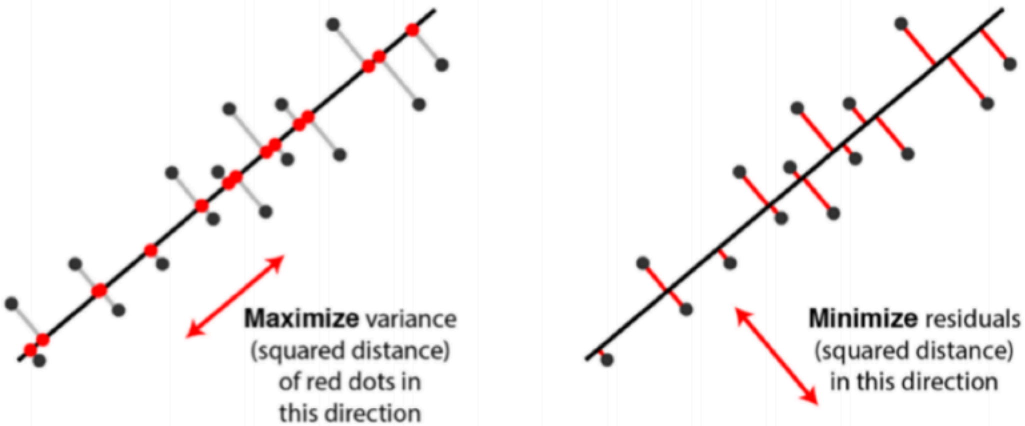
Privacy and Fairness in ML

UNIVERSITY of WASHINGTON

This course so far

How can one predict some value, find structure from data?



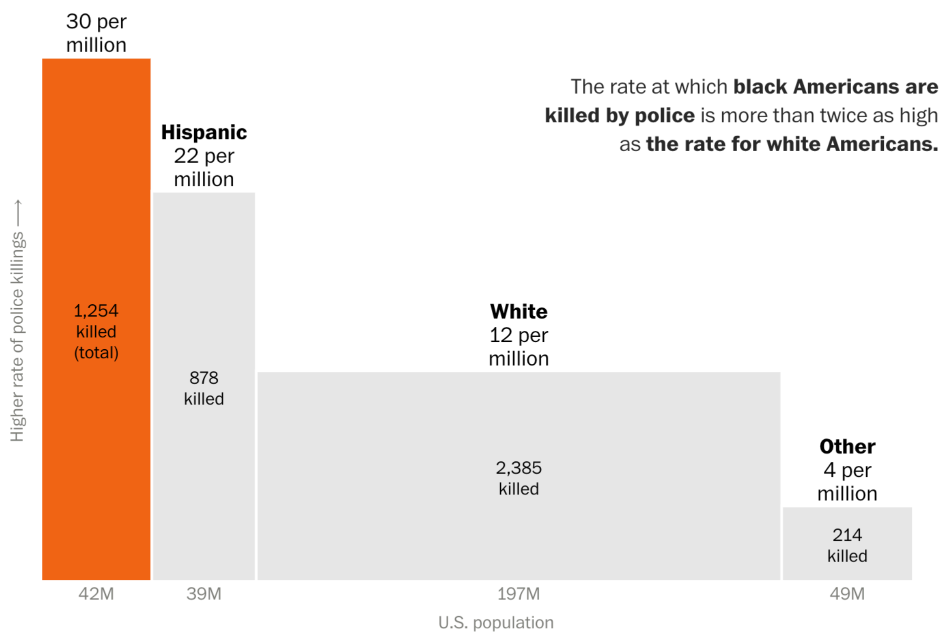
Two equivalent views of principal component analysis.

First, a moment of reflection

Some sobering statistics

Many, many people are killed by police every year (1,003 in the past 12 months)

> half of which aren't reported to the FBI (reporting is voluntary)



Source: an excellent Washington Post interactive database

Design choices

Model Class

Features to collect

Loss function

Regularization

Optimization Method

Constraints

…

Design choices affect your model

Model Class

Features to collect

Loss function

Regularization

Optimization Method

Constraints

…

These will all affect what model you find, and how well and when that structure will generalize

… and that can have consequences "beyond" ML

Model Class

Features to collect

Loss function

Regularization

Optimization Method

Constraints

…

These will all affect what model you find, and how well and when that structure will generalize

Different models will have different kinds of errors, predictions, and failure modes

Consequence #1: Good predictions for whom?

UNIVERSITY of WASHINGTON

The US Criminal Justice System

The US Criminal Justice System

This is a description of how things (largely) are and have been, not how they ought to be.

The US Criminal Justice System

This is a description of how things (largely) are and have been, not how they ought to be.

Any use of ML in the criminal justice system is just a tiny part of the larger CJ ecosystem.

The US Criminal Justice System

This is a description of how things (largely) are and have been, not how they ought to be.

Any use of ML in the criminal justice system is just a tiny part of the larger CJ ecosystem.

Moreover, “reoffend” or “recidivate” are very, very, very far from precise terms in their use (in this lecture, and in industry).

The US Criminal Justice System

The US Criminal Justice System

(Lots and lots and lots and lots of ~~possibly~~ biased processes, including racialized choices of where to police, what behaviors are most dangerous, ...)

The US Criminal Justice System

(Lots and lots and lots and lots of ~~possibly~~ biased processes, including racialized choices of where to police, what behaviors are most dangerous, …)

-> A person is charged with a crime

The US Criminal Justice System

(Lots and lots and lots and lots of ~~possibly~~ biased processes, including racialized choices of where to police, what behaviors are most dangerous, …)

-> A person is charged with a crime

-> A judge determines whether to detain them or release them on bail until their trial

The US Criminal Justice System

(Lots and lots and lots and lots of ~~possibly~~ biased processes, including racialized choices of where to police, what behaviors are most dangerous, …)

-> A person is charged with a crime

-> A judge determines whether to detain them or release them on bail until their trial

-> (Many, many months or years pass)

The US Criminal Justice System

(Lots and lots and lots and lots of ~~possibly~~ biased processes, including racialized choices of where to police, what behaviors are most dangerous, …)

-> A person is charged with a crime

-> A judge determines whether to detain them or release them on bail until their trial

-> (Many, many months or years pass)

-> The trial determines whether the justice system considers a person guilty or innocent of charges

The US Criminal Justice System

(Lots and lots and lots and lots of ~~possibly~~ biased processes, including racialized choices of where to police, what behaviors are most dangerous, …)

-> A person is charged with a crime

-> A judge determines whether to detain them or release them on bail until their trial

-> (Many, many months or years pass)

-> The trial determines whether the justice system considers a person guilty or innocent of charges

-> A sentencing hearing determines the length of prison time, fines, …

The US Criminal Justice System

(Lots and lots and lots and lots of ~~possibly~~ biased processes, including racialized choices of where to police, what behaviors are most dangerous, …)

-> A person is charged with a crime

-> A judge determines whether to detain them or release them on bail until their trial

-> (Many, many months or years pass)

-> The trial determines whether the justice system considers a person guilty or innocent of charges

-> A sentencing hearing determines the length of prison time, fines, …

-> With "good behavior", people in prison can be considered for parole

What does the CJ process have to do with ML?

(Lots and lots and lots and lots of ~~possibly~~ biased processes, including racialized choices of where to police, what behaviors are most dangerous, …)

-> A person is charged with a crime

-> A judge determines whether to detain them or release them on bail until their trial

-> (Many, many months or years pass)

-> The trial determines whether the justice system considers a person guilty or innocent of charges

-> A sentencing hearing determines the length of prison time, fines, …

-> With "good behavior", people in prison can be considered for parole

What does the CJ process have to do with ML?

Predictive policing: determine (statistically) where to send police

-> A person is charged with a crime

-> A judge determines whether to detain them or release them on bail until their trial

-> (Many, many months or years pass)

-> The trial determines whether the justice system considers a person guilty or innocent of charges

-> A sentencing hearing determines the length of prison time, fines, …

-> With "good behavior", people in prison can be considered for parole

What does the CJ process have to do with ML?

Predictive policing: determine (statistically) where to send police

-> A person is charged with a crime

Predict probability of reappearing, being charged with another crime if on parole, …

-> (Many, many months or years pass)

-> The trial determines whether the justice system considers a person guilty or innocent of charges

-> A sentencing hearing determines the length of prison time, fines, …

-> With “good behavior”, people in prison can be considered for parole

What does the CJ process have to do with ML?

Predictive policing: determine (statistically) where to send police

-> A person is charged with a crime

Predict probability of reappearing, being charged with another crime if on parole, …

-> (Many, many months or years pass)

-> The trial determines whether the justice system considers a person guilty or innocent of charges

Predict risk of being charged with another crime, use in sentencing

-> With “good behavior”, people in prison can be considered for parole

What does the CJ process have to do with ML?

Predictive policing: determine (statistically) where to send police

-> A person is charged with a crime

Predict probability of reappearing, being charged with another crime if on parole, …

-> (Many, many months or years pass)

-> The trial determines whether the justice system considers a person guilty or innocent of charges

Predict risk of being charged with another crime, use in sentencing

Predict risk of being charged with another crime, use in parole decisions

The US Criminal Justice System

The US Criminal Justice System

“reoffend” or “recidivate” are very, very, very far from precise terms in their use (in this lecture, and in industry).

The US Criminal Justice System

“reoffend” or “recidivate” are very, very, very far from precise terms in their use (in this lecture, and in industry).

Largely, it corresponds to either being:

The US Criminal Justice System

“reoffend” or “recidivate” are very, very, very far from precise terms in their use (in this lecture, and in industry).

Largely, it corresponds to either being:

- arrested

The US Criminal Justice System

“reoffend” or “recidivate” are very, very, very far from precise terms in their use (in this lecture, and in industry).

Largely, it corresponds to either being:

- **arrested**

- **charged**

The US Criminal Justice System

“reoffend” or “recidivate” are very, very, very far from precise terms in their use (in this lecture, and in industry).

Largely, it corresponds to either being:

- arrested
- charged
- or being found guilty of a crime

The US Criminal Justice System

“reoffend” or “recidivate” are very, very, very far from precise terms in their use (in this lecture, and in industry).

Largely, it corresponds to either being:

- arrested
- charged
- or being found guilty of a crime

usually one which is violent.

Can such predictions be fair with respect to race?

Can such predictions be fair with respect to race?

Attempt #1: don't use race as a feature

  Issue: race is strongly correlated with other features

Attempt #2: Remove any feature correlated with race

  Issue: (nearly) all features correlated with race

"Fairness through unawareness"

What exactly is the goal of this?

Can such predictions be fair with respect to race?

Attempt #1: don’t use race as a feature

    Issue: race is strongly correlated with other features

Attempt #2: Remove any feature correlated with race

    Issue: (nearly) all features correlated with race

“Fairness through unawareness”

What exactly is the goal of this?

To accurately predict which people are most likely commit a crime if released, so they can be released

Can such predictions be fair with respect to race?

Attempt #1: don't use race as a feature

    Issue: race is strongly correlated with other features

Attempt #2: Remove any feature correlated with race

    Issue: (nearly) all features correlated with race

"Fairness through unawareness"

What exactly is the goal of this?

To accurately predict which people are most likely commit a crime if released, so they can be released

Can such predictions be fair with respect to race?

Attempt #1: don't use race as a feature

  Issue: race is strongly correlated with other features

Attempt #2: Remove any feature correlated with race

  Issue: (nearly) all features correlated with race

"Fairness through unawareness"

What exactly is the goal of this?

To accurately predict which people are most likely commit a crime if released, so they can be released

… where these predictions should be accurate? similar? for people of all races.

Can such predictions be fair with respect to race?

Pretrial release risk scale: 1-10
General recidivism scale
Violent recidivism scale

Attempt #1: don't use race as a feature

Issue: race is strongly correlated with other features

Attempt #2: Remove any feature correlated with race

Issue: (nearly) all features correlated with race

"Fairness through unawareness"

What exactly is the goal of this?

To accurately predict which people are most likely commit a crime if released, so they can be released

… where these predictions should be accurate? similar? for people of all races.

Can such predictions be fair with respect to race?

Pretrial release risk scale: 1-10
General recidivism scale
Violent recidivism scale

Attempt #1: don't use race as a feature
    Issue: race is strongly correlated with other features
Attempt #2: Remove any feature correlated with race
    Issue: (nearly) all features correlated with race

"Fairness through unawareness"

What exactly is the goal of this?

To accurately predict which people are most likely commit a crime if released, so they can be released

Concern: not all errors are equally costly.

… where these predictions should be accurate? similar? for people of all races.

Can such predictions be fair with respect to race?

Can such predictions be fair with respect to race?

Pretrial release risk scale: 1-10
General recidivism scale
Violent recidivism scale

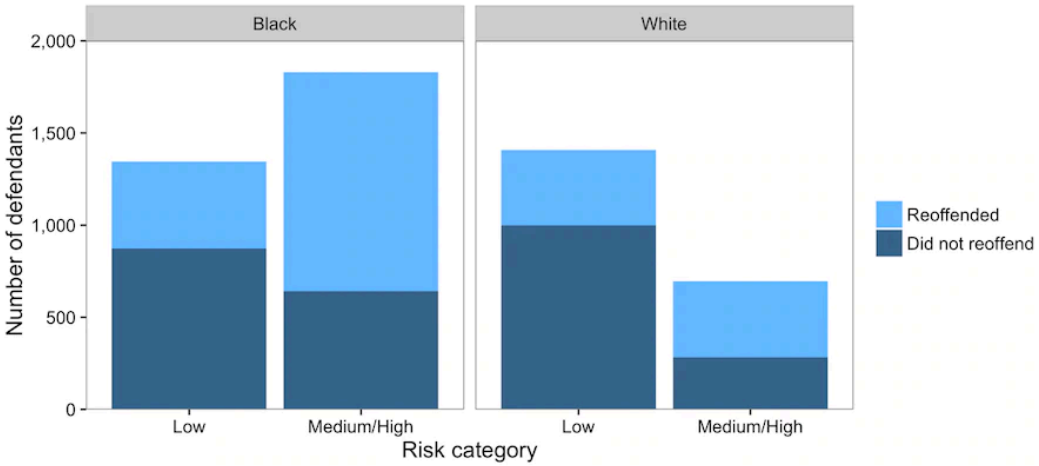
Can such predictions be fair with respect to race?

Pretrial release risk scale: 1-10
General recidivism scale
Violent recidivism scale

Source: Washington Post

Can such predictions be fair with respect to race?

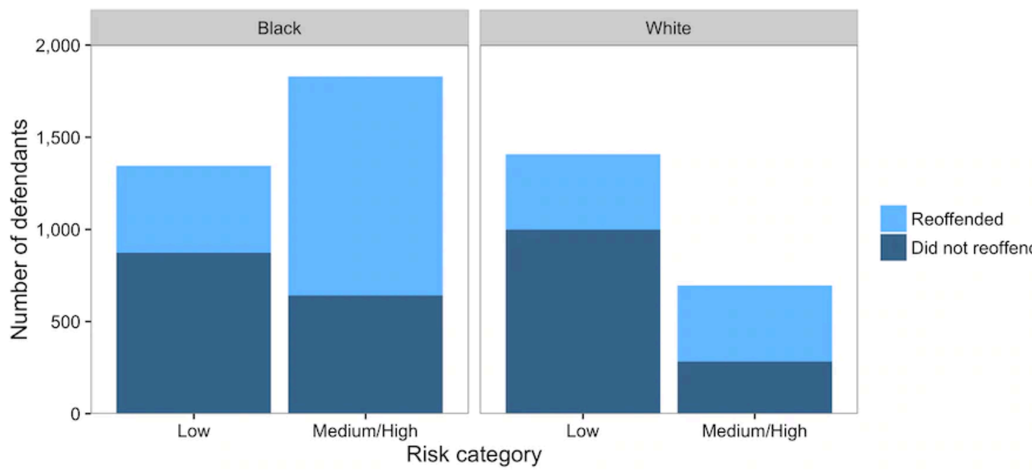
Pretrial release risk scale: 1-10
General recidivism scale
Violent recidivism scale



Source: Washington Post

Can such predictions be fair with respect to race?

Pretrial release risk scale: 1-10
General recidivism scale
Violent recidivism scale



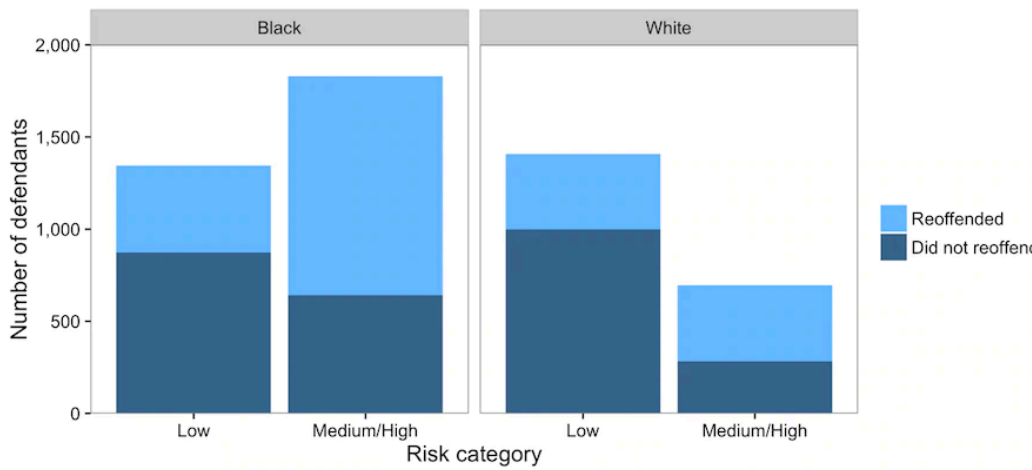
$$\mathbb{P}[\text{does not reoffend}|\text{predicted low risk, white}] <$$
$$\mathbb{P}[\text{does not reoffend}|\text{predicted low risk, black}]$$

Source: Washington Post

Can such predictions be fair with respect to race?

NORTHPOINTE

Pretrial release risk scale: 1-10
General recidivism scale
Violent recidivism scale



$$\mathbb{P}[\text{does not reoffend}|\text{predicted low risk, white}] <$$
$$\mathbb{P}[\text{does not reoffend}|\text{predicted low risk, black}]$$

False positive rate for black defendants higher than for white defendants.

Source: Washington Post

Can such predictions be fair with respect to race?

NORTHPOINTE

Calibration of scores w.r.t. race

Can such predictions be fair with respect to race?

Calibration of scores w.r.t. race

Source: Washington Post

Can such predictions be fair with respect to race?

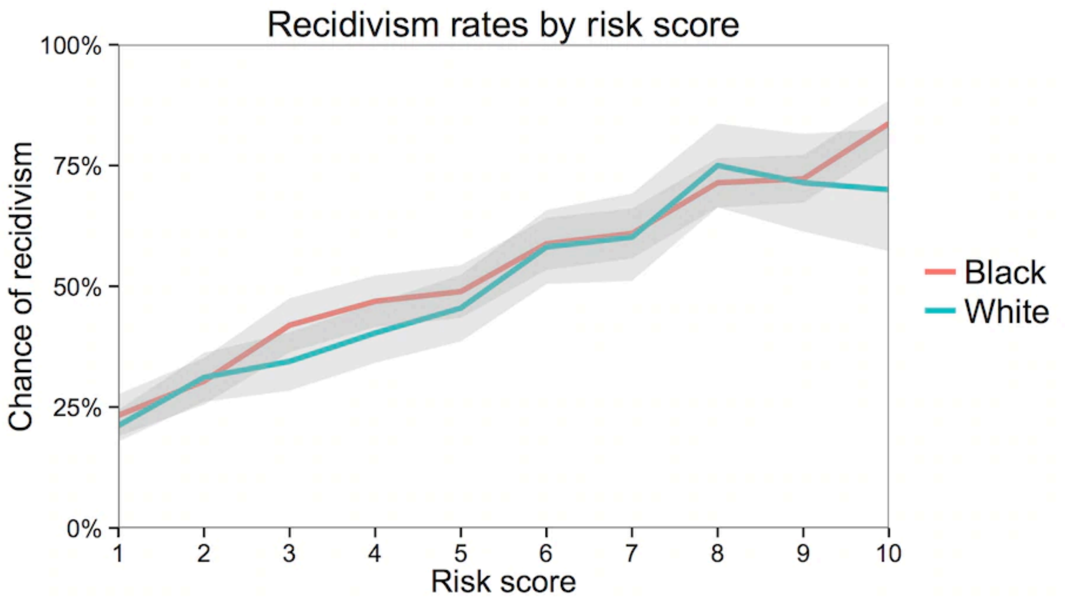
Pretrial release risk scale: 1-10
General recidivism scale
Violent recidivism scale

Calibration of scores w.r.t. race

Source: Washington Post

Can such predictions be fair with respect to race?

Pretrial release risk scale: 1-10
General recidivism scale
Violent recidivism scale

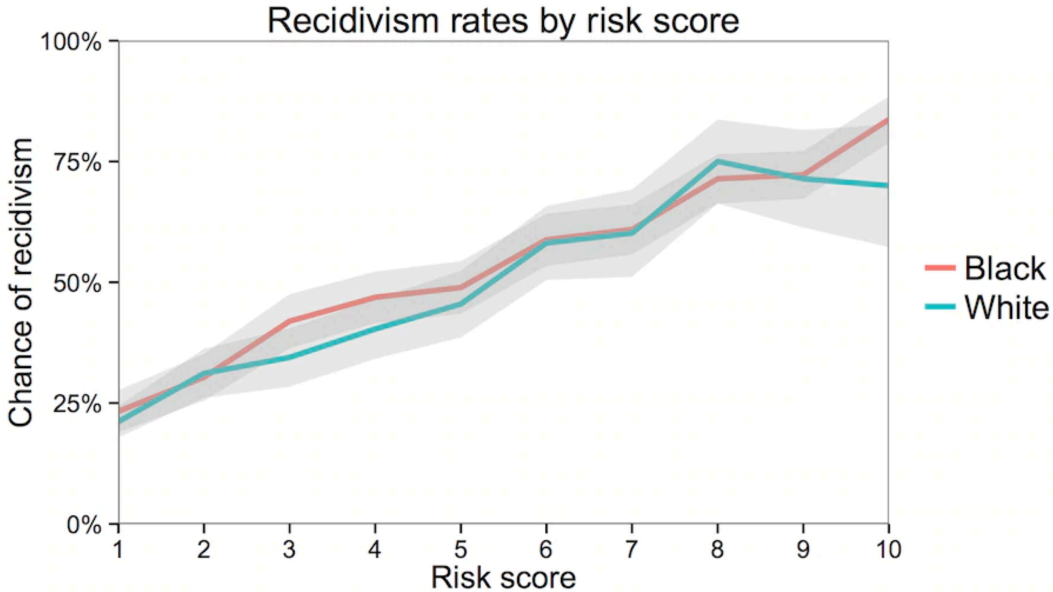


Calibration of scores w.r.t. race

[Source: Washington Post]

Can such predictions be fair with respect to race?

Pretrial release risk scale: 1-10
General recidivism scale
Violent recidivism scale



$$\mathbb{P}[\text{reoffend}|\text{prediction, white}] \approx \mathbb{P}[\text{reoffend}|\text{prediction, black}]$$

Calibration of scores w.r.t. race

Source: Washington Post

Can any predictions be fair with respect to race?

$$\mathbb{P}[\text{does not reoffend}|\text{predicted low risk, white}] <$$
$$\mathbb{P}[\text{does not reoffend}|\text{predicted low risk, black}]$$

while

$$\mathbb{P}[\text{reoffend}|\text{prediction, white}] \approx \mathbb{P}[\text{reoffend}|\text{prediction, black}]$$

How is this possible?

Is this avoidable?

Can any predictions be fair with respect to race?

False positives approximately equal and calibration are mutually exclusive, unless we have perfect predictions or the rate of what we are predicting is equal in every racial group.

… and this is true for more settings than just criminal justice.

Lending
Advertising
…

We are making a choice about how we allocate predictions.

Consequence #2: Leaking Sensitive information

Vignette 1: The Netflix Challenge

Given historical data on how users rated movies in past:

17,700 movies, 480,189 users, 99,072,112 ratings

(Sparsity: 1.2%)

Predict how the same users will rate movies in the future (for $1 million prize)

Alice	1	?	?	4	?	
Bob	?	2	5	?	?	
Carol	?	?	4	5	?	
Dave	5	?	?	?	4	
⋮						

Vignette 1: The Netflix Challenge

Given historical data on how users rated movies in past:

17,700 movies, 480,189 users, 99,072,112 ratings

(Sparsity: 1.2%)

Predict how the same users will rate movies in the future (for $1 million prize)

U1	1	?	?	4	?	
U2	?	2	5	?	?	
U3	?	?	4	5	?	
U4	5	?	?	?	4	
⋮						

Vignette 1: The Netflix Challenge

Given historical data on how users rated movies in past:

17,700 movies, 480,189 users, 99,072,112 ratings

(Sparsity: 1.2%)

Predict how the same users will rate movies in the future (for $1 million prize)

U1	1	?	?	4	?	
U2	?	2	5	?	?	
U3	?	?	4	5	?	
U4	5	?	?	?	4	
⋮						

\+

Vignette 1: The Netflix Challenge

Given historical data on how users rated movies in past:

17,700 movies, 480,189 users, 99,072,112 ratings

NETFLIX

(Sparsity: 1.2%)

Predict how the same users will rate movies in the future (for $1 million prize)

+ IMDb

De-anonymized records

Vignette 1: The Netflix Challenge

Given historical data on how users rated movies in past:

17,700 movies, 480,189 users, 99,072,112 ratings

(Sparsity: 1.2%)

Predict how the same users will rate movies in the future (for $1 million prize)

U1	1	?	?	4	?	…
U2	?	2	5	?	?	
U3	?	?	4	5	?	
U4	5	?	?	?	4	
⋮						

+

Conclusions:

Anonymization can be undone
linkage attacks help in de-anonymizing!

De-anonymized records

Vignette 2: Medical Records

Vignette 2: Medical Records

~ 7% of the US population is hospitalized per year, ~ 1% hospitalized > 1x

328.2 million people living in the US

Vignette 2: Medical Records

~ 7% of the US population is hospitalized per year, ~ 1% hospitalized > 1x

328.2 million people living in the US

In ~28 years, most people will have been hospitalized at least 2x

Vignette 2: Medical Records

~ 7% of the US population is hospitalized per year, ~ 1% hospitalized > 1x

328.2 million people living in the US

In ~28 years, most people will have been hospitalized at least 2x

Vignette 2: Medical Records

~ 7% of the US population is hospitalized per year, ~ 1% hospitalized > 1x

328.2 million people living in the US

In ~28 years, most people will have been hospitalized at least 2x

Suppose I have the dates of admission and nothing else.

Vignette 2: Medical Records

~ 7% of the US population is hospitalized per year, ~ 1% hospitalized > 1x

328.2 million people living in the US

In ~28 years, most people will have been hospitalized at least 2x

Suppose I have the dates of admission and nothing else.

Vignette 2: Medical Records

~ 7% of the US population is hospitalized per year, ~ 1% hospitalized > 1x

328.2 million people living in the US

- In ~28 years, most people will have been hospitalized at least 2x
- Suppose I have the dates of admission and nothing else.
- Does this uniquely identify anyone? If so, how many people?

Vignette 2: Medical Records

~ 7% of the US population is hospitalized per year, ~ 1% hospitalized > 1x

328.2 million people living in the US

- In ~28 years, most people will have been hospitalized at least 2x
- Suppose I have the dates of admission and nothing else.
- Does this uniquely identify anyone? If so, how many people?

Vignette 2: Medical Records

~ 7% of the US population is hospitalized per year, ~ 1% hospitalized > 1x

328.2 million people living in the US

- In ~28 years, most people will have been hospitalized at least 2x
- Suppose I have the dates of admission and nothing else.
- Does this uniquely identify anyone? If so, how many people?
- ~10,000 dates in 28 years, ~660 people admitted per day

Vignette 2: Medical Records

~ 7% of the US population is hospitalized per year, ~ 1% hospitalized > 1x

328.2 million people living in the US

In ~28 years, most people will have been hospitalized at least 2x

Suppose I have the dates of admission and nothing else.

Does this uniquely identify anyone? If so, how many people?

~10,000 dates in 28 years, ~660 people admitted per day

Vignette 2: Medical Records

~ 7% of the US population is hospitalized per year, ~ 1% hospitalized > 1x

328.2 million people living in the US

In ~28 years, most people will have been hospitalized at least 2x

Suppose I have the dates of admission and nothing else.

Does this uniquely identify anyone? If so, how many people?

~10,000 dates in 28 years, ~660 people admitted per day

Of the 659 people admitted with me on my first day, will any of them be admitted w. me on my second day? W.p. 559/9,999, or ~.5%

Vignette 2: Medical Records

~ 7% of the US population is hospitalized per year, ~ 1% hospitalized > 1x

328.2 million people living in the US

In ~28 years, most people will have been hospitalized at least 2x

Suppose I have the dates of admission and nothing else.

Does this uniquely identify anyone? If so, how many people?

~10,000 dates in 28 years, ~660 people admitted per day

Of the 659 people admitted with me on my first day, will any of them be admitted w. me on my second day? W.p. 559/9,999, or ~.5%

Vignette 2: Medical Records

~ 7% of the US population is hospitalized per year, ~ 1% hospitalized > 1x

328.2 million people living in the US

In ~28 years, most people will have been hospitalized at least 2x

Suppose I have the dates of admission and nothing else.

Does this uniquely identify anyone? If so, how many people?

~10,000 dates in 28 years, ~660 people admitted per day

Of the 659 people admitted with me on my first day, will any of them be admitted w. me on my second day? W.p. 559/9,999, or ~.5%

So dates of 2 hospital admissions will uniquely identify many, many people.

Vignette 2: Medical Records

~ 7% of the US population is hospitalized per year, ~ 1% hospitalized > 1x

328.2 million people living in the US

In ~28 years, most people will have been hospitalized at least 2x

Suppose I have the dates of admission and nothing else.

eople?

day

vill any of them be

or ~.5%

y many, many people.

Conclusions:

Anonymizing low dimensional data still isn’t safe… a small number of features can uniquely id people if those features take on many values

Large (#people) datasets aren’t enough to hide users either

A new approach to privacy

perfect privacy

No amount of analysis or linkage can tell you **anything** about **any** user in a database.

Issue: dataset/models learned from it will contain 0 information.

A new approach to privacy

Anonymization isn’t enough

Linkage attacks **will** happen

What should we mean when we say a dataset (or model) protects users’ privacy?

perfect privacy

No amount of analysis or linkage can tell you **anything** about **any** user in a database.

Issue: dataset/models learned from it will contain 0 information.

A new approach to privacy

~~**perfect privacy**~~ **differential privacy**

very much

No amount of analysis or linkage can tell you ~~**anything**~~ about **any** user in a database.

A new approach to privacy

Anonymization isn't enough
Linkage attacks **will** happen
What should we mean when we say a dataset (or model) protects users' privacy?

~~**perfect privacy**~~ **differential privacy**

very much

No amount of analysis or linkage can tell you ~~**anything**~~ about **any** user in a database.

Differential privacy

~~**perfect privacy**~~ **differential privacy**

very much

No amount of analysis or linkage can tell you ~~**anything**~~ about **any** user in a database.

Differential privacy

An algorithm A is ϵ-differentially private if,

~~perfect privacy~~ **differential privacy**

very much

No amount of analysis or linkage can tell you **~~anything~~** about **any** user in a database.

Differential privacy

An algorithm A is ϵ-differentially private if,

for every dataset D

~~perfect privacy~~ **differential privacy**

very much

No amount of analysis or linkage can tell you ~~**anything**~~ about **any** user in a database.

Differential privacy

An algorithm A is ϵ-differentially private if,

for every dataset D

for every user's data $x \in D$, and the dataset $D' = D \setminus \{x\}$

~~**perfect privacy**~~ **differential privacy**

very much

No amount of analysis or linkage can tell you ~~**anything**~~ about **any** user in a database.

Differential privacy

An algorithm A is ϵ-differentially private if,

for every dataset D

for every user's data $x \in D$, and the dataset $D' = D \setminus \{x\}$

for every output $\hat{f} \in \text{range}(A)$

~~perfect privacy~~ **differential privacy**

very much

No amount of analysis or linkage can tell you ~~**anything**~~ about **any** user in a database.

Differential privacy

An algorithm A is ϵ-differentially private if,

for every dataset D

for every user's data $x \in D$, and the dataset $D' = D \setminus \{x\}$

for every output $\hat{f} \in \text{range}(A)$

$$e^{-\epsilon} \geq \frac{\mathbb{P}[A(D)=\hat{f}]}{\mathbb{P}[A(D')=\hat{f}]} \leq e^{\epsilon}$$

~~**perfect privacy**~~ **differential privacy**

very much

No amount of analysis or linkage can tell you ~~**anything**~~ about **any** user in a database.

Differential privacy

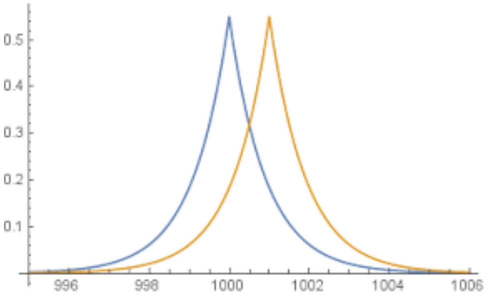
An algorithm A is ϵ-differentially private if,

for every dataset D

for every user's data $x \in D$, and the dataset $D' = D \setminus \{x\}$

for every output $\hat{f} \in \text{range}(A)$

$$e^{-\epsilon} \geq \frac{\mathbb{P}[A(D)=\hat{f}]}{\mathbb{P}[A(D')=\hat{f}]} \leq e^{\epsilon}$$



~~perfect privacy~~ **differential privacy**

very much

No amount of analysis or linkage can tell you ~~**anything**~~ about **any** user in a database.

Vignette 3: The Census

The History of Privacy for the Census