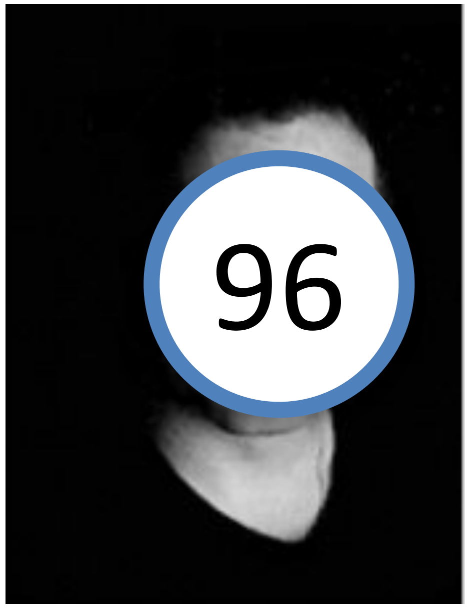
De-identification and Re-identification of PII

Simson L. Garfinkel
Information Access Division
National Institute of Standards and Technology

Paul Ohm
Professor of Law
Georgetown University Law Center



OMB Tech Tuesday

March 8, 2016

GEORGETOWN LAW

National Institute of Standards and Technology
U.S. Department of Commerce

De-identification and Re-identification of PII

Simson L. Garfinkel

Information Access Division

National Institute of Standards and Technology

Paul Ohm

Professor of Law

Georgetown University Law Center

OMB Tech Tuesday

March 8, 2016

GEORGETOWN LAW

NIST
National Institute of Standards and Technology
U.S. Department of Commerce

De-identification and Re-identification of PII

Simson L. Garfinkel
Information Access Division
National Institute of Standards and Technology

Paul Ohm
Professor of Law
Georgetown University Law Center

OMB Tech Tuesday

March 8, 2016

With thanks to Bradley Malin & Daniel Barth-Jones

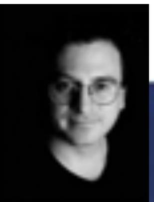
GEORGETOWN LAW

NIST National Institute of Standards and Technology U.S. Department of Commerce

De-Identification: Removing information that can identify

Text:

Images:

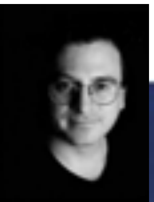


De-Identification: Removing information that can identify

Ulysses S. Grant	April 27, 1822	Hiram Ulysses Grant	18	Point Pleasant	Ohio
Rutherford B. Hayes	October 4, 1822	Rutherford Birchard Hayes	19	Delaware	Ohio
James A. Garfield	November 19, 1831	James Abram Garfield	20	Moreland Hills	Ohio
Chester A. Arthur	October 5, 1829	Chester Alan Arthur	21	Fairfield	Vermont
Grover Cleveland	March 18, 1837	Stephen Grover Cleveland	22	Caldwell	New Jersey
Benjamin Harrison	August 20, 1833		23	North Bend	Ohio
Grover Cleveland	March 18, 1837	Stephen Grover Cleveland	24	Caldwell	New Jersey

Text:

Images:

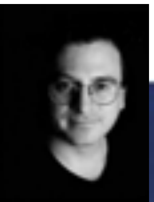


De-Identification: Removing information that can identify

Text:

Ulysses S. Grant	April 27, 1822	Hiram Ulysses Grant	18	Point Pleasant	Ohio
Rutherford B. Hayes	October 4, 1822	Rutherford Birchard Hayes	19	Delaware	Ohio
James A. Garfield	November 19, 1831	James Abram Garfield	20	Moreland Hills	Ohio
Chester A. Arthur	October 5, 1829	Chester Alan Arthur	21	Fairfield	Vermont
Grover Cleveland	March 18, 1837	Stephen Grover Cleveland	22	Caldwell	New Jersey
Benjamin Harrison	August 20, 1833		23	North Bend	Ohio
Grover Cleveland	March 18, 1837	Stephen Grover Cleveland	24	Caldwell	New Jersey

Images:



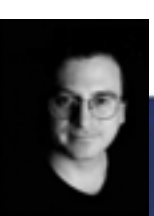
De-Identification: Removing information that can identify

Text:

Ulysses S. Grant	April 27, 1822	Hiram Ulysses Grant	18	Point Pleasant	Ohio
Rutherford B. Hayes	October 4, 1822	Rutherford Birchard Hayes	19	Delaware	Ohio
James A. Garfield	November 19, 1831	James Abram Garfield	20	Moreland Hills	Ohio
Chester A. Arthur	October 5, 1829	Chester Alan Arthur	21	Fairfield	Vermont
Grover Cleveland	March 18, 1837	Stephen Grover Cleveland	22	Caldwell	New Jersey
Benjamin Harrison	August 20, 1833		23	North Bend	Ohio
Grover Cleveland	March 18, 1837	Stephen Grover Cleveland	24	Caldwell	New Jersey

[redacted]	[redacted] 1822	[redacted]	18	Point Pleasant	Ohio
[redacted]	[redacted] 1822	[redacted]	19	Delaware	Ohio
[redacted]	[redacted] 1831	[redacted]	20	Moreland Hills	Ohio
[redacted]	[redacted] 1829	[redacted]	21	Fairfield	Vermont
[redacted]	[redacted] 1837	[redacted]	22	Caldwell	New Jersey
[redacted]	[redacted] 1833	[redacted]	23	North Bend	Ohio
[redacted]	[redacted] 1837	[redacted]	24	Caldwell	New Jersey

Images:



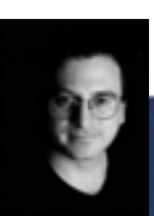
De-Identification: Removing information that can identify

Text:

Ulysses S. Grant	April 27, 1822	Hiram Ulysses Grant	18	Point Pleasant	Ohio
Rutherford B. Hayes	October 4, 1822	Rutherford Birchard Hayes	19	Delaware	Ohio
James A. Garfield	November 19, 1831	James Abram Garfield	20	Moreland Hills	Ohio
Chester A. Arthur	October 5, 1829	Chester Alan Arthur	21	Fairfield	Vermont
Grover Cleveland	March 18, 1837	Stephen Grover Cleveland	22	Caldwell	New Jersey
Benjamin Harrison	August 20, 1833		23	North Bend	Ohio
Grover Cleveland	March 18, 1837	Stephen Grover Cleveland	24	Caldwell	New Jersey

[redacted]	[redacted] 1822	[redacted]	18	Point Pleasant	Ohio
[redacted]	[redacted] 1822	[redacted]	19	Delaware	Ohio
[redacted]	[redacted] 1831	[redacted]	20	Moreland Hills	Ohio
[redacted]	[redacted] 1829	[redacted]	21	Fairfield	Vermont
[redacted]	[redacted] 1837	[redacted]	22	Caldwell	New Jersey
[redacted]	[redacted] 1833	[redacted]	23	North Bend	Ohio
[redacted]	[redacted] 1837	[redacted]	24	Caldwell	New Jersey

Images:



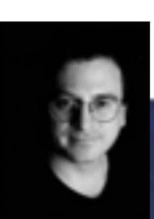
De-Identification: Removing information that can identify

Text:

Ulysses S. Grant	April 27, 1822	Hiram Ulysses Grant	18	Point Pleasant	Ohio
Rutherford B. Hayes	October 4, 1822	Rutherford Birchard Hayes	19	Delaware	Ohio
James A. Garfield	November 19, 1831	James Abram Garfield	20	Moreland Hills	Ohio
Chester A. Arthur	October 5, 1829	Chester Alan Arthur	21	Fairfield	Vermont
Grover Cleveland	March 18, 1837	Stephen Grover Cleveland	22	Caldwell	New Jersey
Benjamin Harrison	August 20, 1833		23	North Bend	Ohio
Grover Cleveland	March 18, 1837	Stephen Grover Cleveland	24	Caldwell	New Jersey

[redacted]	[redacted] 1822	[redacted]	18	Point Pleasant	Ohio
[redacted]	[redacted] 1822	[redacted]	19	Delaware	Ohio
[redacted]	[redacted] 1831	[redacted]	20	Moreland Hills	Ohio
[redacted]	[redacted] 1829	[redacted]	21	Fairfield	Vermont
[redacted]	[redacted] 1837	[redacted]	22	Caldwell	New Jersey
[redacted]	[redacted] 1833	[redacted]	23	North Bend	Ohio
[redacted]	[redacted] 1837	[redacted]	24	Caldwell	New Jersey

Images:



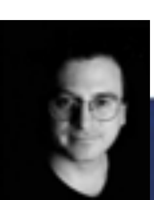
De-Identification: Removing information that can identify

Text:

Ulysses S. Grant	April 27, 1822	Hiram Ulysses Grant	18	Point Pleasant	Ohio
Rutherford B. Hayes	October 4, 1822	Rutherford Birchard Hayes	19	Delaware	Ohio
James A. Garfield	November 19, 1831	James Abram Garfield	20	Moreland Hills	Ohio
Chester A. Arthur	October 5, 1829	Chester Alan Arthur	21	Fairfield	Vermont
Grover Cleveland	March 18, 1837	Stephen Grover Cleveland	22	Caldwell	New Jersey
Benjamin Harrison	August 20, 1833		23	North Bend	Ohio
Grover Cleveland	March 18, 1837	Stephen Grover Cleveland	24	Caldwell	New Jersey

[redacted]	[redacted] 1822	[redacted]	18	Point Pleasant	Ohio
[redacted]	[redacted] 1822	[redacted]	19	Delaware	Ohio
[redacted]	[redacted] 1831	[redacted]	20	Moreland Hills	Ohio
[redacted]	[redacted] 1829	[redacted]	21	Fairfield	Vermont
[redacted]	[redacted] 1837	[redacted]	22	Caldwell	New Jersey
[redacted]	[redacted] 1833	[redacted]	23	North Bend	Ohio
[redacted]	[redacted] 1837	[redacted]	24	Caldwell	New Jersey

Images:



Why de-identify?

Why de-identify?

DATA.GOV

Data Publishing

Why de-identify?

DATA.GOV

Data Publishing

Controlled Sharing

Why de-identify?

Data Publishing

Controlled Sharing

Risk Mitigation

Why de-identify?

DATA.GOV

Data Publishing

Controlled Sharing

Risk Mitigation

Long-term archiving

Why de-identify?

DATA.GOV

Data Publishing

open science

Controlled Sharing

Risk Mitigation

Long-term archiving

Why de-identify?

DATA.GOV

Data Publishing

open science

Controlled Sharing

Risk Mitigation

Oversight

Long-term archiving

De-identification is *not* a single technique.

— *It's a collection of approaches, algorithms, and tools.*

— *Different approaches used with different kinds of data.*

— *Multiple regulations.*

https://pixabay.com/en/drill-milling-milling-machine-444484/

De-identification is about results:

— *No privacy interest in de-identified data (by definition.)*

— *De-identified data can be shared without permission of the data subjects.*

https://pixabay.com/en/child-boy-mask-color-key-188655/

The de-identification problem:
De-identified data can be … re-identified.

[redacted]	[redacted] 1822	[redacted]	18	Point Pleasant	Ohio
[redacted]	[redacted] 1822	[redacted]	19	Delaware	Ohio
[redacted]	[redacted] 1831	[redacted]	20	Moreland Hills	Ohio
[redacted]	[redacted] 1829	[redacted]	21	Fairfield	Vermont
[redacted]	[redacted] 1837	[redacted]	22	Caldwell	New Jersey
[redacted]	[redacted] 1833	[redacted]	23	North Bend	Ohio
[redacted]	[redacted] 1837	[redacted]	24	Caldwell	New Jersey

Re-identification links with another dataset.

Re-identification is rarely 100% certain.

The de-identification problem:
De-identified data can be … re-identified.

[redacted]	[redacted] 1822	[redacted]	18	Point Pleasant	Ohio
[redacted]	[redacted] 1822	[redacted]	19	Delaware	Ohio
[redacted]	[redacted] 1831	[redacted]	20	Moreland Hills	Ohio
[redacted]	[redacted] 1829	[redacted]	21	Fairfield	Vermont
[redacted]	[redacted] 1837	[redacted]	22	Caldwell	New Jersey
[redacted]	[redacted] 1833	[redacted]	23	North Bend	Ohio
[redacted]	[redacted] 1837	[redacted]	24	Caldwell	New Jersey

Re-identification links with another dataset.

Re-identification is rarely 100% certain.

The de-identification problem:
De-identified data can be … re-identified.

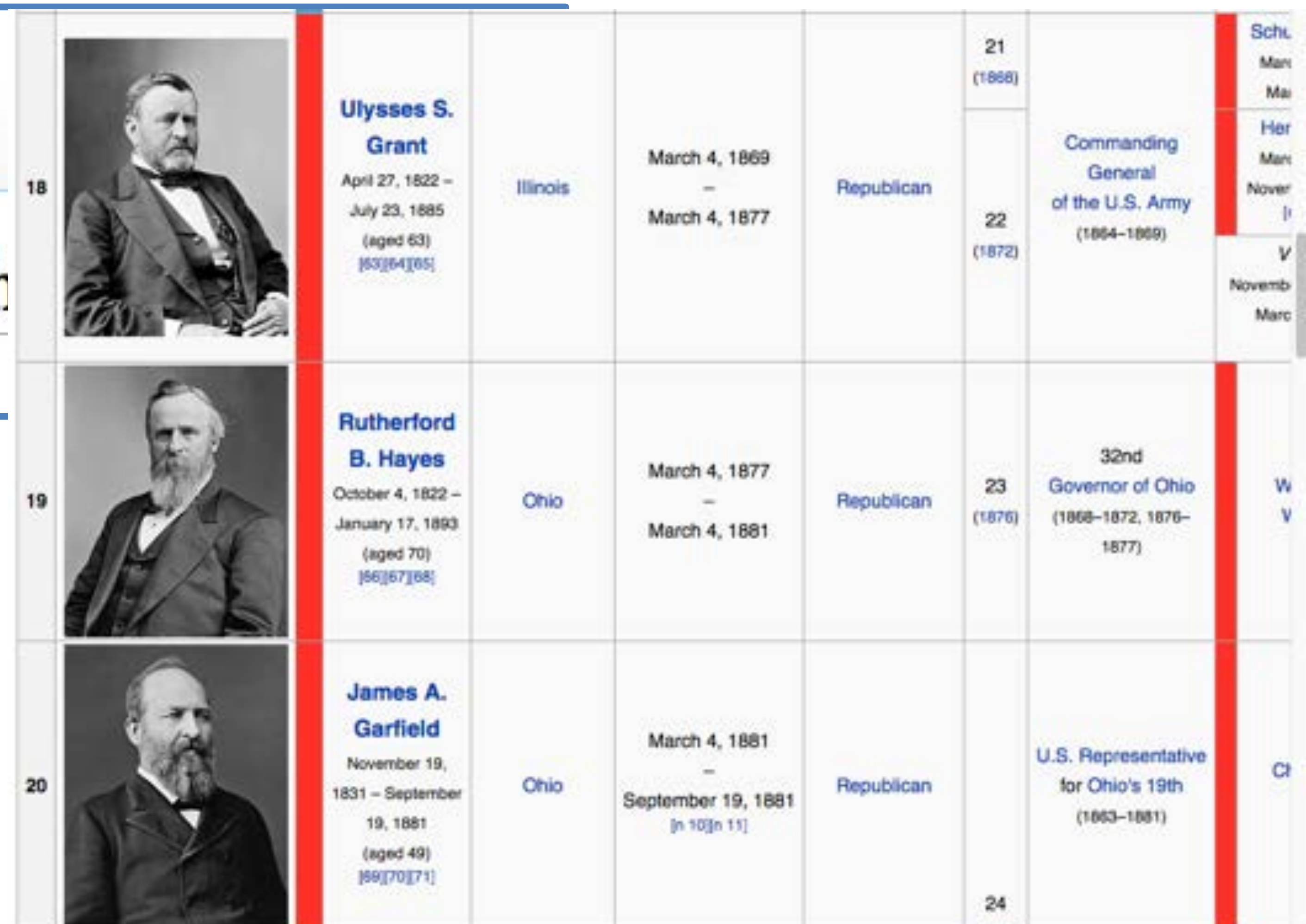
[redacted]	[redacted] 1822	[redacted]	18	Point Pleasant	Ohio
	[redacted] 1822		19	Delaware	Ohio
	[redacted] 1831		20	Moreland Hills	Ohio
	[redacted] 1829		21	Fairfield	Vermont
	[redacted] 1837		22	Caldwell	New Jersey
	[redacted] 1833		23	North Bend	Ohio
	[redacted] 1837		24	Caldwell	New Jersey

WIKIPEDIA
The Free Encyclopedia

Article Talk

List of Presidents of th

From Wikipedia, the free encyclopedia

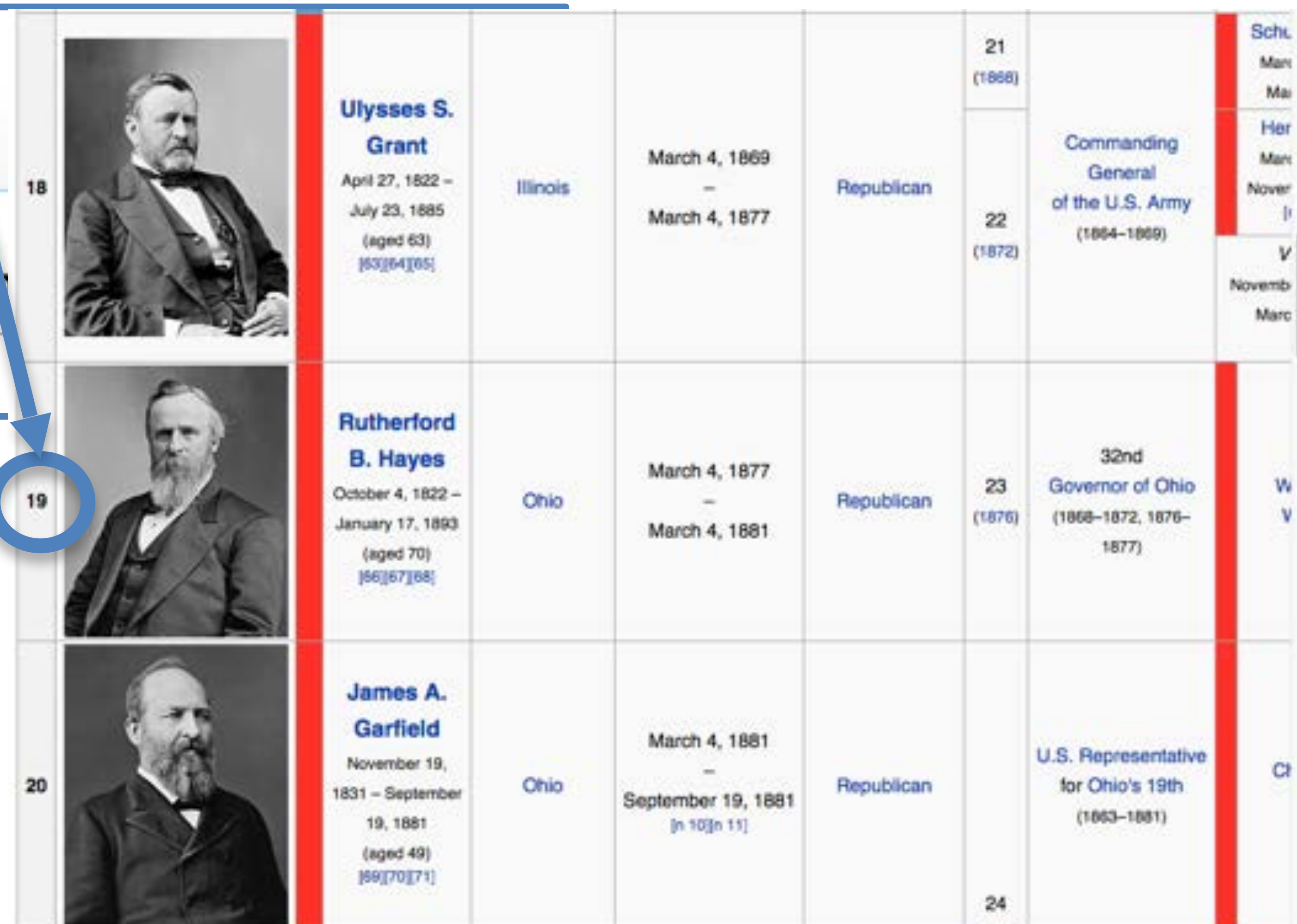


Re-identification links with another dataset.

Re-identification is rarely 100% certain.

The de-identification problem:
De-identified data can be … re-identified.

[redacted]	1822	[redacted]	18	Point Pleasant	Ohio
[redacted]	1822	[redacted]	19	Delaware	Ohio
[redacted]	1831	[redacted]	20	Moreland Hills	Ohio
[redacted]	1829	[redacted]	21	Fairfield	Vermont
[redacted]	1837	[redacted]	22	Caldwell	New Jersey
[redacted]	1833	[redacted]	23	North Bend	Ohio
[redacted]	1837	[redacted]	24	Caldwell	New Jersey

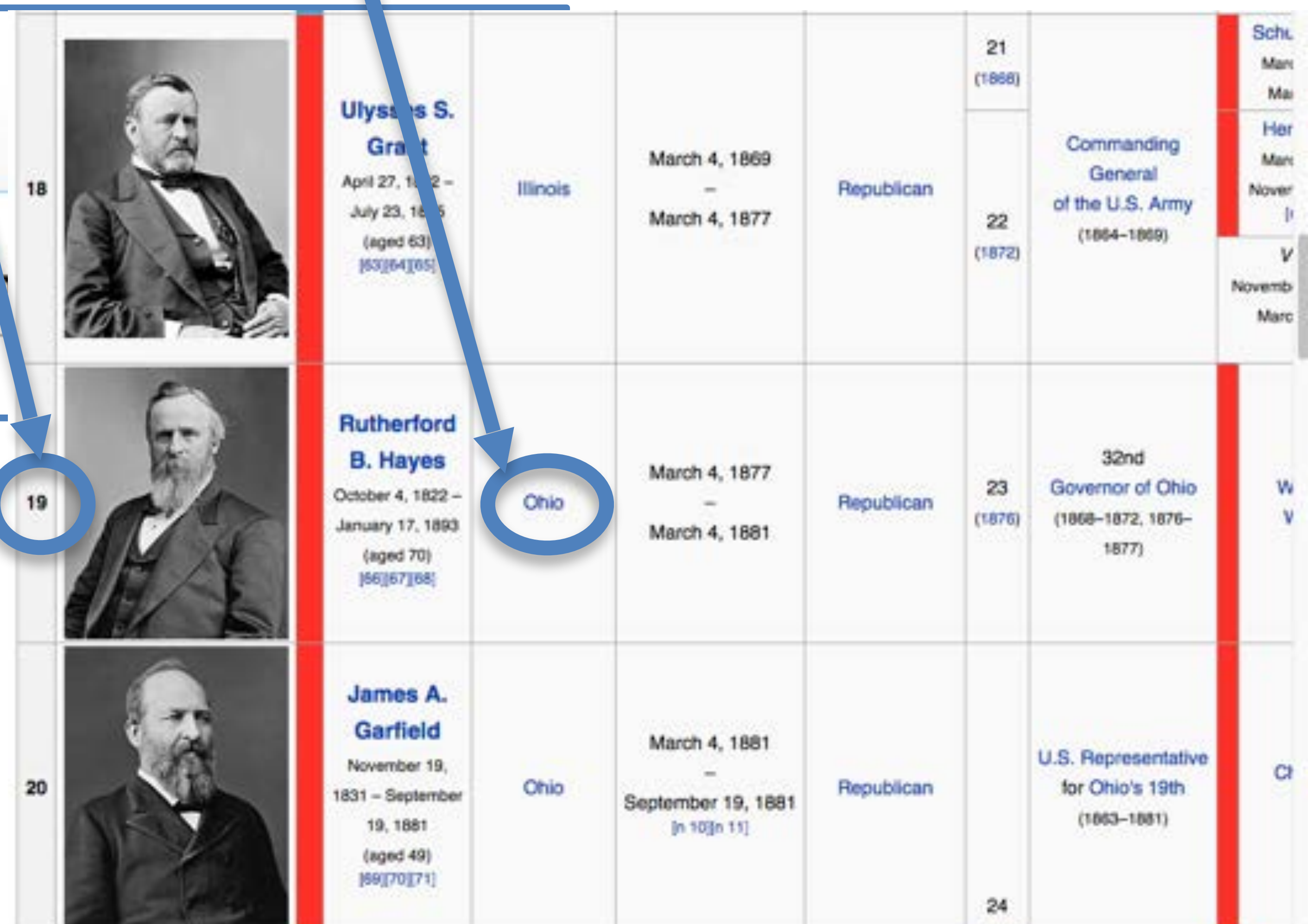


Re-identification links with another dataset.

Re-identification is rarely 100% certain.

The de-identification problem:
De-identified data can be … re-identified.

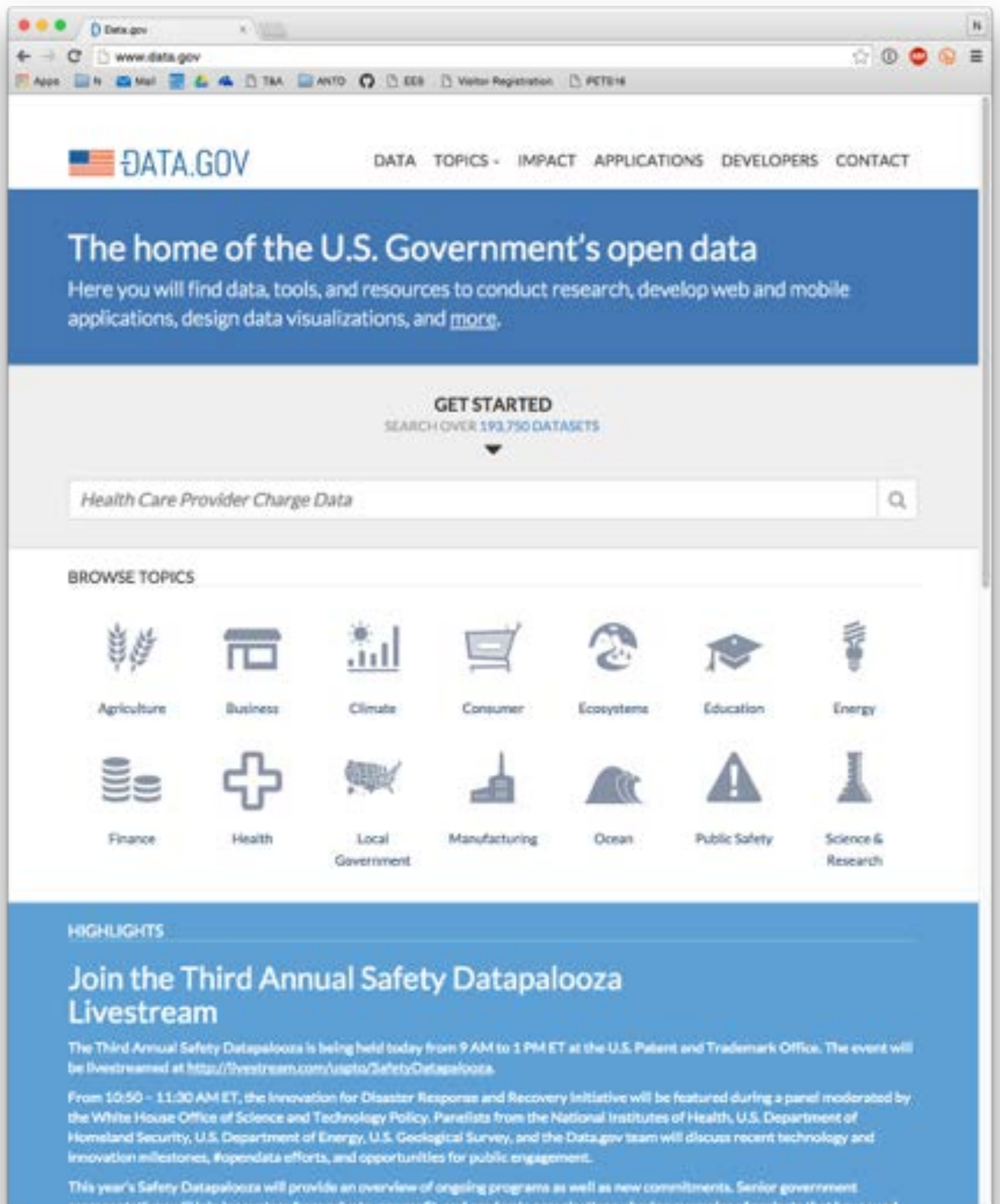
[redacted]	[redacted] 1822	[redacted]	18	Point Pleasant	Ohio
	[redacted] 1822		19	Delaware	Ohio
	[redacted] 1831		20	Moreland Hills	Ohio
	[redacted] 1829		21	Fairfield	Vermont
	[redacted] 1837		22	Caldwell	New Jersey
	[redacted] 1833		23	North Bend	Ohio
	[redacted] 1837		24	Caldwell	New Jersey



Re-identification links with another dataset.

Re-identification is rarely 100% certain.

Public policy is on a collision course:
Open Data vs. Personal Privacy



Privacy

Surveillance

Hackers

Detailed data about individuals is a new "public good." We can use data for medical research!

Stanford MEDICINE | News Center

Email · Share 0 · Tweet

Dangerous side effect of common drug combination discovered by data mining

MAY 25
2011

A widely used combination of two common medications may cause unexpected increases in blood glucose levels, according to a study conducted at the Stanford University School of Medicine, Vanderbilt University and Harvard Medical School. Researchers were surprised at the finding because neither of the two drugs — one, an antidepressant marketed as Paxil, and the other, a cholesterol-lowering medication called Pravachol — has a similar effect alone.

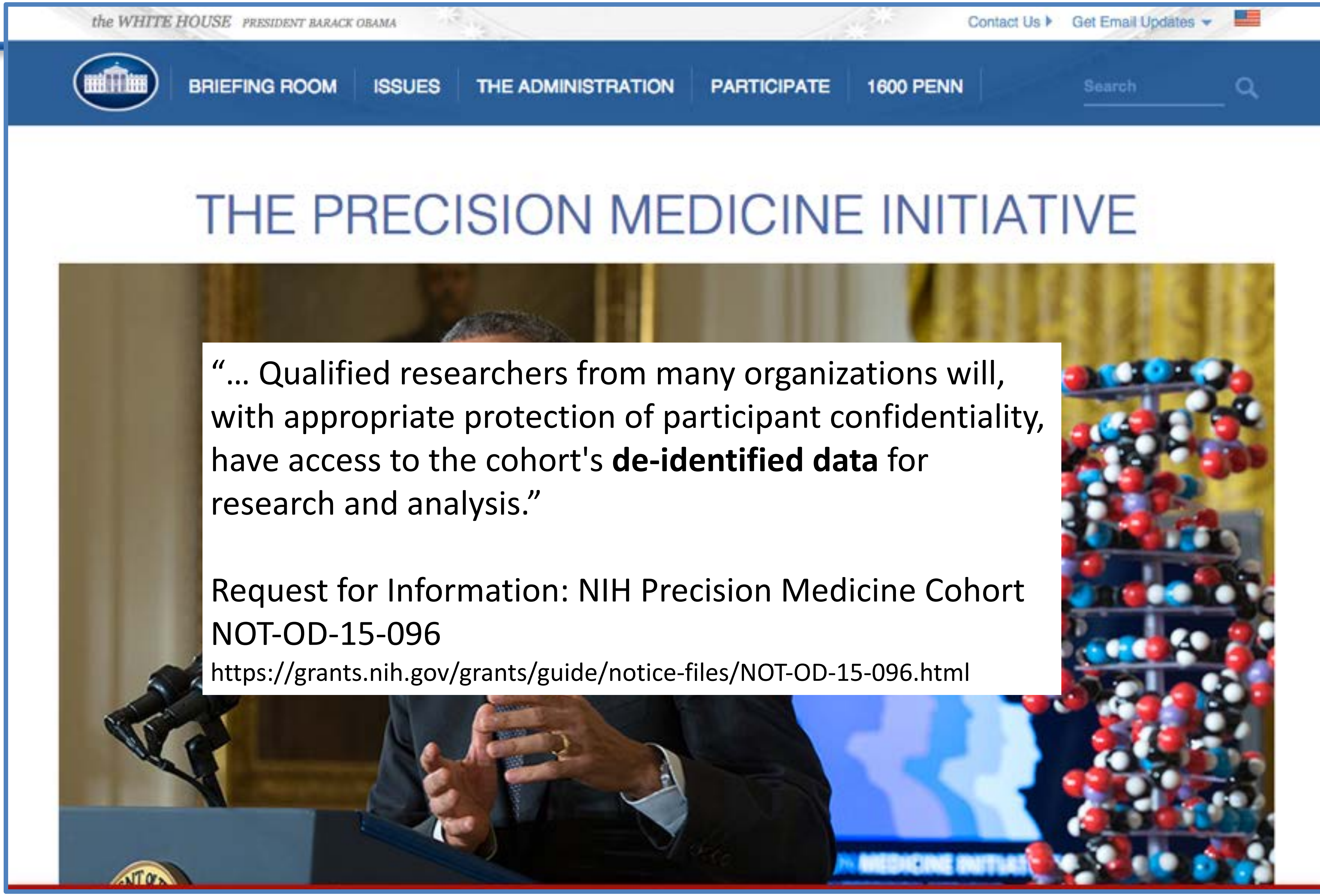

Russ Altman

The increase is more pronounced in people who are diabetic, and in whom the control of blood sugar levels is particularly important. It's also apparent in pre-diabetic laboratory mice exposed to both drugs. The researchers speculate that between 500,000 and 1 million people in this country may be taking the two medications simultaneously.

https://med.stanford.edu/news/all-news/2011/05/dangerous-side-effect-of-common-drug-combination-discovered-by-data-mining.html

Big-data is not a new science—it's the future of all science.

Big-data is not a new science—it's the future of all science.



"… Qualified researchers from many organizations will, with appropriate protection of participant confidentiality, have access to the cohort's **de-identified data** for research and analysis."

Request for Information: NIH Precision Medicine Cohort
NOT-OD-15-096
https://grants.nih.gov/grants/guide/notice-files/NOT-OD-15-096.html

Per-Trip data is the future of transportation planning.

January 13, 2015:

- Uber promises to provide Boston with "anonymized trip-level data by ZIP Code Tabulation Area (ZCTA)."

Data Includes:

- Timestamp
- ZCTA in which trip began
- ZCTA in which trip ended
- Distance traveled
- Duration, in seconds

Uses:

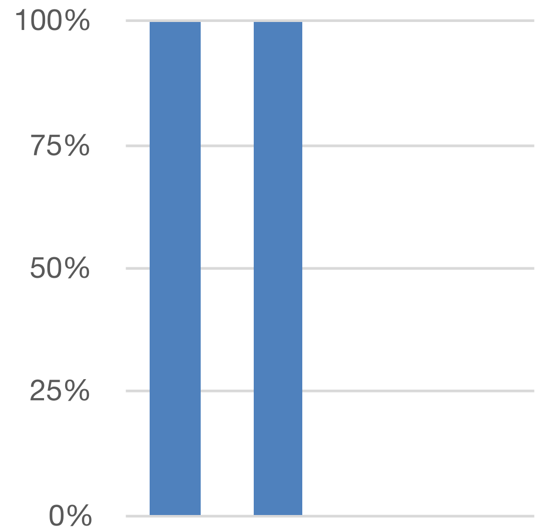
- Traffic analysis
- Detect underserved areas

Pothole Detection:
Using real-time data to avoid the next big thing!

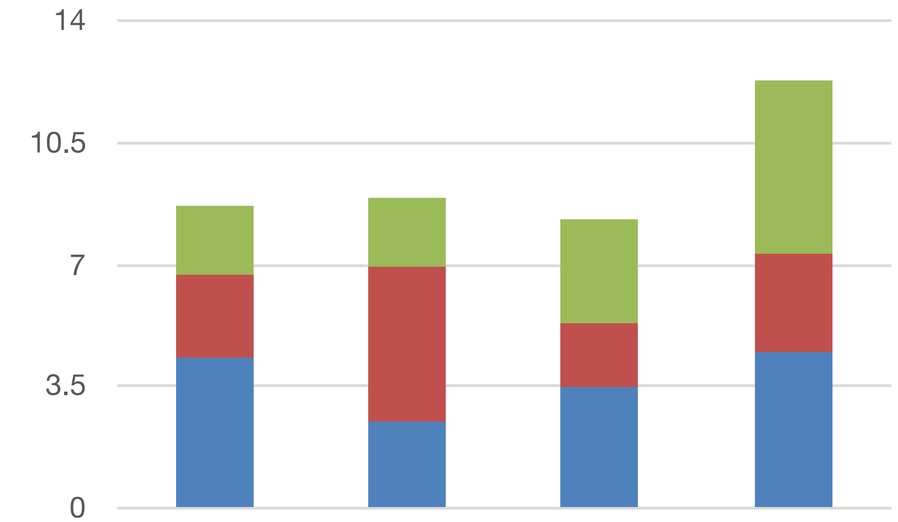
- Share de-identified data with other drivers.
- Alert authorities.

http://www.cheatsheet.com/automobiles/pothole-detection-is-this-the-next-big-car-technology.html/

Education:
Published student-level data allows for re-analysis by unaffiliated third parties (e.g. researchers).



Aggregate data



Re-analyzed

Existing US laws and regulations trust de-identification to protect privacy.

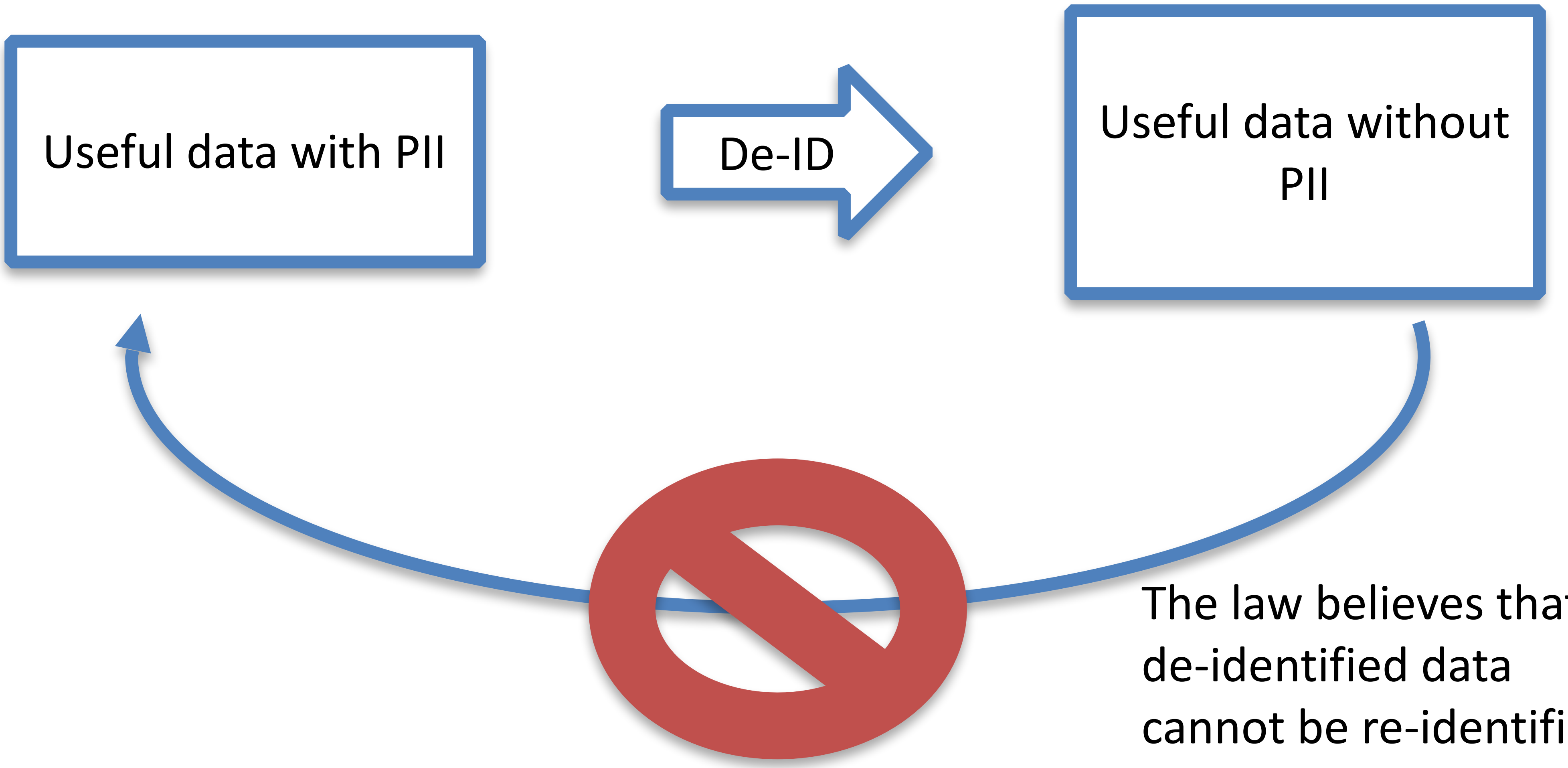
- Educational records can be released if de-identified (FERPA)

- Medical records can be released if de-identified (HIPAA)

- Foodborne Illness Surveillance System allows public release of de-identified aggregate data

- Voluntary safety reports submitted to FAA can be released if the data they contain are de-identified

Our laws assume that perfect de-identification is possible.



The law believes that de-identified data cannot be re-identified.

De-identification questions:

- How do you know if data are properly de-identified?
- What is “anonymized” vs. “de-identified” vs. “pseudonymized?”
- What is the privacy/utility trade-off?

Outline for today's talk

De-identification lets us use data while protecting privacy.

De-identified data can be re-identified.

- Why de-identify? ✔
- Basic de-identification
- Famous re-identification controversies
- De-identification in practice
- Measuring re-identification risk
- De-identification governance
- De-identification @ NIST — Workshop June 29th

President	Birth	Date of Inauguration	Age at Inauguration
XXXXXX	XXXXXX	XXXXXX	57 years, 67 days
XXXXXX	XXXXXX	XXXXXX	61 years, 125 days
XXXXXX	XXXXXX	XXXXXX	57 years, 325 days
XXXXXX	XXXXXX	XXXXXX	57 years, 353 days
XXXXXX	XXXXXX	XXXXXX	58 years, 310 days
XXXXXX	XXXXXX	XXXXXX	57 years, 236 days
XXXXXX	XXXXXX	XXXXXX	61 years, 354 days
XXXXXX	XXXXXX	XXXXXX	54 years, 89 days

Basic De-Identification

- William Weld & Latanya Sweeney
- Identifiers vs. Quasi-Identifiers
- HIPAA Privacy Rule
- Testing the HIPAA Privacy Rule

Original approach:
remove the "directly identifying" information.

Direct Identifiers

President	*Birth*	*Date of Inauguration*	*Age at Inauguration*
George Washington	*February 22, 1732*	*April 30, 1789*	*57 years, 67 days*
John Adams	*October 30, 1735*	*March 4, 1797*	*61 years, 125 days*
Thomas Jefferson	*April 13, 1743*	*March 4, 1801*	*57 years, 325 days*
James Madison	*March 16, 1751*	*March 4, 1809*	*57 years, 353 days*
James Monroe	*April 28, 1758*	*March 4, 1817*	*58 years, 310 days*
John Quincy Adams	*July 11, 1767*	*March 4, 1825*	*57 years, 236 days*
Andrew Jackson	*March 15, 1767*	*March 4, 1829*	*61 years, 354 days*
Martin Van Buren	*December 5, 1782*	*March 4, 1837*	*54 years, 89 days*

Original approach:
remove the "directly identifying" information.

Direct Identifiers

President	*Birth*	*Date of Inauguration*	*Age at Inauguration*
XXXXX	*February 22, 1732*	*April 30, 1789*	*57 years, 67 days*
XXXXX	*October 30, 1735*	*March 4, 1797*	*61 years, 125 days*
XXXXX	*April 13, 1743*	*March 4, 1801*	*57 years, 325 days*
XXXXX	*March 16, 1751*	*March 4, 1809*	*57 years, 353 days*
XXXXX	*April 28, 1758*	*March 4, 1817*	*58 years, 310 days*
XXXXX	*July 11, 1767*	*March 4, 1825*	*57 years, 236 days*
XXXXX	*March 15, 1767*	*March 4, 1829*	*61 years, 354 days*
XXXXX	*December 5, 1782*	*March 4, 1837*	*54 years, 89 days*

The problem: there may be *another database* that includes some of the remaining information.

W List of Presidents of the U ×

Simson

https://en.wikipedia.org/wiki/List_of_Presidents_of_the_United_States_by_date_of_birth

Apps VA wikis apps $ TTD news ref NIST Learn

Create account Log in

Article Talk Read Edit View history Search

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export
Create a book
Download as PDF
Printable version

Languages
Română
中文
Edit links

List of Presidents of the United States by date of birth

From Wikipedia, the free encyclopedia

The following is a list of U.S. Presidents, organized by **date of birth**, plus additional lists of birth related statistics.

Contents [show]

United States Presidents by date of birth [edit]

OB = Order of Birth OP = Order of Presidency AP = Age when assumed Presidency

Note: As Grover Cleveland served two non-consecutive terms, he assumed office twice, as the 22nd and 24th President.

OB	Name	Date of Birth	Birth Name	OP	Birthplace	State of Birth	AP
1	George Washington	February 22, 1732		1	Pope's Creek	Virginia	57
2	John Adams	October 30, 1735	John Adams, Jr.	2	Braintree	Massachusetts	61
3	Thomas Jefferson	April 13, 1743		3	Goochland County	Virginia	57
4	James Madison	March 16, 1751	James Madison, Jr.	4	Port Conway	Virginia	57
5	James Monroe	April 28, 1758		5	Monroe Hall	Virginia	58
7	John Quincy Adams	July 11, 1767		6	Braintree	Massachusetts	57
6	Andrew Jackson	March 15, 1767		7	Waxhaws Region	South/North Carolina	61
9	Martin Van Buren	December 5, 1782		8	Kinderhook	New York	54
8	William Henry Harrison	February 9, 1773		9	Charles City County	Virginia	68
11	John Tyler	March 29, 1790	John Tyler, Jr.	10	Charles City County	Virginia	51
13	James K. Polk	November 2, 1795	James Knox Polk	11	Pineville	North Carolina	49
10	Zachary Taylor	November 24, 1784		12	Barboursville	Virginia	64
14	Millard Fillmore	January 7, 1800		13	Moravia	New York	50
15	Franklin Pierce	November 23, 1804		14	Hillsborough	New Hampshire	48

These two databases can be linked.

President	Birth	Date of Inauguration	Favorite Color
XXXXX	February 22, 1732	April 30, 1789	Red
XXXXX	October 30, 1735	March 4, 1797	Blue
XXXXX	April 13, 1743	March 4, 1801	Green
XXXXX	March 16, 1751	March 4, 1809	Yellow
XXXXX	April 28, 1758	March 4, 1817	Red
XXXXX	July 11, 1767	March 4, 1825	Orange
XXXXX	March 15, 1767	March 4, 1829	Cyan
XXXXX	December 5, 1782	March 4, 1837	Blue

Private Information

These two databases can be linked.

President	Birth	Date of Inauguration	Favorite Color
XXXXX	February 22, 1732	April 30, 1789	Red
XXXXX	October 30, 1735	March 4, 1797	Blue
XXXXX	April 13, 1743	March 4, 1801	Green
XXXXX	March 16, 1751	March 4, 1809	Yellow
XXXXX	April 28, 1758	March 4, 1817	Red
XXXXX	July 11, 1767	March 4, 1825	Orange
XXXXX	March 15, 1767	March 4, 1829	Cyan
XXXXX	December 5, 1782	March 4, 1837	Blue

Private Information

These two databases can be linked.

President	Birth	Date of Inauguration	Favorite Color
XXXXX	February 22, 1732	April 30, 1789	Red
XXXXX	October 30, 1735	March 4, 1797	Blue
XXXXX	April 13, 1743	March 4, 1801	Green
XXXXX	March 16, 1751	March 4, 1809	Yellow
XXXXX	April 28, 1758	March 4, 1817	Red
XXXXX	July 11, 1767	March 4, 1825	Orange
XXXXX	March 15, 1767	March 4, 1829	Cyan
XXXXX	December 5, 1782	March 4, 1837	Blue

These two databases can be linked.

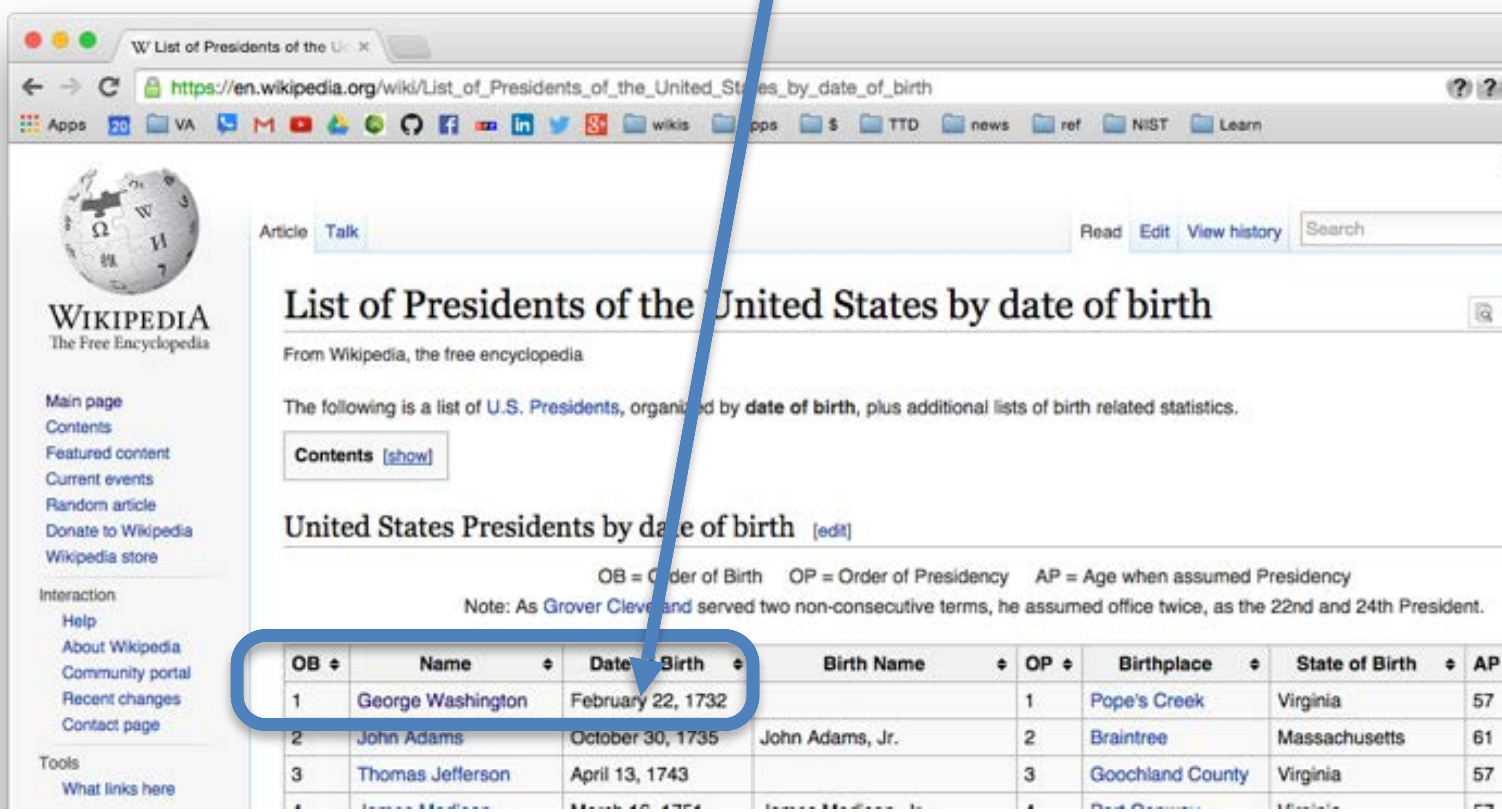
President	Birth	Date of Inauguration	Favorite Color
XXXXX	*February 22, 1732*	*April 30, 1789*	*Red*
XXXXX	*October 30, 1735*	*March 4, 1797*	*Blue*
XXXXX	*April 13, 1743*	*March 4, 1801*	*Green*
XXXXX	*March 16, 1751*	*March 4, 1809*	*Yellow*
XXXXX	*April 28, 1758*	*March 4, 1817*	*Red*
XXXXX	*July 11, 1767*	*March 4, 1825*	*Orange*

This is called a "linkage attack."

"Birth date" is an *indirect identifier.*

Also called a "quasi Identifier."

President	Birth	Date of Inauguration	Favorite Color
XXXXX	February 22, 1732	April 30, 1789	Red
XXXXX	October 30, 1735	March 4, 1797	Blue
XXXXX	April 13, 1743	March 4, 1801	Green
XXXXX	March 16, 1751	March 4, 1809	Yellow
XXXXX	April 28, 1758	March 4, 1817	Red
XXXXX	July 11, 1767	March 4, 1825	Orange
XXXXX	March 15, 1767	March 4, 1829	Cyan
XXXXX	December 5, 1782	March 4, 1837	Blue



In 2000 Latanya Sweeney demonstrated a linkage attack. She re-identified MA governor William Weld's hospital records.

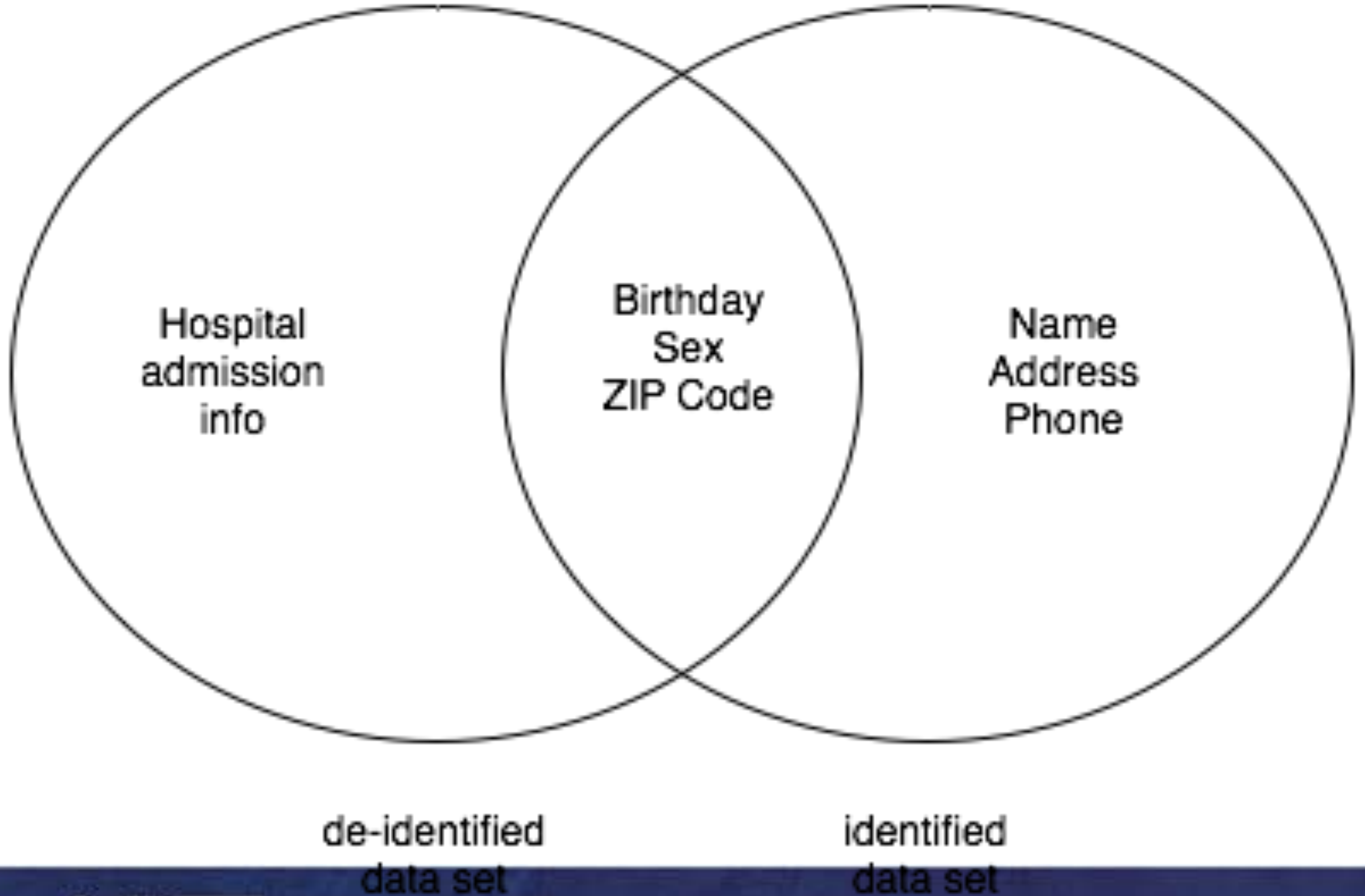
- Weld had fainted in 1996 and was admitted to a hospital.

- State of MA made "de-identified" hospital records of state employees available for research on health care.
  - *Removed name, but left birthday, sex & ZIP code remained.*

Sweeney purchased voter registration records. (Cambridge, MA)

Sweeney found a record in each data set with identical birthday, sex & ZIP

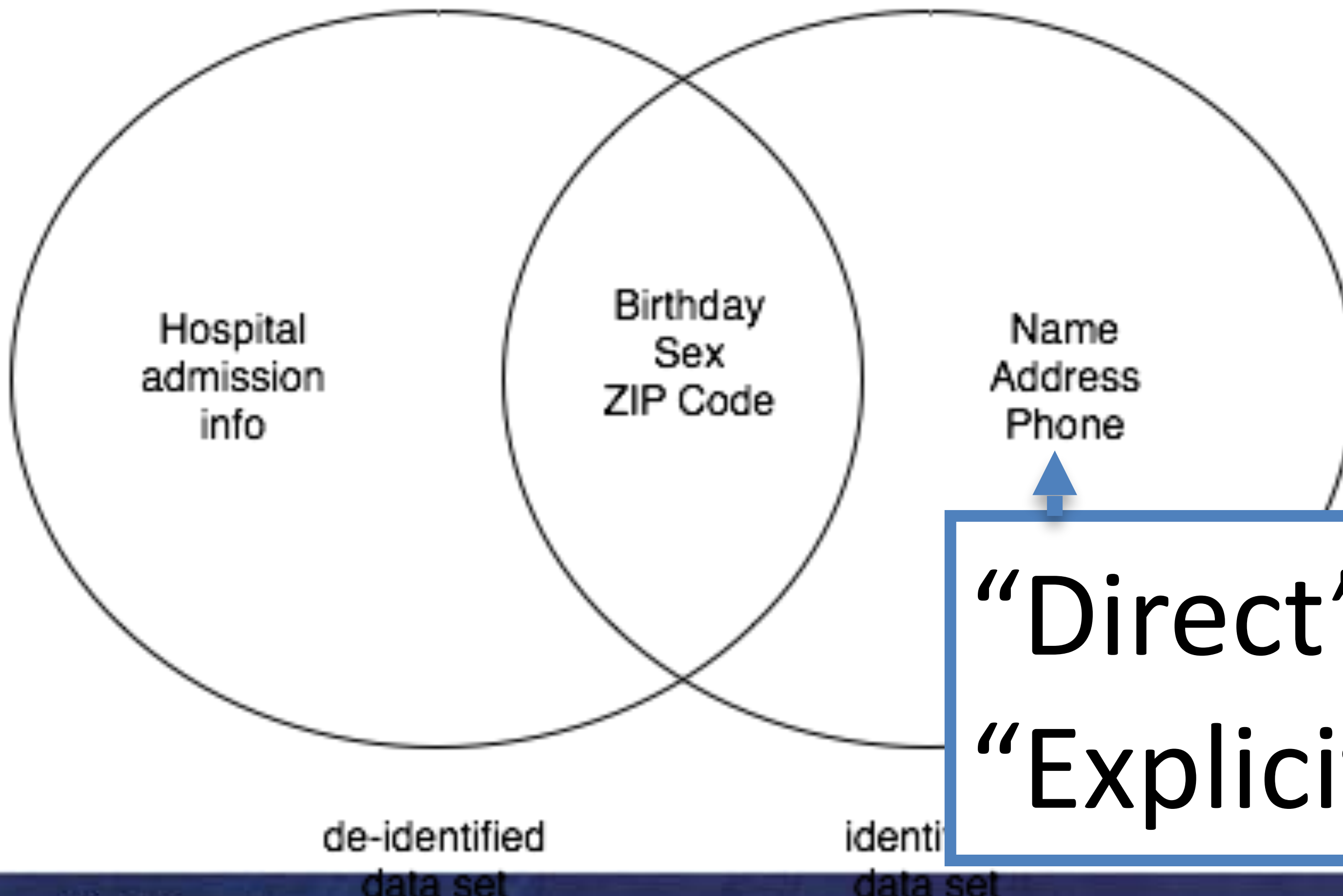
- Weld's records were uniquely identified.
- Sweeney estimated **87%** of US population were uniquely identified by birthday, sex & ZIP



Sweeney purchased voter registration records. (Cambridge, MA)

Sweeney found a record in each data set with identical birthday, sex & ZIP

- Weld’s records were uniquely identified.
- Sweeney estimated **87%** of US population were uniquely identified by birthday, sex & ZIP

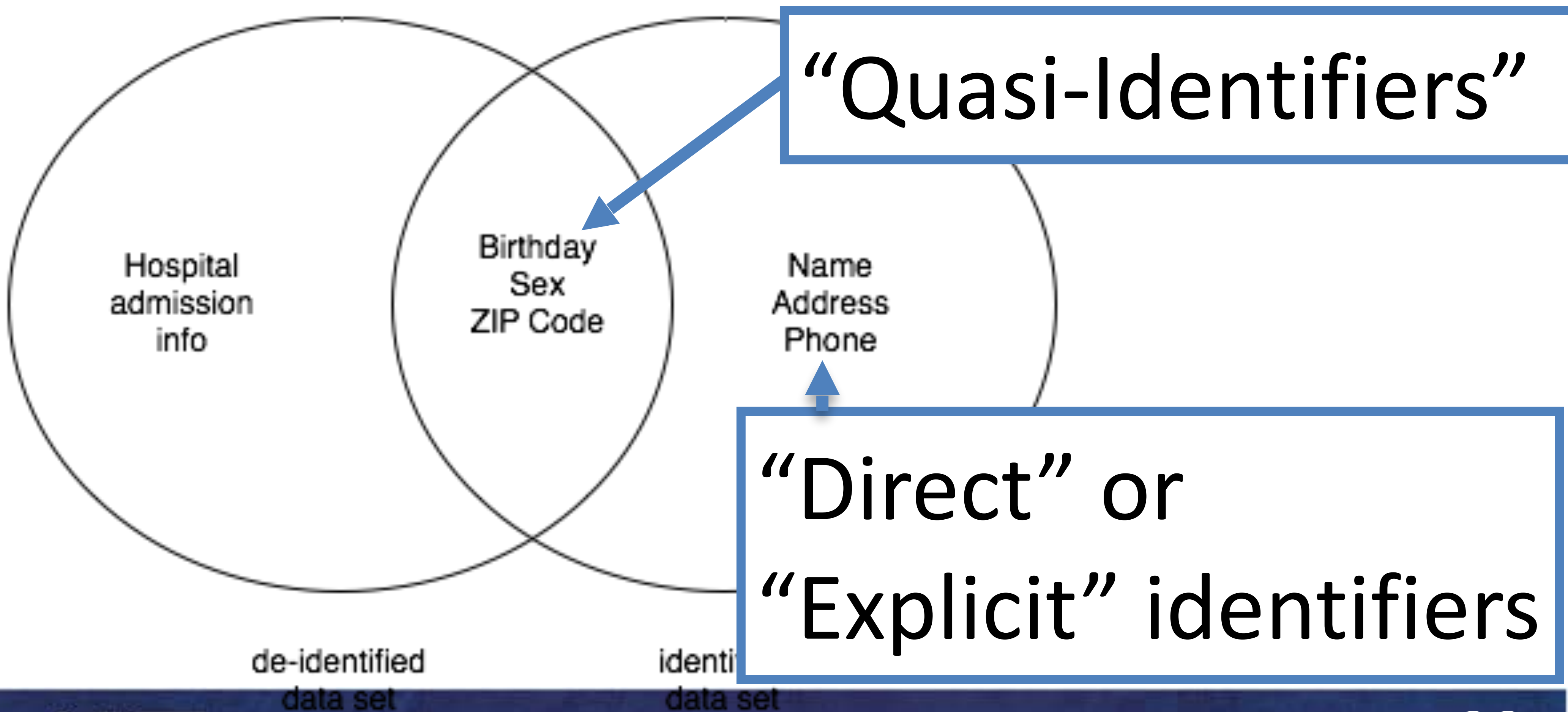


“Direct” or “Explicit” identifiers

Sweeney purchased voter registration records. (Cambridge, MA)

Sweeney found a record in each data set with identical birthday, sex & ZIP

- Weld's records were uniquely identified.
- Sweeney estimated **87%** of US population were uniquely identified by birthday, sex & ZIP



Basic de-identification with Direct Identifiers & Quasi-Identifiers

Direct Identifiers — Main function is to identify people.

- Name
- SSN

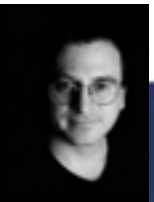
  — *Identifiers must be suppressed*

Quasi-Identifiers — Useful for analysis, but can also identify.

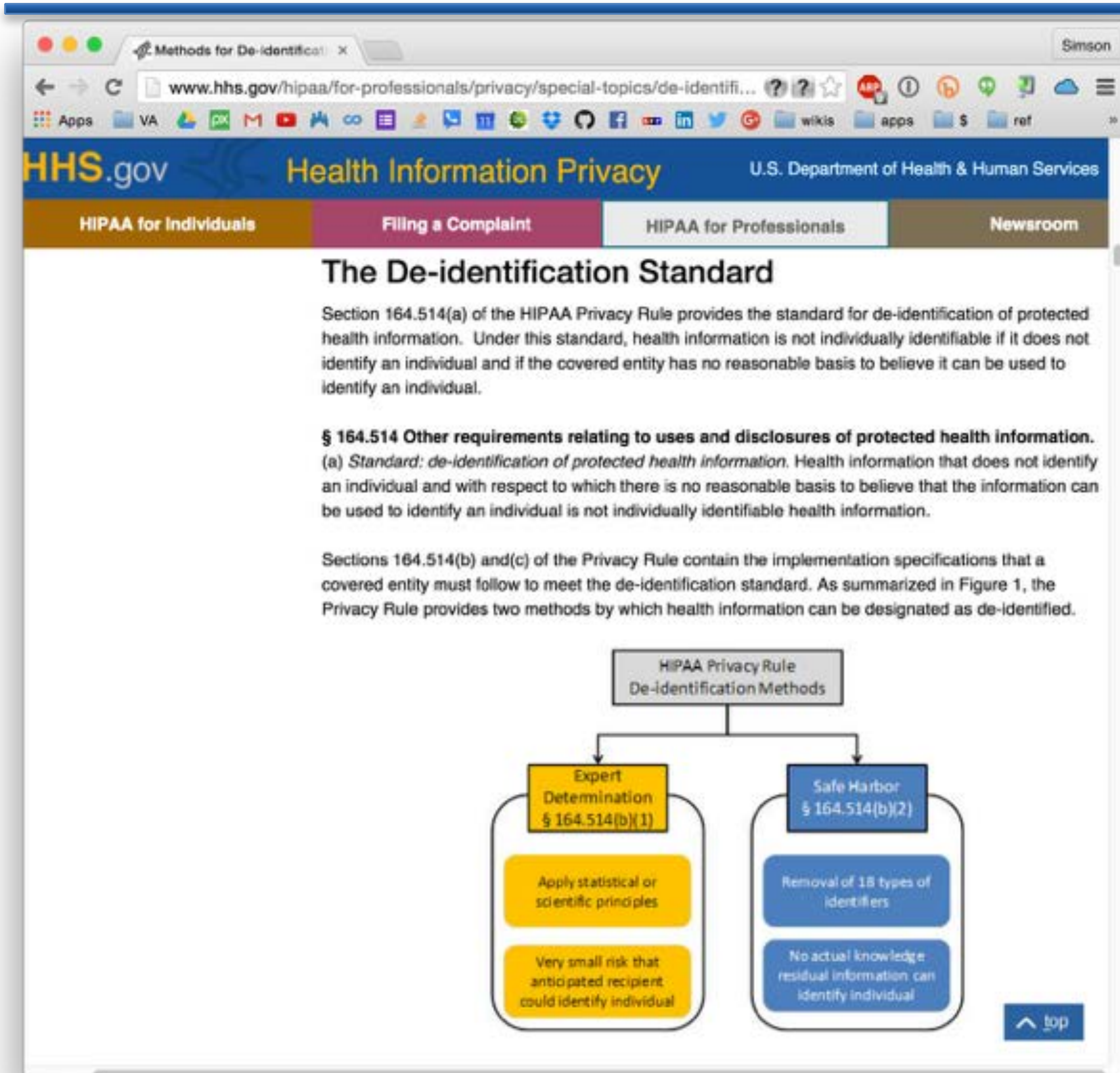
- Date of Birth
- Physical characteristics — height, weight, hair color, etc.
- History, capabilities, etc.

Options for quasi-identifiers:

- **Suppression** January 1, 1980 → XXXXXXXX, 1980
- **Generalization** January 1, 1980 → 1980-1985
- **Swapping** (between people) January 1, 1980 → February 29, 1984



The HIPAA Privacy Rule "Safe Harbor" method is largely based on Sweeney's findings.



HHS.gov
Health Information Privacy
www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/

HIPAA "Safe Harbor" rule:
Medical records are de-identified if 18 data elements are removed

- Must remove:
  - — *Names*
  - — *Geographic subdivisions smaller than a state, except first 3 digits of ZIP, provided the combined ZIP codes contain more than 20,000 people.*
  - — *Dates directly related to an individual (except for "age 90 or older")*
  - — *Individual numbers: phone, fax, SSN, medical record, account #s, etc.*
  - — *Email addresses, IP address, URLs*
  - — *Biometrics: fingerprints, voiceprints, photographs, etc.*
  - — *Any other uniquely identifying number, characteristic or code.*

- Estimated re-identification rate of this rule: 0.01% to 0.25%

HIPAA "Limited Datsets:"
Remove less information / More Useful / Restricted Use.

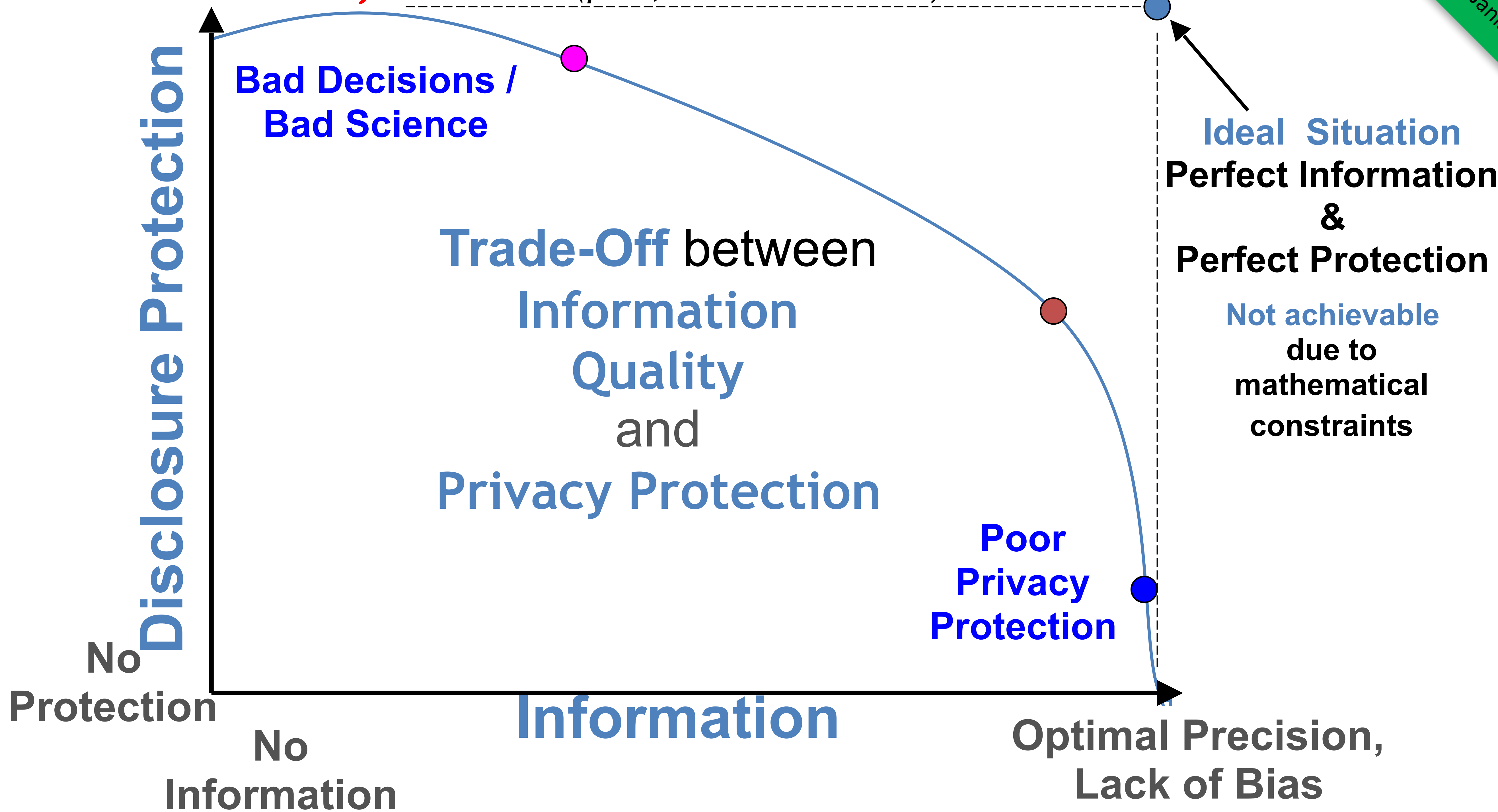
- The same as HIPAA Safe Harbor, except:
  - *Dates may remain (admission, discharge, service, DOB, DOD)*
  - *City, State, 5-digit ZIP code*
  - *Age in years, months, days, or hours*

- May be disclosed to an outside party:
  - *Without a patient's authorization or notification*
  - *But…*

- Must have a **data use agreement** in place:
  - *Cannot release the data set*
  - *Cannot share with others without a DUA*

The Inconvenient Truth:

“*De-identification leads to information loss which may limit the usefulness of the resulting health information*”*(p. 8, HIPAA Guidance)*



Outline for today's talk

- Why de-identify? ✔
- Basic de-identification ✔
- Famous re-identification controversies
- De-identification in practice
- Measuring re-identification risk
- De-identification governance
- De-identification @ NIST — Workshop June 29th

Direct Identifiers

Quasi-Identifiers

Field Suppression

Generalization

Data Swapping

Privacy-Utility tradeoff

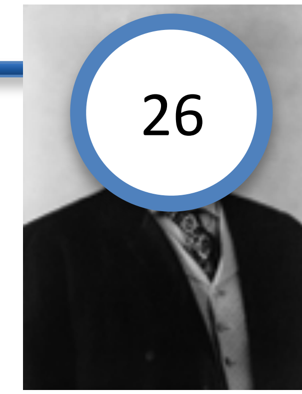
Identifying Quasi Identifiers!
The re-identification controversies.

Re-identification is called a "re-identification attack."

- *The person doing the re-identification is sometimes called a "data intruder."*
- *Motivations:*

Theodore Roosevelt

Re-identification is called a "re-identification attack."



- *The person doing the re-identification is sometimes called a "data intruder."*

- *Motivations:*

test the de-identification

Theodore
Roosevelt

Re-identification is called a "re-identification attack."

— *The person doing the re-identification is sometimes called a "data intruder."*

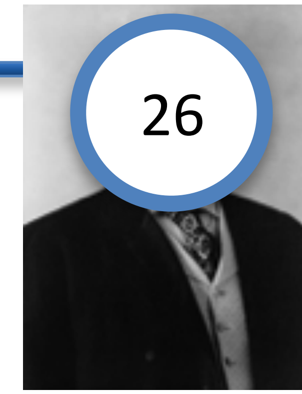
— *Motivations:*

test the de-identification

gain publicity or professional standing

Theodore Roosevelt

Re-identification is called a “re-identification attack.”



— *The person doing the re-identification is sometimes called a “data intruder.”*

— *Motivations:*

test the de-identification

Theodore Roosevelt

gain publicity or professional standing

Harm the data subject

Re-identification is called a “re-identification attack.”

— *The person doing the re-identification is sometimes called a “data intruder.”*

— *Motivations:*

test the de-identification

gain publicity or professional standing

Harm or embarrass the de-identifying organization

Theodore Roosevelt

Harm the data subject

Re-identification is called a "re-identification attack."

— *The person doing the re-identification is sometimes called a "data intruder."*

— *Motivations:*

Commercial Benefit

test the de-identification

Harm or embarrass the de-identifying organization

Theodore Roosevelt

gain publicity or professional standing

Harm the data subject

De-identified data can result in specific harms.

Identity disclosure

- The attacker can link de-identified data to an individual.
- Causes:
  - — *Insufficient de-identification (identifying information remains in the data set)*
  - — *Re-identification by linking*
  - — *Pseudonym reversal*

Attribute disclosure

- The dataset shows that all 20-year-old female patients from Q are left-handed.
  - — *Jane is a 20-year-old female patient from Q.*
  - — *∴Jane is left-handed.*

Inferential disclosure

- Data show correlation between home income and purchase price.
- Knowing Jane purchased a house for $X, we can infer Jane's household income.

De-identified data can result in specific harms.

Identity disclosure
- The attacker can link de-identified data to an individual.
- Causes:
  — *Insufficient de-identification (identifying information remains in the data set)*
  — *Re-identification by linking*
  — *Pseudonym reversal*

De-identification doesn't help against these disclosures

Attribute disclosure
- The dataset shows that all 20-year-old female patients from Q are left-handed.
  — *Jane is a 20-year-old female patient from Q.*
  — *∴Jane is left-handed.*

Inferential disclosure
- Data show correlation between home income and purchase price.
- Knowing Jane purchased a house for $X, we can infer Jane's household income.

Different "release models" can limit opportunities for re-identification.

Release and Forget model

- De-identification data are published on the Internet.

Data Use Agreement (DUA) model:

- Users assert that they will not attempt to re-identify.
- Required for HIPAA "limited" data sets.

Enclave model:

- Users get access to a computer that has the data.
- Users can run queries, but not download the data.

Since 2000, there have been several high-profile incidents in which publicly released de-identified data were re-identified.

Examples include:

- **AOL Search Data**
- **Netflix Prize**
- **Medical Tests**

Credit Card Transactions —

Mobility Traces

Taxi Ride Data —

The AOL Search Log Case of 2006

Goal: Support web information retrieval research

Name	Query	Date	Time
John Doe	Books	1/2/05	16:52
Bob Smith	Payscale	1/4/05	23:41
John Doe	Porn	1/8/05	03:15

The AOL Search Log Case of 2006

Goal: Support web information retrieval research

650k customers, 20 mil. queries, 3 mo. period

Name	Query	Date	Time
John Doe	Books	1/2/05	16:52
Bob Smith	Payscale	1/4/05	23:41
John Doe	Porn	1/8/05	03:15

The AOL Search Log Case of 2006

Goal: Support web information retrieval research

- 650k customers, 20 mil. queries, 3 mo. period
- Names replaced with persistent pseudonyms

Pseudonym	Name	Query	Date	Time
1	[redacted]	Books	1/2/05	16:52
2	[redacted]	Payscale	1/4/05	23:41
1	[redacted]	Porn	1/8/05	03:15

User Queries

User Queries

User 2178
foods to avoid when
breast feeding

User Queries

User 2178
foods to avoid when breast feeding

User 3482401
calorie counting

User Queries

User 3482401
calorie counting

User 2178
foods to avoid when breast feeding

User 3505202
depression and medical leave

User Queries

User 2178
foods to avoid when breast feeding

User 3505202
depression and medical leave

User 3482401
calorie counting

User 7268042
fear that spouse contemplating cheating

User Queries

User 2178
foods to avoid when breast feeding

User 3505202
depression and medical leave

User 47122
Child porno

User 3482401
calorie counting

User 7268042
fear that spouse contemplating cheating

User Queries

User 2178
foods to avoid when breast feeding

User 3505202
depression and medical leave

User 47122
Child porno

User 31350
How to kill oneself with gas

User 3482401
calorie counting

User 7268042
fear that spouse contemplating cheating

User Queries

User 2178
foods to avoid when breast feeding

User 3505202
depression and medical leave

User 47122
Child porno

User 31350
How to kill oneself with gas

User 3482401
calorie counting

User 7268042
fear that spouse contemplating cheating

User 3483689
Time after time

User Queries

User 2178
foods to avoid when breast feeding

User 3505202
depression and medical leave

User 47122
Child porno

User 31350
How to kill oneself with gas

User 3482401
calorie counting

User 7268042
fear that spouse contemplating cheating

User 3483689
Time after time

User 3483689
Wind beneath my wings

User 4417749 issued hundreds of searches

Barbaro & Zeller. A face exposed for AOL searcher no. 4417749. New York Times. Aug 9, 2006.

User 4417749 issued hundreds of searches

Numb fingers

60 single men

Last name = “Arnold”

Dog that urinates on everything

Hand tremors

Landscapers in Lilburn (Georgia)

Homes sold in shadow lack subdivision gwinnett county georgia

Dry mouth

Nicotine effects on the body

bipolar

Barbaro & Zeller. A face exposed for AOL searcher no. 4417749. New York Times. Aug 9, 2006.

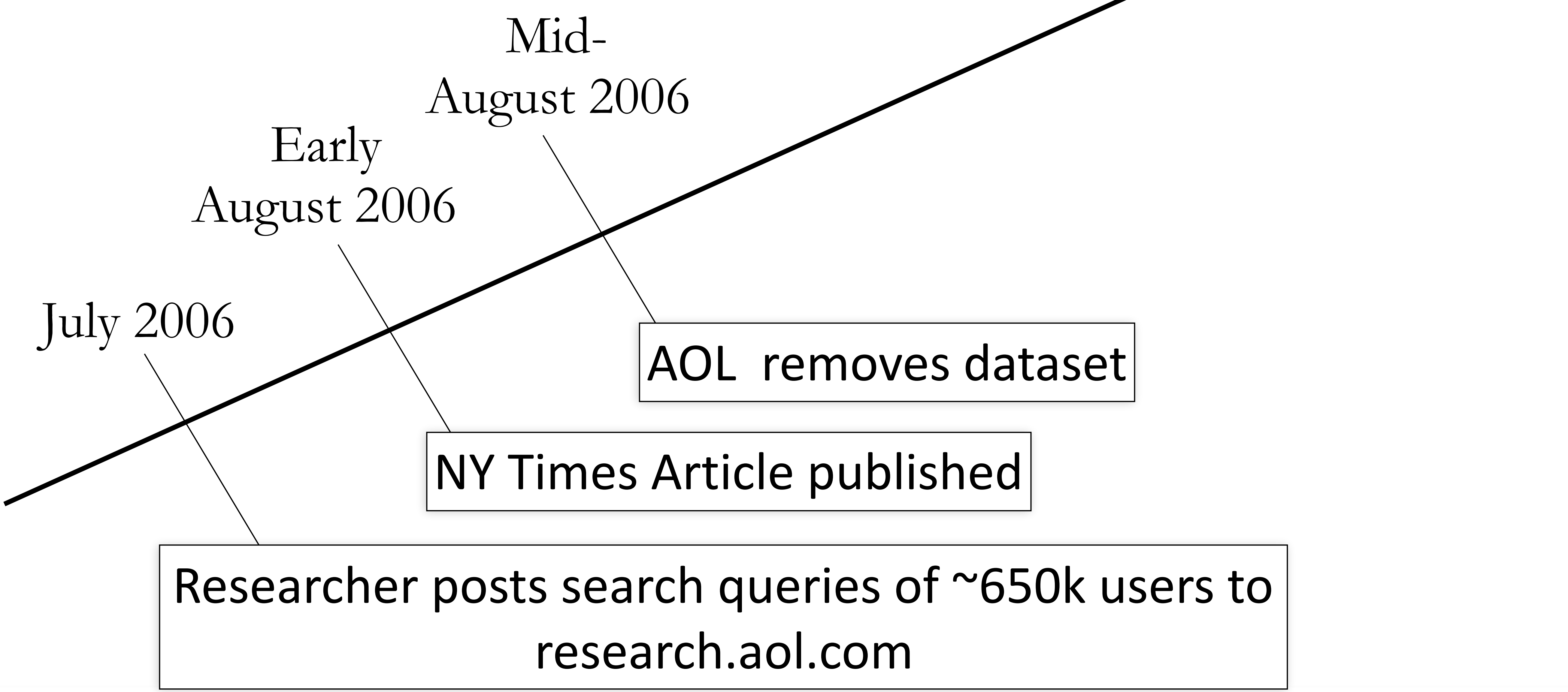
User 4417749 issued hundreds of searches

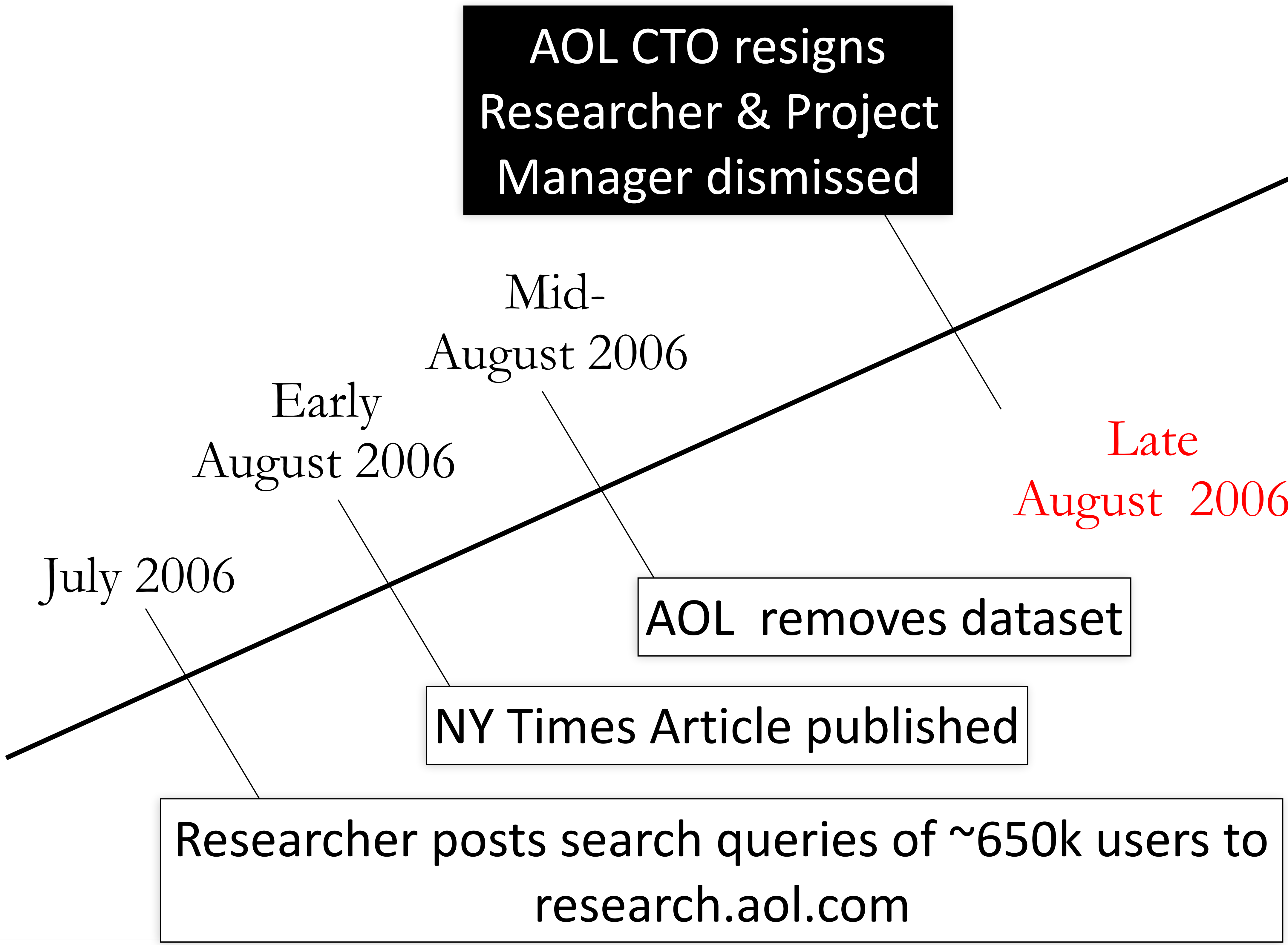
Nu

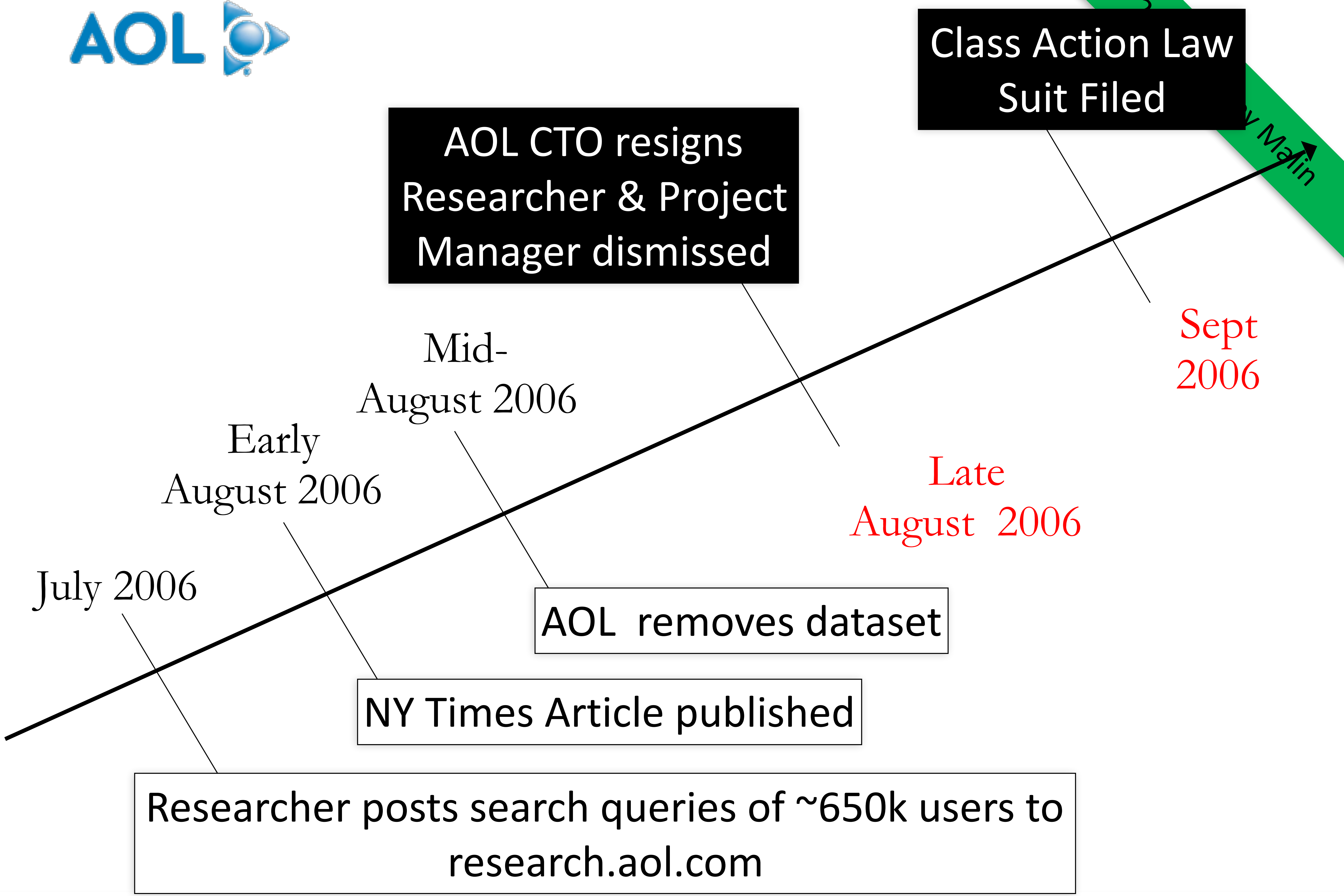
rs

Ho

Thelma Arnold & Dudley


Mid-
August 2006
Early
August 2006
July 2006
AOL removes dataset
NY Times Article published
Researcher posts search queries of ~650k users to research.aol.com

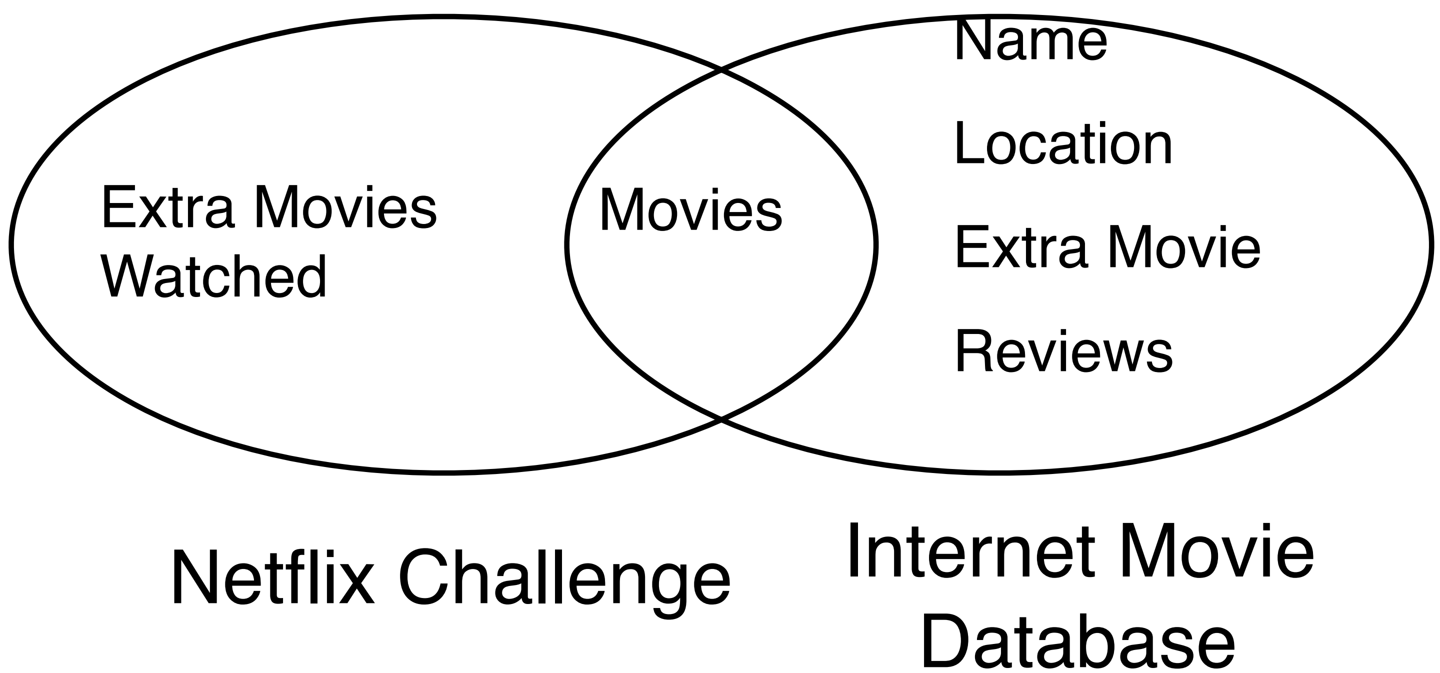

July 2006
Researcher posts search queries of ~650k users to research.aol.com
Early August 2006
NY Times Article published
Mid-August 2006
AOL removes dataset
Late August 2006
AOL CTO resigns
Researcher & Project Manager dismissed


Class Action Law Suit Filed
AOL CTO resigns
Researcher & Project Manager dismissed
Mid-August 2006
Early August 2006
Sept 2006
Late August 2006
July 2006
AOL removes dataset
NY Times Article published
Researcher posts search queries of ~650k users to research.aol.com

The Netflix Challenge (2008-2009)

Netflix published movie selections of ~450,000 pseudonymized subscribers

Re-identification via uniqueness of movie combinations



A. Narayanan & V. Shmatikov. IEEE Security and Privacy Conference. 2008.

Netflix Prize Reidentification

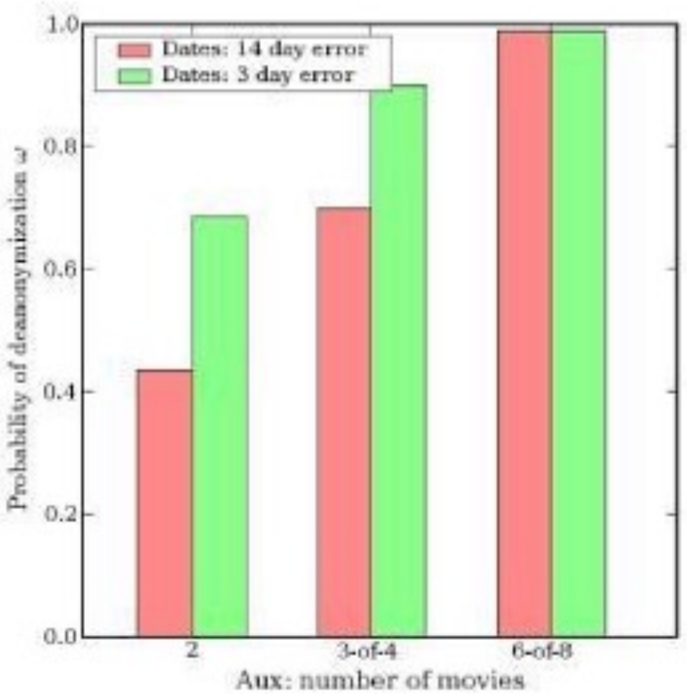


Figure 4. Adversary knows exact ratings and approximate dates.

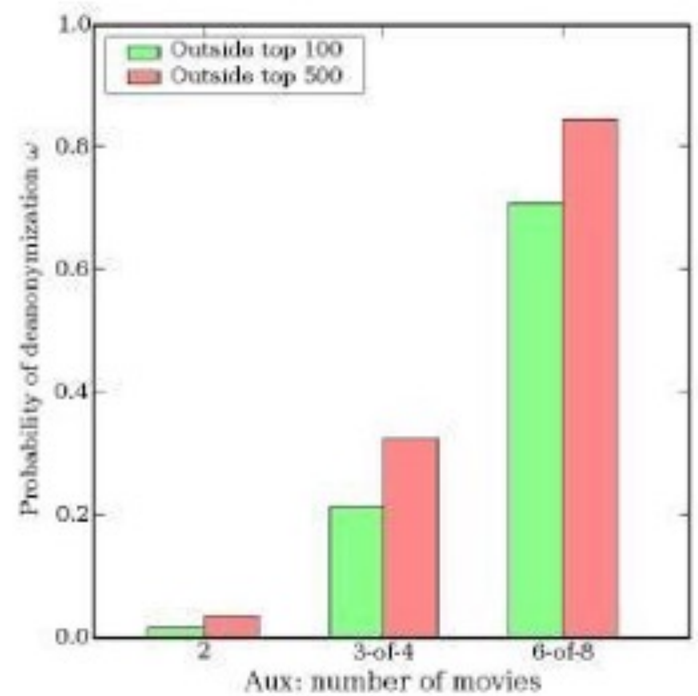


Figure 8. Adversary knows exact ratings but does not know dates at all.

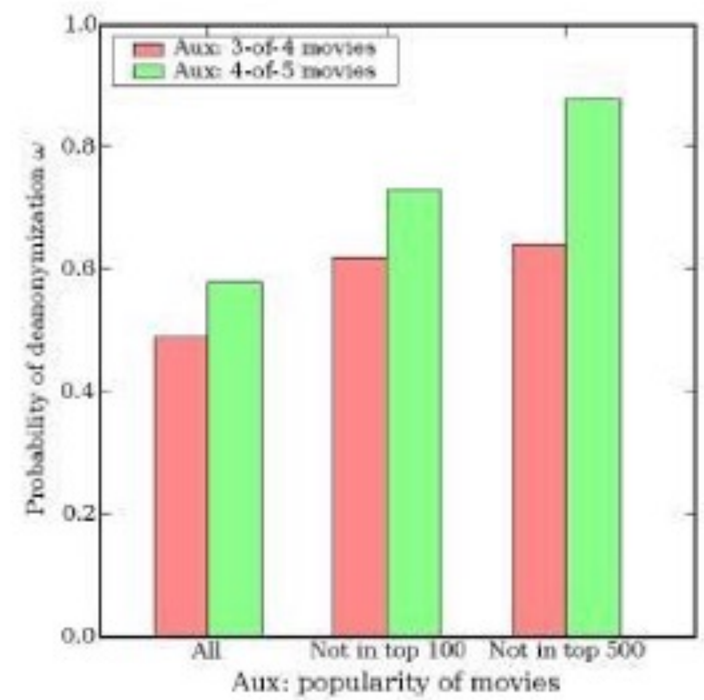


Figure 9. Effect of knowing less popular movies rated by victim. Adversary knows approximate ratings (± 1) and dates (14-day error).

Forbes.com

U.S. EUROPE ASIA

Home Lists Business Tech

Breakth

Welcome Google User

Here are more stories related to your searc

- Netflix Settles Privacy Lawsuit, Canc

See all related stories >

The Firewall

Filtering ideas in the world of security.

Netflix Settles Privacy Lawsuit, Cancels Prize Sequel

March 12, 2010 - 12:35 pm

Taylor Buley Bio | Email

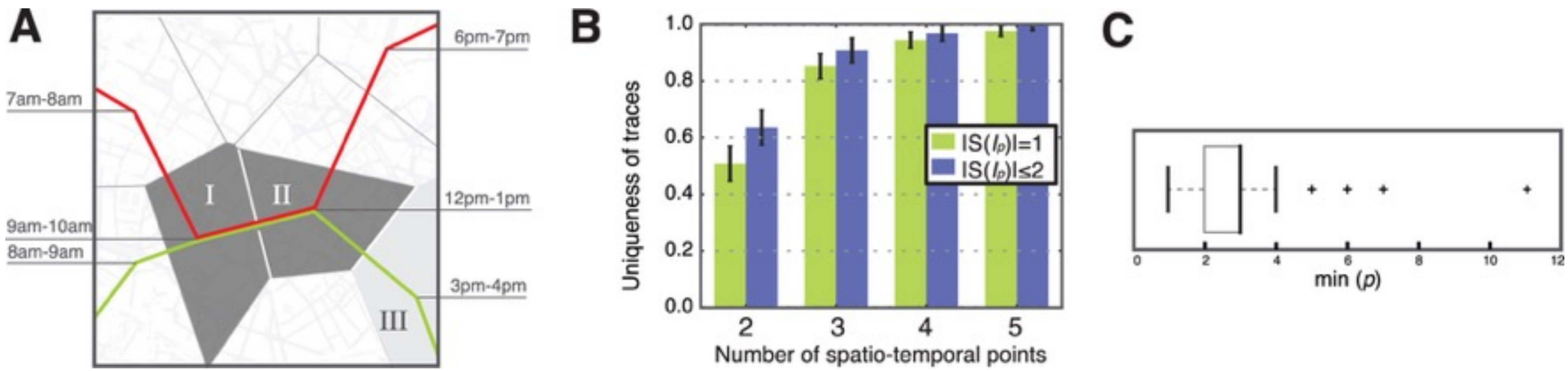
Taylor Buley is a staff writer and editorial developer for Forbes

Share 8

67 retweet

On Friday, Netflix announced on its corporate blog that it has settled a lawsuit related to its Netflix Prize, a $1 million contest that challenged machine learning experts to use Netflix's data to produce better recommendations than the movie giant could serve up themselves.

Cell Phone Location:
4 “spatio-temporal points” uniquely identifies a user in the data set.



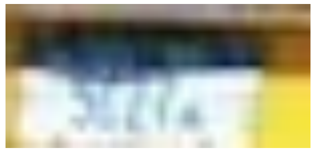
Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen & Vincent D. Blondel,
Unique in the Crowd: The Privacy Bounds of Human Mobility,
NATURE SCIENTIFIC REPORTS, *Oct. 1, 2012.*

Re-identification by flickr:
2014 NYC Taxi Ride data, NYC Taxi and Licensing Commission

In 2014, NYC TLC released taxi ride dataset with the “MD5” of each taxi as a pseudonym

- MD5(“5C27”) = “0f76c35d4a069e0fe76b21d28f009639”
- Every taxi identifiable with a brute force search

An intern at Neustar re-identified 2 rides by searching for photos for taxi licenses and matching MD5 codes and times.



“5C27”

A journalist at Gawker identified 9 other cab rides.

Broken Promises of Privacy

57 UCLA Law Review 1701 (2010)

Underlying theory hasn't changed: intuitions were off.

- Intuitions are still catching up.

Data can be either useful or perfectly anonymous but never both.

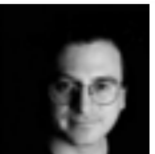
Every privacy law ever written must be rewritten.

Accretion and the database of ruin

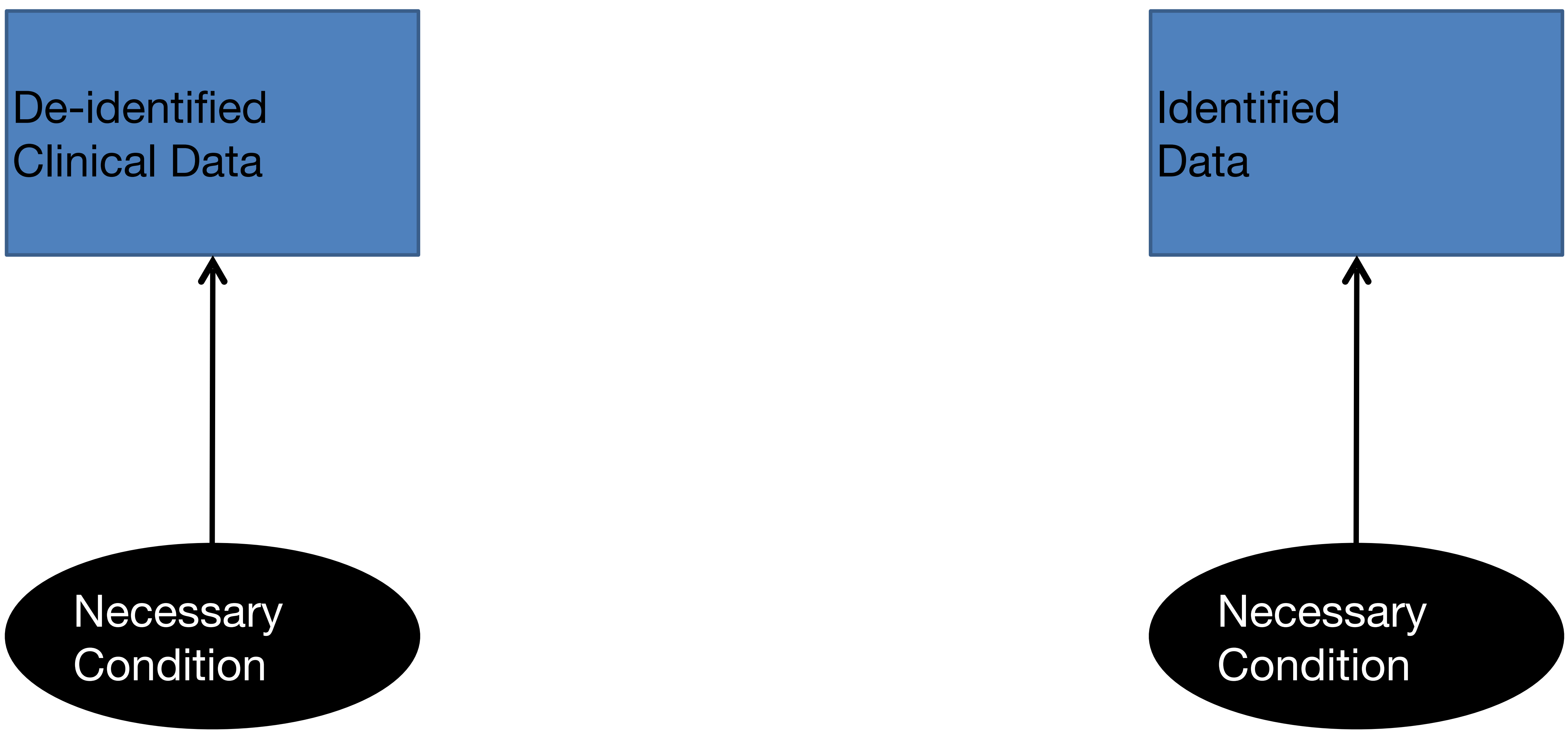
DISTINGUISHABLE

≠

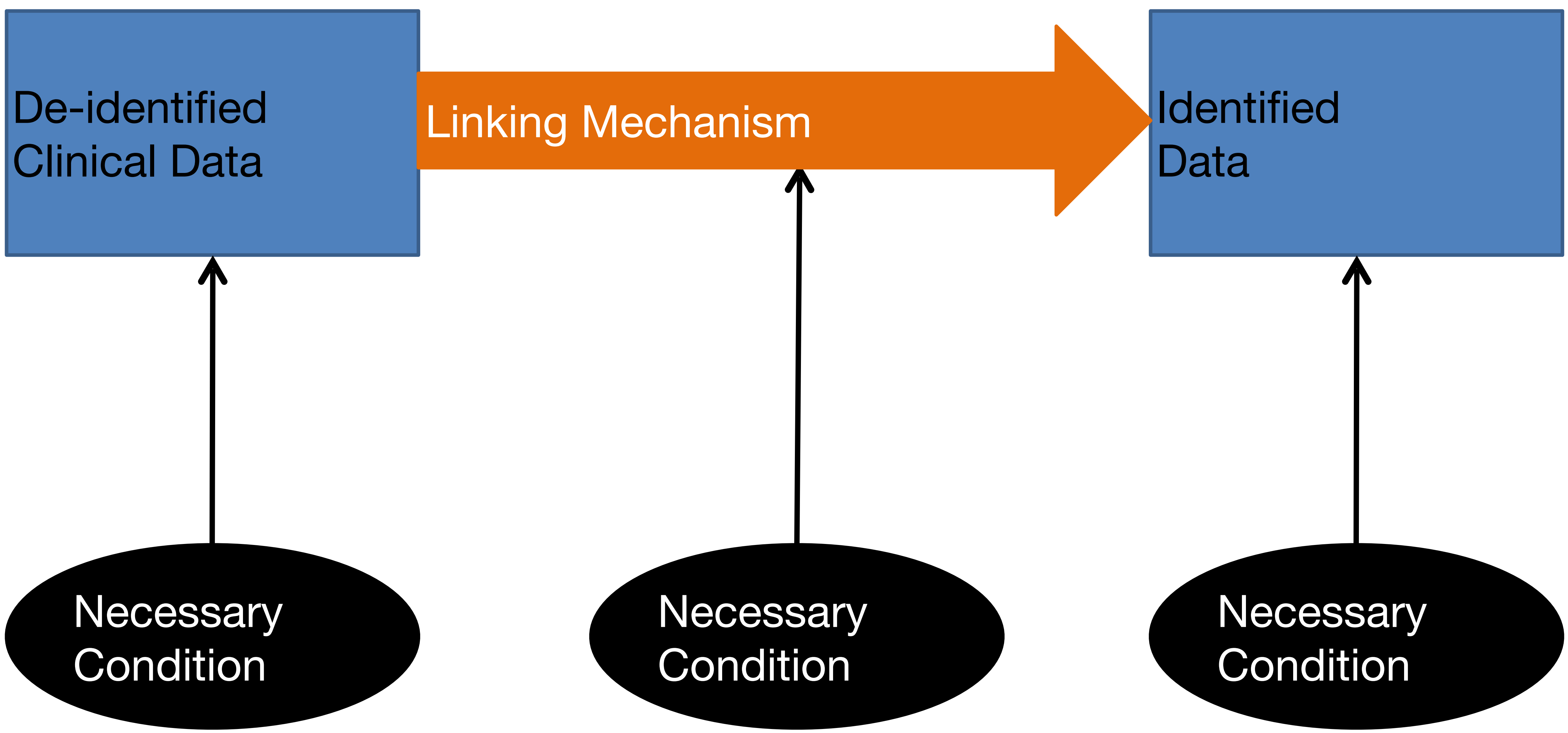
IDENTIFIABLE



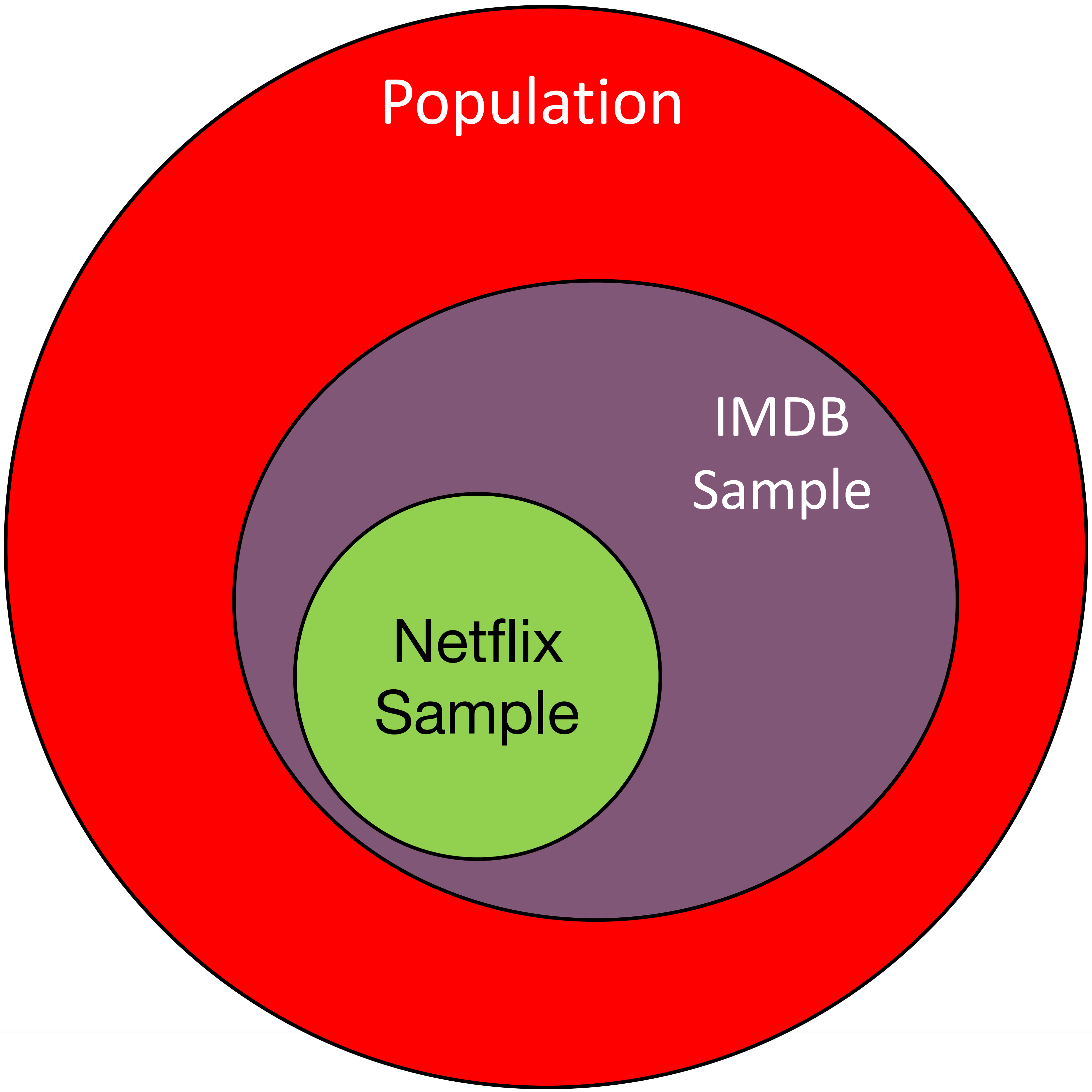
Central Dogma of Re-identification



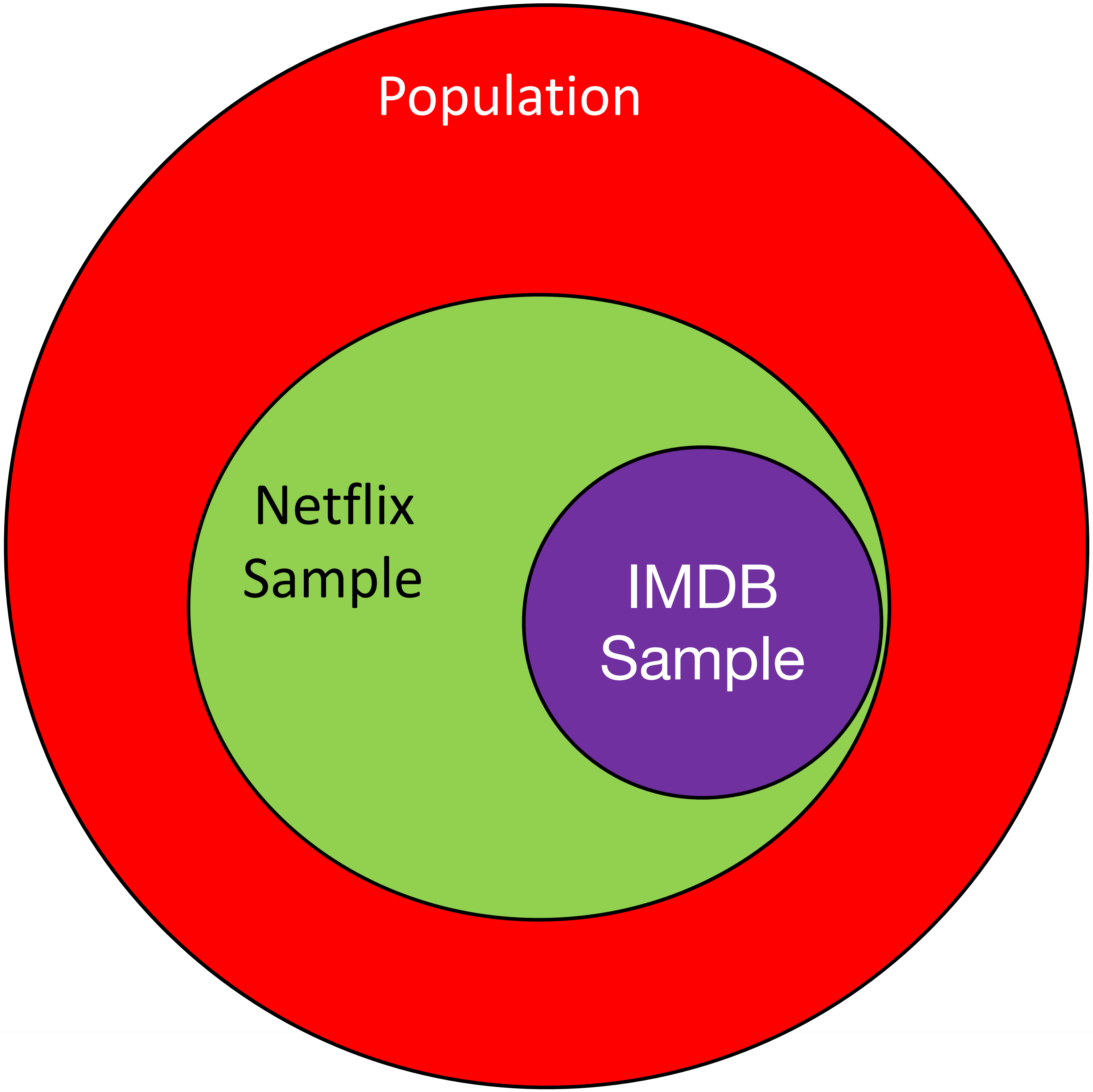
Central Dogma of Re-identification



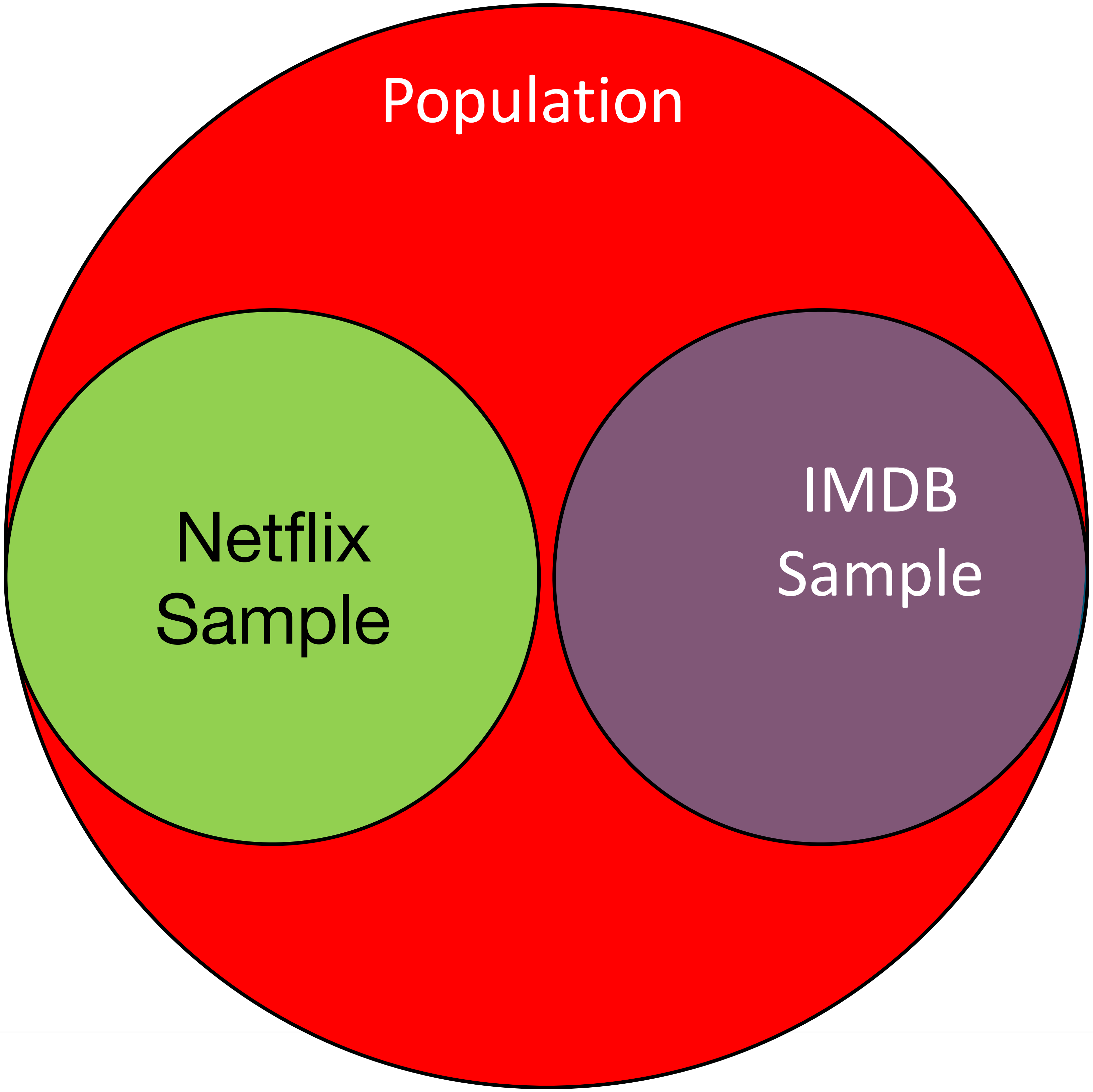
Re-identification ?



Re-identification ?



Re-identification ?



Linking is more complex than it seems!

In order to be 100% linked:

- The person must be present in both data sets.
- The person's records must be "unique" in both data sets.

How "unique" are birthday, sex & ZIP?

- Sweeney estimated 87% of the US population are uniquely distinguished using 1990 Census data.
- Golle computed a 62% re-identification rate using 2000 Census data.
- But **only 55% of Cambridge population was registered to vote in 1996-1997** (Barth-Jones)
  — *So only 55% of Cambridge voters could be identified using voter registration records.*

The Vigorous Debate

Jane Yakowitz, *Tragedy of the Data Commons*, 25 HARV. J.L. & TECH. 1 (2011).

- Data ages
- Reidentification is Hard

Daniel Barth-Jones, *The 'Re-Identification' of Governor William Weld's Medical Information* (working paper).

- Doubt about completeness of the two data sets

Paul M. Schwartz, & Daniel J. Solove, *The PII Problem: Privacy and a New Concept of Personally Identifiable Information*, 86 N.Y.U. LAW REVIEW 1814 (2011).

- Can't just abandon PII
- Seeking half-measures

The Crux of the Debate

Who is making the right predictions about the rate of change of

- Computational power
- Auxiliary information?

Is “statistical breach” a privacy harm /problem?

- This person has a 1/1000 risk of rare disease X unlike the member of the general population with a 0.0000001 risk.

Is perfect de-identification necessary?

Tech Responses: Bad

"Felten's third law" —

- "In technology policy debates, lawyers put too much faith in technical solutions, while technologists put too much faith in legal solutions"

Head in the sand

- "good anonymization" versus "bad anonymization"
- Removing "identifiers"

Tech Responses: Better

Modeling the Risk of Reidentification

- Adversary: Incentives? Time? Resources?
- Auxiliary Information: Reasonably accessible? All possible? Created in the future?
- Organizational controls: trust, audit, security

Mathematically model the degree of de-identification.

Concrete Bottom Line:

- Public release is the worst
- Risk factors at peak
- Controlling risk with data use agreements

Outline for today's talk

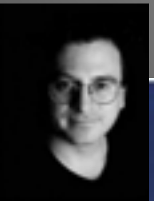
- Why de-identify? ✔
- Basic de-identification ✔
- Famous re-identification controversies ✔
- De-identification in practice
- Measuring re-identification risk
- De-identification governance
- De-identification @ NIST — Workshop June 29th

High-profile re-identifications

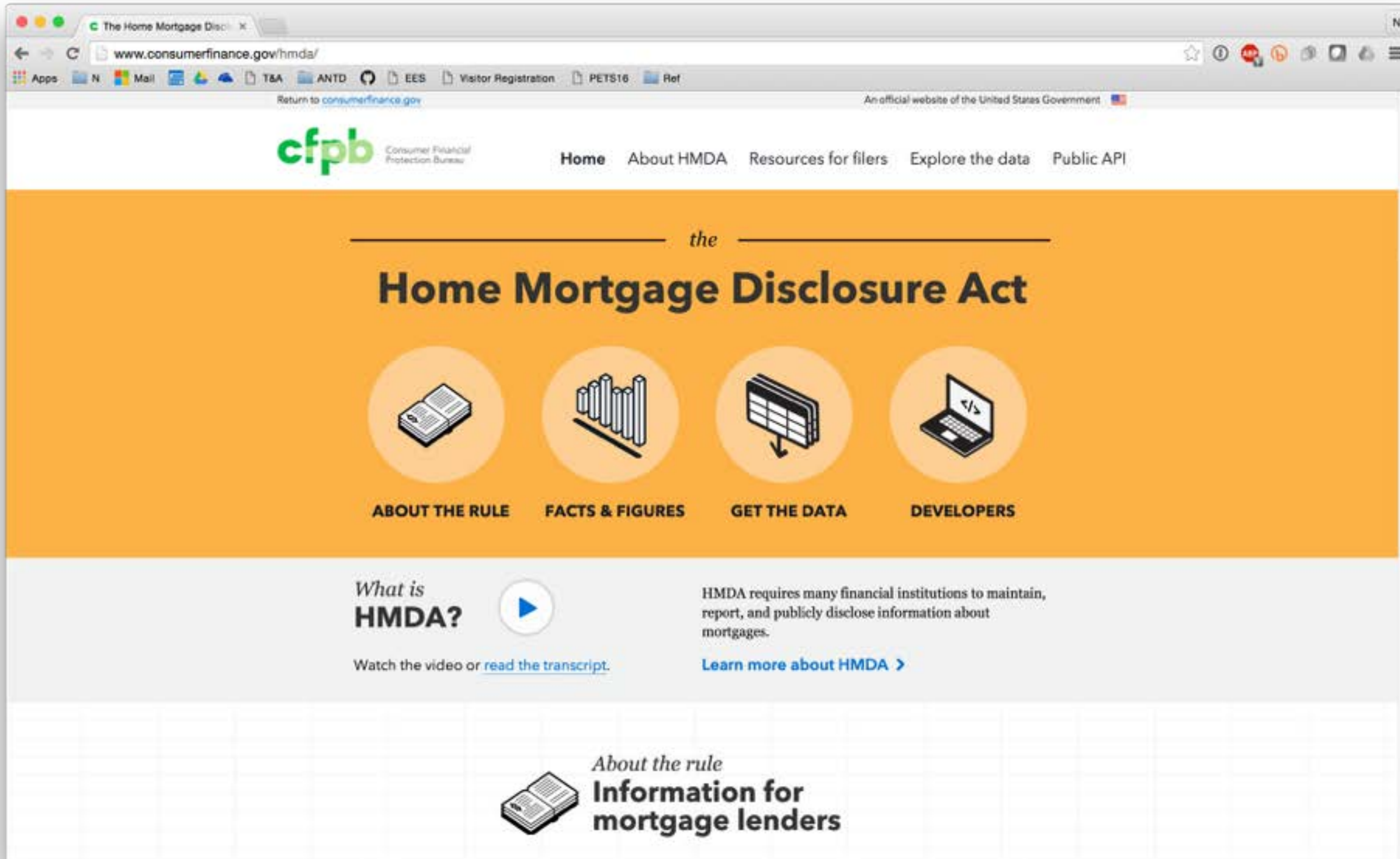
The number of people re-identified was relatively small

Disproportional impact.

De-ID Today



De-identification today:
Consumer Financial Protection Board HMDA



Explore the data

CUSTOM DATASETS SUMMARY TABLES

Important note: Please use caution when analyzing Metropolitan Statistical Areas (MSAs) over multiple years, as the 2014 HMDA data use the updated MSA definitions, released Feb 2013. For example, some MSAs may show the same name and code number in 2014 as previous years, even though the underlying geography has changed.

Filter the data

Select year(s) of data: 2012 ×

Select suggested filters: Select a filter set ▾ ?

Want something more specific? Modify your filters below or download now. [Or start over.]

LOCATION State, metro area, county, and census tract of the property

State: Virginia × ▾ - *or* - **Metro Area:** Select an MSA/MD ▾

County: Arlington County ×

Census tract: 1014.02 ×

PROPERTY Property type and occupancy

Property Type:

- [x] One-to-four family dwelling (other than manufactured housing)
- [x] Manufactured housing
- [ ] Multifamily dwelling

Will the owner use the property as their primary residence?

- [x] Owner-occupied as a principal dwelling
- [ ] Not owner-occupied as a principal dwelling
- [ ] Not applicable

LOAN Loan action, purpose, type, and more

What action was taken on the loan or application?

Loan originated ✖

What is the loan being used for?

- [x] Home purchase
- [ ] Home improvement
- [ ] Refinancing

What type of loan is it?

- [ ] Conventional
- [ ] FHA-insured
- [ ] VA-guaranteed
- [ ] FSA/RHS-guaranteed

What is the loan's lien status?

- [x] Secured by a first lien
- [ ] Secured by a subordinate lien
- [ ] Not secured by a lien
- [ ] Not applicable (purchased loans)

APPLICANT Demographic information for applicants and co-applicants

Applicant Sex: ☑ Male ☐ Female ☐ Not provided ☐ Not applicable

Applicant Race: Select an Applicant R

Applicant Ethnicity: Select an Applicant Ethnicity

Applicant Income: $ Min. ,000 to $ Max. ,000

Show co-applicant filters? Yes No

NEED MORE INSIGHT?

Compare your filtered data across state, loan type, applicant race, and more with a custom summary table.

Create a summary table ›

Preview the results

There are **32** HMDA records from **2012** with the above selected filters.

Preview the first 100 rows

Download raw data

File format:

Spreadsheet (CSV)

Include labels Include labels and codes

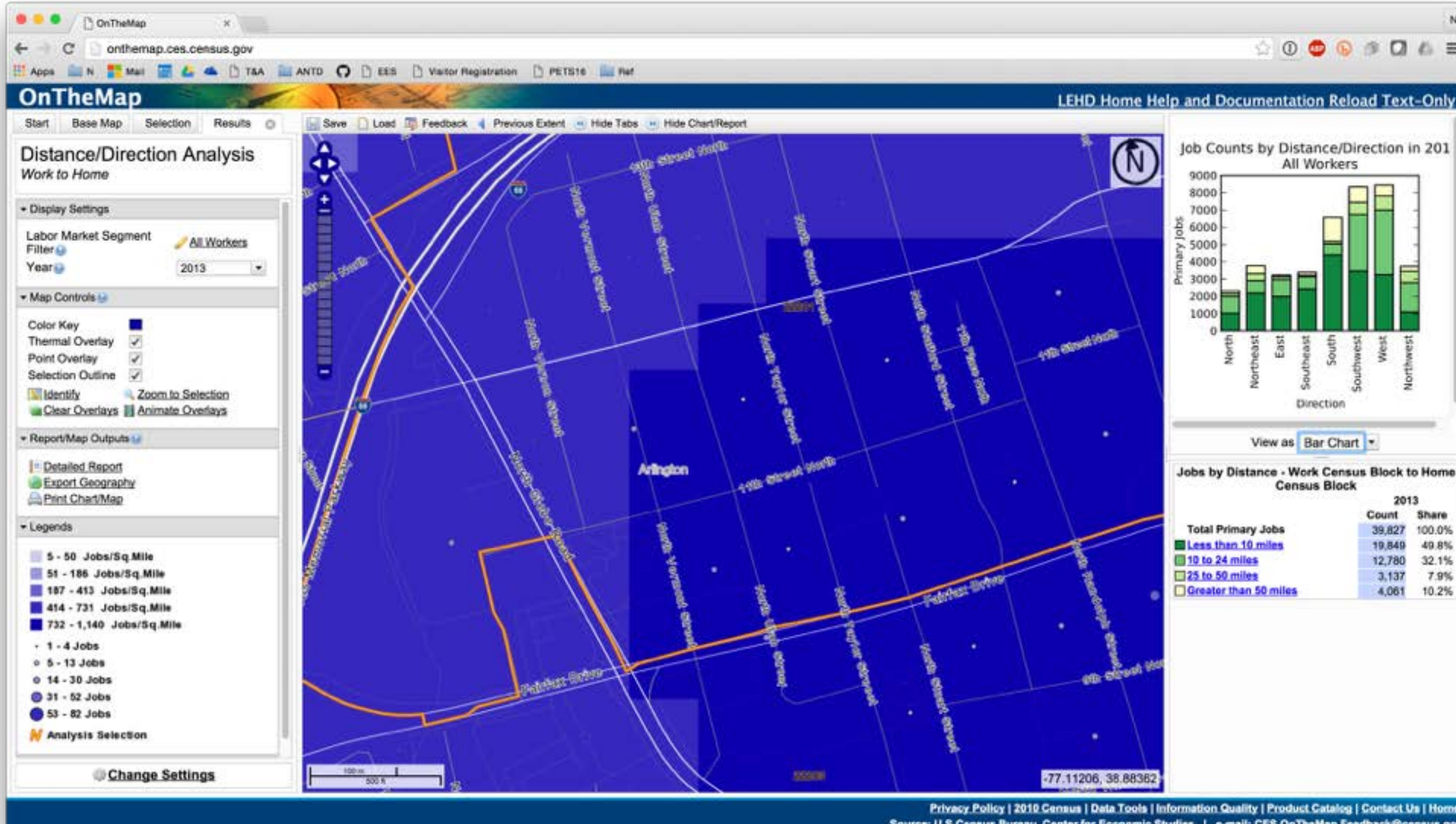
Save & share your work

Save your filters, or share them with a link:

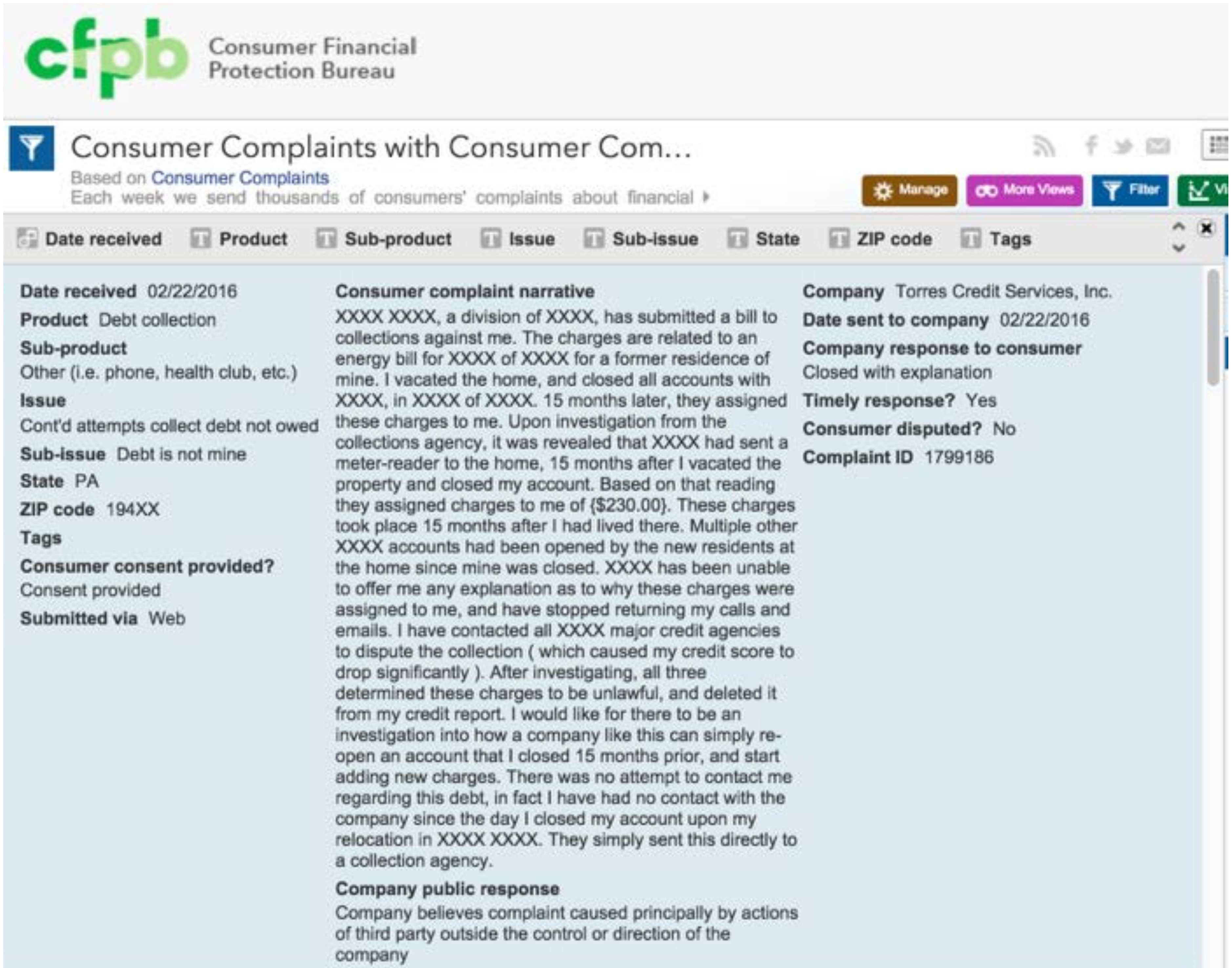
http://www.consumerfinance.gov/hmda/explore#!/as_of_y

loan_amount_000s	co_applicant_sex_name	applicant_race_name_1	applicant_ethnicity_name	co_applicant_race_name_1	co_applicant_ethnicity_name
215	No co-applicant	White	Not Hispanic or Latino	No co-applicant	No co-applicant
225	No co-applicant	White	Not Hispanic or Latino	No co-applicant	No co-applicant
266	No co-applicant	White	Not Hispanic or Latino	No co-applicant	No co-applicant
320	No co-applicant	White	Not Hispanic or Latino	No co-applicant	No co-applicant
335	No co-applicant	White	Not Hispanic or Latino	No co-applicant	No co-applicant
342	No co-applicant	White	Not Hispanic or Latino	No co-applicant	No co-applicant
352	Female				
355	No co-applicant	White	Not Hispanic or Latino	No co-applicant	No co-applicant
355	No co-applicant			No co-applicant	No co-applicant
382	Male	White	Not Hispanic or Latino	White	Not Hispanic or Latino
399	No co-applicant	White	Not Hispanic or Latino	No co-applicant	No co-applicant
400	No co-applicant	White	Not Hispanic or Latino	No co-applicant	No co-applicant
404	No co-applicant	White	Not Hispanic or Latino	No co-applicant	No co-applicant
404	Female	White	Not Hispanic or Latino	White	Not Hispanic or Latino
404	Female	White	Not Hispanic or Latino	White	Not Hispanic or Latino
413	No co-applicant	White	Not Hispanic or Latino	No co-applicant	No co-applicant
416	No co-applicant	Asian	Not Hispanic or Latino	No co-applicant	No co-applicant
417	Female	White	Not Hispanic or Latino	White	Not Hispanic or Latino
417	Male	White	Not Hispanic or Latino	White	Not Hispanic or Latino
428	Female	White	Not Hispanic or Latino	White	Not Hispanic or Latino
444	Female				
450	No co-applicant	Asian	Not Hispanic or Latino	No co-applicant	No co-applicant
464	Female				
477	Female	White	Not Hispanic or Latino	White	Not Hispanic or Latino
486	No co-applicant	White	Not Hispanic or Latino	No co-applicant	No co-applicant
511	Female				
560	Female	White	Not Hispanic or Latino		
588	Female	White	Not Hispanic or Latino	White	Not Hispanic or Latino
604	Female				
618	Female				
634	No co-applicant	White	Not Hispanic or Latino	No co-applicant	No co-applicant
1080	Female	Asian	Not Hispanic or Latino	White	Not Hispanic or Latino

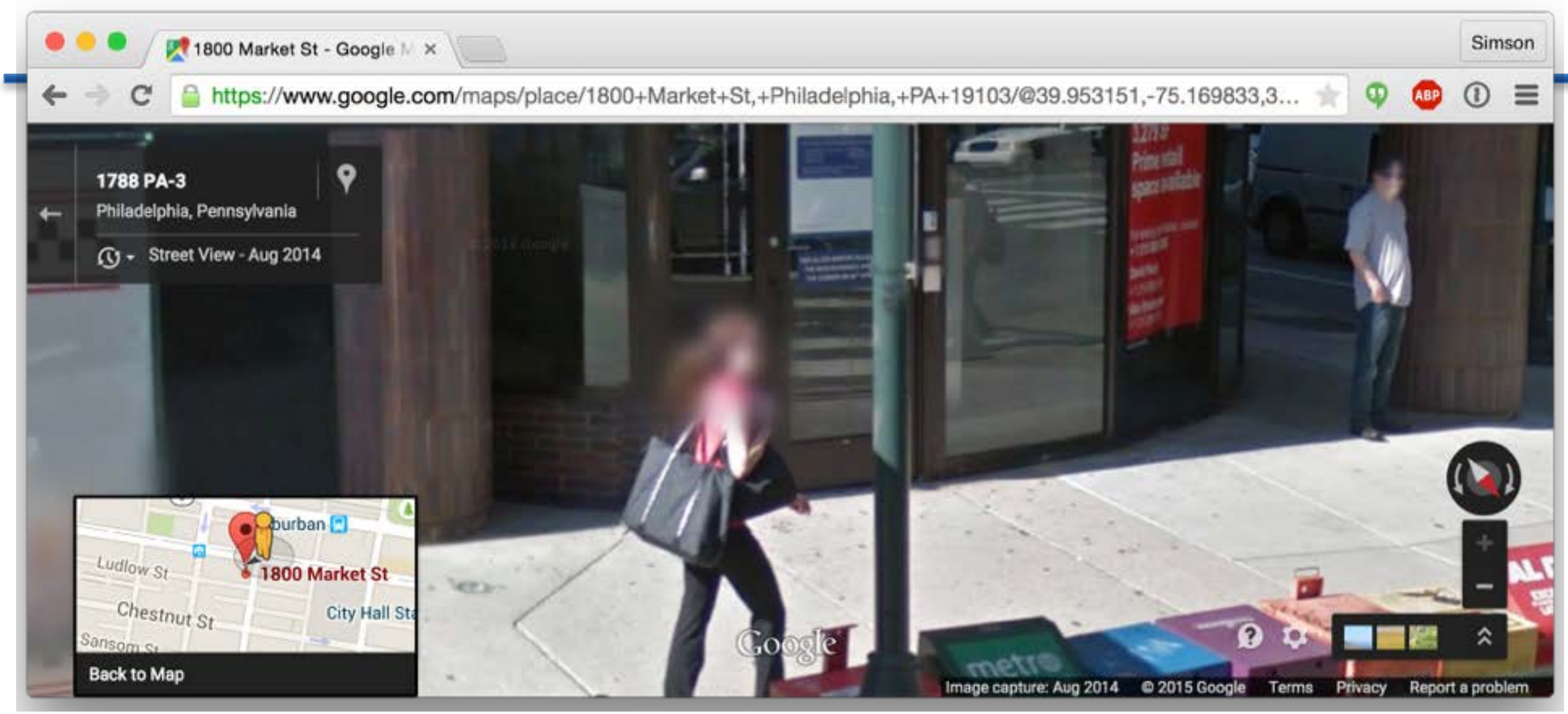
De-identification is being used today:
OnTheMap (Census) — Synthetic Data



De-identification today:
Consumer Complaint Database



Google Street View — faces and license plates



"Large-scale Privacy Protection in Google Street View," Frome et al, 2009

Google claims 90% of faces & 95% of license plates through automated processing.

Multimedia de-identification / redaction

Public release of police body cameras:

http://www.cam.ac.uk/research/news/first-scientific-report-shows-police-body-worn-cameras-can-prevent-unacceptable-use-of-force

Other uses:

- Scientific research; privacy preserving surveillance; data retention

De-identified health datasets are widely distributed. Are they vulnerable?

"A Systematic Review of Re-Identification Attacks on Health Data," El Emam et al, 2011. PLOS One.

Findings:

1. 14 published attacks
2. Few attacks involved health data
3. Most adversaries were researchers
4. Most re-identification attacks were in the US
5. Most re-identification attacks were verified
6. Most re-identified data was not de-identified according to existing standards.

http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0028071

Table 2. A summary of successful re-identification attacks on the evaluation criteria.

ID	Study	Pub Year $	Health data included?	Profession of adversary	Number of individuals re-identified	Country of adversary	Proper de-identification of attacked data ?	Re-identification verified ?
A	[70]	2001	No	Researchers	29 of 273	Germany	"Factually anonymous"	Yes (records containing insurance numbers only)
B	[71]	2001	No	Researchers	75% of 11,000	USA	Direct identifiers removed	No
C	[67]	2002	Yes	Researcher	1 of 135,000	USA	Removal of names and addresses	Yes
	[56]	2003	No	Researchers	219 unique matches, 112 with 2 possibilities, 8 confirmed	UK	Yes	Verified matches, but not identities
D	[22]	2006	No	Journalist	1 of 657,000	USA	No	Yes (with individual)
E	[72]	2006	Yes	Researchers	79% of 550	USA	No	Verified (with original data set)
	[73]	2006	No	Researchers	Of 133 users, 60% of those who mention at least 8 movies	USA	Direct identifiers removed	No
F	[52]	2006	Yes	Expert Witness	18 of 20	USA	Only type of cancer, zip code and date of diagnosis included in request	Yes (verified by the Department of Health)
G	[74]	2007	No	Researchers	2,400 of 4.4 million	USA	Identifying information removed	Verified using original data
	[53]	2007	Yes	Broadcaster	1	Canada	Direct Identifiers removed & possibly other unknown de-id methods used	Yes
H	[23]	2008	No	Researchers	2 of 50	USA	Direct identifiers removed+maybe perturbation	No
I	[75]	2009	Yes	Researcher	1 of 3,510	Canada	Direct identifiers removed	Yes
J	[76]	2009	No	Researchers	30.8% of 150 pairs of nodes	USA	Identifying information removed	Verified using ground-truth mapping of the 2 networks
K	[57,58]???	2010	Yes	Researchers	2 of 15,000	USA	Yes - HIPAA Safe Harbor	Yes

($This is the first year that the report or article appears. Some of the reports we cite have been updated at later dates. Some reports describe re-identification attacks that may have occurred in earlier years. ☧ Since the appearance of the original results in 2010 a second article has been published more recently).

doi:10.1371/journal.pone.0028071.t002

El Emam K, Jonker E, Arbuckle L, Malin B (2011) A Systematic Review of Re-Identification Attacks on Health Data. PLoS ONE 6(12): e28071. doi: 10.1371/journal.pone.0028071
http://journals.plos.org/plosone/article?id=info:doi/10.1371/journal.pone.0028071

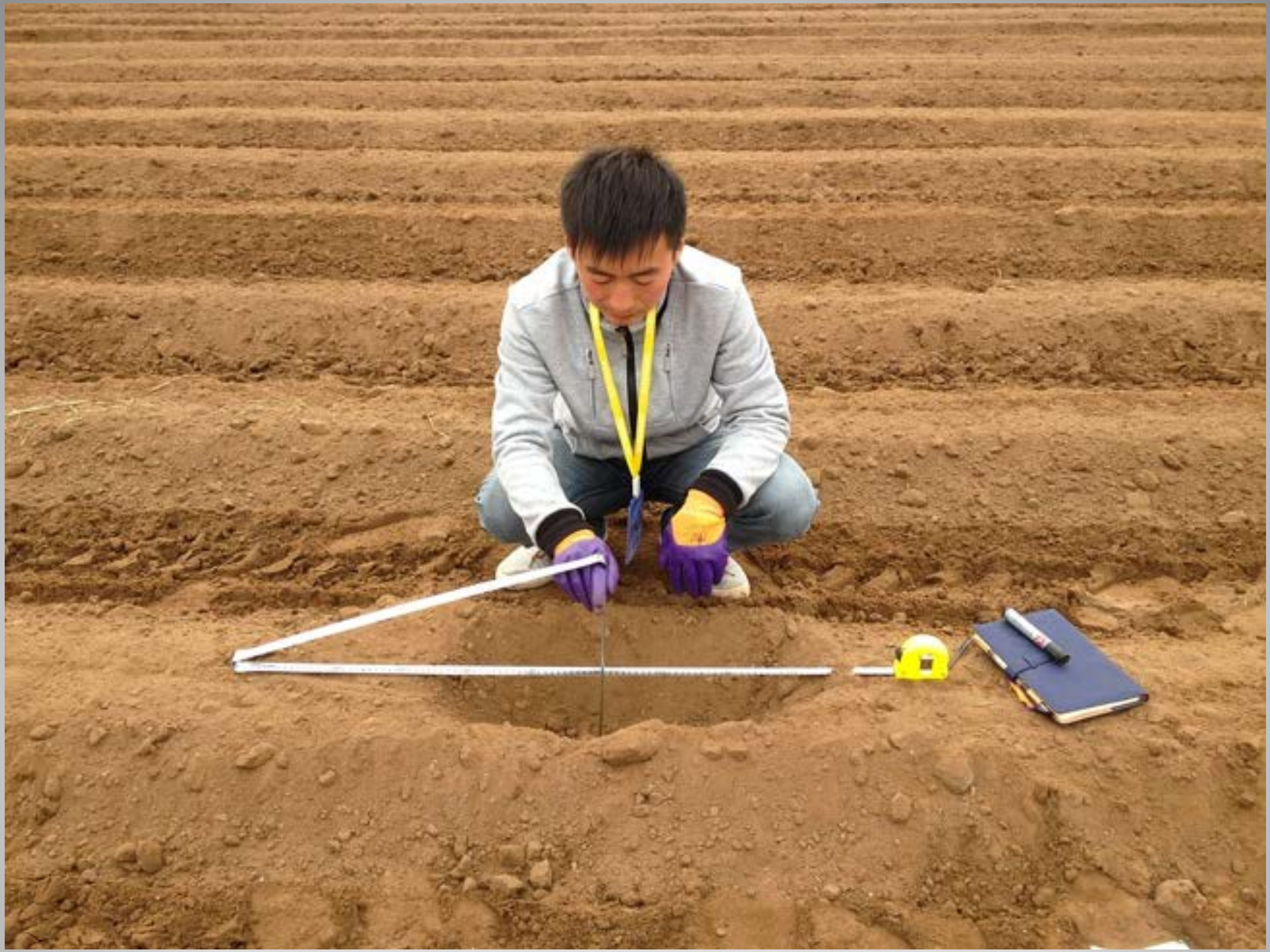
Outline for today's talk

- Why de-identify? ✔
- Basic de-identification ✔
- Famous re-identification controversies ✔
- De-identification in practice ✔
- Measuring re-identification risk
- De-identification governance
- De-identification @ NIST — Workshop June 29th

De-identification is used today.

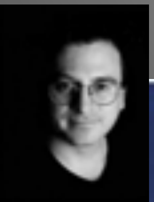
Most published re-identification has been done by researchers.

Re-identification rates are low, but larger than 0



https://pixabay.com/en/measuring-land-character-792513/

Measuring Re-Identification Risk



“Re-identification risk:”

the risk that the **suppressed identifiers** can be learned from the de-identified data.

Various approaches for computing and reporting re-identification risk.

- **Prosecutor Scenario:** Risk that a specific person can be re-identified when the attacker knows the are in the data set.
- **Journalist Scenario:** Risk that at least one person can be re-identified.
- **Marketer Scenario:** The percentage of identities that can be correctly re-identified.
  - The “Class Action Scenario” — Malin

Re-identification risk needs to take into account the ability and resources of the data intruder.

General public — anyone who has access to the data.

Expert — A computer scientist skilled in re-identification.

Insider — A member of the organization that produced the dataset.

Insider Recipient — A member of the organization that received the data and has more background information than the general public.

Information broker — An organization that systematically collects both identified and de-identified information to re-identify.

Nosy Neighbor — Friend or family member with specific info.

K-Anonyminity: A model for re-identification

A dataset that you would like to release:

Race	Birthdate	Sex	Zip	Medication	Diagnosis
Black	9/20/65	M	37203	M1	Gastric Ulcer
Black	2/14/65	M	37203	M1	Gastric Ulcer
Black	10/23/65	F	37215	M1	Gastritis
Black	8/24/65	F	37215	M2	Gastritis
Black	11/7/64	F	37215	M2	Gastritis
Black	12/1/64	F	37215	M2	Stomach Cancer
White	10/23/64	M	37215	M3	Flu
White	3/15/64	F	37217	M3	Flu
White	8/13/64	M	37217	M3	Flu
White	5/5/64	M	37217	M4	Pneumonia
White	2/13/67	M	37215	M4	Pneumonia
White	3/21/67	M	37215	M4	Flu

A dataset is “k-anonymous” if every record is in a set of at least k indistinguishable individuals

Example: k=2

Race	Birthdate	Sex	Zip	Medication	Diagnosis
Black	65	M	37203	M1	Gastric Ulcer
Black	65	M	37203	M1	Gastric Ulcer
Black	65	F	37215	M1	Gastritis
Black	65	F	37215	M2	Gastritis
Black	64	F	37215	M2	Gastritis
Black	64	F	37215	M2	Stomach Cancer
White	64	M	3721-	M3	Flu
White	64	-	37217	M3	Flu
White	64	M	3721-	M3	Flu
White	64	-	37217	M4	Pneumonia
White	67	M	37215	M4	Pneumonia
White	67	M	37215	M4	Flu

The higher “k”, the more privacy.

Attribute disclosure:
We know the Black / 65 / M had a Gastric Ulcer.

Black	65	M	37203	M1	Gastric Ulcer
Black	65	M	37203	M1	Gastric Ulcer
Black	65	F	37215	M1	Gastritis
Black	65	F	37215	M2	Gastritis
Black	64	F	37215	M2	Gastritis
Black	64	F	37215	M2	Stomach Cancer
White	64	M	3721-	M3	Flu
White	64	-	37217	M3	Flu
White	64	M	3721-	M3	Flu
White	64	-	37217	M4	Pneumonia
White	67	M	37215	M4	Pneumonia
White	67	M	37215	M4	Flu

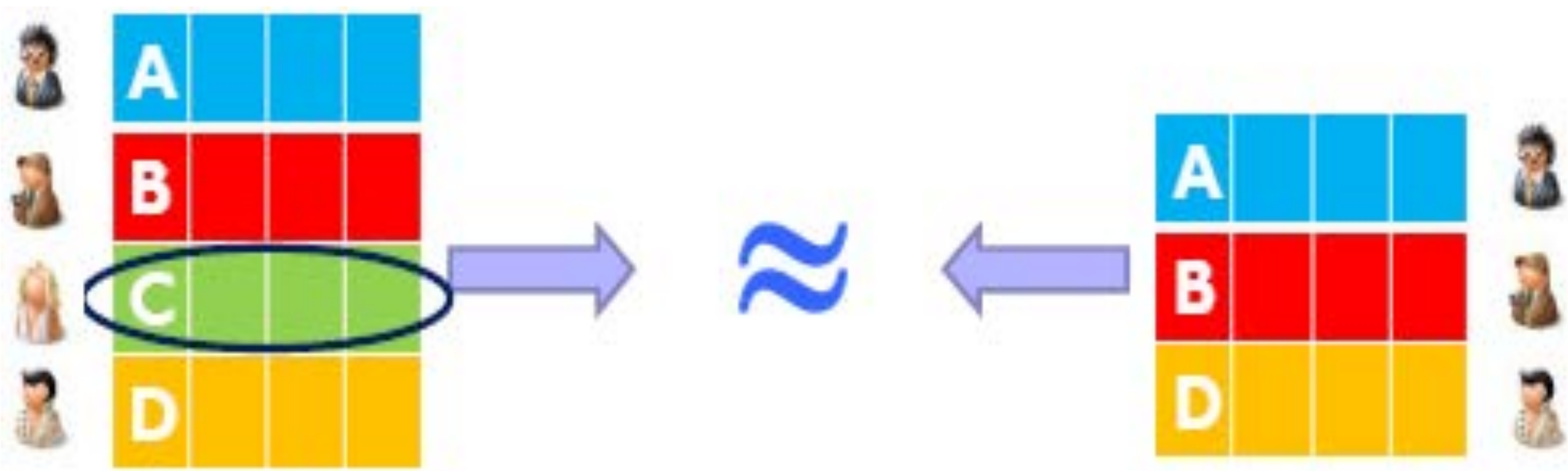
l-diversity solves this problem by assuring "diverseness" of the sensitive values. (This table is not l-diverse.)

Differential Privacy (informal)

Output is similar whether any single individual's record is included in the database or not

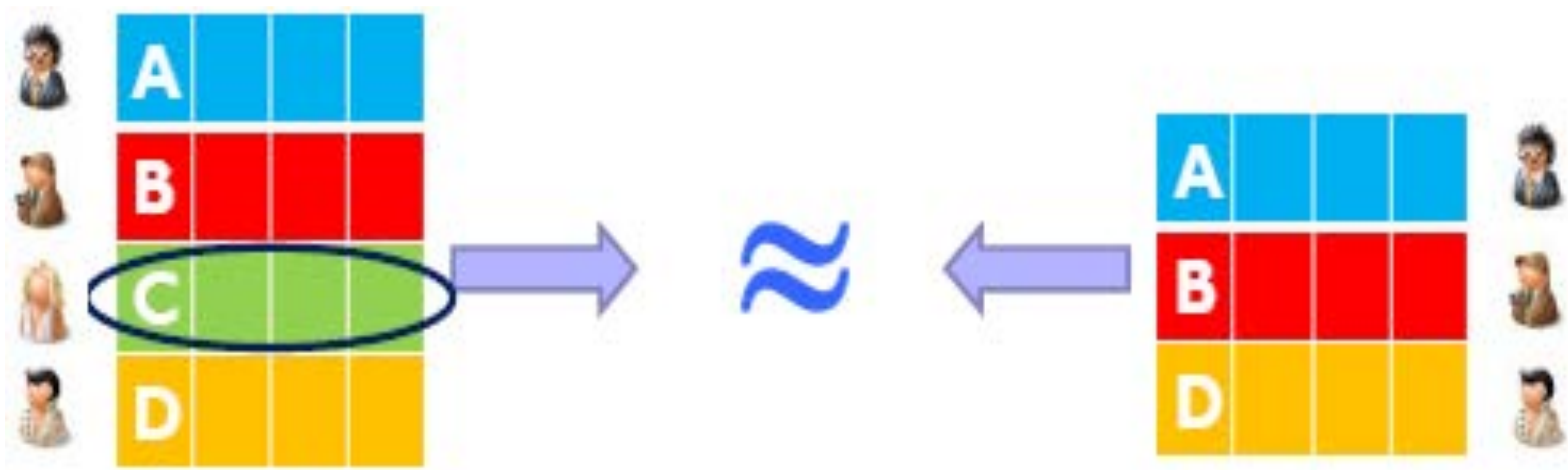
Differential Privacy (informal)

Output is similar whether any single individual's record is included in the database or not



Differential Privacy (informal)

Output is similar whether any single individual's record is included in the database or not

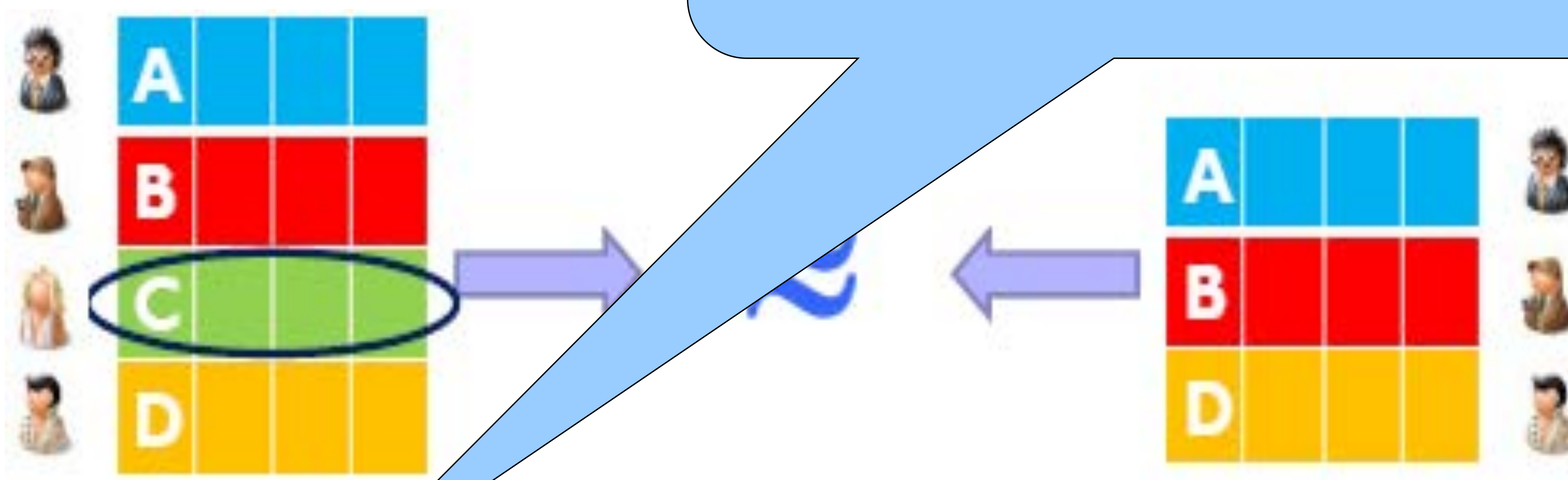


C is no worse off because her record is included in the computation

Differential Privacy (informal)

Output is similar whether any s
is included in the database or n

If there is already <u>some risk</u> of revealing a secret of C by combining auxiliary information and something learned from DB, then that risk is still there but not increased by C's participation in the database



C is no worse off because her record is included in the computation

Differential Privacy is …

… a guarantee intended to encourage individuals to permit their data to be included in socially useful statistical studies

Differential Privacy is …

… a guarantee intended to encourage individuals to permit their data to be included in socially useful statistical studies

The behavior of the system -- probability distribution on outputs -- is essentially unchanged, independent of whether any individual opts in or opts out of the dataset

Differential Privacy is …

… a guarantee intended to encourage individuals to permit their data to be included in socially useful statistical studies

The behavior of the system -- probability distribution on outputs -- is essentially unchanged, independent of whether any individual opts in or opts out of the dataset

… a type of indistinguishability of behavior on neighboring inputs

Suggests other applications:

- Approximate truthfulness as an economics solution concept [MT07, GLMRT]
- As alternative to functional (or syntactic) privacy [GLMRT]

Differential Privacy is …

… a guarantee intended to encourage individuals to permit their data to be included in socially useful statistical studies

- The behavior of the system -- probability distribution on outputs -- is essentially unchanged, independent of whether any individual opts in or opts out of the dataset

… a type of indistinguishability of behavior on neighboring inputs

- Suggests other applications:
  - Approximate truthfulness as an economics solution concept [MT07, GLMRT]
  - As alternative to functional (or syntactic) privacy [GLMRT]

… useless without utility guarantees

- Typically, "one size fits all" measure of utility
- Simultaneously optimal for different priors, loss functions [GRS09]

Sanitization Methods used with Differential Privacy

Input perturbation

Add random noise to database, release

Summary statistics only

Means, variances
Marginal totals
Regression coefficients

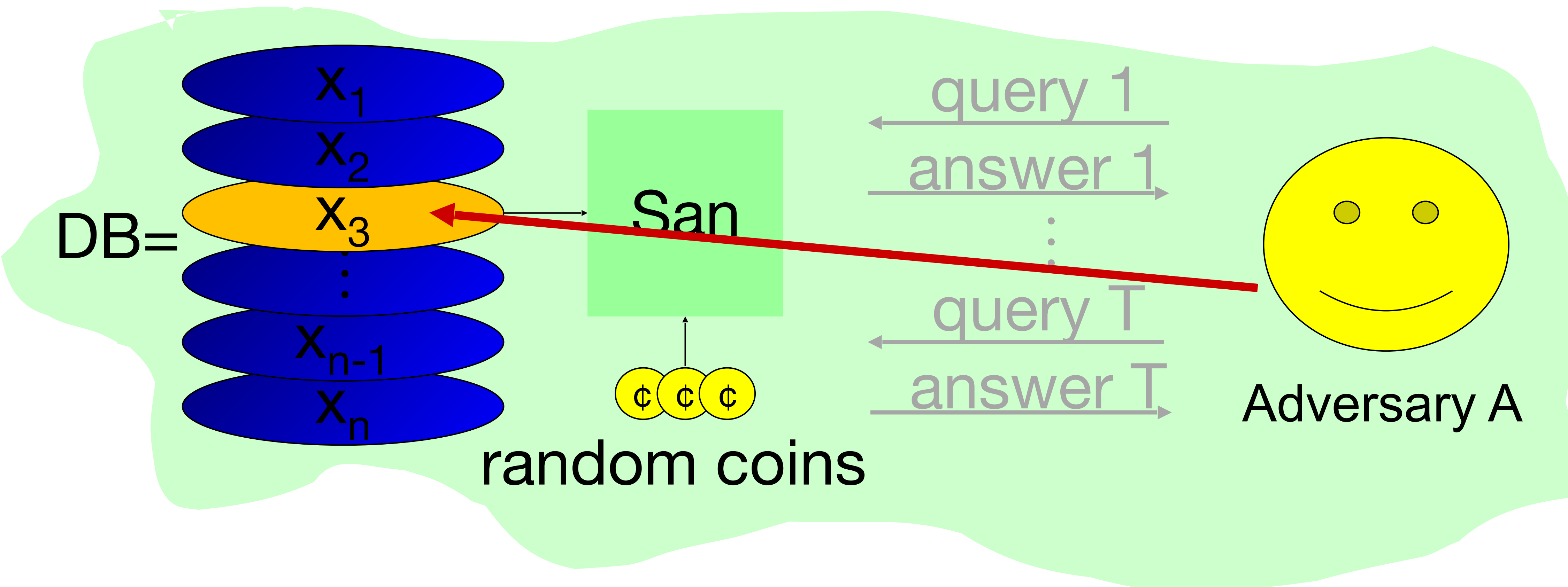
Output perturbation

Summary statistics with noise

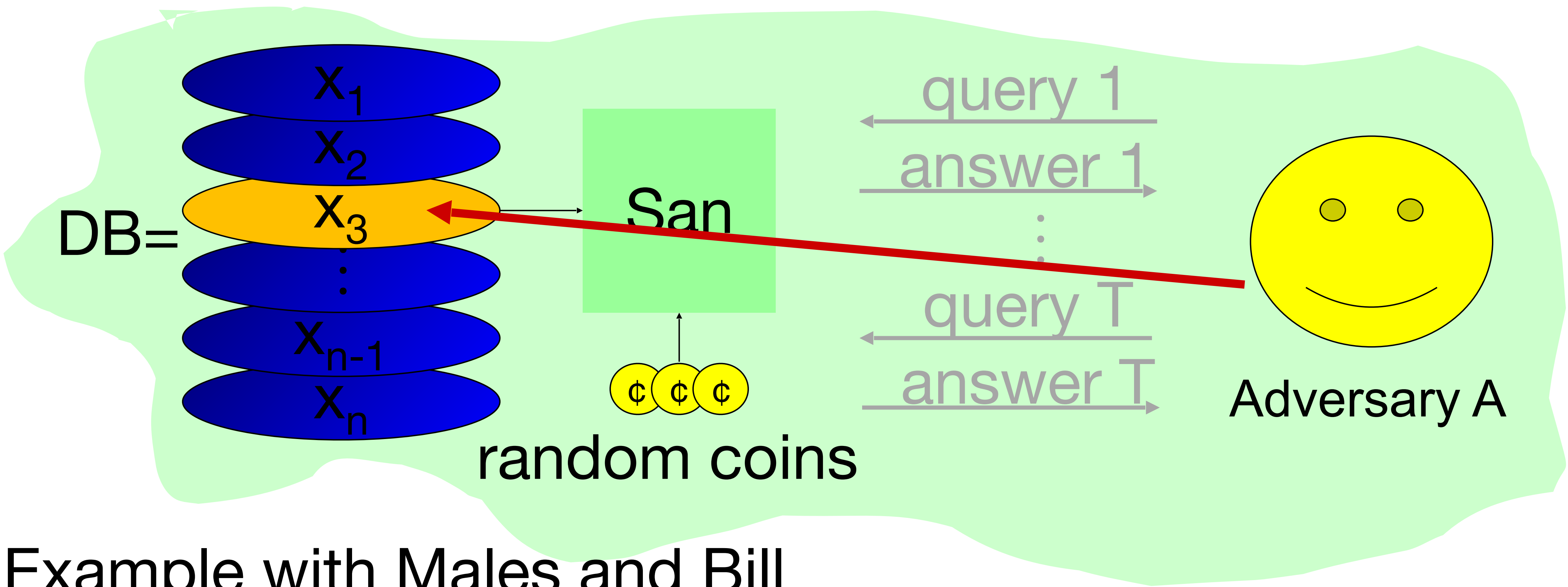
Interactive versions of the above methods

Auditor decides which queries are OK, type of noise

Differential Privacy (1)

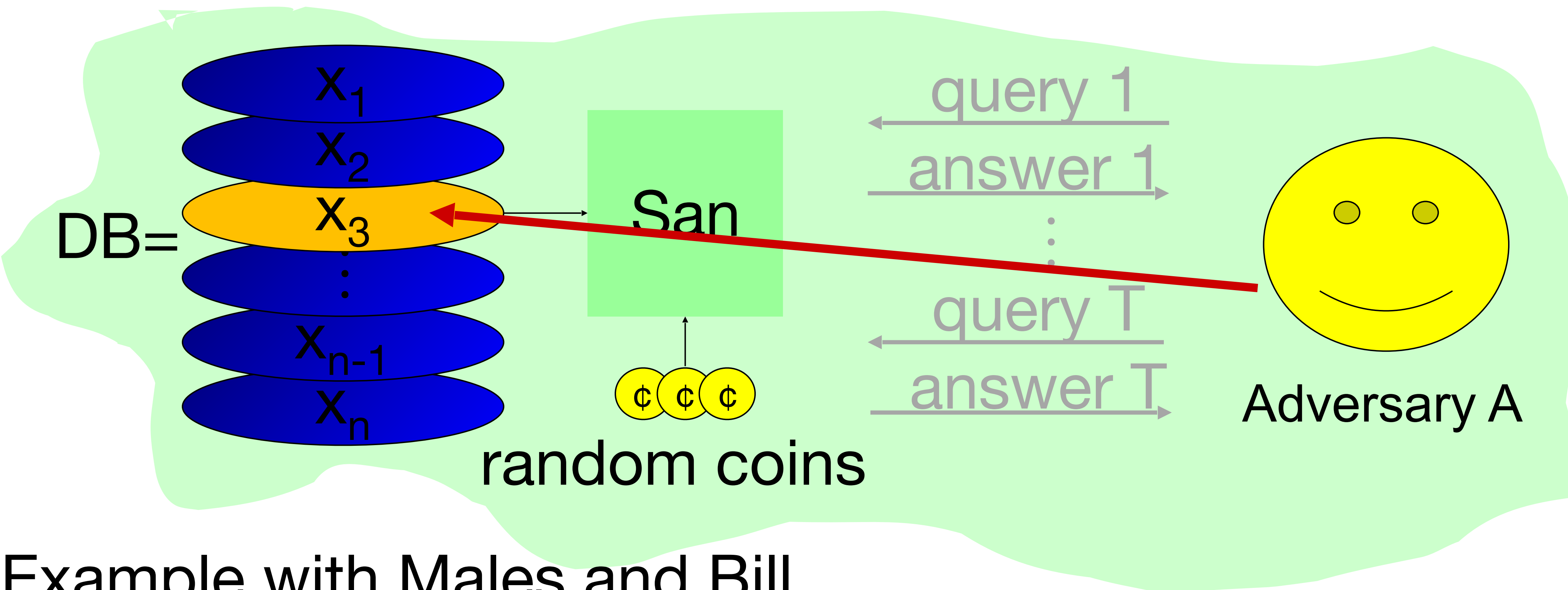


Differential Privacy (1)



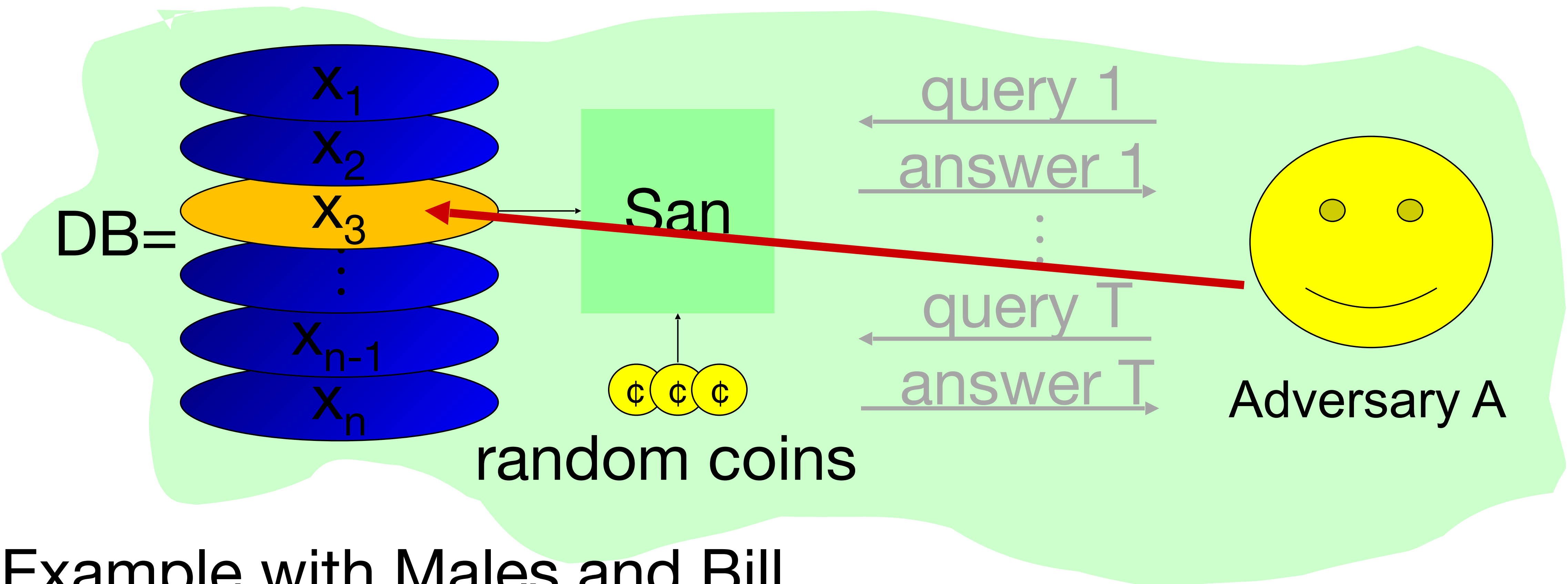
- Example with Males and Bill

Differential Privacy (1)



- Example with Males and Bill
  Adversary learns Bill's height even if he is not in the database

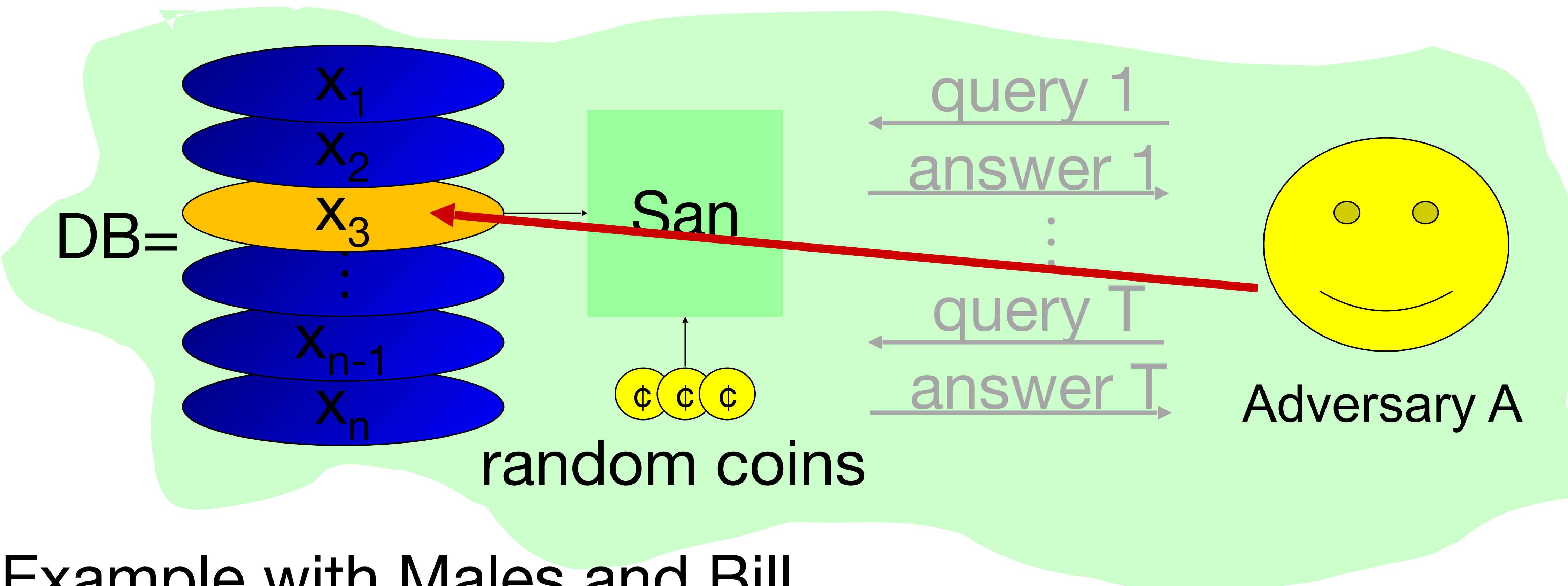
Differential Privacy (1)



- Example with Males and Bill
  Adversary learns Bill's height even if he is not in the database

- Intuition: "Whatever is learned would be learned regardless of whether or not Adam participates"

Differential Privacy (1)



- Example with Males and Bill
  Adversary learns Bill's height even if he is not in the database

- Intuition: "Whatever is learned would be learned regardless of whether or not Adam participates"
  Dual: Whatever is already known, situation won't get worse

Pseudonymization — de-identification that allows re-identification.

De-identified data:

ID	Race	Birthdate	Sex	Zip	Medication	Diagnosis
903	Black	9/20/65	M	37203	M1	Gastric Ulcer
932	Black	2/14/65	M	37203	M1	Gastric Ulcer
119	Black	10/23/65	F	37215	M1	Gastritis
16	Black	8/24/65	F	37215	M2	Gastritis
192	Black	11/7/64	F	37215	M2	Gastritis
50	Black	12/1/64	F	37215	M2	Stomach Cancer
181	White	10/23/64	M	37215	M3	Flu
133	White	3/15/64	F	37217	M3	Flu
374	White	8/13/64	M	37217	M3	Flu
356	White	5/5/64	M	37217	M4	Pneumonia
477	White	2/13/67	M	37215	M4	Pneumonia
499	White	3/21/67	M	37215	M4	Flu

Code Book:

ID	Name
903	Landry
932	Azariah
119	Oakley
16	Lennon
192	Charlie
50	Skyler
181	Dakota
133	Armani
374	Poenix
356	Justice
477	Casey
499	Remy

Pseudonymization —
de-identification that allows re-identification.

De-identified data:

ID	Race	Birthdate	Sex	Zip	Medication	Diagnosis
903	Black	9/20/65	M	37203	M1	Gastric Ulcer
932	Black	2/14/65	M	37203	M1	Gastric Ulcer
119	Black	10/23/65	F	37215	M1	Gastritis
16	Black	8/24/65	F	37215	M2	Gastritis
192	Black	11/7/64	F	37215	M2	Gastritis
50	Black	12/1/64	F	37215	M2	Stomach Cancer
181	White	10/23/64	M	37215	M3	Flu
133	White	3/15/64	F	37217	M3	Flu
374	White	8/13/64	M	37217	M3	Flu
356	White	5/5/64	M	37217	M4	Pneumonia
477	White	2/13/67	M	37215	M4	Pneumonia
499	White	3/21/67	M	37215	M4	Flu

Code Book:

ID	Name
903	Landry
932	Azariah
[redacted]	Oak[redacted]
[redacted]	Le[redacted]
192	[redacted]lie
50	[redacted]kyler
181	[redacted]kota
13[redacted]	[redacted]ni
[redacted]	Po[redacted]
[redacted]	Justi[redacted]
477	Casey
499	Remy

Pseudonymization —
de-identification that allows re-identification.

De-identified data:

ID	Race	Birthdate	Sex	Zip	Medication	Diagnosis
903	Black	9/20/65	M	37203	M1	Gastric Ulcer
932	Black	2/14/65	M	37203	M1	Gastric Ulcer
119	Black	10/23/65	F	37215	M1	Gastritis
16	Black	8/24/65	F	37215	M2	Gastritis
192	Black	11/7/64	F	37215	M2	Gastritis
50	Black	12/1/64	F	37215	M2	Stomach Cancer
181	White	10/23/64	M	37215	M3	Flu
133	White	3/15/64	F	37217	M3	Flu
374	White	8/13/64	M	37217	M3	Flu
356	White	5/5/64	M	37217	M4	Pneumonia
477	White	2/13/67	M	37215	M4	Pneumonia
499	White	3/21/67	M	37215	M4	Flu

Code Book:

ID	Name
903	Landry
932	Azariah
[illegible]	Oakl[illegible]
[illegible]	Le[illegible]
192	[illegible]lie
50	[illegible]kyler
181	[illegible]kota
13[illegible]	[illegible]ni
[illegible]	Po[illegible]
[illegible]	Justi[illegible]
477	Casey
499	Remy

Erasing the map "anonymizes" the data.
(It could still be re-identified!)

Outline for today's talk

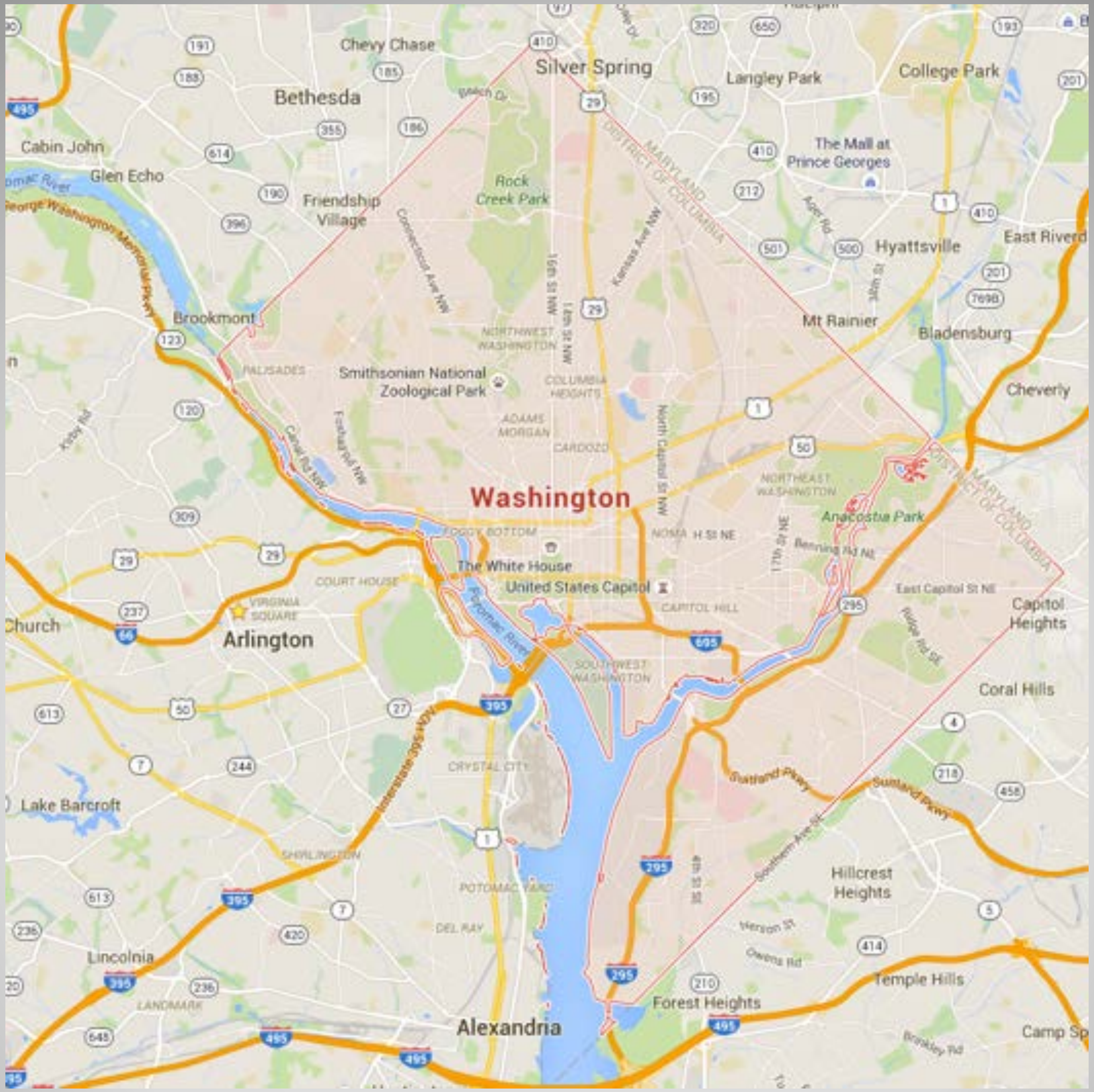
- Why de-identify? ✔
- Basic de-identification ✔
- Famous re-identification controversies ✔
- De-identification in practice ✔
- Measuring re-identification risk ✔
- De-identification governance
- De-identification @ NIST — Workshop June 29th

There are many ways to measure re-identification risk.

K-anonymity measures the # of people that each record could *match.*

Differential privacy adds noise to mask the contribution of each individual

Pseudonymization allows future re-identification



Governance Approaches

Responses: Law

Privacy on the ground versus on the books
(Bamberger and Mulligan, various)

- Say “anonymize,” go to jail.

HIPAA Rule

COPPA Rule: covers all persistent identifiers

Pineda v. William-Sonoma, 51 Cal.4th 524 (Cal. 2011).

- Song-Beverly Credit Card Act: Retailers cannot collect “information concerning the cardholder” as a condition of accepting credit card payment

Responses: Law (cont.)

FTC Privacy Report (March 2012)

data is not “reasonably linkable” to the extent that a company:

1. takes reasonable measures to ensure that the data is de-identified;
   — *This means that the company must achieve a reasonable level of justified confidence that the data cannot reasonably be used to infer information about, or otherwise be linked to, a particular consumer, computer, or other device.*

2. publicly commits not to try to re-identify the data; and

3. contractually prohibits downstream recipients from trying to re-identify the data.

Meanwhile…

“Data is the new oil”

- Every regulation will “kill the Internet”
- The blurring of science and commerce
  — *Who has better data? Census or Facebook?*
  — *Should Facebook get an IRB or should we soften the Common Rule?*

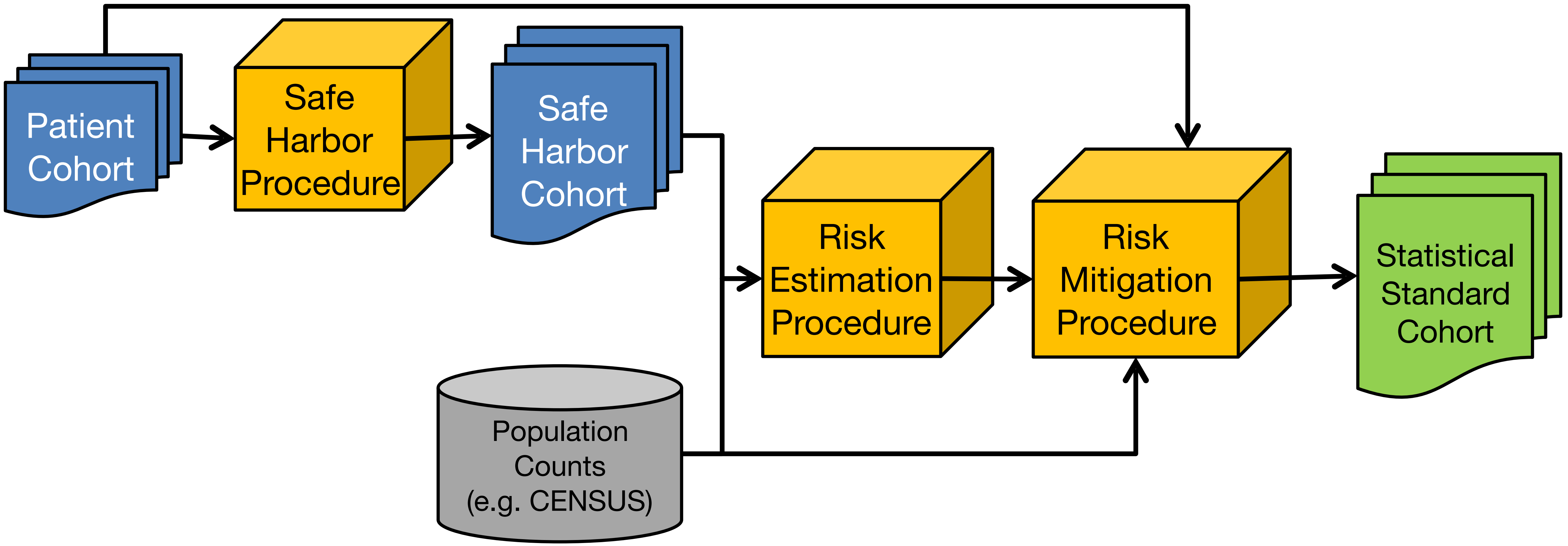
Big Data and Target’s Pregnancy Study

https://pixabay.com/en/large-data-keyboard-computer-895567/

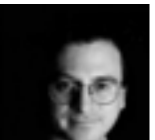
https://pixabay.com/en/pregnant-tummies-heart-244662/

TARGET

Benitez, Loukides & Malin: Discovering de-identification policy alternatives.



K. Benitez, G. Loukides, and B. Malin. Beyond Safe Harbor: automatic discovery of health information de-identification policy alternatives. Proceedings of the ACM International Health Informatics Symposium. 2010: to appear.



Data Release Boards / Data Review Boards

Organizations can use a **Data Release Board** to review data prior to release.

- Composed of experts drawn from different units.
- Can review:
  - *Requests*
  - *Proposed release*
  - *Actual data*

Model used by:

- Department of Education

- Others.

Outline for today's talk

- Why de-identify? ✔
- Basic de-identification ✔
- Famous re-identification controversies ✔
- De-identification in practice ✔
- Measuring re-identification risk ✔
- De-identification governance ✔
- De-identification @ NIST — Workshop June 29th

There are many ways to measure re-identification risk.

K-anonymity measures the # of people that each record could *match.*

Differential privacy adds noise to mask the contribution of each individual


https://pixabay.com/en/ball-http-www-crash-administrator-63527/

For further information…

De-ID@NIST

June 29th — Government-only workshop @ NIST

- Current De-ID practice & requirements
- De-ID tools
- We are looking for participants & speakers.
- deidentification@nist.gov

De-identification evaluation

- Commercial & Open Source tools:
  — *What's available?*
  — *How well do they work?*
- What data sets should we use?
- March – Sept: Pilot Program

De-identification guidance

- June 2016 — Draft document on how to de-id

Questions for federal agencies

What is the acceptable level of re-identification risk?

- 0%?
- HIPAA is ≈ 0.5% (but in a real test, it was 2 out of 12,000)

Who should make the determination?

- Individual scientists?
- Data release boards?
- FOIA Office?
- Privacy Office?
- Legal?

Legal Ramifications for Federal Agencies

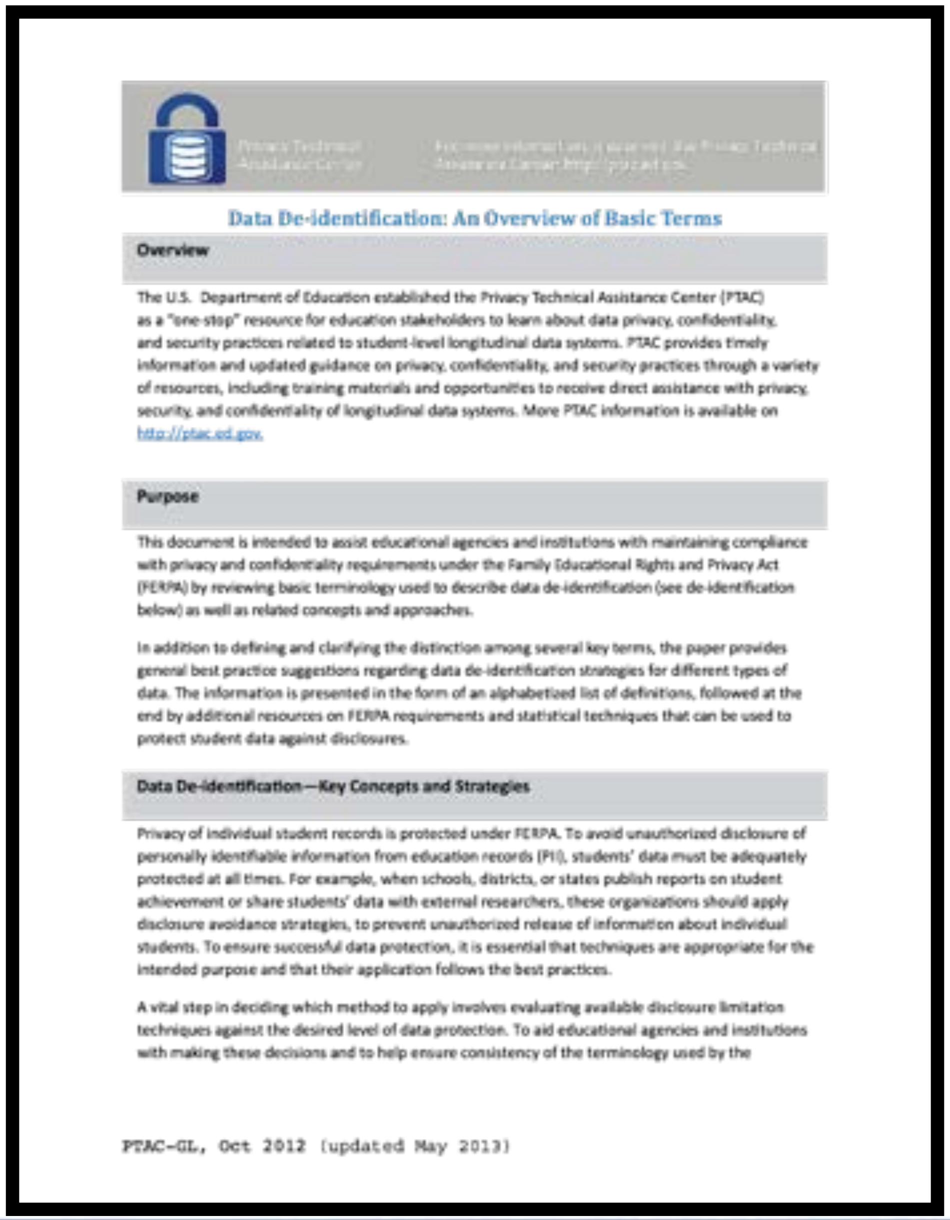
Question: Does deidentification help an agency limit liability or comply with legal requirements?

- E.g. Privacy Act or FOIA?

Answer: No clear answers, but showing reasonable steps to reduce risk of reidentification/privacy harm is a very good idea.

Department of Education & HHS have de-identification guidance.

Privacy Technical Assistance Center
Department of Education
ptac.ed.gov



Data De-identification: An Overview of Basic Terms

Overview

The U.S. Department of Education established the Privacy Technical Assistance Center (PTAC) as a "one-stop" resource for education stakeholders to learn about data privacy, confidentiality, and security practices related to student-level longitudinal data systems. PTAC provides timely information and updated guidance on privacy, confidentiality, and security practices through a variety of resources, including training materials and opportunities to receive direct assistance with privacy, security, and confidentiality of longitudinal data systems. More PTAC information is available on http://ptac.ed.gov.

Purpose

This document is intended to assist educational agencies and institutions with maintaining compliance with privacy and confidentiality requirements under the Family Educational Rights and Privacy Act (FERPA) by reviewing basic terminology used to describe data de-identification (see de-identification below) as well as related concepts and approaches.

In addition to defining and clarifying the distinction among several key terms, the paper provides general best practice suggestions regarding data de-identification strategies for different types of data. The information is presented in the form of an alphabetized list of definitions, followed at the end by additional resources on FERPA requirements and statistical techniques that can be used to protect student data against disclosures.

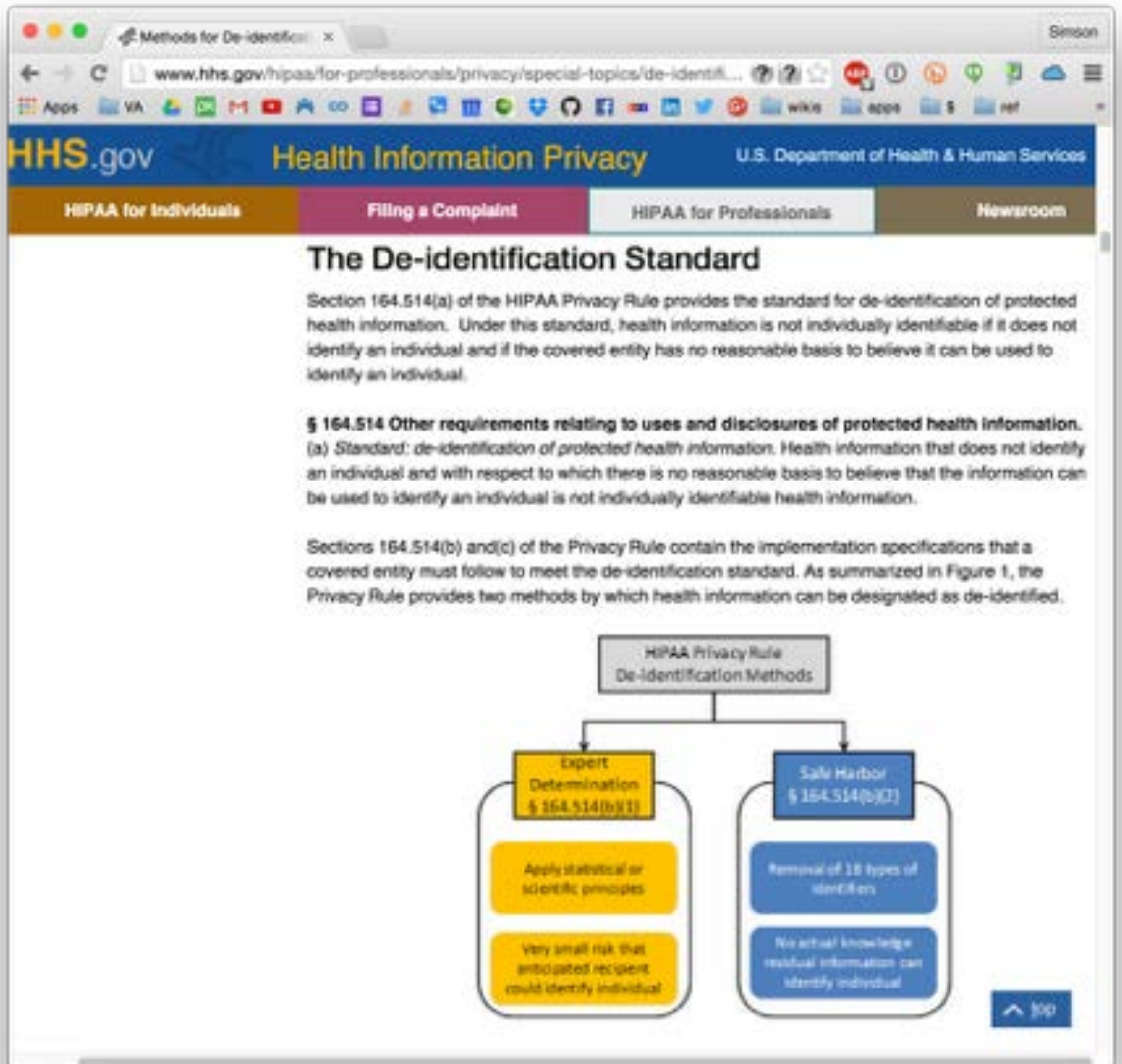
Data De-identification—Key Concepts and Strategies

Privacy of individual student records is protected under FERPA. To avoid unauthorized disclosure of personally identifiable information from education records (PII), students' data must be adequately protected at all times. For example, when schools, districts, or states publish reports on student achievement or share students' data with external researchers, these organizations should apply disclosure avoidance strategies, to prevent unauthorized release of information about individual students. To ensure successful data protection, it is essential that techniques are appropriate for the intended purpose and that their application follows the best practices.

A vital step in deciding which method to apply involves evaluating available disclosure limitation techniques against the desired level of data protection. To aid educational agencies and institutions with making these decisions and to help ensure consistency of the terminology used by the

PTAC-GL, Oct 2012 (updated May 2013)

HHS.gov
Health Information Privacy
www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/



Methods for De-identific
www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identifi...

HHS.gov Health Information Privacy U.S. Department of Health & Human Services

HIPAA for Individuals | Filing a Complaint | HIPAA for Professionals | Newsroom

The De-identification Standard

Section 164.514(a) of the HIPAA Privacy Rule provides the standard for de-identification of protected health information. Under this standard, health information is not individually identifiable if it does not identify an individual and if the covered entity has no reasonable basis to believe it can be used to identify an individual.

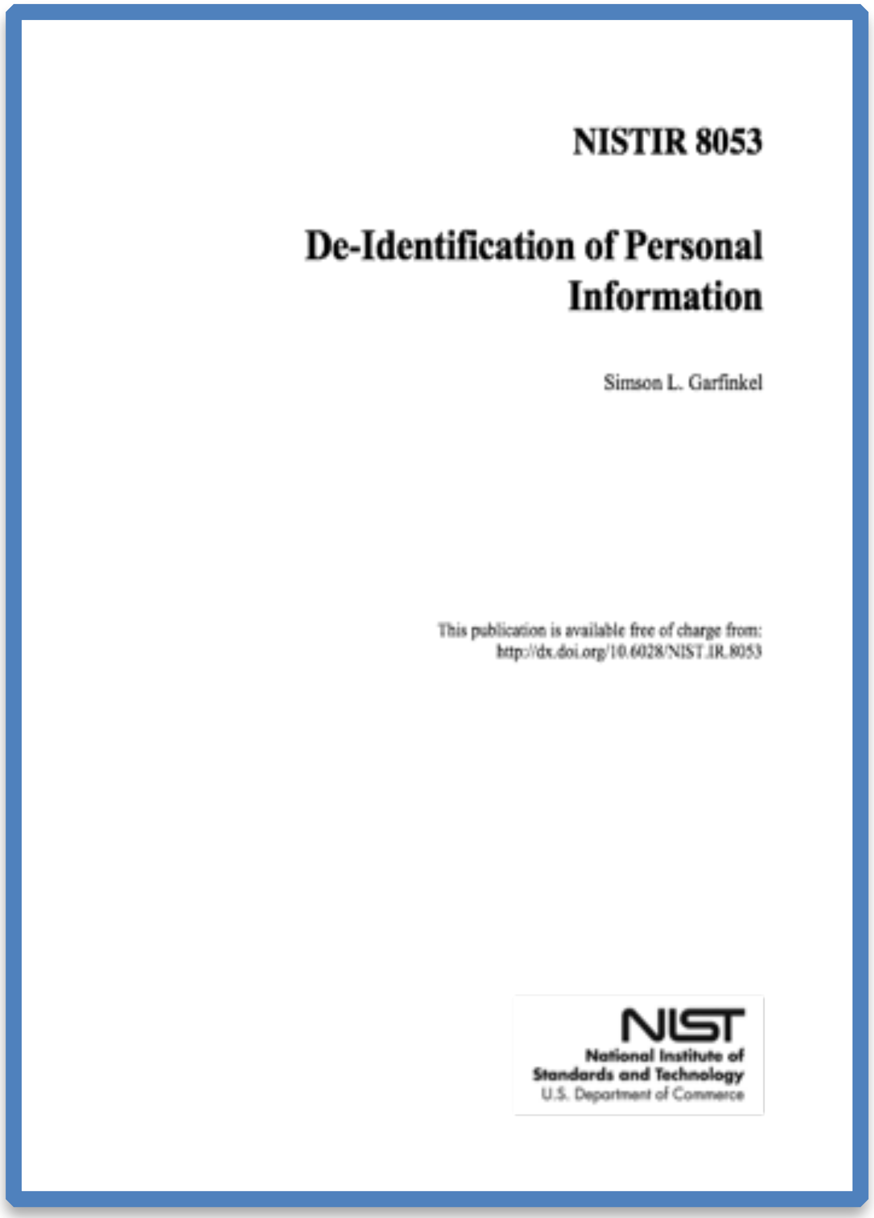
§ 164.514 Other requirements relating to uses and disclosures of protected health information. (a) *Standard: de-identification of protected health information.* Health information that does not identify an individual and with respect to which there is no reasonable basis to believe that the information can be used to identify an individual is not individually identifiable health information.

Sections 164.514(b) and(c) of the Privacy Rule contain the implementation specifications that a covered entity must follow to meet the de-identification standard. As summarized in Figure 1, the Privacy Rule provides two methods by which health information can be designated as de-identified.

This presentation is based in part on NISTIR 8053: De-Identification of Personal Information

Covers:

- Why de-identify?
- De-identification terminology
- Famous re-identification cases
- De-identifying and re-identifying *structured data*
  (e.g. survey data, Census data, etc.)
- Challenges with de-identifying *unstructured data*
  (e.g. medical text, photographs, medical imagery, genetic information)


NISTIR 8053

De-Identification of Personal Information

Simson L. Garfinkel

This publication is available free of charge from:
http://dx.doi.org/10.6028/NIST.IR.8053

NIST
National Institute of Standards and Technology
U.S. Department of Commerce

http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf

October 2015

vi+46 pages

Thanks!