Week 3

MORPH Algorithmic Fairness

Catherine Yeo

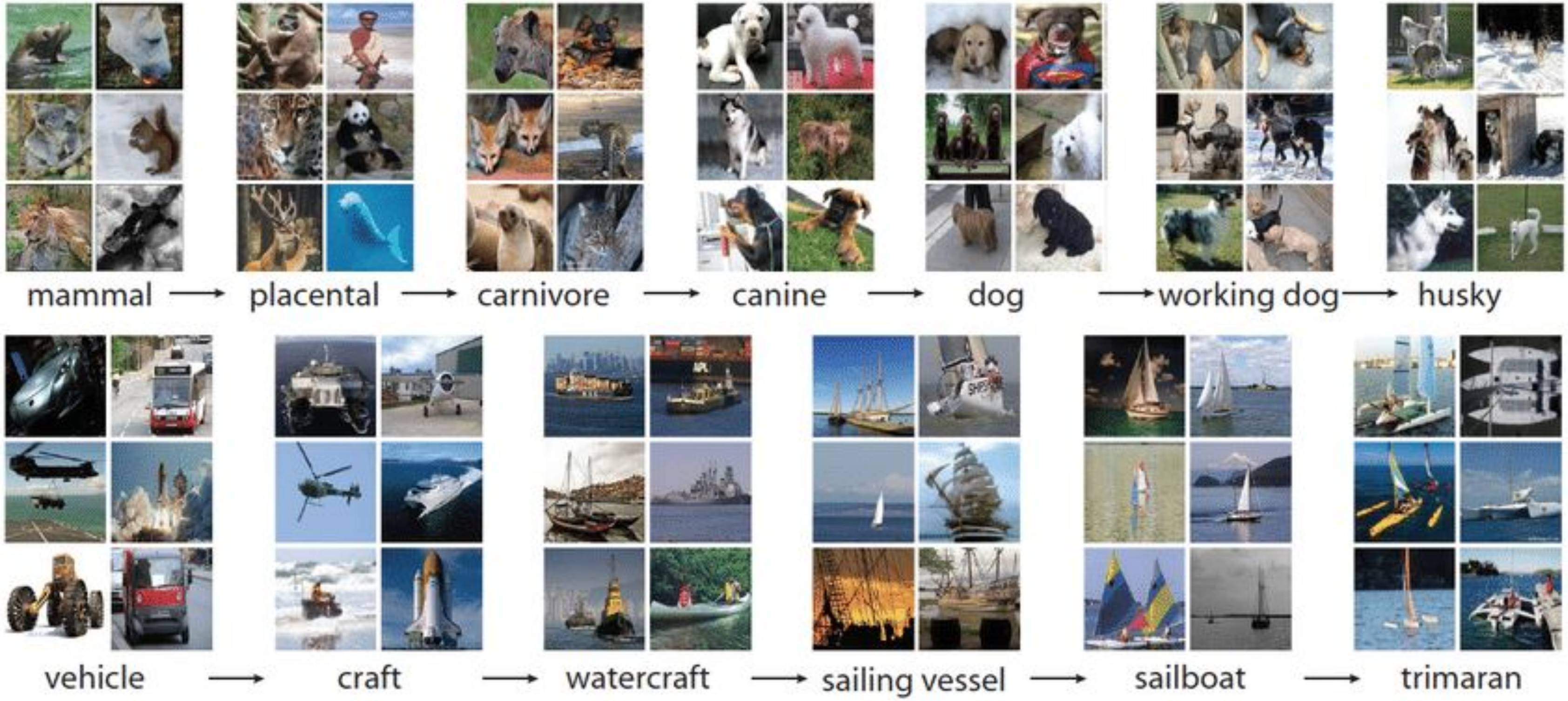
Agenda:
Causality & Counterfactuals

- Feedback
- Interest topic presentations!
- This week in fairness
- Causality
- Counterfactual fairness
- Looking ahead to week 4

Interest Topic Presentations

Connecting any interest of yours to fairness/bias

This Week in Fairness



ImageNet (2009) – 14 million hand-labelled images

QUARTZ

IT'S NOT ABOUT THE ALGORITHM

The data that transformed AI research—and possibly the world

July 26, 2017

By **Dave Gershgorn**
Artificial intelligence reporter

Stanford professor and Google Cloud chief scientist Fei-Fei Li changed everything.

FROM OUR OBSESSION
Machines with Brains
AI is upending companies, industries, and humanity.

ImageNet changed the landscape of DATA. And data -> ML.

Large image datasets: A Pyrrhic win for computer vision?

Vinay Uday Prabhu*
UnifyID AI Labs
vinay@unify.id

Abeba Birhane*
School of Computer Science and Informatics
University College Dublin
abeba.birhane@ucdconnect.ie

July 1, 2020

Abstract

In this paper we investigate problematic practices and consequences of large scale vision datasets. We examine broad issues such as the question of consent and justice as well as specific concerns such as the inclusion of verifiably pornographic images in datasets. Taking the ImageNet-ILSVRC-2012

New paper this week: "Large Image Datasets: A Pyrrhic Win for Computer Vision?"

Large scale image datasets have issues we must aim to mitigate and address in future dataset curation processes.

1) Lack of Consent

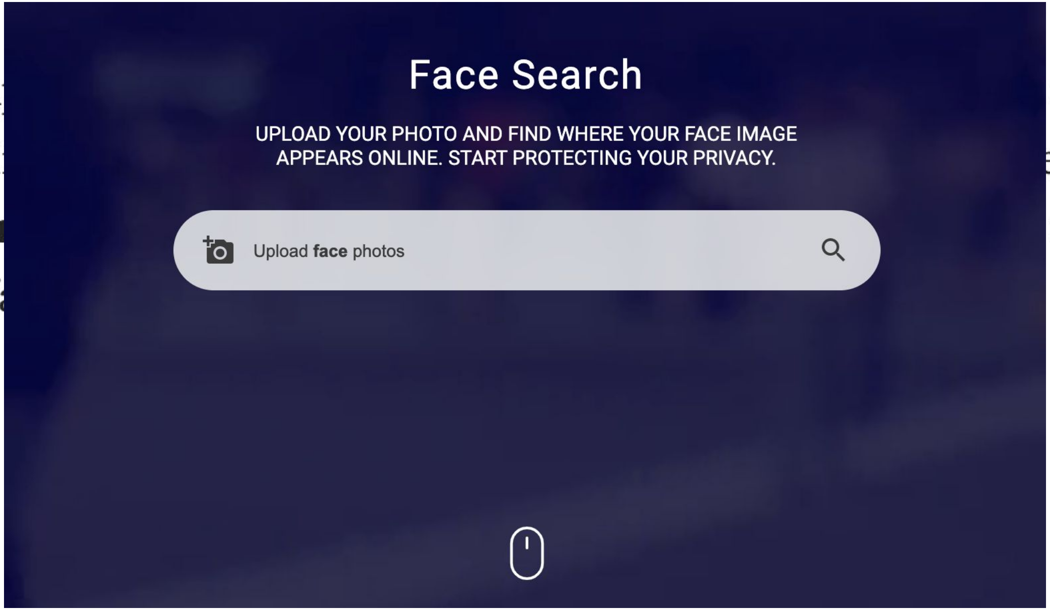
Many of these large-scale datasets freely gather photos, including photos of real people, **without consideration of consent**. In the Open Images V4–5–6 dataset, Prabhu and Birhane found “verifiably non-consensual images” of *children* taken from photo sharing community Flickr.

Photographers don’t upload your photos for the whole world to see without your consent, so why shouldn’t image datasets account for consent?

2) Loss of Privacy

When ImageNet was published, reverse image search did not exist. Now, image scraping tools are widespread, and powerful reverse image search engines (e.g. Google Image Search, PimEyes) allow anyone to be able to uncover real identities of humans/faces in a large image dataset.

With a sim name,
social med er data
points **we r** to give
away our f



3) Perpetuation of Harmful Stereotypes

How a dataset is labelled and curated could lead to us perpetuating what/who is perceived as “desirable”, “normal”, and “acceptable”. Individuals and groups on the margins would then be perceived as “outliers”.

For example, MIT’s 80 Million Tiny Images dataset contains harmful slurs, potentially labeling women as “whores” or “bitches” and minority racial groups with offensive language.

Once trained on biased data, machine learning algorithms can **not only normalize but *amplify* stereotypes**.

{* **ARTIFICIAL INTELLIGENCE** *}

MIT apologizes, permanently pulls offline huge dataset that taught AI systems to use racist, misogynistic slurs

Top uni takes action after *El Reg* highlights concerns by academics

Wed 1 Jul 2020 // 10:55 UTC

88 GOT TIPS?

Katyanna Quach BIO EMAIL TWITTER

SHARE

Special report MIT has taken offline its highly cited dataset that trained AI systems to potentially describe people using racist, misogynistic, and other problematic terms.

The database was removed this week after *The Register* alerted the American super-college. MIT also urged researchers and developers to stop using the training library, and to delete any copies. "We sincerely apologize," a professor told us.

The training set, built by the university, has been used to teach machine-learning models to automatically identify and list the people and objects depicted in still images. For example, if you show one of these systems a photo of a park, it might tell you about the children, adults, pets, picnic spreads, grass, and trees present in the snap. Thanks to MIT's cavalier

// **MOST READ**


MIT apologizes, permanently pulls offline huge dataset that taught AI systems to use racist, misogynistic slurs


I was screwed over by Cisco managers who enforced India's caste hierarchy on me in US HQ, claims engineer


Linux Mint 20 isn't exactly bursting with freshness but, hey, there's kernel 5.4 and it's a long-term support release

Your Thoughts?

- Immediate reactions?
- Should there be regulation of dataset curation?
- If yes, how would it best be done? How can we enforce consent was shared for every image?
- If no, how would you prevent such issues from happening?
- Other thoughts / comments?

If you’re interested in reading more:

Paper:

https://arxiv.org/pdf/2006.16923.pdf

News:

https://www.theregister.com/2020/07/01/mit_dataset_removed/

My Blog Post Summary:

https://medium.com/fair-bytes/we-need-to-change-how-image-datasets-are-curated-b325642394df

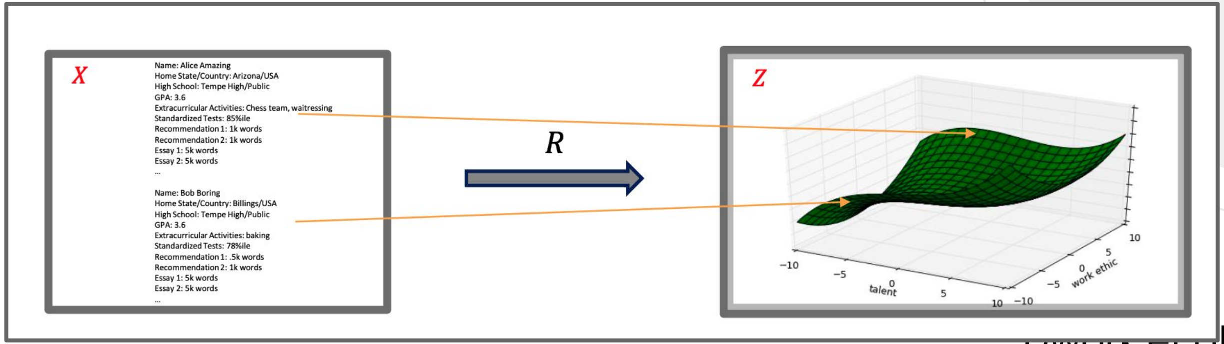
Causality

Key idea in Individual Fairness:

Similar individuals should be treated similarly.

Parameters

- Universe
- Outcome space
- A **representation mapping** between them:
  - Map from individuals to distributions over outcomes



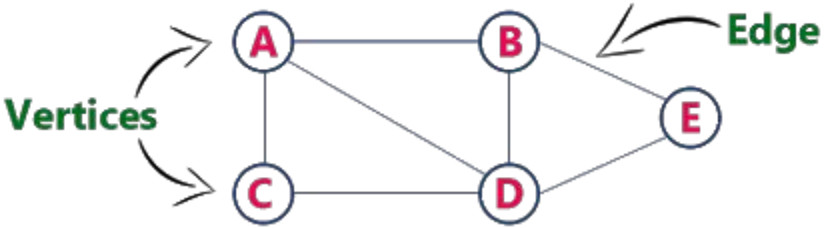
Dwork et al. (2006)

What if we know more about our data?

Causality

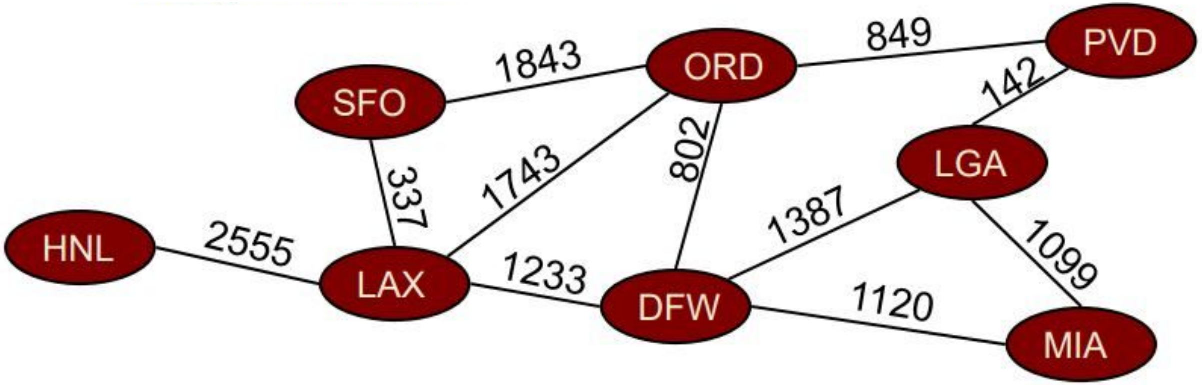
- One thing causes another thing
- We **observe** causes and effects
- Causal inference: the process of drawing a conclusion about a causal connection based on the conditions of the occurrence of an effect
  - Hard to do, because given data, we can't directly show causality — only show correlation
  - So we try to model causality

Graph

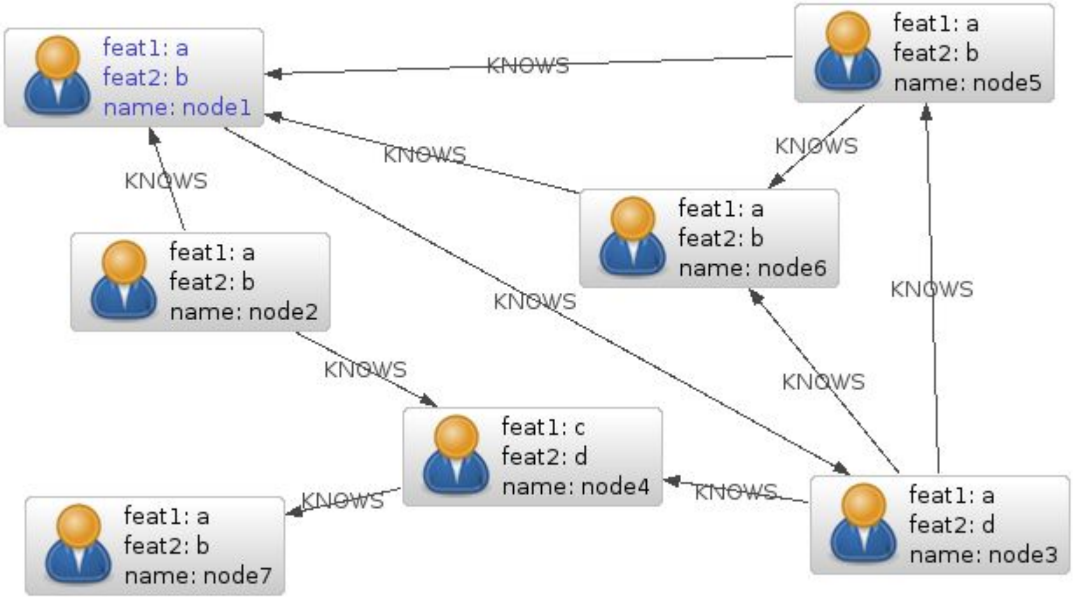


- A graph is a data structure that has:
  - **Nodes/Vertices**
  - **Edges**
- Can be used to represent SO many things:
  - Social network
  - Airport flight tracker
  - A maze
  - Things with different possibilities

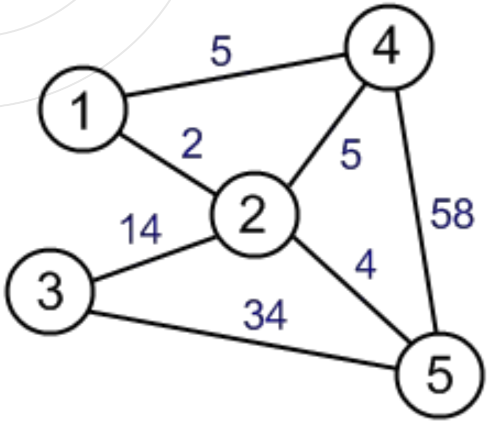
Graph: Example



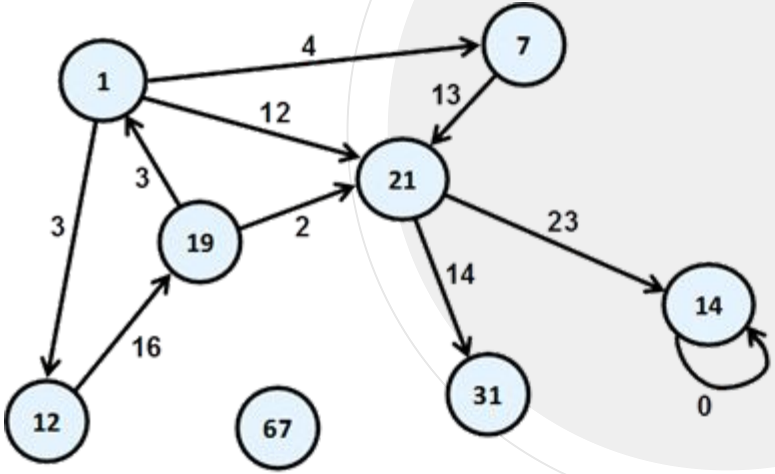
Graph: Example



Graph: Directions

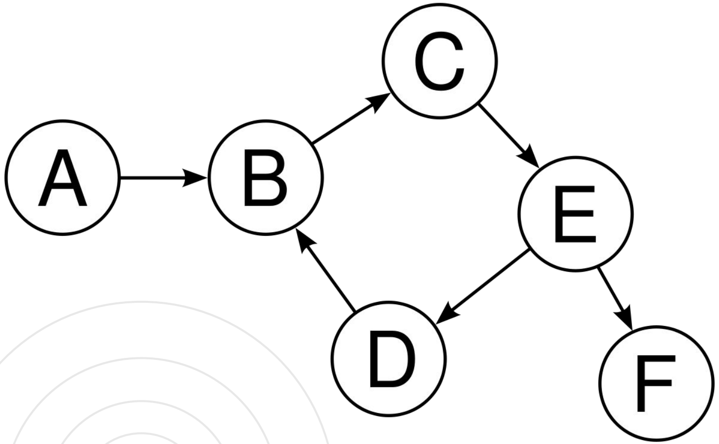


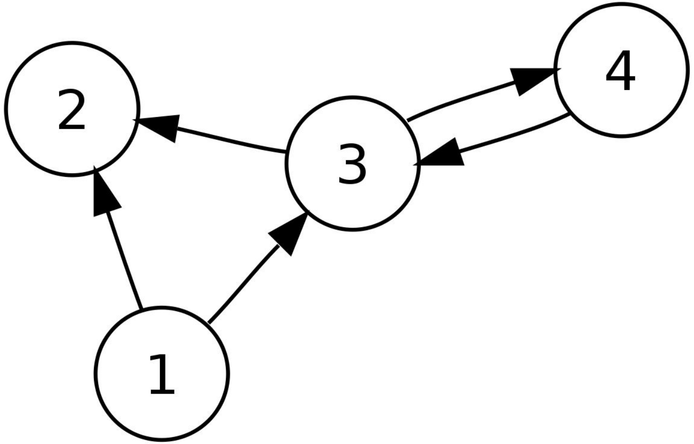
Undirected Graph



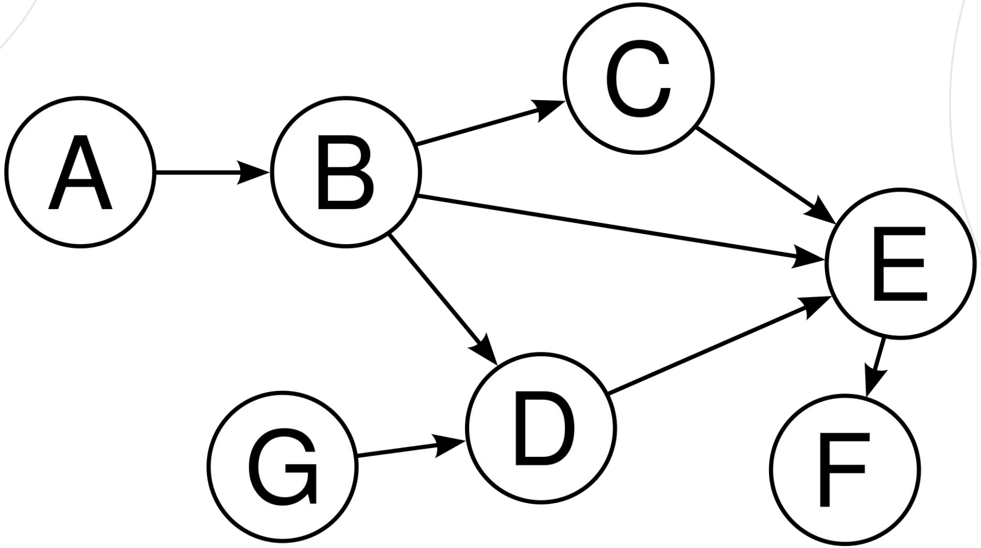
Directed Graph

Graph: Cycles

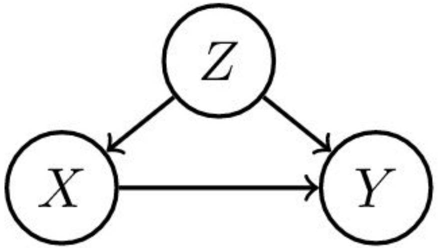




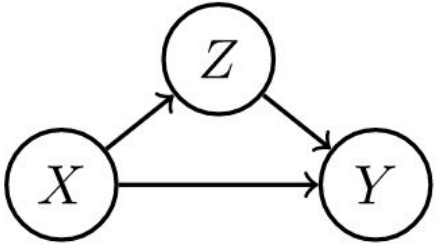
Directed Acyclic Graph (DAG)



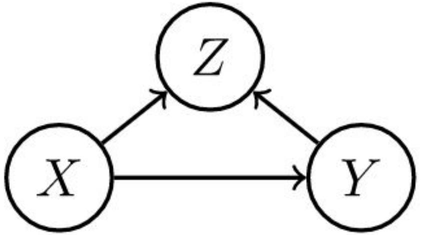
Causal Graphs



Fork



Mediator



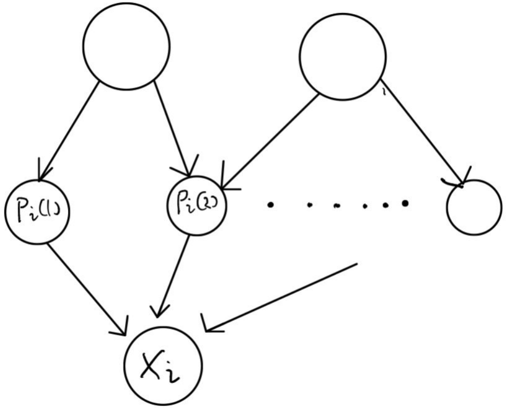
Collider

Causal Model

- A causal model is a triple (U, V, F) of sets:
  - U = set of background variables (factors not caused by anything in V)
  - V = set of observable variables
  - F = set of functions such that $V_i = f_i(pa_i, Upa_i)$
    - V_i is an element of V
    - pa_i (parent of V_i) is in the set V excluding V_i: you can think of it as the graph vertex pointing to V_i

Causal Model

- We represent causal models as a directed acyclic graph: one observable variable causes the next, and so on



Counterfactual Fairness

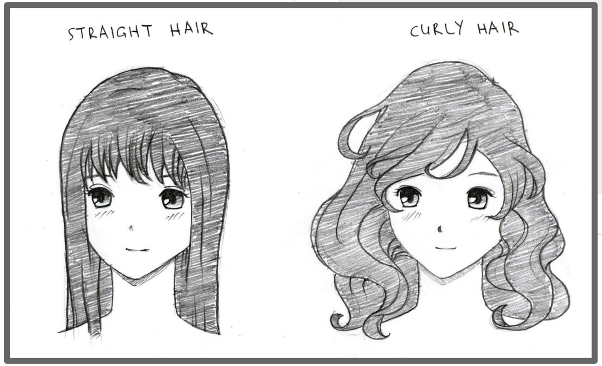
What is a Counterfactual?

- The word “counterfactual” refers to statements or situations that did not happen
  - “If I had arrived there on time…”
  - “If I had bought that instead…”
- For an individual, their **counterfactual** is the same individual in a world with its **sensitive attribute changed**

A machine learning model is fair under counterfactual fairness if it produces the *same prediction* for both an **individual** and its **counterfactual**.

Bank Loan Example

- Suppose our model predicts: will I receive a bank loan or not? (Yes/No)
- Let us choose the sensitive attribute here to be whether a person has curly hair or not, and keep other features the same (or similar)



Bank Loan Example

- Then, a counterfactually fair model would produce the same decision for a person with curly hair and for a person with straight hair
- = they both receive the loan OR neither does

Counterfactual Fairness

Definition 5 (Counterfactual fairness). *Predictor* $\hat{Y}$ *is* **counterfactually fair** *if under any context* $X = x$ *and* $A = a$,

$$P(\hat{Y}_{A\leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A\leftarrow a'}(U) = y \mid X = x, A = a), \tag{1}$$

for all y *and for any value* a' *attainable by* A.

Kusner et al. (2017)

Counterfactual Fairness

Definition 5 (Counterfactual fairness). *Predictor* $\hat{Y}$ *is* **counterfactually fair** *if under any context* $X = x$ *and* $A = a$,

$$P(\boxed{\hat{Y}_{A \leftarrow a}(U) = y} \mid X = x, A = a) = P(\hat{Y}_{A \leftarrow a'}(U) = y \mid X = x, A = a), \qquad (1)$$

for all y *and for any value* a' *attainable by* A.

The decision output (prediction) we want

Kusner et al. (2017)

Counterfactual Fairness

Definition 5 (Counterfactual fairness). *Predictor* $\hat{Y}$ *is* **counterfactually fair** *if under any context* $X = x$ *and* $A = a$,

$$P(\hat{Y}_{A \leftarrow a}(U) = y \mid \boxed{X = x,} A = a) = P(\hat{Y}_{A \leftarrow a'}(U) = y \mid X = x, A = a), \qquad (1)$$

for all y *and for any value* a' *attainable by* A.

Non-sensitive attributes

Kusner et al. (2017)

Counterfactual Fairness

Definition 5 (Counterfactual fairness). *Predictor* $\hat{Y}$ *is* **counterfactually fair** *if under any context* $X = x$ *and* $A = a$,

$$P(\hat{Y}_{A \leftarrow a}(U) = y \mid X = x, \boxed{A = a}) = P(\hat{Y}_{A \leftarrow a'}(U) = y \mid X = x, A = a), \tag{1}$$

for all y *and for any value* a' *attainable by* A.

Sensitive attribute

Kusner et al. (2017)

Counterfactual Fairness

Definition 5 (Counterfactual fairness). *Predictor $\hat{Y}$ is* **counterfactually fair** *if under any context $X = x$ and $A = a$,*

$$P(\hat{Y}_{A\leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A\leftarrow a'}(U) = y \mid X = x, A = a), \qquad (1)$$

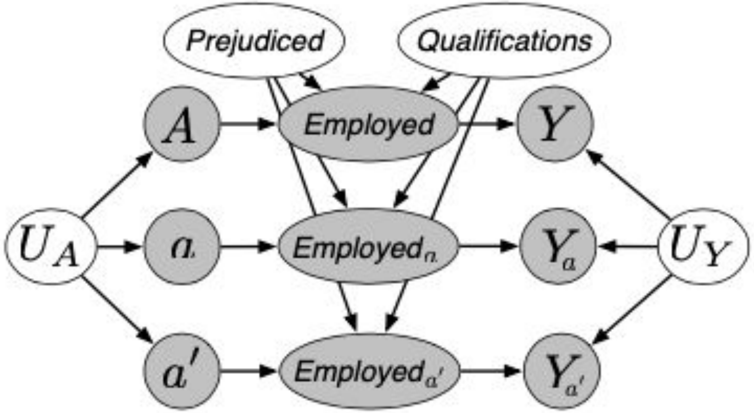
for all y and for any value a' attainable by A.

When we choose sensitive attribute to be *a*

When we choose sensitive attribute to be *a'*

Kusner et al. (2017)

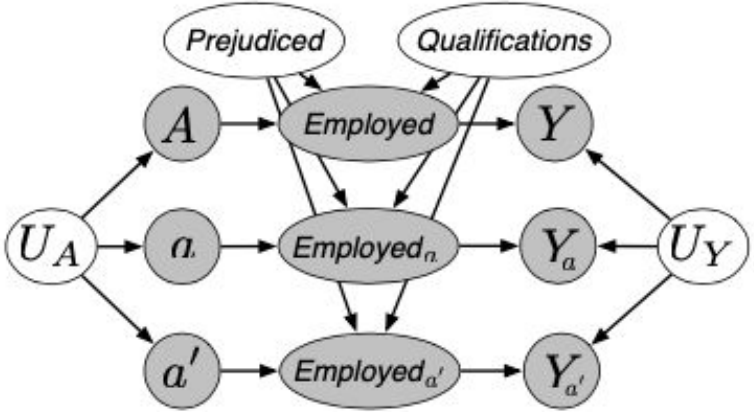
Causal Model



White bubbles = background variables

Kusner et al. (2017)

Causal Model



Gray bubbles = protected attributes (observed variables) -> outcomes

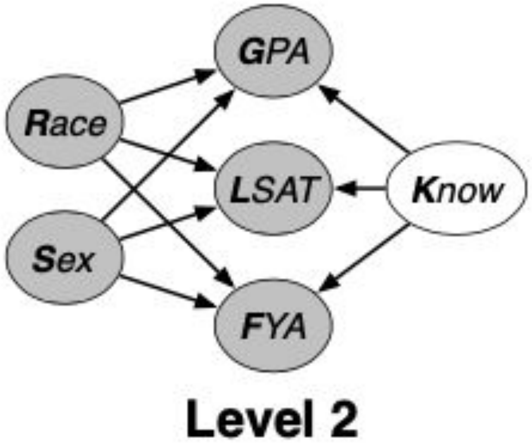
Kusner et al. (2017)

Experiment: Law School Success

- Problem: predict if an applicant will have a high First Year Average (FYA), but ensure it's not biased by race and sex
- LSAT test score, GPA, and FYA scores may all be biased due to social factors
- 3 levels:
    1) Not use any feature related to race & sex
    2) Make race & sex background variables (parents of observed features)
    3) Includes error

Kusner et al. (2017)

Experiment: Law School Success



Kusner et al. (2017)

Counterfactual Fairness in Use

- Crowdsourcing is used in ML to label large-scale datasets (like ImageNet!)
- No good way to measure social worker bias
- Can use the idea of counterfactual fairness!

Ghai et al. (2020)

Counterfactual Fairness in Use

- Definition: A fair social worker would label a query and its counterfactual the same way
- Then take mean absolute difference in the labels/outputs they provided for all pairs of queries and counterfactual queries
  - Higher score = more inherent bias
  - $$WorkerBias = \frac{1}{n}\sum_{i=1}^{n}|Label(Q_i) - Label(CQ_i)|$$

Ghai et al. (2020)

Summary: Fairness should model the causal structure of the world.

Overall thoughts on using causality and counterfactuals?