The Changing Evidence Base of Social Science Research

Gary King

Institute for Quantitative Social Science
Harvard University

(Miller Converse Lecture Series talk, 4/9/09)

What did they know and when did they know it?

What did they know and when did they know it?

1. One-off studies of individual places, people, or events
   - do not scale
   - are not representative
   - do not measure long-term change.
2. Aggregate Government (& other) Statistics
   - Individuals not identified
   - Highly aggregated over time and space
   - No investigator control
   - Little impartiality: Governments, newspapers, NGOs, etc.

Survey Research

Advances:

- Individual level data; no aggregation bias
- Investigator control & survey experiments
- Spawned successful literature on improving survey quality
- The first real information about opinions, attitudes, & identifications
- $\rightsquigarrow$ 1/2 of all quantitative articles in polisci use surveys

Challenges:

- Surveys provide: Occasional snapshots, of random selections, of isolated individuals, from unknown geographic locations
- Interpersonal incomparability, "non-opinions," Hawthorne effects, no direct observation of behavior
- The scientific foundation is crumbling: random selection is no longer possible with cell phone use and nonresponse
- Huge opportunities with web surveys: marginal cost ≈ 0, but what about selection?

The Evidence Base of Social Science Research

The Last 50 Years:

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

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
- Government policies encouraging data collection & experimentation
- The replication movement: academic data sharing
- The march of quantification: through academia, the professions, government, & commerce (SuperCrunchers, The Numerati)
- Advances in statistical methods, informatics, & software

Examples of what's now possible

- Exercise: A survey of how many times you exercised last week ⇝ 100K people carrying cell phones with accelerometers
- Opinions of activists: Sample of a few thousand interviews ⇝ millions of political opinions available every day in the blogosphere
- Social contacts: asking respondents to recall names of their friends over the past year ⇝ a continuous record of social contacts through phone calls, emails, text messages, bluetooth, social media connections, electronic address books
- Economic development in developing countries: Dubious or nonexistent governmental statistics ⇝ satellite images of human-generated light at night, or networks of roads and other infrastructure
- Many more coming. . .

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

1. Large-scale, interdisciplinary research
2. Computer-assisted & quantitative: Traditional approaches infeasible
3. New statistical methods & engineering required

$\rightsquigarrow$ Bigger changes than social science has ever seen

How to Read 100 Million Blogs

(& Classify Deaths without Physicians)

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

  crimson hexagon

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

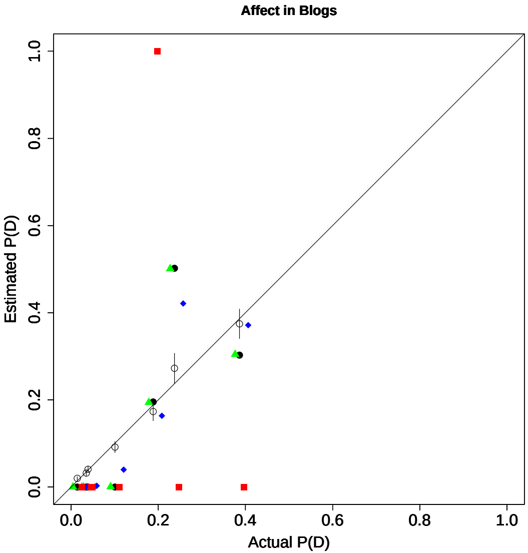
  World Health Organization

- Copies at `http://gking.harvard.edu`

Data and Quantities of Interest

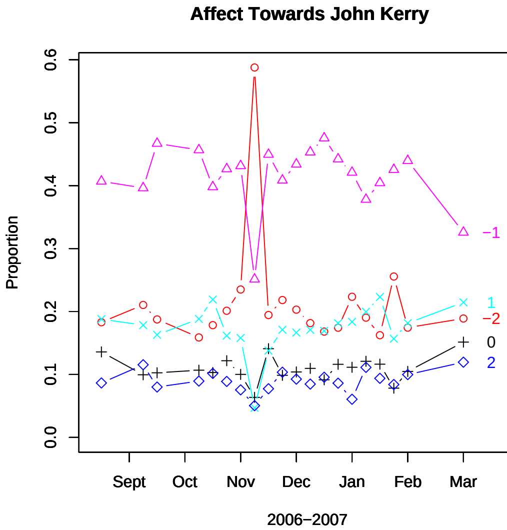
- Input Data:
  - Large set of text documents (e.g., all English language blog posts)
  - Categories (posts about US candidates): extremely negative, negative, neutral, positive, extremely positive, no opinion, not a blog
  - A small "training set" of documents hand-coded into the categories
- Quantities 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 the best 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.

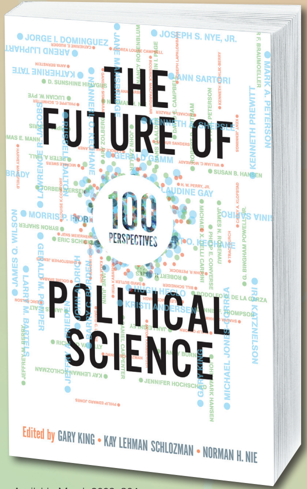


Our Software can Read Better than You!

- Reference: Justin Grimmer and Gary King. “Quantitative Discovery from Qualitative Information: A General-Purpose Document Clustering Methodology”

Why Johnny Can't Classify (Optimally)

- A new goal: Clustering/classification/typologies (no training set)
- 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
- Optimal classification by hand is absurd
- Available compromises pursue different goals:
  - Computer scientists, biologists, statisticians: information retrieval or presenting search results (Google news)
    ⇝ impossible to know in which of our data the methods will work
  - Social scientists: discovery of useful information
    ⇝ We show how to connect substance and method



Available March 2009: 304pp
Pb: 978-0-415-99701-0: **$24.95**
www.routledge.com/politics

THE FUTURE OF POLITICAL SCIENCE

100 Perspectives

Edited by **Gary King**, Harvard University, **Kay Lehman Schlozman**, Boston College and **Norman H. Nie**, Stanford University

"The list of authors in *The Future of Political Science* is a 'who's who' of political science. As I was reading it, I came to think of it as a platter of tasty hors d'oeuvres. It hooked me thoroughly."
—Peter Kingstone, University of Connecticut

"In this one-of-a-kind collection, an eclectic set of contributors offer short but forceful forecasts about the future of the discipline. The resulting assortment is captivating, consistently thought-provoking, often intriguing, and sure to spur discussion and debate."
—Wendy K. Tam Cho, University of Illinois at Urbana-Champaign

"King, Schlozman, and Nie have created a visionary and stimulating volume. The organization of the essays strikes me as nothing less than brilliant. . . It is truly a joy to read."
—Lawrence C. Dodd, Manning J. Dauer Eminent Scholar in Political Science, University of Florida

Routledge
Taylor & Francis Group
an **informa** business

Evaluators' Rate Machine Choices Better Than Their Own

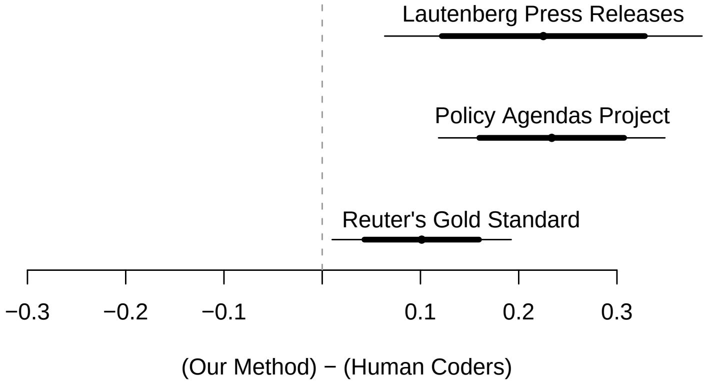
- Scale: (1) unrelated, (2) loosely related, or (3) closely related
- Table reports: mean(scale)

Pairs from	Overall Mean	Evaluator 1	Evaluator 2
Random Selection	1.38	1.16	1.60
Hand-Coded I	1.58	1.48	1.68
Hand-Coded II	2.06	1.88	2.24
Machine	2.24	2.08	2.40

p.s. The hand-coders did the evaluation!

Cluster Quality Experiments

Scale: mean(within clusters) − mean(between clusters)



Lautenberg: 200 Senate Press Releases (appropriations, economy, education, tax, veterans, . . .)

Policy Agendas: 213 quasi-sentences from Bush's State of the Union (agriculture, banking & commerce, civil rights/liberties, defense, . . .)

Reuter's: financial news (trade, earnings, copper, gold, coffee, . . .); "gold

What do Members of Congress Do?

Substantive example of a finding, using our approach

- David Mayhew's (1974) famous typology:
  1. Advertising
  2. Credit Claiming
  3. Position Taking
- We find one more: Partisan Taunting
  - "Senator Lautenberg Blasts Republicans as 'Chicken Hawks' " [Government Oversight]
  - "The scopes trial took place in 1925. Sadly, President Bush's veto today shows that we haven't progressed much since then." [Healthcare]
  - "John Kerry had enough conviction to sign up for the military during wartime, unlike the Vice President, who had a deep conviction to avoid military service" [Government Oversight]
  - ⇝ Is this what it means to be a member of a political party?

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, crowd sourcing
9. Web surfing artifacts: clicks, searches, and advertising clickthroughs. (Google collects 1 petabyte/72 minutes on human behavior!)
10. Government bureaucracies: moving from paper to electronic data bases, increasing availability
11. Governmental policies: requiring more data collection, such e.g., “No Child Left Behind Act” and allowing randomized policy experiments to proliferate
12. 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 even solving problems
- Opportunities require a change in our job descriptions, with new:
  - Large-scale, interdisciplinary research
  - Computer-assisted & quantitative: Traditional approaches infeasible
  - New statistical methods & engineering required
- And then there's you & me:
  - Change comes from replacement not conversion: legislatures, courts, marriages, academic departments, . . .
  - Will you wait to be replaced? or put in the effort to convert and learn how to use the new information to learn about the social and political worlds?

For more information:

http://GKing.Harvard.edu