University of Glasgow | School of Computing Science

THE AWARDS 2020 | UNIVERSITY OF THE YEAR

Human-Centered ML

Dr. Fani Deligianni,

fani.deligianni@glasgow.ac.uk

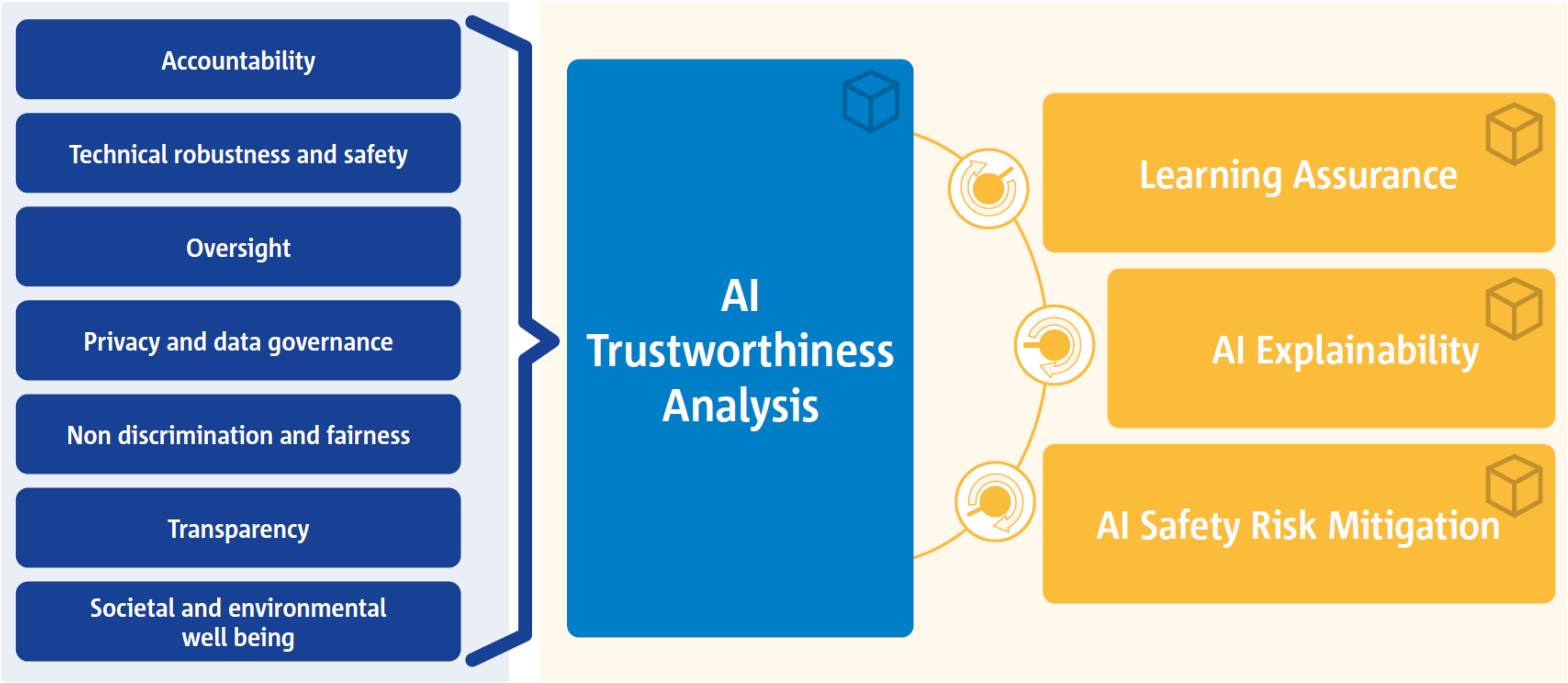
Lecturer (Assistant Professor)

Lead of the Computing Technologies for Healthcare Theme

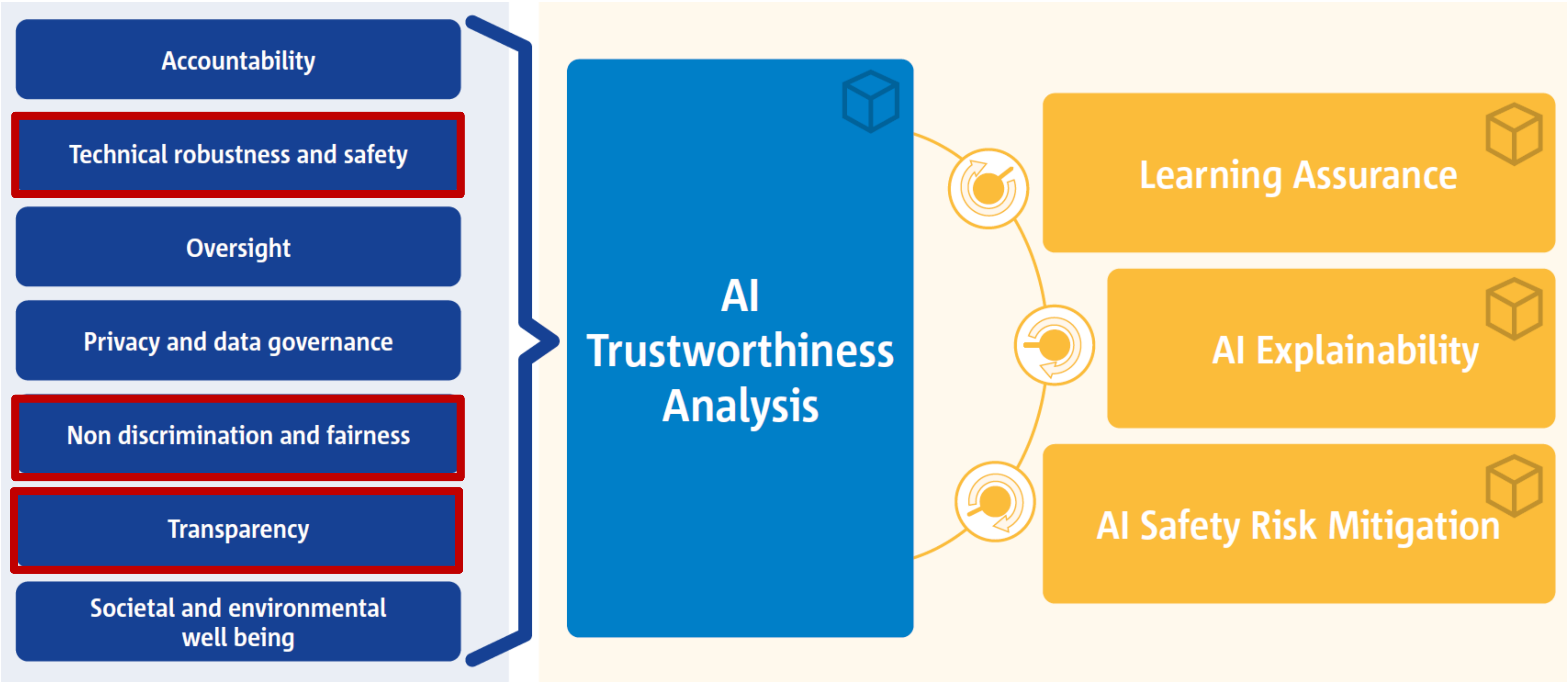
https://www.gla.ac.uk/schools/computing/staff/fanideligianni

WORLD CHANGING GLASGOW

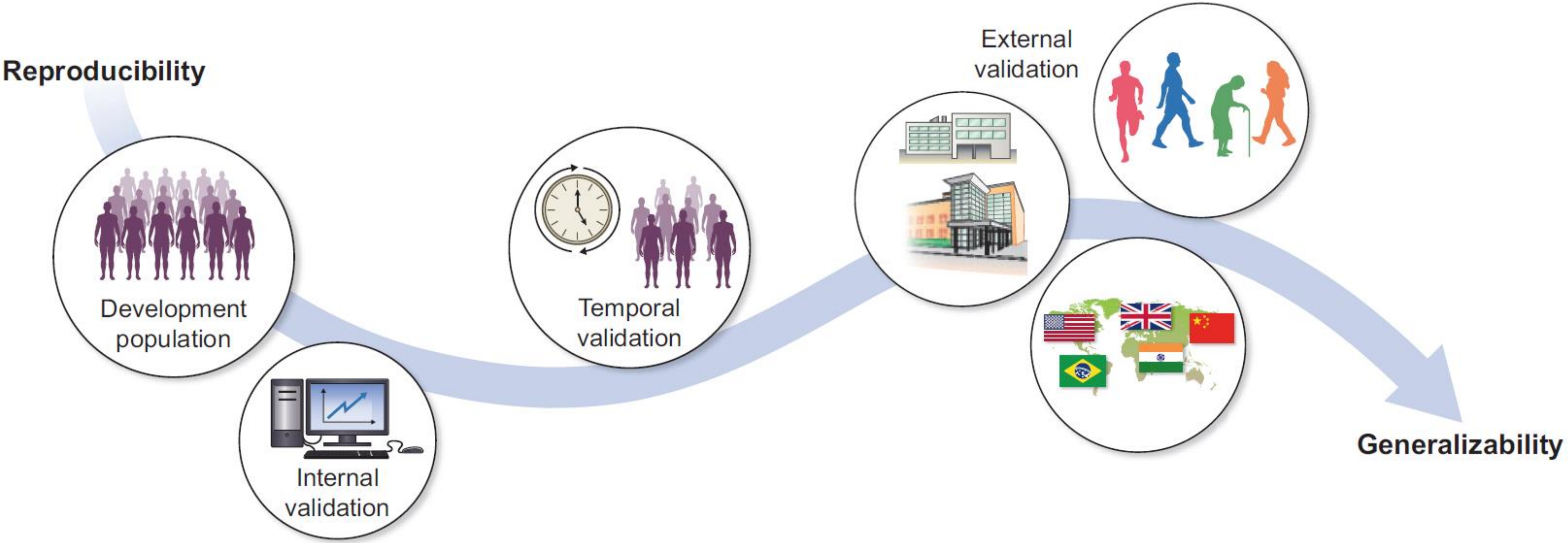
Why do we need Human-Centred AI?



Why do we need Human-Centred AI?



Reproducibility - Generalizability



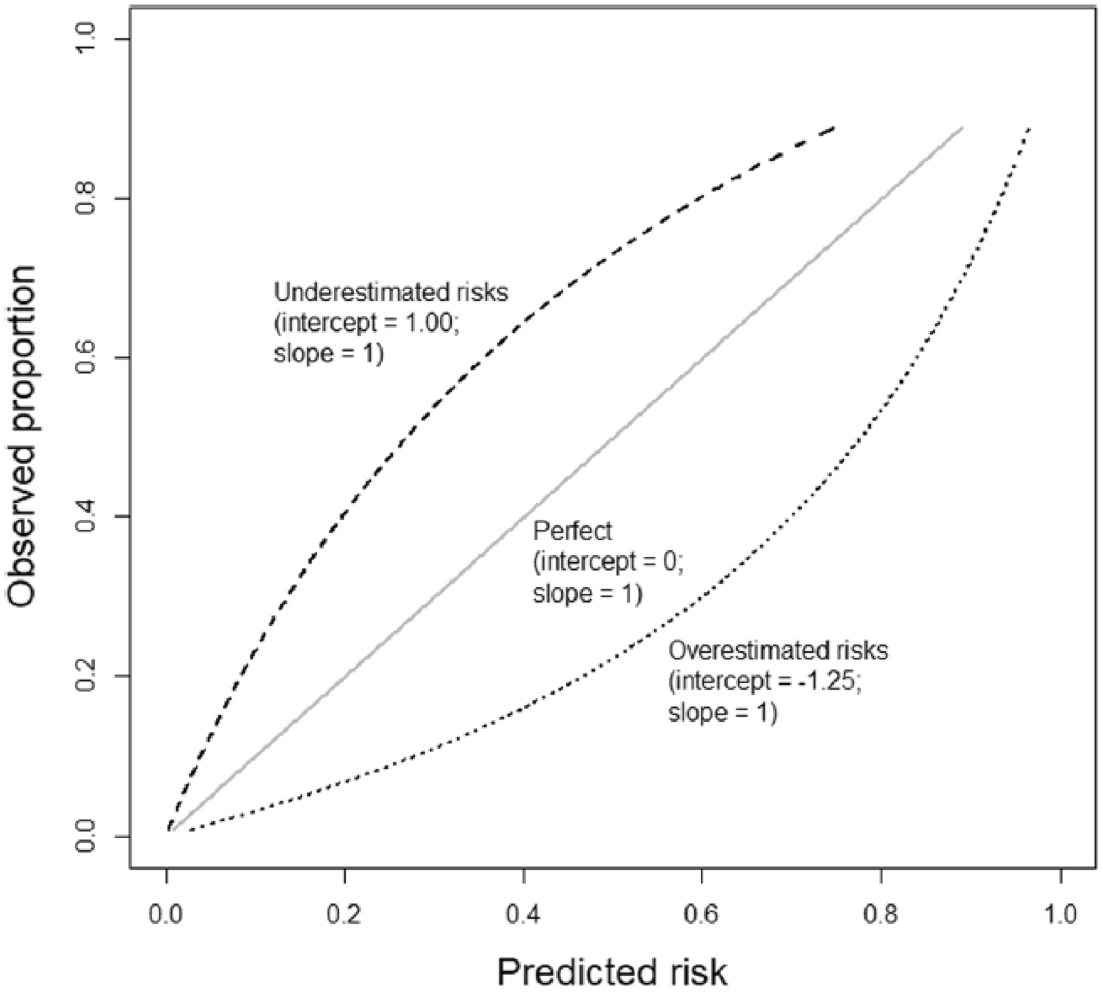
Ramspek et al. ‘External validation of prognostic models: what, why, how, when and where?’, Clinical Kidney Journal, 2021.

ABCD Guide

A. Calibration-in-the-large, or the model intercept
B. Calibration slope
C. Discrimination with the Receiver Operating Characteristic curve
D. Clinical usefulness with decision-curve analysis

Steyerberg et al. ‘Towards better clinical prediction models: seven steps for development and an ABCD for validation’, European Heart Journal, 2014.

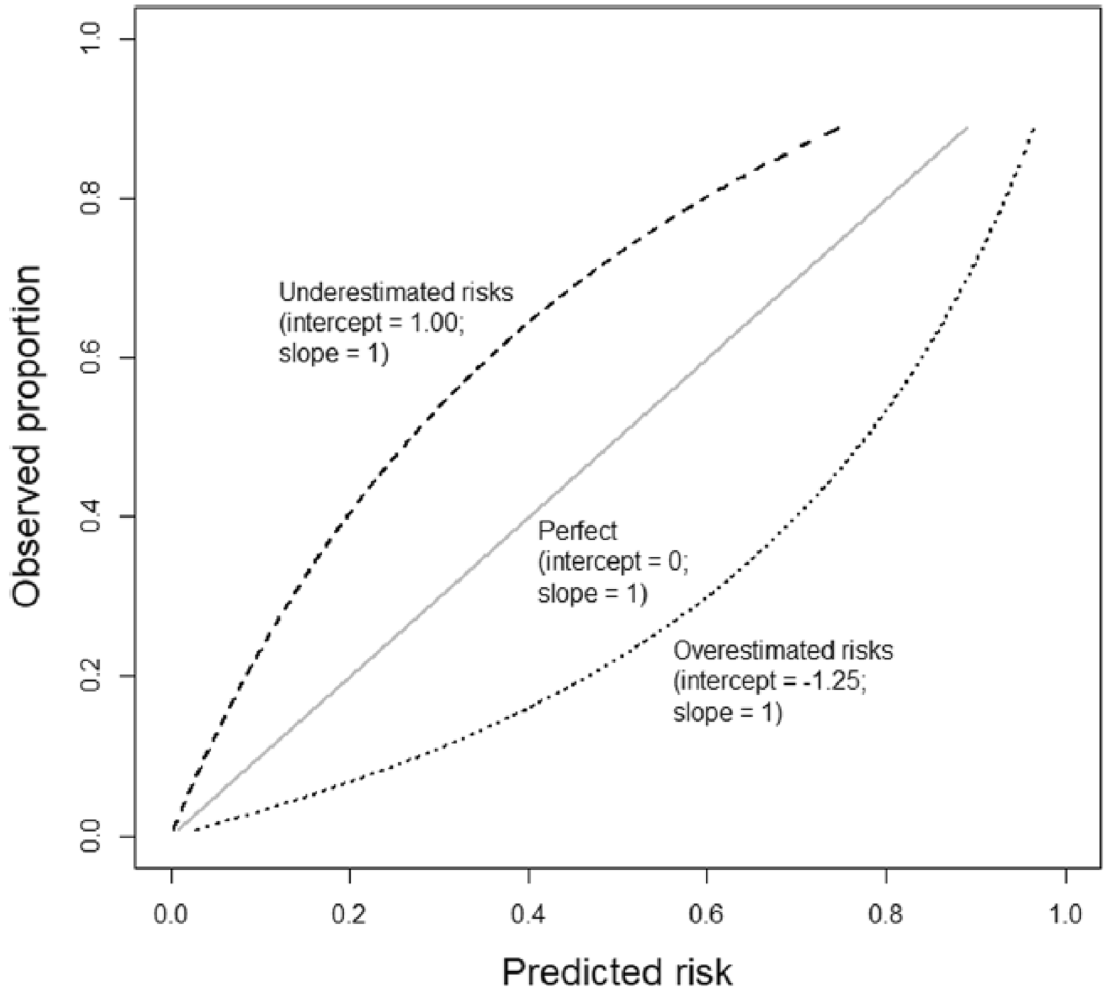
Assessing Calibration

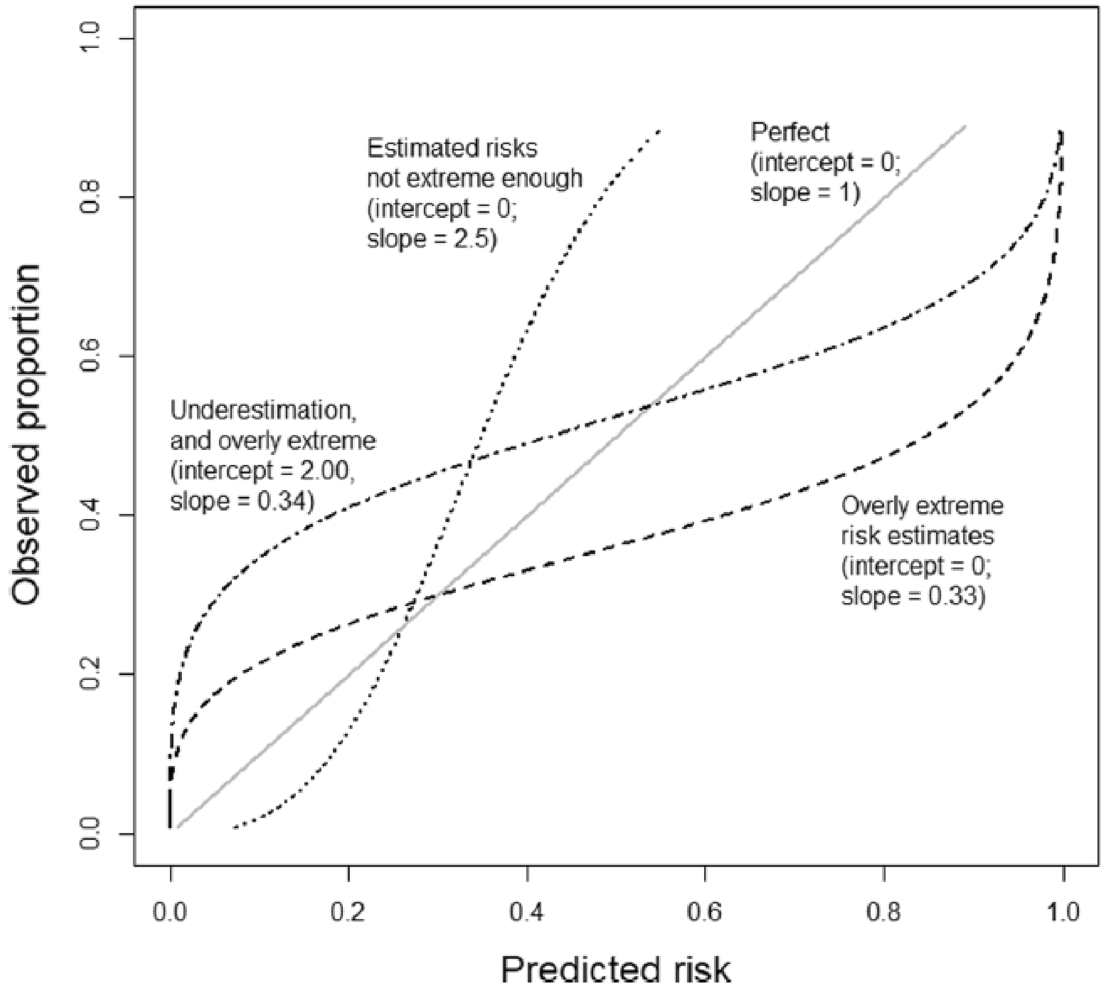


$$\mathbb{P}\left(\hat{Y} = Y \mid \hat{P} = p\right) = p, \quad \forall p \in [0, 1]$$

Calster et al. Calibration: the Achilles heel of predictive analytics, 2019

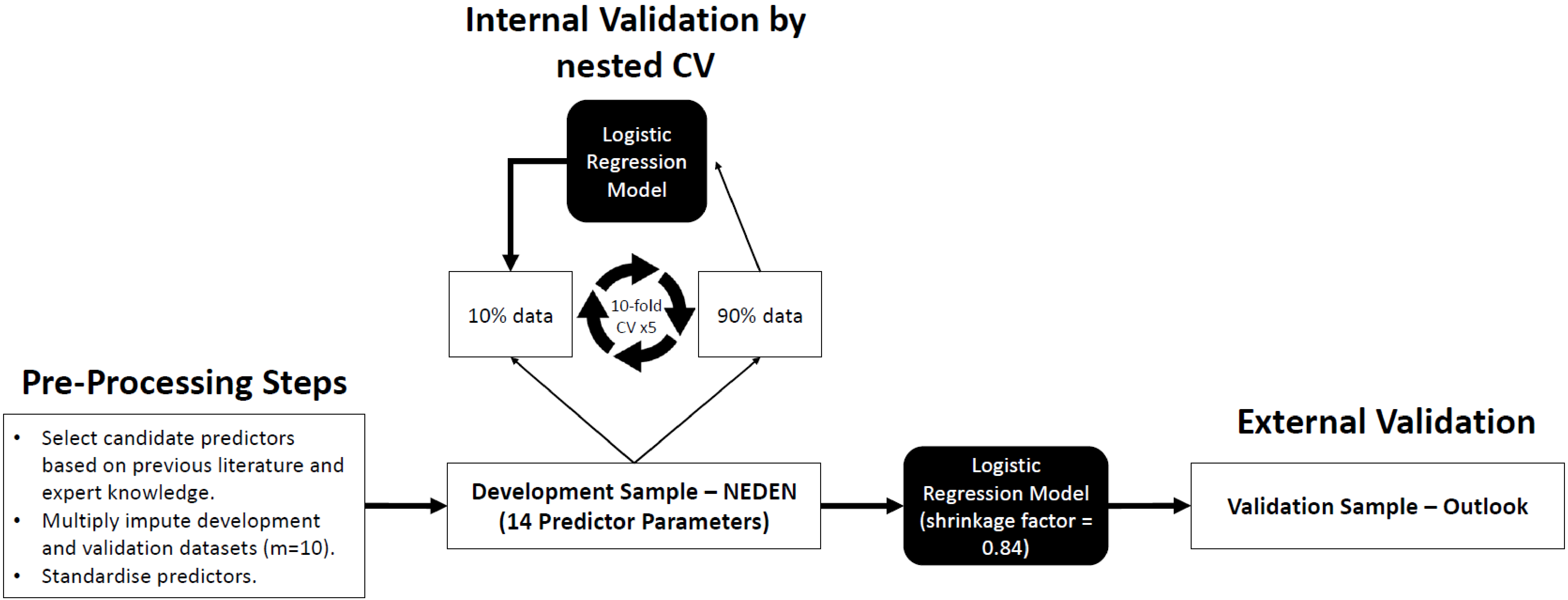
Examples of Extreme Calibration





Calster et al. Calibration: the Achilles heel of predictive analytics, 2019

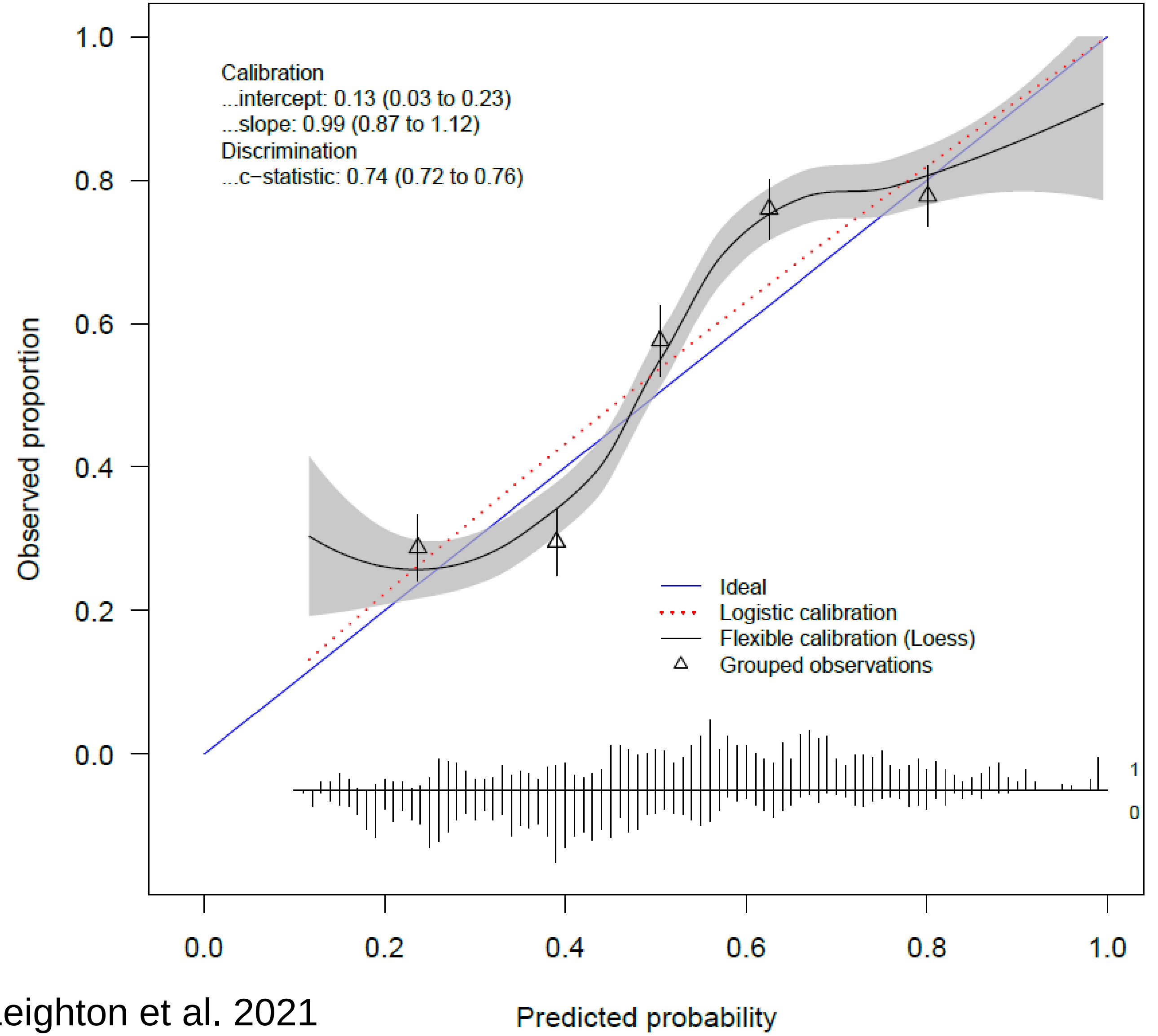
Example in First Episode Psychosis



Leighton et al. Development and validation of a non-remission risk prediction model in First Episode Psychosis: An analysis of two longitudinal studies, 2021

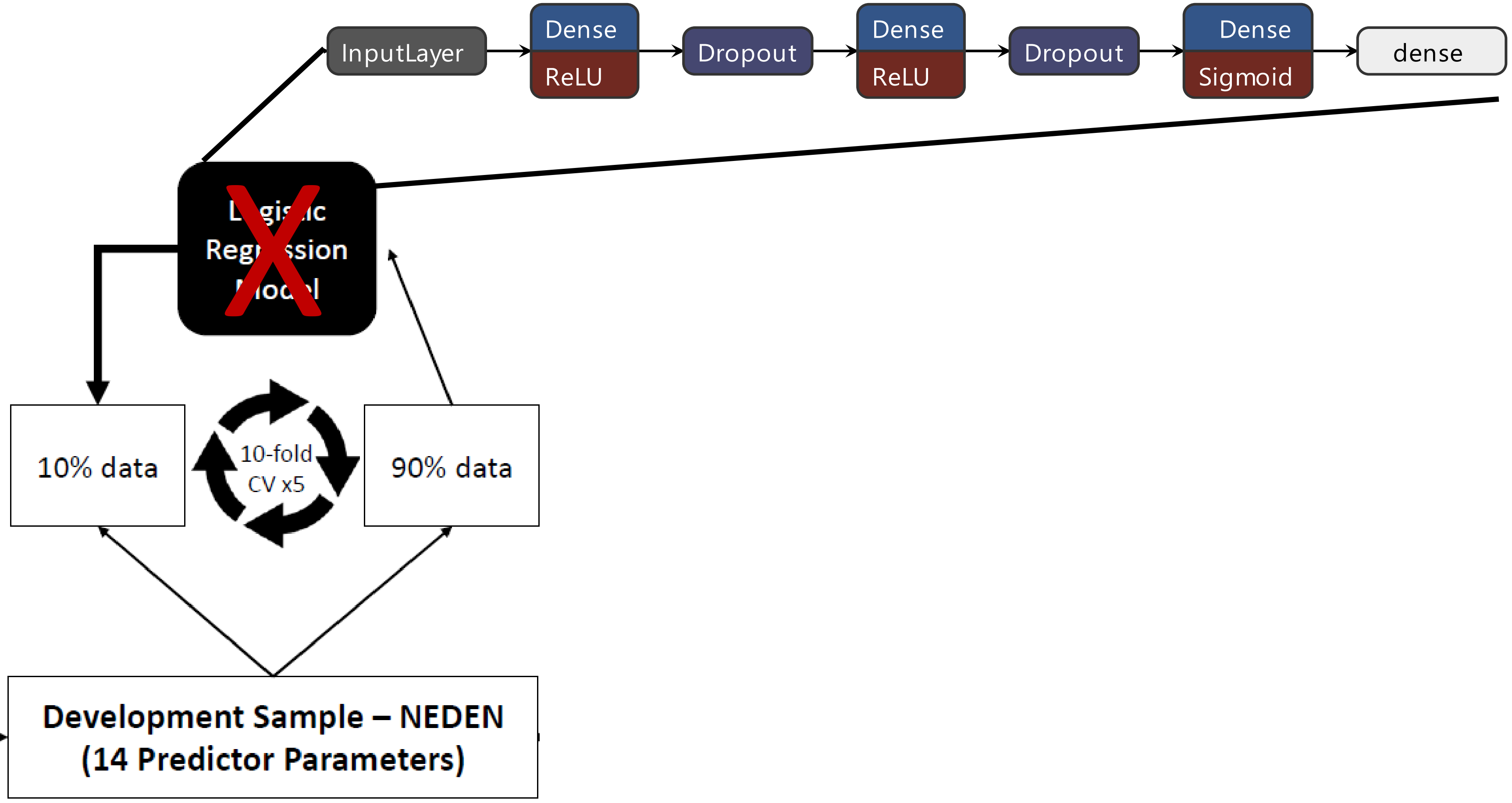
Example Calibration - Discrimination

- Calibration refers to the agreement between observed outcomes and predictions
- Calibration-in-the-large – external validation
- Calibration slope – internal validation



Leighton et al. 2021

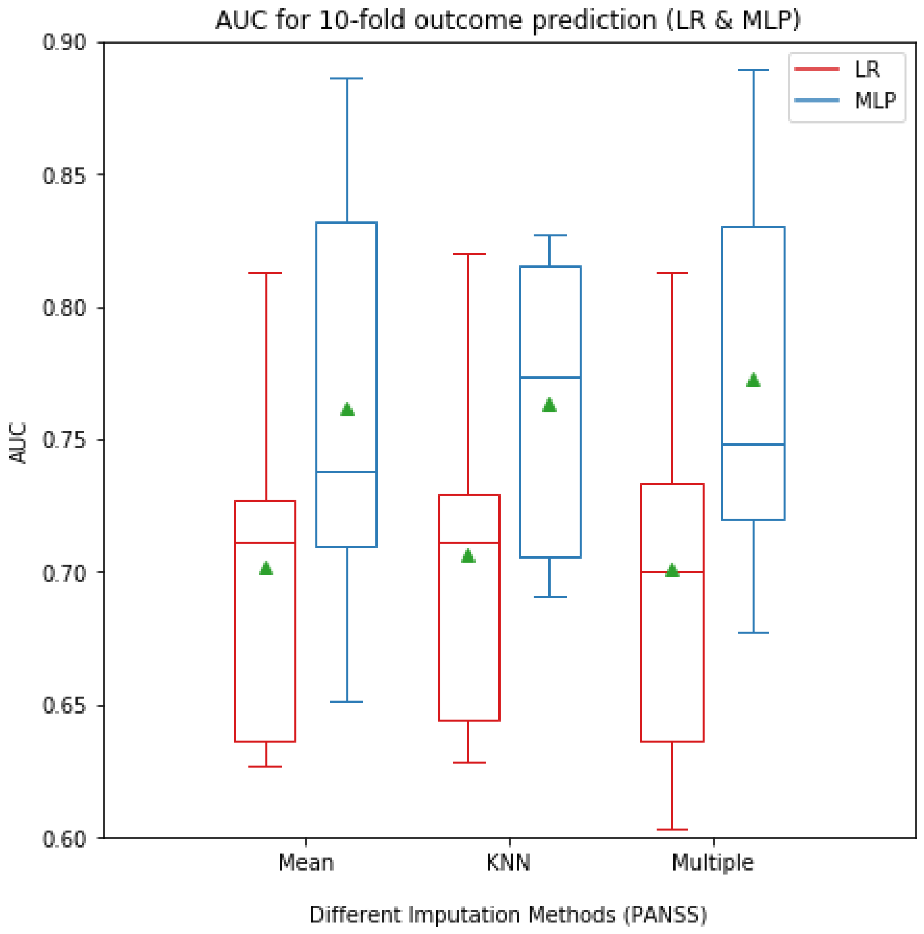
Example Calibration - Discrimination

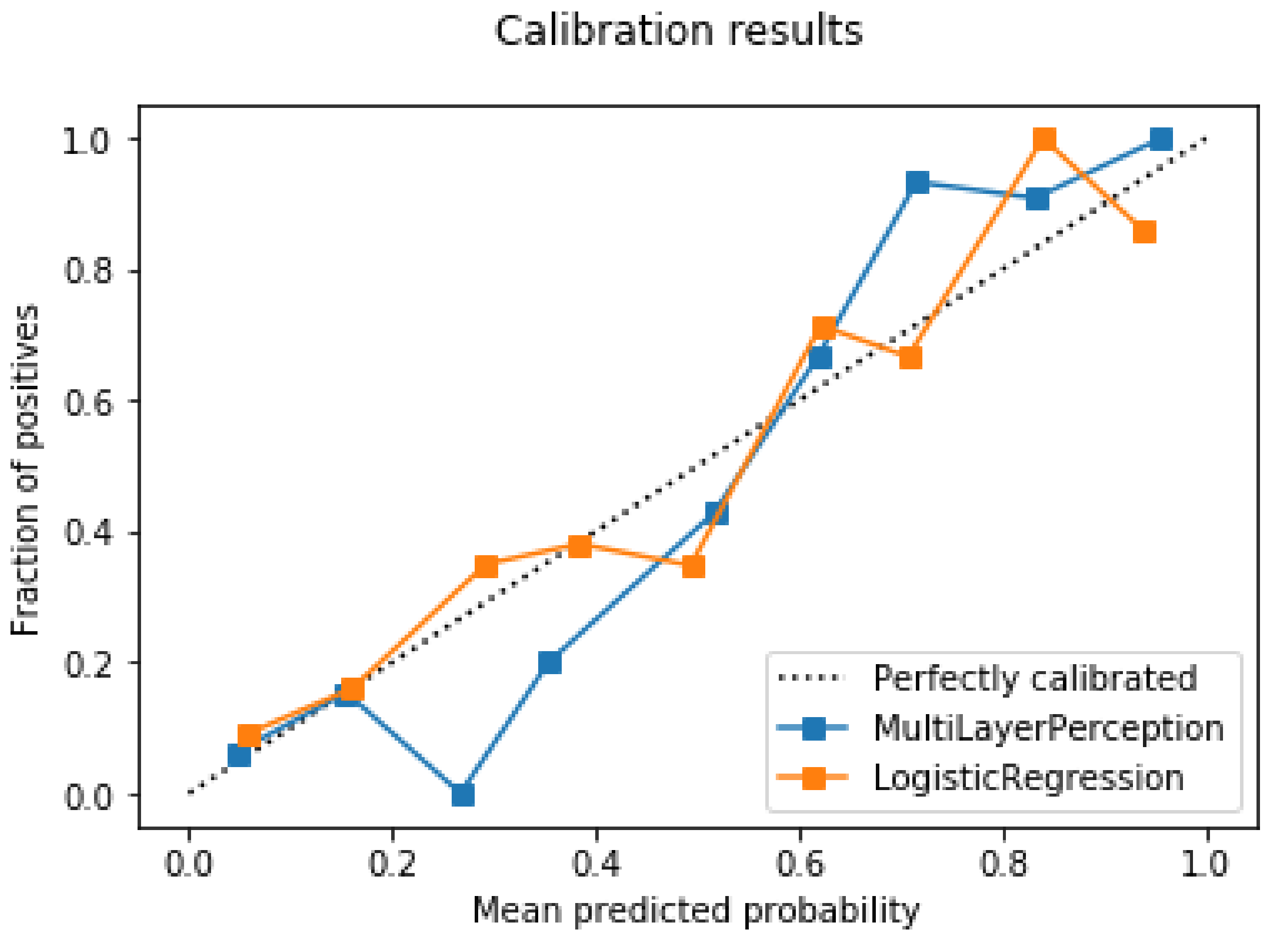


Pre-Processing Steps

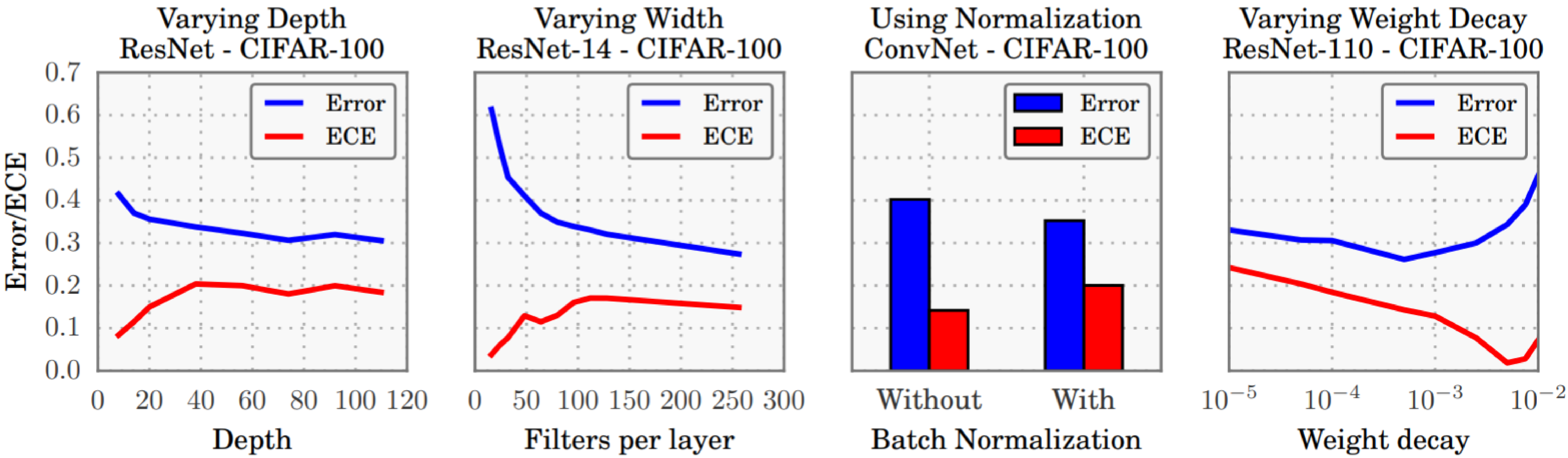
- Select candidate predictors based on previous literature and expert knowledge.
- Multiply impute development and validation datasets (m=10).
- Standardise predictors.

Example Performance - Calibration





Mis-Calibration DNN architectures



$$\mathbb{E}_{\hat{P}}\left[\left|\mathbb{P}\left(\hat{Y}=Y \mid \hat{P}=p\right)-p\right|\right]$$

$$\text{ECE}=\sum_{m=1}^{M}\frac{|B_m|}{n}\left|\text{acc}(B_m)-\text{conf}(B_m)\right|$$

$$\text{acc}(B_m)=\frac{1}{|B_m|}\sum_{i\in B_m}\mathbf{1}(\hat{y}_i=y_i)$$

$$\text{conf}(B_m)=\frac{1}{|B_m|}\sum_{i\in B_m}\hat{p}_i$$

Guo et al. ‘On Calibration of Modern Neural Networks’, 2017

Clinical Consequences

- Breast cancer detection as a use case:
  - A false-negative result is much more harmful than a false-positive result
  - A model with greater specificity but slightly worse sensitivity could have a better AUC
  - Worse choice for a clinical decision system for breast cancer detection

Decision Analysis

- Decision Trees
  - Assign probabilities and
- Explicit valuation of health outcomes
  - Number of complications prevented
  - Quality-adjusted life-years saved

A Simpler Method

For each model:

For p_t in range(a,b):

Calculate the number of true- and false-positive results using p_t as the cut-point for determining a positive or negative result.

$$\boldsymbol{Net\ Benefit} = \frac{\boldsymbol{True\ Positives}}{\boldsymbol{N}} - \frac{\boldsymbol{False\ Positives}}{\boldsymbol{N}}\ \boldsymbol{x}\ \frac{\boldsymbol{Threshold\ Probabilty}}{\boldsymbol{1 - Threshold\ Probability}}$$

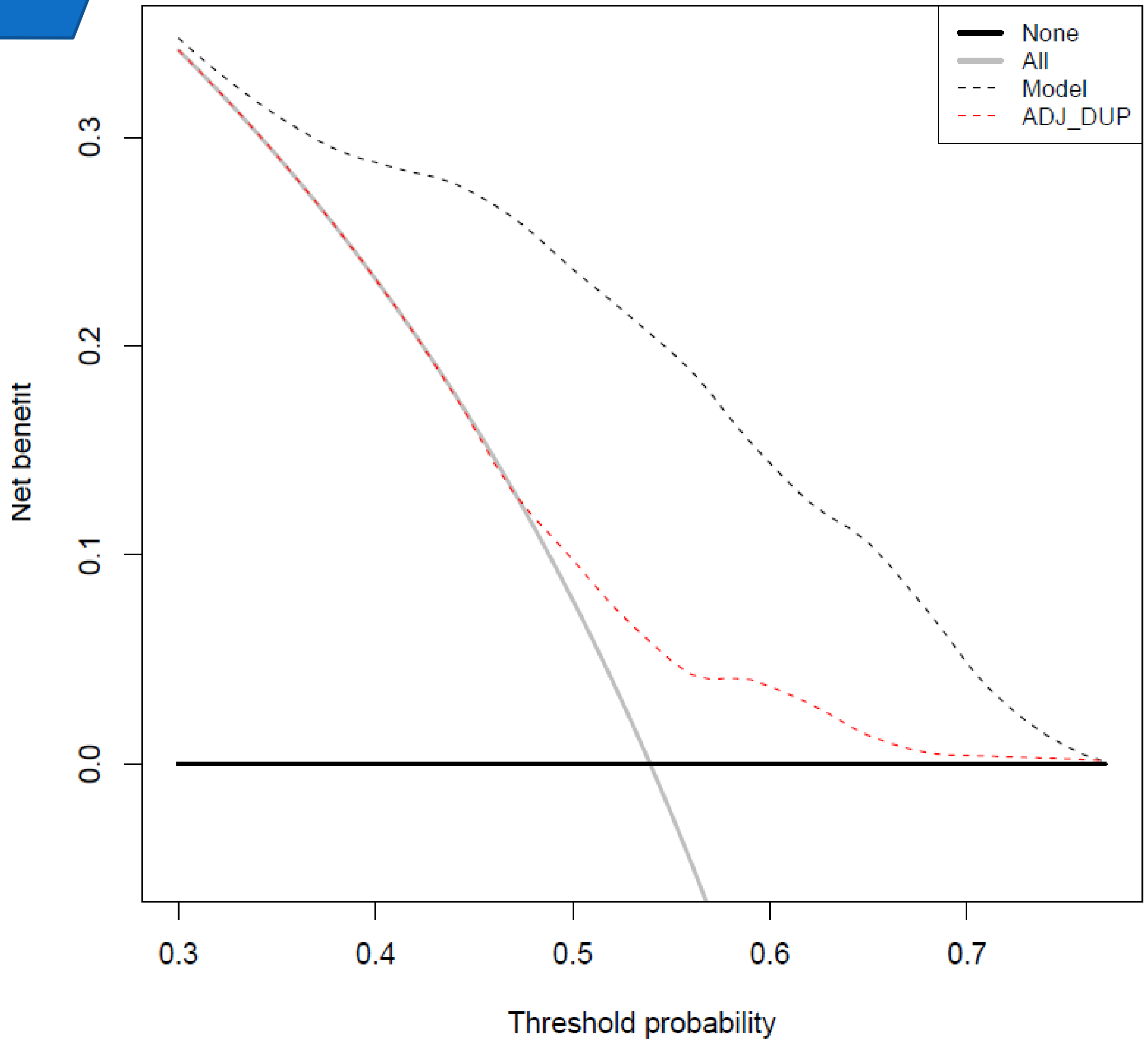
Plot net benefit on the y axis against p_t on the x axis.

Repeat steps assuming all patients are positive

Draw a straight line parallel to the x-axis at y=0 representing the net benefit associated with the strategy of assuming that all patients are negative

Decision Curve Analysis

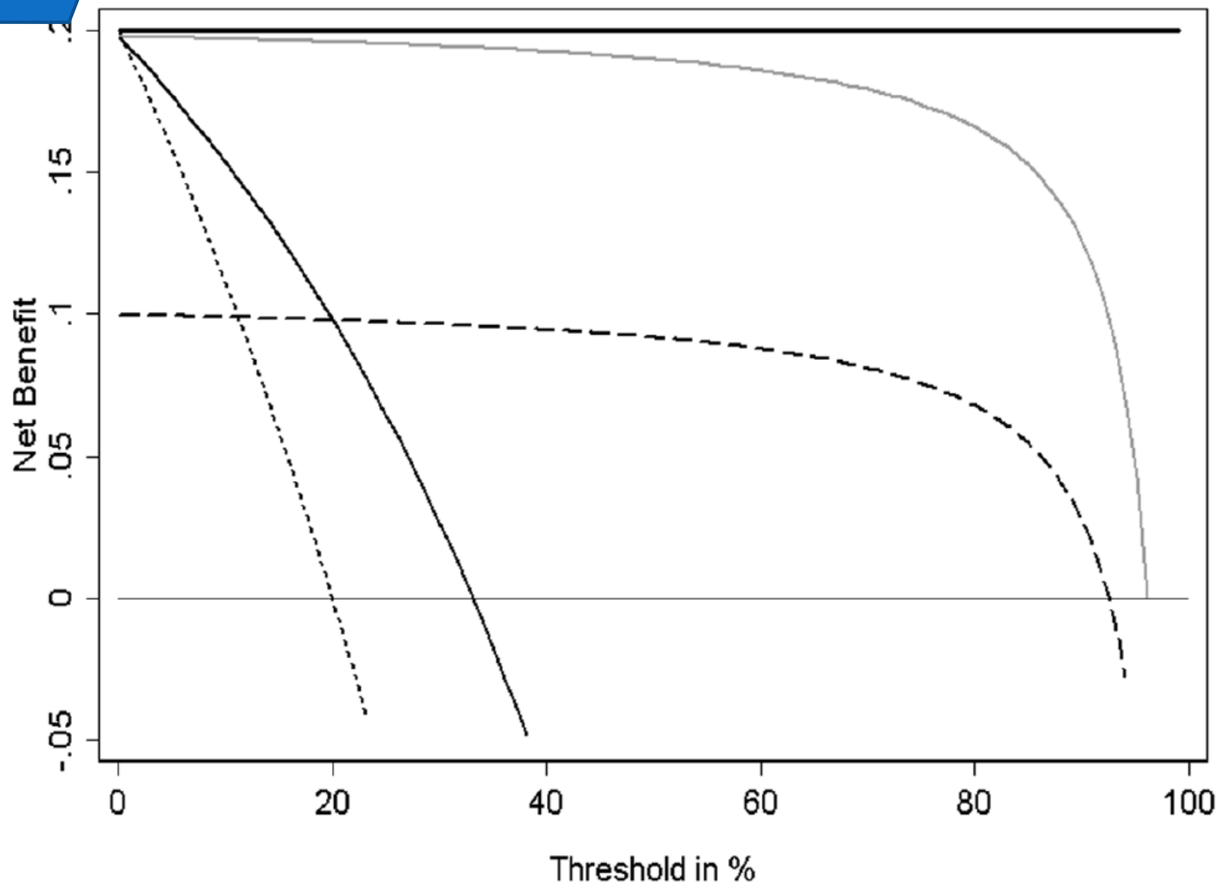
- Compare model with treating all positive, treating none or treating based on the duration of untreated psychosis alone
- Treatment probability threshold: 40-60%



Leighton et al. 2021

Theoretical Examples

- Disease incidence is 20%
- Sensitivity vs Specificity across threshold probabilities

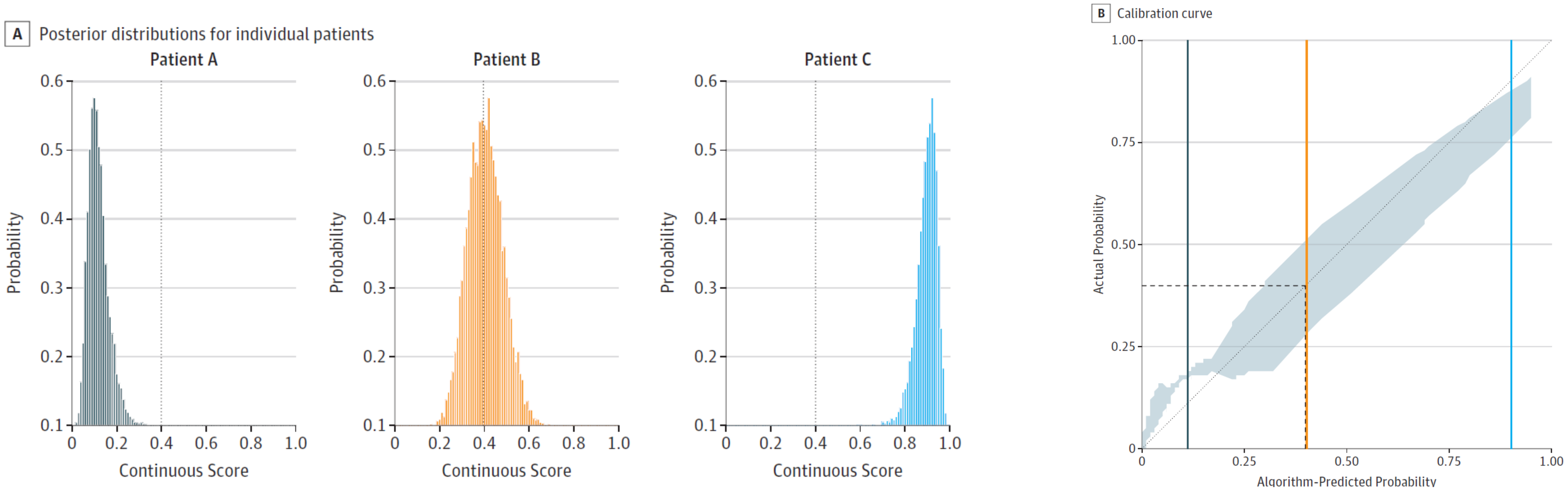


$$Net\ Benefit$$
$$= \text{sensitivity}\ x\ prevalence\ - (1 - specificity) x (1 - prevalance)\ x\ \frac{Threshold\ Probabilty}{1 - Threshold\ Probability}$$

Vickers et al. 'Decision curve analysis: a novel method for evaluating prediction models', 2006

Uncertainty

- Frequentist approach will provide confidence intervals
- Bayesian frameworks offer a principled way to take into account model uncertainty

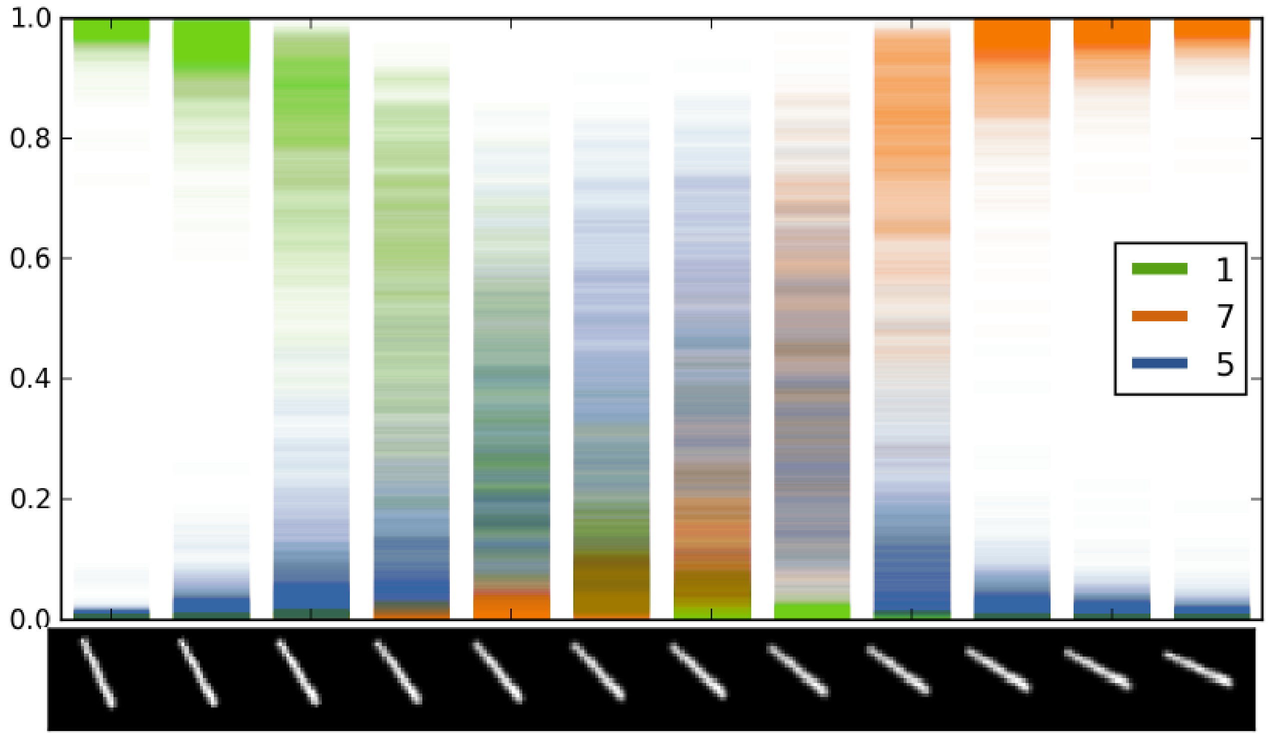


Dan W. Joyce et al. ‘Decision curve analysis: a novel method for evaluating prediction models’, 2006

Uncertainty in DNNs

- Dropout can be interpreted as a Bayesian approximation
- Representing uncertainty in DNNs without compromising computational complexity or test accuracy
- Monte Carlo estimates (MC dropout)
  - Sampling a set of vector of realisations from the Bernoulli distribution $\{\mathbf{W}_1^t, \ldots, \mathbf{W}_L^t\}_{t=1}^T$
  - Approximate variational distribution and estimate uncertainty

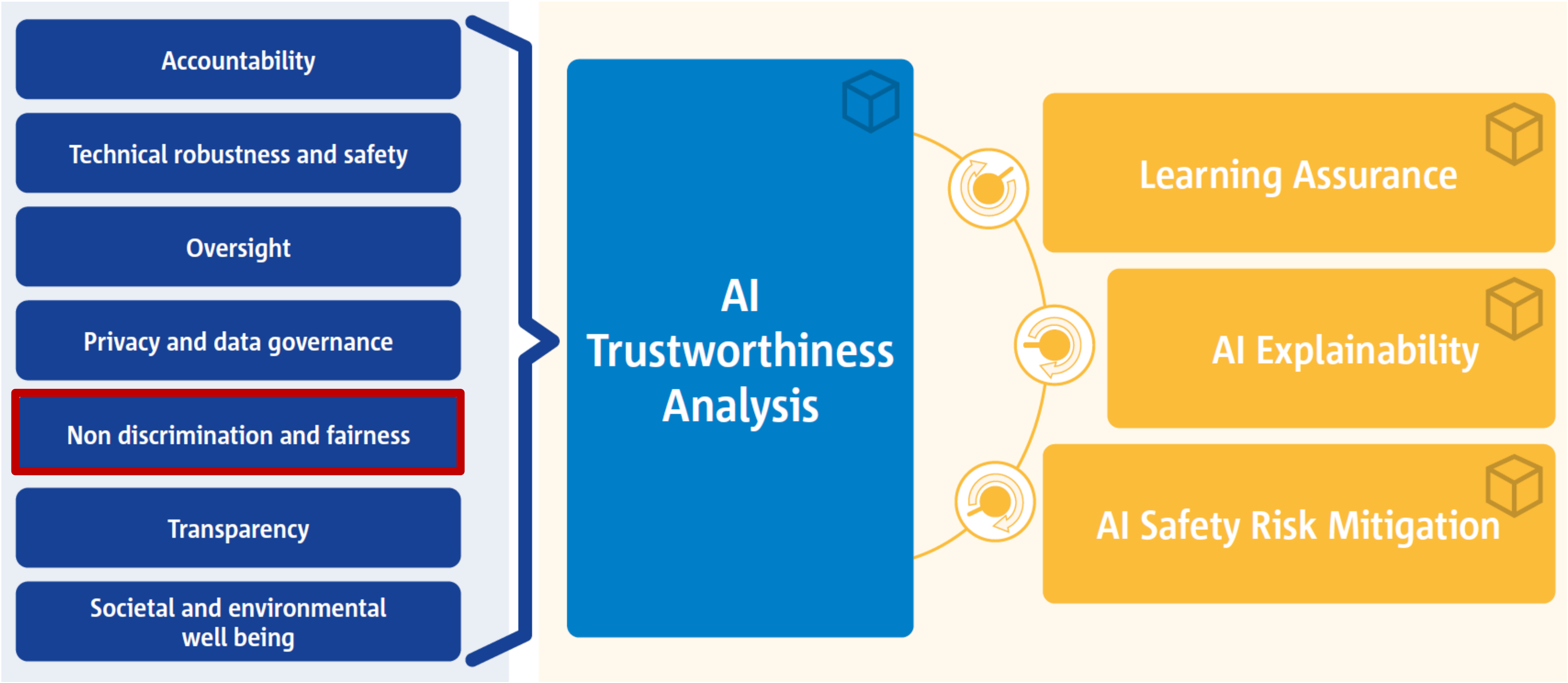
$$\mathbb{E}_{q(\mathbf{y}^*|\mathbf{x}^*)}(\mathbf{y}^*) \approx \frac{1}{T}\sum_{t=1}^{T} \widehat{\mathbf{y}}^*(\mathbf{x}^*, \mathbf{W}_1^t, \ldots, \mathbf{W}_L^t)$$



- Equivalent of performing stochastic forward passes through the network and averaging the results

Gal et al. ‘Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning’, 2016

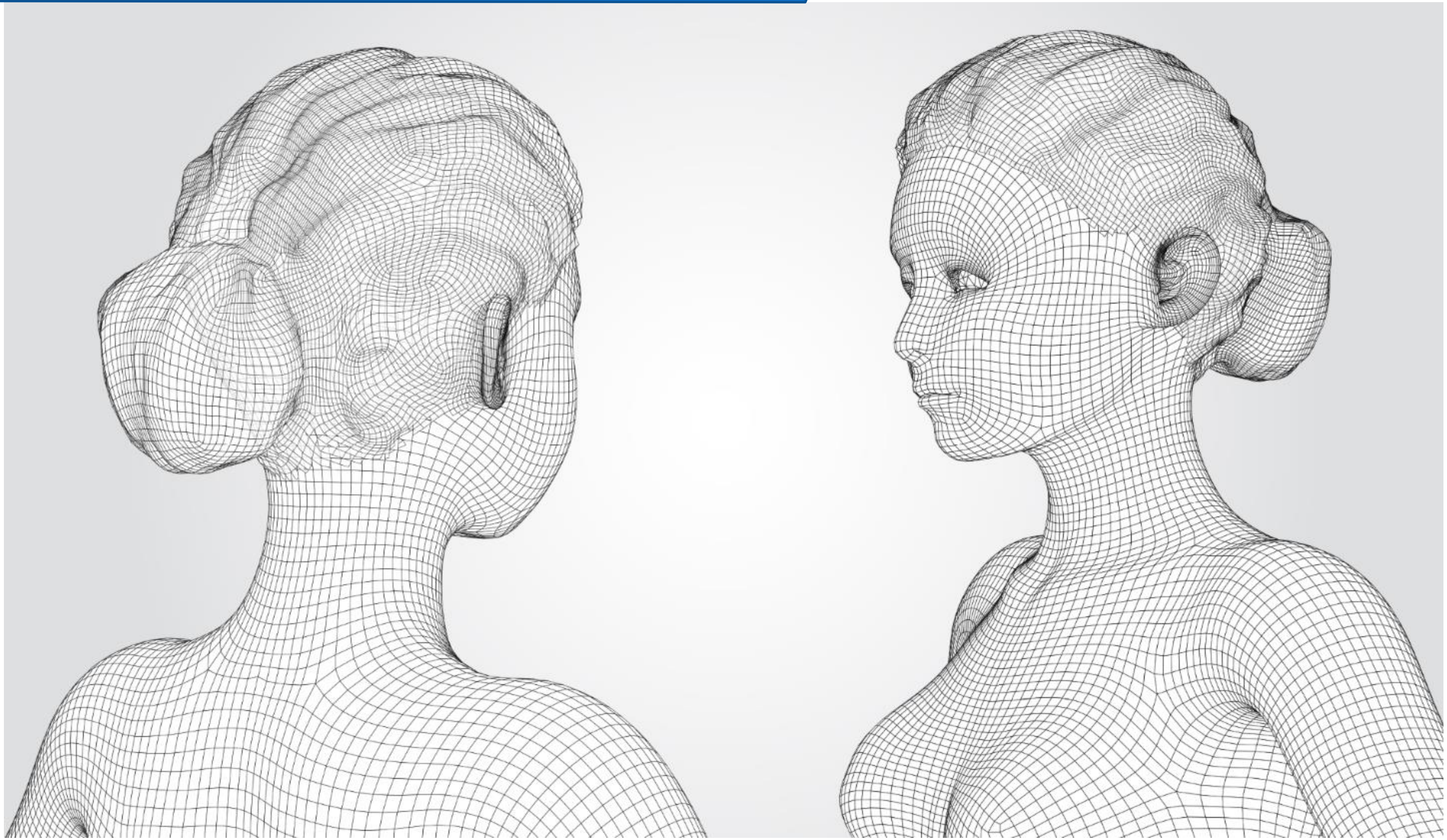
Why do we need Human-Centred CDSS?



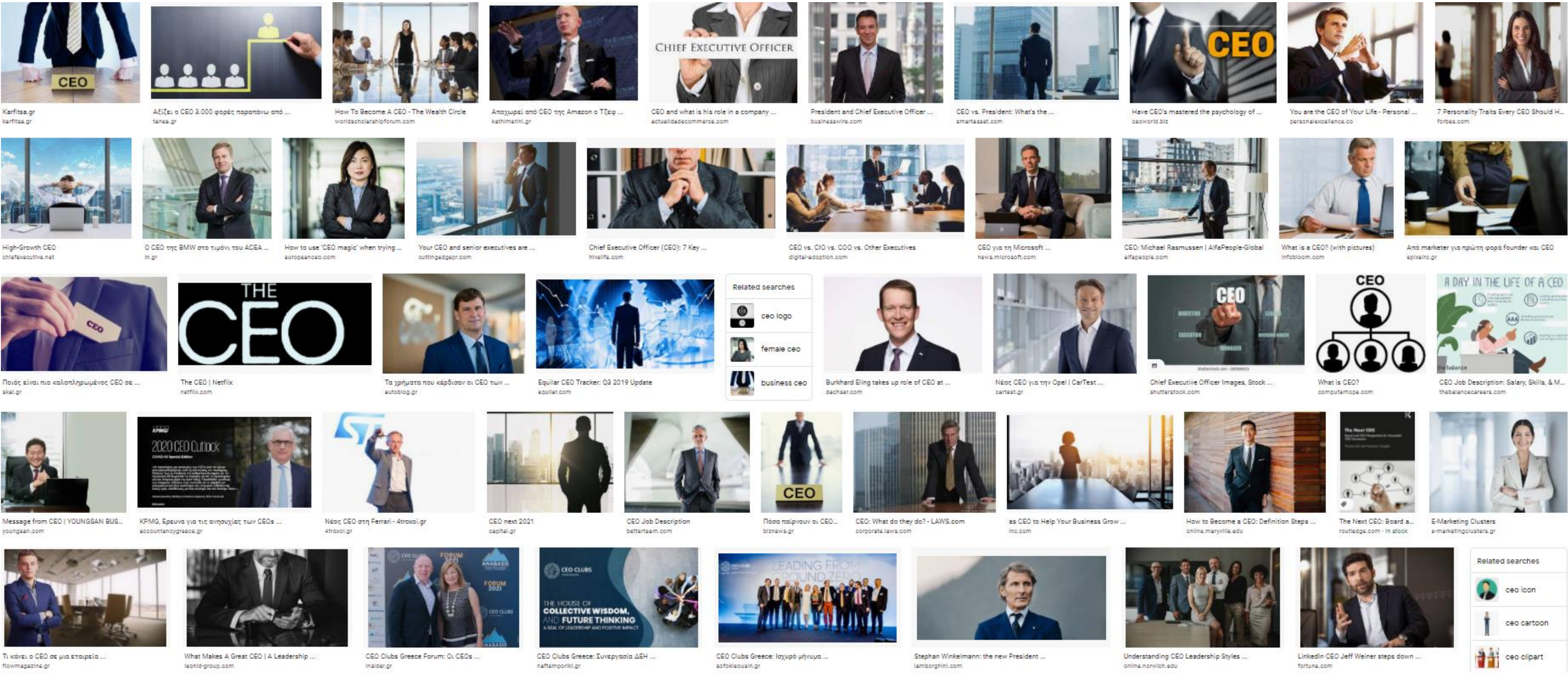
Historic Examples in Algorithmic Bias

Gender and Racial Discrimination has been captured by an algorithm and reproduced

Bias in AI Algorithms



Discrimination in Online Ads



Proxies to sensitive attributes

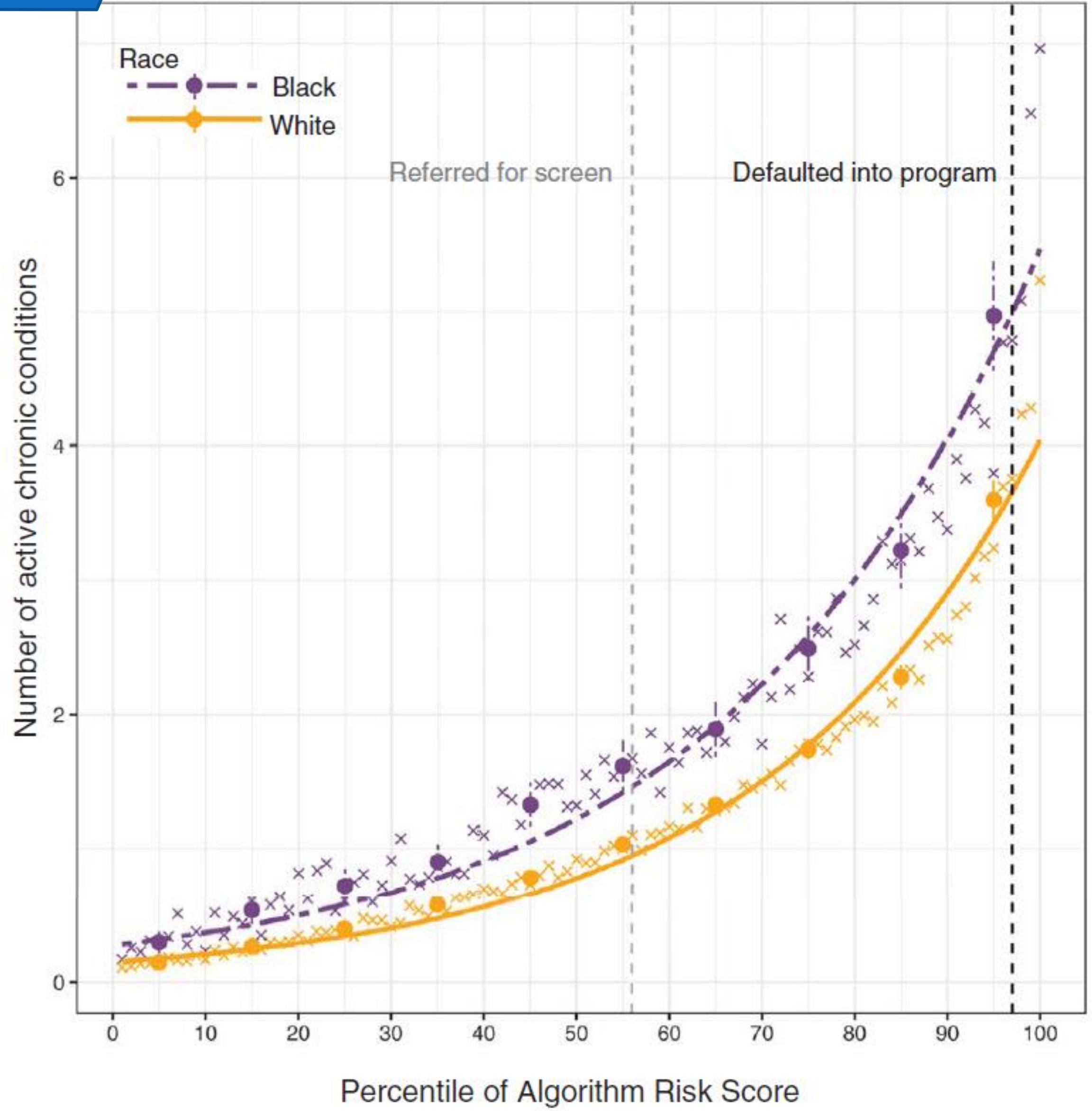
- Anti-discrimination law prohibits unfair treatment based on sensitive attributes, such as gender or race
- Implicit features may correlate with sensitive attributes
- Inherently algorithms inherit the prejudices of prior decision makers

Racial Bias in Healthcare

Obermeyer et al. ‘Dissecting racial bias in an algorithm used to manage the health of populations’, Science, 2019.

Racial Bias in Healthcare

- Quantify bias by plotting algorithmic risk scores against multi-morbidity
- At 97th percentile of risk score blacks have 26.3% more chronic illness than whites
- Significant evidence of disparities that favor white people



Obermeyer et al. 'Dissecting racial bias in an algorithm used to manage the health of populations', Science, 2019.

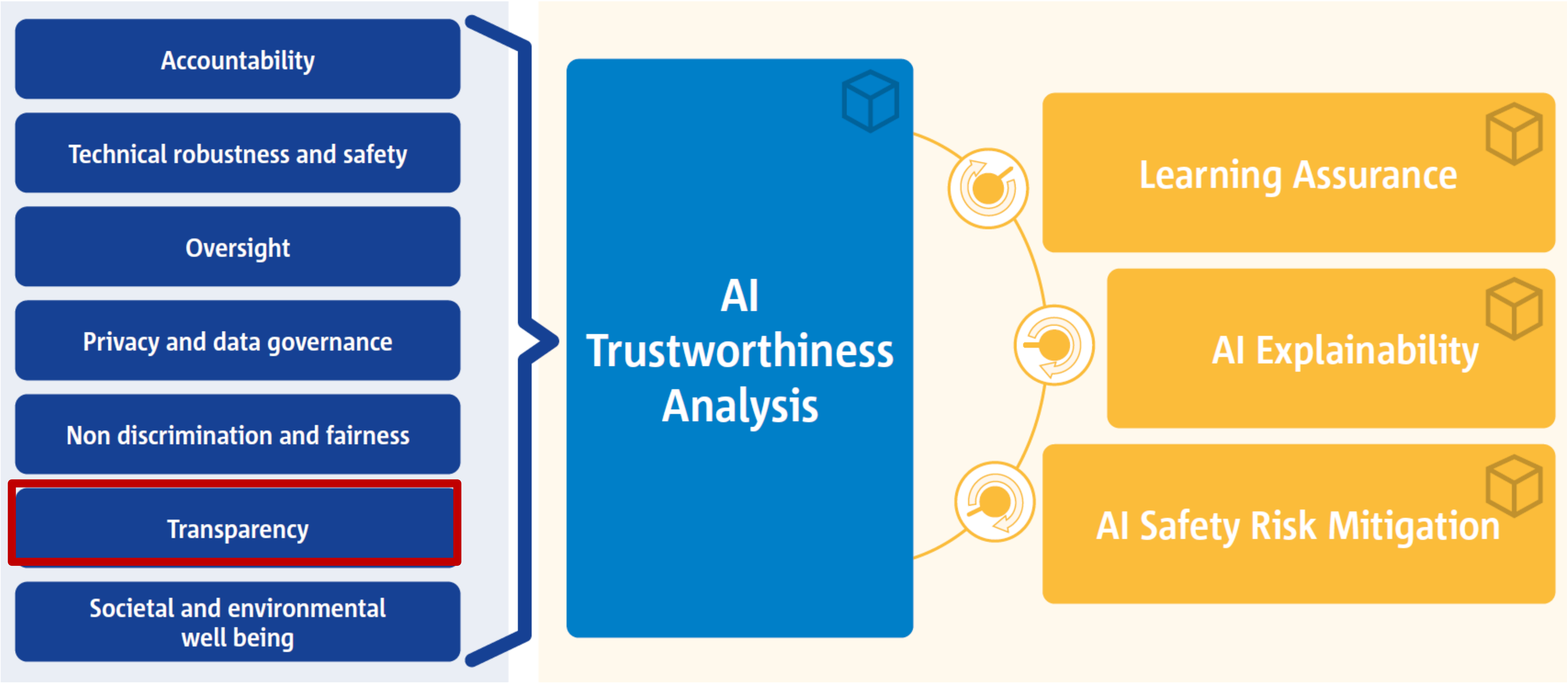
Guarantees Against Discriminatory Bias

- ***Calibration within groups:*** Calibration of algorithmic bias (statistical parity)

$$E[Y|R,W] = E[Y|R,B]$$

- ***Balance for the negative class:*** The average score received by people that are positive with relation to the outcome Y, should be the same in each group
- ***Balance for the positive class:*** The average score received by people that are negative with relation to the outcome Y, should be the same in each group

Why do we need Human-Centred AI?



Importance of ‘Explainability’

- Explainability is required to ensure impartial decision-making process
  - Detect biases, ensure fairness
- Explainability ensures that only meaningful variables infer the output
  - Explain the decision-making process
  - Fundamental human right to know why
- Explainability ensures robustness of the results

Target Audience of Explainability

Target Audience

Who? Clinicians
Why? Trust the model, gain scientific knowledge

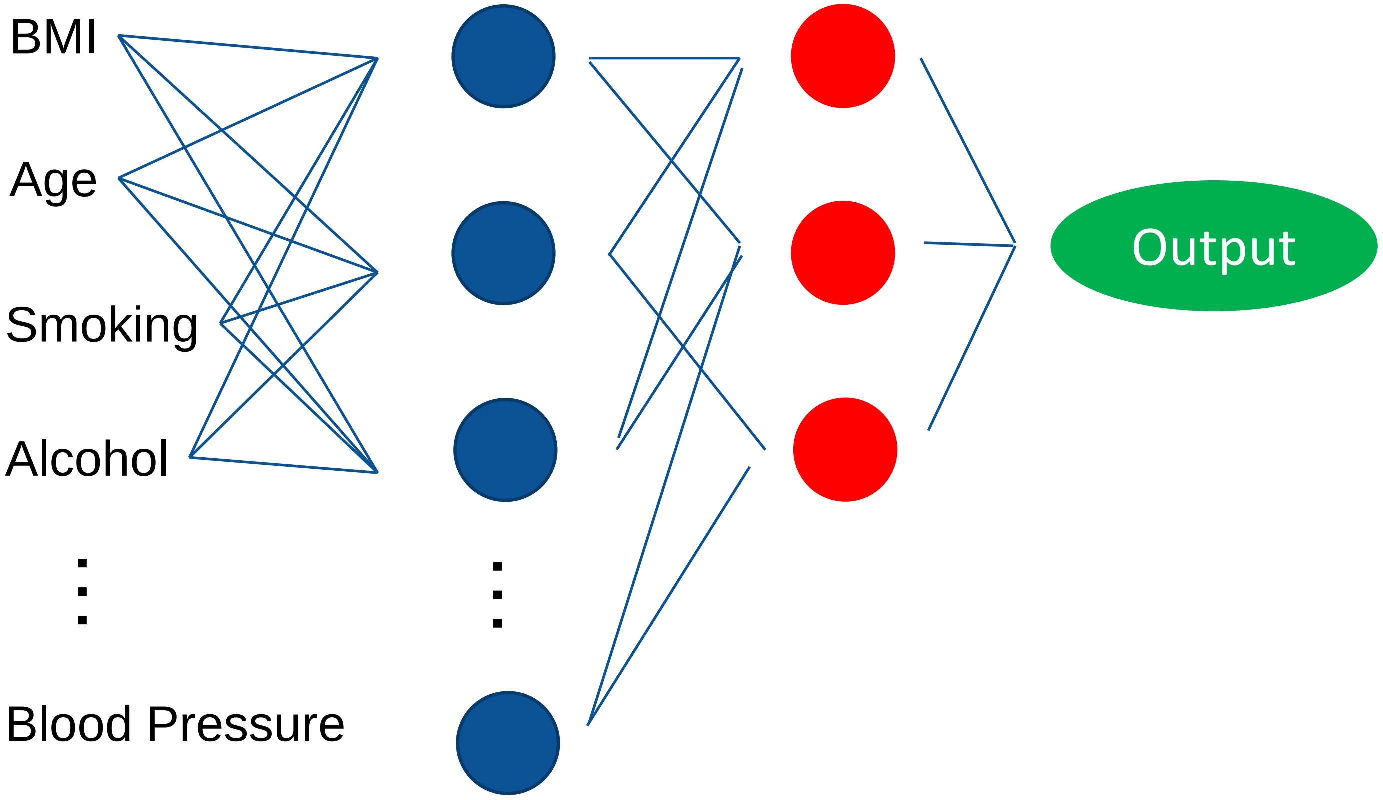
Who? Patients affected by model decisions
Why? Understand situations, verify fair decisions

Who? Regular entities/agencies
Why? Certify model compliance with the legislation

Who? Data Scientists, developers
Why? Improve product efficiency, research etc

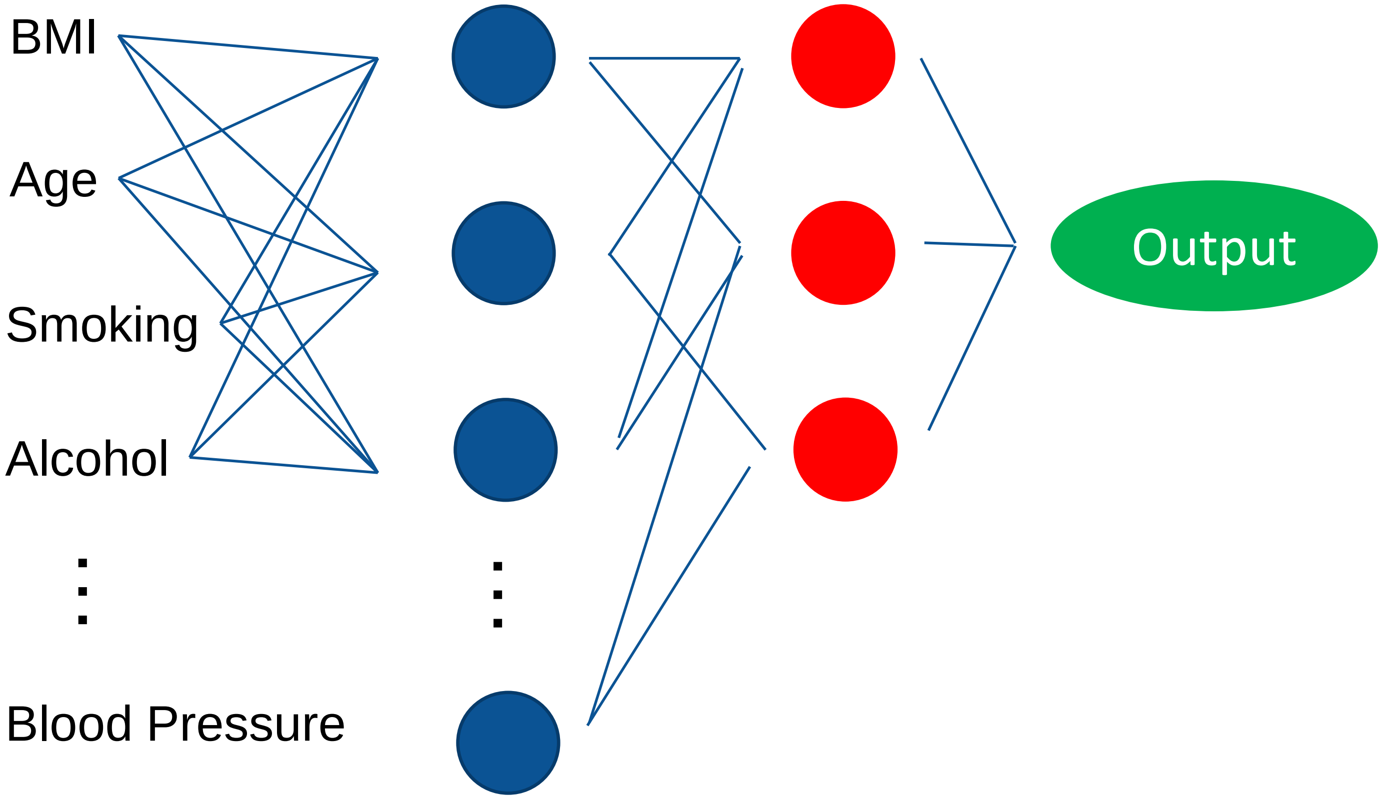
Who? Managers and executive board members
Why? Assess regulatory compliance, understand applications

Explainable Model



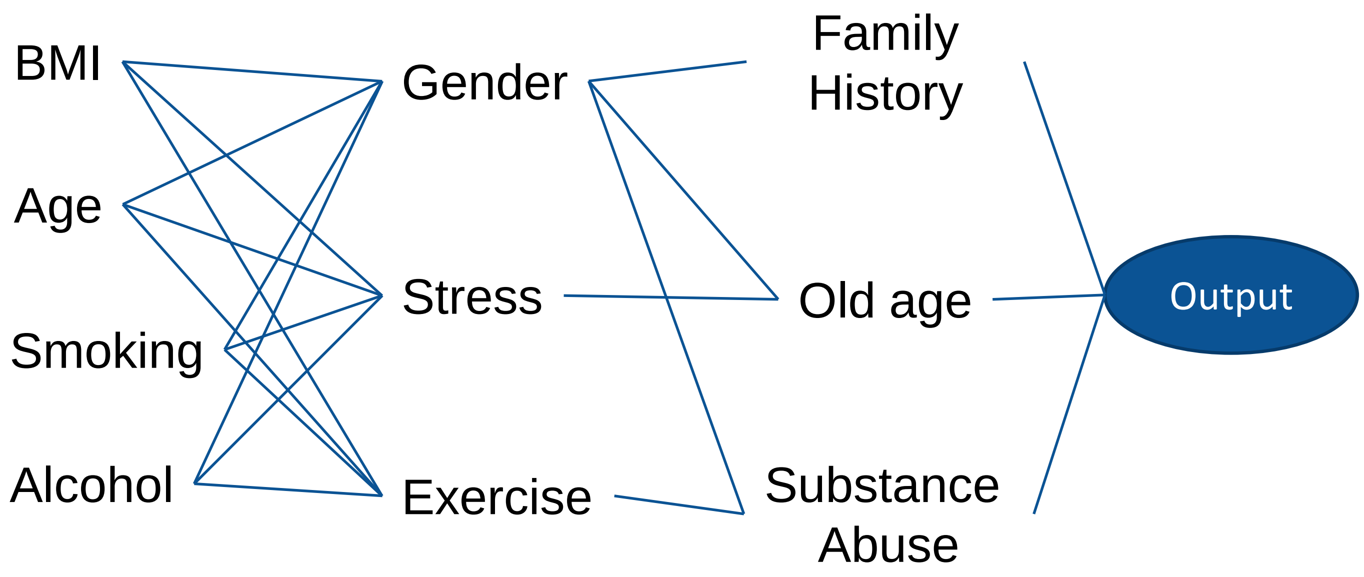
- Do we understand why the model came to this output?
- Do we know the conditions/cases that the model is successful and when it is not?
- Do we know the factors behind this output?

Explainable Model - Factors



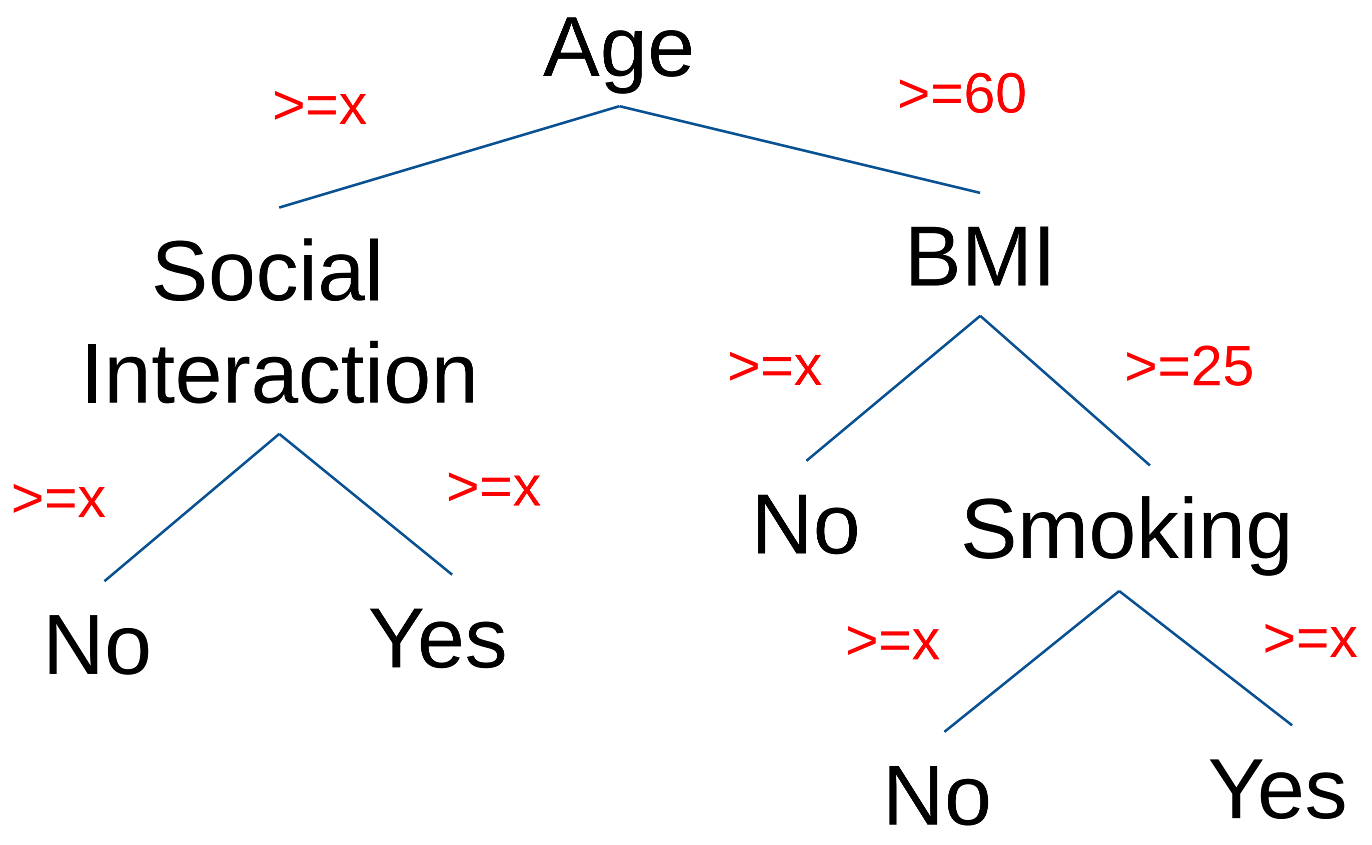
- Age is the most important factor in predicting heart failure.
- Large BMI also increases the probability of a heart attack episode
- History of smoking also increase the probability
- High blood pressure is also associated with heart failure

Explainable Model – Representation Learning



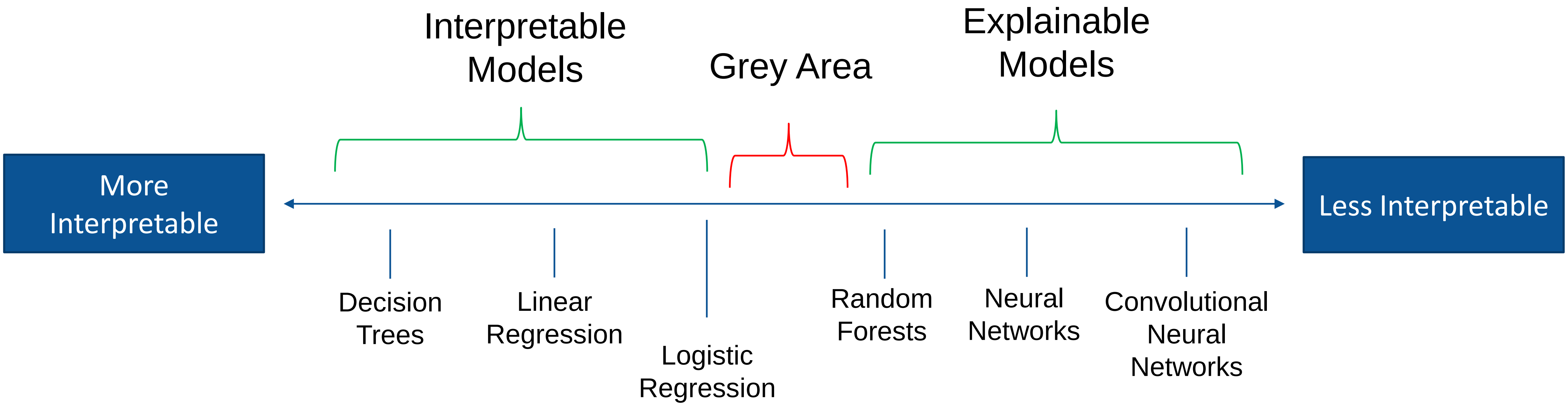
- Knowledge of the what each node represents
- Latent factors that affect the decision process
- How important each node is to the model's performance

Interpretable Models – Decision Trees

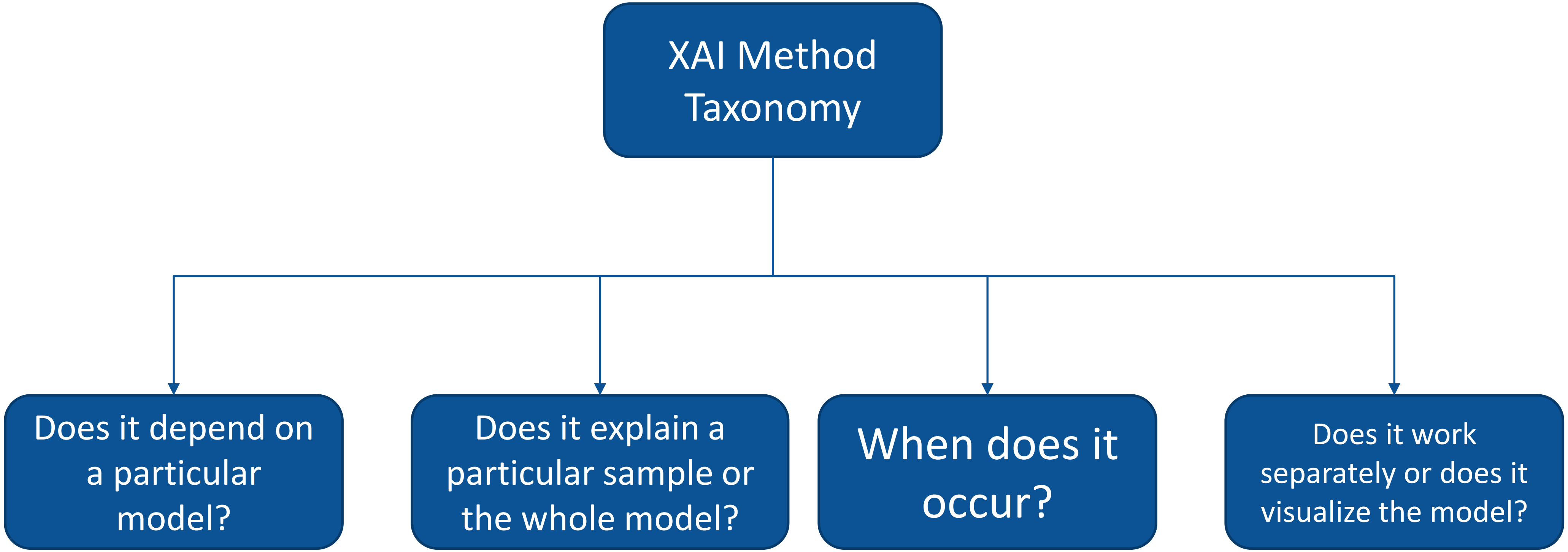


- It is clearly what each node represents
- Easy to visualize and overview the whole decision operation
- Easy to explain to non-specialists
- Results can be tracked and associated with the output of each node

Interpretable vs Explainable Models



Overview



- Model Agnostic
- Model Specific

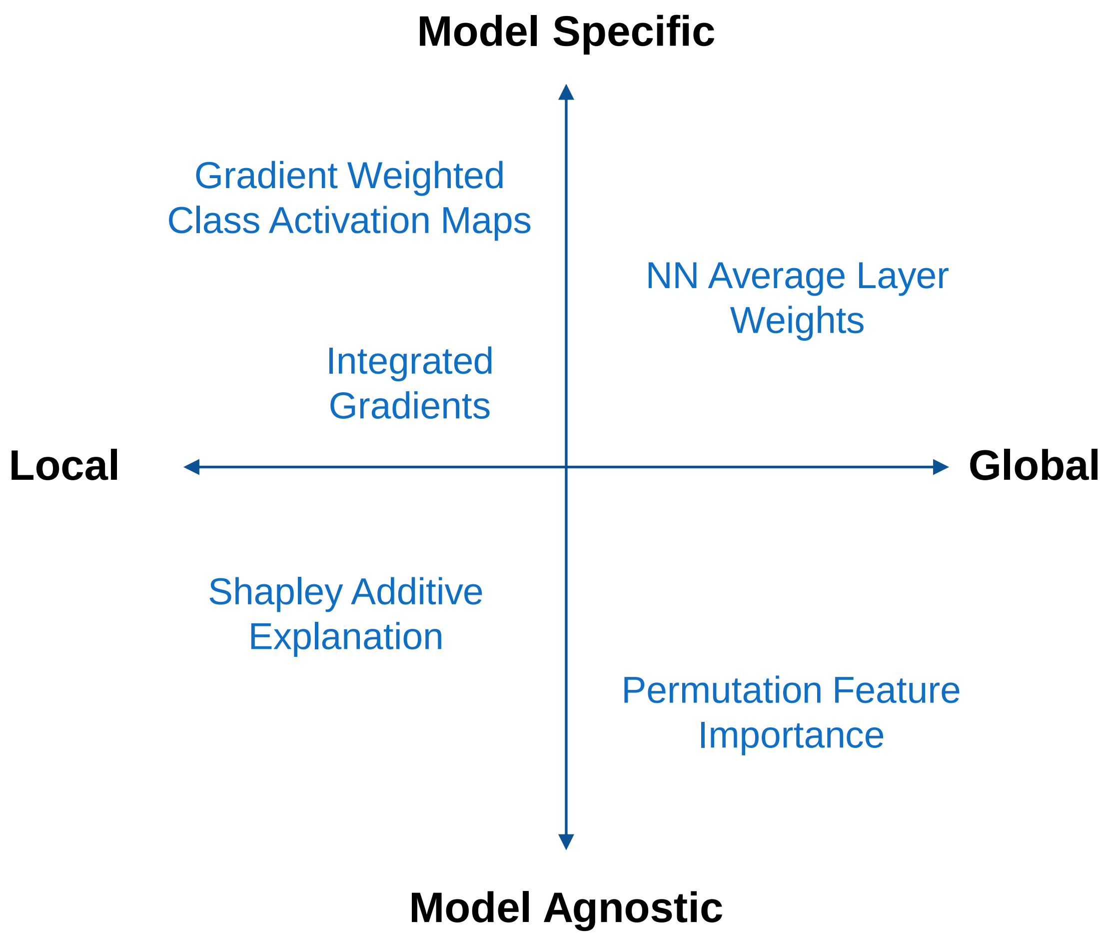
- Global
- Local

- Ad-Hoc
- Post-Hoc

- Surrogate
- Visualization

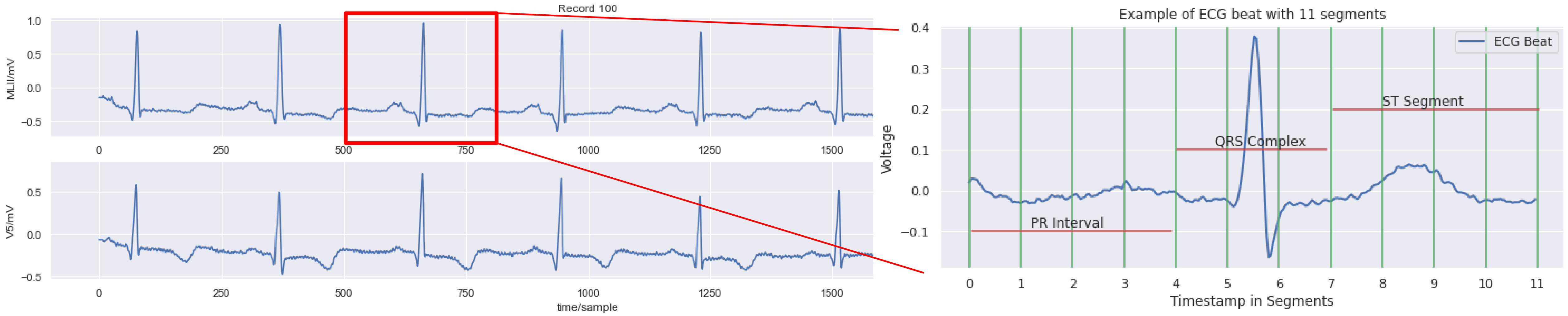
Model Specific Explanations

- Model-specific interpretation tools are limited to specific models.
- Regression weights in a linear model is a model-specific explanation
- Methods based on the activations of deep neural network layers are model-specific

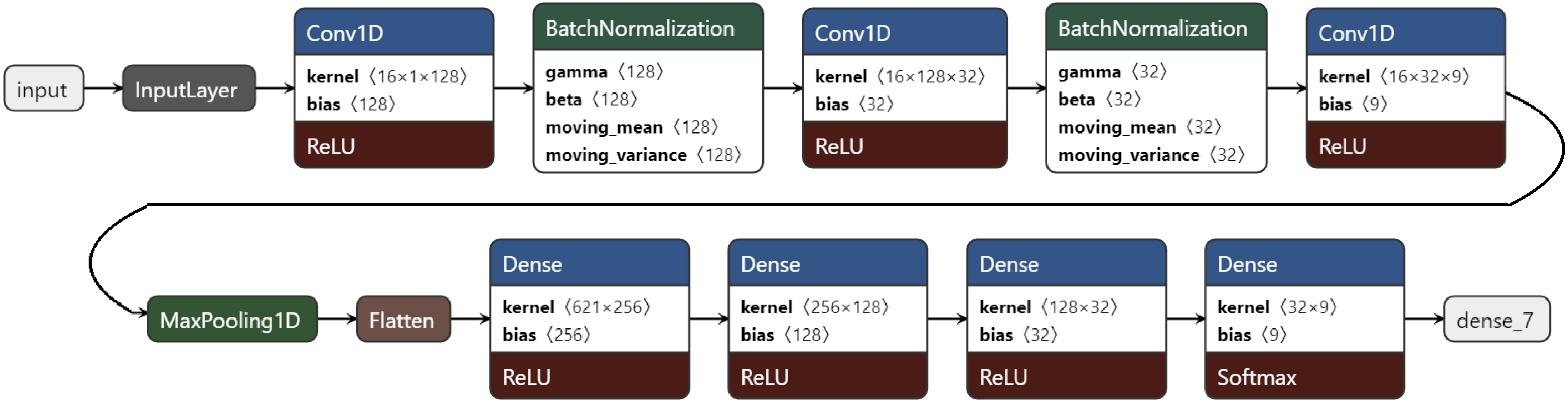


ECG Segmentation

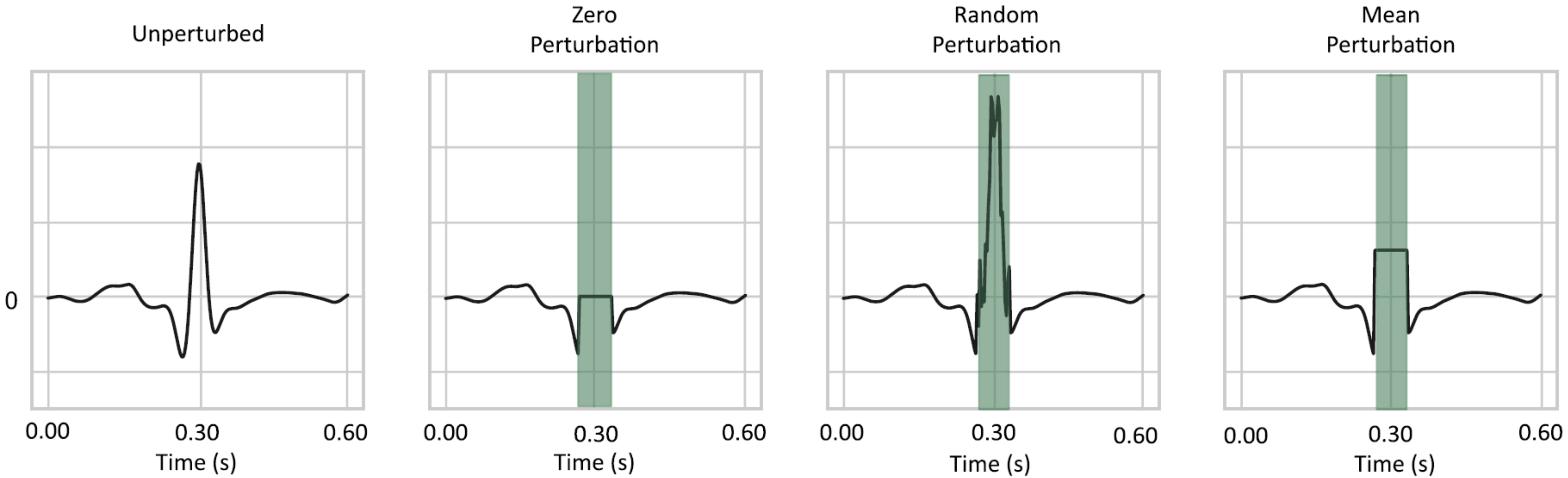
- Segments 1-4 cover the PR interval.
- Segments 5-7 cover the QRS complex
- Segments 8-11 cover the ST segment.
- We expected to see the model focusing on important morphological features of the ECG beat, such as the PR interval, the QRS complex, and the ST segment.



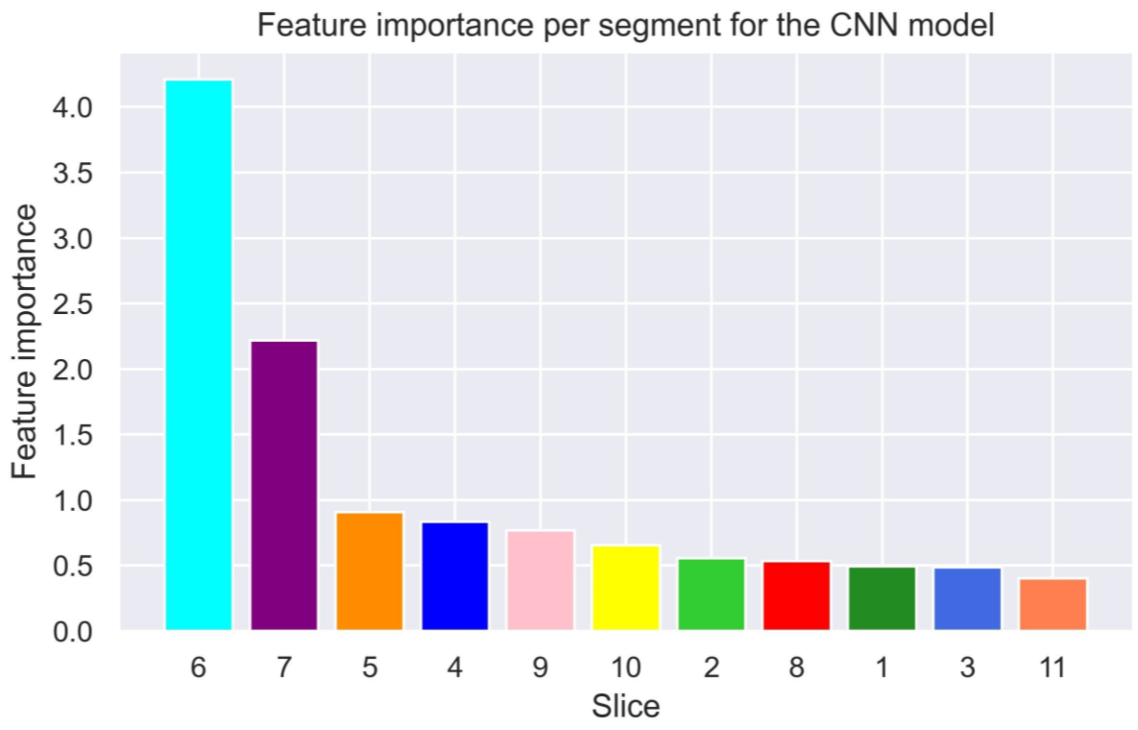
PFI on CNN


input
InputLayer
Conv1D
kernel ⟨16×1×128⟩
bias ⟨128⟩
ReLU
BatchNormalization
gamma ⟨128⟩
beta ⟨128⟩
moving_mean ⟨128⟩
moving_variance ⟨128⟩
Conv1D
kernel ⟨16×128×32⟩
bias ⟨32⟩
ReLU
BatchNormalization
gamma ⟨32⟩
beta ⟨32⟩
moving_mean ⟨32⟩
moving_variance ⟨32⟩
Conv1D
kernel ⟨16×32×9⟩
bias ⟨9⟩
ReLU
MaxPooling1D
Flatten
Dense
kernel ⟨621×256⟩
bias ⟨256⟩
ReLU
Dense
kernel ⟨256×128⟩
bias ⟨128⟩
ReLU
Dense
kernel ⟨128×32⟩
bias ⟨32⟩
ReLU
Dense
kernel ⟨32×9⟩
bias ⟨9⟩
Softmax
dense_7

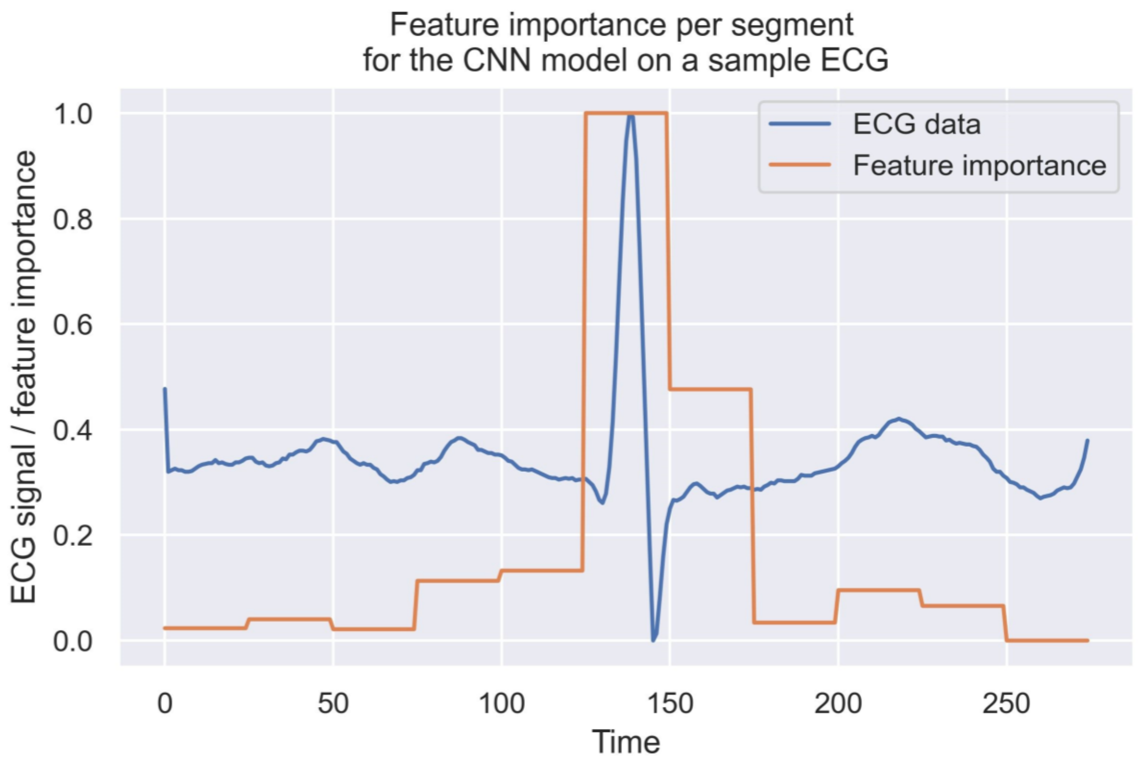
PFI for ECG Classification



PFI on CNN



Feature importance per segment for the CNN model
Feature importance
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
6
7
5
4
9
10
2
8
1
3
11
Slice

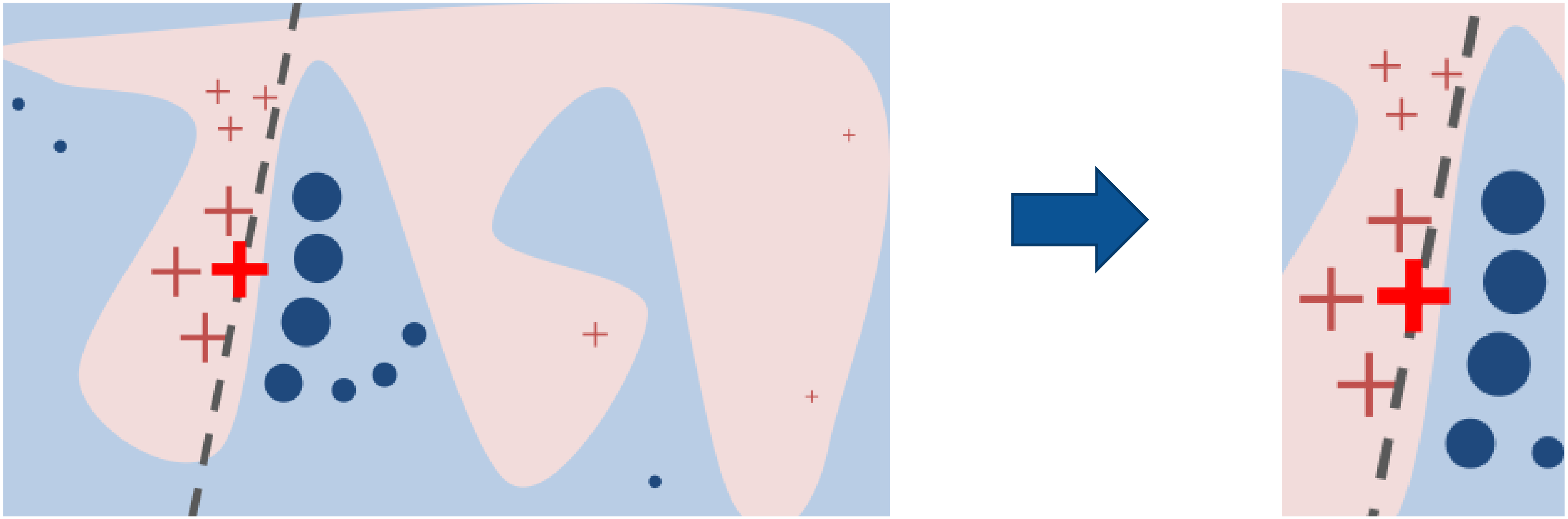


Feature importance per segment
for the CNN model on a sample ECG
ECG data
Feature importance
ECG signal / feature importance
1.0
0.8
0.6
0.4
0.2
0.0
0
50
100
150
200
250
Time

LIME

Local Interpretable Model-agnostic Explanations:

- Locally faithful explanations
- Based on a surrogate (ie. locally linear) model

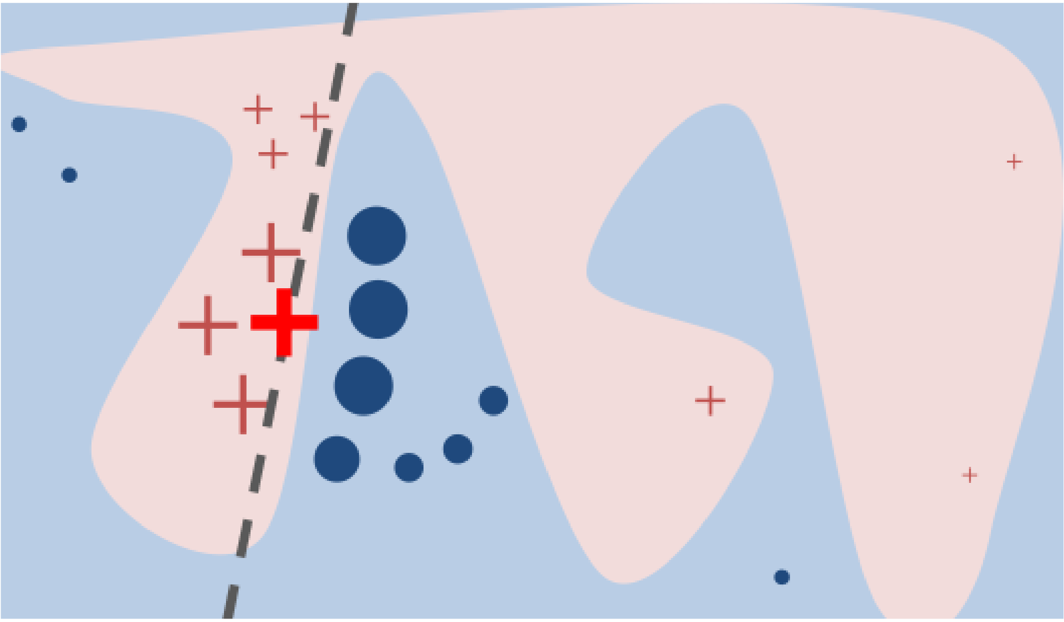


Ribeiro et al. ‘Model-Agnostic Interpretability of Machine Learning’, 2016

LIME - Formulation

Local Interpretable Model-agnostic Explanations:

- Locally faithful explanations
- Based on a surrogate (ie. locally linear) model



Model to explain

$$f : \mathbb{R}^d \rightarrow \mathbb{R}$$

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \; \mathcal{L}(f, g, \Pi_x) + \Omega(g)$$

Explanation

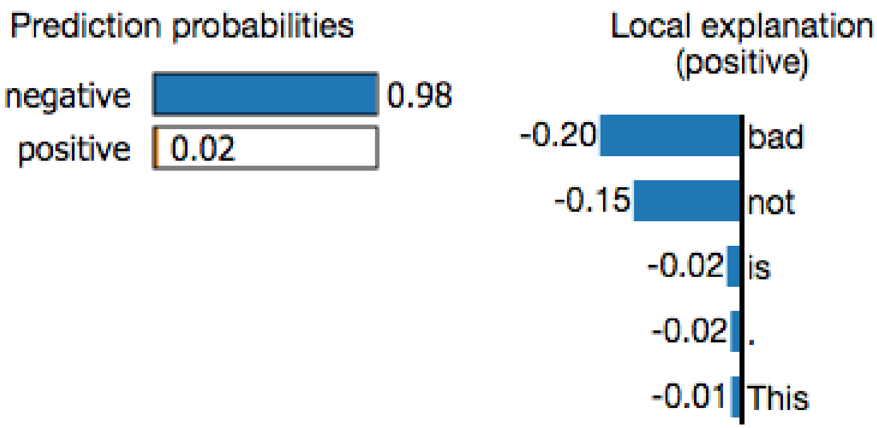
Proximity Measure

Ribeiro et al. ‘Model-Agnostic Interpretability of Machine Learning’, 2016

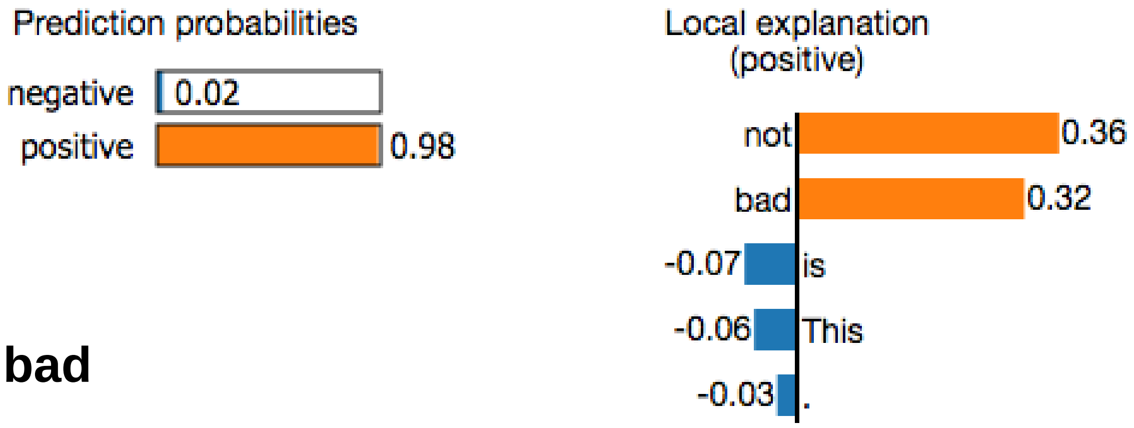
LIME – Explanations

Local Interpretable Model-agnostic Explanations:

- Allow accurate explanations while it retains model flexibility
- The explanation should be accessible even to the non experts
- Small switching costs with relation to changes to the model



This is not too bad



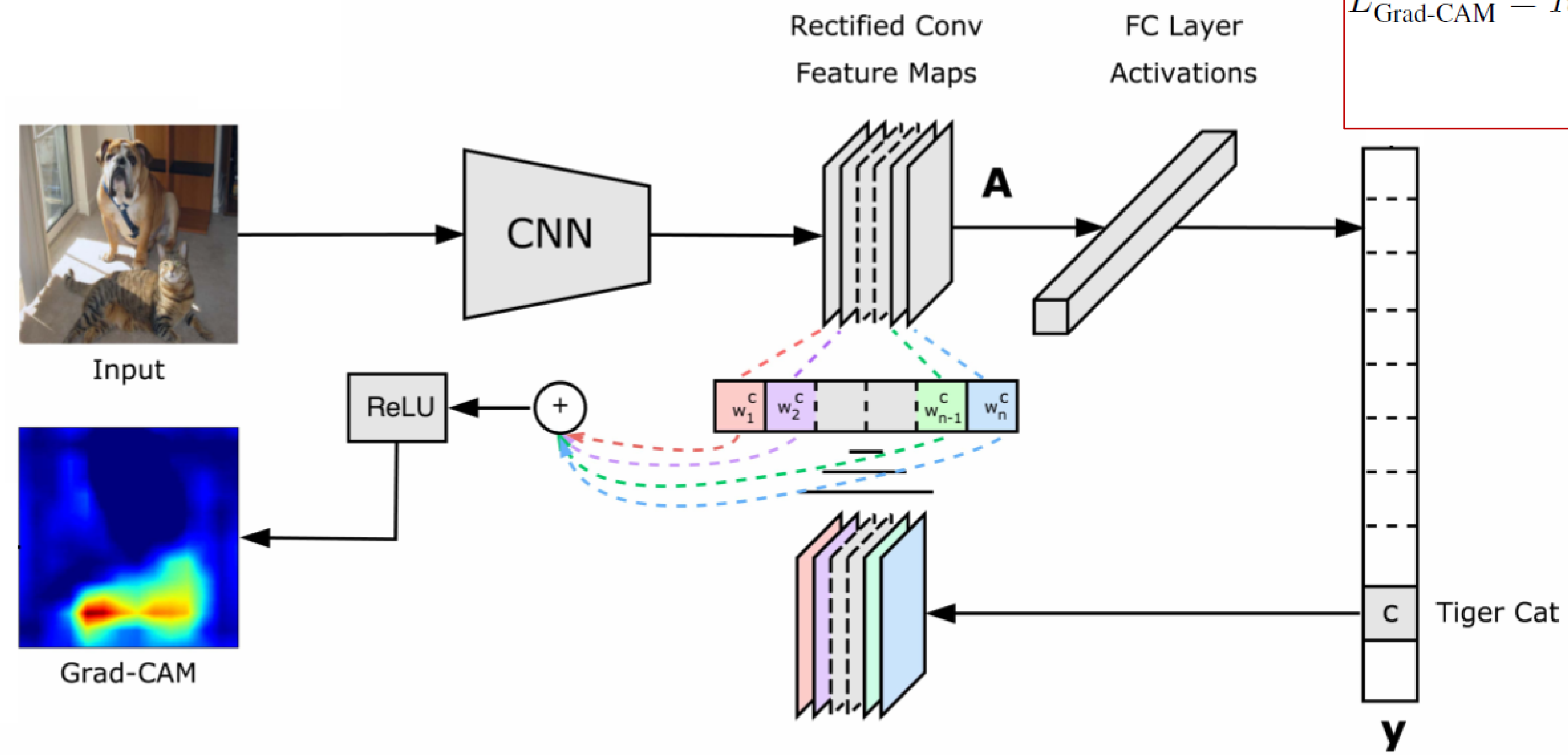
LIME Explanation of a logistic regression

LIME Explanation of an LSTM Model

Gradient Weighted Class Activation Maps

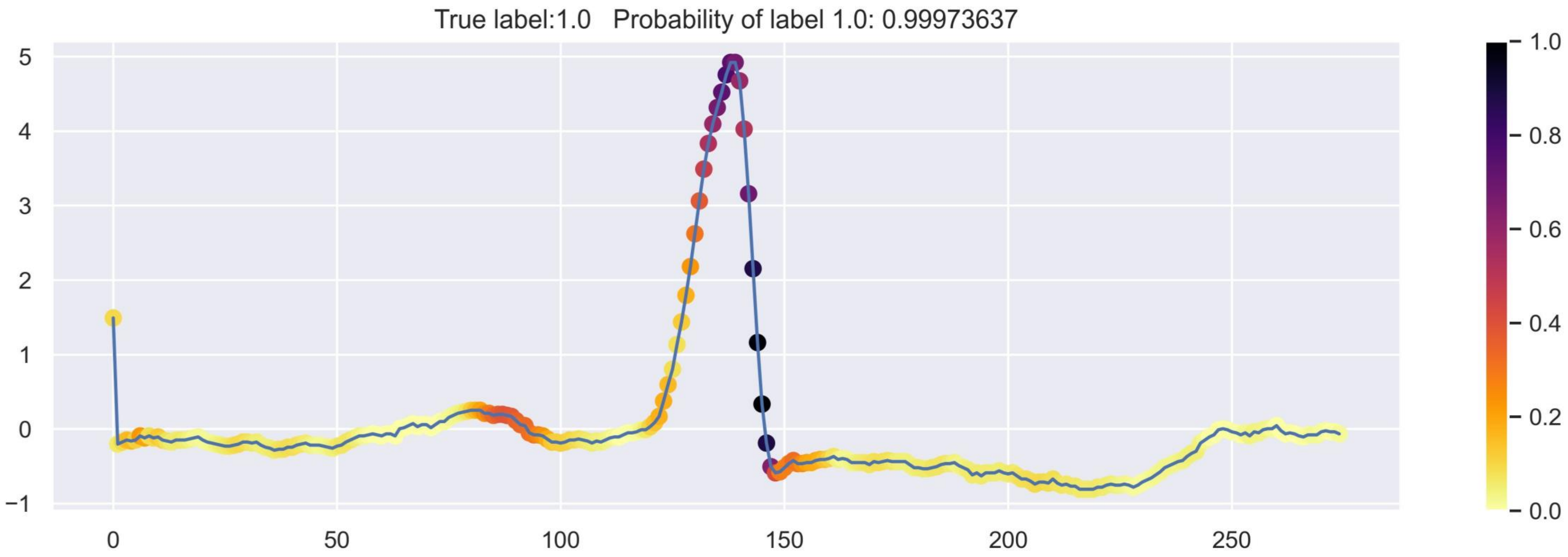
$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$

$$\boxed{L^c_{\text{Grad-CAM}} = ReLU\underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{linear combination}}}$$



Selvaraju et al. ‘Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization’, IJCV, 2021

Grad-CAM Example

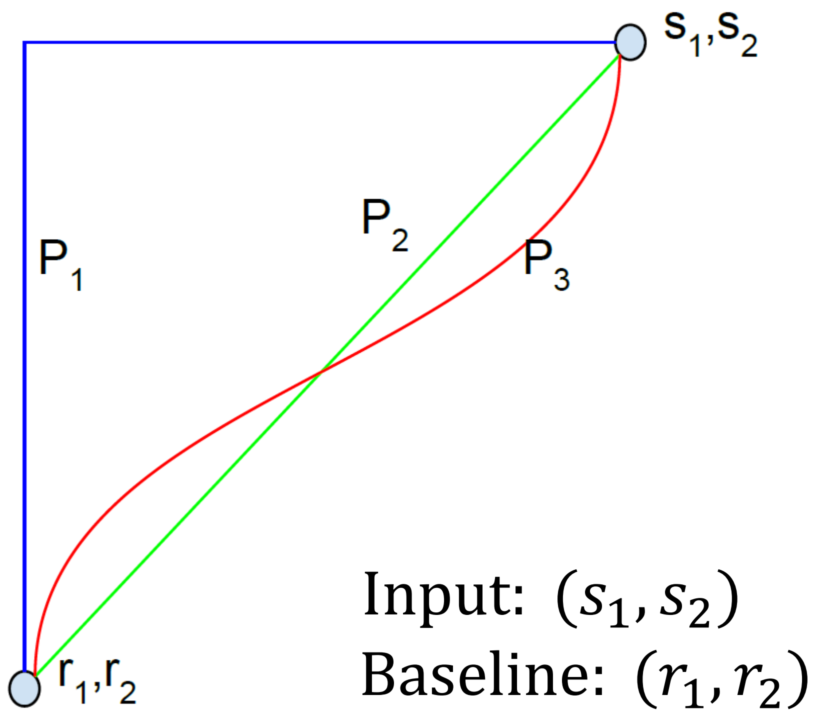


Attributions’ Desirable Properties

- Sensitivity
- Implementation invariance
- Completeness
- Linearity
- Symmetry preserving

Integrated Gradients

- Consider the straight-line path between baseline and input
- Integrate the gradients along this path



Sundararajan et al. ‘Axiomatic Attribution for Deep Networks’, ICML, 2017

Integrated Gradients

1. Create α - array containing m values evenly space between 0 and input.
2. Generate signal interpolations using α.
3. Calculate gradients predictions w.r.t. Input features.
4. Use Riemman’s sum to average the gradients.
5. Re-scale gradient to generate attributions.

$$IntegratedGradients_i^{approx}(x) = \overbrace{(x_i - x_i')}^{5} \times \overbrace{\sum_{k=1}^{m}}^{4} \overbrace{\frac{\partial F \overbrace{\left(x' + \overbrace{\frac{k}{m}}^{1} \times (x - x')\right)}^{2}}{\partial x_i}}^{3} \times \overbrace{\frac{1}{m}}^{4}$$

Sundararajan et al. ‘Axiomatic Attribution for Deep Networks’, ICML, 2017

Expected Gradients

1. Draw samples from training data.
2. Calculate attributions for every sample across all references
3. Average the attributions for samples over all references.

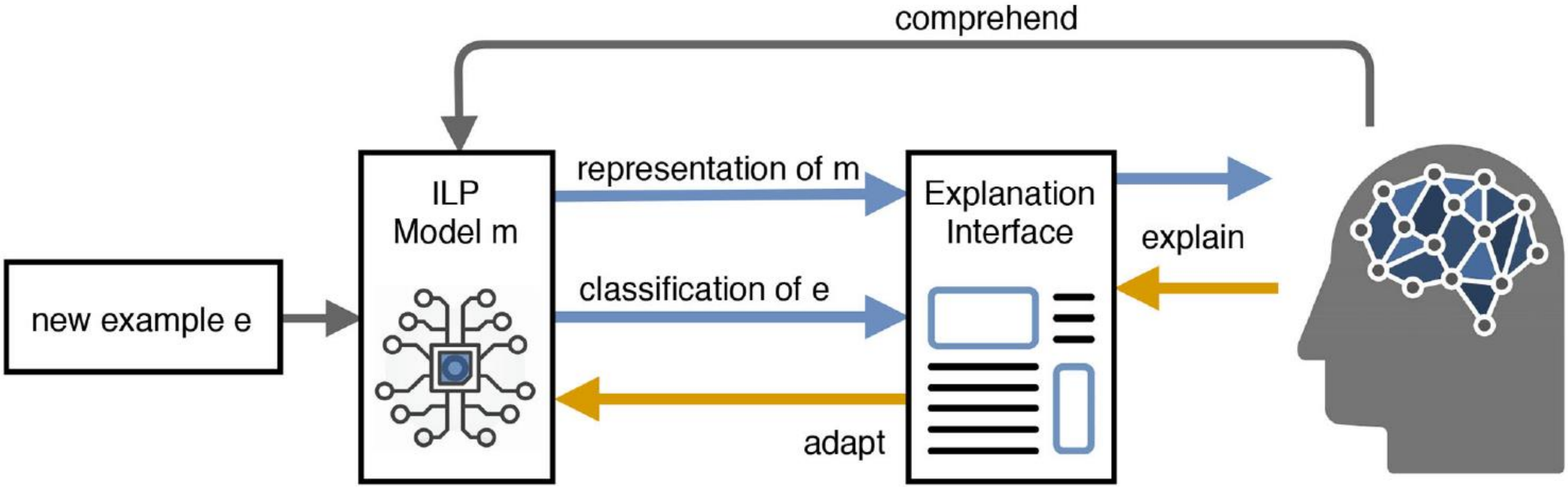
$$ExpectedGradients_i(x) = \overbrace{\underset{\underbrace{x' \sim D}_{1},\, \alpha \sim U(0,1)}{\mathbb{E}}}^{3} \overbrace{\left[(x_i - x_i') \times \frac{\delta f(x' + \alpha \times (x_i - x_i'))}{\delta x'} \right]}^{2}$$

Attributions can be a flexible framework to encode priors:

$$\theta = \mathrm{argmin}_{\theta} \mathcal{L}(\theta; X, y) + \lambda \Omega(\Phi(\theta, X))$$

Erion et al. ‘Improving performance of deep learning models with axiomatic attribution priors and expected gradients’, Nature Machine Intelligence, 2020

Human-Centred ML/AI



Bruckert et al. ‘The Next Generation of Medical Decision Support: A Roadmap Toward Transparent Expert Companions, Frontiers in Artificial Intelligence, 2020.

References

- Coursera course on ‘Informed Clinical Decision Making using Deep Learning Specialization’:
  https://www.coursera.org/specializations/clin-decision-deep-learning
  https://github.com/fd301/CDSS_course
- Horn et al. ‘Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review’, npj Digital Medicine, 2022.
- Bruckert et al. The Next Generation of Medical Decision Support: A Roadmap Toward Transparent Expert Companions, Frontiers in Artificial Intelligence, 2020.
- Murphy et al. ‘Artificial intelligence for good health: a scoping review of the ethics literature’, BMC Medical Ethics, 2021.
- Sutton et al. ‘An overview of clinical decision support systems: benefits, risks, and strategies for success’, NPJ Digital Medicine, 2020.

References

- Rajkomar et al. ‘Ensuring Fairness in Machine Learning to Advance Health Equity’, Annals of Internal Medicine, 2018.
- Kleinberg et al. ‘Inherent Trade-Offs in the Fair Determination of Risk Scores, Proceedings of Innovations in Theoretical Computer Science, 2017.
- Caton et al. Fairness in Machine Learning: A Survey, arXiv:2010.04053, 2020
- Wilkinson et al. ‘Time to reality check the promises of machine learning powered precision medicine’, Lancet Digital Health, 2020.

References

- Arrieta et al. ‘Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI’, Information Fusion, 2020.
- Molnar ‘Interpretable Machine Learning - A Guide for Making Black Box Models Explainable’ https://christophm.github.io/interpretable-ml-book/