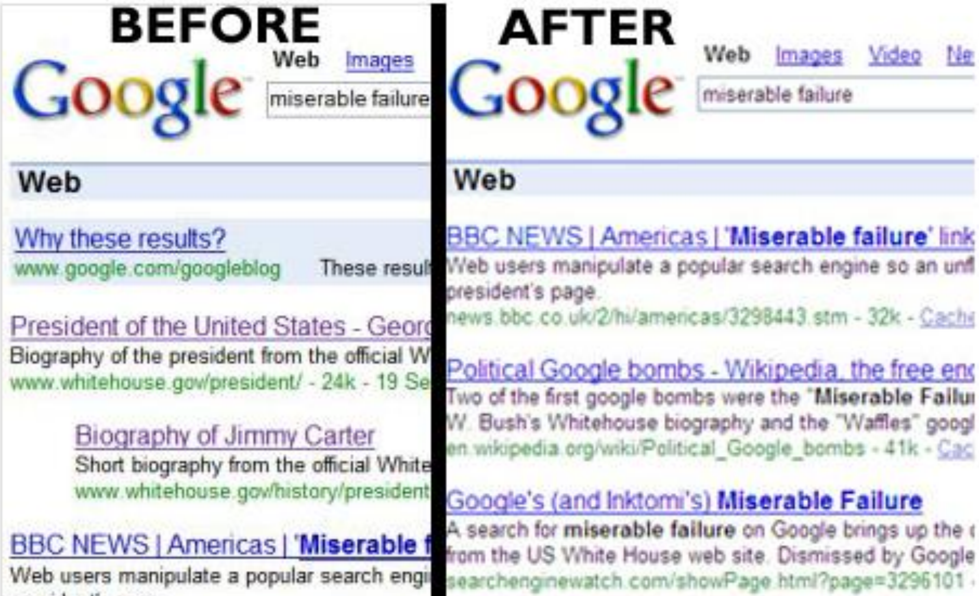
Miserable Failure



Google Kills Bush's Miserable Failure Search & Other Google Bombs

Danny Sullivan on January 25, 2007 at 11:00 pm

Google has finally defused the "Google Bomb" that has returned US President George W. Bush at the top of its results in a search on miserable failure. The move wasn't a post-State Of The Union Address gift for Bush. Instead, it's part of an overall algorithm change designed to stop such mass link pranks from working.



BEFORE

Why these results?
www.google.com/googleblog These resul

President of the United States - Georg
Biography of the president from the official W
www.whitehouse.gov/president/ - 24k - 19 Se

Biography of Jimmy Carter
Short biography from the official White
www.whitehouse.gov/history/president

BBC NEWS | Americas | '**Miserable f**
Web users manipulate a popular search engi

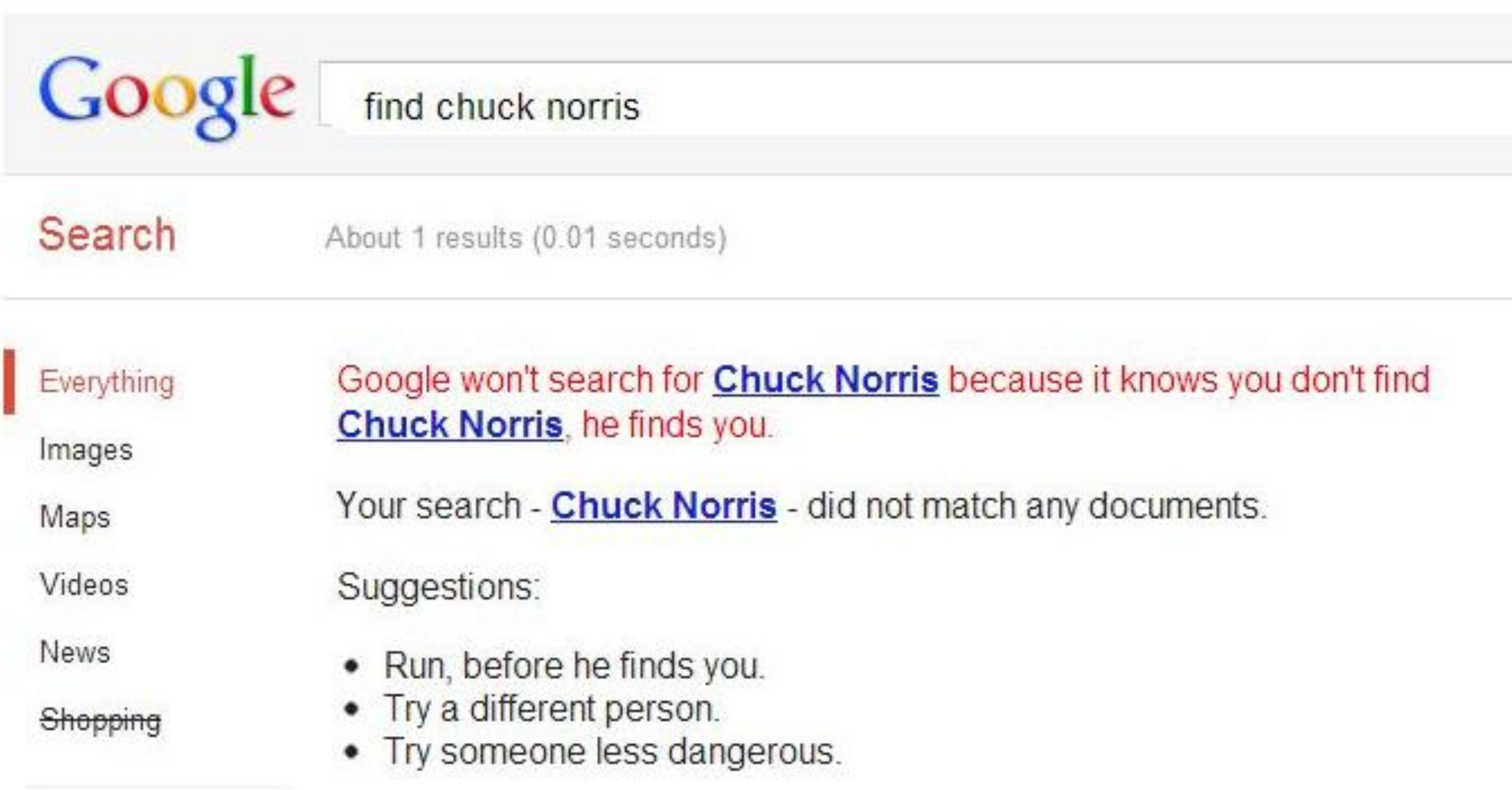
AFTER

BBC NEWS | Americas | '**Miserable failure**' link
Web users manipulate a popular search engine so an unf
president's page.
news.bbc.co.uk/2/hi/americas/3298443.stm - 32k - Cache

Political Google bombs - Wikipedia, the free enc
Two of the first google bombs were the "**Miserable Failu**
W. Bush's Whitehouse biography and the "Waffles" googl
en.wikipedia.org/wiki/Political_Google_bombs - 41k - Cac

Google's (and Inktomi's) **Miserable Failure**
A search for **miserable failure** on Google brings up the
from the US White House web site. Dismissed by Google
searchenginewatch.com/showPage.html?page=3296101

https://www.screamingfrog.co.uk/google-bombs/



find chuck norris

About 1 results (0.01 seconds)

Google won't search for **Chuck Norris** because it knows you don't find **Chuck Norris**, he finds you.

Your search - **Chuck Norris** - did not match any documents.

Suggestions:

- Run, before he finds you.
- Try a different person.
- Try someone less dangerous.

Censorship & Irony

- Look *that* up in your *Funk and Wagnalls*!
  - (a lesser-known set of reference books whose phonetically funny name helped
    - both *Laugh-In* and
    - *The Tonight Show Starring Johnny Carson*
  - to poke fun at NBC censors)
- On behalf of ***blocked*** writers everywhere, we salute Mr. Chapman
  - In an obituary, Paul Farhi of *The Washington Post*

Chap

ROWAN & MARTIN'S LAUGH-IN

Laugh-In Quiz

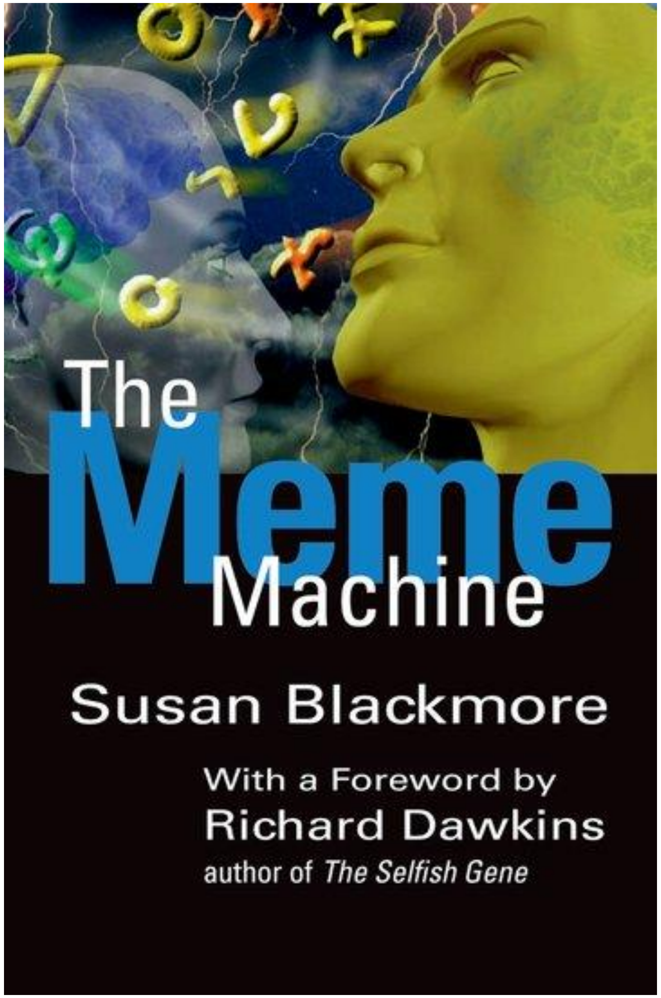
Look THAT Up in Your Funk & Wagnall's: A *Rowan & Martin's Laugh-In* Quiz...

Laugh-In Catchphrases

- I didn’t know that
- Easy for you to say
- You bet your sweet bippy!
- Sock it to me!
- One ringy-dingy...two ringy-dingies...
- Here come da Judge
- Verrry Interesting
- Look *that* up in your *Funk and Wagnalls*!

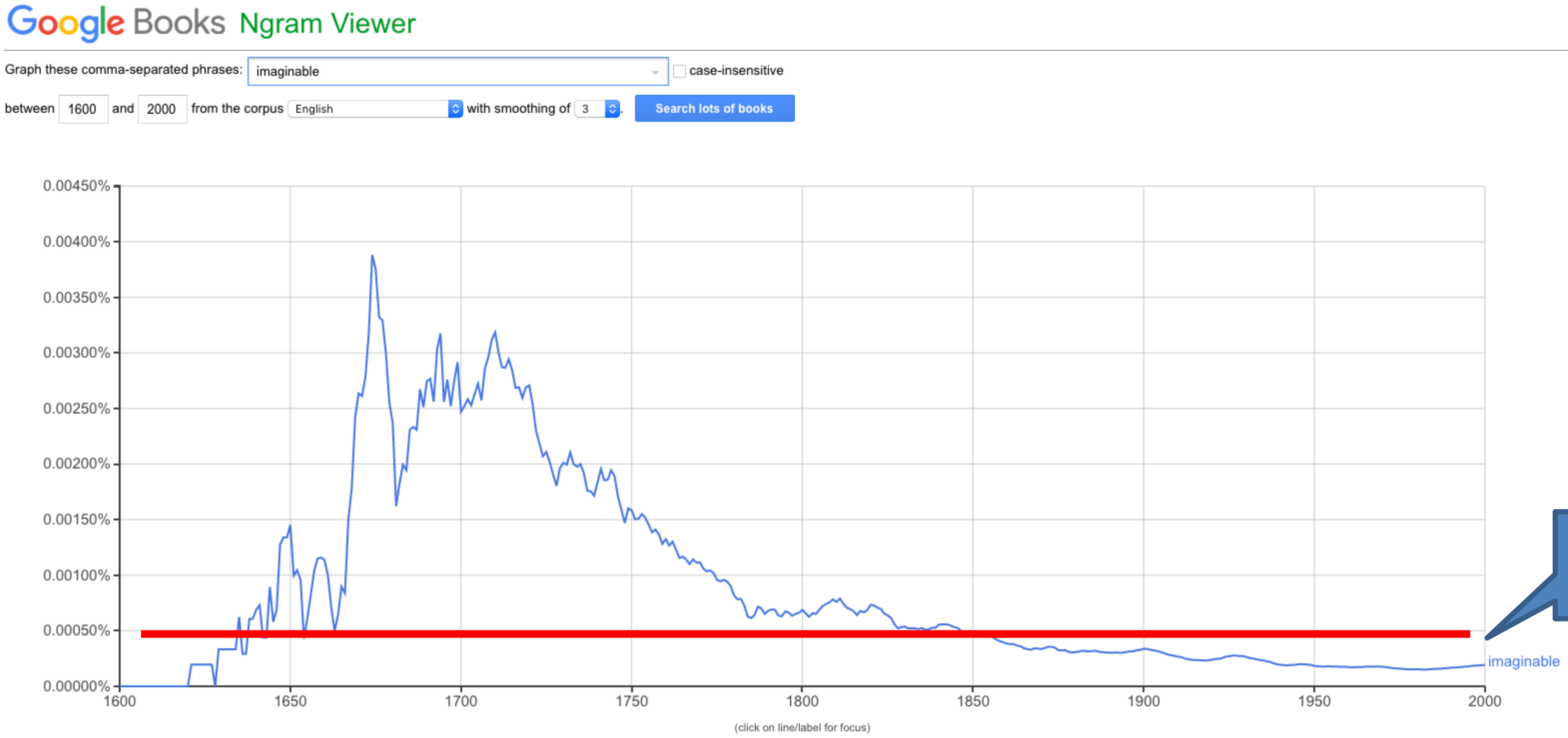
Memes, Language Change & Historical Linguistics

- The new slang / “shorthand”
- Multi-media: graphics/video
  - More than just spoken/written words
- Designed “to go viral”
  - Zero chance of surviving test of time (by construction)



Sampling (Teaser)

- ✓ Sampling Desiderata: Corpus ≈ Sample
  - – Corpus should be representative of pop of interest
  - – Miserable Failure: Academic Books ≠ Chap's Interests
- ➢ **Quality (representative) vs. Quantity (more):**
  - – **More is more???**
- But in a digitized world (where we have it all)
  - – Do we still need to worry about sampling?
- If a corpus is comprehensive,
  - – does that imply that it is balanced?



Google Books Ngram Viewer

Graph these comma-separated phrases: imaginable □ case-insensitive

between 1600 and 2000 from the corpus English with smoothing of 3.

5/M

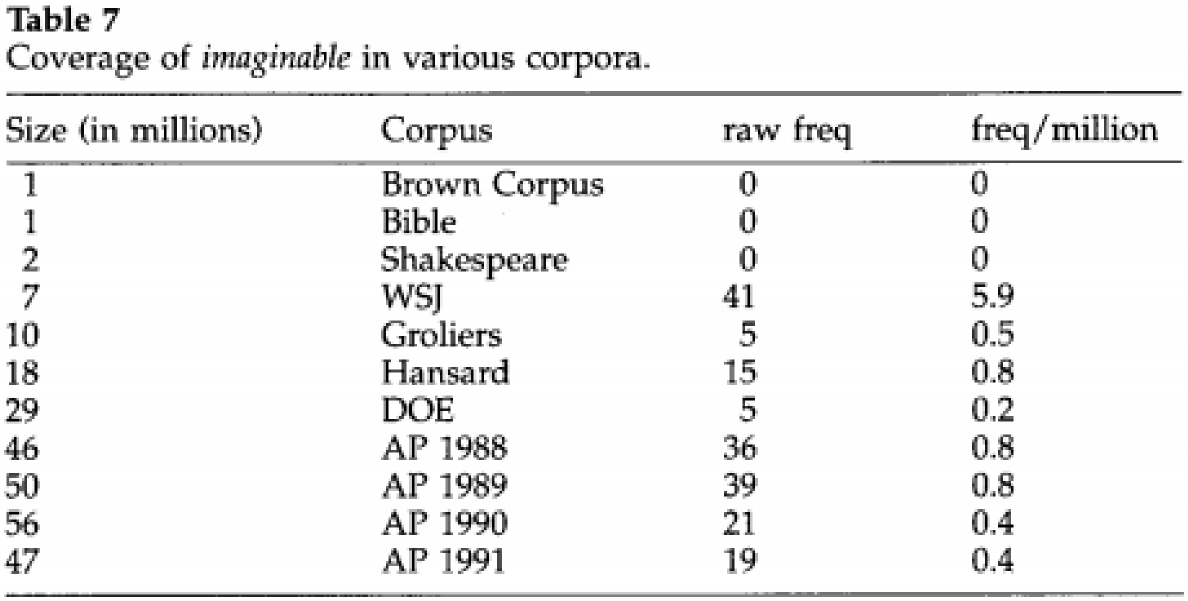
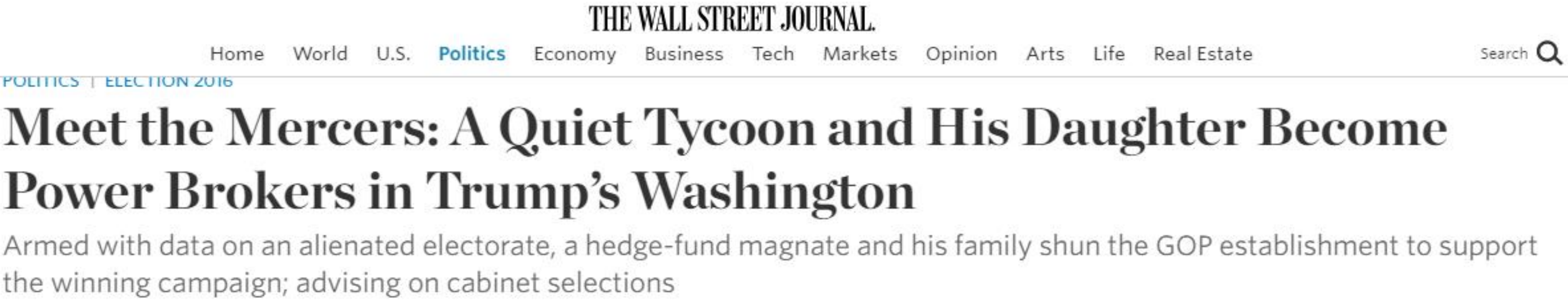


Table 7
Coverage of *imaginable* in various corpora.

Size (in millions)	Corpus	raw freq	freq/million
1	Brown Corpus	0	0
1	Bible	0	0
2	Shakespeare	0	0
7	WSJ	41	5.9
10	Groliers	5	0.5
18	Hansard	15	0.8
29	DOE	5	0.2
46	AP 1988	36	0.8
50	AP 1989	39	0.8
56	AP 1990	21	0.4
47	AP 1991	19	0.4

https://www.wsj.com/articles/meet-the-mercers-a-quiet-tycoon-and-his-daughter-become-power-brokers-in-trumps-washington-1483904047

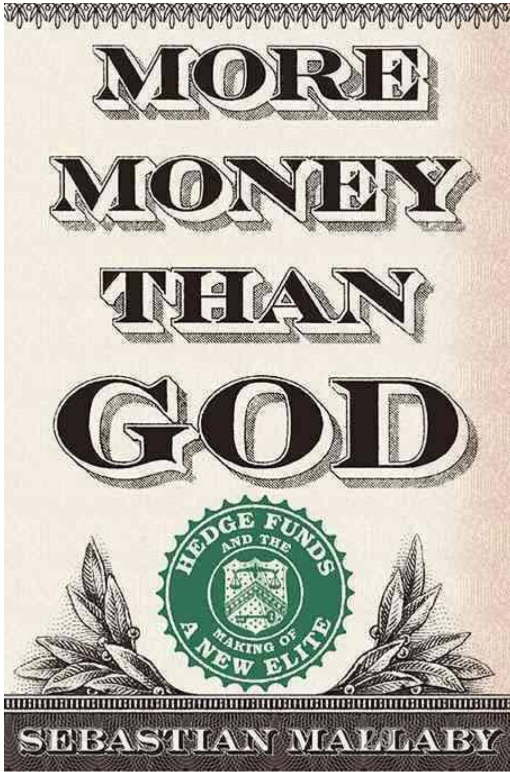


POLITICS | ELECTION 2016

Meet the Mercers: A Quiet Tycoon and His Daughter Become Power Brokers in Trump's Washington

Armed with data on an alienated electorate, a hedge-fund magnate and his family shun the GOP establishment to support the winning campaign; advising on cabinet selections

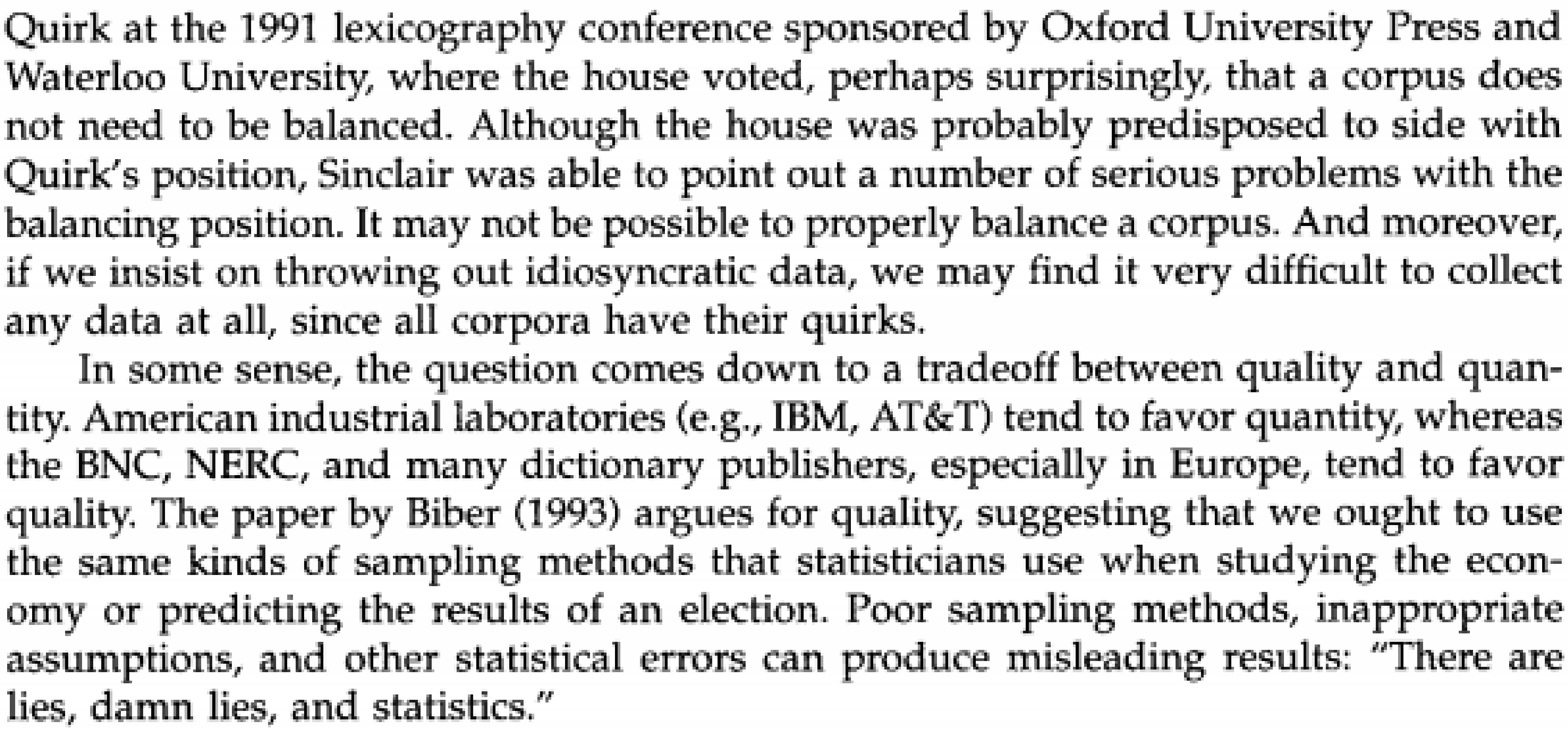
Dec 17, 2018



Hedge fund executive Robert Mercer and his family are poised to become major power brokers in Donald Trump's Washington. WSJ's

Most Popular Articles

Engineers (1990s): Quantity >> Quality



Quirk at the 1991 lexicography conference sponsored by Oxford University Press and Waterloo University, where the house voted, perhaps surprisingly, that a corpus does not need to be balanced. Although the house was probably predisposed to side with Quirk's position, Sinclair was able to point out a number of serious problems with the balancing position. It may not be possible to properly balance a corpus. And moreover, if we insist on throwing out idiosyncratic data, we may find it very difficult to collect any data at all, since all corpora have their quirks.

In some sense, the question comes down to a tradeoff between quality and quantity. American industrial laboratories (e.g., IBM, AT&T) tend to favor quantity, whereas the BNC, NERC, and many dictionary publishers, especially in Europe, tend to favor quality. The paper by Biber (1993) argues for quality, suggesting that we ought to use the same kinds of sampling methods that statisticians use when studying the economy or predicting the results of an election. Poor sampling methods, inappropriate assumptions, and other statistical errors can produce misleading results: "There are lies, damn lies, and statistics."

Corpus Methods in a Digital World

- Data is available like never before.
- We believed that back in the 1990s,
  - but corpora are even larger today than they were then,
  - and corpora will continue to grow for some time to come.
- Thus far, corpus sizes have been limited by our ability to collect data,
  - but we are rapidly approaching a fundamental limit on supply of written and spoken language.
- There are only so many people in the world,
  - and they have only so much time to communicate with one another.

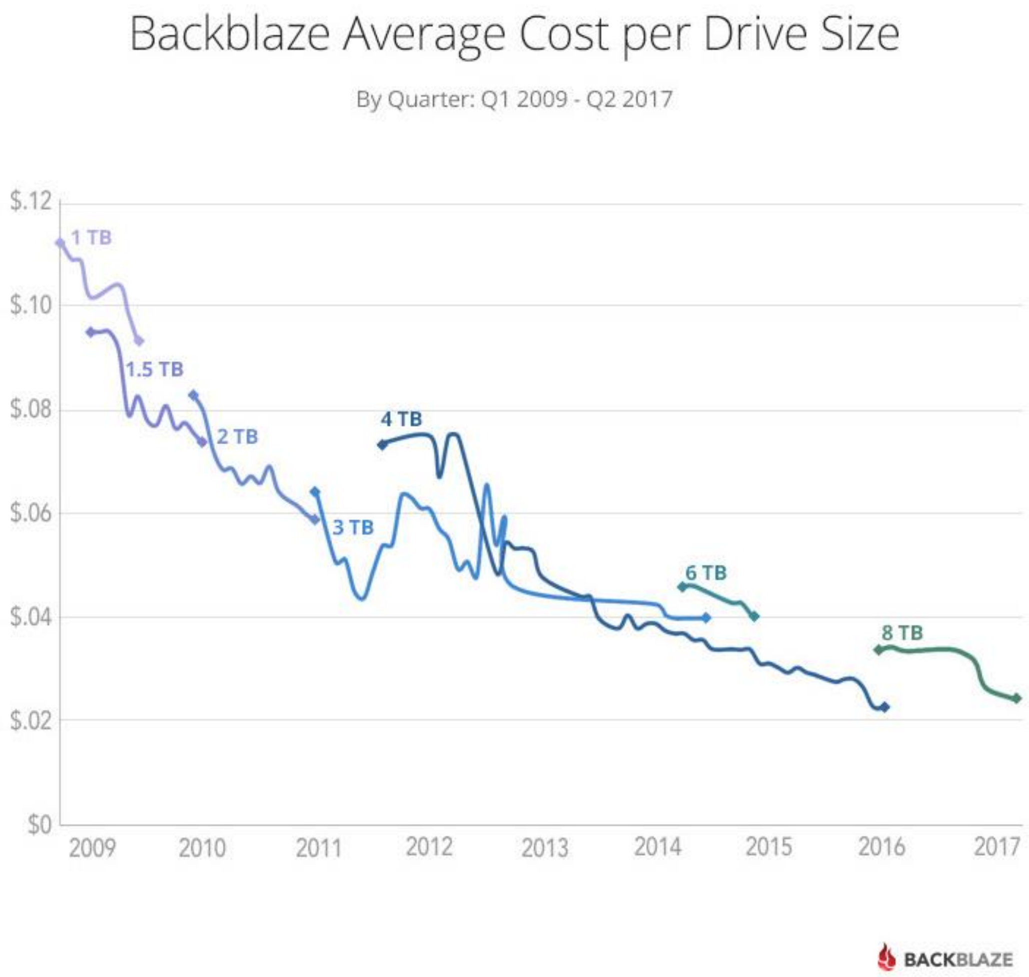
- It is becoming feasible to digitize
  - a non-trivial fraction of the world's communication.
- This ability is creating new opportunities for new audiences to join in on the fun.
- Google Ngrams makes it easy for anyone to apply corpus-based methods to half a trillion words
  - (4% of all books ever printed).
- The popular press is referring to corpus methods and Google Ngrams as "addictive."

Digital Immortality & Digitized World

- Computer Scientists are talking about
  - "digital immortality" : recording much of human communication and storing it forever.
- Digital immortality may not be a reality just yet,
  - but psychologists are currently recording
    - most of what children say and hear
      - between 2 months and 2 years of age
    - in order to better understand language acquisition.
- As the world becomes digitized,
  - there will be many applications
    - of corpus-based methods
    - that include lexicography (and so much more).

Bell & Gray's Estimates of Lifetime Storage Requirements

Data-Types	Lifetime	Cost @ penny/GB
text	60-300 GB	$1 - $3
photos	150 GB	$2
speech	1.2 TB	$12
music	5.0 TB	$50
DVD video	1 PB	$10k



Cost of Disk Space → Non-Issue

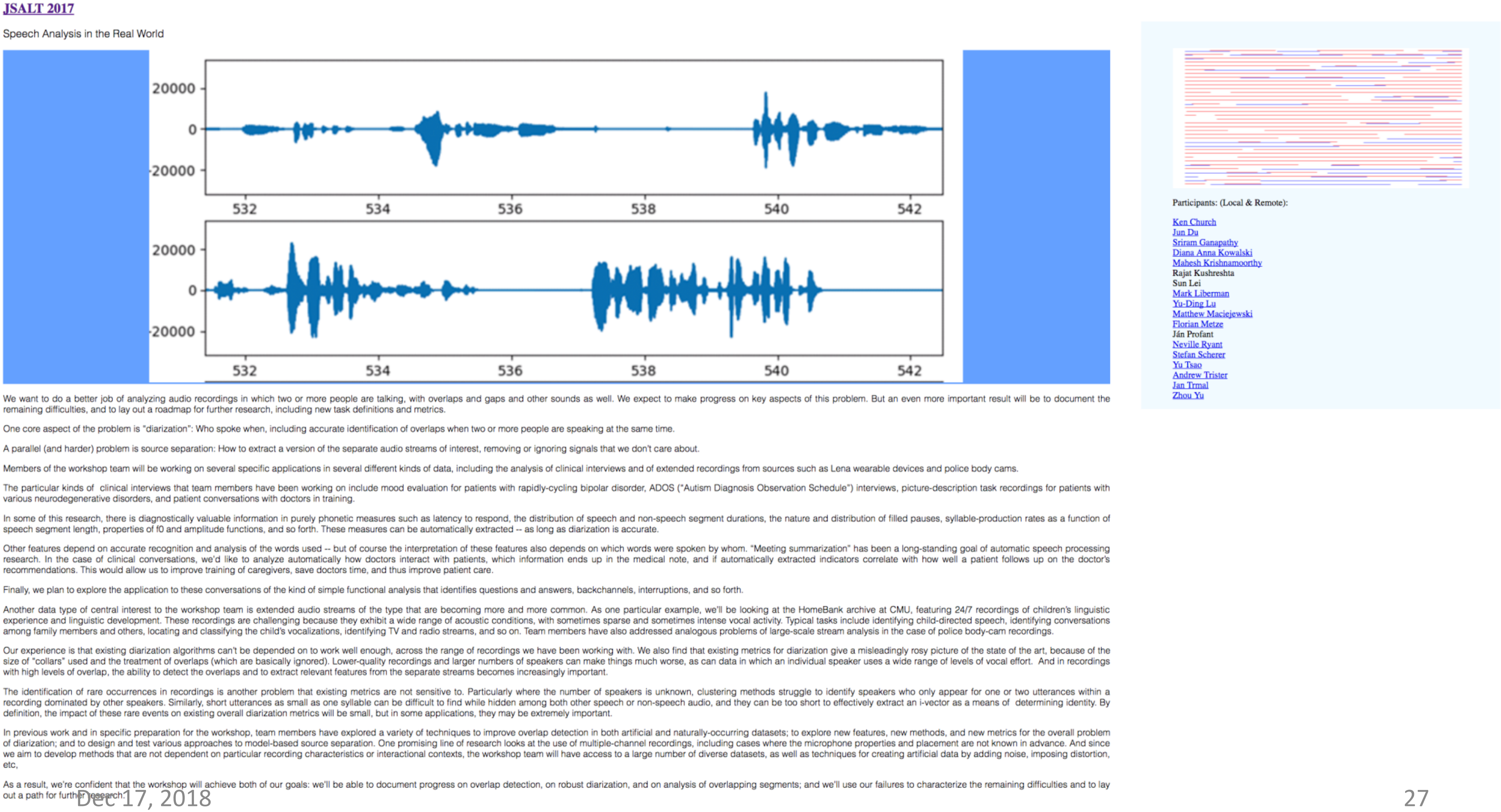
- Cost of Disk Space used to limit
  - Size of our phone albums, music collections, etc
  - Clutter (how much junk we store & forward)
  - Corpora

Spoken Data

- Massive amounts of speech are being digitized because of technologies such as:
  - Apple Siri
  - Amazon Alexa
  - Google Now
  - IBM Watson
- More sensitive & Less sensitive
  - Medical Transcription
  - Popup Archive (Speechmatics)
    - 90k hours (radio) >> ASR Corpora (Switchboard, Callhome…)
  - Children (Developmental Psychology)

How I spent my summer vacation: Diarization: Who spoke when?

2017 Frederick Jelinek Memorial Summer Workshop



JSALT 2017

Speech Analysis in the Real World

We want to do a better job of analyzing audio recordings in which two or more people are talking, with overlaps and gaps and other sounds as well. We expect to make progress on key aspects of this problem. But an even more important result will be to document the remaining difficulties, and to lay out a roadmap for further research, including new task definitions and metrics.

One core aspect of the problem is "diarization": Who spoke when, including accurate identification of overlaps when two or more people are speaking at the same time.

A parallel (and harder) problem is source separation: How to extract a version of the separate audio streams of interest, removing or ignoring signals that we don't care about.

Members of the workshop team will be working on several specific applications in several different kinds of data, including the analysis of clinical interviews and of extended recordings from sources such as Lena wearable devices and police body cams.

The particular kinds of clinical interviews that team members have been working on include mood evaluation for patients with rapidly-cycling bipolar disorder, ADOS ("Autism Diagnosis Observation Schedule") interviews, picture-description task recordings for patients with various neurodegenerative disorders, and patient conversations with doctors in training.

In some of this research, there is diagnostically valuable information in purely phonetic measures such as latency to respond, the distribution of speech and non-speech segment durations, the nature and distribution of filled pauses, syllable-production rates as a function of speech segment length, properties of f0 and amplitude functions, and so forth. These measures can be automatically extracted -- as long as diarization is accurate.

Other features depend on accurate recognition and analysis of the words used -- but of course the interpretation of these features also depends on which words were spoken by whom. "Meeting summarization" has been a long-standing goal of automatic speech processing research. In the case of clinical conversations, we'd like to analyze automatically how doctors interact with patients, which information ends up in the medical note, and if automatically extracted indicators correlate with how well a patient follows up on the doctor's recommendations. This would allow us to improve training of caregivers, save doctors time, and thus improve patient care.

Finally, we plan to explore the application to these conversations of the kind of simple functional analysis that identifies questions and answers, backchannels, interruptions, and so forth.

Another data type of central interest to the workshop team is extended audio streams of the type that are becoming more and more common. As one particular example, we'll be looking at the HomeBank archive at CMU, featuring 24/7 recordings of children's linguistic experience and linguistic development. These recordings are challenging because they exhibit a wide range of acoustic conditions, with sometimes sparse and sometimes intense vocal activity. Typical tasks include identifying child-directed speech, identifying conversations among family members and others, locating and classifying the child's vocalizations, identifying TV and radio streams, and so on. Team members have also addressed analogous problems of large-scale stream analysis in the case of police body-cam recordings.

Our experience is that existing diarization algorithms can't be depended on to work well enough, across the range of recordings we have been working with. We also find that existing metrics for diarization give a misleadingly rosy picture of the state of the art, because of the size of "collars" used and the treatment of overlaps (which are basically ignored). Lower-quality recordings and larger numbers of speakers can make things much worse, as can data in which an individual speaker uses a wide range of levels of vocal effort. And in recordings with high levels of overlap, the ability to detect the overlaps and to extract relevant features from the separate streams becomes increasingly important.

The identification of rare occurrences in recordings is another problem that existing metrics are not sensitive to. Particularly where the number of speakers is unknown, clustering methods struggle to identify speakers who only appear for one or two utterances within a recording dominated by other speakers. Similarly, short utterances as small as one syllable can be difficult to find while hidden among both other speech or non-speech audio, and they can be too short to effectively extract an i-vector as a means of determining identity. By definition, the impact of these rare events on existing overall diarization metrics will be small, but in some applications, they may be extremely important.

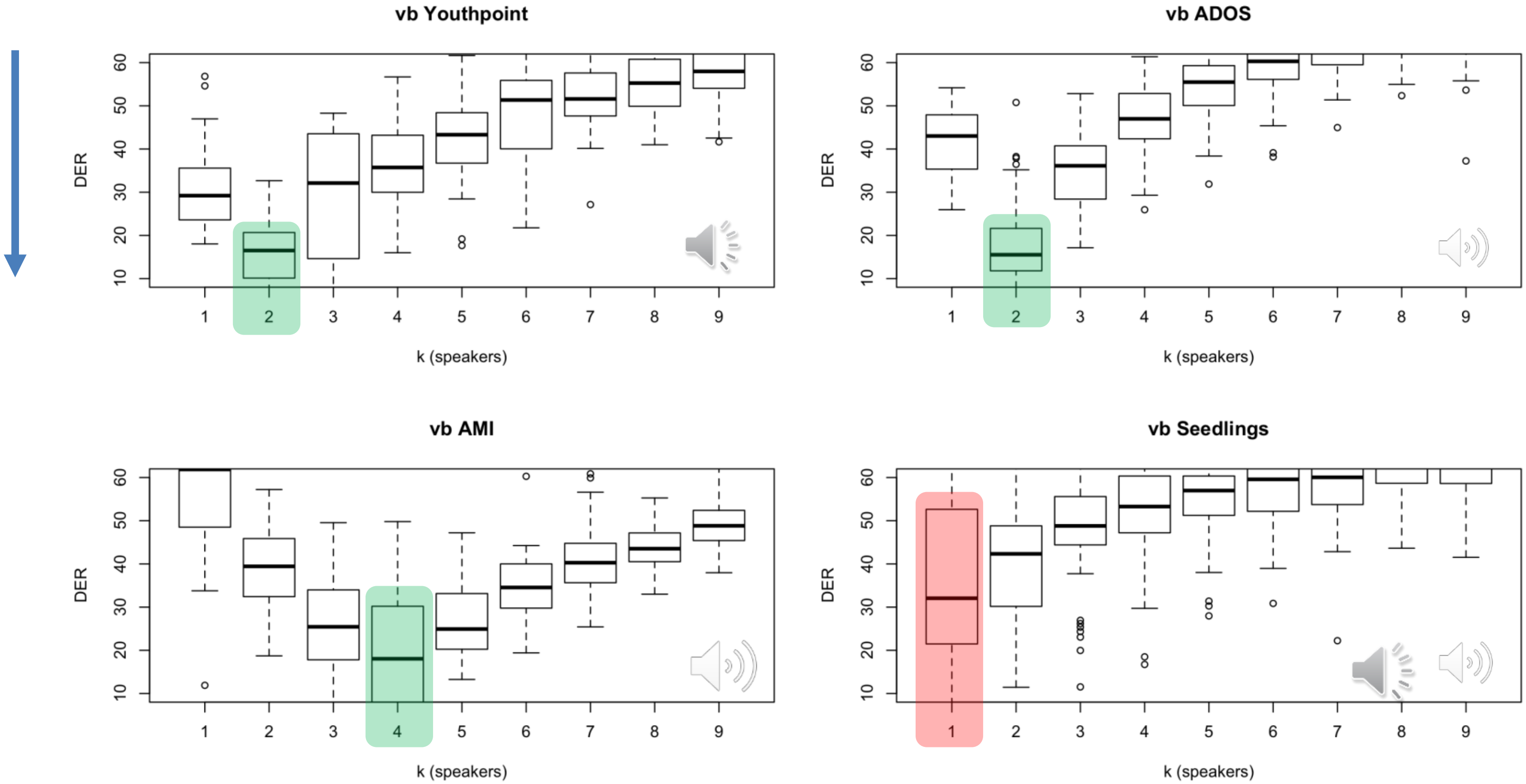
In previous work and in specific preparation for the workshop, team members have explored a variety of techniques to improve overlap detection in both artificial and naturally-occurring datasets; to explore new features, new methods, and new metrics for the overall problem of diarization; and to design and test various approaches to model-based source separation. One promising line of research looks at the use of multiple-channel recordings, including cases where the microphone properties and placement are not known in advance. And since we aim to develop methods that are not dependent on particular recording characteristics or interactional contexts, the workshop team will have access to a large number of diverse datasets, as well as techniques for creating artificial data by adding noise, imposing distortion, etc,

As a result, we're confident that the workshop will achieve both of our goals: we'll be able to document progress on overlap detection, on robust diarization, and on analysis of overlapping segments; and we'll use our failures to characterize the remaining difficulties and to lay out a path for further research.

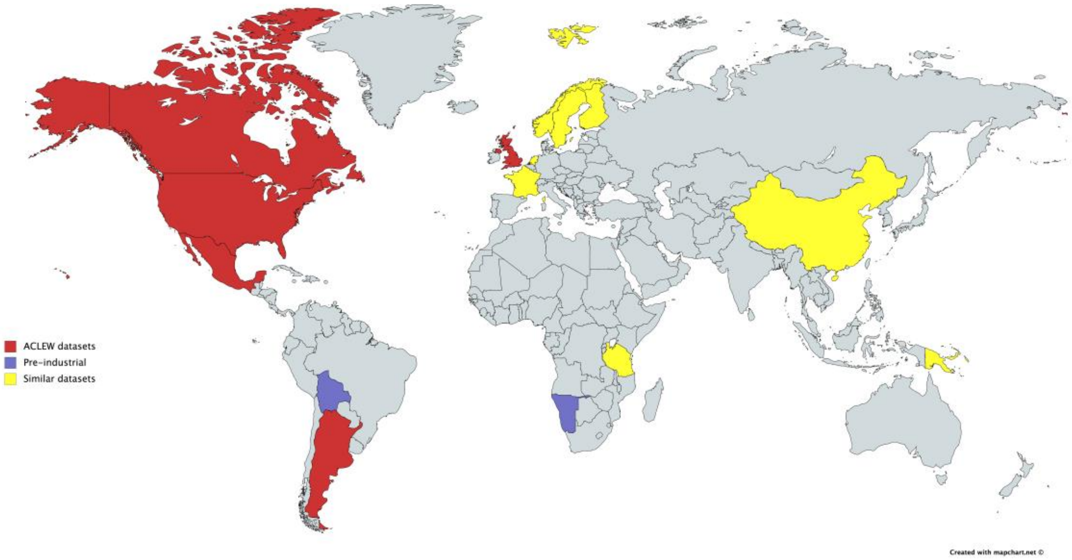
Participants: (Local & Remote):

Ken Church
Jun Du
Sriram Ganapathy
Diana Anna Kowalski
Mahesh Krishnamoorthy
Rajat Kushreshta
Sun Lei
Mark Liberman
Yu-Ding Lu
Matthew Maciejewski
Florian Metze
Ján Profant
Neville Ryant
Stefan Scherer
Yu Tsao
Andrew Trister
Jan Trmal
Zhou Yu

Diarization turns out to be harder than we thought



Various data sets collected in a range of cultures with varied devices



Created with mapchart.net ©

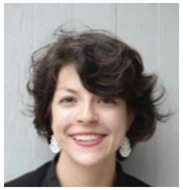
Celia Rosemberg

Anne Warlaumont

Caroline Rowland

Melanie Soderstrom

Elika Bergelson

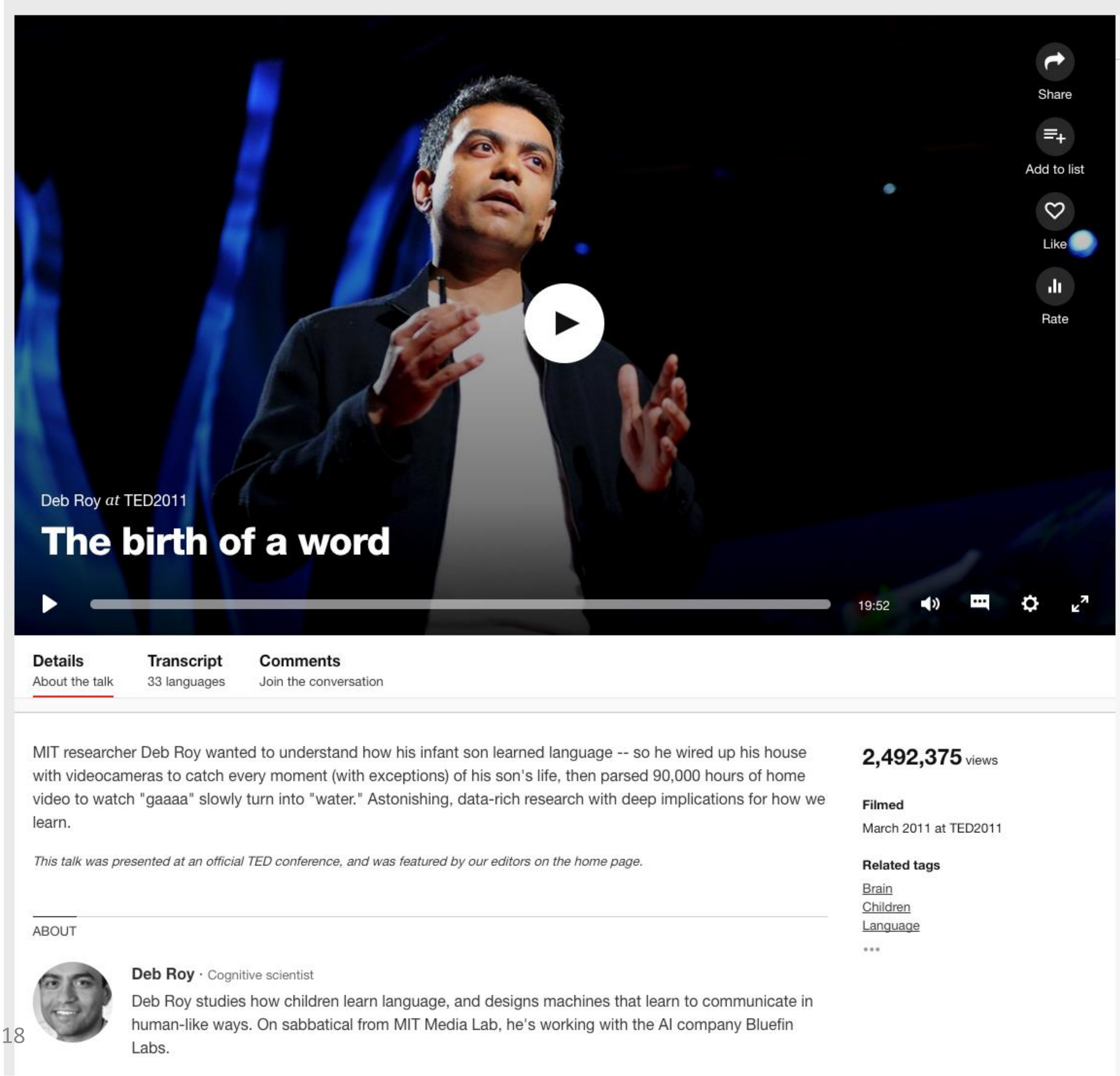


Middy Casillas

Gandhi Yetish

Heidi Colleran

https://sites.google.com/view/aclewdid/home



Deb Roy *at* TED2011

The birth of a word

MIT researcher Deb Roy wanted to understand how his infant son learned language -- so he wired up his house with videocameras to catch every moment (with exceptions) of his son's life, then parsed 90,000 hours of home video to watch "gaaaa" slowly turn into "water." Astonishing, data-rich research with deep implications for how we learn.

This talk was presented at an official TED conference, and was featured by our editors on the home page.

2,492,375 views

Filmed

March 2011 at TED2011

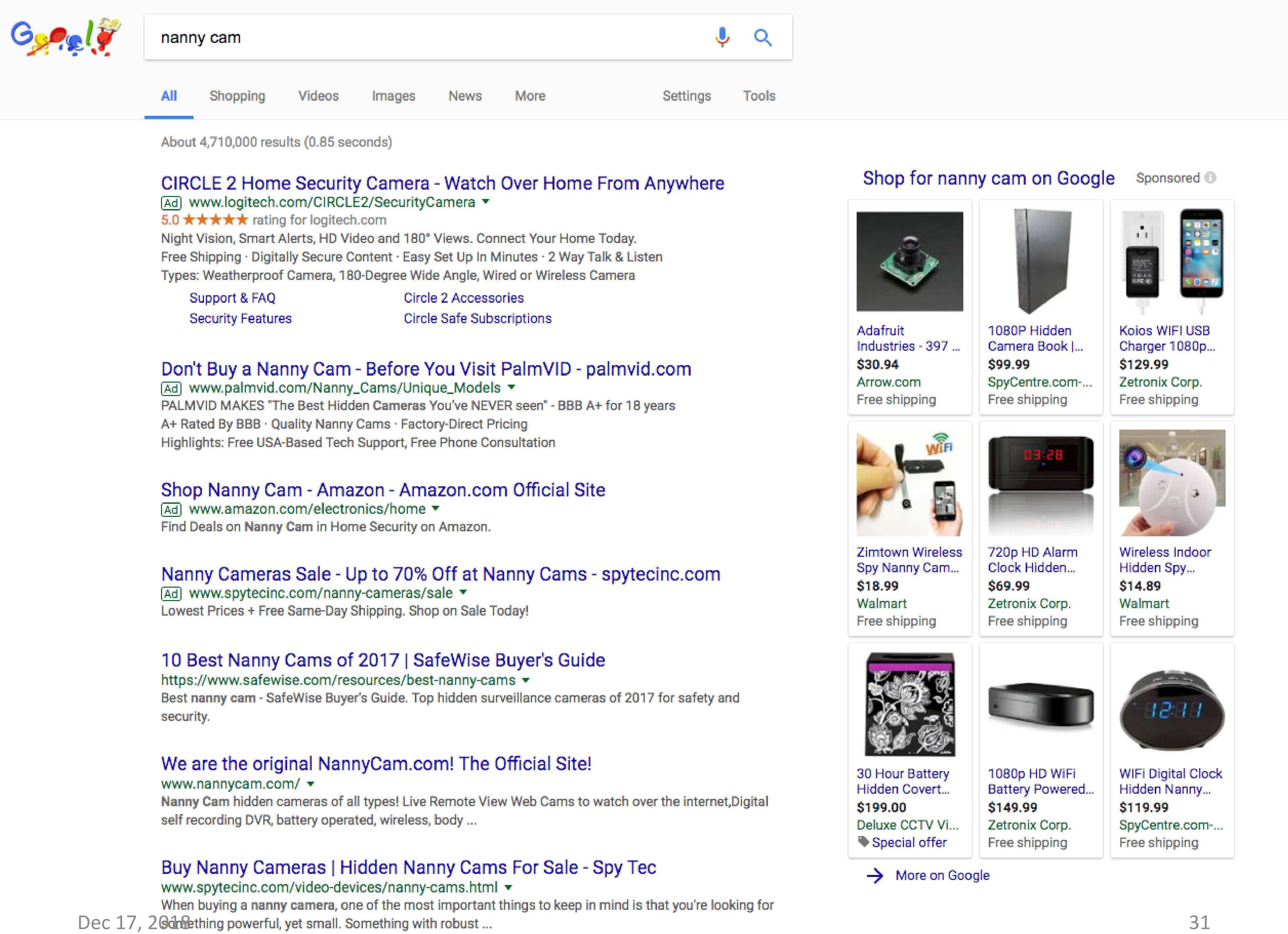
Related tags

Brain
Children
Language
•••

ABOUT

Deb Roy · Cognitive scientist

Deb Roy studies how children learn language, and designs machines that learn to communicate in human-like ways. On sabbatical from MIT Media Lab, he's working with the AI company Bluefin Labs.



About 4,710,000 results (0.85 seconds)

CIRCLE 2 Home Security Camera - Watch Over Home From Anywhere
Ad www.logitech.com/CIRCLE2/SecurityCamera
5.0 ★★★★★ rating for logitech.com
Night Vision, Smart Alerts, HD Video and 180° Views. Connect Your Home Today.
Free Shipping · Digitally Secure Content · Easy Set Up In Minutes · 2 Way Talk & Listen
Types: Weatherproof Camera, 180-Degree Wide Angle, Wired or Wireless Camera
Support & FAQ
Circle 2 Accessories
Security Features
Circle Safe Subscriptions

Don't Buy a Nanny Cam - Before You Visit PalmVID - palmvid.com
Ad www.palmvid.com/Nanny_Cams/Unique_Models
PALMVID MAKES "The Best Hidden **Cameras** You've NEVER seen" - BBB A+ for 18 years
A+ Rated By BBB · Quality Nanny Cams · Factory-Direct Pricing
Highlights: Free USA-Based Tech Support, Free Phone Consultation

Shop Nanny Cam - Amazon - Amazon.com Official Site
Ad www.amazon.com/electronics/home
Find Deals on **Nanny Cam** in Home Security on Amazon.

Nanny Cameras Sale - Up to 70% Off at Nanny Cams - spytecinc.com
Ad www.spytecinc.com/nanny-cameras/sale
Lowest Prices + Free Same-Day Shipping. Shop on Sale Today!

10 Best Nanny Cams of 2017 | SafeWise Buyer's Guide
https://www.safewise.com/resources/best-nanny-cams
Best **nanny cam** - SafeWise Buyer's Guide. Top hidden surveillance cameras of 2017 for safety and security.

We are the original NannyCam.com! The Official Site!
www.nannycam.com/
Nanny Cam hidden cameras of all types! Live Remote View Web Cams to watch over the internet,Digital self recording DVR, battery operated, wireless, body ...

Buy Nanny Cameras | Hidden Nanny Cams For Sale - Spy Tec
www.spytecinc.com/video-devices/nanny-cams.html
When buying a **nanny camera**, one of the most important things to keep in mind is that you're looking for something powerful, yet small. Something with robust ...

Shop for nanny cam on Google
Sponsored

Adafruit Industries - 397 ...
$30.94
Arrow.com
Free shipping

1080P Hidden Camera Book |...
$99.99
SpyCentre.com-...
Free shipping

Koios WIFI USB Charger 1080p...
$129.99
Zetronix Corp.
Free shipping

Zimtown Wireless Spy Nanny Cam...
$18.99
Walmart
Free shipping

720p HD Alarm Clock Hidden...
$69.99
Zetronix Corp.
Free shipping

Wireless Indoor Hidden Spy...
$14.89
Walmart
Free shipping

30 Hour Battery Hidden Covert...
$199.00
Deluxe CCTV Vi...
Special offer

1080p HD WiFi Battery Powered...
$149.99
Zetronix Corp.
Free shipping

WIFi Digital Clock Hidden Nanny...
$119.99
SpyCentre.com-...
Free shipping

→ More on Google

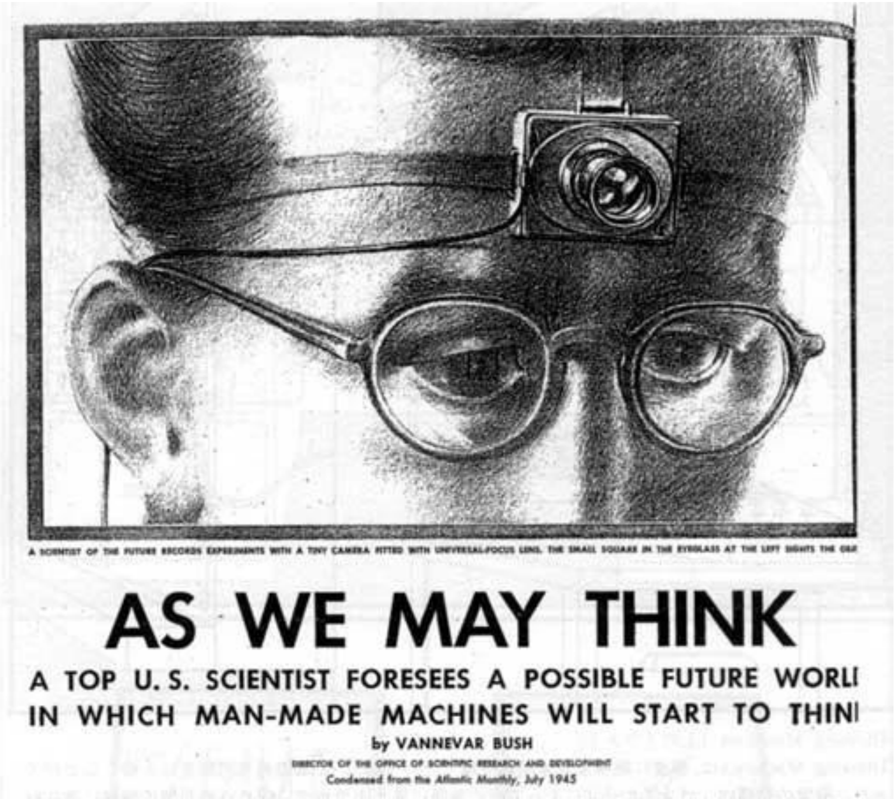
Data collection is becoming easier

- Deb Roy popularized the idea of digitizing first few years of a child's life.
  - His Ted Talk13 has 2.4M views.
- When the Human Speechome Project14 started,
  - it was necessary to install a machine room
  - in Deb Roy's basement.
- Since then, the technology has made considerable progress.

- A community is developing around DARCLE
  - (Daylong Audio Recordings of Children's Linguistic Environments)
- There is interest in collecting audio of child development across a wide range of diverse languages and social backgrounds.
- Ambitious scope → Consortia
  - HomeBank
  - TalkBank
  - CLARIN
- New directions: video, aphasia

Digital Libraries

- As we may think (1945) → WayBack Machine
- Web is large
  - But not the largest corpus
- Healthy eco-system:
  - more readers than writers
  - usage logs >> crawls



A SCIENTIST OF THE FUTURE RECORDS EXPERIMENTS WITH A TINY CAMERA FITTED WITH UNIVERSAL-FOCUS LENS. THE SMALL SQUARE IN THE EYEGLASS AT THE LEFT SIGHTS THE OB

AS WE MAY THINK

A TOP U. S. SCIENTIST FORESEES A POSSIBLE FUTURE WORLI
IN WHICH MAN-MADE MACHINES WILL START TO THINI

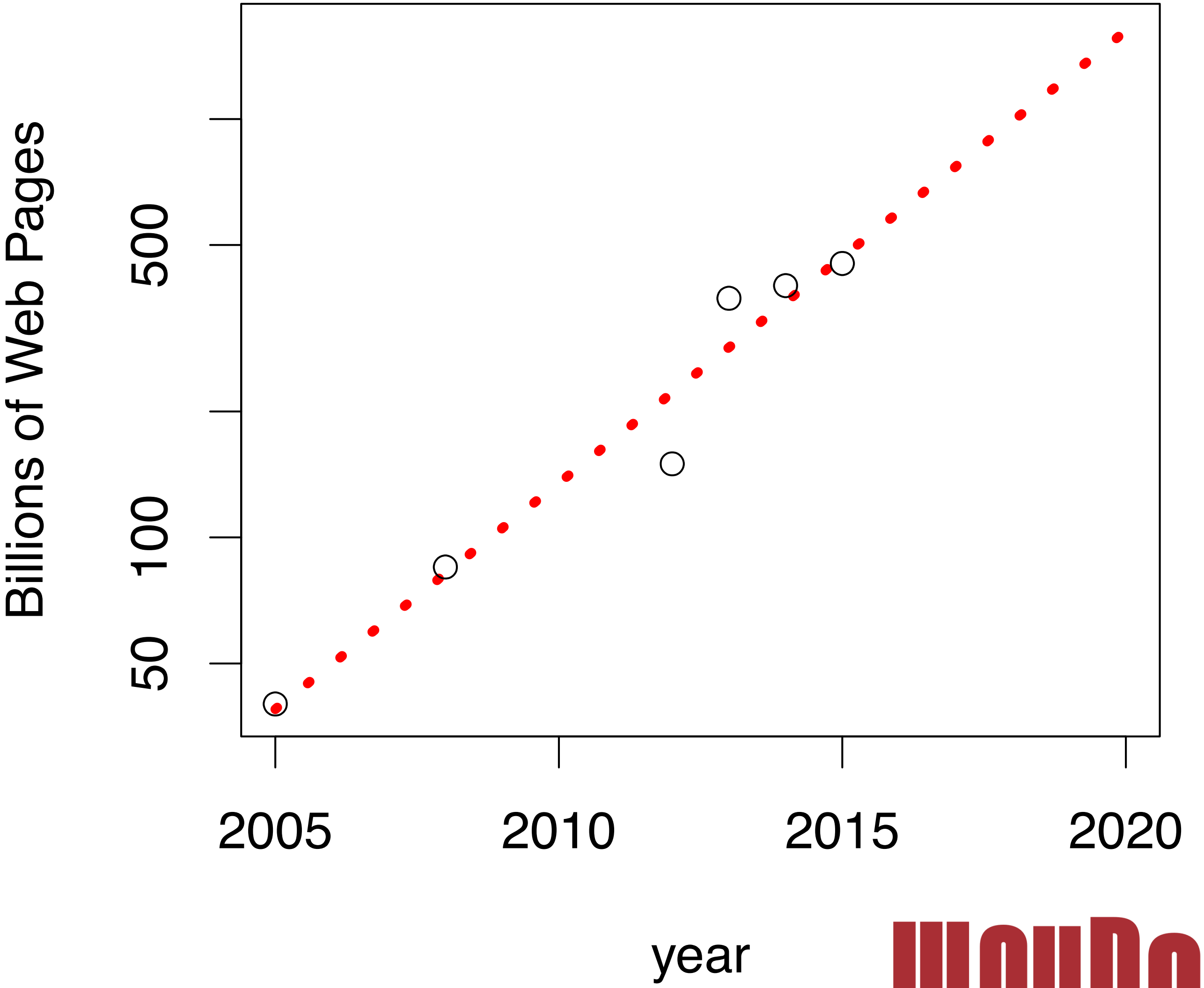
by VANNEVAR BUSH

DIRECTOR OF THE OFFICE OF SCIENTIFIC RESEARCH AND DEVELOPMENT

Condensed from the *Atlantic Monthly*, July 1945

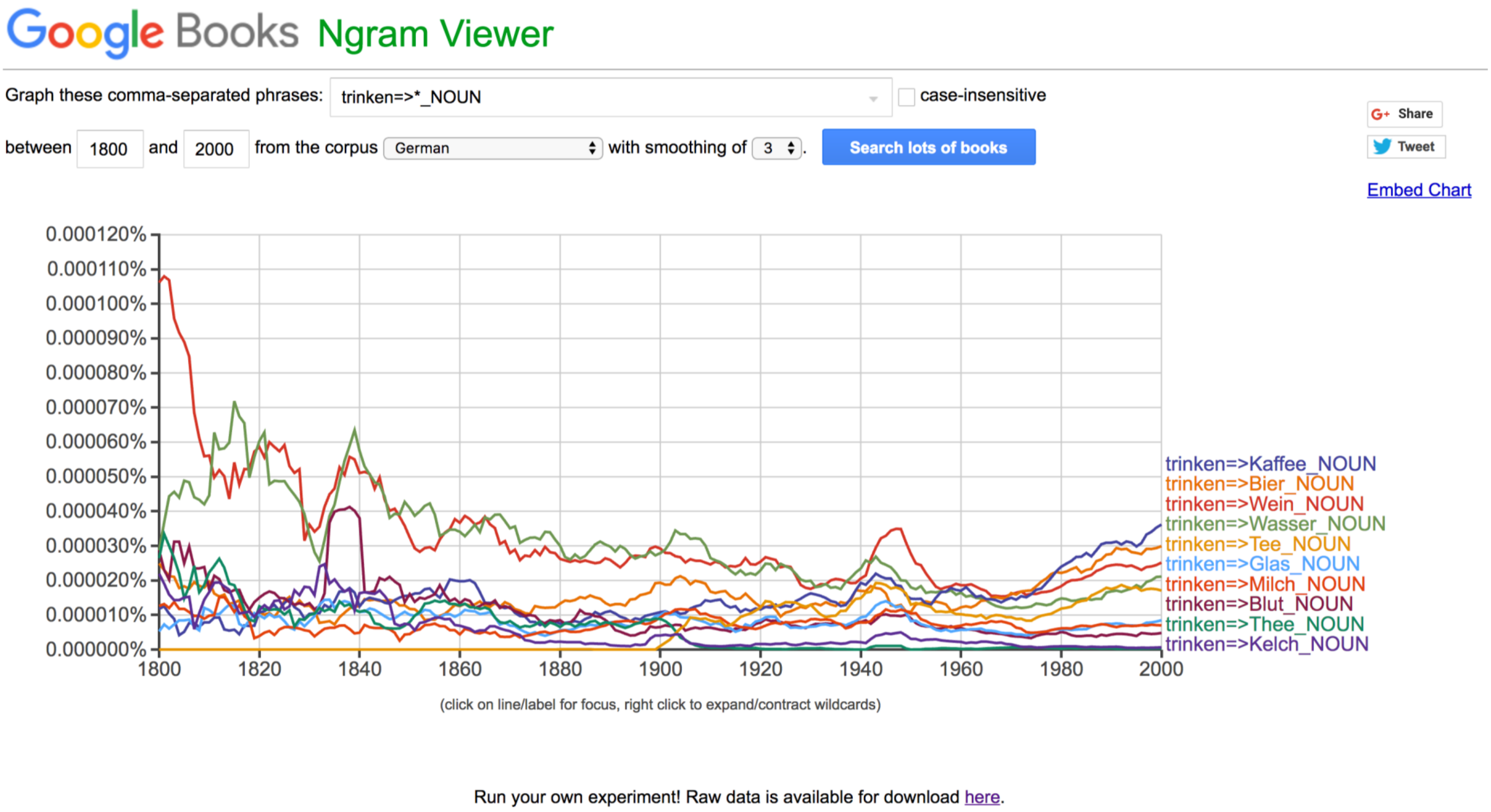
WayBackMachine

~1 Trillion Web Pages in the Wayback Mach



Dec 17, 2018

Google Books:
4% of All Books ≈ ½ Trillion Words



Google Books Ngram Viewer

Graph these comma-separated phrases: trinken=>*_NOUN □ case-insensitive

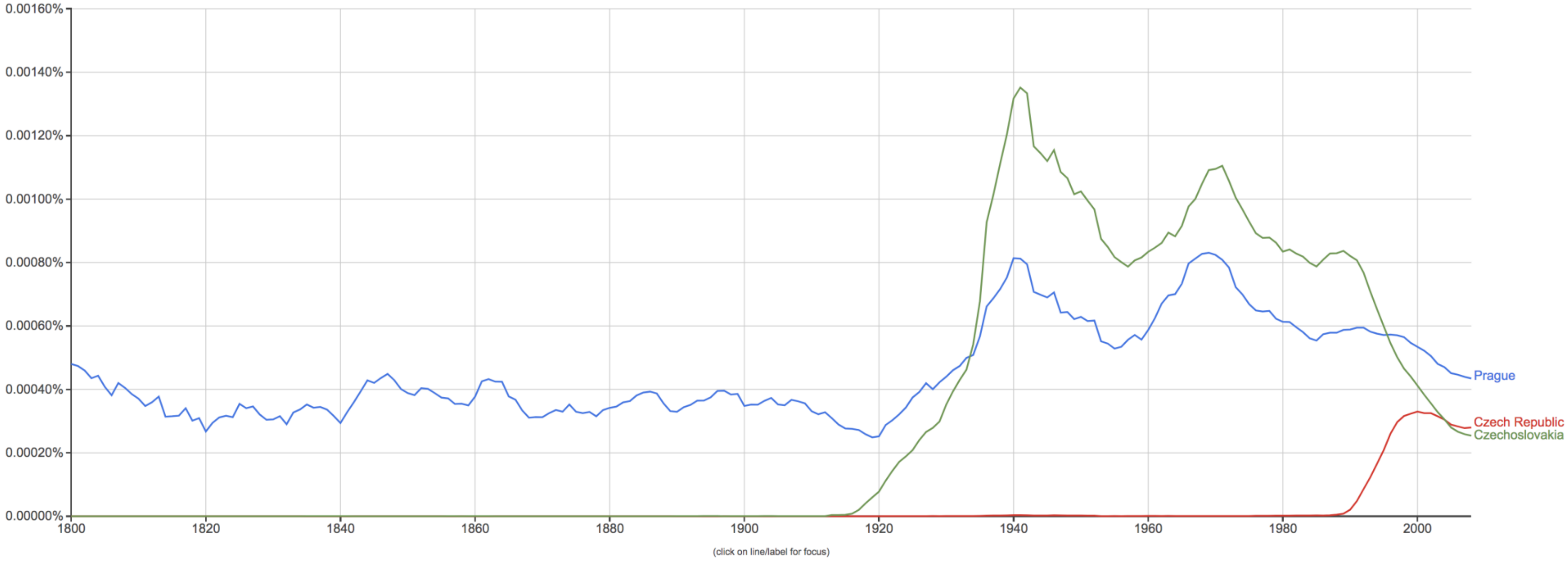
between 1800 and 2000 from the corpus German with smoothing of 3.

(click on line/label for focus, right click to expand/contract wildcards)

Run your own experiment! Raw data is available for download here.

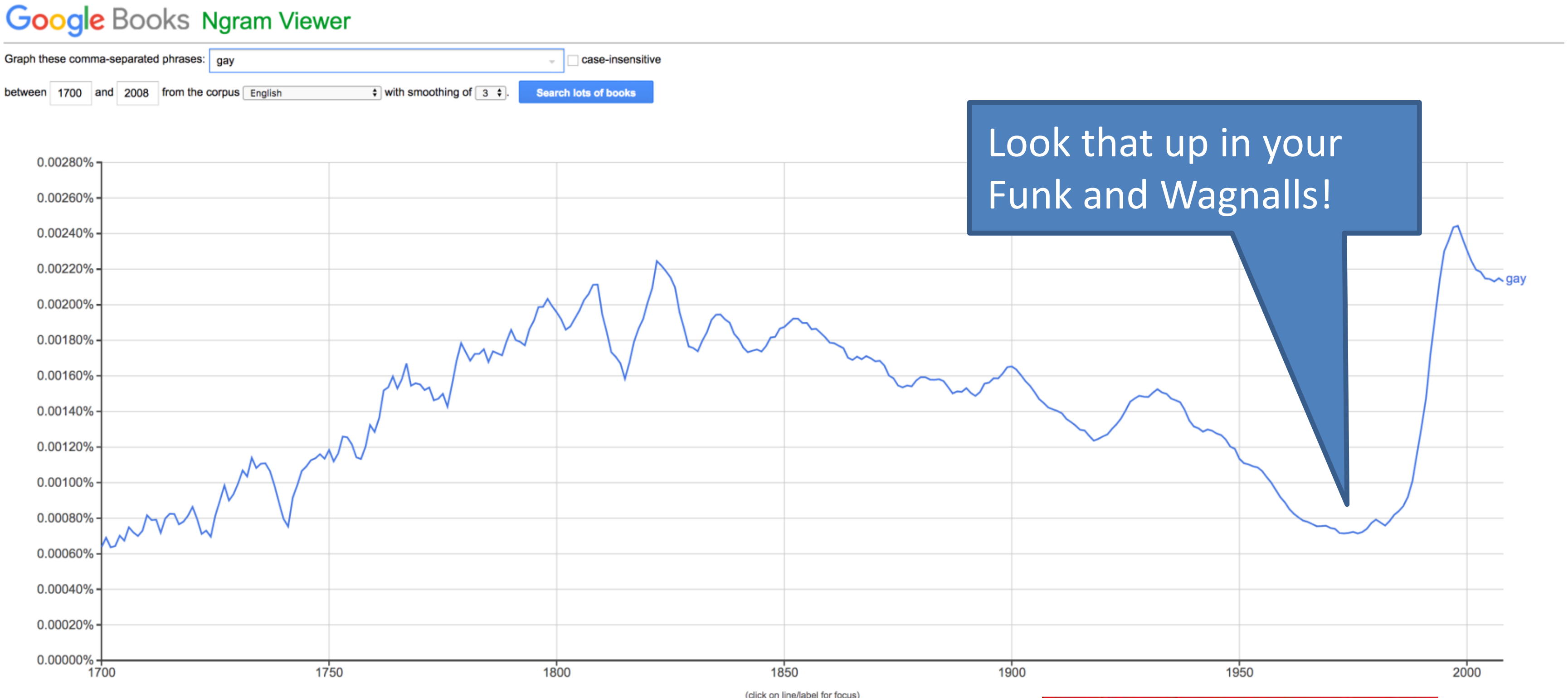
Sanity Check:

Prague, Czech Republic, Czechoslovakia



Example of Change: Gay

http://thestarryeye.typepad.com/gay/2015/03/before-gay-meant-gay.html



Google Books Ngram Viewer

Graph these comma-separated phrases: gay □ case-insensitive

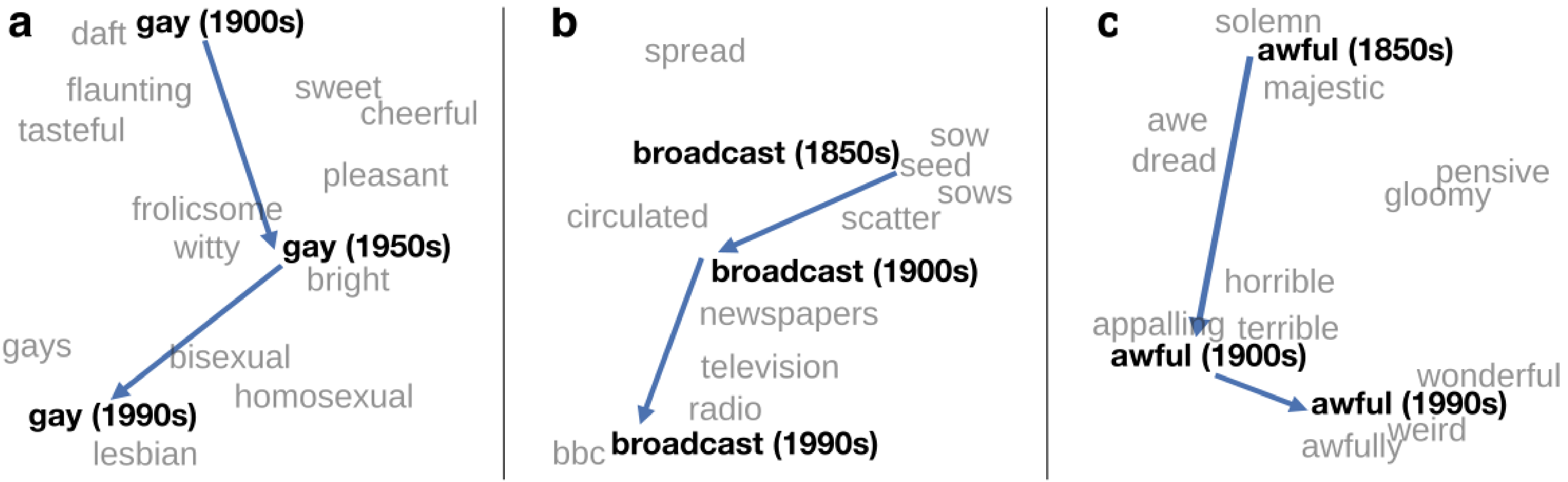
between 1700 and 2008 from the corpus English with smoothing of 3. Search lots of books

Look that up in your Funk and Wagnalls!

1962: *Let me tell you about a place*
Somewhere up-a New York way
*Where the people are so **gay***

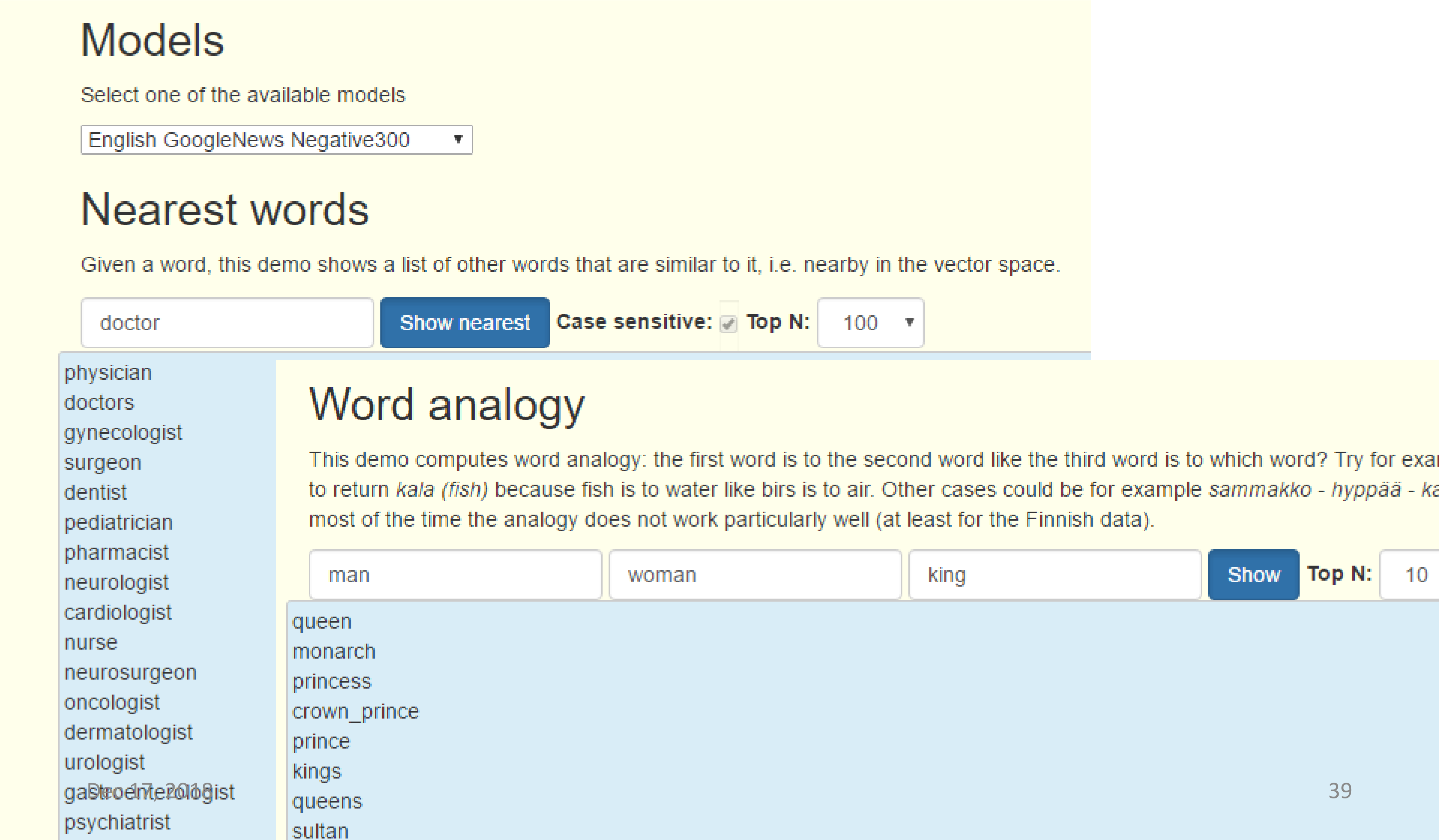
HistWords: Word Embeddings for Historical Text

William L. Hamilton, Jure Leskovec, Dan Jurafsky



Word2vec

http://bionlp-www.utu.fi/wv_demo/



Models

Select one of the available models

English GoogleNews Negative300 ▾

Nearest words

Given a word, this demo shows a list of other words that are similar to it, i.e. nearby in the vector space.

doctor

Case sensitive: ☑ Top N: 100 ▾

physician
doctors
gynecologist
surgeon
dentist
pediatrician
pharmacist
neurologist
cardiologist
nurse
neurosurgeon
oncologist
dermatologist
urologist
gastroenterologist
psychiatrist

Word analogy

This demo computes word analogy: the first word is to the second word like the third word is to which word? Try for exa
to return *kala (fish)* because fish is to water like birs is to air. Other cases could be for example *sammakko - hyppää - ka*
most of the time the analogy does not work particularly well (at least for the Finnish data).

man | woman | king

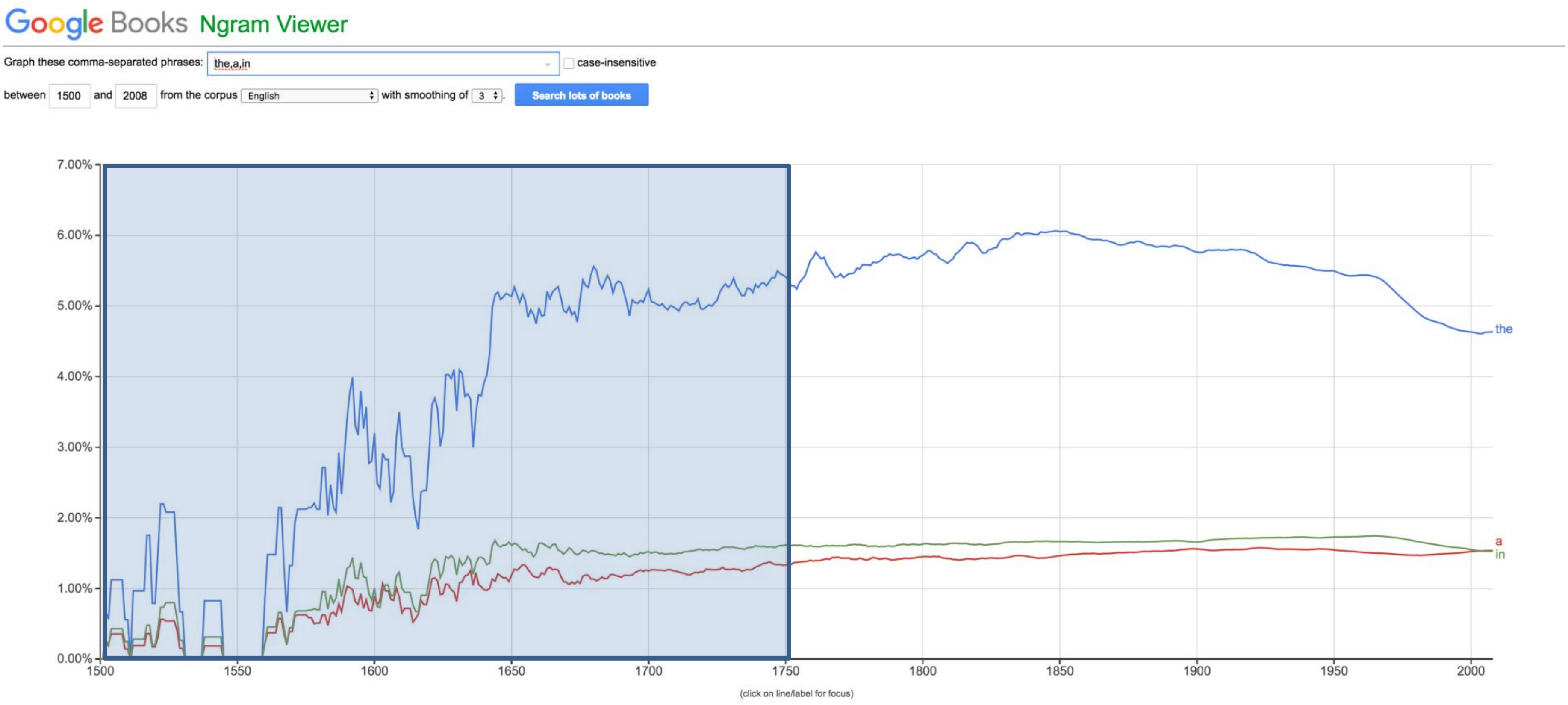
Top N: 10

queen
monarch
princess
crown_prince
prince
kings
queens
sultan

Sampling (Teaser)

- ✓ Sampling Desiderata: Corpus ≈ Sample
  - ✓ Corpus should be representative of population of interest
- ✓ Quality (representative) vs. Quantity (more):
  - ✓ More is more???
- ➢ **But in a digitized world (where we have it all)**
  - – **Do we still need to worry about sampling?**
- **If a corpus is comprehensive,**
  - – **does that imply that it is balanced?**

Google Ngrams: More Recent Estimates are More Reliable (Most books were published recently) 1500 –2000: Frequency of function words: the, a, in

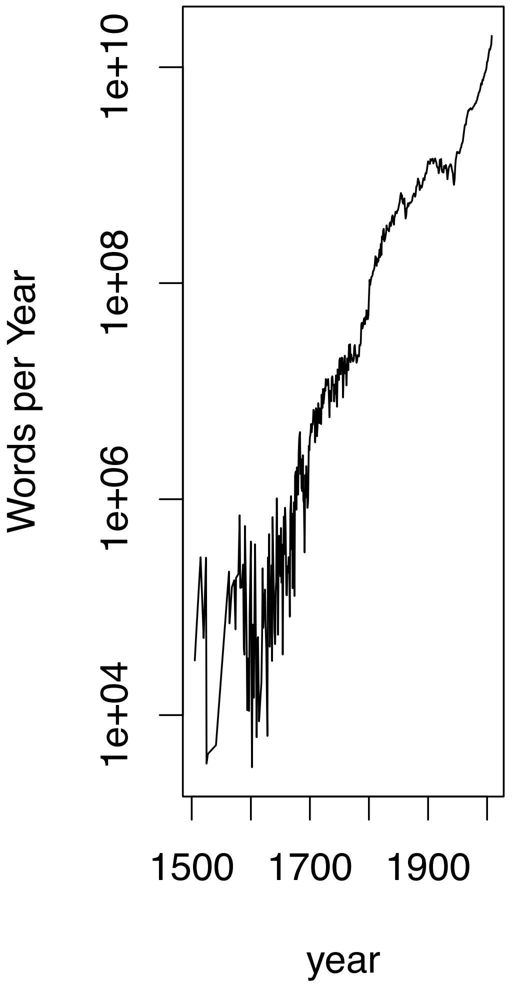


Graph these comma-separated phrases: the,a,in □ case-insensitive

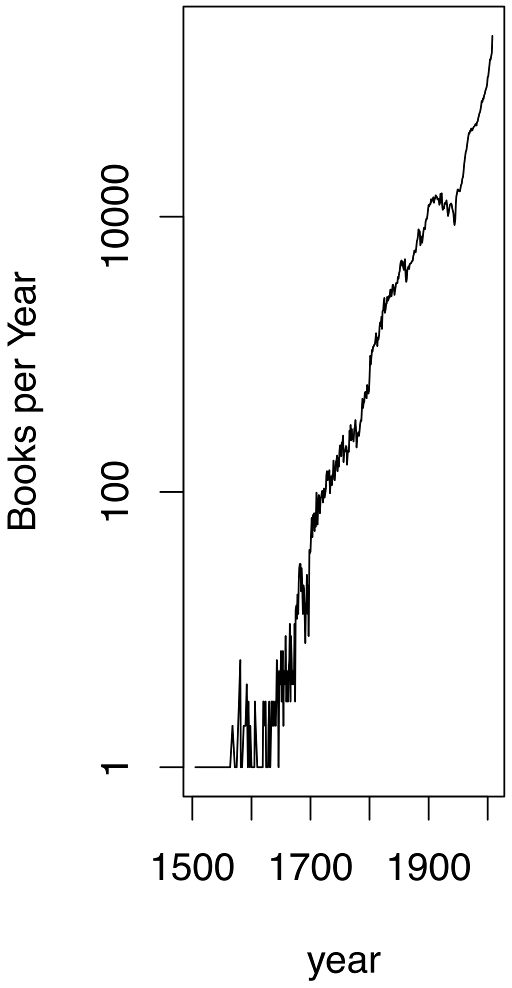
between 1500 and 2008 from the corpus English with smoothing of 3 .

Google Books: 35% Growth per Decade
More Recent Estimates are More Reliable

~0.5 Trillion Words



~5 Million Books



Conclusions

- Endless debate: quantity vs. quality
- Engineers tend to favor quantity
  - But we need your help to keep the debate going
- In a digitized world, even if we had it all,
  - Sampling is still an issue
- Practical Apps:
  - Predict future (test set)
  - From past (training data)
- Need to sample the past so it is representative of the population of interest (future)

- Future depends on many factors including:
  - Speaker/Audience
    - Adult ≠ Child ≠ Teenager
  - Register: Slang ≠ Formal
  - Spoken ≠ Written
  - Language Change:
    - Past ≠ Future
    - Time & Space