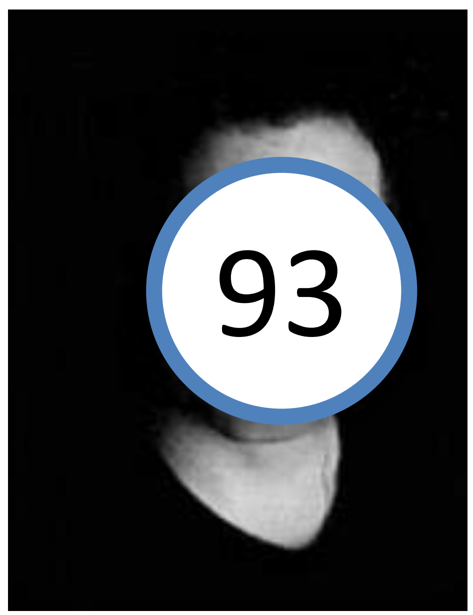
De-Identification and the Health Insurance Portability and Accountability Act (HIPAA)

Overview and framing of current issues



Simson L. Garfinkel, Ph.D.
Information Technology Laboratory
National Institute of Standards and Technology

Subcommittee on Privacy, Confidentiality & Security
National Committee on Vital and Health Statistics
May 24, 2016

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

National Institute of Standards and Technology

Founded in 1901

Non-regulatory federal laboratory.

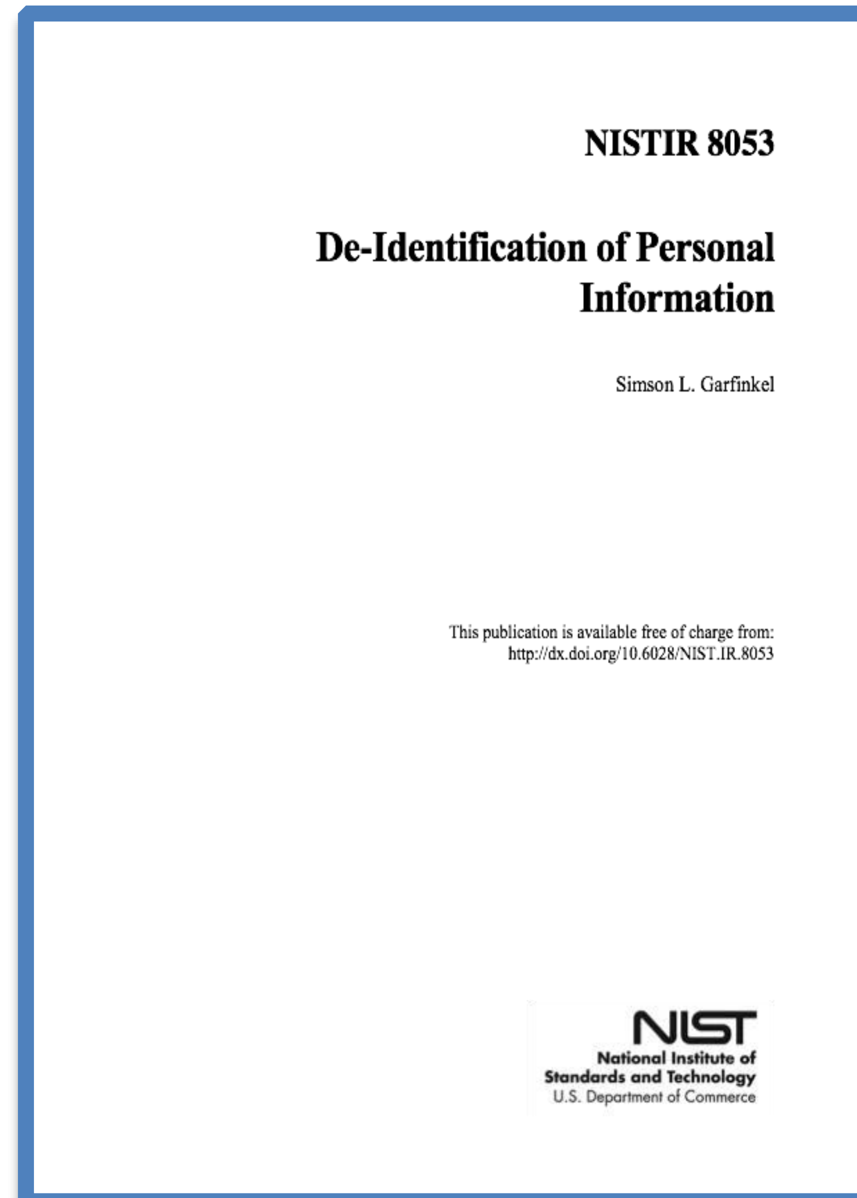
Mission:

“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.”

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

Today there is a significant and growing interest in de-identification.

Controlled Sharing

Open Science

DATA.GOV

Data Publishing

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

Under the current HIPAA Privacy Rule, de-identified Protected Health Information can be distributed without restriction.

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

Medical Records

De-identification

✗ **name**
✗ **address**
✗ **birthday**
✗ **medical record number**
etc.

https://commons.wikimedia.org/wiki/File:Applications-internet.svg

Public Internet

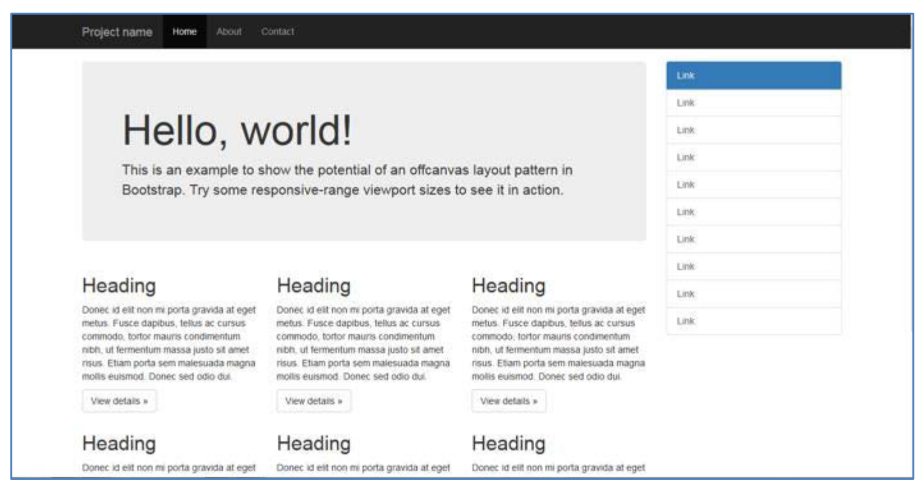
Interest in de-identification extends far beyond healthcare.

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

Social Science Data

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

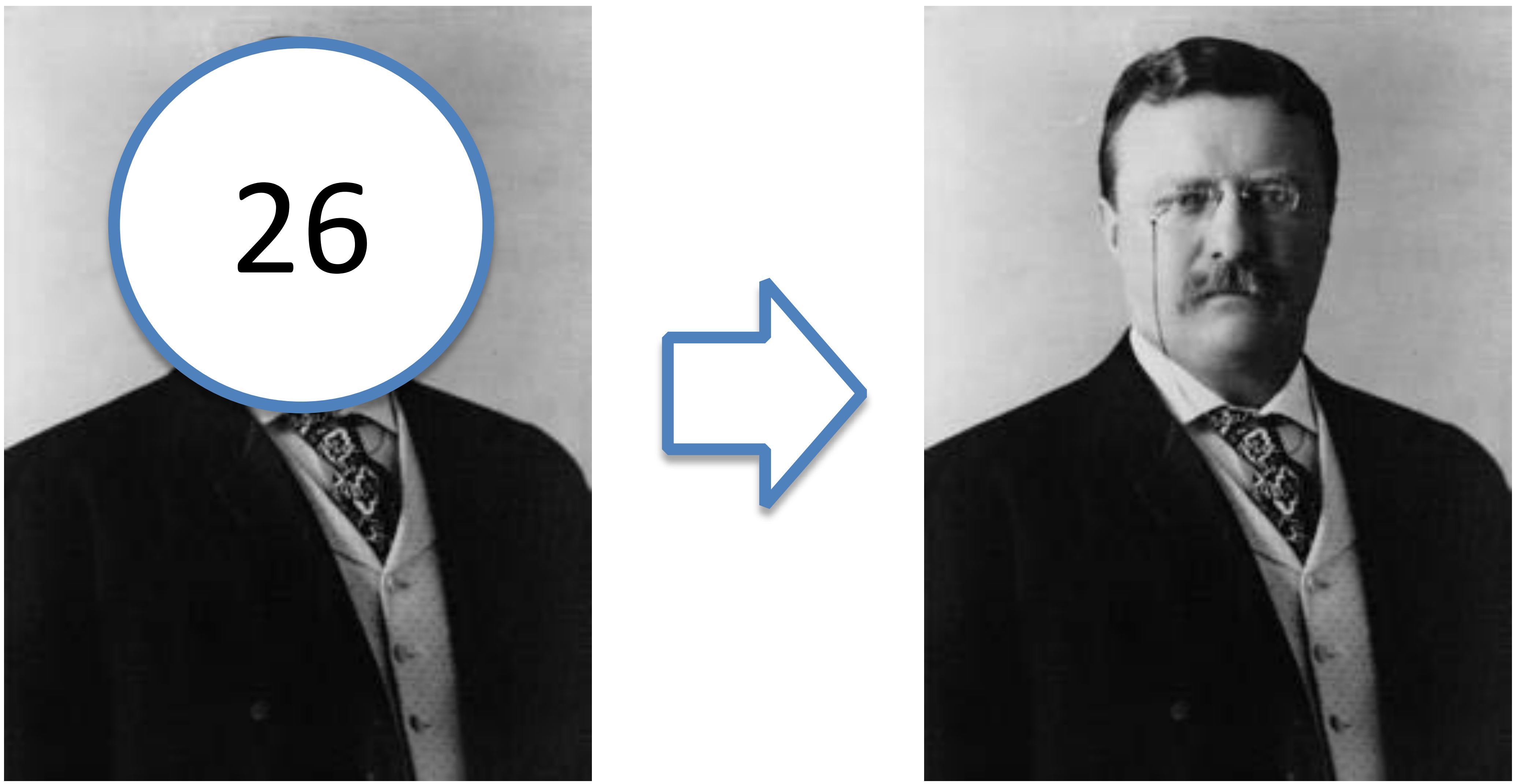
Consumer Financial Data



Website

“We will never share your personal information…”

De-identified data can be re-identified



Sometimes data are not properly de-identified.

De-identified data can be re-identified

24	Grover Cleveland
25	William McKinley
26	**Theodore Roosevelt**
27	William Howard Taft
28	Woodrow Wilson

Sometimes de-identified data can be *linked* to another dataset

Simple statistics can be identifying.

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 data.

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
…

…

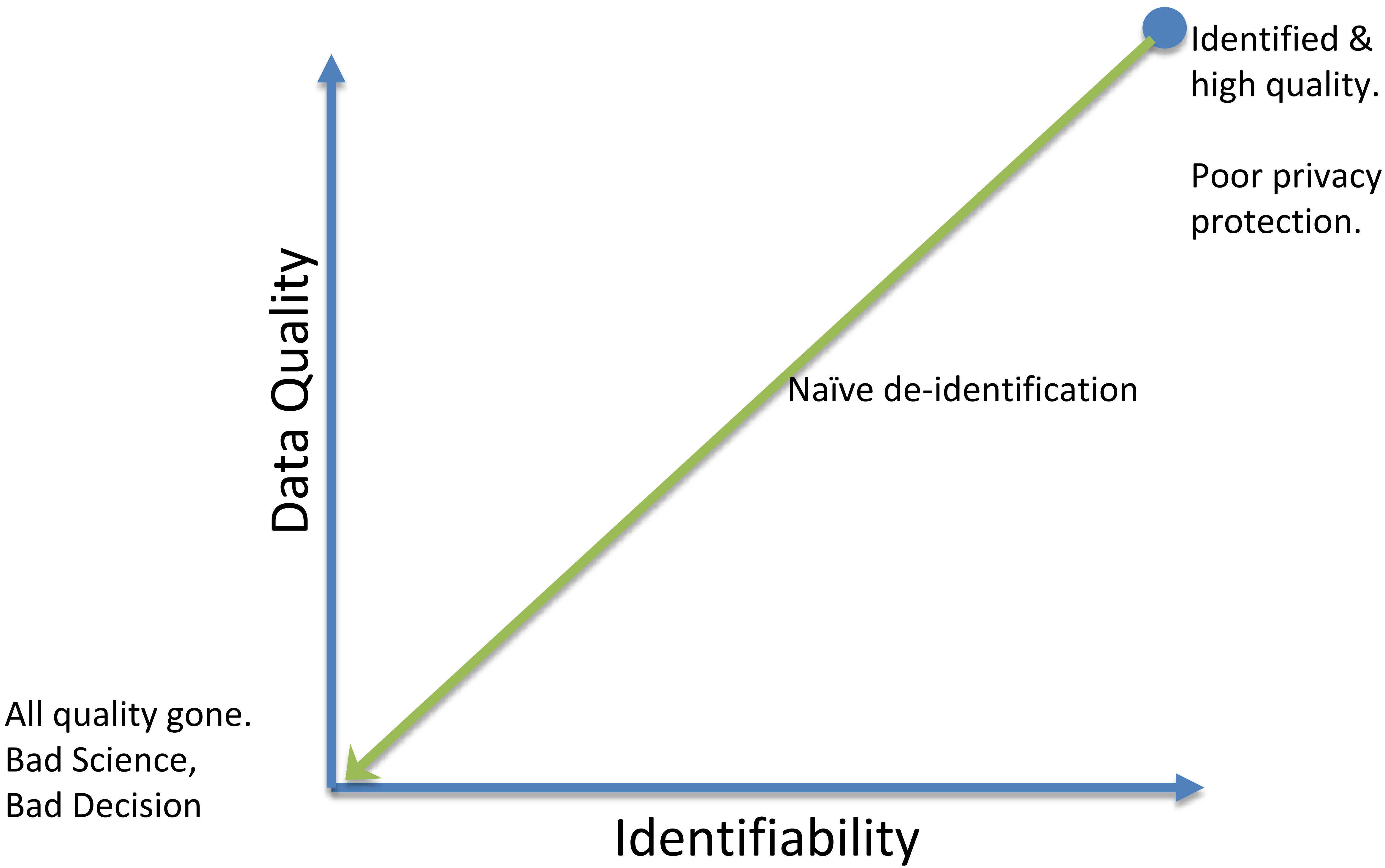
Ebola Patients			ICD-10	Diagnosis
Alice	30	F	A98.4	Ebola
Bob	35	M	A98.4	Ebola
Carol	40	F	A98.4	Ebola

Techniques for limiting identity 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

Lowering identifiability lowers data quality.



HIPAA Privacy Rule “Safe Harbor” Provision: Medical records are de-identified if 18 data elements are removed

Direct Identifiers:

— *Names*
— *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.*

Indirect Identifiers:

— *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”)*

Geographic information requires special attention

Indirect identifiers			***Direct identifier***		
Title	**Age**	**Sex**	**Address**	**ICD-10**	**Diagnosis**
Lab Tech	35	M	100 Utah St. Anytown, 20124	K25.0	Gastric Ulcer with hemorrhage
Lab Tech	56	F	653 Pleasant St. Uptown, 20321	J00	Acute nasopharyngitis [Common Cold]
Professor	35	M	564 Main St. Nassis, 25312	T25.332S	Burn of third degree of left toe
Professor	69	F	202 Sky Lane Katap, 20134	C64.1	Malignant neoplasm of right kidney
Contracts Specialist	52	F	956 Diablo Rd. Quirky, 23990	L30.9	Dermatitis, unspecified [Eczema]
University President	56	F	451 Termo Dr. Boltz, 25333	C64.1	Malignant neoplasm of right kidney

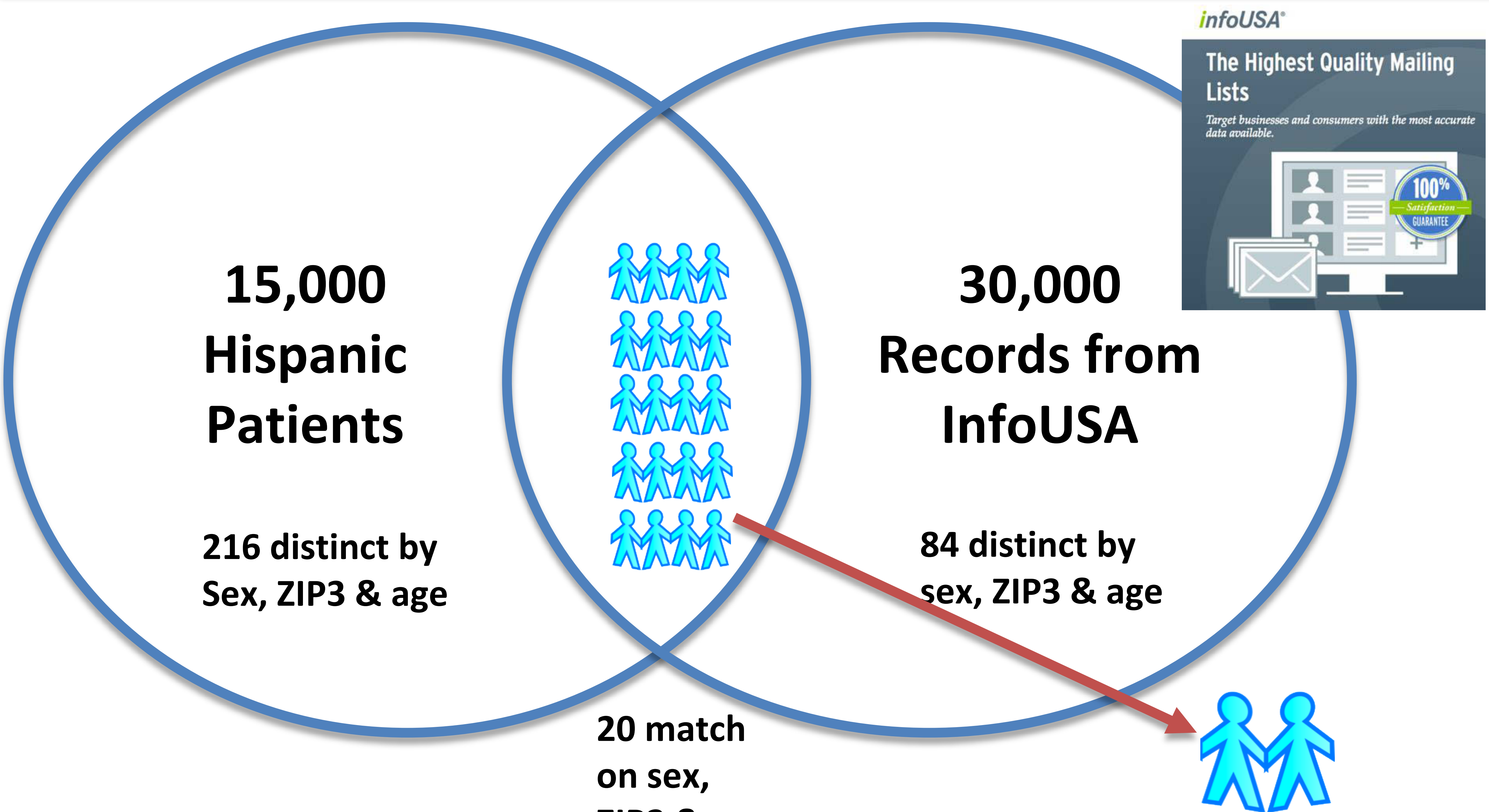
Hypothetical dataset from university healthcare system

Safe Harbor allows ZIP3
(assuming there are 20,000 people living in the area)

Indirect identifiers			***Direct identifier***		
Title	**Age**	**Sex**	**Address**	**ICD-10**	**Diagnosis**
Lab Tech	35	M	[redacted] 201XX	K25.0	Gastric Ulcer with hemorrhage
Lab Tech	56	F	[redacted] 203XX	J00	Acute nasopharyngitis [Common Cold]
Professor	35	M	[redacted] 253XX	T25.332S	Burn of third degree of left toe
Professor	69	F	[redacted] 201XX	C64.1	Malignant neoplasm of right kidney
Contracts Specialist	52	F	[redacted] 239XX	L30.9	Dermatitis, unspecified [Eczema]
University President	56	F	[redacted] 253XX	C64.1	Malignant neoplasm of right kidney

Hypothetical dataset from university healthcare system

Results of the 2010 Office of the National Coordinator for Health Information Technology Safe Harbor Re-Identification Test:



2 actual matches on last name, street address, and phone

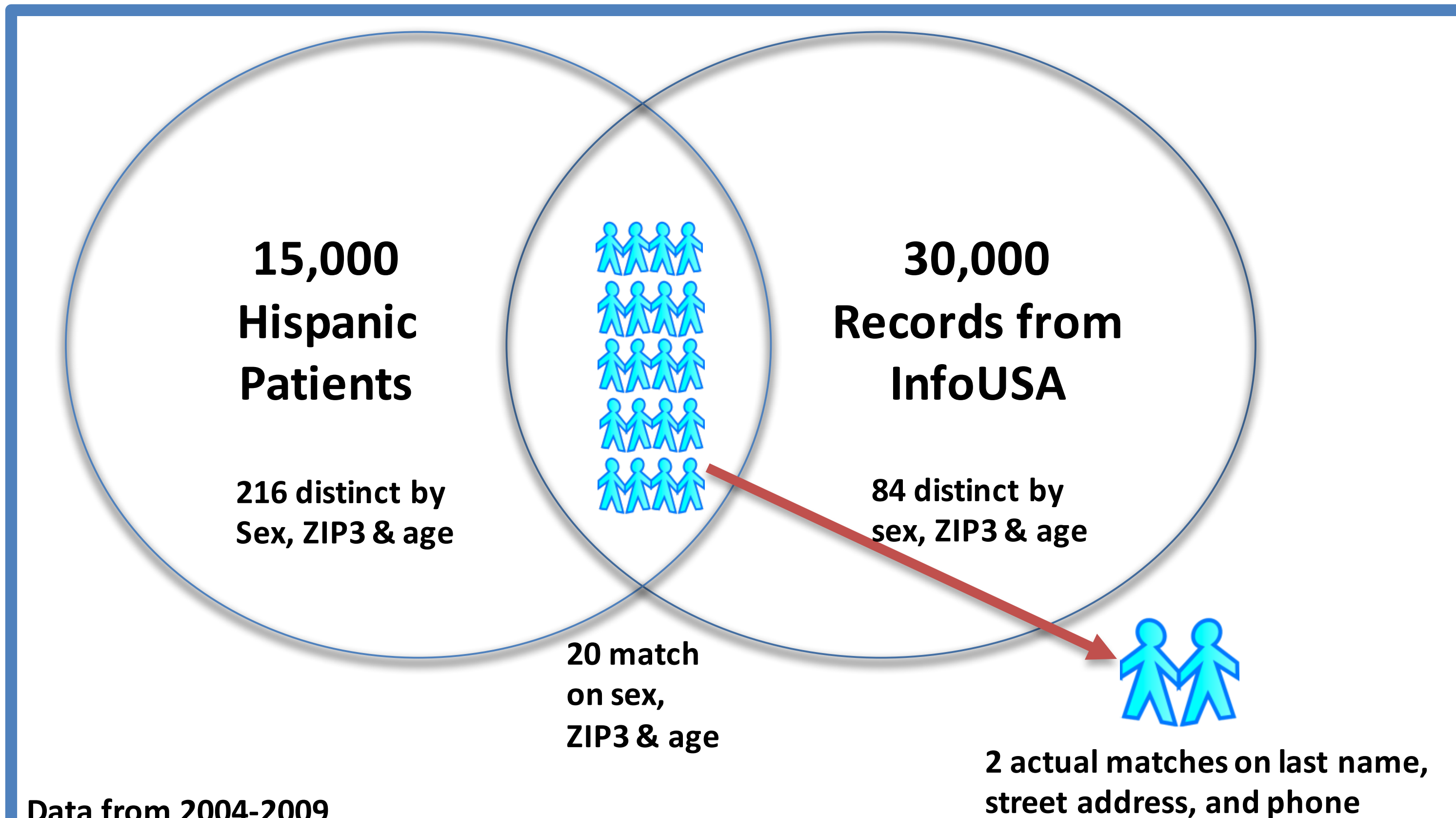
Data from 2004-2009

K-anonymity: assure at least “k” records have the same set of indirect identifiers.

Indirect identifiers			*Direct identifier*		
Title	**Age**	**Sex**	**Address**	**ICD-10**	**Diagnosis**
Lab Tech	35	M	[redacted] 201XX	K25.0	Gastric Ulcer with hemorrhage
Lab Tech	56	F	[redacted] 203[redacted]XX	J00	Acute nasopharyngitis [Common Cold]
Professor	35	M	[redacted] 253XX	T25.332S	Burn of third degree of left toe
Professor	69	F	[redacted] 201XX	C64.1	Malignant neoplasm of right kidney
Contracts Specialist	52	F	[redacted] 239XX	L30.9	Dermatitis, unspecified [Eczema]
University President	56	F	[redacted] 253XX	C64.1	Malignant neoplasm of right kidney

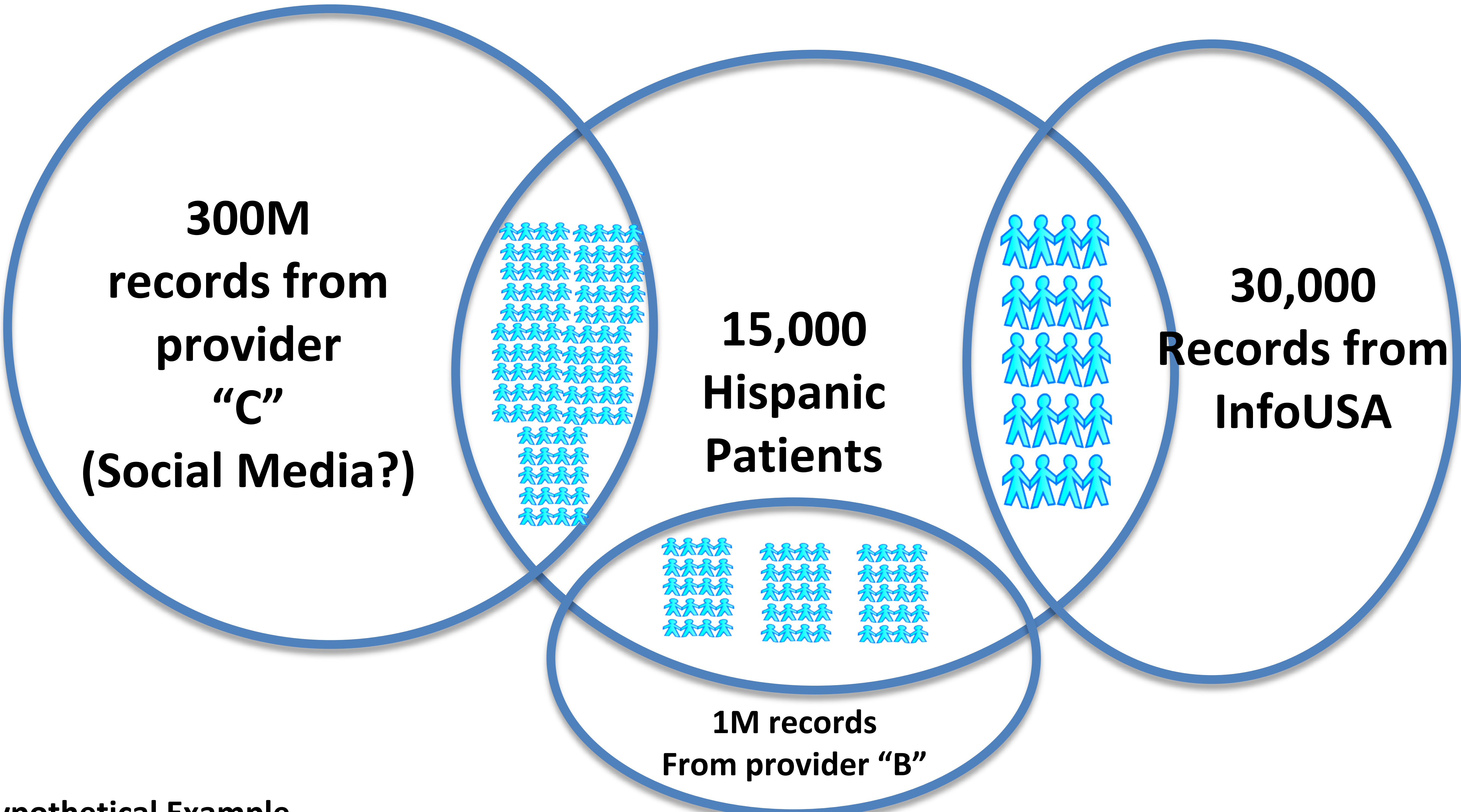
Color background indicates values modified for k=2 k-anonymity

“Tiger Teams” are another way to test re-identification.



Estimated Re-identification rate:
No verification: 20 in 15,000
Verification: 2 in 15,000

Re-identification tests assume data available to match. As more data become available, re-identification gets easier.



Hypothetical Example

A constellation of diseases can be an identifier

De-identified medical records from provider "N"

Smallpox

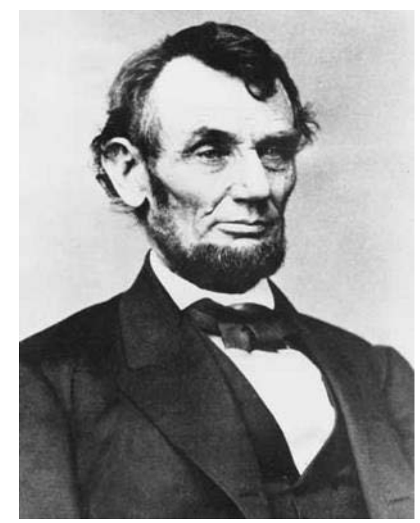
Concussion at age 9

Malaria at age 21

Depression

Fractured jaw

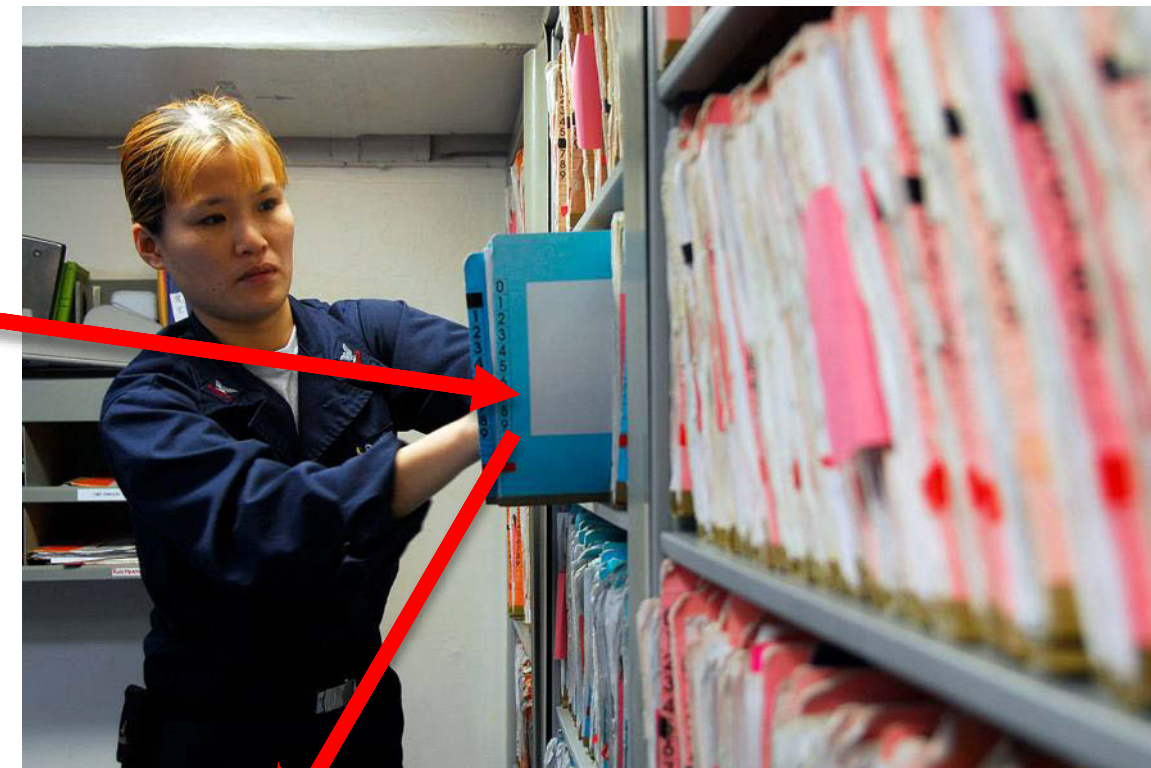
Linked records

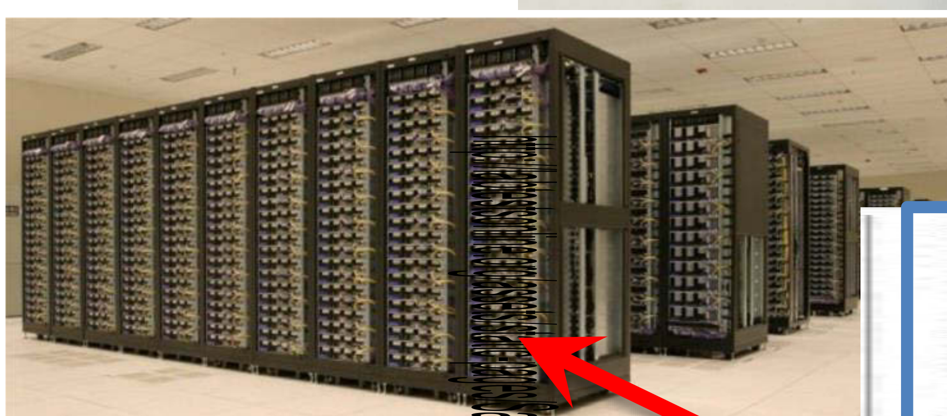


Hypothetical Example

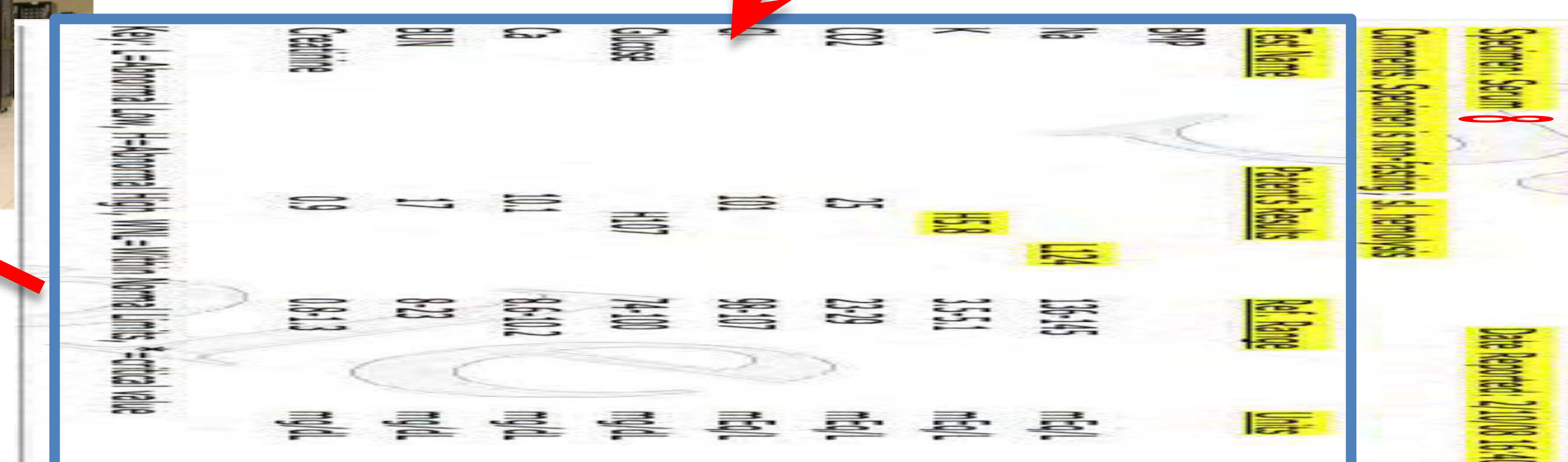
Atreya, Smith, McCoy, Malin & Miller (2013)
"Reducing patient re-identification risk for laboratory results within research datasets."

Medical tests





Database with linked test results



A single identified blood test can be the link to dozens of de-identified records

Blood tests can be de-identified by adding noise

Example Lab Report

Patient copy

1 University Medical Center, Dept. of Pathology — 02/14/2008 2
123 University Way, City, ST 12345 — 16:13

Doe, Mr. John Q. 3

Patient ID No. 987654321 3 — D.O.B. 01/01/1941 — 67Y
Ordering MD: Smith, Peter MD 4 — Physician Copy for Dr: Smith, Jane
PT medications: multivitamins 5

Specimen(s) Collected: 2/10/08 14:30 — Lab Acc'n No. 2234
Specimen: Serum — Date Reported: 2/10/08 16:
Comments: Specimen is non-fasting; sl. hemolysis

Test Name	Patient's Results	Ref. Range	Units
BMP			
Na	L124	136-145	mEq/L
K	H5.8	3.5-5.1	mEq/L
CO2	25	23-29	mEq/L
Cl	101	98-107	mEq/L
Glucose	H107	74-100	mg/dL
Ca	10.1	8.6-10.2	mg/dL
BUN	17	8-23	mg/dL
Creatinine	0.9	0.8-1.3	mg/dL

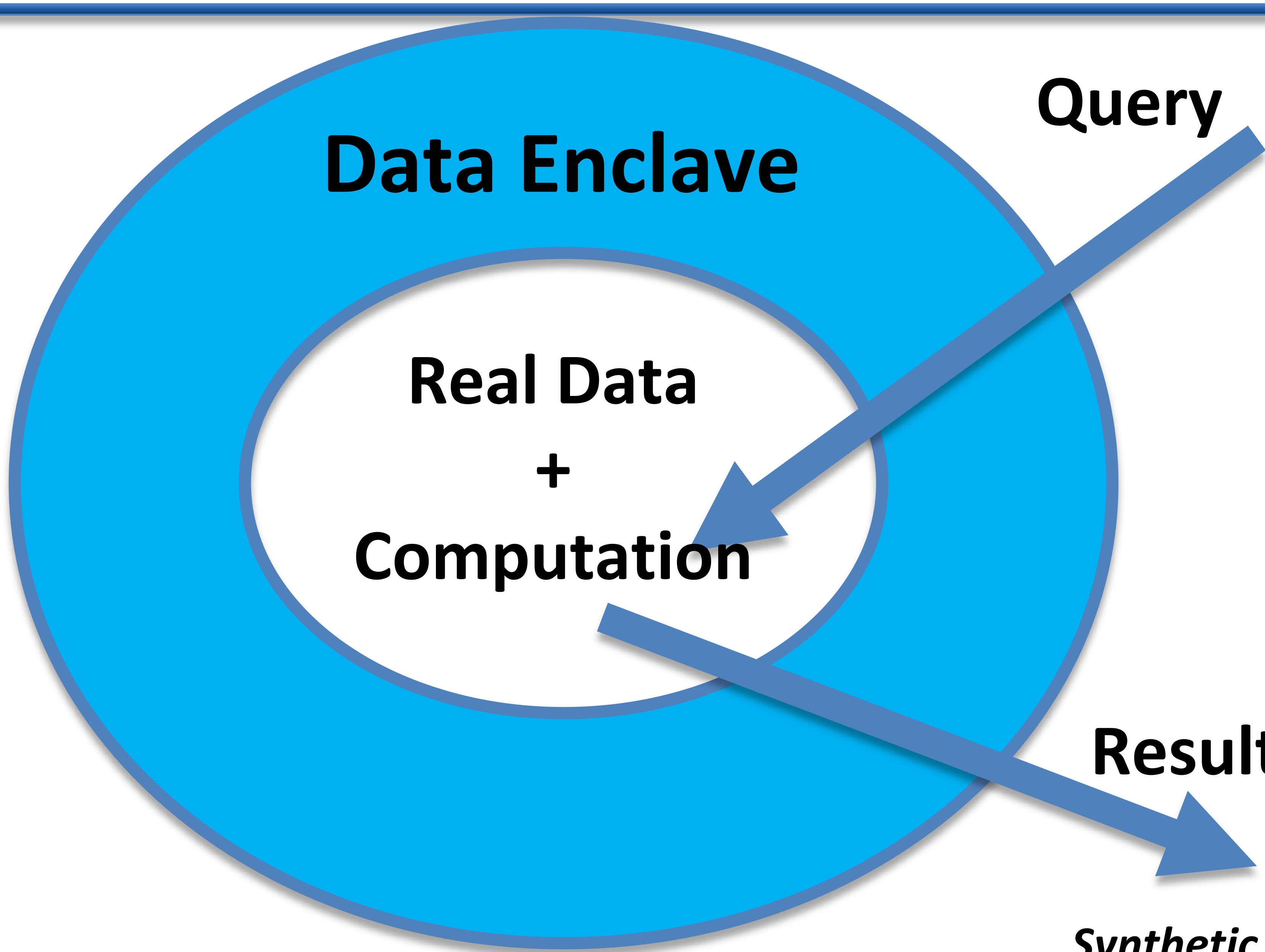
Key: L=Abnormal Low, H=Abnormal High, WNL=Within Normal Limits, *=critical value

Na: 124 ⇒ 126
K: 5.8 ⇒ 5.9
CO2: 25 ⇒ 24
Cl: 101 ⇒ 104
4
Glucose: 107 ⇒ 110
Ca: 10.1 ⇒ 9.9
BUN: 17 ⇒ 17
Creatinine: 0.9 ⇒ 1.0

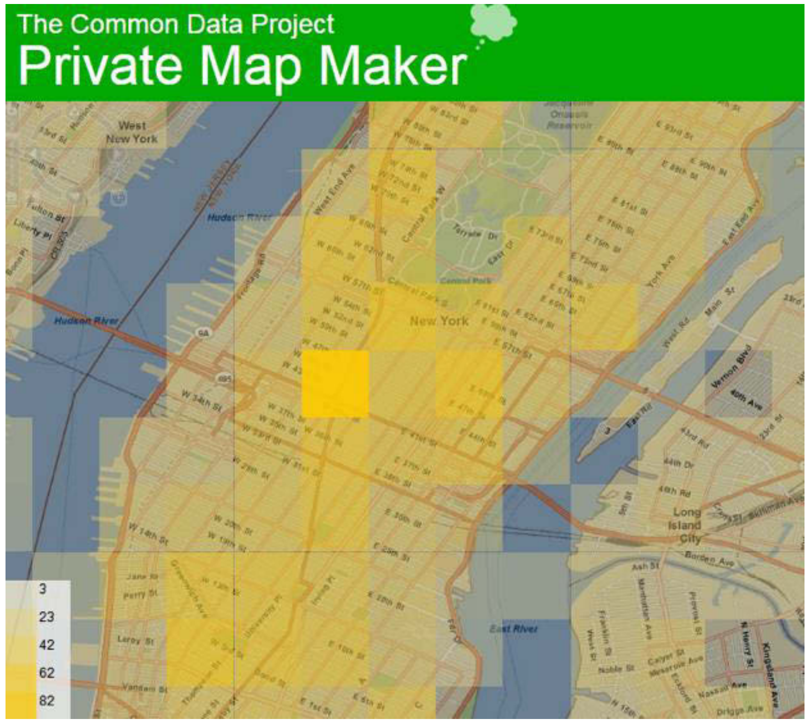
(values for demonstration only)

Research database

"Differential Privacy" adds systematic noise to query results



Synthetic Data



Key concepts: Privacy Budget & Noise

The Census Bureau distributes synthetic data to protect privacy while preserving some data quality.

Can synthetic datasets designed to enable research also be used to promote accountability and transparency?

Animated encounter data

Fare Basis	Airline	Booking Class	Trip Type	Fare	Cabin	Effective Date	Expiration Date	Min / Max Stay	Adv Purchase Req
TA14A0SP	DL	T	One-Way	189.00(USD)	E		02/16/11		14
TA03A0SG	DL	T	One-Way	209.00(USD)	E		02/28/11		03
UA10A0UY	DL	U	One-Way	236.00(USD)	E				10
LA10A0VY	DL	L	One-Way	251.00(USD)	E				10
LA07A0NY	DL	L	One-Way	286.00(USD)	E				07
KA07A0UY	DL	K	One-Way	326.00(USD)	E				07
QA03A0NP	DL	Q	One-Way	396.00(USD)	E				03
HA00A0NY	DL	H	One-Way	466.00(USD)	E				
UC14A0NJ	DL	U	Round-Trip	482.00(USD)	E			V / 30	14
HA00A0UY	DL	H	One-Way	616.00(USD)	E				
MA00A0RY	DL	M	One-Way	694.00(USD)	E				
MA00UPNY	DL		One-Way	1076.00(USD)	B				
BA00UPRQ	DL		One-Way	1348.00(USD)	B				
Y0	DL	Y	One-Way	1435.00(USD)	E				
YUP	DL		One-Way	1527.00(USD)	B				
F0	DL	F	One-Way	1826.00(USD)	F				
Y	DL	Y	One-Way	1889.00(USD)	E				
C	DL	C	One-Way	2213.00(USD)	B				
F	DL	F	One-Way	2766.00(USD)	F				

Synthetic tabular data

Body-worn camera video with replaced faces

De-identification strategies should be formally evaluated.

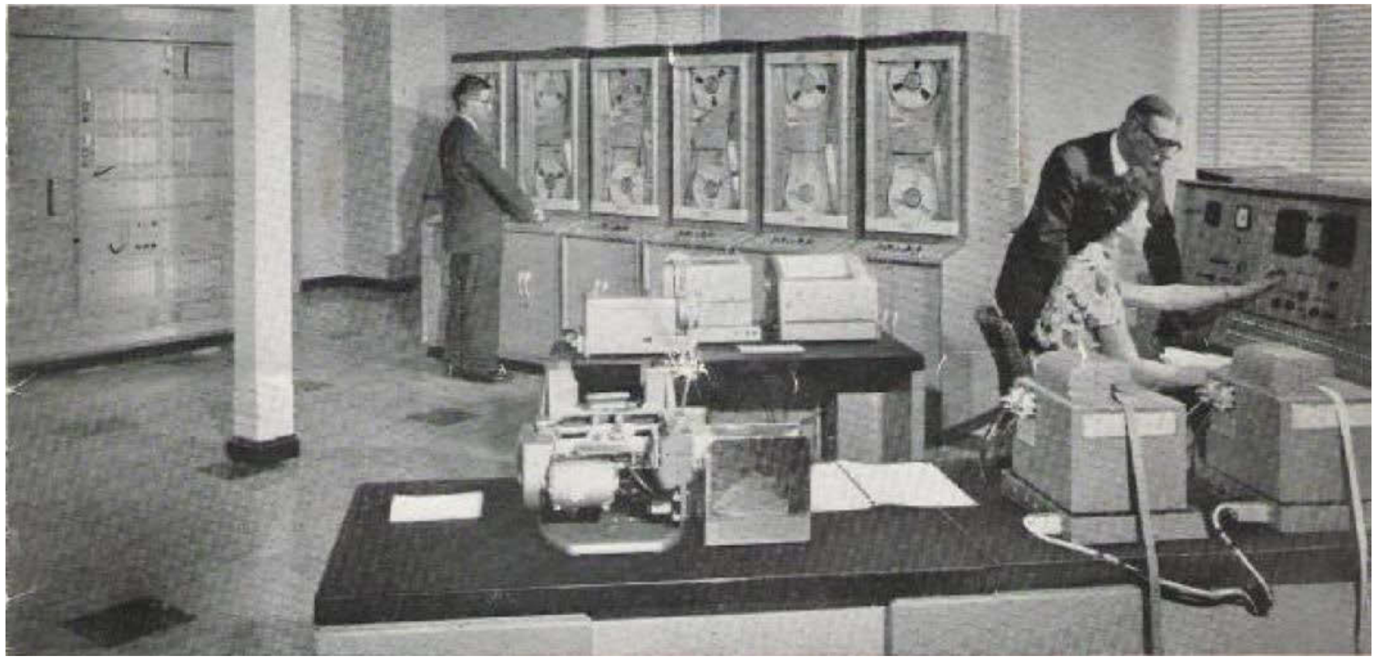
Do they meet the stated policy goals?

Does the software faithfully implement the stated algorithm?

Are the statistical privacy guarantees actually met?

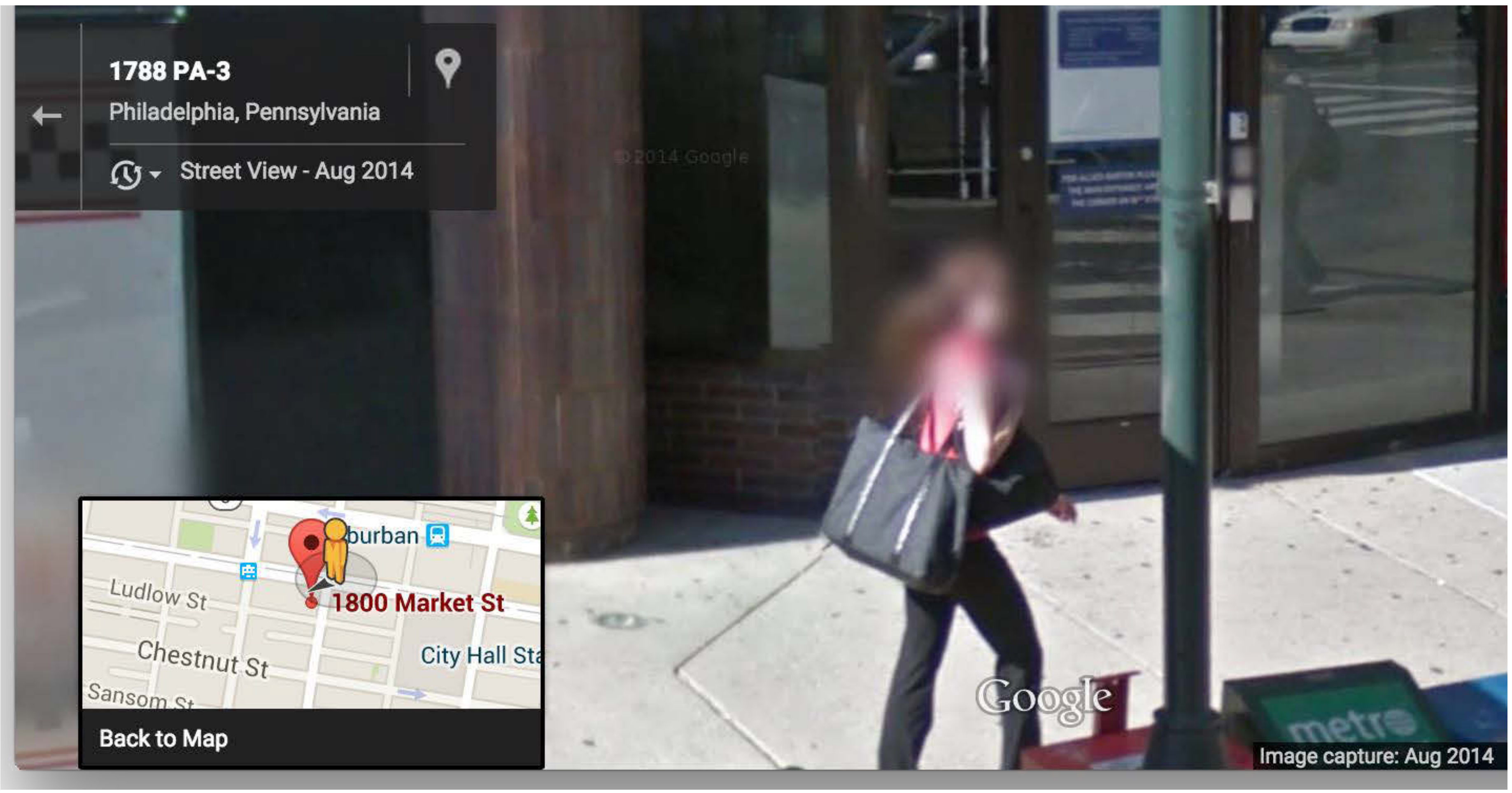
Is the necessary training in place?

Will there be monitoring and auditing?



https://en.wikipedia.org/wiki/EMIDEC_1100

De-identification of non-tabular data poses special problems.

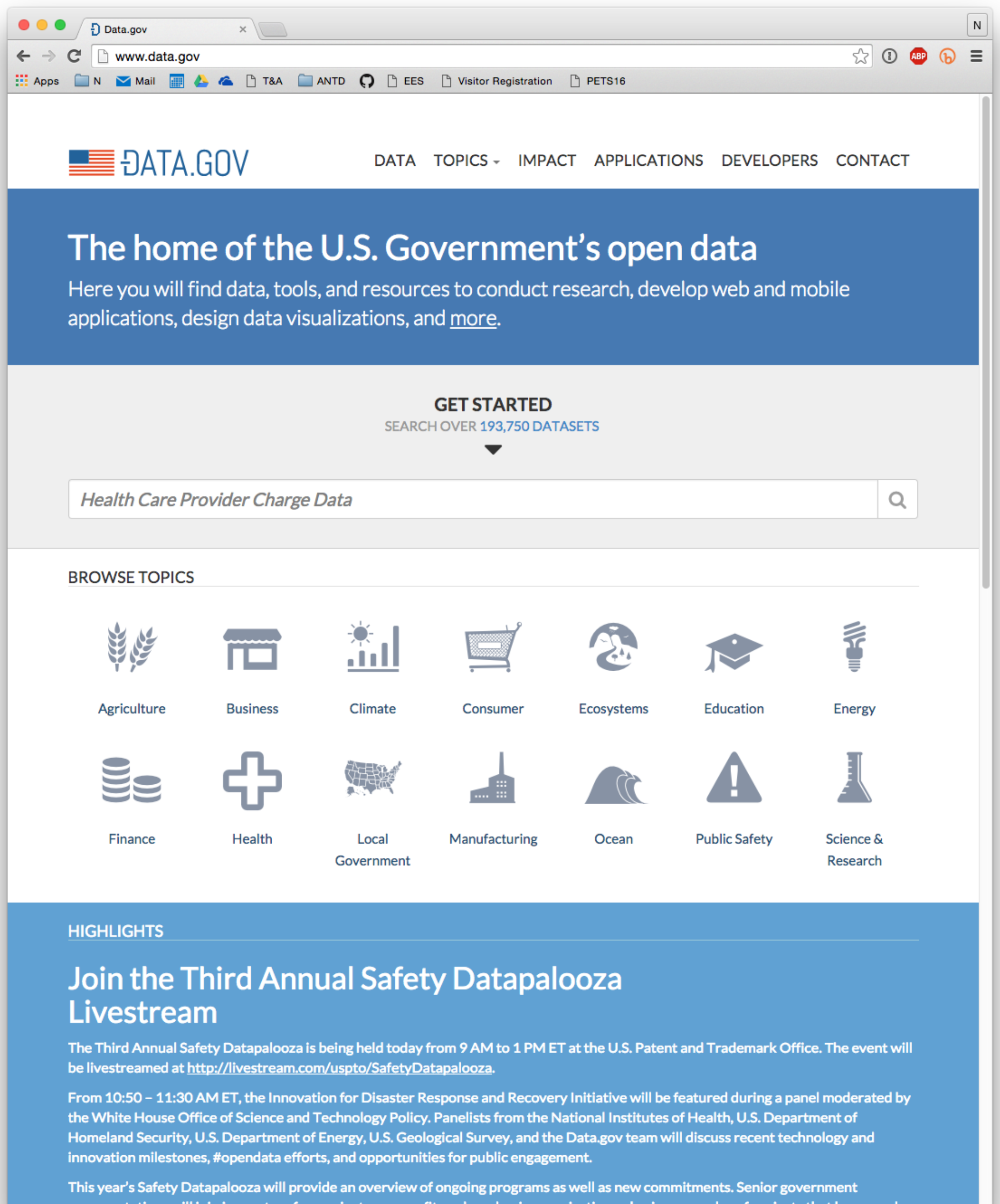


Google claims 90% of faces and 95% of license plates removed through automated processing.

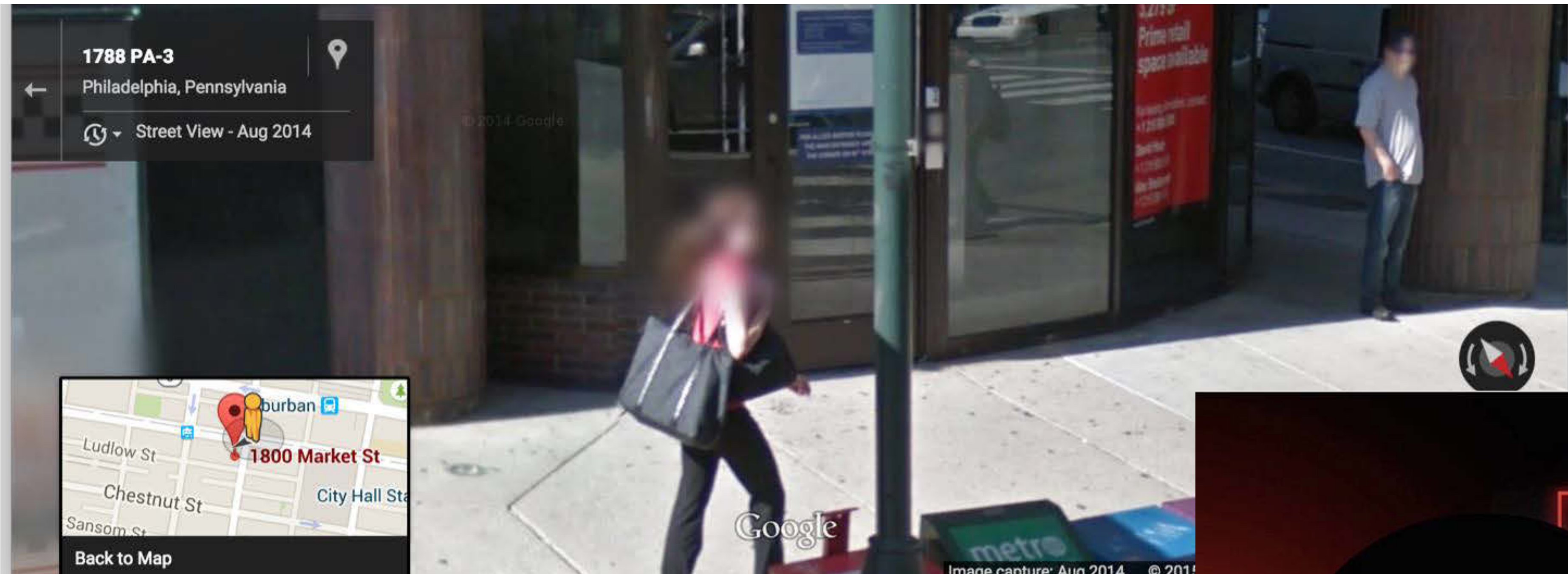
http://www.randomhistory.com/photos/2014/scoliosis-xray.jpg

Medical imagery can be highly identifying.

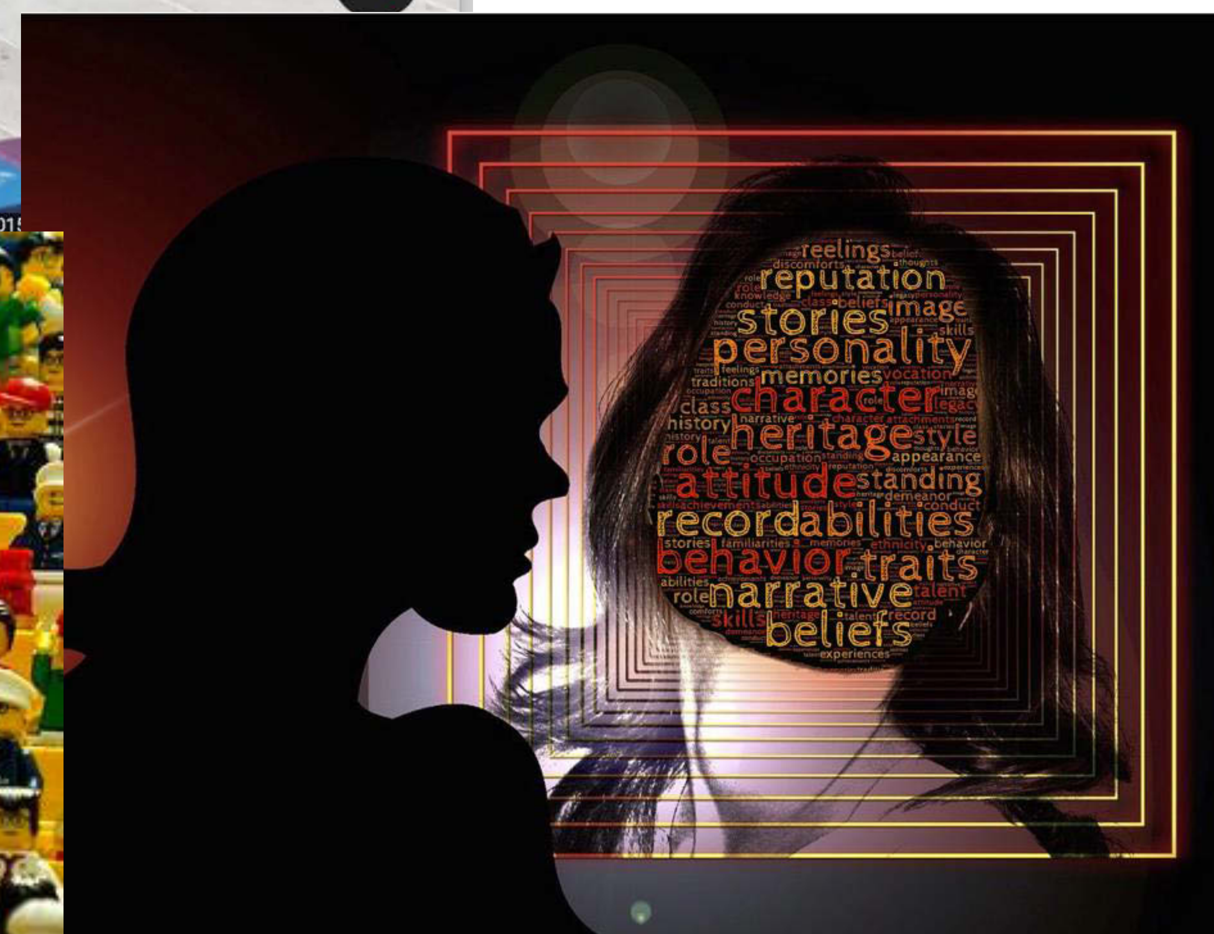
More research is needed to determine if systems can protect privacy and allow for unlimited use of data.



Can raw data be transformed so completely that individuals cannot recognize their own data once they are in a crowd?



https://pixabay.com/en/lego-doll-the-per-amphitheatre-1044891/



https://pixabay.com/en/self-self-image-image-identity-792365/

In summary:
We have learned a lot about de-identification in recent years.

The de-identification “toolkit” has several options

- suppression, generalization — *commonly used in healthcare*
- field swapping, noise addition — *commonly used in vital statistics*

K-anonymity and Differential Privacy are formal models for evaluating the quality of de-identification

We increasingly have the ability to:

- Modify data so that the data subjects’ identity is rem leaving information that is somewhat useful.
- But the more useful it is,
  the more likely it can be re-identified

We need procedures for:

- Evaluating the effectiveness of de-identification
- Evaluating the usefulness of the data that remain.

We need these techniques for a wide range of

- Structured data, text, medical, video

Lowering identifiability lowers data quality.