

ONTARIO GOVERNMENT USE OF BIG DATA ANALYTICS

David Goodis

Assistant Commissioner, Ontario IPC

David Weinkauff, Ph.D.

Senior Policy and Technology Advisor, Ontario IPC



OUTLINE

- Big data and Ontario's privacy laws (David Goodis)
- Ontario IPC's "Big Data Guidelines" (David Weinkauff)
- Comments from a government perspective (John Roberts)
- 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 (2)

- May still be possible to conduct big data under *FIPPA* if:
 - collection of personal information (PI) is **expressly authorized by statute** [s. 38(2)]
 - 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
 - inherent tension between big data and principle of data minimization
 - what is now known as “data mining” was originally called “data fishing”
 - 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



COLLECTION (2)

- 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
 - PI collected for the purpose of administering a program can be used for secondary purpose of fulfilling the policy analysis function of the program
 - 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



ANALYSIS

- 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 (2)

- 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



PROFILING

- Issue: **lack of transparency**
 - profiling not only processes PI but generates it as well
 - evaluation or prediction of PI happens in the background
 - individuals may not understand the consequences
- BP: individuals should be **informed of the nature of the predictive model** or profile being used, including:
 - the use of profiling and the fields of PI generated by it
 - a plain-language description of the logic employed by the model
 - the implications or potential consequences of the profiling on individuals



PROFILING (2)

- Issue: **individuals as objects**
 - profiling takes reductive approach to understanding where individuals only amount to the sum of their parts
 - even if accurate, individuals may feel a loss of dignity from being subjected to profiling
 - 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**

