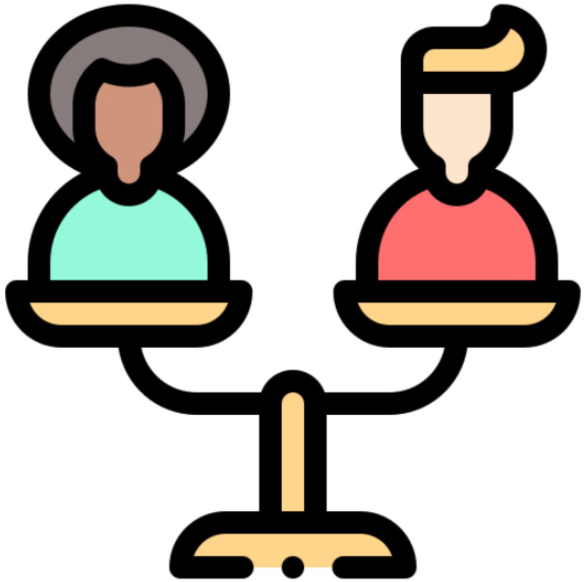
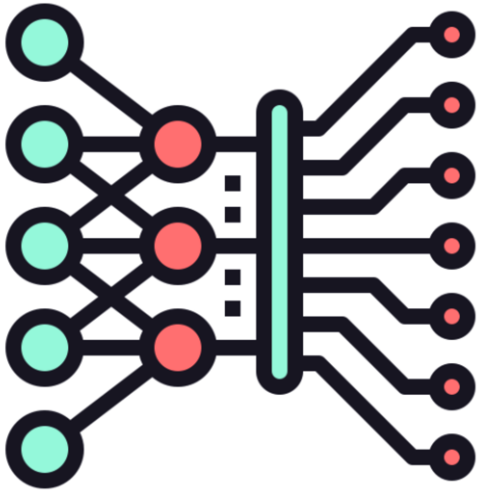
The Algorithmic Foundations of Ethical Machine Learning





About me: Juba Ziani

- Joined GT ISyE this August
- Education:
  - Postdoc ’21 (Upenn), Computer Science
  - PhD ‘19 (Caltech), Computer Science/OR
  - MSc ‘13 (Columbia), IEOR
  - MSc ‘12 (Supélec), Information Sciences
- Research interests:
  - Game theory and mechanism design
  - Data and online markets
  - Differential privacy and fairness
  - Machine learning

Logistics

- Instructor: Prof. Juba Ziani (me)
- Times: Tues/Thurs, 9:30-10:45am
- Location: Groseclose, Room 119 (here)
- Email: jziani3@gatech.edu

Workload and grading

- 3 problem sets: 15% each (total 45%)
- Paper reading, class presentations, and written summary:
  - Privacy: 5% presentation, 10% summary
  - Fairness: 5% presentation, 10% summary
- Research project: 5% proposal, 5% presentation, 10% write-up (total 20%)
- Class participation: 5%
- No exam

Workload and grading

- 3 problem sets: 15% each (total 45%)
- Paper reading, class presentations, and written summary:
  - Privacy: 5% presentation, 10% summary
  - Fairness: 5% presentation, 10% summary
- Research project: 5% proposal, 5% presentation, 10% write-up (total 20%)
- Class participation: 5%
- No exam

Collaboration policy:

- Problem sets: can discuss, but write solutions separately.
- Paper reading/written summary/presentation: alone or group of 2.
- Project: working in groups of ≤ 3 is encouraged.

Workload and grading

- 3 problem sets: 15% each (total 45%)
- Paper reading, class presentations, and written summary:
  - Privacy: 5% presentation, 10% summary
  - Fairness: 5% presentation, 10% summary
- Research project: 5% proposal, 5% presentation, 10% write-up (total 20%)
- Class participation: 5%
- No exam

Late policy:

- 3 tokens throughout the course
- Each token = 24h extension, no question asked
- Major emergencies: email me

About covid

What are you comfortable with?

A few caveats:

- I have to teach in-person, I cannot teach on zoom. But, will make lecture notes available + will be happy to set up online office hours to answer Q's
- I cannot force you to wear a mask or ask about your vaccination status; however I **urge** you to do what you can to protect your fellow classmates.

Office hours

Wednesdays 3:00 – 4:00pm

3:00 – 5:00pm on weeks problem sets are due

Office hours

Wednesdays 3:00 – 4:00pm

3:00 – 5:00pm on weeks problem sets are due

Uncomfortable with in-person interactions

➔ email me at jziani3@gatech.edu to set an online meeting

Other class policies

Academic honor code:

- Georgia Tech's Academic Honor Code here: http://osi.gatech.edu/content/honor-code

Office of Disability Services:

- Georgia Tech has policies regarding disability accommodations (http://disabilityservices.gatech.edu/).
- If you require special accommodations, please notify me ASAP

Focus and goals of this course

Focus:

- Privacy and fairness issues that arise in ML
- 1-2 lecture of motivation and context for each
- Mostly algorithmic & technical tools to analyze and understand these issues: *differential privacy, algorithmic fairness*

Focus and goals of this course

Focus:

- Privacy and fairness issues that arise in ML
- 1-2 lecture of motivation and context for each
- Mostly algorithmic & technical tools to analyze and understand these issues: *differential privacy, algorithmic fairness*

Main objectives:

1. Understanding the motivation behind privacy and fairness
2. Understanding how technical tools can help address these issues
3. Acquiring the basic toolkit and understanding of research areas to perform research in privacy and/or fairness

Focus and goals of this course

This is a mathematically and technically oriented class.

Pre-requisites:

- Probability
- Algorithms
- Basic understanding of ML: regression and classification
- Proof-based math (problem sets will be proof-based)

Topics covered

Part I: Differential Privacy (DP)

1. Why differential privacy? Previous privacy failures, how DP addresses them.

2. Formal definitions and properties of differential privacy.

3. Algorithms and mechanisms for differential privacy + formal guarantees.

4. Applications and advanced privacy techniques.

Topics covered

Part II: Fairness in ML

1. Why fair ML? What happens when fairness not directly taken into account?

2. Formal definitions of algorithmic fairness.

3. Overview of research/techniques in fairness in ML.

4. Applications.

Course material

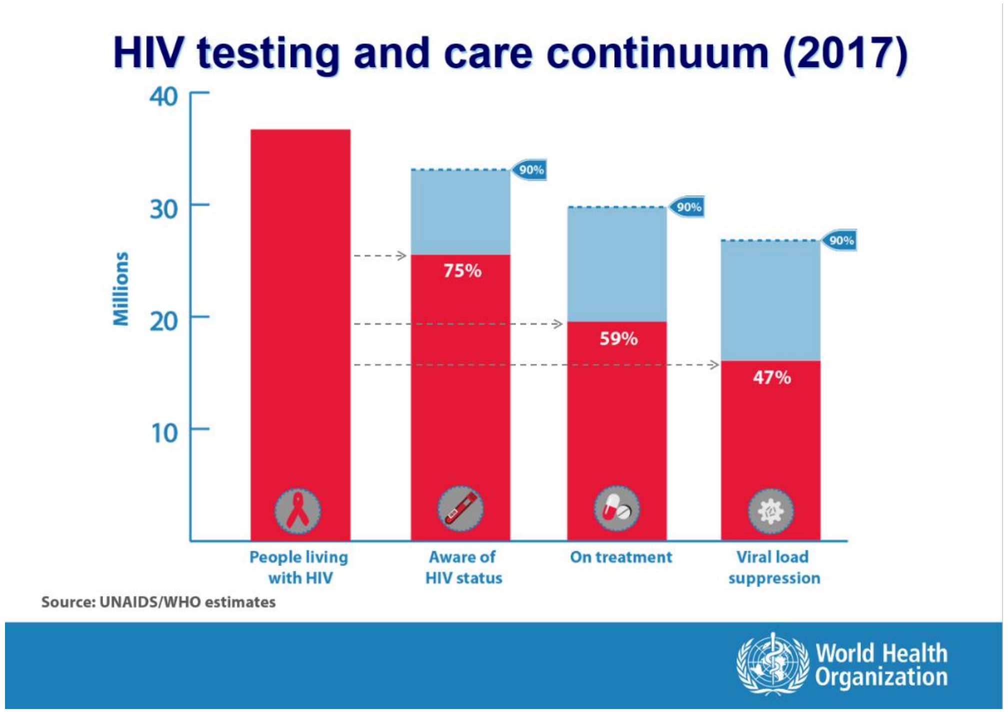
Books:

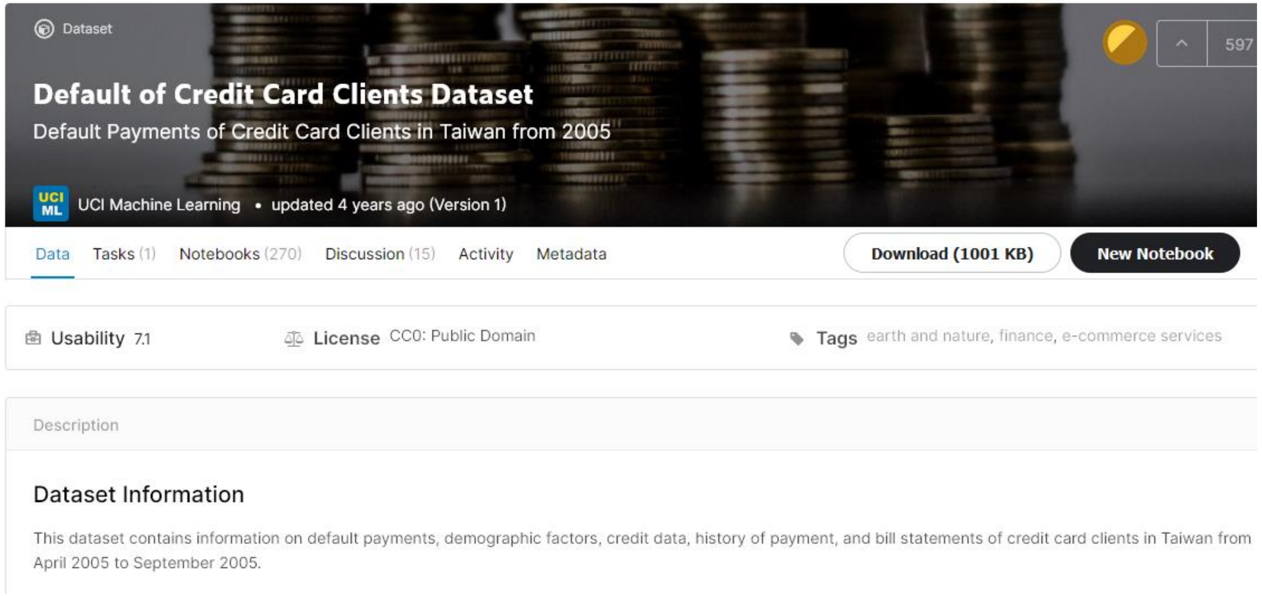
- “The Algorithmic Foundations of Differential Privacy” by Cynthia Dwork and Aaron Roth: https://www.cis.upenn.edu/~aaroth/privacybook.html
- “Fairness and Machine Learning: Limitations and Opportunities” by S. Barocas, M, Hardt, A. Naranayan: https://fairmlbook.org/
- “The Ethical Algorithm” by Michael Kearns and Aaron Roth (Optional)
- Research papers, references provided throughout the course

(Differential) privacy

Why is privacy important?

Why is privacy important?





Data privacy is *hard*

Many privacy failures over the past 15 years

Why so many failures?

Common approaches: not properly formalized, or too ad-hoc:

1. Naïve definitions of privacy. Intuitive $\neq$ good. Ex: anonymization
2. Trying to anticipate specific attacks, and prevent those

Data privacy is *hard*

Many privacy failures over the past 15 years

Why so many failures?

Common approaches: not properly formalized, or too ad-hoc:

1. Naïve definitions of privacy. Intuitive $\neq$ good. Ex: anonymization
2. Trying to anticipate specific attacks, and prevent those

Problem:

1. Not thinking carefully about what is a good definition/what it protects against
2. No protection against reconstruction attacks that have not been predicted/anticipated

Failures of data privacy: anonymization

What is data anonymization?

Name	DOB	Gender	State/zip code	Has cancer?
Juba Ziani	Come on guys	Male	GA 30309	No
Marge Simpson	04/19/1987	Female	SP 75234	No
Rick Sanchez	01/15/1943	Male	WA 98101	Yes
Misty	04/01/1983	Female	KT 16983	No

Failures of data privacy: anonymization

What is data anonymization?

Name	DOB	Gender	State/zip code	Has cancer?
~~Juba Ziani~~	Come on guys	Male	GA 30309	No
~~Marge Simpson~~	04/19/1987	Female	SP 75234	No
~~Rick Sanchez~~	01/15/1943	Male	WA 98101	Yes
~~Misty~~	04/01/1983	Female	KT 16983	No

Failures of data privacy: anonymization

What is data anonymization?

Name	DOB	Gender	State/zip code	Has cancer?
1das4fg5d5as2	Come on guys	Male	GA 30309	No
345fa4f331t43	04/19/1987	Female	SP 75234	No
254jrtul42f4sf1	01/15/1943	Male	WA 98101	Yes
175dsa4f6jz68d	04/01/1983	Female	KT 16983	No

Failures of data privacy: anonymization

So what's the problem?

"Simple Demographics Often Identify People Uniquely"; L. Sweeney 2000

- A few attributes are enough to uniquely identify most of the US population
- (Zip, gender, date of birth) → identifies **87%** of US population

Failures of data privacy: anonymization

So what's the problem?

"Simple Demographics Often Identify People Uniquely"; L. Sweeney 2000

- A few attributes are enough to uniquely identify most of the US population
- (Zip, gender, date of birth) → identifies **87%** of US population

Name	DOB	Gender	State/zip code	
1das4fg5d5as2	Come on guys	Male	GA 30309	
345fa4f331t43	04/19/1987	Female	SP 75234	
254jrtul42f4sf1	**01/15/1943**	**Male**	**WA 98101**	
175dsa4f6jz68d	04/01/1983	Female	KT 16983	

Failures of data privacy: anonymization

So what's the problem?

"Simple Demographics Often Identify People Uniquely"; L. Sweeney 2000

- A few attributes are enough to uniquely identify most of the US population
- (Zip, gender, date of birth) → identifies **87%** of US population

Name	DOB	Gender	State/zip code	
1das4fg5d5as2	Come on guys	Male	GA 30309	
345fa4f331t43	04/19/1987	Female	SP 75234	
Rick Sanchez	**01/15/1943**	**Male**	**WA 98101**	
175dsa4f6jz68d	04/01/1983	Female	KT 16983	

Failures of data privacy: anonymization

So what's the problem?

"Simple Demographics Often Identify People Uniquely"; L. Sweeney 2000

- A few attributes are enough to uniquely identify most of the US population
- (Zip, gender, date of birth) → identifies **87%** of US population

Name	DOB	Gender	State/zip code	Has cancer?
1das4fg5d5as2	Come on guys	Male	GA 30309	No
345fa4f331t43	04/19/1987	Female	SP 75234	No
Rick Sanchez	**01/15/1943**	**Male**	**WA 98101**	**Yes**
175dsa4f6jz68d	04/01/1983	Female	KT 16983	No

Failures of data privacy: anonymization

Very easy attack:

- Sweeney spend only $20 for public DOB/gender/zip codes info in Cambridge. Bought voter rolls.

Failures of data privacy: anonymization

Very easy attack:

- Sweeney spend only $20 for public DOB/gender/zip codes info in Cambridge. Bought voter rolls.
- Same birthday as the governor of Mass: 6 people in Cambridge

Failures of data privacy: anonymization

Very easy attack:

- Sweeney spend only \$20 for public DOB/gender/zip codes info in Cambridge. Bought voter rolls.
- Same birthday as the governor of Mass: 6 people in Cambridge
- Only 3 were male

Failures of data privacy: anonymization

Very easy attack:

- Sweeney spend only $20 for public DOB/gender/zip codes info in Cambridge. Bought voter rolls.
- Same birthday as the governor of Mass: 6 people in Cambridge
- Only 3 were male
- Only 1 had the right zip code

Failures of data privacy: anonymization

Very easy attack:

- Sweeney spend only $20 for public DOB/gender/zip codes info in Cambridge. Bought voter rolls.
- Same birthday as the governor of Mass: 6 people in Cambridge
- Only 3 were male
- Only 1 had the right zip code

➔ Sweeney was able to ***uniquely identify the governor's medical records***! Sent them to his office.

Failures of data privacy: anonymization

Very easy attack:

- Sweeney spend only $20 for public DOB/gender/zip codes info in Cambridge. Bought voter rolls.
- Same birthday as the governor of Mass: 6 people in Cambridge
- Only 3 were male
- Only 1 had the right zip code

➔ Sweeney was able to ***uniquely identify the governor's medical records***! Sent them to his office.

Solution: hide identifying attributes? Ad-hoc and risky.

Location data

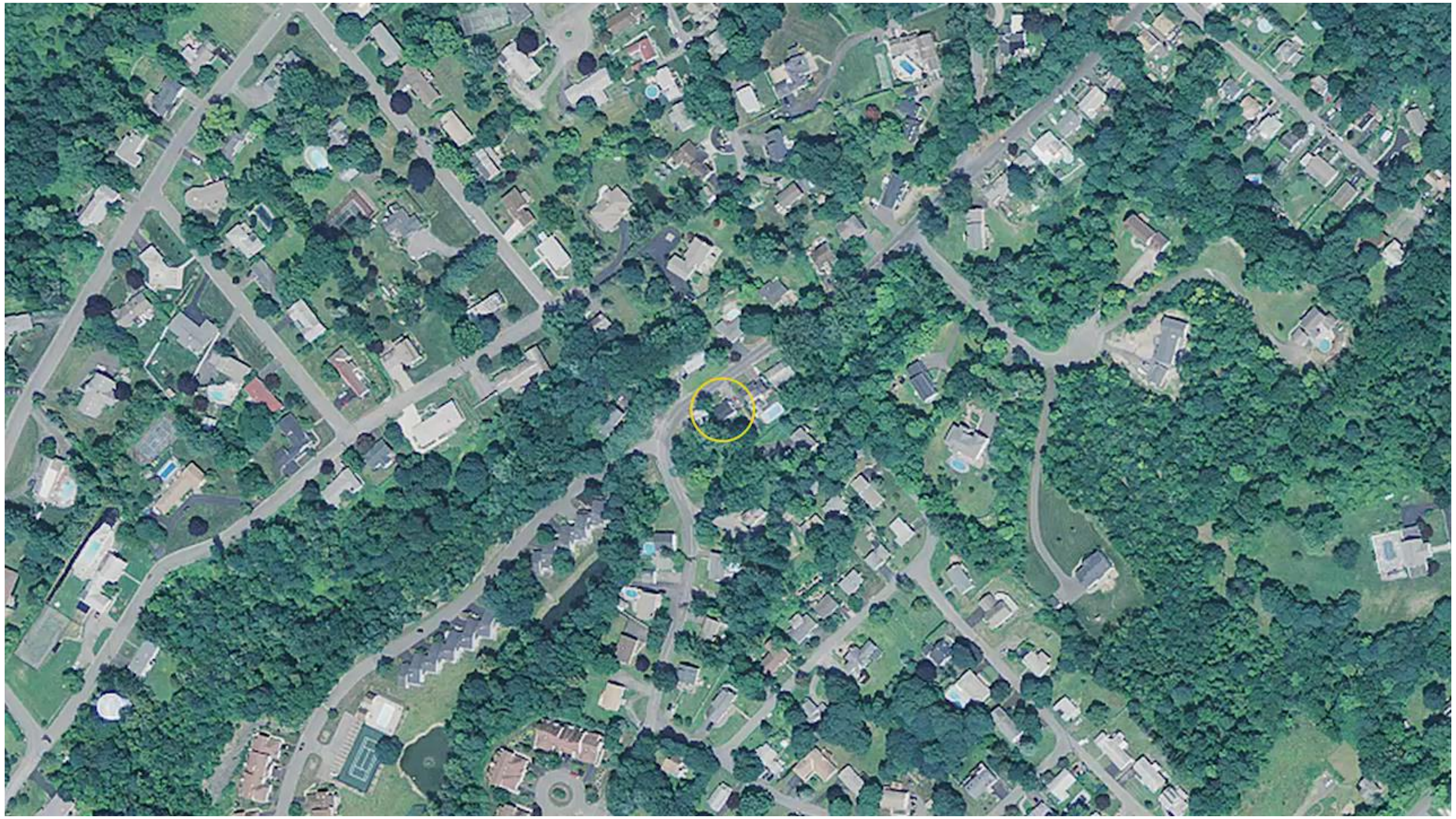
Location data can be used to breach your privacy:

- Your phone/apps can track your location data
- Often, this location data is anonymized/every agent in the database as a randomized ID then used or re-sold to other businesses
- But location data can reveal your identity easily…

Example: New York Times' study

- Was able to obtain one company's database
- > 1 million phones in the NY area, anonymized IDs

Location data



“(…) leaves a house in upstate New York at 7 a.m. and travels to a middle school 14 miles away, staying until late afternoon each school day. **Only one person makes that trip: Lisa Magrin, a 46-year-old math teacher**.”

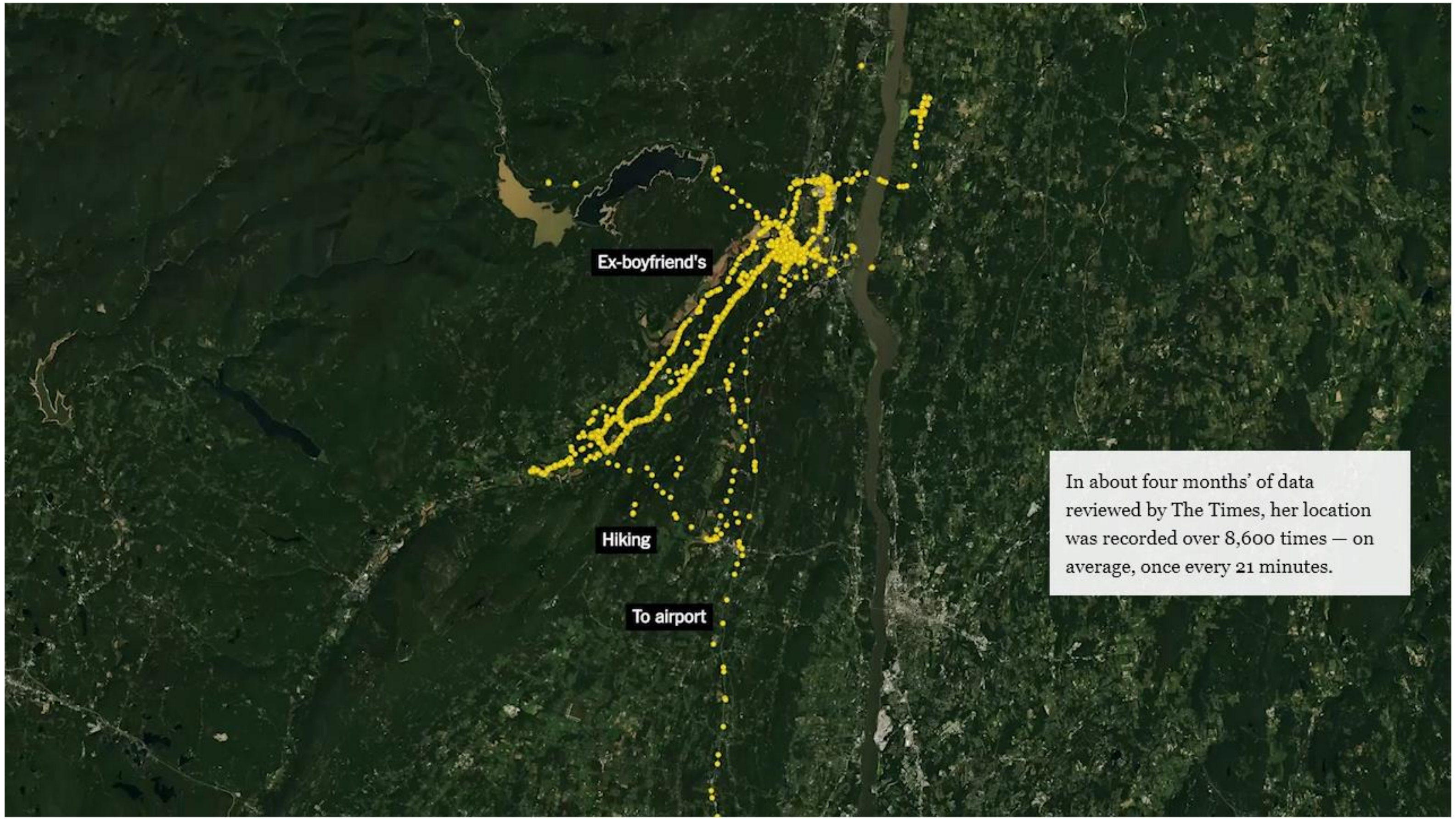
Not so bad, already known information about her. But what about rest of her location data?

Location data

Can learn:

- Medical information about her

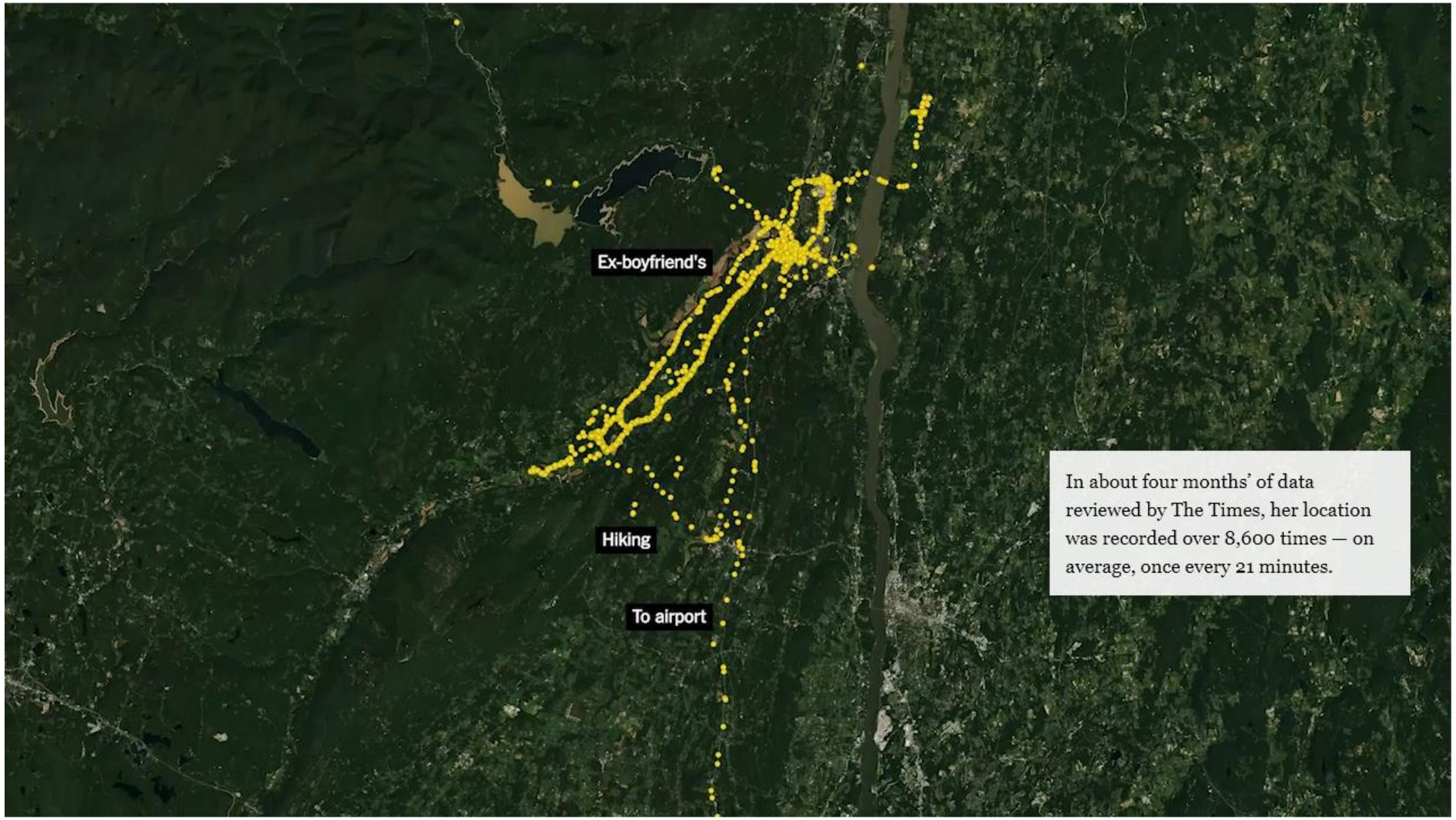
Location data



Can learn:

- Medical information about her
- Travel information
- Visits to ex-boyfriend
- When/where she hikes

Location data



Can learn:

- Medical information about her
- Travel information
- Visits to ex-boyfriend
- When/where she hikes

What is the harm?

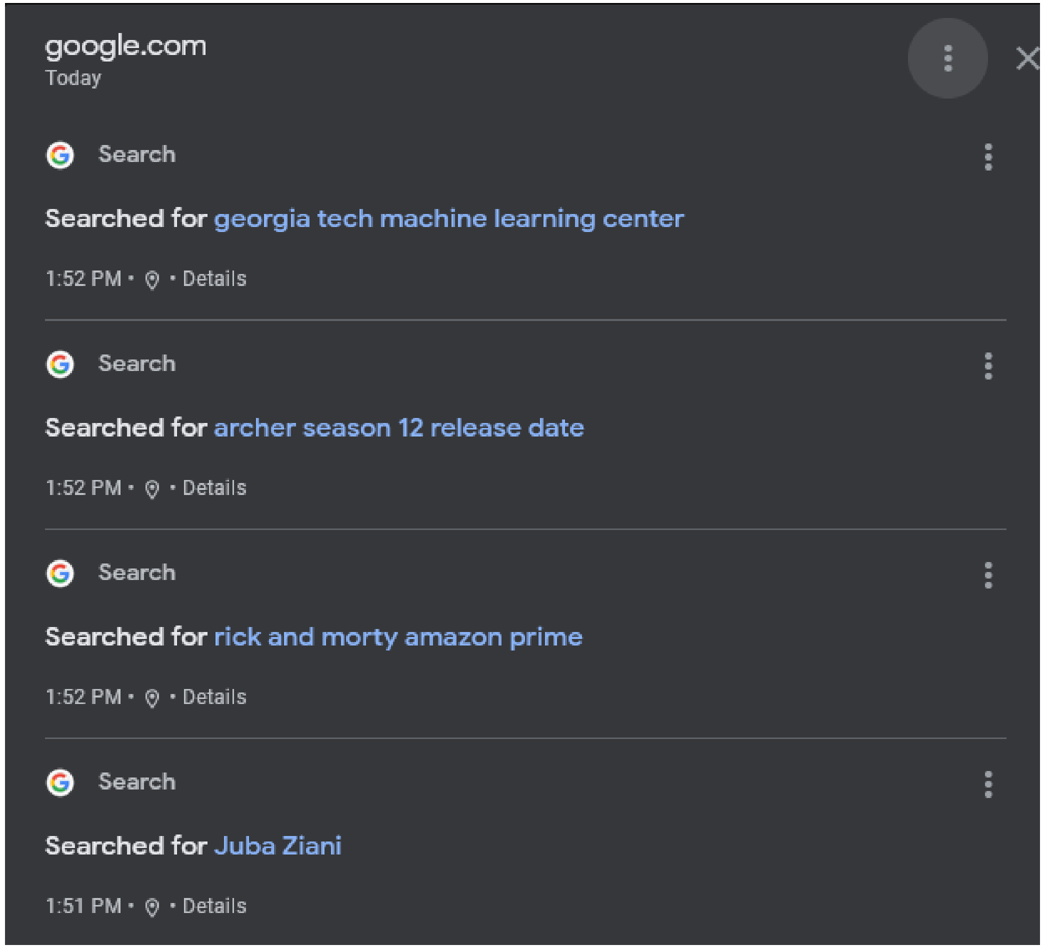
- At best, creepy.
- At worst, hurtful (know when to rob your place, how to blackmail you, etc.)

Location data

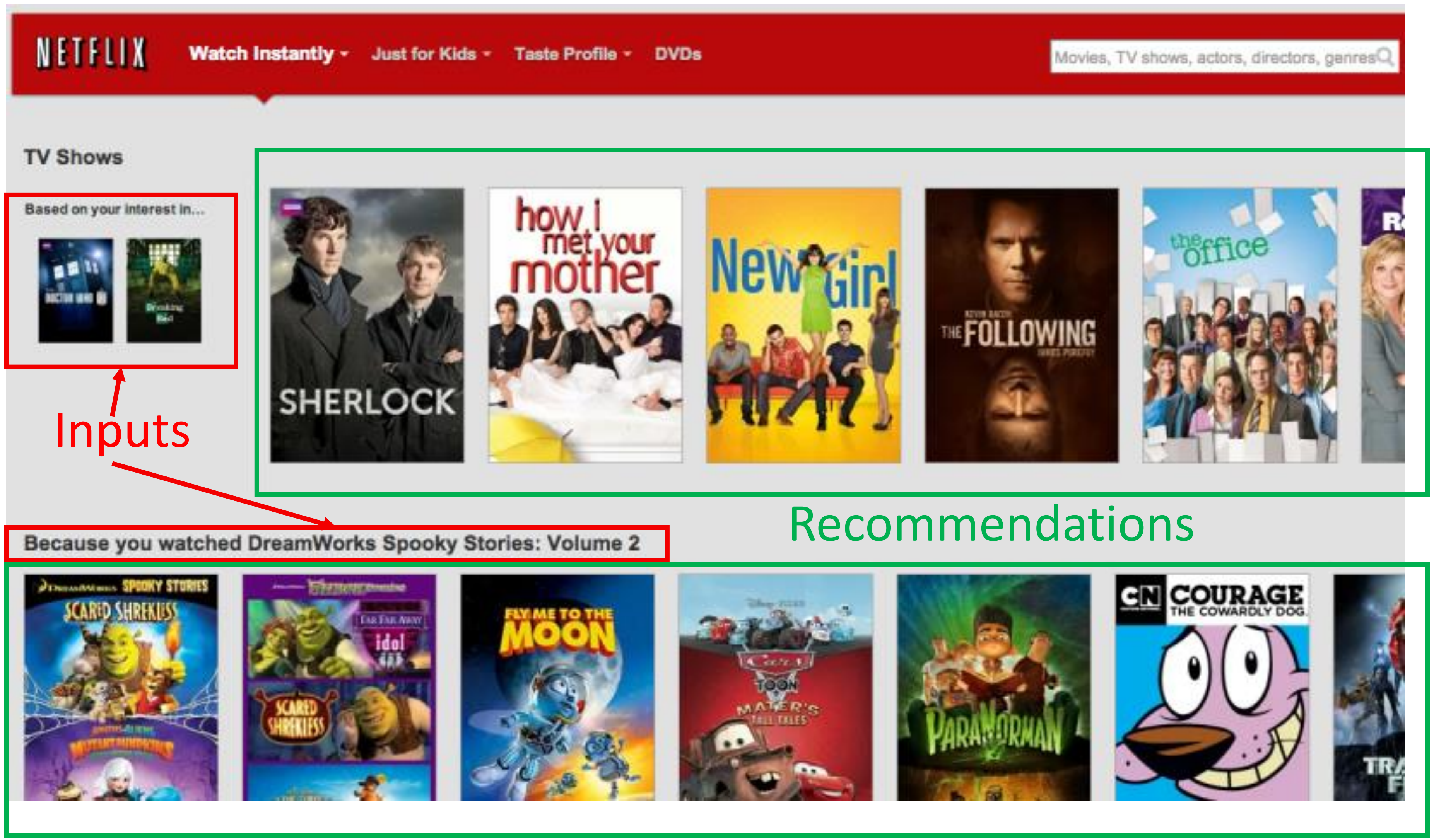
Other studies/breaches from location data:

- Tracked a person working with the mayor of NY
- Tracked workers dealing with sensitive techs/working in nuclear plants
- Nurse tracked to the main operating room at her hospital. Expressed concerns about her privacy and the privacy of her patients
- Tracking people to Planned Parenthood and abortion clinics
- Etc.

Search history



A Deeper Problem: the Netflix Competition



The Netflix Competition

How to improve recommendation system?

- Machine learning competition
- Try to predict user ratings from historical data as well as possible
- Provide "*anonymized*" data to participating teams

Netflix provided more than just anonymization:

- Only small subsets of the full data; reduced the number of attributes
- Deleted some of the ratings
- Modified dates/temporal data

The Netflix Competition

- **“How To Break Anonymity of the Netflix Prize Dataset”, Arvind Narayanan and Vitaly Shmatikov, 2006**
- Only 2 weeks after the Netflix competition

The Netflix Competition

- **"How To Break Anonymity of the Netflix Prize Dataset", Arvind Narayanan and Vitaly Shmatikov, 2006**
- Only 2 weeks after the Netflix competition

What they show:

- Only need imperfect info:
    1. approx. dates of rating ($\pm$2 weeks) for 6 movies
    2. 2 ratings and dates (with a 3-day error)

- Can uniquely identify the person:
    1. 99% of the time
    2. 68% of the time

The Netflix Competition

- **“How To Break Anonymity of the Netflix Prize Dataset”, Arvind Narayanan and Vitaly Shmatikov, 2006**
- Only 2 weeks after the Netflix competition

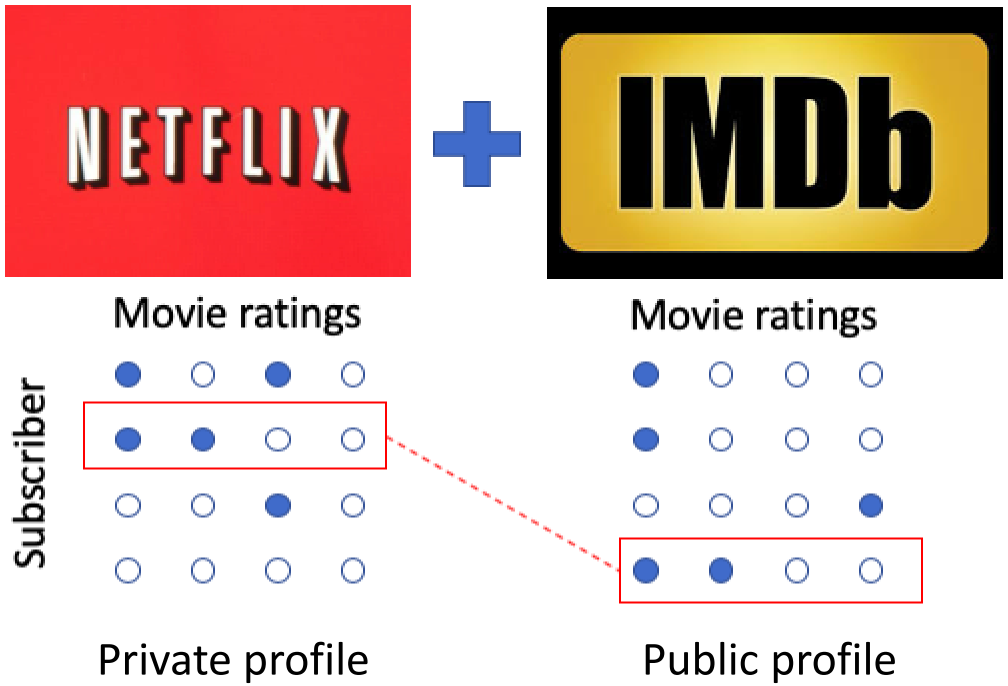
What they show:

- Only need imperfect info:
  1. approx. dates of rating ($\pm$2 weeks) for 6 movies
  2. 2 ratings and dates (with a 3-day error)

- Can uniquely identify the person:
  1. 99% of the time
  2. 68% of the time

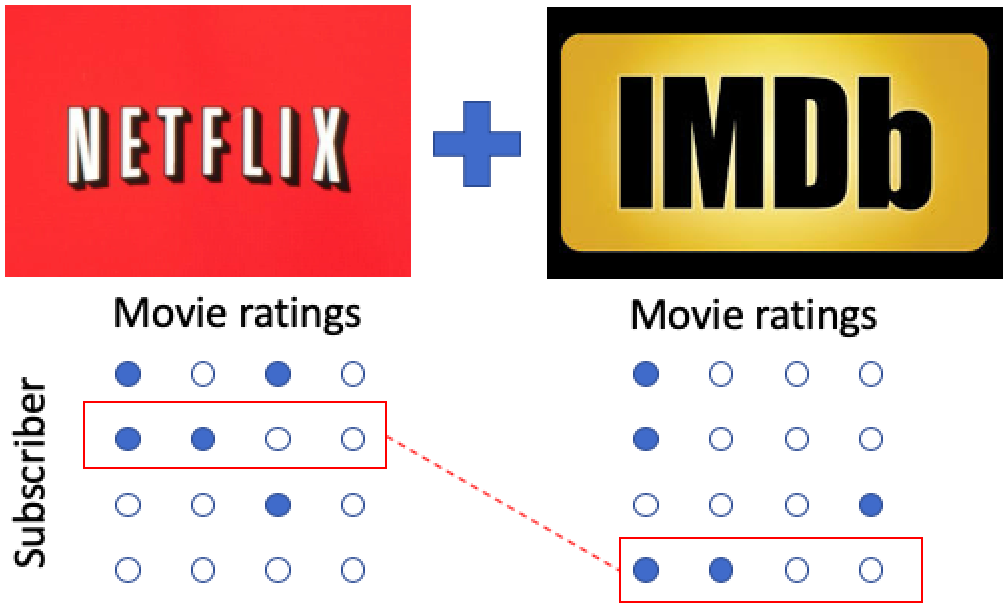
The Netflix Competition

How did they do it?



The Netflix Competition

How did they do it?



Why is it bad?

- Netflix watch history: more expansive and private than imdb public rating
- Link imdb and Netflix profile ➔ learn private watch history on Netflix
- Gay mother sued Netflix: watch history could reveal her sexual orientation to others

Hiding identifiable features: k-anonymization

Hospital X's data

Name	Age	Gender	Zip Code	Smoker	Diagnosis
Richard	64	Male	19146	Yes	**Heart disease**
Susan	61	Female	19118	No	**Arthritis**
Matthew	67	Male	19104	Yes	**Lung cancer**
Alice	63	Female	19146	No	**Crohn's disease**
Rebecca	56	Female	19103	Yes	**HIV**
Lisa	55	Female	19146	Yes	**Ulcerative colitis**

(from *The Ethical Algorithm*, by Michael Kearns and Aaron Roth)

Hiding identifiable features: k-anonymization

Hospital X's data

Name	Age	Gender	Zip Code	Smoker	Diagnosis
*	60-70	Male	191**	Yes	Heart disease
*	60-70	Female	191**	No	Arthritis
*	60-70	Male	191**	Yes	Lung cancer
*	60-70	Female	191**	No	Crohn's disease
*	50-60	Female	191**	Yes	HIV
*	50-60	Female	191**	Yes	Ulcerative colitis

(from *The Ethical Algorithm*, by Michael Kearns and Aaron Roth)

Hiding identifiable features: k-anonymization

Hospital X's data

Name	Age	Gender	Zip Code	Smoker	Diagnosis
*	60-70	Male	191**	Yes	Heart disease
*	60-70	Female	191**	No	Arthritis
*	60-70	Male	191**	Yes	Lung cancer
*	60-70	Female	191**	No	Crohn's disease
*	50-60	Female	191**	Yes	HIV
*	50-60	Female	191**	Yes	Ulcerative colitis

(from *The Ethical Algorithm*, by Michael Kearns and Aaron Roth)

k-anonymization – Issue #1

Hospital X’s data

Name	Age	Gender	Zip Code	Smoker	Diagnosis
Richard?	60-70	Male	191**	Yes	**Heart disease**
Not Richard	60-70	Female	191**	No	Arthritis
Richard?	60-70	Male	191**	Yes	**Lung cancer**
Not Richard	60-70	Female	191**	No	Crohn’s disease
Not Richard	50-60	Female	191**	Yes	HIV
Not Richard	50-60	Female	191**	Yes	Ulcerative colitis

- Don’t know Richard’s exact medical condition
- But, know Richard has a serious medical condition (either lung cancer or heart disease)

k-anonymization – Issue #2

Additional information! Hospital Y's data:

Name	Age	Gender	Zip Code	Diagnosis
*	50-60	Female	191**	HIV
*	50-60	Female	191**	Lupus
*	50-60	Female	191**	Hip fracture
*	…	…	…	…

- In hospital X, only 2 females between age 50-60. Only one has HIV.
- Imagine we know *Rebecca went to both hospitals X and Y.*
  ➔ The 50-60 female with HIV is the only person in both X and Y.
  ➔ Must be Rebecca
  ➔ Rebecca has HIV!

k-anonymization – Issue #2

Name	Age	Gender	Zip Code	Diagnosis
*	50-60	Female	191**	HIV
*	50-60	Female	191**	Lupus
*	50-60	Female	191**	Hip fracture
*	…	…	…	…

3-anon!

Cross-referencing several k-anonymous databases breaks k-anonymity!

What lesson did we learn from failures of anonymization ?

Data aggregation

Idea:

- Only release aggregated statistics/model.
- Examples:
  - Population-level statistics such as averages, etc.
  - Neural net (only see the final model, not the training data)

Why should it naively work?

- No individual-level details or features!
- Cannot identify a single row in a DB/no access to such row-by-row data

Data aggregation: genomic data

Genome-wide association studies:

G A T A T T C G T A C G G A T T

G A T G T T C G T A C T G A A T

G A T A T T C G T A C G G A T T

G A T A T T C G T A C G G A A T

G A T G T T C G T A C T G A A T

G A T G T T C G T A C T G A A T

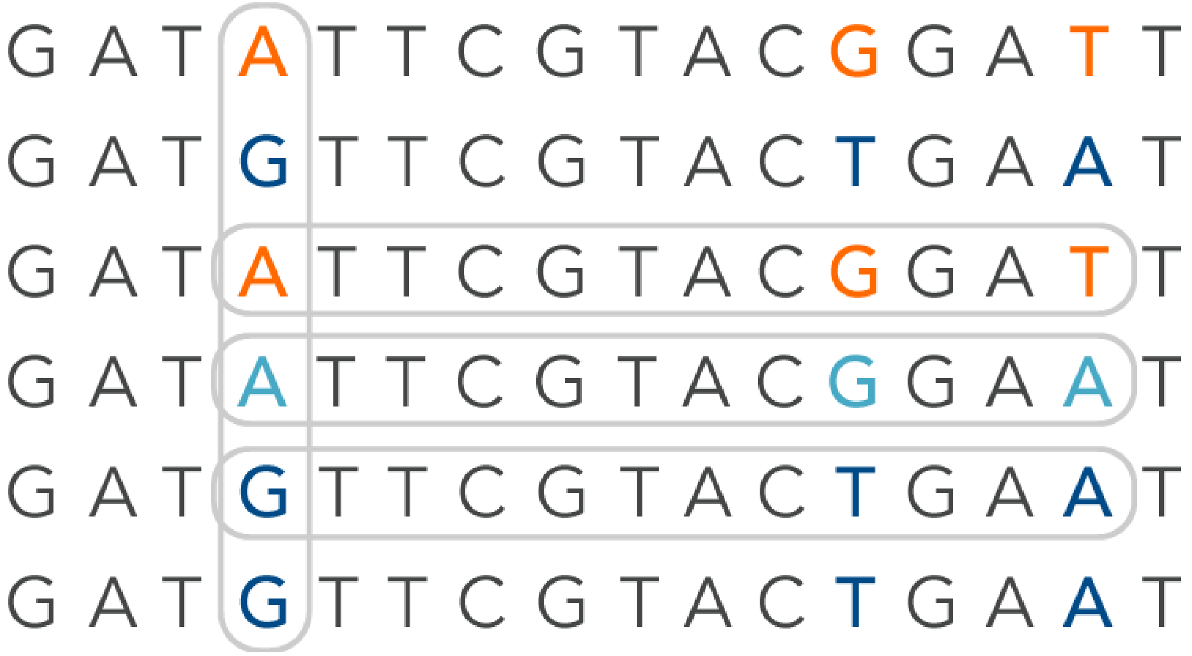
Data aggregation: genomic data

Genome-wide association studies:

Single nucleotide
polymorphism (SNP)

Data aggregation: genomic data

Genome-wide association studies:



Single nucleotide polymorphism (SNP)

Disease associated to specific SNPs?

Data aggregation: genomic data

“Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays”, Homer et al., 2008

Can tell whether an individual with known genotype appears in a certain mixture of DNA samples

How?

- Statistical analysis: correlation on SNPs/alleles between i) individual’s data and ii) distribution of alleles in relevant population
- Minimal correlation for a single SNPs…
- … But thousands of SNPs ➔ strong correlation

Data aggregation: genomic data

Is this a problem?

- Need to already know an individual's SNPs to run this attack
- Only learn whether the individual's genetic data was used in study

Answer: Yes.

- Genomic data is more and more commonplace (ancestry tests, etc.)
- What if study only contains cancer patients/tries to link alleles to some rare disease? Can learn that you have a rare disease!

Data aggregation: neural nets

"The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks", Carlini et al., 2019

Data aggregation: neural nets

WHEN YOU TRAIN PREDICTIVE MODELS ON INPUT FROM YOUR USERS, IT CAN LEAK INFORMATION IN UNEXPECTED WAYS.

“The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks”, Carlini et al., 2019

Predictive models tend to memorize:

- Imperfect generalization/overfitting to dataset
- More obvious in language models:
  - Work by memorizing characters/word associations
  - Can repeat word associations from training data

Data aggregation: neural nets

WHEN YOU TRAIN PREDICTIVE MODELS ON INPUT FROM YOUR USERS, IT CAN LEAK INFORMATION IN UNEXPECTED WAYS.

"The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks", Carlini et al., 2019

Predictive models tend to memorize:

- Imperfect generalization/overfitting to dataset
- More obvious in language models:
  - Work by memorizing characters/word associations
  - Can repeat word associations from training data

Potential attack:

- Predict next word: "My SSN is…"
- Recovers some SSN used in training data

Beyond aggregating: adding noise

Answering queries exactly is not enough for privacy, even if queries aggregate a lot of data

Beyond aggregating: adding noise

Answering queries exactly is not enough for privacy, even if queries aggregate a lot of data

Natural next step:

- Do not answer queries exactly!
- Add noise/randomness to data or to queries

Q: Is this enough?

Beyond aggregating: adding noise

Answering queries exactly is not enough for privacy, even if queries aggregate a lot of data

Natural next step:

- Do not answer queries exactly!
- Add noise/randomness to data or to queries

Q: Is this enough?

A: You have to be careful ***how and how much noise*** you add

Adding noise inadequately: Aircloak's failures

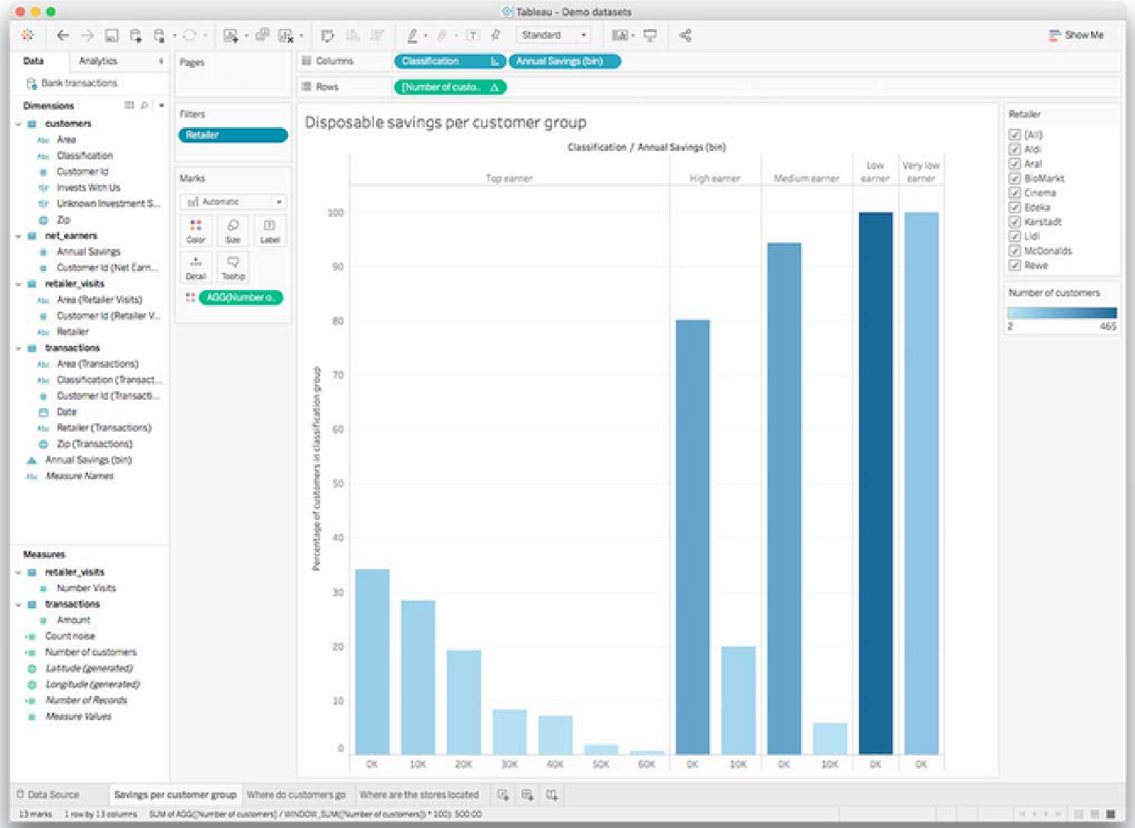
First True Anonymization Solution to Provide High-Quality Analytics

Aircloak's unique approach ensures the existing primary database is not modified in any way. Aircloak handles all data types including unstructured text.

Aircloak offers analysts a rich explorative SQL database interface. Analysts submit SQL queries and interact with the existing database to extract the requested data.

Both the queries and responses are dynamically modified by Aircloak to ensure anonymity while still providing high accuracy.

Read more about Aircloak Insights' features



Adding noise inadequately: Aircloak's failures

aircloak

Home Solutions Background Company Blog

Aircloak Attack Challenge

The first bounty program for anonymized data re-identification. As part of its commitment to transparency and strong anonymization, Aircloak offers the world's first bounty program for re-identification of anonymized data.

Adding noise inadequately: Aircloak’s failures

Differential privacy researcher were able to recover perfectly 90% of the SSNs in the database:

- Aloni Cohen and Kobbi Nissim, 2017

Adding noise inadequately: Aircloak's failures

Differential privacy researcher were able to recover perfectly 90% of the SSNs in the database:

- Aloni Cohen and Kobbi Nissim, 2017
- Aircloak used ad-hoc fix: reduced query language to prevent specific attack, without addressing amount of noise added…

Adding noise inadequately: Aircloak's failures

Differential privacy researcher were able to recover perfectly 90% of the SSNs in the database:

- Aloni Cohen and Kobbi Nissim, 2017
- Aircloak used ad-hoc fix: reduced query language to prevent specific attack, without addressing amount of noise added…
- Travis Dick, Matthew Joseph, Zachary Schutzman, 2020. Very slight modification of 2017 attack!

Adding noise inadequately: Aircloak's failures

Differential privacy researcher were able to recover perfectly 90% of the SSNs in the database:

- Aloni Cohen and Kobbi Nissim, 2017
- Aircloak used ad-hoc fix: reduced query language to prevent specific attack, without addressing amount of noise added…
- Travis Dick, Matthew Joseph, Zachary Schutzman, 2020. Very slight modification of 2017 attack!

Even worse:

They used a simple reconstruction attack known since… 2003!

"Revealing information while preserving privacy", Irit Dinur and Kobbi Nissim

What does the Dinur-Nissim paper show?

Goal: try to recover secret bits of users in a database of n users

Theorem 1: There exists a reconstruction attack that issues 2^n queries, obtains answers with error αn, and reconstruct the secret bits of all but $4\alpha n$ users.

What does the Dinur-Nissim paper show?

Goal: try to recover secret bits of users in a database of n users

Theorem 1: There exists a reconstruction attack that issues 2^n queries, obtains answers with error αn, and reconstruct the secret bits of all but $4\alpha n$ users.

How bad is this?

- $\alpha = O(1/n)$ ➔ all but $O(1)$ users! Effectively everyone!
- $\alpha = O(1/n^{1/2})$ ➔ all but $O\left(n^{1/2}\right)$ users, out of n. Almost everyone!

What does the Dinur-Nissim paper show?

Goal: try to recover secret bits of users in a database of n users

Theorem 1: There exists a reconstruction attack that issues 2^n queries, obtains answers with error αn, and reconstruct the secret bits of all but $4\alpha n$ users.

How bad is this?

- $\alpha = O(1/n)$ ➔ all but $O(1)$ users! Effectively everyone!
- $\alpha = O(1/n^{1/2})$ ➔ all but $O\left(n^{1/2}\right)$ users, out of n. Almost everyone!

But this is an inefficient attack. Requires exponentially (in n) many queries!

What does the Dinur-Nissim paper show?

Goal: try to recover secret bits of users in a database of n users

Theorem 2: There exists a reconstruction attack that issues $O(n)$ (random) queries, obtains answers with error αn, and reconstruct the secret bits of all but $O(\alpha^2 n^2)$ users.

What does the Dinur-Nissim paper show?

Goal: try to recover secret bits of users in a database of n users

Theorem 2: There exists a reconstruction attack that issues $O(n)$ (random) queries, obtains answers with error αn, and reconstruct the secret bits of all but $O(\alpha^2 n^2)$ users.

How bad is this?

- $\alpha = O(1/n^\wedge\alpha)$ ➔ all but $O\left(n^{2-2\alpha}\right)$ users! Almost everyone for $\alpha < 1$, n large
- $\alpha = O(1/n^{1/2})$ ➔ all but $O(1)$ users. Almost nobody.

What does the Dinur-Nissim paper show?

Goal: try to recover secret bits of users in a database of n users

Theorem 2: There exists a reconstruction attack that issues $O(n)$ (random) queries, obtains answers with error αn, and reconstruct the secret bits of all but $O(\alpha^2 n^2)$ users.

How bad is this?

- $\alpha = O(1/n\hat{}\alpha)$ ➔ all but $O(n^{2-2\alpha})$ users! Almost everyone for $\alpha < 1$, n large
- $\alpha = O(1/n^{1/2})$ ➔ all but $O(1)$ users. Almost nobody.

To protect privacy on most of the database against computationally efficient attacks, need noise of the order of at least $n^{1/2}$.

What does the Dinur-Nissim paper show?

For a summary of how to perform these attacks:

- https://differentialprivacy.org/reconstruction-theory/
- https://differentialprivacy.org/diffix-attack/

Link to the full paper:

- https://crypto.stanford.edu/seclab/sem-03-04/psd.pdf

What we have learned so far?

Examples of failures of privacy techniques:

1. Anonymization allows simple re-identification through public features
2. Aggregation is still not enough. Cannot answer statistical queries exactly. Vulnerable to reconstruction attacks.
3. Adding noise is the right direction, but this noise needs to be calibrated carefully.

Overall message:

- Intuitive or ad-hoc privacy measures that anticipate specific attacks do not work.
- Vulnerable to unanticipated, and sometimes very simple attacks.

Differential privacy is the only known framework to rigorously prevent such reconstruction attacks and privacy violations