# DATA7001
# INTRODUCTION TO DATA SCIENCE

## Module 2 Getting the Data I Need

# Module Topics

- Types of Data
- Data Ingestion
- **Managing Data Privacy**
- Sampling Big Data

Patient data is private → Patient data has insights for scientific research on biomedicine, drug innovation, public health …

CCTV cameras (can) record private conversations from general public → Audio data can contain critical information on criminal, and terrorist activity

If you had access to data on people’s movements, behaviors and social habits … what would you do with it?

# Ethical use of big data

**Some considerations..**

Fair benefit Sharing

Reciprocity

Privacy

Potential Discrimination

Cultural diversity

Confidentiality

Commercialization

Equity

Consent

Conflict of Interest

Ownership

Intellectual Property rights

# Legal use of big data

## Information Privacy Law

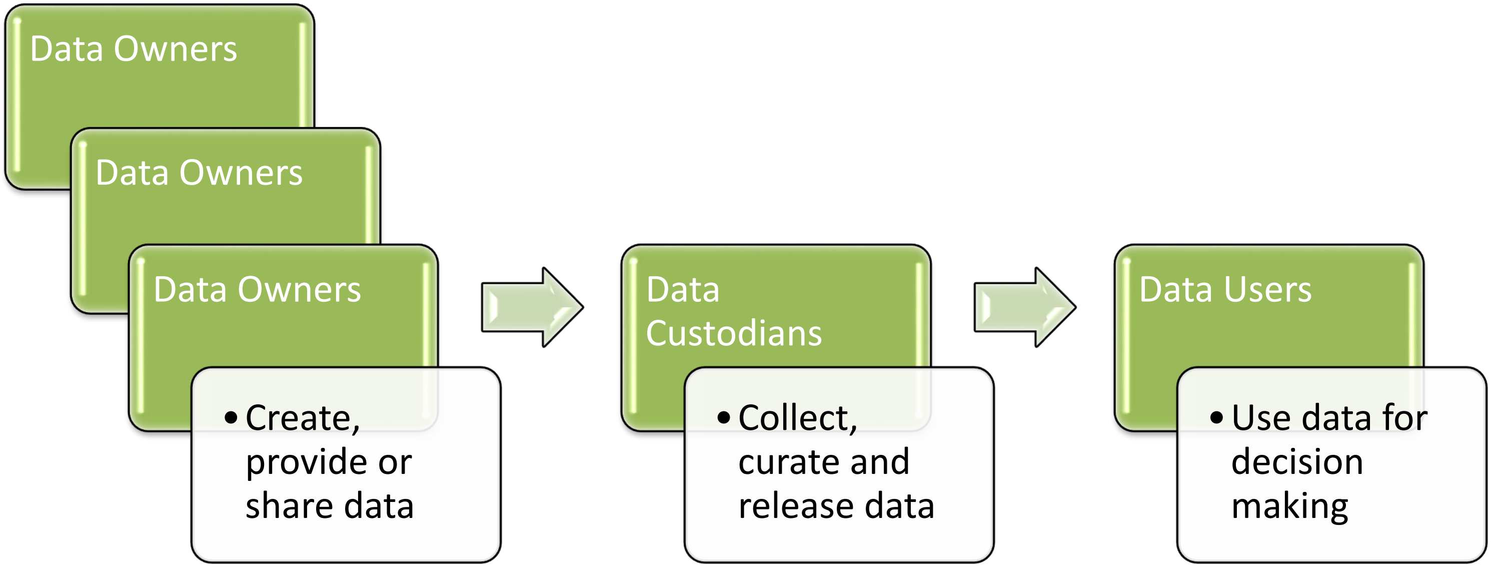
- Information privacy or data protection law is based on control concepts of privacy
  - Alan Westin, *Privacy and Freedom* (1967)
  - *Privacy is the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others*
- Information privacy law is therefore about….the mechanics of personal information exchange
  1. Individuals have limited rights of control over collected personal information.
  2. Collecting organisations have legal obligations on how personal information is collected, stored and used.
- The law tries to balance individual rights and organisational requirements by providing
  - Fairness protections for individuals in the form of information privacy principles
  - Collecting organisations with flexibility to use and exchange personal information

## Questions for Data Scientists

- Does information privacy law apply to my de-identified data sets?
- Do I need to tell individuals about how I'm using their personal information?
- Do I have to keep my data sets up to date and secure?
- I can use my data sets for any purpose, right?

Acknowledgments: Mark Burdon (UQ Law School) and Andrew Crowden (UQ Philosophy)

# Privacy preserving data release



Data Owners

Data Owners

Data Owners

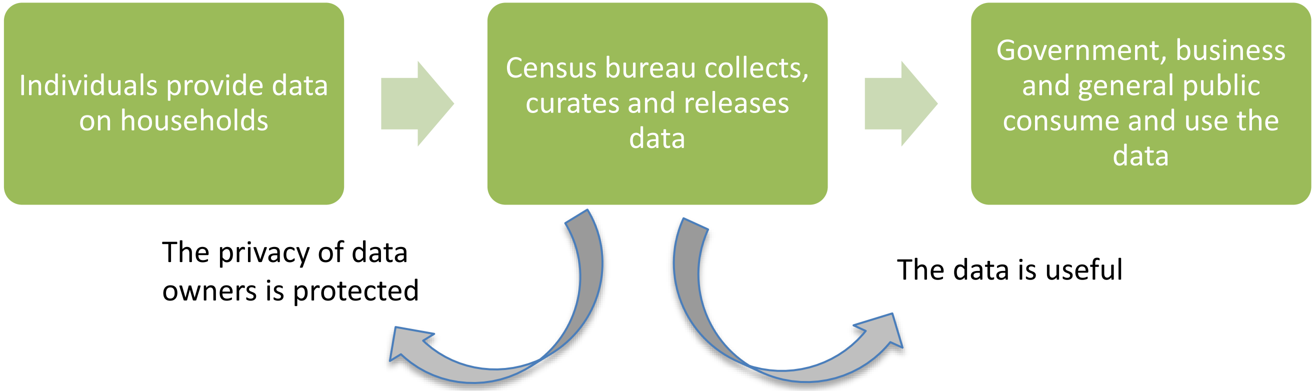
- Create, provide or share data

Data Custodians

- Collect, curate and release data

Data Users

- Use data for decision making



Individuals provide data on households

Census bureau collects, curates and releases data

Government, business and general public consume and use the data

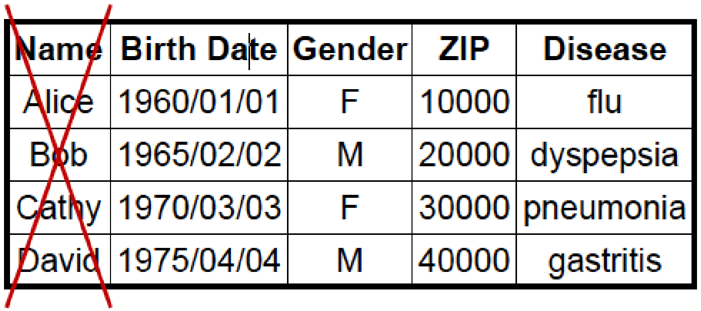
The privacy of data owners is protected

The data is useful

Acknowledgements Xiaokui Xiao, Nanyan Technological University

# Anonymisation Failure

- Massachusetts Group Insurance Commission
  - Medical records of state employees



| ~~Name~~ | Birth Date | Gender | ZIP | Disease |
|---|---|---|---|---|
| ~~Alice~~ | 1960/01/01 | F | 10000 | flu |
| ~~Bob~~ | 1965/02/02 | M | 20000 | dyspepsia |
| ~~Cathy~~ | 1970/03/03 | F | 30000 | pneumonia |
| ~~David~~ | 1975/04/04 | M | 40000 | gastritis |

Medical Records

At the time MGIC released the data, William Weld, then Governor of Massachusetts, assured the public that GIC had protected patient privacy by deleting identifiers. In response, then-graduate student **Latanya Sweeney** started hunting for the Governor's hospital records in the GIC data. She knew that Governor Weld resided in Cambridge, Massachusetts, a city of 54,000 residents and seven ZIP codes.

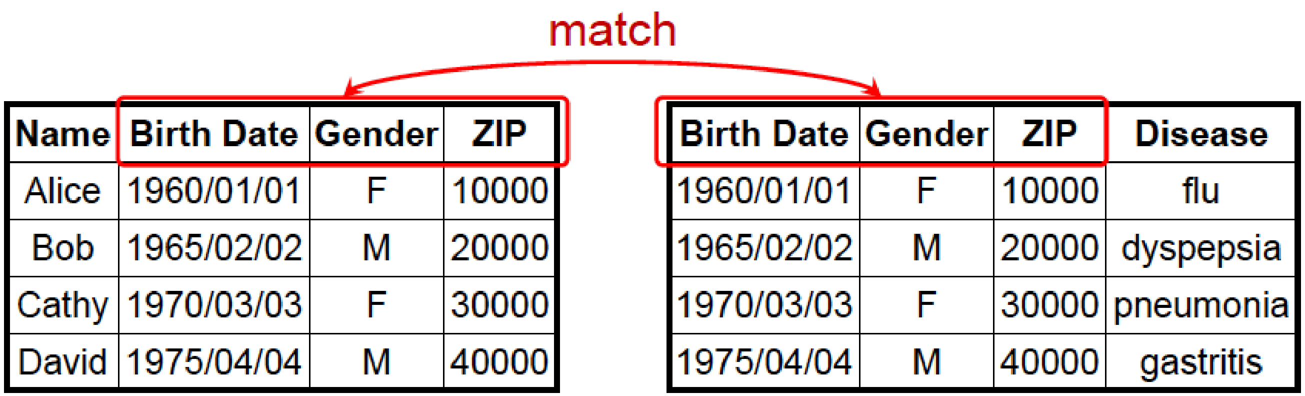
For twenty dollars, she purchased the complete voter rolls from the city of Cambridge, a database containing, among other things, the name, address, ZIP code, birth date, and sex of every voter. By combining this data with the GIC records, Sweeney found Governor Weld with ease. Only six people in Cambridge shared his birth date, only three of them men, and of them, only he lived in his ZIP code. In a theatrical flourish, Dr. Sweeney sent the Governor's health records (which included diagnoses and prescriptions) to his office.

In 2000, Sweeney showed that 87 percent of all Americans could be uniquely identified using only three bits of information: ZIP code, birthdate, and gender.

https://arstechnica.com/tech-policy/2009/09/your-secrets-live-online-in-databases-of-ruin/

# Annonymisation Failure

- Massachusetts Group Insurance Commission
  - – Medical records of state employees



| Name | Birth Date | Gender | ZIP |
|---|---|---|---|
| Alice | 1960/01/01 | F | 10000 |
| Bob | 1965/02/02 | M | 20000 |
| Cathy | 1970/03/03 | F | 30000 |
| David | 1975/04/04 | M | 40000 |

Voter Registration List

| Birth Date | Gender | ZIP | Disease |
|---|---|---|---|
| 1960/01/01 | F | 10000 | flu |
| 1965/02/02 | M | 20000 | dyspepsia |
| 1970/03/03 | F | 30000 | pneumonia |
| 1975/04/04 | M | 40000 | gastritis |

Medical Records

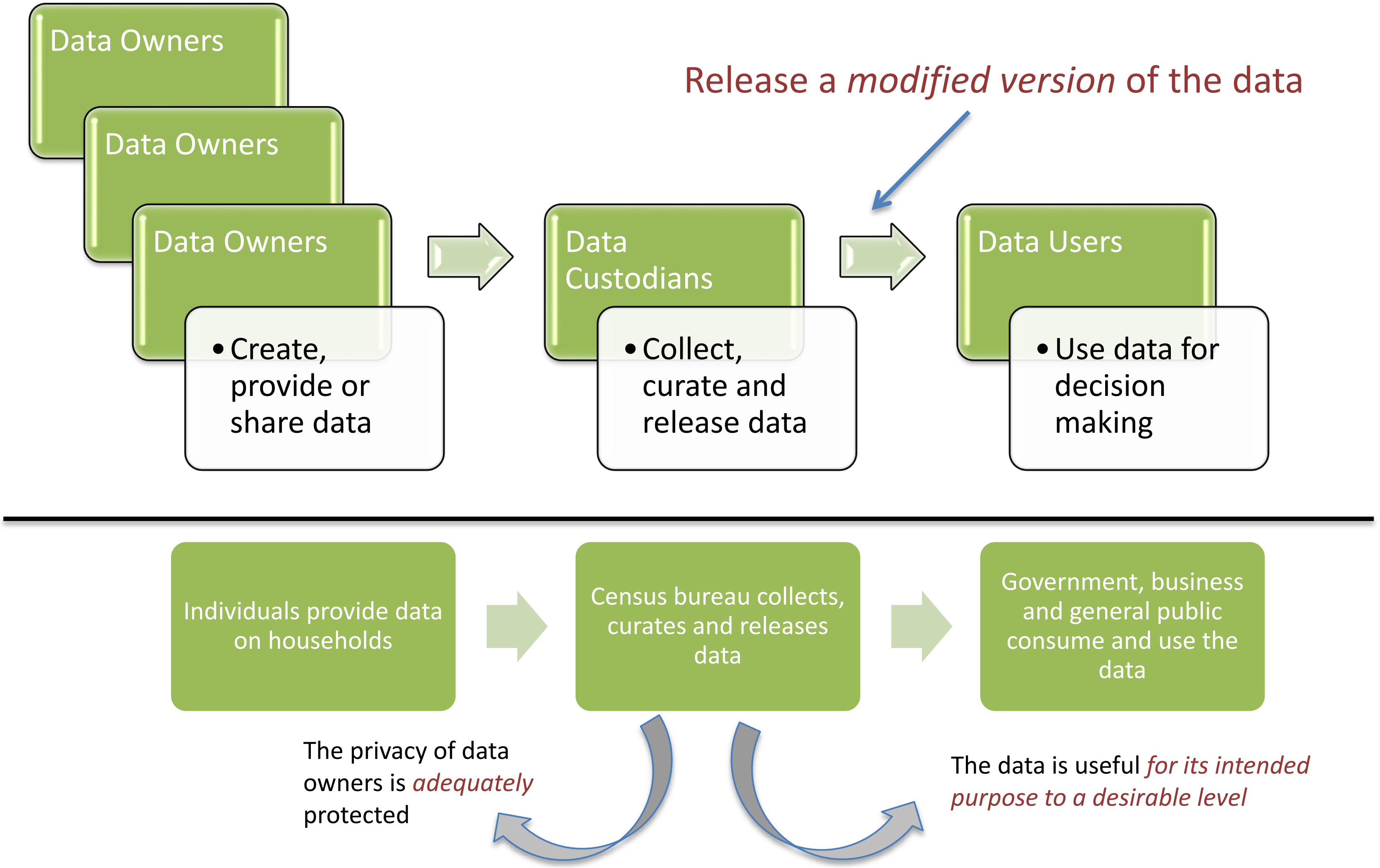
# Another Privacy Breach - AOL

Log record: < User ID, Query, … >
Example: < 4417749, “UQ”, … >

Method:

- Find all log entries for AOL user 4417749
- Many queries for businesses and services in Lilburn, GA (population 11K)
- A number of queries for different persons with the last name Arnold
- Lilburn has 14 people with the last name Arnold
- The **New York Times** contacted them and found that AOL User 4417749 is Thelma Arnold

# Privacy preserving data release



Data Owners

Data Owners

Data Owners

- Create, provide or share data

Data Custodians

- Collect, curate and release data

Release a *modified version* of the data

Data Users

- Use data for decision making

Individuals provide data on households

Census bureau collects, curates and releases data

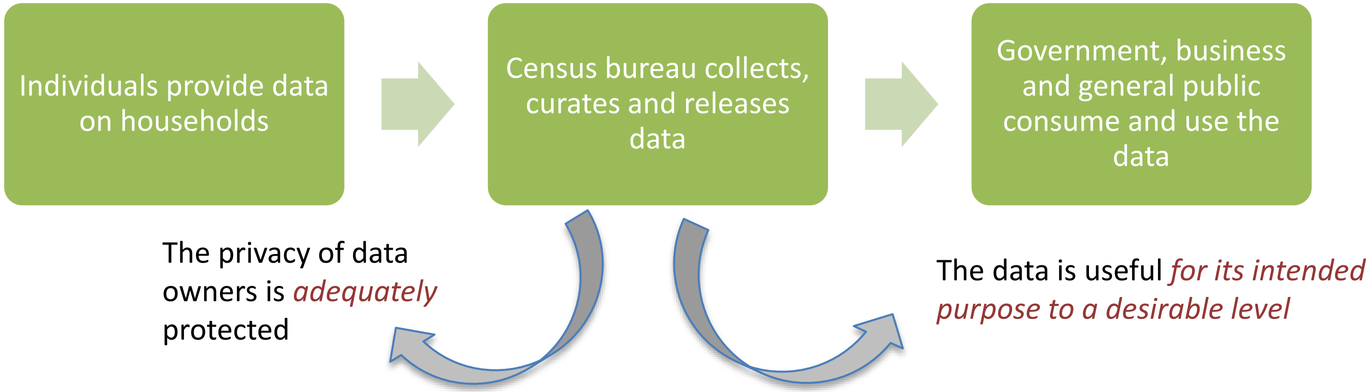
Government, business and general public consume and use the data

The privacy of data owners is *adequately* protected

The data is useful *for its intended purpose to a desirable level*

Acknowledgements Xiaokui Xiao, Nanyan Technological University

- privacy principle: what do we mean that privacy is by “adequately”protected?
- modification method: how should we modify the data to ensure adequate privacy while maximizing usefulness of the data for its intended purpose ?



Individuals provide data on households

Census bureau collects, curates and releases data

Government, business and general public consume and use the data

The privacy of data owners is *adequately* protected

The data is useful *for its intended purpose to a desirable level*

# Existing solutions (post 2000)

- ***K*-Anonymity**
- *l*-diversity
- Differential privacy

# *k*-Anonymity

Example: We want to release medical records

| **Name** | **Age** | **Zip** | **Disease** |
|---|---|---|---|
| John | 20 | 1000 | dyspepsia |
| Bob | 30 | 2000 | dyspepsia |
| Cathy | 40 | 3000 | pneumonia |
| Jane | 50 | 4000 | gastritis |

Latanya Sweeney. 2002. *k*-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* 10, 5 (October 2002), 557-570.

# *k*-Anonymity

*k*-anonymity [Sweeney, 2002] requires that <Age, ZIP> combination can be matched to at least *k* patients.

This is done by making Age and ZIP less specific.

| Name | Age | Zip |
|---|---|---|
| John | 20 | 1000 |
| Bob | 30 | 2000 |
| Cathy | 40 | 3000 |
| Jane | 50 | 4000 |

adversary's knowledge

| Age | Zip | Disease |
|---|---|---|
| 20 | 1000 | dyspepsia |
| 30 | 2000 | dyspepsia |
| 40 | 3000 | pneumonia |
| 50 | 4000 | gastritis |

medical records

# *k*-Anonymity

*k*-anonymity [Sweeney, 2002] requires that <Age, ZIP> combination can be matched to at least *k* patients. This is done by making Age and ZIP less specific.

| Name | Age | Zip |
|---|---|---|
| John | 20 | 1000 |
| Bob | 30 | 2000 |
| Cathy | 40 | 3000 |
| Jane | 50 | 4000 |

adversary's knowledge

generalization

| Age | Zip | Disease |
|---|---|---|
| [20, 30] | [1000, 2000] | dyspepsia |
| [20, 30] | [1000, 2000] | dyspepsia |
| [40, 50] | [3000, 4000] | pneumonia |
| [40, 50] | [3000, 4000] | gastritis |

medical records

2-anonymous table

# *k*-Anonymity

The general approach for *k*-anonymity required identification of attributes that an adversary may know e.g. Age and ZIP. These are called *Quasi-identifiers (QI)*.

You then divide the tuples into sizes of at least *k* and generalize the QI values of each group to make them identical.

QI

| **Age** | **Zip** | **Disease** |
|---|---|---|
| [20, 30] | [1000, 2000] | dyspepsia |
| [20, 30] | [1000, 2000] | dyspepsia |
| [40, 50] | [3000, 4000] | pneumonia |
| [40, 50] | [3000, 4000] | gastritis |

*k* groups

medical records

# *k*-Anonymity

*k*-anonymity requires that each combination of quasi-identifiers (QI) is hidden in a group of at least size *k.*

But what about the remaining attributes?

| Name | Age | Zip |
|---|---|---|
| John | 20 | 1000 |
| Bob | 30 | 2000 |
| Cathy | 40 | 3000 |
| Jane | 50 | 4000 |

adversary’s knowledge

| Age | Zip | Disease |
|---|---|---|
| [20, 30] | [1000, 2000] | dyspepsia |
| [20, 30] | [1000, 2000] | dyspepsia |
| [40, 50] | [3000, 4000] | pneumonia |
| [40, 50] | [3000, 4000] | gastritis |

medical records

sensitive attribute!

# *k*-Anonymity

*What do you know about John?*

| Name | Age | Zip |
|---|---|---|
| John | 20 | 1000 |
| Bob | 30 | 2000 |
| Cathy | 40 | 3000 |
| Jane | 50 | 4000 |

adversary's knowledge

| Age | Zip | Disease |
|---|---|---|
| [20, 30] | [1000, 2000] | dyspepsia |
| [20, 30] | [1000, 2000] | dyspepsia |
| [40, 50] | [3000, 4000] | pneumonia |
| [40, 50] | [3000, 4000] | gastritis |

medical records

# Vulnerability of Privacy Preserving Algorithms

- *k*-anonymity has been abandoned due to its vulnerability – disclosure of sensitive attributes is possible [Machanavajjhala et al. 2006].
- New algorithms have been proposed …
  - *l*-diversity
  - Differential privacy

# Task and Discussion

What are the two principles of private data release?

# POLL QUESTIONS - PRIVACY