«Human-Centered Data Science»

Approaching Fairness in Data Science Workflows

Prof. Dr. Claudia Müller-Birn

Human-Centered Computing, Institute of Computer Science

Freie Universität Berlin

June 9, 2022

Lecture Overview

Recap

Defining Fairness and Introducing a Scenario (data set, learner protected variable)

Approaches to Measure Fairness (Statistical-based measures)

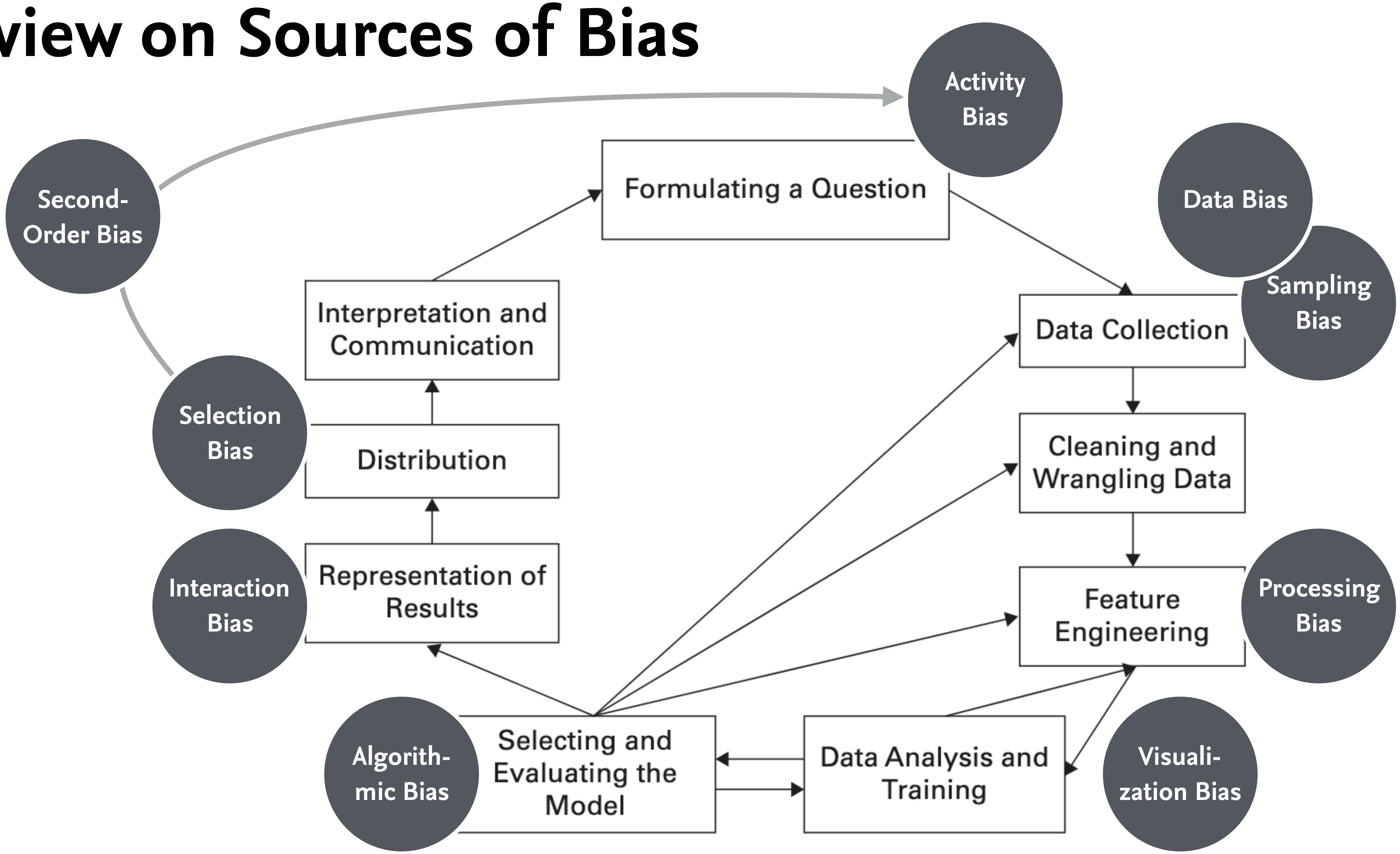
☕ **Break**

Approaches to Measure Fairness (similarity-based measures, causal reasoning)

Case Study Reflection

Recap

Overview on Sources of Bias



Protected Variables in Law

Protected Variables	Art. 3 GG	§§ 1 u.a. AGG	Erwg. 71 DSGVO	Art. 9 DSGVO
"Race" or Ethnicity	yes	yes	yes	yes
Ancestry, home, origin	yes			
Gender	yes	yes		
Language	yes			
Political opinion or attitude and other attitudes	yes		yes	yes
Religion and world view	yes	yes	yes	yes
Disability	yes	yes		
Age		yes		
Union membership	yes*		yes	yes
Genetic characteristics or predispositions and health status	yes		yes	yes
Biometrics				yes
Sexual life, sexual identity or orientation		yes	yes	yes

Orwat, C. (2019), Diskriminierungsrisiken durch Verwendung von Algorithmen , Nomos , Baden-Baden.

Discrimination: Treatment vs Impact

As seen, modern legal frameworks offer various levels of protection for being discriminated by belonging to a particular class of: gender, age, ethnicity, nationality, disability, religious beliefs, and/or sexual orientation.

Often two concepts are differentiated:

- **Disparate treatment:** Treatment depends on class membership
- **Disparate impact:** Outcome depends on class membership

Discrimination from a Computer Science Perspective

The focus is on quantification. In the context of an algorithm generating a prediction:

» Predictions for people with similar non-protected attributes should be similar

» Differences should be mostly explainable by non-protected attributes

Two basic frameworks for measuring discrimination:

» **Discrimination at the individual level:** consistency or individual fairness

» **Discrimination at the group level:** statistical parity

Hajian, Sara, Francesco Bonchi, und Carlos Castillo. 2016. Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining. KDD '16. the 22nd ACM SIGKDD International Conference. New York, New York, USA: SIGMOD, ACM Special Interest Group on Management of Data. https://doi.org/10.1145/2939672.2945386.
Žliobaitė I. 2015. A survey on measuring indirect discrimination in machine learning. arXiv pre-print.

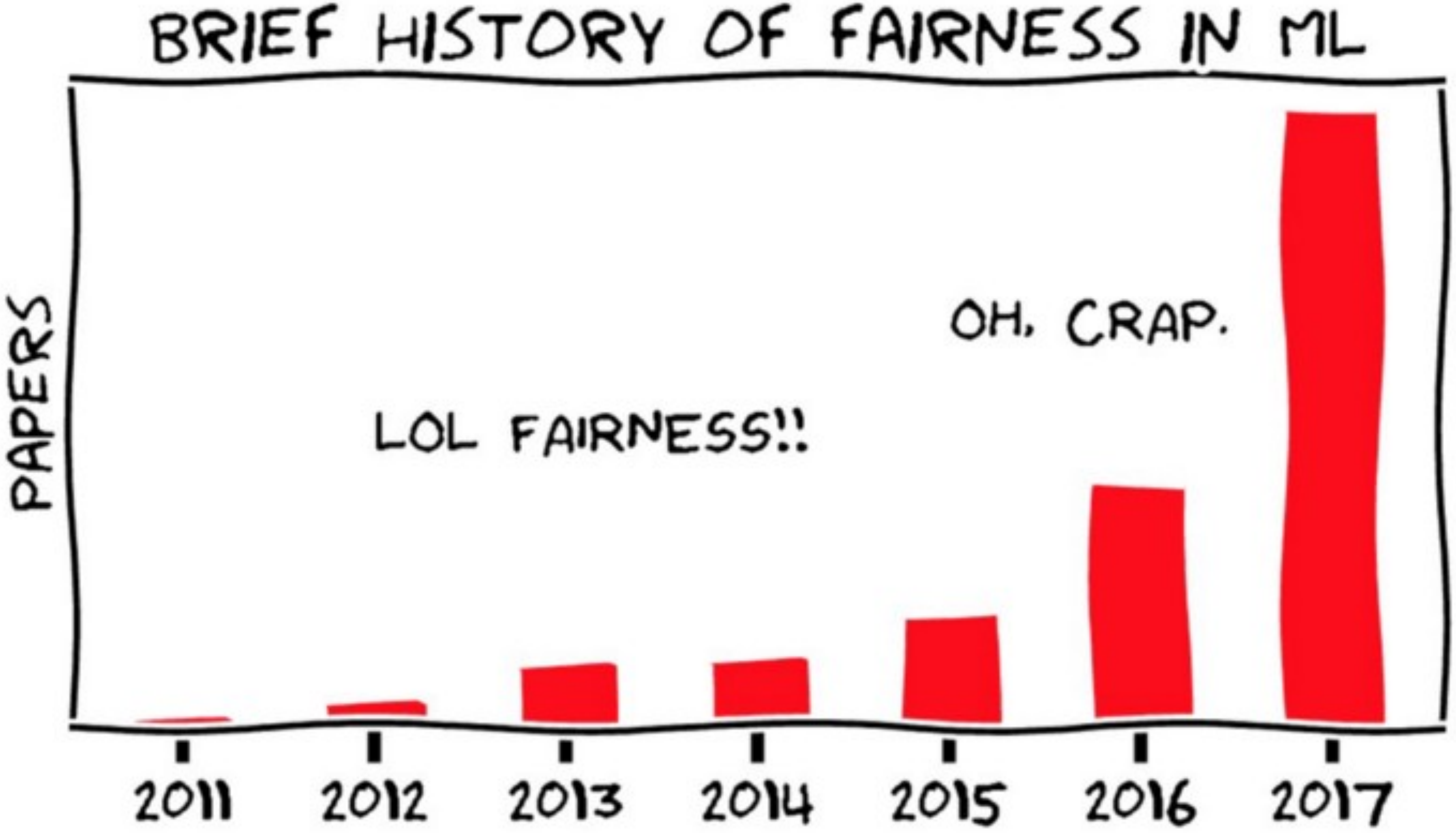


Diagram created by Moritz Hardt (unknown first time use)

Definitions of Fairness

> “In machine learning, a given algorithm is said to be fair, or to have fairness, if its results are independent of given variables, especially those considered sensitive, such as the traits of individuals which should not correlate with the outcome (i.e. gender, ethnicity, sexual orientation, disability, etc.).”

Definition taken from Wikipedia https://en.wikipedia.org/w/index.php?title=Fairness_(machine_learning)&oldid=985721580

Definitions of Fairness

> “In machine learning, a given algorithm is said to be fair, or to have fairness, if its results are independent of given variables, especially those considered sensitive, such as the traits of individuals which should not correlate with the outcome (i.e. gender, ethnicity, sexual orientation, disability, etc.).”

Definition taken from Wikipedia https://en.wikipedia.org/w/index.php?title=Fairness_(machine_learning)&oldid=985721580

> “In the context of decision-making, fairness is the absence of any prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics. Thus, an unfair algorithm is one whose decisions are skewed toward a particular group of people.”

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. arXiv preprint arXiv:1908.09635.

Open-Source Libraries on Fairness

AI Fairness 360

AI Fairness 360 (IBM) Users can examine, report, and mitigate discrimination and bias in machine learning models.

Aequitas
Bias & Fairness Audit

Aequitas (Carnegie Mellon University) For both data scientists and policymakers, with a Python library and a website to upload data for bias analysis.

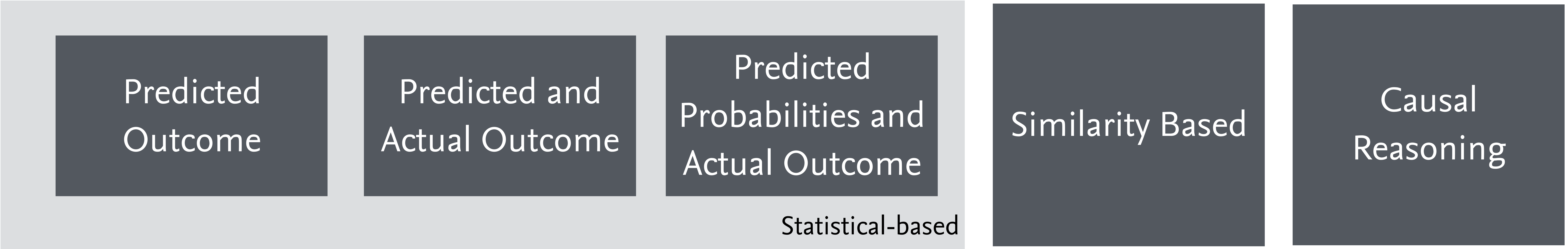
FairML

FairML Auditing tool for blackbox predictive models by quantifying the relative effects of various input on the model's predictions.

…

Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). Association for Computing Machinery, New York, NY, USA, 59–68. DOI:https://doi.org/10.1145/3287560.3287598

Overview on Approaches to Define Fairness



Predicted Outcome

Predicted and Actual Outcome

Predicted Probabilities and Actual Outcome

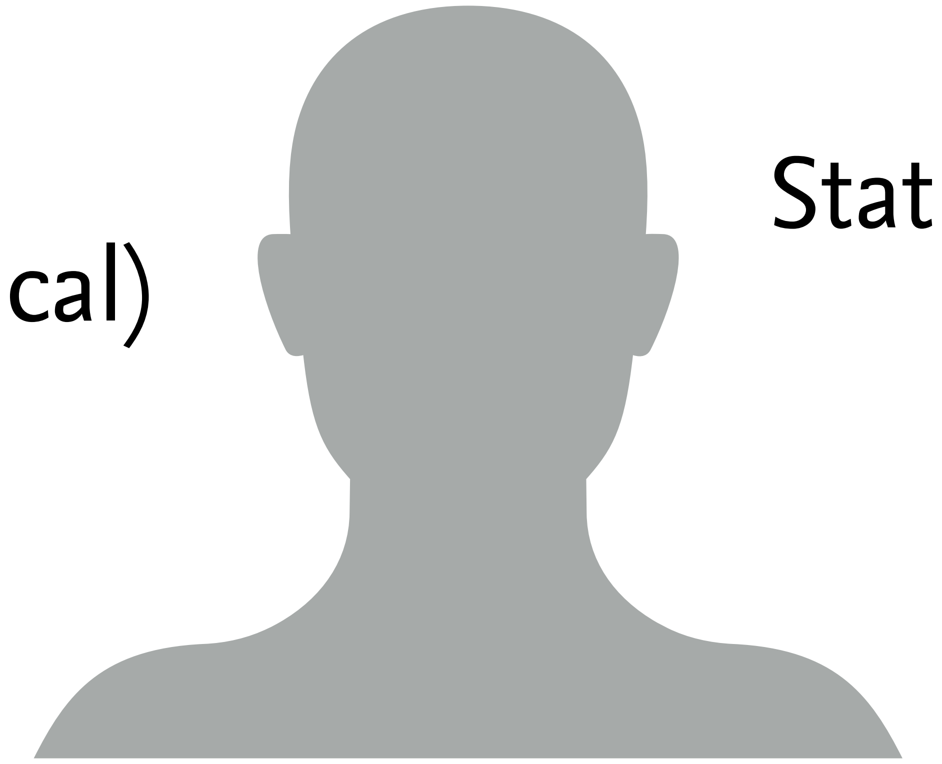
Statistical-based

Similarity Based

Causal Reasoning

Verma, S., & Rubin, J. (2018). Fairness definitions explained. FairWare@ICSE, 1–7. http://doi.org/10.1145/3194770.3194776

German Credit Dataset



Period of present residency (numerical)

Credit purpose (categorical)

Employment (categorical)

Credit duration (numerical)

Telephone (binary)

Credit amount (numerical)

Status of existing checking account (categorical)

Status of savings accounts and bonds (categorical)

Residence (categorical)

Credit history (categorical)

Number of existing credits (numerical)

Installment plans (categorical)

Property (categorical)

Installment rate (numerical)

Foreign worker (binary)

Age (numerical)

Employment length (categorical)

Personal status and gender (categorical)

Credit score (binary)

Other debtors (categorical)

Dependents (numerical)

https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29

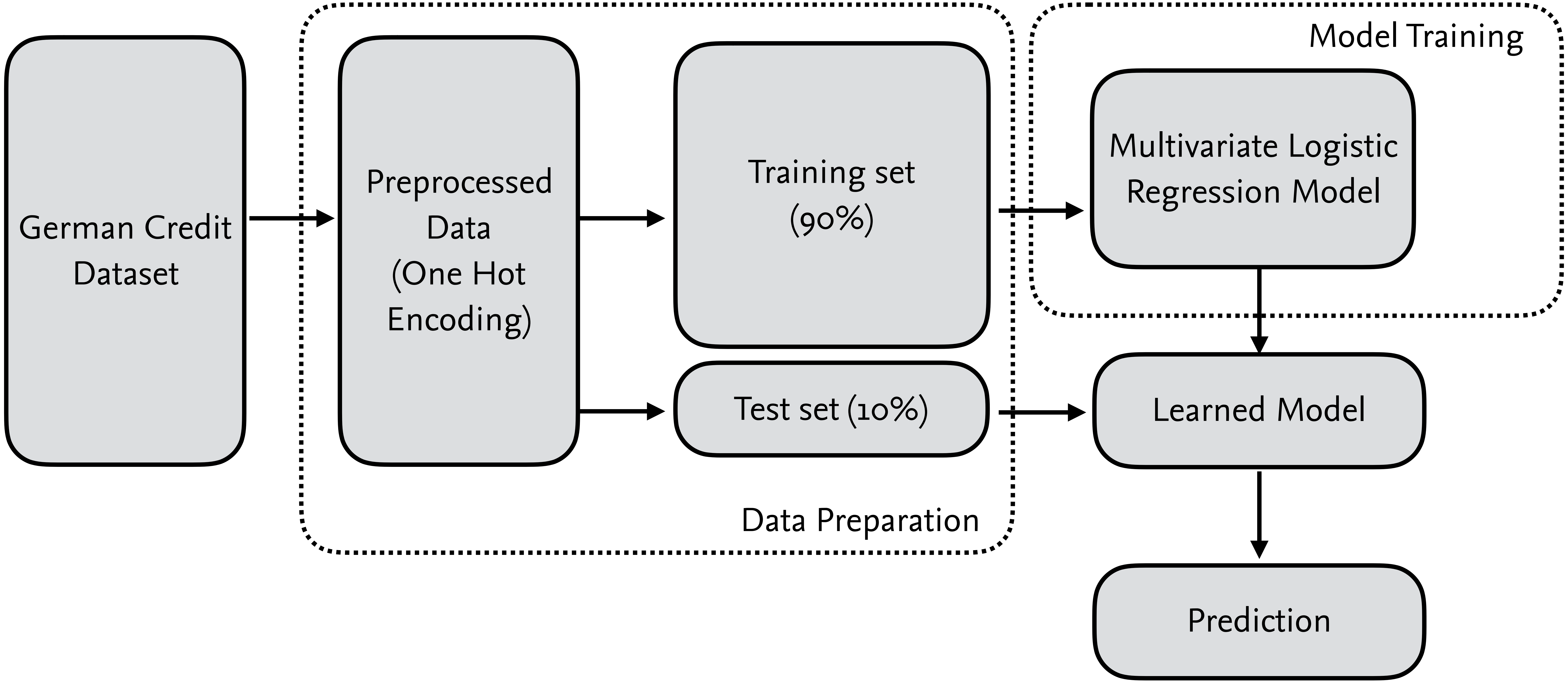
Scenario

Anna is requesting a loan amount of 1,567 DM for a duration of 12 months for the purpose of purchasing a television, with a positive checking account balance that is smaller than 200 DM, having less than 100 DM in savings account, and having one existing credit at this bank. She duly paid existing credits at the bank till now and has no other installment plan. She possesses a car and owns a house, has been living at the present residence for one year and has a registered telephone. She is a skilled employee, working in the present employment for past four years. She is a 22-year-old married female and is a German citizen. She has one dependent and no guarantors.

https://thispersondoesnotexist.com/ Imagined by a GAN (generative adversarial network) StyleGAN2 (Dec 2019) - Karras et al. and Nvidia

Is this fair compared to other requests?

Implementing the Decision Maker



Verma, S., & Rubin, J. (2018). Fairness definitions explained. FairWare@ICSE, 1–7. http://doi.org/10.1145/3194770.3194776

What is the Protected Variable?

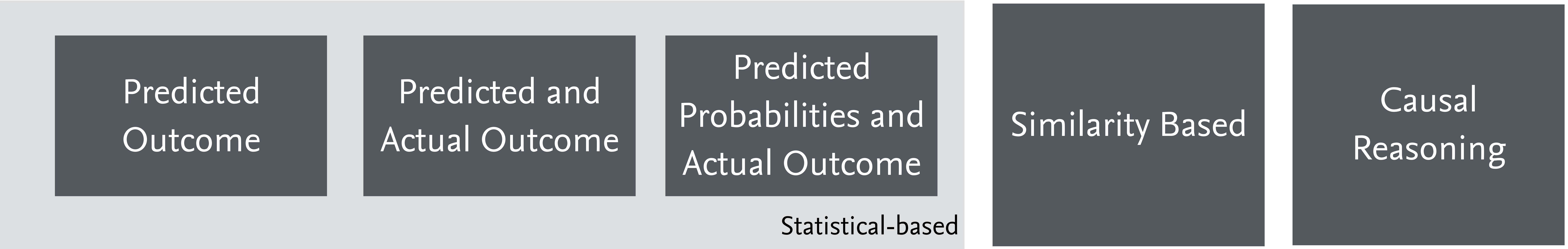
Personal *status* and *gender* (categorical)

- » male :: SINGLE :: MARRIED/WIDOWED :: DIVORCED/SEPARATED
- » female :: DIVORCED/SEPARATED/MARRIED

Attribute	**Coefficient**
Personal status and gender: single male	0.16
Personal status and gender: married male	-0.04
Personal status and gender: married/divorced female	-0.08
Personal status and gender: divorced male	-0.14

Verma, S., & Rubin, J. (2018). Fairness definitions explained. FairWare@ICSE, 1–7. http://doi.org/10.1145/3194770.3194776

Overview on Approaches to Measure Fairness



Predicted Outcome

Predicted and Actual Outcome

Predicted Probabilities and Actual Outcome

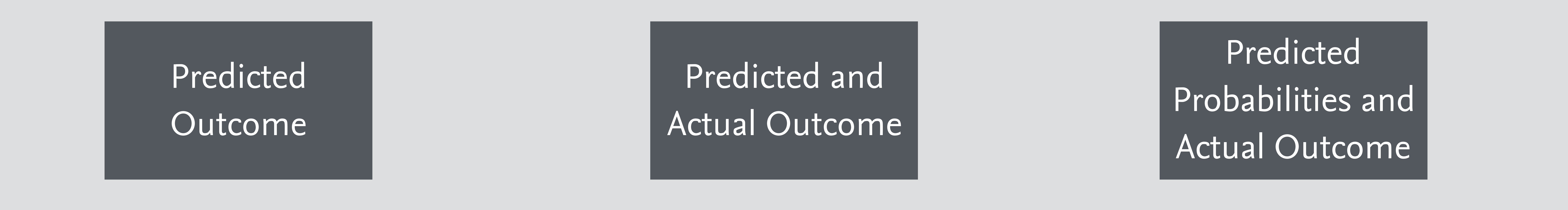
Statistical-based

Similarity Based

Causal Reasoning

Verma, S., & Rubin, J. (2018). Fairness definitions explained. FairWare@ICSE, 1–7. http://doi.org/10.1145/3194770.3194776

Statistical-based Definitions on Fairness


Predicted Outcome
Predicted and Actual Outcome
Predicted Probabilities and Actual Outcome

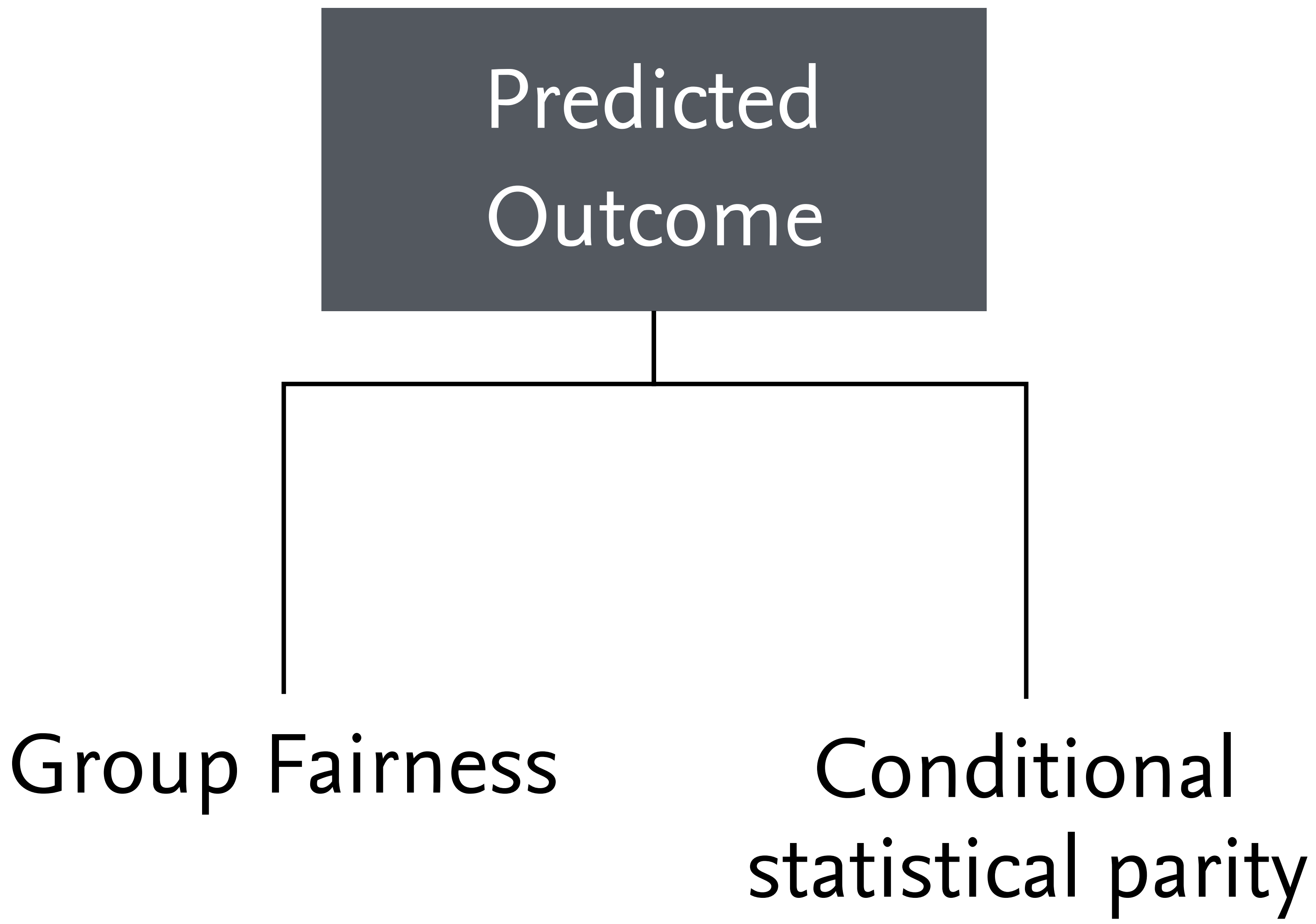
Predicted outcome definitions solely rely on a model's predictions.

Predicted and actual outcome combine a model's predictions with the true labels.

Predicted probabilities and actual outcome employ the predicted probabilities instead of the predicted outcomes.

Verma, S., & Rubin, J. (2018). Fairness definitions explained. FairWare@ICSE, 1–7. http://doi.org/10.1145/3194770.3194776

Statistical-based Definitions on Fairness



G: Protected or sensitive attribute for which non-discrimination should be established.

X: All additional attributes describing the individual.

Y: The actual classification result (here, good or bad credit score of an applicant)

S: Predicted probability for a certain classification c , P(Y=c|G,X) (here, predicted probability of having a good or bad credit score).

d: Predicted decision (category) for the individual (here, predicted credit score for an applicant - good or bad); d is usually derived from S, e.g., d = 1 when S is above a certain threshold.

Verma, S., & Rubin, J. (2018). Fairness definitions explained. FairWare@ICSE, 1–7. http://doi.org/10.1145/3194770.3194776

Group Fairness

Definition

A classifier satisfies the definition if subjects in both protected and unprotected groups have equal probability of being assigned to the positive predicted class.

Example

Probability for male and female applicants to have good predicted credit score:

$$P(d = 1 | G = m) = P(d = 1 | G = f)$$

Results from the classifier:

- $P(d = 1 | G = m) = 0.81$
- $P(d = 1 | G = f) = 0.75$

✗

Conditional Statistical Parity

Definition

A classifier satisfies the definition if subjects in both protected and unprotected groups have equal probability of being assigned to the positive predicted class, controlling for a set of legitimate factors L.

Example

Legitimate factors: credit amount, applicant's credit history, employment, and age.

Probability for male and female applicants to have good predicted credit score:

$$P(d = 1 | L = l, G = m) = P(d = 1 | L = l, G = f)$$

Results from the classifier:

» $P(d = 1 | G = m) = 0.46$

» $P(d = 1 | G = f) = 0.49$

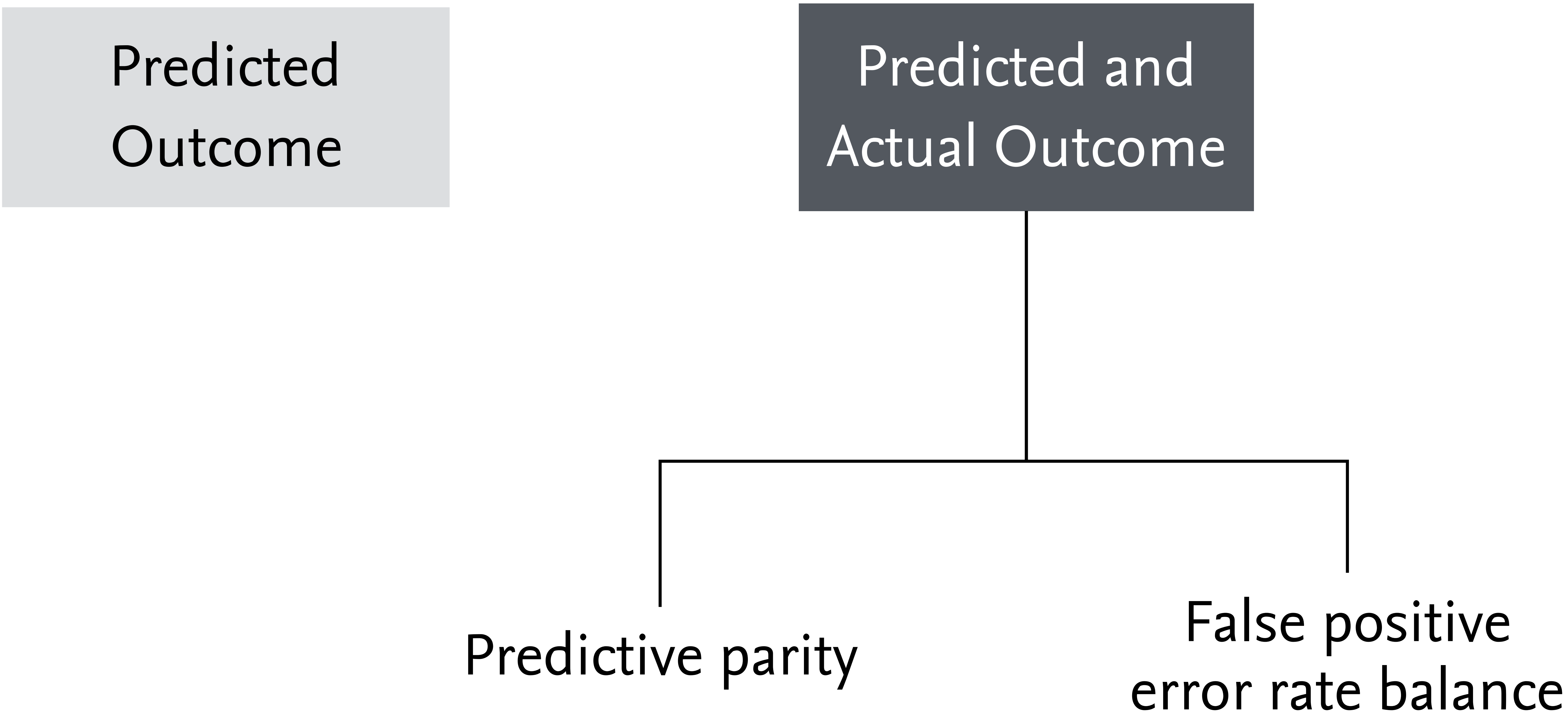
✓

Recap: Statistical Performance Metrics

True Condition
based on the labeled data

		TRUE	FALSE
Predicted Condition by the ML-pipeline	TRUE	**True Positives (TP)** $PPV = \frac{TP}{TP + FP}$	**False Positives (FP)** $FPR = \frac{FP}{FP + TN}$
	FALSE	**False Negatives (FN)**	**True Negatives (TN)**

Statistical-based Definitions on Fairness


Predicted Outcome
Predicted and Actual Outcome
Predictive parity
False positive error rate balance

Verma, S., & Rubin, J. (2018). Fairness definitions explained. FairWare@ICSE, 1–7. http://doi.org/10.1145/3194770.3194776

Predictive Parity

Definition

A classifier satisfies this definition if both protected and unprotected groups have equal PPV – the probability of a subject with positive predictive value to truly belong to the positive class.

Example

For both male and female applicants, the probability of an applicant with a good predicted credit score to actually have a good credit score should be the same:

$$P(Y = 1|d = 1, G = m) = P(Y = 1|d = 1, G = f).$$

Results:

- » $PPV_{male} = 0.73$
- » $PPV_{female} = 0.74$

✓

False Positive Error Rate Balance

Definition

A classifier satisfies this definition if both protected and unprotected groups have equal FPR – the probability of a subject in the negative class to have a positive predictive value.

Example

Probability of an applicant with an actual bad credit score to be incorrectly assigned a good predicted credit score should be the same for both male and female applicants:

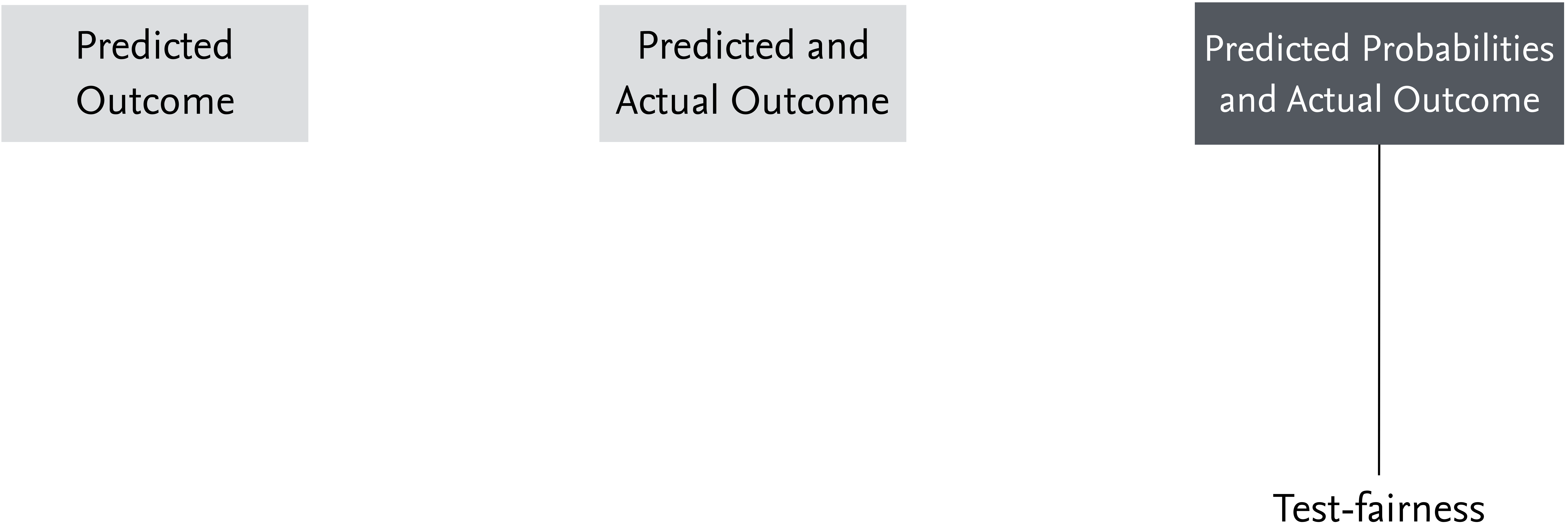
$$P(d = 1|Y = 0, G = m) = P(d = 1|Y = 0, G = f).$$

Results:

- $FPR_{male} = 0.70$
- $FPR_{female} = 0.55$

✗

Statistical-based Definitions on Fairness



Predicted Outcome

Predicted and Actual Outcome

Predicted Probabilities and Actual Outcome

Test-fairness

Verma, S., & Rubin, J. (2018). Fairness definitions explained. FairWare@ICSE, 1–7. http://doi.org/10.1145/3194770.3194776

Test Fairness

Definition

A classifier satisfies this definition if for any predicted probability score S, subjects in both protected and unprotected groups have equal probability to truly belong to the positive class.

Example

For any given predicted probability score s in [0, 1], the probability of having actually a good credit score should be equal for both male and female applicants:

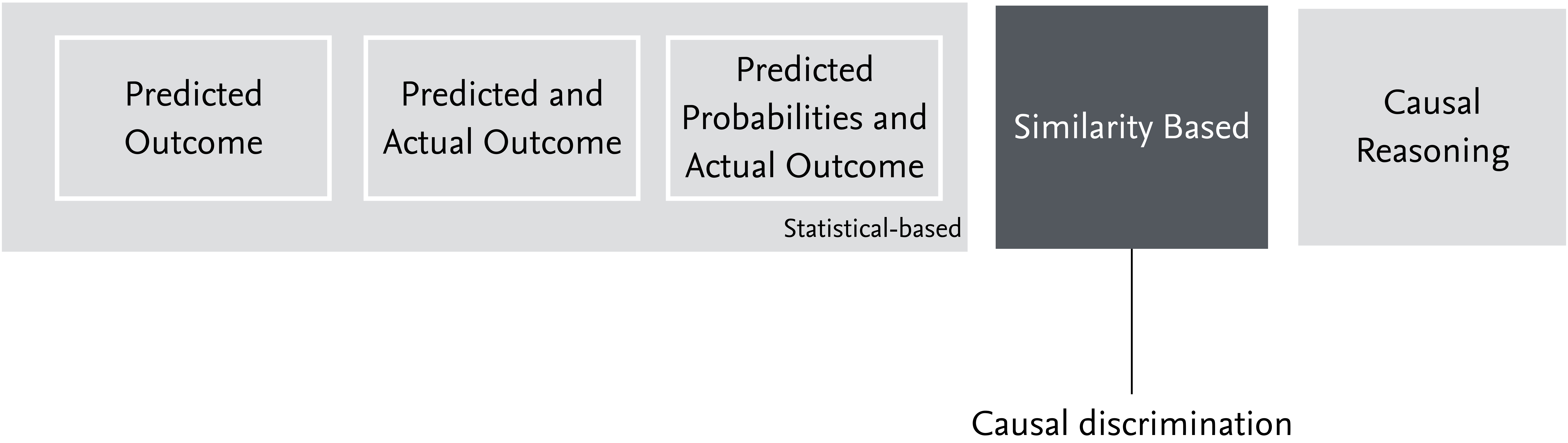
$$P(Y = 1|S = s, G = m) = P(Y = 1|S = s, G = f).$$

S	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$P(Y = 1\|S = s, G = m)$	1.0	1.0	0.3	0.3	0.4	0.6	0.6	0.7	0.8	0.8	1.0
$P(Y = 1\|S = s, G = f)$	0.5	0.3	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0

✓ ✗

5 minutes break

Similarity-based Definitions on Fairness



Verma, S., & Rubin, J. (2018). Fairness definitions explained. FairWare@ICSE, 1–7. http://doi.org/10.1145/3194770.3194776

Fairness through Causal Discrimination

Definition

A classifier satisfies this definition if it produces the same classification for any two subjects with the exact same attributes X.

Example

A male and female applicants who otherwise have the same attributes X will either both be assigned a good credit score or both assigned a bad credit score:

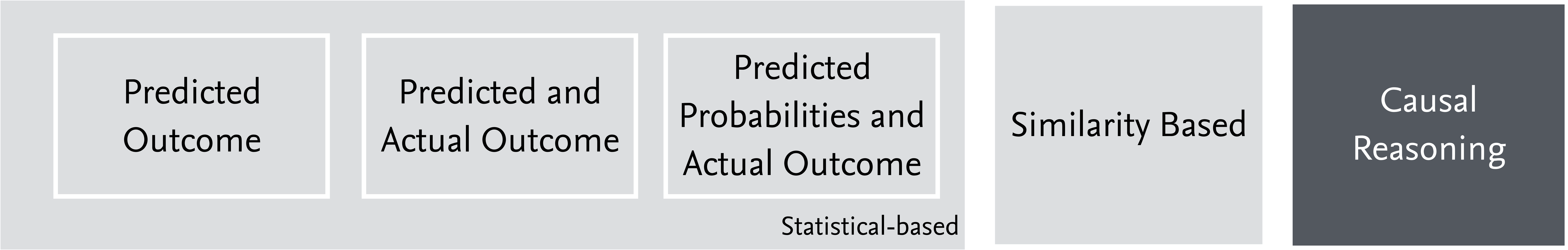
$$(X_f = X_m \wedge G_f != G_m) \rightarrow d_f = d_m.$$

Results:

For 8.8% of male and female applicants, the output classification was not same.

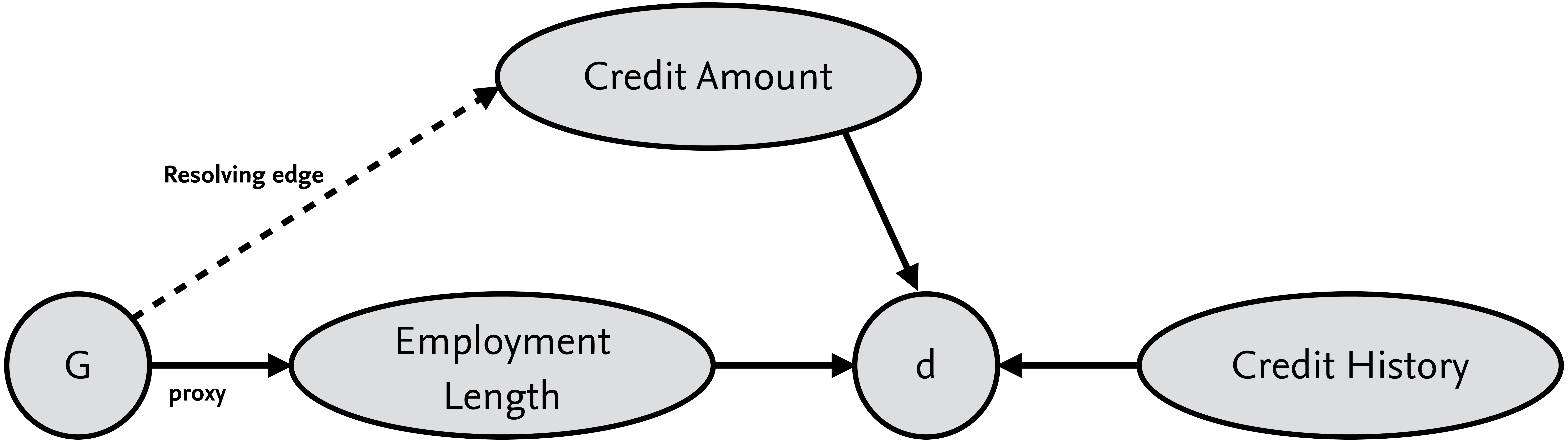
✗

Causal Reasoning-based Definitions on Fairness


Predicted Outcome
Predicted and Actual Outcome
Predicted Probabilities and Actual Outcome
Statistical-based
Similarity Based
Causal Reasoning

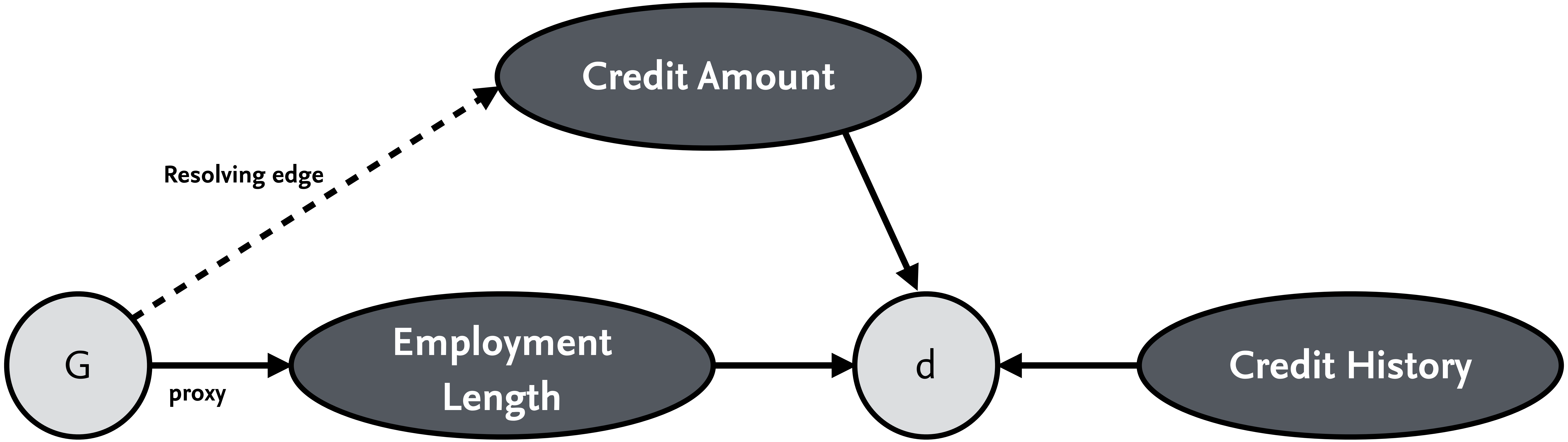
Verma, S., & Rubin, J. (2018). Fairness definitions explained. FairWare@ICSE, 1–7. http://doi.org/10.1145/3194770.3194776

Example of an Causal Graph



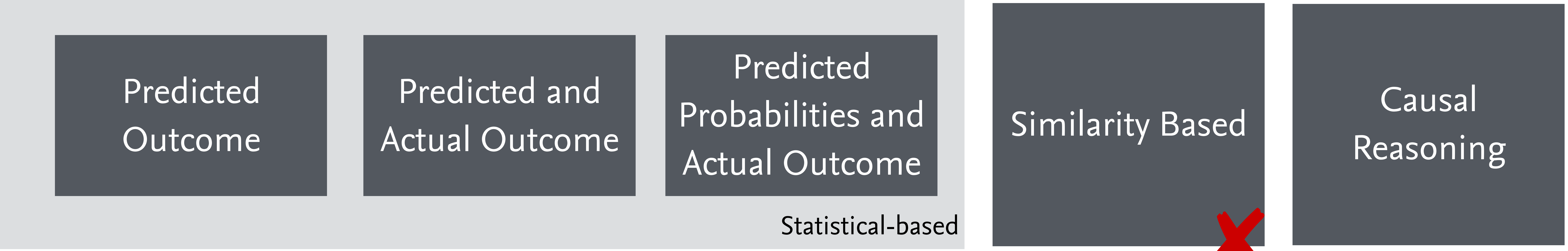
Verma, S., & Rubin, J. (2018). Fairness definitions explained. FairWare@ICSE, 1–7. http://doi.org/10.1145/3194770.3194776

Counterfactual Fairness



Verma, S., & Rubin, J. (2018). Fairness definitions explained. FairWare@ICSE, 1–7. http://doi.org/10.1145/3194770.3194776

What does it mean for our Scenario?



Predicted Outcome

Predicted and Actual Outcome

Predicted Probabilities and Actual Outcome

Statistical-based

Similarity Based ✗

Causal Reasoning

Group Fairness ✓

Predictive parity ✓

Test-fairness ✗ ✓

Conditional statistical parity ✗

False positive error rate balance ✗

Verma, S., & Rubin, J. (2018). Fairness definitions explained. FairWare@ICSE, 1–7. http://doi.org/10.1145/3194770.3194776

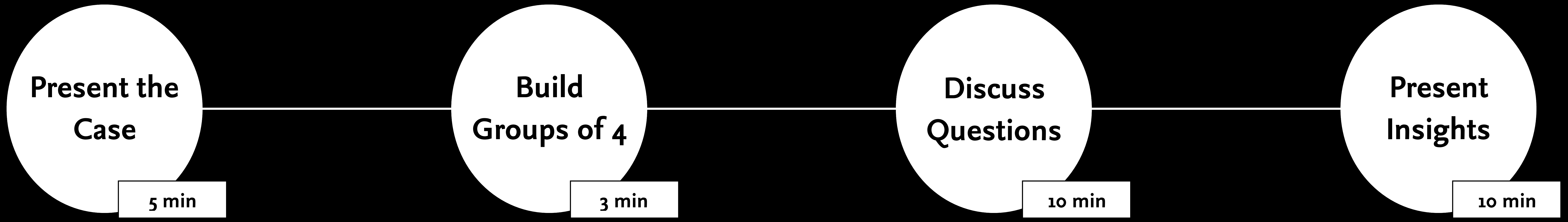
Your Reflexive Practice for Considering Fairness

Data-driven systems require a nuanced understanding of the relevant social context. It should be known in advance how a system affects the social context in a predictable way.

The social and technical requirements of the social context should be modeled appropriately. Fairness is a social requirement that is translated into a <u>technical</u> one.

Thus, such translation needs to be contestable, i.e., a person should be able to challenge machine predictions. Contestability allows people to provide evidence why they disagree with a prediction. A prerequisite is transparency and explainability :)

Case Study Reflection



Present the Case

5 min

Build Groups of 4

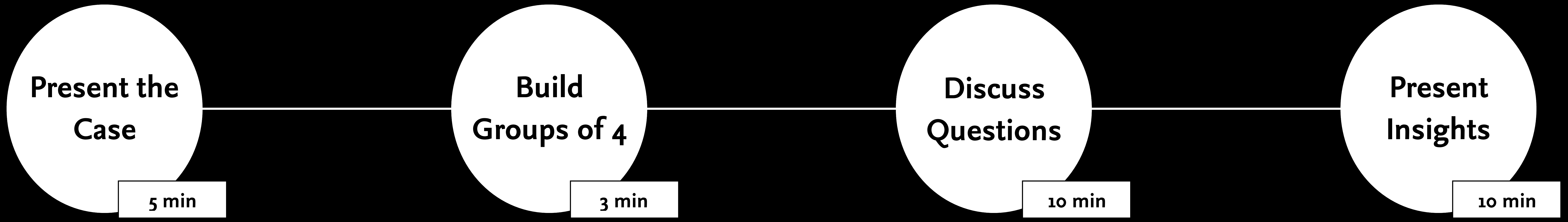
3 min

Discuss Questions

10 min

Present Insights

10 min



Present the Case

5 min

Build Groups of 4

3 min

Discuss Questions

10 min

Present Insights

10 min

1. Should Paula and her team have rejected machine learning in the selection of methods, if only for ethical reasons?

2. How do you evaluate the way learning data is generated?

3. As a mechanical engineer, Andreas doesn't have a lot of experience with the algorithms his computing system applies. He simply trusts the results that are delivered. Is he in charge of the results or the entire team?

4. Is a true positive rate of 97% really a good value for systems that affect human lives?

Check your Insights

What is the relation of discrimination and fairness?

How do you define fairness?

What group of fairness measures have we discussed and how do they differ?

How do you find an appropriate measure for fairness? What do you need to consider?

Human-Centered Computing
Research Group

Freie Universität Berlin

«Human-Centered Data Science»

Next week: Ensuring Transparency in your Data Science Workflow

Prof. Dr. Claudia Müller-Birn

Human-Centered Computing, Institute of Computer Science

Freie Universität Berlin

June 16, 2022