Holistic AI | The Alan Turing Institute

Implementation Workshop

Sachin Beepath

sachin.beepath@holisticai.com

About us

Fiona Thendean

AI Risk Management Officer (Holistic AI)

Before joining HAI, Fiona studied Statistics and Machine Learning at Carnegie Mellon University. She moved to London to pursue an MSc in Artificial Intelligence for Sustainable Development at UCL, where she focused on Explainability for Deep Reinforcement Learning for her dissertation.

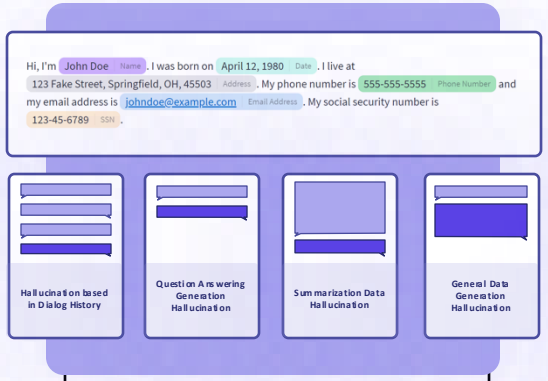
Sachin Beepath

AI Assurance Officer (Holistic AI)

Prior to working for HAI, Sachin studied astrophysics at the University of Waterloo in Canada, where he is originally from. Sachin moved to London to pursue an MSc in Data Science and Machine Learning at UCL where he did his dissertation on Bias in Facial Recognition. In his free time Sachin enjoys producing music and DJing.

Holistic AI is an AI Governance company that aims to empower enterprises to adopt and scale AI confidently.

Holistic AI Platform

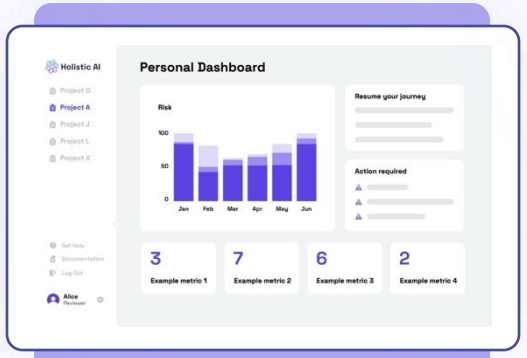


Safeguard

Governance and monitoring of ChatGPT and other language models

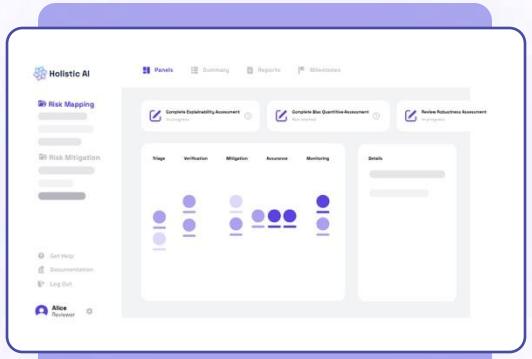
Tracker

All-in-one repository of all things related to AI regulations and harms



AI Governance Platform

Enterprise-ready application for safe AI adoption and usage



AI Auditing

Third-party conformity assessment and reporting for specific AI risks

Holistic AI

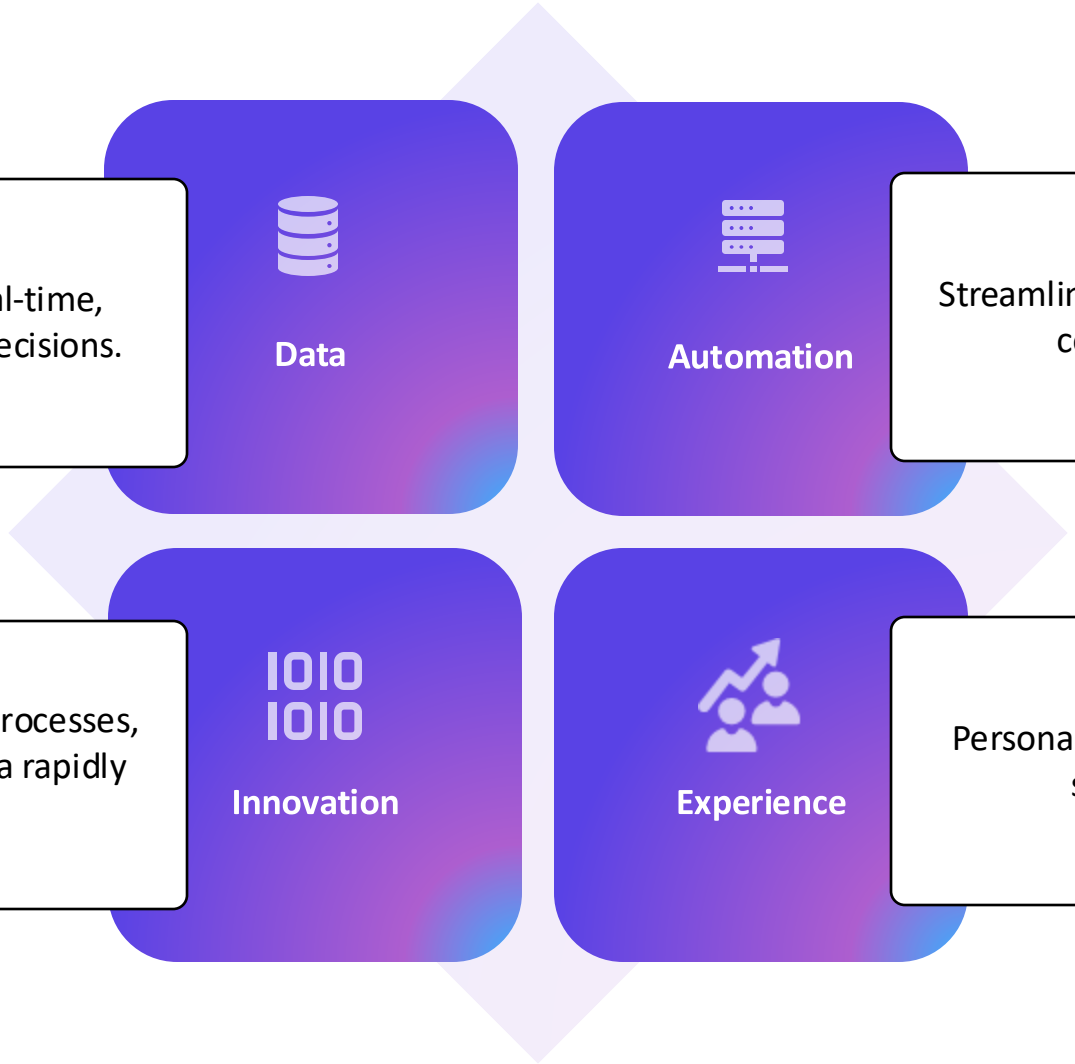
Agenda

- Introduction
- Sources of Bias
- Concepts of Fairness
- Measuring and Mitigating Bias
- Trade-offs with other AI risk verticals

How is AI Being Used?

Why are businesses adopting AI?

Holistic AI



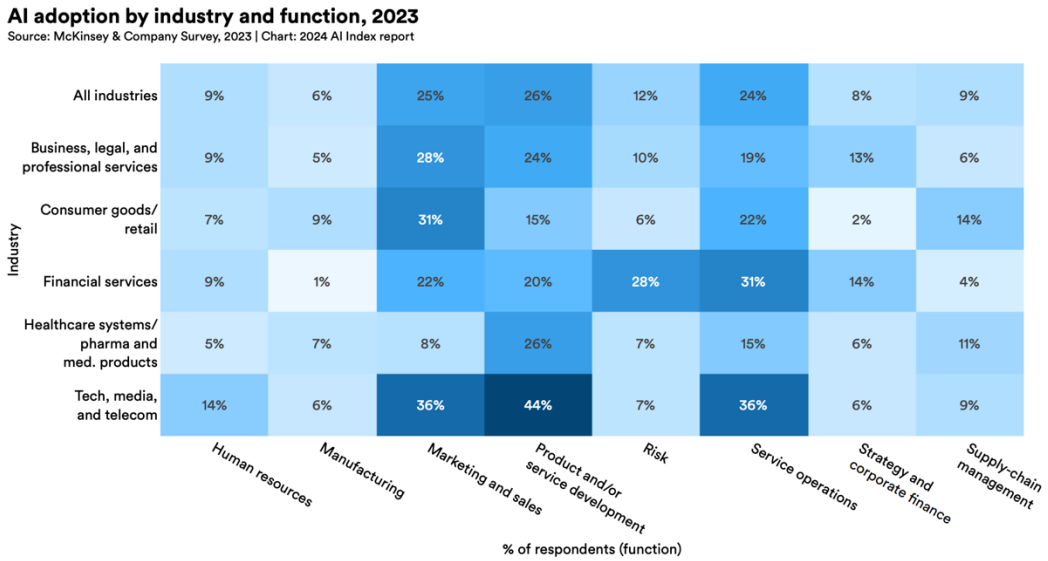
Analyse vast amounts of data in real-time, make more informed and strategic decisions.

Streamline repetitive tasks, reduce operational costs and improve productivity.

Develop smarter products, optimise processes, and stay ahead of industry trends in a rapidly evolving market.

Personalise customer interactions, increasing satisfaction and engagement.

What is happening?



Emerging Regulations and Frameworks

Holistic AI

- UK AI Regulation White Paper
- EU AI Act
- NIST AI RMF
- ISO 420001

The Future

UK AI Market

is predicted to grow to over $1 trillion by 2035

Global GDP

is estimated to grow by $15.7 trillion by 2030

Explosion in AI vendors

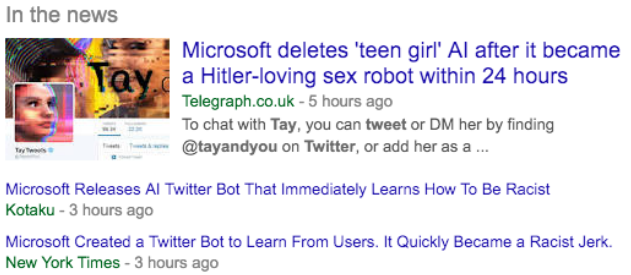
OpenAI | ANTHROP\C | deepseek

Risks Associated with AI

General set of AI risks

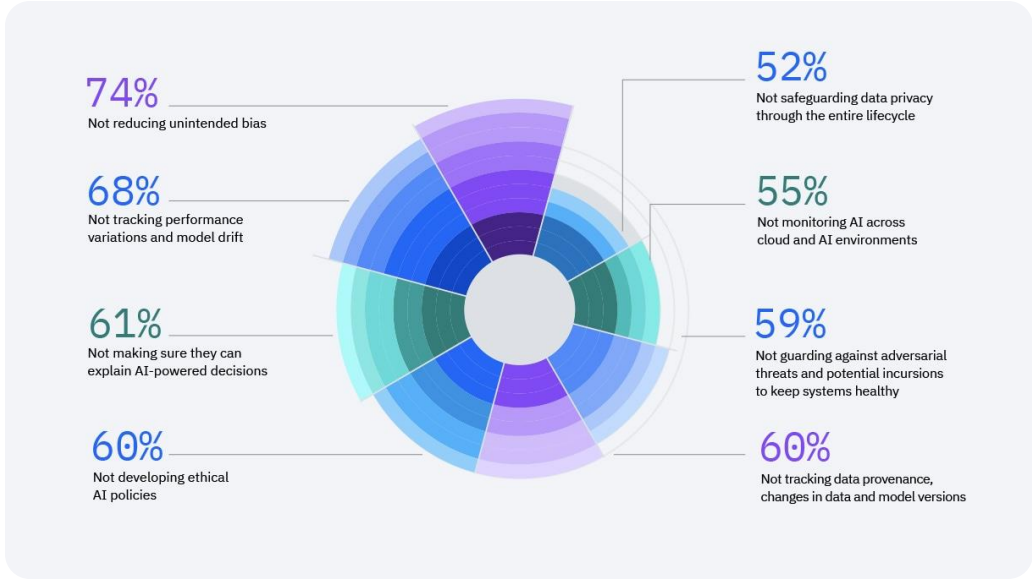
Security and Privacy	Data minimization principles as well as adopt privacy-enhancing techniques to mitigate personal or critical data leakage.
Bias and Discrimination	Avoid unfair treatment of individuals given their protected characteristics.
Interpretability and Explainability	Provide decisions or suggestions that can be understood by their users and developers.
Performance and Robustness	Be safe and secure, not vulnerable to tampering or compromising of the data they are trained on.

To avoid these cases



Why Bias is the key risk?

Reputational damage	Instances of biased AI garner significant public attention and can lead to public backlash against the technology, as well as increased scrutiny and regulation.
Challenging risk to deal with	Compared to other key risks, Bias is undoubtedly the area where data scientists and other professionals are least prepared to deal
Regulatory requirements	Most laws and regulations are targeted towards bias, for example anti-discrimination laws.
General awareness	Anecdotally, Privacy is for Data what Bias is for AI



https://www.ibm.com/watson/resources/ai-adoption

Societal Implications of Bias in AI

Holistic AI

Historical Discrimination

Society's historical bias reflected in the data that we use to train algorithmic systems. will be perpetuated by the decision-making system.

Impact on Life Chances

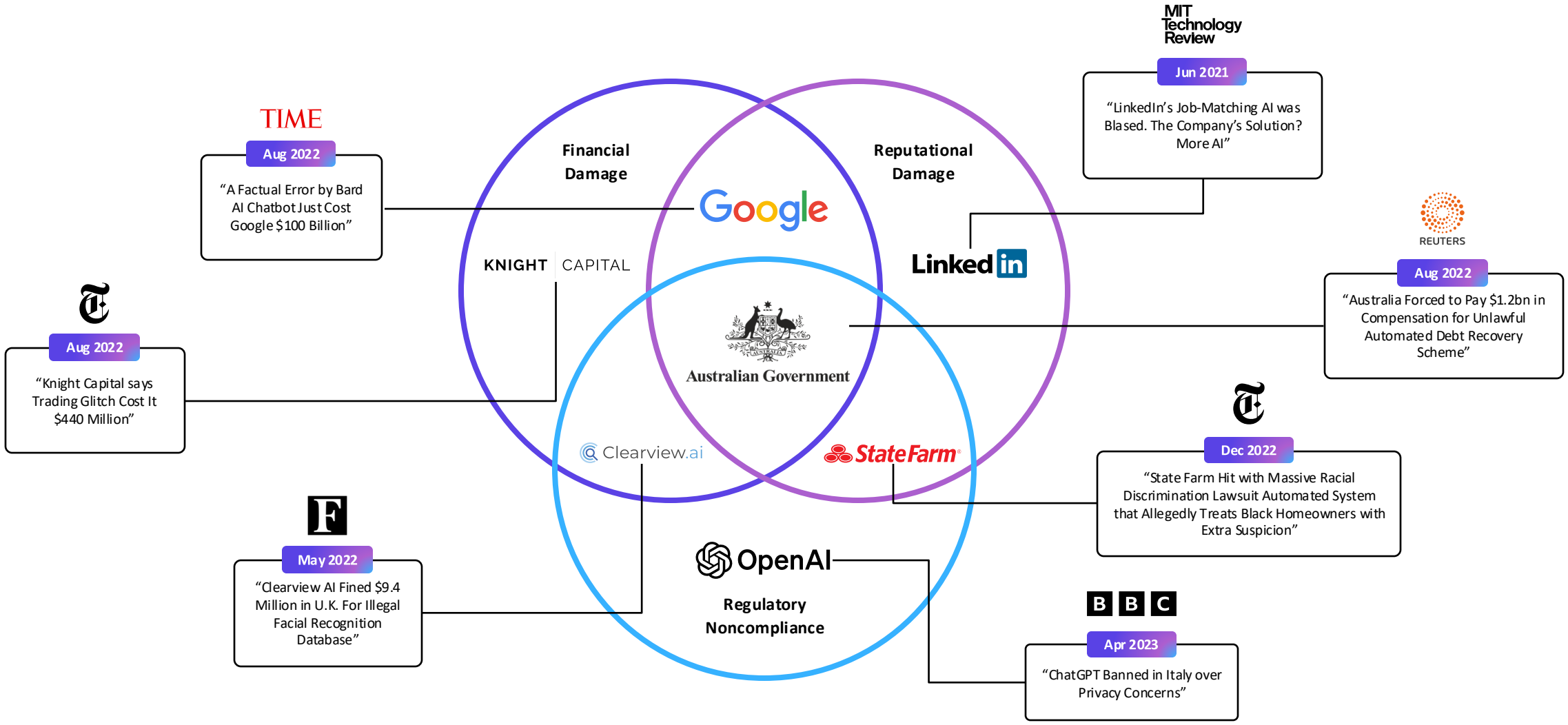
Algorithmic systems are being used for decision making in a wide range of areas such as criminal justice, healthcare, education, employment, consumer lending, etc.

Responsibility

When developing decision-making systems, we are responsible for minimizing harm.

Business Implications of Bias in AI

Holistic AI



Risks and issues with HR technologies

With the pervasive usage, side-effects such as bias, transparency, cheating will increase

Holistic AI

Tutoring firm settles US agency's first bias lawsuit involving AI software

EEOC Settles First AI-Bias Case, but Many More Are Likely

Amazon scrapped 'sexiest AI' tool

Lawsuit Claims Workday's AI And Screening Tools Discriminate Against Those Black, Disabled Or Over Age 40

HireVue drops facial monitoring amid A.I. algorithm audit

Candidates using ChatGPT to cheat their way into a job

Real-World Examples of Bias in AI

ML can be biased

Amazon recruitment

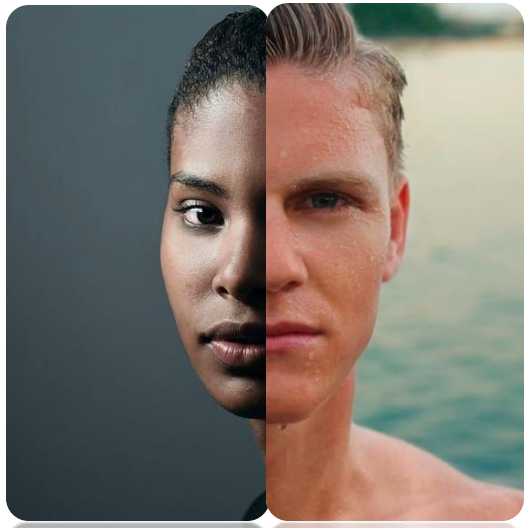
RESUME

2023: AWESOME

2020: WOMEN'S COLLEGE

Gender recognition ("gender shades")

Up to ~35% error rate



Up to ~1% error rate

The COMPAS example

- Used in US court to predict risk of recidivism
- ProPublica study in 2016:

	White	Black
Labelled higher risk – but didn't re-offend	23%	45%
Labelled low risk – did re-offend	48%	28%

- Northpointe defence: accuracy for white and black is the same (~60 %)

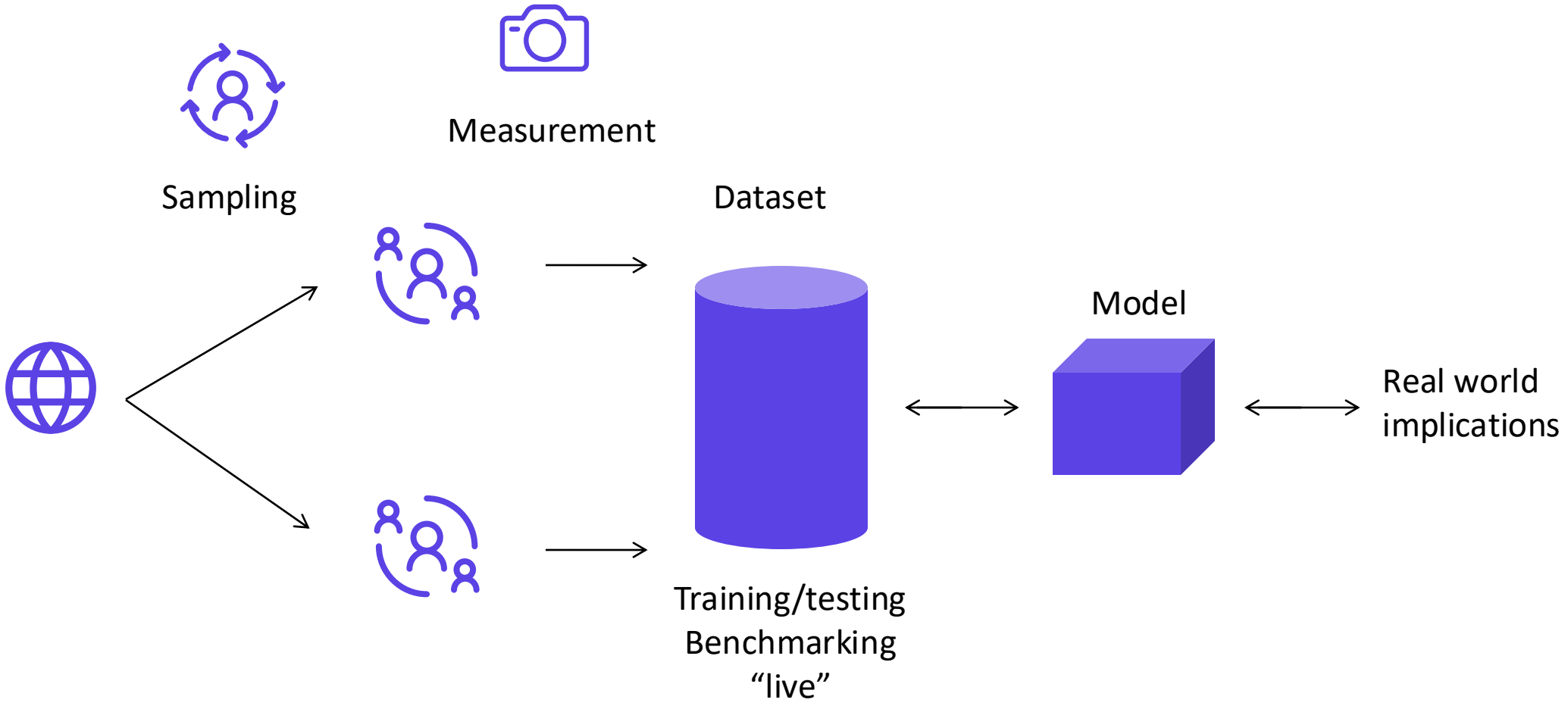
GenAI

- UNESCO study reveal that popular LLMs like gpt3.5 and Llama 2 exhibit bias against women in its outputs.
  - Richer narratives in stories about men
  - Homophobic attitudes and racial stereotyping

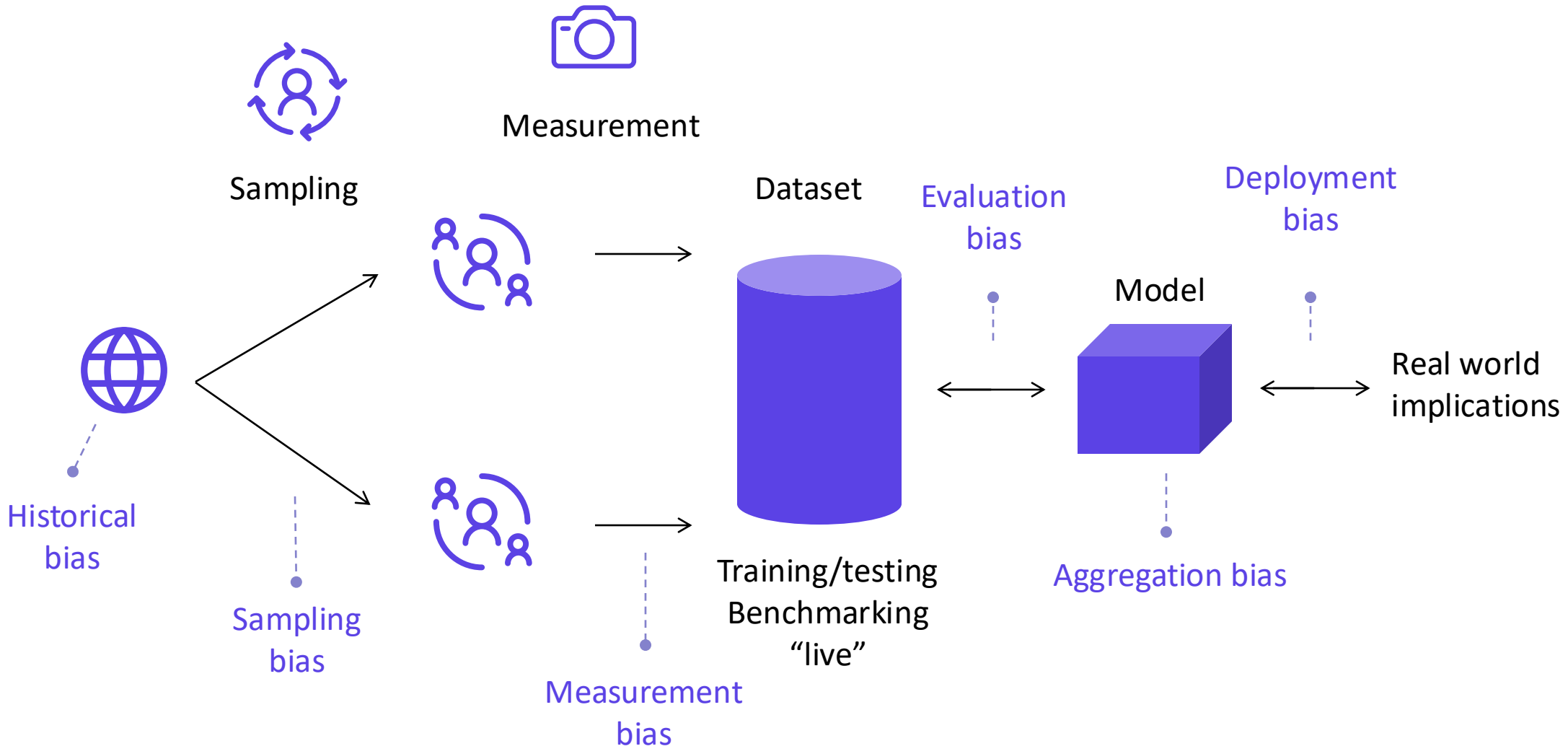
Holistic AI | The Alan Turing Institute

Sources of Bias

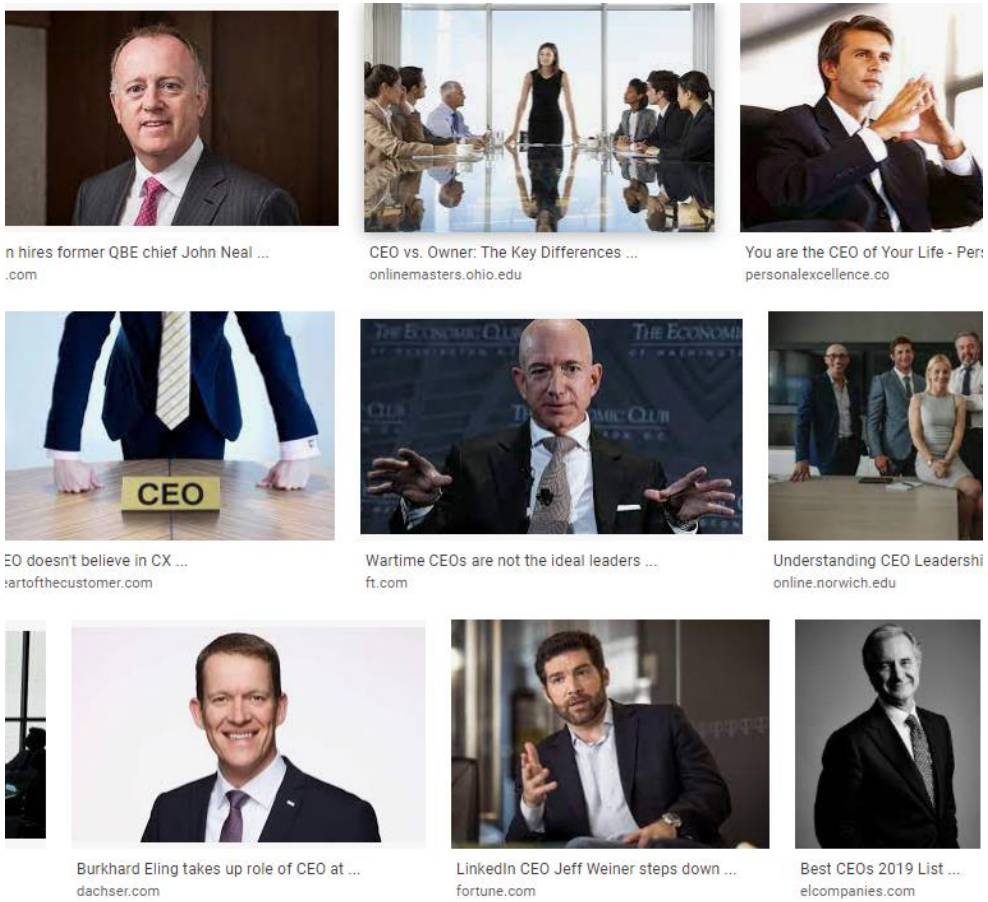
Machine Learning lifecycle



Sources of Bias



Historical Bias

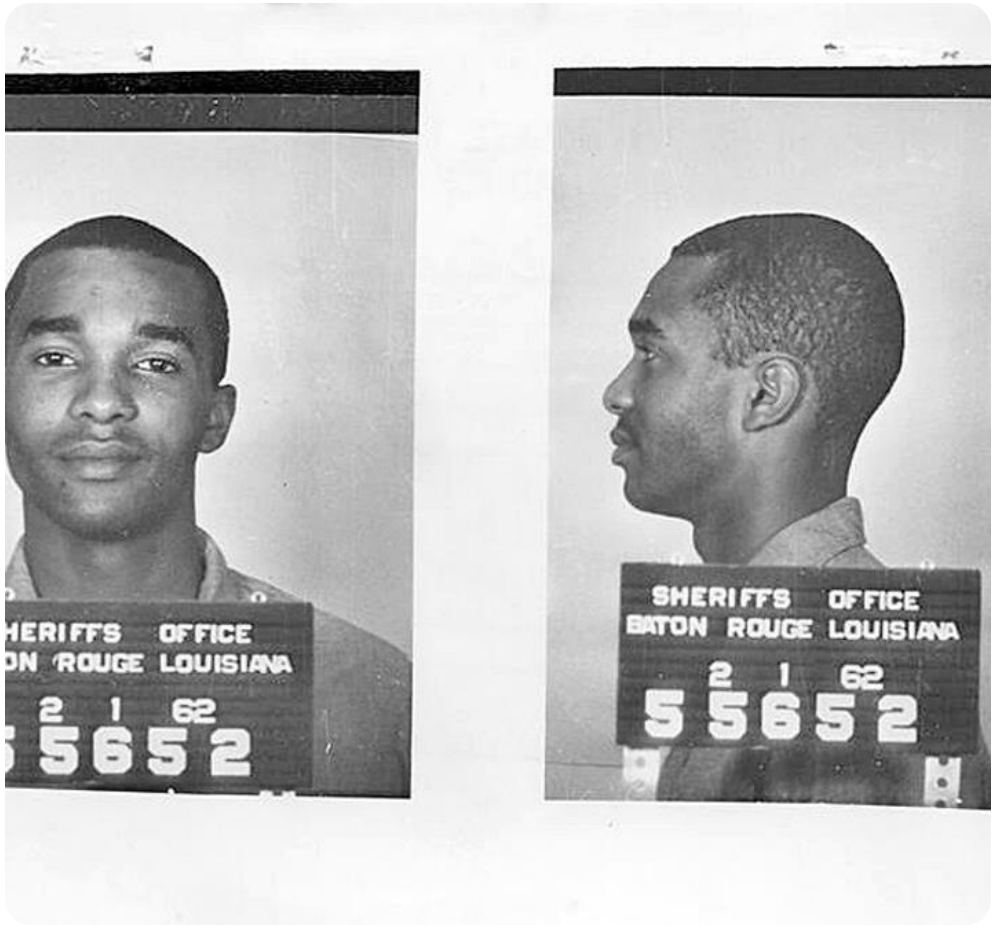


- Pre-existing bias reflected in the data
- Example: In 2018, 5% of Fortune 500 CEOs were women.
- What should be the results of an image search for ‘CEO’? Google recently changed image search result to include higher proportion of women

Sampling Bias

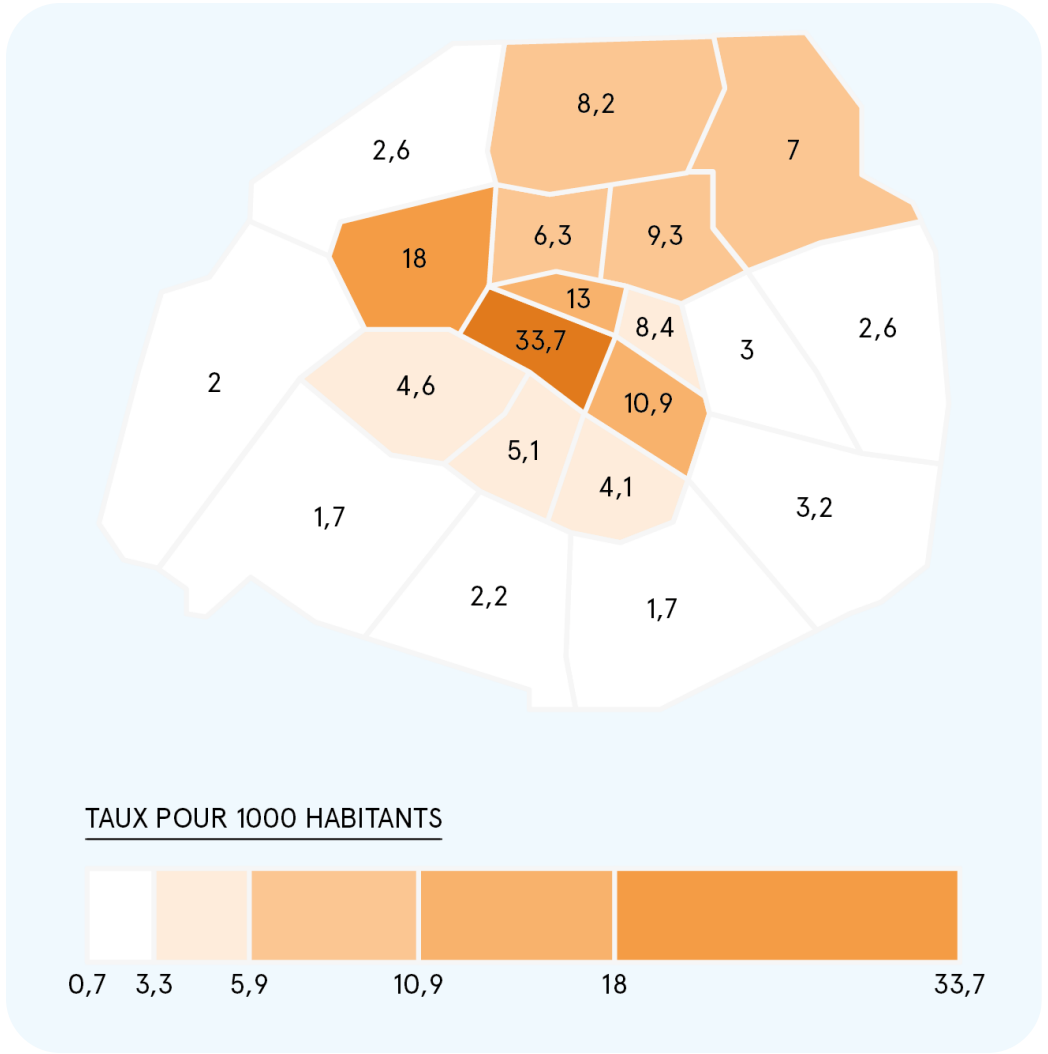
- Occurs when certain parts of the input space are underrepresented
  - sampling methods only reach a portion of the population
  - population of interest ≠ training data
- Example: ImageNet. 45% from US. 1% from China.
- Classifier for ‘bride’ performs worse in under-represented countries

Measurement bias



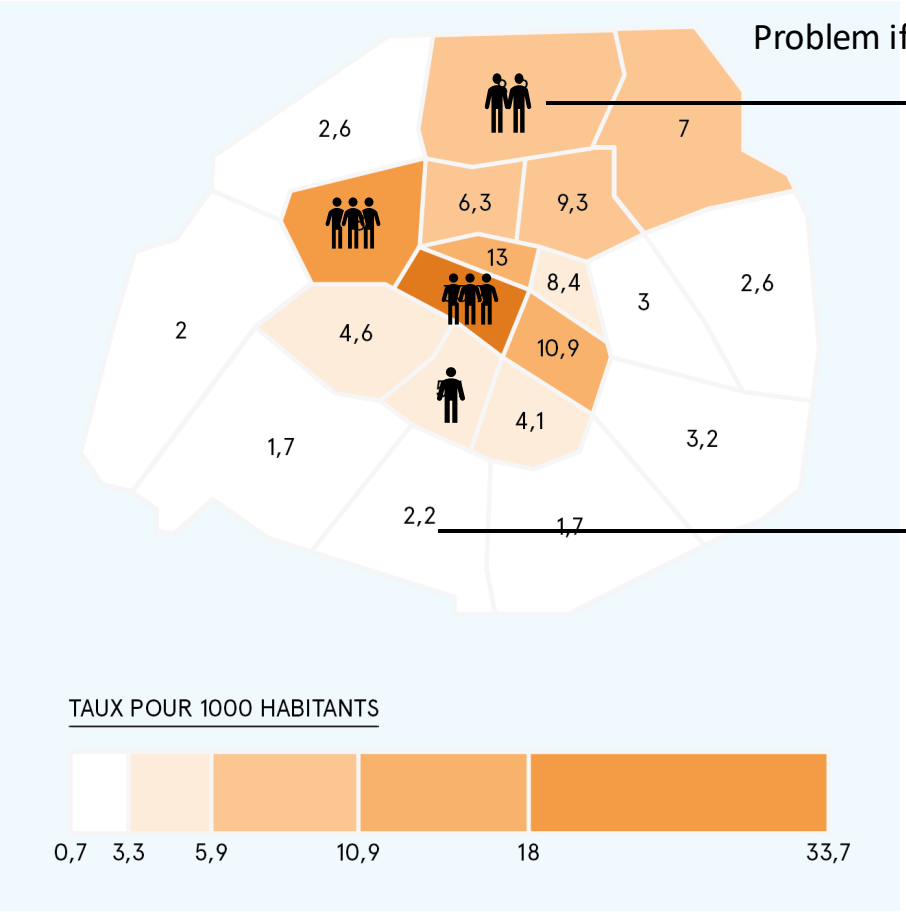
- Occurs when selecting or collecting features.
- Measurement process/ data quality may vary across groups
- Example:
  - Arrest rates used as a measure for crime rate.
  - Number of times someone has been arrested may depend on the reinforced presence of police forces in some neighbourhoods.

Measurement bias



https://bonadvisor.com/

Measurement bias



Problem if neighborhood with predominant ethnicity

Arrest database

1

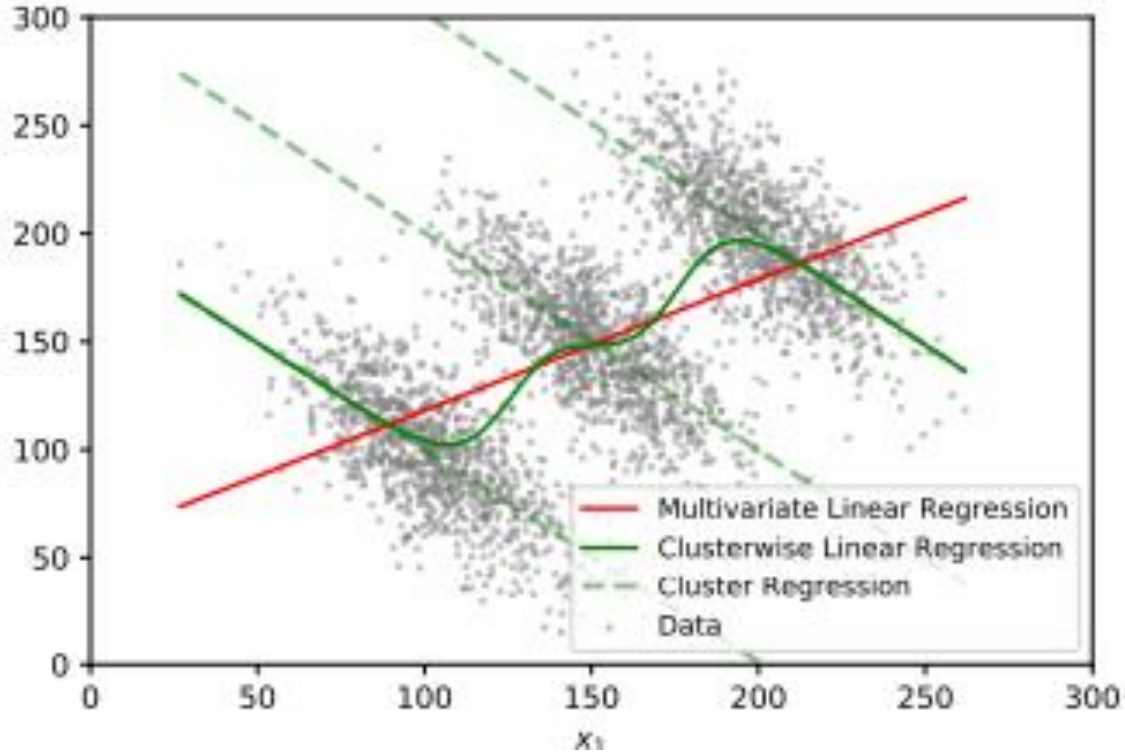
0

COMPAS: prior arrests + friend/family arrests to measure likelihood of recidivism

https://bonadvisor.com/

Aggregation bias

- One-size fit all model
- Difference across groups might require several models
- Example: HbA1c levels (used to monitor diabetes) differ across gender and ethnicity



Mehrabi et al. (2019). A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.

Evaluation bias

The Algorithmic Justice League by Joy Buolamwini

- When testing on benchmarks that are unbalanced compared to target population
- Example: Adience and IJB-A benchmark datasets in Gender Shades (>70% lighter skinned subjects)

Deployment bias

- Model was trained and tested for use case X but is used for application Y
- Example: model was trained to identify good candidates for investment banking jobs, and now being used to identify junior level lawyers

Holistic AI | The Alan Turing Institute

Bias and Fairness in AI

Bias vs. Fairness

- **Bias**: Systematic errors in the decision-making process of an algorithm that lead to inaccurate or unfair outcomes.
- **Fairness**: Ensuring equity in the decision-making process of an algorithm across individuals and groups.

What is Fairness?

Individual Fairness: Seeks for similar individuals to be treated similarly

- "Am I as an individual being treated as another with similar qualifications?"
  - Fairness through unawareness
  - Fairness through awareness
  - Counterfactual fairness

Group Fairness: Split a population into groups defined by protected attributes (e.g. male, female) and seeks for some measure to be as equal as possible across groups

- "Is the minority group to which I belong being treated as the majority group?"
  - **Equality of Outcome.** Distribution across groups should be the same
  - **Equality of Opportunity.** AI Performance across groups should be the same

Individual Fairness: Fairness Through Unawareness

- Protected attributes A, are not explicitly used in the input features X

$$\mathrm{A} \cap \mathrm{X} = \emptyset$$

- What happens when features are highly correlated to protected attributes?

  Proxies: Zipcode → Ethnicity

Individual Fairness: Fairness Through Awareness

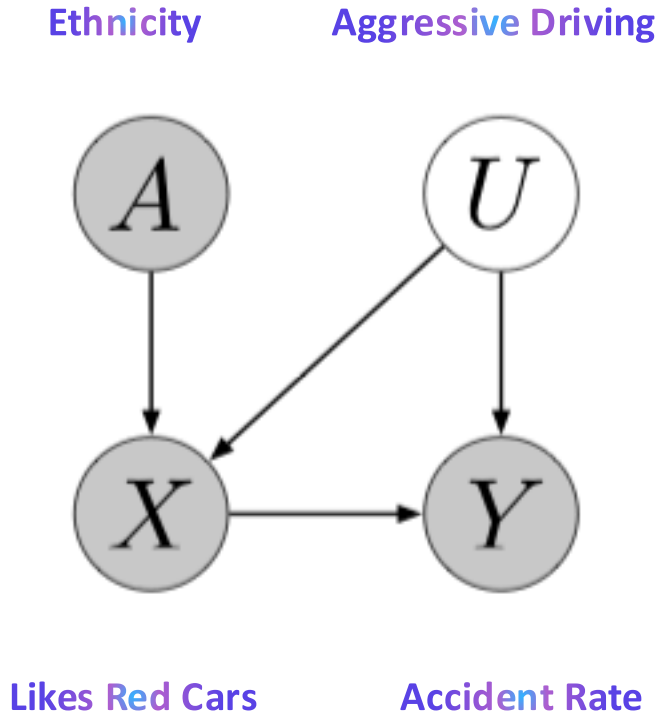
- Similar predictions are given to similar individuals (w.r.t. similarity/distance metric)
- For individuals x and y, the model M is fair if the Lipschitz condition is satisfied:

$$D\big(M(x), M(y)\big) \leq D(x, y)$$

- How do we choose a similarity/distance metric?

Individual Fairness: Counterfactual Fairness

Represent the problem using a causal graph. The system is fair if the outcome Y does not depend on a descendent of the protected attribute A



Individual Fairness: Counterfactual Fairness

Holistic AI

- The predictor $\hat{Y}$ is counter factually fair if under any context $X = x, A = a$, for all y and any value a' attainable by A

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

- How to determine the causal graph?

Group Fairness

"Is the minority group to which I belong, being treated the same as the majority group?"

- **Equality of Outcome:** The likelihood of a positive outcome is the same for every group
- **Equality of Opportunity:** The probability of assigning a member of the minority group to the positive class is the same as assigning a non-member to the positive class

Holistic AI

Group Fairness: Equality of Outcome

- **Definition:** Equal likelihood of positive predicted outcome.
- **Mathematically:**

$$P(\hat{Y} = 1|A = a) = P(\hat{Y} = 1|A = a')$$

- **Example metrics:**

  Disparate Impact = $\frac{SR_{female}}{SR_{male}}$ → ideal value of 1

  Statistical Parity = $SR_{female} - SR_{male}$ → ideal value of 0

- **Potential problem: Positive discrimination?**

Group Fairness: Equality of Opportunity

- **Definition:** This should be equal for all groups:
  - probability of a person in the positive class being correctly assigned a positive outcome (TPR)
  - probability of a person in a negative class being incorrectly assigned a positive outcome (FPR)
- **Mathematically:**

$$P(\hat{Y}=1 \mid A=m, Y=y)=P(\hat{Y}=1 \mid A=f, Y=y),\ y \in\{0,1\}$$

- **Example metric:** Average Odds Difference

$$AOD=1/2[(\llbracket FPR \rrbracket\_f-\llbracket FPR \rrbracket\_m)+(\llbracket TPR \rrbracket\_f-\llbracket TPR \rrbracket\_m)]$$

- **Potential problem:** Where does ground-truth come from, is it a valid unbiased measure?

Holistic AI | The Alan Turing Institute

Measuring Bias

The Problem

Holistic AI

50 Male and 50 Female candidates for a job opening. The company uses a round of automated video assessment to filter the candidates.

Pool of candidates

50 Male

50 Female

Video assessment

Successful candidates

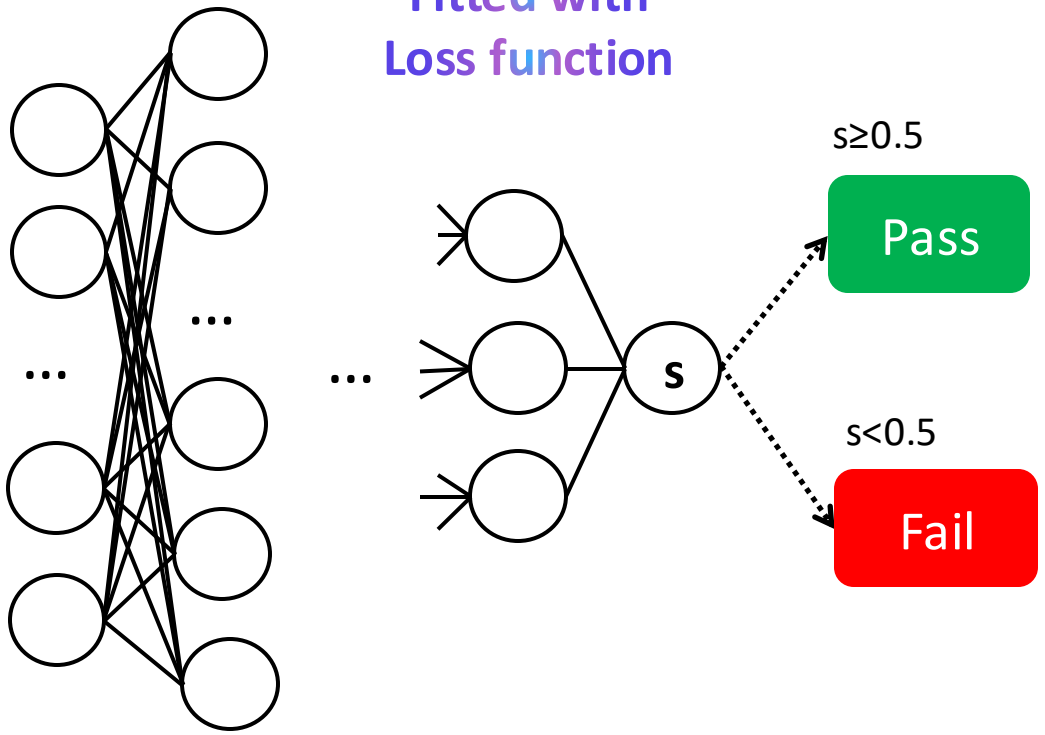
?

The Model

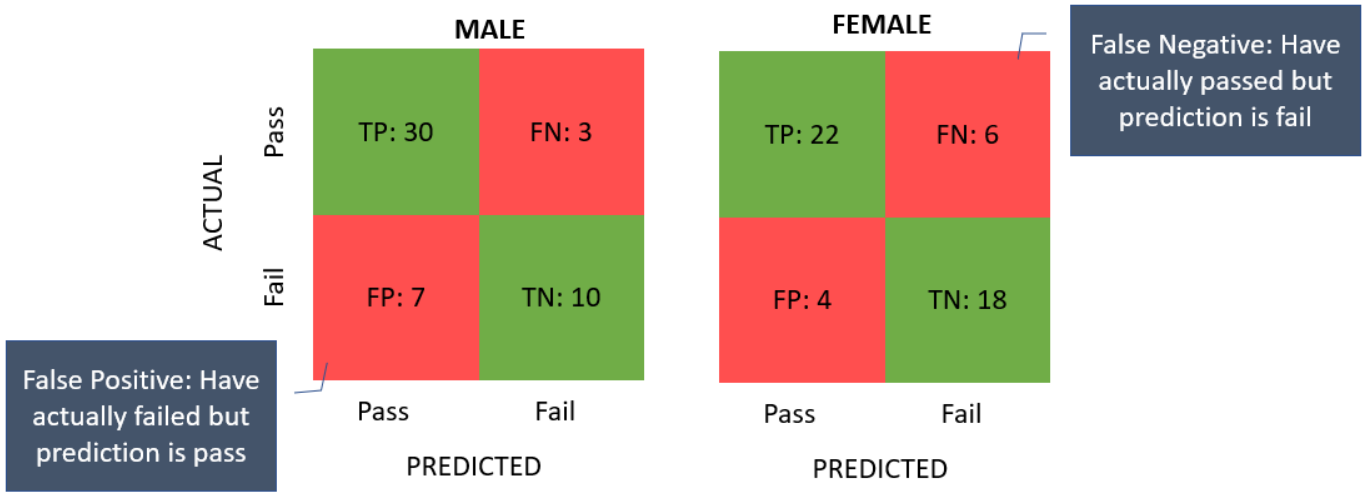
Holistic AI

- Personality assessment
- Cognitive ability
- Experience
- Technical skills

1,000 features



Equality of Outcome



$$\boldsymbol{P(\hat{Y} = 1|A = male) = P(\hat{Y} = 1|A = female)}$$

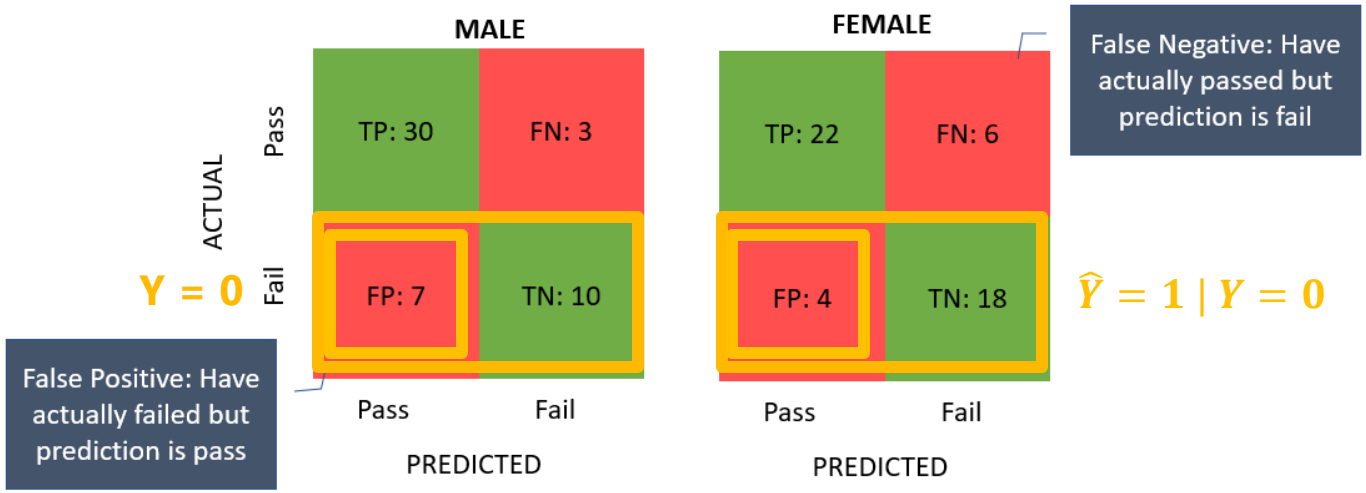
$$SR_{Male} = \frac{37}{50} = 0.74$$

$$SR_{Female} = \frac{26}{50} = 0.52$$

$$SP = 0.52 - 0.74 = \mathbf{-0.22}$$

$$DI = \frac{0.52}{0.74} = \mathbf{0.7}$$

Equality of Opportunity

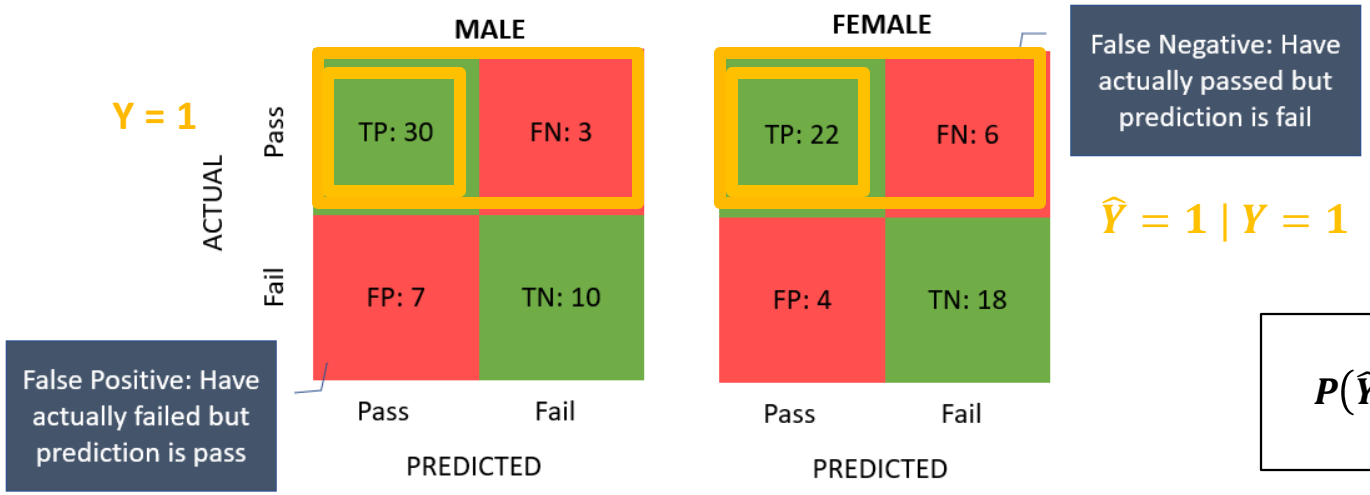


$$P(\hat{Y} = 1 | A = male, Y = 0) = P(\hat{Y} = 1 | A = female, Y = 0)$$

$$P(\hat{Y} = 1 | A = m, Y = 0) = FPR_{Male} = \frac{FP}{FP + TN} = \frac{7}{7 + 10} = 0.41$$

$$P(\hat{Y} = 1 | A = f, Y = 0) = FPR_{Female} = \frac{4}{4 + 18} = 0.18$$

Equality of Opportunity



$$P(\hat{Y} = 1 | A = male, Y = 1) = P(\hat{Y} = 1 | A = female, Y = 1)$$

$$P(\hat{Y} = 1 | A = m, Y = 1) = TPR_{Male} = \frac{TP}{TP + FN} = \frac{30}{30 + 3} = 0.91$$

$$P(\hat{Y} = 1 | A = f, Y = 0) = FPR_{Female} = \frac{22}{22 + 6} = 0.79$$

Equality of Opportunity

Holistic AI

$$AOD = \frac{1}{2}[(FPR_{Female} - FPR_{Male}) + (TPR_{Female} - TPR_{Male})$$

$$AOD = \frac{1}{2}[(0.18 - 0.41) + (0.79 - 0.91)$$

$$AOD = \mathbf{-0.175}$$

Holistic AI | The Alan Turing Institute

Choosing a Fairness Metric

Choosing a Fairness Metric

Holistic AI

Too many metrics

Not mathematically possible to construct an algorithm that simultaneously satisfies all reasonable definitions of a "fair" or "unbiased" algorithm.

Governance structures

Deciding which definition to use must be done in accordance with governance structures.

Regulatory requirements

Every jurisdiction dictates what equality or non-discrimination is. Deciding which definition to use must be done in accordance with governance structures.

Choosing a Fairness Metric

Holistic AI

Equality of Opportunity

Worldview says that the score correlates well with future success and there is a way to use the score to correctly compare the abilities of applicants.

Equality of Outcome

Worldview says that the score may contain structural biases so its distribution being different across groups should not be mistaken for a difference in distribution in ability.

Choosing a Fairness Metric

Idea of construct space:

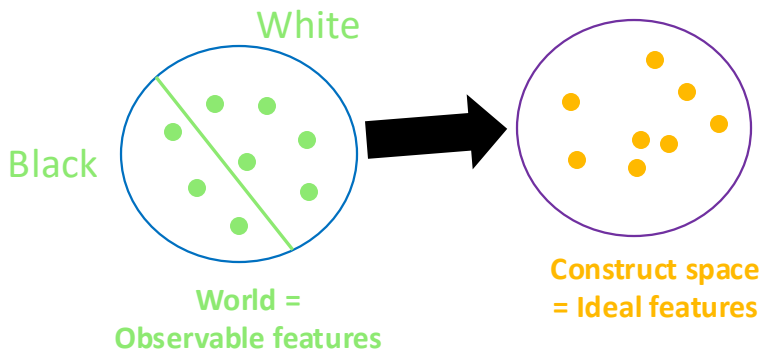
Equality of Opportunity

What you see is what you get (WYSIWYG) → construct space and observed space are essentially the same

Equality of Outcome

We're all equal (WAE) → Structural bias. Observed space not a good representation of construct space.

[Friedler et al., 2016]



Choosing a Fairness Metric

There is no universal measure of algorithm fairness. Use domain knowledge to identify protected groups:

- Groups subject to historic prejudice
- Groups for which the impact of an erroneous outcome would be highest

Validate that your mitigating action has not adversely impacted other groups

“What gets measured, gets managed” – calculate and present a wide range of fairness measures to understand the impact of a model on individuals and groups

Example: Facial Recognition

- No selection process that may require equal representativeness
- Ground-truth label is trusted

→ **Equality of Opportunity**

Example: Hiring

Holistic AI

Algorithm selects top 100 candidates.

The problem: more white candidates selected in proportion (Disparate Impact = 0.7). But equality of opportunity metrics are good (Average odds difference).

- **Option 1:** Do nothing (Equality of opportunity) → darker skinned candidates complain that the data used to train the model contains bias.
- **Option 2:** Mitigate (Equality of Outcome) → a non-selected white candidate complains he was more qualified than a darker skinned candidates who was selected.

Which option to choose?

How to justify a course of action?

Rule of Thumb

If you trust the ground-truth labels
→ Equality of Opportunity

If you don’t or if the label may contain bias too? Equality of Outcome or Equality of Opportunity?

Holistic AI | The Alan Turing Institute

Mitigating Bias

Choosing an Intervention Method

Holistic AI

Inadequate training data

A model approximates a function $f^*(X)$ that maps the input data X to the target label or value Y. Training data consists of "noisy, approximate examples of $f^*(X)$ evaluated at different training points" (Goodfellow et al., 2016) and as such may not always accurately represent the function we are aiming to approximate.

Issues with the data collection process or the distribution of historic outcomes can mean training data contains unwanted bias.

Unwanted bias caused by model generalisations

Many common machine learning models work by minimising the sum of prediction errors. It is therefore less costly to misclassify the minority than it is the majority?

Model design choices prioritised without considering fairness

Common misconception: algorithms are neutral; ‘the data doesn’t lie’. In reality, model design involves choices of inputs, feature engineering, architecture and hyper-parameters that impact outcomes (Mittelstadt et al., 2016).

Technical Solutions for Bias Mitigation

Holistic AI

Regardless of the measure used, bias can be mitigated at different points in a modelling pipeline

Pre-processing

- Reweighing
- Learning Fair Representations
- Correlation Remover

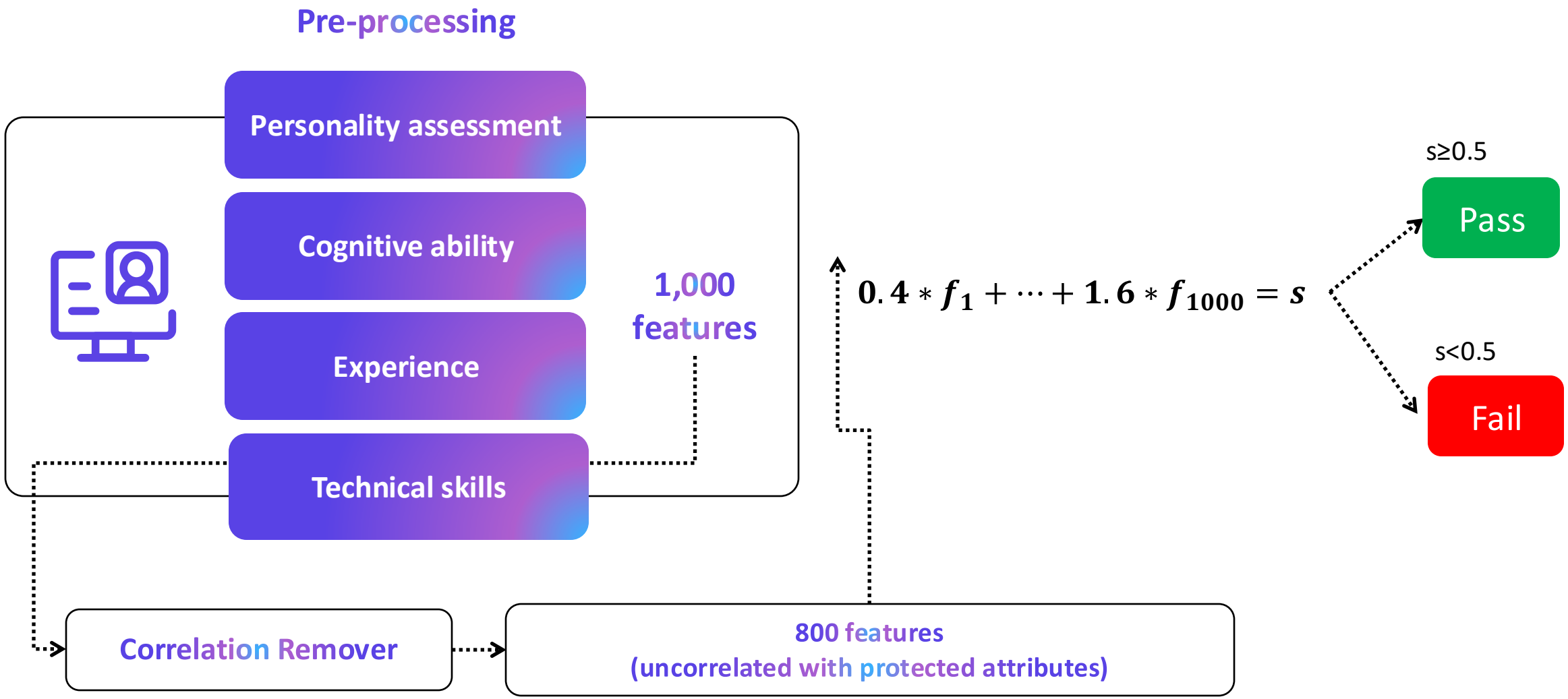
In-processing

- Adversarial Debiasing
- Fairness Constraints
- Gradient Based Methods

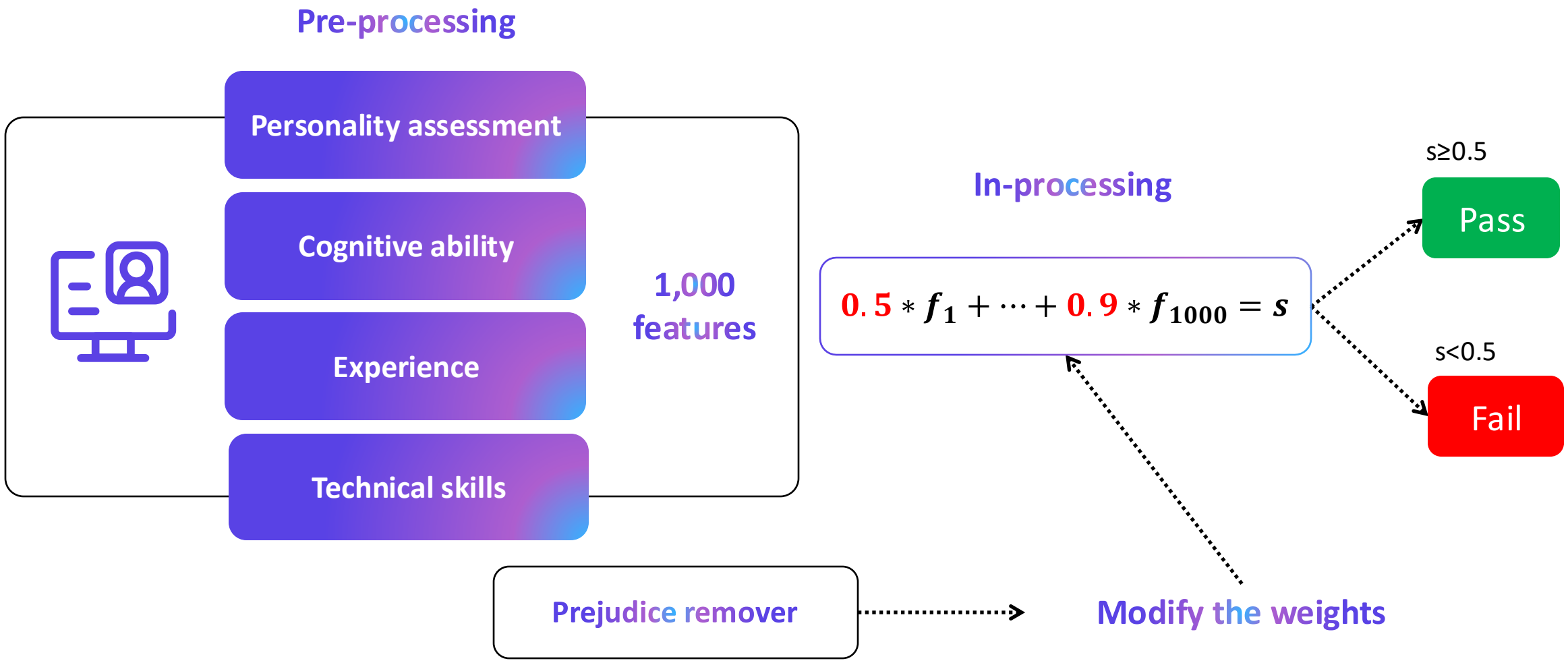
Post-processing

- Calibrated Equalized Odds
- Reject Option Classification

Pre-Processing: Correlation Remover

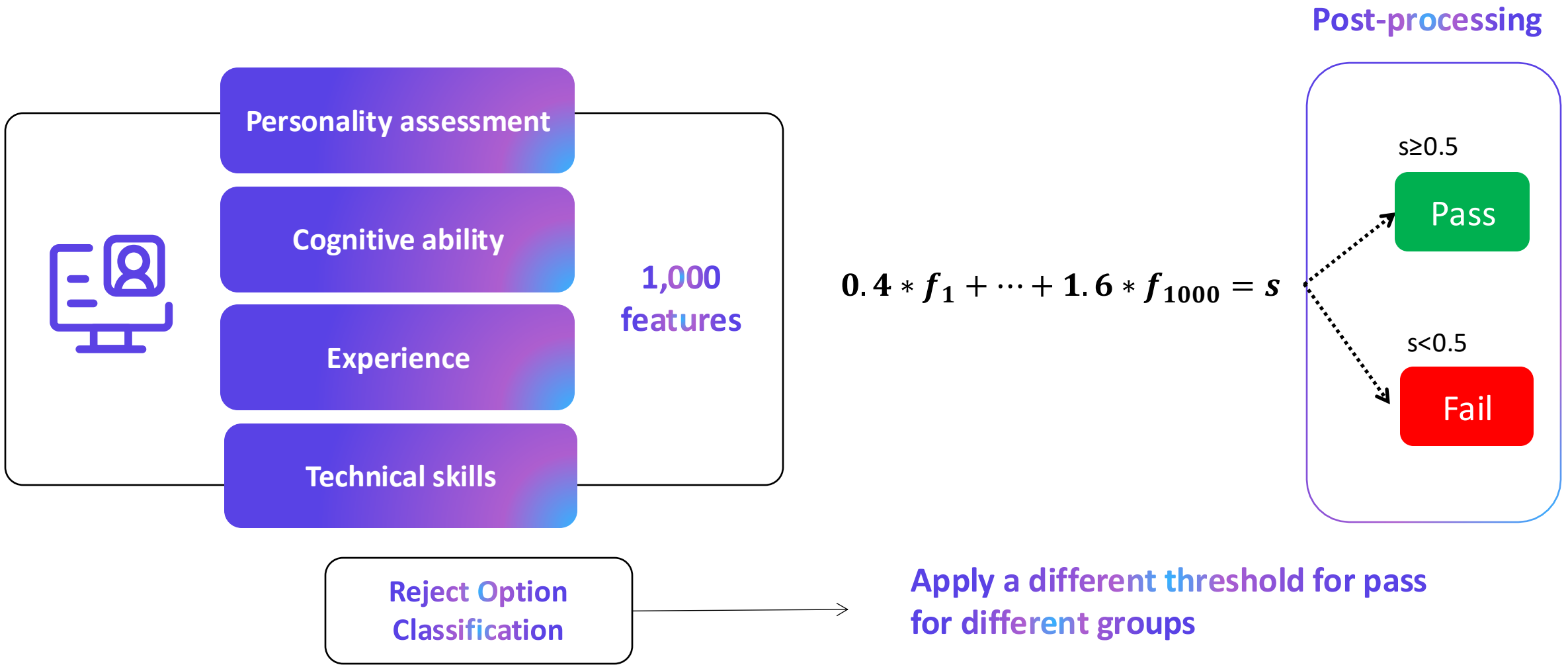


In-Processing: Prejudice Remover



Post-Processing

Holistic AI



Our Example Using ‘Adult’ Dataset

Holistic AI

Binary classification: Predict whether annual income exceeds $50K based on census data.

Training data:
$X_i \in \mathbb{R}^d$ (large number of feature variables describing each individual candidate)
$Y_i \in \{0,1\}$ (exceeds $50k/year or not)

Protected attributes:
Sex: Female, Male

Pipeline:
Standardize input features (excluding protected attributes)
Logistic Regression Model

Train the model on the data and establish a bassline for both bias and efficacy metrics

https://archive.ics.uci.edu/dataset/2/adult

Establishing a Baseline

Holistic AI

- The efficacy assessment shows we can predict salary class with accuracy (0.85).
- There is a clear bias with respect to the equality of outcome metrics.

Metric	Value
Accuracy	0.85
Balanced Accuracy	0.76
Precision	0.73
Recall	0.60
F1-Score	0.66

Metric	Value	Reference
Statistical Parity	-0.18	0
Disparate Impact	0.31	1
Equality of Opportunity Difference	-0.06	0
Average Odds Difference	-0.07	0

Our Example Mitigation Techniques

Holistic AI

Pre-processing example:

- Reweighing (Kamiran and Calders 2012)

In-processing example:

- Grid Search Reduction (Argarwal et al 2018)

Post-processing example:

- Equalised Odds Postprocessing (Hardt et al 2016)

Pre-Processing Example: Reweighing

Holistic AI

Reweighing applies a weight to each tuple in the training data. Each weight is calculated such that the training data, with weights applied, achieves statistical parity between a protected group compared with the privileged group.

- **Inputs:** Training data, Protected Attributes
- **Approach:** Assigns lower weights to examples that have been either ‘deprived’ or ‘favoured’ in the training set
- **Output:** Weights associated with training data that achieves statistical parity between a protected group compared with the privileged group.

Pre-Processing Example: Reweighing

Holistic AI

- Improved performance with respect to bias metrics
- Minimal decline in efficacy metrics

Metric	Value
Accuracy	0.84
Balanced Accuracy	0.75
Precision	0.73
Recall	0.55
F1-Score	0.63

Metric	Value	Reference
Statistical Parity	-0.09	0
Disparate Impact	0.55	1
Equality of Opportunity Difference	0.16	0
Average Odds Difference	0.06	0

In-Processing Example: Grid Search Reduction

Holistic AI

GSR (Agarwal et al 2018) reduces classification to a series of cost-sensitive, weighted classification problems. The optimal solution is a deterministic classifier with the lowest empirical error subject to fair classification constraints amongst the candidates searched.

- **Inputs:** Training Data, Machine Learning Algorithm, Protected Attributes
- **Approach:** Consider a grid of constraints and calculate the best approach for each.
- **Output:** Classifier that minimizes prediction error subject to the selected fairness constraints

In-Processing Example: Grid Search Reduction

Holistic AI

- Improved performance with respect to Equality of Outcome (SP, DI) but decreased with respect to Equality of Opportunity (EOD, AOD)
- Minimal decline in efficacy metrics

Metric	Value
Accuracy	0.83
Balanced Accuracy	0.71
Precision	0.72
Recall	0.47
F1-Score	0.57

Metric	Value	Reference
Statistical Parity	-0.005	0
Disparate Impact	0.97	1
Equality of Opportunity Difference	0.34	0
Average Odds Difference	0.19	0

Post-Processing Example: Equalised Odds

The Equalised Odds Post-Processing approach (Hardt et al 2016) flips predictions outputted from our base model until the desired error rate distribution between the protected group and the privileged group is achieved.

- **Inputs:** Predicted Probabilities from base-model, Target Labels, Protected Attributes
- **Approach:** Convex optimisation to find the optimum rates at which to flip predictions, using the True Positive Rates and False Positive Rates for each group.
- **Output:** Predictions that achieve equalised odds (Equality of Opportunity).

Post-Processing Example: Equalised Odds

Holistic AI

- Improved performance with respect to bias metrics, especially for Equality of Opportunity
- Minimal decline in efficacy metrics

Metric	Value
Accuracy	0.83
Balanced Accuracy	0.73
Precision	0.66
Recall	0.55
F1-Score	0.61

Metric	Value	Reference
Statistical Parity	-0.09	0
Disparate Impact	0.58	1
Equality of Opportunity Difference	0.02	0
Average Odds Difference	0.005	0

Considerations

Holistic AI

- Consider fairness throughout the full Machine Learning lifecycle
- Fairness is not just a data scientist problem:
  - A definition and understanding of impacts on individuals and groups must have been agreed upon with accountable experts
- There are some trade-offs one must account for when considering bias and mitigation of bias

Holistic AI | The Alan Turing Institute

Other verticals of Trustworthy AI & Trade-offs

Transparency

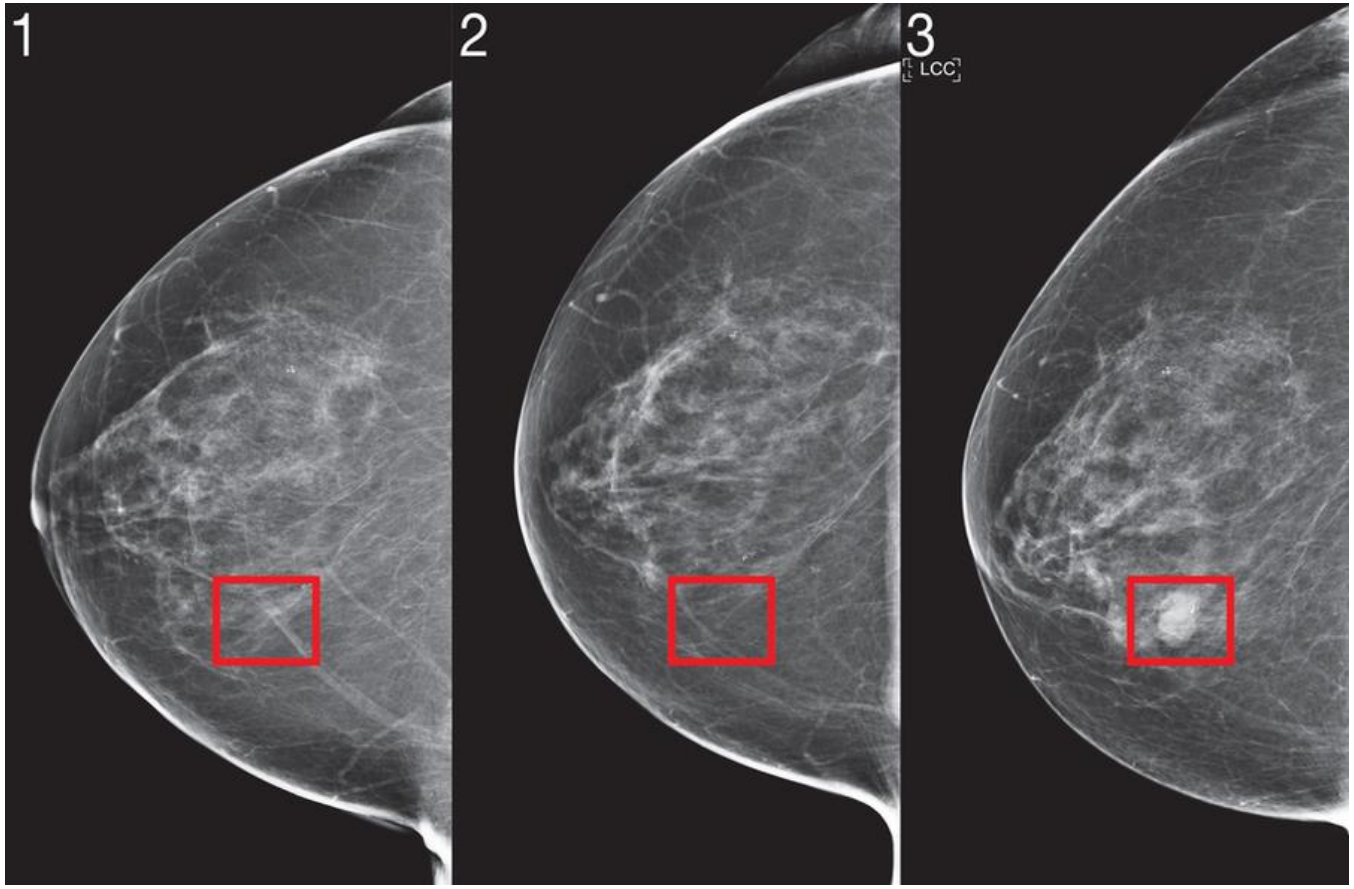
Holistic AI

- **Interpretability**: "the degree to which a human can understand the cause of a decision“ [Miller, 2018]
- **Explainability**: the degree to which the inner mechanics of an algorithm are understood by a human
- **Transparency** is an umbrella term that encompasses concepts such as Explainability and interpretability

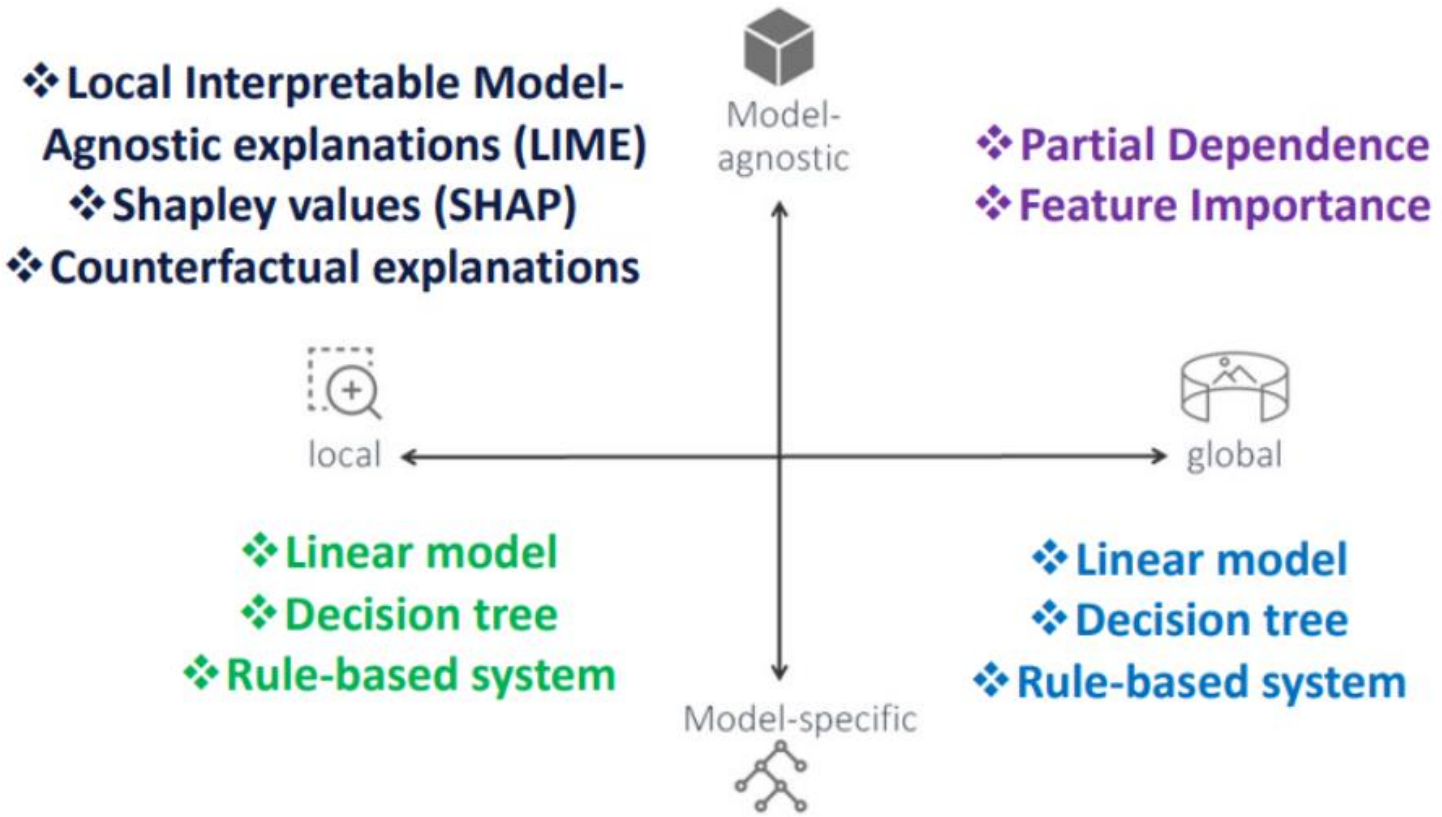
Examples

- Cancer diagnosis (or any other medical diagnosis algorithm)

- Automated loan approval/refusal



Overview of Explainability Techniques



[Koshiyama et al.,2021]

Feature Importance

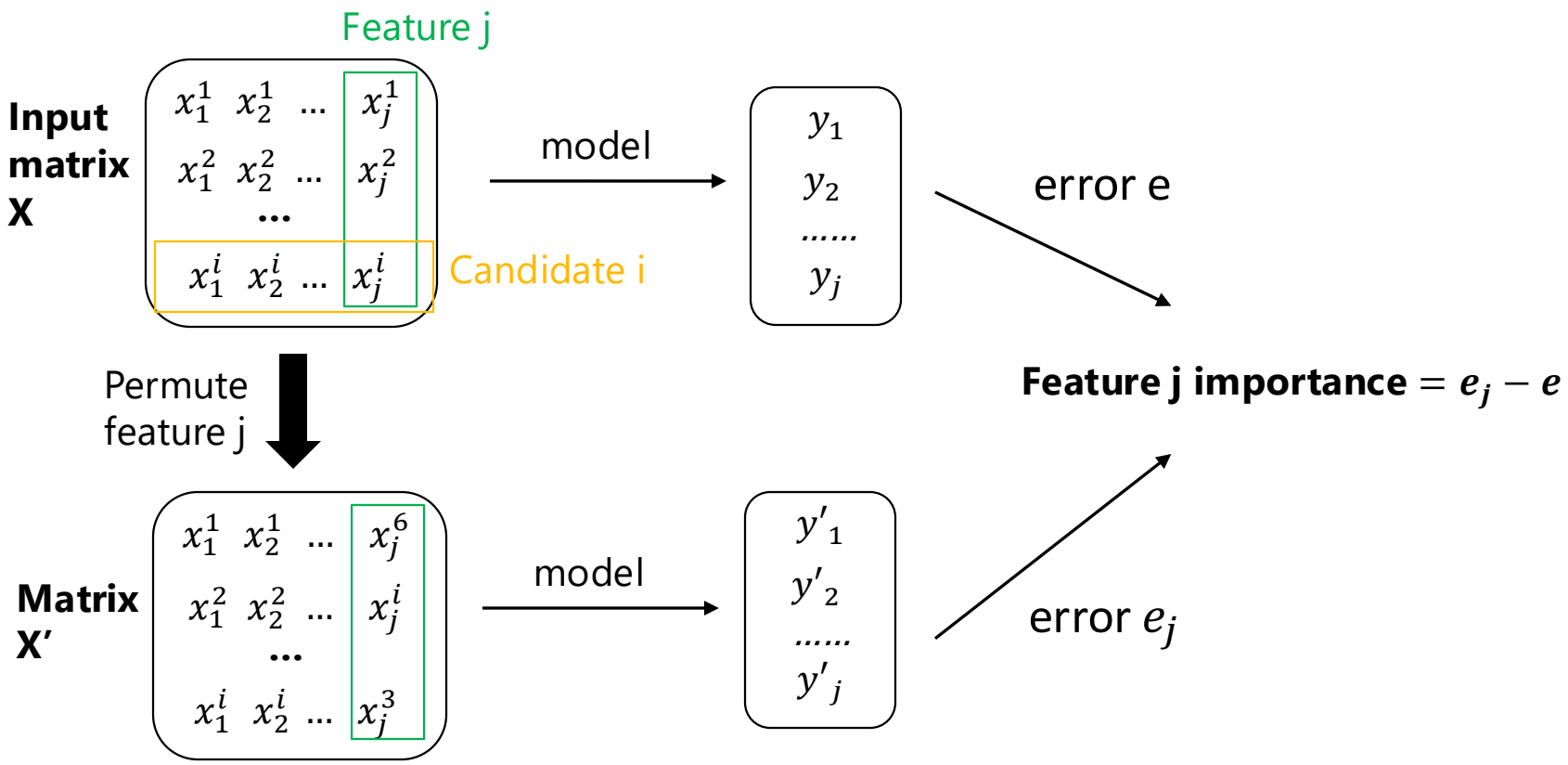
Holistic AI

- Looks to assign a score to each feature relative to its importance in the prediction.

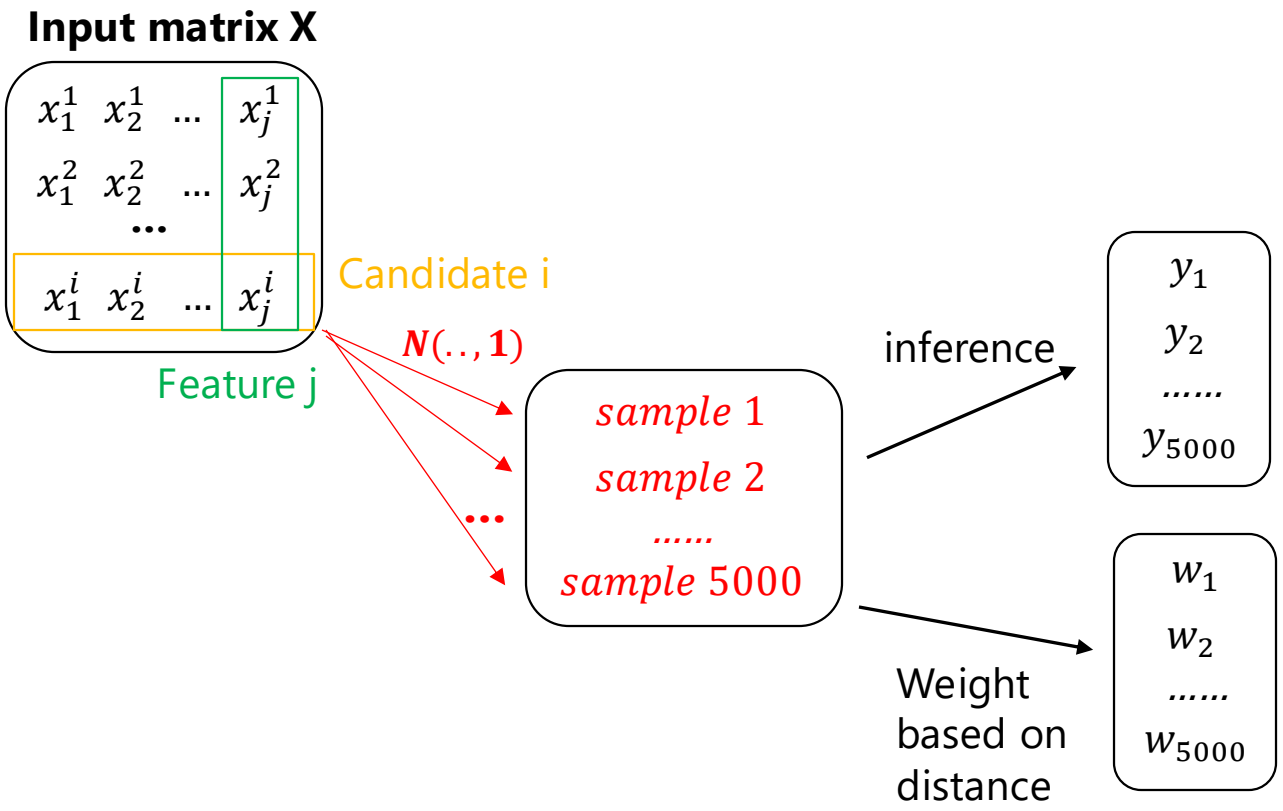
- Simple example in linear regression:

$$y = \underbrace{w_1 * x_1}_{\text{Importance of feature 1 in outcome y}} + w_2 * x_2 + \ldots + w_n * x_n$$

Permutation Feature Importance



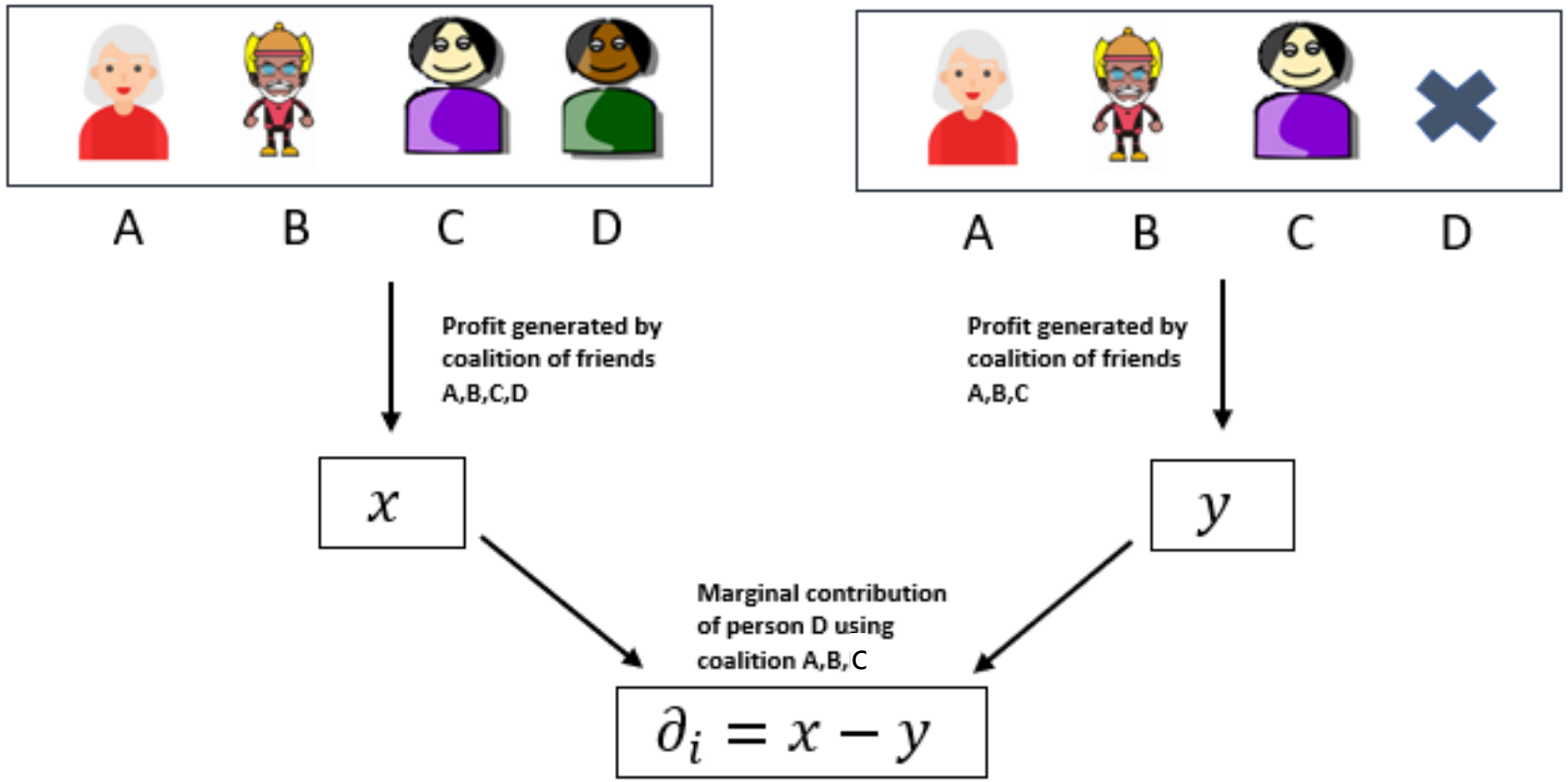
LIME (Local Interpretable Model-agnostic explanations)



Explain results for candidate i:

- Sample 5,000 times candidate vector i using a normal distribution
- Predict output
- Assign weight base on distance
- Fit interpretable model locally with weights on new dataset
- Interpret local model.

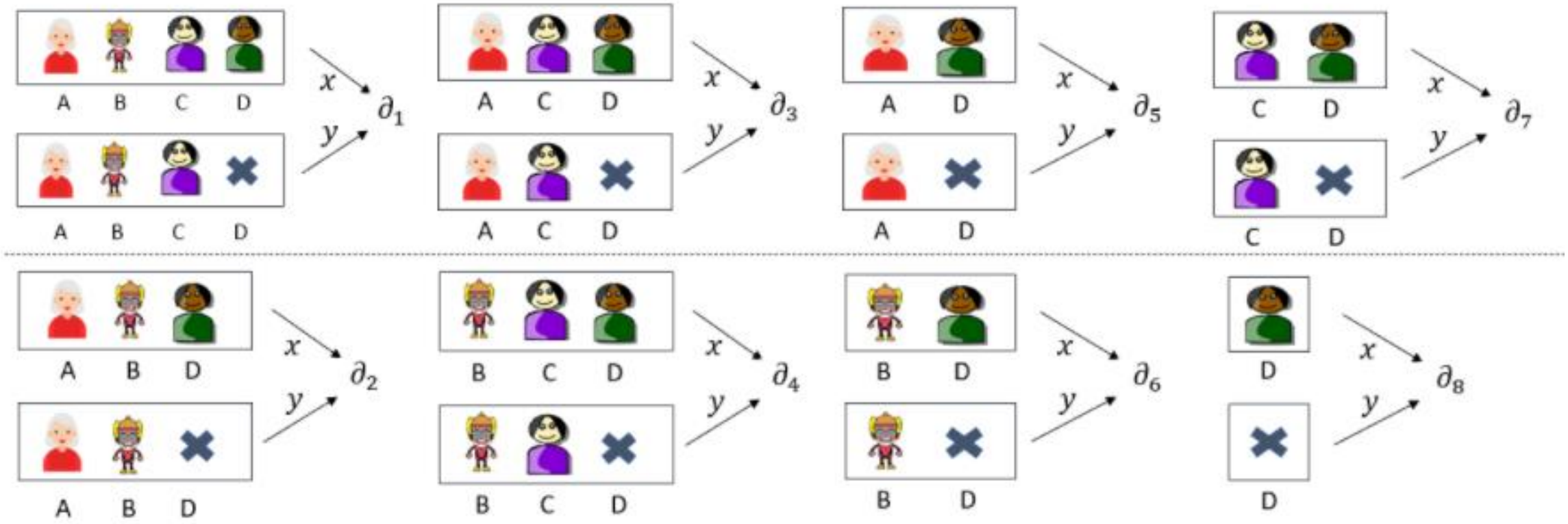
SHAPLEY Values



Adapted from https://towardsdatascience.com/shap-shapley-additive-explanations-5a2a271ed9c3 by Fernando Lopez

SHAPLEY Values

Holistic AI

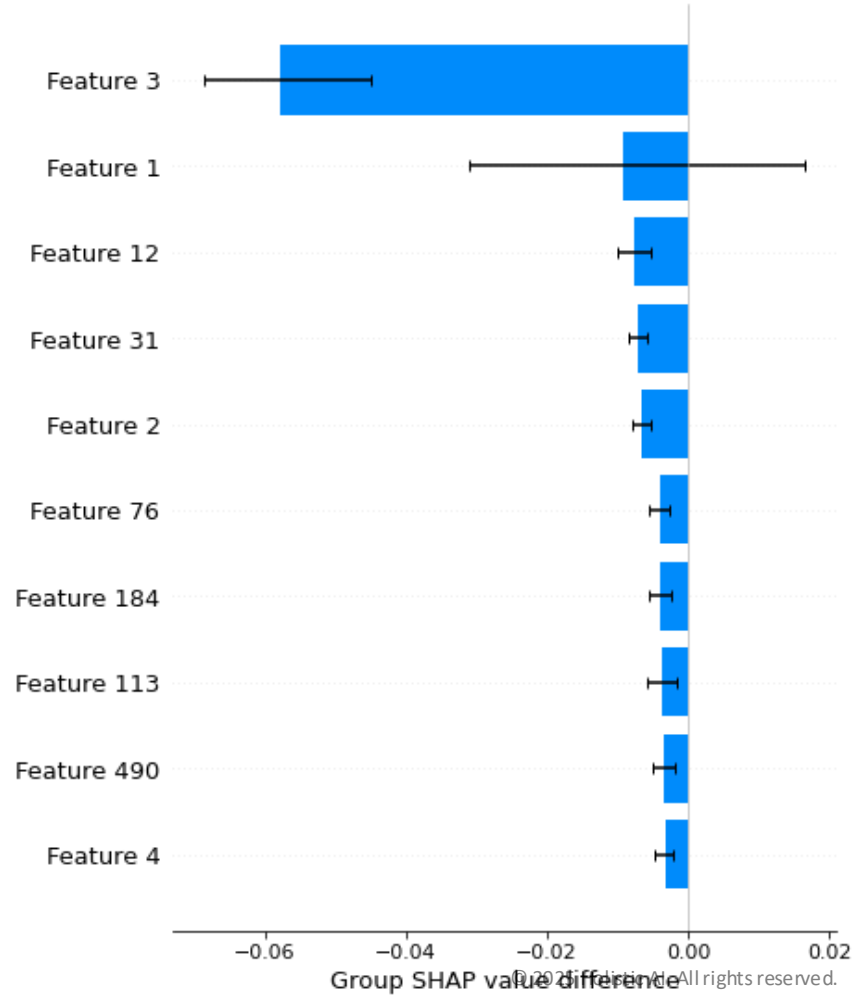


The shapley value for person D is therefore: $\Phi_D = \frac{\delta_1+\delta_2+\delta_3+\delta_4+\delta_5+\delta_6+\delta_7+\delta_8}{8}$

Positive Interactions with Fairness

Holistic AI

- Are the most influential factors reasonable? Are they proxy for a protected characteristics?
- Are the influential factors the same across different groups ?
- Instead of explaining output → explain fairness metric
  - Example: effect of permutation importance on Disparate Impact metric.



Negative Interactions with Fairness

Holistic AI

- In some other context, mitigating for bias may make a model less explainable
- Example: Unlocking via facial recognition. Retraining the model by adding more fairness constraints can make the decision boundary become more complex therefore making the model less explainable.
- In this example, interpretable methods (black box) can still be used for different groups.

Robustness

Technical Robustness and Safety (EU guidelines):

- Resilience to attack and security
- Fallback plan and general safety
- Accuracy
- Reliability and Reproducibility

EU-HLEG. (2019). Ethics guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Resilience to attack and security

- Quality of a system to be safe, not vulnerable to tampering.
- Protect against hacking: data poisoning, model leakage or the infrastructure, both software and hardware.

Fallback plan and general safety

- Safeguards that enable a fallback plan in case of problem
- Continue operation with minimisation of unintended consequences and errors: human-in-the-loop, switching to rule-based,…

Accuracy

- High level of accuracy desirable
- Explicit and well-formed development and evaluation process

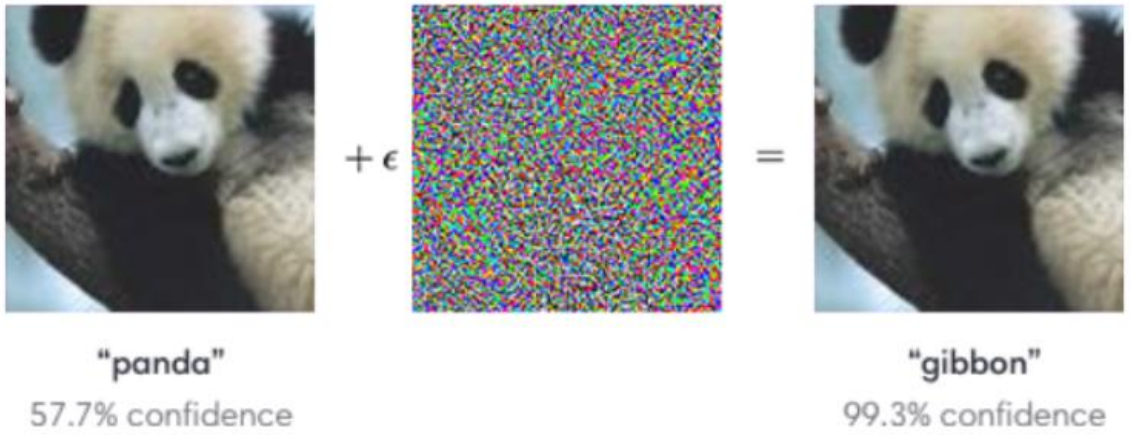
Reliability and Reproducibility

- Works properly with a range of inputs and in a range of situations
- Same behaviour when repeated under the same conditions

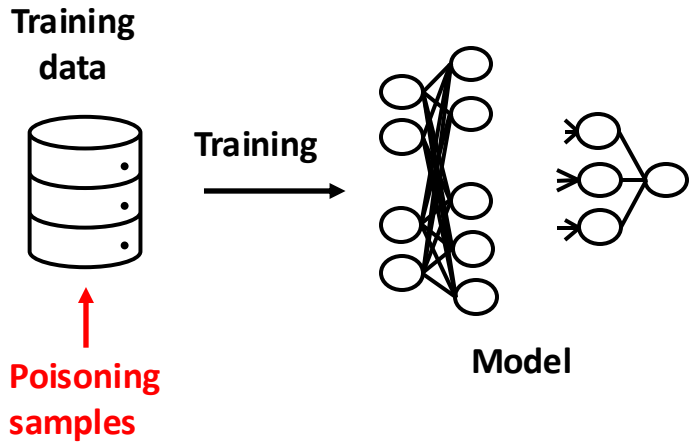
ML attacks

Evasion attacks.

- manipulating input data to evade a trained classifier at test time
- do not compromise normal system operation



[Goodfellow et al., 2015]



Poisoning attacks.

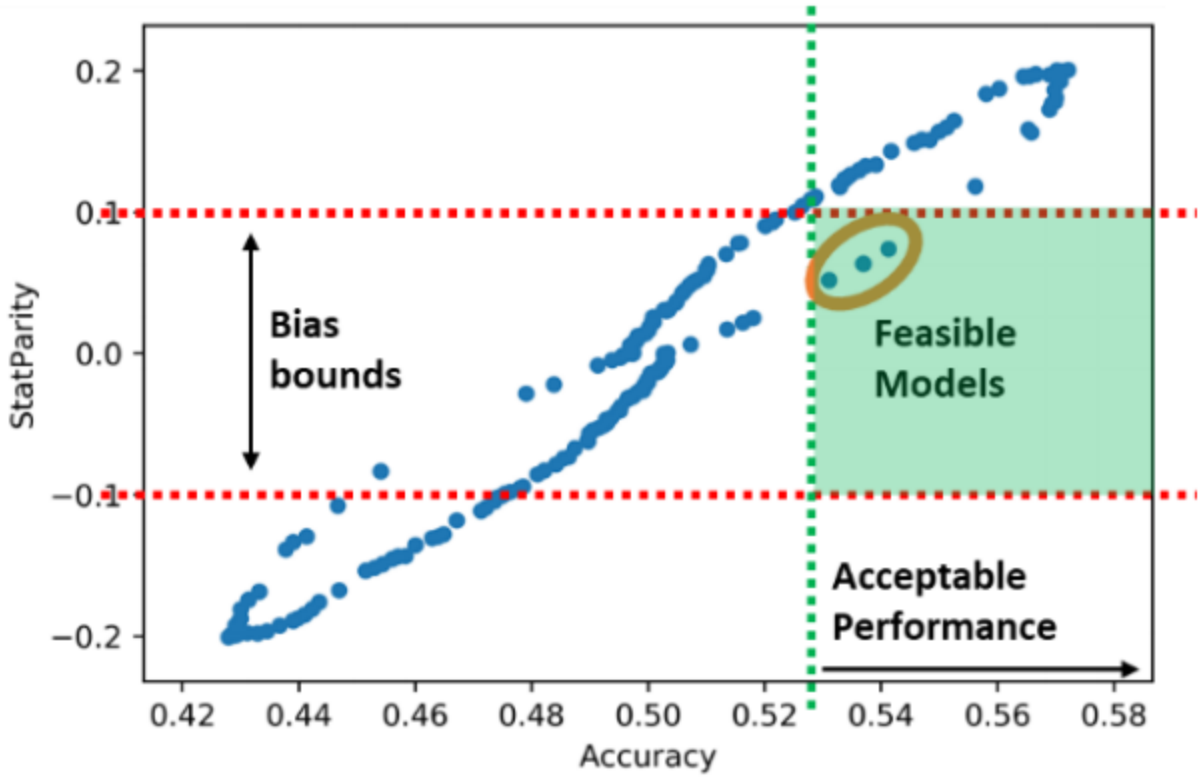
- injecting a small fraction of poisoning samples into the training data (occur during the training phase) to increase misclassification at test time
- compromise normal system operation

Defences against these attacks?

Holistic AI

- Adversarial training
- Regularization
- Outlier detection
- Ensemble learning

Bias vs Accuracy



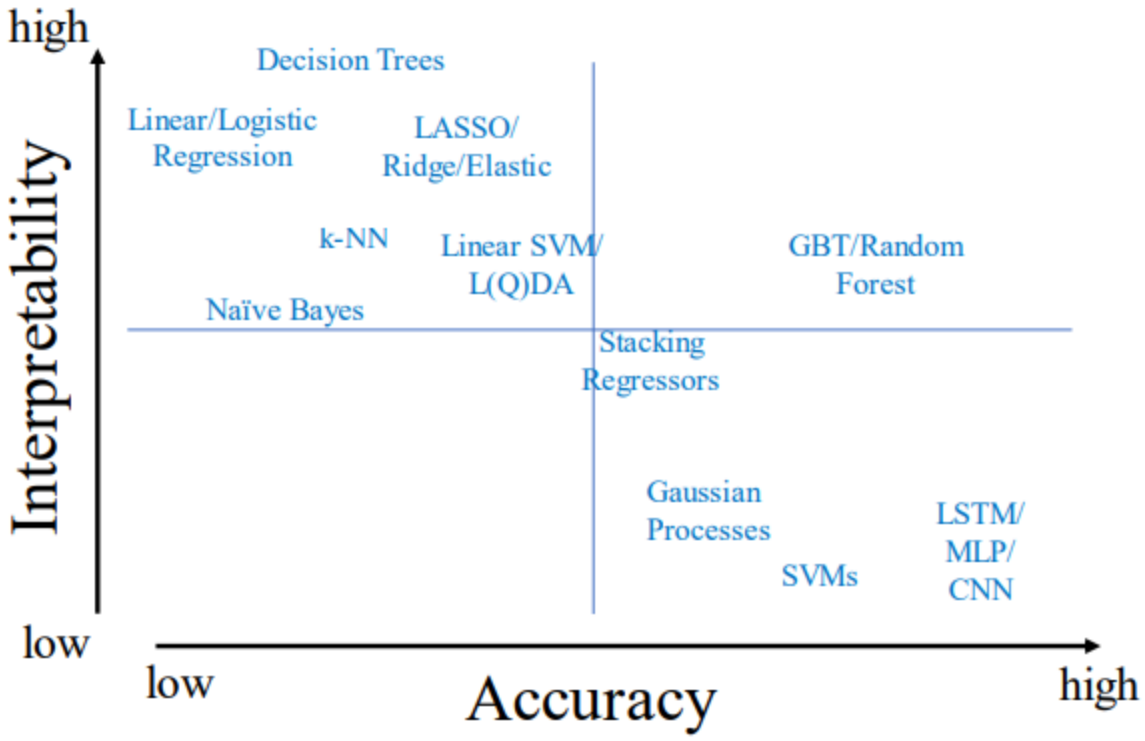
Towards Algorithm Auditing
by Koshiyama et al., 2021

Trade-offs: Adversarial Robustness vs Fairness

Holistic AI

- Paper by Xu et al. 2021
- “robust fairness” problem of adversarial training: large disparity of accuracy and robustness among different classes (not observed in natural training)
- adversarial training -> tendency to "favor the accuracy of the classes which are “easier” to be predicted."

Trade-offs:
Accuracy vs Explainability



Towards Algorithm Auditing by Koshiyama et al., 2021

Privacy and Security

Holistic AI

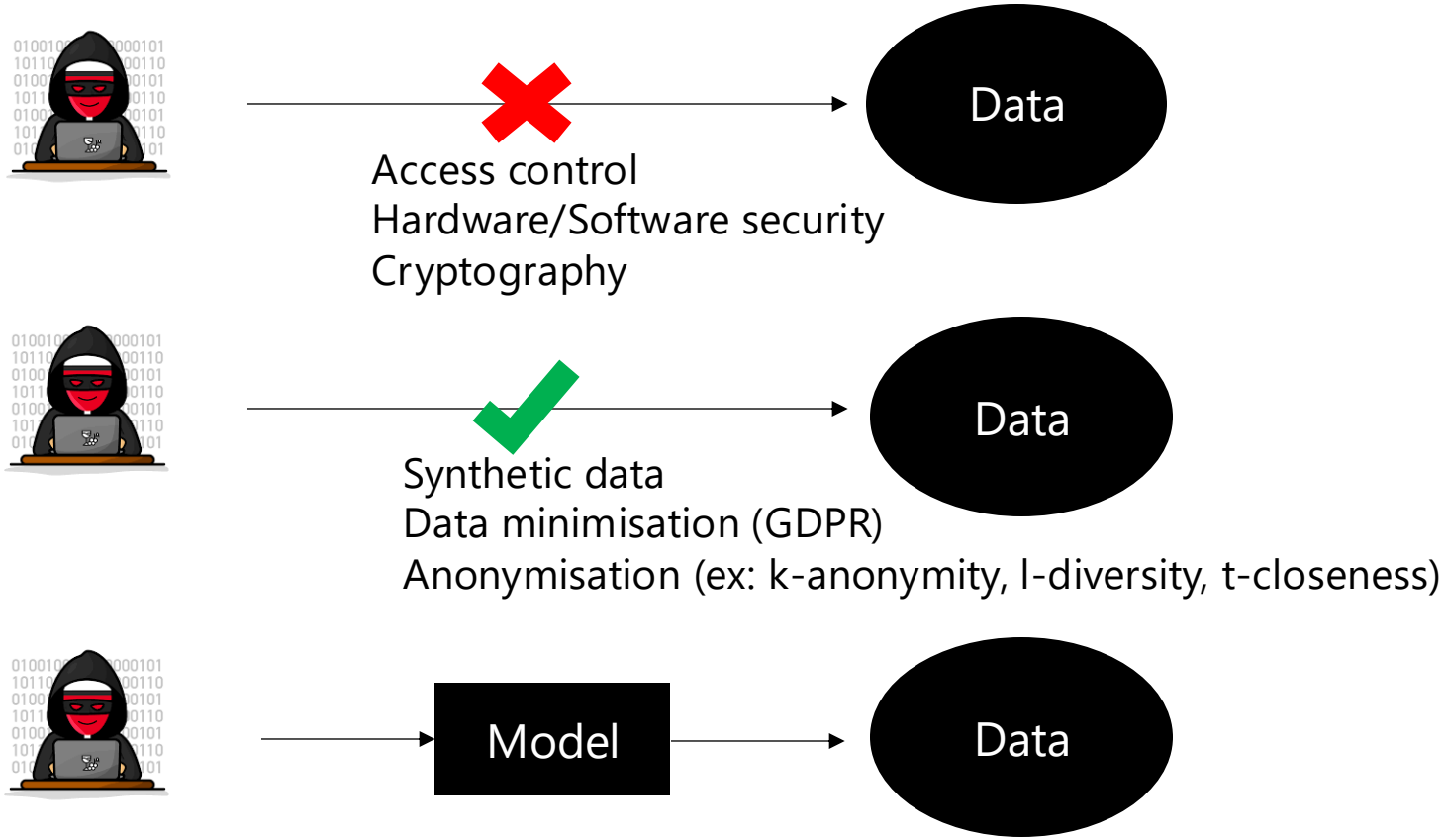
Algorithms must guarantee data protection throughout a system's entire lifecycle, whether this is user information provided by the user or generated by the system.

European Commission, *Ethics guidelines for trustworthy AI*

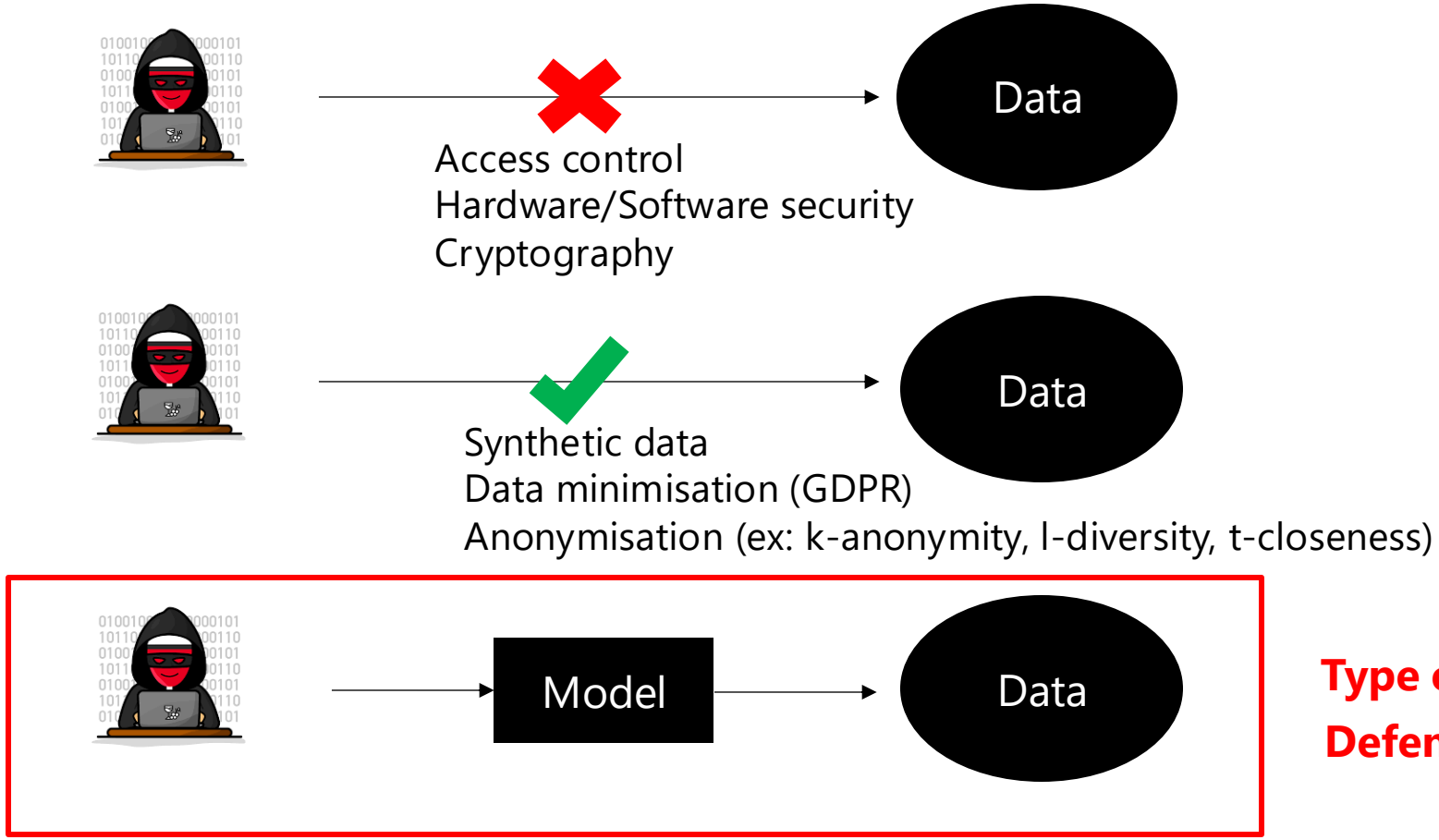
Examples

- Uber data breach
- SolarWinds, develops enterprise IT and cybersecurity software

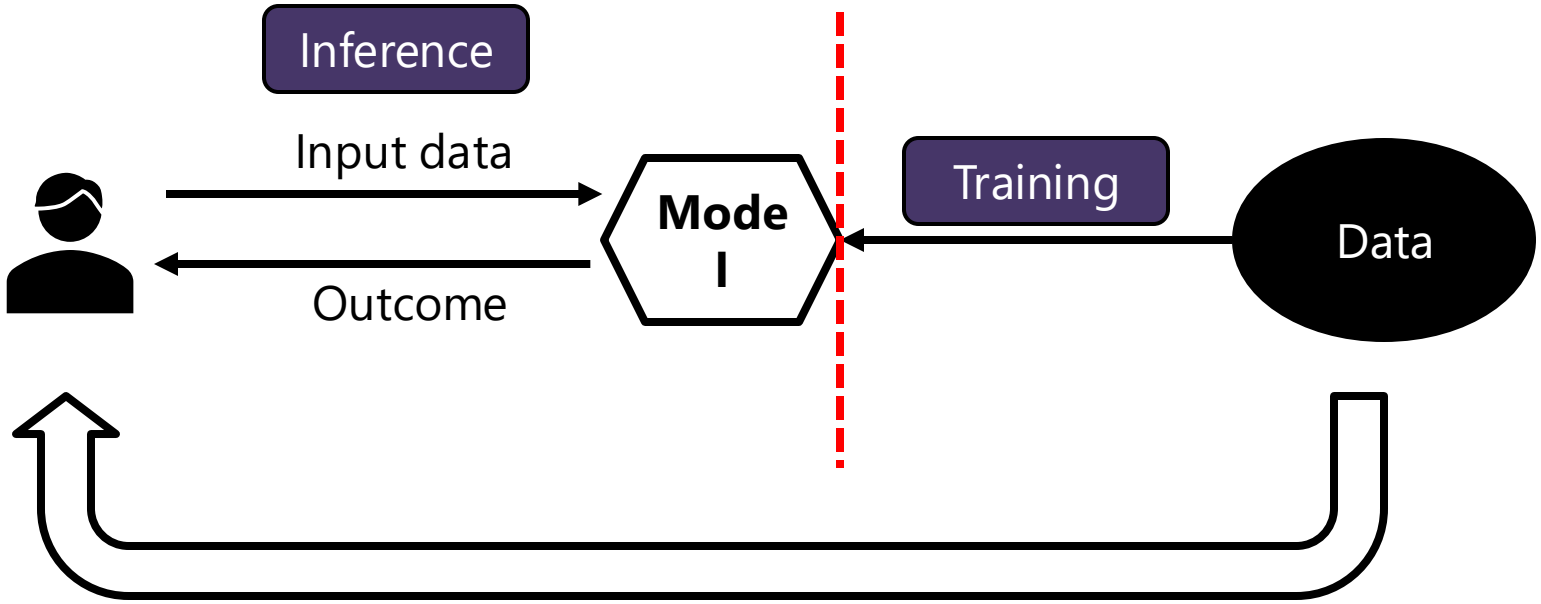
Threats depend on attacker access



Our focus today



Typical black-box setting



Can the attacker infer something from the data with this set up?

Two types of attacks

Membership Inference

- Is this sample part of the training data?
- Attack eased when more information about training data is “encoded” into the model itself, ex. when overfitting occurs.
- Example: Medical study about Alzheimer disease. If we know a certain hospital record has been used to train a model ➔ reveals this patient has Alzheimer.

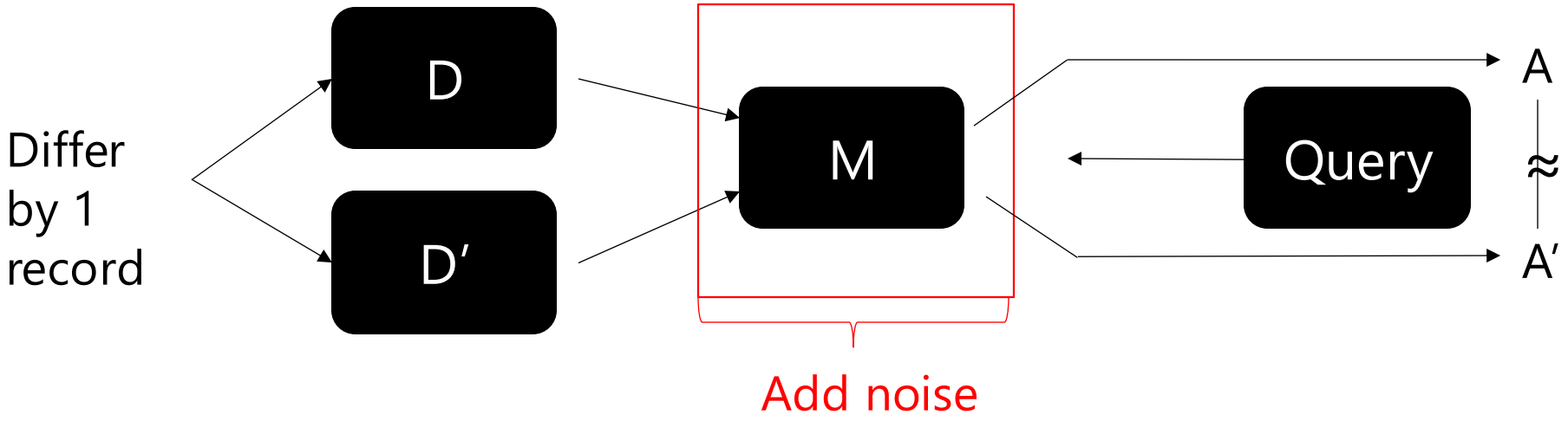
Model extraction

- attacker tries to create a mock-up model by querying the models for various inputs
- Complex models harder to reconstruct.
- Overfitting prevents attack.
- mock up model can then be used for other adversarial attacks

See other types of attacks in [Rigaki & Garcia, 2021]

Against Membership Inference: Differential Privacy

- When we cannot determine whether a particular individual has been used in training.



Privacy <> Fairness

- Sensitive information: sex, gender, religion, ethnicity, etc.
- Highly overlaps with information required to measure/mitigate group fairness
- Quasi-Identifiers that could help re-identification attacks

Privacy <> Transparency

- Model can leak information about training data
- But model transparency helps with explainability & interpretability, which itself helps with fairness

Holistic AI

Questions?

Holistic AI

holisticai.com

we@holisticai.com