UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Ethical Aspects of Data

Fairness, Bias, Mitigation

Frederic Precioso

21/11/2023

(MAASAI, Joint Research Group INRIA-CNRS-UniCA)

frederic.precioso@univ-cotedazur.fr

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Overview

- **Context**
- Definition & Taxonomy
- Bias in ML
- Ethics In AI is not only about fairness

UNIVERSITÉ
CÔTE D'AZUR

univ-cotedazur.fr

CONTEXT

Biases in Data

Margaret Mitchell

Biases in Data

Selection Bias: Selection does not reflect a random sample

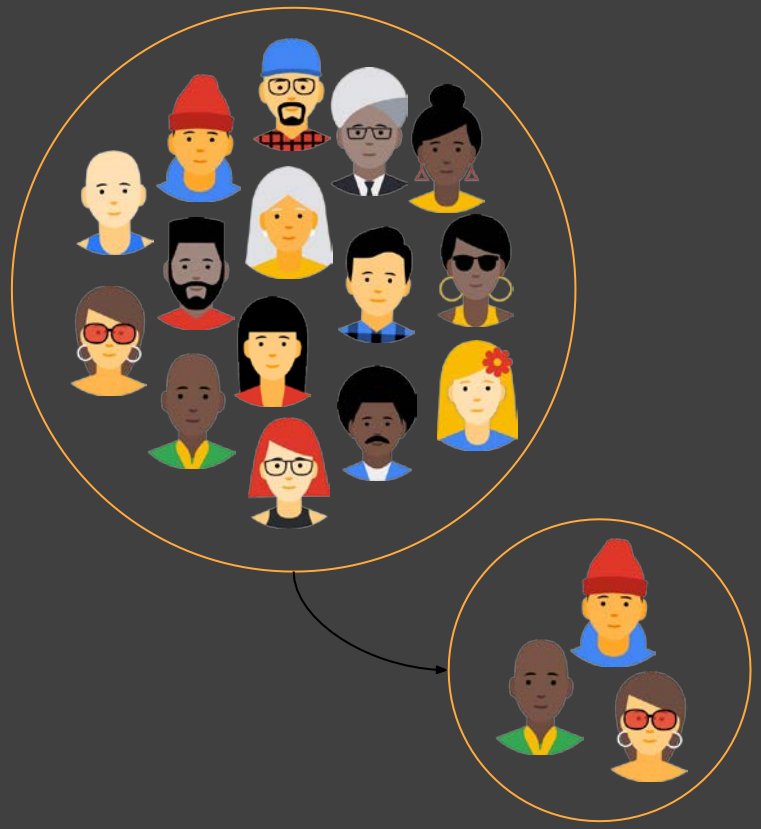
CREDIT

Biases in Data

Out-group homogeneity bias: Tendency to see outgroup members as more alike than ingroup members

Biases in Data →
Biased Data Representation

It's possible that you have an appropriate amount of data for every group you can think of but that some groups are represented less positively than others.



Margaret Mitchell

Biases in Data → Biased Labels

Annotations in your dataset will reflect the worldviews of your annotators.

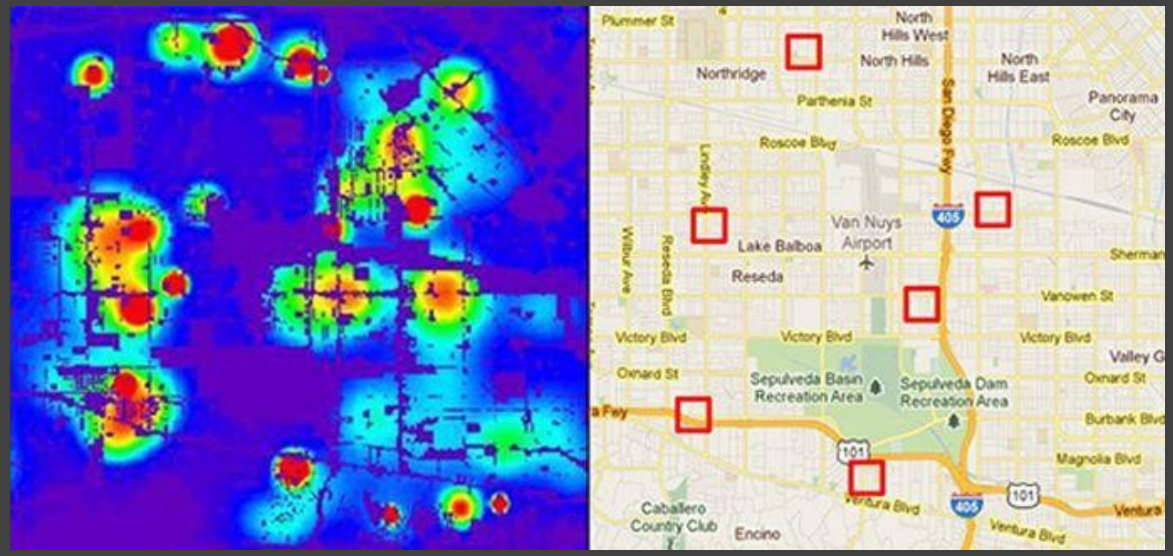
https://ai.googleblog.com/2018/09/introducing-inclusive-images-competition.html

Predicting Future Criminal Behavior

Margaret Mitchell

Predicting Policing

- Algorithms identify potential crime hot-spots
- Based on where crime is previously reported, not where it is known to have occurred
- Predicts future events from past



CREDIT

Smithsonian. *Artificial Intelligence Is Now Used to Predict Crime. But Is It Biased?* 2018

Predicting Sentencing

- Prater (who is white) rated **low risk** after shoplifting, despite two armed robberies; one attempted armed robbery.
- Borden (who is black) rated **high risk** after she and a friend took (but returned before police arrived) a bike and scooter sitting outside.
- Two years later, Borden has not been charged with any new crimes. Prater serving 8-year prison term for grand theft.

CREDIT

ProPublica. Northpointe: Risk in Criminal Sentencing. 2016.

Predicting Criminality

Israeli startup, Faception

> *"Faception is first-to-technology and first-to-market with proprietary computer vision and machine learning technology for profiling people and **revealing their personality based only on their facial image**."*

Offering specialized engines for recognizing "High IQ", "White-Collar Offender", "Pedophile", and "Terrorist" from a face image.

Main clients are in homeland security and public safety.

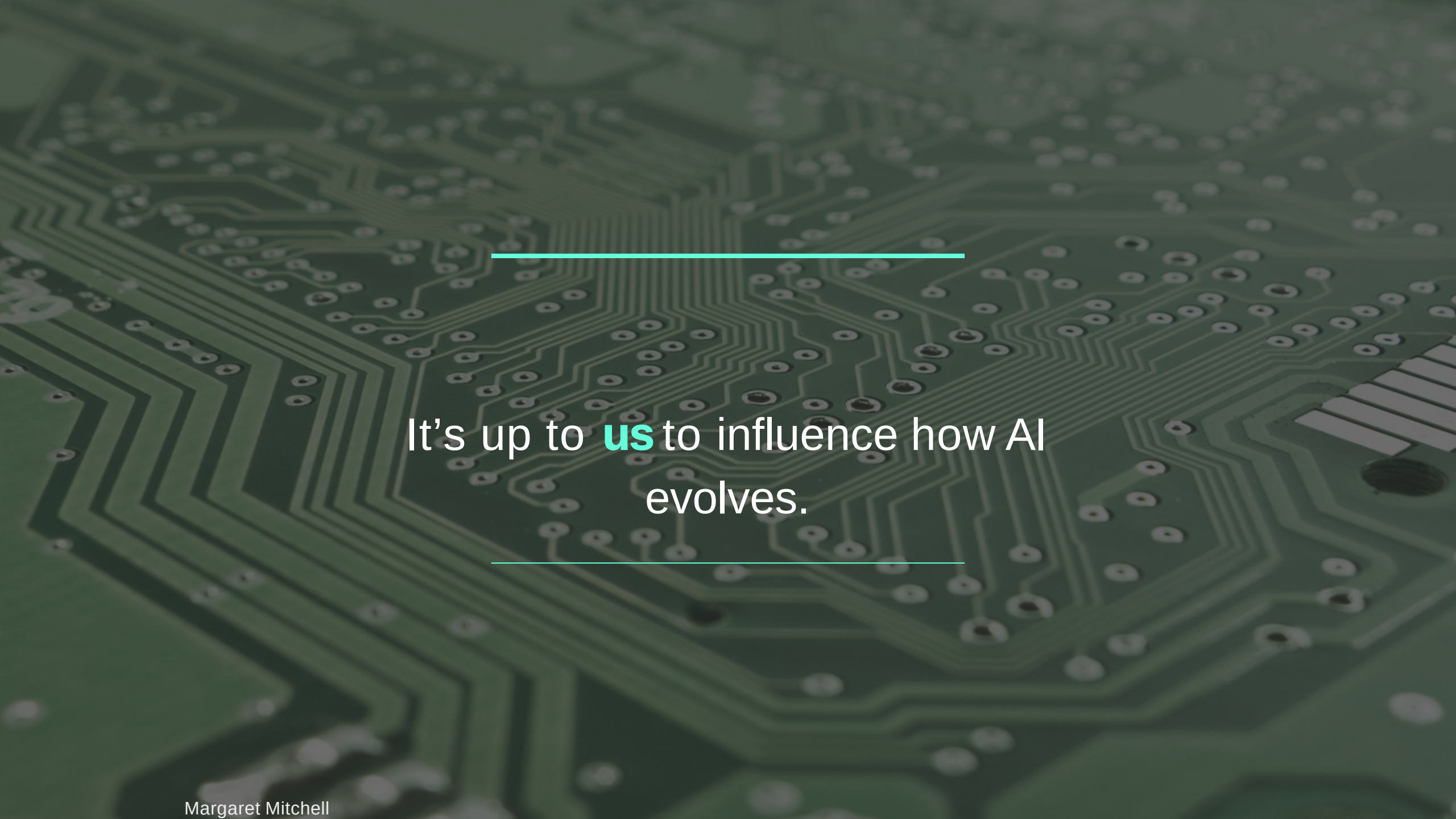
Predicting Criminality

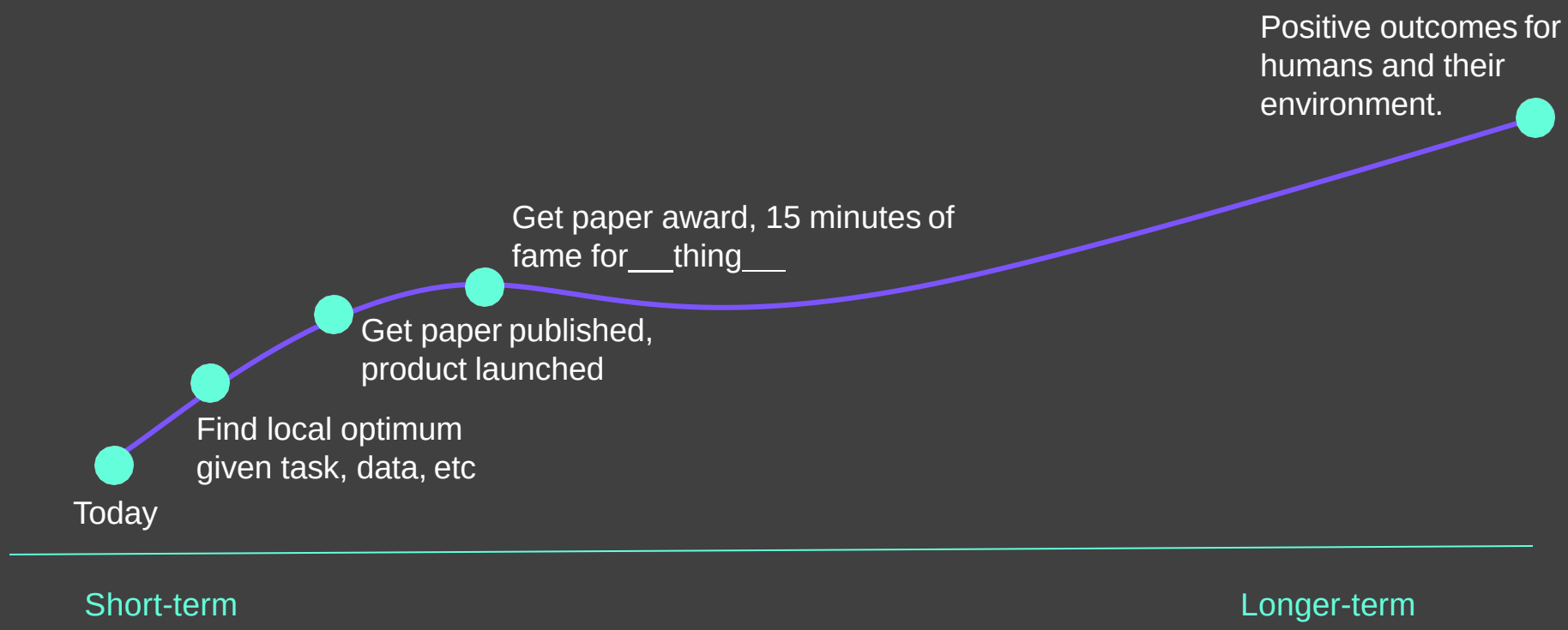
"Automated Inference on Criminality using Face Images" Wu and Zhang, 2016. arXiv

1,856 closely cropped images of faces; Includes "wanted suspect" ID pictures from specific regions.

"[…] angle θ from nose tip to two mouth corners is on average 19.6% smaller for criminals than for non-criminals ..."

See our longer piece on Medium, "Physiognomy's New Clothes"


It’s up to us to influence how AI evolves.
Margaret Mitchell



Margaret Mitchell

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Overview

- Context
- **Definition & Taxonomy**
- Bias in ML
- Ethics In AI is not only about fairness

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

DEFINITIONS & TAXONOMY

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Early taxonomy - 1996

Pre-existing Bias

Roots in

- Social institutions
- Social attitudes
- Social practices

Technical Bias

Arising from

- Technical constraints
- Technical considerations

Emergent

From changing

- Societal knowledge
- Population
- Cultural values

Batya Friedman and Helen Nissenbaum. "Bias in Computer Systems". In: ACM Trans. Inf. Syst.14.3 (July 1996).

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Early taxonomy - 1996

Pre-existing bias:

Pre-existing bias can enter a system either through the **explicit** and **conscious efforts** of **individuals** or **institutions**, or **implicitly** and **unconsciously**, even **in spite of the best of intentions**.

1.1 Individual Bias that originates from individuals who have significant input in to the design of the system, such as the client commissioning the design or the system designer (e.g.,a client embeds personal racial biases into the specifications for loan approval software).

1.2 Societal Bias that originates from society at large, such as from organizations (e.g., industry), institutions (e.g., legal systems), or culture at large (e.g., gender biases present in the larger society that lead to the development of educational software that overall appeals more to boys than girls).

Batya Friedman and Helen Nissenbaum. "Bias in Computer Systems". In: ACM Trans. Inf. Syst.14.3 (July 1996).

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Early taxonomy - 1996

Technical bias:

Technical bias arises from technical constraints or technical considerations.

2.1 Computer Tools, from a limitation of the computer technology including hardware, software, and peripherals

e.g., in a database for matching organ donors with potential transplant recipients certain individuals retrieved and displayed on initial screens are favored systematically for a match over individuals displayed on later screens.

2.2 Decontextualized Algorithms, from the use of an algorithm that fails to treat all groups fairly under all significant conditions

e.g., a scheduling algorithm for airplanes take-off relies on the alphabetic airline listing to rank order flights ready within a given period of time.

2.3 Random Number Generation, from imperfections in pseudo random number generation or in the misuse of pseudo random numbers

e.g., an imperfection in a random-number generator used to select recipients for a scarce drug leads systematically to favoring individuals toward the end of the database.

2.4 Formalization of Human Constructs, from attempts to make human constructs such as discourse, judgments, or intuitions amenable to computers: when we quantify the qualitative, discretize the continuous, or formalize the nonformal

e.g., a legal expert system advises defendants on whether or not to plea bargain by assuming that law can be spelled out in an unambiguous manner that is not subject to human and humane interpretations in context.

Batya Friedman and Helen Nissenbaum. "Bias in Computer Systems". In: ACM Trans. Inf. Syst.14.3 (July 1996).

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Early taxonomy - 1996

Emergent Bias:

Emergent bias arises in a context of use with real users. This bias typically emerges some time after a design is completed, as a result of changing societal knowledge, population, or cultural values. User interfaces are likely to be particularly prone to emergent bias because interfaces by design seek to reflect the capacities, character, and habits of prospective users. Thus, a shift in context of use may well create difficulties for a new set of users.

3.1 New Societal Knowledge, from the emergence of new knowledge in society that cannot be or is not incorporated into the system design

e.g., a medical expert system for AIDS patients has no mechanism for incorporating cutting-edge medical discoveries that affect how individuals with certain symptoms should be treated.

3.2 Mismatch between Users and System Design, when the population using the system differs on some significant dimension from the population assumed as users in the design.

3.2.1 Different Expertise, when the system is used by a population with a different knowledge base from that assumed in the design

e.g., an ATM with an interface that makes extensive use of written instructions—"place the card, magnetic tape side down, in the slot to your left"—is installed in a neighborhood with primarily a nonliterate population.

3.2.2 Different Values, when the system is used by a population with different values than those assumed in the design

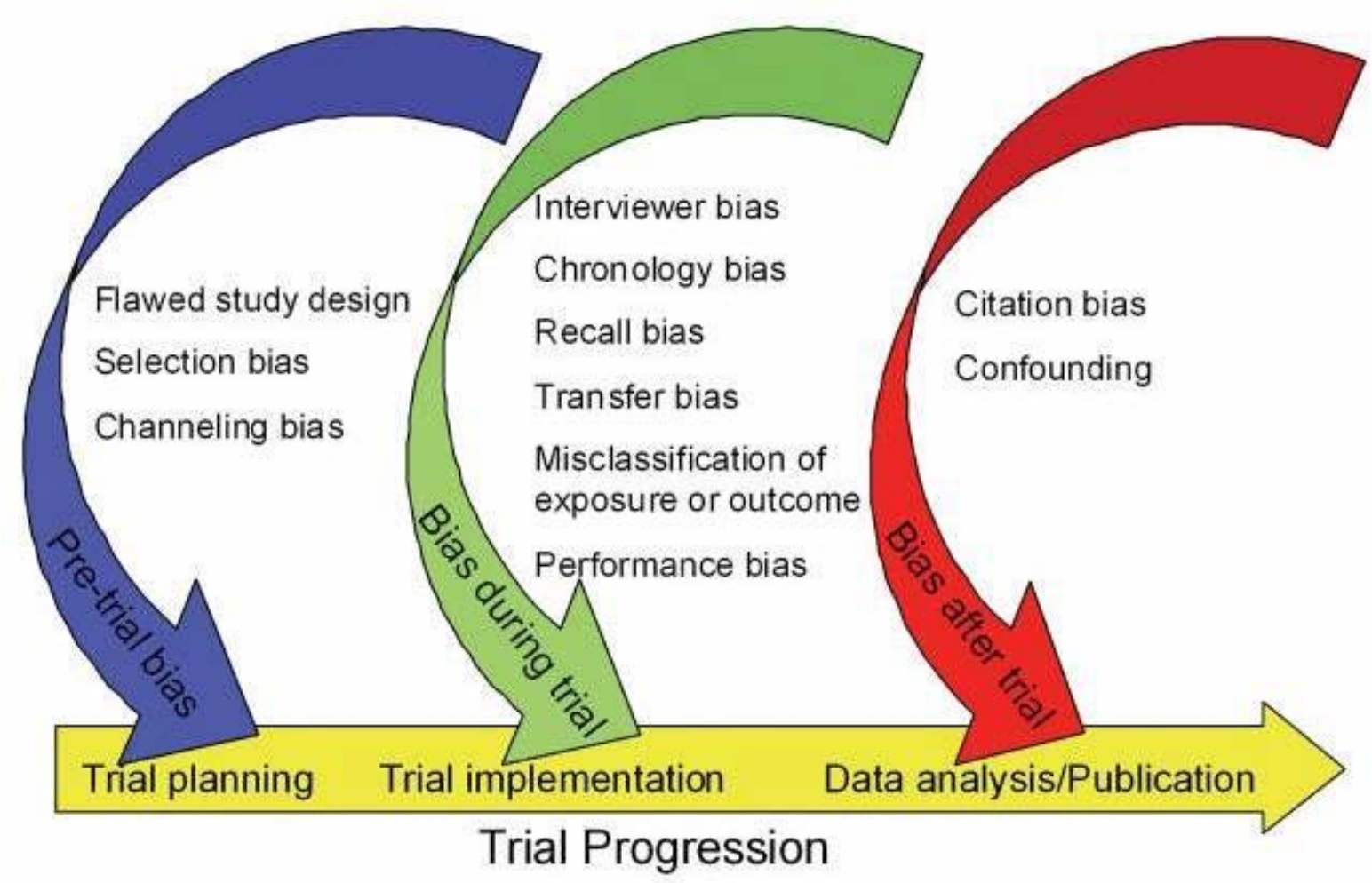
e.g., educational software to teach mathematics concepts is embedded in a game situation that rewards individualistic and competitive strategies but is used by students with a cultural background that largely eschews competition and instead promotes cooperative endeavors.

Batya Friedman and Helen Nissenbaum. "Bias in Computer Systems". In: ACM Trans. Inf. Syst.14.3 (July 1996).

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Major Biases in Clinical Research - 2010



Christopher Pannucci and Edwin Wilkins. "Identifying and Avoiding Bias in Research". In: Plastic and reconstructive surgery 126 (Aug. 2010)

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Major Biases in Clinical Research - 2010

Type of Bias	How to Avoid
Pre-trial bias	
Flawed study design	• Clearly define risk and outcome, preferably with objective or validated methods. Standardize and blind data collection.
Selection bias	• Select patients using rigorous criteria to avoid confounding results. Patients should originate from the same general population. Well designed, prospective studies help to avoid selection bias as outcome is unknown at time of enrollment.
Channeling bias	• Assign patients to study cohorts using rigorous criteria.
Bias after trial	
Citation bias	• Register trial with an accepted clinical trials registry. Check registries for similar unpublished or in-progress trials prior to publication.
Confounding	• Known confounders can be controlled with study design (case control design or randomization) or during data analysis (regression). Unknown confounders can only be controlled with randomization.

Christopher Pannucci and Edwin Wilkins. "Identifying and Avoiding Bias in Research". In: Plastic and reconstructive surgery 126 (Aug. 2010)

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Major Biases in Clinical Research - 2010

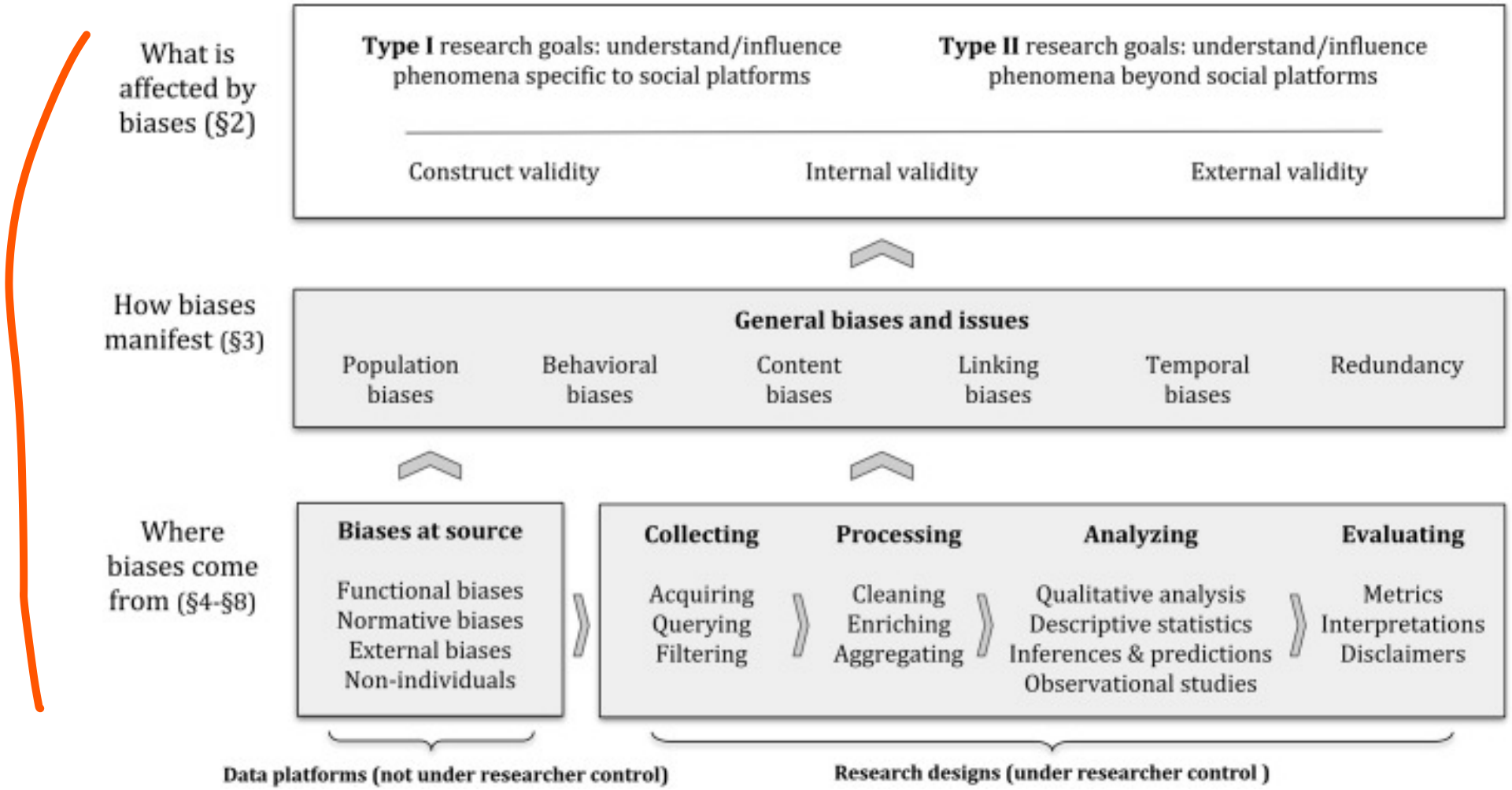
Type of Bias	How to Avoid
Bias during trial	
Interviewer bias	• Standardize interviewer's interaction with patient. Blind interviewer to exposure status.
Chronology bias	• Prospective studies can eliminate chronology bias. Avoid using historic controls (confounding by secular trends).
Recall bias	• Use objective data sources whenever possible. When using subjective data sources, corroborate with medical record. Conduct prospective studies because outcome is unknown at time of patient enrollment.
Transfer bias	• Carefully design plan for lost-to-followup patients prior to the study.
Exposure Misclassification	• Clearly define exposure prior to study. Avoid using proxies of exposure.
Outcome Misclassification	• Use objective diagnostic studies or validated measures as primary outcome.

Christopher Pannucci and Edwin Wilkins. "Identifying and Avoiding Bias in Research". In: Plastic and reconstructive surgery 126 (Aug. 2010)

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Bias in social data - 2019



Alexandra Olteanu et al. "Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries". In: Frontiers in Big Data2 (2019).

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Many different biases!

- Population Bias
- Behavioral Bias
- Content Production Bias
- Linking Bias
- Temporal Bias
- Sampling Bias
- Popularity Bias
- Algorithmic Bias
- User Interaction Bias
- Historical Bias
- Representation Bias
- Measurement Bias
- Evaluation Bias
- Aggregation Bias
- Simpson's Paradox
- Longitudinal Data Fallacy
- Emergent Bias
- Reporting Bias
- Automation Bias
- Selection Bias
- Coverage Bias
- Participation Bias
- Group Attribution Bias
- Implicit Bias
- Confirmation Bias
- Experimenter's Bias
- Presentation Bias
- Ranking Bias
- Social Bias
- Self-Selection Bias
- Omitted Variable Bias
- Cause-Effect Bias
- Observer Bias
- Funding Bias

Ninareh Mehrabi et al. "A Survey on Bias and Fairness in Machine Learning" (2019)

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

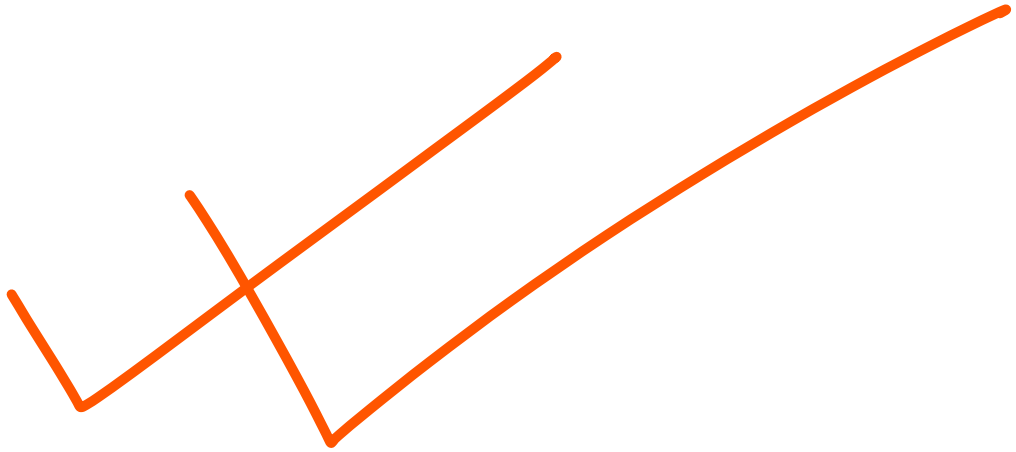
Overview

- Context
- Definition & Taxonomy
- **Bias in ML**
  - **Different forms of Biases**
  - **Bias Metrics**
  - **Mitigation**
- Ethics In AI is not only about fairness

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

BIAS IN ML

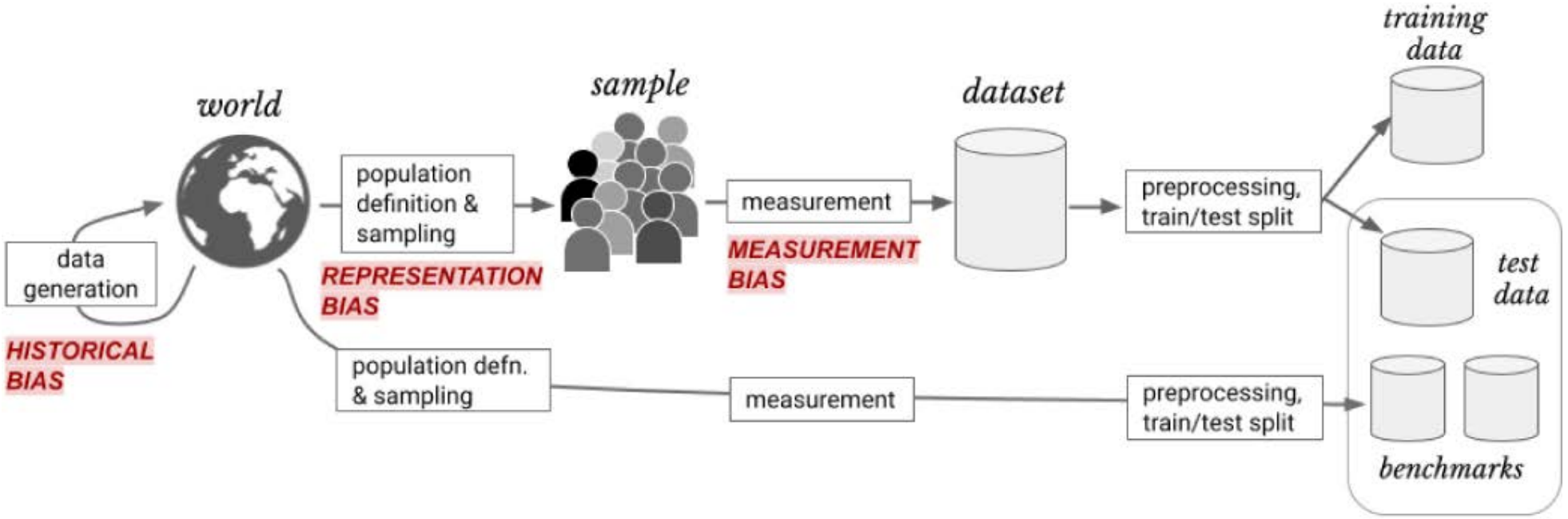


Different forms of Biases

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Bias in ML pipeline - 2021



(a) Data Generation

Harini Suresh and John V. Guttag. "A framework for understanding sources of harm throughout the machine learning life cycle". In Equity and access in algorithms, mechanisms, and optimization (pp. 1-9), 2021.

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Historical bias

- **Historical bias arises even if data is perfectly measured and sampled, if the world *as it is* or *was* leads to a model that produces harmful outcomes**. Such a system, even if it reflects the world accurately, can still inflict harm on a population. Considerations of historical bias often involve evaluating the representational harm (such as reinforcing a stereotype) to a particular group.
- *Examples:*
  - *5% of Fortune 500 CEO's are women. What should image search for CEO return?*
  - *How does Google change search algorithm? Should Google change algorithm? Who makes this decision?*

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

https://www.google.com/search?q=teacher

Search in June 2016

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

https://www.google.com/search?q=teacher

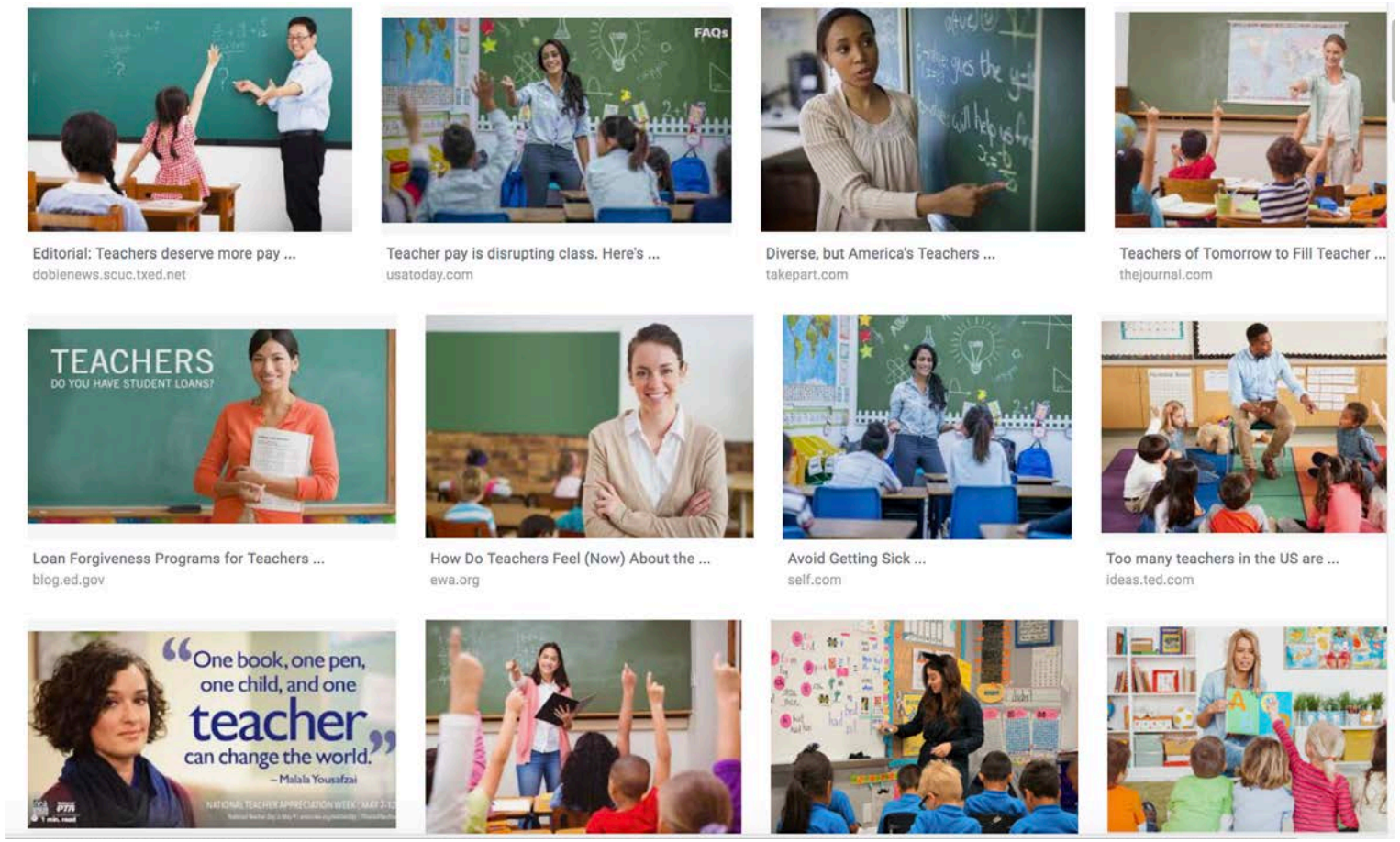
Search in January 2017

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

https://www.google.com/search?q=teacher

Search in February 2019

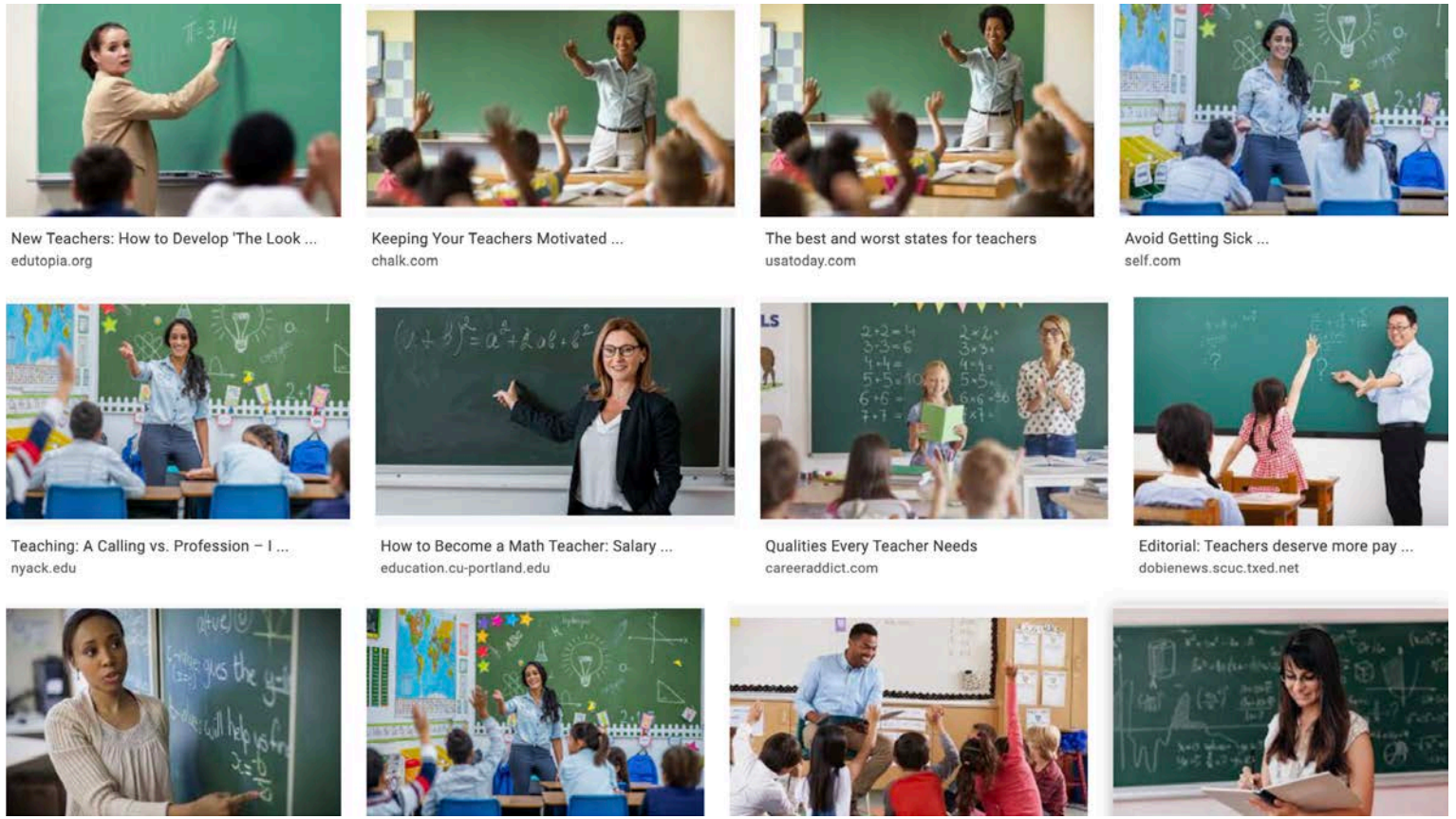


UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

https://www.google.com/search?q=teacher

Search in February 2020



UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

https://www.google.com/search?q=professor

Search in June 2016

Credits: Duke, Compsci 342, Algorithmic Bias Spring 2020

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

https://www.google.com/search?q=professor

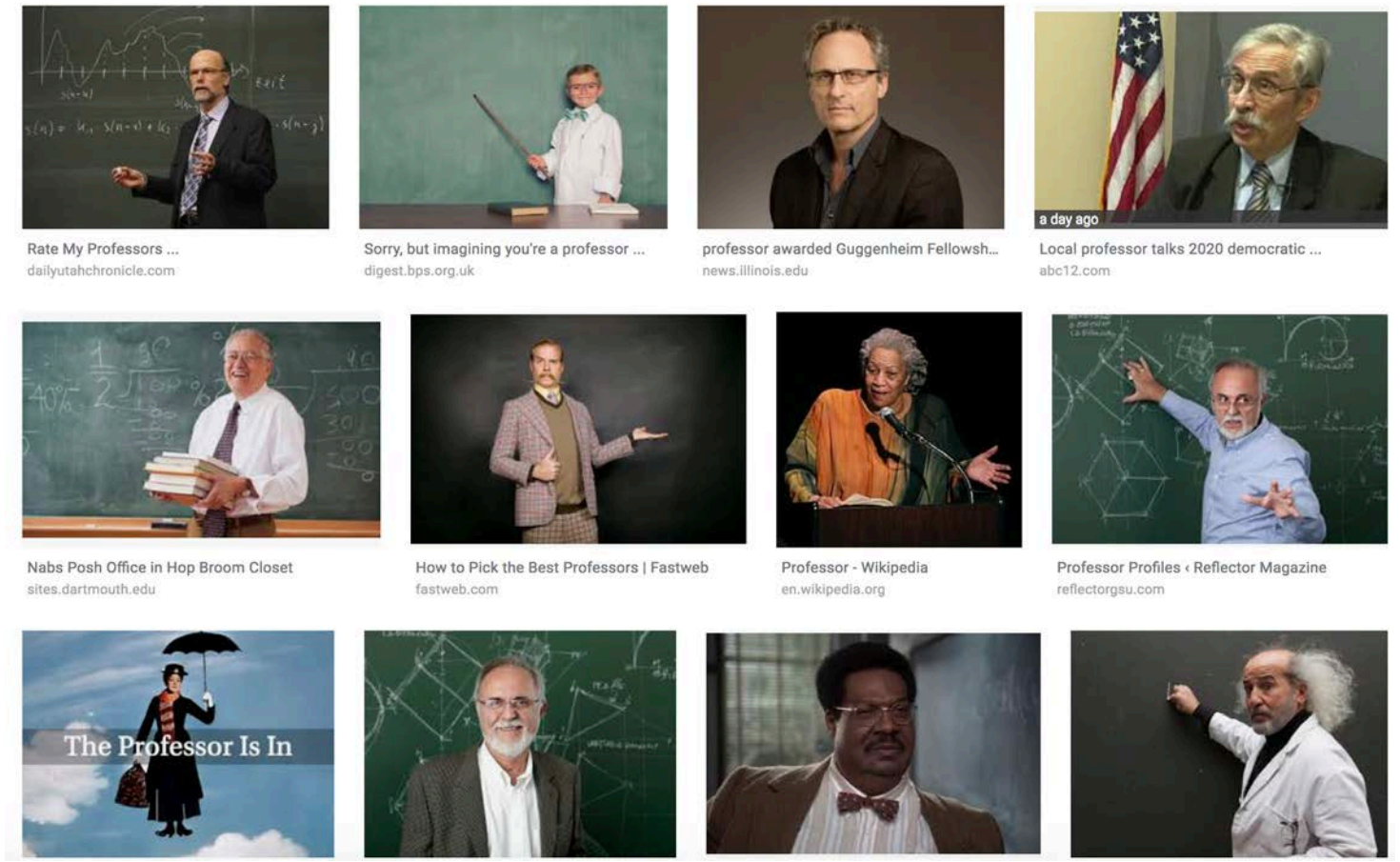
Search in January 2017

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

https://www.google.com/search?q=professor

Search in February 2019

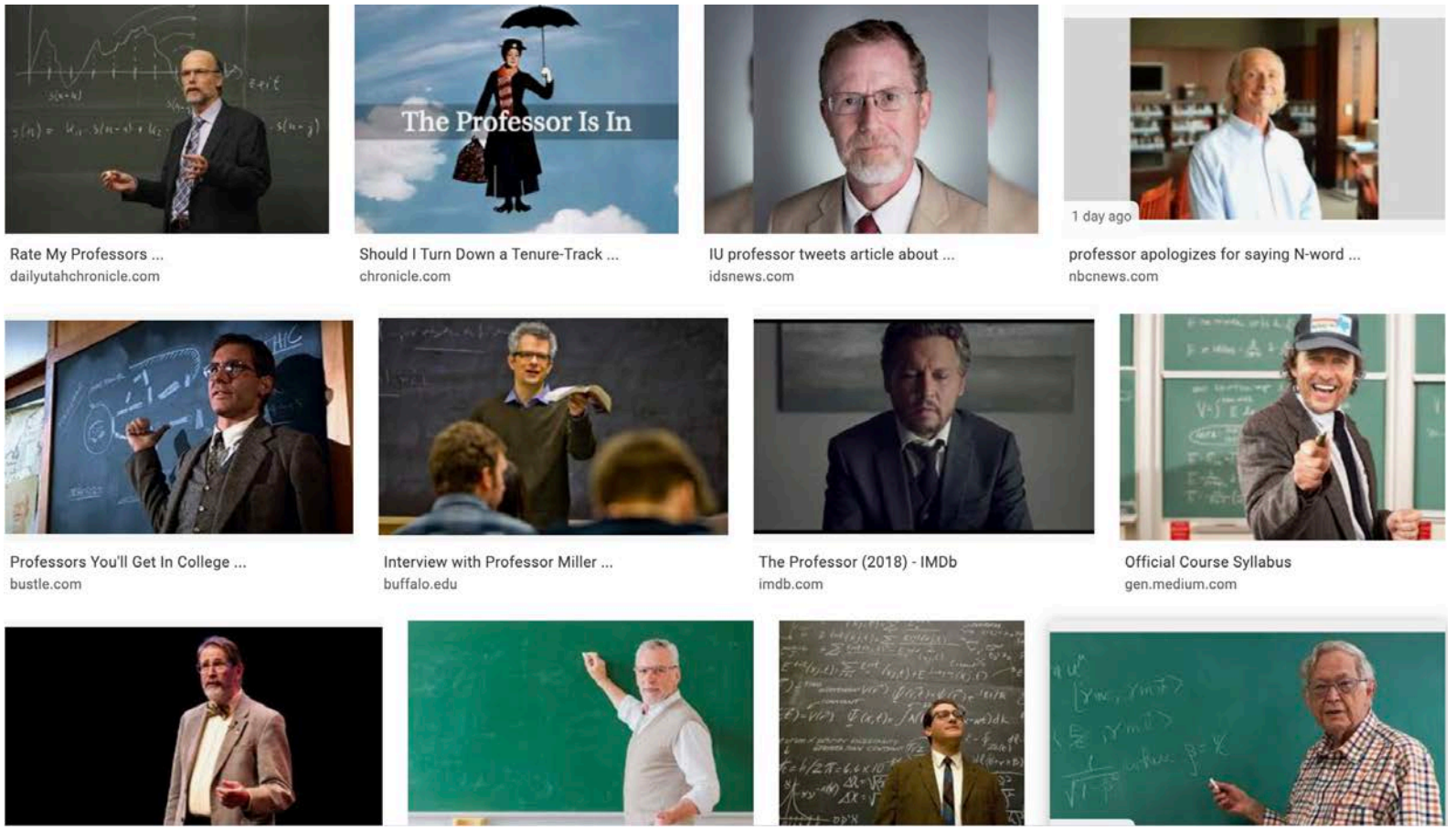


UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

https://www.google.com/search?q=professor

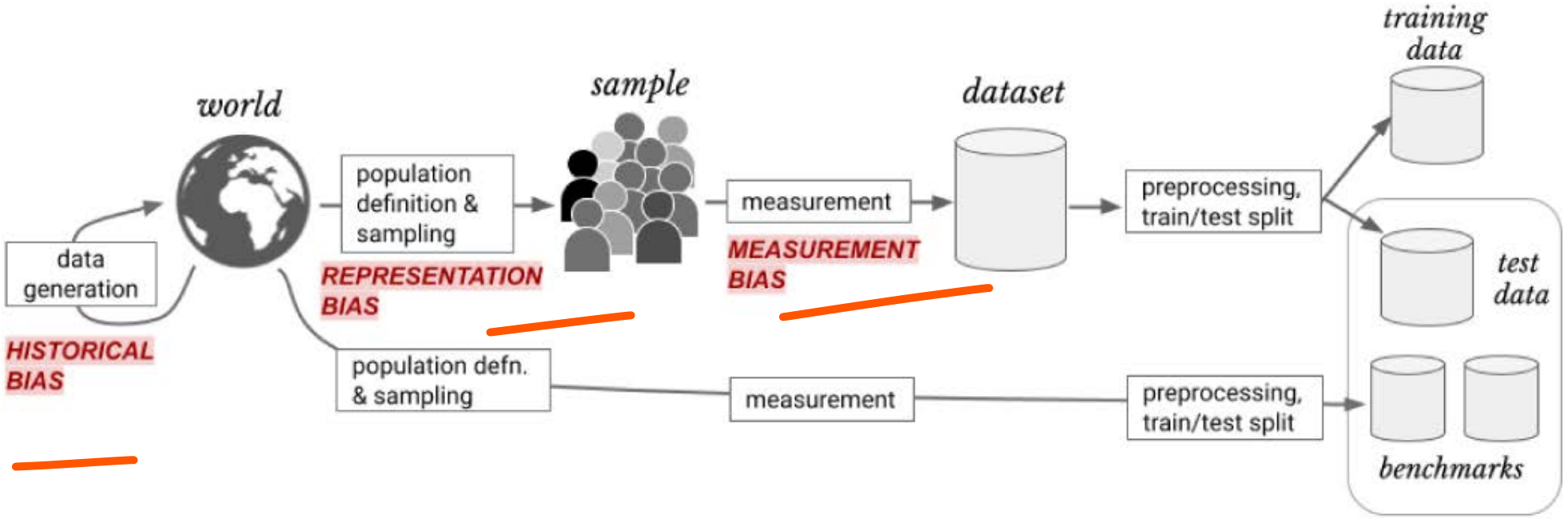
Search in February 2020



UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Bias in ML pipeline - 2021



(a) Data Generation

Harini Suresh and John V. Guttag. "A framework for understanding sources of harm throughout the machine learning life cycle". In Equity and access in algorithms, mechanisms, and optimization (pp. 1-9), 2021.

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Representation bias

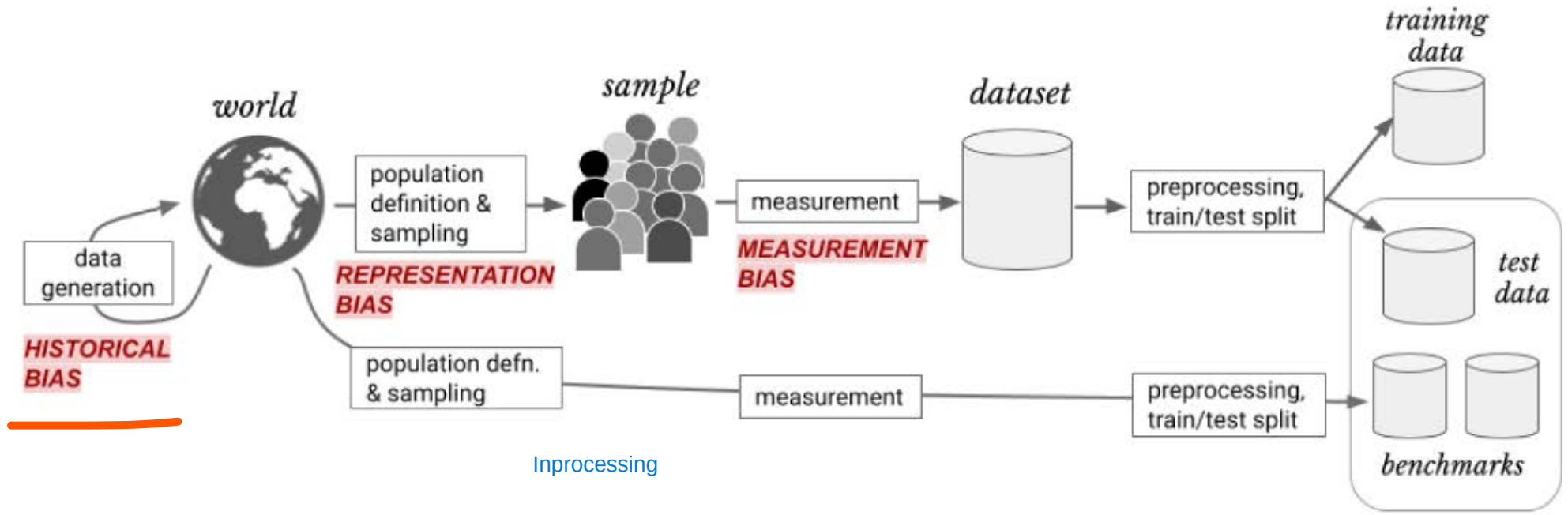
- Representation bias occurs when the development **sample under-represents some part of the population**, and subsequently **fails to generalize well** for a subset of the use population.
  Representation bias can arise in several ways:
  - ***When defining the target population, if it does not reflect the use population.***
  - ***When defining the target population, if contains under-represented groups.***
  - ***When sampling from the target population, if the sampling method is limited or uneven.***
- *Examples:*
  - Where a corpus of images originates -> ImageNet: 45% from United States
  - What about changing populations? -> Measure today, model tomorrow?

Credits: Duke, Compsci 342, Algorithmic Bias Spring 2020

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Bias in ML pipeline - 2021



(a) Data Generation

Harini Suresh and John V. Guttag. "A framework for understanding sources of harm throughout the machine learning life cycle". In Equity and access in algorithms, mechanisms, and optimization (pp. 1-9), 2021.

Measurement bias

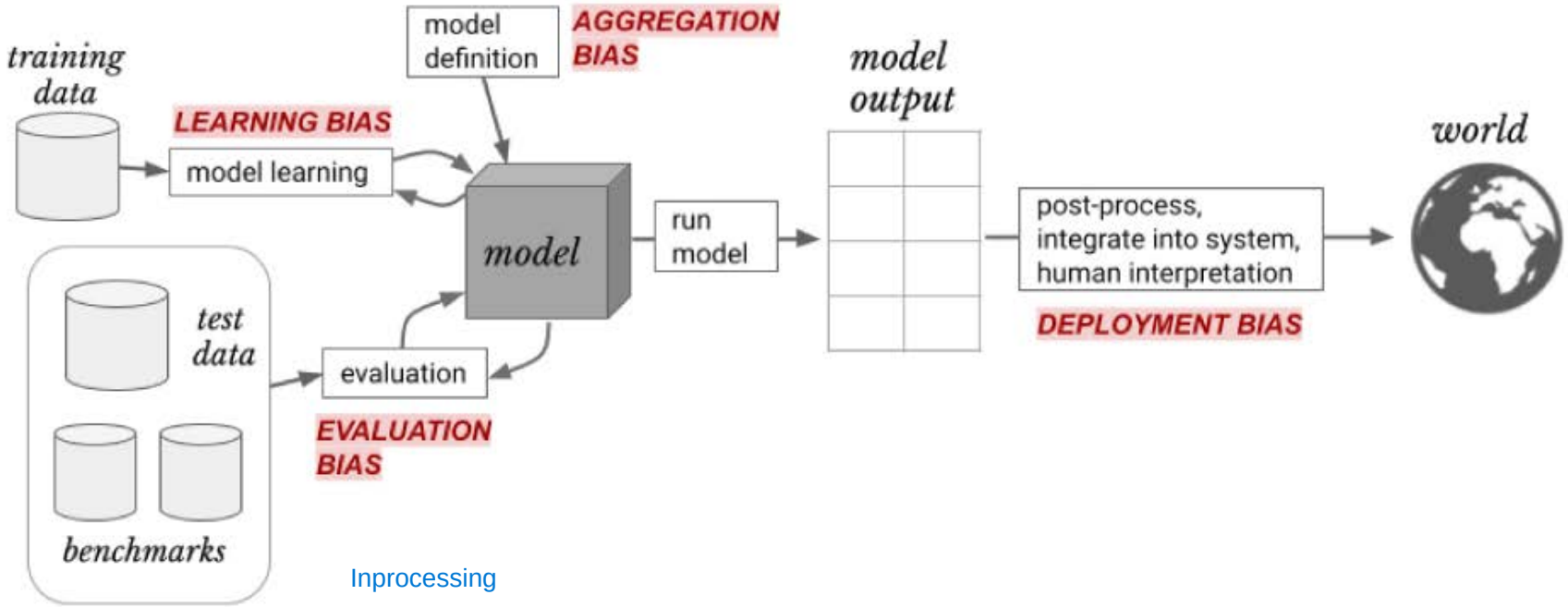
- Measurement bias occurs when choosing, collecting, or computing features and labels to use in a prediction problem. Typically, a feature or label is a proxy (a concrete measurement) chosen to approximate some construct (an idea or concept) that is not directly encoded or observable. Proxies become problematic when they are poor reflections or the target construct and/or are generated differently across groups, which can when:
  - ***The proxy is an oversimplification of a more complex construct.***
  - ***The method of measurement varies across groups.***
  - ***The accuracy of measurement varies across groups.***
- *Examples:*
  - Proxy labels: arrest used for crime -> What populations are arrested more? Is this useful for predictive policing in terms of crime? Re-arrest as proxy for recidivism in sentencing and parole automation
  - How do US measure "successful student“ -> GPA?

Credits: Duke, Compsci 342, Algorithmic Bias Spring 2020

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Bias in ML pipeline - 2021



(b) Model Building and Implementation

Harini Suresh and John V. Guttag. "A framework for understanding sources of harm throughout the machine learning life cycle". In Equity and access in algorithms, mechanisms, and optimization (pp. 1-9), 2021.

Evaluation Bias

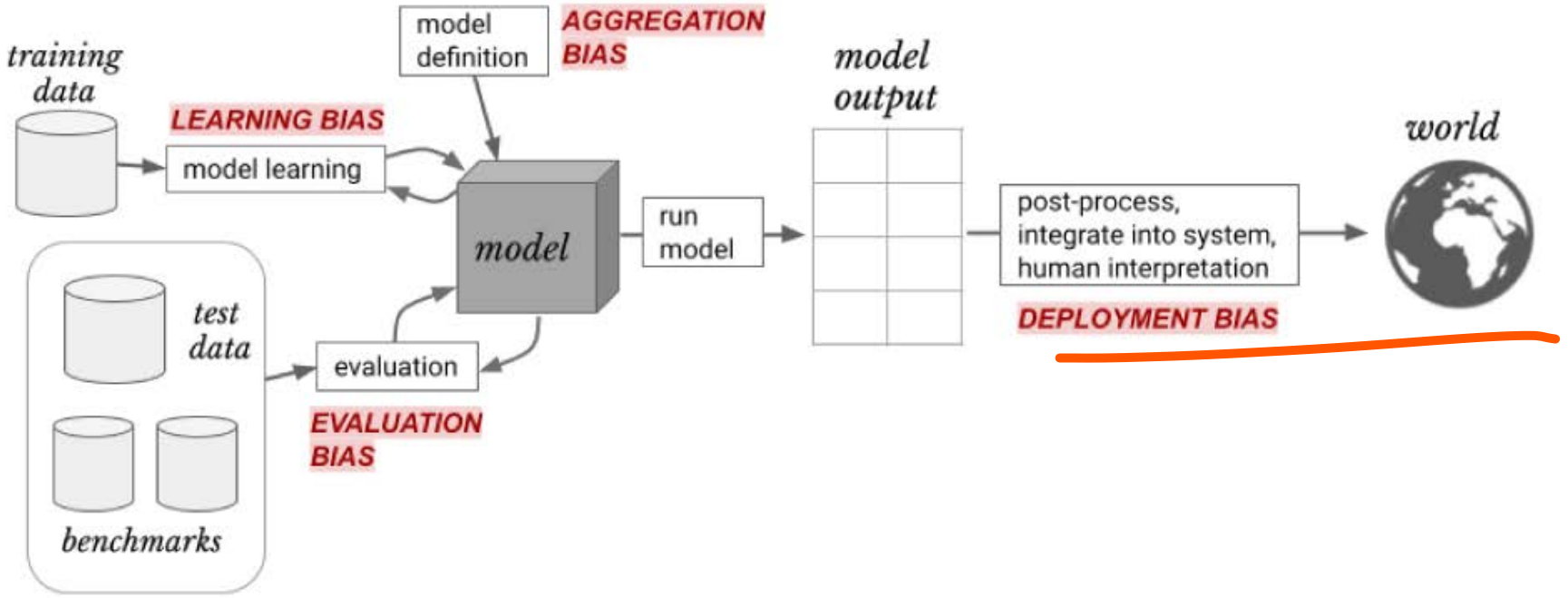
- **Evaluation bias occurs when the benchmark data used for a particular task does not represent the use population.**
- Evaluation bias ultimately arises because of a desire to quantitatively compare models against each other on a set of external datasets, but is often extended to make general statements about how good a model is.
- Such generalizations are often not statistically valid, and can lead to overfitting to a particular benchmark (and so the model is invalid on the other external datasets).
- **Evaluation bias can also be exacerbated by the choice of metrics that are used to report performance.** For example, aggregate measures can hide subgroup underperformance, but these singular measures are often used because they make it more straightforward to compare models and make a judgment about which one is "better." Just looking at one type of metric (e.g., accuracy) can also hide disparities in other types of errors (e.g., false negative rate).
- Examples:
  - Benchmark data doesn't represent population -> Algorithm optimized on training data, then benchmarked on a different population

Credits: Duke, Compsci 342, Algorithmic Bias Spring 2020

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Bias in ML pipeline - 2021



(b) Model Building and Implementation

Harini Suresh and John V. Guttag. "A framework for understanding sources of harm throughout the machine learning life cycle". In Equity and access in algorithms, mechanisms, and optimization (pp. 1-9), 2021.

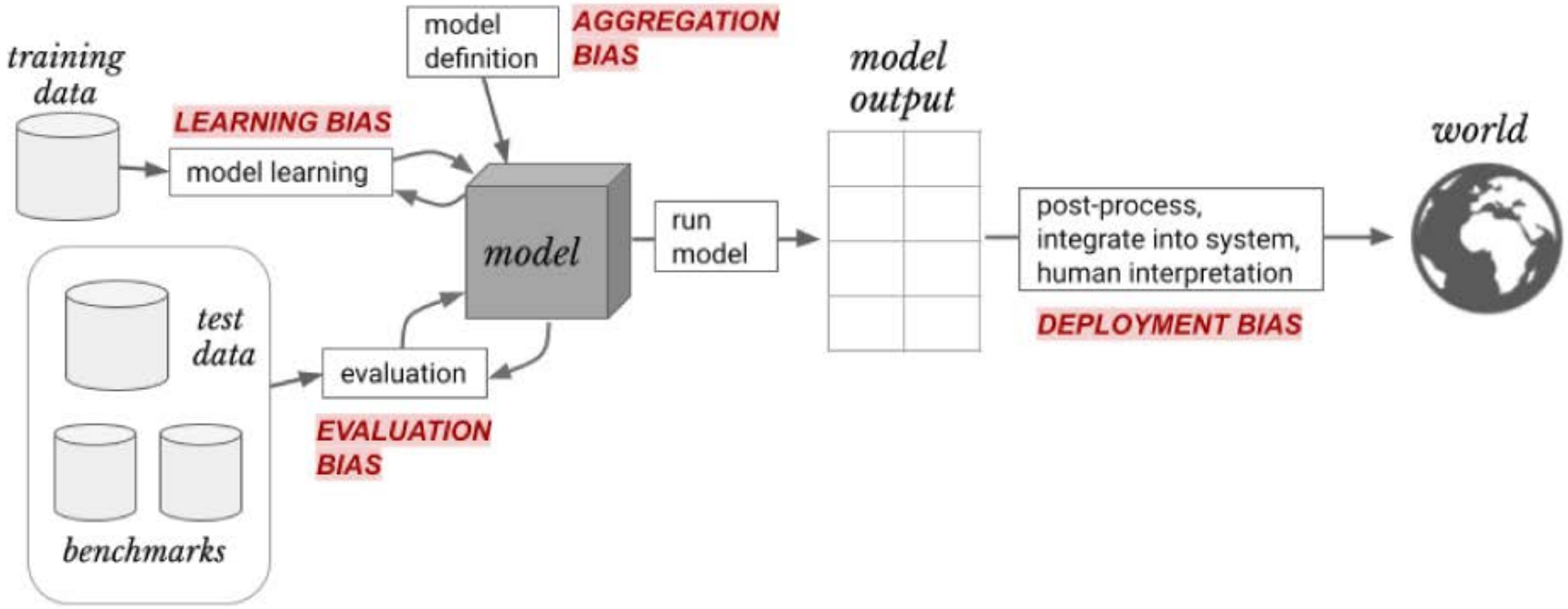
Aggregation bias

- Aggregation bias arises when a one-size-fits-all model is used for data in which there are underlying groups or types of examples that should be considered differently. Underlying aggregation bias is an assumption that the mapping from inputs to labels is consistent across subsets of the data. In reality, this is often not the case.
  A particular dataset might represent people or groups with different backgrounds, cultures or norms, and a given variable can mean something quite different across them. Aggregation bias can lead to a model that is not optimal for any group, or a model that is fit to the dominant population (e.g., if there is also representation bias).
- *Examples:*
  - Don't use a single model across multiple groups
    - Diabetes across ethnicities and gender
    - Must factor different subgroups into account when aggregating data/modeling

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Bias in ML pipeline - 2021



(b) Model Building and Implementation

Harini Suresh and John V. Guttag. "A framework for understanding sources of harm throughout the machine learning life cycle". In Equity and access in algorithms, mechanisms, and optimization (pp. 1-9), 2021.

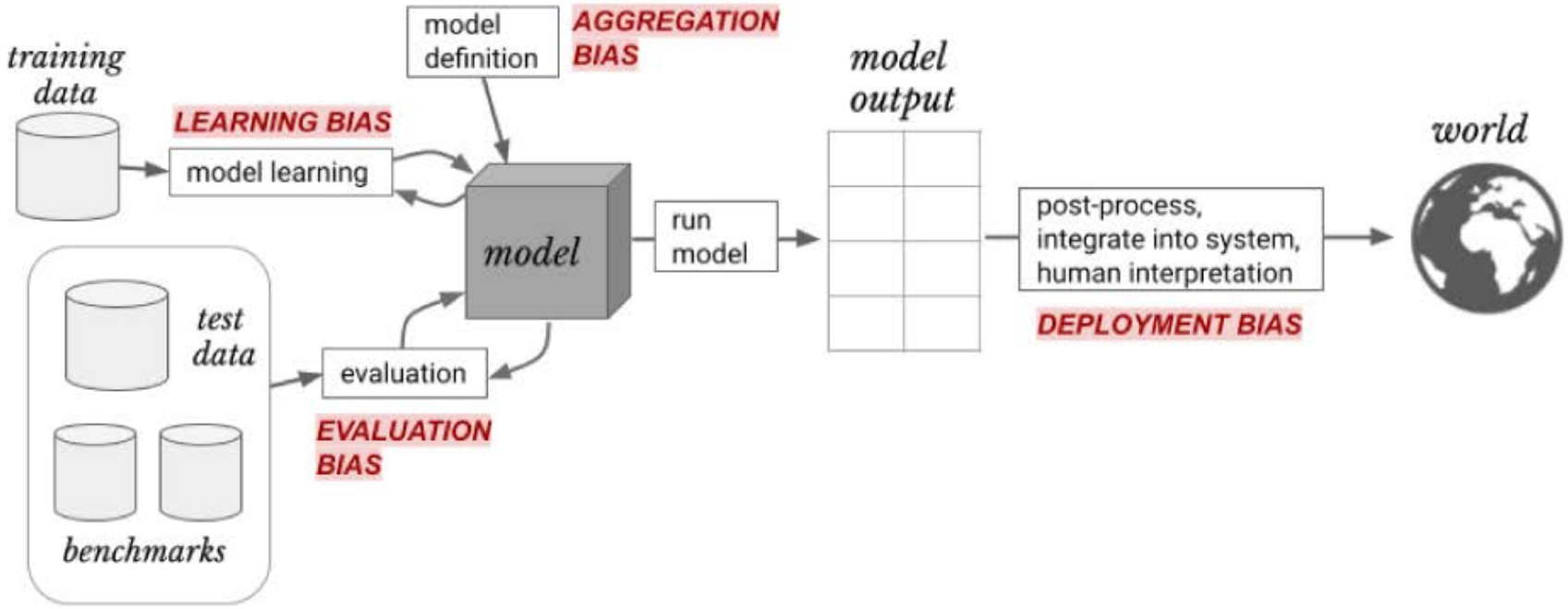
Learning bias

- Learning bias arises when modeling choices amplify performance disparities across different examples in the data. For example, an important modeling choice is the objective function that an ML algorithm learns to optimize during training. Typically, these functions encode some measure of accuracy on the task (e.g., cross-entropy loss for classification problems or mean squared error for regression problems). However, issues can arise when prioritizing one objective (e.g., overall accuracy) damages another (e.g., disparate impact). For example, minimizing cross-entropy loss when building a classifier might inadvertently lead to a model with more false positives than might be desirable in many contexts.
- *Examples:*
  - Don't use a single model across multiple groups
    - Diabetes across ethnicities and gender
    - Must factor different subgroups into account when aggregating data/modeling

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Bias in ML pipeline - 2021



(b) Model Building and Implementation

Harini Suresh and John V. Guttag. "A framework for understanding sources of harm throughout the machine learning life cycle". In Equity and access in algorithms, mechanisms, and optimization (pp. 1-9), 2021.

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Deployment bias

- Deployment bias arises when there is a mismatch between the problem a model is intended to solve and the way in which it is actually used. This often occurs when a system is built and evaluated as if it were fully autonomous, while in reality, it operates in a complicated sociotechnical system moderated by institutional structures and human decision-makers.
  In some cases, for example, systems produce results that must first be interpreted by human decision-makers. Despite good performance in isolation, they may end up causing harmful consequences because of phenomena such as automation or confirmation bias.
- *Examples:*
  - Algorithmic risk assessment tools in the criminal justice context are models intended to predict a person's likelihood of committing a future crime.

Bias Metrics

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Basic fairness metric terminology

- **Sensitive attribute: a.k.a. A "Protected field" is an attribute that may need to comply with a particular fairness metric.**
  It's important to note that not all attributes are sensitive by definition.
  For example, US federal law protects individuals from discrimination or harassment based on the following nine protected classes: sex (including sexual orientation and gender identity), race, age, disability, color, creed, national origin, religion, or genetic information (added in 2008).
- **Proxy attribute: This means attributes that correlate to a sensitive attribute.**
  For example, let's assume that for a specific use case, race and ethnicity are sensitive attributes. The dataset may also contain highly correlated fields for these two. For example, an individual's postal code might be highly correlated with them, making it a proxy attribute to the sensitive attribute value.

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Basic fairness metric terminology

- **Parity: A parity measure is a simple observational criterion that requires the evaluation metrics to be independent of the salient group A.**
  Such measures are static and don't take changing populations into account. If a parity criterion is satisfied, it does not mean that the algorithm is fair. These criteria are reasonable for surfacing potential inequities and are particularly useful in monitoring live decision-making systems.
- **Confusion matrix: Fairness metrics are most commonly applied to classification use cases.**
  One should be familiar with the basics of a confusion matrix to understand the basic mathematics behind the most common fairness metrics. In a binary classification case, we will refer to one of the predicted classes as "Positive" while the other will be "Negative." Given this configuration, common performance metrics can be derived. To generalize for a multiclass classification use case, we refer to the sensitive class (meaning a specific value out of the possible classification classes) as positive, while all others are negative.

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Confusion matrix and metrics

		True condition			
	Total population	Condition positive	Condition negative	$= \frac{\sum \text{Condition positive}}{\sum \text{Total population}}$ Prevalence	Accuracy (ACC) = $\frac{\sum \text{True positive} + \text{True negative}}{\sum \text{Total population}}$
Predicted condition	Predicted condition positive	True positive	False positive Type I error	Positive predictive value (PPV), precision = $\frac{\sum \text{True positive}}{\sum \text{Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\sum \text{False positive}}{\sum \text{Predicted condition positive}}$
	Predicted condition negative	False negative Type II error	True negative	False omission rate (FOR) = $\frac{\sum \text{False negative}}{\sum \text{Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\sum \text{True negative}}{\sum \text{Predicted condition negative}}$
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power $= \frac{\sum \text{True positive}}{\sum \text{Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\sum \text{True negative}}{\sum \text{Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR-}}$
		False negative rate (FNR), Miss rate $\frac{\sum \text{False negative}}{\sum \text{Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\sum \text{True negative}}{\sum \text{Condition negative}}$	Negative likelihood ratio (LR-) $= \frac{\text{FNR}}{\text{TNR}}$	F_1 score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$

To view the complete list of possible metrics, see https://en.wikipedia.org/wiki/Confusion_matrix.

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Fairness Metrics

Setup & notation

- To illustrate fairness metrics in practice, we will use two real-life use cases of a loan approval process and an additional use case where the model attempts to predict the likelihood of an individual being convicted of criminal activity and sent to jail. **A binary classification** model predicts if we should give **a positive answer, i.e., loan approval/going to jail**, and, for our examples, **the protected attribute will be gender**.
- Given this setup, we will use the following notation:
  - A = possible values of the sensitive attributes (e.g., gender)
  - R = the output/prediction of the algorithm (e.g., positive R=1)
  - Y = the ground truth

The graph illustrations of each example show the threshold of positive model outputs for each group, where every output above the threshold will indicate a positive outcome.

To view the complete list of possible metrics, see https://en.wikipedia.org/wiki/Confusion_matrix.

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Metric 1: Group unawareness

- Group unawareness means that we don't want the gender of a person to impact in any way our decision to approve or reject a loan. In simple, intuitive words, we don't want the model to "use" gender as an attribute.
- Group unawareness is a straightforward and intuitive way to define fairness. An ML equivalent term to such a condition could be feature attribution.
- When we refer to unawareness, we are technically referring to the fact that we want 0 attribution for gender attributes.
- This can easily be measured using interpretation tools like *Lime*, *SHAP*, etc.

To view the complete list of possible metrics, see https://en.wikipedia.org/wiki/Confusion_matrix.

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Metric 1: Group unawareness

- There may be a "proxy sensitive attribute" in the data, which means attributes that correlate to a sensitive attribute. For example, an individual's postal code may be a proxy for income, race, or ethnicity. So even if we willfully ignore the direct sensitive group, we may still have hidden awareness.
- The training data may contain historical biases. For example, women's work histories are more likely to have gaps. Which may lower their loan requests' probability of being approved.

To view the complete list of possible metrics, see https://en.wikipedia.org/wiki/Confusion_matrix.

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Metric 2a: Demographic parity

A = possible values of the sensitive attributes (e.g., gender)
R = the output/prediction of the algorithm (e.g., positive R=1)
Y = the ground truth

- If we want to ensure the same approval rate for male and female applicants, we can use demographic parity. Demographic parity states that the proportion of each segment of a protected class (e.g., gender) should receive a positive outcome at equal rates.
- Let's say A is a protected class (in our case, sex):

$$\forall w, m \in A \qquad P\left(R = 1 | A = w\right) = P\left(R = 1 | A = m\right)$$

- Another way to look at it is to require that the prediction be statistically independent of the protected attribute.

$$\forall a \in A \qquad P\left(R = 1 | A = a\right) = P\left(R = 1\right), \forall a \in A$$

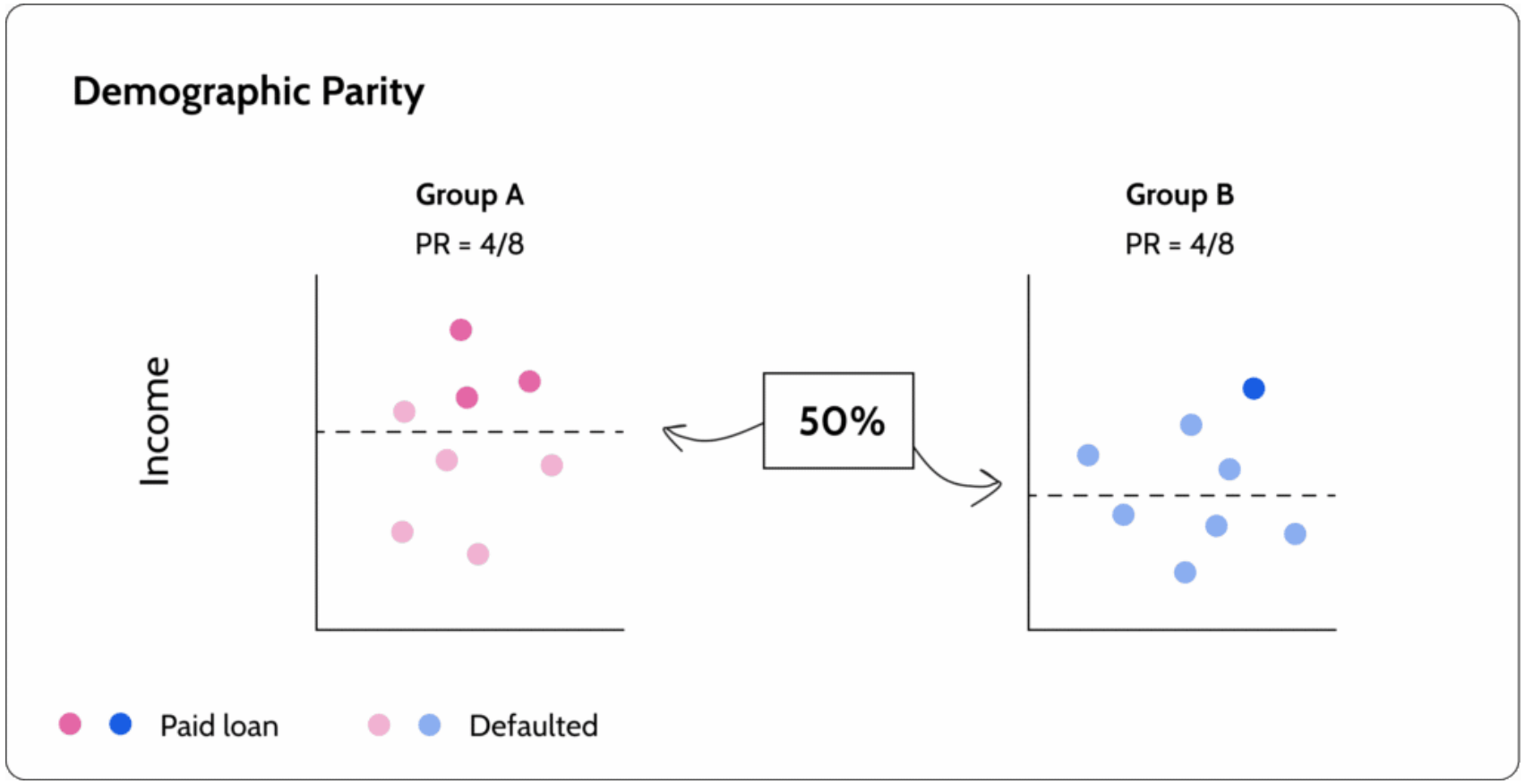
or

$$P\left(R = 0 | A = a\right) = P\left(R = 0\right), \forall a \in A$$

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Metric 2a: Demographic parity



In this illustration, we can see that half of each group is approved.

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Metric 2b: Disparate impact

A = possible values of the sensitive attributes (e.g., gender)
R = the output/prediction of the algorithm (e.g., positive R=1)
Y = the ground truth

- **Disparate impact is a variation of demographic parity.** The calculation is the same as demographic parity, but instead of aiming for an equal approval rate, it aims to achieve a higher-than-specified ratio. It is commonly used when some privileged group is in play (a group that has a higher probability of getting a positive answer when there is no fairness problem).
  For example, let's say that people with high incomes have a higher likelihood of returning the loan. On the face of things, it's OK for this group to have a higher approval rate than people with low income.

$$\left| \frac{P(R=1|A=unprivileged)}{P(R=1|A=privileged)} \right| \geq 1-\epsilon, \epsilon \in [0,1)$$

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Metric 2b: Disparate impact

- On the flip side, we want to ensure that the approval ratio between these groups does not become too high. The industry standard here is a 4/5 rule. If the unprivileged group receives a positive outcome of less than 80% of the proportion of the privileged group, this is a disparate impact.

- Enforcing a specific ratio between groups may result in very qualified applicants not being approved or applicants with a low probability of returning the loan to be approved in the name of maintaining the ratio.

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

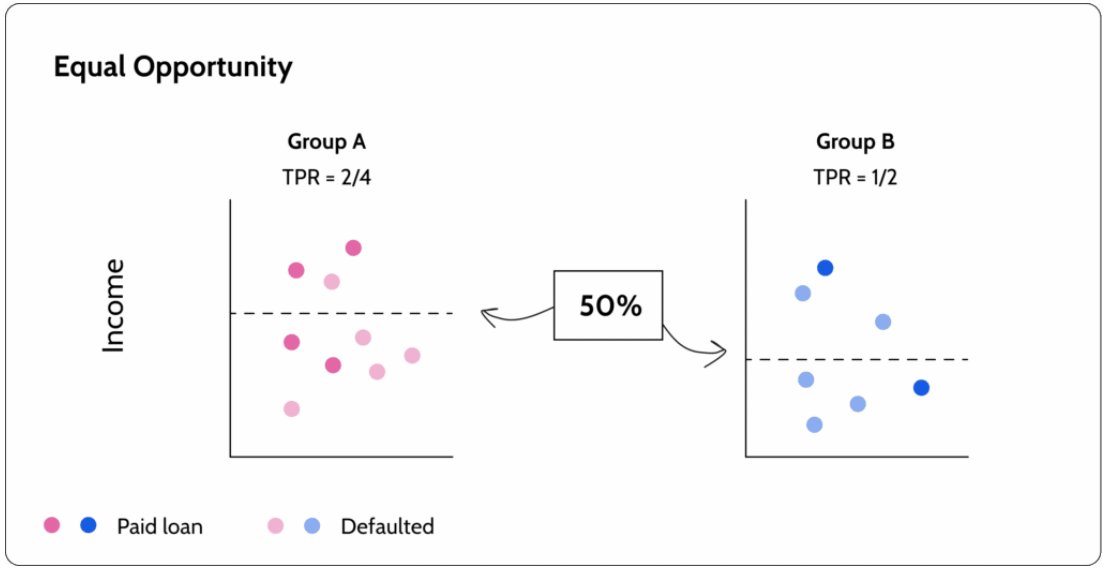
Metric 3: Equal opportunity

A = possible values of the sensitive attributes (e.g., gender)
R = the output/prediction of the algorithm (e.g., positive R=1)
Y = the ground truth

- Equal opportunity examines the true positive rate (TPR). Meaning it looks at the number of people that returned the loan to see the ratio of them that the model approved.

$$P(R = +|Y = +, A = a) = P(R = +|Y = +, A = b), \forall a, b \in A$$

In the illustration, we can see that both groups have the same TPR.
Meaning that we are basing our fairness calculation on the actual, not the prediction output.
This helps us mitigate some of the overbalancing risks inherent to demographic parity.



UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

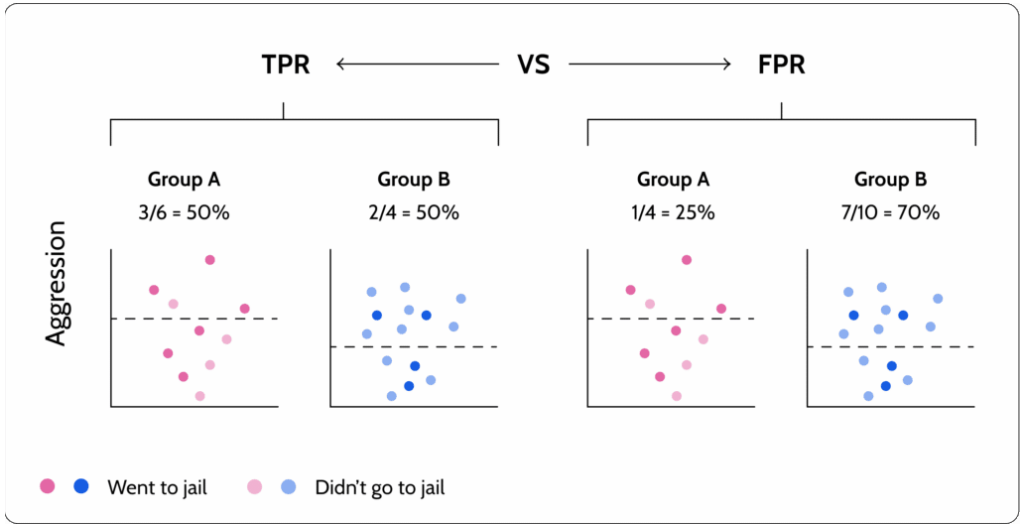
Metric 3: Equal opportunity

- It may not help close an existing gap between two groups:

  Let's look at a model that predicts applicants who qualify for a job. Let's say Group A has 100 applicants, and 58 are qualified. Group B also has 100 applicants, but only 2 are qualified. If the company decides it needs 30 applicants, the model will offer 29 applicants from Group A, and only 1 from Group B as the TPR for both groups is ½.

Metric 4: Equal odds

- To understand why equal opportunity is not good enough in some cases, let's look at a use case where we have a model that predicts if someone is a criminal **where the protected attribute is ethnicity**.
- Based on equal opportunity, the ratio of criminals that go to jail is equal between two ethnicities (see image).

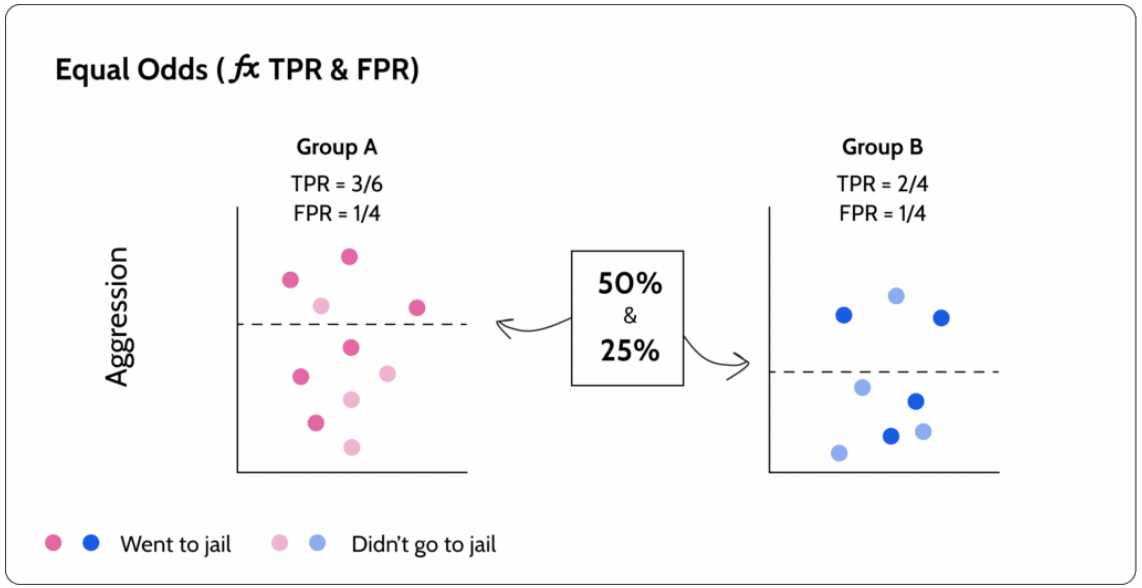


- However, we see that the FPR between the groups is different. The implication is that more innocent people from a specific sensitive group are sent to jail. In cases where we want to enforce equal TPR and FPR, we will use equal odds.
- **Equal odds is a more restrictive version of equal opportunity.** In addition to looking at the TPR, it also looks at the FPR to find an equilibrium between them.

Metric 4: Equal odds

A = possible values of the sensitive attributes (e.g., ethnicity)
R = the output/prediction of the algorithm (e.g., positive R=1)
Y = the ground truth

$$P(R = +|Y = y, A = a) = P(R = +|Y = y, A = b), y \in \{+, -\}, \forall a, b \in A$$



- Equal odds is a very restrictive metric because it tries to achieve equal TPR and FPR for each group. Therefore, it may cause the model to have poor performance.

UNIVERSITÉ CÔTE D'AZUR

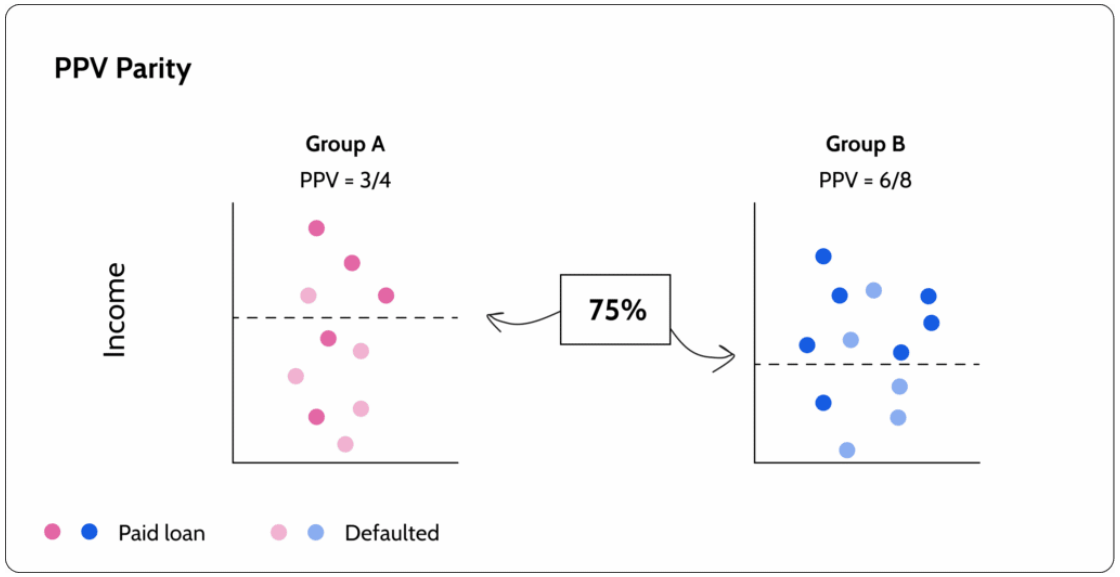
univ-cotedazur.fr

Metric 5: PPV-parity

A = possible values of the sensitive attributes (e.g., ethnicity)
R = the output/prediction of the algorithm (e.g., positive R=1)
Y = the ground truth

- **PPV-parity (positive predicted value) equalizes the chance of success, given a positive prediction.** In our example, we ensure that the ratio of people that actually return the loan out of all the people the model approved is the same in both groups. Meaning we want to see that the two groups have the same positive predicted value (PPV).

$$P\left(Y = +|R = +, A = a\right) = P\left(Y = +|R = +, A = b\right) \forall a, b \in A$$



UNIVERSITÉ CÔTE D'AZUR

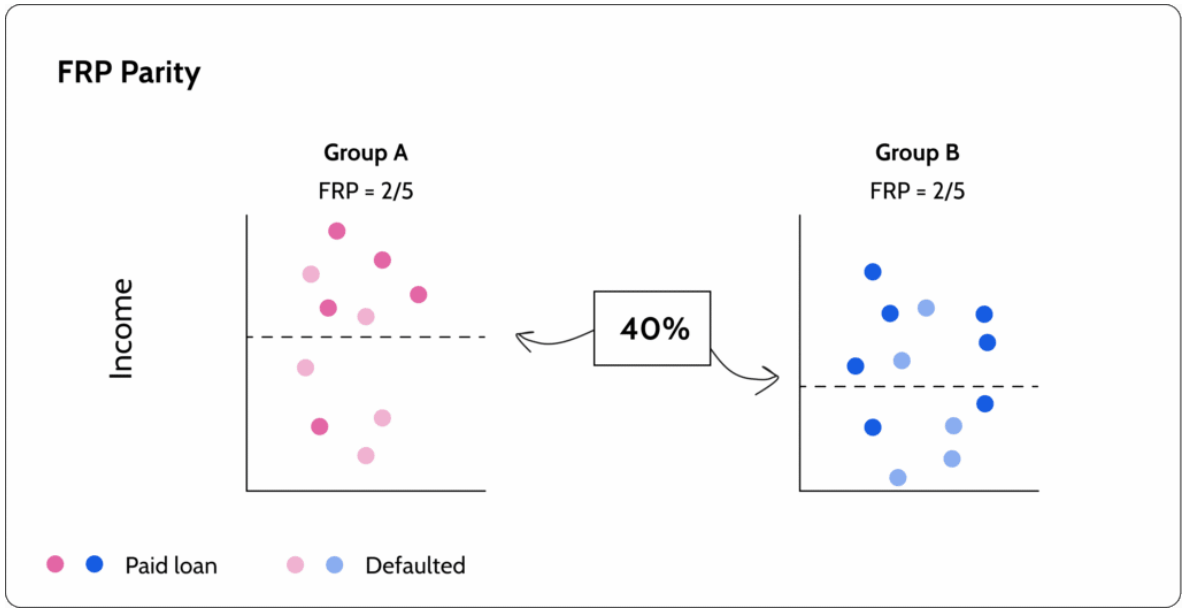
univ-cotedazur.fr

Metric 6: FPR-parity

A = possible values of the sensitive attributes (e.g., ethnicity)
R = the output/prediction of the algorithm (e.g., positive R=1)
Y = the ground truth

- **FPR-parity (false positive rate) is the exact opposite, it wants to ensure that the two groups have the same false positive rate (FPR).** In our example, we ensure that the ratio of people that the model approved but defaulted on their loan out of all the people that defaulted on their loan is the same in both groups.

$$P\left(R = +|Y = -, A = a\right) P\left(R = +|Y = -, A = b\right) \forall a, b \in A$$



UNIVERSITÉ CÔTE D'AZUR

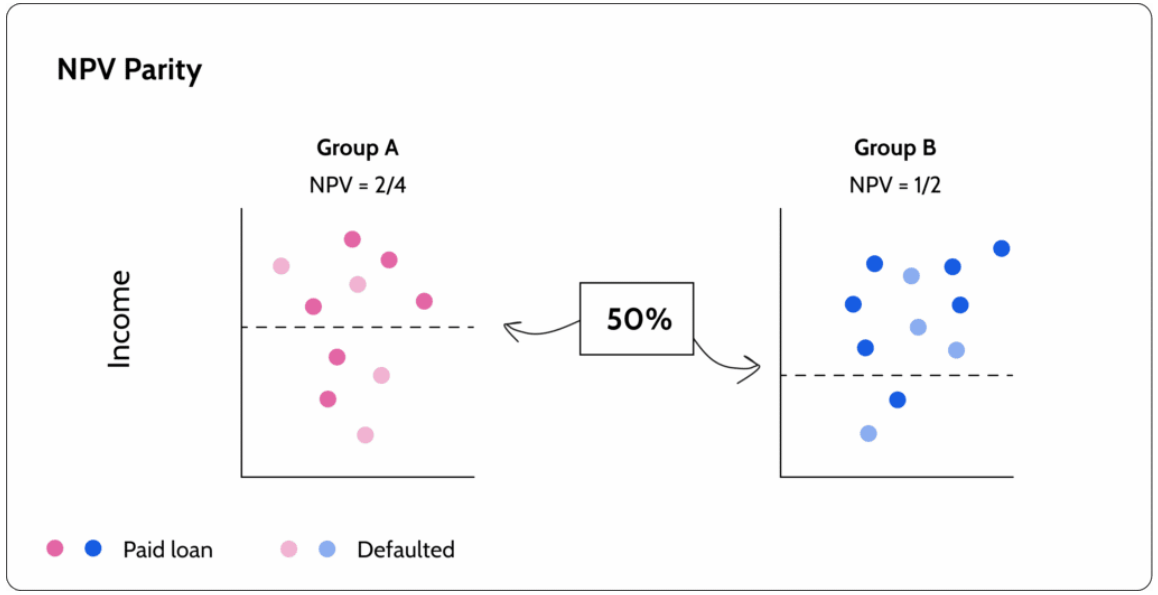
univ-cotedazur.fr

Metric 7: NPV-parity

A = possible values of the sensitive attributes (e.g., ethnicity)
R = the output/prediction of the algorithm (e.g., positive R=1)
Y = the ground truth

- **NPV-parity (negative predicted value) says that we should ensure that the ratio of correctly rejecting people out of all the people the model has rejected is the same for each group.** We want the two groups to have the same negative predicted value (NPV).

$$P\left(R=-|Y=-,A=a\right)=P\left(R=-|Y=-,A=b\right)\forall a,b\in A$$



UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Open-source fairness metrics libraries

Open source library	Notes
AIF360	Provides a comprehensive set of metrics for datasets and models to test for biases and algorithms to mitigate bias in datasets and models.
Fairness Measures	Provides several fairness metrics, including difference of means, disparate impact, and odds ratio. It also provides datasets, but some are not in the public domain and require explicit permission from the owners to access or use the data.
FairML	Provides an auditing tool for predictive models by quantifying the relative effects of various inputs on a model's predictions, which can be used to assess the model's fairness.
FairTest	Checks for associations between predicted labels and protected attributes. The methodology also provides a way to identify regions of the input space where an algorithm might incur unusually high errors. This toolkit also includes a rich catalog of datasets
Aequitas	This is an auditing toolkit for data scientists as well as policymakers; it has a Python library and website where data can be uploaded for bias analysis. It offers several fairness metrics, including demographic, statistical parity, and disparate impact, along with a "fairness tree" to help users identify the correct metric to use for their particular situation. Aequitas's license does not allow commercial use.
Themis	An open-source bias toolbox that automatically generates test suites to measure discrimination in decisions made by a predictive system.
Themis-ML	Provides fairness metrics, such as mean difference, some bias mitigation algorithms, additive counterfactually fair estimator, and reject option classification.
Fairness Comparison	Includes several bias detection metrics as well as bias mitigation methods, including disparate impact remover, prejudice remover, and two-Naive Bayes. Written primarily as a test bed to allow different bias metrics and algorithms to be compared in a consistent way, it also allows additional algorithms and datasets.

And many others: Fairlens,…

Mitigation

Mitigation techniques

Pre-processing

Agnostic to

ML approach

In-processing

Modification of the algorithm to remove discrimination

Some promising techniques are based on adversarial networks

Post-processing

Agnostic to

ML approach

Policies

Datasheets and data statements

Impact assessments

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Policies & Responsibility

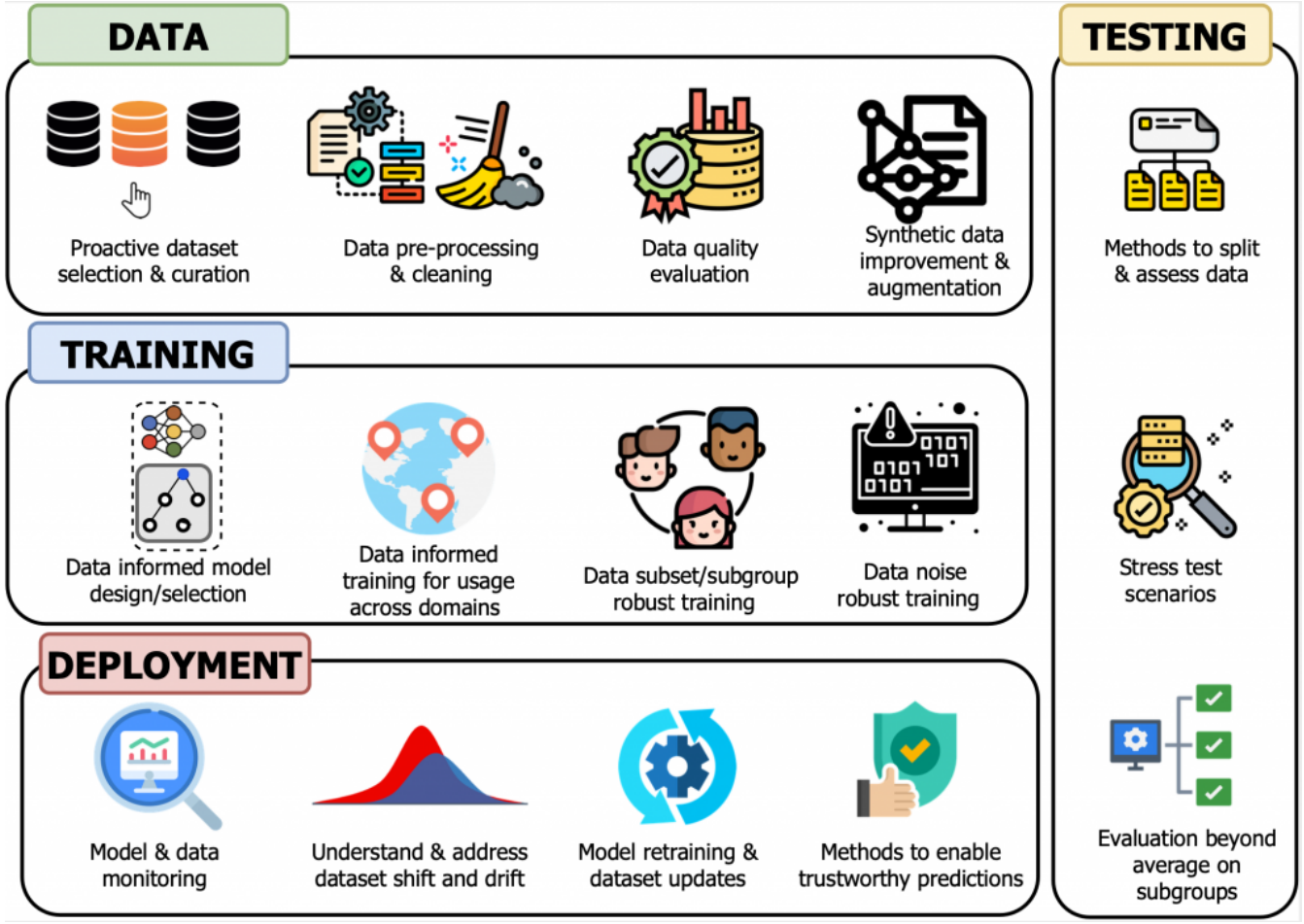
- Amazon is responsible if its facial recognition system is not used according to guidelines
  - Guidelines given, who is responsible for ensuring the guidelines are followed?

- Because of abuse and false-positives, predictive policing and law-enforcement should be prohibited from using this software
  - Safety first

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

From Model-centric AI to Data-centric AI



From Prof. Michela Van Shaar website: https://www.vanderschaar-lab.com/data-centric-ai/

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Overview

- Context
- Definition & Taxonomy
- Bias in ML
- **Ethics In AI is not only about fairness**

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

ETHICS IN AI IS NOT ONLY ABOUT FAIRNESS

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Beyond the question of fairness…

Timnit Gebru

https://www.youtube.com/watch?v=T2oZvzgriII

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Beyond the question of fairness…

Lucile Sasstelli
Professor at University Cote d'Azur,
I3S CNRS Lab
Scientific Director of EFELIA
(French School of AI in Cote d'Azur)

Is data fixable? On the need of socially-informed practices in ML research and education (part 1)
Part 1: Deployment failures and approaches to data

Is data fixable? On the need of socially-informed practices in research and education (part 2)
Part 2: A more holistic perspective on data creation and expectations

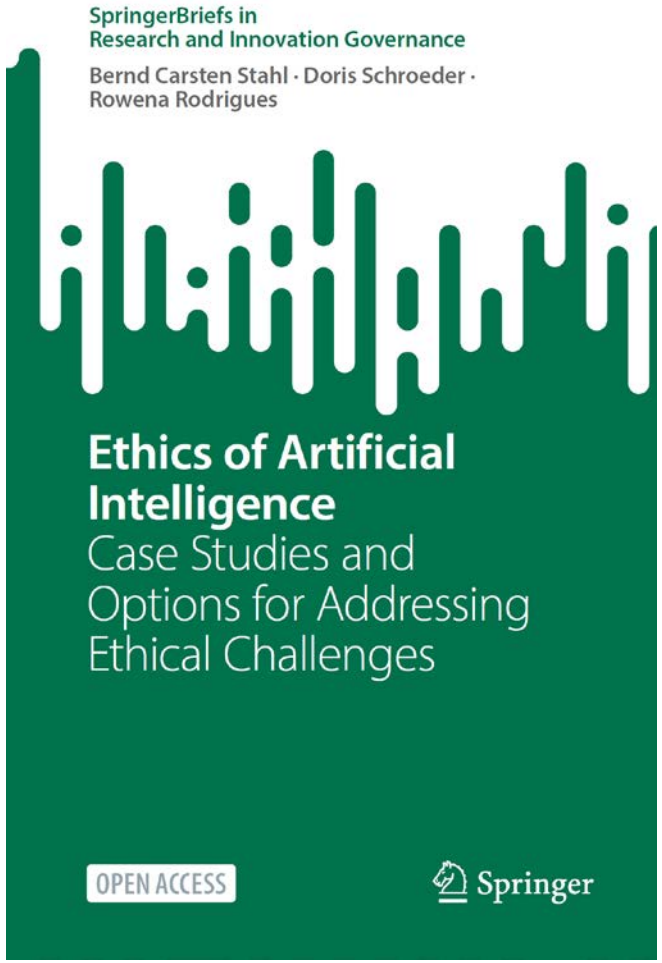
Is data fixable? On the need of socially-informed practices in ML research and education (part 3)
Part 3: AI ethics and our ML education practices

https://webusers.i3s.unice.fr/~sassatelli/

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

The Backbone of the module



https://link.springer.com/book/10.1007/978-3-031-17040-9

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Table of contents

UNIVERSITÉ
CÔTE D'AZUR

Any Question?