SURV622/SURVMETH622: Privacy and Confidentiality

James Wagner
February 10, 2025

Context

Smart Cities

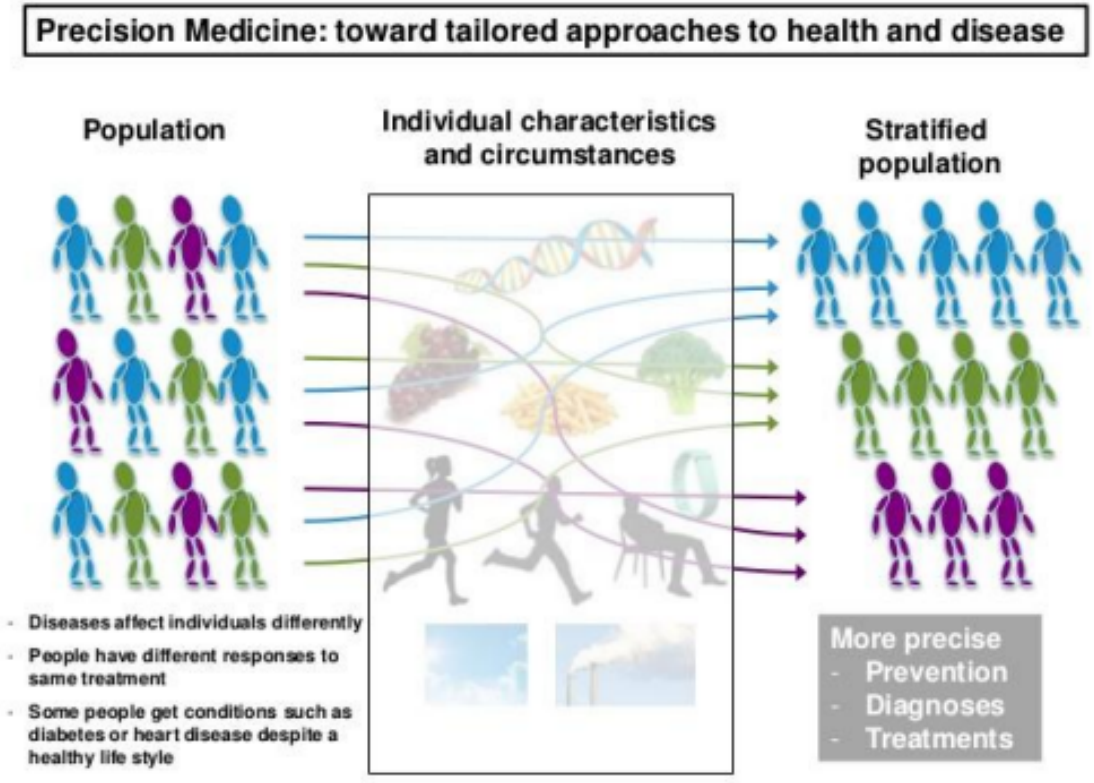
Three Types of Data:

-- **Digitized administrative datasets** on people and places (records of services for people, taxes paid, land transactions, police encounters)

-- **Sensors, wireless networks, video cameras,** etc. can monitor people and things throughout a city and the "Internet of Things" makes it possible to control things.

-- **Internet data** such as Google Street View, Zillow (real-estate), Yelp (reviews of retailers)

Precision Medicine



Issues:

-- What about people's privacy?

-- Can we optimize the systems?

-- Who controls the systems?

-- What happens when systems fail?

In Big Data Era

- Most data **no longer collected** by the government
  - Instead by internet search logs, social media, supermarket scanners, etc.

- Question: **how to share collected information** without violating privacy guarantees becomes more relevant

Additional Problems

- What is the legal framework when the ownership of data is unclear?
  - Collection and analysis often no longer within the same entity
  - Ownership of data less clear

- Who has the legal authority to make decisions about permission, access, and dissemination and under what circumstances?
  - The challenge in the case of big data is that data sources are often combined, collected for one purpose and used for another and users often have no good understanding of it or how their data will be used.

Concepts Out of Date

- Notification is either comprehensive or comprehensible, but not both (Nissenbaum, 2011)
- Understanding of the nature of harm has diffused over time.
- Consumers value their own privacy in variously flawed ways. (Acquisit, 2014)

Privacy Paradox

- Studies have shown that people say they care about privacy, but make decisions that contradict this.
  - “Our initial hypothesis that users’ privacy concern impede the depth and breadth of truthful online interaction was not confirmed. *In contrast, participants displayed a surprising readiness to reveal private and even highly personal information and to let themselves be ‘drawn into’ communication with the anthropomorphic 3-D bot*” (Spiekermann 2011)

  - “*Even within the subset of participants who expressed the highest degree of concern* over strangers being able to easily find out their sexual orientation, political views, and partners’ names, *48% did in fact publicly reveal their sexual orientation online, 47% revealed their political orientation, and 21% revealed their current partner’s name*.” (Acuisti, 2015)

Case Study: Issues with Consent

- Opt-in vs. opt-out wording
- Gain vs. loss framing
- Front vs. back placement

- *These design factors should not change respondents' concerns over privacy, yet the change consent rates*

*The data you **already provided** to us whould be **much more (gain frame) / much less (loss frame)** valuable if you would allow us to link them with .... Do you agree?*

Web	Back	Total
% agree: gain	62.4	520
% agree: loss	75.4	489
Total	498	1009

Phone	Front	Back	Total n
% agree	90.8	78.7	598

Web	Front	Back	Total
% agree	82.6	62.4	520

*The data you are about to provide (front) / already provided (back) to us would be **much more** valuable if you would allow us to link them with .... Do you agree?*

Sakshaug et al. 2018

Case Studies: Issues with Anonymization

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Kashmir Hill
FORBES STAFF

Welcome to The Not So Private Parts where technology & privacy collide

FOLLOW ON FORBES (2881)

FULL BIO >

Opinions expressed by Forbes Contributors are their own.

TARGET

Target has got you in its aim

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.

Case Studies: Issues with Anonymization

A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2006

E-MAIL
PRINT
REPRINTS

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.

Erik S. Lesser for The New York Times

Thelma Arnold's identity was betrayed by AOL records of her Web searches, like ones for her dog, Dudley, who clearly has a problem.

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on everything."

And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for "landscapers in Lilburn, Ga," several people with the last name Arnold and "homes sold in shadow lake subdivision gwinnett county georgia."

It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga., frequently researches her friends' medical ailments and loves her three dogs. "Those are my searches," she said, after a reporter read part of the list to her.

Privacy and Confidentiality

Privacy and Confidentiality

- What is *privacy*?

- What is *confidentiality?*

Privacy and Confidentiality

- What is *privacy*?
- **Privacy** refers to the rights that individuals have to keep information about themselves private
  - Government should not collect unnecessary information about individuals
  - There can be considerable disagreement about what that means
- What is *confidentiality*?

Privacy and Confidentiality

- What is *privacy*?
- **Privacy** refers to the rights that individuals have to keep information about themselves private
  - Government should not collect unnecessary information about individuals
  - There can be considerable disagreement about what that means
- What is *confidentiality*?
- **Confidentiality** refers to the duty of those who hold information about others that has been collected for a necessary purpose to protect that information

Privacy and Confidentiality

- Not always obvious what information an individual will consider sensitive
  - Information about some things very likely to be sensitive in context of our societies (e.g., income, health, or sexual behavior)
  - Information on other things could be sensitive under certain circumstances (e.g., location history, time and mode of travel for commute to work on reference daty, library record, Google search history, year of birth, or educational attainment)
- Threats to confidentiality may include threats related to both *data security* and *inferential disclosure*


I would like to present my client's internet search history from that evening.
I'd rather just confess to the murder.

Source: NBC

Data Security

- Holders of confidential individual-level data have a responsibility to adopt strong measures to prevent data breaches
- Common statistical agency practices for protecting data that cannot be made public include:
  - Limit data access to employees with sworn agents
  - Authorize access for employees and agents only to the data they need to fulfill their responsibilities
  - Maintain access logs so that unauthorized access can be detected
  - Require that data be accessed only via secure systems
  - Require passwords and/or multi-factor authentication for access
  - Require annual training for employees and agents on security practices

Data Security

- Linked data files more sensitive than separate input data files
  - Contain more information about each data subject
  - May be a richer and more tempting target for a potential attack
- Risk reduced if linkage limited to data files that include *only* the information needed to meet specific objectives and *only* for the time required for those objectives to be realized
  - Idea behind the National Security Data Service model recommended by the Commission on Evidence-Building Policymaking (2017)
- Requirements for this approach to be feasible in practice
  - Standardized procedures for requesting and obtaining permissions for data linkage, so that process does not take months or years
  - Standardized protocols for preserving source data and linkage code, so that analyses based on inked data can be replicated

Disclosure, Risks, and Control

What is Disclosure?

There are three types of disclosure: identity, attribute, and residual

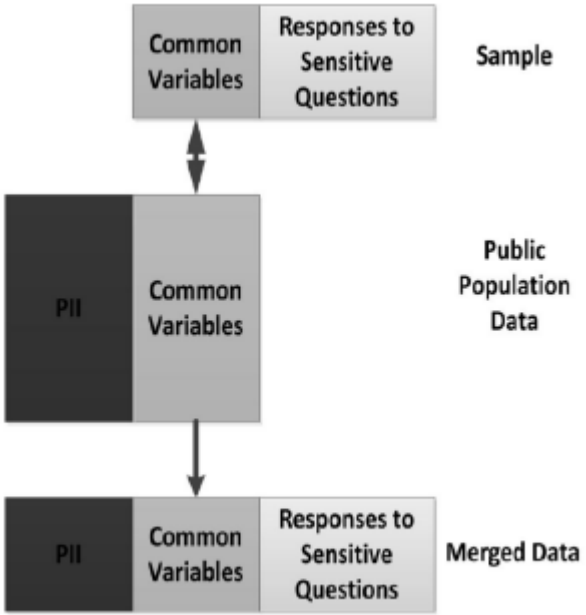
- **Identity** disclosure occurs when an individual can be identified from the released output, leading to information being provided about that identified subject
- **Attribute** disclosure occurs when confidential information is revealed and can be attributed to an individual. It is not necessary for a specific individual to be identified or for a specific value to be given for attribute disclosure to occur. For example, publishing a narrow range for the salary of persons exercising a particular profession in one region may constitute a disclosure.
- **Residual** disclosure can occur when released information can be combined to obtain confidential data. Care must be taken to examine all output to be released. While a table on its own might not disclose confidential information, disclosure can occur by combining information from several sources, including external ones (e.g., suppressed data in one table can be derived from other tables.)

Source: Guide for Researchers under Agreement with Statistics Canada, October 2005

Statistical Disclosure Risks: Microdata

- Statistical agencies commonly have released public use microdata from which direct identifiers have been removed
- Problem: based on even a small number of characteristics, many people are unique in the population
  - Example: in 1990, 87% (216 million of 248 million) of the population in the US had reported characteristics that likely made them unique based only on 5-digit ZIP code, gender, and date of birth (Sweeney 2002)
  - Example: In 2006, Netflix released a database containing histories of movie ratings given by subscribers. Researchers estimated that knowing the approximate dates (plus or minus two weeks) on which a person had rated six obscure movies would allow them to be identified in the database 99 percent of the time.

Statistical Disclosure Risks: Microdata



- Disclosure may occur if variables on sample file can be matched to same variables in public records or other accessible information
  - Matching may allow individuals to be identified
  - Example records: Births, marriages
- Data breaches that increase amount of publicly available information increase risk of a disclosure

Krenzke and Li (2019)

Statistical Disclosure Risks: Microdata

- Example: Massachusetts Group Insurance Commission released data on individual hospital stays during 1996 to researchers.
  - File contained birth data, sex, and ZIP code.
  - Researchers Latanya Sweeney identified records for Governor William Weld by linking to voting records that contained name, address, and ZIP code, birth data, and sex of every voter
- Example: State of Washington released hospital records satisfying HIPAA safe harbor standards (removal of 18 types of identifiers) to researchers.
  - Records contained no name, SSN, or addresses, but did include year and first three digits of ZIP codes with more than 20,000 residentes.
  - Sweeney identified individuals on file by matching to newspaper stories about accidents leading to hospitalizations.

Statistical Disclosure Risks: Microdata

- Example: AOL posted three months of search queries for 650,000 users, suppressing obviously identifying information such as username and IP address.
  - Two New York Times reporters were able to track down a user living in Georgia based on these search queries, which contained text such as “landscapers in Lilburn, GA,’ several people with the last name Arnold and ‘homes sold in shadow lake subdivision gwinnett county Georgia” (Ohm 2010)

Statistical Disclosure Risks: Tabular Data

- Allowing multiple queries against a database or publishing many tables, even if each individual table is safe, may create disclosure risks.
- Example: Disclosure of information about earned income based on responses to successive queries
  - First inquiry asks about all working women age 40-49. In the 100 records returned, average income is $39,800.
  - Second inquiry asks about all working women age 40-49 who were born in the US. In the 98 records returned, average income is $40,000.
  - Can infer average income for the two working women age 40-49 who were not born in the US
    - Solve for $x in 39800=(98*$40000 + 2*$x)/100 ⇒ $x=$30,000
  - In addition, either of the two working age women age 40-49 who were not born in the US could infer the income of the other exactly
  - On their own, either of the first two tabulations would pass typical disclosure reviews, but release of both tabulations inadvertently discloses a statistic than generally would not be releasable

Statistical Disclosure Risks: Tabular Data

- Example: Disclosure of information about individual characteristics based on multiple published tables
  - Statistical agencies often release large numbers of tabulations based on any given data collection
    - Example: More than 7.7 billion linearly independent statistics–or about 25 statistics per person–published based on data collected as part of the 2010 Census (Garfinkel, Abowd, and Martindale 2018)
  - Comparisons across tables can reveal information about patterns of characteristics at the individual level
  - If sufficiently many tables are published, may be possible to recreated the underlying microdata (data reconstruction theorem)

Practicalities of Disclosure Control

- The aim of disclosure control is to ensure that no unauthorized individual, technically competent with public data and private information could:
  - Identify any information not already public knowledge with a reasonable degree of confidence, and
  - Associate that information with the supplier of the information
- Modern privacy protection “injects noise” into the data
  - Allows the agency to set the level of tolerable risk
    - Tradeoff between utility of data and reduction of risk
  - Can produce detectable inconsistencies in the data
  - Can attenuate relationships in the data
  - Differential privacy protection applied to the 2020 Decennial Census