Explainable Artificial Intelligence (XAI)

Shabab Akhter
External Lecturer
RUC

Contents

1. Background & fundamentals of XAI
    a. What is XAI and why do we need it?
    b. Key concepts such as explainability, interpretability, etc.
    c. Intro to techniques
2. Basic explainability techniques & methods
    a. Explaining statistical models
    b. Visualization techniques, Feature importance, partial dependence, etc.
    c. Exercise: Train statistical model and explain the model
3. Advanced explainability techniques & methods
    a. Explaining advanced deep learning and vision models
    b. Salience maps, SHAP, Lime, etc.
    c. Exercise: Train XGBoost and Neural network and explain the model
4. Real world considerations
    a. Is it really possible to ***explain?!***
    b. Correlation vs Causation

Contents

1. **Background & fundamentals of XAI**
   a. **What is XAI and why do we need it?**
   b. **Key concepts such as explainability, interpretability, etc.**
   c. **Intro to techniques**
2. Basic explainability techniques & methods
   a. Explaining statistical models
   b. Visualization techniques, Feature importance, partial dependence, etc.
   c. Exercise: Train statistical model and explain the model
3. Advanced explainability techniques & methods
   a. Explaining advanced deep learning and vision models
   b. Salience maps, SHAP, Lime, etc.
   c. Exercise: Train XGBoost and Neural network and explain the model
4. Real world considerations
   a. Is it really possible to ***explain?!***
   b. Correlation vs Causation

Introduction to AI

Brief history of Artificial Intelligence (AI):

Late 1950s: solve algebraic problems, prove theorems and learn languages.

1980s and early 1990s: human interaction and activities such as image recognition.

Mid and late 90s: predictive modelling and machine learning (ML).

Past decade: Deep learning & ML solutions for high stakes decisions such as cancer prediction, recidivism, self driving cars, etc.

Key challenge:

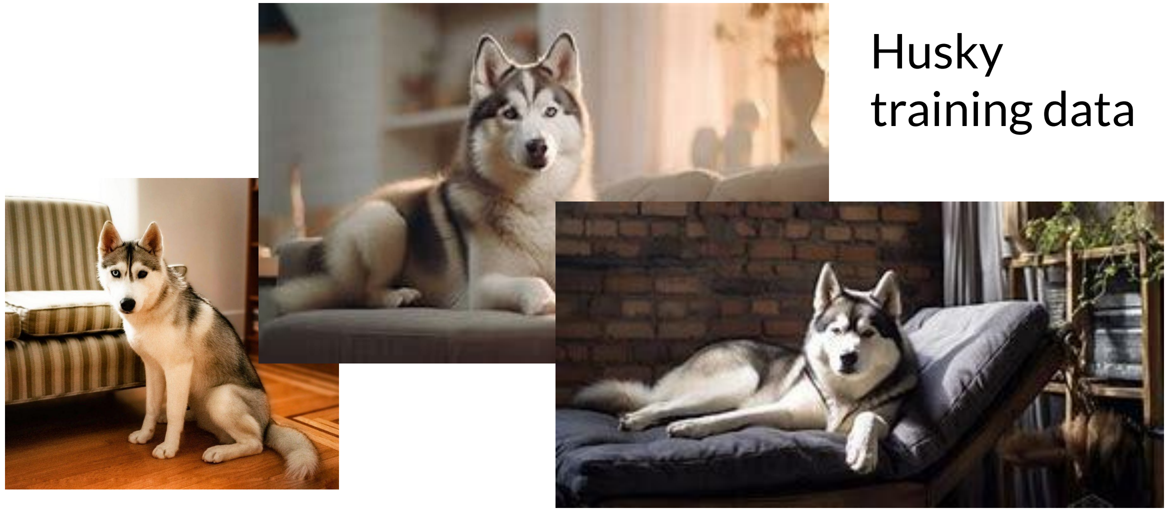
Deep learning methods learn extremely complex patterns. The pattern learned is not intuitive. Hence it is almost impossible to understand why a certain prediction is made.

Examples:

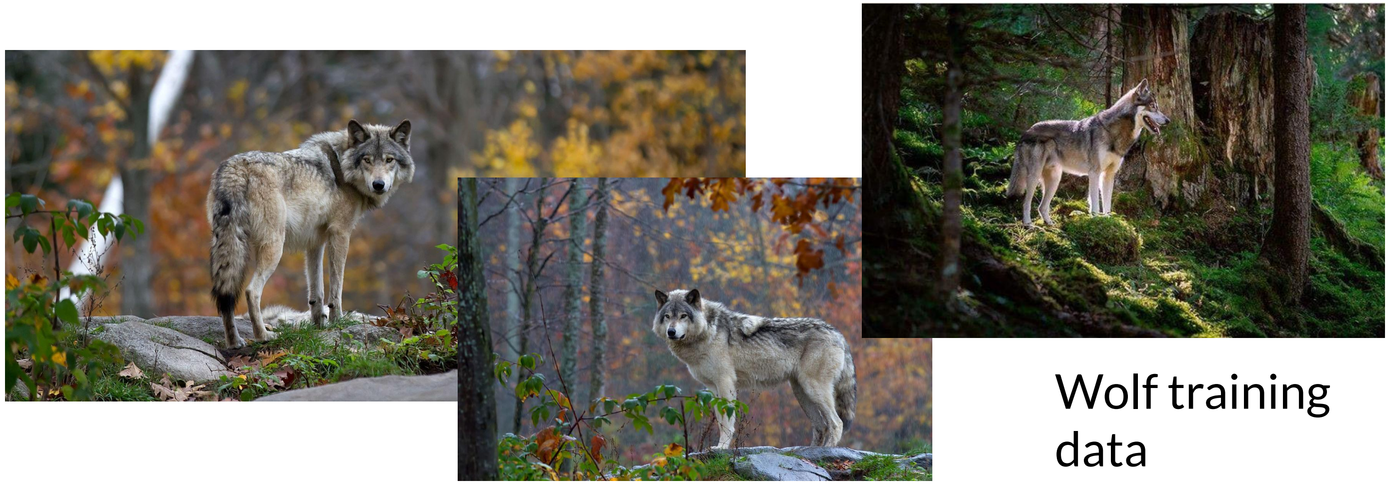
1. Wakabayashi (2018) describe how Uber's self driving car **killed a pedestrian**.
2. Chen (2018) explains how IBM Watson's cancer support AI application was providing **wrong treatment classifications**.
3. The **husky-wolf detector** experiment carried out by Ribeiro et al. (2016)

The husky – wolf experiment

1. Thousands of images of huskies and wolves used to train a classifier
2. Model performed very well on test data
3. But in real life setting, when deployed, it was highly inaccurate



Husky training data



Wolf training data

What do you think went wrong?

Introduction to XAI

Explainable Artificial Intelligence (XAI):

XAI is a broadly defined term but generally, it is understood as *any activity that makes the inner workings of a machine or deep learning model more humanly interpretable.*

The term is relatively new and is gaining attention in the community in recent times as highlighted in the figure.

But the concept has been around since the 80s (maybe even earlier).

Key subfield:

Model explainability: deals with making predictive models more explainabile.

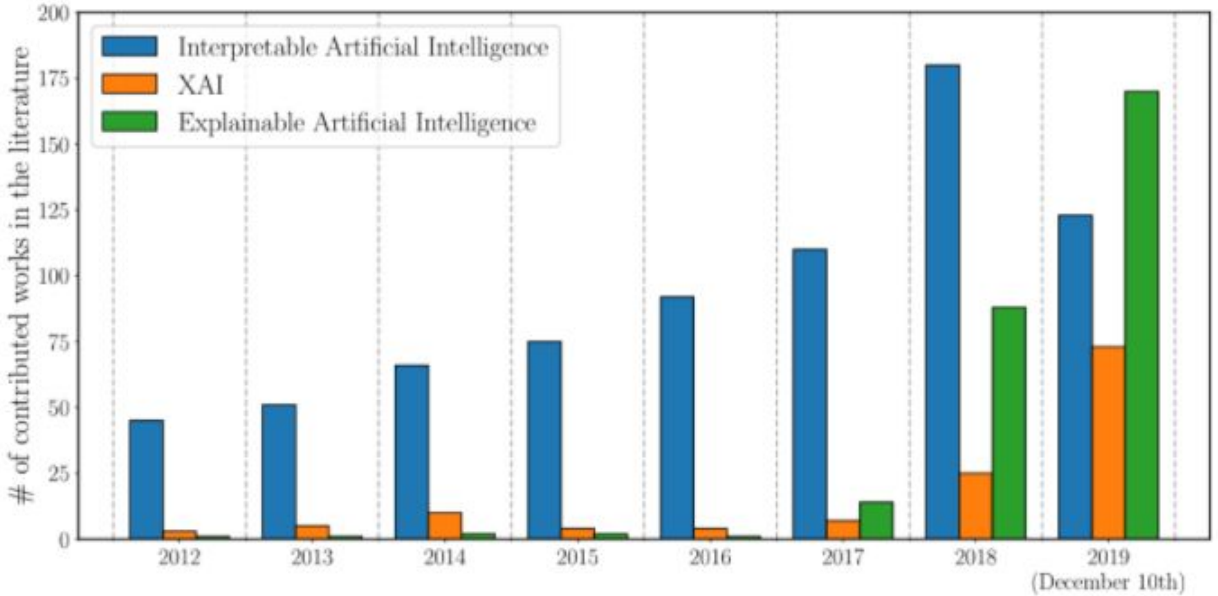


Figure 1: Evolution of the number of total publications whose title, abstract and/or keywords refer to the field of XAI during the last years. Data retrieved from Scopus (December 10th, 2019) by using the search terms indicated in the legend when querying this database. It is interesting to note the latent need for interpretable AI models over time (which conforms to intuition, as interpretability is a requirement in many scenarios), yet it has not been until 2017 when the interest in techniques to explain AI models has permeated throughout the research community. *Figure is by Arrieta et al. (2019).*

Goals of XAI

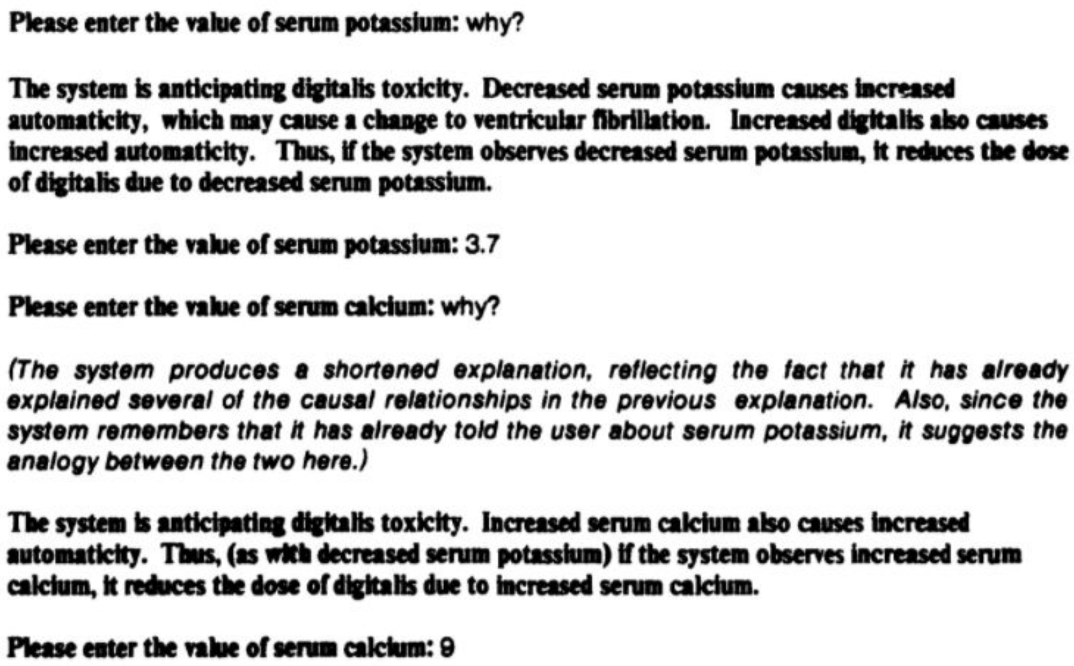
Arrieta et al. (2019) identified several goals:

1. **Trustworthiness**
2. **Causality**
3. **Transferability**
4. **Informativeness**
5. **Confidence**
6. **Fairness**
7. **Accessibility**
8. **Interactivity**
9. **Privacy awareness**

However, the three most important goals as identified by several others such as *Shin and Park (2019)*, *Mittlestadt and Floridi (2017)* and also the *2020 EU White paper on AI* are:

1. **Fairness**: algorithms should not exhibit discriminatory behaviour. Such traits can be learned based on the implicit bias present in the humans creating the solution.
2. **Accountability:** an entity needs to be responsible for the actions of the model. There is currently an accountability loophole.
3. **Transparency**: how clear or open the process is to users. Ideally, a transparent algorithm is one whose inner workings is clear to the end user,

Ways of explaining models (I)



Please enter the value of serum potassium: why?

The system is anticipating digitalis toxicity. Decreased serum potassium causes increased automaticity, which may cause a change to ventricular fibrillation. Increased digitalis also causes increased automaticity. Thus, if the system observes decreased serum potassium, it reduces the dose of digitalis due to decreased serum potassium.

Please enter the value of serum potassium: 3.7

Please enter the value of serum calcium: why?

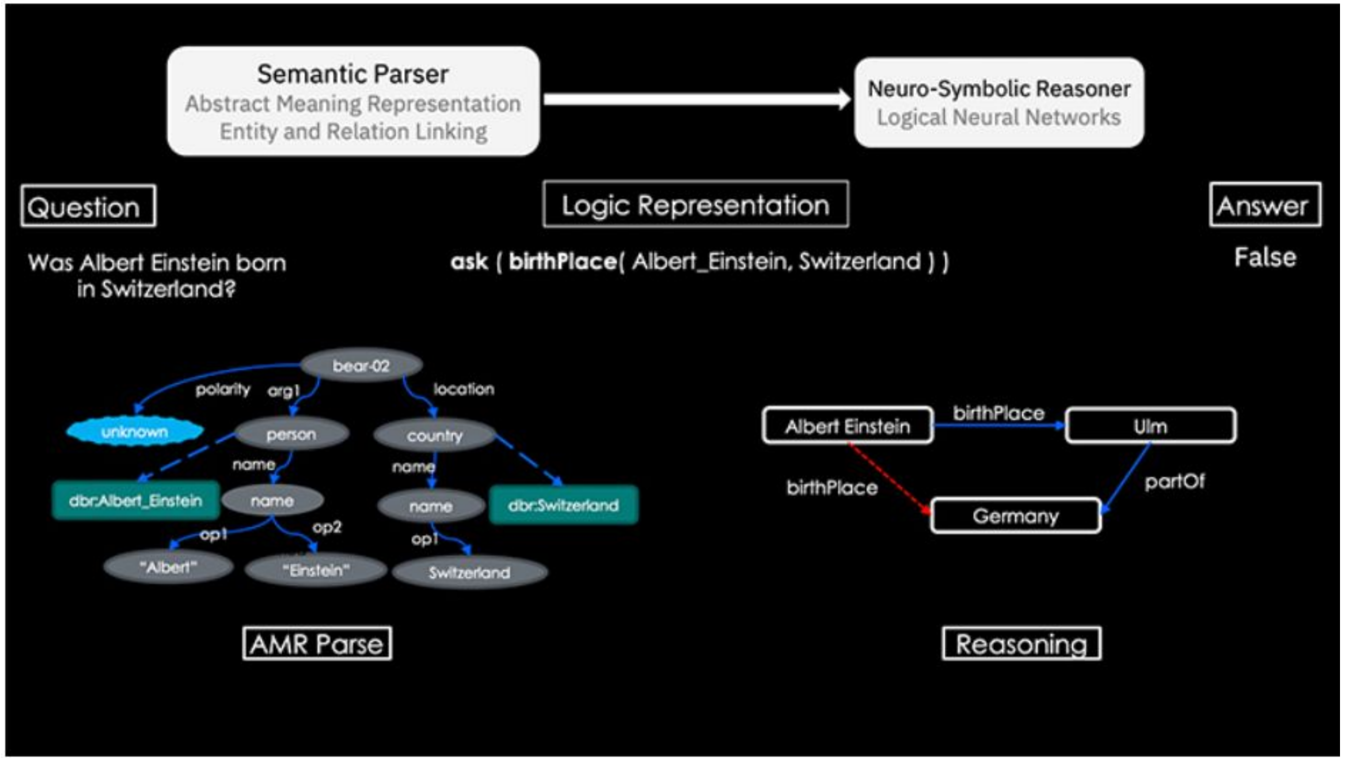
(The system produces a shortened explanation, reflecting the fact that it has already explained several of the causal relationships in the previous explanation. Also, since the system remembers that it has already told the user about serum potassium, it suggests the analogy between the two here.)

The system is anticipating digitalis toxicity. Increased serum calcium also causes increased automaticity. Thus, (as with decreased serum potassium) if the system observes increased serum calcium, it reduces the dose of digitalis due to increased serum calcium.

Please enter the value of serum calcium: 9

XPLAIN: *Swartout (1983)* developed a solution that can justify the actions of a program. They do this by providing explanations for any step in the program, the explanations are provided in English.
US Army event logger: *Van Lent et al. (2004)* presented this stand alone implementation where the actions of the AI player in a combat game is explained using an event logger.

Ways of explaining models (II)



Causal inference:

A causal relationship is of the form: *"if A happens then B will happen"*. In an ML model it can mean: *"the predicted class is true because of the high value of feature A"*.

Ways of explaining models (III)

Several types of methods:

Feature importance: identify how much a certain feature impacts the model outcome.

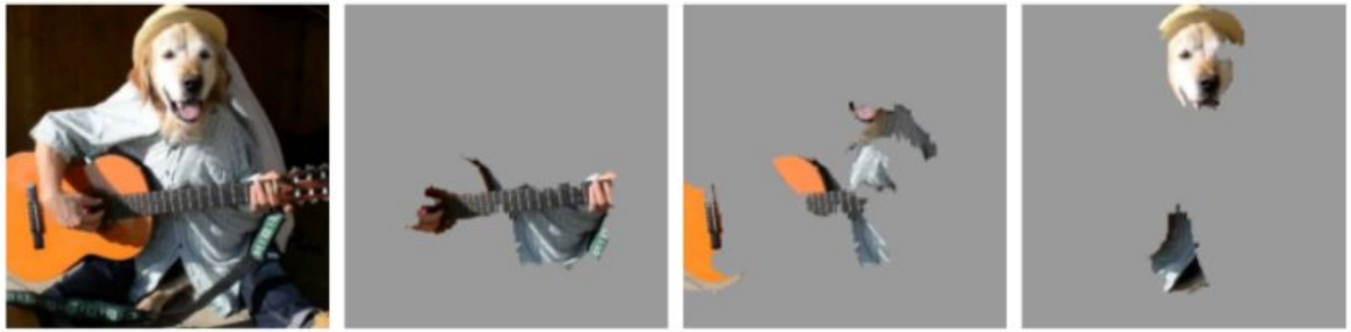
Visual representations: partial dependence plots, correlation matrix, etc.

Text explanations: explain prediction from a model in natural language eg - English.

Saliency maps: identify & highlight pixels that impacted the decision by the model.

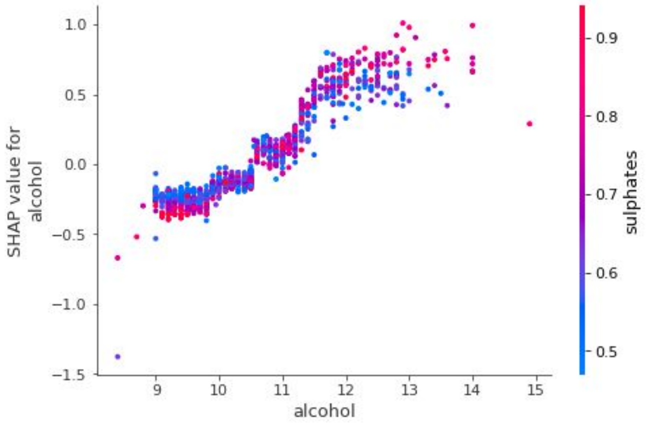
Methods can be:

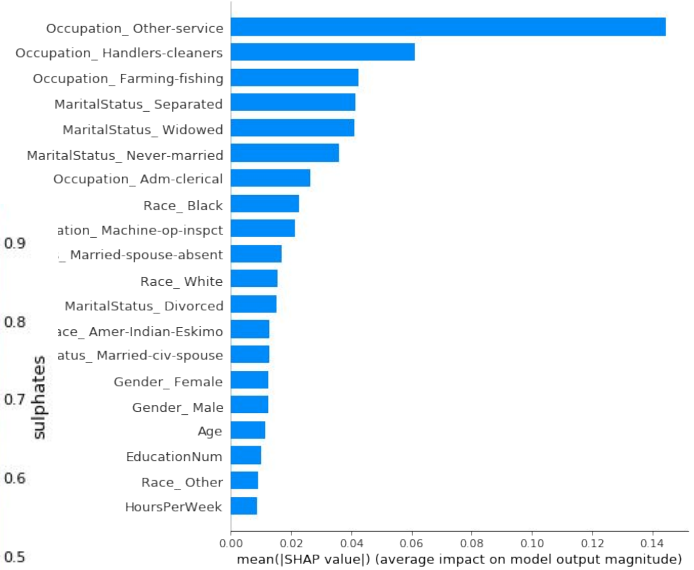
1. **Global:** feature importance of the model.
2. **Local:** feature importance of specific row.
3. **Model agnostic**: method can be applied to several types of models.
4. **Post hoc**: method builds another model on top of model being explained.



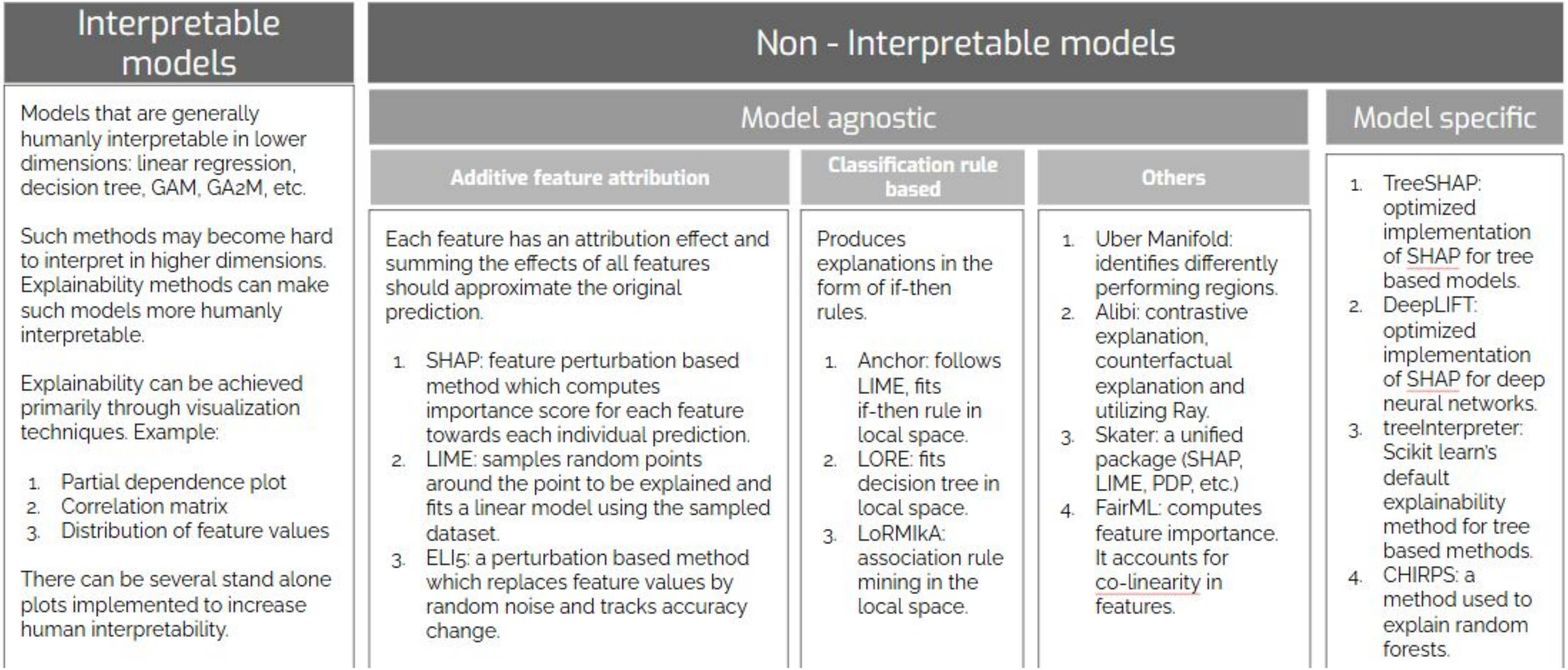
(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

"This instance is classified as Apple because it is red and sweet."





Overview of explainability methods



Interpretable models	Non - Interpretable models			
	Model agnostic			Model specific
	Additive feature attribution	Classification rule based	Others	
Models that are generally humanly interpretable in lower dimensions: linear regression, decision tree, GAM, GA2M, etc. Such methods may become hard to interpret in higher dimensions. Explainability methods can make such models more humanly interpretable. Explainability can be achieved primarily through visualization techniques. Example: 1. Partial dependence plot 2. Correlation matrix 3. Distribution of feature values There can be several stand alone plots implemented to increase human interpretability.	Each feature has an attribution effect and summing the effects of all features should approximate the original prediction. 1. SHAP: feature perturbation based method which computes importance score for each feature towards each individual prediction. 2. LIME: samples random points around the point to be explained and fits a linear model using the sampled dataset. 3. ELI5: a perturbation based method which replaces feature values by random noise and tracks accuracy change.	Produces explanations in the form of if-then rules. 1. Anchor: follows LIME, fits if-then rule in local space. 2. LORE: fits decision tree in local space. 3. LoRMIkA: association rule mining in the local space.	1. Uber Manifold: identifies differently performing regions. 2. Alibi: contrastive explanation, counterfactual explanation and utilizing Ray. 3. Skater: a unified package (SHAP, LIME, PDP, etc.) 4. FairML: computes feature importance. It accounts for co-linearity in features.	1. TreeSHAP: optimized implementation of SHAP for tree based models. 2. DeepLIFT: optimized implementation of SHAP for deep neural networks. 3. treeInterpreter: Scikit learn's default explainability method for tree based methods. 4. CHIRPS: a method used to explain random forests.

Categorization of models

General categorization

Opaque - cannot be understood by anyone, not even the domain expert.

Intelligible - can be understood but some basic level of domain expertise / intellect is needed.

Symbolic - can be understood by almost anyone as the output will include text or visual representations which act like a form of reasoning.

User based categorization

Understood by domain expert - an ML model is understood by the domain expert. The domain expert is someone who has knowledge of the domain in question.

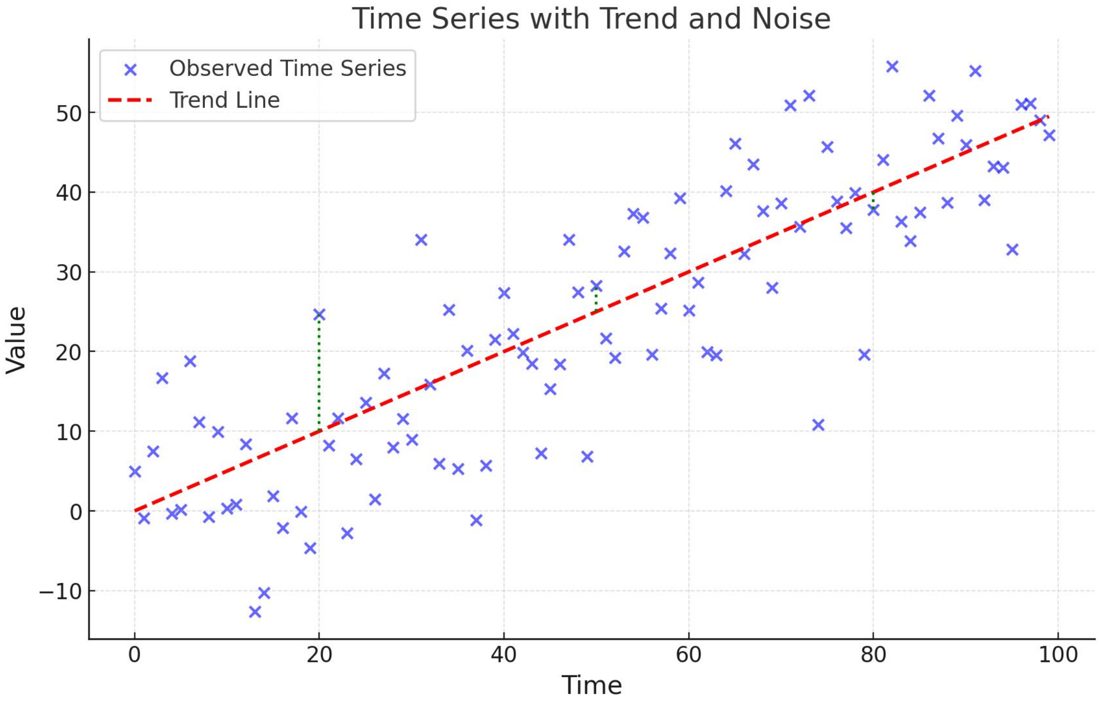
Understood by developer - an ML model is understood by its developer. The developer is someone or a group who are involved in the development phase.

Understood by end user - an ML model is understood by the end user. The end user is the person who will use the model in practice.

Contents

1. Background & fundamentals of XAI
    a. What is XAI and why do we need it?
    b. Key concepts such as explainability, interpretability, etc.
    c. Intro to techniques
2. **Basic explainability techniques & methods**
    a. **Explaining statistical models**
    b. **Visualization techniques, Feature importance, partial dependence, etc.**
    c. **Exercise: Train statistical model and explain the model**
3. Advanced explainability techniques & methods
    a. Explaining advanced deep learning and vision models
    b. Salience maps, SHAP, Lime, etc.
    c. Exercise: Train XGBoost and Neural network and explain the model
4. Real world considerations
    a. Is it really possible to ***explain?!***
    b. Correlation vs Causation

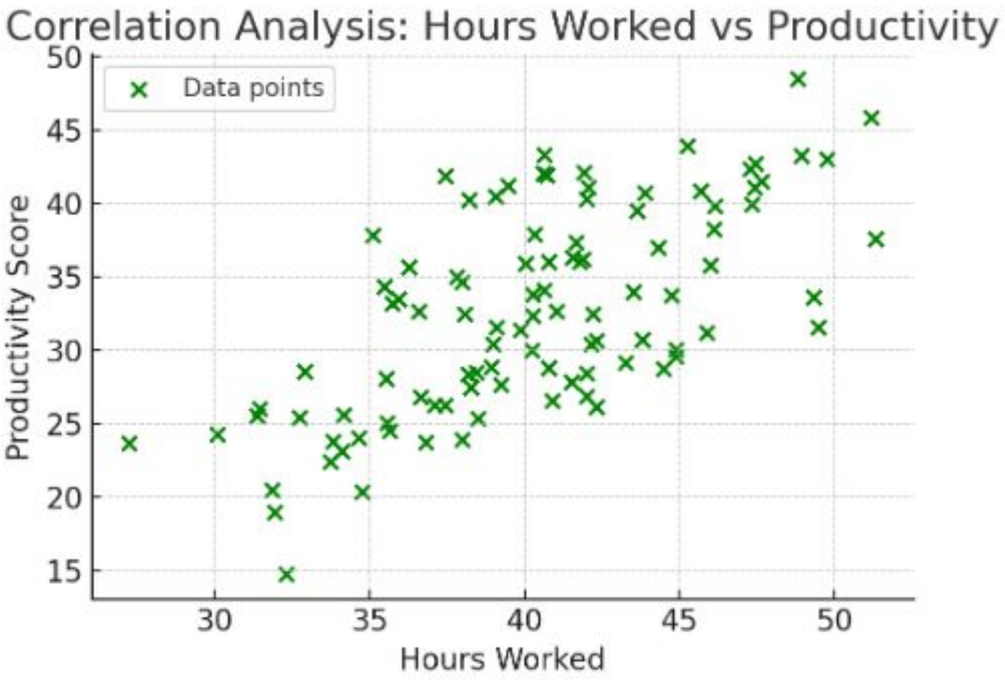
Explaining statistical models



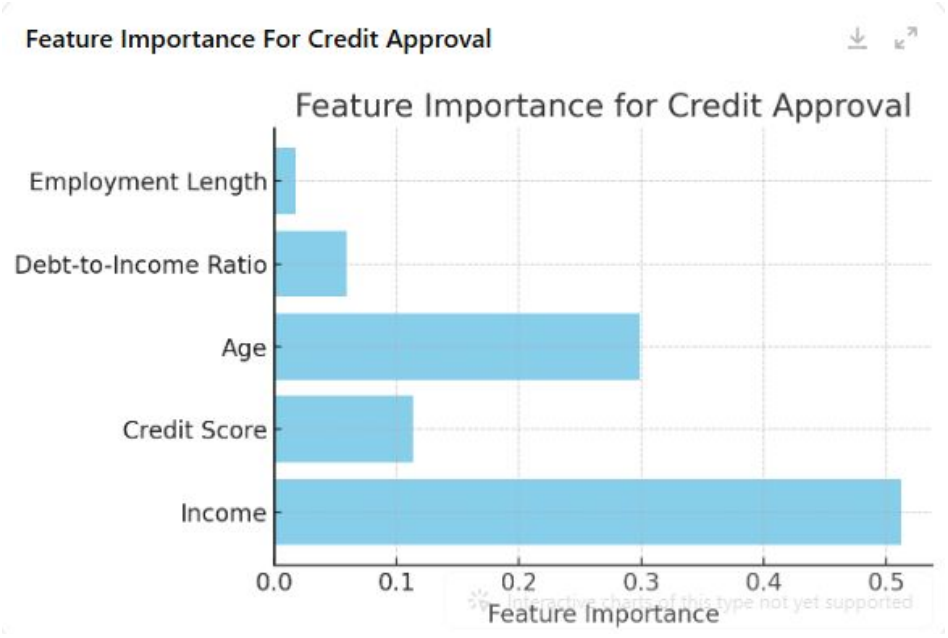
It is simple to explain statistical models:

1. You can look at trends of individuals variables
2. You can do various types of plots to uncover interactions
3. The model follows simpler logic and mathematical formula under the hood so it is easy to keep track of what is happening

How are statistical models explained?

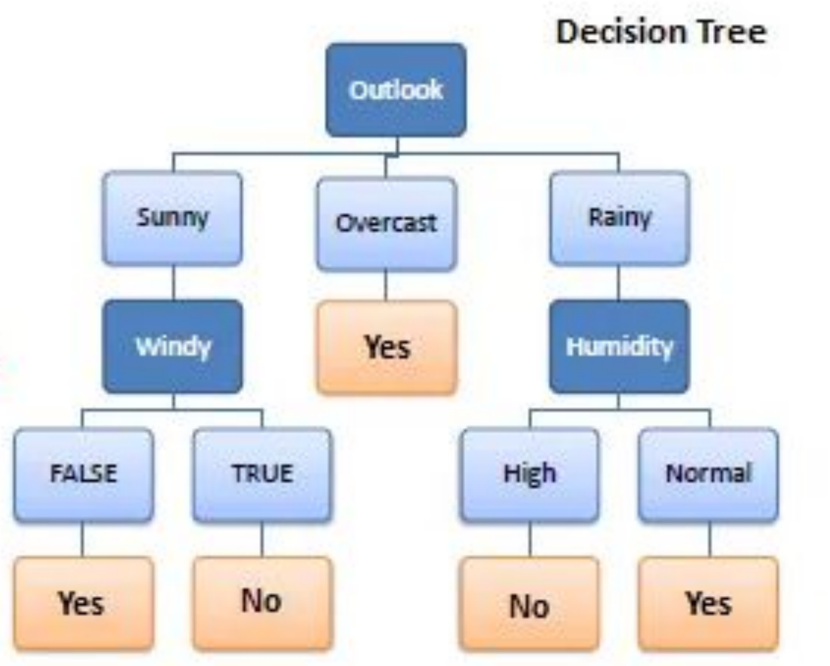


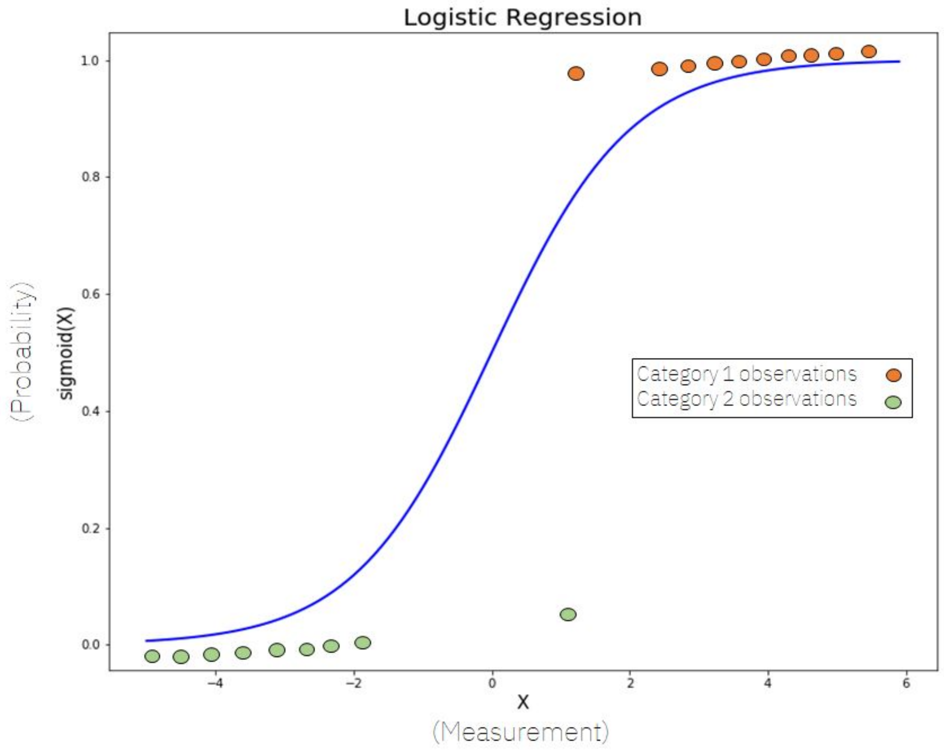
With linear models, you can do trend analysis and correlation analysis to uncover how the model works



For some classification models you can find feature importance

Statistical models are self explanatory





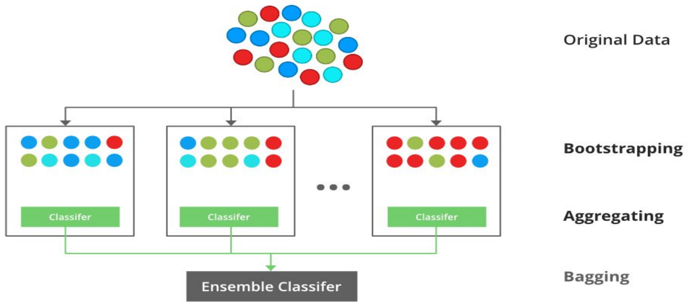
Exercise

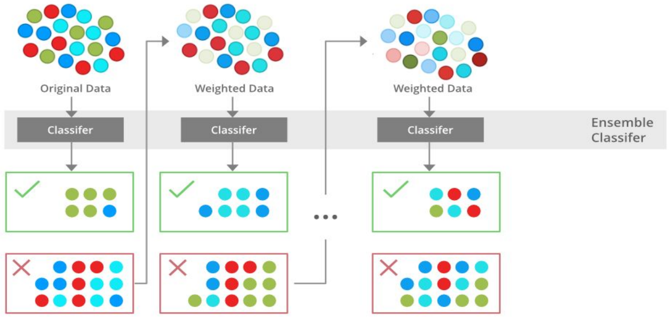
1. Let us see some statistical models and how they can be explained
2. Your turn:
    1. Train at least two basic models on the energy dataset from TSA_Example notebook. You may choose to use only the last 5 years of the dataset to reduce data size
    2. Explain the patterns the model has learned via various plotting capabilities

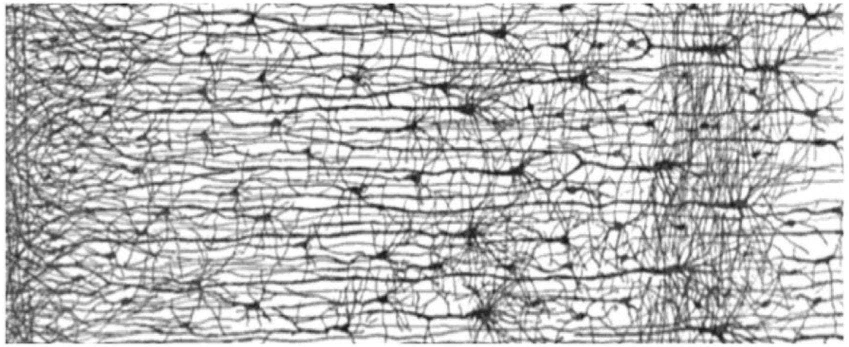
Contents

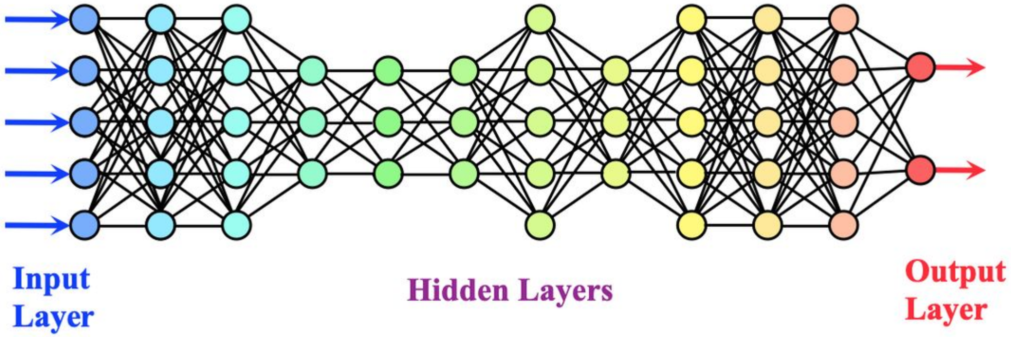
1. Background & fundamentals of XAI
    a. What is XAI and why do we need it?
    b. Key concepts such as explainability, interpretability, etc.
    c. Intro to techniques
2. Basic explainability techniques & methods
    a. Explaining statistical models
    b. Visualization techniques, Feature importance, partial dependence, etc.
    c. Exercise: Train statistical model and explain the model
3. **Advanced explainability techniques & methods**
    a. **Explaining advanced deep learning and vision models**
    b. **Salience maps, SHAP, Lime, etc.**
    c. **Exercise: Train XGBoost and Neural network and explain the model**
4. Real world considerations
    a. Is it really possible to ***explain?!***
    b. Correlation vs Causation

Deep learning models are not easy to understand





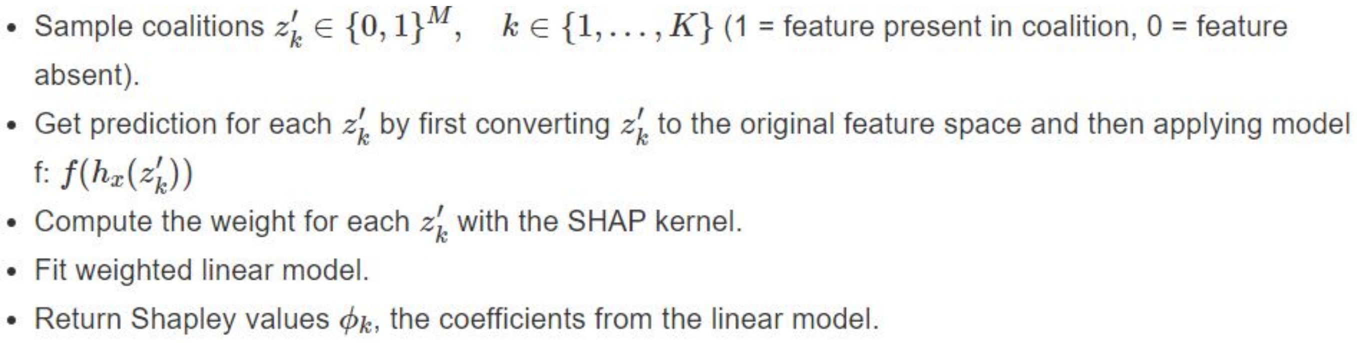




Here we have examples of XGBoost & Neural networks.

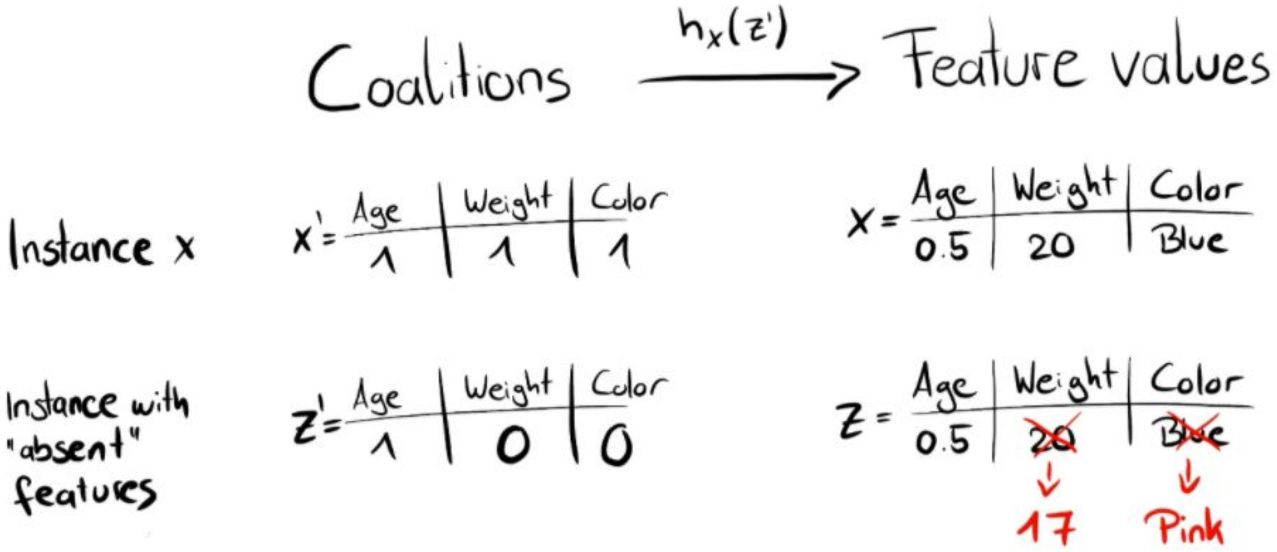
The model architecture is highly complex and not understood by us humans very easily

But there are methods to explain - SHAP



- Sample coalitions $z'_k \in \{0,1\}^M, \quad k \in \{1,\ldots,K\}$ (1 = feature present in coalition, 0 = feature absent).
- Get prediction for each z'_k by first converting z'_k to the original feature space and then applying model f: $f(h_x(z'_k))$
- Compute the weight for each z'_k with the SHAP kernel.
- Fit weighted linear model.
- Return Shapley values ϕ_k, the coefficients from the linear model.

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j$$



$x' =$

Age	Weight	Color
1	1	1

$x =$

Age	Weight	Color
0.5	20	Blue

$z' =$

Age	Weight	Color
1	0	0

$z =$

Age	Weight	Color
0.5	~~20~~ ↓ 17	~~Blue~~ ↓ Pink

Theory behind SHAP

SHAP values originate from the world of Economics and more specifically *Cooperative Game Theory*:

The core idea is to fairly distribute the total payout among players who work together to achieve an outcome. It ensures that each player gets credit for their contribution.

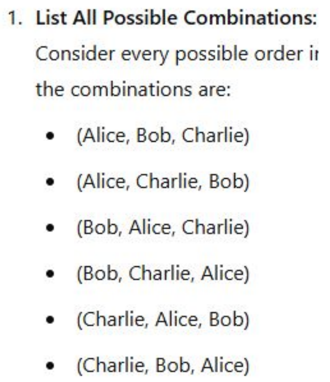
Example:

A Pizza Delivery Team Game 🍕

Imagine three friends—Alice, Bob, and Charlie—team up to deliver pizzas. Together, they earn $300 for a night's work. But not all of them contribute equally:

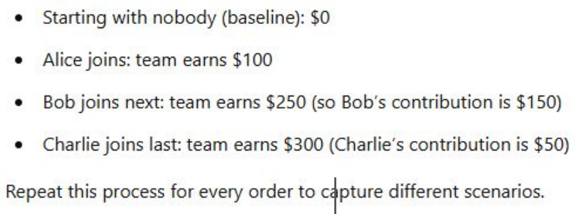
1. Alice has a car, so deliveries are faster.
2. Bob knows the neighborhood well, so he finds the quickest routes.
3. Charlie is great with customers, so they get better tips.The challenge:

How should they split the $300 fairly?



1. **List All Possible Combinations:**
   Consider every possible order ir
   the combinations are:
   - (Alice, Bob, Charlie)
   - (Alice, Charlie, Bob)
   - (Bob, Alice, Charlie)
   - (Bob, Charlie, Alice)
   - (Charlie, Alice, Bob)
   - (Charlie, Bob, Alice)

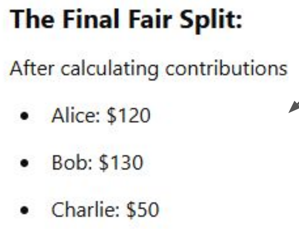
Calculate each players contribution



- Starting with nobody (baseline): $0
- Alice joins: team earns $100
- Bob joins next: team earns $250 (so Bob's contribution is $150)
- Charlie joins last: team earns $300 (Charlie's contribution is $50)

Repeat this process for every order to capture different scenarios.

Average the contributions
Since the order can change, Shapley values average each player's contribution across all scenarios. This ensures fairness, considering both their unique skills and how those skills complement others.

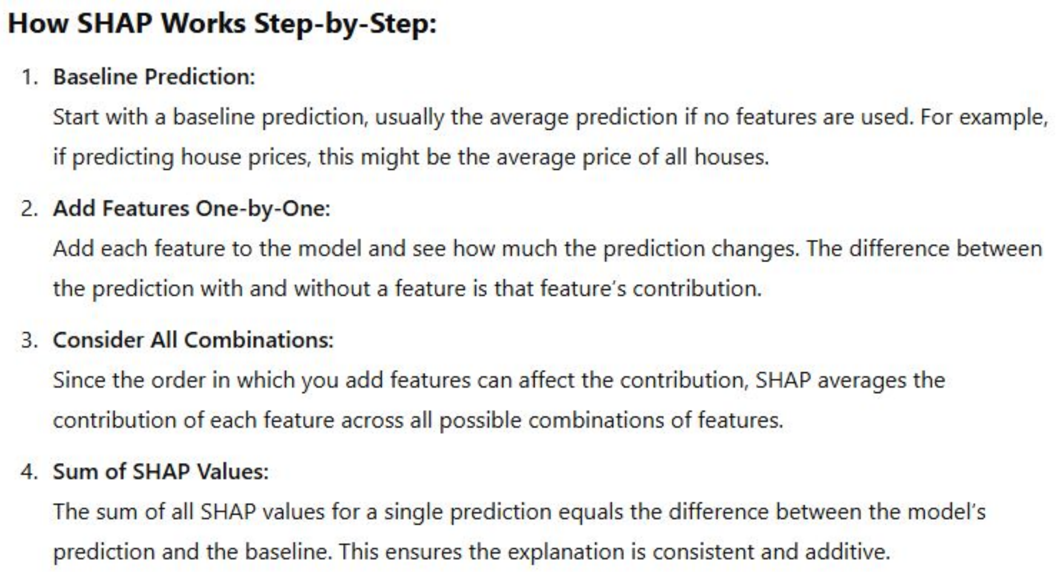


The Final Fair Split:

After calculating contributions

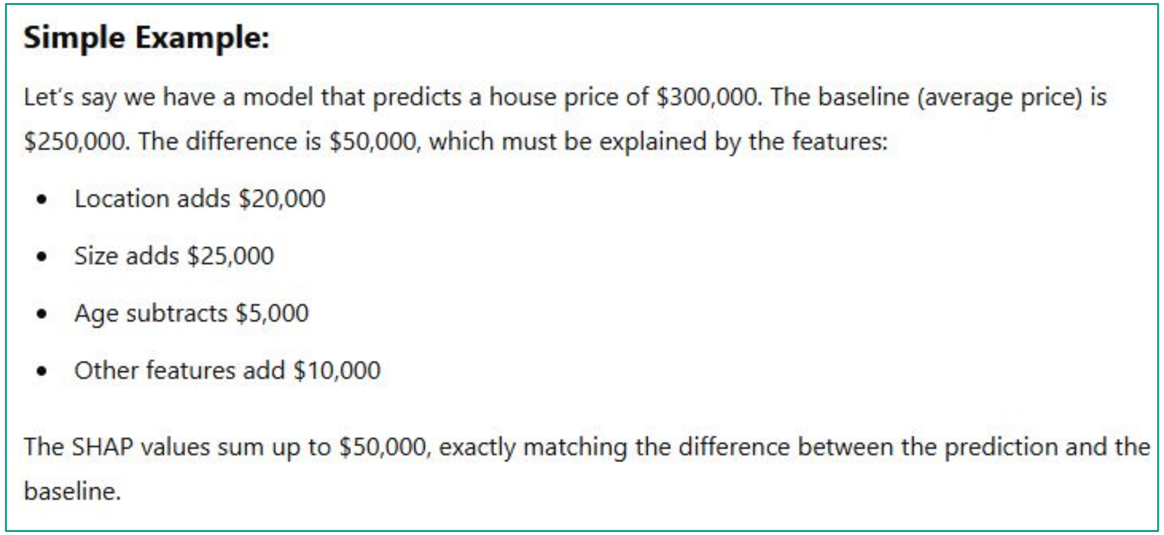
- Alice: $120
- Bob: $130
- Charlie: $50

How does SHAP work in Machine Learning?



How SHAP Works Step-by-Step:

1. **Baseline Prediction:**
   Start with a baseline prediction, usually the average prediction if no features are used. For example, if predicting house prices, this might be the average price of all houses.
2. **Add Features One-by-One:**
   Add each feature to the model and see how much the prediction changes. The difference between the prediction with and without a feature is that feature's contribution.
3. **Consider All Combinations:**
   Since the order in which you add features can affect the contribution, SHAP averages the contribution of each feature across all possible combinations of features.
4. **Sum of SHAP Values:**
   The sum of all SHAP values for a single prediction equals the difference between the model's prediction and the baseline. This ensures the explanation is consistent and additive.

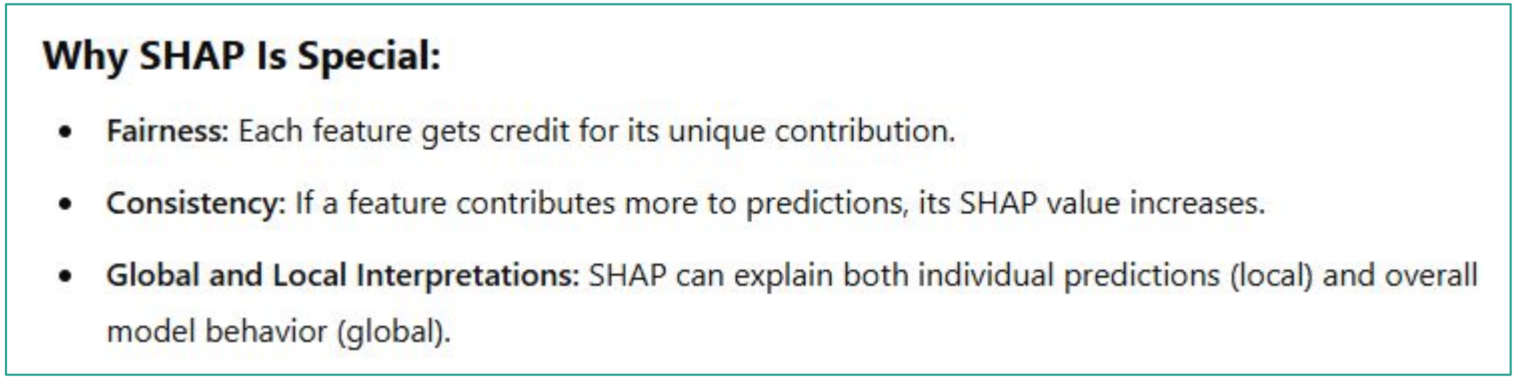


Simple Example:

Let's say we have a model that predicts a house price of $300,000. The baseline (average price) is $250,000. The difference is $50,000, which must be explained by the features:

- Location adds $20,000
- Size adds $25,000
- Age subtracts $5,000
- Other features add $10,000

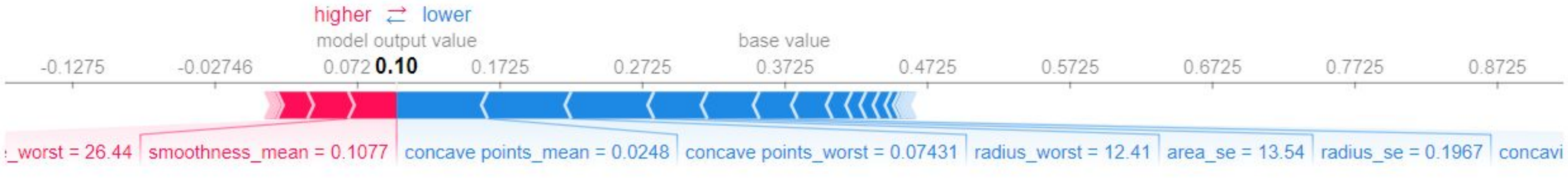
The SHAP values sum up to $50,000, exactly matching the difference between the prediction and the baseline.

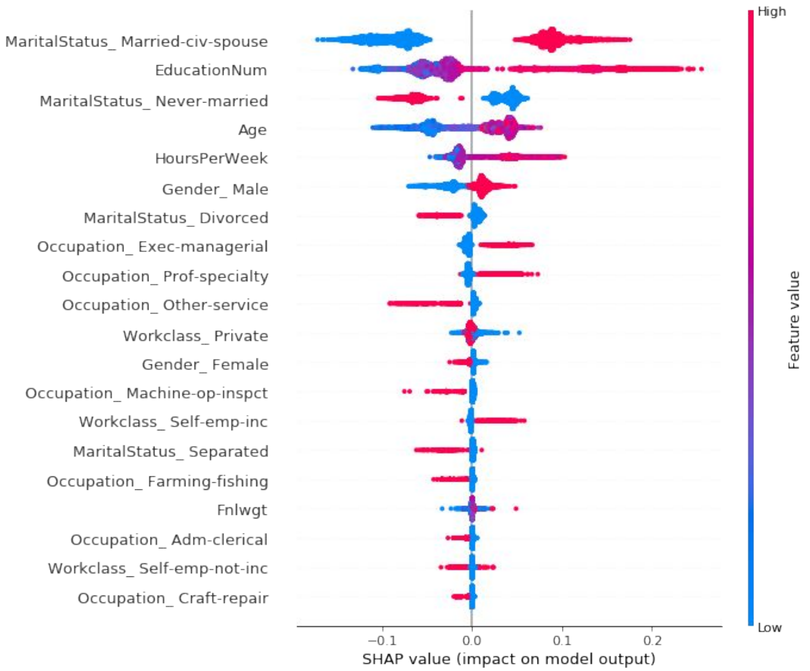


Why SHAP Is Special:

- **Fairness:** Each feature gets credit for its unique contribution.
- **Consistency:** If a feature contributes more to predictions, its SHAP value increases.
- **Global and Local Interpretations:** SHAP can explain both individual predictions (local) and overall model behavior (global).

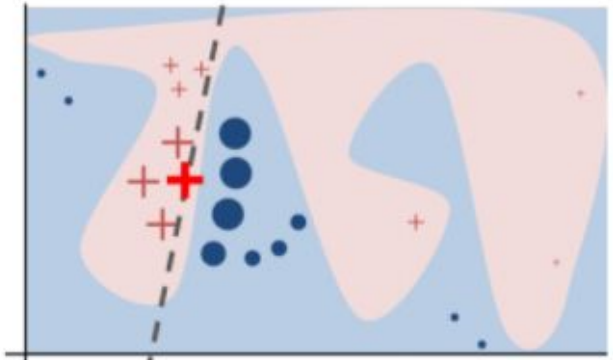
SHAP examples



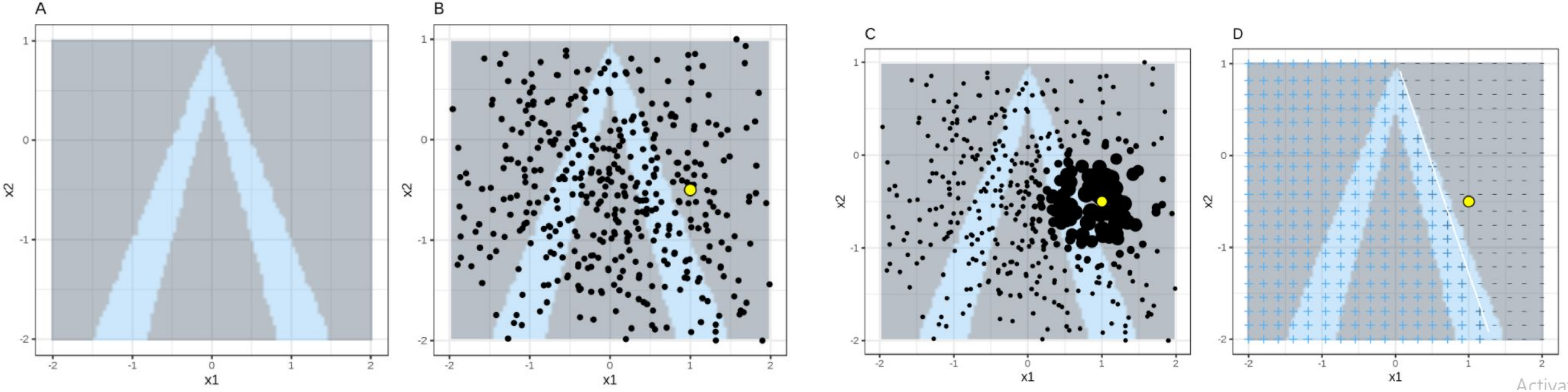


1. Shapley values can be computed for both local and global scenario of a model
2. The top example represent a local situation
3. The example on the left illustrates the overall SHAP values for entire model

Another method - LIME



- Select your instance of interest for which you want to have an explanation of its black box prediction.
- Perturb your dataset and get the black box predictions for these new points.
- Weight the new samples according to their proximity to the instance of interest.
- Train a weighted, interpretable model on the dataset with the variations.
- Explain the prediction by interpreting the local model.



Exercise

1. Let us see some deep learning models and how they can be explained
2. Your turn:
    1. Train at least two advanced models on the energy dataset from TSA_Example notebook. You may choose to use only the last 5 years of the dataset to reduce data size
    2. Explain the patterns the model has learned using SHAP local and global plots
    3. Explain the patterns using LIME
    4. Compare the answers from LIME & SHAP (global & some local) – what do you see? Is it the same?

Contents

1. Background & fundamentals of XAI
   a. What is XAI and why do we need it?
   b. Key concepts such as explainability, interpretability, etc.
   c. Intro to techniques
2. Basic explainability techniques & methods
   a. Explaining statistical models
   b. Visualization techniques, Feature importance, partial dependence, etc.
   c. Exercise: Train statistical model and explain the model
3. Advanced explainability techniques & methods
   a. Explaining advanced deep learning and vision models
   b. Salience maps, SHAP, Lime, etc.
   c. Exercise: Train XGBoost and Neural network and explain the model
4. **Real world considerations**
   a. **Is it really possible to *explain?!***
   b. **Correlation vs Causation**

Is it possible to explain?

Let’s use an analogy to try and answer this question:

1. Imagine a three friends, Alice, Bob & Charlie go camping in the woods during summer
2. Alice says that it would be fun to have a bonfire. Charlie agrees with Alice and sets up the bonfire in an attempt to impress her
3. While things started well, after about 20 minutes, somehow the bonfire spirals out of control and sets fire to some trees
4. Within moments, there is a wildfire, and the entire forest is burning down

How do you explain this fire and who is responsible?

Is it Alice for wanting the bonfire?

Is it Charlie for actually creating the bonfire?

Correlation vs Causation

Example 1:
Ice Cream & Drowning Correlation:
Data might show that ice cream sales and drowning incidents both increase in the summer.

Does Ice Cream Cause Drowning? No!

The actual cause is hot weather—more people swim, and more people buy ice cream.

Example 2:
Number of Churches & Crime Rate in Cities:
Data might show that cities with more churches tend to have higher crime rates.

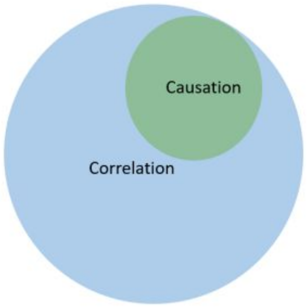
Does Having More Churches Cause Crime? No!

The real reason is that larger cities have more of everything—more churches, more people, and more crime.

Example 3:
Shoe size & intelligence in kids:
Data might show that kids with bigger shoe sizes tend to score better on IQ tests.

Does Bigger Feet Make Kids Smarter? No!

Older kids naturally have bigger feet and also tend to do better on IQ tests.



Correlation vs. Causation: What's the Difference?

Correlation means two things happen together, but one does not necessarily cause the other. Causation means one thing directly causes the other.

→

Lesson: Just because two things rise together doesn’t mean one causes the other.

To summarize, we cannot truly uncover causal relationships in machine learning, we only uncover correlations

Wrap up

1. Background & fundamentals of XAI
   a. What is XAI and why do we need it?
   b. Key concepts such as explainability, interpretability, etc.
   c. Intro to techniques
2. Basic explainability techniques & methods
   a. Explaining statistical models
   b. Visualization techniques, Feature importance, partial dependence, etc.
   c. Exercise: Train statistical model and explain the model
3. Advanced explainability techniques & methods
   a. Explaining advanced deep learning and vision models
   b. Salience maps, SHAP, Lime, etc.
   c. Exercise: Train XGBoost and Neural network and explain the model
4. Real world considerations
   a. Is it really possible to ***explain?!***
   b. Correlation vs Causation