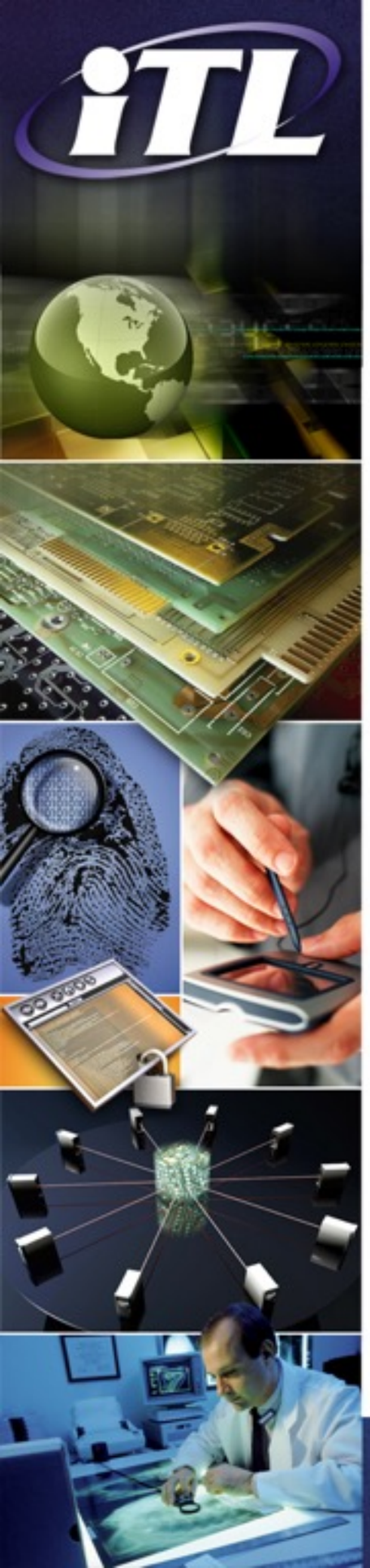


De-identification of personal Information

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

Oct. 26, 2016
Privacy+Security Forum
Washington DC

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

De-identification is a tool for protecting privacy.

De-identification removes information from a dataset so that individuals cannot be identified.

Organizations have *lots of data.*

- Much of this data has personal information.
- Privacy concerns put limits on data use.

Examples:

- Public health data. (Disease incidences)
- Educational data.
- Marketing data (clickstreams; shopping carts)

The goal of de-identification is to:

- Remove the “sensitive” personal data.
- Allow use of the remaining “non-sensitive” data.

But de-identification can go wrong.
Sometimes de-identified data can be re-identified.

March 2014 — New York City Taxi & License Commission tweets a "TAXI FACTS" Infographic:

2014 — Chris Whong files a "Freedom of Information Law" request for *all the data* that was used to create this graphic.

TLC provides Chris Whong with all the data

175 million trips

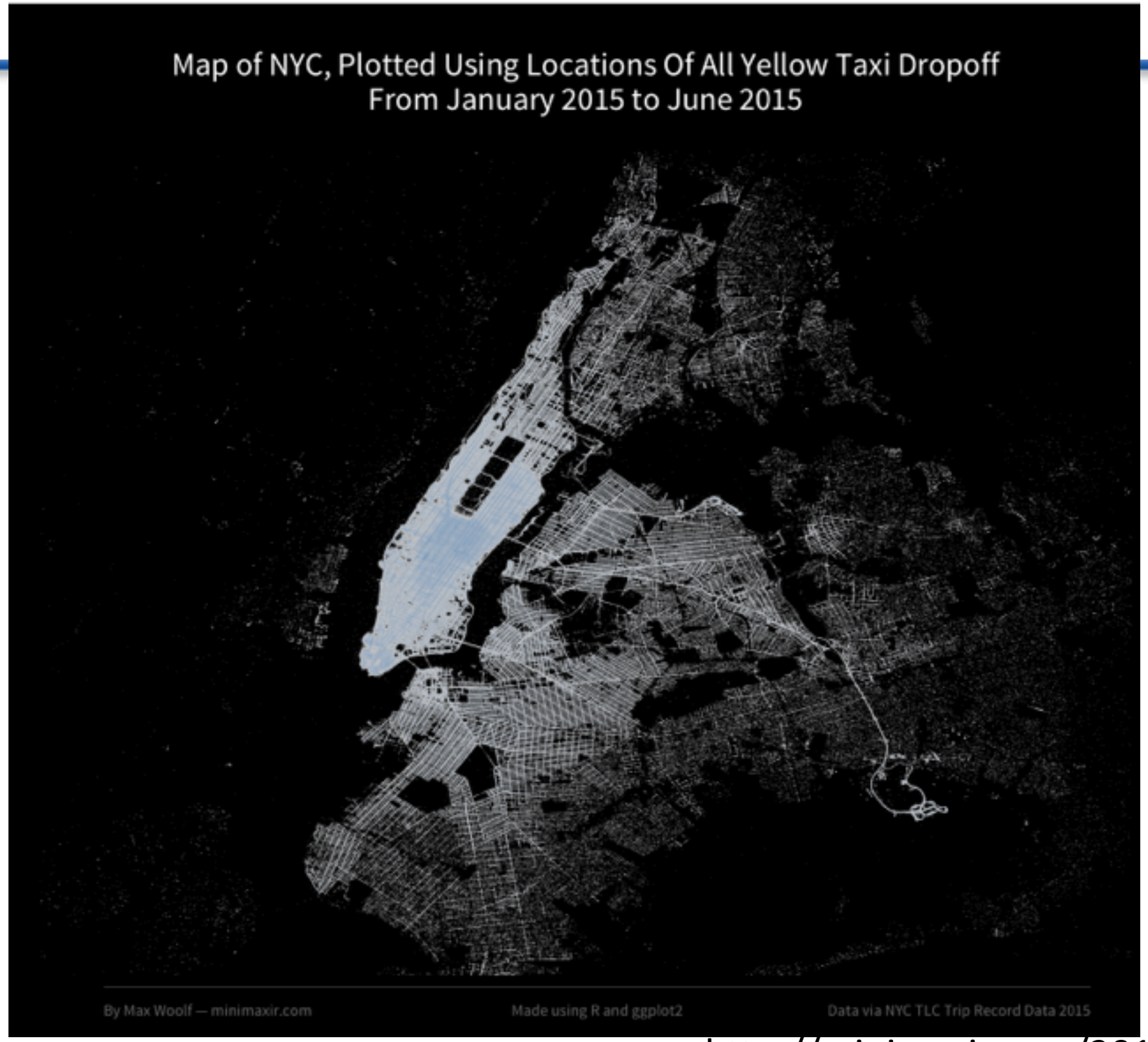
	A	B	C	D	E	F	G	H	I	J	K
1	medallion	hack_license	vendor_id	pickup_datetime	payment_type	fare_amoun	surcharge	mta_tax	tip_amount	tolls_amoun	total_amount
2	89D227B655E5C82AECF13C3F	BA96DE419E711691B944	CMT	1/1/13 15:11	CSH	6.5	0	0.5	0	0	7
3	0BD7C8F5BA12B88E0B67BED	9FD8F69F0804BDB5549F	CMT	1/6/13 0:18	CSH	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	CCD4367B417ED6634D98	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.

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

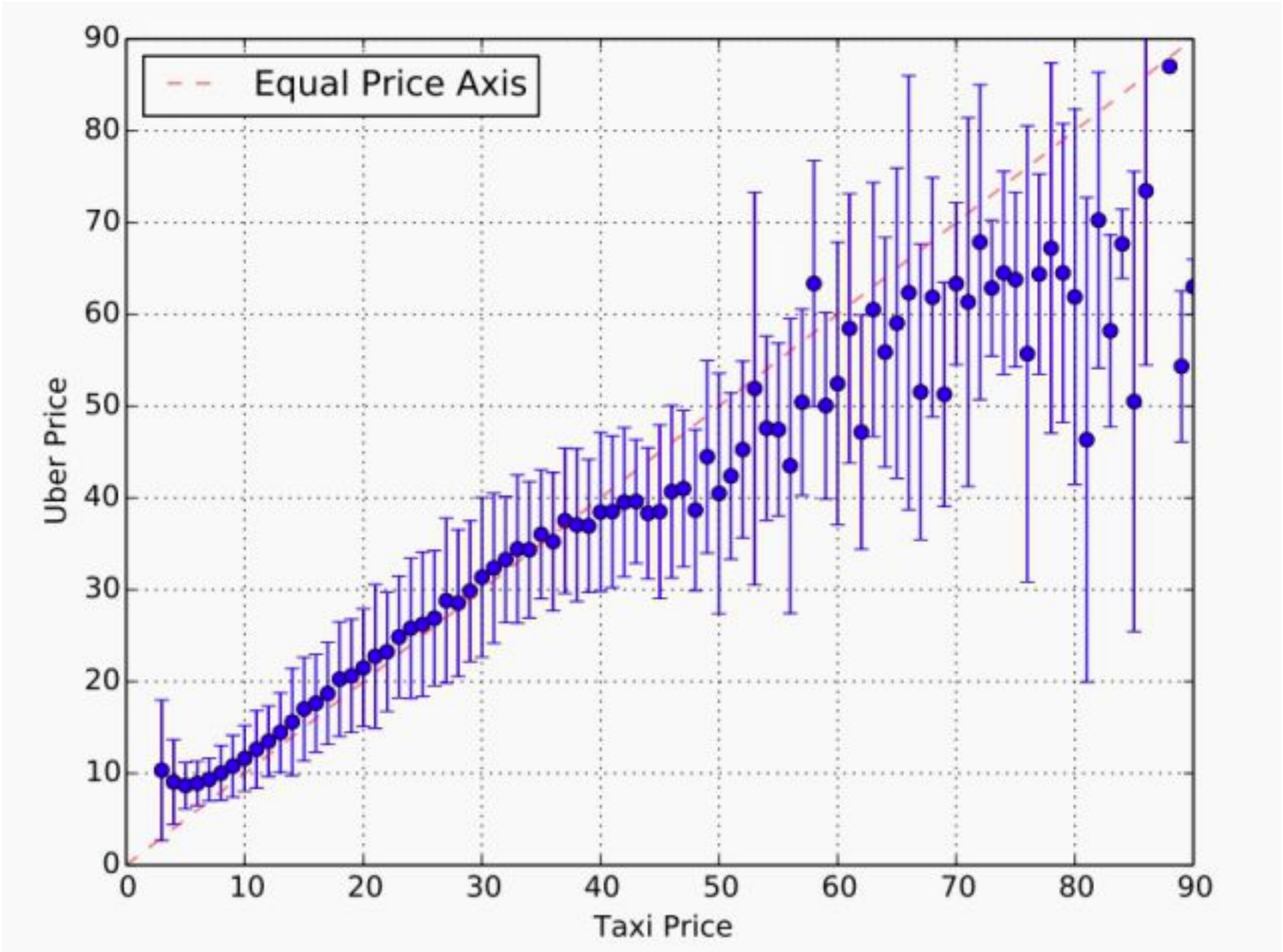
Map of NYC Taxi Service:



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

Comparison of Taxi prices and Uber prices:

Uber more expensive



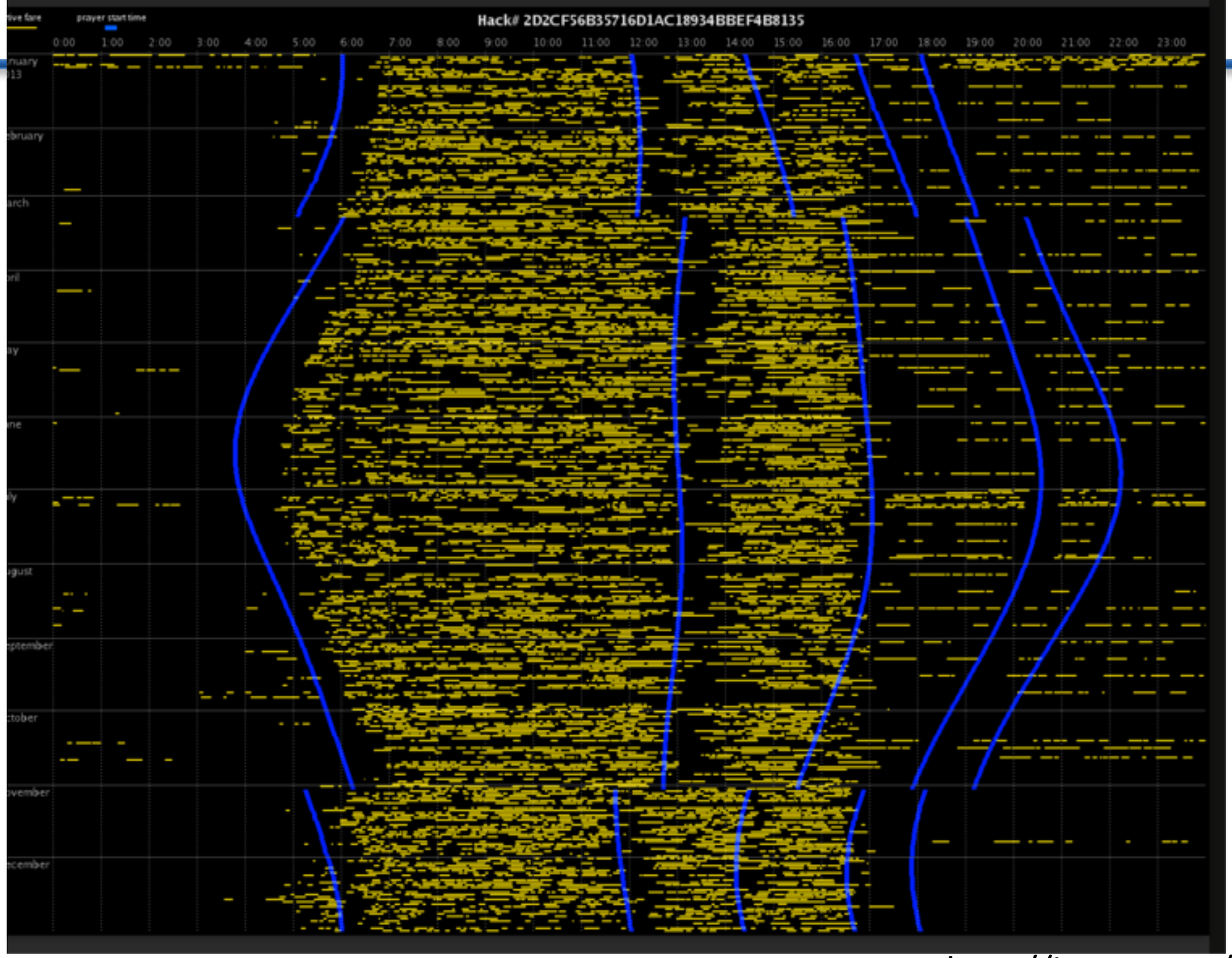
Taxi more expensive

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

Identifying Muslim cabbies from trip data and prayer times [OC]

(source) · 10 months ago

Next Post >



http://imgur.com/a/GzduB

The taxi medallion numbers were not properly de-identified.

Medallions were encoded with MD5.

- e.g. MD5(“5C27”) = “0f76c35d4a069e0fe76b21d28f009639”

The MD5 algorithm can't be reversed, but "data intruders" can do a "brute force search" on all possible values

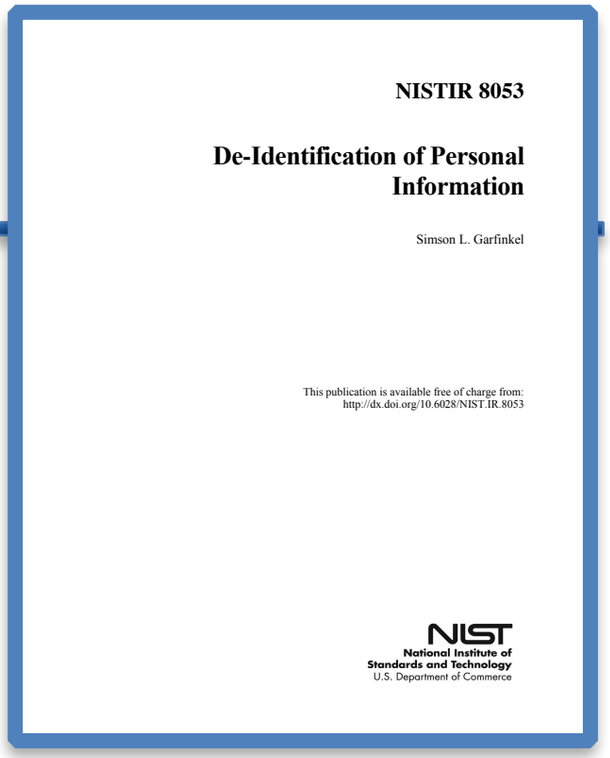
- MD5("5C26") = 244bec24d51fb45fc0f3dadd28fd3cf1
- MD5("5C27") = **0f76c35d4a069e0fe76b21d28f009639**
- MD5("5C28") = be9f314926dd314b36496d926e42f4db

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

"5C27"

A journalist at Gawker identified 9 other cab rides.

NIST Interagency Report 8053: "De-identification of personal 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

Report is based on:

- Review of literature and interviews.
- Public comments received April 6 - May 15, 2015
- Currently in review

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

Covers:

- Why de-identify?
- De-identification terminology
- 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*
- Famous re-identification attacks

The Taxi database was not properly de-identified.

- The encoded "medallion" numbers could be reversed.

	A	B
1	medallion	hack_license
2	89D227B655E5C82AECF13C3F	BA96DE419E711691B944
3	0BD7C8F5BA12B88E0B67BED	9FD8F69F0804BDB5549F
4	0BD7C8F5BA12B88E0B67BED	9FD8F69F0804BDB5549F
5	DFD2202EE08F7A8DC9A57B0	51EE87E3205C985EF843
6	DFD2202EE08F7A8DC9A57B0	51EE87E3205C985EF843
7	20D9ECB2CA0767CF7A01564	598CCE5B9C1918568DEE
8	496644932DF3932605C22C7	513189AD756FF14FE670
9	0B57B9633A2FECD3D3B1944	CCD4367B417ED6634D9
10	2C0E91FF20A856C891483ED	1DA2F6543A62B8ED934

- De-identification Step #1: Remove Identifiers

Removing the identifiers is not enough!

Massachusetts' Weld Collapses at Commencement
May 19, 1996 | From Associated Press

WALTHAM, Mass. — Massachusetts Gov. William F. Weld collapsed Saturday during commencement at Bentley College, but doctors said they found nothing seriously wrong with him.

The 50-year-old governor had just received an honorary doctorate of law when he fainted.

"He fell headfirst [toward the podium], but they caught him," said Bill Petras, a graduating senior who sat five rows from the stage.

Weld was briefly unconscious but was alert by the time he was lifted onto a stretcher and taken to an ambulance. Moments before fainting, Weld had started shaking as he approached the podium, Petras said.

Massachusetts Group Insurance Commission released records of state employees hospital admissions for research.

Commonwealth of Massachusetts
Group Insurance Commission
Your Benefits Connection

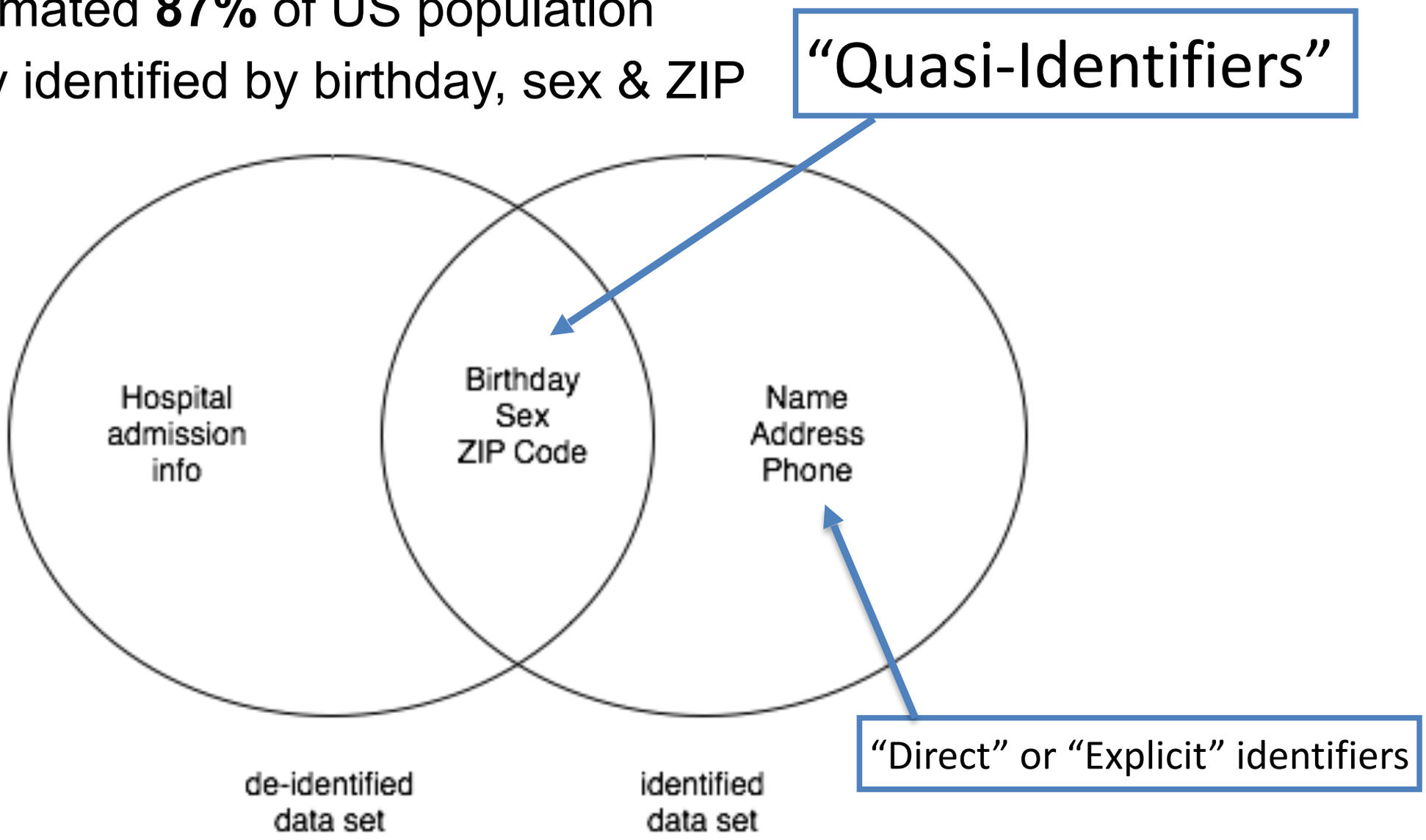
Names and addresses removed to protect privacy.

Latanya 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.

"Linkage Attack"
Identifies 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

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.

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

Black	65	M	37203	M1	Gastric Ulcer
Black	65	M	37205	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.

It may be that *there are only* identifiers and quasi-identifiers.

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

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

test the de-identification

Harm or embarrass the de-identifying organization

Commercial Benefit

Theodore Roosevelt

gain publicity or professional standing

Harm the data subject

“Re-identification risk” is the measure of the risk that the identifiers and other information about individuals can be learned from the de-identified data.

There are 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.

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.

Re-identification can result in specific harms.

Identity disclosure

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

De-identification doesn't help against these disclosures

Attribute disclosure

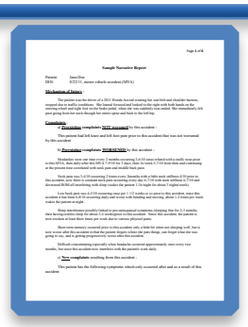
- The dataset shows that all 20-year-old female patients from Q are left-handed.
  - *Jane is a 20-year-old female patient from Q.*
  - *∴Jane is left-handed.*

Inferential disclosure

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

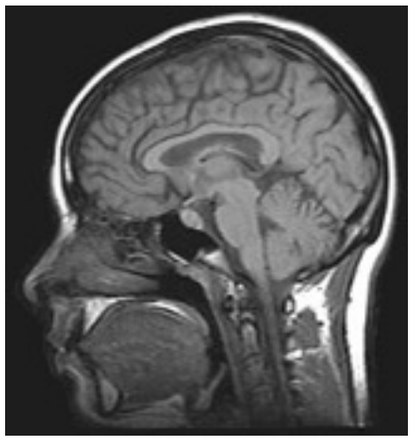
Unstructured data is poses significant de-identification challenges.

— *Free-form medical text*



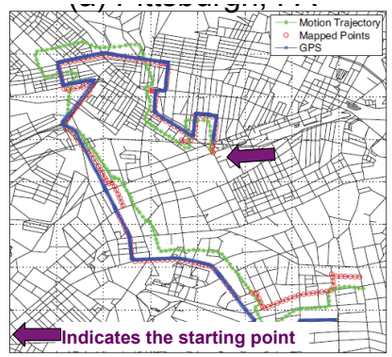
— *Photographs and Video*

— *Medical Imagery*



— *Genetic information*

— *Geographic and map data*



De-identification challenges are similar across modality.

- Only the specific de-identification and re-identification techniques are different.

There is a trade-off between de-identification privacy and utility.

- The more that's removed, the less useful are the data that remain.

In conclusion…

De-identification:

- Can be applied to many different data **modalities**
- Different techniques, similar kinds of risks:
  - *Improper de-identification*
  - *Linkage attacks — Attribute disclosure*
  - *Inferential disclosure*
- Non-structured data and multi-media pose significant challenges

Re-identification:

- Usually involves linking with another data set.
- Can be performed at any time in the future.
- Hard to calculate the re-identification risk, since it depends on future data.