MACHINE LEARNING

ENGINEERING BOOTCAMP

MING LUN **ONG**

ABOUT ME

My name is Ming Lun. I'm currently a data scientist with DSAID/MOE, and I'm currently working on education use-cases.

Previously, I've done research in applied machine learning (explainability of healthcare predictions), and in data science (a privacy-preserving method at a research institute).

I'm interested in making ML explainable and reliable to end-users.

Telegram: @mung_lin

OVERVIEW

1. Introduction to Machine Learning (10 mins)
2. Introduction to Explainable ML (5 mins)
3. Explainable ML Techniques (10 mins)
4. Technical Demonstration (15 mins)
5. Going Forward (10 mins)
6. Open Discussion – How to Productionise? (10 mins)

WHAT WE ARE GOING TO DO

- Discuss an overview of Machine Learning, Data Science, and how it relates to Artificial Intelligence
- Debunk some buzzwords, and brainstorm the productionisation of ML
- Explainability of ML and the importance of such
- Exploration of data science problems (i.e. the MNIST dataset)

FUNDAMENTALS OF MACHINE LEARNING

BUZZWORDS

Data Science

Machine Learning

Artificial Intelligence

Deep Learning

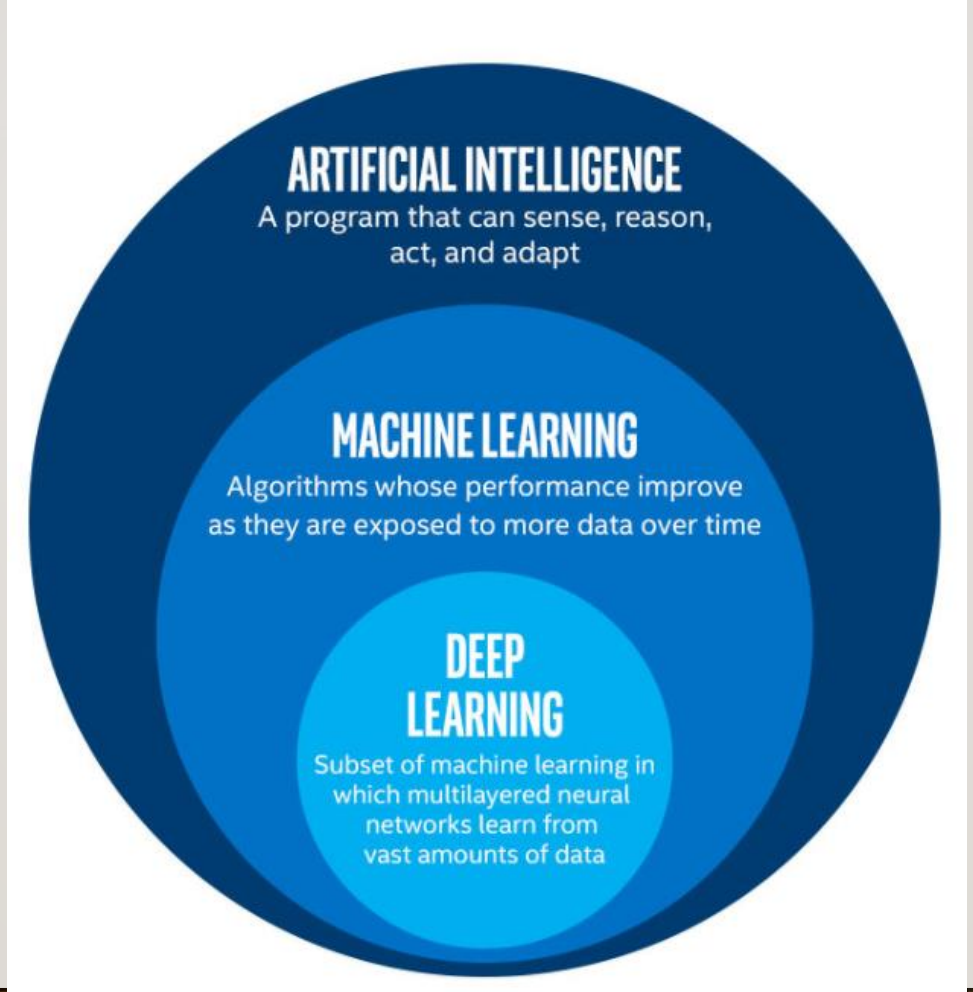
Neural Networks

Computer Vision

Natural Language Processing

BUZZWORDS

Source: Stack Exchange

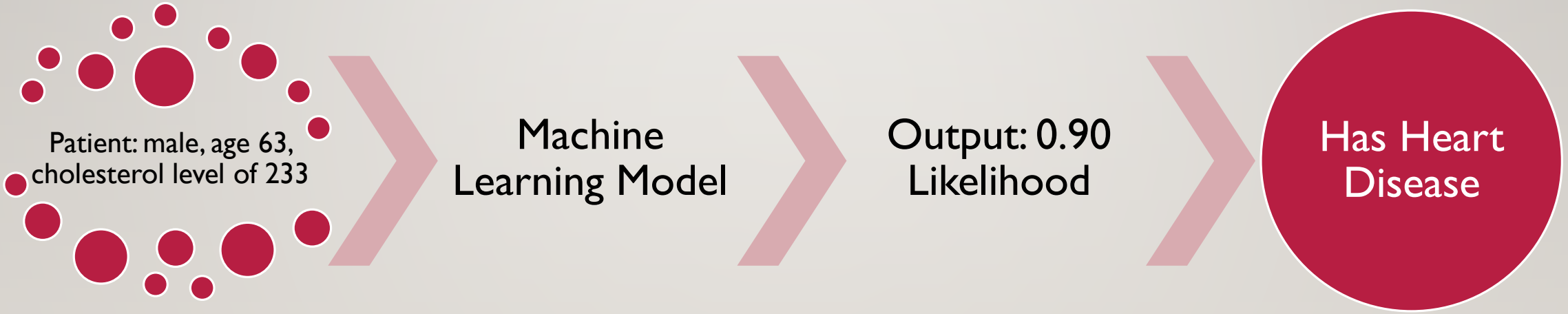


THE ARTIFICIAL INTELLIGENCE BOOM

- AI has garnered a lot of interest over the last few years
- A major breakthrough was in 2012 where AlexNet achieved a prediction accuracy of 84.7% on the ImageNet dataset.
- The boom in applications such as CV and NLP has been fuelled by computational power, collected and organised data, and improved digitalisation

THE BASIC MACHINE LEARNING PIPELINE

Inputs **Model** **Likelihood** **Prediction**



GENERAL QUESTIONS

I can take questions about:

- The machine learning pipeline and interesting developments
- Formulating data science problems and using machine learning to solve them
- How to go deeper into DS/ML, and some foundations
- Whether AI will cause the technological singularity and replace humans

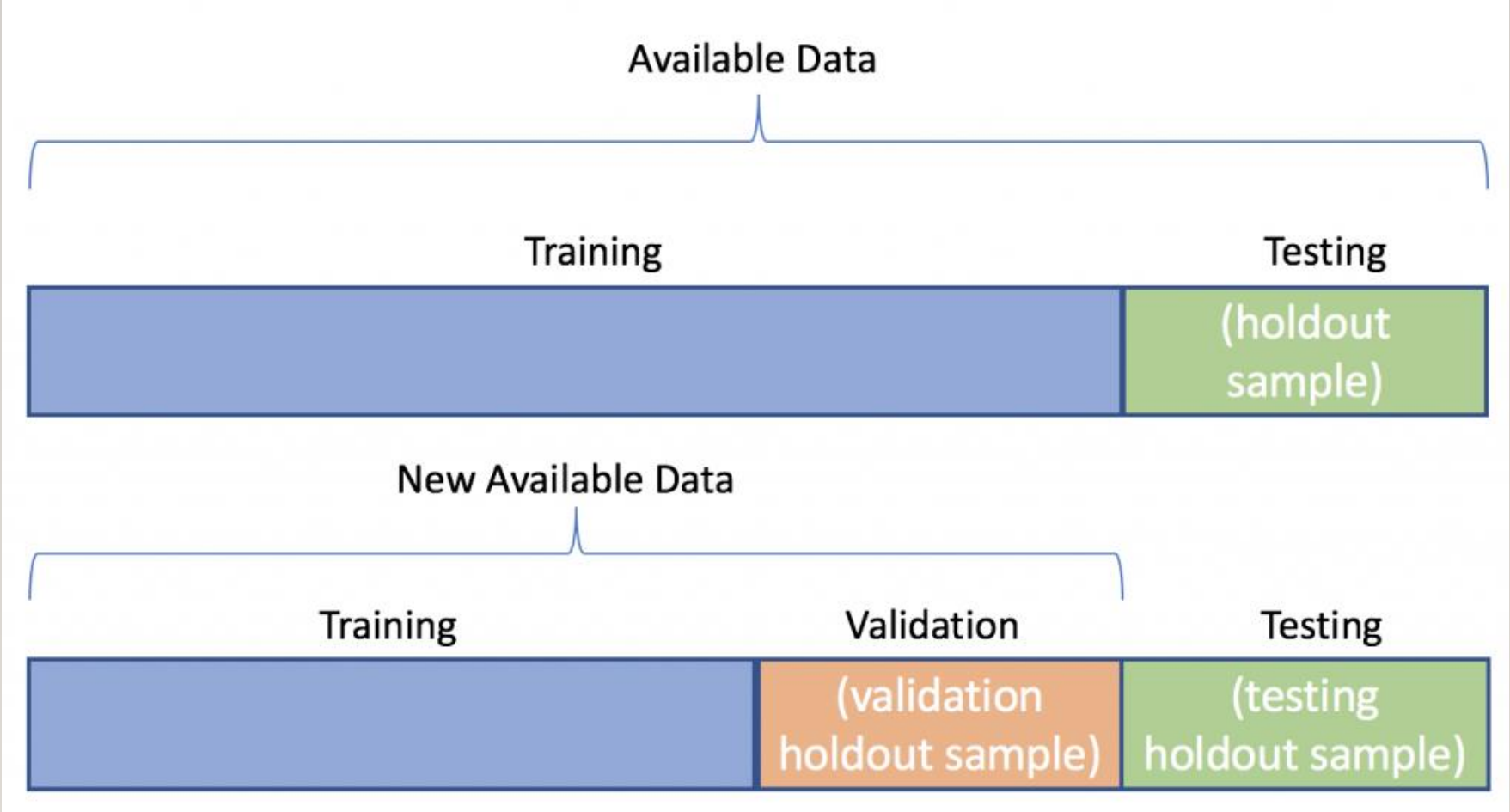
MACHINE LEARNING CONCEPTS


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

Source: xkcd

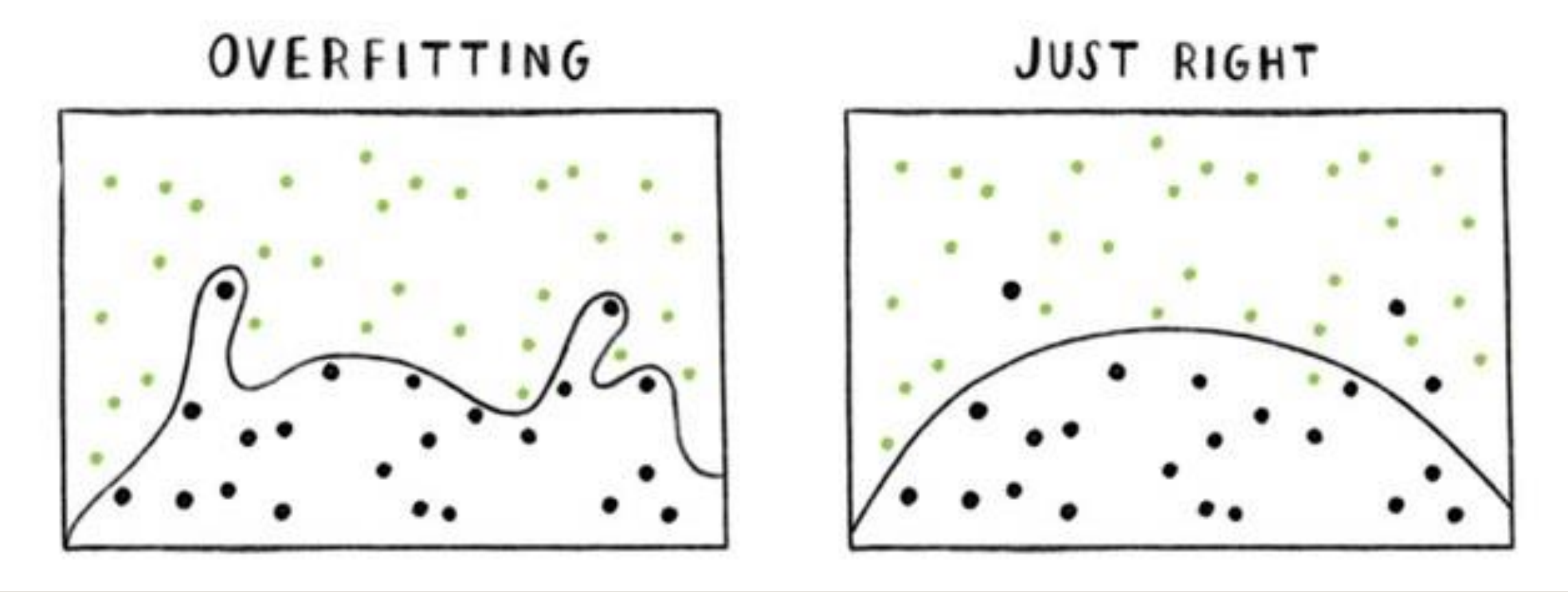
TRAIN-TEST SPLIT

Source: AlgoTrading 101 Blog



OVERFITTING

Source: KDNuggets



DEEP MODELS VS SIMPLE MODELS

SETTING UP THE ENVIRONMENT

Setting up the scripts:

Open Anaconda Prompt

Type “conda install tensorflow”

Wait

Type “conda activate tf”

EXPLAINABLE MACHINE LEARNING

TAKEAWAYS

Requirements:

1. Basic understanding of ML models (*as input-output functions*)
2. Appreciation for motivations behind explainable ML
3. Rough understanding of game theory, perturbations

Outcome:

1. Interpretability Methods (LR, DT, RF)
2. Explainability Methods (LIME, SHAP, example-based methods)
3. Apply packages in datasets

ADVANCED CONTENT

If folks are willing, I am game to discuss:

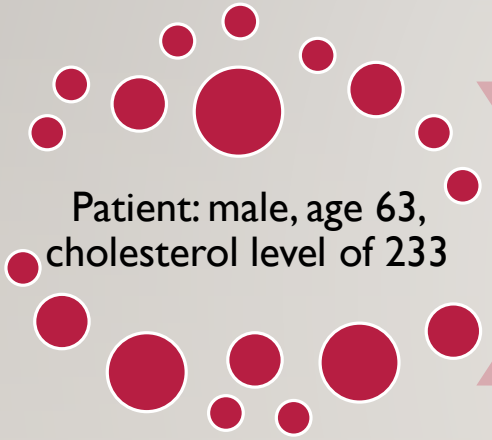
1. Mathematical formulations of SHAP (game theory), and the three axioms from Shapley values
2. Limitations (LIME not deterministic, SHAP's independence assumption, RF tree interpreter not peer-reviewed)
3. My research (Explainability Curves, Causal Hierarchy, Human-in-the-Loop Methods)
4. Other ex-ML methods (Prototypes and Criticisms)

① INTRODUCTION

WHY IS EXPLAINABILITY IMPORTANT?

GENERAL PREDICTION FRAMEWORK

Inputs **Model** **Likelihood** **Prediction**



Machine Learning Model

Output: 0.90 Likelihood

Has Heart Disease

Why did the model return the particular outcome?

What contributed to this prediction having a high likelihood?

ABOUT EXPLAINABILITY

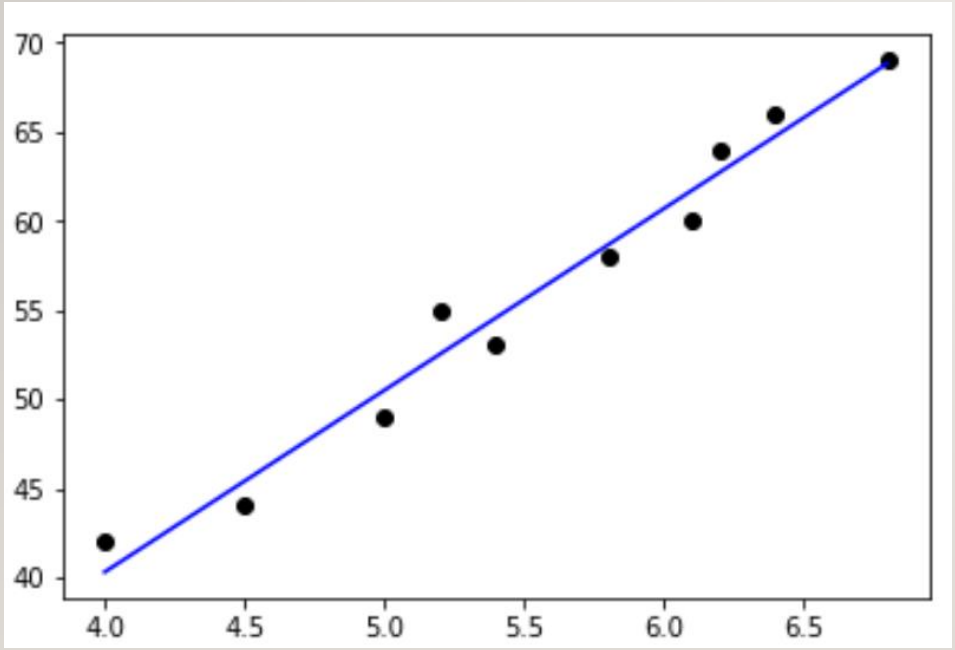
Explainable Machine Learning tells us how and why Machine Learning models behave and function:

- Which features (variables) are most important?
- What is an example of a positive/negative class?

MOTIVATION 1: DEMYSTIFYING MODEL COMPLEXITY

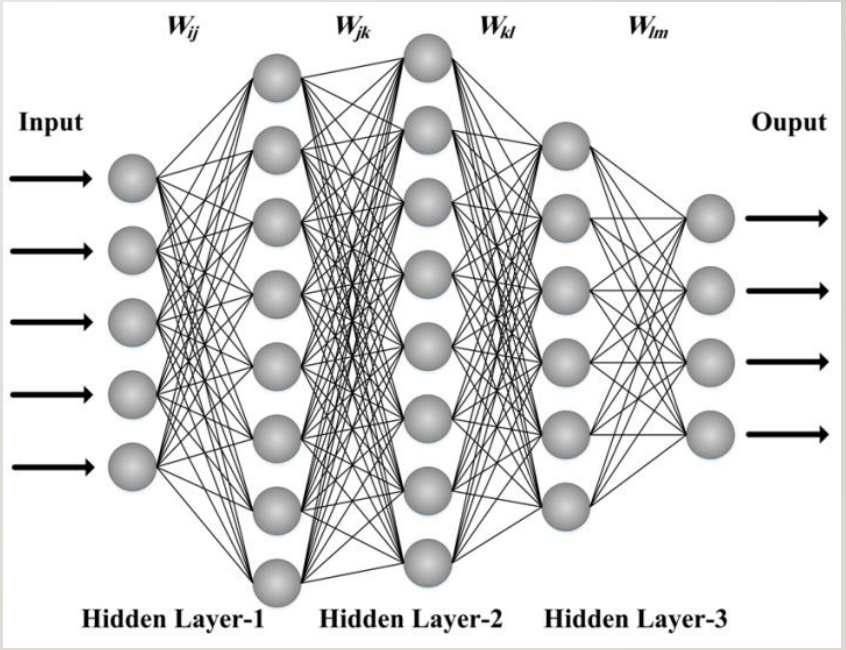
Simple Models

(eg. Linear Models, Decision Trees)



Complex Models

(eg. Convolutional/Recurrent Neural Networks, Random Forests, Attention Models)



MOTIVATION 2: TRUST

Users can trust the model, if we can explain why a certain prediction is made

- ML is deployed in a variety of scenarios (healthcare, finance, fraud detection), but end-users may not understand quantitative outputs
- End-users can understand if ML aligns with their own expertise.

MOTIVATION 3: TRANSPARENCY

When deploying ML, there is a need to tell patients, customers and citizens why certain predictions are being made.

- The new EU General Data Protection Regulation (2016) requires a **"right to explanation"**

② METHODS

WHAT TECHNIQUES ARE AVAILABLE?

LOGISTIC REGRESSION

Linear Models

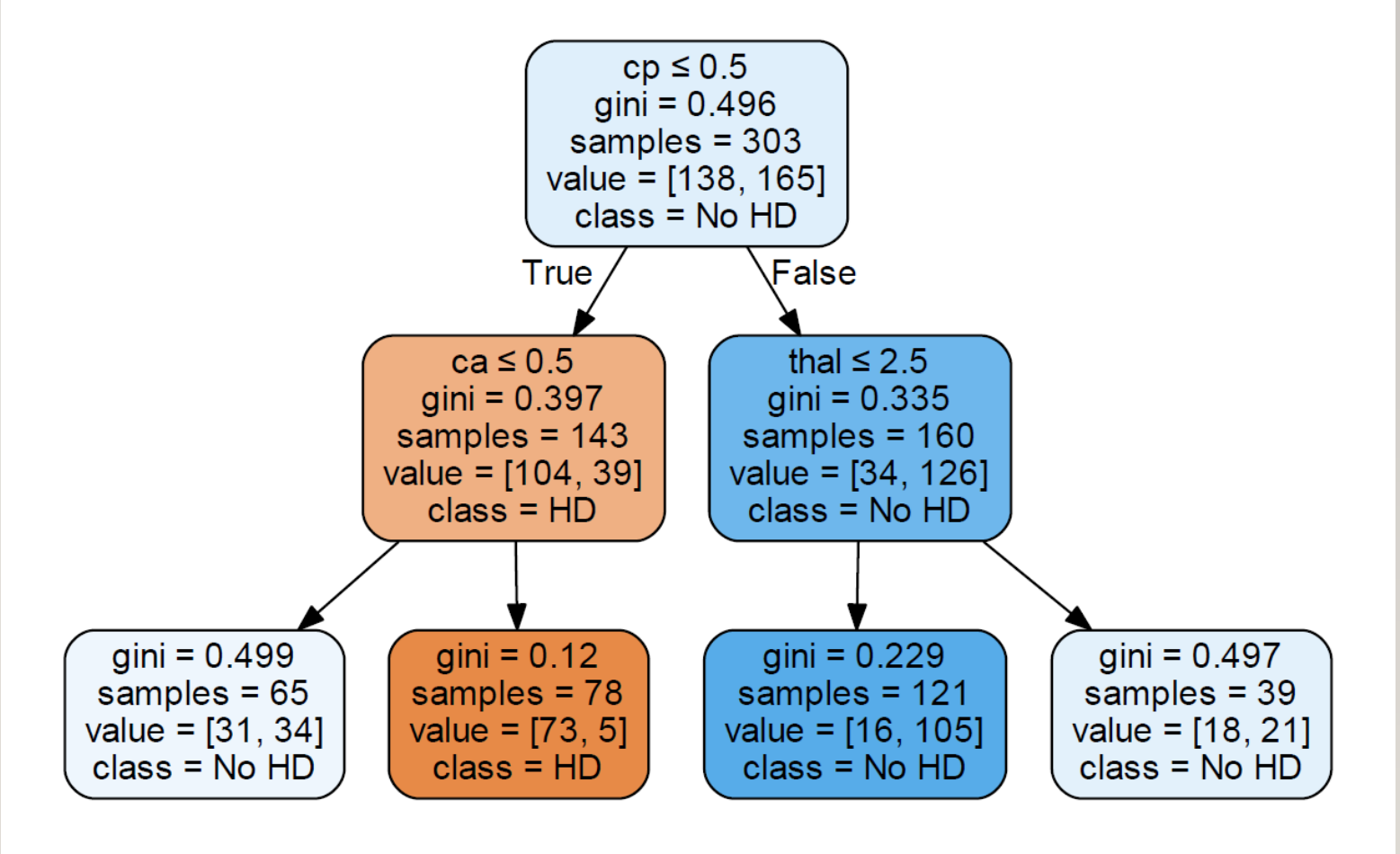
Linear models such as logistic regression (for classification problems) and linear regression (for regression problems) are formulated through a linear combination of input features x_i and variable weights β_i. For the Linear Regression model in Equation 2.1, the coefficients β_i directly represent the impact of a single variable. Conversely, the logistic regression Equation 2.2 models the log-odds through a linear combination of the individual i features.

$$y_{LinReg} = f(x) = \alpha + \sum_{i=1}^{N} \beta_i x_i \tag{2.1}$$

$$y_{LogReg} = f(x) = \log\left(\frac{\pi}{1-\pi}\right) = \alpha + \sum_{i=1}^{N} \beta_i x_i \tag{2.2}$$

DECISION TREE

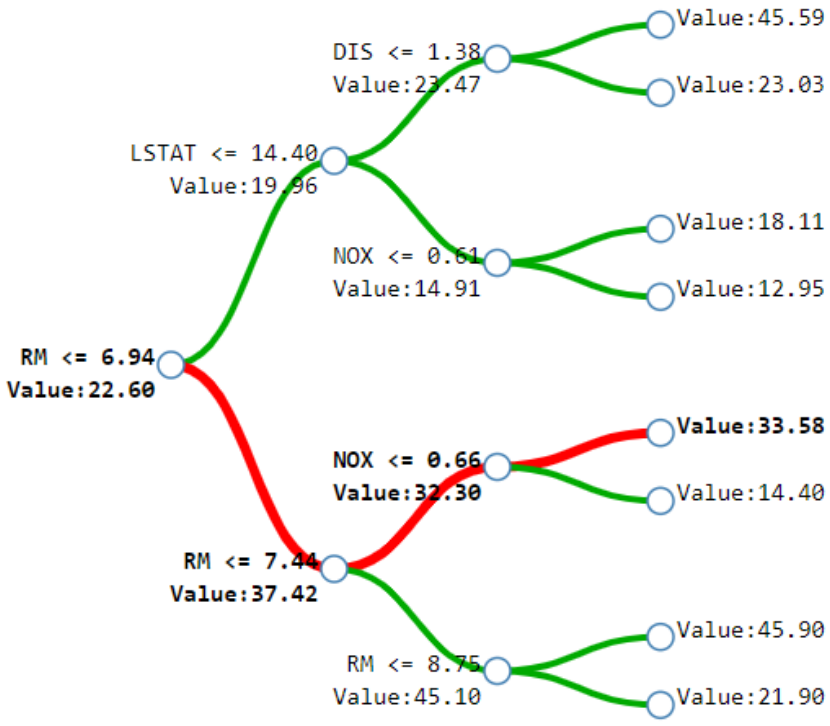
Decision Trees are commonly thought to be interpretable, because they can be visualised.



RANDOM FOREST

The TreeInterpreter method maps the decision path from the tree root to nodes.

The value returned quantifies the feature which provides the maximal information on the made classification, based on entropy.



TreeInterpreter Blog Post (2014)

Prediction: **33.58** ≈ **22.60** (trainset mean) + **5.12**(gain from RM) - **5.12**(loss from RM) + **1.28**(gain from NOX)

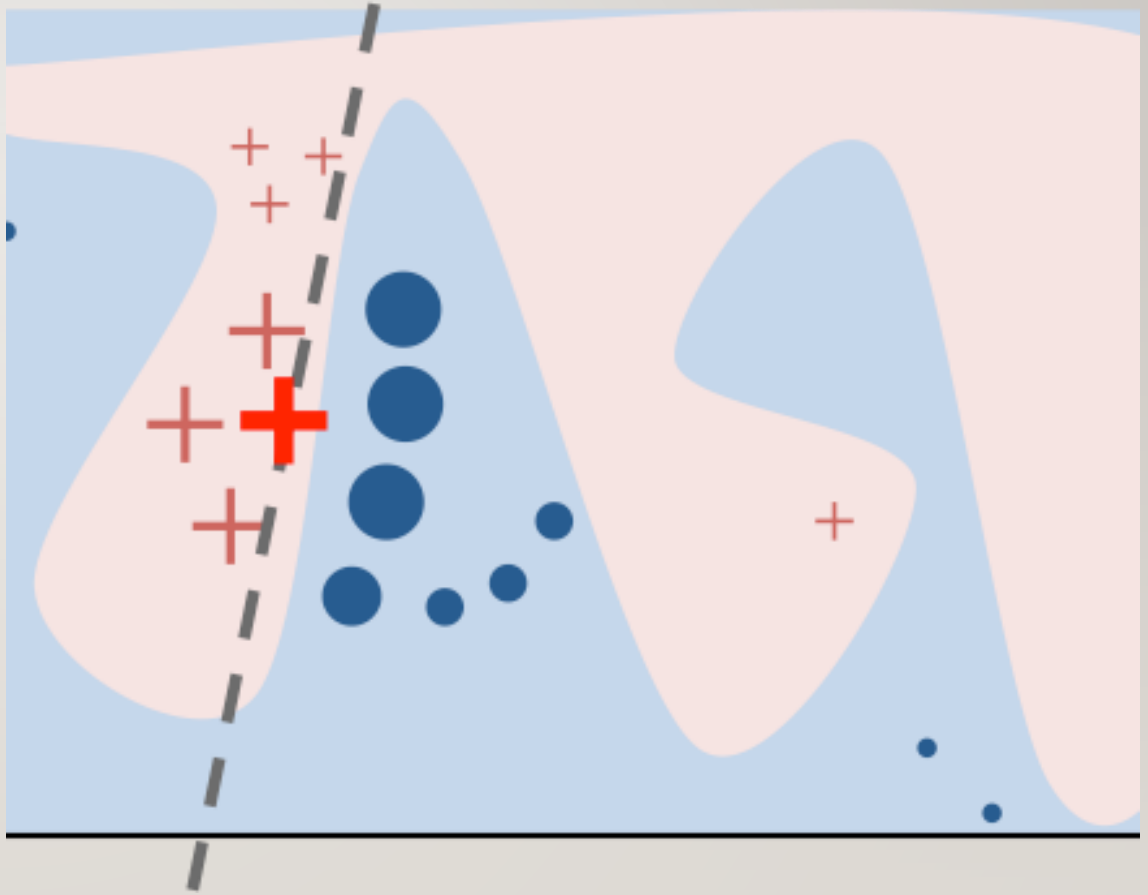
LIME

LIME creates approximations around a given prediction, and fits a linear model around these.

The simpler linear model is used as an explanation.

Loss function:

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

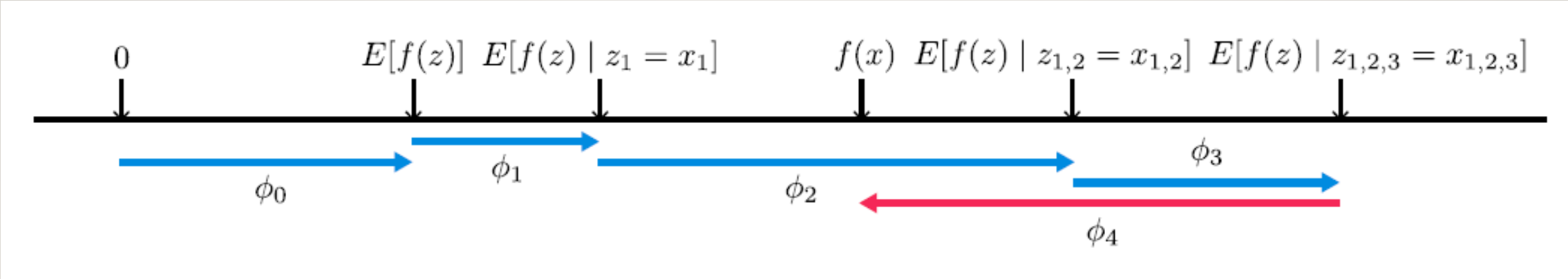


"Why Should I Trust You?", Ribeiro et al. (2016)

SHAP

SHAP represents the change in the model output when conditioned on the particular feature value.

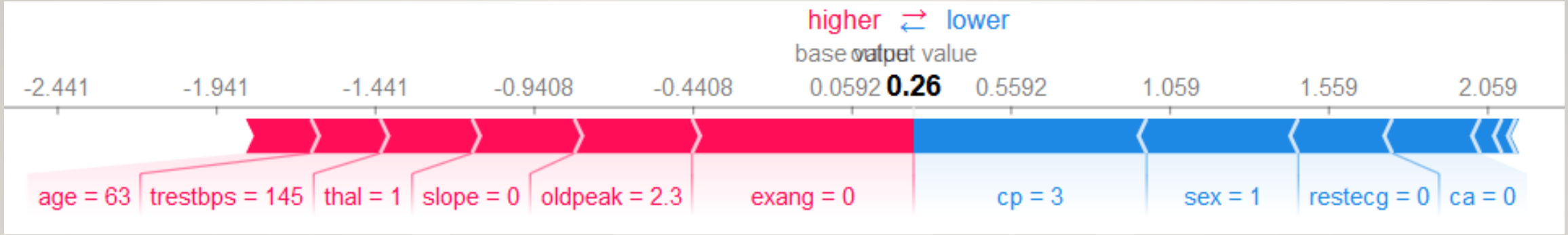
$$f(x_j) = \phi_0 + \sum_{i=1}^{M} \phi_{i,j,f} x'_{i,j}, i = 1, 2, ...M$$



A Unified Approach to Interpreting Model Predictions, Lundberg et al. (2017)

SHAP

What can a SHAP plot tell us?



A SHAP explanation for a single instance (patient)

SHAPLEY VALUES FROM GAME THEORY

Assume a game played by 3 players, A, B and C. Each coalition plays the game for a monetary reward of between \$0 and \$100. The set of rewards are described as follows:

Players	Outcome (\$)
None	0
A	10
B	20
C	30
AB	50
AC	60
BC	70
ABC	100

Calculating the outcomes for player A:

$A_{Missing}$	$Outcome_{Missing}$	$A_{Present}$	$Outcome_{Present}$	*Change*	*Permutations*	*No.*
None	0	A	10	10	ABC, ACB	2
B	20	AB	50	30	BAC	1
C	30	AC	60	30	CAB	1
BC	70	ABC	100	30	BCA, CBA	2

We average the individual contributions across the 6 sets of permutations:

- $\phi_A = \frac{1}{6}(10 * 2 + 30 + 30 + 30 * 2) = 23.33$
- $\phi_B = \frac{1}{6}(20 * 2 + 40 + 40 + 40 * 2) = 33.33$
- $\phi_C = \frac{1}{6}(30 * 2 + 50 + 50 + 50 * 2) = 43.33$

My Thesis (2021)

INTERPRETABILITY & EXPLAINABILITY

Interpretability:

- The ability for cause-effect to be observed within an ML system

Ex-ML is about Cause-effect analyses

Explainability:

- The post-hoc explanations designed for otherwise un-interpretable models

Ex-ML assumes that models are black-box – but is that true?

③ DEMONSTRATION

HOW TO USE THESE EXPLANATIONS?

④ GOING FOWARD

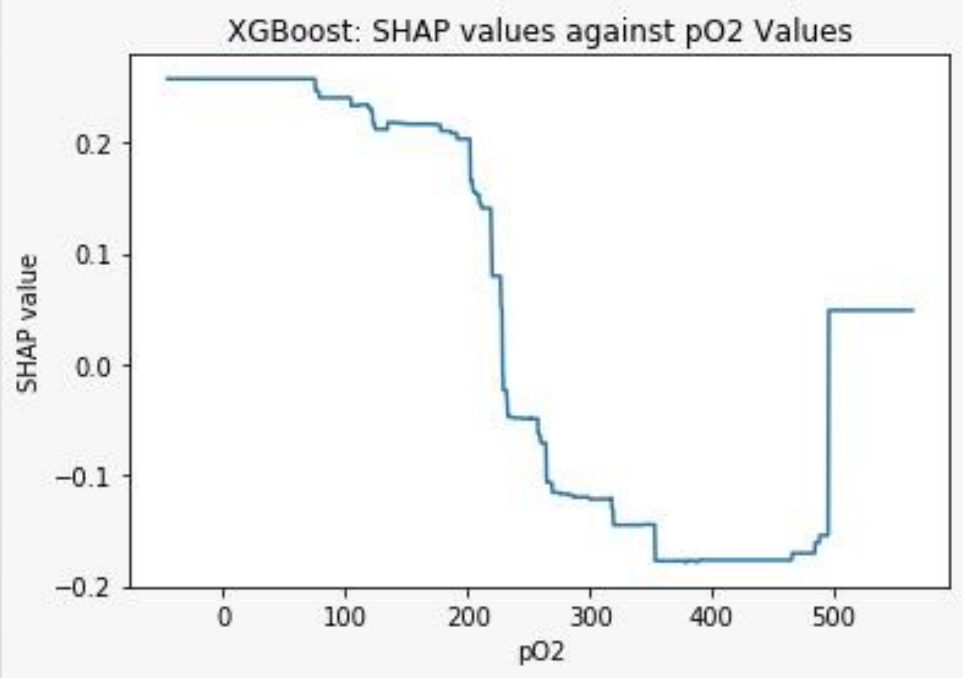
WHERE TO BEST USE EX-ML

SOME RESEARCH WORK

Focus: Explainable Machine Learning for Healthcare

Problems:

- How do we get doctors to trust ML models? Can ML replicate medical guidelines?
- Actionability: Can we tell doctors which features to act on?

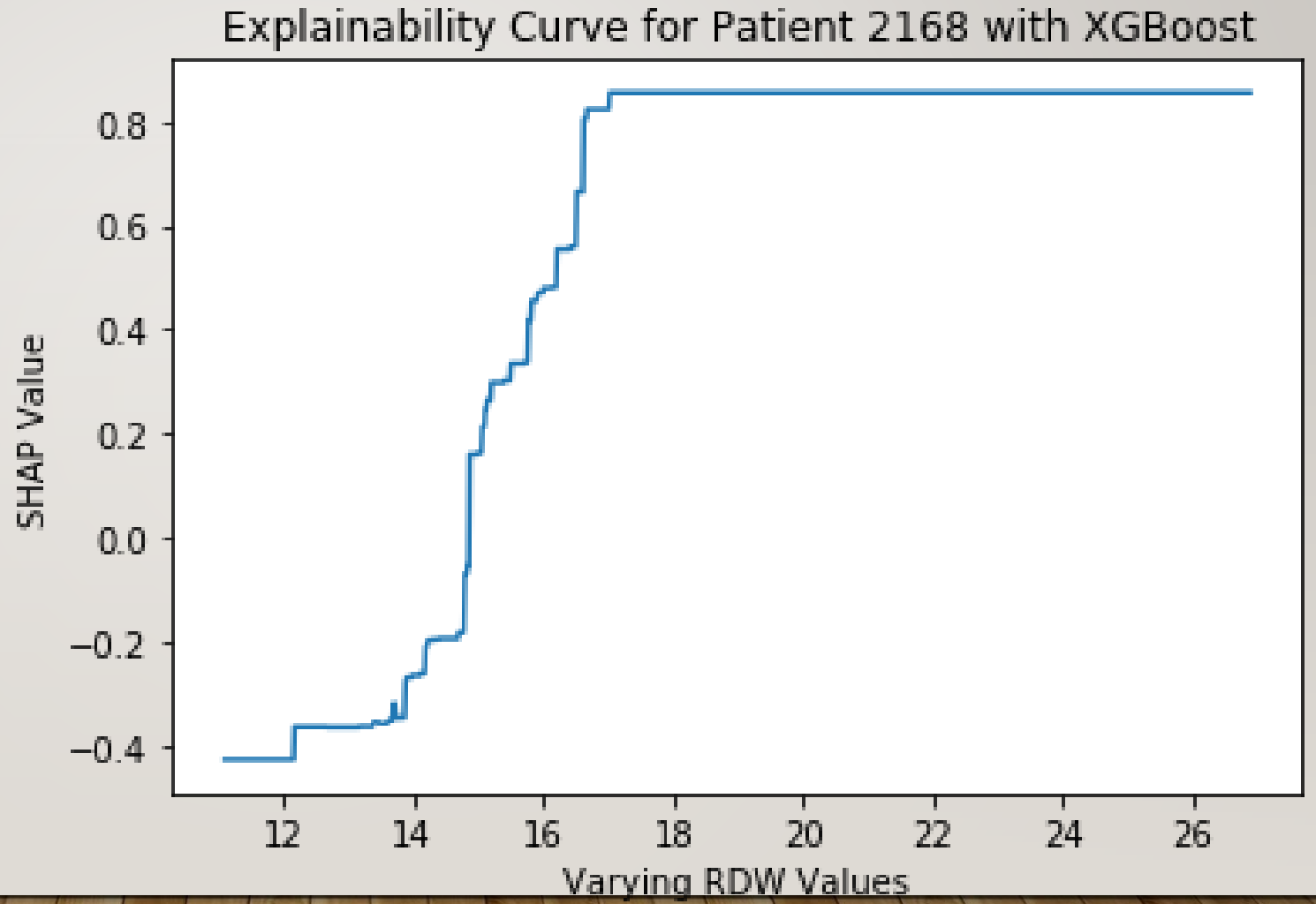


Explainability Curves

CLINICAL EXPLAINABILITY FOR AN INSTANCE

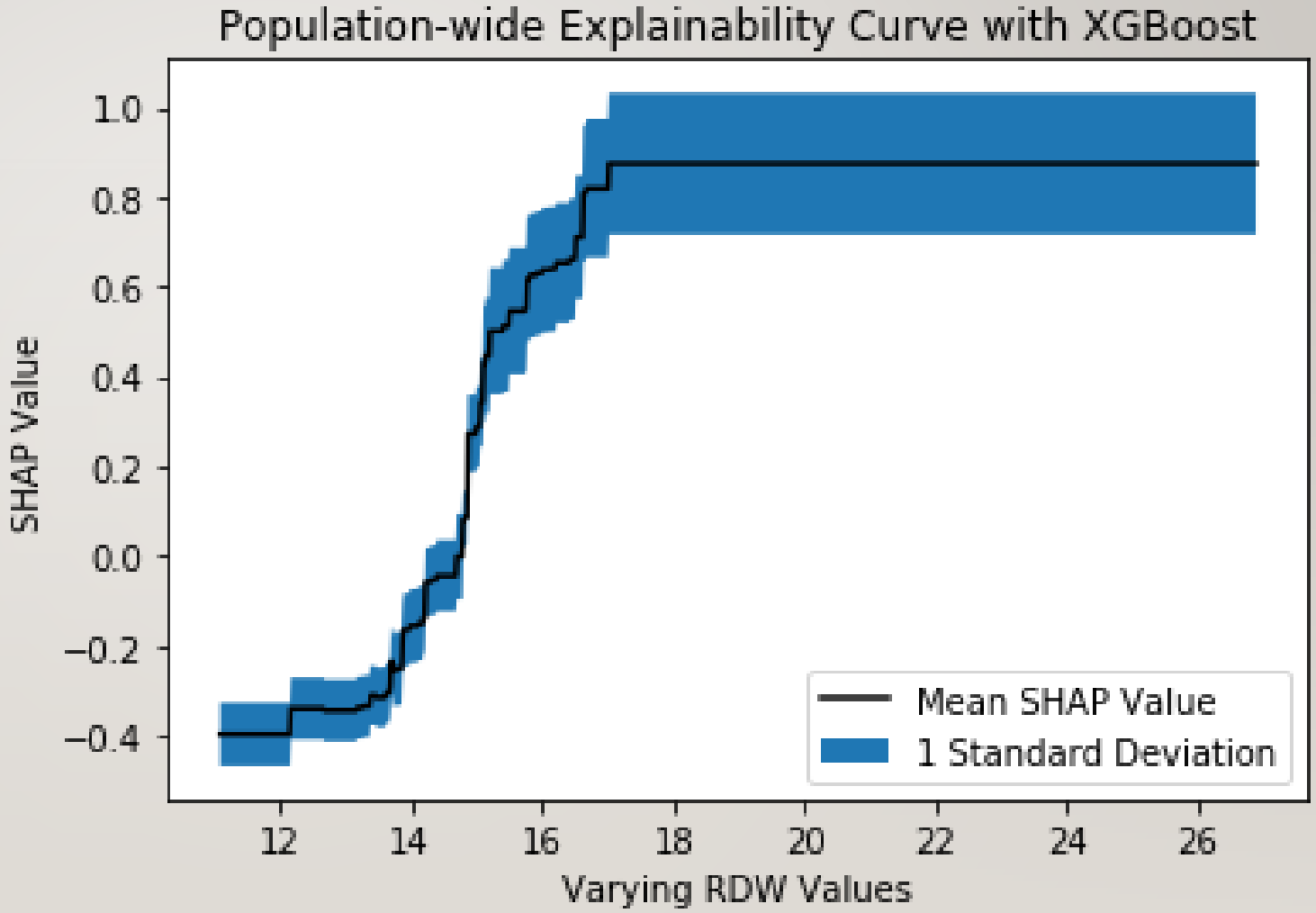
Our method commences by expanding the explanation to all possible feature values.

This clinical explanation shows feature importances, over the entire range of possible feature values.



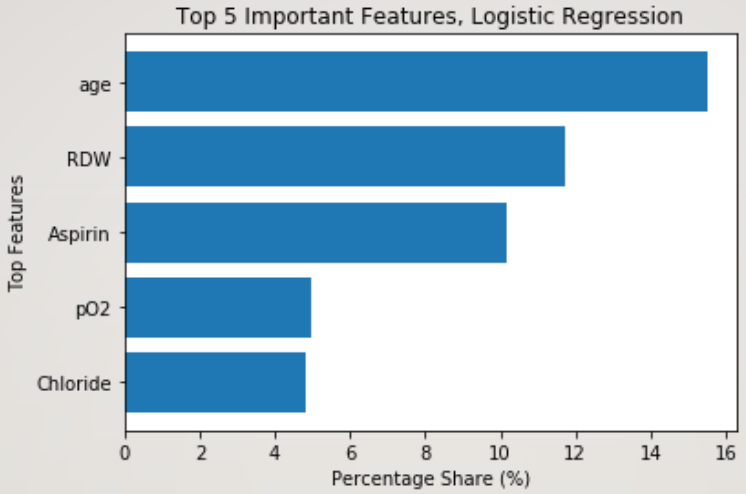
CLINICAL EXPLAINABILITY FOR A POPULATION

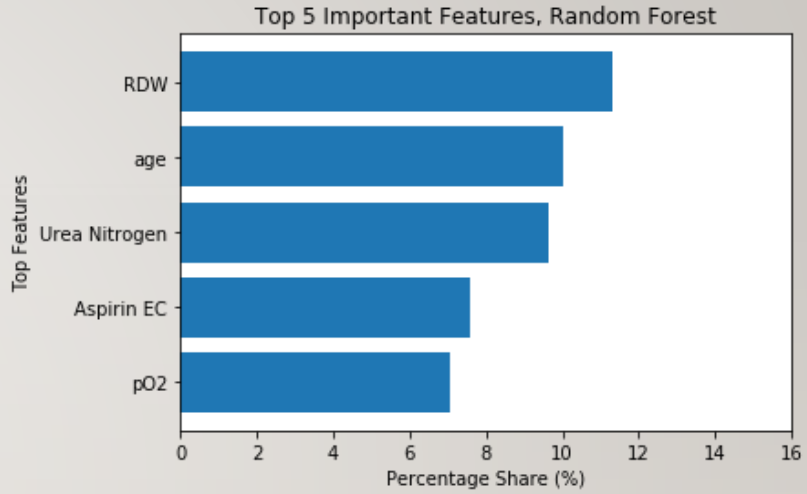
We obtain a distribution of which features are important across the whole population of patients.

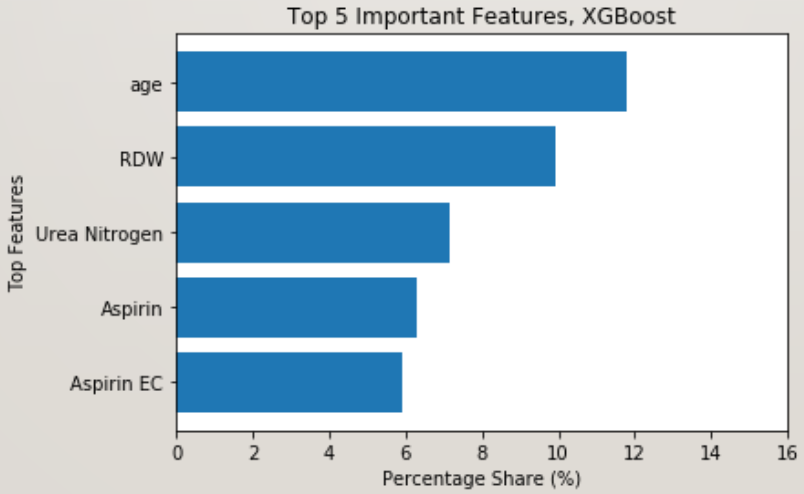


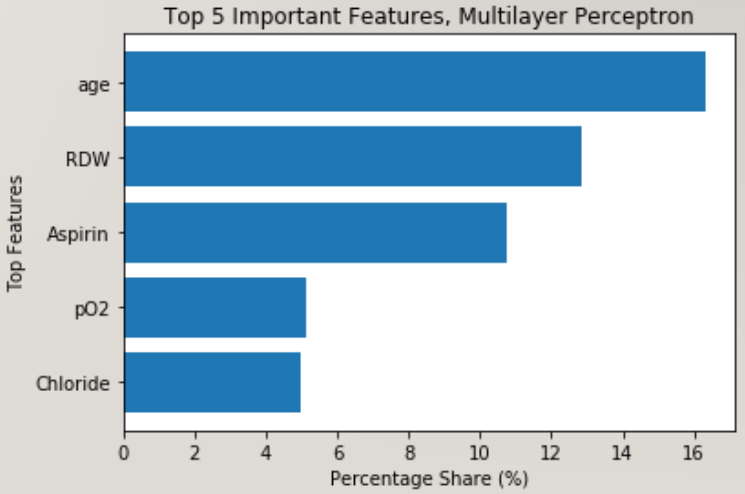
FEATURE IMPORTANCE VARIES ACROSS MODELS

Different ML models function in different ways: the features which are regarded as important differ.



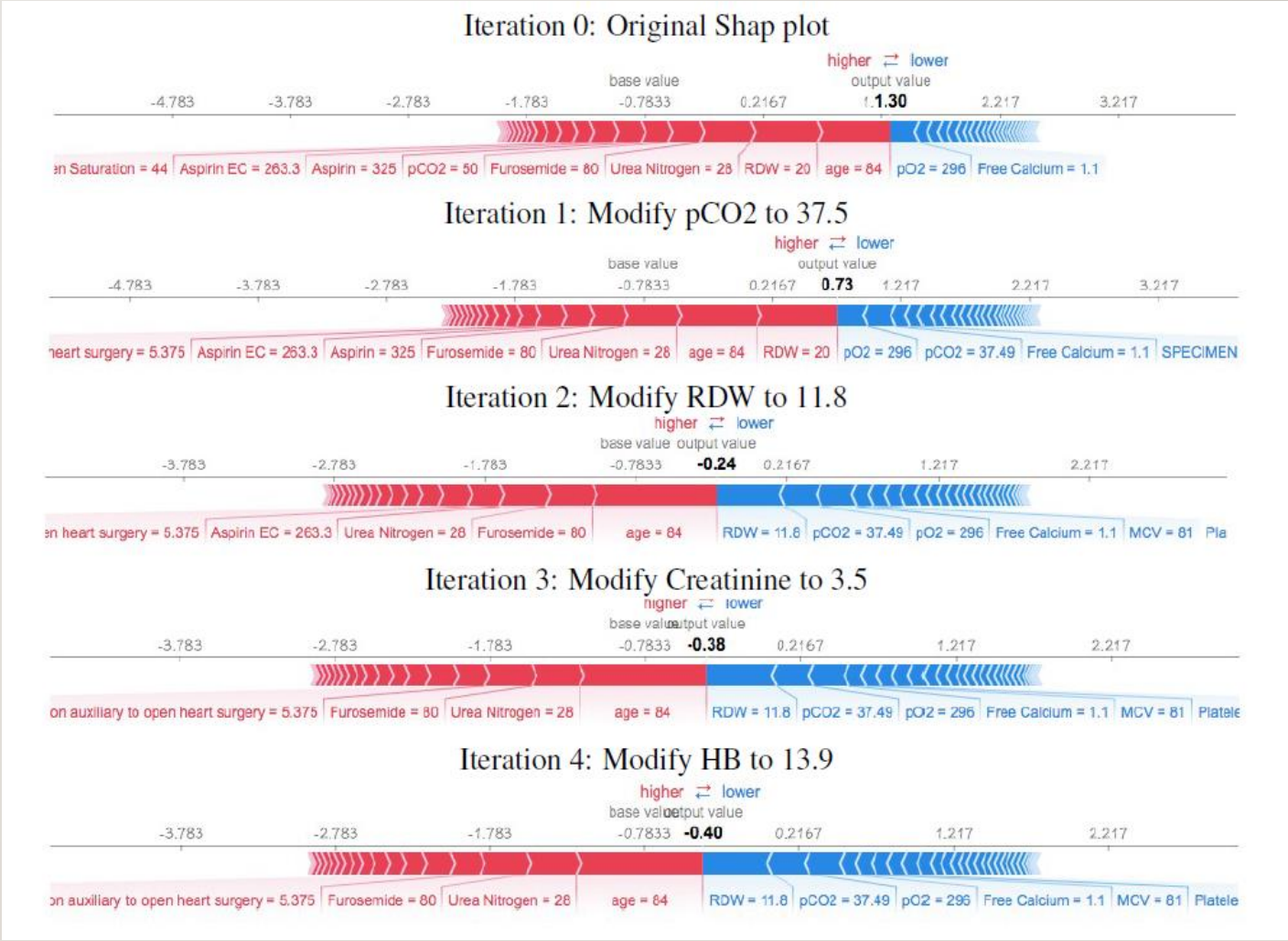






ACTIONABILITY FOR ONE INDIVIDUAL

We want to act on features to reduce a patient's likelihood of disease.



ACTIONABILITY ACROSS AN ENTIRE POPULATION

Across an entire population, we want the incidence of disease to be reduced. As such, we address the features that, when acted upon, reduce disease prevalence in a population.

STV Rank	**Feature**	**Actionable?**	**Target Value**	Mean Prediction
1	age	No	-	0.277
2	RDW	Yes	11.8	0.201
3	Urea Nitrogen	Yes	8.56	0.152
4	Aspirin	Yes	64.2	0.131
5	Furosemide	Yes	0	0.076
6	Acetaminophen	Yes	648.3	0.047
7	pO2	Yes	378.7	0.025
8	MCHC	No	35.9	0.021

Table 3: Actions globally taken on $k = 8$ features

CONCLUSIONS

Overview

Explainability methods are a promising solution to the black-box problem.

However, more work needs to be done to expand explanations, and address the lack of actionable insight.

Where do we go?

Moving from explainability, we want ML models to be actionable, so that they can motivate and recommend action.

To strengthen the case for actionability, further work is needed to establish causality in machine learning.

THANK YOU!

MING LUN **ONG**

EMAIL: MINGLUN.ONG@GMAIL.COM

TELEGRAM: @MUNG_LIN