Black Box Model Explainability

Data Sciences Institute
Topics in Deep Learning

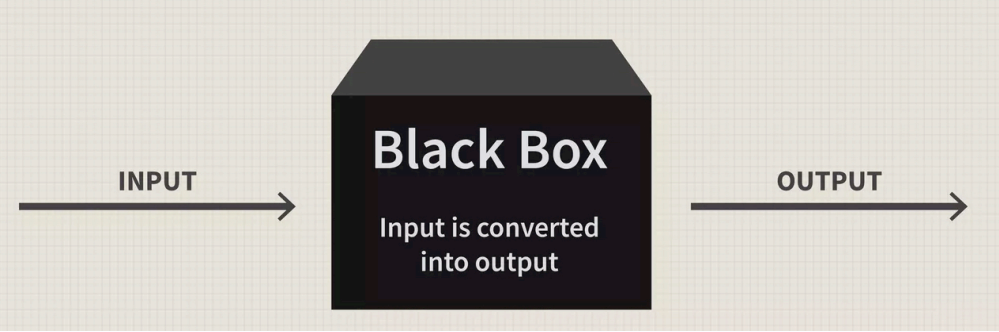
Outline

- Black box models
- Local methods
  - SHAP
  - LIME
- Global methods
  - Global surrogate models
  - HRT
- Model specific methods
  - Grad cam
  - Attention
- Mechanistic interpretability

Black Box Models

What is a black box?

- In general parlance, a "black box" refers to some function whose internal workings are unknown or opaque
- Like a machine model, it simply maps an input to an output: $f(x) = y$, where $f : \mathbb{R}^p \to \mathbb{R}^k$

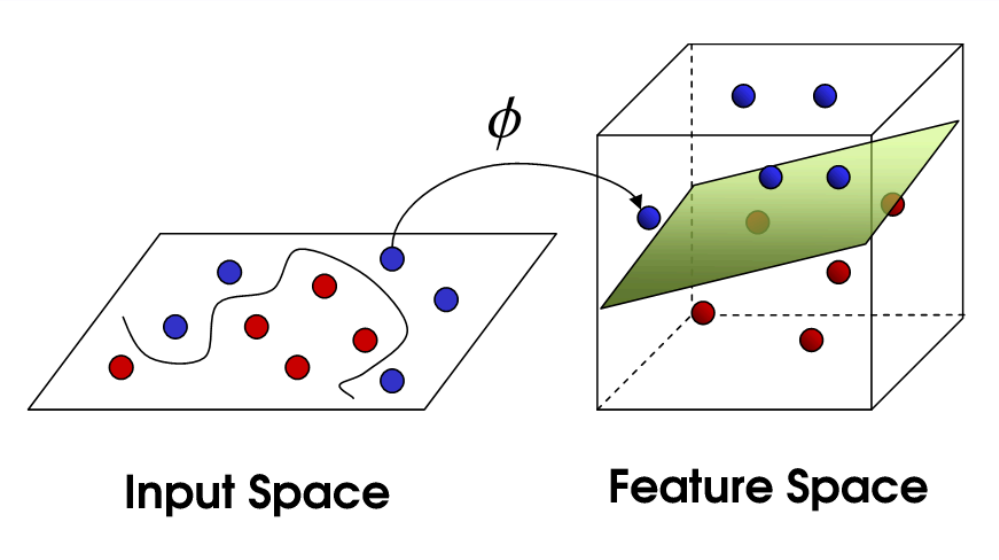


What is a black box in ML?

- In machine learning, we use black to refer to:
  - Algorithm classes "whose internal workings are unknown or opaque"
  - Methods that work for any arbitrary function (i.e. as long as you can perform inference or possibly take a gradient from $f_\theta(x)$)
- Questions:
  - What are some ML algorithms you would consider "black boxes"?
  - What are some methods that work for arbitrary functions?

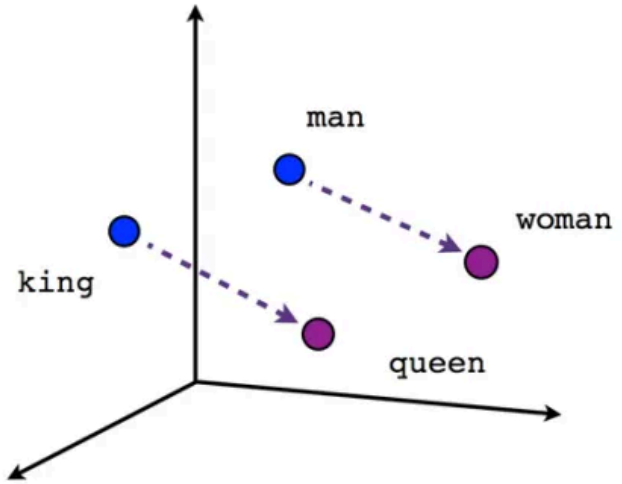
Why use a black box model?

- Flexibly model arbirtary functions

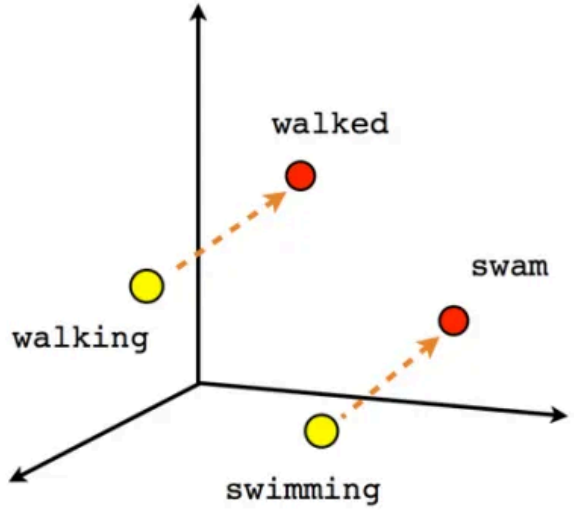


Why use a black box model?

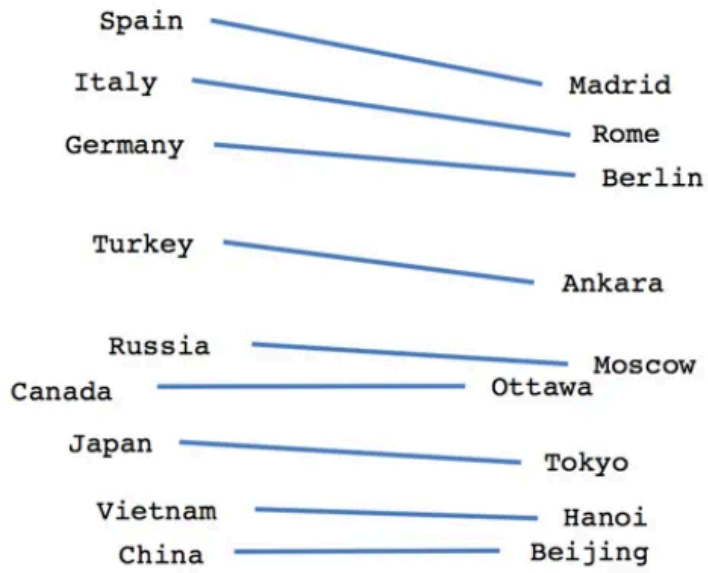
- Learned features >> hand-crafted features (usually)



Male-Female



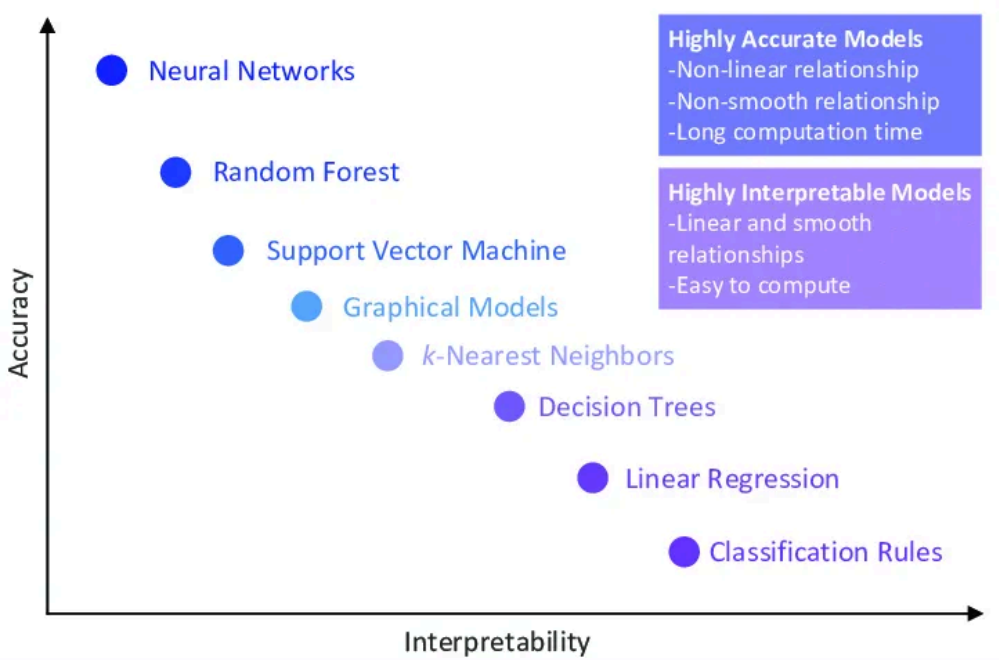
Verb tense



Country-Capital

Complexity vs interpretability

- Many ML models, and all DL modelsm are capable of modeling highly complex, non-linear relationships
- There is usually (but not always) a trade-off between model compexlity and performane



Source: Morocho-Cayamcela et. al (2019)

Challenges of black box models

- **Lack of transparency**: how do models generate their predictions?
- **Poor interpretability**: given a certain input, why was a particular prediction made?
- **Trust issues**: stakeholders may find it challenging to trust a well-performing model that they cannot understand
- **Accountability issues**: if a black box model's behaviour results in serious issues such as death, who should be held accountable?

Importance of explainability

- Explainability is crucial for models deployed in high-stakes environments such as healtchare

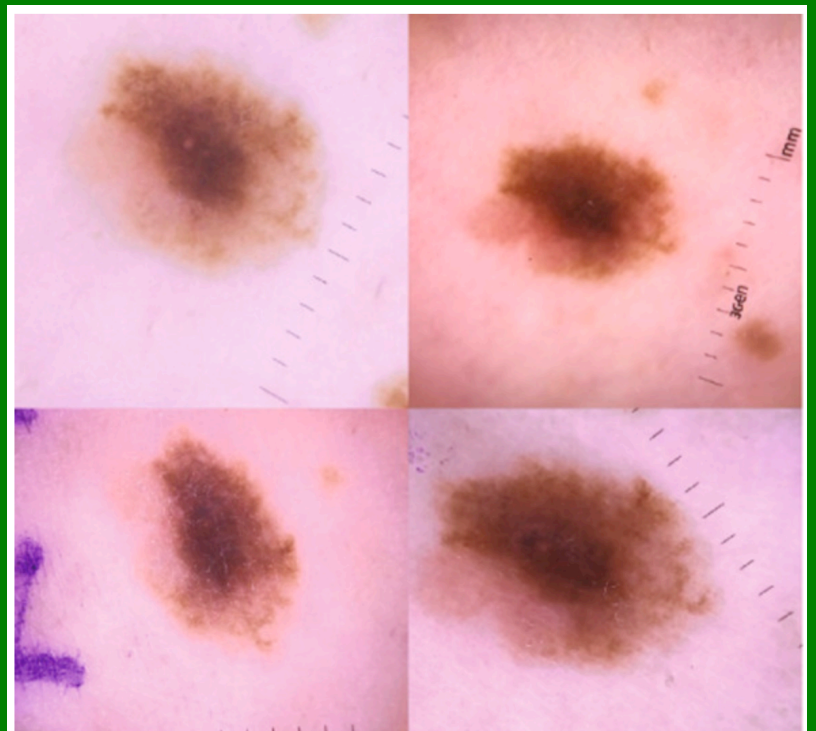
- The more we understand a model, the more we can:
  - Build trust among stake-holders
  - Foster ethical AI practices
  - Ensure regulatory compliance
  - Facilitate and contextualise model debugging and improvement

Lesson objective

- Explore different methods for elucidating understanding how complex, non-linear models work

Breakout #1

Suppose there is a melanoma classifier that uses a CNN. As a potential future patient, how would you want this classifier to explain its "prediction" about whether you had melanoma or not from your picture?



Local Methods

Why is this person at high risk of suffering from cardiovascular disease?

Understanding individual predictions

- **Local explainability methods** offer insights into individual predictions made by black box models
  - They focus on explaining why a particular prediction was made for a specific instance or region of the input space

Why does it matter?

- DL models are trained on datasets that may not be representative of the entire population
  - If the training data contains biases, such as overrepresentation or underrepresentation of certain demographic groups, these biases can be learned by the model

- Understanding why a model behaves differently for each individual or subgroup can help stakeholders identify and address algorithmic bias and unintended behaviours

Methodological approaches

In this lesson, we will go over two different approaches for local explainability:

1. Variable attribution (SHAP): how does an individual prediction differ from the average, and how can this difference be attributed among the input variables?

2. Surrogate models of behaviour (LIME): can we model black box behaviour locally using an easy to interpret white box model?

Variable Attribution: SHAP

SHapley Additive exPlanations (SHAP)

- SHAP is a method for explaining the output of machine learning models by quantifying the contribution of each feature to the prediction

- This is based on the concept of Shapley values from cooperative game theory: given a set of players (features), how do we distribute the payout (prediction) resulting from a collaborative game (prediction task)

Calculating variable contribution

- SHAP considers all possible subsets of features, known as coalitions, for a given instance
  - Each coalition represents a different combination of features
- For each feature value within a coalition, SHAP calculates its marginal contribution by comparing model predictions with and without the feature value included in the coalition
  - This captures how much the inclusion of that feature value changes the prediction
- The Shapley value for each feature value is computed as the **average of its marginal contributions** across all possible coalitions

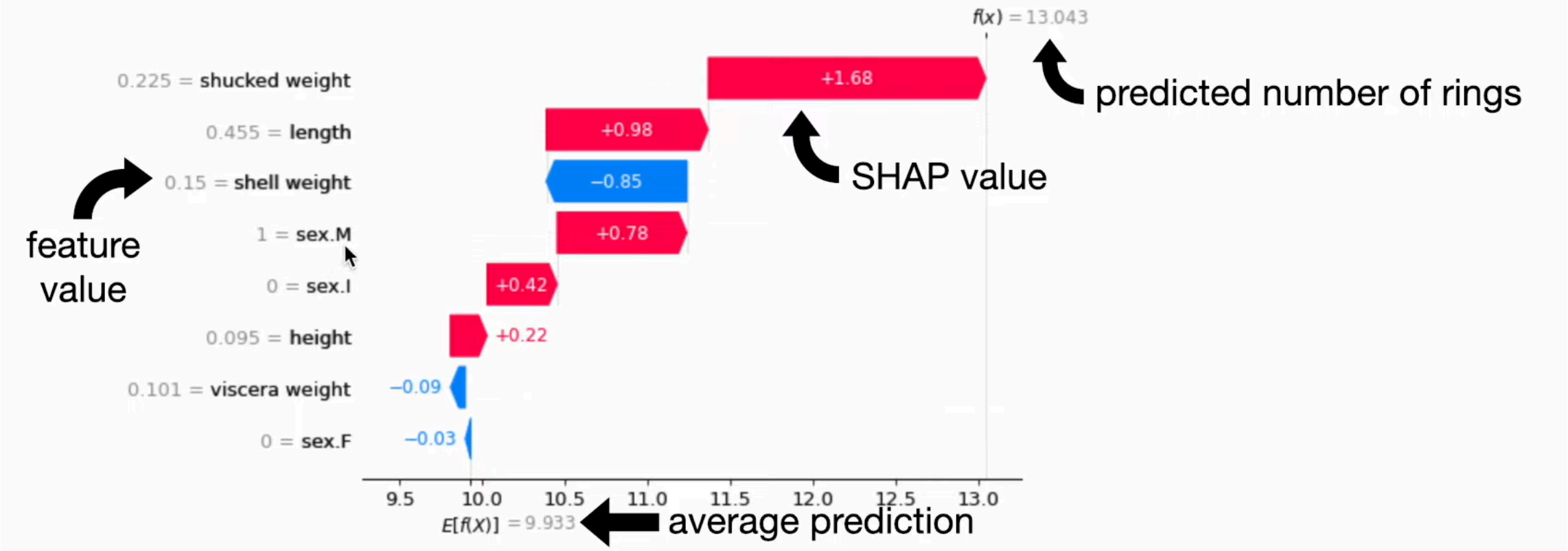
Calculating variable contribution

- Consider a model predicting risk of heart disease based on **age, cholesterol levels, and smoking status**
- SHAP generates all possible combinations of these features, ranging from no features to all three features included in the coalition
- For each feature value within a coalition, SHAP calculates its marginal contribution by comparing model predictions with and without that feature value included
    - For instance, it measures how much adding the cholesterol level feature to a coalition changes the model's prediction compared to when it's absent

Interpreting SHAP values

- **Sign**: positive SHAP values indicate that a feature value increases the prediction, while negative values indicate a decrease
- **Magnitude**: the magnitude of the SHAP value represents the importance or impact of that feature value on the prediction
- **Additive property**: the sum of SHAP values across all features equals the difference between instance and average predictions
- **Visual interpretation**: SHAP values can be visualized using various plots, such as the waterfall plot, which displays how individual feature values push the prediction of an instance away from the average value

SHAP waterfall plot: predicting the number of rings in an abalone shell



Limitations of SHAP

- **Computationally expensive**: considering all coallitions can be computationally intensive, especially in complex contexts
- **Assumption of independence**: considering all possible coallitions equally may does not reflect feature interdependence, which indicate that certain coallitions are more likely than others in real life
- **Potential misinterpretation**: Users may sometimes misinterpret SHAP values, assuming causality or feature importance and producing false conclusions

Surrogate models: LIME

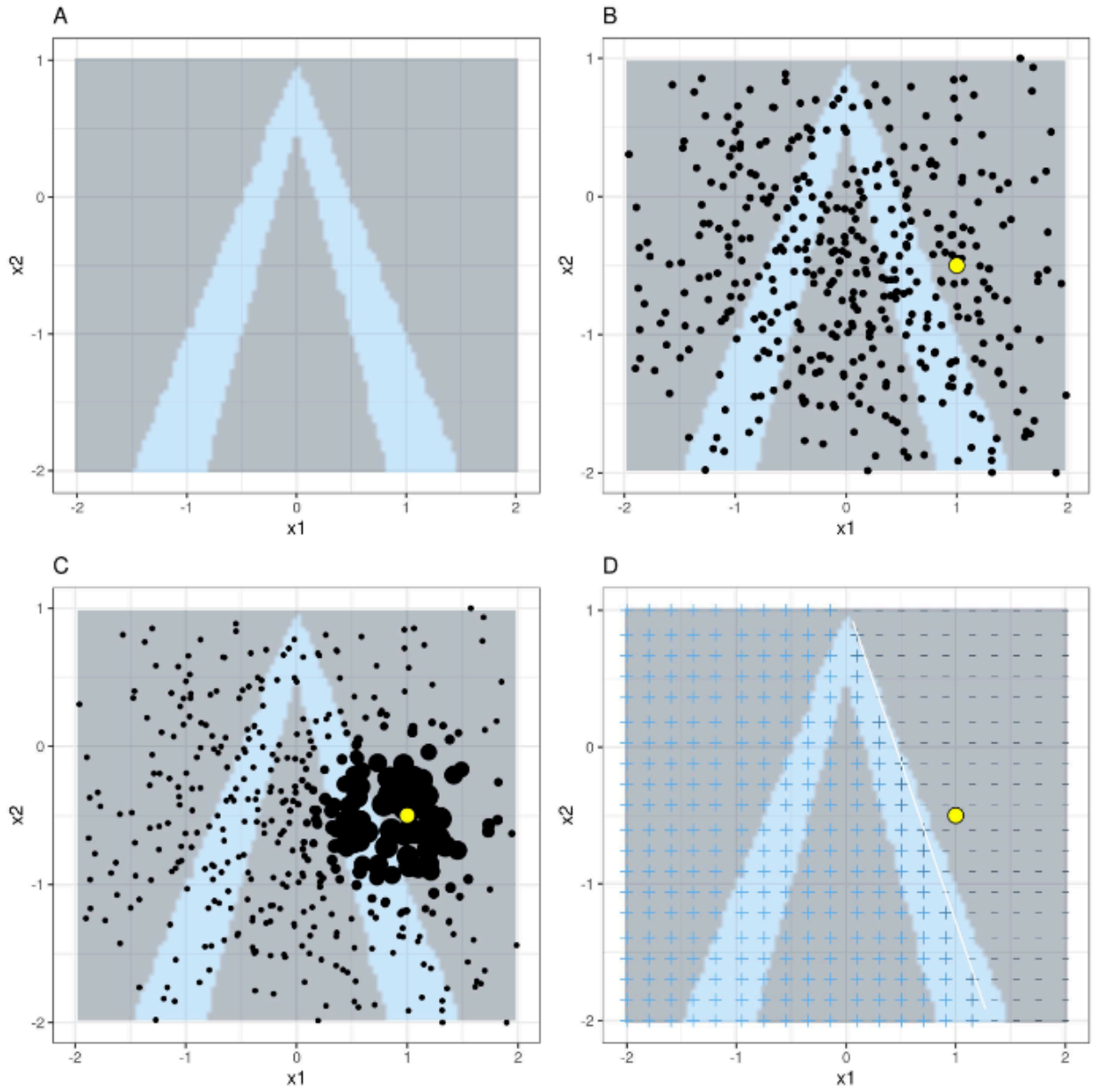
Local Interpretable Model-agnostic Explanations (LIME)

- LIME is a technique for explaining individual predictions of black box machine learning models at a local level
- It approximates the behavior of the black box model by training interpretable surrogate models on perturbed instances around the prediction of interest
- LIME provides insights into why a specific prediction was made by highlighting the contribution of different features for that instance

Mechanistic overview

Given an original instance of interest, LIME does the following:

1. Sample multiple new instances around the original neighbourhood
2. Weight each new instance according to their proximity to the original
3. Train a simple, interpretable model (such as linear regression) on the neighbourhood data
4. Explain the original instance's prediction by interpreting the surrogate model



A. Black box model predictions given features x1 and x2

B. Instance of interest (big yellow dot) and data sampled from a normal distribution (small dots)

C. Assign higher weight to points near the instance of interest.

D. Signs of the grid show the classifications of the locally learned model from the weighted samples. The white line marks the decision boundary (P(class=1) = 0.5

Limitations of LIME

- **No single correct way of defining a neighbourhood**: the reweighing function used for new sampled instances based on their distance from the original is variable and can have important impacts in downstream results
- **Generation of unlikely samples**: sampling a neighbourhood by using a normal distribution around the instance of interest may generate samples that wouldn't exist in real data, leading to surrogate models that do not adequately represent the real underlying data distribution
- **Model dependence**: interpretability results heavily depend on the choice of both the black box model and the surrogate model

Global Methods

Which risk factors are generally associated with higher risk of suffering from cardiovascular disease?

Understanding overall model behaviour

- **Global explainability methods** offer insights into average model behaviour and general data characteristics

Why does it matter?

- Global explainability enhances our general understanding of a model's decision-making process across an entire dataset, enhancing methodological transparency and increasing trust amongst stakeholders

- It also facilitates model debugging and improvement by identifying unexpected behaviours and potential areas of improvement, such as feature selection

Methodological approaches

In this lesson we will go over two different global explainability approaches:

- **Global surrogates**
- **Holdout Randomization Test (HRT)**

Global Surrogates

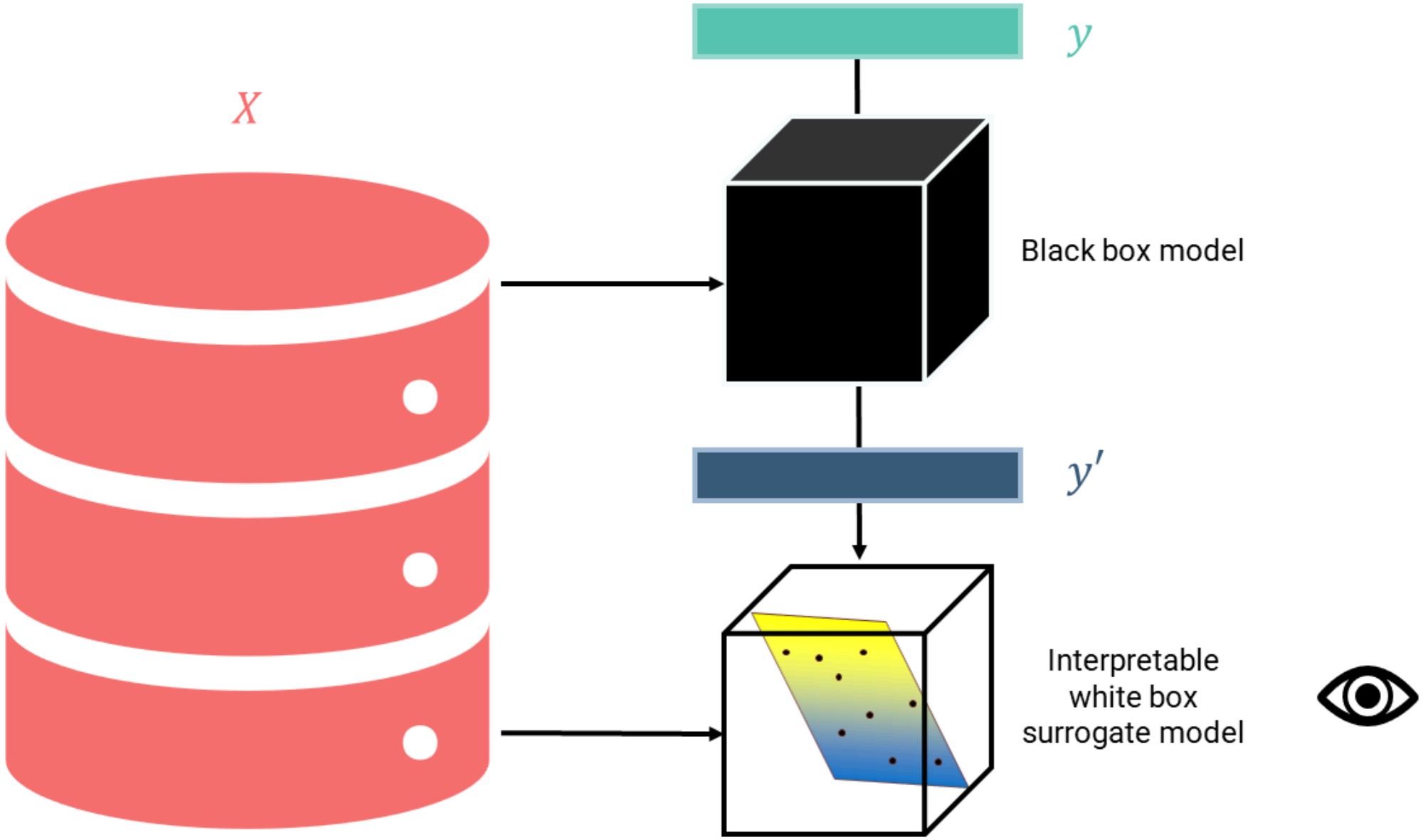
Global surrogate models

- A global surrogate is a simple, interpretable model (e.g., linear regression or decision tree) trained to approximate the predictions of a black box model
  - We are using simple machine learning to model the behaviour of more complex DL algorithms

Basic principle

A global surrogate model can be obtained and interpreted as follows:

1. Define a dataset X
2. Obtain prediction outputs of X using the black box model
3. Select and train an interpretable model using X as input and the black box predictions as output
4. Measure how closely the predictions of both models align
5. Interpret the surrogate model (e.g., which features have the most important coefficients in linear regression)


y
X
Black box model
y'
Interpretable
white box
surrogate model

Limitations

- **Misinterpretation**: the insights gained from global surrogate models are related to model behaviour, **NOT** to the characteristics of the data itself
- **Susceptibility to choice of training data for surrogate model**: the surrogate model can be trained in any data of similar distribution to that used by the black box model in training. It can happen that surrogate models can model black box behaviour better for some data subsets than others
- **How good is good enough?**: there are no clear rules to determine how similar the surrogate model predictions have to be to its black box counterpart to be considered an acceptable approximation of behaviour

Holdout Randomization Test

Holdout Randomization Test (HRT)

- Given a trained model and a held out test set, HRT repeatedly evaluates model performance on the test set following individual feature perturbations
  - Measuring the impact of these perturbations on model predictions serves as a proxy of overall feature importance

- These measures provide insights into feature interactions and overall model behaviour

HRT algorithm

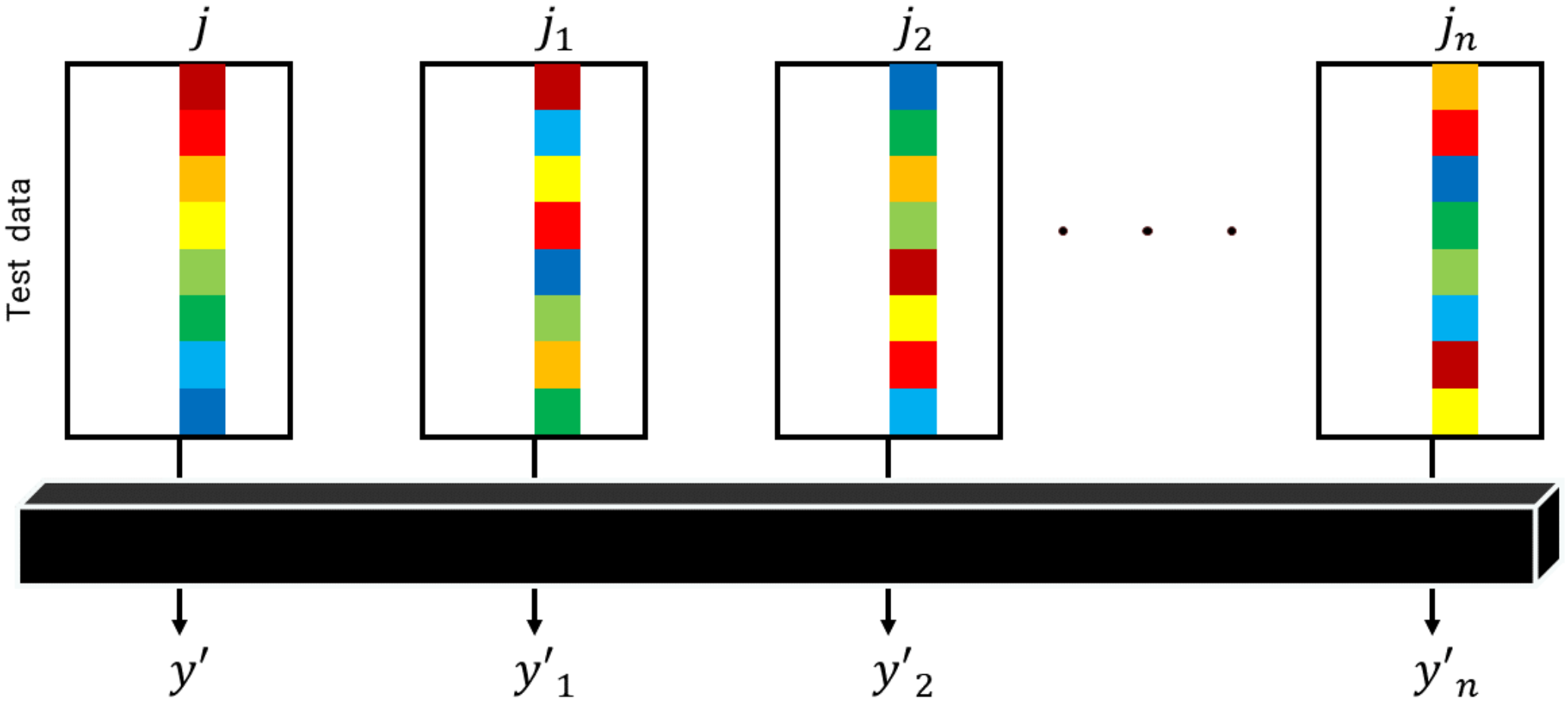
Given a trained model and a test set HRT can be implemented as follows:

1. Compute baseline performance on the test set
2. For each feature in the test set:
    - Shuffle the feature of interest
    - Evaluate test set performance following this shuffle
    - Repeat this process multiple times to generate a distribution of performance given the shuffled feature
    - Compute a test statistic to determine whether or not the disturbance of this feature led to worse test set performance

Interpreting HRT results

- At a high level, HRT conducts a conditional independece test for each feature X_j, with the null hypothesis stating that an outcome y is independent of feature X_j given all other features

- Intuitively, if X_j is predictive of y, perturbing this feature in isolation will break down its relationship to y and lead to drops in performance



$$H_0: loss(y', y) = \frac{1}{n}\sum_{i=1}^{n} loss(y'_n, y)$$

Limitations of HRT

- **Sensitivity to test set size**: effectiveness of HRT may vary depending on the size of the holdout set, with smaller holdout sets potentially leading to less reliable assessments of feature importance
- **Limited interpretability**: while HRT provides insights into feature importance stability, it may not offer detailed explanations for why certain features are deemed important or how they contribute to model prediction
- **Assumption of echangeability**: HRT assumes that feature values are exchangeable, which may not hold true in all datasets, potentially leading to biased assessments of feature importance
    - The act of shuffling features in isolation may introduce unrealistic data upon which feature importance is calculated

References

(1) Molnar, C. (2022). Interpretable Machine Learning:
A Guide for Making Black Box Models Explainable (2nd ed.). Available online

(2) A Data Odyssey. (2023, March 20). SHAP with Python (Code and Explanations) [Video]. YouTube. https://www.youtube.com/watch?v=L8_sVRhBDLU

(3) Tansey, W., Veitch, V., Zhang, H., Rabadan, R., & Blei, D. M. (2018, November 1). The Holdout randomization test for feature selection in black box models. arXiv.org. Available online

(4) Spector, A., & Janson, L. (2020, November 30). Powerful knockoffs via minimizing reconstructability. arXiv.org. Available online