# Causal Inference

Ritik Dutta

January 27, 2020

IIT Gandhinagar

# An Example

- Correlation vs. Causality

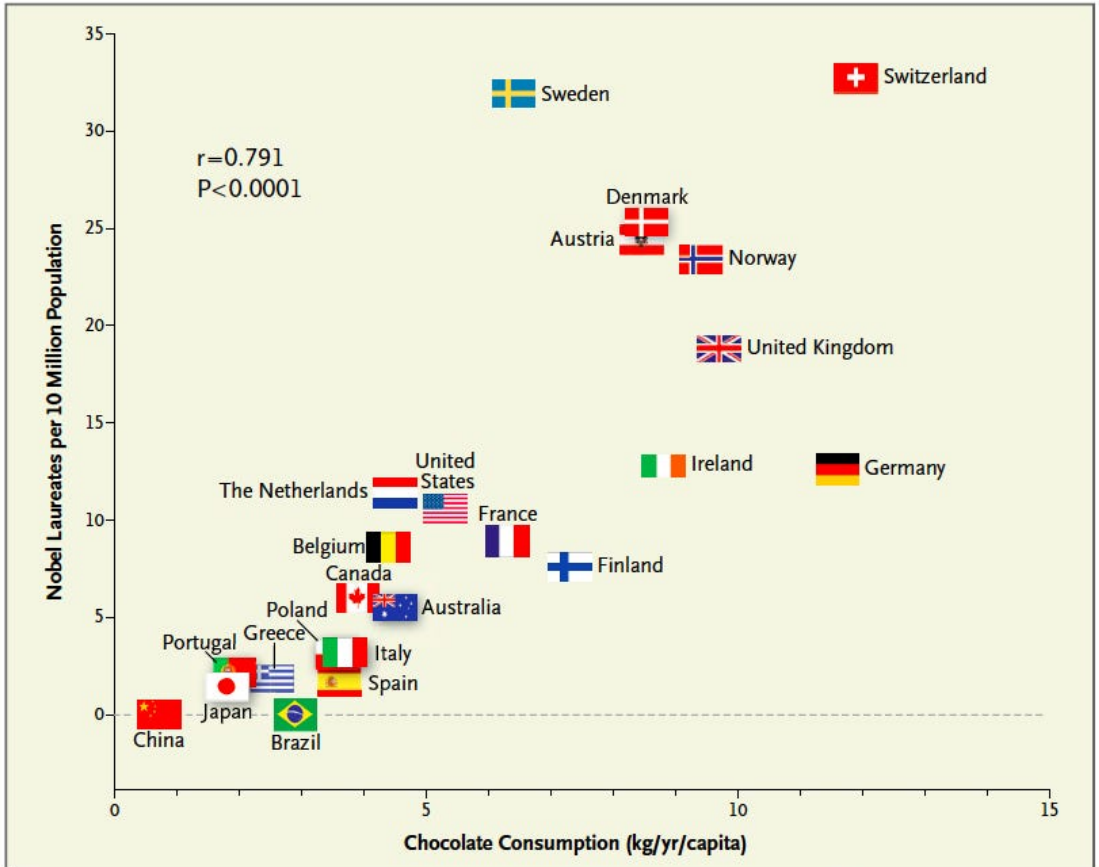



**Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

# Why is Causality Hard?

- No single definition
- No fail-proof method for finding it
- Observational data

# Why is Causality Hard?

There are different frameworks for causality -

- For time-series data: Granger causality
- Potential Outcomes / Counterfactuals framework
- Pearl's structural equation models - causal graph models
- Additive models, Dawid's decision-oriented approach, Information Geometry, ...

# Why is Causality Hard?

There are different frameworks for causality -

- For time-series data: Granger causality
- Potential Outcomes / Counterfactuals framework
- **Pearl's structural equation models - causal graph models**
- Additive models, Dawid's decision-oriented approach, Information Geometry, ...

# Independence of Random Variables

| S: Heavy smoker | C: Lung cancer before 60 |
|---|---|
| 0 | 0 |
| 1 | 1 |
| 0 | 1 |
| 1 .... | 1 .... |

# Independence of Random Variables

| S: Heavy smoker | C: Lung cancer before 60 |
|---|---|
| 0 | 0 |
| 1 | 1 |
| 0 | 1 |
| 1 .... | 1 .... |

Suppose correlation is 0, no causal link between the features, right?

# Independence of Random Variables

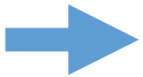
| S: Heavy smoker | C: Lung cancer before 60 |
|---|---|
| 0 | 0 |
| 1 | 1 |
| 0 | 1 |
| 1 .... | 1 .... |

Suppose correlation is 0, no causal link between the features, right?

**Uncorrelation doesn't imply independence!**

# Joint PDF and Independence

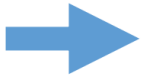
| S: Heavy smoker | C: Lung cancer before 60 |
|---|---|
| 0 | 0 |
| 1 | 1 |
| 0 | 1 |
| 1 .... | 1 .... |



| | S=0 | S=1 |
|---|---|---|
| C=0 | 30/100 | 10/100 |
| C=1 | 20/100 | 40/100 |

# Joint PDF and Independence

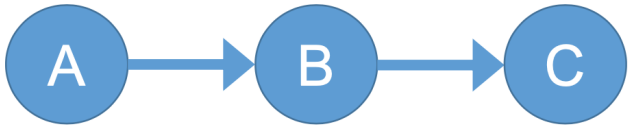
| S: Heavy smoker | C: Lung cancer before 60 |
|---|---|
| 0 | 0 |
| 1 | 1 |
| 0 | 1 |
| 1 .... | 1 .... |



| | S=0 | S=1 | |
|---|---|---|---|
| C=0 | 30/100 | 10/100 | 0.4 |
| C=1 | 20/100 | 40/100 | 0.6 |
| | 0.5 | 0.5 | |

# Directed Graphical Models

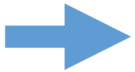
- Given data on A,B,C we can estimate the joint PDF $P(A, B, C)$
- See if it factorizes as $P(A, B, C) = P(A)P(B|A)P(C|B)$ i.e. has some conditional independencies.
- A directed graphical model describes all distributions that have a given set of conditional independencies.




| A | B | C |
|---|---|---|
| 0 | 1 | 0 |
| 1 | 1 | 1 |
| ... | ... | ... |

# Smoking Causes Cancer

| S: Heavy smoker | C: Lung cancer before 60 |
|---|---|
| 0 | 0 |
| 1 | 1 |
| 0 | 1 |
| 1 .... | 1 .... |



| | S=0 | S=1 |
|---|---|---|
| C=0 | 30/100 | 10/100 |
| C=1 | 20/100 | 40/100 |

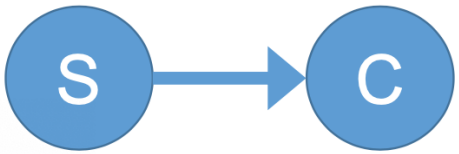
# Universe 1

| S=0 | 0.5 |
|---|---|
| S=1 | 0.5 |

Pr(S)

| | S=0 | S=1 |
|---|---|---|
| C=0 | 30/50 | 10/50 |
| C=1 | 20/50 | 40/50 |

Pr(C/S)




Pr(S,C)

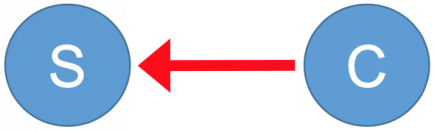
# Universe 2

| C=0 | 0.4 |
|---|---|
| C=1 | 0.6 |

Pr(C)

| | C=0 | C=1 |
|---|---|---|
| S=0 | 30/(100*0.4) = 0.75 | 20/(100*0.6) = 0.33 |
| S=1 | 10/(100*0.4) = 0.25 | 40/(100*0.6) = 0.66 |

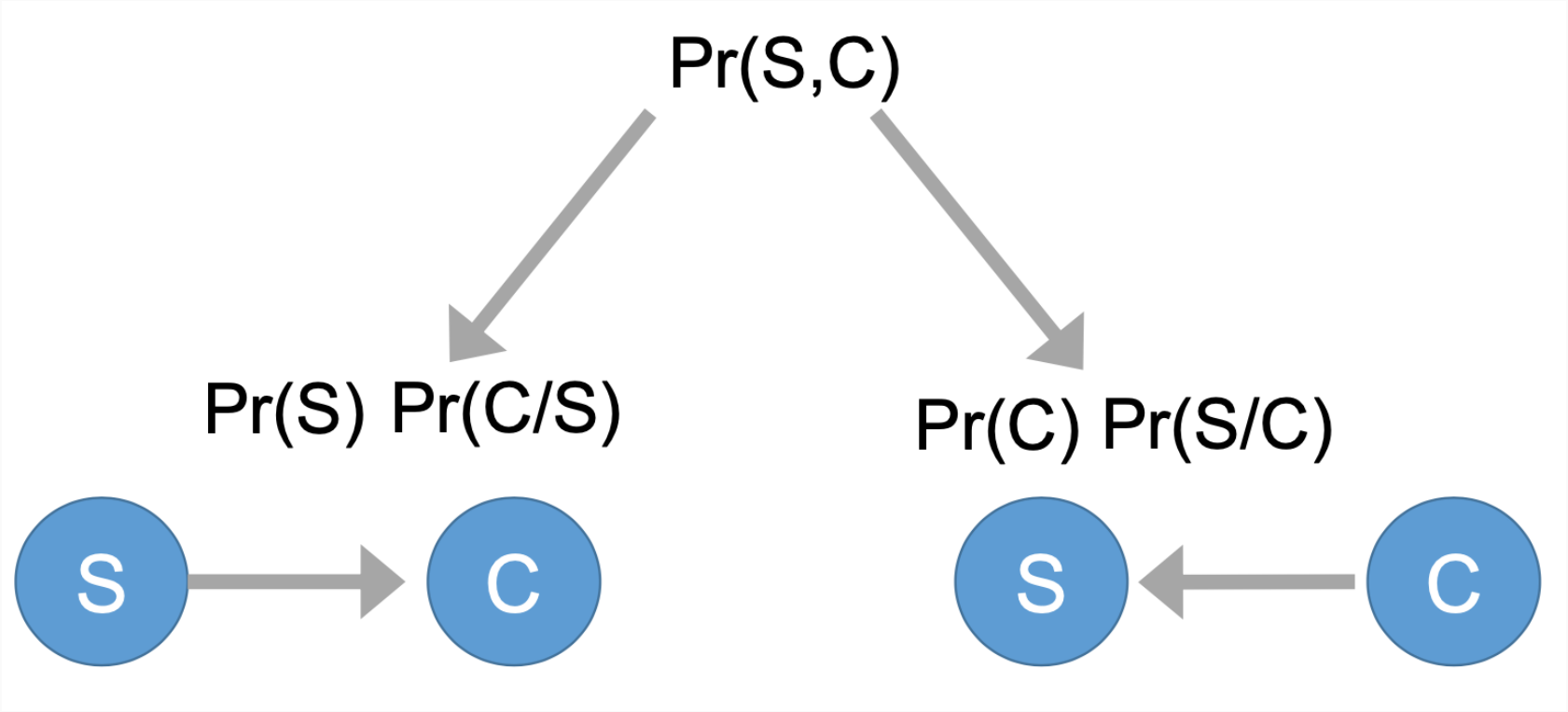
Pr(S/C)




Pr(S,C)

S=F(C,E)
E ⫫ C

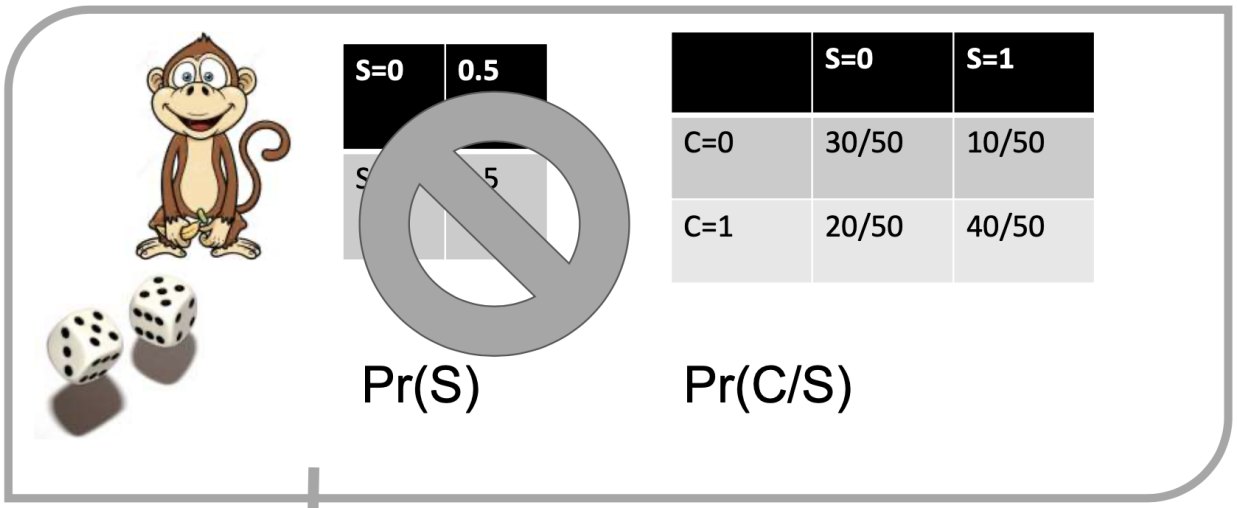
# How to find the Causal Direction?



# How to find the Causal Direction?

- Is it impossible to find true causal direction from observational data for two random variables.
- **Interventions** required to fill the gap

# Intervention - Force People to Smoke

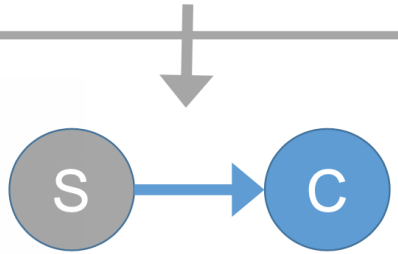


| S=0 | 0.5 |
|---|---|
| S | 5 |

Pr(S)

| | S=0 | S=1 |
|---|---|---|
| C=0 | 30/50 | 10/50 |
| C=1 | 20/50 | 40/50 |

Pr(C/S)




- Flip coin and force each person to smoke or not, with prob ½.
- In Universe1 (i.e. Under S→C) ,
- new joint pdf stays same as before intervention.

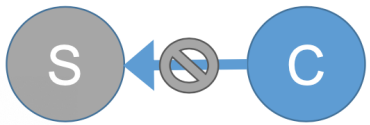
# Intervention - Force People to Smoke

| C=0 | 0.4 |
|---|---|
| C=1 | 0.6 |

Pr(C)

| | C=0 | C=1 |
|---|---|---|
| S=0 | ) = 0.75 | 20/(100*0.6) = 0.33 |
| S=1 | 0*0 = 0.25 | 40/(100*0.6) = 0.66 |

Pr(S/C)

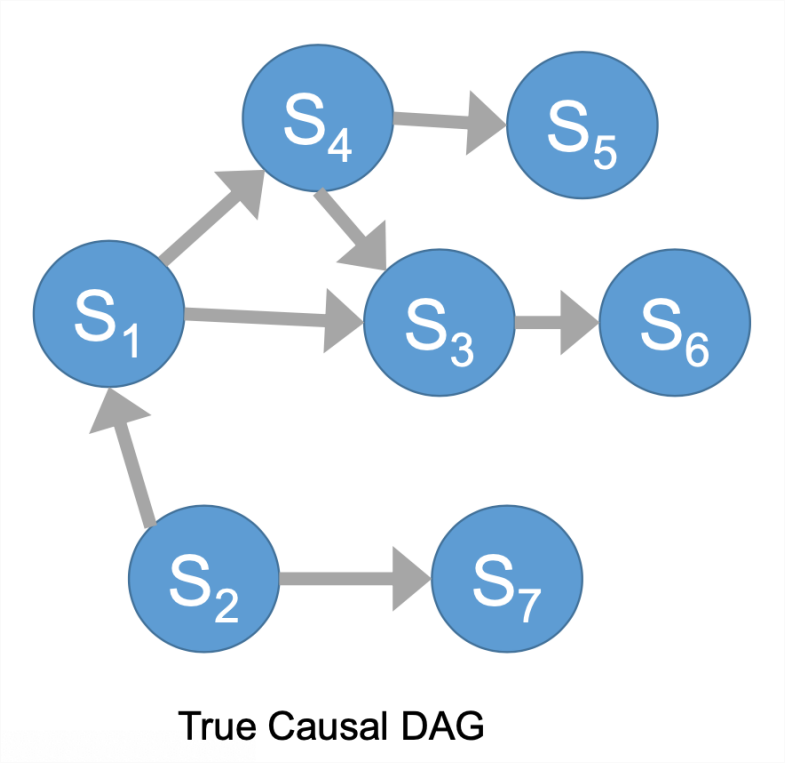



- Flip coin and force each person to smoke or not, with prob ½.
- In Universe 2 (Under C→S)
- S, C will become independent after intervention.
- So check correlation on data after intervention and find true direction!

# Interventions Are Not Always Possible

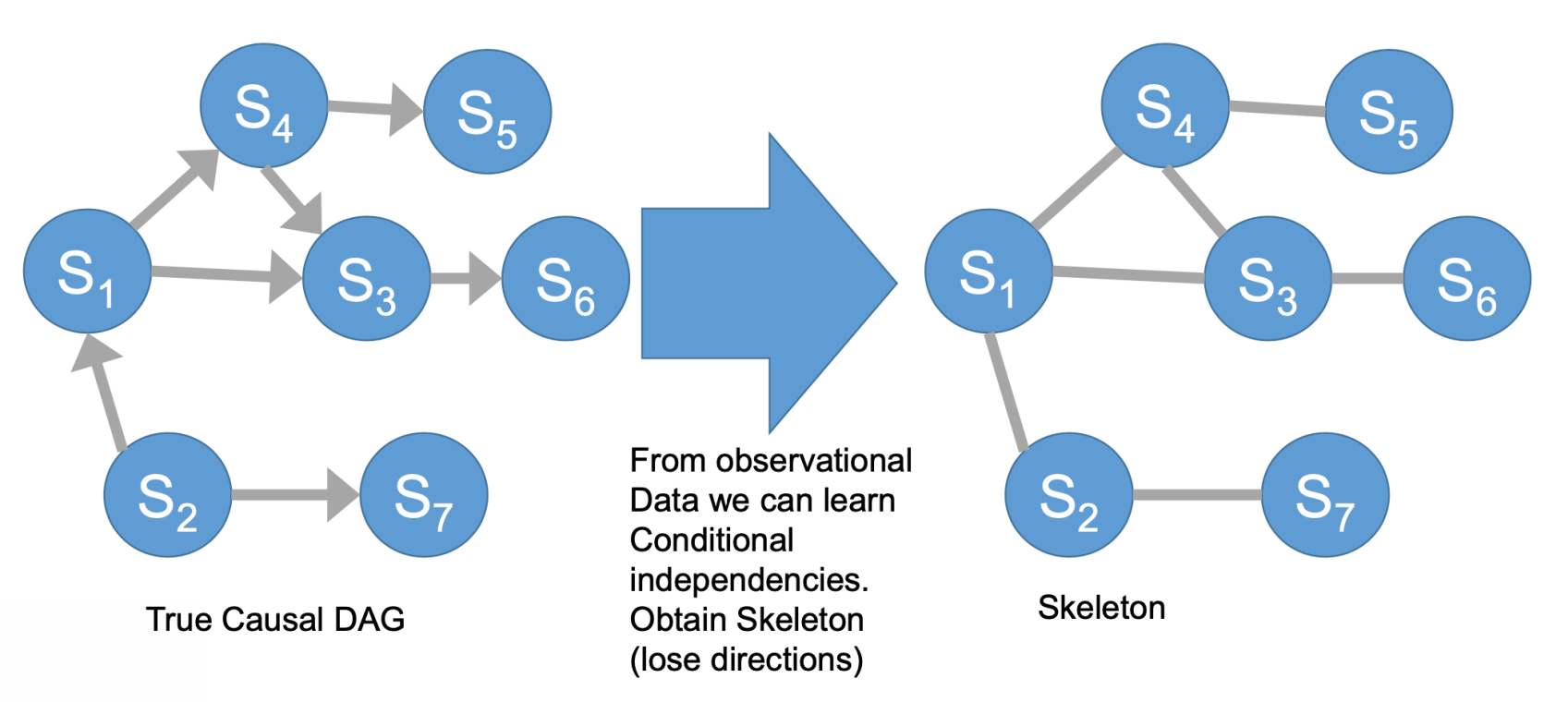
"You are giving dying people sugar pills?"

# More Variables




True Causal DAG

# More Variables

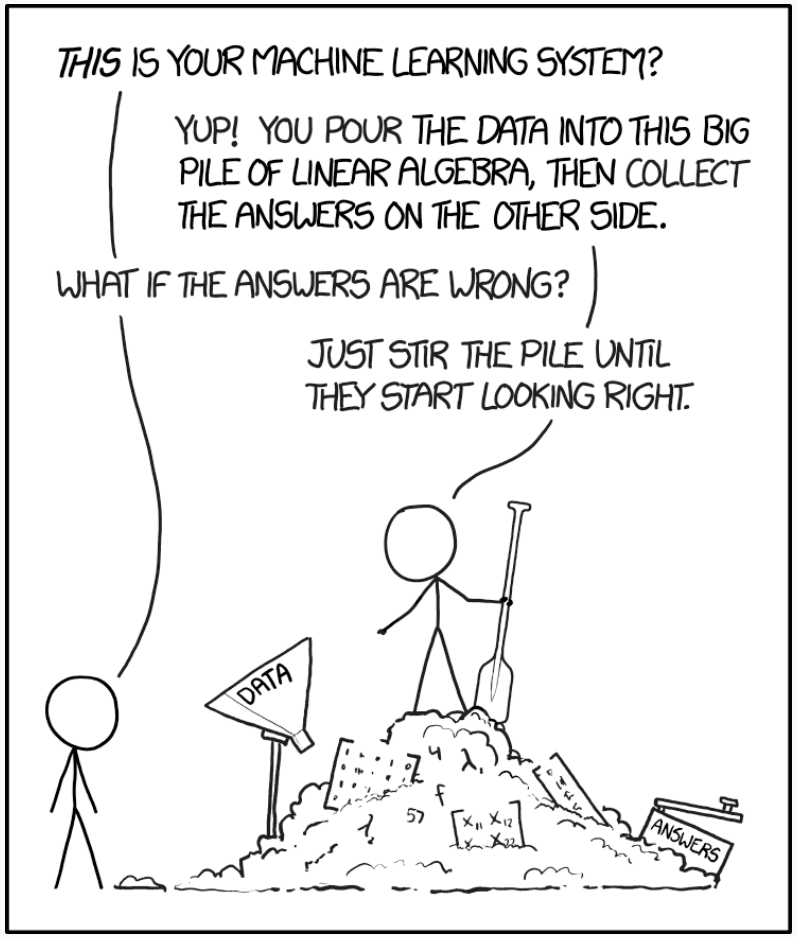



$$(A \perp\!\!\!\perp B) \mid C \quad \iff \quad \Pr(A \cap B \mid C) = \Pr(A \mid C) \Pr(B \mid C)$$

# Next Steps?

- PC Algorithm
- NCC Algorithm
- SAM
- ...

# ML in Practice




https://xkcd.com/1838/

# Racist Robots in the News

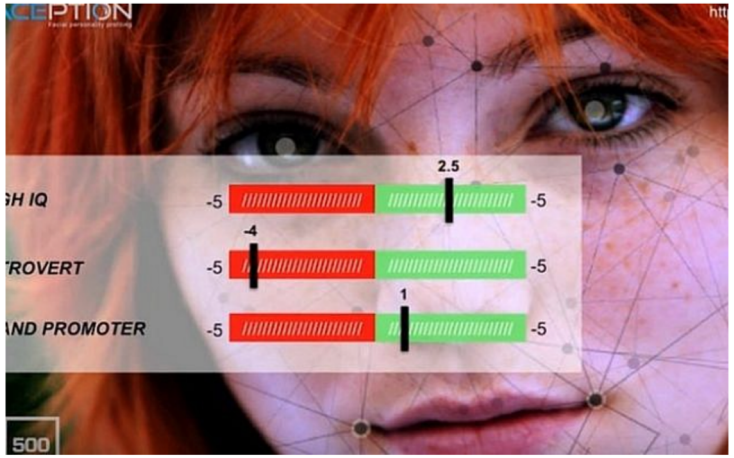
FACEPTION 'CAN MATCH AN INDIVIDUAL WITH VARIOUS PERSONALITY TRAITS AND TYPES WITH A HIGH LEVEL OF ACCURACY'

## New Israeli facial imaging claims to identify terrorists and pedophiles

Tel Aviv start-up Faception says its face 'classifiers' can spot criminals and even great poker players in a split second, but the experts are not convinced

By SUE SURKES
24 May 2016, 10:52 pm | 9

3K shares




An image taken from a May 2016 presentation by Faception co-founder Shai Gilboa (screen capture: YouTube)

A Tel-Aviv based start-up company says it has developed a program to identify personality types such as terrorists, pedophiles, white collar offenders and even great poker players from facial analysis that takes just a fraction of a second.

OPINION | TECH

## 'Gaydar' Shows How Creepy Algorithms Can Get

Imagine what an oppressive government could do with it.

By Cathy O'Neil
409 September 25, 2017, 6:30 AM EDT

Watch out. *Photographer: Jin Lee/Bloomberg*

Artificial intelligence keeps getting creepier. In one controversial study, researchers at Stanford University have demonstrated that facial recognition technology can identify gay people with surprising precision, although many caveats apply. Imagine how that could be used in the many countries where homosexuality is a criminal offense.

Nikon S630


GOOGLE

## Google Photos Mistakenly Labels Black People 'Gorillas'

BY CONOR DOUGHERTY JULY 1, 2015 7:01 PM 41

Email
Share
Tweet
Save
More

Google continued to apologize Wednesday for a flaw in Google Photos, which was released to great fanfare in May, that led the new application to mistakenly label photos of black people as "gorillas."

The company said it had fixed the problem and was working to figure out exactly how it happened.

"We're appalled and genuinely sorry that this happened," said a Google representative in an emailed statement. "We are taking immediate action to prevent this type of result from appearing."

From self-driving cars to photos, Google, like every technology company, is constantly releasing cutting-edge technologies with the understanding that problems will arise and that it will have to fix them as it goes. The idea is that you never know what problems might arise until you get the technologies in the hands of real-world users.

In the case of the Google Photos app — which uses a combination of advanced computer vision and machine learning techniques to help users collect, search and categorize photos — errors are easy to spot. When the app was unveiled at the company's annual developer show, executives went through carefully staged demonstrations to show how it can recognize landmarks like the Eiffel Tower and give users the ability to search their photos for people, places or things — even things as specific as a particular dog breed.

# Facial Recognition Software Is Bad At Identifying Darker Skinned People

**Computing**
The Observer

Sun 28 May 2017 08.27 EDT

3,817 | 498

**Interview**

## 'A white mask worked better': why algorithms are not colour blind

*By Ian Tucker*

**When Joy Buolamwini found that a robot recognised her face better when she wore a white mask, she knew a problem needed fixing**

▲ Joy Buolamwini gives her TED talk on the bias of algorithms Photograph: TED

Joy Buolamwini is a graduate researcher at the MIT Media Lab and founder of the Algorithmic Justice League - an organisation that aims to challenge the biases in decision-making software. She grew up in Mississippi, gained a Rhodes scholarship, and she is also a Fulbright fellow, an Astronaut scholar and a Google Anita Borg scholar. Earlier this year she won a $50,000 scholarship funded by the makers of the film *Hidden Figures* for her work fighting coded discrimination.

# ProPublica's Study of NorthPointe Software

PRO PUBLICA

Donate

## Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

*by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica*
May 23, 2016

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

Just as the 18-year-old girls were realizing they were too big for the tiny conveyances — which belonged to a 6-year-old boy — a woman came running after them saying, "That's my kid's stuff." Borden and her friend immediately dropped the bike and scooter and walked away.

But it was too late — a neighbor who witnessed the heist had already called the police. Borden and her friend were arrested and charged with burglary and petty theft for the items, which were valued at a total of $80.

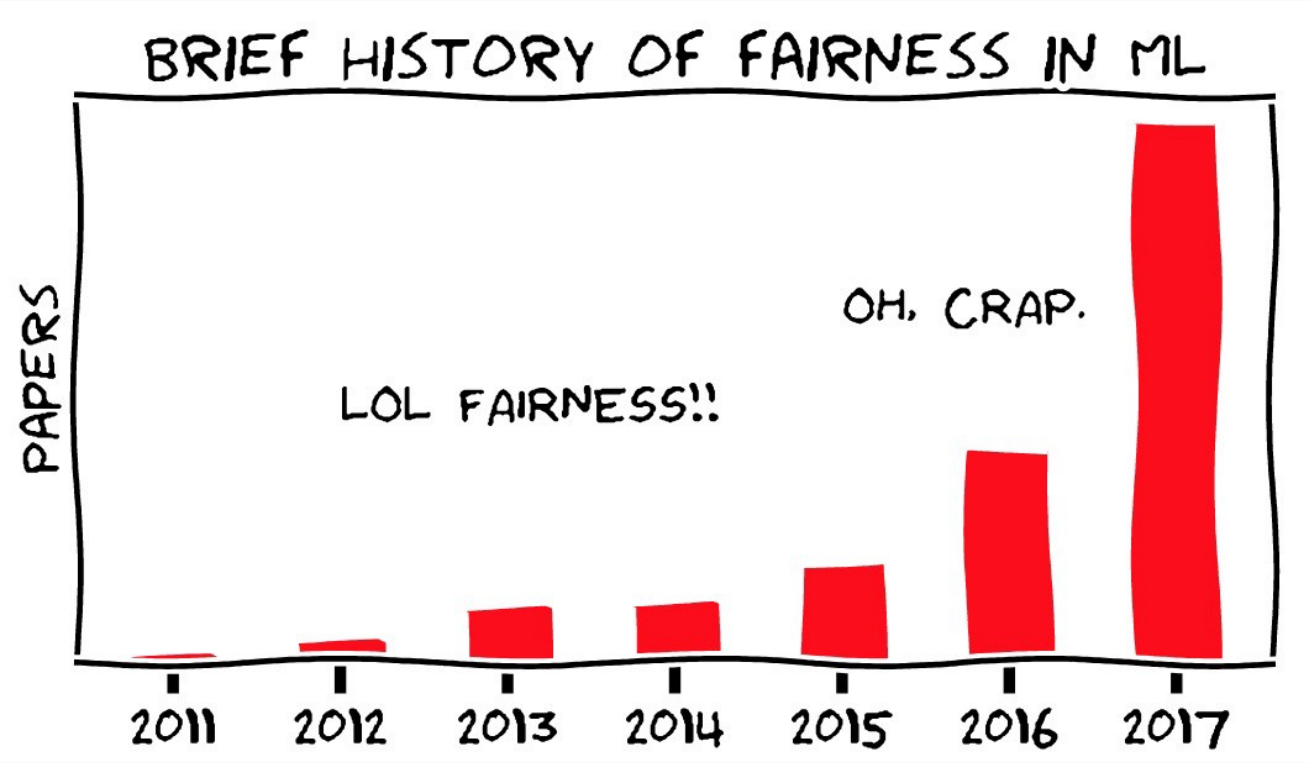
Compare their crime with a similar one: The previous summer, 41-year-old Vernon Prater was picked up for shoplifting $86.35 worth of tools from a nearby Home Depot store.

### Prediction Fails Differently for Black Defendants

| | WHITE | AFRICAN AMERICAN |
|---|---|---|
| Labeled Higher Risk, But Didn't Re-Offend | 23.5% | 44.9% |
| Labeled Lower Risk, Yet Did Re-Offend | 47.7% | 28.0% |

***Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes.*** *(Source: ProPublica analysis of data from Broward County, Fla.)*

# Fairness


BRIEF HISTORY OF FAIRNESS IN ML
PAPERS
LOL FAIRNESS!!
OH. CRAP.
2011
2012
2013
2014
2015
2016
2017

# Bias in Computer Systems

The primary sources of bias -

- Pre-existing bias - from social institutions, practices, and attitudes
- Technical bias - from technical constraints or considerations
- Emergent bias - from context of use

# How to define Fairness?

- Probabilistically
- Lots of fairness definitions
    - Decisions should be in some sense probabilistically independent of sensitive feature values (such as gender, race)

# Lot of Parity Definitions

(Probabilistic definitions of different kinds of fairness)

- Demographic parity
- Accuracy parity
- True positive parity
- False positive parity
- Positive rate parity
- Precision parity
- Positive predictive value parity
- Negative predictive value parity
- Predictive value parity
- ...

# Lot of Parity Definitions

(Probabilistic definitions of different kinds of fairness)

- Demographic parity: The output of the classifier does not depend on the sensitive attribute (e.g., gender, race, education level, etc.) $P(C|A = 0) = P(C|A = 1)$
- Accuracy parity
- True positive parity
- False positive parity
- Positive rate parity
- Precision parity
- Predictive value parity
- ...

# Lot of Parity Definitions

(Probabilistic definitions of different kinds of fairness)

- Demographic parity
- Accuracy parity: The accuracy of the classifier does not depend on the sensitive attribute (e.g., gender, race, education level, etc.) $P(C = Y|A = 0) = P(C = Y|A = 1)$
- True positive parity
- False positive parity
- Positive rate parity
- Precision parity
- Predictive value parity
- ...

# Some Definitions

- X contains features of an individual (e.g., medical records)
- A is a sensitive attribute (e.g., race, gender, . . . )
- Y is the true outcome (the ground truth, e.g., whether a patient actually has cancer)
- C is the machine learning algorithm that uses X and A to predict the value of Y (e.g., predict whether the patient has cancer)

# Some Simplifying Assumptions

- The sensitive attribute A divides the population into two groups *a* (e.g., whites) and *b* (e.g., non-whites)
- The machine learning algorithm C outputs 0 (e.g., predicts not cancer) or 1 (e.g., predicts cancer)
- The true outcome Y is 0 (e.g., not cancer) or 1 (e.g., cancer)

# Impossibility Results

- Assume differing base rates - i.e., $Pr_a(Y = 1) \neq Pr_b(Y = 1)$– and an imperfect machine learning algorithm ($C \neq Y$), then you can not simultaneously achieve
  a) Precision parity: $Pr_a(Y = 1|C = 1) = Pr_b(Y = 1|C = 1)$
  b) True positive parity: $Pr_a(Y = 1|C = 1) = Pr_b(Y = 1|C = 1)$
  c) False positive parity: $Pr_a(C = 1|Y = 0) = Pr_b(C = 1|Y = 0)$

# An Algorithm

Train an algorithm on the X variables to predict Y - Fairness through Unawareness

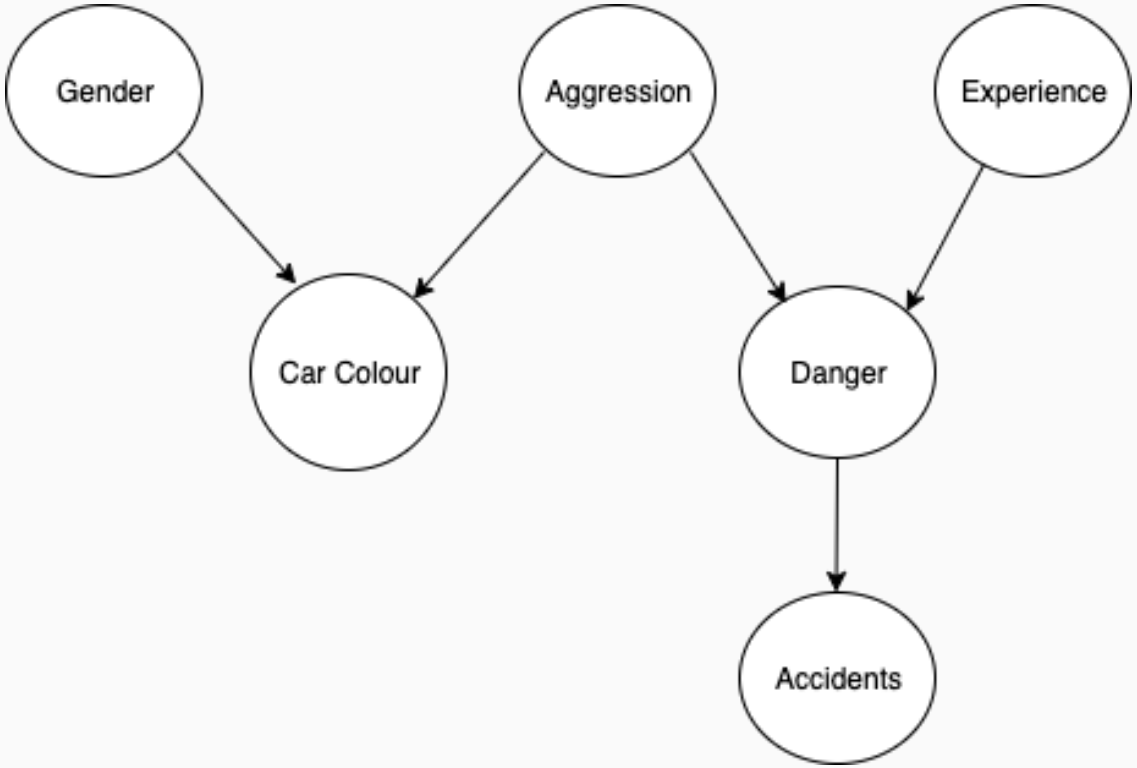
# An Algorithm

Train an algorithm on the X variables to predict Y - Fairness through Unawareness - **Features in X could be correlated with A!**

# Counterfactual Fairness

- A causality-based notion of fairness
- Uses causal graph to determine which attributes are affected when the sensitive attribute is changed
- Attributes affected by the sensitive attribute should be excluded by the learning algorithm to make the model counterfactually fair

# Counterfactual Fairness



# Further Reading

- Counterfactual Fairness Lecture
- Fair ML UC Berkeley Course
- Fair ML Book

# References


1. Causal Inference: A Friendly Introduction, Alex Dimakis, UT Austin
2. Counterfactual Fairness, Matt Kusner
3. A Tutorial on Fairness in Machine Learning
4. CS 294: Fairness in Machine Learning, UC Berkeley