The Social Science Data Revolution

Gary King

Department of Government
Harvard University

(Horizons in Political Science talk, Government Department, Harvard University, 3/30/11)

The Changing Evidence Base of Social Science Research

The Last 50 Years:

- Survey research
- Aggregate government statistics
- In depth studies of individual places, people, or events

The Next 50 Years: Spectacular increases in new data sources, due to...

- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- The replication movement: academic data sharing (e.g., Dataverse)
- Governments encouraging data collection, distribution, experimentation (e.g., GovData)
- Advances in statistical methods, informatics, & software
- *The march of quantification*: through academia, professions, government, & commerce (*SuperCrunchers*, *The Numerati*, *MoneyBall*)

Examples of what's now possible

- Opinions of activists: A few thousand interviews $\rightsquigarrow$ millions of political opinions in social media posts (1B tweets/week)
- Exercise: A survey: "How many times did you exercise last week? $\rightsquigarrow$ 500K people carrying cell phones with accelerometers
- Social contacts: A survey: "Please tell me your 5 best friends" $\rightsquigarrow$ continuous record of phone calls, emails, text messages, bluetooth, social media connections, electronic address books
- Economic development in developing countries: Dubious or nonexistent governmental statistics $\rightsquigarrow$ satellite images of human-generated light at night, or networks of roads and other infrastructure
- Expert-vs-Statistician contests: Whenever enough information is quantified (& a right answer exists), stats wins every time
- Many, many more...

How to make progress in the new data-rich world?

1. Computer-assisted methods: Traditional quantitative-only or qualitative-only approaches are infeasible
2. Large-scale, interdisciplinary, collaborative research
3. New statistical methods & engineering required
4. Better theory: to respond to massive new evidence, privacy challenges, data-driven science

$\rightsquigarrow$ Bigger changes in the practice of social science then ever before

Two Examples

of Automated Text Analysis

Example 1: How to Read a Billion Blog Posts

(& Classify Deaths without Physicians)

- Daniel Hopkins and Gary King. "Extracting Systematic Social Science Meaning from Text" *AJPS*, $\leadsto$ commercialized via:

crimson hexagon $\leadsto$

$\leadsto$

Fast Company Names Crimson Hexagon Number Seven on "The 10 Most Innovative Companies in Web" List Leading Social Intelligence Firm Recognized For Revolutionary Measurement of Consumer Opinions in Social Media

Published: Wednesday, 16 Mar 2011 | 9:20 AM ET Text Size

CAMBRIDGE, Mass., Mar 16, 2011 (BUSINESS WIRE) -- Fast Company named

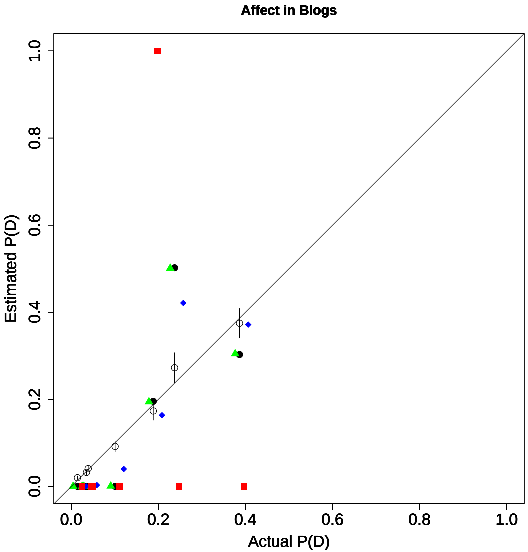
- Gary King and Ying Lu. "Verbal Autopsy Methods with Multiple Causes of Death," *Statistical Science* $\leadsto$ used by (among others):

World Health Organization

Data and Quantities of Interest

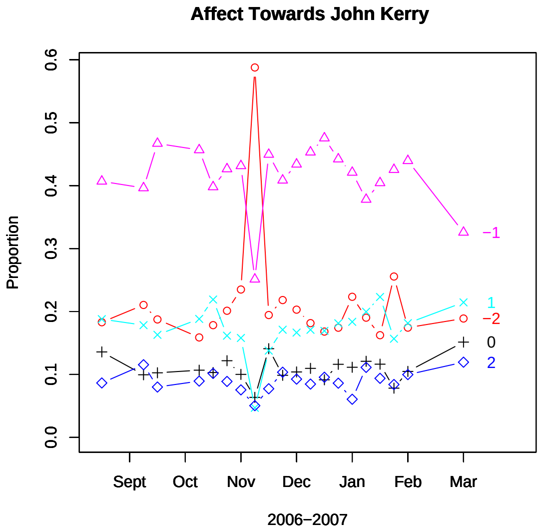
- Input Data:
  - Large set of text documents (e.g., all English language blog posts)
  - Categories (posts about US candidates): extremely negative, negative, neutral, positive, extremely positive, no opinion, not a blog
  - A small "training set" of documents hand-coded into the categories
- Quantities of interest
  - Computer science: individual document classifications (spam filters, Google searches)
  - Social Science: proportion in each category (proportion of email which is spam; proportion extremely negative comment about Pres Bush)
- Estimation
  - *Can* get the 2nd by counting the 1st (if 1st is accurate)
  - High classification accuracy $\not\Rightarrow$ unbiased category proportions
  - 70% classification accuracy is high $\Rightarrow$ disaster for category proportions
  - New methodology: unbiased category proportions, even when classification accuracy is low

Out-of-sample Comparison: 60 Seconds vs. 8.7 Days



Reactions to John Kerry's Botched Joke

You know, education — if you make the most of it . . . you can do well. If you don't, you get stuck in Iraq.



Example 2: Computer-Assisted "Reading"

- Justin Grimmer and Gary King. 2011. "General-Purpose Clustering and Conceptualization" *Proceedings of the National Academy of Sciences*.
- Conceptualization through Classification: "one of the most central and generic of all our conceptual exercises. . . . Without classification, there could be no advanced conceptualization, reasoning, language, data analysis or,for that matter, social science research." (Bailey, 1994).
- Cluster Analysis: simultaneously (1) invents categories and (2) assigns documents to categories

What's Hard about Clustering?

(Why Johnny Can't Classify)

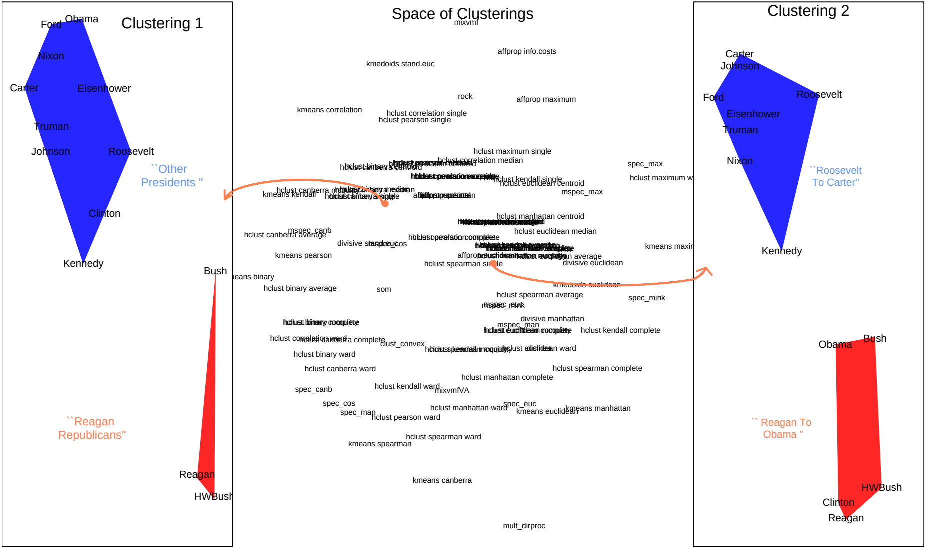
- Goal: Computer-assisted conceptualization & clustering
- Bell(n) $=$ number of ways of partitioning n objects
- Bell(2) $= 2$ (AB, A B)
- Bell(3) $= 5$ (ABC, AB C, A BC, AC B, A B C)
- Bell(5) $= 52$
- Bell(100) $\approx 10^{28} \times$ Number of elementary particles in the universe
- Now imagine choosing the *optimal* classification scheme by hand!
- Available compromises pursue different goals:
  - Standard Approach: Fully automated methods $\rightsquigarrow$ no method works well in general; impossible to know which to apply!
  - Our Approach: Computer-assisted methods $\rightsquigarrow$ You, not some computer algorithm, decides what's important, but with help

Switch from Fully Automated to Computer Assisted

- Computer-Assisted Clustering
  - Easy in theory: list all clusterings; choose the best
  - Impossible in practice: Too hard for us mere humans!
  - An organized list will make the search possible
  - Insight: Many clusterings are perceptually identical
  - E.g.,: consider two clusterings that differ only because one document (of 10,000) moves from category 5 to 6
- Question: How to organize clusterings so humans can understand?

How to Zoom Out while Reading

You choose one (or more) clustering, based on insight, discovery, useful information,. . .



Evaluation: More Informative Discoveries

- Found 2 scholars analyzing lots of textual data for their work
- Created 6 clusterings:
  - 2 clusterings selected with our method (biased against us)
  - 2 clusterings from each of 2 other methods (varying tuning parameters)
- Created info packet on each clustering (for each cluster: exemplar document, automated content summary)
- Asked for $\binom{6}{2}=15$ pairwise comparisons
- User chooses $\Rightarrow$ only care about the one clustering that wins
- Both cases a Condorcet winner:

"Immigration":

Our Method 1 $\rightarrow$ vMF 1 $\rightarrow$ vMF 2 $\rightarrow$ Our Method 2 $\rightarrow$ K-Means 1 $\rightarrow$ K-Means 2

"Genetic testing":

Our Method 1 $\rightarrow$ {Our Method 2, K-Means 1, K-means 2} $\rightarrow$ Dir Proc. 1 $\rightarrow$ Dir Proc. 2

Evaluation: What Do Members of Congress Do?

Example Discovery

Taunting ruins deliberation

Sen. Lautenberg
on Senate Floor
4/29/04

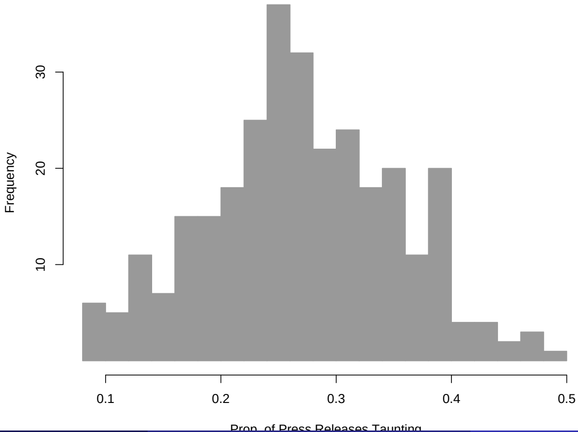
- "Senator Lautenberg Blasts Republicans as 'Chicken Hawks' " [Government Oversight]

Out of Sample Confirmation of Partisan Taunting

- Discovered using 200 press releases; 1 senator.

Out of Sample Confirmation of Partisan Taunting

- Discovered using 200 press releases; 1 senator.
- Confirmed using 64,033 press releases; 301 senator-years.
- Apply supervised learning method: measure proportion of press releases a senator taunts other party



Normative Implications of Taunting

- Partisan taunting:
  - Very common
  - Makes deliberation less likely
  - Occurs more often in homogeneously partisan districts (i.e., when preaching to the choir)
- Incompatibility of the principles of representative democracy
  - To get <u>reflection</u>: Homogeneous (noncompetitive) districts
  - To get <u>deliberation</u> (no taunting): Heterogeneous (competitive) districts
  - $\rightsquigarrow$ you can't have both!

Some New Data Types

1. Unstructured text: emails (1 LOC every 10 minutes), speeches, government reports, blogs, social media updates, web pages, newspapers, scholarly literature
2. Commercial activity: credit cards, sales data, and real estate transactions, product RFIDs
3. Geographic location: cell phones, Fastlane or EZPass transponders, garage cameras
4. Health information: digital medical records, hospital admittances, google/MS health, and accelerometers and other devices being included in cell phones
5. Biological sciences: effectively becoming social sciences as genomics, proteomics, metabolomics, and brain imaging produce huge numbers of *person-level variables*.
6. Satellite imagery: increasing in scope, resolution, and availability.
7. Electoral activity: ballot images, precinct-level results, individual-level registration, primary participation, and campaign contributions

Some More New Data Examples

8. Social media: facebook, twitter, social bookmarking, blog comments, product reviews, virtual worlds, game behavior, crowd sourcing
9. Web surfing artifacts: clicks, searches, and advertising clickthroughs. (Google collects 1 petabyte/72 minutes on human behavior!)
10. Multiplayer web games and virtual worlds: Billions of highly controlled experiments on human behavior
11. Government bureaucracies: moving from paper to electronic data bases, increasing availability
12. Governmental policies: requiring more data collection, such e.g., "No Child Left Behind Act"; allowing randomized policy experiments; Obama pushing data distribution
13. Scholarly data: the replication movement in academia, led in part by political science, is massively increasing data sharing

Enormous Emerging Opportunities for Social Scientists

- For the first time: technologies, policies, data, and methods are making it feasible to attack some of the most vexing problems that afflict human society
- A massive change from studying problems to understanding and solving problems
- Opportunities require a change in our job descriptions, with new:
  1. Computer-assisted methods: Traditional quantitative-only or qualitative-only approaches are infeasible
  2. Large-scale, interdisciplinary, collaborative research
  3. New statistical methods & engineering required
  4. Better theory: to respond to massive new evidence, privacy challenges, data-driven science
- And then there's you & me:
  - In most legislatures, courts, academic departments, . . . , change comes from replacement not conversion
  - Will we wait to be replaced? or put in the effort to convert and learn how to use the new information?

For more information

http://GKing.Harvard.edu