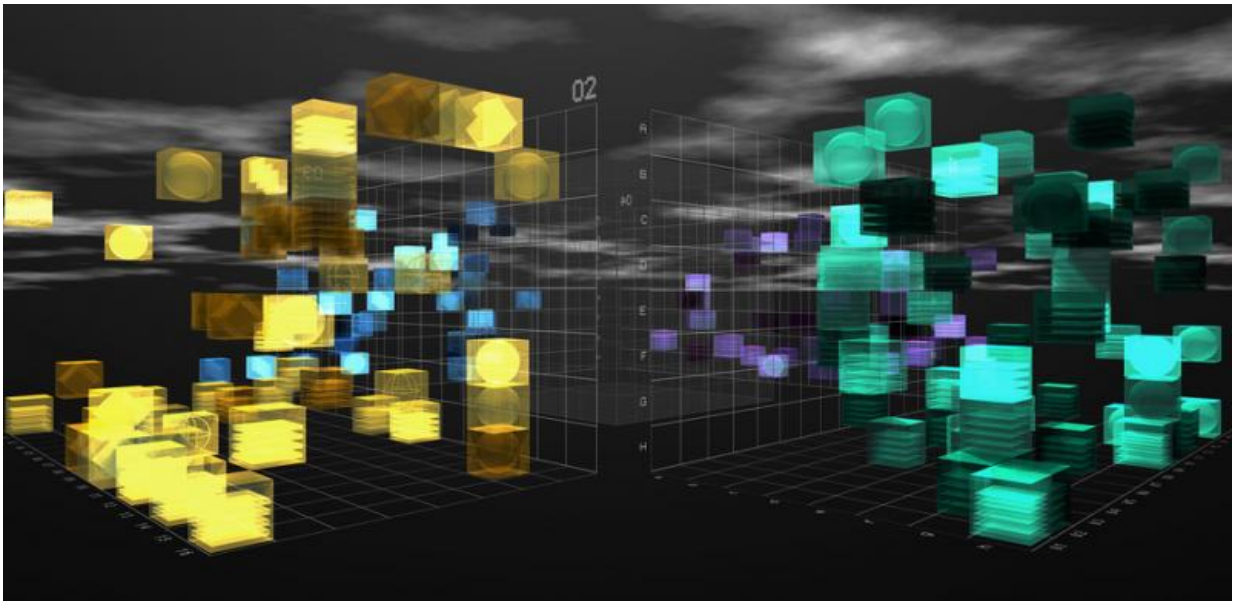
Big Data is useful, but we need to protect your privacy too

May 8 2015, by Christine O'keefe



Protecting your privacy when your data collected for one use might have a secondary use for other researchers. Credit: Flickr/, CC BY-NC-ND

These days, massive volumes of data about us are collected from censuses and surveys, computers and mobile devices, as well as scanning machines and sensors of many kinds. But this data can also reveal personal and sensitive information about us, raising some serious privacy concerns.

Data are routinely collected when we shop, use public transport, visit our

GP or access government services in person or online. There's also data from using our smart phones and fitness monitoring devices.

These data are generally collected for a purpose, called the "primary purpose". For example, having purchased goods delivered, catching a bus from home to work, having a health check, obtaining a Medicare refund, navigating or searching our local area, as well as logging our fitness regime.

But in addition to being used for such primary purposes, many data are stored and used for other purposes, called "secondary purposes". This includes research to help inform decision-making and debate within government and the community.

For example, data from Medicare, the Pharmaceutical Benefits Scheme and hospitals can be used to identify potential adverse drug reactions much faster than is currently possible.

What about privacy?

But these data can also reveal highly sensitive information about us, such as about our preferences, behaviours, friends and whether we have a disease or not.

Given the rapid change in the volume and nature of data in the digital age, it is timely to ask whether the existing ethics frameworks for the secondary use of such data are still adequate. Do they address the right ethical issues associated with research using the data? In particular, how will an individual's privacy be protected?

There have been two important responses to these issues. A group of researchers, supported by the University of Melbourne and the Carlton Connect Initiative, explored these issues through workshops, desk

research and many consultations.

They produced the Guidelines for the Ethical Use of Digital Data in Human Research. It's a work in progress, requiring ongoing practice and revision, rather than a definitive set of prescriptions.

A team at CSIRO and the Sax Institute also addressed the deeper ethical issue of protecting privacy in the secondary use of health data. This work will be developed into Guidelines for Confidentiality Protection in Public Health Research Results.

Ethical issues for digital data

In the first of the guidelines, five key categories of ethical issues are identified as highly relevant to digital data and require additional consideration when using digital data.

1. Consent: making sure that participants can make informed decisions about their participation in the research
2. Privacy and confidentiality: privacy is the control that individuals have over who can access their personal information. Confidentiality is the principle that only authorised persons should have access to information
3. Ownership and authorship: who has responsibility for the data, and at what point does the individual give up their right to control their personal data?
4. Data sharing – assessing the social benefits of research: data matching and re-use of data from one source or research project in another
5. Governance and custodianship: oversight and implementation of the management, organisation, access and preservation of digital data.

The voluntary guidelines were developed to help people conducting research and to assist ethics committees to assess research involving digital data.

Without such guidelines, there is a risk that new ethical issues involving digital data will not adequately be considered and managed by researchers and ethics committees.

Privacy risks from the data

Traditionally, the data custodians responsible for granting access to data sets have sought to protect people's confidentiality by only providing access to approved researchers. They also restricted the detail of the data released, such as replacing age or date of birth by month or year of birth.

More recently, data custodians are increasingly being asked for highly flexible access to more and more details about individual persons from an expanded range of data collections.

Custodians are responding by developing a new flexible range of access modes or mechanisms, including remote analysis systems and virtual data centres.

Under remote analysis, a researcher does not have access to any of the data but submits queries and receives analysis results through a secure webpage.

A virtual data centre is less restrictive than a remote analysis system. It enables researchers to interact directly with data, submit queries and receive results through a secure interface.

But the results of statistical analysis as released by a virtual data centre may still reveal personal information. For example, if a result such as an

average is computed on a very small number of people then it is probably very close to the value for each of those people.

By following such voluntary guidelines, researchers can maintain confidentiality while ensuring that society can benefit from their work.

The rapid technological advances in our society are creating more and more data archives of many different types. It's vital that we continue to assess the ethical and privacy risks from secondary use of this data if researchers are to reap the potential benefits from access to the information.

This story is published courtesy of The Conversation *(under Creative Commons-Attribution/No derivatives).*

Source: The Conversation