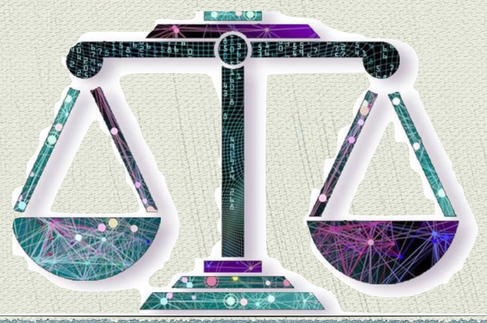
GEORGETOWN UNIVERSITY
MASSIVE DATA INSTITUTE
McCourt School of Public Policy

When is AI biased? Measuring fairness in machine learning models

Agoritsa Polyzou

Day 1



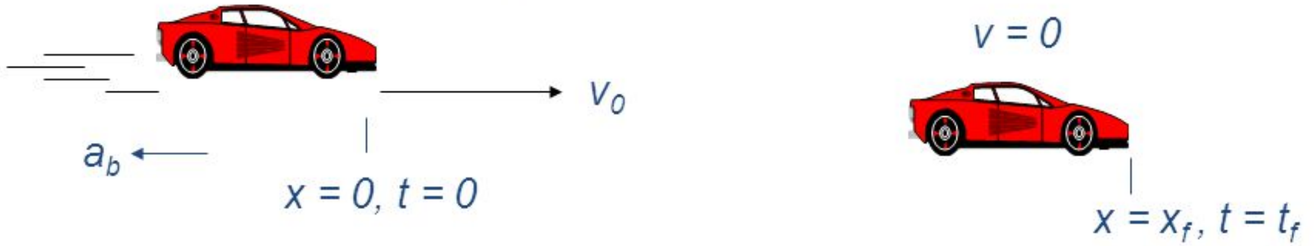
Introductions

- Who am I?
- Who are you?
- What do we care about?

Problem solving

Example

- A car is traveling with an initial velocity v_0. At $t = 0$, the driver puts on the brakes, which slows the car at a rate of a_b. At what time t_f does the car stop, and how much farther x_f does it travel?



In this case, we have:

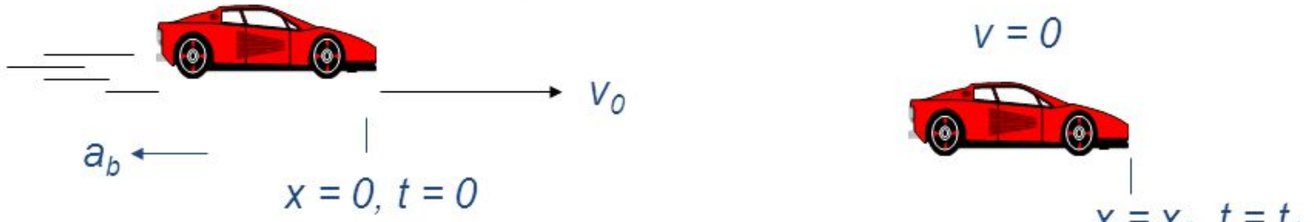
- few features
- good understanding of the environment (assumptions)
- good understanding of its impact on our task

⇒ we can come up with a good mathematical model!

Problem solving

Example

- A car is traveling with an initial velocity v_0. At $t = 0$, the driver puts on the brakes, which slows the car at a rate of a_b. At what time t_f does the car stop, and how much farther x_f does it travel?



In this case, we have:

- few features
- good understanding of the environment (assumptions)
- good understanding of its impact on our task

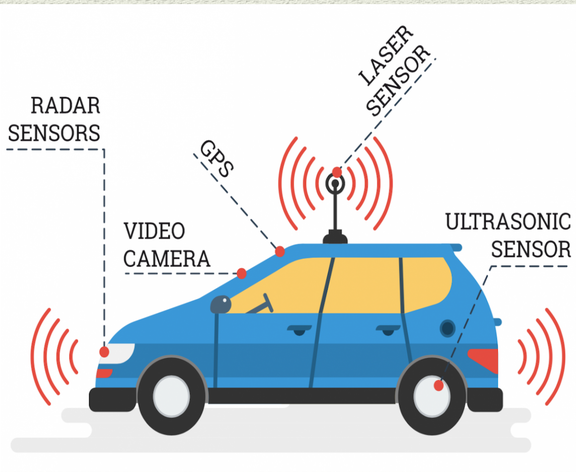
⇒ we can come up with a good mathematical model!

- Above, we derived: $v = v_0 + at$
- Realize that $a = -a_b$
- Also realizing that $v = 0$ at $t = t_f$:
  find $0 = v_0 - a_b t_f$ or

$t_f = v_0/a_b$

What if we have a self-driving car?

What if we have a self-driving car?



In this case, do we have :

… a full understanding of the environment? ✗

… true model? ✗

… few features? ✗

But we do have … data ! ✓

What if we have a self-driving car?

In this case, do we have :

… a full understanding of the environment? ❌

… true model? ❌

… few features? ❌

But we do have … data ! ✔

Data is the key to unlocking **machine learning,**
and machine learning is the key to unlocking **hidden insights** on the data!

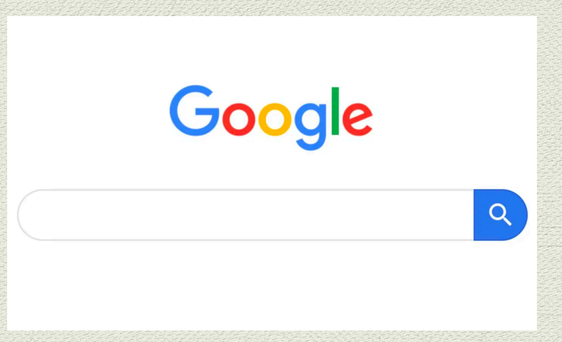
💡 Machine learning models can learn from data, identify patterns and support decisions. We don't have to explicitly formulate the solution!

Machine learning is …

About algorithms that give computers the tools and technology we can utilize to analyze massive amounts of data, find patterns, make predictions, and answer questions.

- Computers are good at storing, organizing, and processing **data**.
- Data: numbers, words, images, clicks, sounds, etc…

Applications



Search engines
(Understand text)

Transportation
(Self-driving cars)

Social Media
(Sentiment Analysis)

Criminal justice
(Recidivism risk)

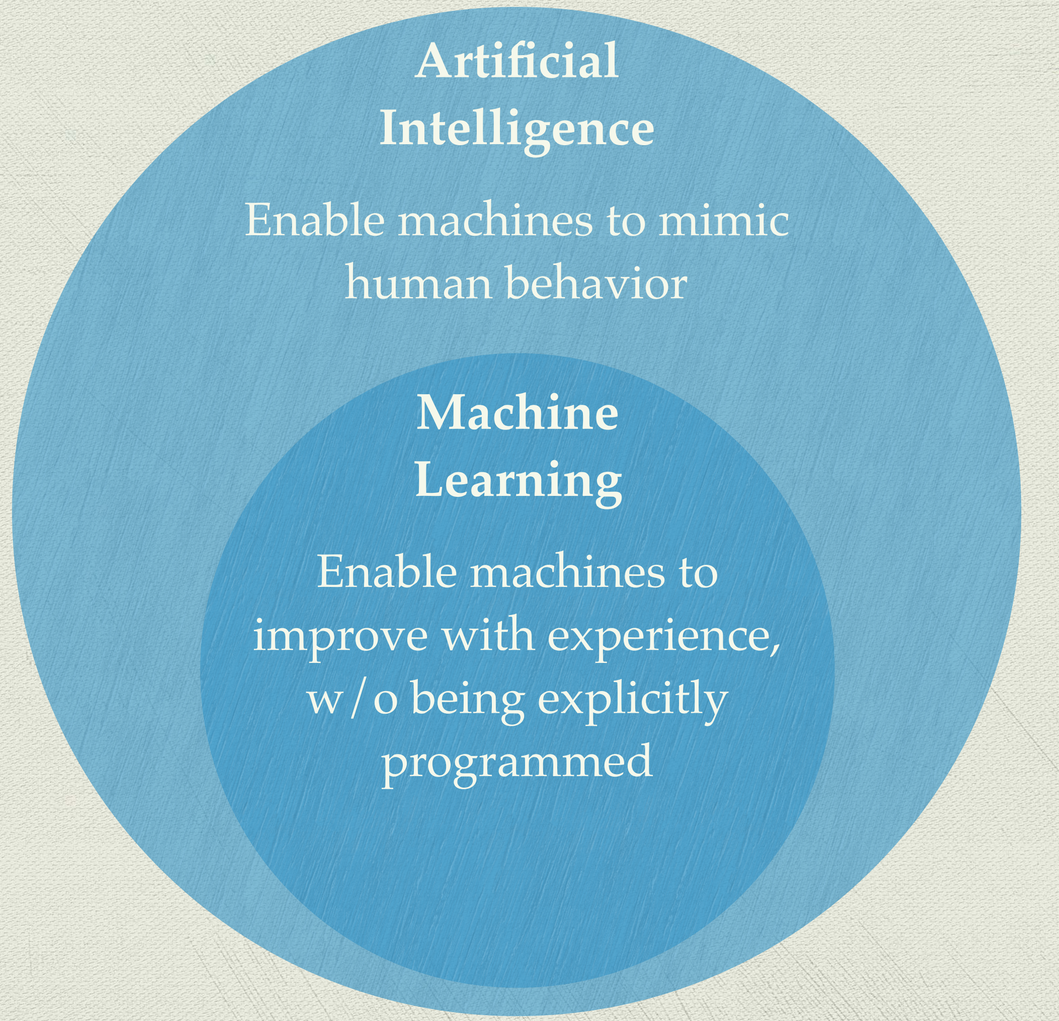
Finance
(Fraud Detection)

Government
(Handwriting recognition)

Healthcare
(Predict diagnosis)

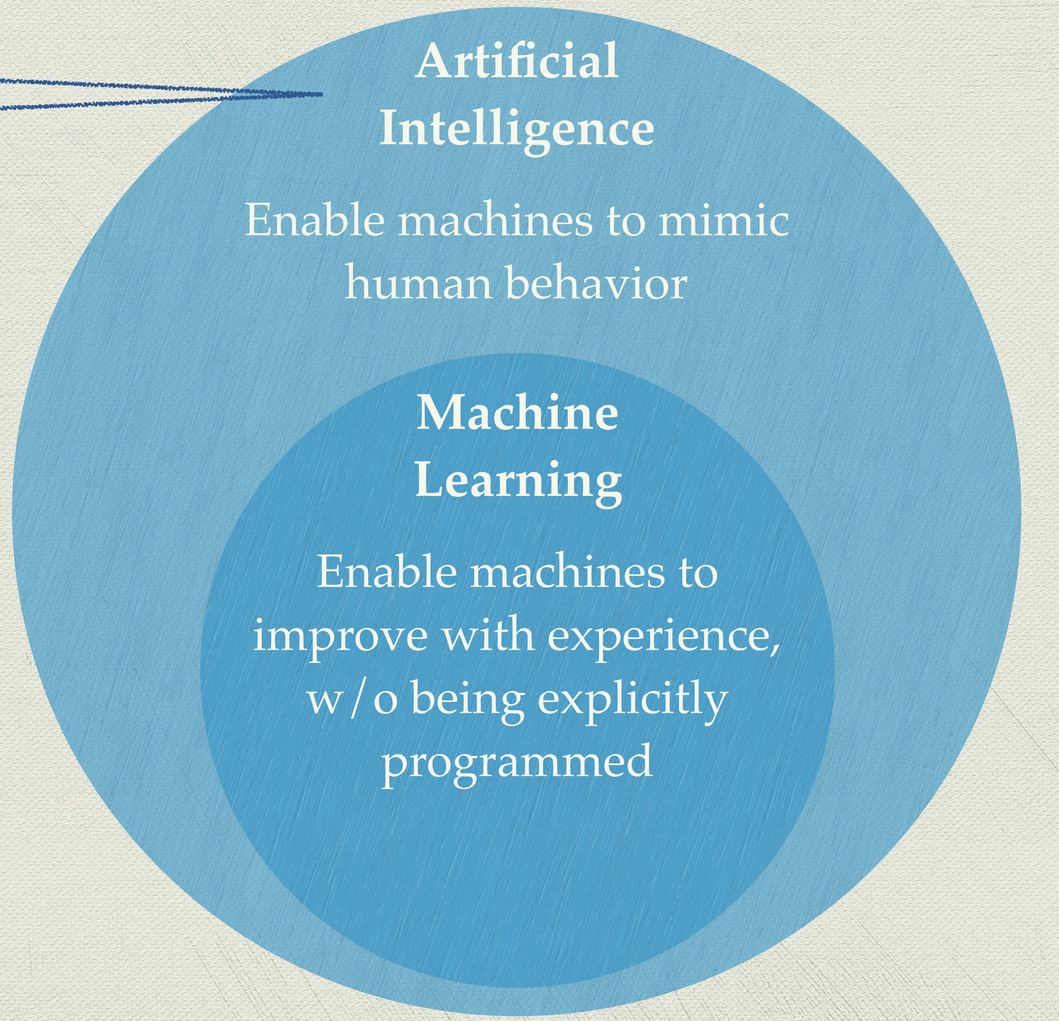
Shopping
(Product recom.)

AI and ML


Artificial Intelligence
Enable machines to mimic human behavior
Machine Learning
Enable machines to improve with experience, w/o being explicitly programmed

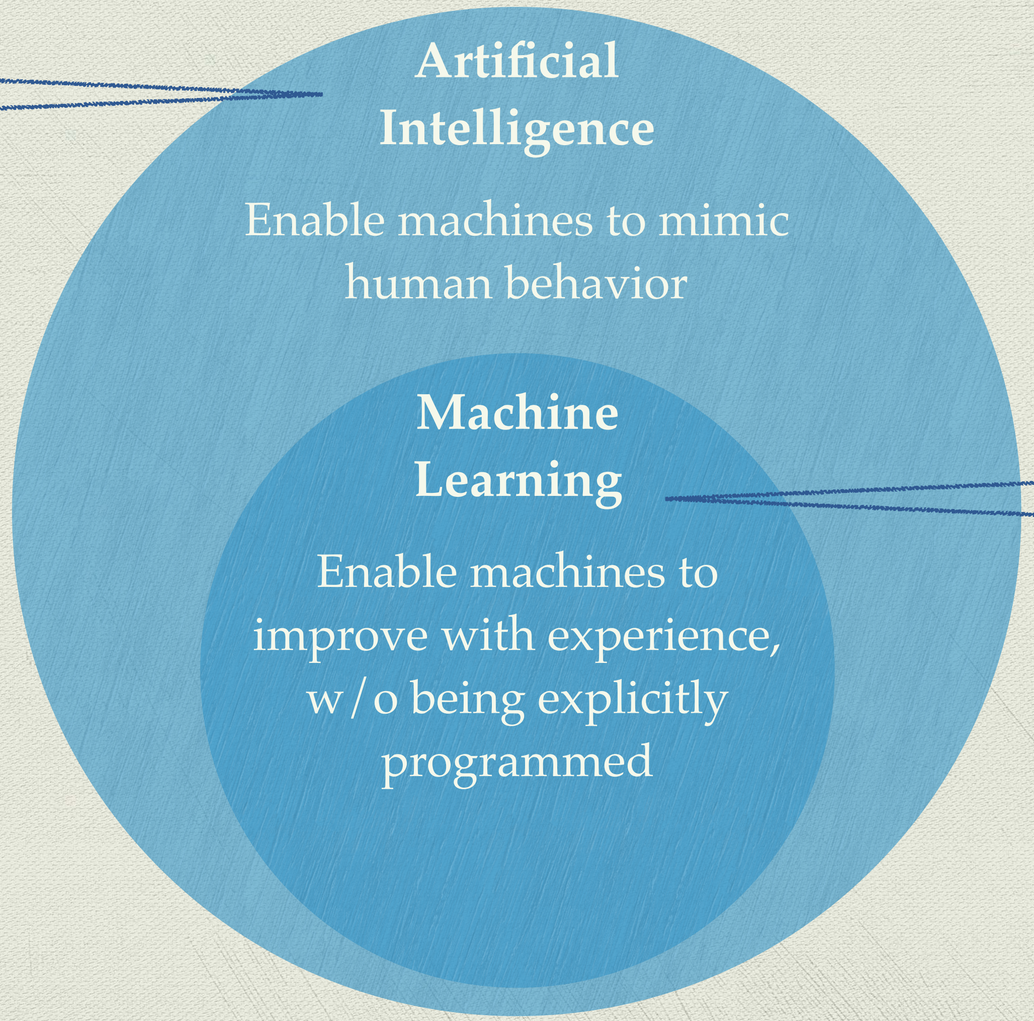
AI and ML

learning from rules or logical induction

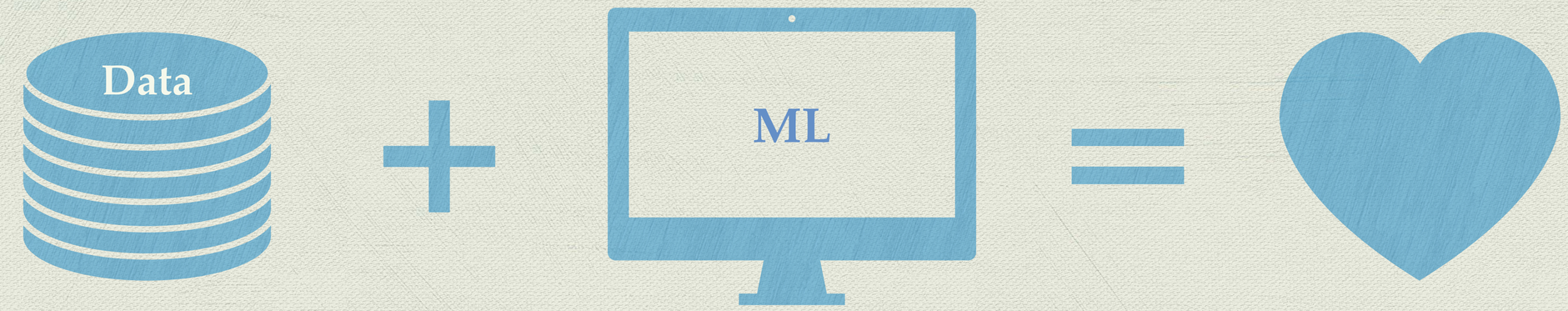


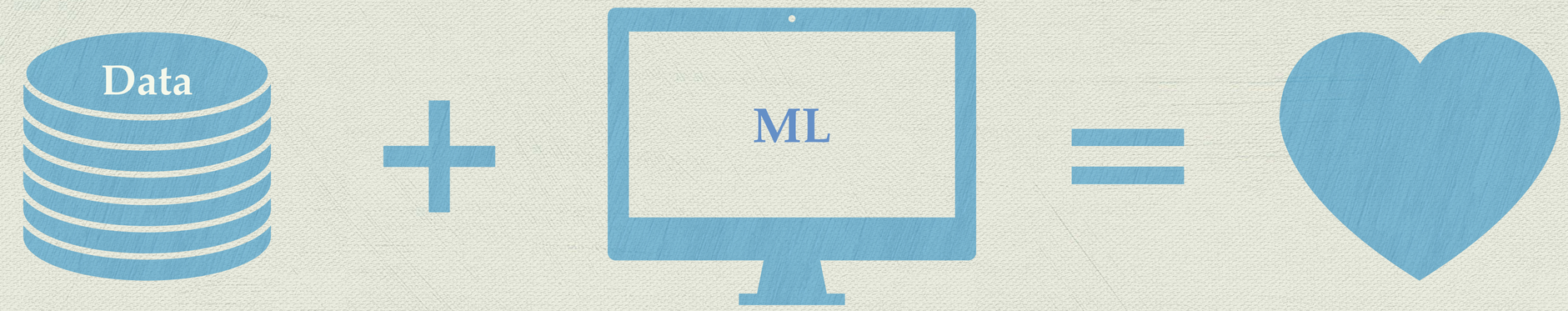
AI and ML

learning from rules or logical induction



just give them the data and let them learn


Data
ML

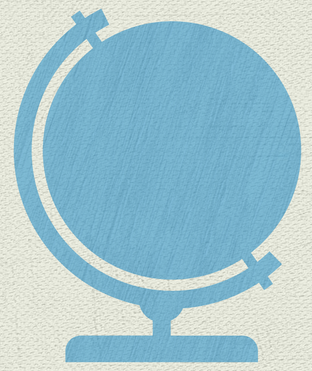


But…

- data influenced by the real world, come from human engineers or by observing the real world,
- models are likely to pick up the prejudices, biases and flaws of human reasoning.
- Results end up being systemically prejudiced.

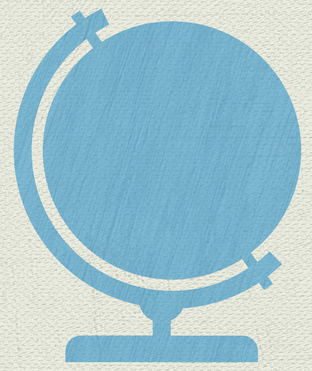
⇒ **ML depends on the quality, objectivity, and size of the data.**

World full of biases



Mercator Map

World full of biases



≠

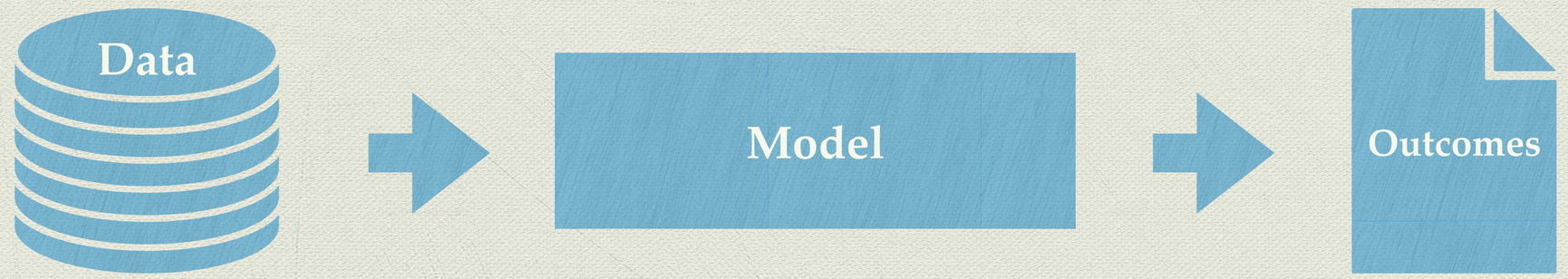
Mercator Map

- Africa is 14 times larger than Greenland.
- South America is twice the size of Europe.

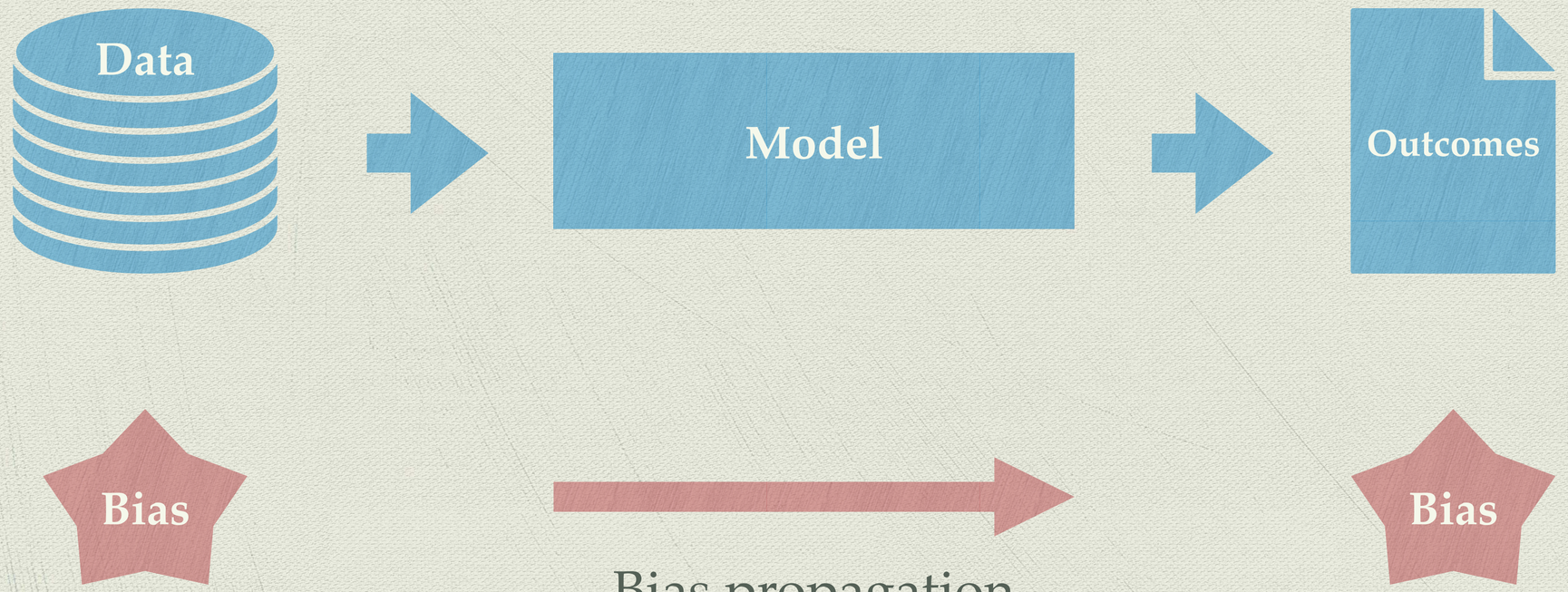
Map bias (like any other type of bias) can have unexpected consequences!

https://borgenproject.org/map-bias/

Biased Data ⇒ Biased ML models



Biased Data ⇒ Biased ML models



Biased ML models in the news

Microsoft silences its new A.I. bot Tay, after Twitter users teach it racism [Updated]

Sarah Perez @sarahintampa / 10:16 AM EDT • March 24, 2016 Comment

Microsoft's newly launched A.I.-powered bot called Tay, which was responding to tweets and chats on GroupMe and Kik, has already been shut down due to concerns with its inability to recognize when it was making offensive or racist statements. Of course, the bot wasn't *coded* to be racist, but it "learns" from those it interacts with. And naturally, given that this is the Internet, one of the first things online users taught Tay was how to be racist, and how to spout back ill-informed or inflammatory political opinions. [**Update**: Microsoft now says it's "making adjustments" to Tay in light of this problem.]

Two Petty Theft Arrests

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

Study finds gender and skin-type bias in commercial artificial-intelligence systems

Examination of facial-analysis software shows error rate of 0.8 percent for light-skinned men, 34.7 percent for dark-skinned women.

Watch Video

Larry Hardesty | MIT News Office
February 11, 2018

TECH AMAZON ARTIFICIAL INTELLIGENCE

Amazon reportedly scraps internal AI recruiting tool that was biased against women

The secret program penalized applications that contained the word "women's"

By James Vincent | Oct 10, 2018, 7:09am EDT

https://techcrunch.com/2016/03/24/microsoft-silences-its-new-a-i-bot-tay-after-twitter-users-teach-it-racism/ , https://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212 , https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing , https://www.theverge.com/2018/10/10/17958784/ai-recruiting-tool-bias-amazon-report

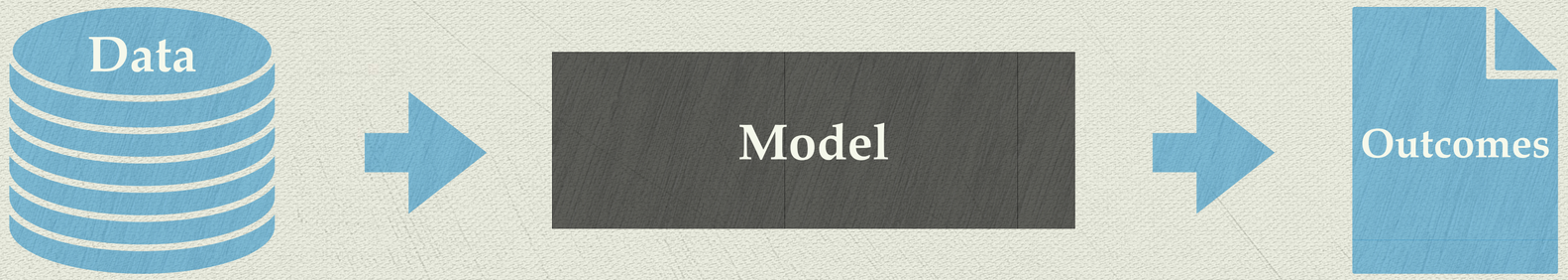
Sensitive Attributes / Features

The following are some of the attributes / features which could result in bias:

- Race
- Gender
- Color
- Religion
- National origin
- Marital status
- Sexual orientation
- Education background
- Source of income
- Age

Why is fair ML a challenging problem?

- Amount of data - too big to explore, analyze, and find biases.
- Unknown unknowns.
- Black-box ML models.



- Wide range of unexpected & unpredictable consequences, including complex social repercussions.
- Domain specific; different social context.

It is not enough to know that bias exists..

We need to understand the mechanics of how it arises in the first place,

we need to be able to fix it.

It is not enough to know that bias exists..

Day 1

We need to understand the mechanics of how it arises in the first place,

we need to be able to fix it.

Day 2

Optimism in the beginning

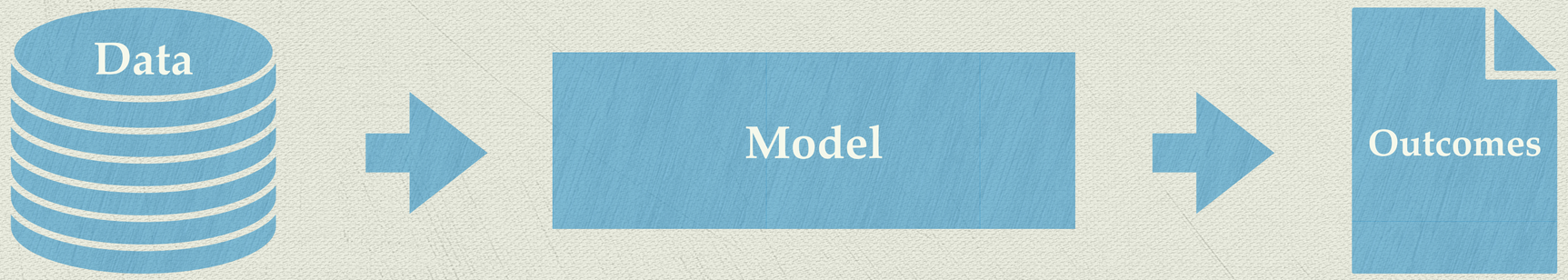
When we started using ML algorithms for prediction tasks,

there was an initial superficial feeling of optimism

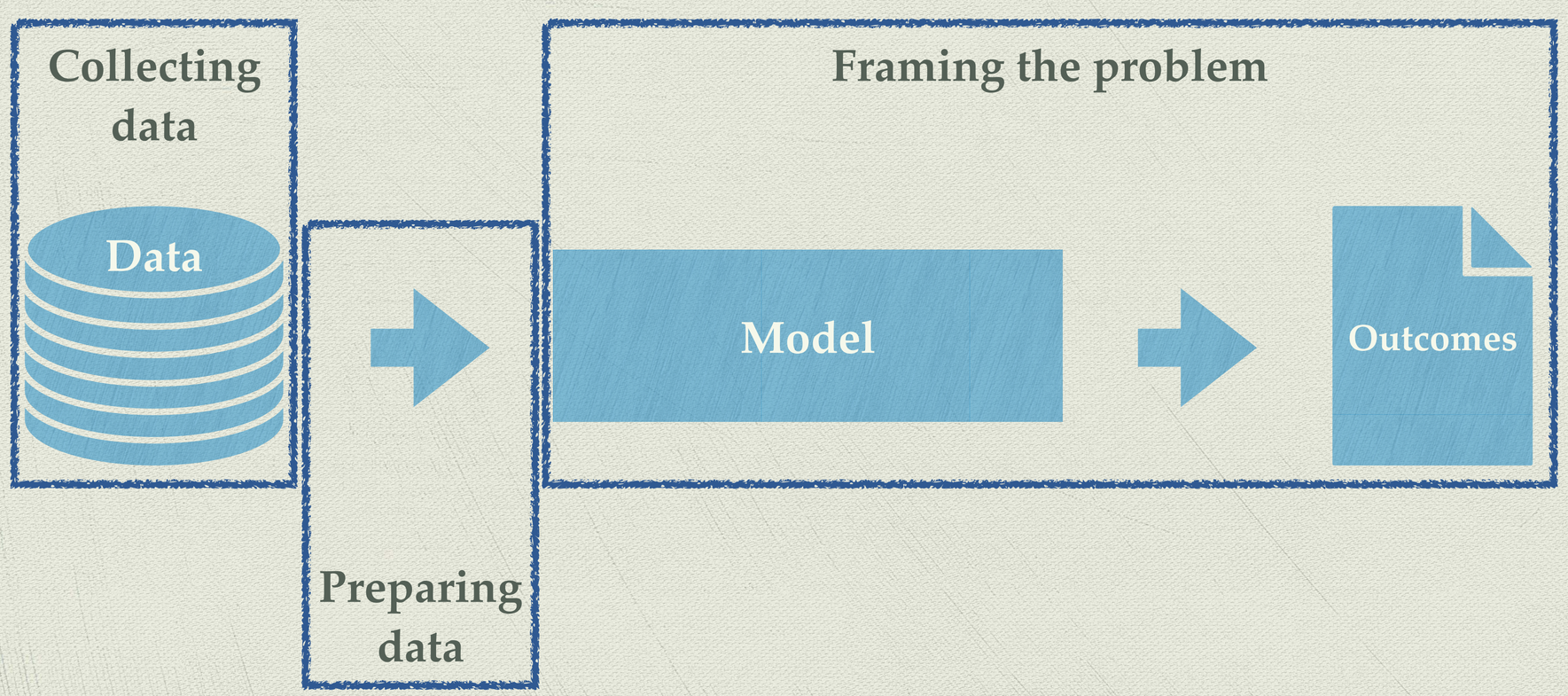
because **algorithms have no direct incentive to exhibit bias**.

Despite that, there are many sources of potential bias…

How is bias introduced in ML?



How is bias introduced in ML?



How is bias introduced in ML? (1)

Framing the problem

- The problem might be vague, uncertain, with high-level objectives.
- How the prediction task is defined and measured is influenced by profit goals, without considering fairness or discrimination.
- An algorithm might **unintentionally** engage in discriminatory behavior, even though if that was not the initial intention.

How is bias introduced in ML? (1)

Framing the problem

- The problem might be vague, uncertain, with high-level objectives.
- How the prediction task is defined and measured is influenced by profit goals, without considering fairness or discrimination.
- An algorithm might **unintentionally** engage in discriminatory behavior, even though if that was not the initial intention.

Example:

- Task: Predict creditworthiness for a credit card company
- Goal: maximize profit ⇒ approve subprime loans (vs. maximize the loans that get repaid)
- Outcome: people with good credit history are rejected, as they would get lower rates.

How is bias introduced in ML? (2)

Collecting the data

In two different ways:

1) the data is unrepresentative of reality, or

2) the data reflects existing prejudices.

How is bias introduced in ML? (2)

Collecting the data

In two different ways:

1) the data is unrepresentative of reality, or

2) the data reflects existing prejudices.

Example:

1) If an algorithm is fed more photos of light-skinned faces than dark-skinned faces, the resulting face recognition system would inevitably be worse at recognizing darker-skinned faces.
2) Amazon's internal recruiting tool was dismissing female candidates. Because it was trained on historical hiring decisions, which favored men over women, it learned to do the same.

How is bias introduced in ML? (3)

Preparing the data

- when selecting which attributes you want the algorithm to consider (during data cleaning).
- Choosing which attributes to consider/ignore can significantly influence prediction accuracy.
- While its impact on accuracy is easy to measure, its impact on the model's bias is not.

How is bias introduced in ML? (3)

Preparing the data

- when selecting which attributes you want the algorithm to consider (during data cleaning).
- Choosing which attributes to consider/ignore can significantly influence prediction accuracy.
- While its impact on accuracy is easy to measure, its impact on the model's bias is not.

Example:

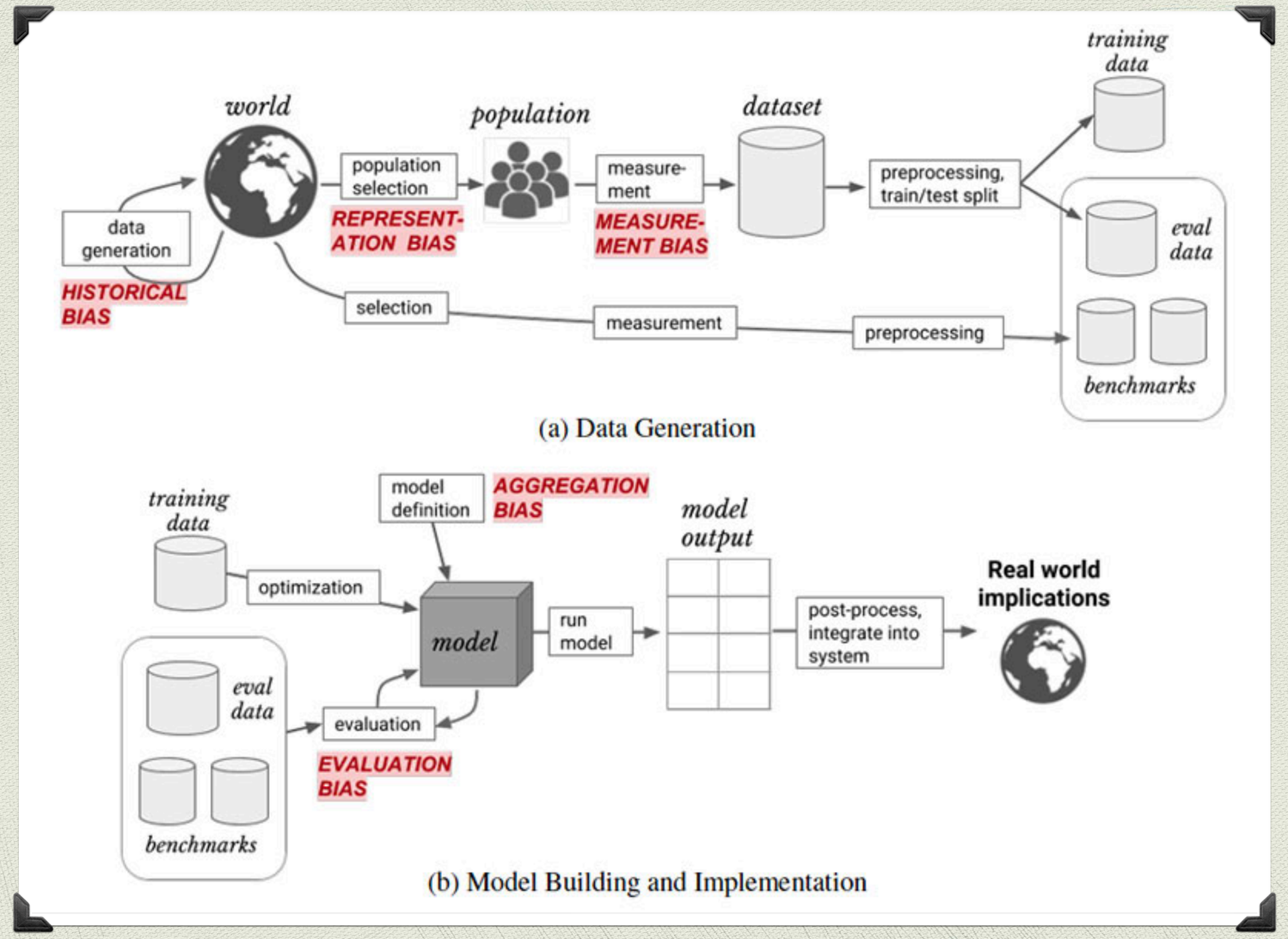
- Task: Titanic Survival prediction; who survived and who didn't.
- One might disregard the passenger id of the travelers, thinking that it is completely irrelevant.
- Titanic passengers were assigned rooms according to their passenger id. The smaller the id, the closer their assigned rooms are to the lifeboats. Thus, id ↑, survival ↓.

Activity 1

1) **Find an interesting ML application that you like.***
    a. Goal
    b. Data used
    c. Users (and other stakeholders)
2) **Think of ways that an algorithm for this application could discriminate its users.**

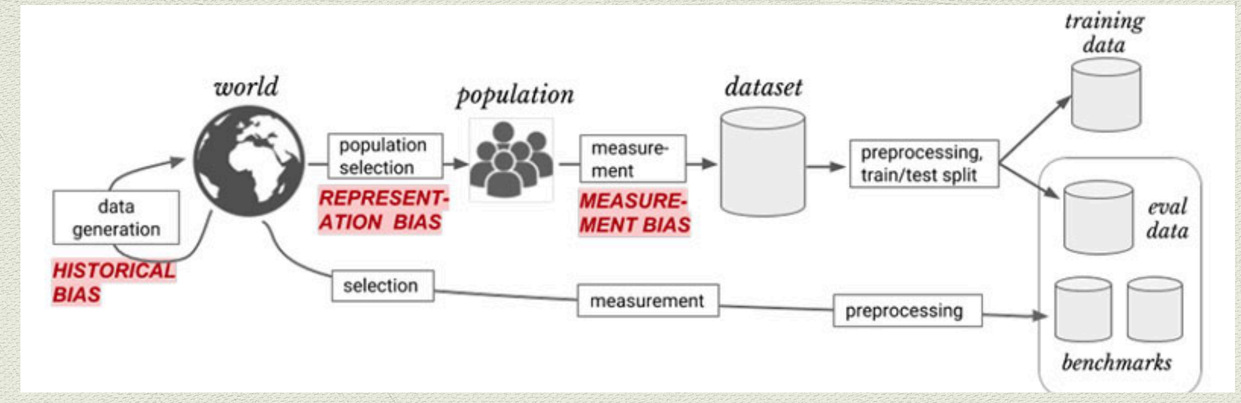
* or look for areas/domains where you think that ML is unlikely to be applied.

The different forms of bias



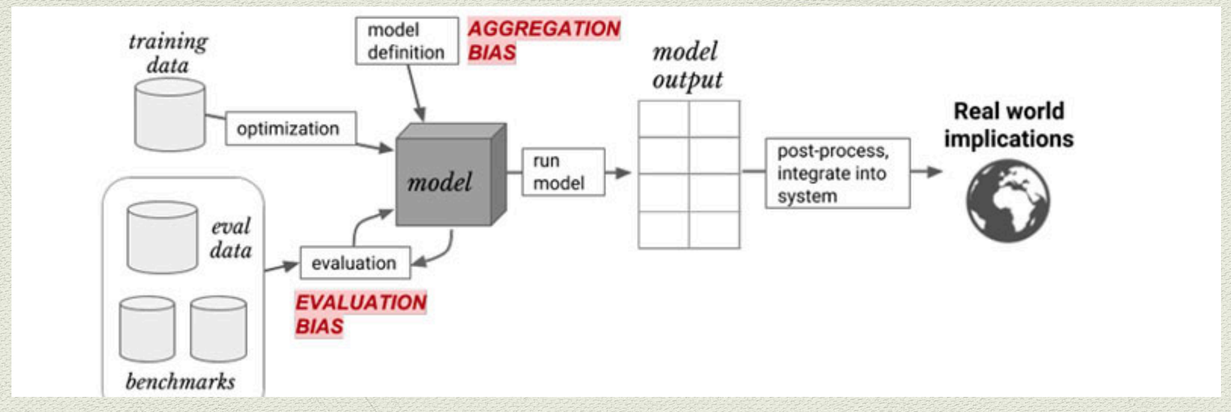
A Framework for Understanding Unintended Consequences of Machine Learning, Harini Suresh, John V. Guttag, 2020.

Forms of biases (1)



- **Historical** bias
  - when the world, as it is, is biased
- **Prejudice/stereotype** bias
  - When data is consciously or unconsciously reflecting stereotypes; view the world as biased because of their experiences & cultural influences, while it isn't.
- **Population/representation/selection/sample/coverage** bias
  - when a dataset does not reflect the realities of the environment in which a model will run
- **Measurement** bias
  - when the collected data differs from that collected in the real world, or when faulty measurements result in data distortion.
- **Exclusion** bias
  - in preprocessing, discard valuable data thought to be unimportant

Forms of biases (2)



- **Aggregation** bias
  - during model construction, when distinct populations are inappropriately combined
- **Evaluation** bias
  - during model evaluation, when the testing or external benchmark populations do not equally represent the real population, or when the performance metrics are not appropriate for the way in which the model will be used.
- **Algorithmic** bias
  - systematic and repeatable errors in a computer system that creates unfair outcomes; when algorithms pick up biases and propagate them into the ouput.

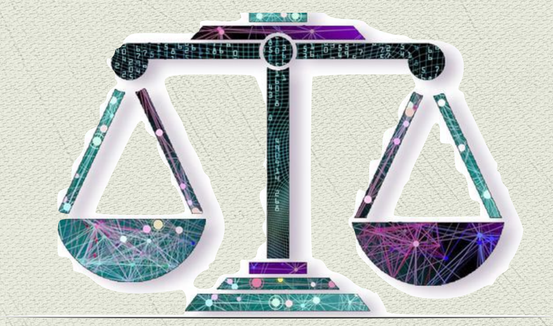
Activity 2: Audit search engines

1) Search images in different search engines.
2) Find the counts of male/female/both/NA in the top 20 images.
3) Which are the associated biases in this exercise?

- **Historical** bias
  - when the world, as it is, is biased
- **Prejudice/stereotype** bias
  - When data is consciously or unconsciously reflecting stereotypes; view the world as biased because of their experiences & cultural influences, while it isn't.
- **Population/representation/selection/sample/coverage** bias
  - when a dataset does not reflect the realities of the environment in which a model will run

Next Step:

Define & measure fairness



Problem set up

Let's assume that we have a classification problem.

- $x \in \mathbb{R}^p$: visible attributes of an individual, $x = (x_p, x_u)$
- a_0, a_1 : two possible outcomes to predict
- $d : \mathbb{R}^p \rightarrow \{0,1\}$: Model/decision algorithm
- $d(x_i) = k$: predict a_k for an individual

Goal: fairly predict a_0 or a_1

Anti-classification

We have algorithmic fairness when:

decisions do not consider protected attributes.

$$d(x) = d(x') \quad \text{for all} \quad x, x' \quad \text{such that} \quad x_u = x'_u$$

Anti-classification

We have algorithmic fairness when:

decisions do not consider protected attributes.

$d(x) = d(x')$ **Predictions** for all x, x' **Pairs of individuals** such that $x_u = x'_u$ **Same unprotected attributes**

Let’s just remove sensitive features!

Let’s just remove sensitive features!

Let's just remove sensitive features!

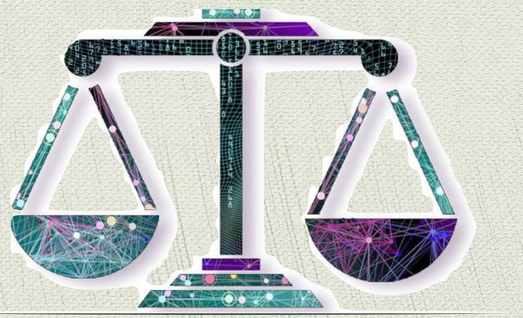
This will not work !

1) Models can learn the latent representation of those features from other provided features.
2) The exclusion of any information can lead to discriminatory decisions.

Anti-classification in not enough…

Tune in tomorrow for…

- More definitions of fairness
- Different ways to improve fairness
- Hands-on activity!
- Fairness for other prediction tasks, e.g., recommender systems



See you tomorrow!!