Big Data
Big Privacy

Privacy Awareness Week
29 April 2013
SPEAKING NOTES

Stephen Wilson
Lockstep Group

Setting the scene

- Practical experience shows a gap in the understanding that "technologists" as a class have regarding data privacy.
- A gap between technology and the law is perpetuated to some extent by the popular impression that the law has not kept up with the march of technology. As a technologist, I am more optimistic; I actually find that principles-based data privacy law anticipates almost all of the current controversies in cyberspace (though not quite all).
- IT professionals hear the well-meaning slogan "Privacy Is Not A Technology Issue" and they say 'thank god – that's one thing I don't need to worry about'. Conversely privacy laws are written with some naivety about how information flows in modern IT.

Technicalities

- People tend to think intuitively that Personal Information is the stuff of forms and questionnaires and call centres. Technologists can be surprised that the definition of Personal Information covers a great deal more.
- If metadata or event logs in an IT system are personally identifiable, then they constitute Personal Information regardless of whether they are untouched by human hands.

Personal Information

Information or an opinion, whether true or not, about an individual whose identity is apparent, or can reasonably be ascertained, from the information or opinion

Privacy Act 1988

Technicalities cont.

- Our privacy legislation is technology neutral with regards the manner of collection. Indeed, the term "collection" is not specially defined in the law. So if Personal Information has wound up in an information system, it doesn't matter if it was gathered directly from the subject, or has instead been imported or found in the public domain or generated almost from scratch by some algorithm: it has been *collected* and as such is covered by the Privacy Act.

Collection

An organisation must not collect PI unless the information is necessary for one or more of its functions or activities

National Privacy Principle NPP 1

Case: Google StreetView Wi-Fi collectio

- Google StreetView cars collect Wi-Fi hub locations, which on their own are unidentified. It was found that the StreetView software was also inadvertently collecting Wi-Fi network traffic, some of which contained Personal Information. Australian and Dutch Privacy Commissioners found Google was in breach.
- Many technologists argued that Wi-Fi data in the "public domain" is not private, and that Google was within its rights to do whatever it liked with it. But the argument fails to grasp the technicality that our privacy laws basically do not distinguish public from "private". In fact the words "public" and "private" are not operable in the Privacy Act (indeed there is a view that the Privacy Act is really a data protection law).
- Lesson for Big Data privacy: it doesn't much matter if data is sourced from the public domain: you are still subject to *Collection* and *Use Limitation* principles.

Case: Facebook facial recognition

- Facebook photo tagging creates biometric templates used to subsequently generate tag suggestions. Before displaying suggestions, Facebook's facial recognition algorithms run in the background over all photo albums. When they make a putative match and record a deduced name against a hitherto anonymous piece of image data, they have collected Personal Information.
- European privacy regulators found biometric data collection without consent to be a serious breach, and forced Facebook to shut down facial recognition and tag suggestions in the EU.
- Lesson for Big Data privacy: it doesn't much matter if you *generate* Personal Information using sophisticated algorithms: you are still subject to *Collection* and *Use Limitation* principles.

Case: Target pregnancy predictor

- Target in the US was found to be experimenting with statistical methods for identifying that a regular customer is likely to be pregnant, by looking for trends in her buying habits.
- In Australia, the privacy implications would be amplified by the fact that tagging someone in a database as pregnant [even if that prediction is wrong!] relates to health and therefore represents a collection of *Sensitive Information.* The Privacy Act requires express consent in advance of collecting Sensitive Information.
- Stores may need to disclose to their customers up front that Big Data processes can produce health information by mining their buying habits. Consent may be required.
- Note a present problem in Australia for grocery stores that sell medicinals online. St Johns Wort for example seems innocuous but it indicates that a customer has (or believes they have) depression. IT security managers might not have thought about the implications of logging mental health information in regular web servers and databases.

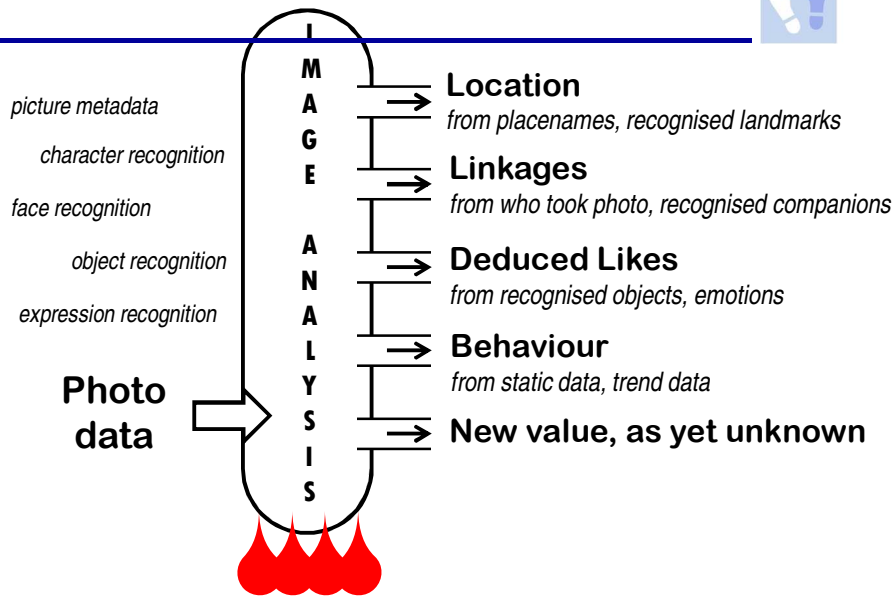
Case: "DNA Hacking"

- In February 2013 research was published where a subset of anonymous donors to a DNA research program were identified by cross-matching genes to data in public genealogy databases.
- All of a sudden, the ethics of re-identifying genetic material has become a hot topic. A lot of attention is focusing on the nature of the informed consent; different initiatives (like the *Personal Genome Project* and *1,000 Genomes*) give different levels of comfort about the possibility of re-identification. Absolute anonymity is typically disclaimed but re-identification is said to be difficult.
- But a nice legal problem is that regardless of the consent given by a Subject (1st party) to a researcher (2nd party), when a third party takes anonymous data and re-identifies it without consent, they have collected Personal Information, as per the principles discussed above. Lockstep believes, following the European facial recognition precedent, that re-identification of DNA without consent is likely to be ruled problematic (if not unlawful) in some jurisdictions, and is therefore unethical in *all* jurisdictions.

Big Data's privacy challenge

- Principles-based data protection laws have proven to be relevant and powerful in the cases of Google's StreetView Wi-Fi collection and Facebook's facial recognition; and they seem to govern data mining for health information, and DNA re-identification. But there is one area where the principles may struggle to cope with Big Data.
- Orthodox privacy management involves telling individuals What information you collect about them, Why you need it, When you collect it, and How. But with Big Data, even if a company wanted to be completely transparent, it may not know what Personal Information lies waiting to be mined and discovered in the data, nor when exactly this might be done.
- An underlying theme in Big Data business models is data mining, or perhaps more accurately, *data refining*, as shown in the next page. An increasing array of data processing techniques are applied to vast stores of raw information (like image data in the example) to extract metadata and increasingly valuable knowledge.

Data as crude oil



Big Data's challenge cont.

- There is nothing intrinsically wrong with a business model that extracts value from raw information, even if it converts anonymous data into Personal Information. But the privacy promise enshrined in OECD data protection laws – to be open with individuals about what you know about them and why – can be hard to honour.
- There is a bargain at the heart of most social media companies today, in which Personal Information is traded for a rich array of free services. The bargain is opaque; the "infomopolies" are coy about the value they attach to the Personal Information of their members.
- If OSNs were more open about their business models, it seems likely that most of members would still be happy with the bargain. After all, Google, Facebook, Twitter et al have become indispensable for many of us. They do deliver fantastic value. But the Personal Information trade needs to be transparent.

"Big Privacy"!

- **Exercise constraint**
  Remember privacy is essentially about restraint. If a business knows me, then privacy means they are restrained in how they use that knowledge.
- **Meta transparency**
  We're at the very start of Big Data. Who knows what lies ahead. *Meta transparency* means not only being open about what Personal Information is collected and why, but also being open about the business model and the emerging tools.
- **Engage customers in a fair value deal**
  Most savvy digital citizens appreciate there is no such thing as a free lunch; they already know at some level that "free" digital services are paid for by trading Personal Information. Some individuals manage their own privacy in an ad hoc way by deliberately obfuscating or manipulating the details they divulge. Ultimately consumers and businesses alike will do better by engaging in a real deal that sets out how PI is truly valued and leveraged.
- **Dynamic consent models**
  The most important area for law and policy to catch up with technology seems to be in consent. As businesses discover new ways to refine raw data to generate value, individuals need to be offered better visibility of what's going on, and new ways to opt out and opt back in again depending on how they gauge the returns on offer.

Further reading

- *Re-identification of DNA may need ethics approval*
  http://lockstep.com.au/blog/2013/02/08/dna-privacy-letter-to-science
- *It's not too late for privacy*
  http://lockstep.com.au/blog/2012/10/29/not-too-late-for-privacy
- *Photo data as crude oil*
  http://lockstep.com.au/blog/2012/04/11/photos-as-crude-oil
- *What stops Target telling you're pregnant?*
  http://lockstep.com.au/blog/2012/03/07/target-tells-youre-pregnant

swilson@lockstep.com.au
http://lockstep.com.au