The perils of “big data”

Rolfe Bozier

17-Sep-2014

Agenda

- What do we mean by “big data”?
- What do people do with these datasets?
- What could go wrong?
- The future

What do I mean by big data?

- Personal information – things not meant to be released
  - Medical records
  - Smart electricity meters
  - Internet browsing history
  - Business transaction records
  - Bank details
  - Tax records
  - Criminal records
- Third party – information we relinquish
  - Browser cookies
  - Search results
  - Internet commerce transactions
- Behavioural – fallout from our activities
  - Social media: Linked-in, Facebook, Twitter
  - Web forum activity
  - Flickr, Picasa, blogs, etc.
- Public – information we need to provide
  - Census
  - Electoral roll
  - Telephone number
  - Company records

What do I mean by big data?

- Characterizing datasets – some dimensions
  - Availability
    - Who can access the data?
    - How secure is it?
    - E.g. private, 3rd party, public
  - Sensitivity
    - How important is the content to you?
    - "Everyone has a `database of ruin'"
  - Identification
    - How closely is it tied to **you**?
    - E.g. identified, de-identified, aggregated
- A lot of the risks with big datasets involve altering their position along one of these dimensions

Some relevant datasets

- Some datasets you are almost certainly included in:
  - Facebook, Linked-in
  - Web forums, mailing lists, chat sites etc.
  - Internet purchases (Amazon, eBay, ...)
  - Search engine activities
  - Web site analytics
  - Loyalty cards (Everyday Rewards, FlyBuys, MYER one, …)
  - Toll roads, Opal card, air travel records
  - Census, electoral roll
  - Every time you put your email address on a web form
  - Tax records, bank transactions
  - Medical records

Data analytics – what can you do with data?

- Brief answer: a lot more than people think!
  - Machine learning / data analytics / pattern analysis is very big in this area
- Derive personal attributes from your behaviour
  - Targeted ads relevant to your recent searches
  - Track your activity from searching for something through to an online transaction
  - Amazon suggestions based on previous purchases and other people’s activity
  - Determine that you [your daughter] is pregnant [http://goo.gl/JrCt3P]
  - Develop a personal profile based on your shopping habits
  - Develop a personal profile based on your likes/dislikes
- Collect and on-sell your data
  - US FTC looked at 12 mobile fitness apps and found they were sharing data with 76 third parties [http://goo.gl/o0gBfr]
  - Why do you think that mobile app you just installed was “free”?
    - What permissions did you grant it?
- Dataset linking
  - Combine two datasets using a common identifier
  - Combine two datasets by identifying correlations between them
  - Two personal examples:
    - NRMA road service / NRMA MotorServe
    - Vodafone billing

Data analytics – what can you do with data?

- Re-identification from anonymized data
  - How many independent bits of data to identify a person?
  - In Australia, it is probably around 25 bits
  - Could be supplied by: data-of-birth, postcode, gender
- Data scientists are starting to think that it just isn't possible to really anonymize datasets
  - "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization", Paul Ohm
  - "No silver bullet: De-identification still doesn't work", Arvind Narayanan
- NYC released data on 173 million taxi trips as part of an FOI request [http://goo.gl/7Qp9N3]
  - Licence plate and driver ID information was anonymized by hashing with MD5
  - But these are predictable, so just hash all possible values to de-anonymize the data
- Massachusetts Group Insurance Commission released hospital visit on all state employees for research purposes [http://goo.gl/41spHl]
  - User information was removed, but DOB, zip-code and gender were retained
  - A postgrad obtained this data from the voter rolls and got the medical records for the State Governor
- Netflix released a huge database of movie recommendations for a competition to predict when movie viewers would like to watch [http://goo.gl/CpCyZg]
  - 2 computer scientists correlated records with users in the Internet Movie Database IMDB
- AOL released a large dataset of search queries [http://goo.gl/MEyOaT]
  - But searches will occasionally be quite personal and distinctive
  - A news crew were able to track down and interview one person in the dataset based on her query strings

What could go wrong?

- Undesired outcomes
  - Receive targeted advertising (online or physical)
  - Public release of private details
    - Example: iCloud nude photos
- Commercial disadvantage
  - Companies can engage in dynamic pricing – pay more because
    - You live in a wealthy neighborhood
    - You are using a Mac
    - Your profile indicates that you make discretionary purchases
  - Companies engage in risk management
    - Your premiums went up because of adverse behaviour
    - Your credit rate is higher because of perceived risk
    - Example: a man in the US had his credit limit reduced because he made a purchase in a “risky location”
  - Job application rejection
- Target of criminal activity
  - identity theft
  - robbery, fraud,...
- Some outcomes are difficult to avoid
  - Government surveillance [“metadata collection”]
  - Legal data acquisition [Subpoenas of your phone records etc.]

How does it affect me?

- What can I do – as a consumer?
  - Just accept it / look at the bright side
  - “if you aren’t paying for the product, you are the product”
  - Stop voluntarily giving away information
    - Or at least think about it
    - What’s wrong with that free mobile app…?
  - Some concrete steps:
    - Don’t just install that mobile app
    - Think hard about social media
    - Stop using the loyalty card
    - Disable third party trackers in your browser
- What can I do – as a developer?
  - Be aware of the ramifications of having people use your software
  - What personal data will they provide?
    - It is protected?
    - What about inadvertent data collection?
  - There are different barriers to data collection:
    - Researchers typically have to write a proposal for an ethics committee before they collect data
    - Software developers just go right ahead
  - Data collection and protection is difficult
  - Data scrubbing is really hard to do right

The future

- Risks look likely to increase in the future
  - More data sharing / selling
  - More social applications
  - Internet of Things / smart meters
  - More storage of data off-site (in the cloud)
- Technology improvements
  - Facial / gait recognition
  - Video object detection
  - Photo content recognition
  - Licence plate reading
  - Machine learning
  - Geotagged data
- Data is permanent

No longer true…

"On the Internet, nobody knows you're a dog."