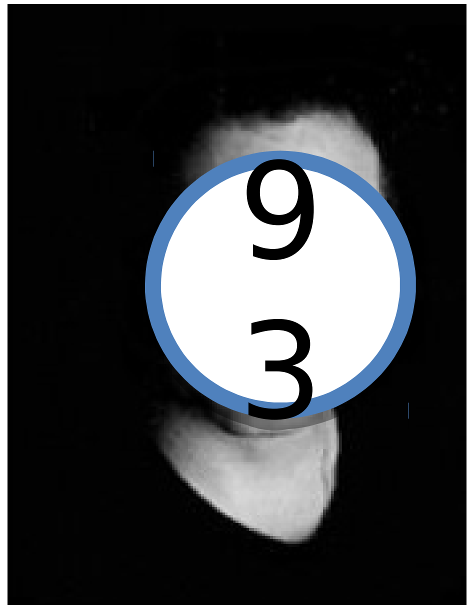
Data De-Identification

Overview and framing of current issues



June 9: Emerging Methods: Part II

11:00am

Simson L. Garfinkel, Ph.D.

Information Technology Laboratory

National Institute of Standards and Technology

Berkeley Initiative for Transparency in the Social Sciences

Summer Institute—Transparency and Reproducibility Methods for Social Science Research

National Institute of Standards and Technology

Founded in 1901

Non-regulatory federal laboratory.

Mission:

- "To promote US innovation and industrial competitiveness by advancing measurement science, standards, and technology in ways that enhance economic security and improve our quality of life."

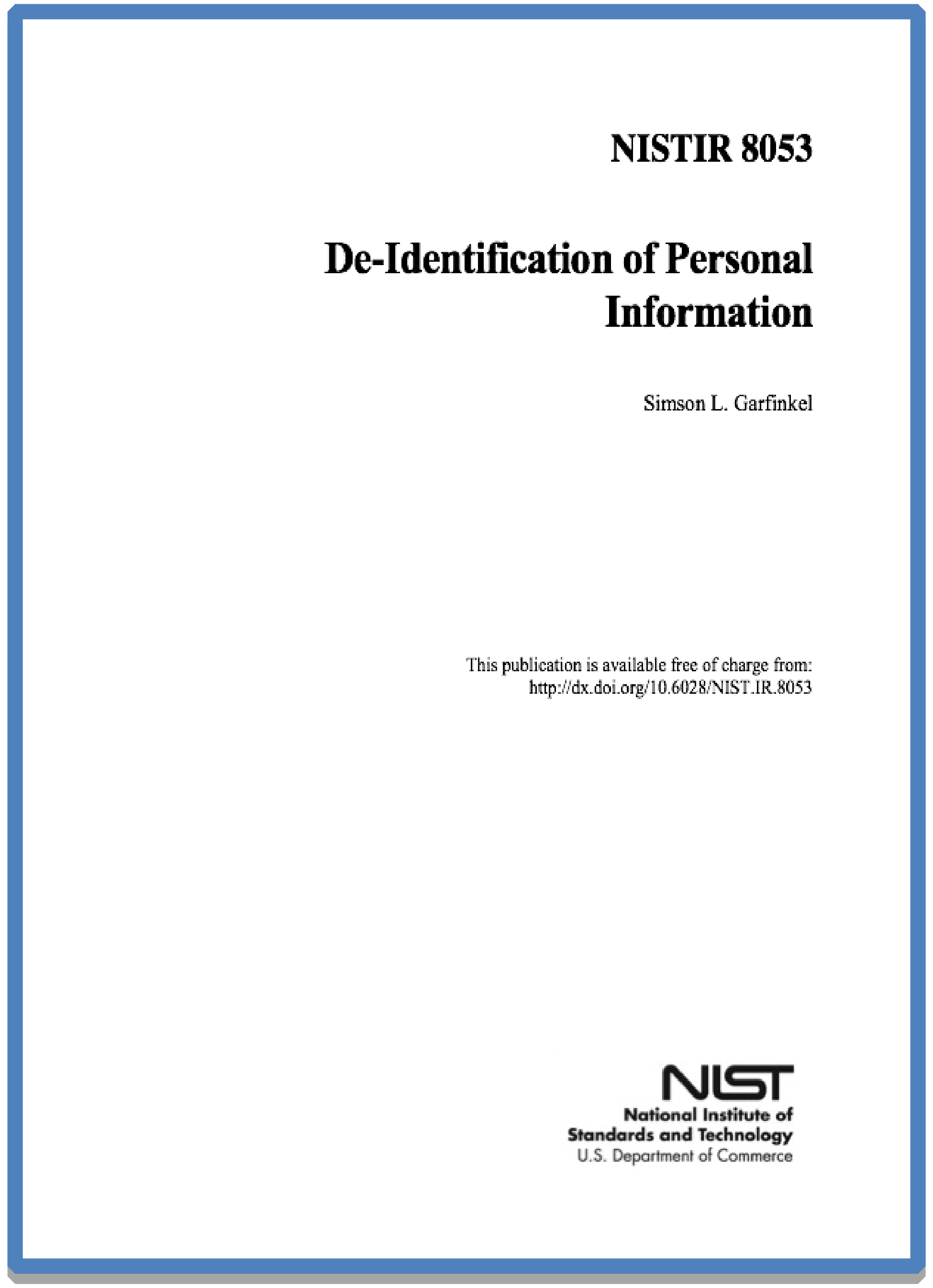
K20 Reference Kilogram:

http://www.nist.gov/pml/si-redef/kg_intro.cfm

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

Contents:

- 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

De-Identification: Removing information that can identify

Grover Cleveland	March 18, 1837	Stephen Grover Cleveland	Caldwell	New Jersey	24
William McKinley	January 29, 1843	William McKinley, Jr.	Niles	Ohio	25
Theodore Roosevelt	October 27, 1858	Theodore Roosevelt, Jr.	New York City	New York	26
William Howard Taft	September 15, 1857		Cincinnati	Ohio	27
Woodrow Wilson	December 28, 1856	Thomas Woodrow Wilson	Staunton	Virginia	28

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

Text:

[redacted]	[redacted] 1837	[redacted]	Caldwell	New Jersey	24
[redacted]	[redacted] 1843	[redacted]	Niles	Ohio	25
[redacted]	[redacted] 1858	[redacted]	New York City	New York	26
[redacted]	[redacted], 1857	[redacted]	Cincinnati	Ohio	27
[redacted]	[redacted] 1856	[redacted]	Staunton	Virginia	28

Images:

There is a significant and growing interest in de-identification.

Controlled Sharing

Open Science

Risk
Mitigation

DATA.GOV

Data Publishing

Oversight

Long-term
archiving

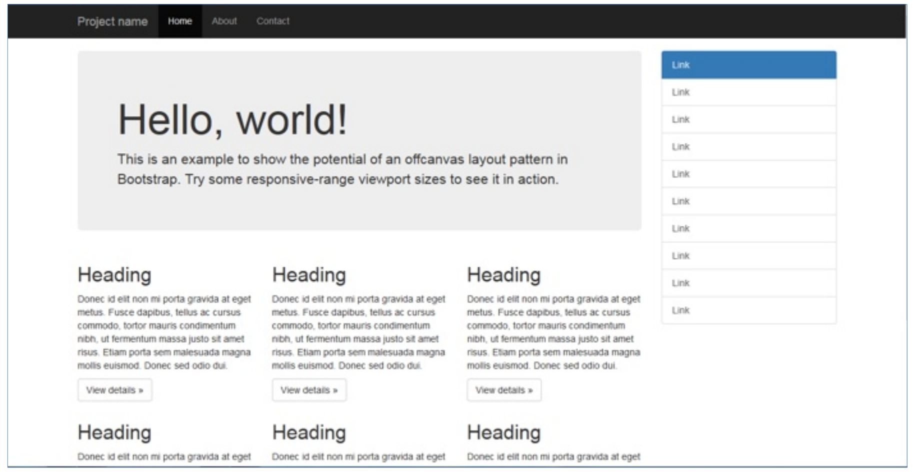
Interest in de-identification extends far beyond healthcare.

https://www.flickr.com/photos/usdagov/4423599680

Social Science Data

https://pixabay.com/en/credit-card-bill-bank-statement-1104961/

Consumer Financial Data



Website visitor data

“We will never share your personal information…”

De-identification is not a single technique.

De-identification: “general term for any process of removing the association between a set of identifying data and the data subject”

- ISO/TS 25237:2008(E)

“De-identification is a process that reduces the risk of identification of entries in a data set.”

- John Moehrke

“De-identification is a tool that organizations can use to remove personal information from data that they collect, use, archive, and share with other organizations.”

- NISTIR 8053

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

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

— *Multiple regulations.*

https://pixabay.com/en/tools-technique-open-end-wrench-1093117/

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.

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.

Russ Altman

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

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).

The fundamental de-identification problem: information can be *identifying* without being an *identifier*.

- **Identifier:** "information used to claim an identity, before a potential corroboration by a corresponding authenticator"

  — *(ISO/TS 25237:2008)*

Simply removing identifiers does not necessarily de-identify.

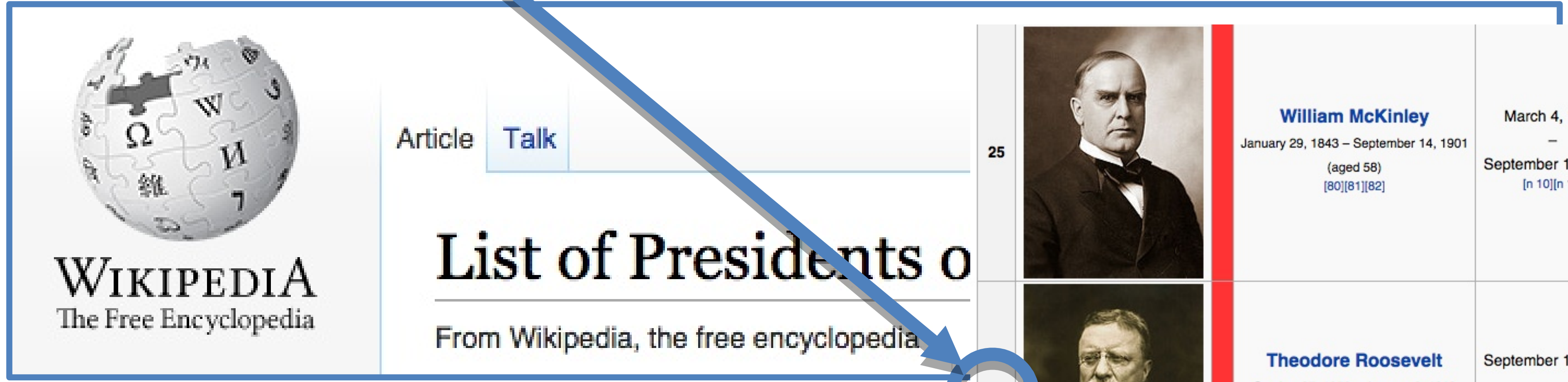
Subject 26 Photo:

Subject 26 Narrative:

XXXXXXXXXXXXXXXX ([a] XXXXXXX XX, XXXX – XXXXXXX X, XXXX), often referred to by his initials XX, was an American statesman, author, explorer, soldier, naturalist, and reformer who served as the XXth President of the United States.

We can use auxiliary information to figure out the identity of #26.

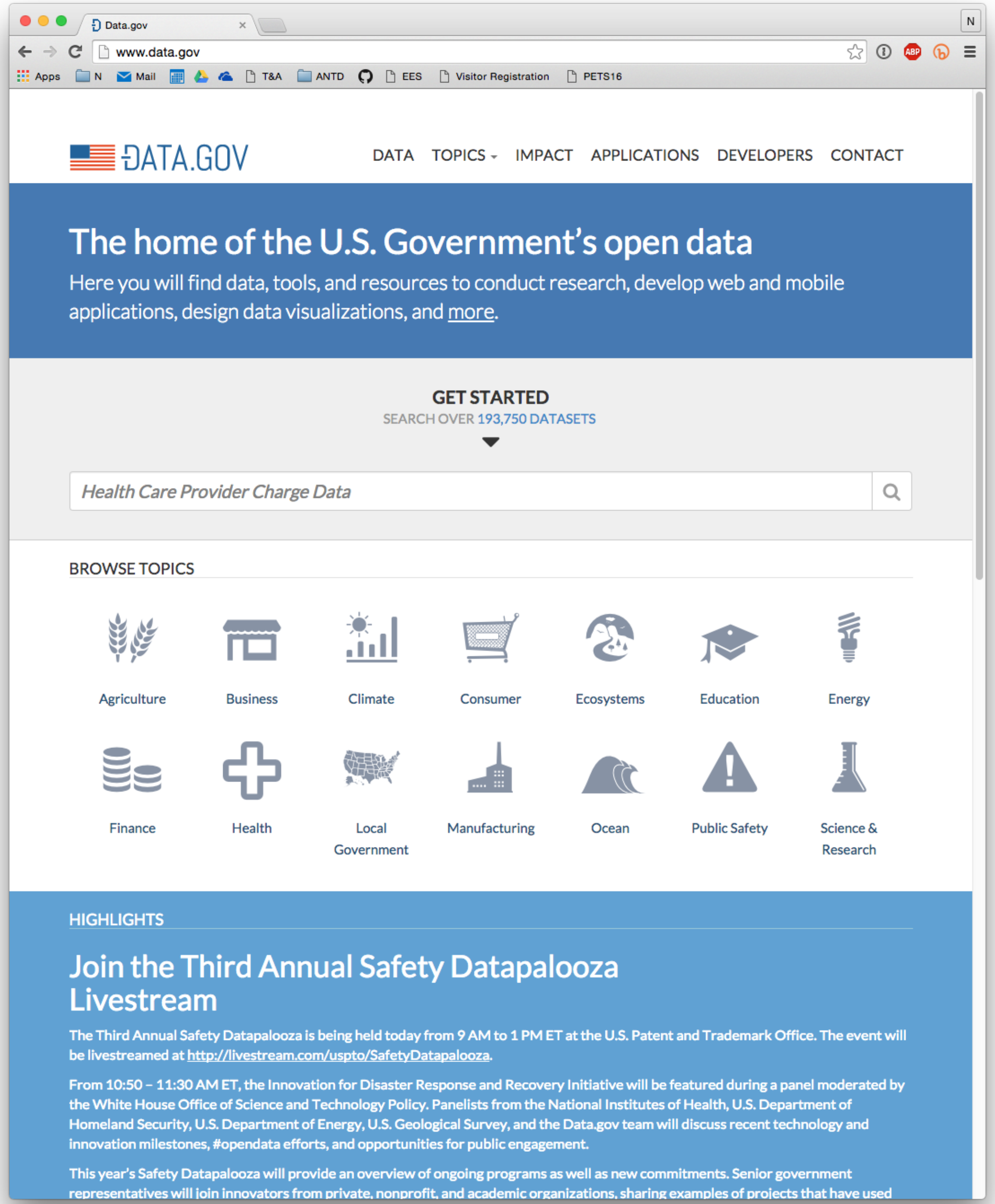
Many kinds of data can be used for a “linkage attack.”



25		William McKinley January 29, 1843 – September 14, 1901 (aged 58) [80][81][82]	March 4, 1897 – September 14, 1901 [n 10][n 11]	Republican	28 (1896)	39th Governor of Ohio (1892–1896)
26		Theodore Roosevelt October 27, 1858 – January 6, 1919 (aged 60) [83][84][85]	September 14, 1901 – March 4, 1909 [n 7]	Republican	29 (1900) 30 (1904)	25th Vice President of the United States
27		William Howard Taft September 15, 1857 – March 8, 1930 (aged 72) [86][87][88]	March 4, 1909 – March 4, 1913 [n 3]	Republican	31 (1908)	42nd United States Secretary of War (1904–1908)

Re-identification is rarely 100% certain.

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

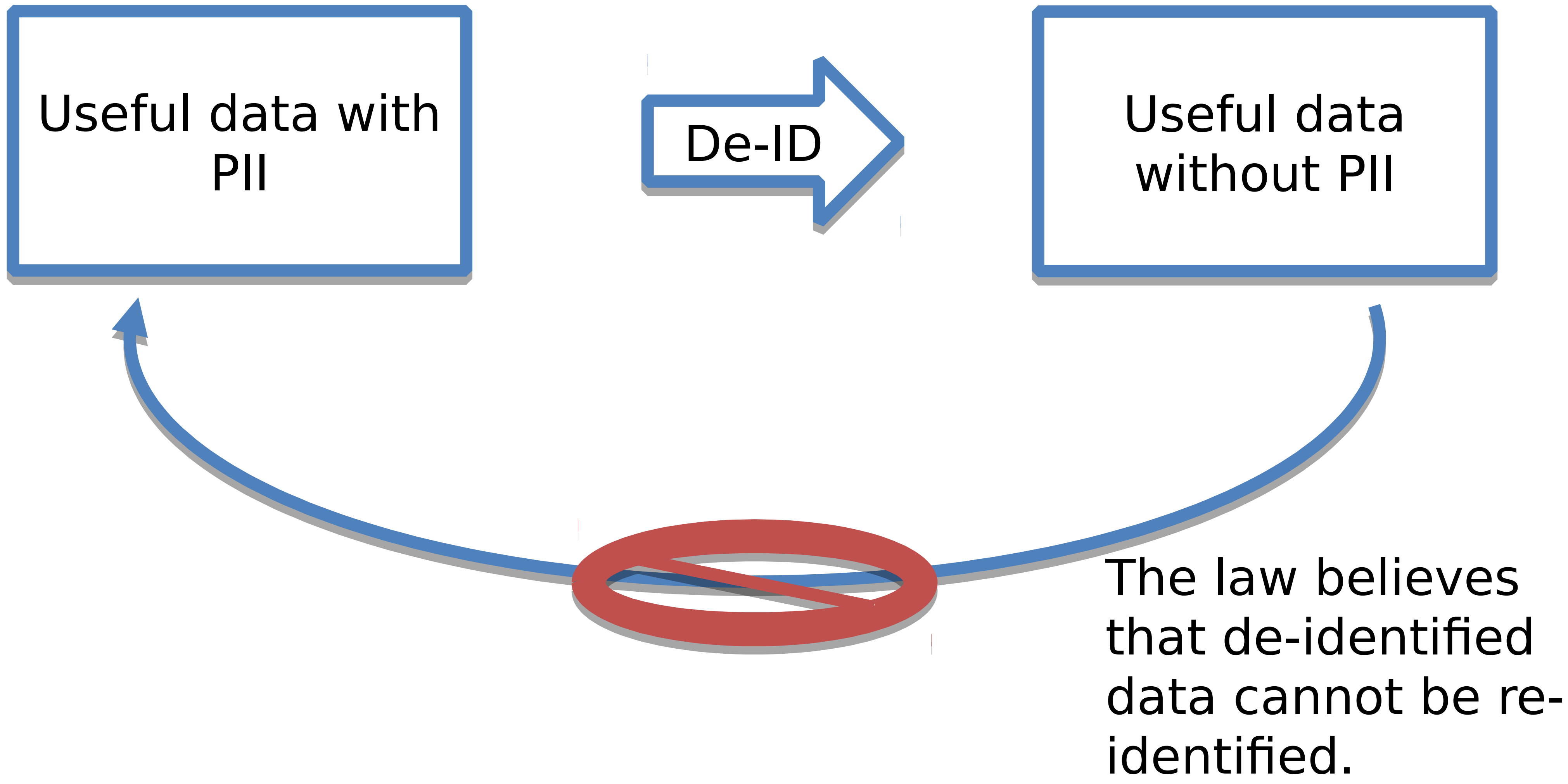


Protect Privacy

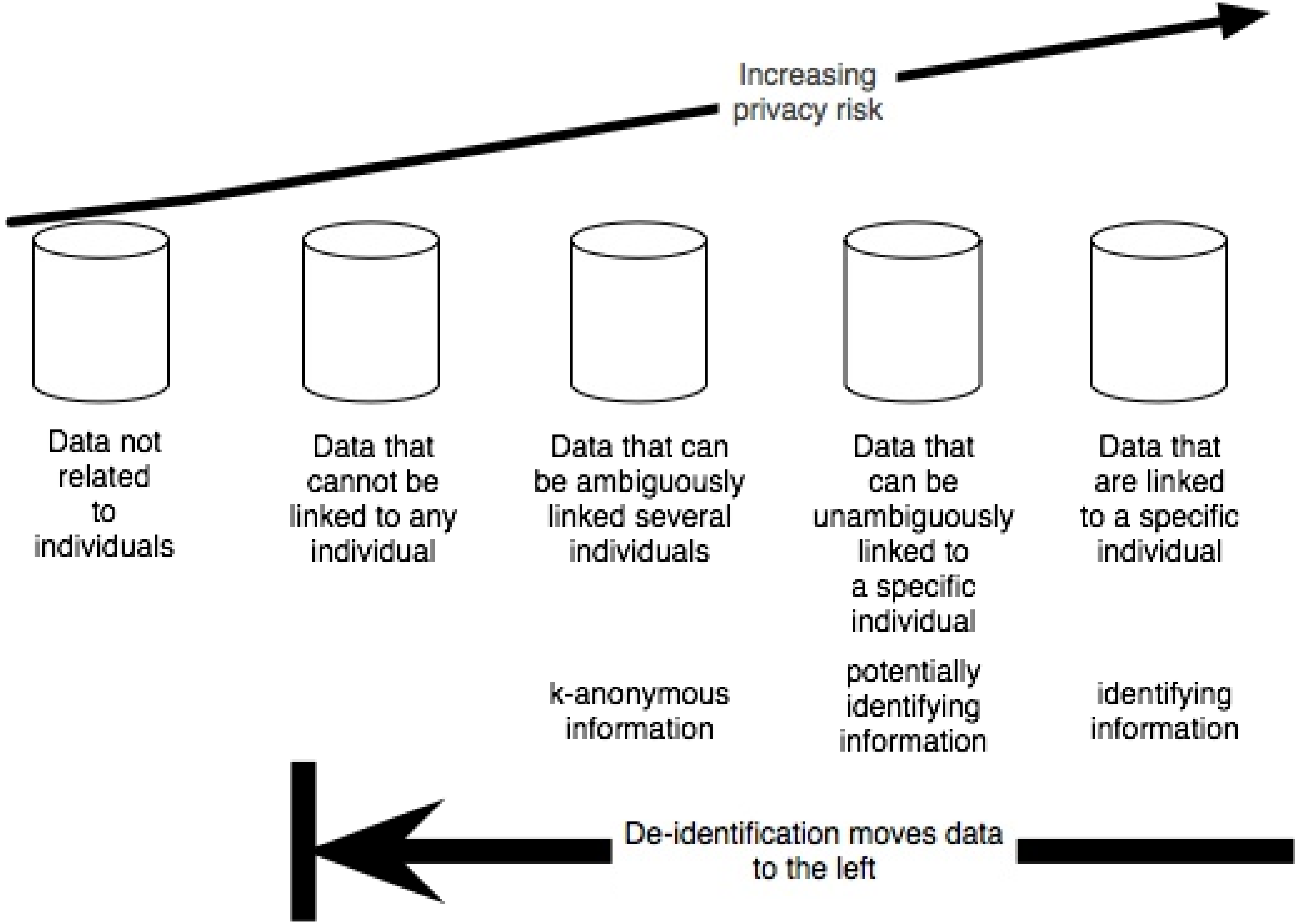
Avoid Surveillance

Foil "Data Intruders"

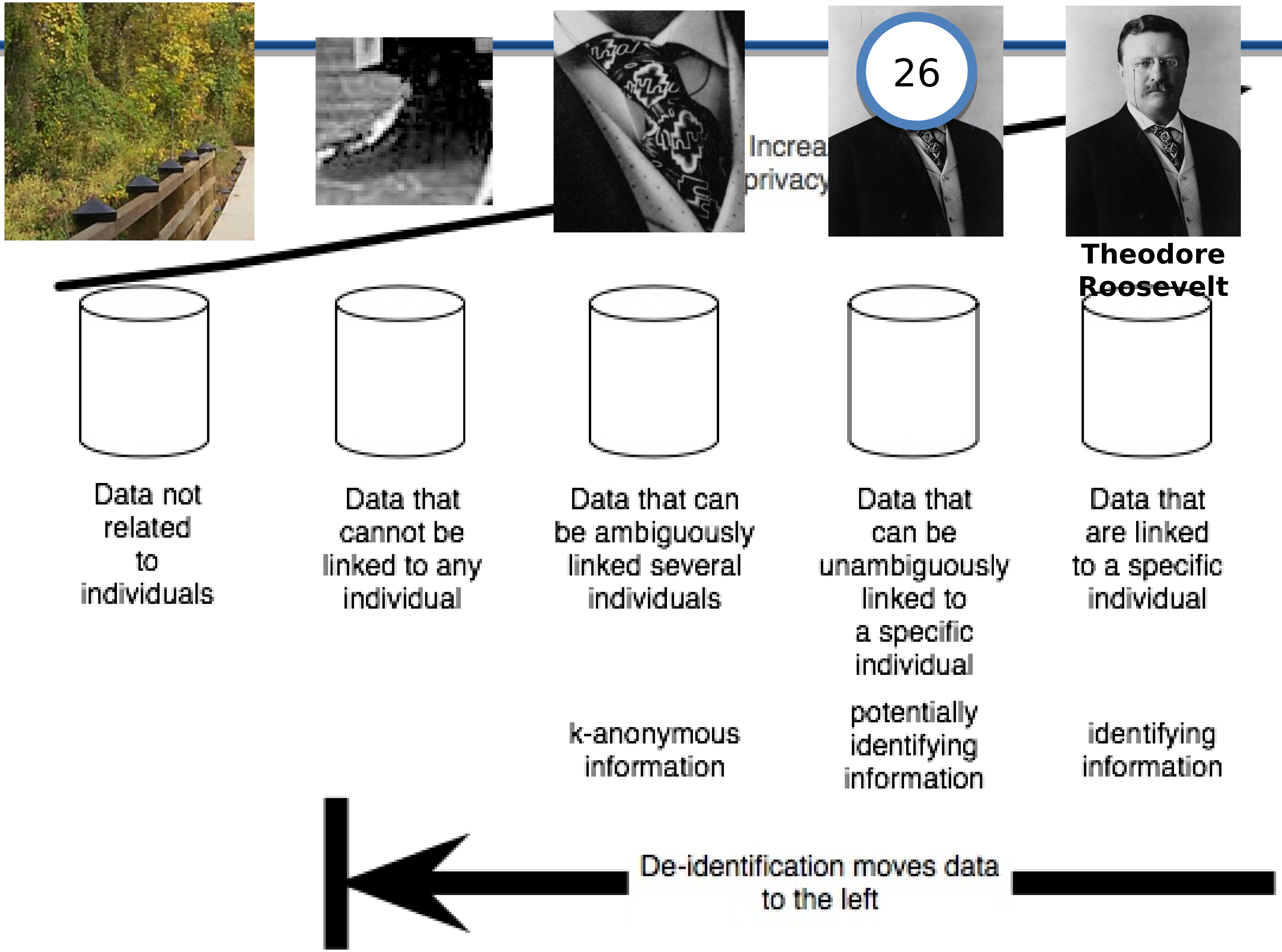
Our laws assume that perfect de-identification is possible.



NISTIR 8053 proposes an "identifiability spectrum" for data:



We can put photos on the identifiability spectrum



De-identification questions:

How do you know if data are properly de-identified?

What is “anonymized” vs. “de-identified” vs. “pseudonymized?”

What is the trade-off between identifiability and data quality?

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

For further information.

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

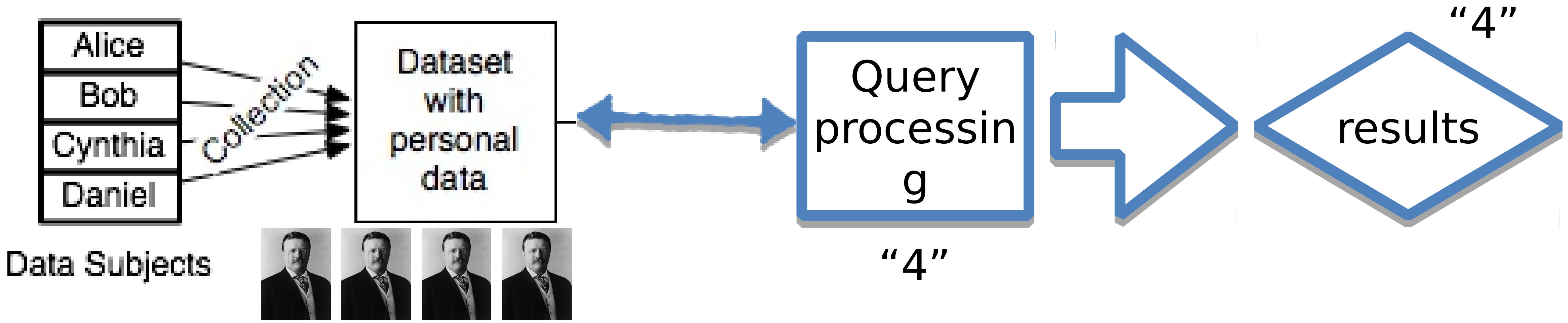
Basic De-I

eeney

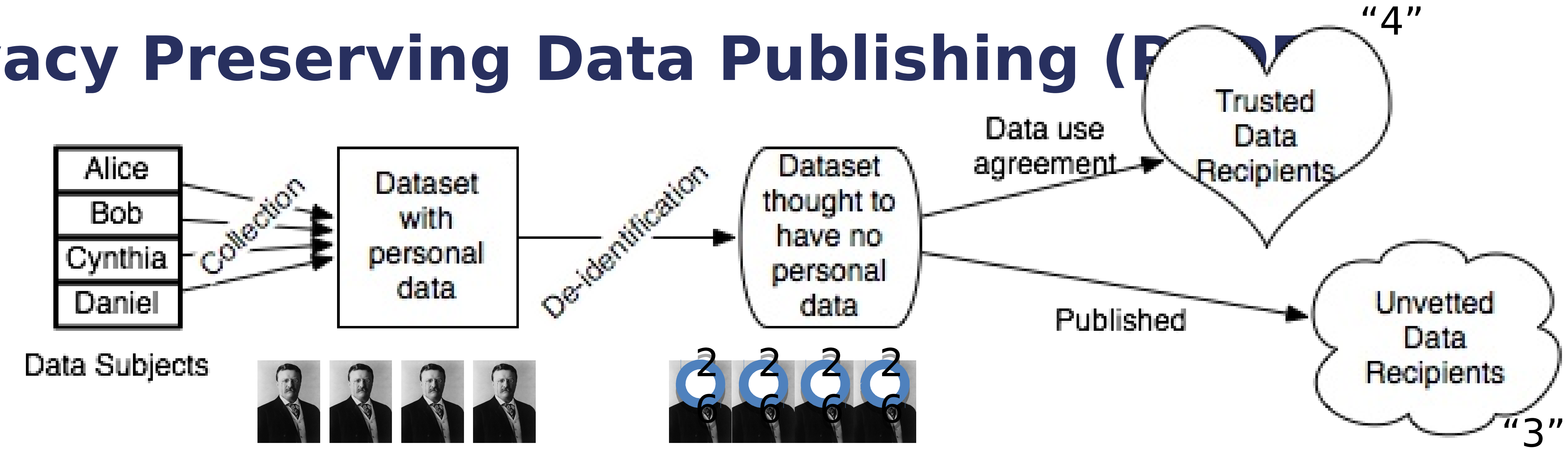
s

There are two approaches for privacy-sensitive data processing.

#1: Privacy Preserving Data Mining (PPDM)



#2: Privacy Preserving Data Publishing (P



#1 — Privacy Preserving Data Mining

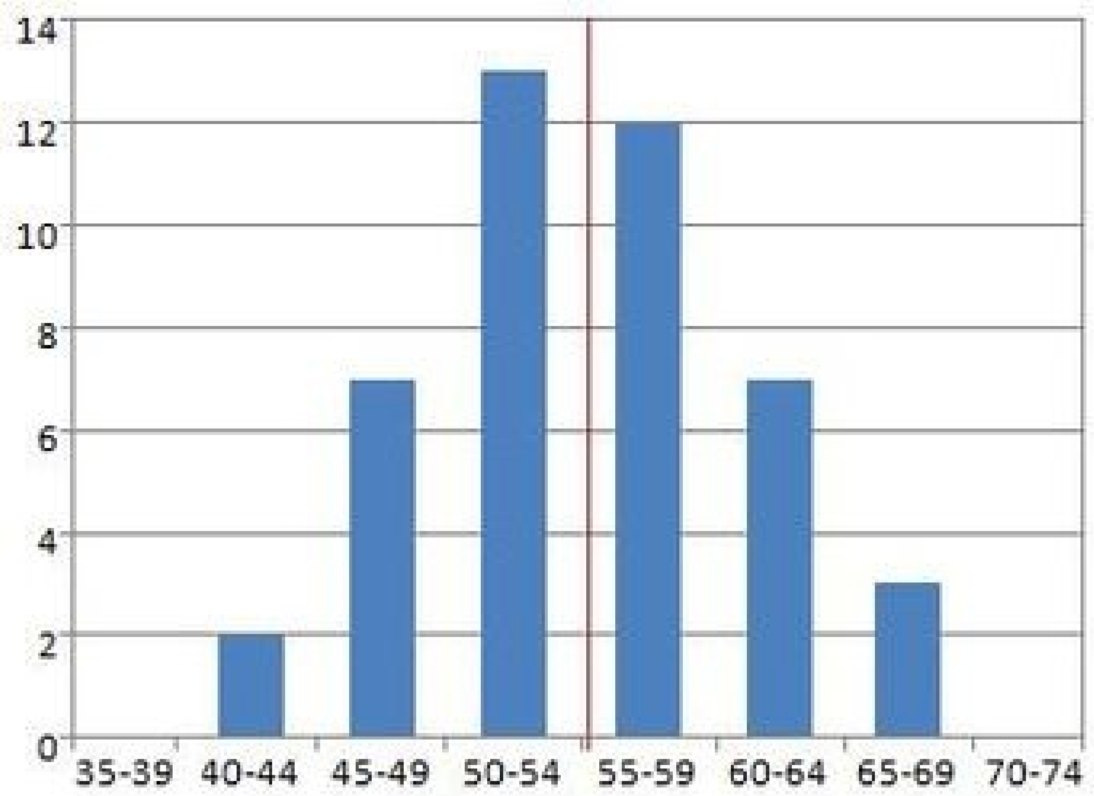
Data are used for statistical processing and machine learning

Data are not released

- Statistical tables, classifiers, other kinds of results
- “The average age at accession of a US president is 54 years and 11 months”

Techniques:

- Statistical Disclosure Control
- Differential Privacy



#2 — Privacy Preserving Data Publishing

Data are released in some form that protects privacy.

- De-identification
  - — *Field suppression, generalization, field swapping*
- Synthetic data generation

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

Start by removing the “directly identifying” information.

Direct Identifiers

Sensitive Values

President	*Birth*	*Estimated IQ*	*Favorite Color*
XXXXX	***February 22, 1732***	***132.5***	***red***
XXXXX	***October 30, 1735***	***142.5***	***blue***
XXXXX	***April 13, 1743***	***153.75***	***green***
XXXXX	***March 16, 1751***	***141.25***	***yellow***
XXXXX	***April 28, 1758***	***124.125***	***red***
XXXXX	***July 11, 1767***	***168.75***	***orange***
XXXXX	***March 15, 1767***	***126.25***	***cyan***
XXXXX	***December 5, 1782***	***133.35***	***blue***

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

Create account Log in

Article Talk Read Edit View history Search

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

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

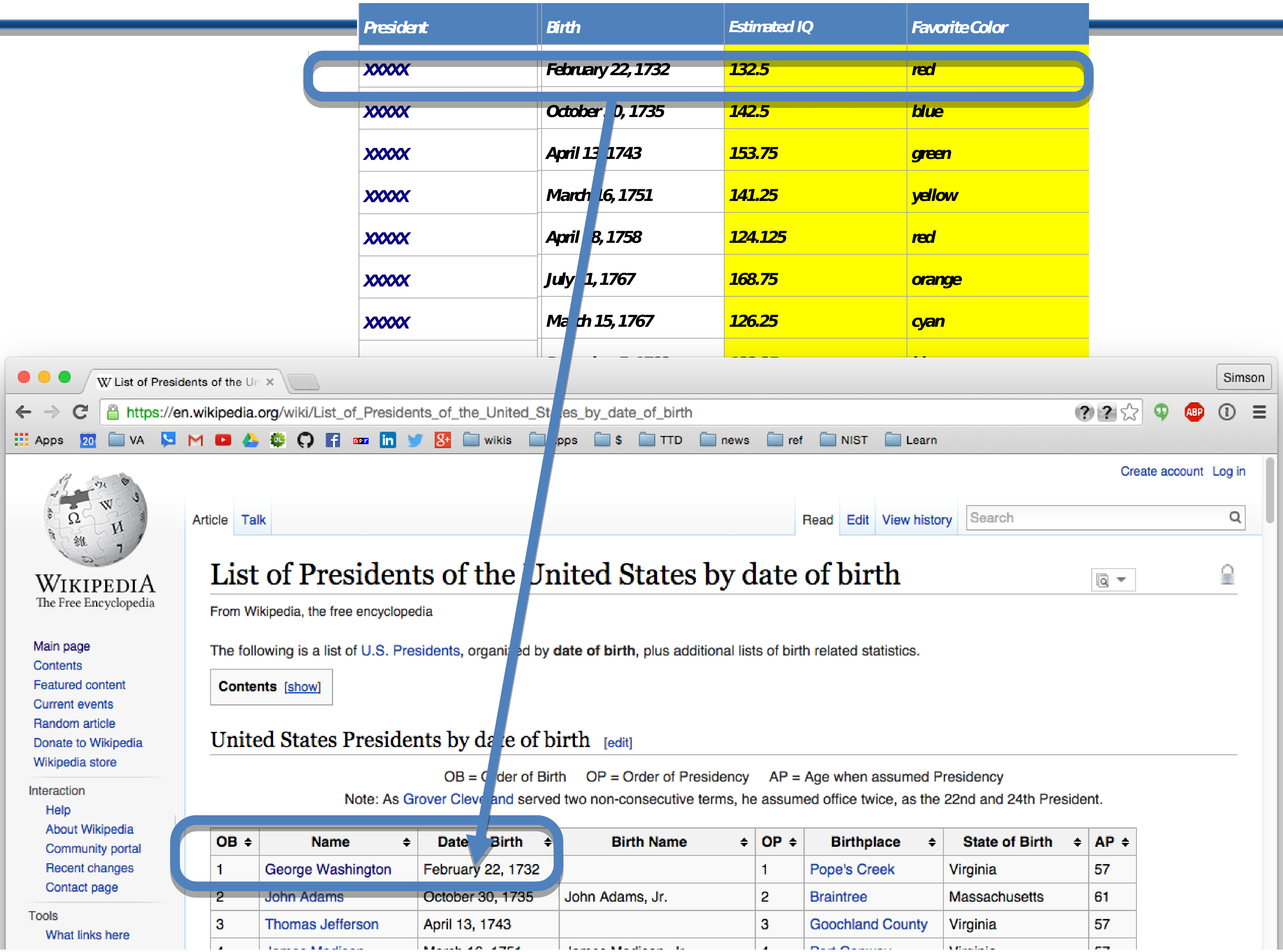
Print/export
Create a book
Download as PDF
Printable version

Languages
Română
中文
Edit links

This is called a “linkage attack.”

“Birth date” is an *indirect identifier.*

Also called a “quasi Identifier.”



Latanya Sweeney performed a linkage attack to re-identify Governor William Weld's hospital records. (2000)

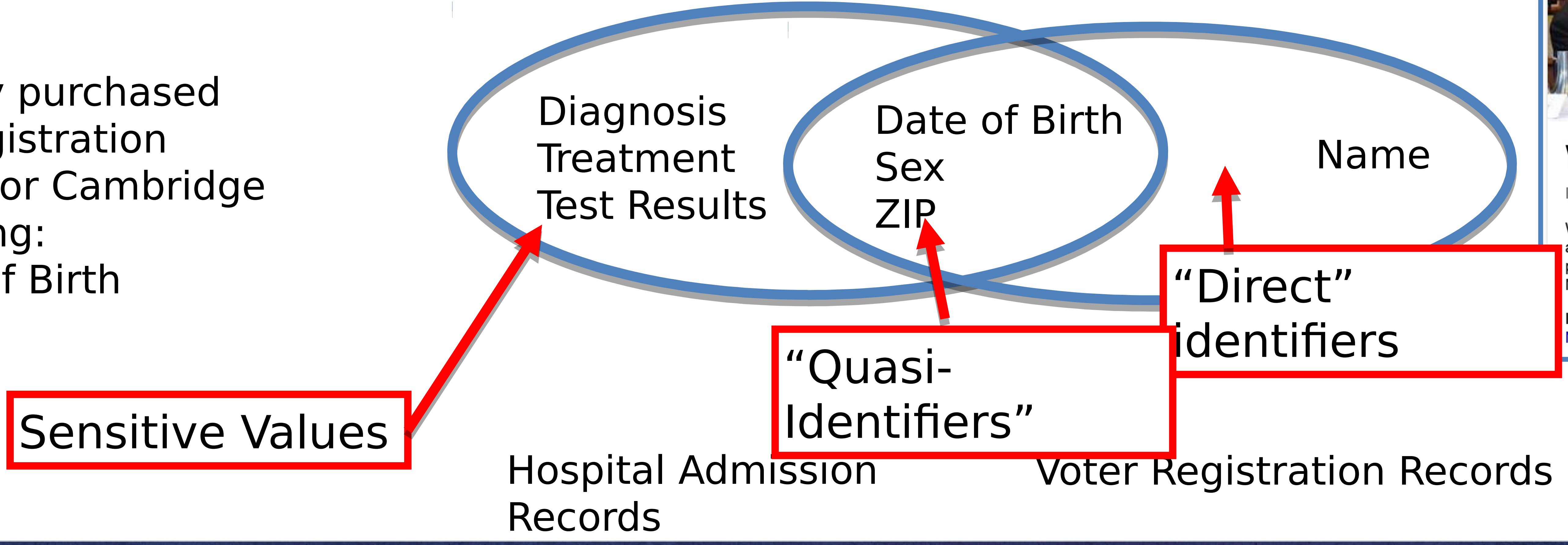
Governor Weld fainted in 1996 at a college graduation and was admitted to a hospital

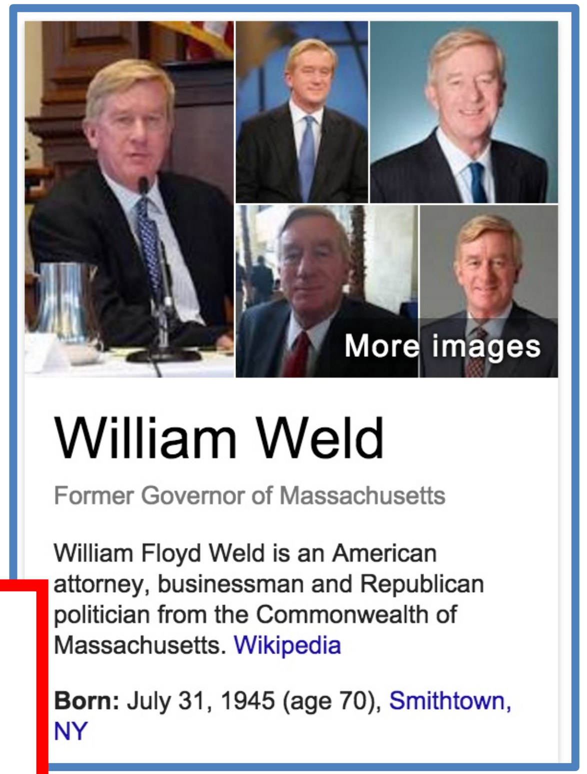
State of MA made "de-identified" hospital records of state employees available for research on health care

- MA Removed name; left Date of Birth, Sex & ZIP

Sweeney purchased voter registration records for Cambridge containing:

- Date of Birth
- Sex
- ZIP





To reduce the risk of re-identification: Remove the DIs; manipulate or remove the QIs.

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
- **Noise Addition** January 1, 1980 → December 21, 1979

The identifiability of a quasi-identifier depends on the availability of additional data.

Researchers examining cancer at a university get this data set from the university's insurance company:

Title	Age	Sex	Address	ICD-10	Diagnosis
	…		[redacted]		…
Lab Tech	35	M	[redacted]	K25.0	Gastric Ulcer with hemorrhage
Lab Tech	56	F	[redacted]	J00	Acute nasopharyngitis [Common Cold]
Professor	35	M	[redacted]	C64.1	Malignant neoplasm of right kidney
Professor	69	F	[redacted]	C64.1	Malignant neoplasm of right kidney
Contracts Specialist	52	F	[redacted]	L30.9	Dermatitis, unspecified [Eczema]
University President	56	F	[redacted]	C64.1	Malignant neoplasm of right kidney
	. . .		[redacted]		. . .

(Hypothetical dataset from university healthcare system)

Re-identified information can link with other, sensitive data.

De-identified research database:

Patient 234-334-11
Diagnostic Codes: A98.4, J00, L30.9

Patient 234-334-11
Age: 35
Genetic History: …

Patient 234-334-11
Psychological Records
…

Patient 234-334-11
Social Services History
…

…

Ebola Patients			ICD-10	Diagnosis
Alice	30	F	A98.4	Ebola
Bob	35	M	A98.4	Ebola
Carol	40	F	A98.4	Ebola

There are four main techniques for modifying data to limit data disclosure.

Title	Age	Sex	Address	ICD-10	Diagnosis
University President	56	F	[redacted]	C64.1	Malignant neoplasm of right kidney

Generalization: University President ⇒ Senior Administrator
Age: 56 ⇒ Age: 50-59

Field Swapping: Age: 52 ⇒ Age: 56
Age: 56 ⇒ Age: 52

Noise addition: University President ⇒ VP Finance
Age: 56 ⇒ Age: 58 ±5

Suppression: University President ⇒ XXXXXXXXXXXXX
Age: 56 ⇒ Age: XXX

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%

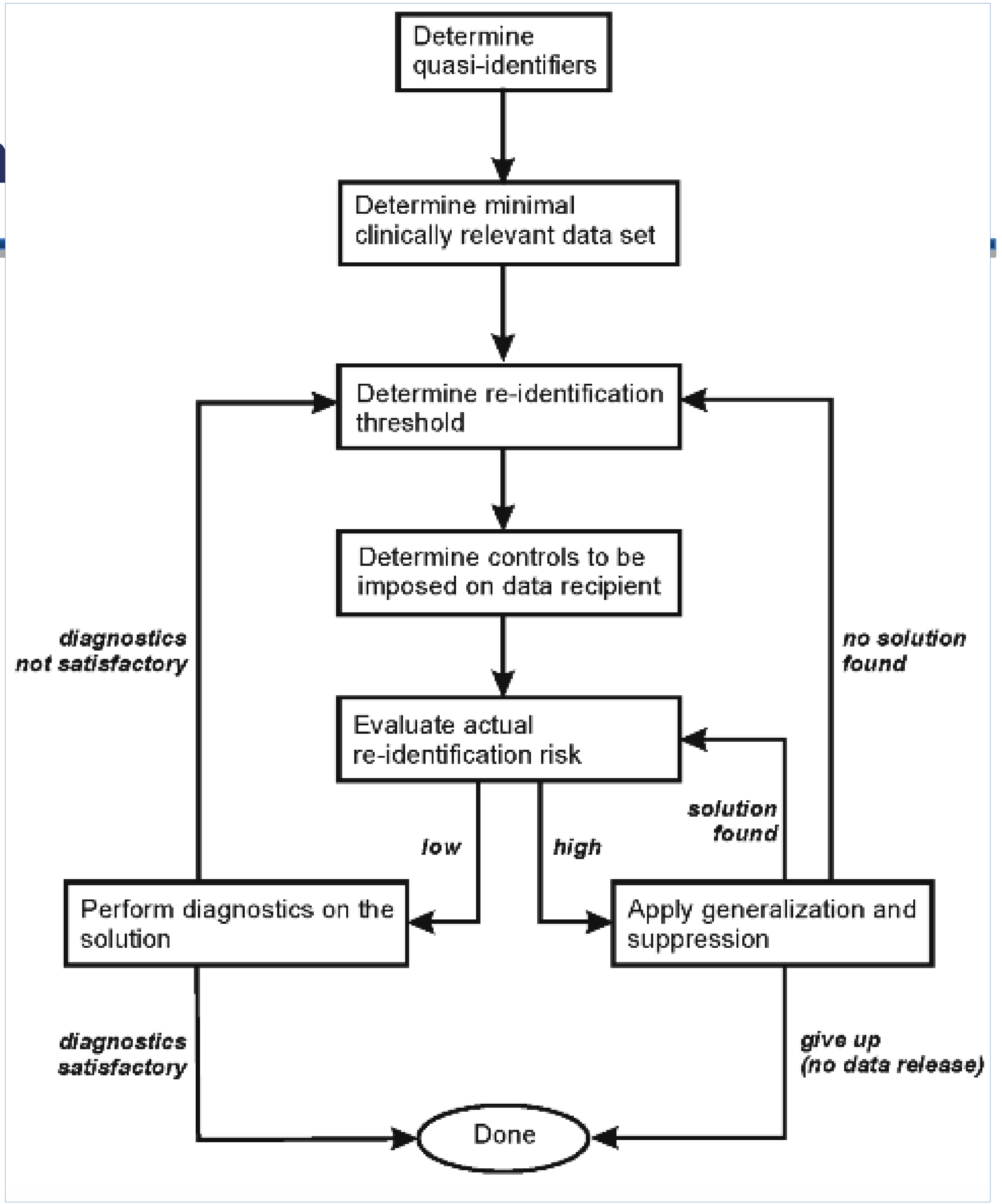
Calculating re-identification risk:
There are several risk assessment m

Risk of record re-identification = 1 / ---------------------------- # possible matching records in population

Must be calculated for every record.

Key issues:

- Definition of “matching”
- Definition of “population”



Evaluating the Risk of Re-identification of Patients from Hospital Prescription Records, El Emam et al, Can J Hosp Pharm 2009;62(4):307–319

HIPAA “Limited Dataset:” Removes less information / 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:

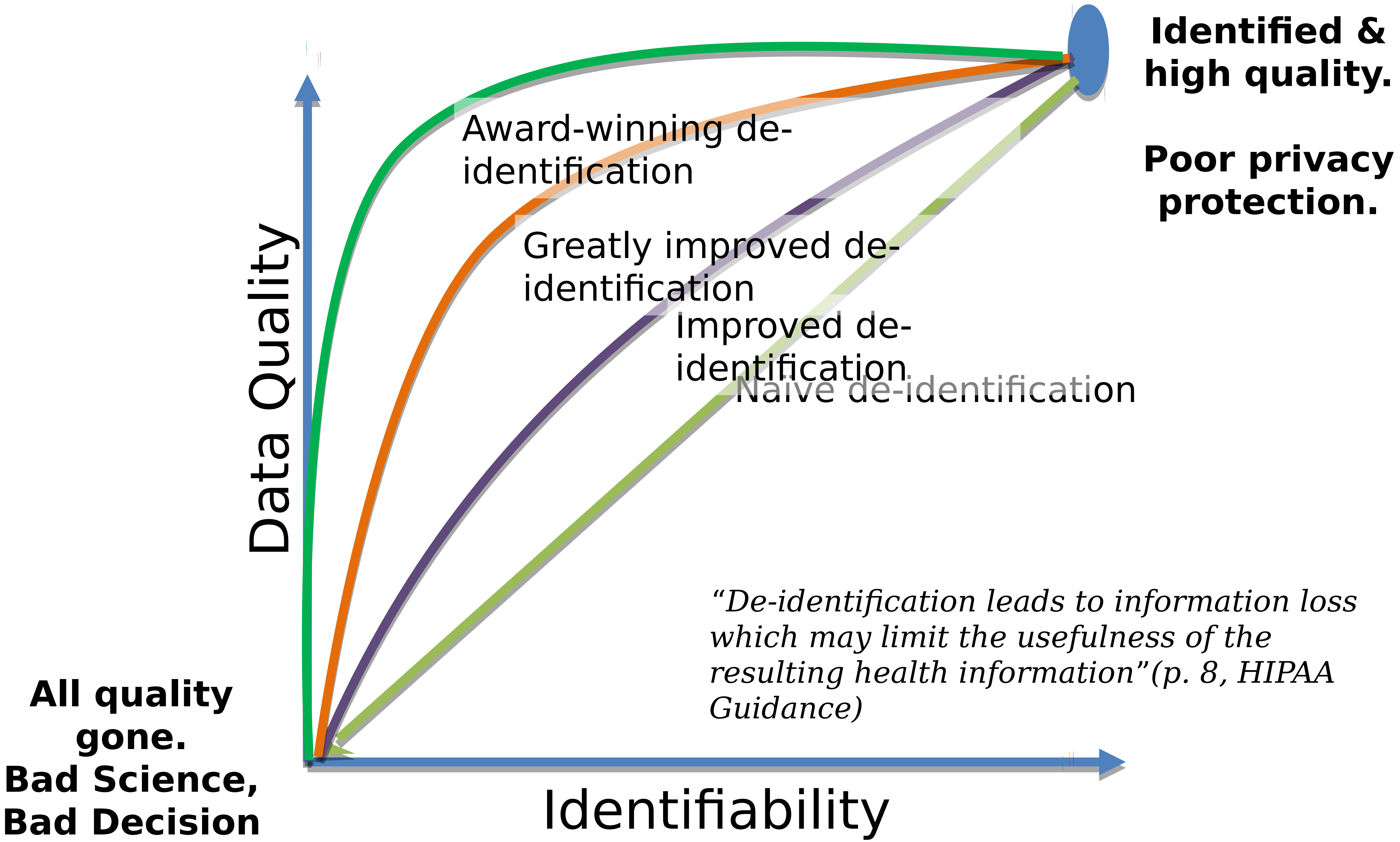
- *Without a patient’s authorization or notification*
- *But…*

Must have a **data use agreement** in place:

- *Cannot release the data set*
- *Cannot share with others without a DUA*

Higher data quality
&
Higher identifiability

Lowering identifiability lowers data quality.



"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

For further information

Direct Identifiers

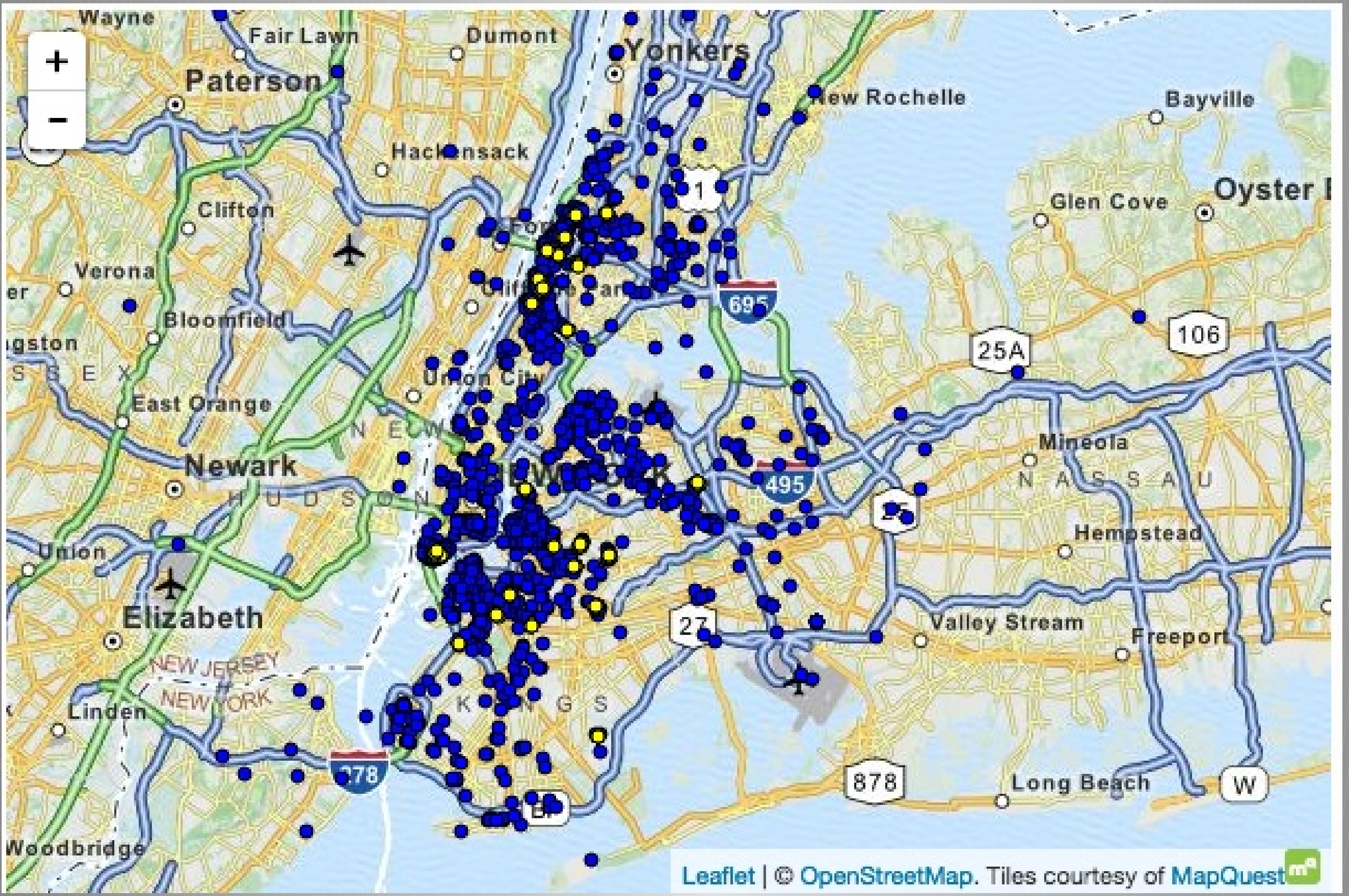
Quasi-Identifiers

Field Suppression

Generalization

Data Swapping

Data quality / Identifiability tradeoff



Famous re-identification controversies.

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

Commercial Benefit

- gain publicity or professional standing

Harm or embarrass the de-identifying organization

Harm the data subject

Theodore Roosevelt

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 have cancer.
  - *Jane is a 20-year-old female patient from Q.*
  - *∴Jane has cancer.*

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.
- Risks: someone/anyone might try to re-identify

Data Use Agreement (DUA) model:

- Users assert that they will not attempt to re-identify.
- Risks: rogue insider; inadvertent re-identification; data breach.

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 transacti
(Montjoy et al.)

Cell phone mobility tra
(Montjoy et al.)

Taxi ride data —
NYC Taxi & License Co

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]			
1	Bob Smith	Payscale	1/4/05	23:41
	John Doe	Popcorn	1/8/0	03:1

For each user, AOL released their “query string” and other information.

User 2178
foods to avoid when breast feeding

User 3505202
depression and medical leave

User 47122
Child porno

User 31350
How to kill oneself by natural gas

User 3482401
calorie counting

User 7268042
fear that spouse contemplating cheating

User 3483689
Time after time

User 3483689
Wind beneath my wings

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

http://www.nytimes.com/2006/08/09/technology/09aol.html

User 44177

Nu

rs

su

Thelma Arnold
& Dudley

Timeline of the controversy

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

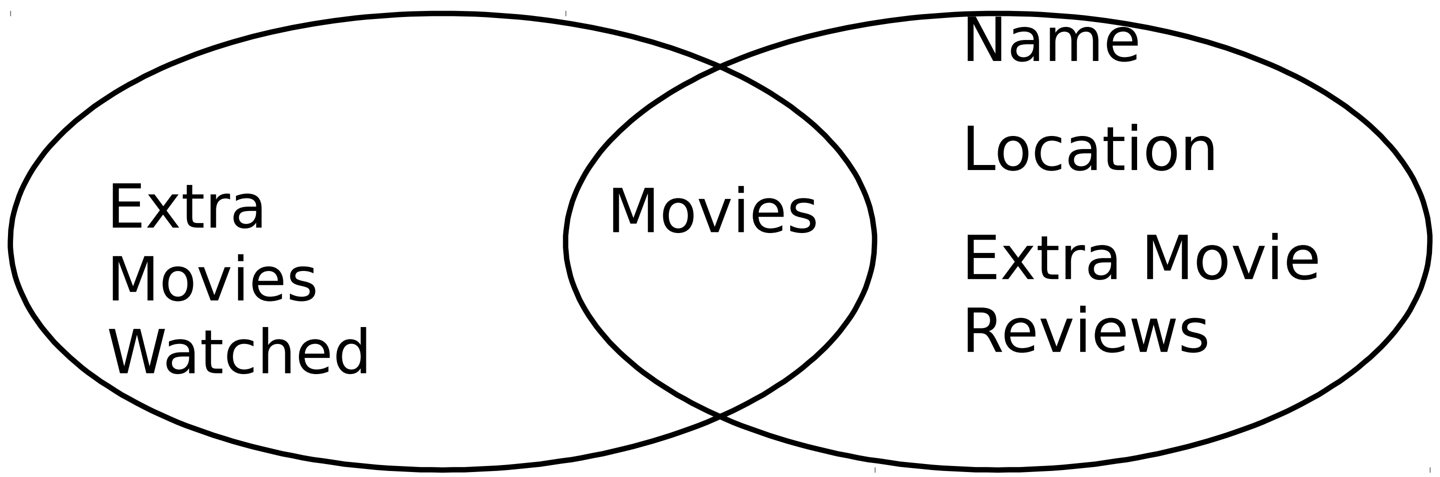
Sept 2006

Class Action Law Suit Filed

The Netflix Challenge (2008-2009)

Netflix published movie selections of ~450,000 pseudonymized subscribers

Re-identification via uniqueness of movie combinations



Netflix Challenge

Internet Movie Database

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

Forbes.com
U.S. EUROPE ASIA

Home Lists Business Tech

Breakthr

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

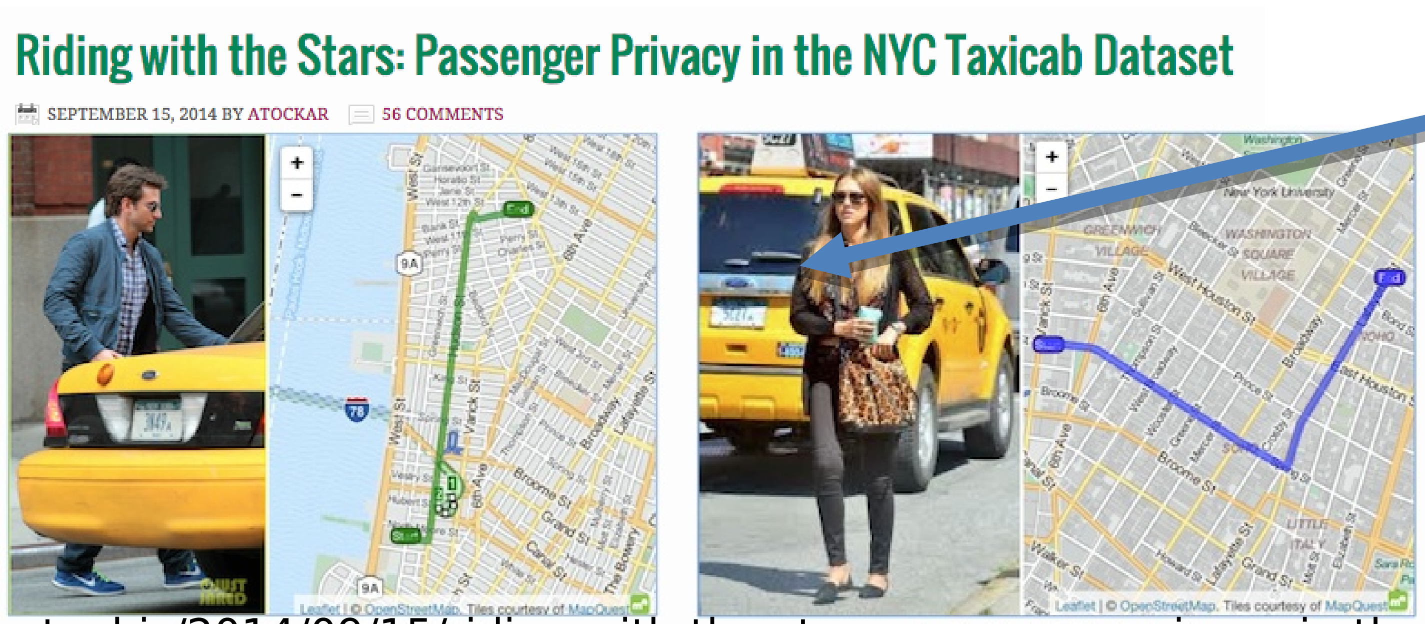
On Friday, Netflix announced on its corporate blog that it has settled a lawsuit related to its Netflix Prize, a $1 million contest that challenged machine learning experts to use Netflix's data to produce better recommendations than the movie giant could serve up themselves.

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”

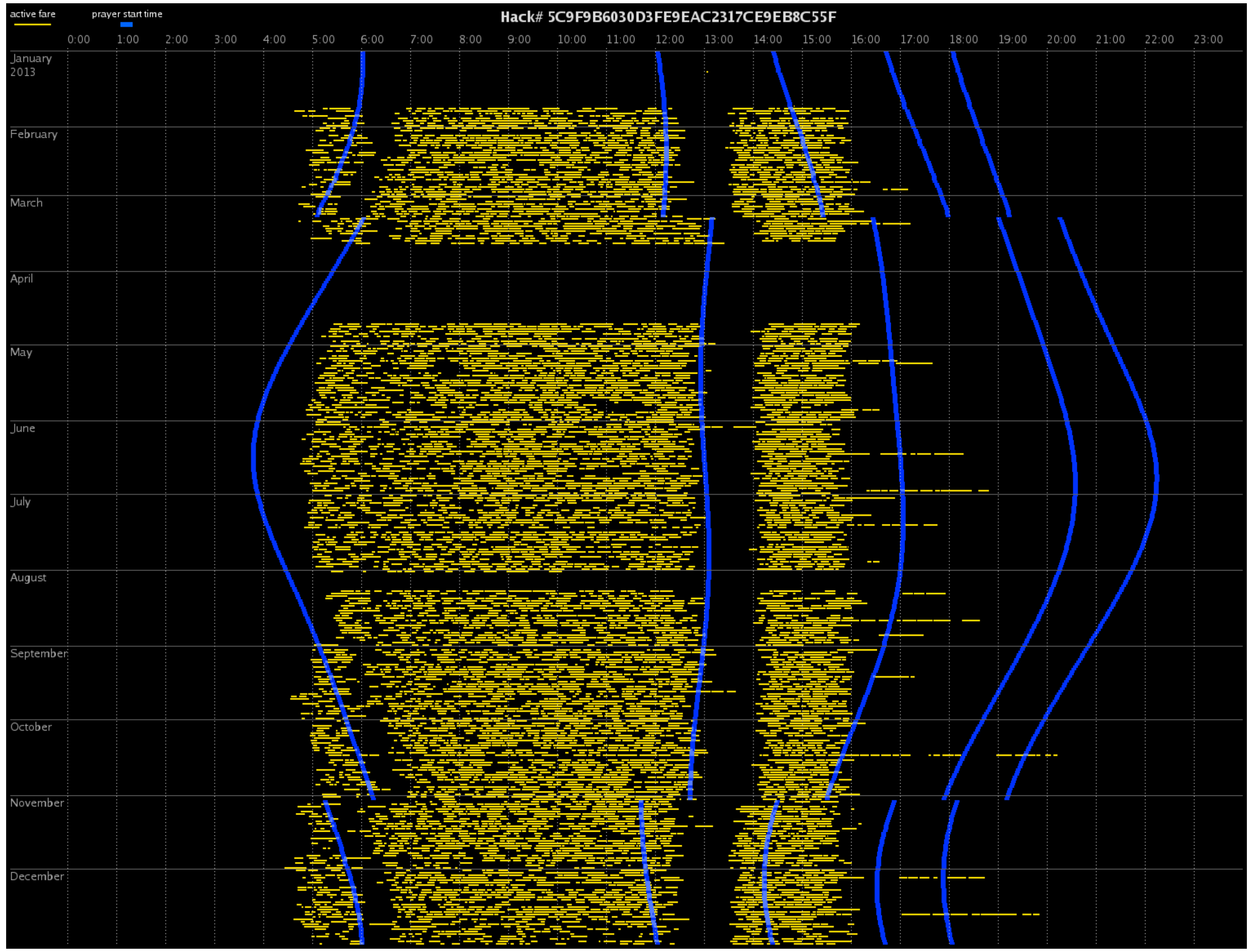
A journalist at Gawker identified 9 other cab rides.

https://research.neustar.biz/2014/09/15/riding-with-the-stars-passenger-privacy-in-the-nyc-taxicab-dataset/

47

Time series data can have unanticipated revelations. Breaks in taxi driving “pinpoint” Muslim cab drivers

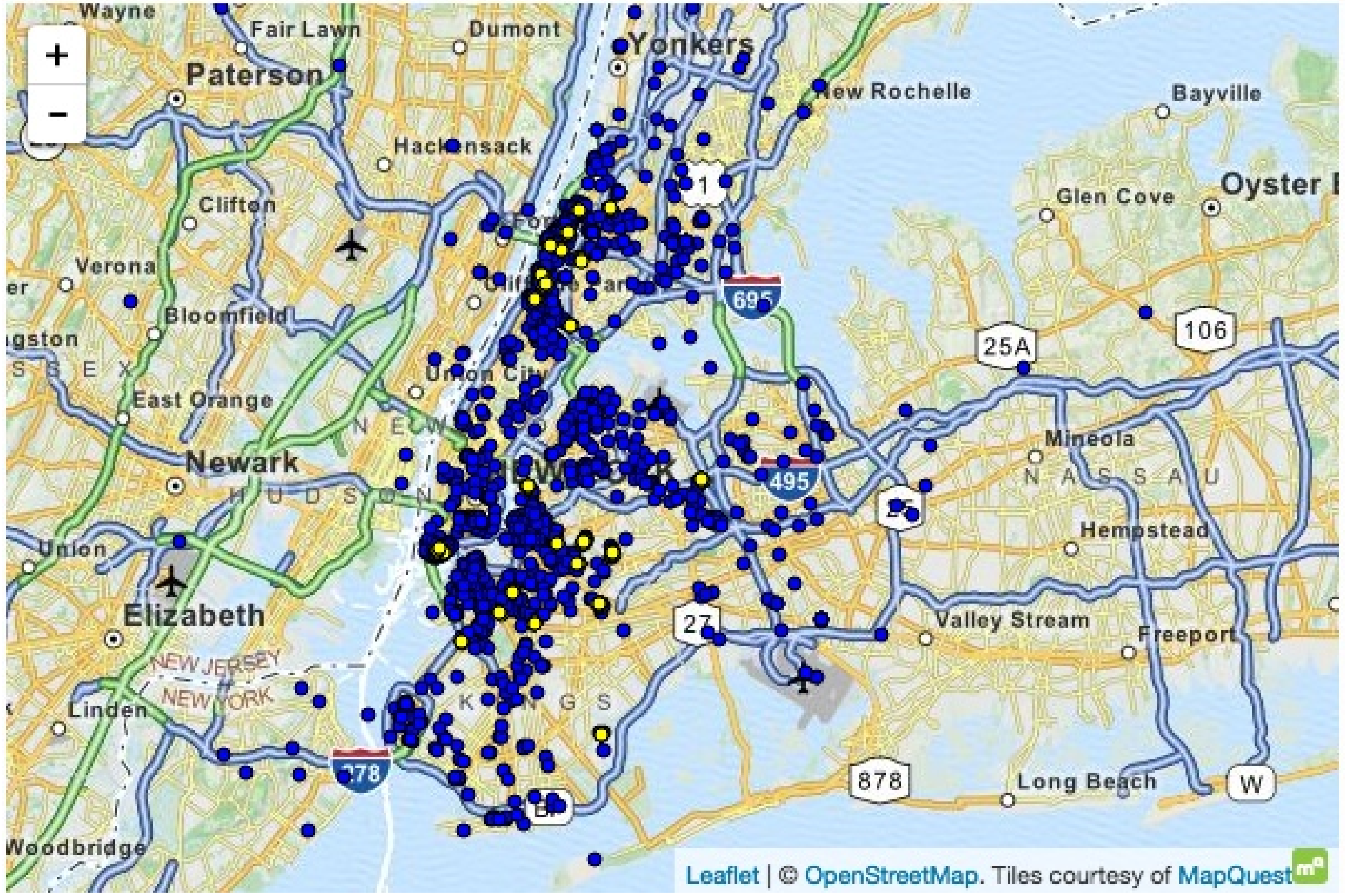
Half of all taxi drivers in NYC are Muslim, but there is no obvious correlation of taxi trips with call to prayer times:



For some drivers, there is an obvious correlation.

http://mashable.com/2015/01/28/redditor-muslim-cab-drivers

The trips alone identify pickups and drop-offs at Larry Flynt's Hustler Club



http://content.research.neustar.biz/blog/differential-privacy/stripRaw.html

Frequent customer

Occasional customer

Google Street view

Re-identification by linking is more complex than

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.*

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

Keep these points in mind when evaluating a re-identification attack…

Sample unique ≠ population unique

- Re-identification attacks are based on using quasi-identifiers to link “uniques”
- Being “unique” in a sample does *not* imply being unique in the population.

To be effective, person must exist in the linked data set.

To be accurate, the attack must be verified.

- A test of the HIPAA standard found 20 matches in 15,000, but only 2 of the matches were real.

Outline for today’s talk

Why de-identify? ✔

Basic de-identification ✔

Famous re-identification controversies ✔

De-identification in practice

Measuring re-identification risk

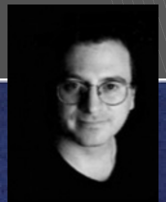
For further information.

High-profile re-identifications

The number of people re-identified was relatively small

Disproportional impact.

De-identification in practice

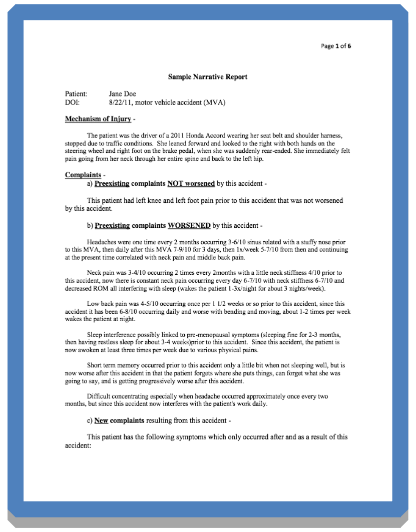


NISTIR 8053 discusses de-identification of many kinds of unstructured data.

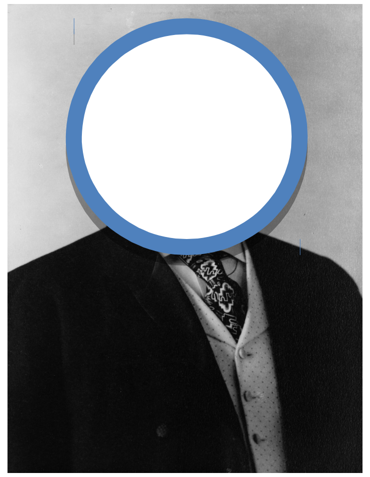
OB	Name	Date of Birth	Birth Name	OP	Birthplace	State of Birth	AP
1	George Washington	February 22, 1732		1	Pope's Creek	Virginia	57
2	John Adams	October 30, 1735	John Adams, Jr.	2	Braintree	Massachusetts	61
3	Thomas Jefferson	April 13, 1743		3	Goochland County	Virginia	57

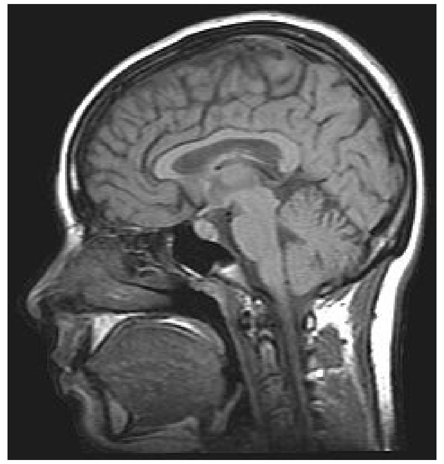
Tabular information (structured data)

Free-form medical text


Sample Narrative Report

Photographs and Video

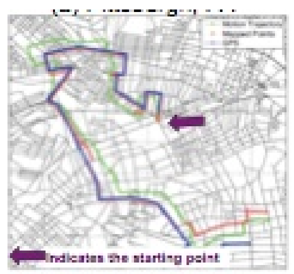




Medical Imagery

Genetic information

Geographic and map data



Medical text — de-identifying medical narratives

Challenges:

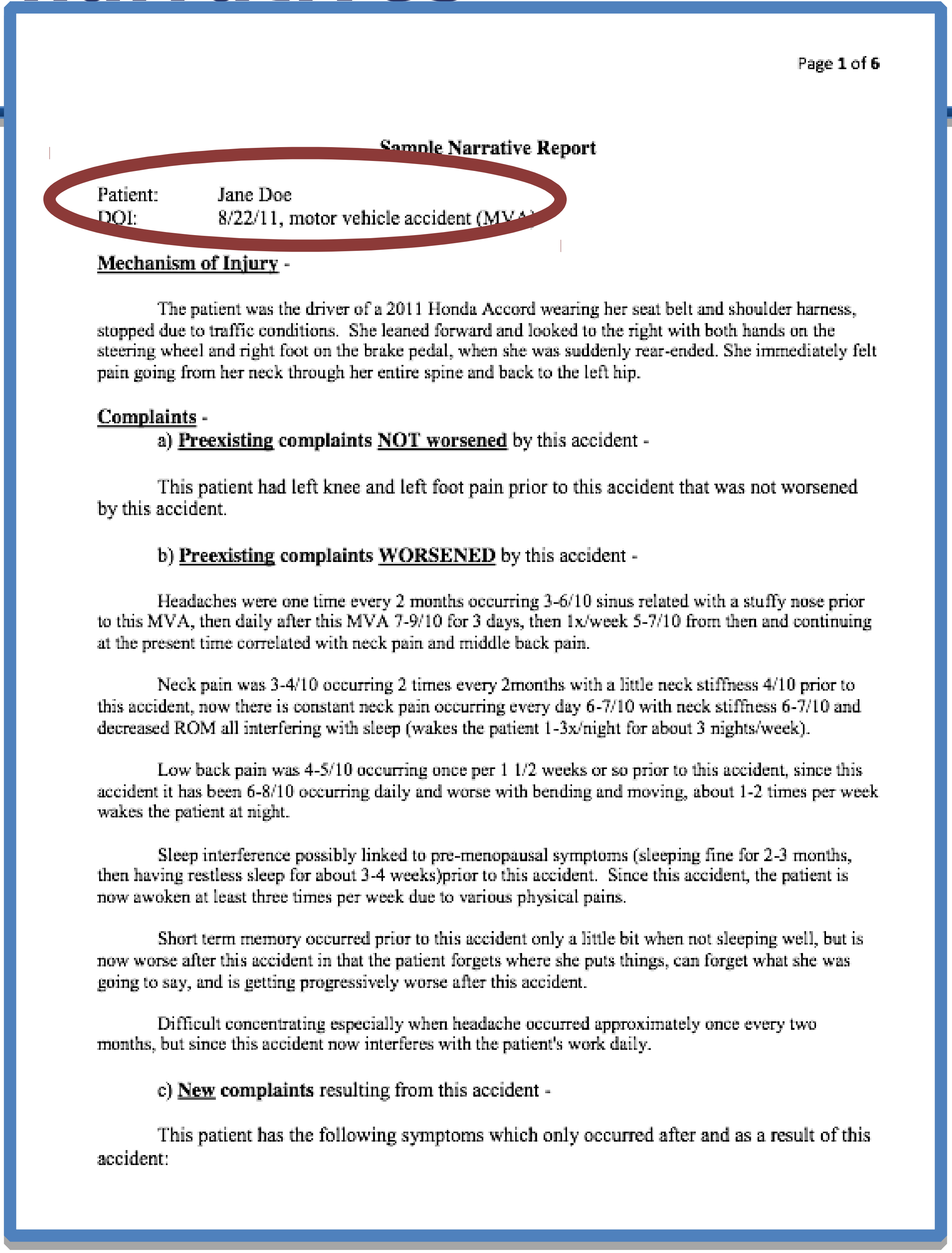
- Finding the direct identifiers
- Not removing important medical information like eponyms. (e.g. “Addison’s Disease”)

NL Approaches:

- Rule-based (e.g. regex)
- Statistical machine learning.

Several evaluations.

Success rate ≈ 95%



Page 1 of 6

Sample Narrative Report

Patient: Jane Doe
DOI: 8/22/11, motor vehicle accident (MVA)

Mechanism of Injury -

The patient was the driver of a 2011 Honda Accord wearing her seat belt and shoulder harness, stopped due to traffic conditions. She leaned forward and looked to the right with both hands on the steering wheel and right foot on the brake pedal, when she was suddenly rear-ended. She immediately felt pain going from her neck through her entire spine and back to the left hip.

Complaints -

a) **Preexisting complaints NOT worsened** by this accident -

This patient had left knee and left foot pain prior to this accident that was not worsened by this accident.

b) **Preexisting complaints WORSENED** by this accident -

Headaches were one time every 2 months occurring 3-6/10 sinus related with a stuffy nose prior to this MVA, then daily after this MVA 7-9/10 for 3 days, then 1x/week 5-7/10 from then and continuing at the present time correlated with neck pain and middle back pain.

Neck pain was 3-4/10 occurring 2 times every 2months with a little neck stiffness 4/10 prior to this accident, now there is constant neck pain occurring every day 6-7/10 with neck stiffness 6-7/10 and decreased ROM all interfering with sleep (wakes the patient 1-3x/night for about 3 nights/week).

Low back pain was 4-5/10 occurring once per 1 1/2 weeks or so prior to this accident, since this accident it has been 6-8/10 occurring daily and worse with bending and moving, about 1-2 times per week wakes the patient at night.

Sleep interference possibly linked to pre-menopausal symptoms (sleeping fine for 2-3 months, then having restless sleep for about 3-4 weeks)prior to this accident. Since this accident, the patient is now awoken at least three times per week due to various physical pains.

Short term memory occurred prior to this accident only a little bit when not sleeping well, but is now worse after this accident in that the patient forgets where she puts things, can forget what she was going to say, and is getting progressively worse after this accident.

Difficult concentrating especially when headache occurred approximately once every two months, but since this accident now interferes with the patient's work daily.

c) **New complaints** resulting from this accident -

This patient has the following symptoms which only occurred after and as a result of this accident:

“Hiding in plain sight” approach replaces identifiers with fake identifiers.

Conventional

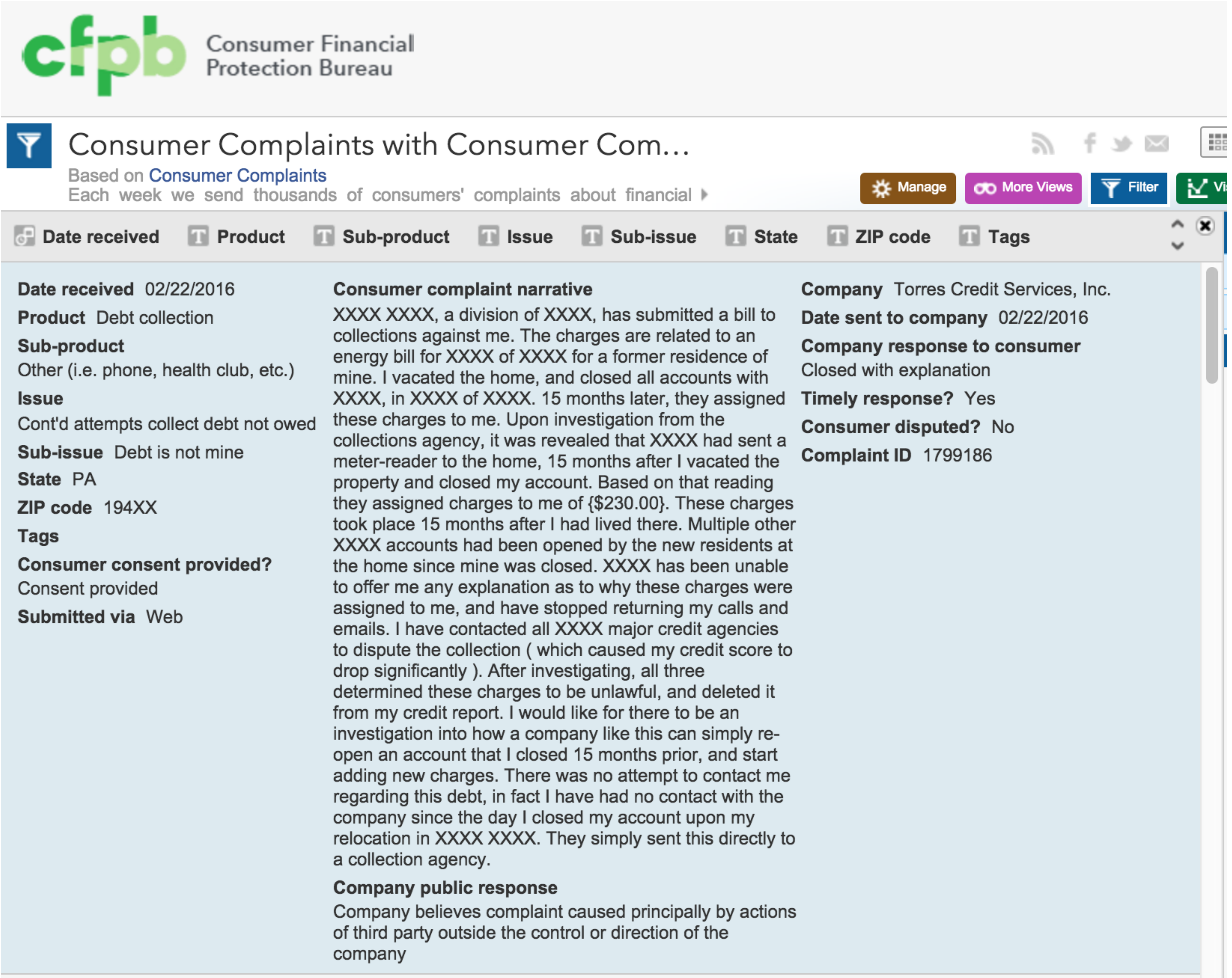
HISTORY OF PRESENT ILLNESS: The patient is a 77-year-old woman with long standing hypertension who presented as a walk-in to me at the Oak Valley Health Center on July 9th. Recent had been started q.o.d. on Clonidine since May 5th to tape off of the drug. Was told to start Zestril 20 mg. q.d. again. The patient was sent to the Smith Cardiac Unit for direct admission for cardioversion and anticoagulation, with the Cardiologist, Dr. Pearson to follow

Hiding in plain sight

HISTORY OF PRESENT ILLNESS: The patient is a 77-year-old woman with long standing hypertension who presented as a walk-in to me at the Janice Joplin Outpatient Center on March 15th. Recent had been started q.o.d. on Clonidine since January 10th to tape off of the drug. Was told to start Zestril 20 mg. q.d. again. The patient was sent to the Boston City Hospital for direct admission for cardioversion and anticoagulation, with the Cardiologist, Dr. Hand to follow

Carrell D, Malin B, Aberdeen J, et al. Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text. Journal of the American Medical Informatics Association : JAMIA. 2013;20(2):342-348. doi:10.1136/amiajnl-2012-001034.

Text De-identification today: Consumer Complaint Database



Multimedia de-identification / redaction is an area of growing concern.

The primary interest is 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

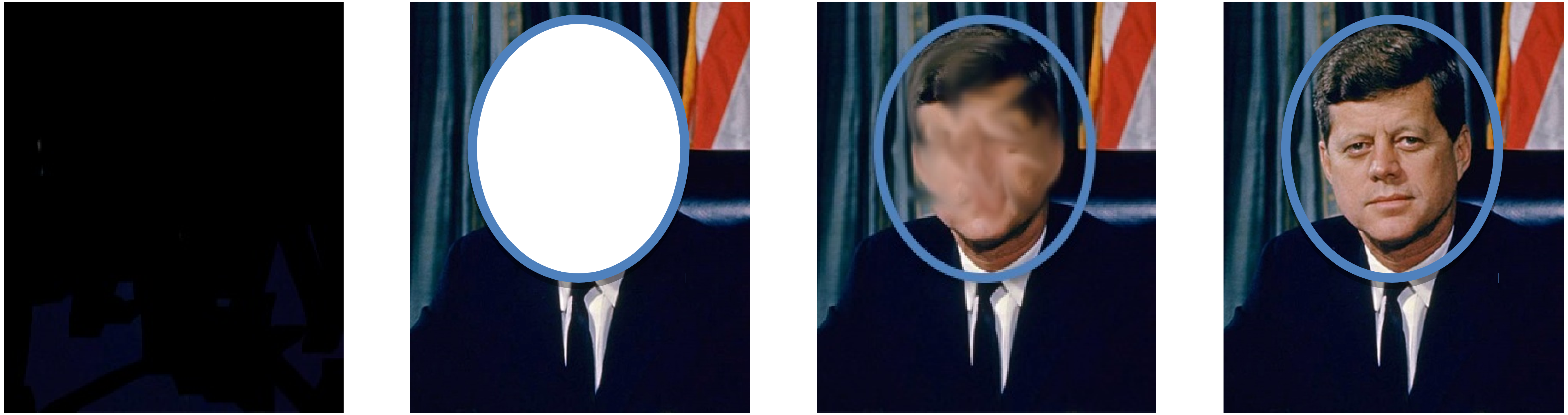
Other uses:

- Scientific research; privacy preserving surveillance; data retention

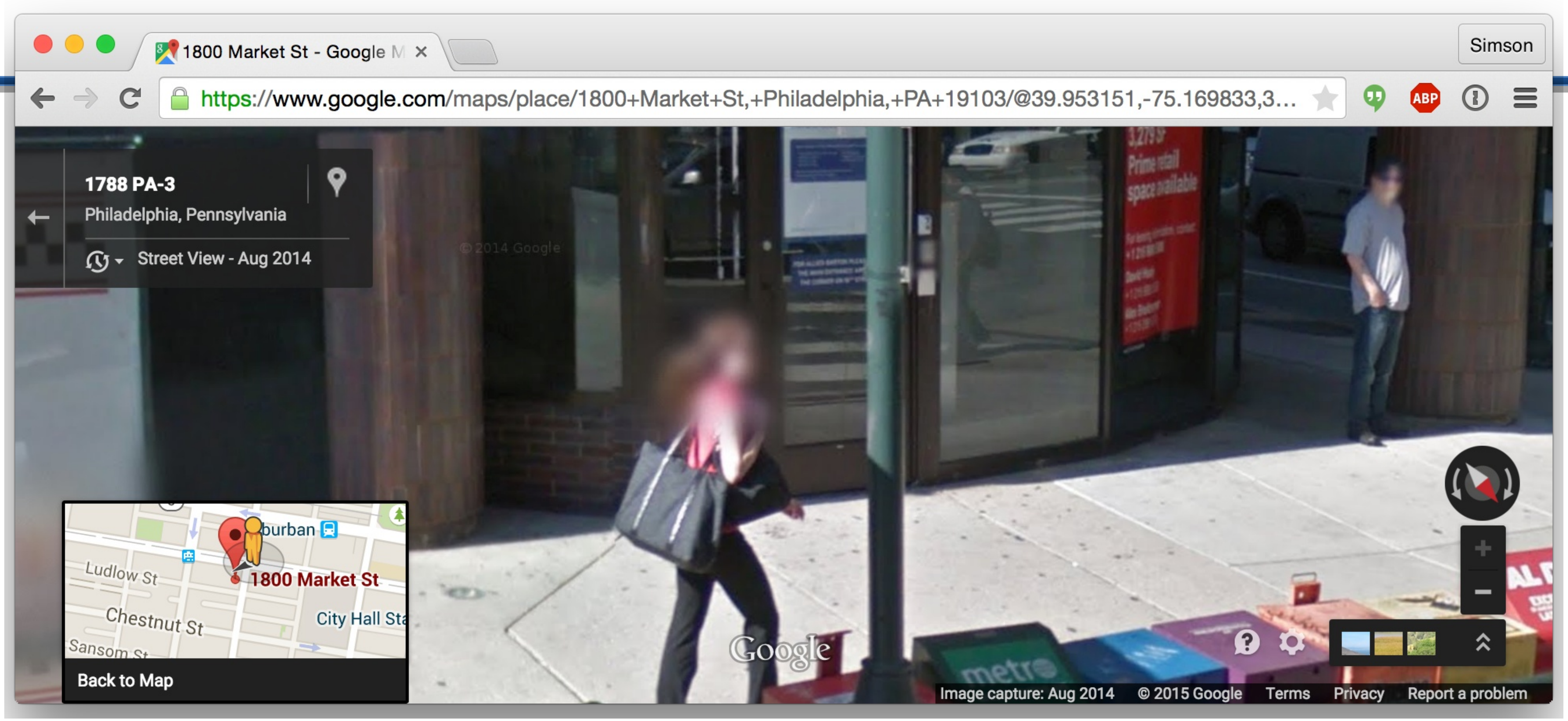
De-identifying photographs and video

Key challenges:

- What to remove?
- Usefulness of de-identified imagery
- Evaluation of the de-identification techniques / software / specific effort



Step 1: Detect *what* to obscure:

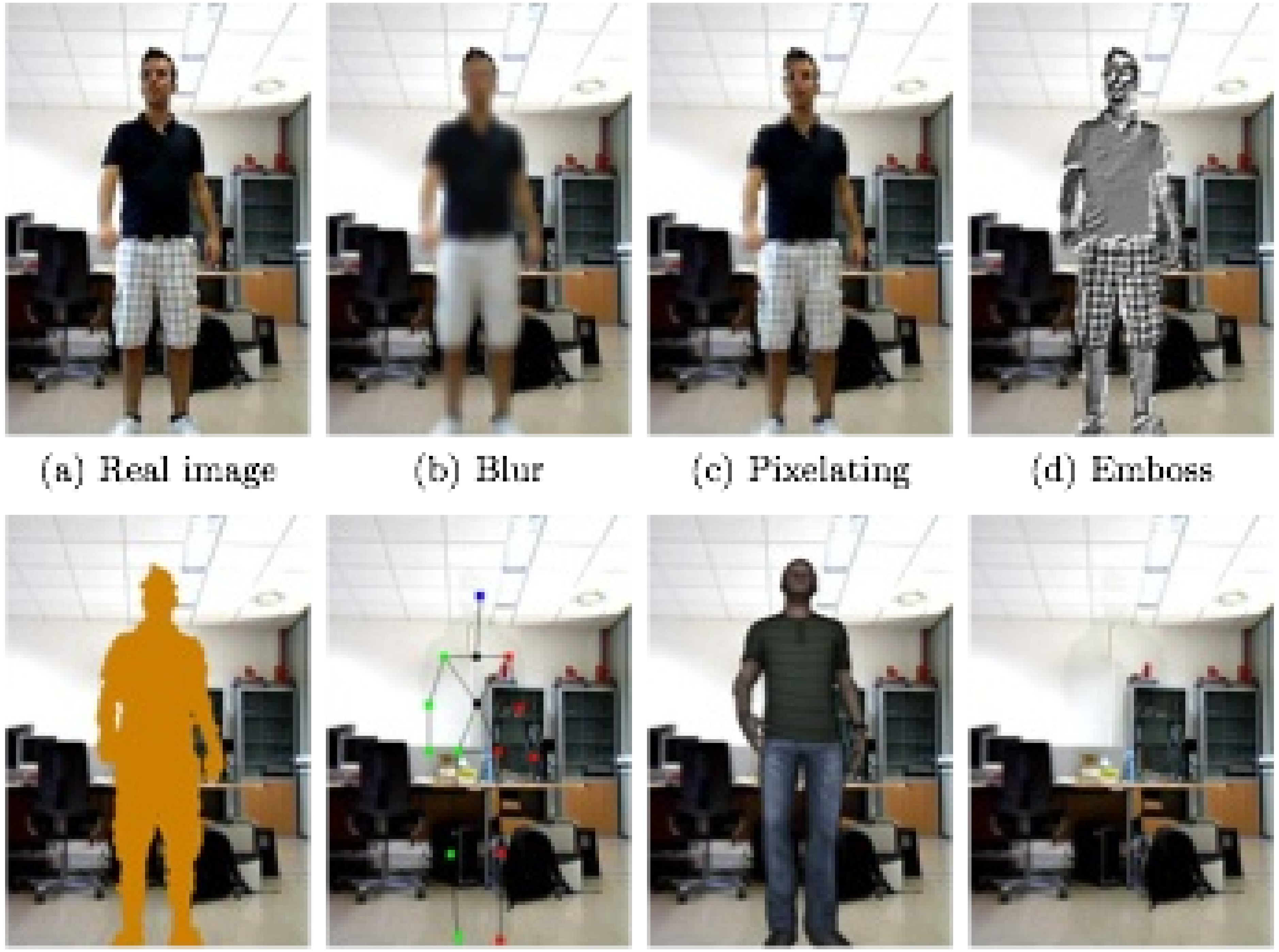


"Large-scale Privacy Protection in Google Street View," Frome et al, 2009

Most research has focused on faces and license plates

- Google's Street View — 90% of faces; 95% of license plates

Step 2: Determine *how* to obscure:



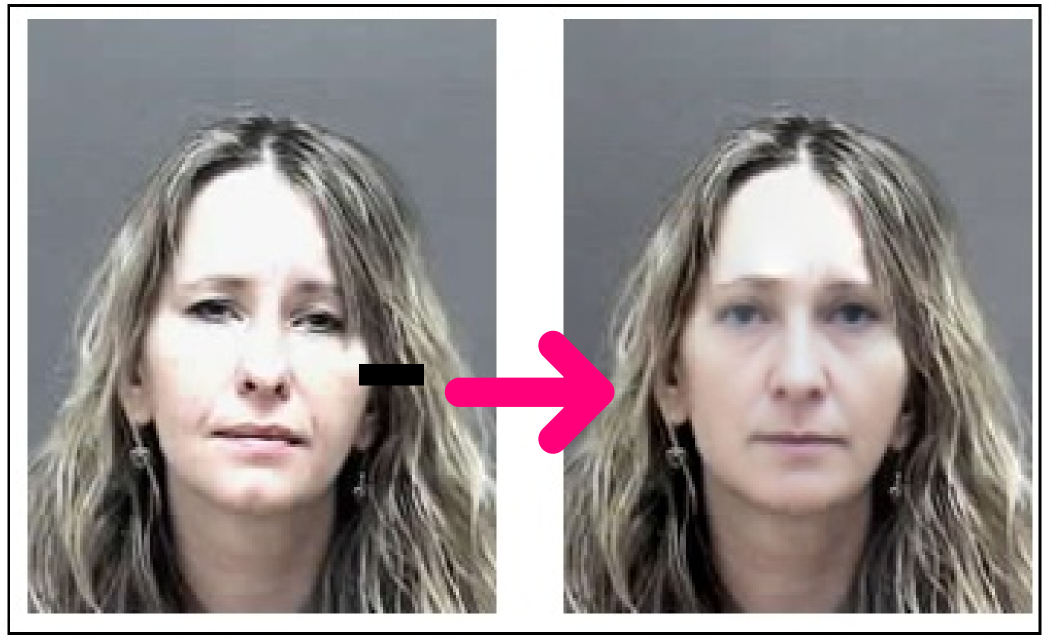
(a) Real image (b) Blur (c) Pixelating (d) Emboss

(e) Solid silhouette (f) Skeleton (g) 3D avatar (h) Invisibility

Obscuring with synthetic faces: preserves context, prevents automated identification

These techniques can preserve:

- Gender
- Race
- Age



White/Female/Middle-aged

Effectiveness:

- + Stops automated face identification.
- - Humans can still identify people they know

Black/Male/Youth

De-identifying medical imagery: Imagery may contain identifying information

Three kinds of identifying information:

- Metadata (DCOM)
- “Burned in”
- Biometrics

http://www.randomhistory.com/photos/2014/scoliosis-xray.jpg

Genetic identification:
People can be identified without being sequenced!

sign in subscribe jobs US edition

theguardian

US world opinion sports soccer tech arts lifestyle fashion business

home › science

Genetics

Teenager finds sperm donor dad on internet

Ian Sample, science correspondent

@iansample

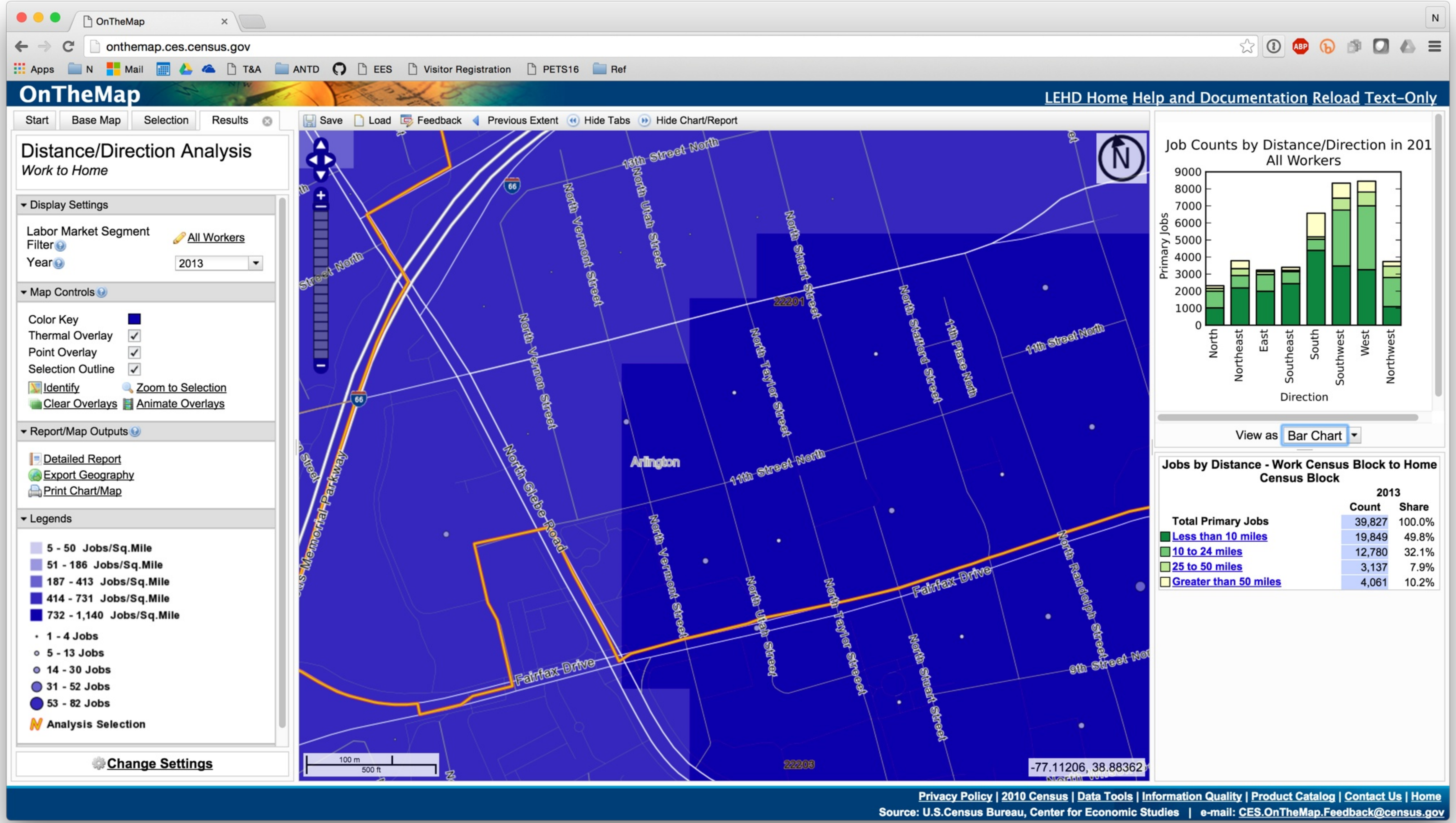
Wednesday 2 November 2005 20.42 EST

Save for later Shares 3

Using nothing more than a swab of saliva and the internet, a 15-year-old boy has tracked down his anonymous sperm donor father, according to details released today.

By sending a swab taken from the inside of his cheek for genetic testing, the teenager was able to use genealogy websites to trace his father by looking for men with a matching Y-chromosome, which is passed down the male line.

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



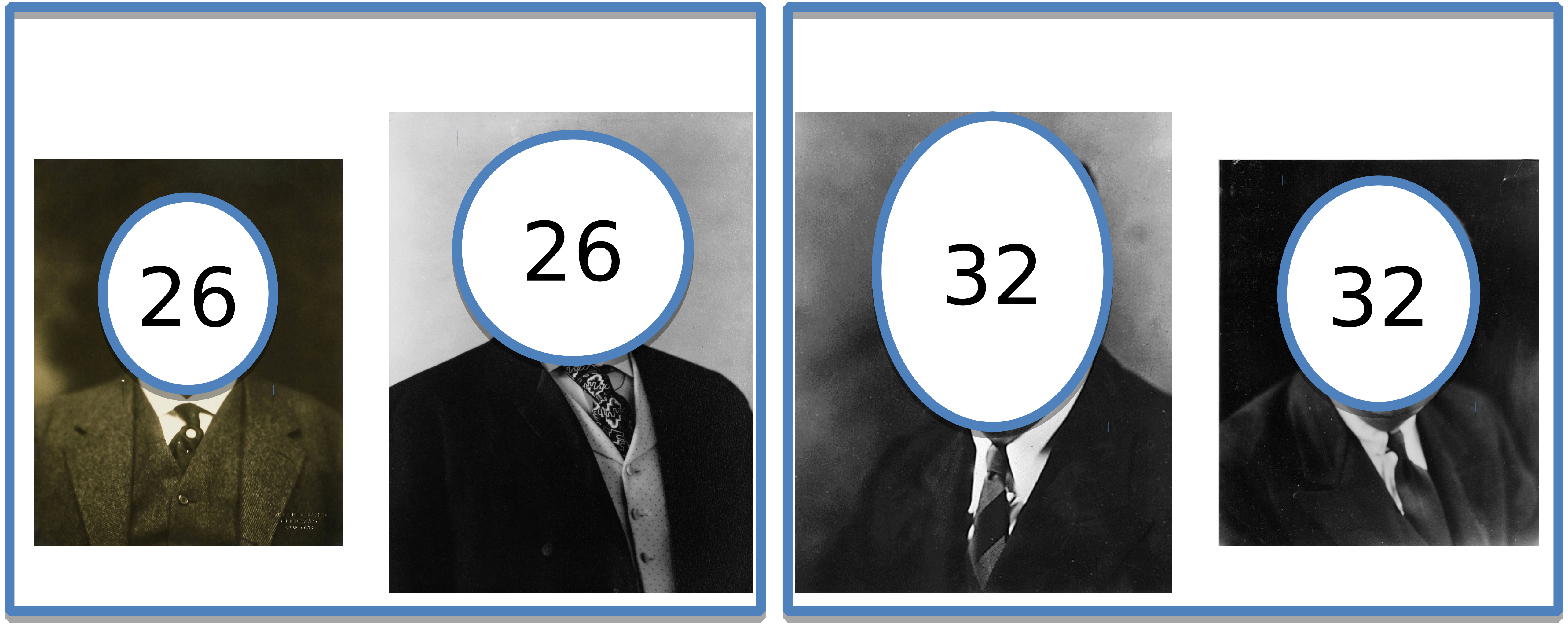
Pseudonymization — de-identification that allows re-identification.

Identifiers are replaced with pseudonyms.

- Sometimes called “coded data.”

Pseudonyms match multiple records belonging to the same individual.

Useful for time series data.



The “code book” can be used to re-identify.

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
[redacted]	Azariah
[redacted]	[redacted]
[redacted]	[redacted]
192	[redacted]
50	[redacted]
181	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
477	Casey
499	Remy

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

Khaled El Emam's de-identification protocol

1. 1: Classify variables
2. 2: Pseudonymize or Remove Direct Identifiers
3. 3: K-Anonymize the Indirect Identifiers
4. 4: Perform a Motivated Intruder Test
5. 5: Update the De-identification

6. https://iapp.org/news/a/a-de-identification-protocol-for-open-data/

Outline for today's talk

Why de-identify? ✔

Basic de-identification ✔

Famous re-identification controversies ✔

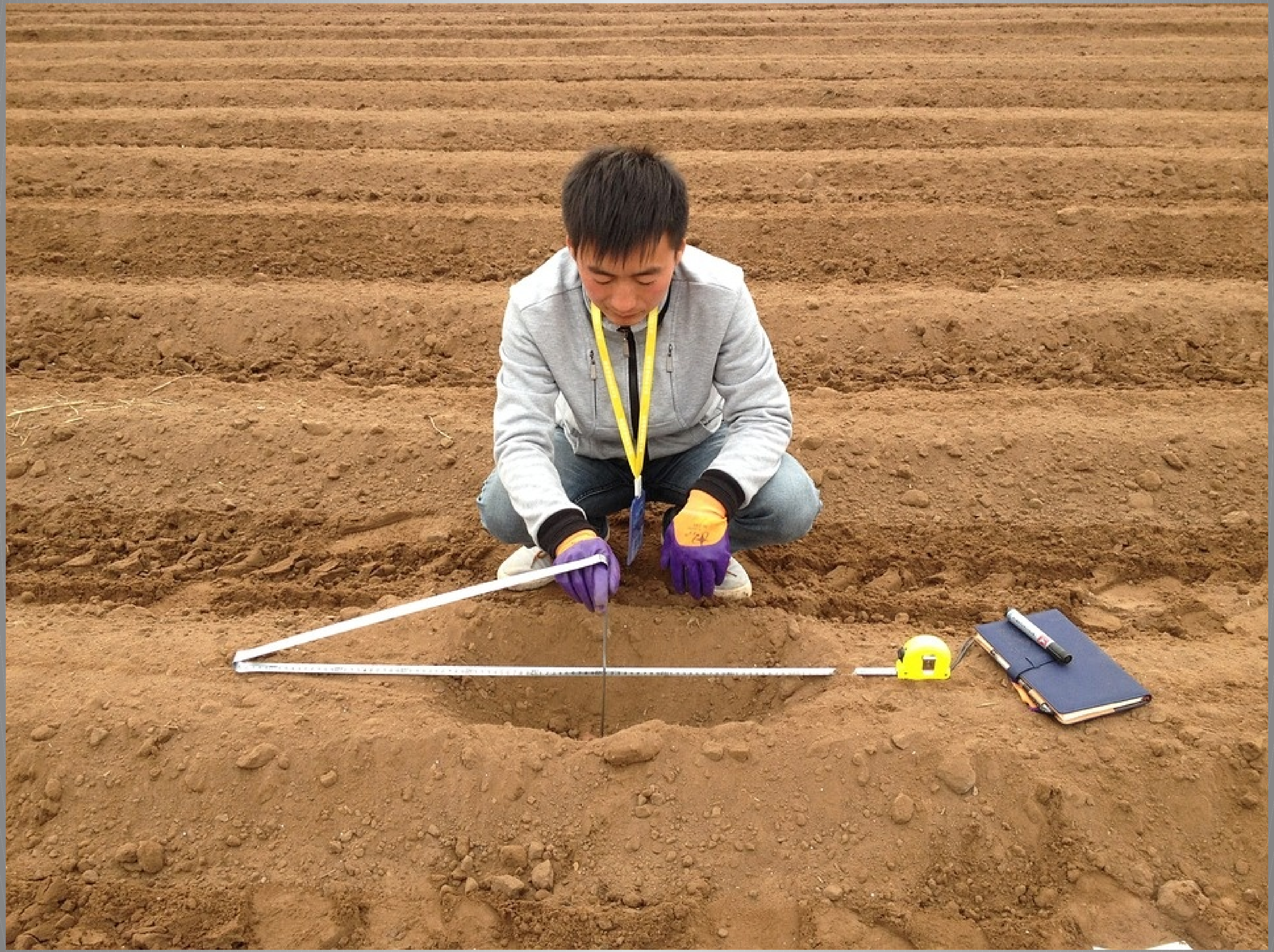
De-identification in practice ✔

Measuring re-identification risk

For further information

De-identification is used today.

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. “self-reidentification”

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

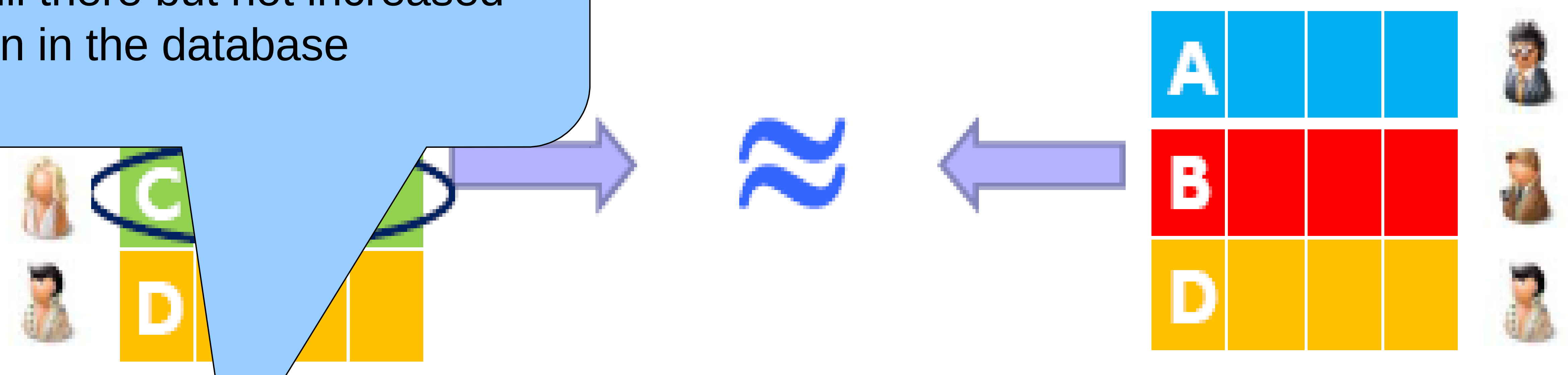
l-diversity solves this problem by assuring "diverseness" of the sensitive values.

(This table is not l-diverse.)

Differential Privacy (informal)

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

If there is already some risk 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

- 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 data quality guarantees

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

Statistical methods used with Differential Privacy

Input perturbation

- Add random noise to database, release

Summary statistics only

- Means, variances
- Marginal totals
- Regression coefficients

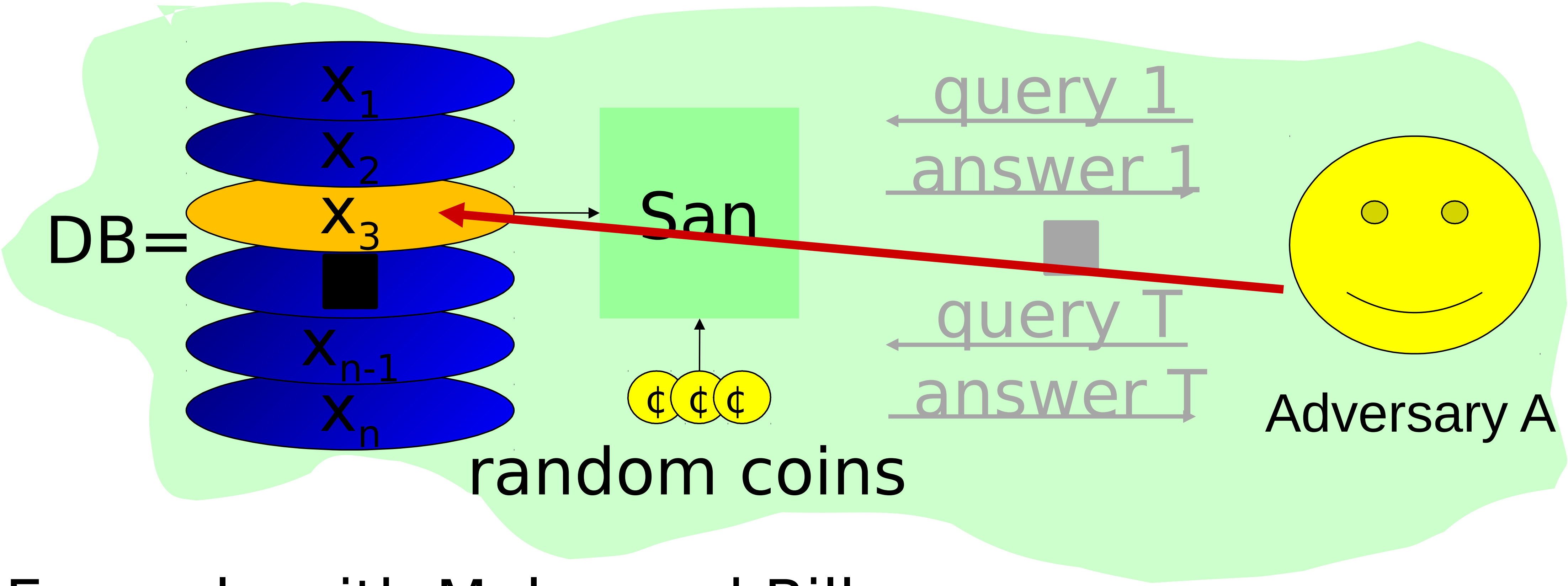
Output perturbation

- Summary statistics with noise

Interactive versions of the above methods

- Auditor decides which queries are OK, type of noise

Differential Privacy (1)



- 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

Outline for today’s talk

Why de-identify? ✔

Basic de-identification ✔

Famous re-identification controversies ✔

De-identification in practice ✔

Measuring re-identification risk ✔

For further information.

There are many ways to measure re-identification risk.

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

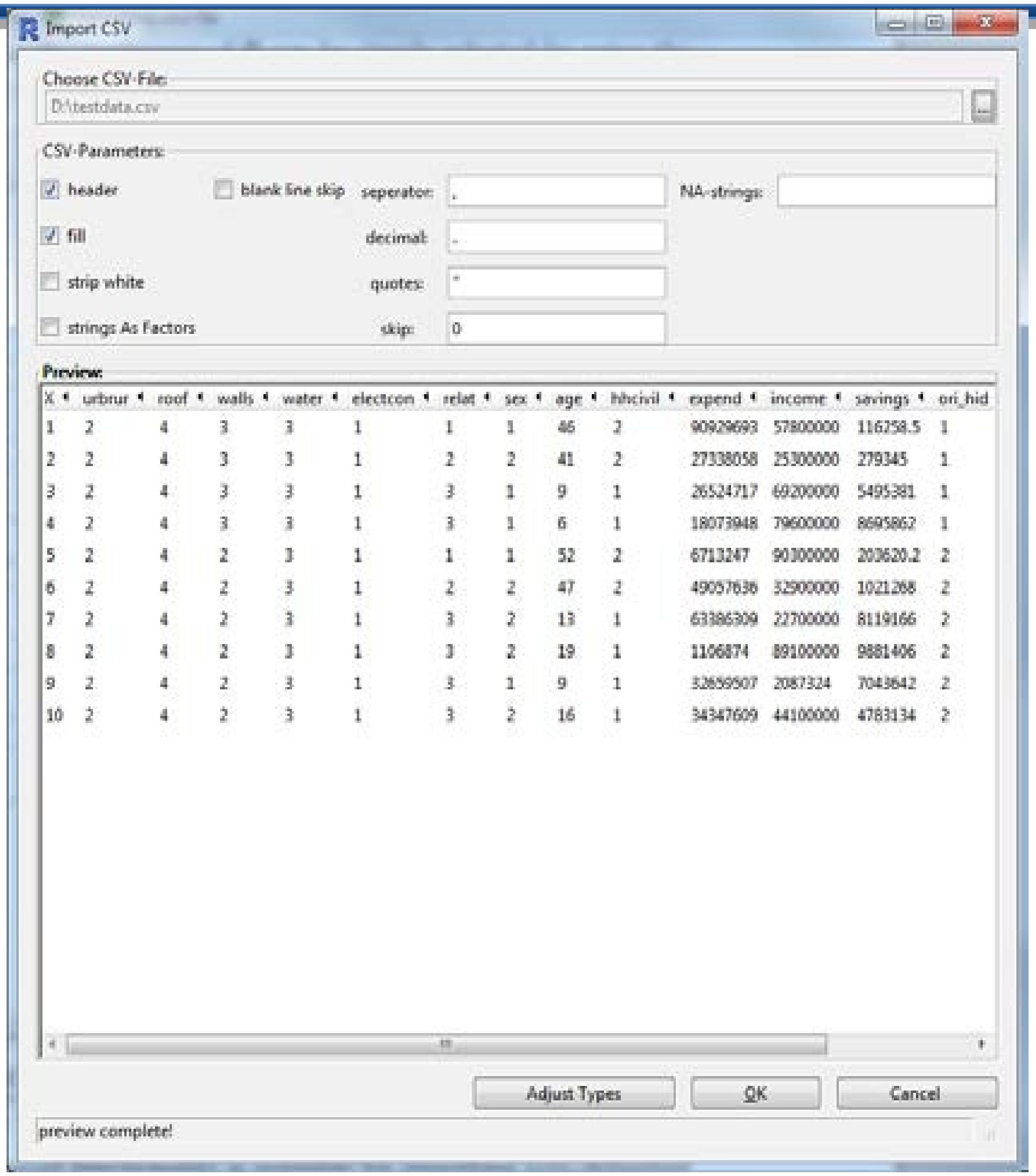
Differential privacy adds noise to mask the contribution of each individual

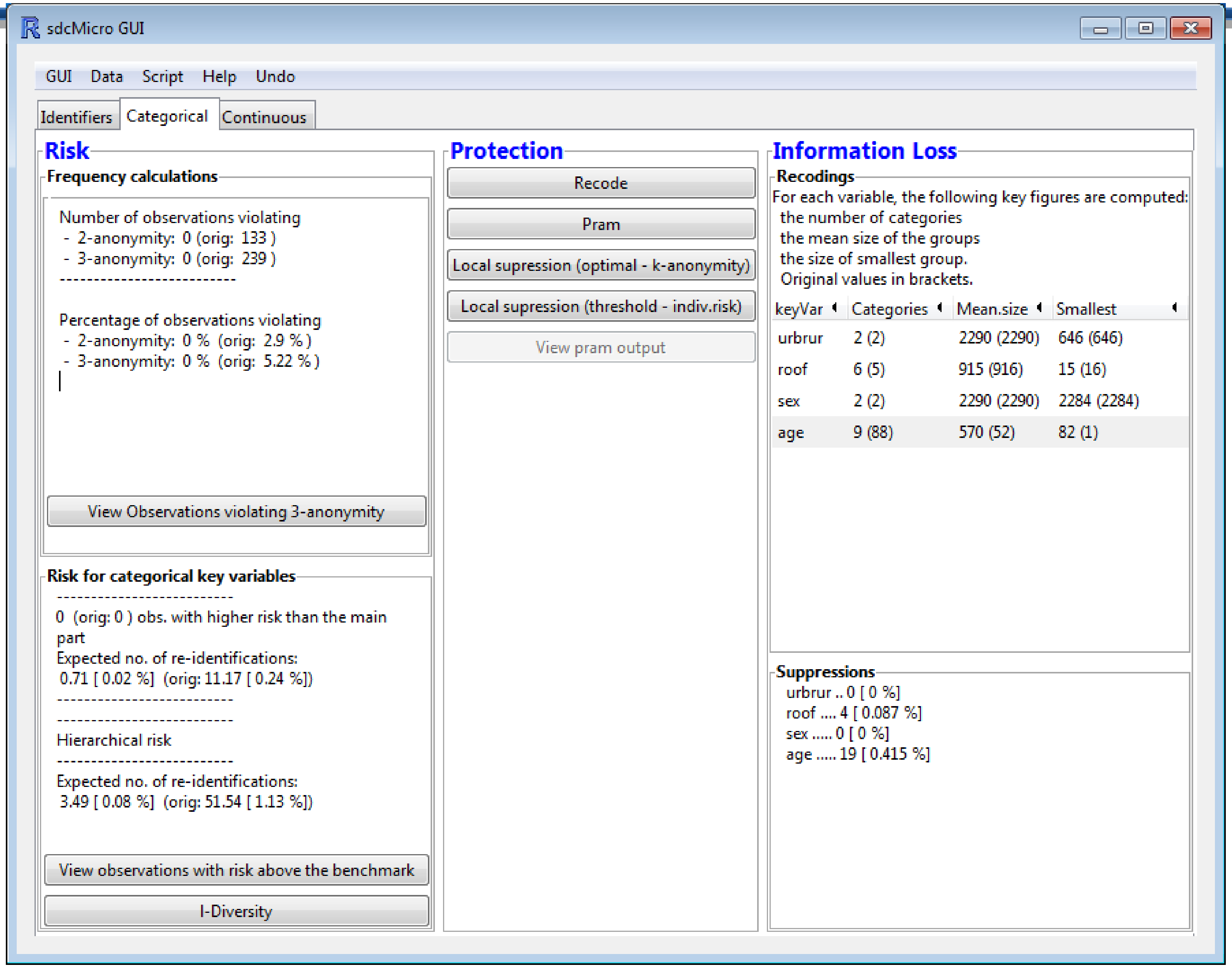
Pseudonymization allows future re-identification

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

For further information…

sdcMicro — Statistical Disclosure Control for “R”



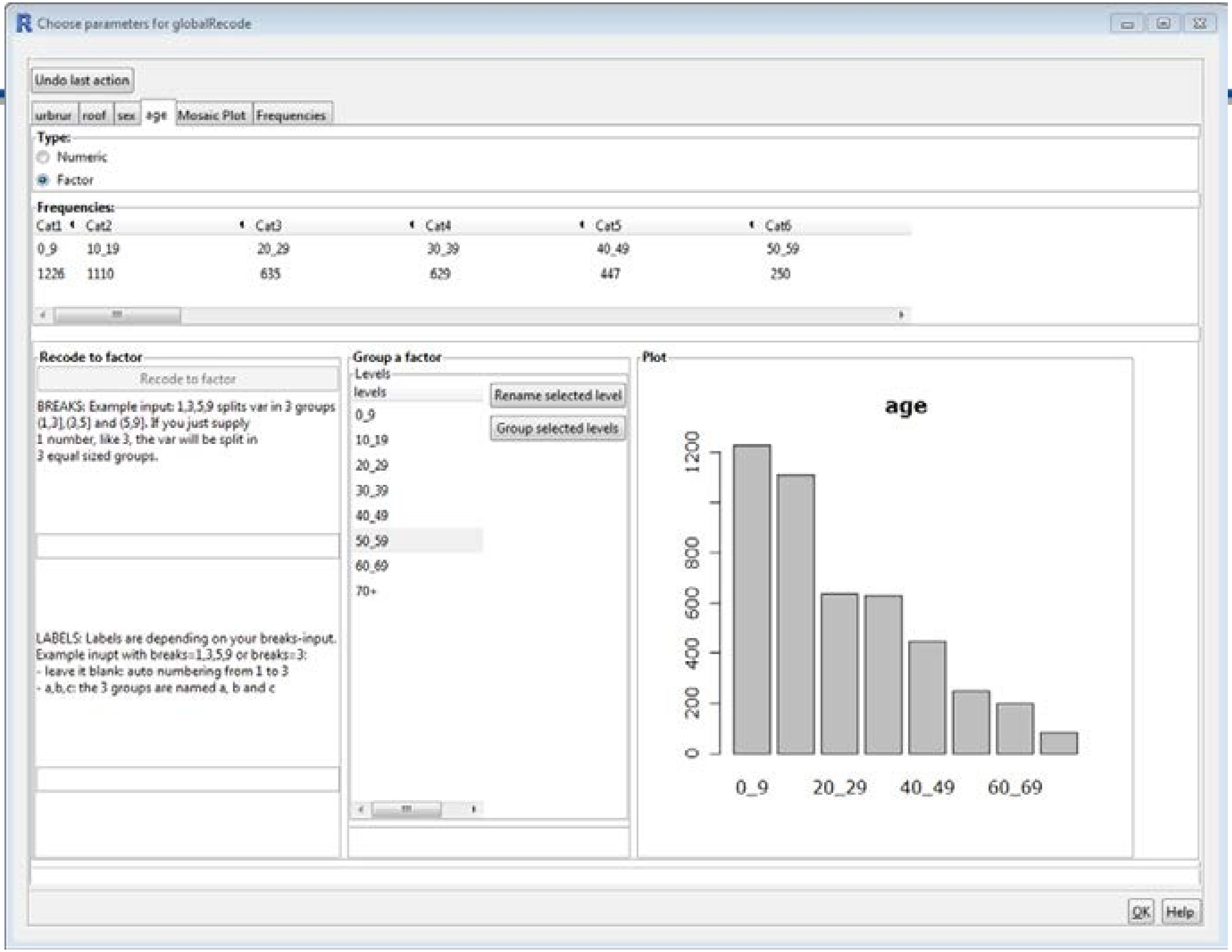


– http://www.ihsn.org/home/sites/default/files/resources/Tutorial%20sdcMicroGUI%20v6.pdf

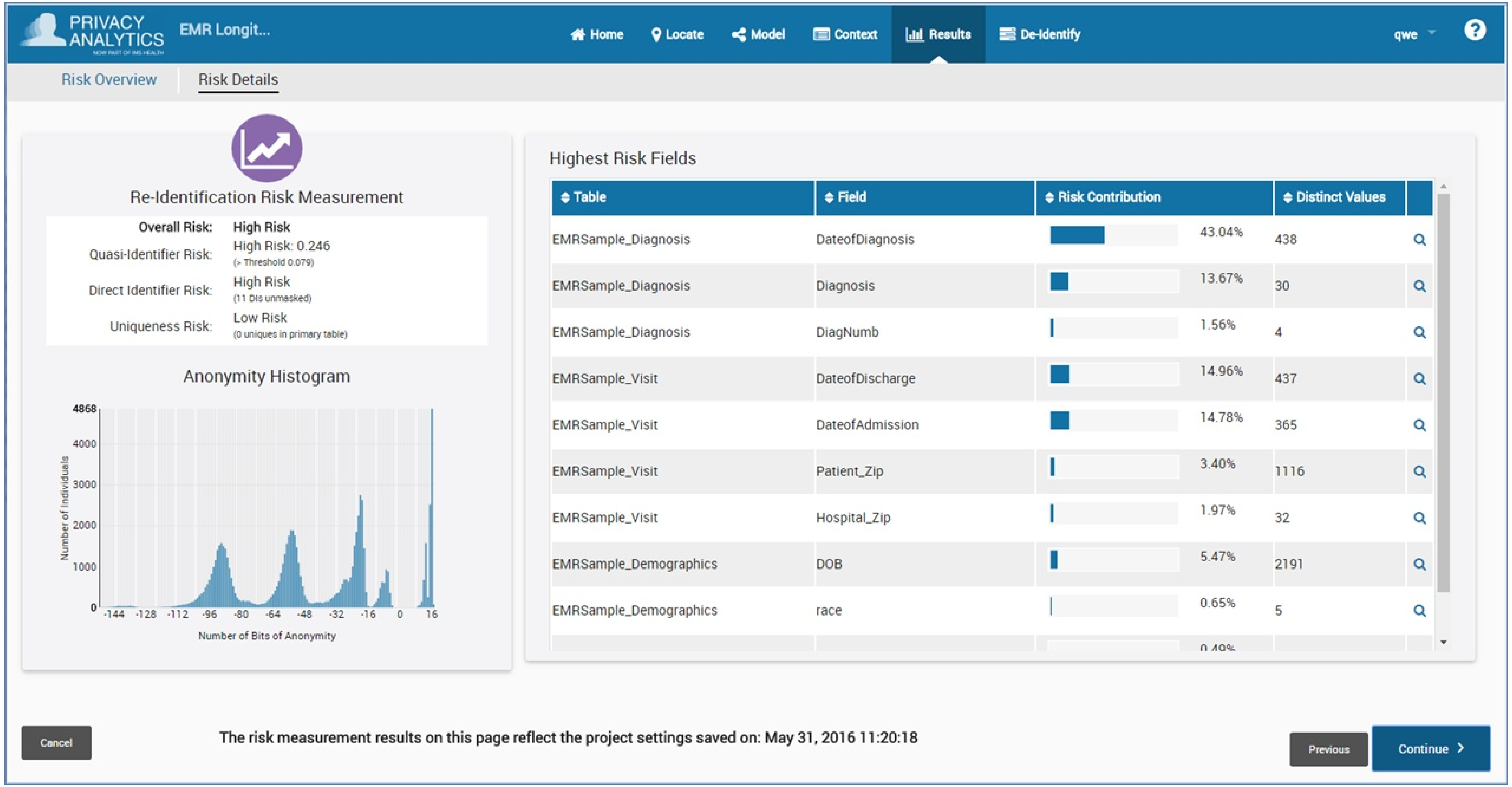
sdcMicro cont.

Limitations:

- Only a single table.
- Only a single CPU.
- No support.

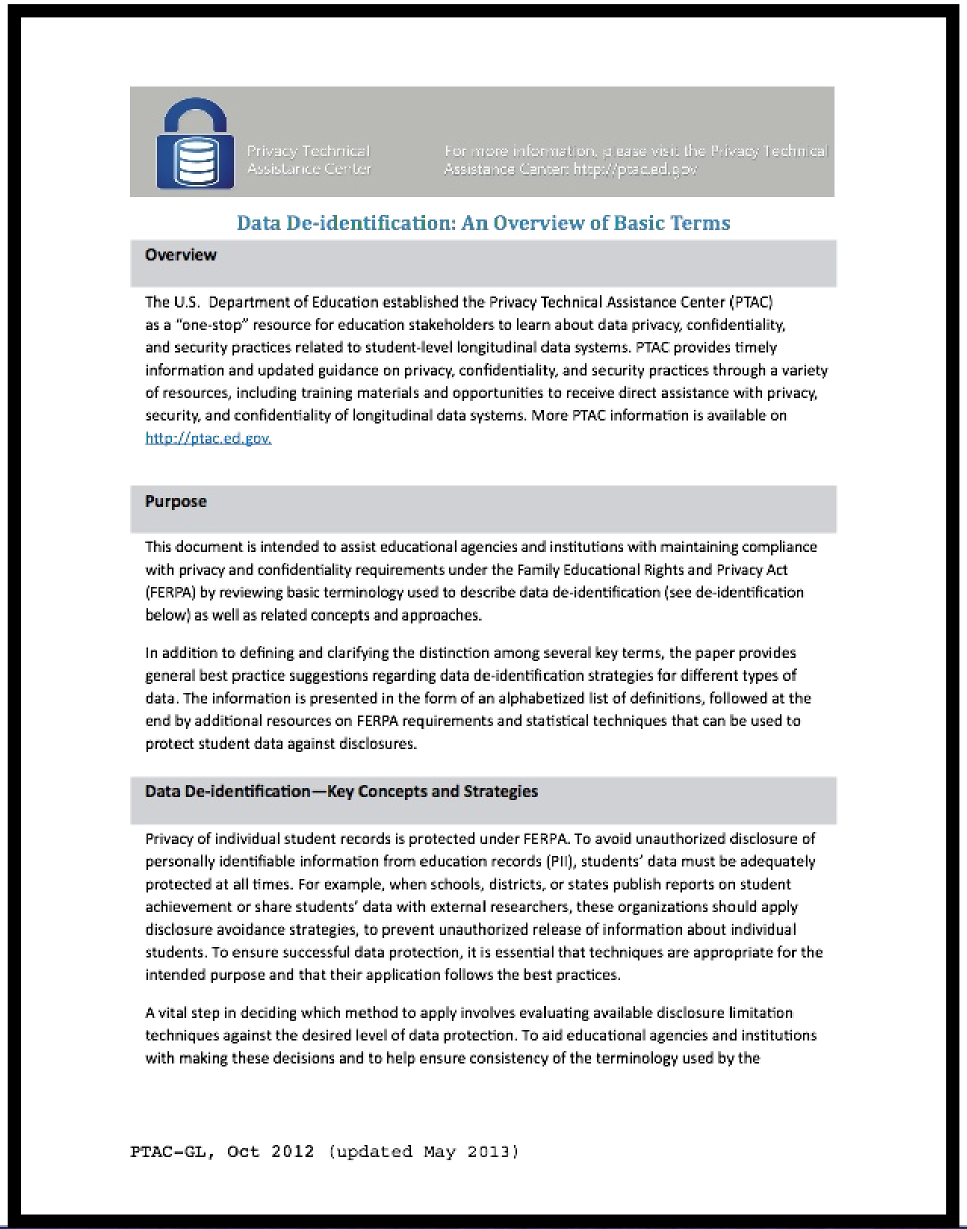


Privacy Analytics Eclipse de-identification engine.



Department of Education & HHS have de-identification guidance.

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



Privacy Technical Assistance Center

For more information, please visit the Privacy Technical Assistance Center: http://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.

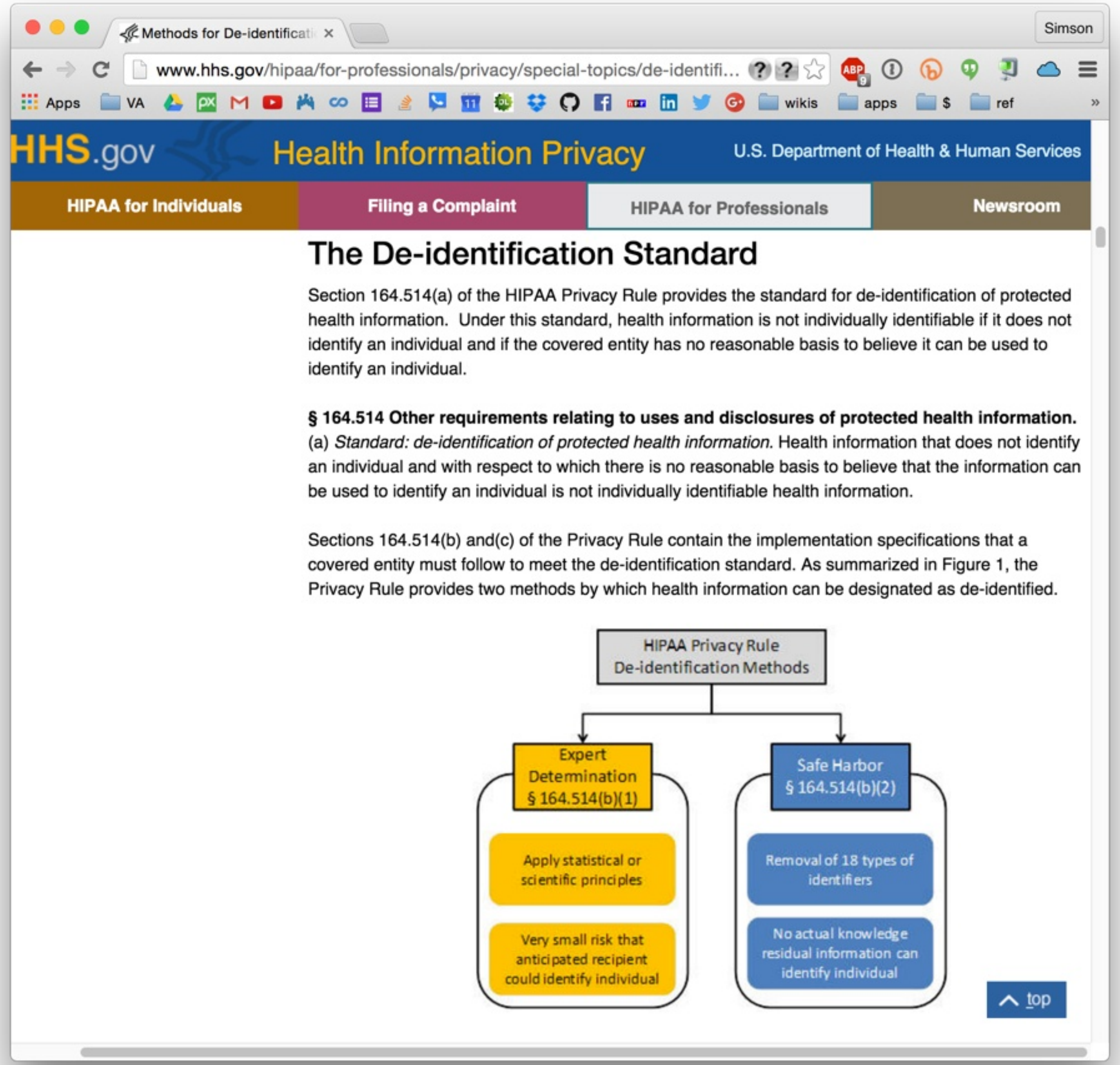
Data De-identification—Key Concepts and Strategies

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/



88

Introduction to Statistical Disclosure

IHSN

INTERNATIONAL HOUSEHOLD SURVEY NETWORK

Introduction to Statistical Disclosure Control (SDC)

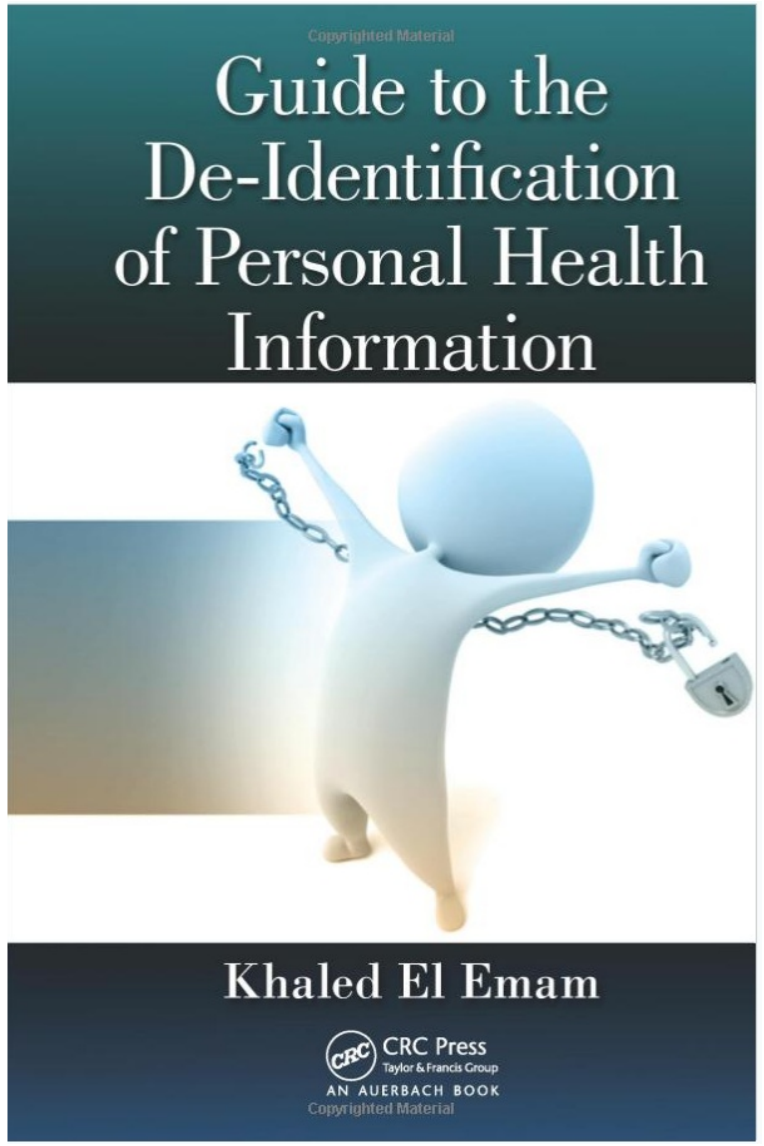
Matthias Templ, Bernhard Meindl, Alexander Kowarik and Shuang Chen

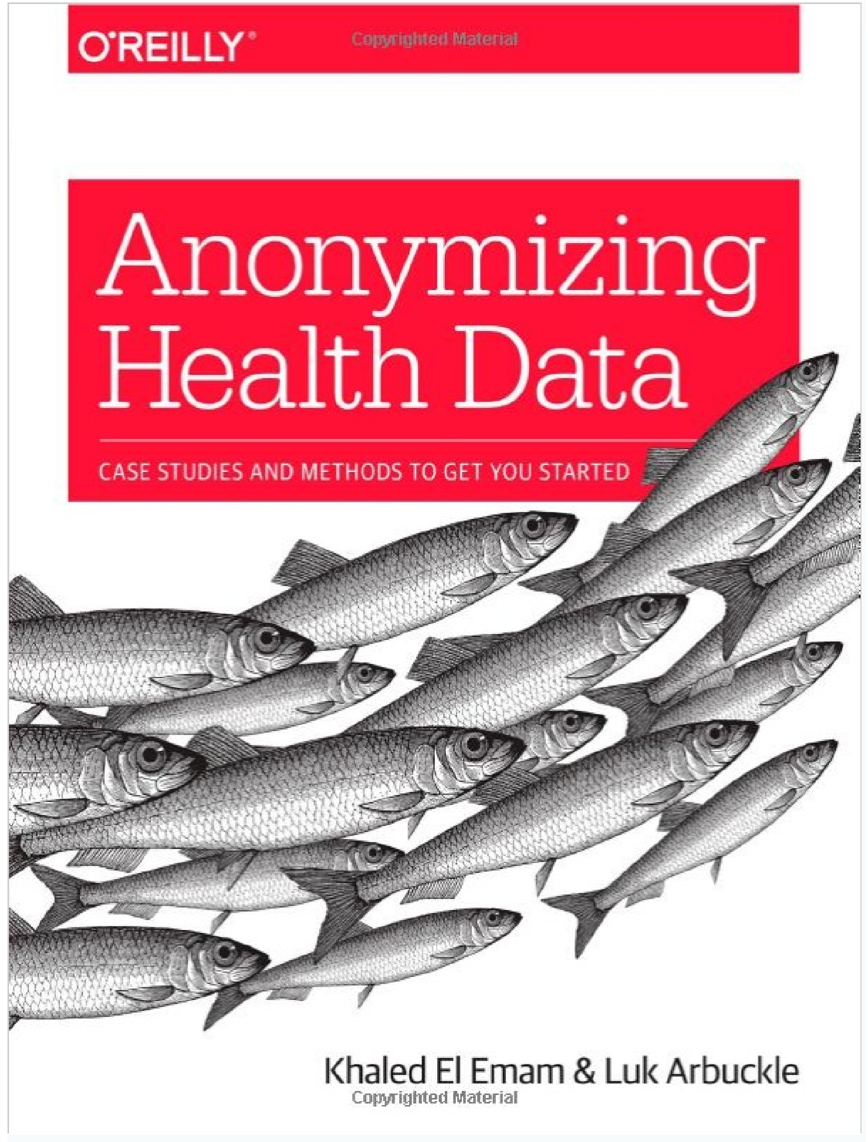
IHSN Working Paper No 007
August 2014

www.ihsn.org

— http://www.ihsn.org/home/sites/default/files/resources/ihsn-working-paper-007-Oct27.pdf

Books!


Copyrighted Material
Guide to the De-Identification of Personal Health Information
Khaled El Emam
CRC Press
Taylor & Francis Group
AN AUERBACH BOOK
Copyrighted Material

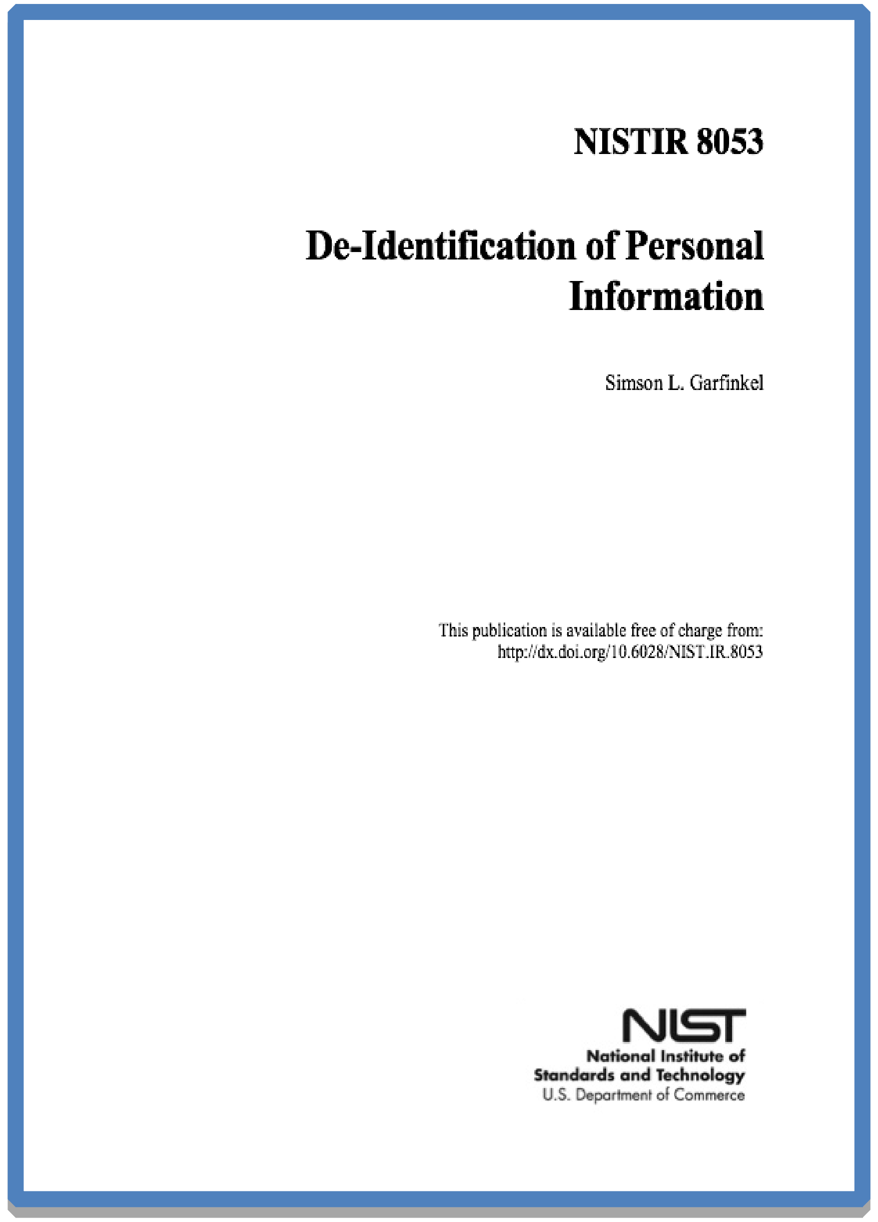

O'REILLY
Copyrighted Material
Anonymizing Health Data
CASE STUDIES AND METHODS TO GET YOU STARTED
Khaled El Emam & Luk Arbuckle
Copyrighted Material


Sharing Clinical Trial Data
MAXIMIZING BENEFITS, MINIMIZING RISK
INSTITUTE OF MEDICINE
OF THE NATIONAL ACADEMIES

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!

91