Hazarding A Guess: The Dangers of Mining Big Data

OECD TECHNOLOGY FORESIGHT FORUM: “EXPLOITING THE DATA-DRIVEN ECONOMY”
OCTOBER 22, 2012

SOLON BAROCAS
FELLOW, INTERNET SOCIETY
DOCTORAL CANDIDATE, MEDIA, CULTURE, AND COMMUNICATION, NEW YORK UNIVERSITY
Data mining involves more than the rationalization of an existing, more ad hoc procedure; instead, it tends to change the activity to which it is applied.

- e.g., the profound change policing experienced with the introduction of profiling: from a general concern with crime reduction (efficacy) to a narrower concern with the likelihood that each police action would result in crime detection (efficiency) (Harcourt 2007)

Which is to say that data mining simultaneously allows an organization to pursue certain goals more effectively while changing those goals in the process.
Errors and Bias

- Matching errors
  - e.g., Ted Kennedy

- Type I and type II errors
  - Trade-offs between false positives and false negatives
  - Assumptions concerning the distribution of variance
  - Benefits outweigh the costs
    - Proportionality?

- Nonuniversal generalizations and non-distributive group profiles

- Concept and population Drift
Costs of Non-Erroneous and Lawful Discrimination

- Discriminatory by design
  - Non-erroneous discrimination?
  - Lawful discrimination?
    - Absence of animus

- Fairness
  - extend different options and opportunities to individuals and groups according to their estimated value
    - produce unequal access to information, goods, and services

- Inequality
  - Constrain life chances
  - Exacerbate historical inequalities

- Ratchet effect (Harcourt 2007)
  - Stratification
  - Stigmatization
Constrained Worldview and Behavior

- **Soft cage of customization**
  - The primacy of historical correlations

- **Self-reinforcing feedback loop**
  - Paradoxically, individuals may lose control of their own preferences by relying on a system that attempts to cater to them
Frustrating the Fair Information Practice Principles

- ‘Personal data’ as the trigger for protection
- Latent facts
  - Ability to infer sensitive details from seemingly innocuous information
  - The challenges of providing notice
    - “you must first discover what you would want to hide” (Hildebrandt 2009)
- The fundamental incompatibility of purpose specification and use limitation
- Trade-off between transparency and accuracy
- The uselessness of anonymity
Substantive Regulations

- Supplement privacy with other normative concerns
  - Autonomy
  - Non-discrimianion
  - Fairness
- But even these might be insufficient as the distinction between equality of opportunity and equality of outcome begins to blur
  - From procedural fairness to distributive justice?