"Big Data" and the Challenge of Forecasting Atrocities

Philip A. Schrodt

Parus Analytics
Charlottesville, VA 22901
schrodt735@gmail.com

Leila and Melville Straus 1960 Family Symposium
Dartmouth Dickey Center for International Understanding
6 - 7 October 2014

Examples of "Big Data" in this domain

- Good Judgement Project: in one year, team members shared links to around 32,000 unique news articles and other texts
- Political Instability Task Force: "Merged Data Set" of structural variables has 2,700 variables
  - Though typically only about 1% of these are used
- ICEWS and Phoenix event data sets generate about 3,500 events per day within the CAMEO coding ontology
  - CAMEO is focused on conflict; a more extensive ontology that included cooperative events would probably generate about 10,000 events per day
- A combination of the European Media Monitor database and Wikipedia standardized biographical data allow all name variations and a political biography of political actors to be generated in a couple seconds, replacing perhaps an hour of library work.
- Systematic models consistently predict with 80% accuracy, vs the "dart-throwing chimp" of typical expert forecasts

New environment of open source analytical tools

- R and Python
  - Example: `Goose` and `newspaper` libraries in Python will extract the headline and text from almost any html-formatted news site
- Stanford CoreNLP suite for natural language processing

Decentralized open collaboration environments

- GitHub
  - After about two decades of experience, the sociology of open-source projects is fairly well understood. It's not anarchy, and it's not just for hippies
- Near-real-time collaboration with geographically dispersed teams is now routine
- Example from recent Open Event Data Alliance work
  - ICEWS, using a centralized team and proprietary tools, required about $20M and three years to develop a parser-based coding system for near-real-time coding
  - Using open source tools, OEDA did the same thing in about 6 months and around $100,000 in resources
  - (The systems probably—but not necessarily—are comparable, though we still don't have access to sufficient ICEWS data to make a one-to-one comparison)

Cautions

"Big Data" is looking a lot like the 1980s "artificial intelligence" craze

- A lot of hype and a lot of urban legends
  - Google searches predicted a flu outbreak in one case but this did not hold up in later tests
  - Other cases didn't even work once: they were just made up

- Too many people in suits

  "At [our venture capital firm], we saw [the collapse of the "clean energy" firms] coming. The most obvious clue was satorial: clean-tech executives were running around wearing suits and ties. This was a huge red flag, because real technologists wear T-shirts and jeans. So we instituted a blanket rule: pass on any company whose founders dressed up for pitch meetings."
  Peter Theil, Zero to One

-

Not all problems are equally hard

For example, automated geocoding turns out to be really, really hard: try auto-coding “San Jose” or “Sidi Musa.”

But some problems *can* be solved

- A quarter century after the "end" of artificial intelligence, I can be in the middle of nowhere, pull a little gadget from my shirt pocket, speak (!) into it, and within a few seconds get step-by-step driving instructions to the location of a 110-foot scrap lumber dinosaur sculpture in rural Vermont
- And a machine can win at the game Jeopardy.
- These tasks actually far exceed the expectations of 1980s AI, which was still looking at chess.

The Twenty Year Rule

- As a rule of thumb, new technologies—the electric motor and the personal computer are the archetypical examples—take about two decades (probably not coincidentally, a human generation) before they are effectively incorporated into organizational operations
- We are at about that twenty-year point with at least four key technologies
  - World Wide Web
  - open source software
  - automated coding of event data
  - policy-oriented quantitative forecasting models: Political Instability Task Force

Theory is important

- Atrocities–thankfully–are a rare-events situation
  - We aren't doing product recommendations for Wal-Mart, Amazon or Netflix
  - We aren't doing election predictions, which are a regular event
- Theory provides—or at least could provide—some suggestions for what to look at based on a very long baseline of human history, provided those theories are adequately tested and not simply hindsight illusions
- Atheoretical "data mining," in contrast, seems far less promising
- Bayesian methods using informed priors based on information provided by experts should work better than frequentist methods which are based on implausible null hypotheses. But only limited work has been done with this

Thank you

Email: schrodt735@gmail.com

Open Event Data Alliance: http://openeventdata.github.io

Slides:
http://eventdata.parusanalytics.com/presentations.dir/presentations.html

Forecasting papers:
http://eventdata.parusanalytics.com/papers.html