The IPC’s Big Data Guidelines
Privacy, Fairness and Ethics

David Goodis
Assistant Commissioner, IPC

David Weinkauf, Ph.D.
Senior Policy and Technology Advisor, IPC

Ontario Connections Conference
June 7, 2017
Outline

• Big data and Ontario’s privacy laws (David Goodis)

• IPC’s “Big Data Guidelines” (David Weinkauf)

• Questions
Big Data and Ontario’s Privacy Laws

• *FIPPA/MFIPPA* not designed with big data in mind; not possible when proclaimed in 1988/1991:
  - world wide web not yet invented (1989)
  - information technology was less prevalent
  - types of data and analytics were less complex
  - uses of personal information were discrete and determinate

• Current legislative framework treats government institutions as silos:
  - collection of personal information must be “necessary”
  - secondary uses are restricted
  - information sharing is limited
Big Data and Ontario’s Privacy Laws (Cont’d)

• May still be possible to conduct big data under FIPPA if:
  o collection of personal information (PI) is **expressly authorized by statute** [s. 38(2)]
  o disclosures are for purpose of **complying with a statute** [s. 42(1)(e)]

• Such cases should be the exception, not the rule
• To support big data in general, we need a **new legislative framework**
Ontario IPC’s Big Data Guidelines

• Designed to inform institutions of key issues, best practices when conducting big data projects involving PI

• Divides big data into four stages; each stage raises a number of concerns (14 total)

• Institutions should avoid uses of PI that may be unexpected, invasive, inaccurate, discriminatory or disrespectful of individuals

• Today we will discuss a selection of points raised in paper
What Is Big Data?

- The term “big data” generally refers to the 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
Collection

• Issue: *speculation of need rather than necessity*
  o inherent tension between big data and principle of data minimization
  o what is now known as “data mining” was originally called “data fishing”
  o analyze data first and ask “why” later

• Best practice (BP): proposed collection of PI should be **reviewed and approved** by a research ethics board (REB) or similar body
Issue: privacy of publicly available information

- potential uses and insights derivable from a piece of information are no longer discrete and recognizable in advance
- innocuous PI can be collected, integrated and analyzed with other PI to reveal hidden patterns and correlations that only an advanced algorithm can uncover

BP: any publicly available PI should be treated the same as non-public PI
Integration

• **Issue:** *inadequate separation of policy analysis and administrative functions*
  
  o PI collected for the purpose of administering a program can be used for secondary purpose of fulfilling the policy analysis function of the program
  
  o however, in general the reverse is not the case

• **BP:** integrated data sets should be **de-identified** before analysis to ensure adequate separation

• De-identification also helps to address the inherent tension between big data and principle of data minimization
• **Issue: biased data sets**
  - even if “all” data is collected, the practices that generate the data may contain implicit biases that over- or underrepresent certain people
  - also, the conditions under which a data set is generated may cause some members of the target population to be excluded

• **BP: assess whether the information analyzed is representative of the target population by considering whether:**
  - the practices that generated the data set allowed for discretionary decisions
  - the design of a program or service contained overly restrictive requirements
Analysis (Cont’d)

• Issue: discriminatory proxies
  – Charter guarantees every individual a right to “equal protection and benefit of the law without discrimination”
  – variables in a data set that are not explicitly protected may correlate with protected attribute

• BP: ensure analysis of integrated data set does not result in any variables being used as proxies for prohibited discrimination

• Outcome of analysis may need to be reviewed by REB or similar body to determine its potential for such discrimination
Analysis (Cont’d)

• Issue: spurious correlations
  o with so many combinations of variables at play, there are likely to be some that appear to be meaningful without actually being so
  o however, correlation does not imply causation
  o two variables may relate by chance or to a third variable
• BP: ensure any patterns discovered in the analysis are meaningful
• You may need to verify results of the analysis in a manner that is independent of the procedure used
Profiling

• Issue: lack of transparency
  o profiling not only processes PI but generates it as well
  o evaluation or prediction of PI happens in the background
  o individuals may not understand the consequences

• BP: individuals should be informed of the nature of the predictive model or profile being used, including:
  o the use of profiling and the fields of PI generated by it
  o a plain-language description of the logic employed by the model
  o the implications or potential consequences of the profiling on individuals
Profiling (Cont’d)

• Issue: individuals as objects
  o profiling takes reductive approach to understanding where individuals only amount to the sum of their parts
  o even if accurate, individuals may feel a loss of dignity from being subjected to profiling
  o extension of profiling to too many aspects of society or individuals’ lives would have serious consequences, such as loss of autonomy, serendipity and exposure to a variety of perspectives

• BP: the public and civil society organizations should be consulted regarding the appropriateness and impact of proposed profiling