cogs 105 this week

today: basics, and culturomics

Where I Was

CAP/UC Feb 27-28th, 2015

Modeling Social Coordination & Joint-Action Workshop

Center for Cognition, Action & Perception
University of Cincinnati, Feb 27-28th, 2015

A fundamental feature of social behavior is the co-present interactions that characterize everyday social activity. The success of such interactions, whether measured in terms of social connection, goal achievement, or the ability of an individual or group of individuals to understand and predict the meaningful intentions and behaviors of others, is not only dependent on the processes of social cognition, but also on the between-person perceptual-motor coordination that makes such co-present interactions possible. Understanding and modeling the dynamics of social motor coordination, including how it emerges and is maintained over time, as well as how its stable states are activated, dissolved, transformed, and exchanged over time, is therefore an extremely important endeavor. Accordingly, this workshop is part of a larger National Institutes of Health research project (R01GM105045-01) that is directed towards *developing a dynamical modeling strategy for capturing the self-organized behavioral dynamics of goal-directed physical activity among socially coordinated human agents and how the dynamics of such tasks are influenced by physical, informational, and task-goal properties*. The overall aim of the proposed project is to build models of the temporal and spatial patterns that dynamically emerge during a number of different cooperative and competitive movement based joint-action tasks. More specifically, the project aims to:

1. Model the behavioral dynamics of both deliberate and spontaneous inter-agent coordination patterns—specifically those in which co-actors coordinate complementary and goal-directed movements, and take on different action roles (e.g., leader-follower, speaker-listener, attacker-defender)—as self-organized and emerging from relationships among the physical and informational characteristics of both the agents themselves and their task environments.
2. Model not only the steady-state properties of the interaction structures of social motor coordination (e.g., how stable they are and how resistant they are to perturbation), but also—because coordination structures in natural social tasks are typically time-varying (e.g., in terms of individual sub-roles and leader/follower relationships)—model and understand the symmetry-breaking bifurcations and dynamics of switching between multiple behavioral steady states of social coordination.

Big Data

- Today: the basics, and culturomics.
- Thursday: neurosynth.
- Next week: language analysis and modeling.

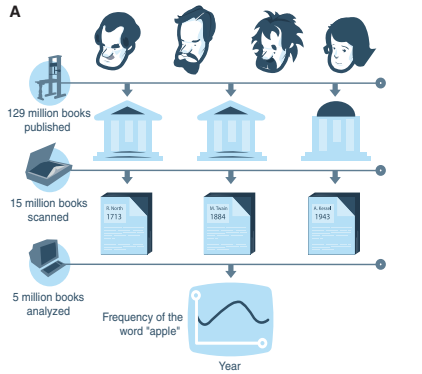
BIGDATA

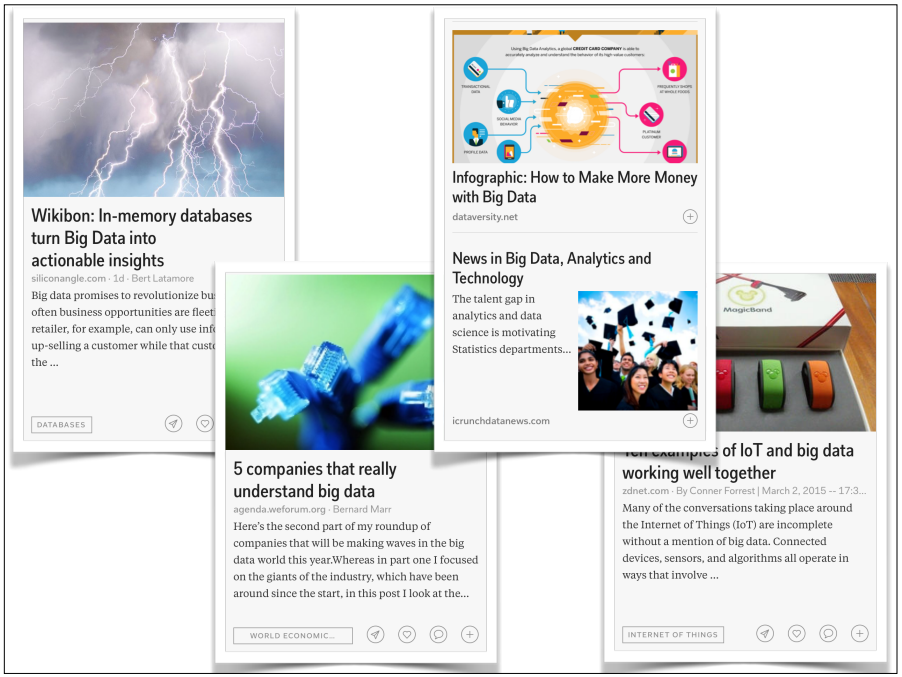
culture (Google Ngram)

the brain (Neurosynth)

BIGDATA

culture (Google Ngram)





Harvesting Naturally Occurring Data Sets to Further Theory-building in Cognitive Science

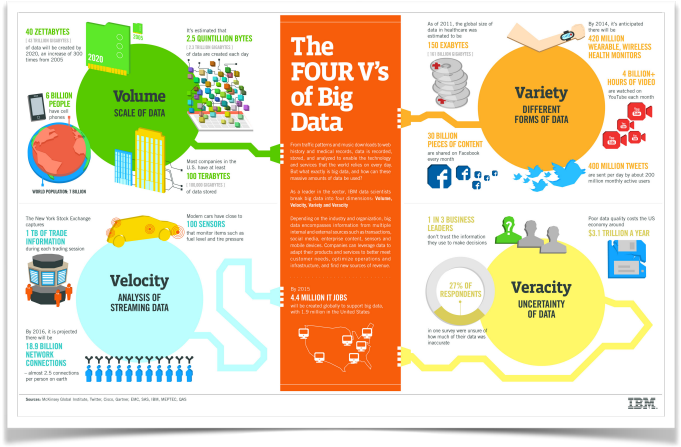
The very expertise with which psychologists wield their tools for achieving laboratory control may have had the unwelcome effect of blinding us to the possibilities of discovering principles of behavior without conducting experiments. When creatively interrogated, a diverse range of large, real-world data sets provides powerful diagnostic tools for revealing principles of human judgment, perception, categorization, decision making, language use, inference, problem solving, and representation. Examples of these data sets include web site linkages, dictionaries, logs of group interactions, image statistics, large corpora of texts, history of financial transactions, photograph repository tags and contents, trends in twitter tag usage and propagation, patent use, consumer product sales, performance in high-stakes sporting events, dialect maps over time, and scientific citations. The goal of this issue is to present some exemplary case studies of mining large data sets to reveal important principles and phenomena in cognitive science, and to discuss some of the underlying issues involved with conducting traditional experiments, large data analyses, and the synthesis of both methods.

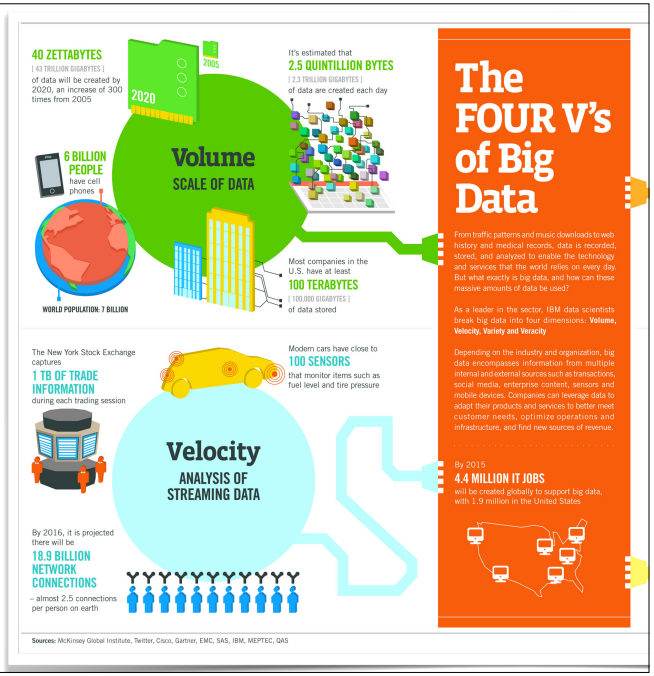
Led by Goldstone, Indiana U.
mathematical symposium in Pasadena

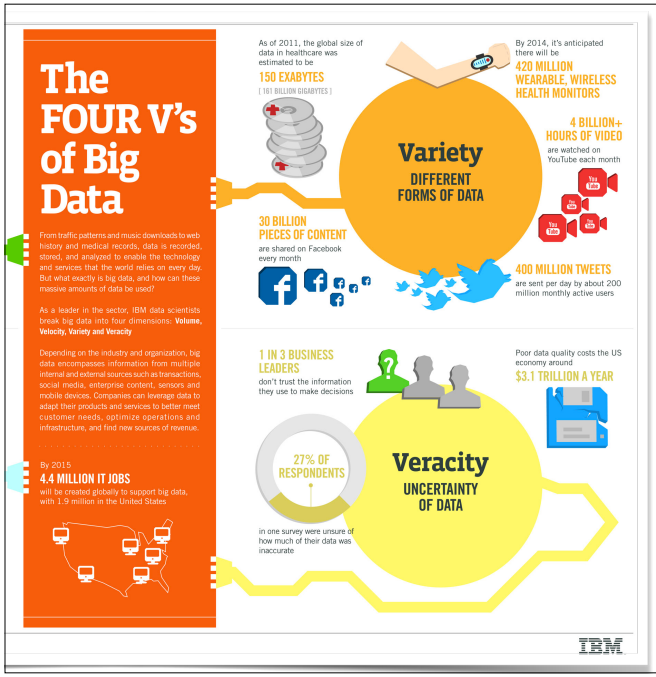
"Big Data"

- Big data is a **broad trend** in technology and science that is seeing professionals utilizing massive amounts of data in order to test new ideas, render new predictions and analyses, and so on.
- "Big data" = **data big enough that it is awkward, in some way, to work with**.
- But "big data" does **not** have to be defined purely in terms of volume.
  - IBM's data science / big data group has a nice figure describing general issues with big data:

4 V's of Big Data







Why Big Data?

- Big data means… big data — we have **more power** to test sometimes subtle cognitive hypotheses.
  - This is precisely interpretable in the "statistical power" sense as we discussed in the first part of class.
- We can **bridge cultural analysis with cognitive theory** — such as in culturomics.
- Big data **can sometimes** be even more **ecologically valid** than laboratory data.
- Big data can sometimes be **very easy to acquire**, especially given increasing volume and velocity.

A Recent Manifesto

- Tom Griffiths, UC Berkeley

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Manifesto for a new (computational) cognitive revolution

Thomas L. Griffiths *

Department of Psychology, University of California, Berkeley, United States

ARTICLE INFO

Article history:
Available online xxxx

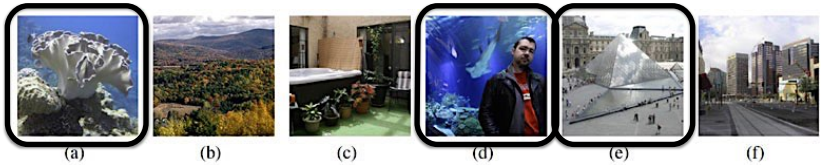
Keywords:
Computational modeling
Big data
Crowdsourcing

ABSTRACT

The cognitive revolution offered an alternative to merely analyzing human behavior, using the notion of computation to rigorously express hypotheses about the mind. Computation also gives us new tools for testing these hypotheses – large behavioral databases generated by human interactions with computers and with one another. This kind of data is typically analyzed by computer scientists, who focus on predicting people's behavior based on their history. A new cognitive revolution is needed, demonstrating the value of minds as intervening variables in these analyses and using the results to evaluate models of human cognition.

An Example

- What makes an image memorable?



"In addition to applications in the gaming and entertainment industries (e.g for building visual memory games and more efficient mnemonic techniques), the impact of such algorithms in education, clinical research and re-education could be phenomenal: measuring visual memory degradation, and understanding more precisely what aspects of visual memory are deficient in specific psychological or brain disorders is an expanding area of research."

Aude Oliva, MIT

http://cvcl.mit.edu/memorableImages.html

Construct Validity?

- All the concerns we raised about construct validity in the last part of the course can be raised here.
- How do we connect big data — sometimes at a very large scale — to questions about how we process information?
- **Scale cascade thesis**: decisions even by individuals in their local, peculiar context get aggregated across millions of people making those decisions, and so the individual scale cascades into a very large population scale, permitting us to see patterns across millions of individuals.
  - Big data = **"macroscope"**

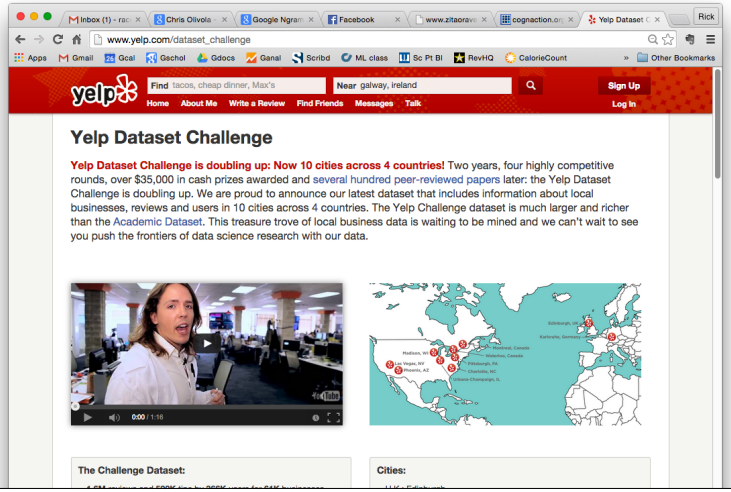
But How?

- Areas that use big data are in need of cognitive scientists like us to help **frame and explore new questions about human behavior and cognition**.
- But, there is no precise recipe for a big-data project.
- We have to apply the **cascade thesis**:
  - If people do X in situation Y, what patterns Z would we expect in data that measures thousands upon thousands (or millions!) of people?
  - Where can I get big data that will let me have a "macroscope" onto the relationship between X/Y/Z?

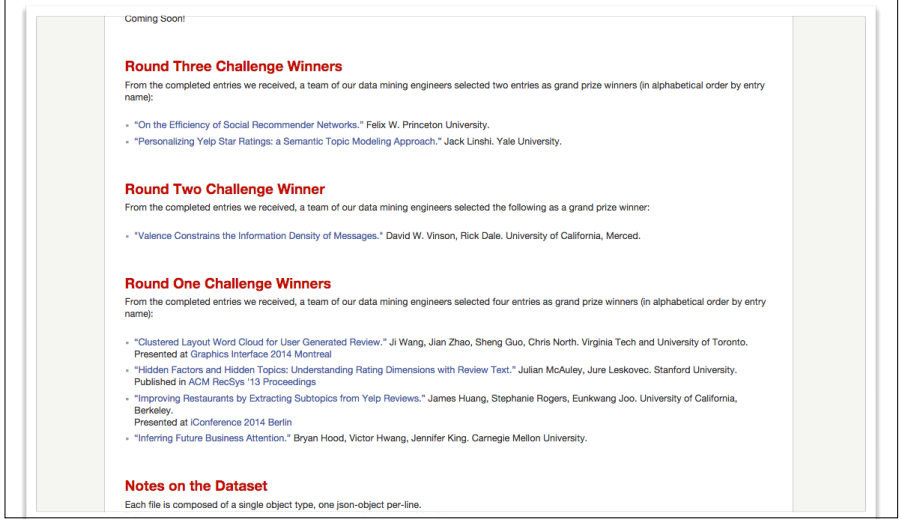
Get Your Macroscope On

- Online databases
  - **Public**, such as Google Ngram — something made widely available and easy to use.
  - **Private**: Sometimes companies will allow you access to their data under various conditions; need to arrange a contract or set of conditions.
- Internet scraping ("do it yourself")
  - **Simple aggregation**, such as downloading a large number of web pages from one site. Simple and easy to do so.
  - **Content identification** is more complicated; might have to write some programming to "crawl" through content of interest (e.g., news comments, sports statistics, etc.).

Another Example



Previous Winners



Coming Soon!

Round Three Challenge Winners

From the completed entries we received, a team of our data mining engineers selected two entries as grand prize winners (in alphabetical order by entry name):

- "On the Efficiency of Social Recommender Networks." Felix W. Princeton University.
- "Personalizing Yelp Star Ratings: a Semantic Topic Modeling Approach." Jack Linshi. Yale University.

Round Two Challenge Winner

From the completed entries we received, a team of our data mining engineers selected the following as a grand prize winner:

- "Valence Constrains the Information Density of Messages." David W. Vinson, Rick Dale. University of California, Merced.

Round One Challenge Winners

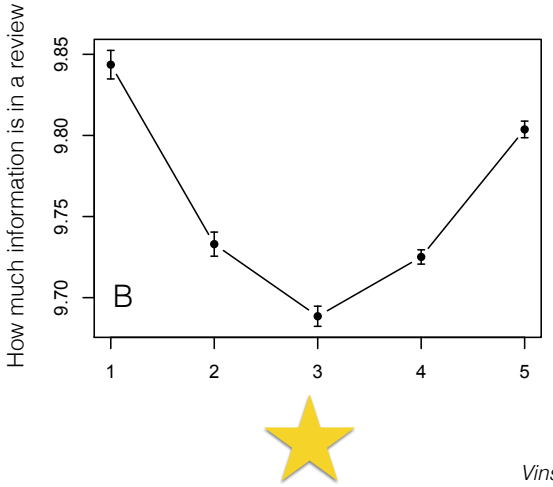
From the completed entries we received, a team of our data mining engineers selected four entries as grand prize winners (in alphabetical order by entry name):

- "Clustered Layout Word Cloud for User Generated Review." Ji Wang, Jian Zhao, Sheng Guo, Chris North. Virginia Tech and University of Toronto. Presented at Graphics Interface 2014 Montreal
- "Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text." Julian McAuley, Jure Leskovec. Stanford University. Published in ACM RecSys '13 Proceedings
- "Improving Restaurants by Extracting Subtopics from Yelp Reviews." James Huang, Stephanie Rogers, Eunkwang Joo. University of California, Berkeley. Presented at iConference 2014 Berlin
- "Inferring Future Business Attention." Bryan Hood, Victor Hwang, Jennifer King. Carnegie Mellon University.

Notes on the Dataset

Each file is composed of a single object type, one json-object per-line.

What We Found

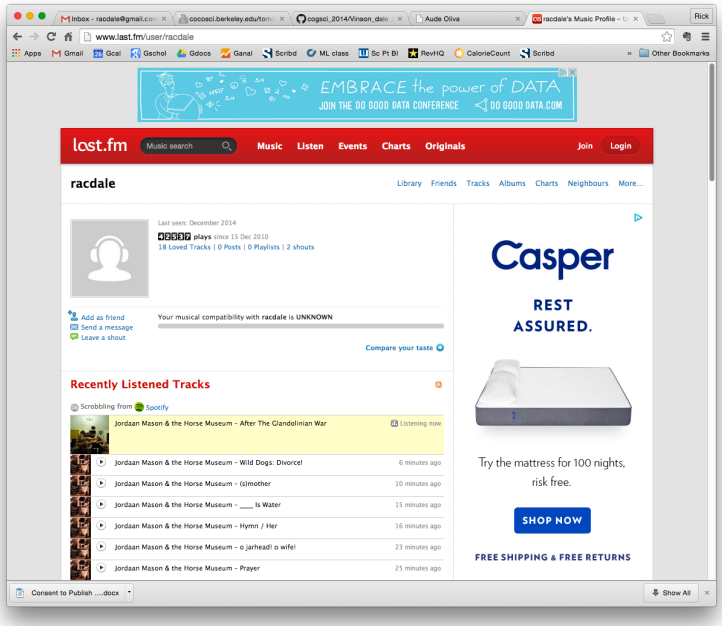


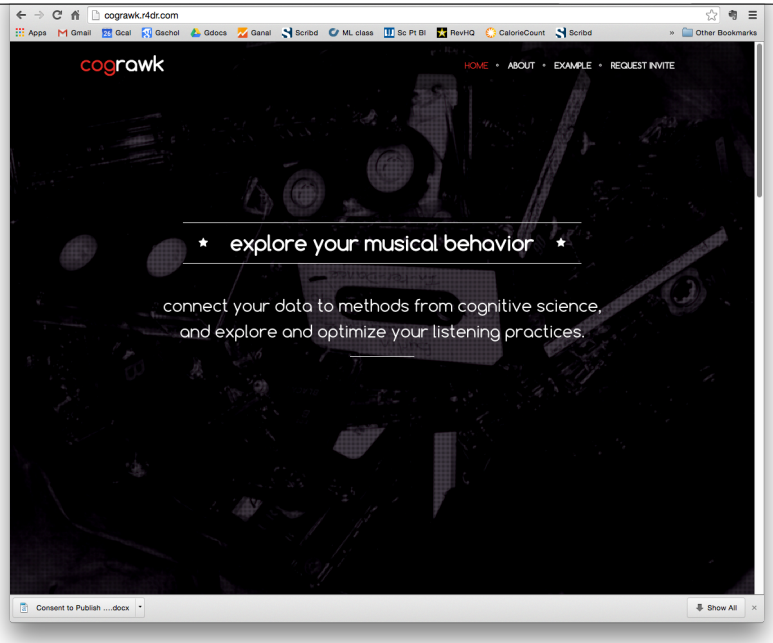
Vinson & Dale, 2014

Another Example

- At UC Merced we are interested in dynamics and human behavior. Why not apply our trade to some areas of broad interest?
- We can use Big Data to study ourselves…
- Side project: last.fm aggregation.

last.fm

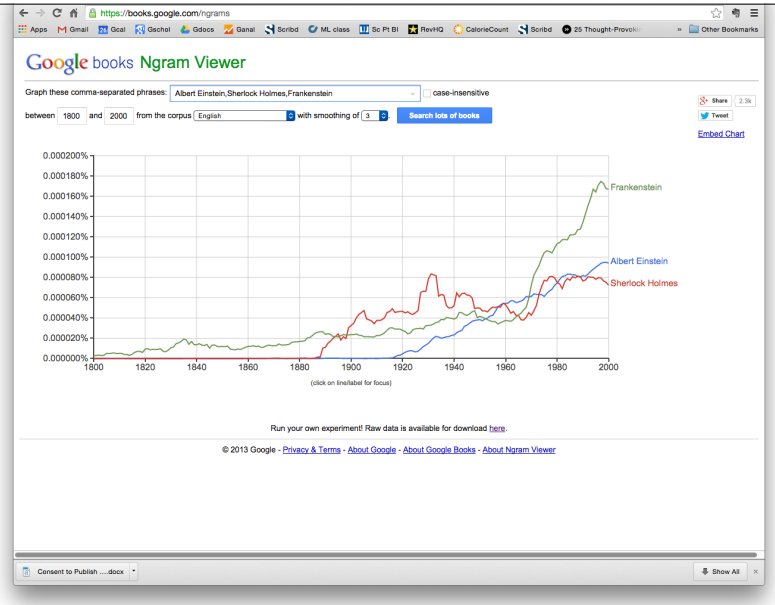




Culturomics

Culturomics

- Google teamed up with researchers to release 4% of all published books in history, digitized into n-grams for us to explore and analyze.
- Represents massive **volume**, but also **increasing velocity**.
- Could there be problems of variety and veracity in Google Ngram?
  - Of course, because we have many different kinds of publications and literatures, and also words may be used in diverse contexts.
  - For example, are words being “used” or “mentioned.”



What’s an n-gram?

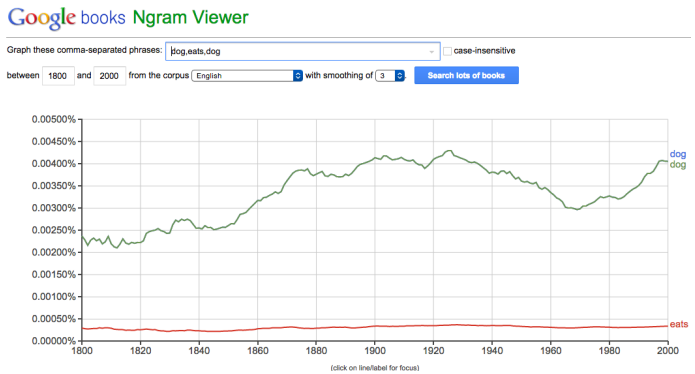
- An n-gram is a sequence of words of length n:
  - Here are four 1-grams (or, “unigrams”): [Spivey], [eating], [run], [fresh]
  - Three 2-grams (or “bigrams”): [Spivey pontificates], [eating apples],[Yoshimi expostulates]
  - Two 3-grams (or “trigram”): [the cat puked], [Ronald McDonald sweats]

What's an n-gram?

- n-grams have associated *frequencies*.
- Unigrams have frequencies that equal the frequency of the one word itself, of course.
- Question: Would bigrams be more or less frequent, on average, than unigrams?
  - Bigrams are much less frequent, on average, than unigrams.
  - E.g., [the person] is less frequent than [the]
- n-grams can also have frequency 0, of course! E.g., [apple cake freshened] probably has a frequency of 0, but it is still a trigram that we can investigate. Who freshened the apple cake, after all?

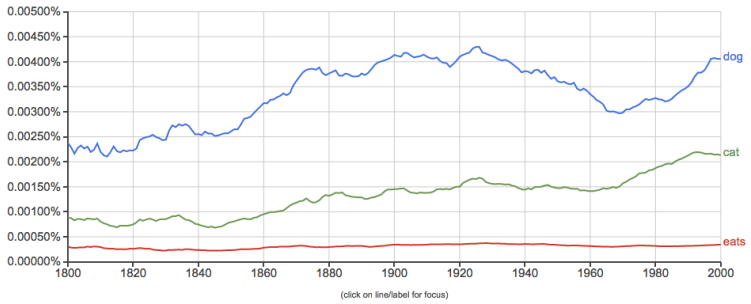
Google Ngram

- So Google Ngram lets us specify n-grams of interest, up to a length of 5.



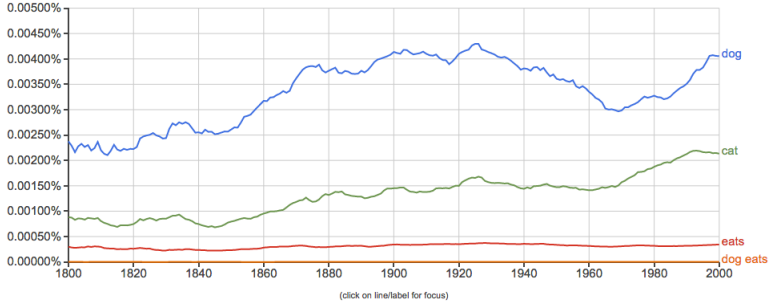
Google Ngram

Three unigrams: dog, cat, eats



Google Ngram

Add a bigram: dog eats

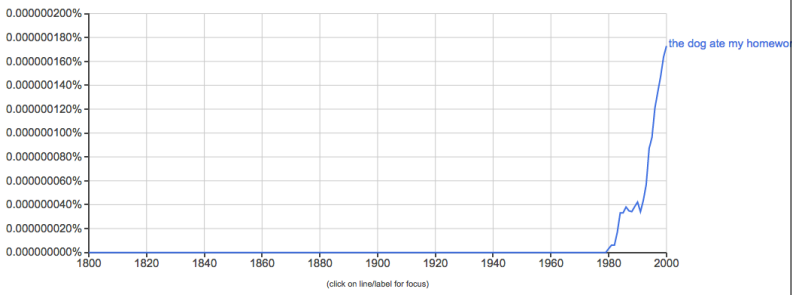


Google Ngram

A bigram and trigram: dog eats, dog eats cat

Google Ngram

Here's a 5-gram for you: the dog ate my homework



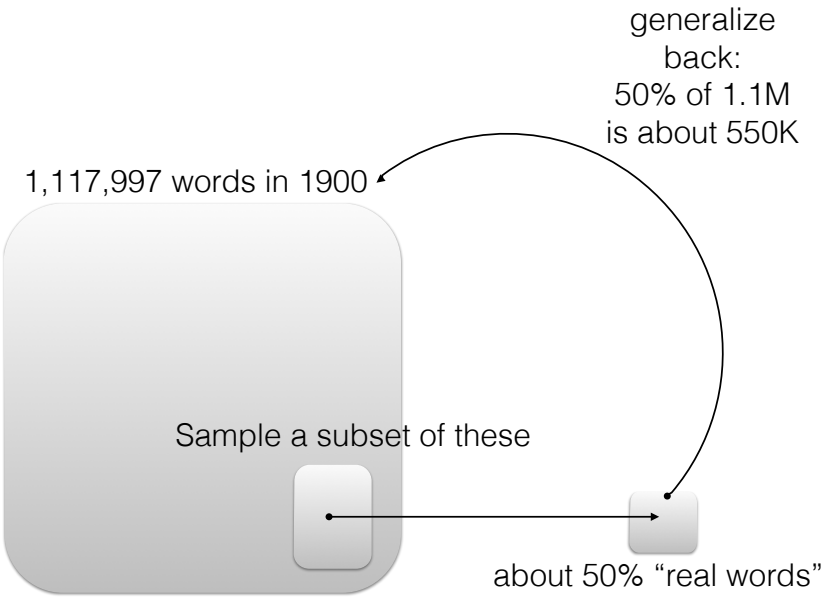
Research Questions

Remember the macroscope principle. Individual decisions cascade over millions of people and generate patterns at a higher level. What is Google Ngram good for?

How big is the English lexicon?

How Big?

- In Michel et al., they compiled a list of all common 1-grams in 1900, 1950, and 2000.
  - 1,117,997, 1,102,920, and 1,489,337, 1-grams respectively in 1900, 1950, and 2000.
  - Lots of these were junk of course (numbers, stray letters, etc.).
- So: They **sampled a subset of the unigrams (words)** in each year, and estimated the % of that sample that were "real" words; they can then generalize that % to the total count.

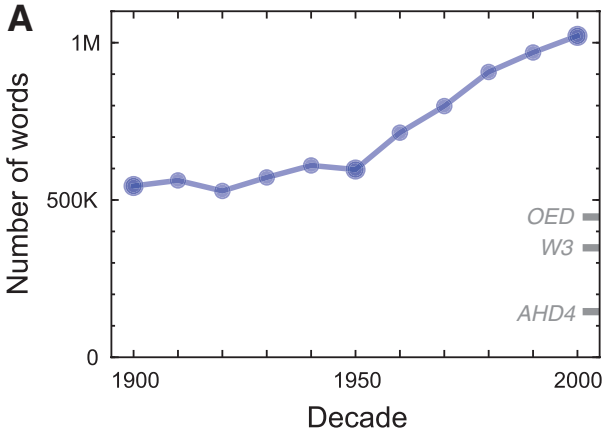


Result

Using this technique, we estimated the number of words in the English lexicon as 544,000 in 1900, 597,000 in 1950, and 1,022,000 in 2000. The lexicon is enjoying a period of enormous growth: The addition of ~8500 words/year has increased the size of the language by over 70% during the past 50 years (Fig. 2A).

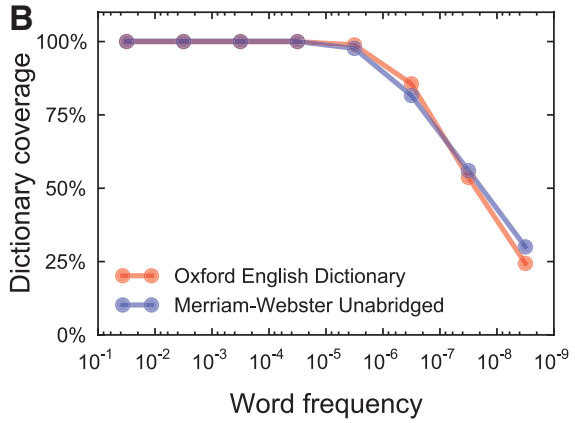
From Michel et al. reading

Result



From Michel et al. reading

Result

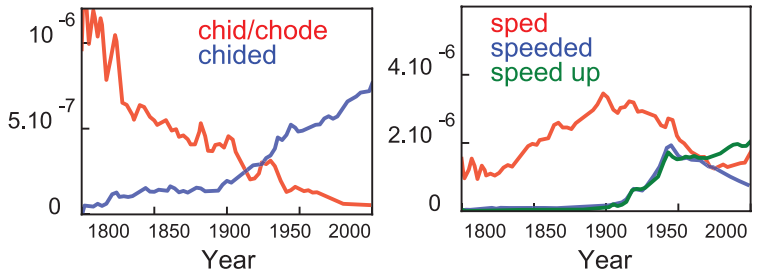


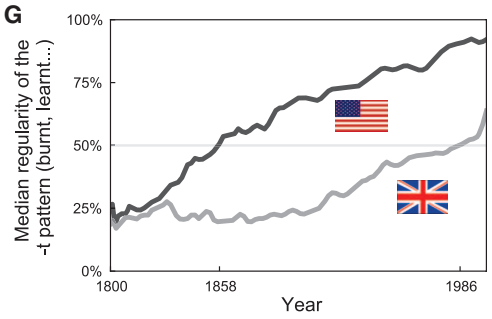
From Michel et al. reading

Research Questions

Remember the macroscope principle. Individual decisions cascade over millions of people and generate patterns at a higher level. What is Google Ngram good for?

How is grammar changing?



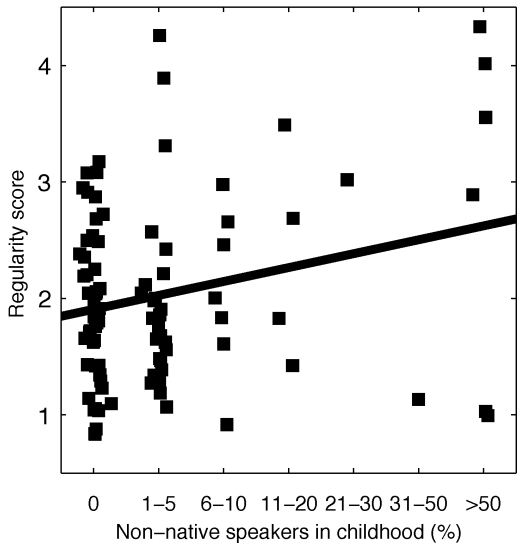


US regularized at a faster rate than the UK… why?

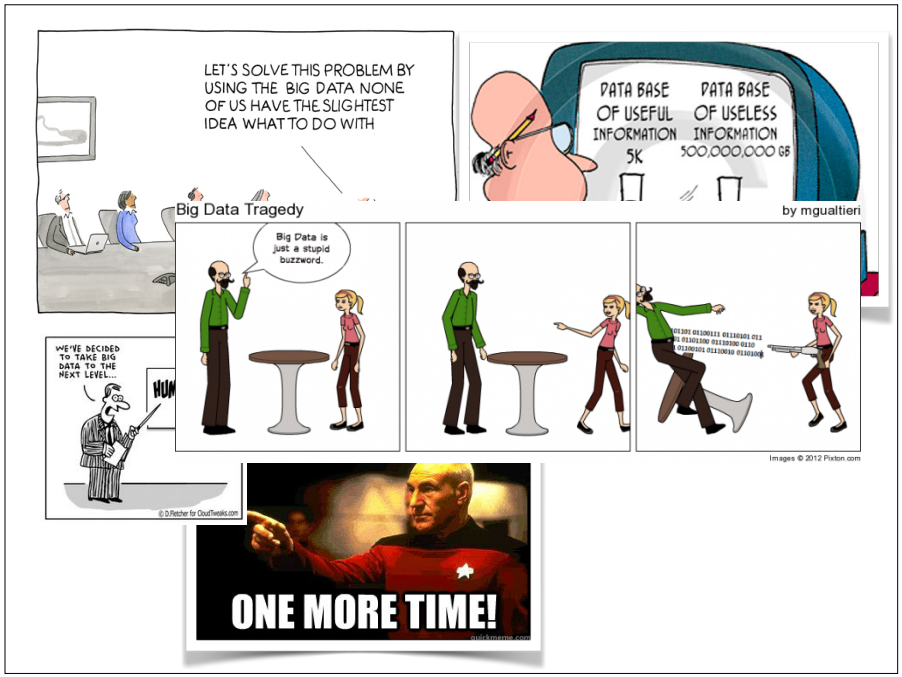
From Big Data to Behavioral Study

- Big Data can help us make predictions for testing outside of the lab.
- We used the finding from Michel et al. to see if people in densely populated regions are more likely to endorse regularization: slept —> sleeped.
- Crowdsourced many participants online and asked how many people were non-native English speakers around them.
- If you hear non-standard English during your days growing up, are you more willing to endorse a grammatical regularization?

Dale & Lupyan, 2012

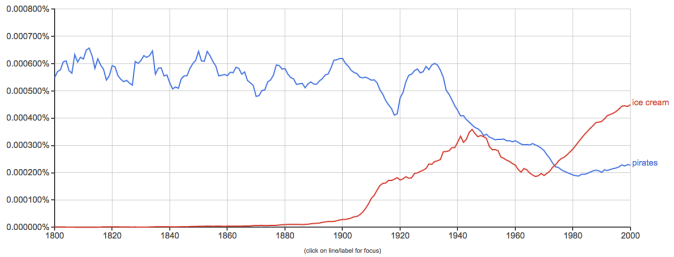


Dale & Lupyan, 2012



Dangers

- A more tangible danger other than trying to "seem fashionable" is that Big Data can produce **countless potentially trivial or uninteresting relationships**.
- Put simply: **With so much data, even the smallest correlations may come out statistically significant**; how do we separate the wheat from the chaff?



Big Data

- Today: the basics, and culturomics.
- Thursday: neurosynth.
- Next week: language analysis and modeling.