You Are Not Anonymous

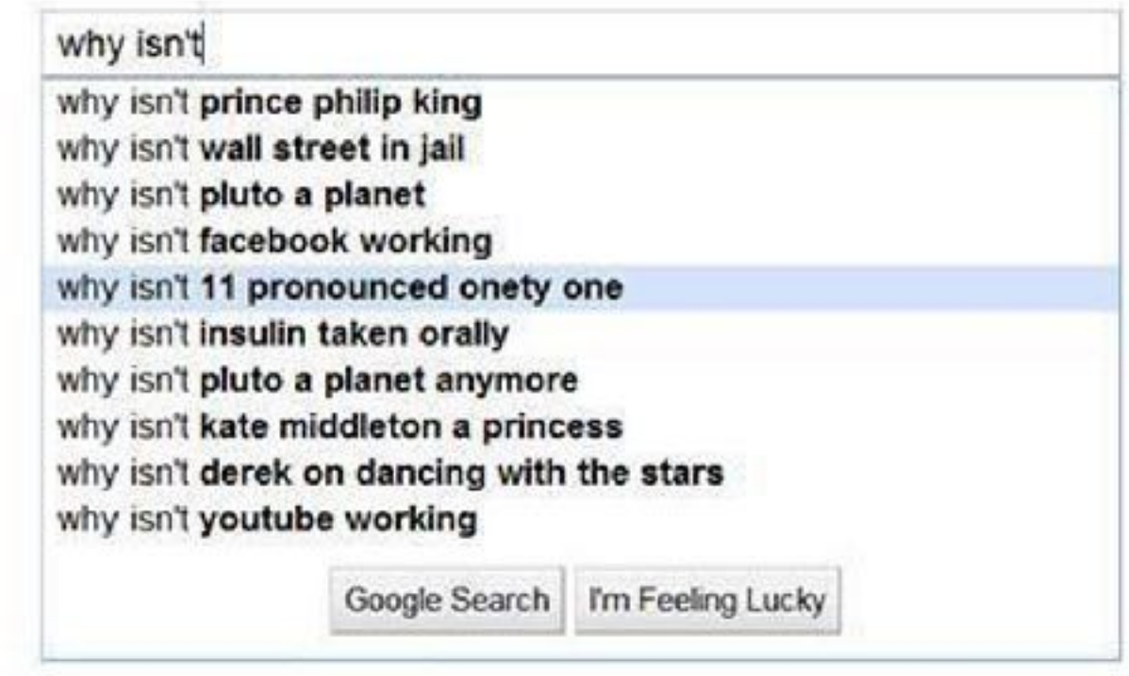
Jessica Su
Santa Clara University
April 13, 2018

Anonymity is important

Anonymity is important

There are many circumstances where you'd like to be anonymous

When searching on Google



Anonymity is important

In 2006, AOL released anonymous search logs from 650,000 users

98280	prayers to break curses	2006-04-09 5
98280	prayers for cleansing	2006-04-09 2
98280	prayers for defeating enemy	2006-04-09 1
98280	bible scriptures for defeating the enemy	2006-04-09 4
98280	prayers to plead the blood of jesus against problems	2006-04-09 2
98280	prayers to plead the blood of jesus against problems	2006-04-09 1
98280	prayers to plead the blood of jesus against problems	2006-04-09 3
98280	how does a male's cocaine use affect a fetus	2006-04-10 1
98280	how does a male's cocaine use affect a fetus	2006-04-10 5
98280	birth defects caused by father's cocaine use	2006-04-10 1
98280	birth defects caused by father's cocaine use	2006-04-10 4
98280	are chainletter scams ever successful	2006-04-10 0

Anonymity is important

In 2006, AOL released anonymous search logs from 650,000 users

A Face Is Exposed for AOL Searcher No. 4417749

By **MICHAEL BARBARO** and **TOM ZELLER Jr.** AUG. 9, 2006

Some users were quickly deanonymized

Anonymity is important

In 2006, AOL released anonymous search logs from 650,000 users

TECH INDUSTRY

AOL apologizes for release of user search data

Search log information originally intended for use on new research site; company calls data posting a mistake.

BY **DAWN KAWAMOTO** / AUGUST 9, 2006 5:38 AM PDT

Anonymity is important

In 2006, AOL released anonymous search logs from 650,000 users

TECHNOLOGY

AOL executive quits after posting of search data - Technology - International Herald Tribune

By TOM ZELLER JR. AUG. 22, 2006

AOL has announced the resignation of its chief technology officer, two weeks after the company came under intense criticism from privacy advocates for releasing hundreds of thousands of its customers' Web search queries.

An AOL researcher who put the queries online and a manager overseeing the project were dismissed, according to an AOL employee

Anonymity is important

In 2006, AOL released anonymous search logs from 650,000 users

CULTURE

AOL sued over Web search data release

BY **ELINOR MILLS** / SEPTEMBER 25, 2006 12:17 PM PDT

Anonymity can be broken

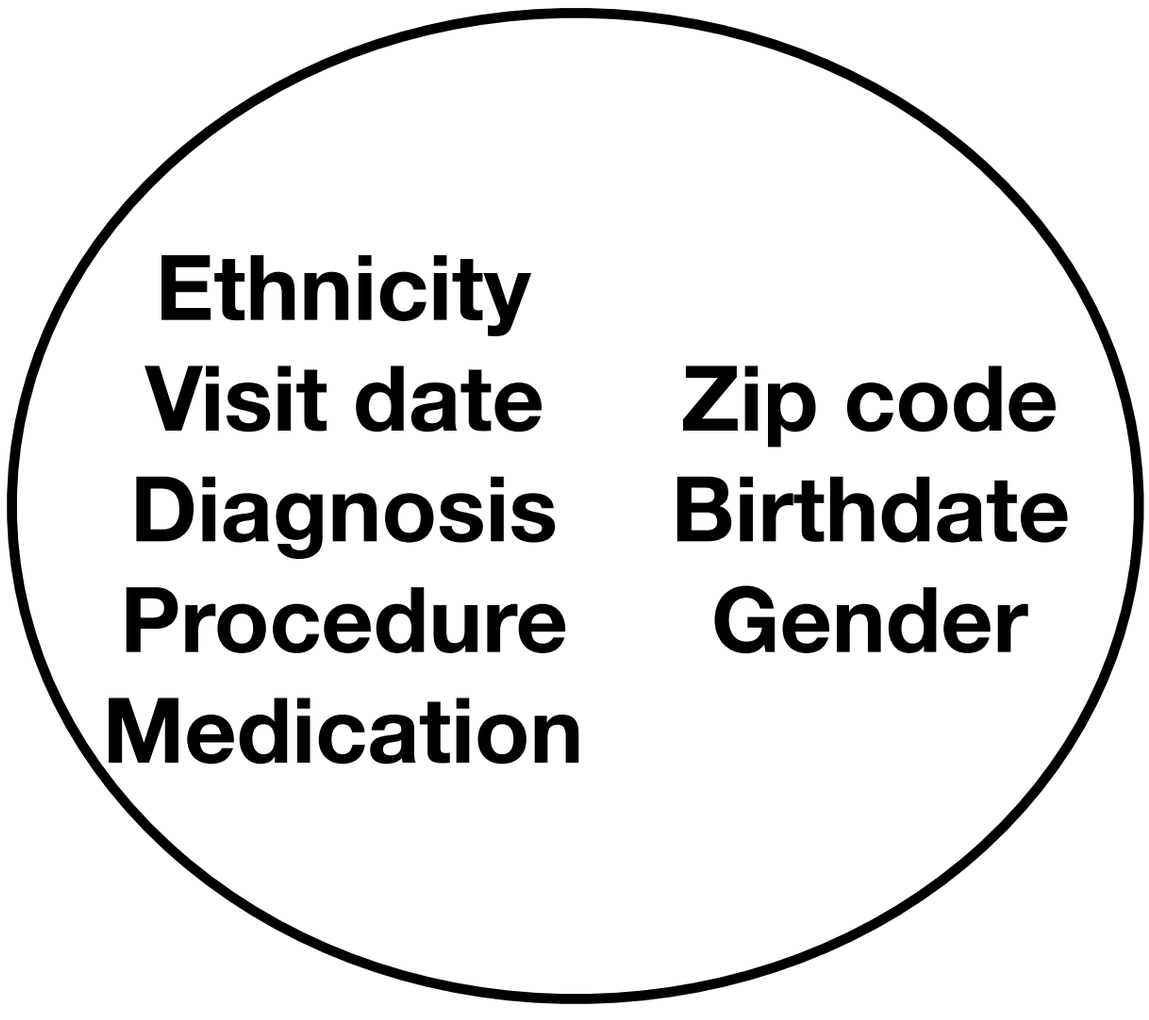
(or "How Latanya Sweeney accessed the medical records of the governor of Massachusetts")

Anonymity can be broken

You can break anonymity by linking anonymous datasets to datasets with PII (personally identifiable information)

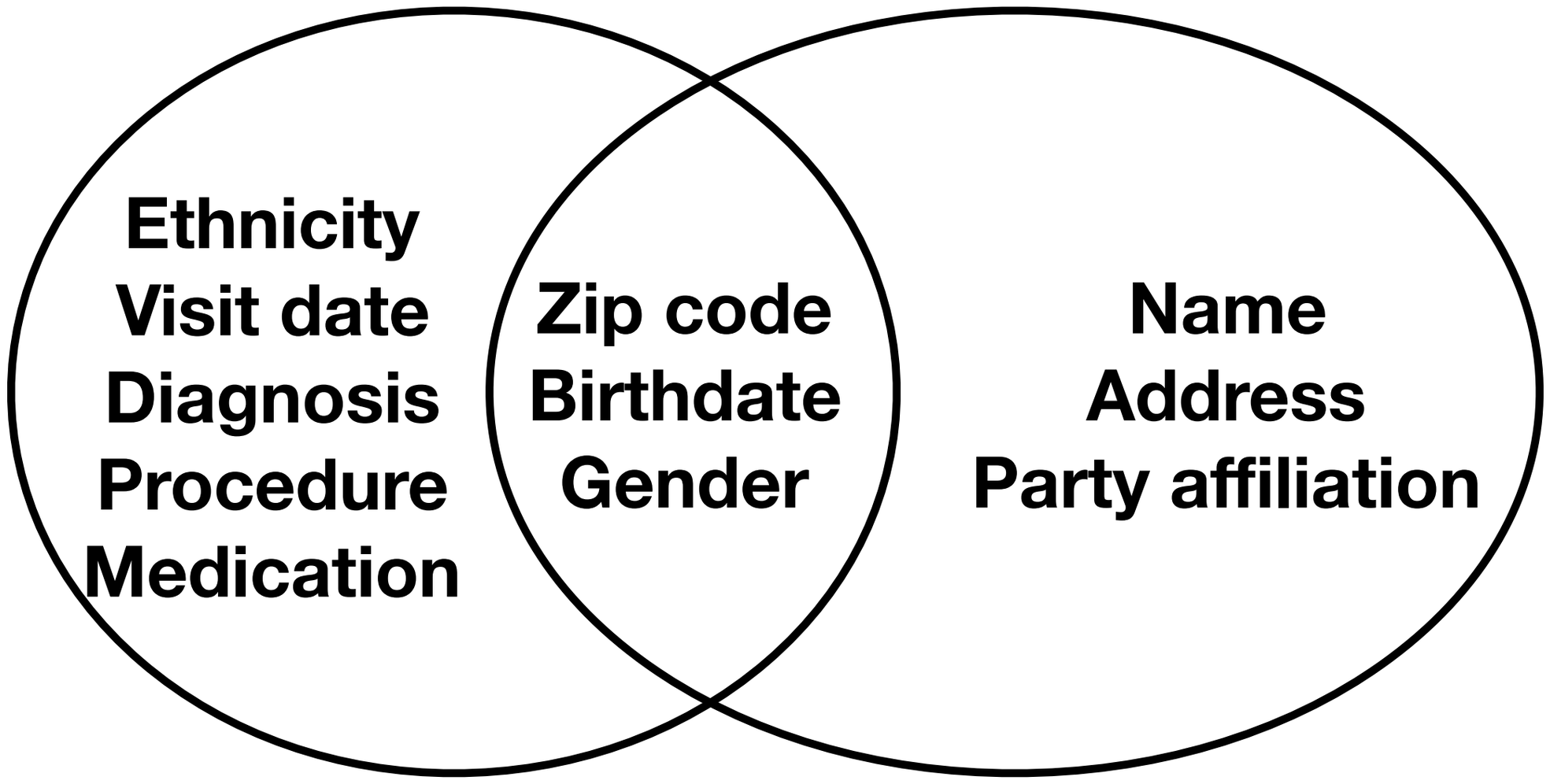
Anonymity can be broken

Latanya Sweeney had anonymous health insurance data



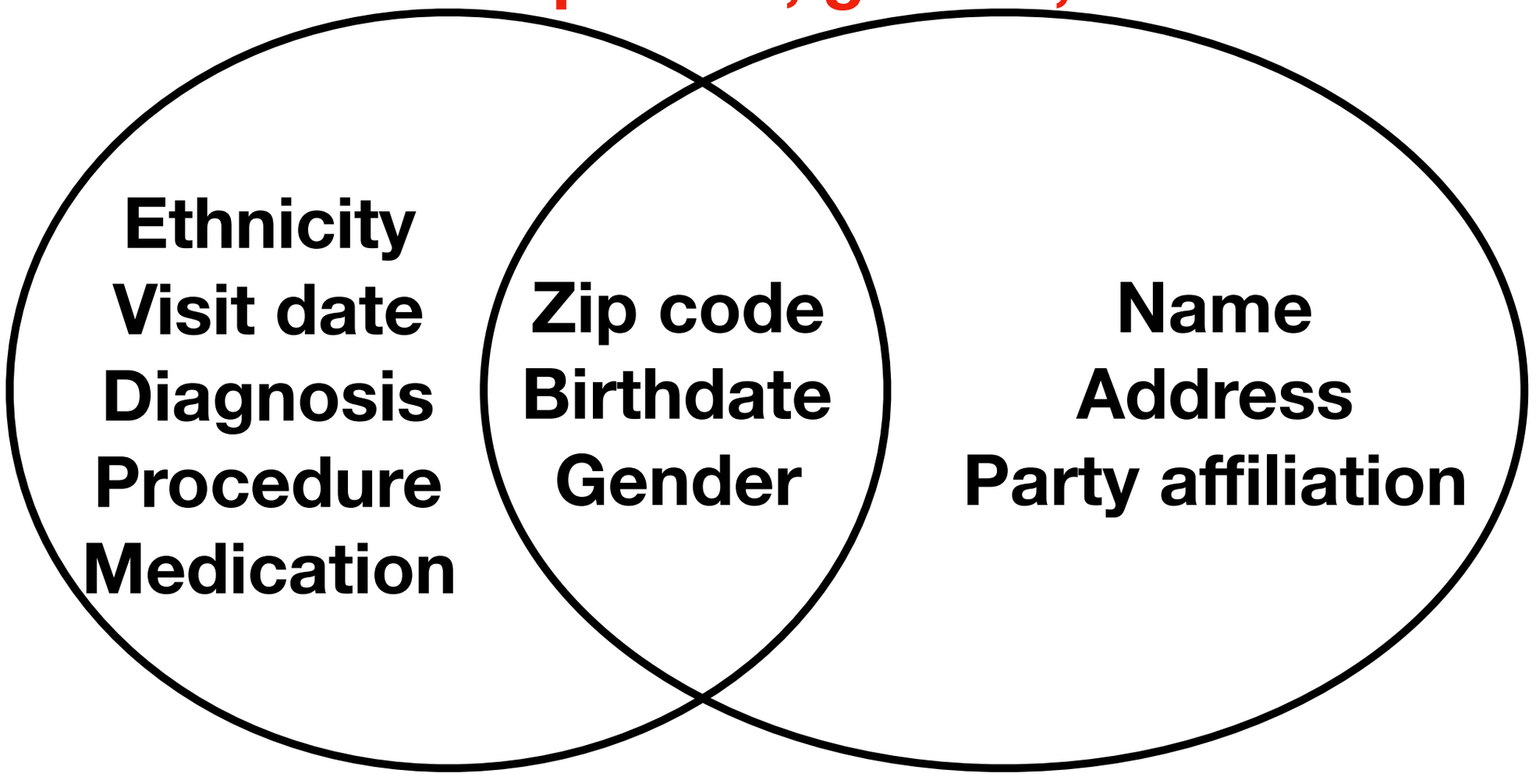
Anonymity can be broken

She bought publicly available voter registration data for $20 that contained some of the same fields



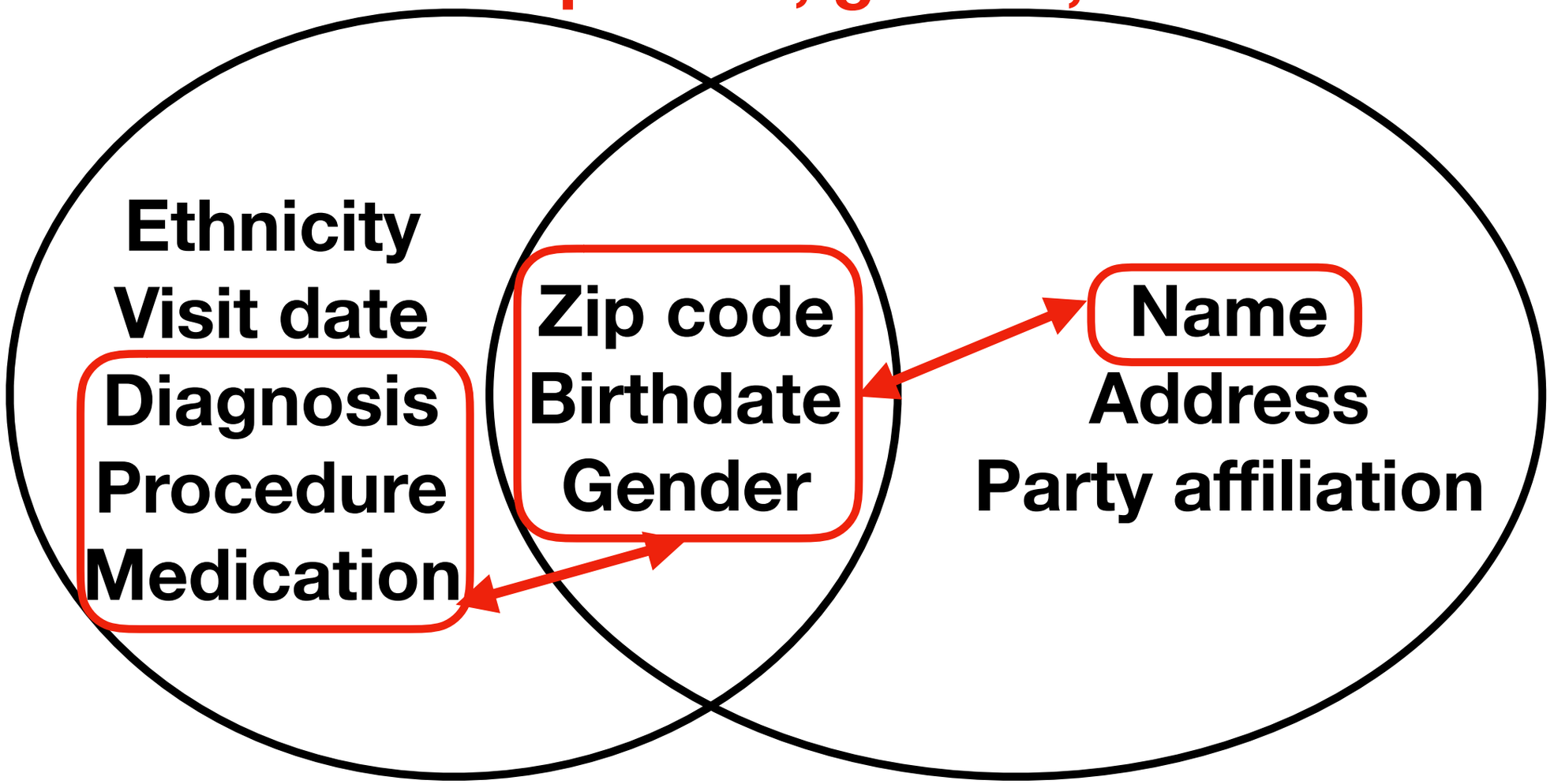
Anonymity can be broken

Fact: 87% of Americans are uniquely identifiable based on their zip code, gender, and birthdate



Anonymity can be broken

Fact: 87% of Americans are uniquely identifiable based on their zip code, gender, and birthdate



This means you can link people's real names to their sensitive medical histories!

Anonymity can be broken

Professor Latanya Sweeney, director of the Data Privacy Lab at Harvard, along with her research assistant and two students scraped data on 1,130 people of the now more than 2,500 who have shared their DNA data for the Personal Genome Project. Church's project posts information about the volunteers on the Internet to help researchers gain new insights about human health and disease. Their names do not appear, but the profiles list medical conditions including abortions, illegal drug use, alcoholism, depression, sexually transmitted diseases, medications and their DNA sequence.

Of the 1,130 volunteers Sweeney and her team reviewed, about 579 provided zip code, date of birth and gender, the three key pieces of information she needs to identify anonymous people combined with information from voter rolls or other public records. Of these, Sweeney succeeded in naming 241, or 42% of the total. The Personal Genome Project confirmed that 97% of the names matched those in its database if nicknames and first name variations were included. She describes her findings here.

Anonymity can be broken

Natural response:

Treat (ZIP, gender, birthdate) tuple as personally identifying information

Make sure each combination of personally identifying attributes appears at least twice

Anonymity can be broken

Natural response:

Treat (ZIP, gender, birthdate) tuple as personally identifying information

Make sure each combination of personally identifying attributes appears at least twice

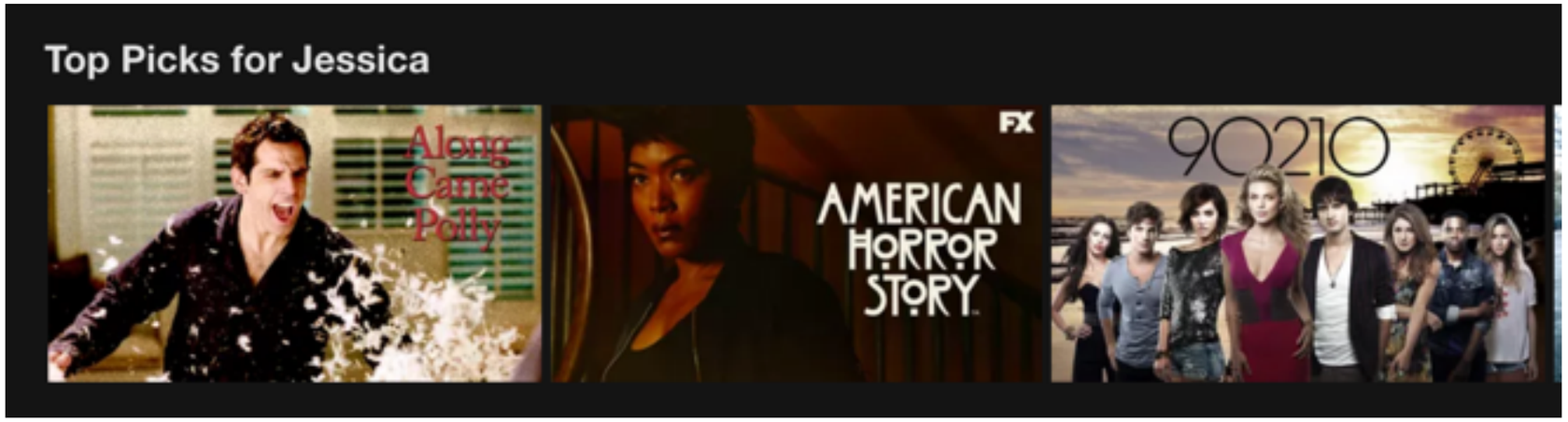
Problem: No clear separation between personally identifying attributes and non-identifying attributes

The Netflix deanonymization study

(or "How to make sensitive inferences from boring data")

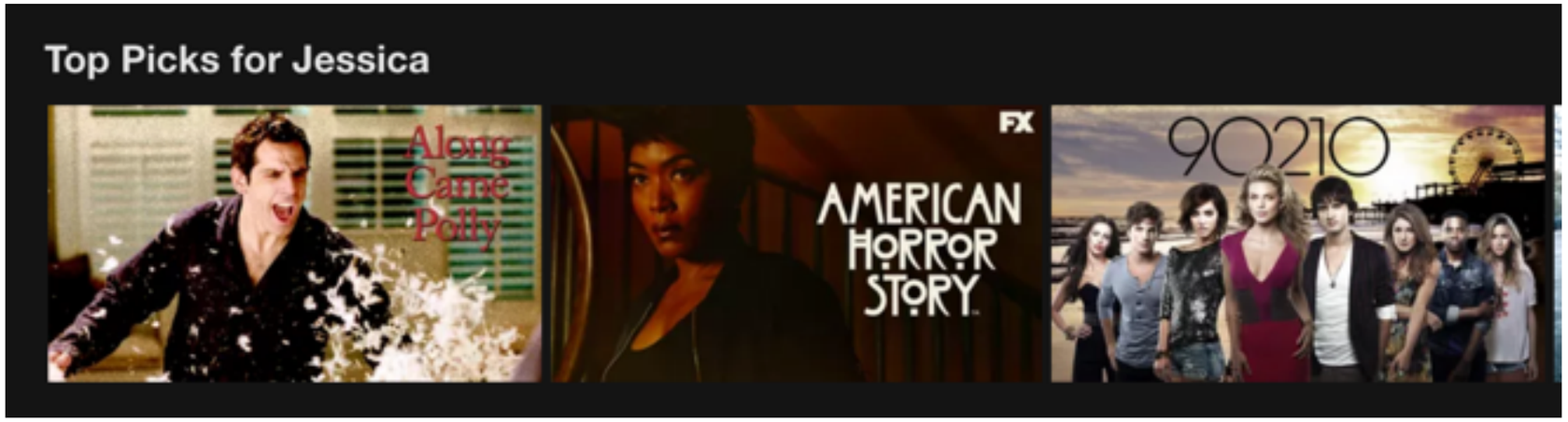
The Netflix challenge

Netflix has a service that recommends movies to people



The Netflix challenge

Netflix has a service that recommends movies to people



One key part of this was their algorithm to predict how users would rate movies

The Netflix challenge

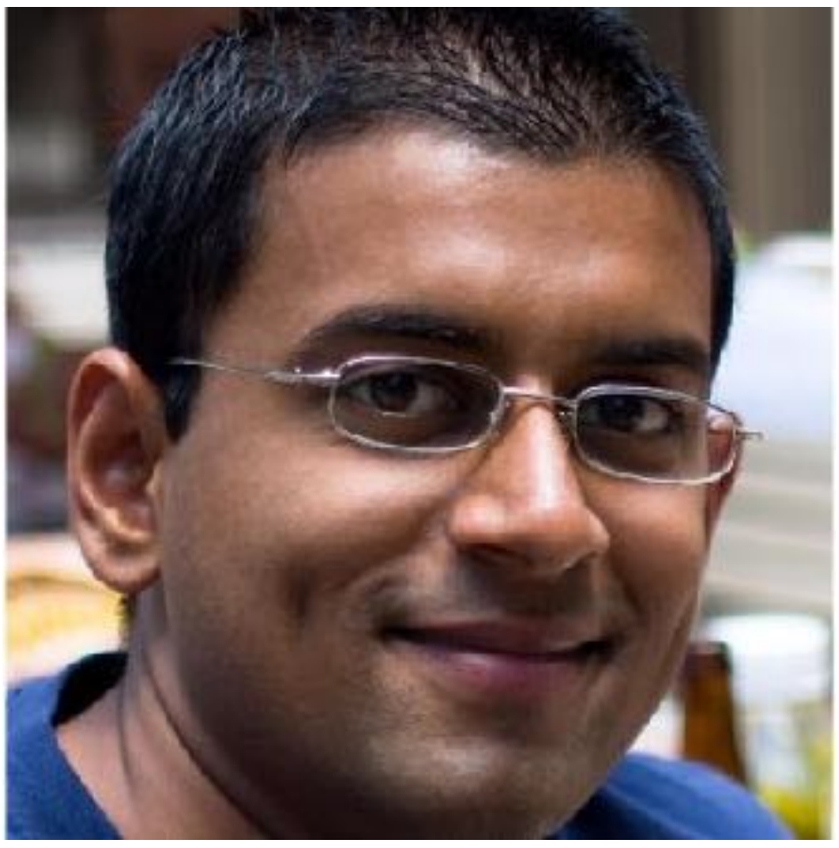
In 2006, Netflix announced a $1 million prize for the first team that could improve their algorithm's performance by 10%

The Netflix challenge

As part of the contest, Netflix released a dataset of anonymous movie ratings

The Netflix challenge

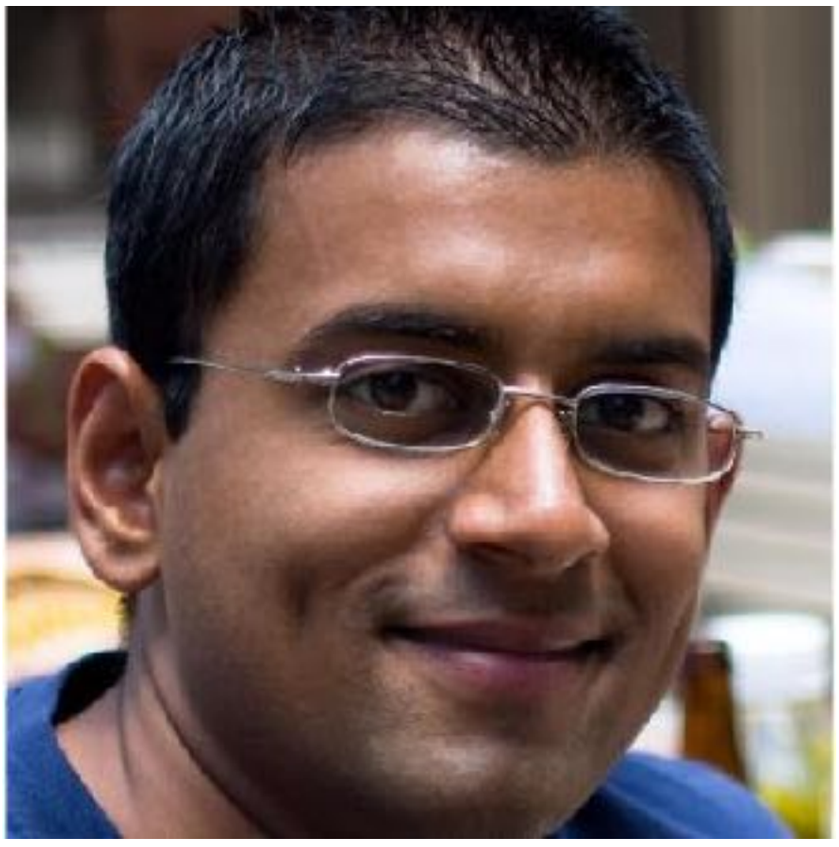
As part of the contest, Netflix released a dataset of anonymous movie ratings



Arvind Narayanan, anonymity expert

The Netflix challenge

As part of the contest, Netflix released a dataset of anonymous movie ratings



Arvind Narayanan, anonymity expert

"Can we figure out who these ratings belong to?"

Two key ingredients of deanonymization

Two key ingredients of deanonymization

1) Users must have distinctive, uniquely identifying attributes

Two key ingredients of deanonymization

1) Users must have distinctive, uniquely identifying attributes

(e.g. ZIP code, gender, birthdate)

Two key ingredients of deanonymization

1) Users must have distinctive, uniquely identifying attributes

(e.g. ZIP code, gender, birthdate)

2) Those attributes must also appear in a less anonymous dataset

Two key ingredients of deanonymization

1) Users must have distinctive, uniquely identifying attributes

99% of users are uniquely identifiable if you know a randomly selected subset of 8 of their movie ratings

Two key ingredients of deanonymization

1) Users must have distinctive, uniquely identifying attributes

99% of users are uniquely identifiable if you know a randomly selected subset of 8 of their movie ratings

2) Those attributes must also appear in a less anonymous dataset

Two key ingredients of deanonymization

1) Users must have distinctive, uniquely identifying attributes

99% of users are uniquely identifiable if you know a randomly selected subset of 8 of their movie ratings

2) Those attributes must also appear in a less anonymous dataset

Ratings on the Internet Movie Database are attached to people's online identities

Deanonymization results

Two Netflix users were linked to their IMDB profiles

Deanonymization results

Two Netflix users were linked to their IMDB profiles

Movies viewed included
"Jesus of Nazareth"
"Power and Terror: Noam Chomsky in Our Times"
"Fahrenheit 9/11"

How did they do it?

Suppose Mary is an IMDB user.

Naive approach: search for a Netflix user who has rated all of the movies that Mary reviewed.

How did they do it?

Suppose Mary is an IMDB user.

Naive approach: search for a Netflix user who has rated all of the movies that Mary reviewed.

Problem: there is a lot of noise in the dataset, and IMDB and Netflix records do not perfectly correspond.

How did they do it?

Instead, use a scoring function that softly penalizes a Netflix user for deviating from Mary's IMDB ratings

How did they do it?

Instead, use a scoring function that softly penalizes a Netflix user for deviating from Mary's IMDB ratings

If the highest score is much higher than the second-highest score, return the highest score

Otherwise, there is no match

What have we learned?

We can't divide the data into "public" and "sensitive" attributes

All movie ratings are sensitive when combined with other movie ratings

What have we learned?

We can't divide the data into "public" and "sensitive" attributes

All movie ratings are sensitive when combined with other movie ratings

This is a general problem with sparse, high-dimensional data

What have we learned?

Anonymous data is not safe to release

RYAN SINGEL SECURITY 12.17.09 04:29 PM

NETFLIX SPILLED YOUR BROKEBACK MOUNTAIN SECRET, LAWSUIT CLAIMS

Identifying authors on the Internet

People make "anonymous" posts on the Internet

People make "anonymous" posts on the Internet

Can we figure out who they are based on their writing style?

Experiment design

Compare the anonymous posts to posts that are written under people's names

Experiment design

Compare the anonymous posts to posts that are written under people's names

Koppel et al:

Experiment design

Compare the anonymous posts to posts that are written under people's names

Koppel et al:

10000 blogs from blogger.com

Experiment design

Compare the anonymous posts to posts that are written under people's names

Koppel et al:

- **10000 blogs from blogger.com**
- **Divide each blog into 2000 words of "known text" and a 500-word anonymous "snippet"**

Experiment design

Compare the anonymous posts to posts that are written under people's names

Koppel et al:

- **10000 blogs from blogger.com**
- **Divide each blog into 2000 words of "known text" and a 500-word anonymous "snippet"**
- **Match the anonymous snippets to the authors of the known texts**

Feature selection

Each post is represented by a vector of numerical features

Feature selection

Each post is represented by a vector of numerical features

Question: What are some examples of good features?

Feature selection

Each post is represented by a vector of numerical features

Question: What are some examples of good features?

Koppel used the numbers of "space-free character 4-grams"

Space-free character 4-grams

Example: Always buy rugs not drugs

Space-free character 4-grams

Example: Always buy rugs not drugs

Space-free character 4-grams:

Alwa	buy	drug
lway	rugs	rugs
ways	not	

Space-free character 4-grams

Example: Always buy rugs not drugs

Space-free character 4-grams:

Alwa	buy	drug
lway	rugs	rugs
ways	not	

Feature vector:

[1	0	1	1	0	1	2	1]
Alwa	bugs	drug	lway	mugs	not	rugs	ways

How to deanonymize

Cosine similarity measures how similar two vectors are

$$similarity(u, v) = \frac{u \cdot v}{||u||\ ||v||}$$

dot product

magnitude of vector

How to deanonymize

Cosine similarity measures how similar two vectors are

$$similarity(u, v) = \frac{u \cdot v}{||u||\ ||v||}$$

dot product

magnitude of vector

Find the cosine similarities between the feature vector of the anonymous document and the feature vectors of the named documents

Pick the author who wrote the document with the highest cosine similarity

Improvements

46% of the snippets were correctly assigned

Improvements

46% of the snippets were correctly assigned

To improve this, run the deanonymization algorithm using only a randomly sampled subset of the features

Improvements

46% of the snippets were correctly assigned

To improve this, run the deanonymization algorithm using only a randomly sampled subset of the features

Do this on many different subsets

Improvements

46% of the snippets were correctly assigned

To improve this, run the deanonymization algorithm using only a randomly sampled subset of the features

Do this on many different subsets

If enough of the subsets agree that the snippet was written by someone, return that person

Improvements

46% of the snippets were correctly assigned

To improve this, run the deanonymization algorithm using only a randomly sampled subset of the features

Do this on many different subsets

If enough of the subsets agree that the snippet was written by someone, return that person

Idea: It's harder to be #1 across many subsets of the feature set than it is to be #1 on the full feature set

Another idea

Narayanan et al: Can we train a machine learning classifier to predict which author wrote a document?

Another idea

Narayanan et al: Can we train a machine learning classifier to predict which author wrote a document?

Category	Description	Count
Length	number of words/characters in post	2
Vocabulary richness	Yule's K[3] and frequency of *hapax legomena*, *dis legomena*, etc.	11
Word shape	frequency of words with different combinations of upper and lower case letters.[4]	5
Word length	frequency of words that have 1–20 characters	20
Letters	frequency of *a* to *z*, ignoring case	26
Digits	frequency of 0 to 9	10
Punctuation	frequency of `. ? ! , ; : ( ) " - '`	11
Special characters	frequency of other special characters `` ` ~ @ # $ % ^ & * _ + = [ ] { } \ | / < > ``	21
Function words	frequency of words like 'the', 'of', and 'then'	293
Syntactic category pairs	frequency of every pair (A, B), where A is the parent of B in the parse tree	789

Table I
THE FEATURES USED FOR CLASSIFICATION. MOST TAKE THE FORM OF FREQUENCIES, AND ALL ARE REAL-VALUED.

In conclusion

There are a whole bunch of these studies

In conclusion

There are a whole bunch of these studies

- **Can deanonymize location data, credit card data, web browsing history data, etc.**

In conclusion

There are a whole bunch of these studies

- **Can deanonymize location data, credit card data, web browsing history data, etc.**

Lesson: be careful when releasing "anonymous" data

In conclusion

There are a whole bunch of these studies

- **Can deanonymize location data, credit card data, web browsing history data, etc.**

Lesson: be careful when releasing "anonymous" data

- **Often you can link it back to people's real identities**

Thanks for listening

- Latanya Sweeney. k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05):557–570, 2002.
- Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In 2008 IEEE Symposium on Security and Privacy (sp 2008), pages 111–125. IEEE, 2008.
- Moshe Koppel, Jonathan Schler, and Shlomo Argamon. "Authorship attribution in the wild." Language Resources and Evaluation 45.1 (2011): 83-94.
- Arvind Narayanan, Hristo Paskov, Neil Zhenqiang Gong, John Bethencourt, Emil Stefanov, Eui Chul Richard Shin, and Dawn Song. On the feasibility of internet-scale author identification. In IEEE Symposium on Security and Privacy, 2012.