Privacy Considerations at Each Stage of the Big Data Lifecycle

David Weinkauf, Ph.D.
Senior Policy and Technology Advisor
Office of the Information and Privacy Commissioner of Ontario

Access, Privacy and Records and Information Management (RIM) Symposium

October 17, 2016

Information and Privacy Commissioner of Ontario
Commissaire à l'information et à la protection de la vie privée de l'Ontario

Outline

- What is big data?
- Privacy concerns at each stage of the big data lifecycle
  - Collection
  - Integration
  - Data mining / analytics
  - Use
- Next steps

What Is Big Data?

- The term “big data” generally refers to the convergence and combined use of a number of advancements in computing and technology, including
  - *New sources and methods of data collection*
  - *Virtually unlimited capacity to store data*
  - *Improved record linkage techniques*
  - *Algorithms that learn from and make predictions on data*
- The net effect is a newfound ability to analyze large, complex data sets; uncover hidden patterns and correlations in them; and use these patterns to derive rules that allow for automated decision-making and the prediction of future results

Big Data Lifecycle

- The big data lifecycle can be divided into four stages
  1. Collection
  2. Integration
  3. Data mining / analytics
  4. Use
- Each stage raises a number of privacy concerns / issues
- Issues need to be addressed to prevent unexpected, invasive and discriminatory uses of personal information (PI)

Stage 1: Collection

- **Issues:** indirect collection and secondary purposes
  - Data sharing runs counter to two fundamental privacy principles:
    - PI should be collected directly from the individual to whom it pertains
    - PI should only be used for the purpose for which it was collected (with limited exceptions)
- **Issue**: speculation of need rather than necessity
  - What is now known as “data mining” was originally called “data fishing”
  - Analyze data first and ask “why” later

Stage 2: Integration

- **Issue**: false positives from probabilistic or "fuzzy" linkages
  - Section 40 (2) of *FIPPA* states that "The head of institution shall take reasonable steps to ensure that personal information on the records of the institution is not used unless it is accurate and up to date."
- **Issue**: inadequate separation of functions
  - From the U.S. Privacy Protection Study Commission's 1977 *Personal Privacy in an Information Society* report: "Even where organizational separation exists […] individually identifiable information and records used for research or statistical purposes can be commingled with information and records used for administrative purposes."

Stage 3: Data Mining / Analytics

- **Issue**: poor quality data
  - Information may be lacking, incorrect or outdated
- **Issue**: algorithmic biases
  - Feedback-loop problem
  - Variables may act as proxies for discrimination
  - Underrepresentation or overrepresentation of certain populations
- **Issue**: spurious correlations
  - Variables may occur together without a causal relation
  - Large enough data sets tend to have meaningless correlations

Stage 4: Use

- **Issue**: generation of new personal information
  - Exposure of sensitive information—e.g., Target's "pregnancy prediction algorithm"
- **Issue**: non-transparent logic of algorithms
  - Complex and opaque
  - Confidential and proprietary
- **Issue**: lack of human intervention / individual recourse
  - Adverse actions of automated decision-making
  - Data fundamentalism

Next Steps

- The challenge is to ensure adequate measures protect the privacy of individuals while enabling big data initiatives
- Such measures could include
  - Legislative authority to integrate data sets containing PI
  - Independent review / approval of projects
  - Transparency of approved projects
  - Secure process for linking
  - De-identification
  - Verification of accuracy and non-bias of results
  - Allow affected individuals to challenge or respond to automated decisions

Questions?