Social Data and Models in 2030

Philip Schrodt
Parus Analytics

Workshop on Methodology Training in
Political Science and Sociology
McMaster University

11 May 2017

Use of data-driven models will increase

- Relevant data are increasingly in digital form, either directly or digital versions can be easily extracted (text, images)
- Increased computational capacity (storage and processing) create opportunities for new and more effective methods ("deep learning" is just the latest such innovation), most of these coming from the private sector
- Most humans are not very good at most data-intensive problems (Tetlock, Kahneman)
- "Computer programming" is ubiquitous and relatively easily learned in a variety of guises (e.g. user interfaces vs. analytics)

The Economist

MAY 6TH–12TH 2017

Theresa May v Brussels

Ten years on: banking after the crisis

South Korea's unfinished revolution

Biology, but without the cells

The world's most valuable resource

Data and the new rules of competition

amazon
Uber
Microsoft
Google
f
Tesla

These methods will primarily be "machine learning" (ML) rather than classical "statistics"

- ML methods are more flexible, ad hoc and easy to develop. ML community is essentially self-documenting
- Classical null-hypothesis approach is counter-intuitive (and not particularly effective) except in some limited domains
- Most widely-used statistical methods are not well adapted to large sets of heterogeneous but correlated independent variables
- Arguably, most people who forty years ago would be working with equations are now working with algorithms. They are also working for Google and Microsoft Research, not in universities (sorry…).

The very finite set of widely used ML methods

- Support vector machines
- Clustering, typically using k-means
- Random forests
  - These are a relatively recent ensemble variation on the older method of decision trees
- Neural networks
  - A very old method which is now being used with vastly greater hardware and a few new algorithmic tricks to create "deep learning"
- Logistic regression
  - yes, logistic regression
  - Which not infrequently is "embarassingly effective"

Fundamentals of social science methodology remain very important

- Basic structure of concepts, categories/typologies, theory, hypotheses, indicators, measurement, random error
- Problems and approaches to assessing causality, which are quite subtle
    - In particular, humans frequently base their behavior on expectations, weakening the strict temporal ordering of cause and effect (endogeneity)
- Design and measurement biases do not disappear when one has a lot of data: they may get worse
- More generally, social scientists have about 50 years of systematic experience with issues many "data scientists" are only now encountering

Challenge: assessing robustness of models

- "Reproducibility crisis" is a huge issue
    - Existing methods, particularly regression-based, are very brittle
    - Significance tests are used and interpreted inappropriately
    - Publication biases (and data hoarding) are rampant
    - Fraud and sloppiness are probably more common than we'd like to admit
- ML methods are potentially far worse on this because the methods are new, constantly changing, and most are so parameter-rich it is difficult to assess what is driving the model
- Policies based on incorrect assessments of data can be really expensive. Or lethal

Challenge: Visualization

- Ability to display data has far outpaced standards for doing so in a fashion that can be meaningfully interpreted
  - Far too many visualizations appear to have been produced by kittens playing with spaghetti, M&Ms and food coloring
- There is a surprisingly large gap between the expectations of most analysts on data visualization and the degree to which those displays reach their intended audience
  - Visualizations having no impact at all are *probably* a greater issue than visualizations being misleading
- Cartesian coordinate displays (in various forms) are alien to most people

Challenge: Role of mathematics

- Mathematical notation is primarily now a method of communication (and not infrequently, simple intimidation), not an analytical means of deriving results: computers handle that now
- Nonetheless, the ability to "idiomatically" read mathematical notation remains very useful
  - Algebra and linear algebra
  - Common functions such as polynomials, exponentials and logs
- Classical statistical distributions and related concepts such as long-tails and bi-modality remain useful because they arise naturally in data

Thank you

Email: schrodt735@gmail.com

Blog: asecondmouse.org

Github: philip-schrodt

Web site: philipschrodt.org