What is Explainable AI (XAI) and why should you care?

What, Why, For Whom, and How of Explainability

Naseeb Shaik

Mounika Vanka

Rajanie Prabha

Agenda

What is XAI?

Examples of using XAI

Why is XAI needed?

Goals of XAI

XAI Techniques - Before During and After Modelling Explainability

Challenges with XAI

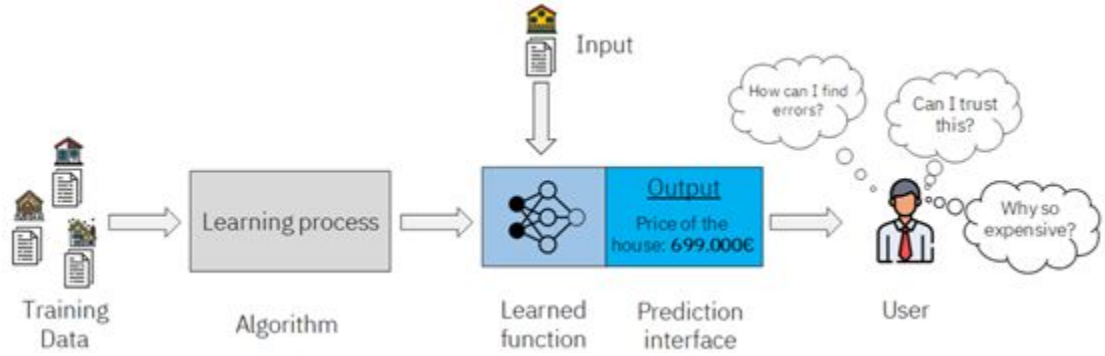
Explainable AI in the Industry

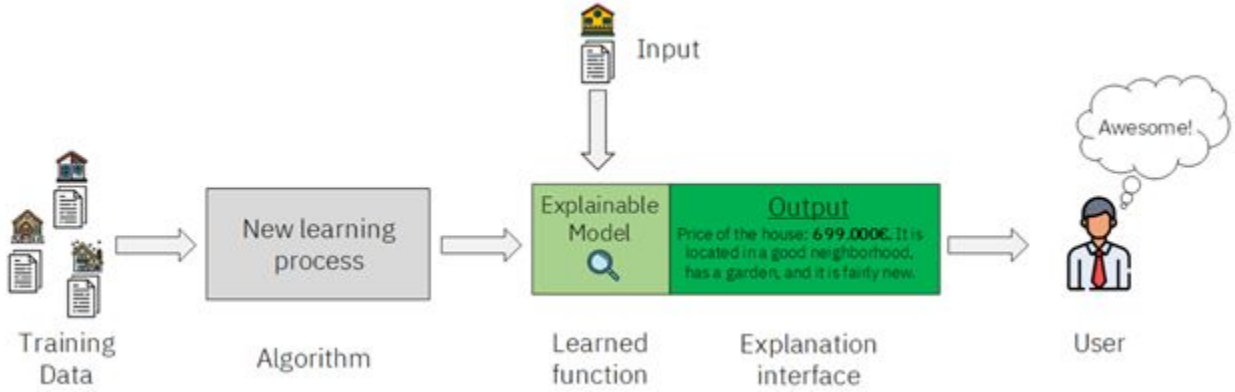
Workshop Activity

Q & A

What ? Defining explainable AI

- Explainable AI refers to the concept of how AI works and how it arrives at those decisions being made clear to humans.
- It is concerned with explaining input variables and the decision-making stages of a model
- Enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners.


Input
How can I find errors?
Can I trust this?
Why so expensive?
Learning process
Output
Price of the house: 699.000€
Training Data
Algorithm
Learned function
Prediction interface
User


Input
Awesome!
New learning process
Explainable Model
Output
Price of the house: 699.000€. It is located in a good neighborhood, has a garden, and it is fairly new.
Training Data
Algorithm
Learned function
Explanation interface
User

Why ? Need for Explainable AI

1. Bias in decision support systems like COMPAS

   COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) : It is a decision support tool that assesses the sentencing and parole of convicts

	White	African American
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

(https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing)

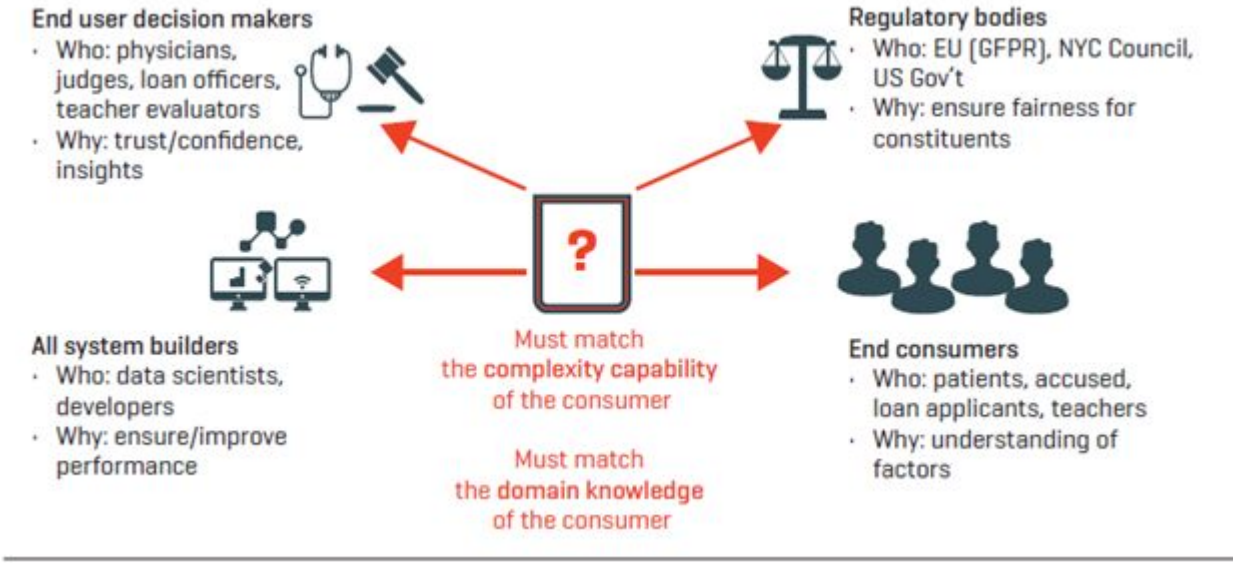
2. Bias in Mortgage charges

A new University of California, Berkeley study has found that both online and face-to-face lenders charge higher interest rates to African American and Latino borrowers than White and Asian

- Black and Latino borrowers pay 5.6 to 8.6 basis points higher interest on purchase loans than White and Asian ethnicity borrowers do, and 3 basis points more on refinance loans.
- For borrowers, these disparities cost them $250M to $500M annually.
- For lenders, this amounts to 11 percent to 17 percent higher profits on purchase loans to minorities, based on the industry average 50-basis-point profit on loan issuance.

(https://news.berkeley.edu/story_jump/mortgage-algorithms-perpetuate-racial-bias-in-lending-study-finds/)

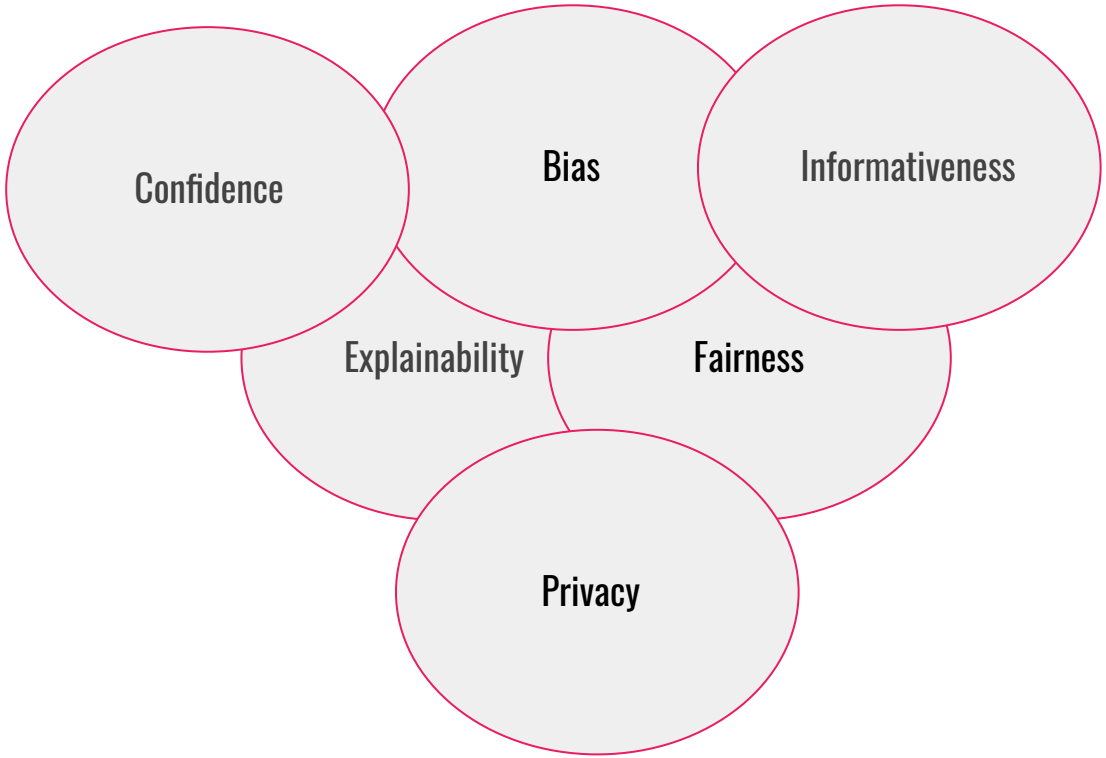
For Whom? An Explanation for Whom?



What For? Goals of XAI

- Trustworthiness
- Transferability
- Informativeness
- Confidence
- Fairness
- Accessibility
- Interactivity
- Privacy awareness

Goals of XAI - Interlinked



XAI Techniques

Before During and Post-hoc modelling

XAI Techniques - Before, During and After

01	**Pre-modelling Explainability**	• Exploratory data analysis • Dataset summarisation • Dataset description standardization • Explainable Feature Engineering
02	**Explainable modelling/ During modelling**	• Regularization • Adopting explainable models • Joint prediction and explanation • Adopting hybrid models
03	**Post-modelling Explainability**	• Proxy models • Activation Optimization • Backward propagation

XAI Techniques - Before Model Deployment

1. **Exploratory Data Analysis:**

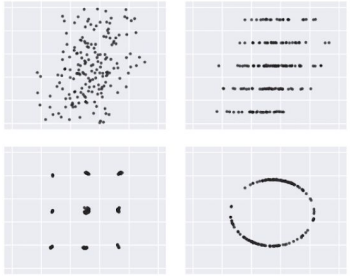
   a.Extracting statistical information like S.D, mean from dataset.

   b.Methods to visualize higher dimensional data:

   1. Parallel Coordinate Plots

   2. PCA

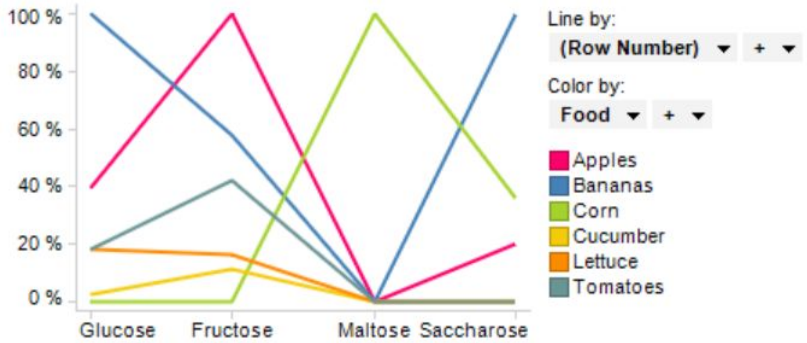
   3. UMAP



Visualization Example : Parallel Coordinate Plot

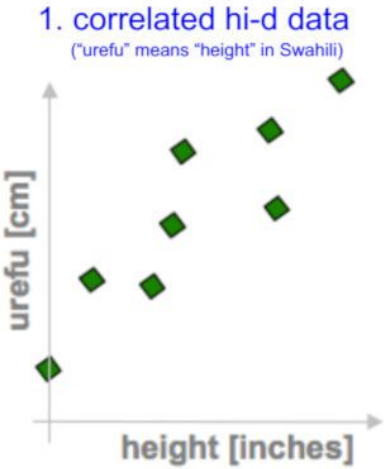
Each row corresponds to a line plot in the graph and the values are all normalized to allow for easy comparison.

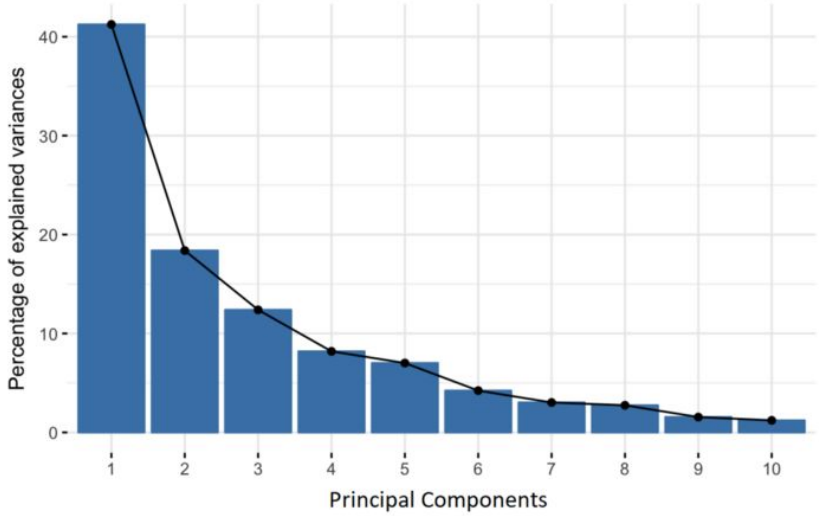
Food	Glucose	Fructose	Maltose	Saccharose
Apples	2.10	4.50	0.00	1.30
Bananas	4.40	2.70	0.00	6.40
Corn	0.60	0.20	0.30	2.30
Cucumber	0.70	0.70	0.00	0.00
Lettuce	1.30	0.90	0.00	0.00
Tomatoes	1.30	2.00	0.00	0.00

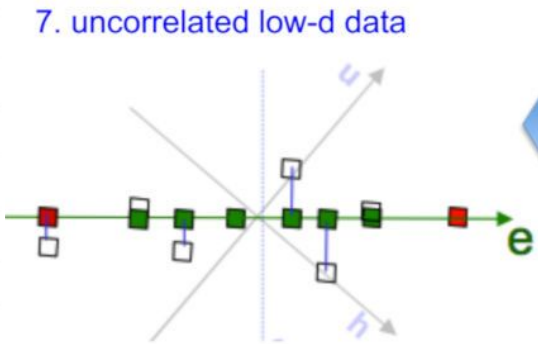


Visualization Example: PCA (Principal Component Analysis)

All datasets have a lot of variables (millions, sometimes), it becomes incomprehensible to understand which variables are the most important. PCA comes to rescue here.







XAI Techniques - Before Model Deployment

2. Dataset Description Standardisation:

Proper communication between the dataset owners and the dataset users.

Dataset Nutrition Label framework as a standard for providing a distilled yet comprehensive overview of datasets.

Module Name	Description	Contents
Metadata	Meta information. This module is the only required module. It represents the absolute minimum information to be presented	Filename, file format, URL, domain, keywords, type, dataset size, % of missing cells, license, release date, collection range, description
Provenance	Information regarding the origin and lineage of the dataset	Source and author contact information with version history
Variables	Descriptions of each variable (column) in the dataset	Textual descriptions
Statistics	Simple statistics for all variables, in addition to stratifications into ordinal, nominal, continuous, and discrete	Least/most frequent entries, min/max, median, mean,.etc
Pair Plots	Distributions and linear correlations between 2 chosen variables	Histograms and heatmaps
Probabilistic Model	Synthetic data generated using distribution hypotheses from which the data was drawn - leverages a probabilistic programming backend	Histograms and other statistical plots
Ground Truth Correlations	Linear correlations between a chosen variable in the dataset and variables from other datasets considered to be "ground truth", such as Census Data	Heatmaps

XAI Techniques - Before Model Deployment

3. Dataset Summarisation:

Clearly explain subsets/common patterns in very large datasets. Provides a minimal subset in a large dataset that is representative of the whole dataset.

4. Explainable Feature Engineering:

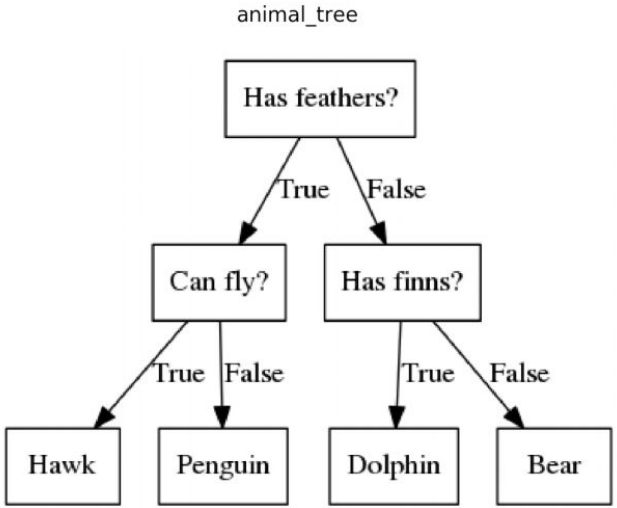
Mathematical models to uncover underlying structure of a dataset.

Example: An expert identifies the features in the dataset and explains them.

XAI Techniques - During Model Deployment

1. Explainable modelling:

Linear Regression, **decision trees**, rule sets, generalized additive models, case-based reasoning methods.



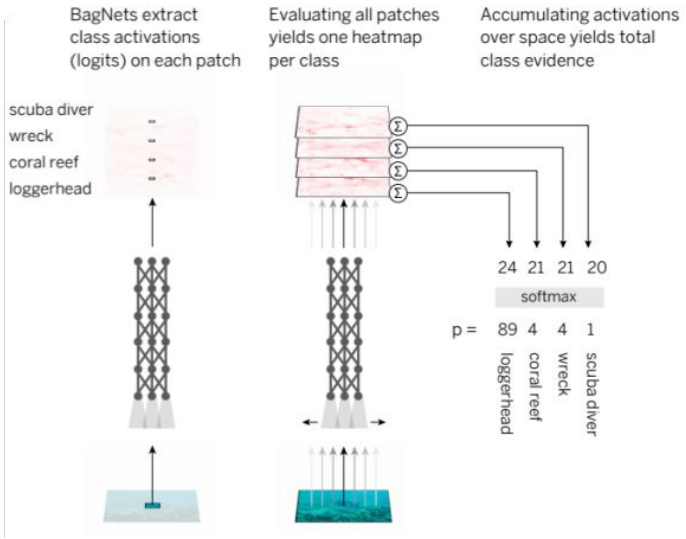
XAI Techniques - During Model Deployment

2. Hybrid models:

    deep k-Nearest Neighbour (requires storing the hidden representation of entire dataset), Self-Explaining Neural Network, Conceptual Explanation Networks, **BagNets**.

Example - During Model Deployment

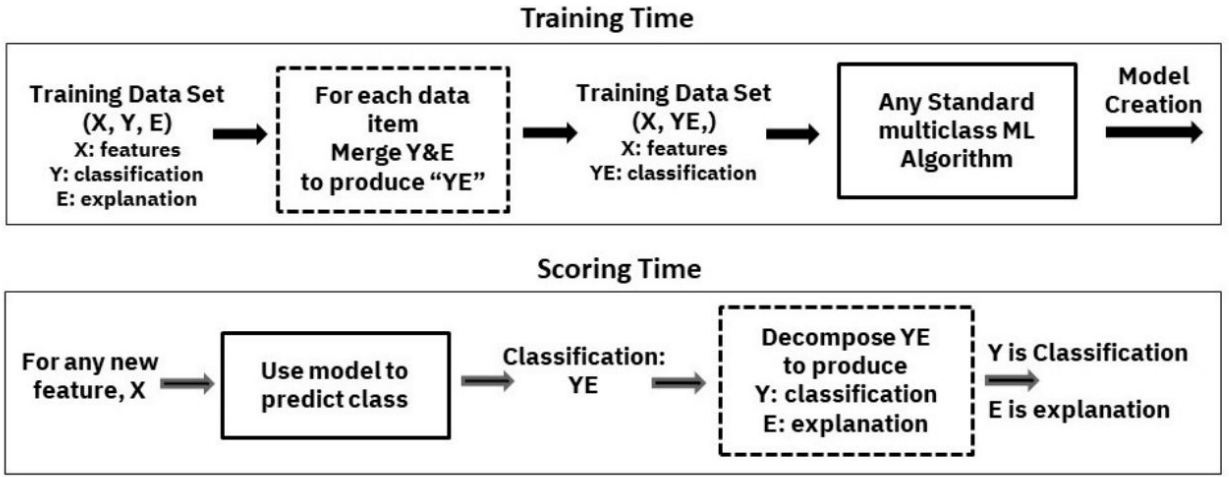
BagNets



Example - During Model Deployment

3. Joint Prediction and Explainability

AI model that can provide both an explanation and also a prediction.

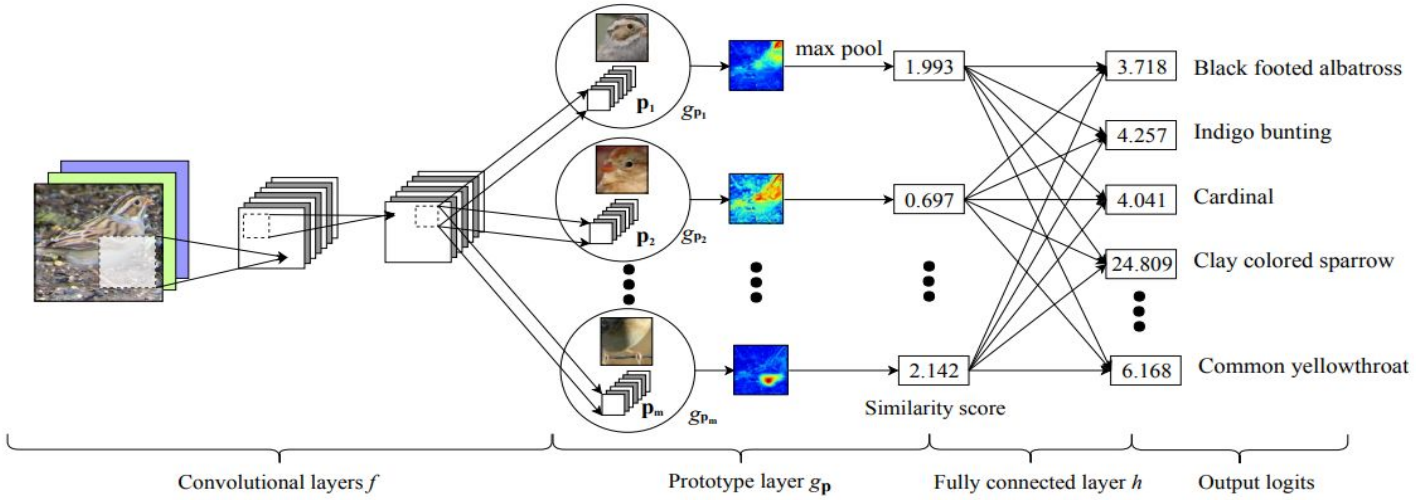


TED (Teaching Explanations for Decisions)

XAI Techniques - During Model Deployment

4. Explainability through Architectural Adjustments

"This looks like that" -Addition of a layer that looks for specific features in the image to help explain the model.



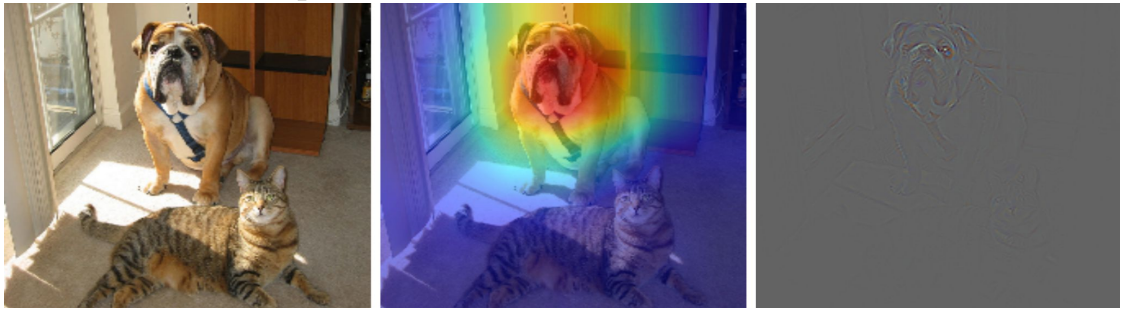
XAI Techniques - During Model Deployment

5. Explainability through Regularization:

Using a simple decision tree, by adding a novel regularization feature to the model, **Saliency Heatmaps.**

Examples - During Model Deployment

Saliency Heatmaps



Example: Shapely additive Explanation.

1. Estimating the importance of each input feature using saliency scores
2. Using decision trees, dependency plots.

Note: These can only be used for simple classification problems, not for complex models.

Drawbacks - During Model Deployment

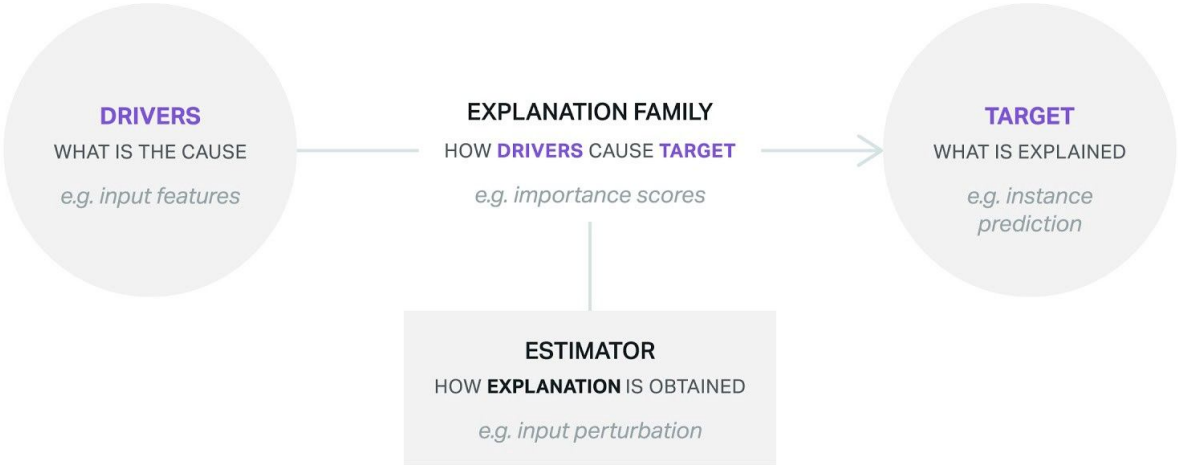
Drawbacks:

1. Assume availability of explanations in training dataset.

2. Explanations not reflective of model behaviour but rather human perception.

XAI Techniques - Before, During and After

Post HOC explainability:



XAI Techniques - Before, During and After

Examples of Target:

1. Predictions / output

Depends on the purpose of the models for the end users.

Examples of Drivers:

1. Training Samples

2. Hyperparameter settings

3. Choice optimization algorithm

XAI Techniques - Before, During and After

Explanation Family:

The type of explanation chosen to help explain the decisions of the model.

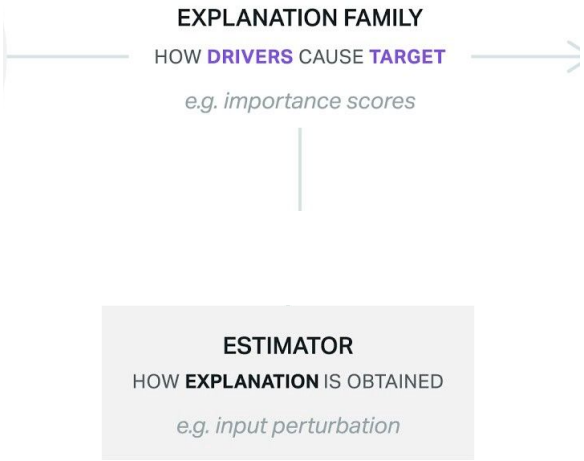
These explanations should be as faithful as possible to the complexity of the model.

Example, Decision Trees, decision scores, etc.

Estimator:

This involves passing a set of values to the model to look at the output.

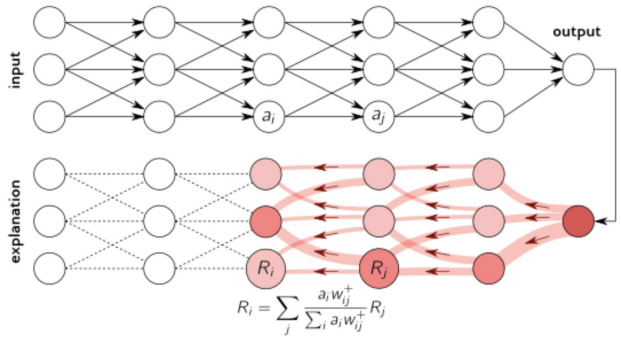
Example: Backward Propagation



Example - Before, During and After

Backward Propagation:

Contribution of all the input layers are obtained, by calculating the contribution of the neurons in the layer previous to the target.



CHALLENGES with XAI

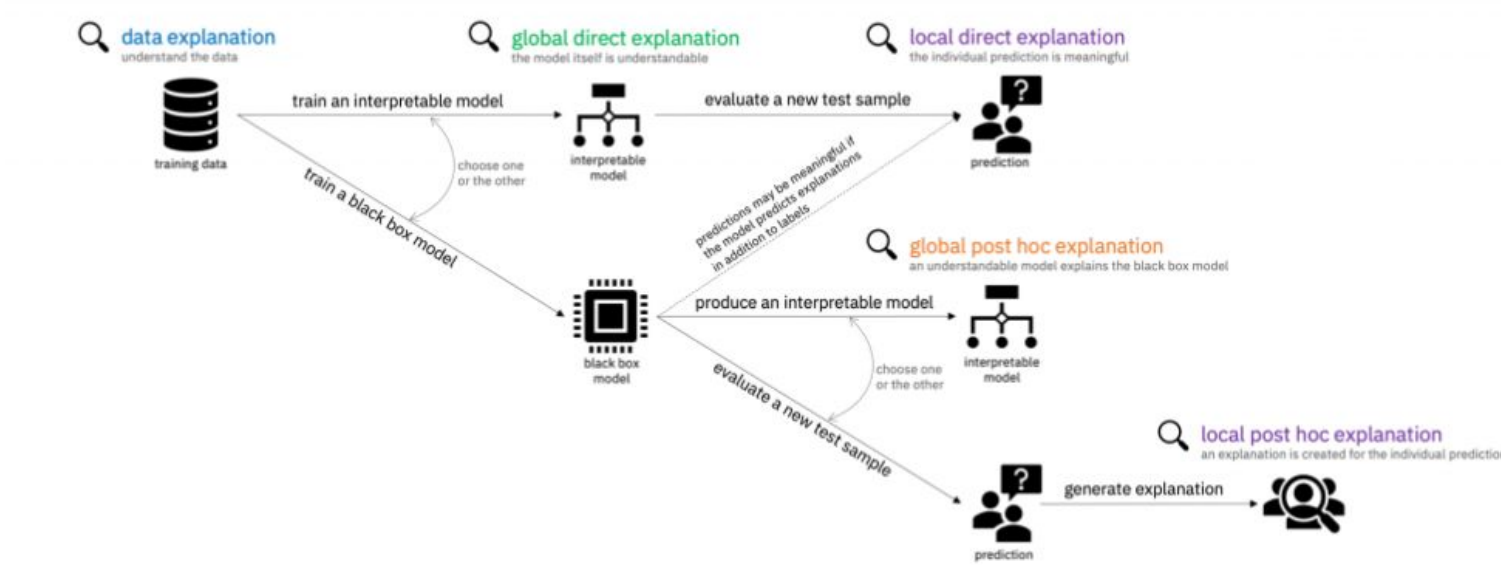
1. Confidentiality

2. Complexity of the algorithms

3. Unavailability of a quantitative measurement of the visualization maps

Explainable AI in the industry

IBM



AI Explainability 360 Usage Diagram

https://www.ibm.com/blogs/research/2019/08/ai-explainability-360/

Explainable AI in the industry

Google Facets:

https://www.ibm.com/blogs/research/2019/08/ai-explainability-360/

Accenture AI

https://www.accenture.com/us-en/services/digital/what-artificial-intelligence-really?c=acn_glb_brandexpressiongoogle_12192026&n=psgs_0521&gclid=Cj0KCQjwraqHBhDsARIsAKuGZeFRG94GoIgbRWlON9Mgoq-jUh7JYV6UifsqscLl-af6sqYHlNomLgwaAs5JEALw_wcB&gclsrc=aw.ds

DARPA

Colab workbook

– – –

https://colab.research.google.com/drive/1TyX1QRM7-ab3HjSwc2OWji7jyUEhXa0C?usp=sharing

Image Classification: Flower Dataset

5 categories: Daisy, Tulips, Dandelions, Roses, Sunflowers

Resources

1. Alejandro Barredo Arrieta, Natalia D´ıaz-Rodr´ıguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjaminsh, Raja Chatila, and Francisco Herrera, *Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI*, Dec 2019 arXiv:1910.10045v2 [cs.AI] 26 Dec 2019.
2. https://www.youtube.com/watch?v=hTi7ZW6rlM0&ab_channel=GoogleCloudTech
3. https://cloud.google.com/ai-platform/prediction/docs/ai-explanations/overview
4. https://www.youtube.com/watch?v=6Mqh-JeKHR8&ab_channel=BerilSirmacek
5. https://blog.usejournal.com/explainability-of-ai-the-challenges-and-possible-workarounds-14d8389d2515
6. https://towardsdatascience.com/what-explainable-ai-fails-to-explain-and-how-we-fix-that-1e35e37bee07
7. https://towardsai.net/p/programming/decision-trees-explained-with-a-practical-example-fe47872d3b53
8. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
9. https://news.berkeley.edu/story_jump/mortgage-algorithms-perpetuate-racial-bias-in-lending-study-finds/