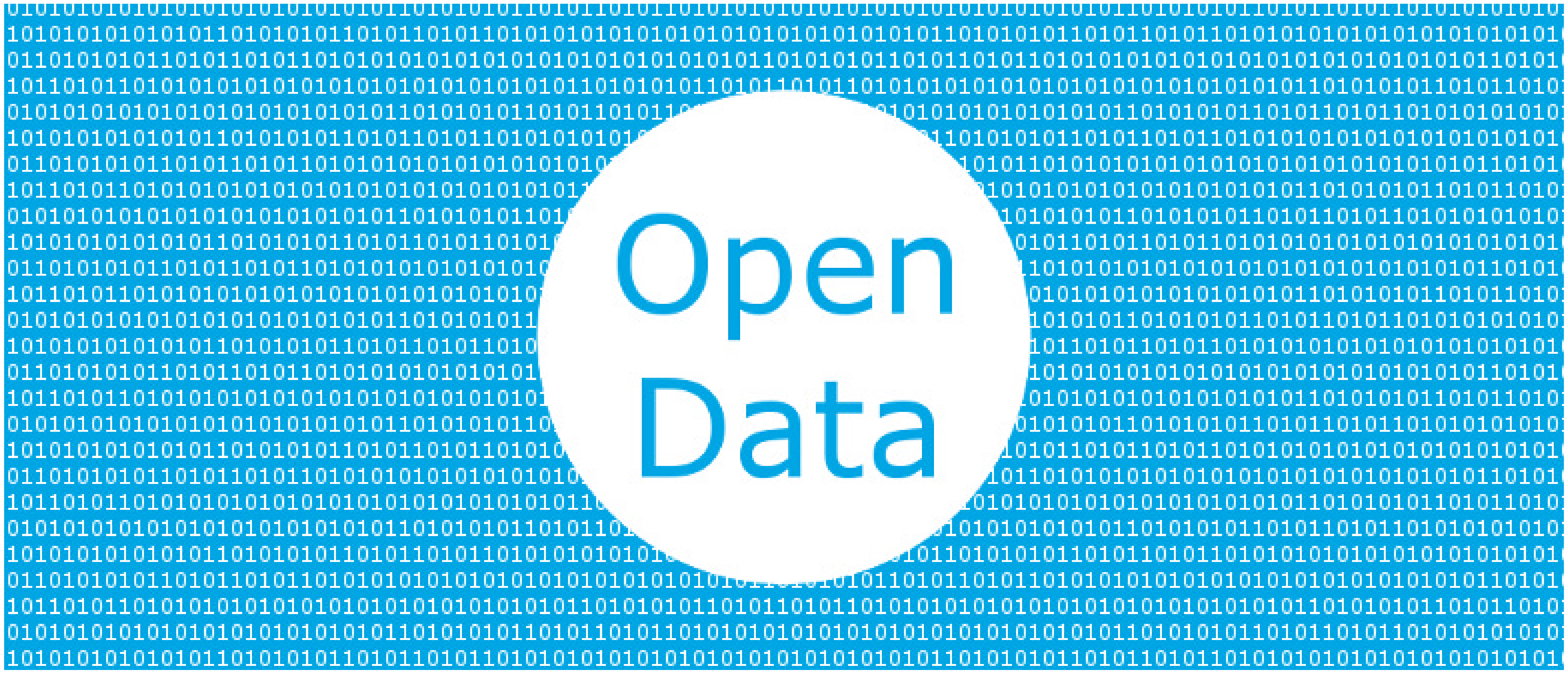
De-Identification and Differential Privacy

Special Topics on Privacy and Public Auditability
National Institute of Standards and Technology
January 27, 2020

Simson L. Garfinkel
Senior Computer Scientist for Confidentiality and Data Access
Associate Directorate for Research and Methodology
U.S. Census Bureau

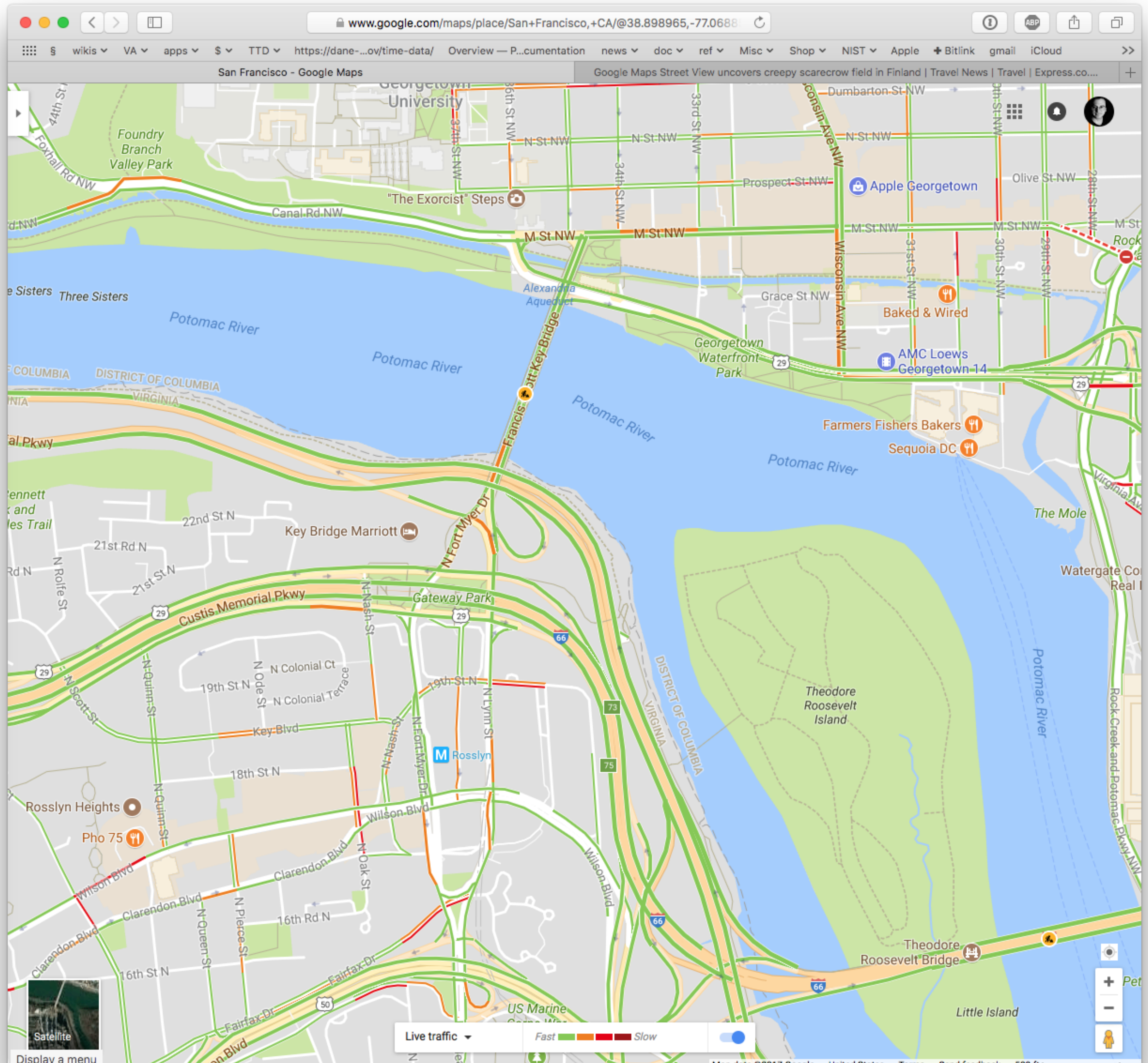
Outline for today's talk

1. Privacy risks of Open Data
2. Addressing privacy risks with de-identification
3. De-identification techniques
4. De-identification failings
5. Database reconstruction
6. Differential Privacy

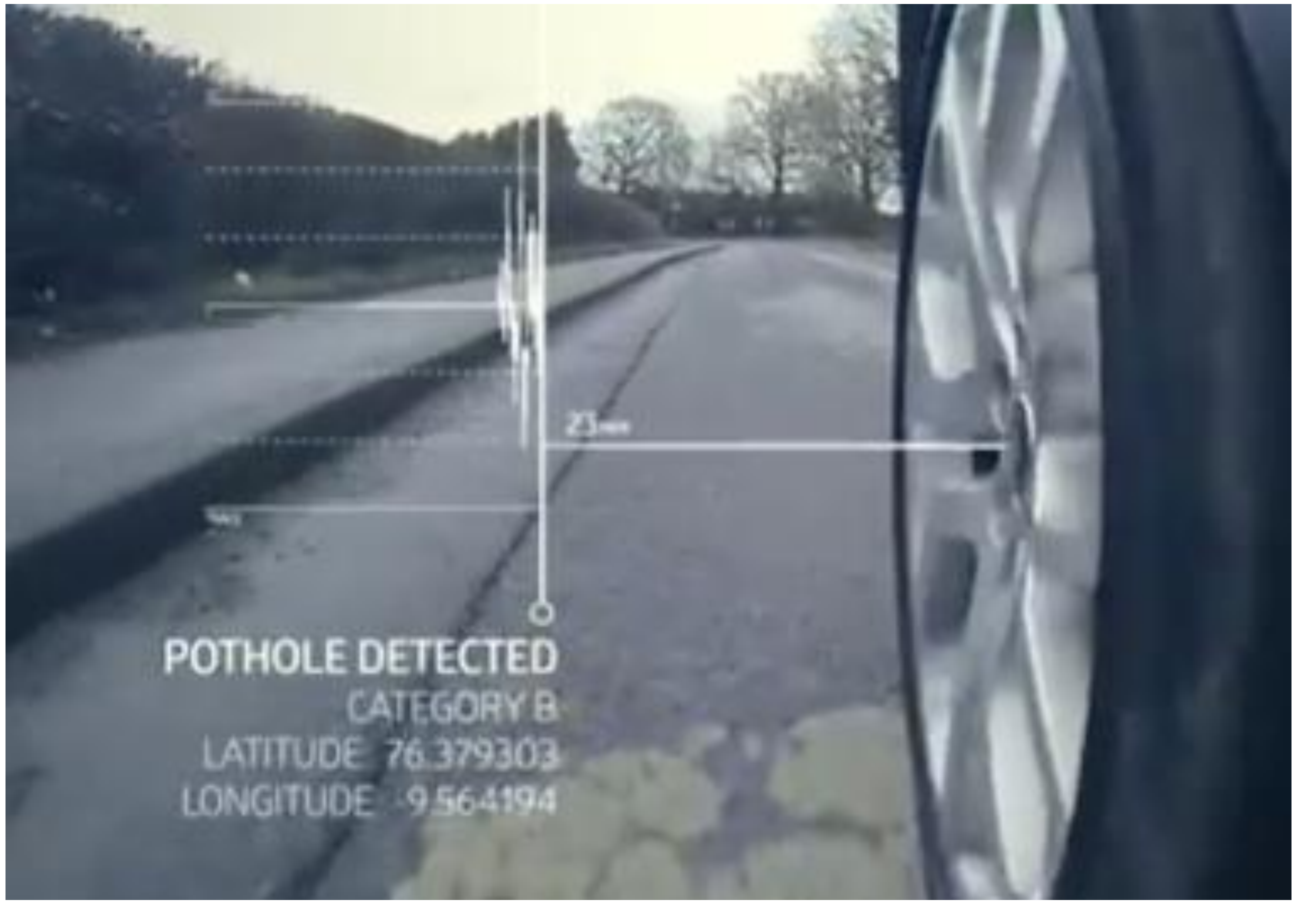


Lesson 1: People want our data

Open Data and data products using information from individuals holds great promise…



Land Rover plans to use real-time data to help you avoid the next big thing!



- “Using cellular communication, we can take anonymized data, so that the location and the severity of the pothole.
- “And we can share that in the cloud
- “As other users come along, we can provide that back to them to warn other drivers. “
- -Dr. Mike Bell, Global Connected Car Director, Jaguar Land Rover

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

Open Data concerns are shared by Federal, State, Local Governments & Organizations.

Sections

The Washington Post
Democracy Dies in Darkness

The Switch

Why the names of six people who complained of sexual assault were published online by Dallas police

By **Andrea Peterson** April 29, 2016

17

Dallas Police Department/Handout via Reuters

Dallas Open Records: What went wrong?

"The Dallas Police Department made public the names, ages, and home addresses of some alleged sexual assault victims on an official website, an incident that highlights how the push to put more police records online may also be inadvertently leaving victims exposed.

…

"The Dallas Police Department's online incident database does not appear to have included reports categorized as sexual assaults. In at least six other cases, though, the victim complained of a sexual assault or an attempted sexual assault and the incidents were labelled as "Class C Assault offenses" or simply "Injured Person." In these cases, the name and age of that victim is listed online. A few times, the home address was included as well.

"In one instance, a note says a "suspect sexually and physically assaulted" the alleged victim. Another says an "unknown suspect had unwanted sexual contact with the complainant." In some cases, the records seem to indicate that the alleged victims received follow-up sexual assault care from the department's Victim Services unit.

Open Data Lessons

1. People want our data — but the data can cause harms.

Outline for today's talk

1. Privacy risks of Open Data ☺
2. Addressing privacy risks with de-identification
3. De-identification techniques
4. De-identification failings
5. Database reconstruction
6. Differential Privacy

Lesson 2:
We can de-identify data by removing names.

Open Data is US Policy.

EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF MANAGEMENT AND BUDGET
WASHINGTON, D.C. 20503

THE DIRECTOR

May 9, 2013

M-13-13

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: Sylvia M. Burwell
Director

Steven VanRoekel
Federal Chief Information Officer

Todd Park
U.S. Chief Technology Officer

Dominic J. Mancini
Acting Administrator, Office of Information and Regulatory Affairs

SUBJECT: Open Data Policy – Managing Information as an Asset

M-13-13: Open Data must be consistent with privacy

"Consistent with OMB's Open Government Directive, agencies must adopt a presumption in favor of openness to the extent permitted by law and subject to privacy, confidentiality, security, or other valid restrictions."

In practice, do not release:

- Data that would be exempt from FOIA.
- Data that could harm an individual

Especially an issue for proactive data releases.

M-13-13 warns that data may be identifiable

Personally identifiable information: As defined in OMB Memorandum M-10-23,[17] "personally identifiable information" (PII) refers to information that can be used to distinguish or trace an individual's identity, either alone or when combined with other personal or identifying information that is linked or linkable to a specific individual. The definition of PII is not anchored to any single category of information or technology. Rather, it requires a case-by-case assessment of the specific risk that an individual can be identified. In performing this assessment, it is important for an agency to recognize that non-PII can become PII whenever additional information is made publicly available (in any medium and from any source) that, when combined with other available information, could be used to identify an individual.

Mosaic effect: The mosaic effect occurs when the information in an individual dataset, in isolation, may not pose a risk of identifying an individual (or threatening some other important interest such as security), but when combined with other available information, could pose such risk. Before disclosing potential PII or other potentially sensitive information, agencies must consider other publicly available data – in any medium and from any source – to determine whether some combination of existing data and the data intended to be publicly released could allow for the identification of an individual or pose another security concern.

Corporations have taken a similar approach. It sometimes doesn't work.

In 2006 America Online (AOL) published "search logs" to help the research community.

```
500k User Session Collection
----------------------------------------------
This collection is distributed for NON-COMMERCIAL RESEARCH USE ONLY.
Any application of this collection for commercial purposes is STRICTLY PROHIBITED.

Brief description:
This collection consists of ~20M web queries collected from ~650k users over three months.
The data is sorted by anonymous user ID and sequentially arranged.

The goal of this collection is to provide real query log data that is based on real users.
It could be used for personalization, query reformulation or other types of search
research.

The data set includes {AnonID, Query, QueryTime, ItemRank, ClickURL}.
        AnonID    - an anonymous user ID number.
        Query     - the query issued by the user, case shifted with
                    most punctuation removed.
        QueryTime - the time at which the query was submitted for search.
        ItemRank  - if the user clicked on a search result, the rank of the
                    item on which they clicked is listed.
        ClickURL  - if the user clicked on a search result, the domain portion of
                    the URL in the clicked result is listed.

Each line in the data represents one of two types of events:
        1. A query that was NOT followed by the user clicking on a result item.
        2. A click through on an item in the result list returned from a query/
```

AOL's logs were de-identified…

Anon ID	Query

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

Thelma Arnold & Dudley

Government agencies have also inadvertently revealed personal information.

In March 2014, the New York City Taxi & License Commission tweeted a "TAXI FACTS" infographic:

Chris Whong files a "Freedom of Information Law" request for *all the data* used to create the graphic.

NYC TLC provided Chris Whong with all of the data

175 million trips:

	A	B	C	D	E	F	G	H	I	J	K
1	medallion	hack_license	vendor_id	pickup_datetime	payment_type	fare_amoun	surcharge	mta_tax	tip_amount	tolls_amoun	total_amount
2	89D227B655E5C82AECF13C3F	BA96DE419E711691B944	CMT	1/1/13 15:11	CSH	6.5	0	0.5	0	0	7
3	0BD7C8F5BA12B88E0B67BED	9FD8F69F0804BDB5549F	CMT	1/6/13 0:18	CSH	6	0.5	0.5	0	0	7
4	0BD7C8F5BA12B88E0B67BED	9FD8F69F0804BDB5549F	CMT	1/5/13 18:49	CSH	5.5	1	0.5	0	0	7
5	DFD2202EE08F7A8DC9A57B0	51EE87E3205C985EF843	CMT	1/7/13 23:54	CSH	5	0.5	0.5	0	0	6
6	DFD2202EE08F7A8DC9A57B0	51EE87E3205C985EF843	CMT	1/7/13 23:25	CSH	9.5	0.5	0.5	0	0	10.5
7	20D9ECB2CA0767CF7A01564	598CCE5B9C1918568DEE	CMT	1/7/13 15:27	CSH	9.5	0	0.5	0	0	10
8	496644932DF3932605C22C79	513189AD756FF14FE670	CMT	1/8/13 11:01	CSH	6	0	0.5	0	0	6.5
9	0B57B9633A2FECD3D3B1944	CCD4367B417ED6634D9	CMT	1/7/13 12:39	CSH	34	0	0.5	0	4.8	39.3
10	2C0E91FF20A856C891483ED6	1DA2F6543A62B8ED9347	CMT	1/7/13 18:15	CSH	5.5	1	0.5	0	0	7

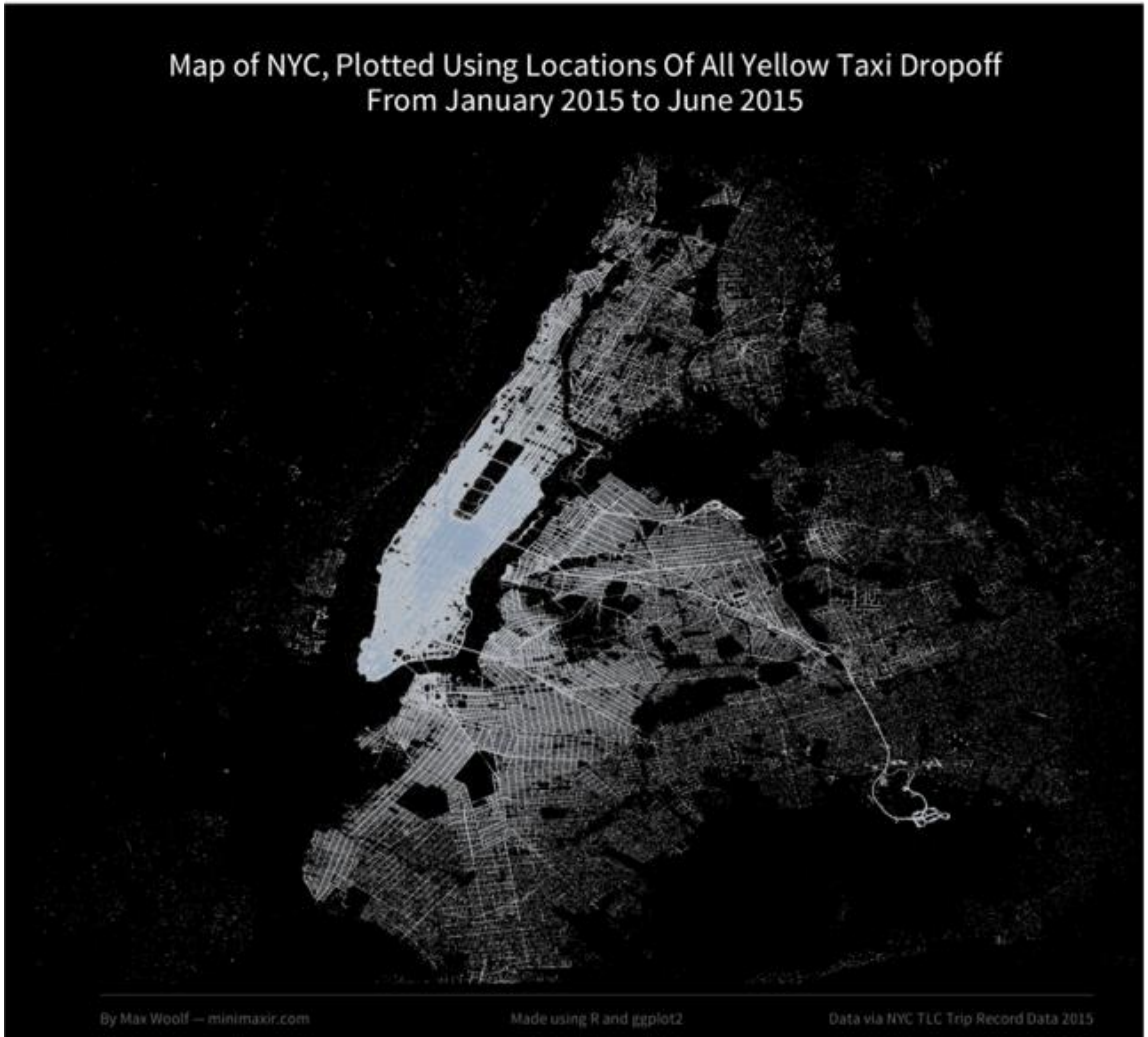
Every trip:

- Pickup date, time & GPS
- Drop-off date, time & GPS
- Fare & tip
- Encoded medallion number

Chris Whong published it on the Internet…

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

With this data, you can make a map of NYC Taxi Service



http://minimaxir.com/2015/08/nyc-map/

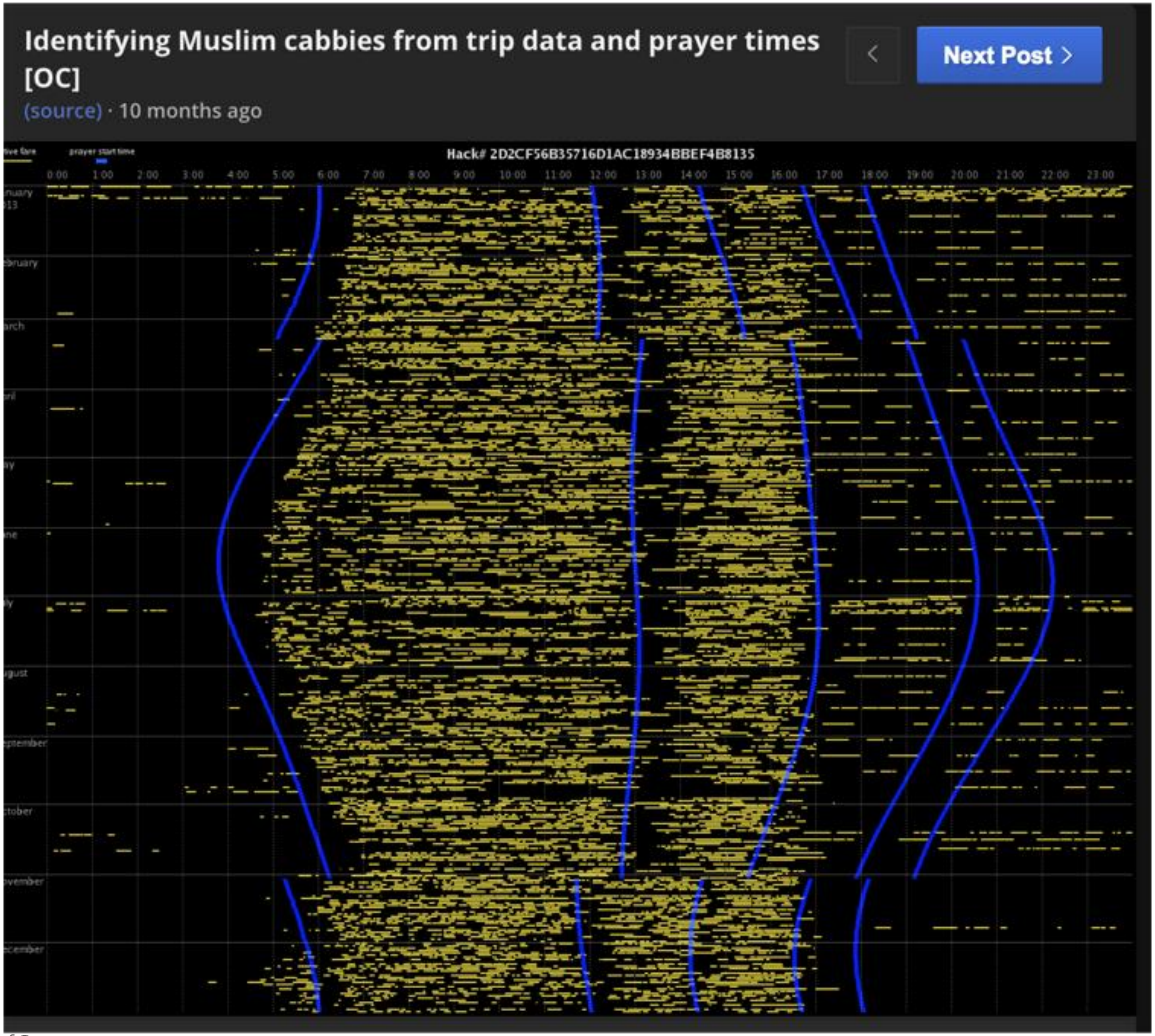
Compare taxi prices and Uber prices:

Uber more expensive

Taxi more expensive

http://qz.com/363759/data-proves-that-often-a-yellow-taxi-is-a-better-deal-than-an-uber/

Each taxi has a pseudonym, which allows taxi rides to be linked.



The taxi medallion numbers were not properly de-identified.

Pseudonym
0f76c35d4a069e0fe76b21d28f009639
be9f314926dd314b36496d926e42f4db
9ee993809f648d39d24f5ba8f862d7f1
23f7e8636fb9099822aa381054d215d4

The pseudonyms looked suspicious to Anthony Tockar, an intern at Neustar Research.

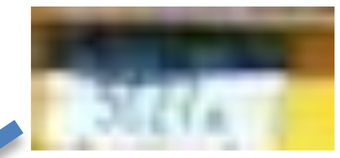
Tockar realized that the pseudonyms were MD5 hashes

MD5("5C27") = be9f314926dd314b36496d926e42f4db

Tockar performed a “brute force” attack on the hashes.

Anthony Tockar identified the medallion number the records.

He searched for photos in flickr that showed movie stars at taxis where he could read the medallion number.




Riding with the Stars: Passenger Privacy in the NYC Taxicab Dataset

SEPTEMBER 15, 2014 BY ATOCKAR 56 COMMENTS

A journalist at Gawker identified 9 other cab rides.

Open Data Lessons

1. People want our data
   — but the data can cause harms.
2. We can de-identify data by removing names
   — but people can still be identified.

Outline for today's talk

1. Privacy risks of Open Data ☺
2. Addressing privacy risks with de-identification ☺
3. De-identification techniques
4. De-identification failings
5. Database reconstruction
6. Differential Privacy

Google Street View

Lesson 3:
Beyond names, *all direct identifiers must be removed.*

Agencies want to release data to researchers

Identifying Data
Names & Address

Sensitive Data
Medical Records

Commonwealth of Massachusetts
Group Insurance Commission
Your Benefits Connection

May 18, 1996: Massachusetts Governor William Weld Collapses at Bentley College Commencement

Massachusetts Governor Doing Well After Collapse

WALTHAM, Massachusetts (CNN, May 18) -- Gov. William Weld collapsed during a graduation ceremony at Bentley College, but doctors said he was doing well.

The governor was taken to Deaconess-Waltham Hospital, where he was undergoing a battery of tests, according to Bentley College spokeswoman Katherine Blake. Weld will remain in the hospital overnight for observation, she said.

Doctors said they performed an electrocardiogram, chest X-ray and blood tests, but found no immediate cause for concern.

"With all this testing we have done, nothing acute is showing," said Dr. Rifat Dweik.

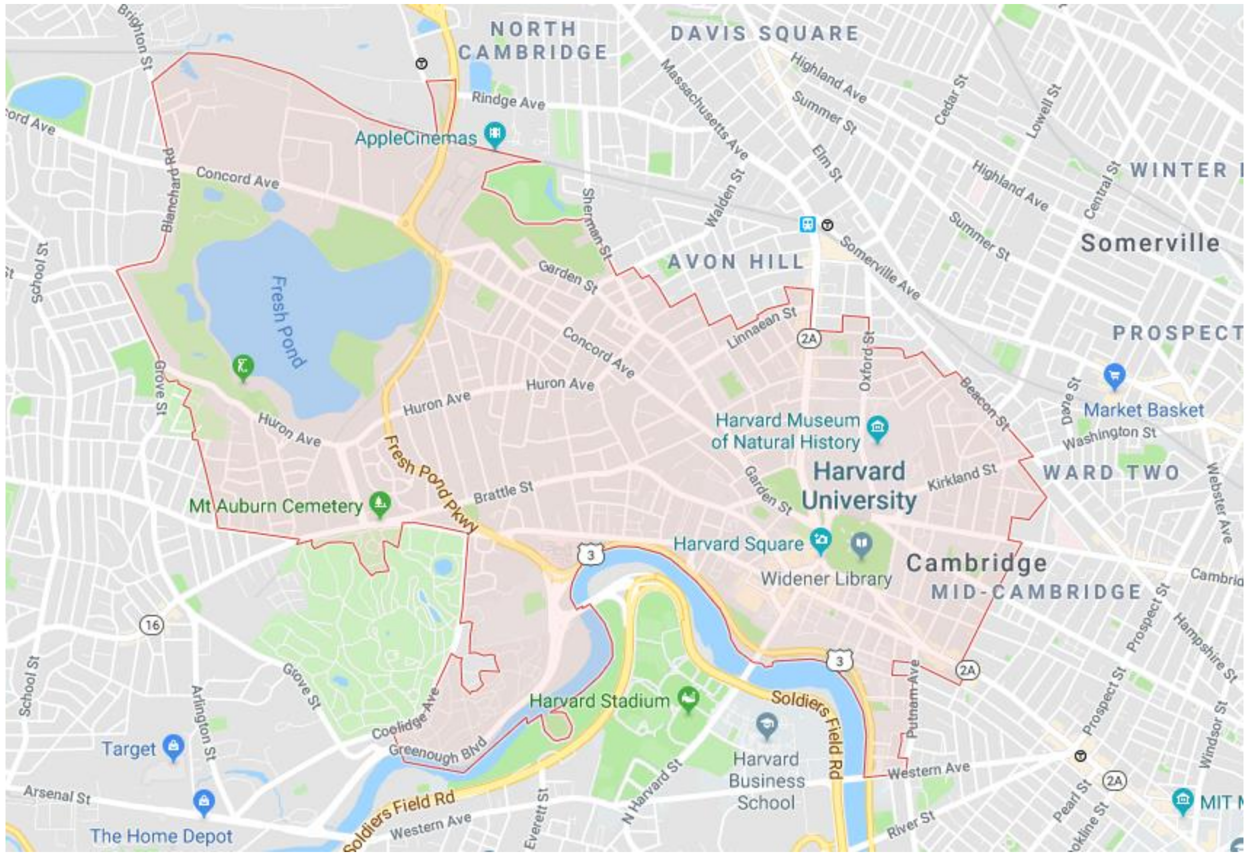
"Right now, it looks like maybe the flu," said Pam Jonah, one of Weld's press aides.

Weld was receiving an honorary doctorate of law at 11 a.m. EDT when he was stricken, according to Blake.

In 1997, MIT Graduate Student Latanya Sweeney decided to search for William Weld’s medical records in the GIC data.

Sweeney obtains GIC dataset and looks for Weld's data.

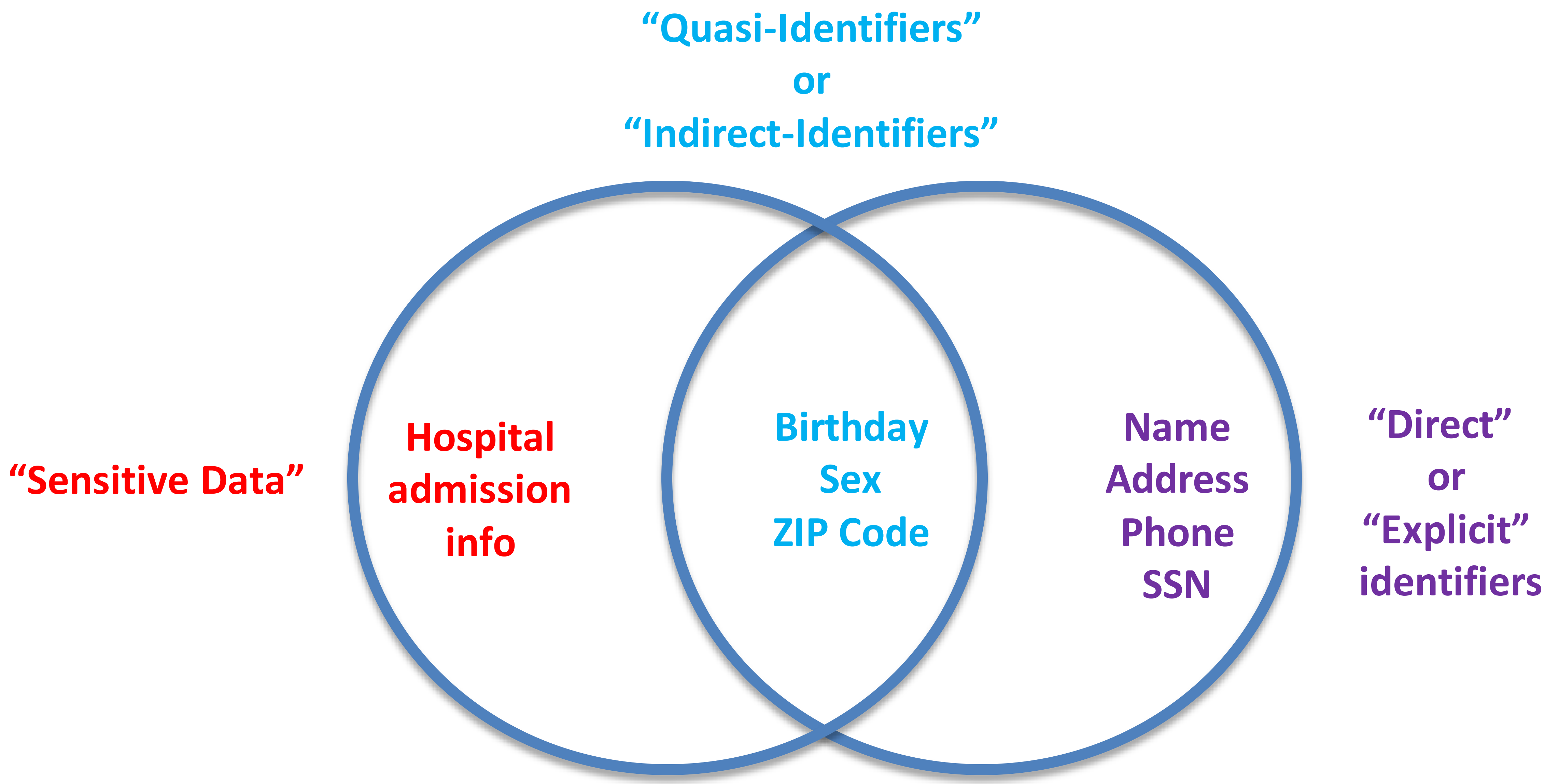
- She knew that Weld lived in Cambridge, MA.
- Sweeney purchased Cambridge voter rolls for $20.
- Six people had the same birthday (July 31, 1945)
- Three were men
- One person had the same ZIP code.



02138

"Linkage Attack"
Matching records using *quasi-identifiers*

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



Sweeney invented K-Anonymity
A model for de-identifying structured data.

A dataset that you would like to release:

Name	Race	Birthdate	Sex	Zip	Medication	Diagnosis
Alice	Black	9/20/65	M	37203	M1	Gastric Ulcer
Bob	Black	2/14/65	M	37203	M1	Gastric Ulcer
Candice	Black	10/23/65	F	37215	M1	Gastritis
Dan	Black	8/24/65	F	37215	M2	Gastritis
Eliza	Black	11/7/64	F	37215	M2	Gastritis
Felix	Black	12/1/64	F	37215	M2	Stomach Cancer
Gazelle	White	10/23/64	M	37215	M3	Flu
Harry	White	3/15/64	F	37217	M3	Flu
Irene	White	8/13/64	M	37217	M3	Flu
Jack	White	5/5/64	M	37217	M4	Pneumonia
Kelly	White	2/13/67	M	37215	M4	Pneumonia
Lenny	White	3/21/67	M	37215	M4	Flu

First you remove the identifiers…

Sweeney invented K-Anonymity
A model for de-identifying structured data.

A dataset that you would like to release:

Identifiers | Quasi Identifiers

Identifiers	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

Next, you manipulate the quasi-identifiers to remove unicity.

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.

K- anonyminity does not prevent attribute disclosure: We know all [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.

Removing or transforming direct identifiers

Removal and replacement with NULL value

Masking with a repeating character, e.g. XXXXXXXXX

Encryption

Hashing with a keyed hash

Replacing with keywords,

- "George Washington" → "PATIENT"

Replacement with realistic surrogates

- "George Washington" → "Lenny Wilkins"

Transforming quasi-identifiers

- Top and bottom coding
- Micro aggregation
- Generalization categories with small values
- Data suppression
- Blanking and imputing
- Attribute or record swapping
- Noise infusion

De-identification Caveats — what can go wrong

Mistakes happen:

- Metadata may contain identifiers.
- Direct identifiers can be missed.
- Hard to determine what's a quasi-identifier.

Open Data Lessons

1. People want our data
   — but the data can cause harms.
2. We can de-identify data by removing names
   — but people can still be identified.
3. Beyond names, *all direct identifiers must be removed.*
   Quasi-identifiers (indirect identifiers) must be manipulated.

Outline for today's talk

1. Privacy risks of Open Data ☺
2. Addressing privacy risks with de-identification ☺
3. De-identification techniques ☺
4. De-identification failings
5. Database reconstruction
6. Differential Privacy

Netflix Awards $1 Million Prize and Starts a New Contest

BY STEVE LOHR SEPTEMBER 21, 2009 10:15 AM

Jason Kempin/Getty Images Netflix prize winners, from left: Yehuda Koren, Martin Chabbert, Martin Piotte, Michael Jahrer, Andreas Toscher, Chris Volinsky and Robert Bell.

All data are potentially identifying.

The Netflix Challenge (2008-2009)

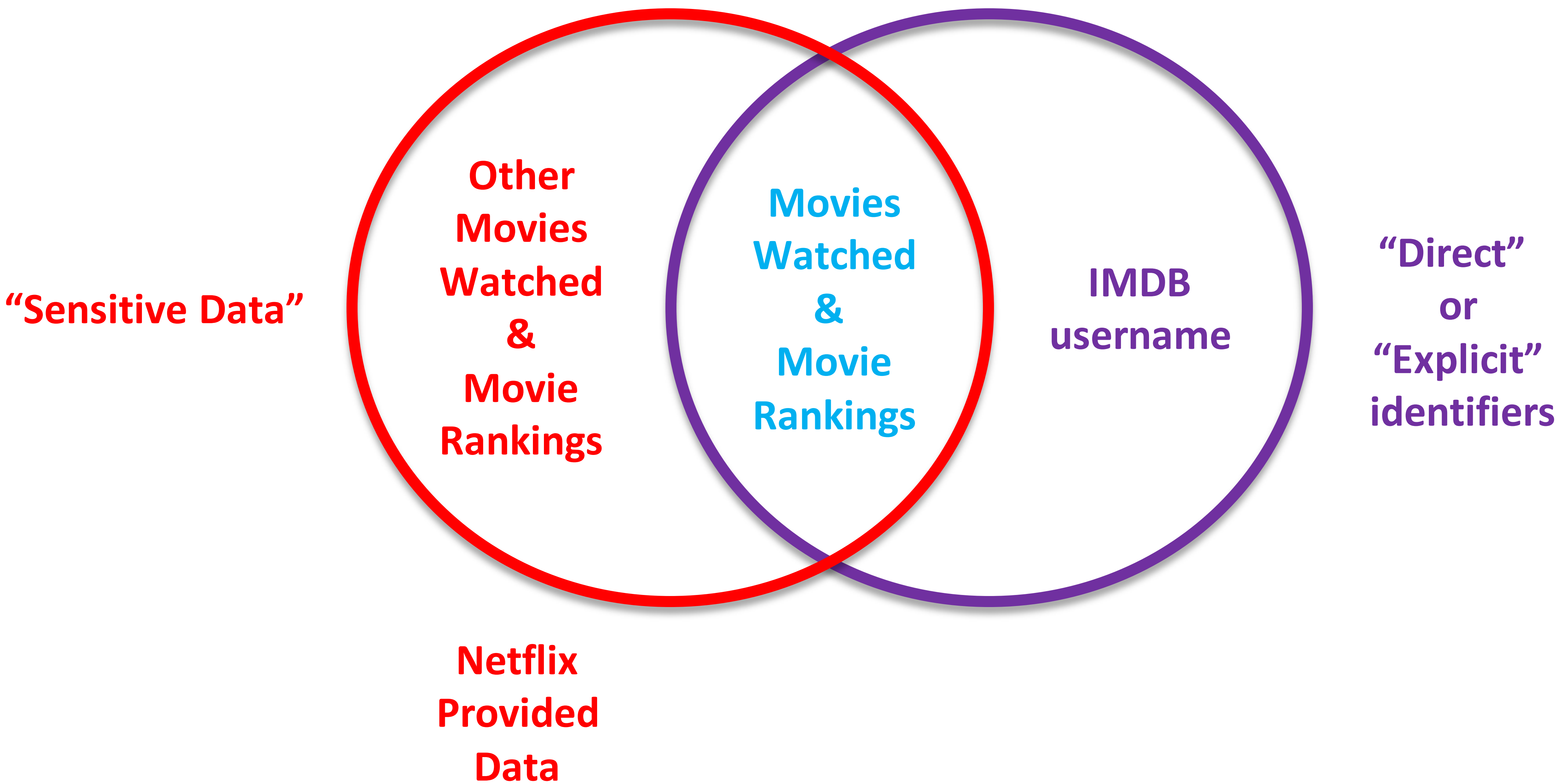
Netflix published movie data for ~450,000 subscribers:

- Pseudonymized username
- Information on movies watched:
  - *Movie Title*
  - *Date watched*
  - *Rating*

Challenge: Improve Netflix recommendation algorithm

Unintentional Challenge: Identify Netflix subscribers!

Re-identifying the Netflix Challenge Victims



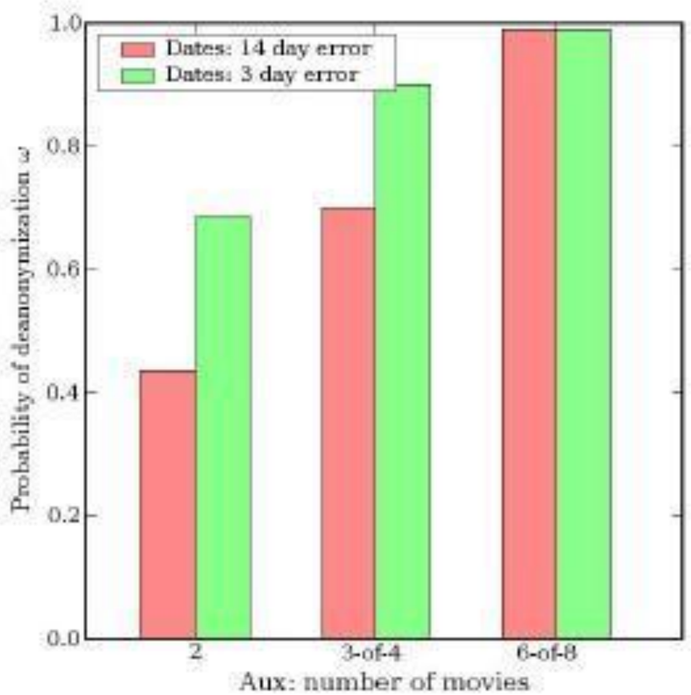


Figure 4. Adversary knows exact ratings and approximate dates.

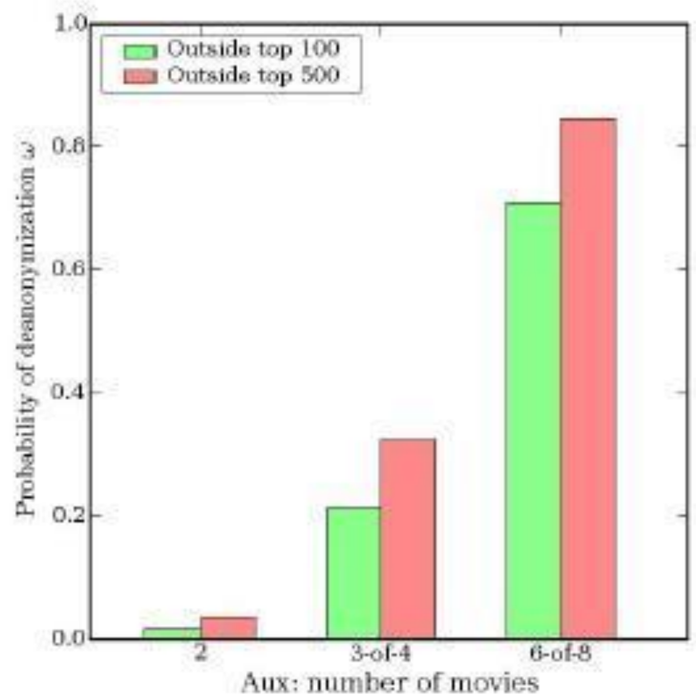


Figure 8. Adversary knows exact ratings but does not know dates at all.

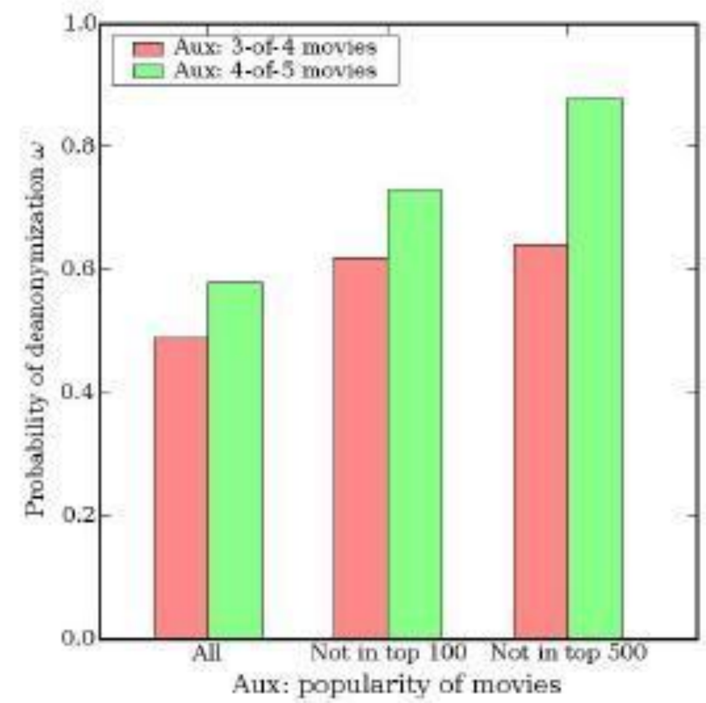


Figure 9. Effect of knowing less popular movies rated by victim. Adversary knows approximate ratings (± 1) and dates (14-day error).

MAR 12, 2010 @ 12:35 PM

Netflix Settles Privacy Lawsuit, Cancels Prize Sequel

The Firewall

the world of security **FULL BIO**

Opinions expressed by Forbes Contributors are their own.

Taylor Buley, Contributor

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.

The lawsuit called attention to academic research that suggests that Netflix indirectly exposed the movie preferences of its users by publishing anonymized customer data. In the suit, plaintiff Paul Navarro and others sought an injunction preventing Netflix from going through the so-called "Netflix Prize II," a follow-up challenge that Netflix promised would offer up even more personal data such as genders and zipcodes.

"Netflix is not going to pursue a sequel to the Netflix Prize," says spokesman Steve Swasey. "We looked at this, we heard some dissension and so we've settled it, resolved the issues and are moving on."

Netflix's decision to forestall the so-called "Netflix Prize II" was part of the settlement agreement, says Scott Kamber, the plaintiff's attorney. Also part of the settlement and per industry norms, Netflix is not admitting any wrongdoing.

Open Data Lessons

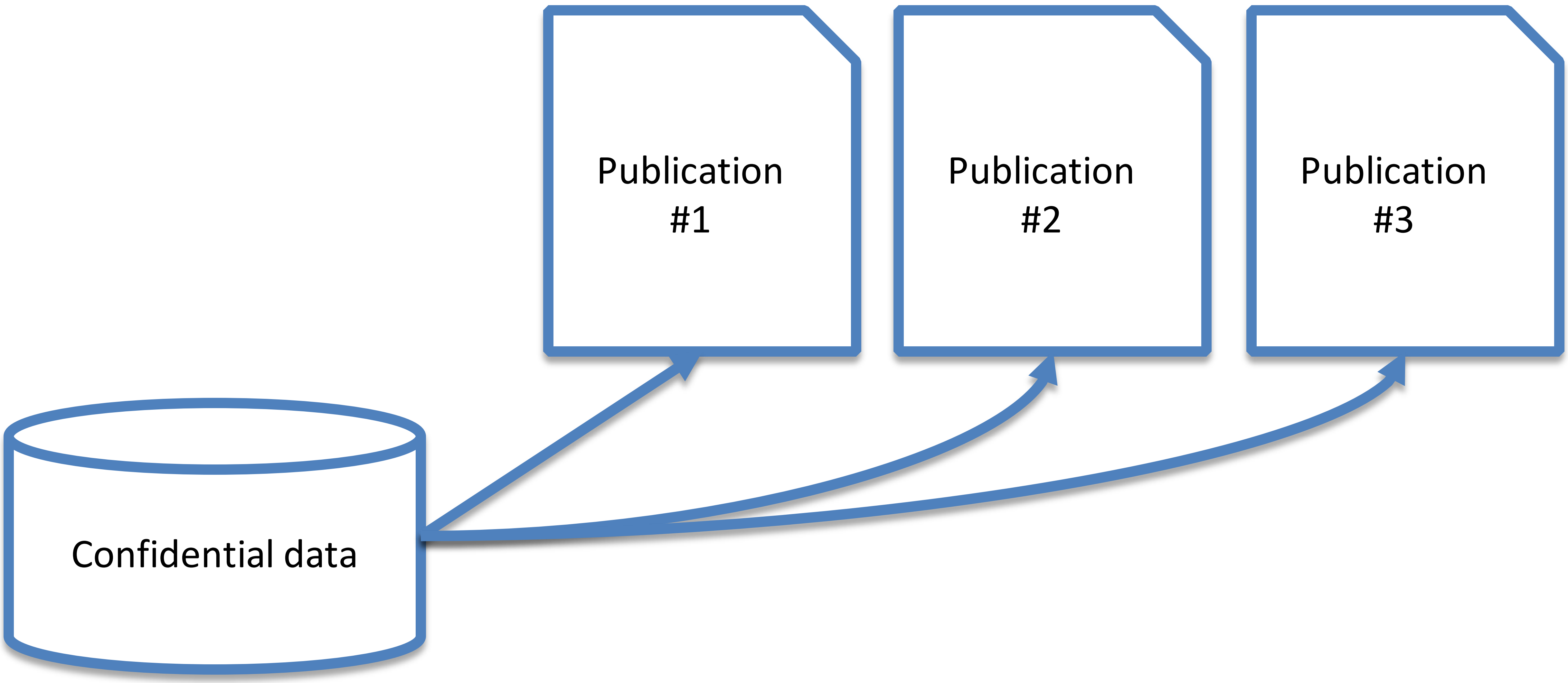
1. People want our data
   — but the data can cause harms.
2. We can de-identify data by removing names
   — but people can still be identified.
3. Beyond names, *all direct identifiers must be removed.*
   Quasi-identifiers (indirect identifiers) must be manipulated.
4. ***All data are potentially identifying.***

Outline for today's talk

1. Privacy risks of Open Data ☺
2. Addressing privacy risks with de-identification ☺
3. De-identification techniques ☺
4. De-identification failings ☺
5. Database reconstruction
6. Differential Privacy

Lesson 5: Database Reconstruction is real.

Every time an agency publishes statistics based on confidential data, a little bit of information is revealed.



At some point, enough statistics are released that the confidential database can be "reconstructed."

Even when values are suppressed, confidential data can be reconstructed.

Item	Group	Number	Average Age
1A	Individuals	10	40
1B	Males	5	34
1C	Females	5	46
1D	Whites	5	50
1E	Blacks	5	30
2A	Children (0-17)	3	10
2D	White children	1	[redacted]
2E	Black children	2	10
3A	Parents	4	32.5
3B	Male parents	2	30
3C	Female parents	2	35
3F	Parents over 40	0	–
4A	Grandparents	3	80
4B	Male grandparents	1	[redacted]
4C	Female grandparents	2	75
4D	White grandparents	2	80
4E	Black grandparents	1	[redacted]
5A	Households	2	40
5B	Tri-generational households	0	–
5C	Single-parent households	0	–
5D	Childless households	1	[redacted]

One way to reconstruct:
create simultaneous equations consistent with published data.

ID	Household	Age	Sex	Race	Generation
1	H1	A1	S1	R1	G1
2	H2	A2	S2	R2	G2
3	H3	A3	S3	R3	G3
4	H4	A4	S4	R4	G4
5	H5	A5	S5	R5	G5
6	H6	A6	S6	R6	G6
7	H7	A7	S7	R7	G7
8	H8	A8	S8	R8	G8
9	H9	A9	S9	R9	G9
10	H10	A10	S10	R10	G10

Key	Value
Male	0
Female	1
White	0
Black	1
Child	0
Parent	1
Grandparent	2

"Average age is 40:"

$$\frac{A1 + A2 + A3 + A4 + \ldots + A10}{10} = 40$$

In 2003, Dinur & Nissim showed that statistical databases can be reconstructed with far less data than was previously thought.

Revealing Information while Preserving Privacy

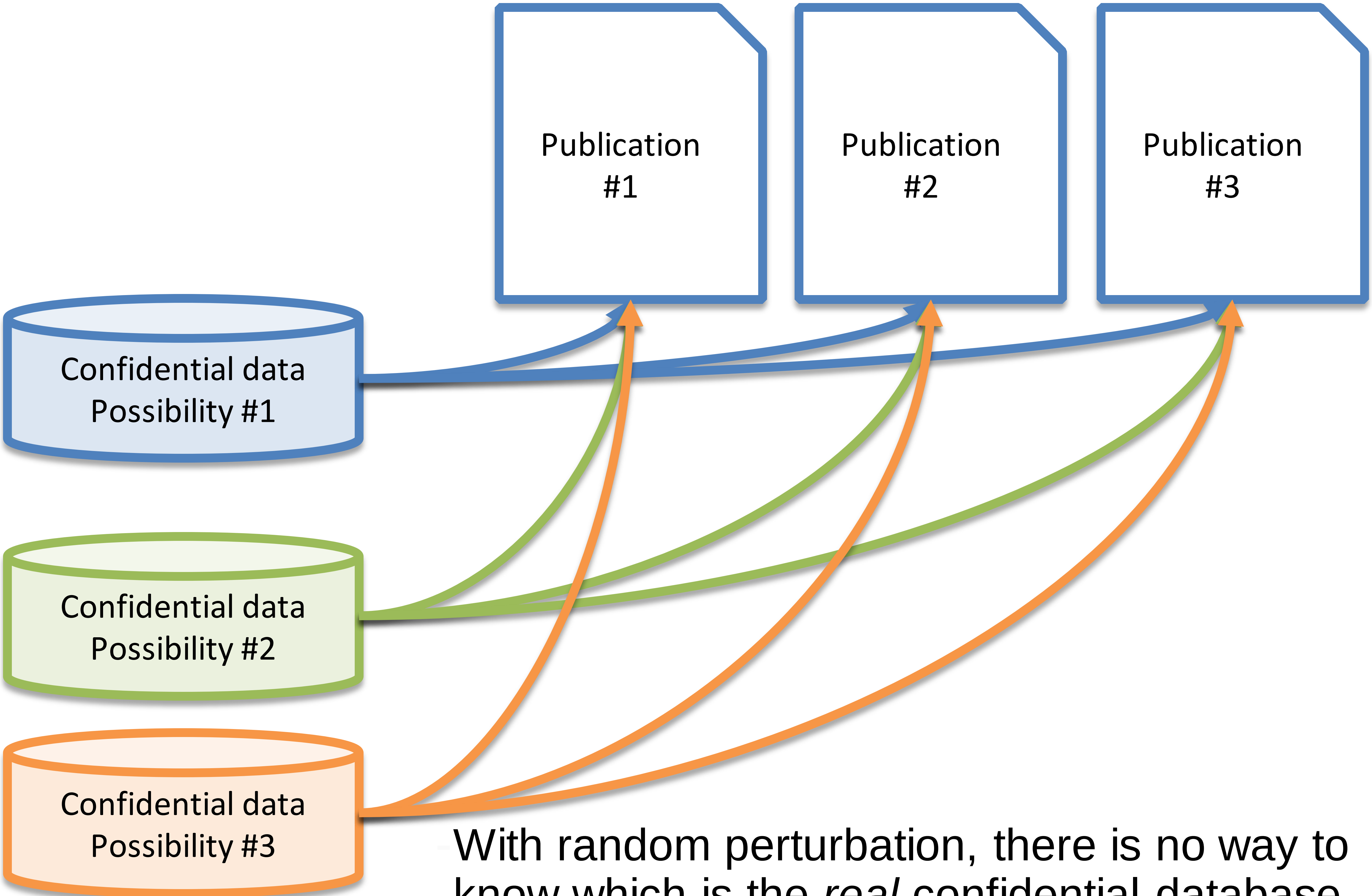
Irit Dinur Kobbi Nissim*
NEC Research Institute
4 Independence Way
Princeton, NJ 08540
{iritd,kobbi}@research.nj.nec.com

ABSTRACT

We examine the tradeoff between privacy and usability of statistical databases. We model a statistical database by an n-bit string $d_1, .., d_n$, with a query being a subset $q \subseteq [n]$ to be answered by $\sum_{i \in q} d_i$. Our main result is a polynomial reconstruction algorithm of data from noisy (perturbed) subset sums. Applying this reconstruction algorithm to statistical databases we show that in order to achieve privacy one has to add perturbation of magnitude $\Omega(\sqrt{n})$. That is, smaller perturbation always results in a strong violation of privacy. We show that this result is tight by exemplifying access algorithms for statistical databases that preserve privacy while adding perturbation of magnitude $\tilde{O}(\sqrt{n})$.

For time-$\mathcal{T}$ bounded adversaries we demonstrate a privacy-preserving access algorithm whose perturbation magnitude is $\approx \sqrt{\mathcal{T}}$.

Perturbation (noise infusion) is the only way to protect against database reconstruction.



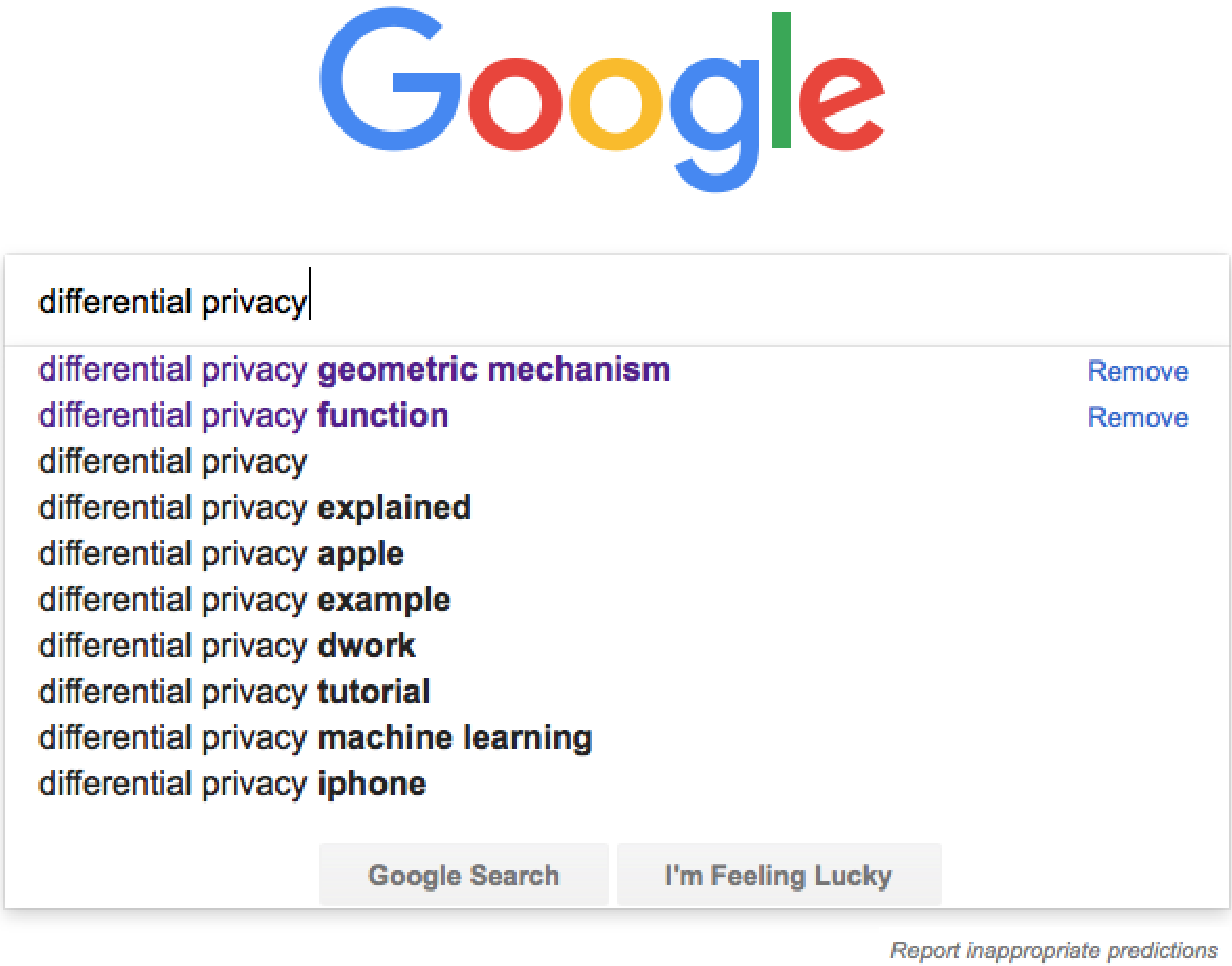
With random perturbation, there is no way to know which is the *real* confidential database.

Open Data Lessons

1. People want our data
   — but the data can cause harms.
2. We can de-identify data by removing names
   — but people can still be identified.
3. Beyond names, *all direct identifiers must be removed.*
   Quasi-identifiers (indirect identifiers) must be manipulated.
4. ***All data are potentially identifying.***
5. ***Database reconstruction is a real threat* —** but it can be addressed using perturbation.

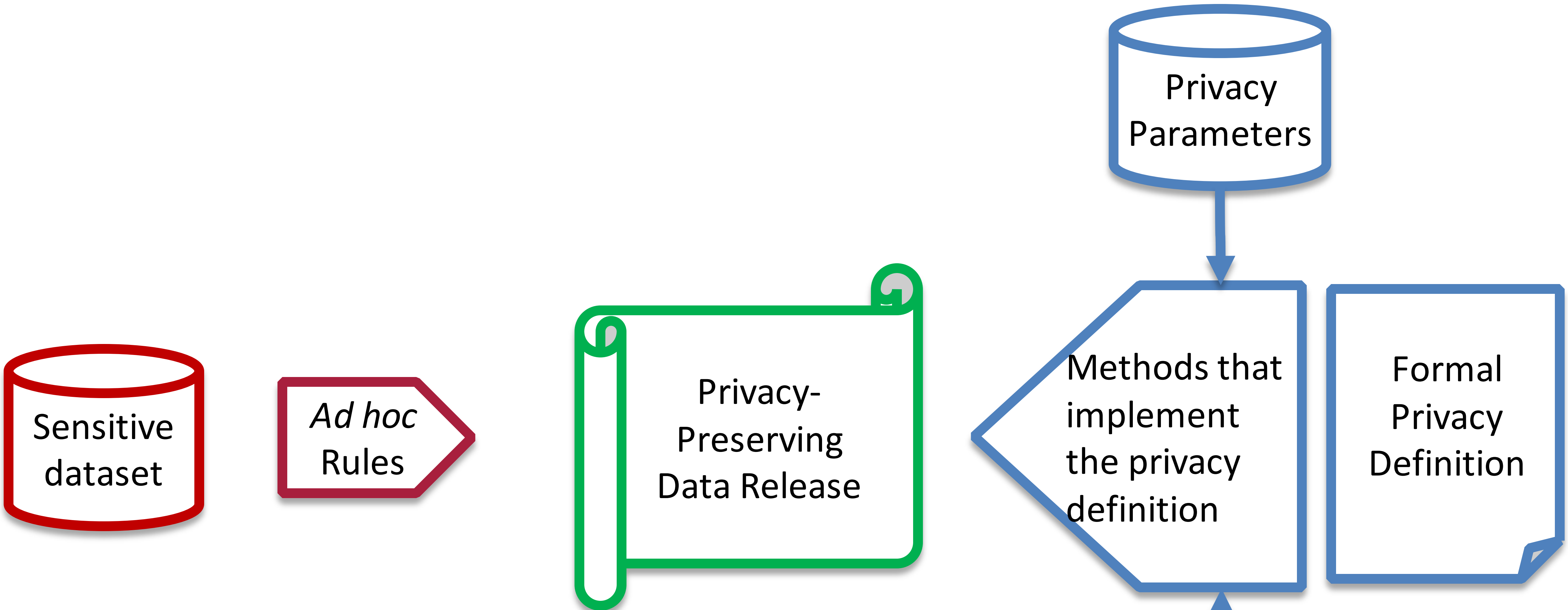
Outline for today's talk

1. Privacy risks of Open Data ☺
2. Addressing privacy risks with de-identification ☺
3. De-identification techniques ☺
4. De-identification failings ☺
5. Database reconstruction☺
6. Differential Privacy



Differential Privacy: The Big Idea

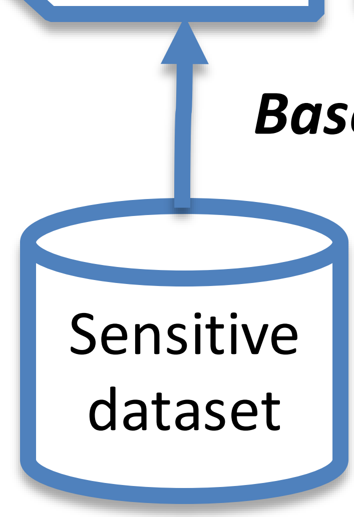
Differential privacy is a new approach for assuring privacy in the release of statistical data.



Based on hope and assumptions.

1. Data are identify, quasi-identifying, or not-identifying
2. Future data sets will not be released that can be linked with previously released data
3. Adversaries have limited resources to pursue re-identification attacks

Based on math.



In traditional data publications, there are many ways that the contributions of an individual can leak out

January

Name	Affect	Grade
Alex	Sad	30
Bobbie	Sad	50
Casey	Happy	80
Harper	Happy	100

Statistical Tabulation

Students: 4
Percent Happy: 50%
Average Grade: 65

February

Name	Affect	Grade
Alex	Sad	30
Bobbie	Sad	50
Casey	Happy	80
Emerson	Sad	90
Harper	Happy	100

Statistical Tabulation

Students: 5
Percent Happy: 40%
Average Grade: 70

It's pretty easy to determine that the new kid is sad and has a 90.

Differential privacy's core idea:
Create uncertainty regarding the presence any person in the dataset.

Noise is added to mask an individual's contribution

January

Name	Affect	Grade
Alex	Sad	30
Bobbie	Sad	50
Casey	Happy	80
Harper	Happy	100

Statistical Tabulation + noise

Students: 4
Percent Happy: 45%
Average Grade: 50

February

Name	Affect	Grade
Alex	Sad	30
Bobbie	Sad	50
Casey	Happy	80
Emerson	Sad	90
Harper	Happy	100

Statistical Tabulation + noise

Students: 5
Percent Happy: 60%
Average Grade: 75

If we ran the statistics different times, we would get different results

January

Name	Affect	Grade
Alex	Sad	30
Bobbie	Sad	50
Casey	Happy	80
Harper	Happy	100

Statistical Tabulation + noise

Students: 4
Percent Happy: 45%
Average Grade: 50

January

Name	Affect	Grade
Alex	Sad	30
Bobbie	Sad	50
Casey	Happy	80
Harper	Happy	100

Statistical Tabulation + noise

Students: 4
Percent Happy: 55%
Average Grade: 75

January

Name	Affect	Grade
Alex	Sad	30
Bobbie	Sad	50
Casey	Happy	80
Harper	Happy	100

Statistical Tabulation + noise

Students: 4
Percent Happy: 51%
Average Grade: 60

In this example, a *policy decision* requires that the number of students be accurately reported.

Data users understand that noise has been added.

January

Name	Affect	Grade
Alex	Sad	30
Bobbie	Sad	50
Casey	Happy	80
Harper	Happy	100

Statistical Tabulation + noise

Students: 3
Percent Happy: 40%
Average Grade: 50

January

Name	Affect	Grade
Alex	Sad	30
Bobbie	Sad	50
Casey	Happy	80
Harper	Happy	100

Statistical Tabulation + noise

Students: 6
Percent Happy: 45%
Average Grade: 45

January

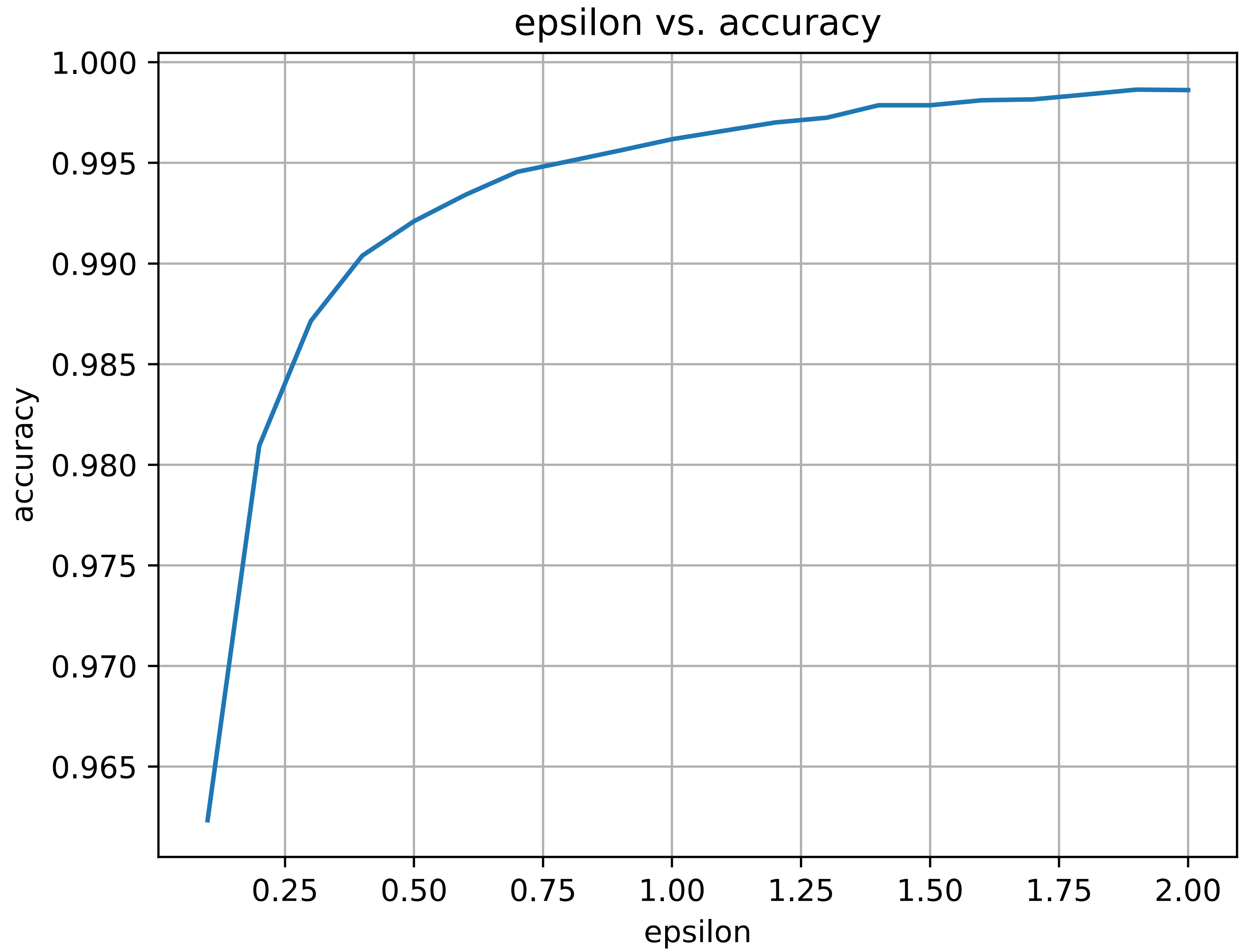
Name	Affect	Grade
Alex	Sad	30
Bobbie	Sad	50
Casey	Happy	80
Harper	Happy	100

Statistical Tabulation + noise

Students: 5
Percent Happy: 51%
Average Grade: 60

In this example, a *policy decision* requires that the exact number of students in the class be confidential.

How much noise do we add?
That is a policy decision



Differential privacy uses the parameter ε (epsilon) to describe the privacy/accuracy tradeoff.

ε = 0 — No accuracy, full privacy

ε = ∞ — No privacy, full accuracy

Noise can be added in two places:
1) When data are collected. 2) When statistics are produced.

Input noise infusion:

Name	Affect	Grade
Alex	Sad + NOISE	30 + NOISE
Bobbie	Sad + NOISE	50 + NOISE
Casey	Happy + NOISE	80 + NOISE
Harper	Happy + NOISE	100 + NOISE

Statistical Tabulation

Students: 4
Percent Happy: 30..70
Average Grade: 50..80

Advantages:

- » Tabulator need not be trusted.
- » More statistics do not pose additional privacy threats.

Output noise infusion:

Name	Affect	Grade
Alex	Sad	30
Bobbie	Sad	50
Casey	Happy	80
Harper	Happy	100

Statistical Tabulation

Students: 4
Percent Happy: 40..60
Average Grade: 60..70

Advantages:

- » More accurate for the same level of privacy
- » Allows uses of confidential data that do not involve publication.

Other choices for policy makers

Where should the accuracy be spent?

What values should be reported exactly (with no privacy)

What are the possible bounds (sensitivity) of a person's data?

e.g. If reporting average student age, can students be 5..18 or 5..115?

How do we convey privacy guarantees to public?

Final problem: what do we do about microdata?

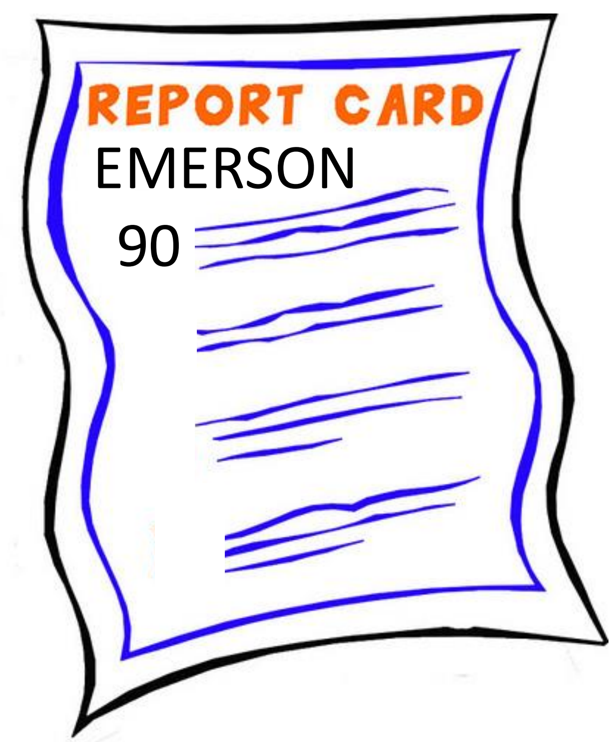
Let's say we want to publish this microdata:

Name	Affect	Grade
Alex	Sad	30
Bobbie	Sad	50
Casey	Happy	80
Emerson	Sad	90
Harper	Happy	100

NOT DIFFERENTIALLY PRIVATE

ID#	Affect	Grade
1	Sad	30
2	Sad	50
3	Happy	80
4	Sad	90
5	Happy	100

Now say Emerson's report card is lost on the way home:



As a result of the data release, Emerson's affect can be determined from the microdata.

The only solution is to add noise to microdata or produce synthetic microdata.

Differential privacy was invented in 2006 by Dwork, McSherry, Nissim and Smith

Differential privacy is just 11 years old.

Today's public key cryptography was invented in 1976-1978

Remember public key cryptography in 1989?

- No standardized implementations. No SSL/TLS. No S/MIME or PGP.
- Very few people knew how to build systems that used crypto.

Open Data Lessons

1. People want our data
   — but the data can cause harms.
2. We can de-identify data by removing names
   — but people can still be identified.
3. Beyond names, *all direct identifiers must be removed.*
   Quasi-identifiers (indirect identifiers) must be manipulated.
4. ***All data are potentially identifying.***
5. ***Database reconstruction is a real threat* —** but it can be addressed using perturbation.
6. Differential privacy provides mathematical guarantees for privacy:
   - *Requires that we accept privacy/accuracy trade-off.*
   - *Requires determining the amount of privacy (or accuracy)*
   - *Makes releasing microdata <u>really hard</u>.*
   - *We are just beginning to learn how to use it.*

Simson.L.Garfinkel@census.gov