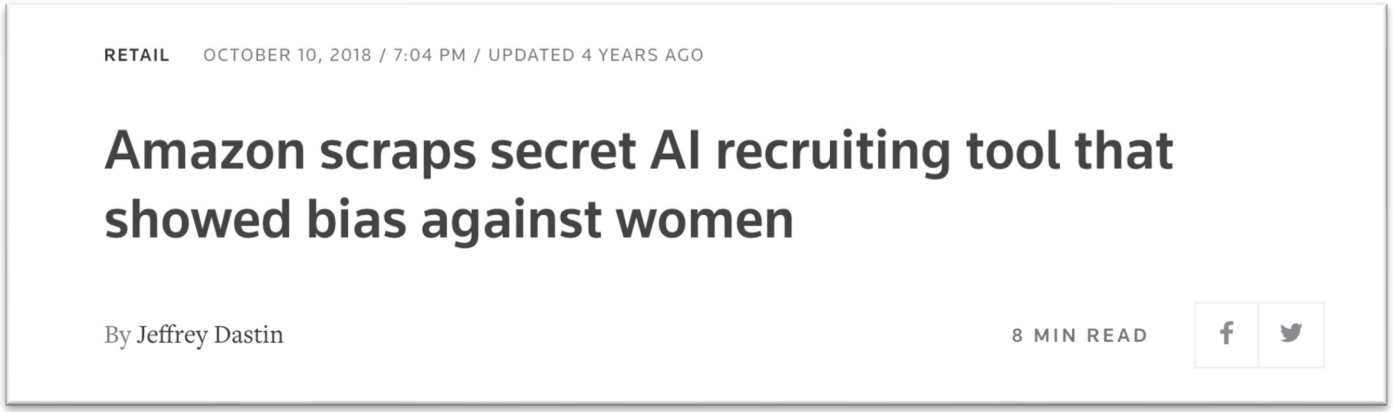
Explaining Models and Predictions

MGL 7811: Ingénierie logicielle des systèmes d'intelligence artificielle

Diego Elias Costa, PhD
Université du Québec à Montréal

The case for explainability

- We need to understand AI models to **find their problems**



RETAIL OCTOBER 10, 2018 / 7:04 PM / UPDATED 4 YEARS AGO

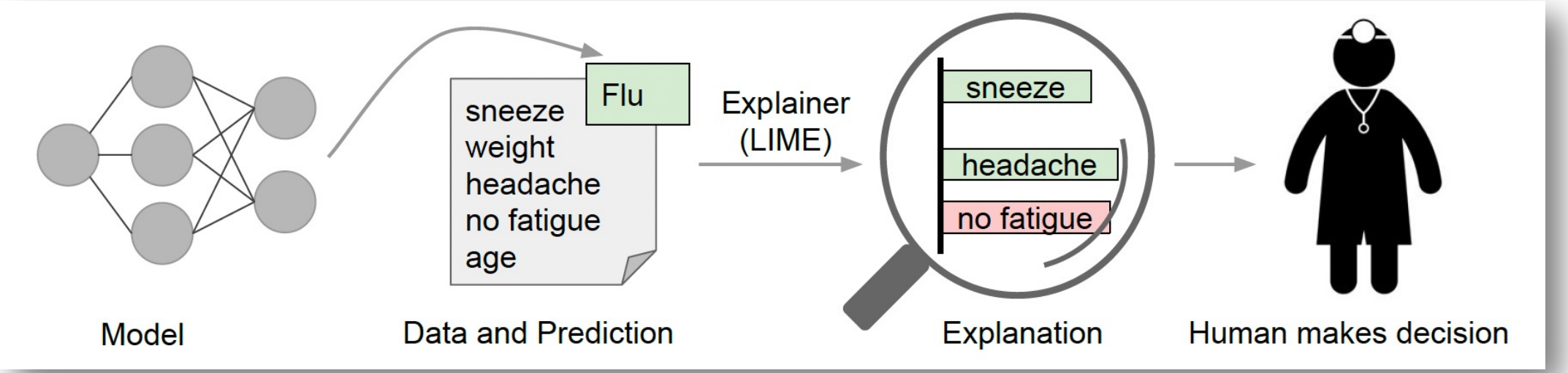
Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin 8 MIN READ

https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

The case for explainability

- We need to understand AI predictions to **make decisions**



The case for explainability

- We need explainability to **fine-tune our AI solutions**

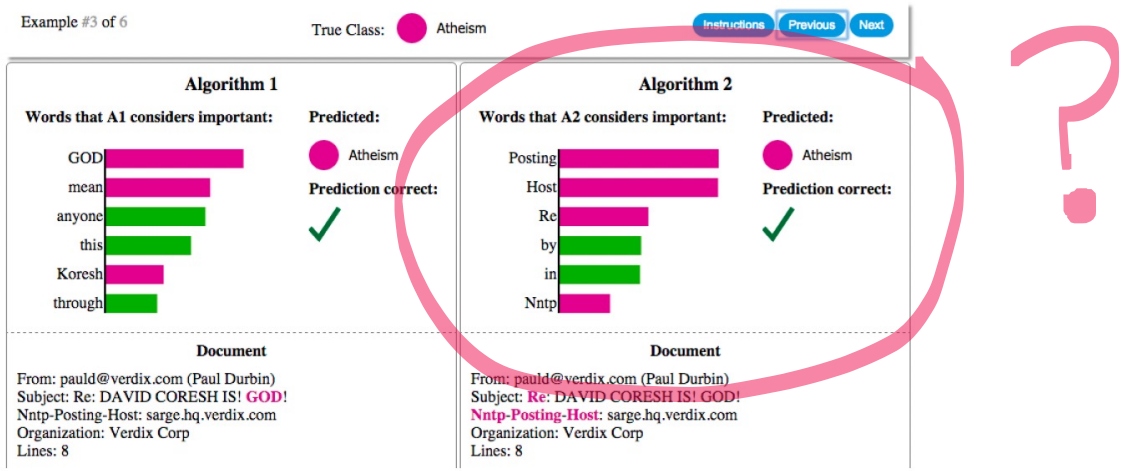
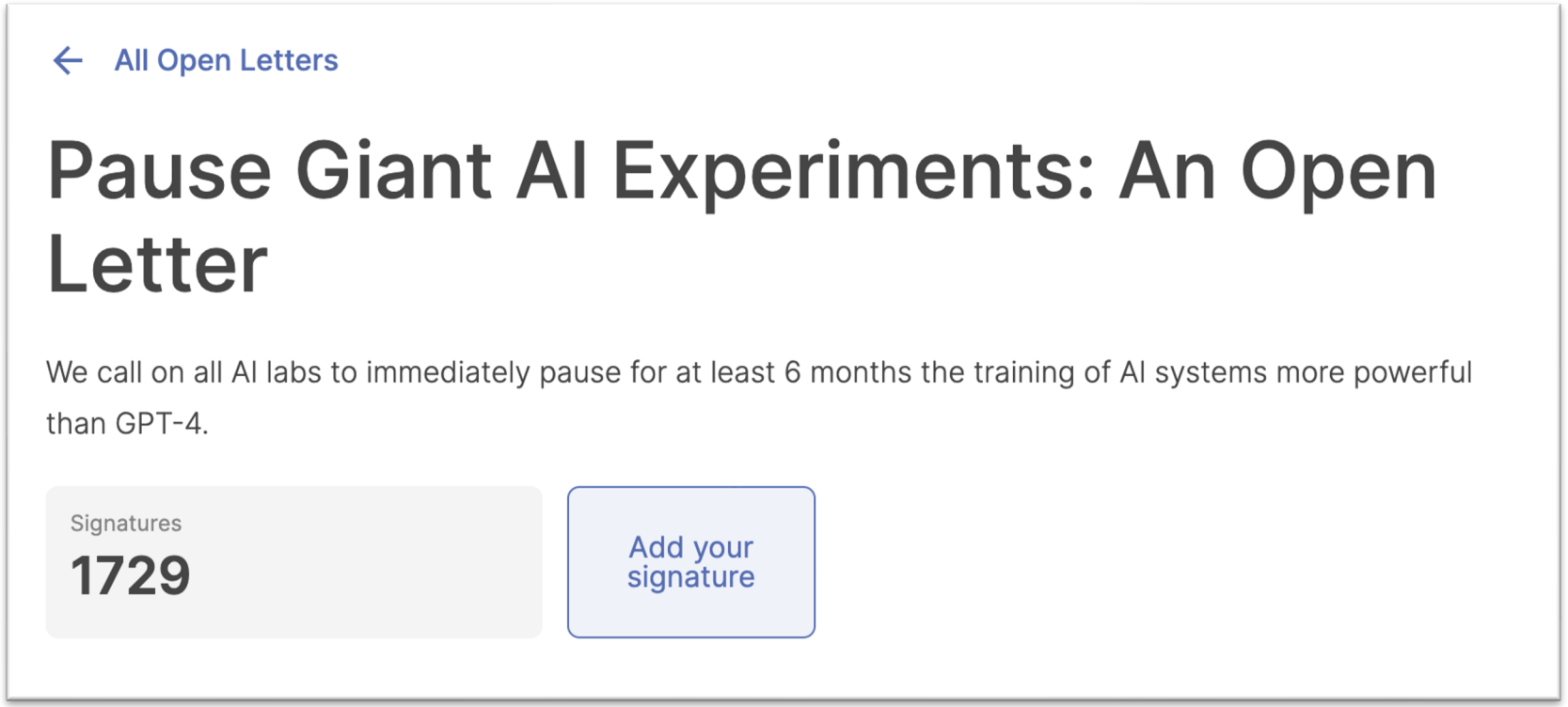


Figure 2: Explaining individual predictions of competing classifiers trying to determine if a document is about "Christianity" or "Atheism". The bar chart represents the importance given to the most relevant words, also highlighted in the text. Color indicates which class the word contributes to (green for "Christianity", magenta for "Atheism").

The case for explainability

- We need explainability to **reliably control** AI solutions



https://futureoflife.org/open-letter/pause-giant-ai-experiments/

If users to not trust a model or a prediction, they will not use it.

If engineers do not trust their model, they should not deploy it.

Benefits for Explaining Models and Predictions

For engineers

- Better understand the **model**
  - Validate model's decision with domain experts
  - Debug model's behavior
  - Assess the quality of the model beyond just accuracy metrics
  - Trust the model will behave reasonably on real-world data
  - Learn from the model

For end-users

- Trust an AI system's **decision** (prediction)
  - Comprehend the contributing factors
  - Have enough information to act
  - Treats the user fairly
  - Uses the provided information responsibly
- Trust the AI system is well developed (model)

Benefits for Explaining Models and Predictions

For engineers

- Better understand the **model**
  - Validate model's decision with domain experts
  - Debug model's behavior
  - Assess the quality of the model beyond just accuracy metrics
  - Trust the model will behave reasonably on real-world data
  - Learn from the model

Trusting the model

For end-users

- Trust an AI system's **decision** (prediction)
  - Comprehend the contributing factors
  - Have enough information to act
  - Treats the user fairly
  - Uses the provided information responsibly
- Trust the AI system is well developed (model)

Benefits for Explaining Models and Predictions

For engineers

- Better understand the **model**
  - Validate model's decision with domain experts
  - Debug model's behavior
  - Assess the quality of the model beyond just accuracy metrics
  - Trust the model will behave reasonably on real-world data
  - Learn from the model

Trusting the model

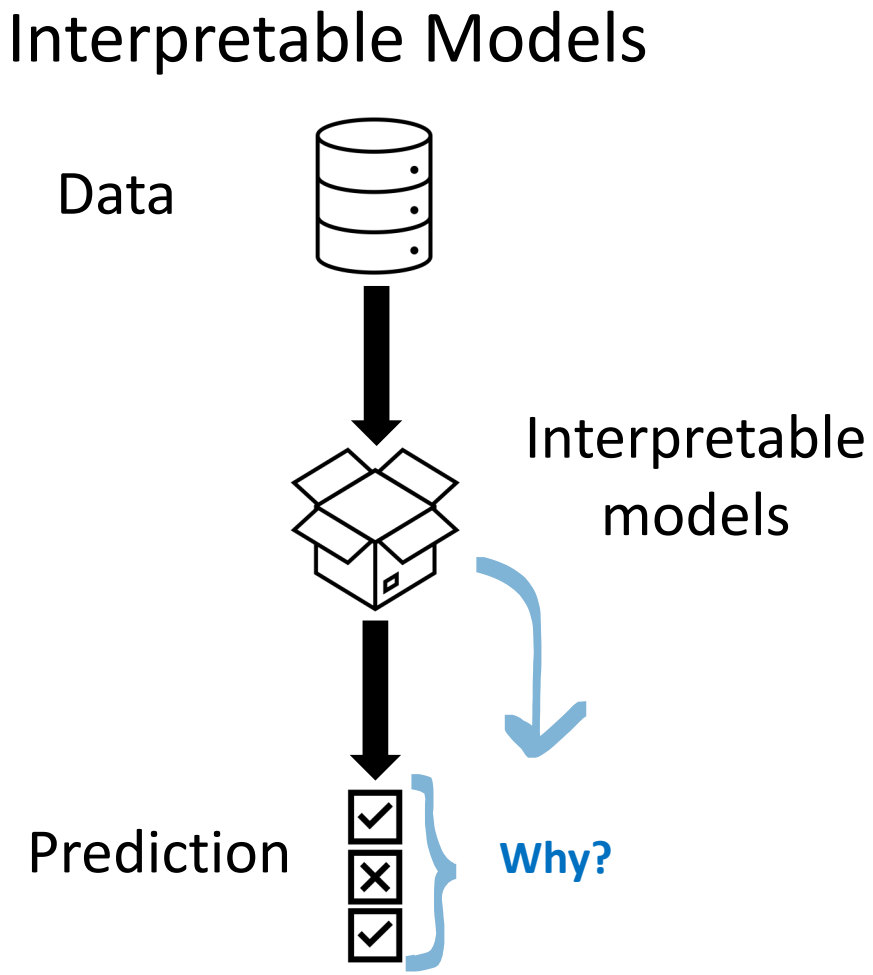
For end-users

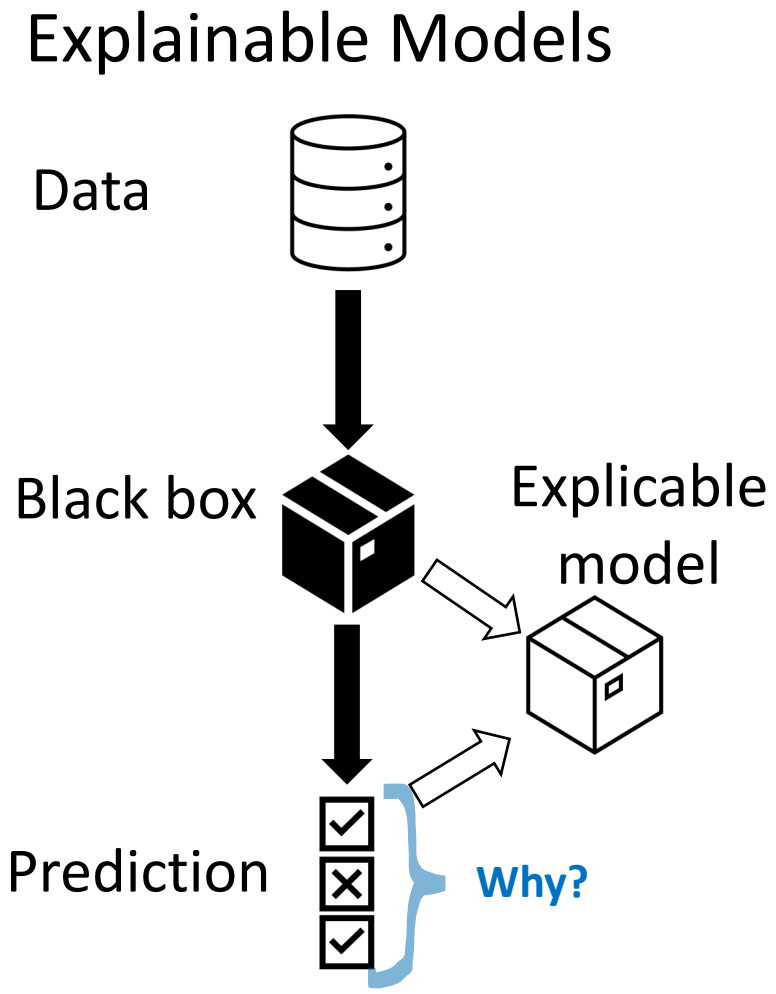
- Trust an AI system's **decision** (prediction)
  - Comprehend the contributing factors
  - Have enough information to act
  - Treats the user fairly
  - Uses the provided information responsibly

Trusting the prediction

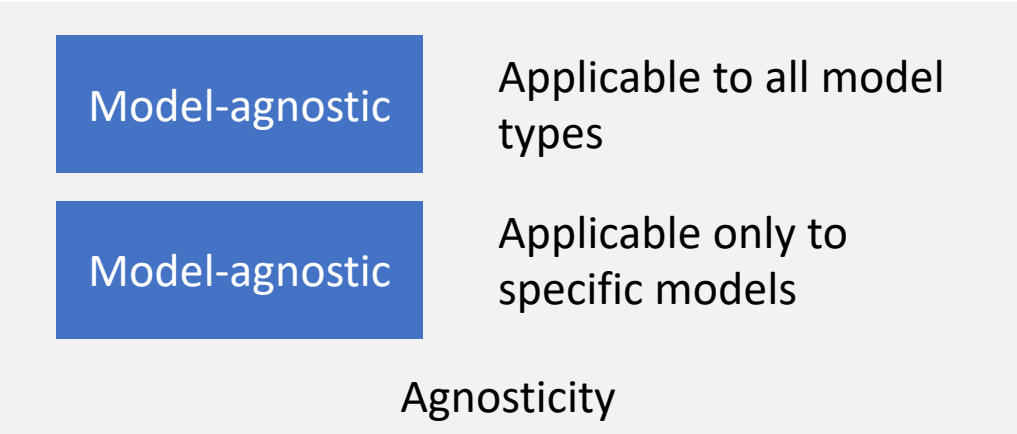
- Trust the AI system is well developed (model)

Difference between Explainable and Interpretable Models?





Categorization of Interpretable/Explainable methods



Source: DeepFindr YoutTube video - Explainable AI explained! #1 Introduction

Categorization of Interpretable/Explainable methods

Model-agnostic	Applicable to all model types
Model-agnostic	Applicable only to specific models

Agnosticity

Global explanation	Explains the whole model
Local explanation	Explains individual predictions

Scope

[Source: DeepFindr YoutTube video - Explainable AI explained! #1 Introduction]

Categorization of Interpretable/Explainable methods

Model-agnostic	Applicable to all model types
Model-agnostic	Applicable only to specific models

Agnosticity

Global explanation	Explains the whole model
Local explanation	Explains individual predictions

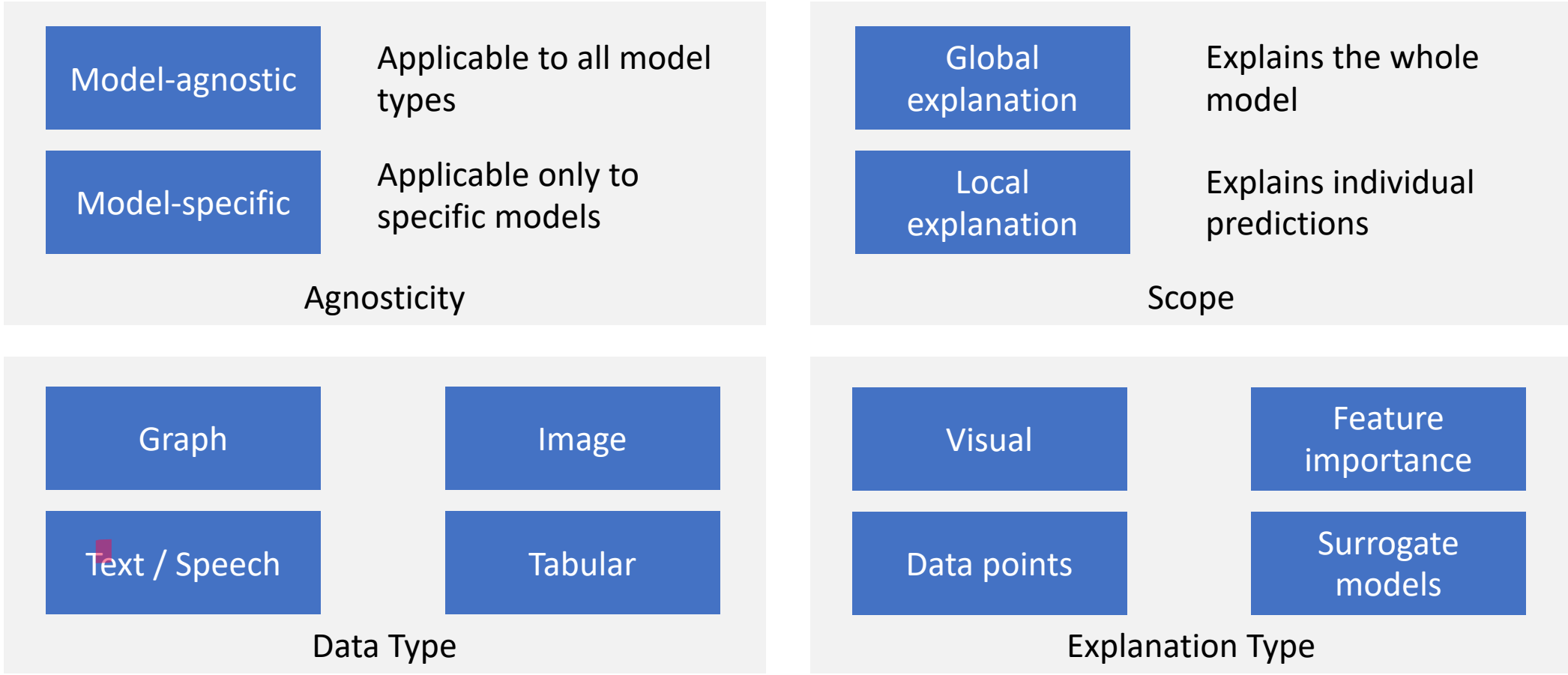
Scope

Graph	Image
Text / Speech	Tabular

Data Type

[Source: DeepFindr YoutTube video - Explainable AI explained! #1 Introduction]

Categorization of Interpretable/Explainable methods

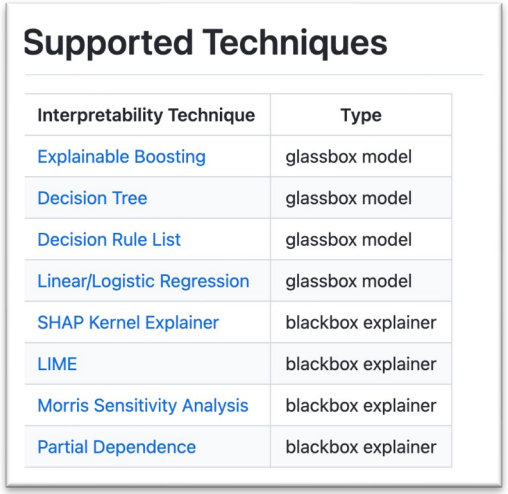


[Source: DeepFindr YoutTube video - Explainable AI explained! #1 Introduction]

Great set of resources

- Awesome Interpretable Machine Learning
  - Curated list of resources of interpretable machine learning
  - H20.ai MLI resource

- Recommended Python libraries
  - Interpret Library:
    - Developed by Microsoft
    - Compatible with Scikit-learn API
    - User interface integration
  - ELI5
    - Easy to use + compatible with scikit-learn



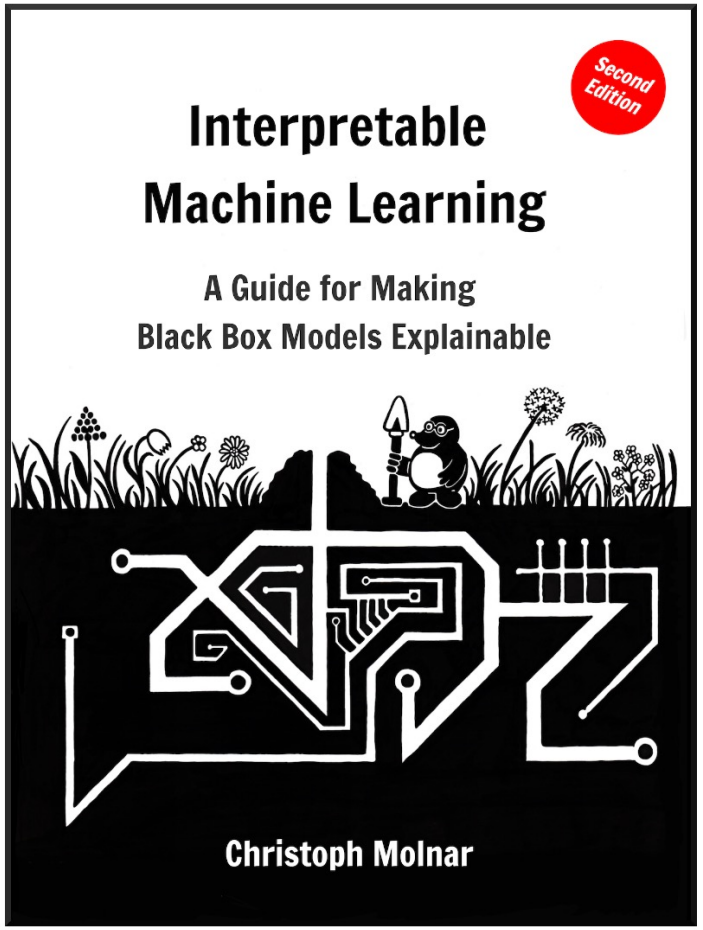
Supported Techniques

Interpretability Technique	Type
Explainable Boosting	glassbox model
Decision Tree	glassbox model
Decision Rule List	glassbox model
Linear/Logistic Regression	glassbox model
SHAP Kernel Explainer	blackbox explainer
LIME	blackbox explainer
Morris Sensitivity Analysis	blackbox explainer
Partial Dependence	blackbox explainer

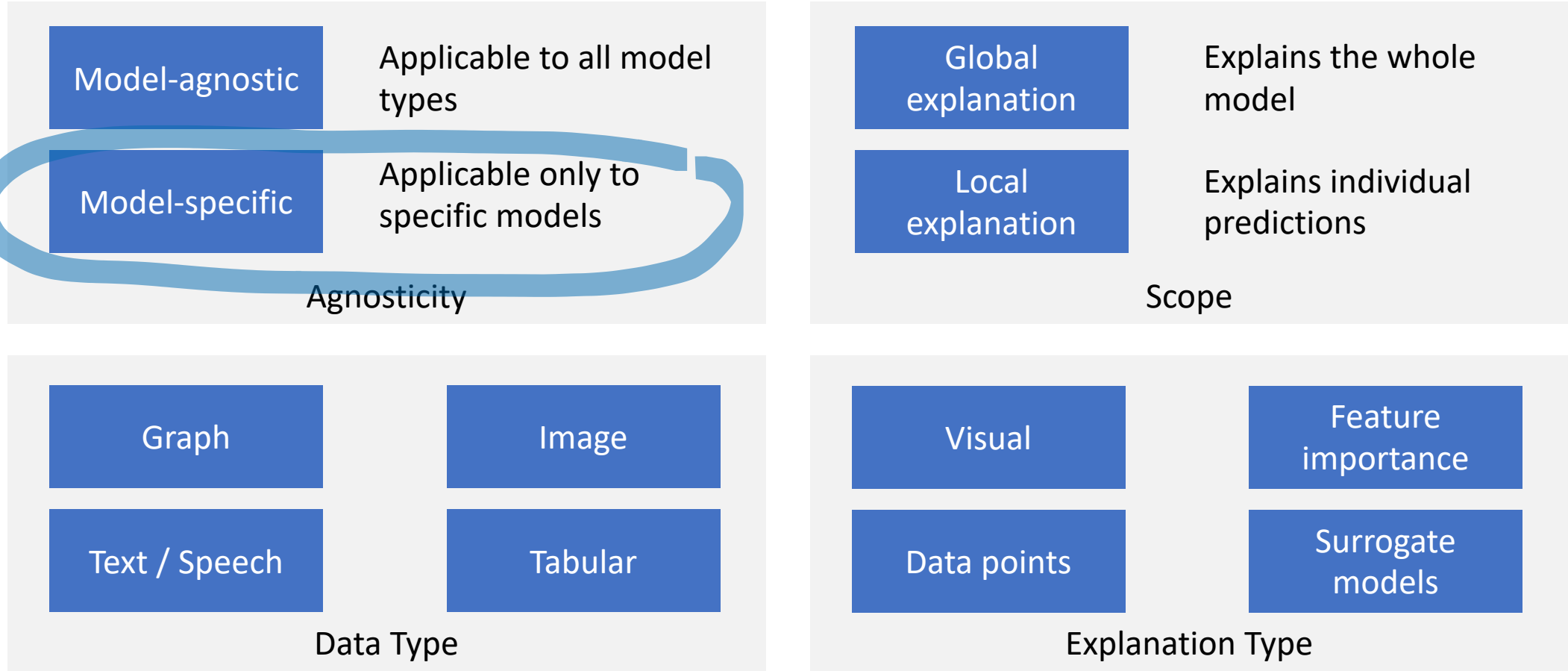
Great set of resources

- Interpretable Machine Learning
  - A Guide For Making Black Box Models Explainable

- Quite comprehensive
  - Describe numerous methods
  - Includes code and tooling source



Categorization of Interpretable/Explainable methods



[Source: DeepFindr YoutTube video - Explainable AI explained! #1 Introduction]

Trade-off between interpretability x accuracy

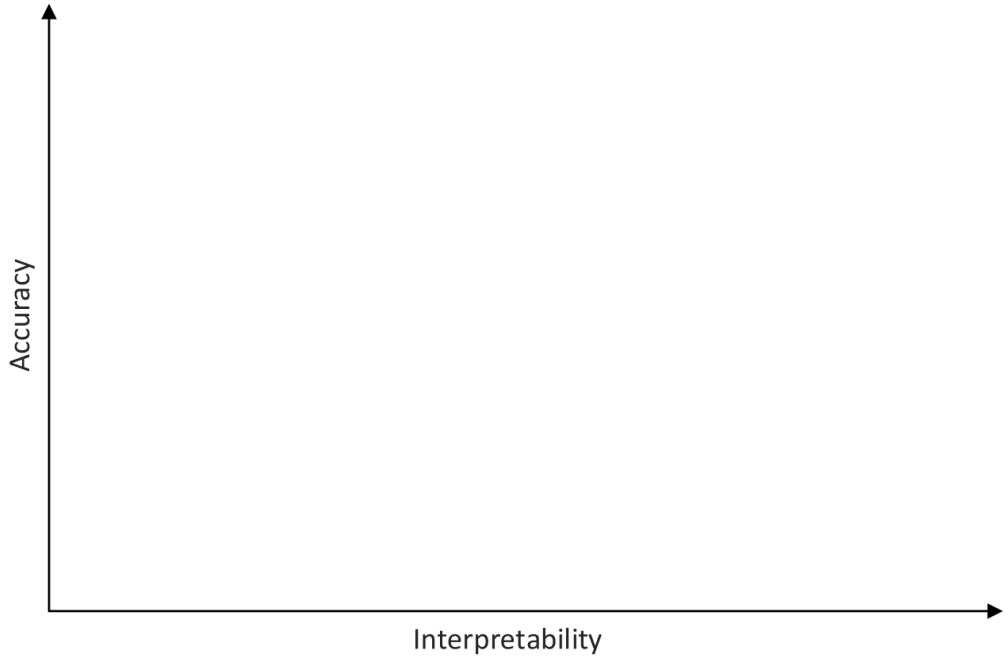


FIGURE 11. The trade-off between interpretability and accuracy of some relevant ML models. Highly interpretable algorithms such as classification rules, or linear regression, are often inaccurate. Very accurate DNNs are a classic example of black boxes.

Source: Machine Learning for 5G/B5G Mobile and Wireless Communications: Potential, Limitations, and Future Directions

Trade-off between interpretability x accuracy

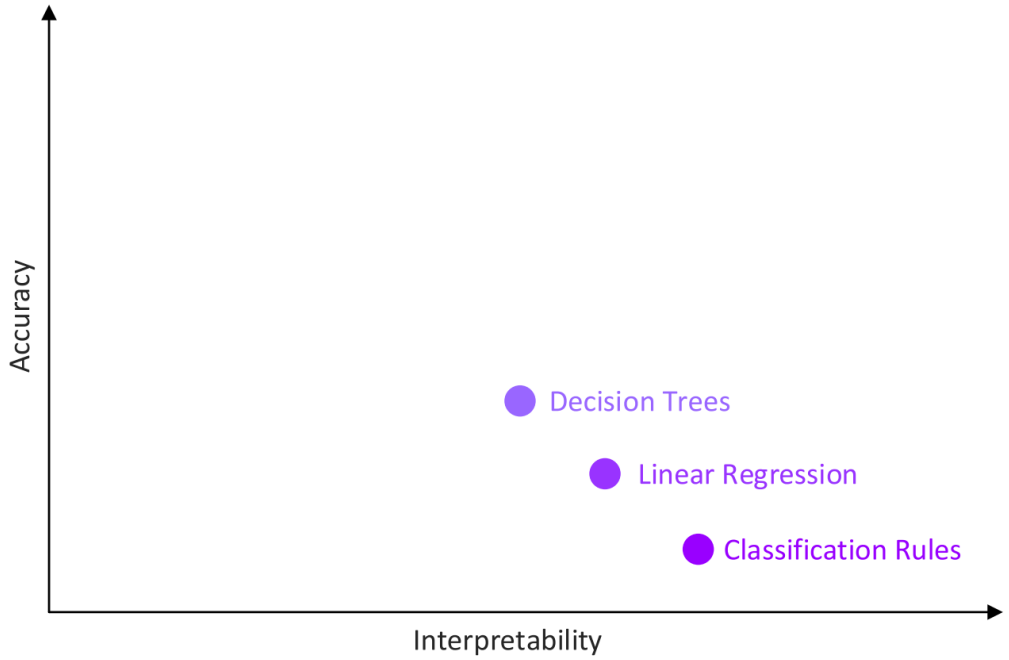


FIGURE 11. **The trade-off between interpretability and accuracy of some relevant ML models. Highly interpretable algorithms such as classification rules, or linear regression, are often inaccurate. Very accurate DNNs are a classic example of black boxes.**

Source: Machine Learning for 5G/B5G Mobile and Wireless Communications: Potential, Limitations, and Future Directions

Trade-off between interpretability x accuracy

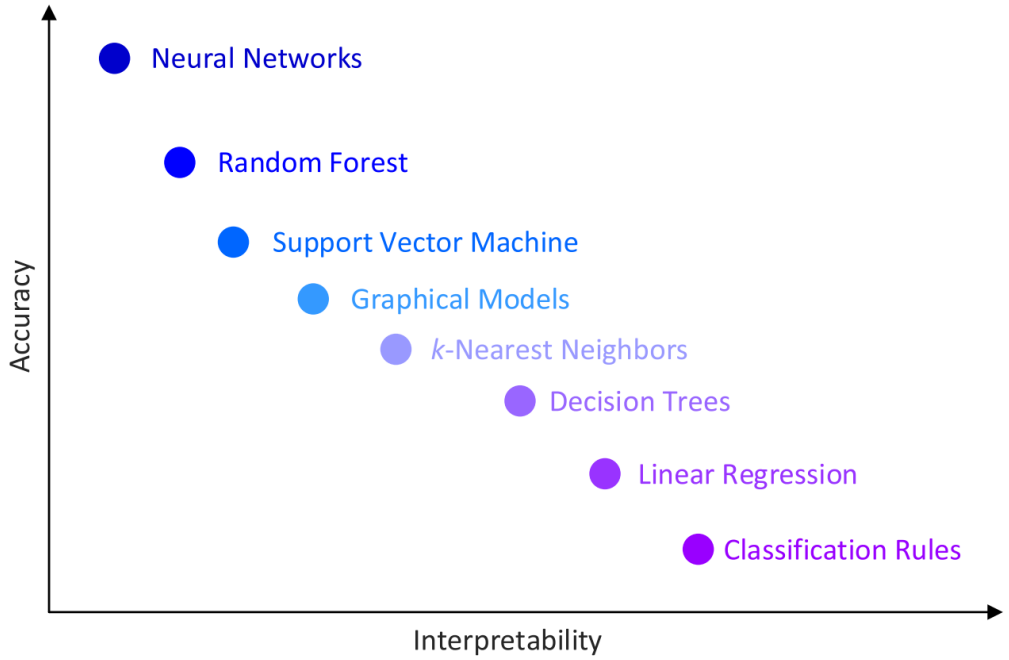


FIGURE 11. **The trade-off between interpretability and accuracy of some relevant ML models. Highly interpretable algorithms such as classification rules, or linear regression, are often inaccurate. Very accurate DNNs are a classic example of black boxes.**

Source: Machine Learning for 5G/B5G Mobile and Wireless Communications: Potential, Limitations, and Future Directions

Trade-off between interpretability x accuracy

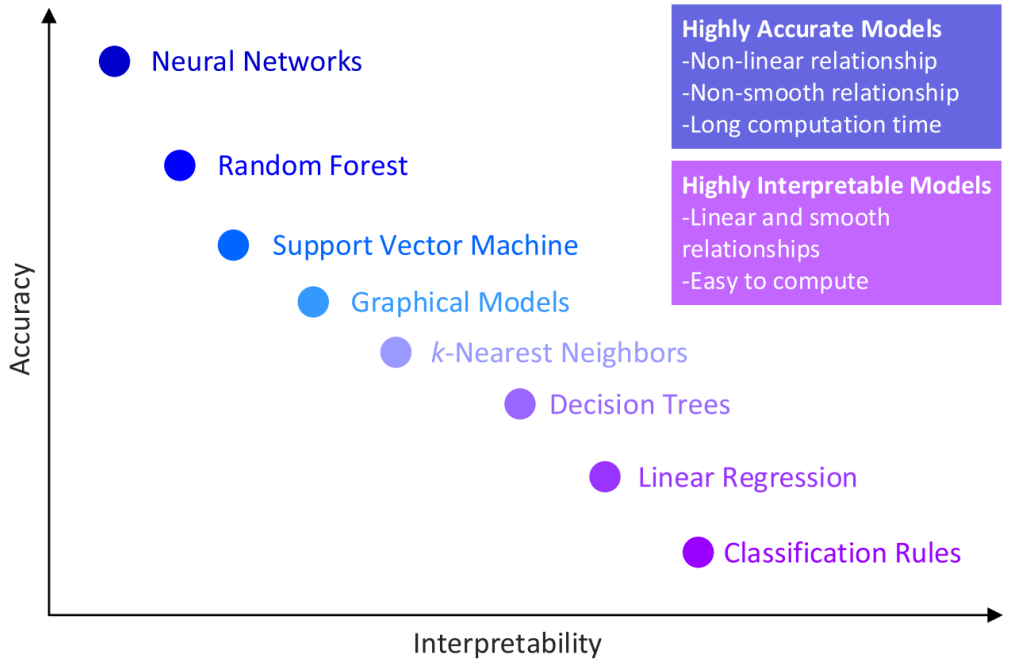


FIGURE 11. The trade-off between interpretability and accuracy of some relevant ML models. Highly interpretable algorithms such as classification rules, or linear regression, are often inaccurate. Very accurate DNNs are a classic example of black boxes.

Source: Machine Learning for 5G/B5G Mobile and Wireless Communications: Potential, Limitations, and Future Directions

Trade-off between interpretability x accuracy

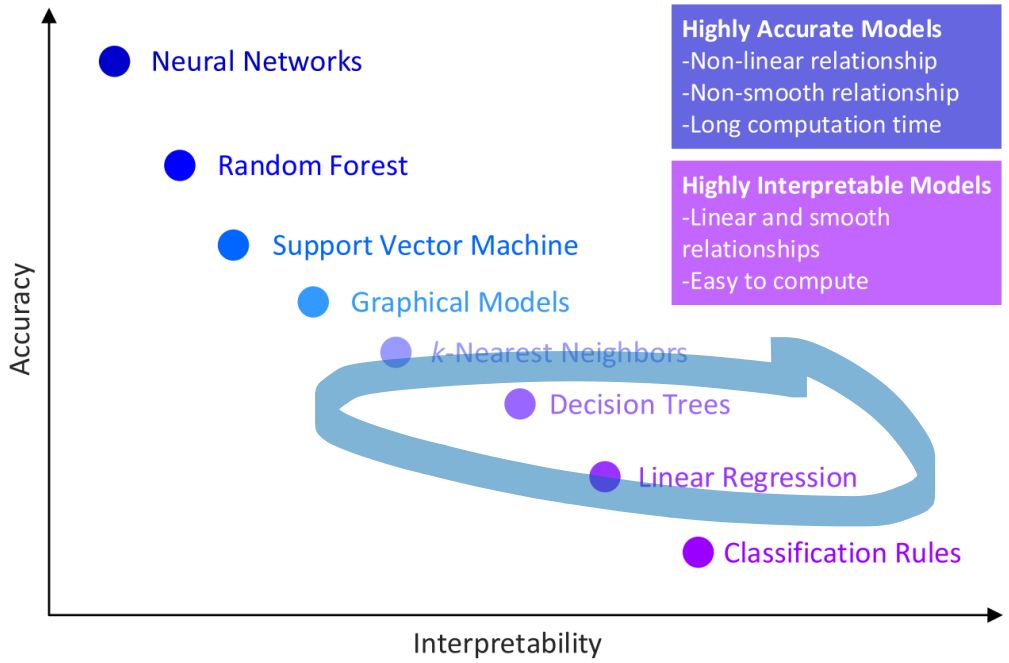


FIGURE 11. **The trade-off between interpretability and accuracy of some relevant ML models. Highly interpretable algorithms such as classification rules, or linear regression, are often inaccurate. Very accurate DNNs are a classic example of black boxes.**

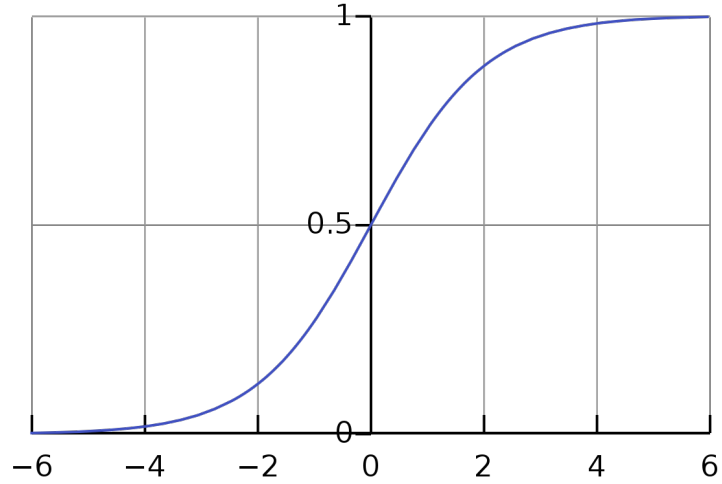
Source: Machine Learning for 5G/B5G Mobile and Wireless Communications: Potential, Limitations, and Future Directions

Logistic Regression

- Applicable for binary classification problems
  - Inherently interpretable

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

Credit approved



Credit not approved

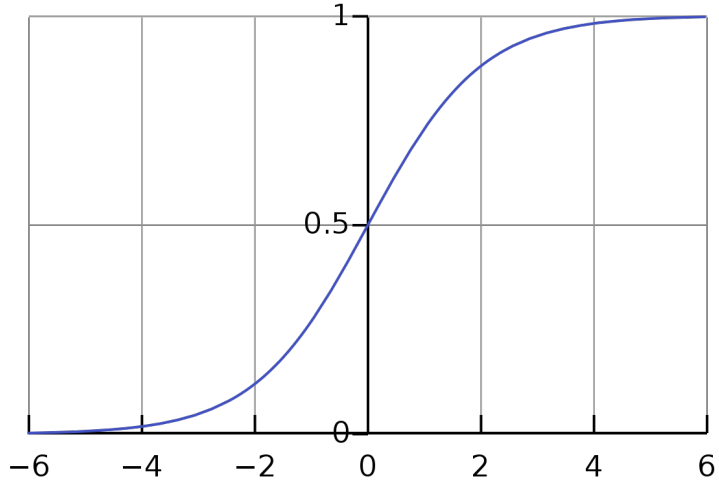
Logistic Regression

- Applicable for binary classification problems
  - Inherently interpretable

Features

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

Credit approved



Credit not approved

Logistic Regression

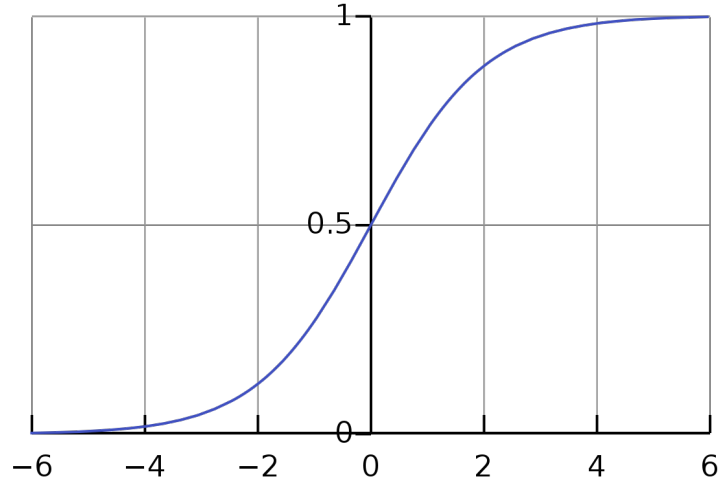
- Applicable for binary classification problems
  - Inherently interpretable

Features

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

Weights

Credit approved



Credit not approved

Logistic Regression

- Method: **Odds Ratio** (global model explanation)
  - How much a **unit increase** would change the outcome probability
  - Odds ratio of a feature (β): e^{β}

Weight[2]	Feature
+1.055	Checking account_no_inf
+0.906	Age_cat_Elder
+0.569	Purpose_radio/TV
+0.547	Housing_own
+0.545	<BIAS>
+0.514	Saving accounts_rich
+0.392	Checking account_rich
+0.383	Purpose_furniture/equipment
+0.298	Credit amount
+0.279	Age_cat_Senior
+0.265	Saving accounts_no_inf
+0.170	Housing_rent
+0.163	Saving accounts_quite rich
+0.049	Purpose_business
+0.026	Purpose_vacation/others
+0.000	Purpose_car
-0.036	Purpose_domestic appliances
-0.037	Duration
-0.087	Age_cat_Adult
-0.103	Saving accounts_moderate
-0.150	Purpose_repairs
-0.172	Housing_free
-0.174	Job
-0.287	Checking account_moderate
-0.294	Saving accounts_little
-0.295	Sex_female
-0.296	Purpose_education
-0.553	Age_cat_Young Adult
-0.615	Checking account_little

Credit Approved

Credit not Approved

Being of sex female increase **the odds of not getting the credit by** : $e^{0.295} = 1.34$

Discrimination by sex!

Logistic Regression

- Use weights to explain the prediction (Eli5)

```
[43]: eli5.show_prediction(classifier, X_test[2], feature_names=feature_names.values, show_feature_values=True
```

y=1 (probability **0.876**, score **1.960**) top features

Contribution	Feature	Value
+1.060	Credit amount	3.554
+1.055	Checking account_no_inf	1.000
+0.545	<BIAS>	1.000
+0.170	Housing_rent	1.000
+0.000	Purpose_car	1.000
-0.087	Age_cat_Adult	1.000
-0.103	Saving accounts_moderate	1.000
-0.332	Duration	9.000
-0.348	Job	2.000

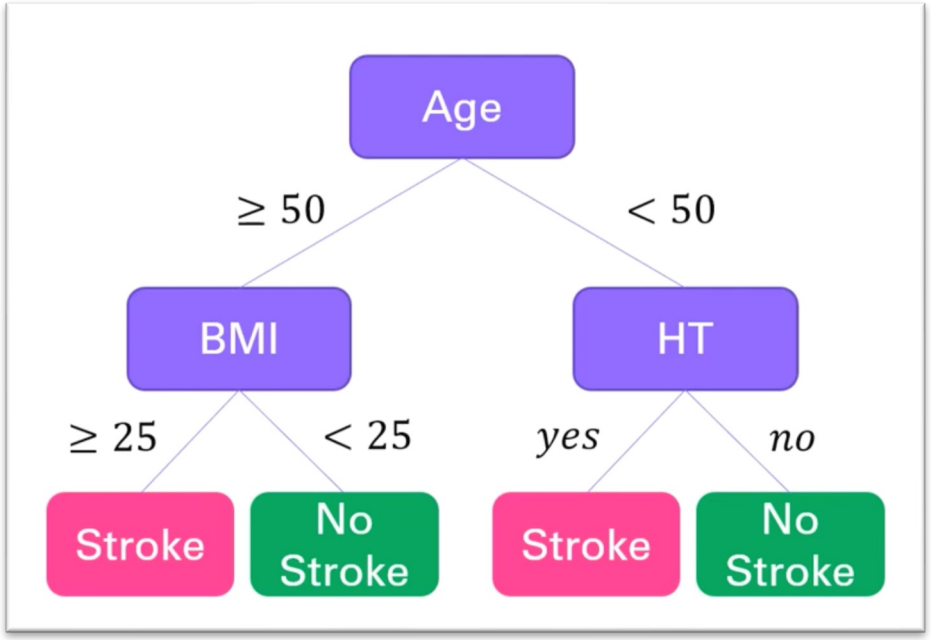
Credit approved (y=1)

Probability of 87.6% of paying back.
Contributing factors:

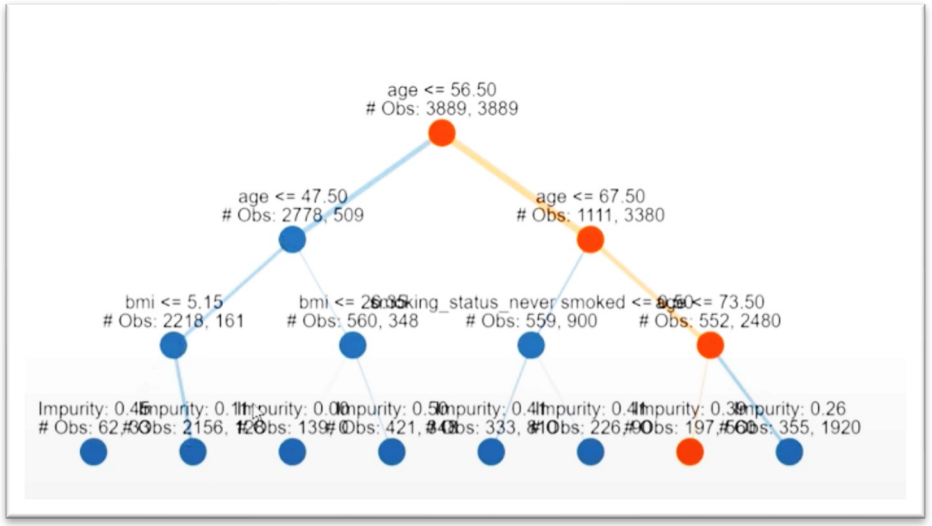
- Credit amount (+)
- No information on checking account (+)
- Rent a house (+)
- Has 2 jobs (-)
- Rent a house (+)

Decision Trees

- Applicable to multi-class problem
  - Inherently interpretable



Explaining the model

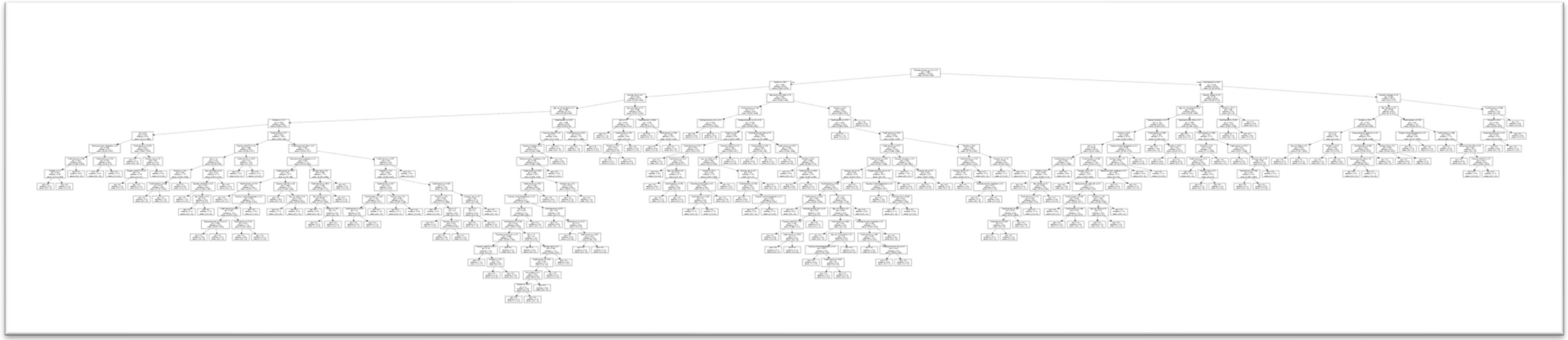


Explaining the prediction (tree visualizer)

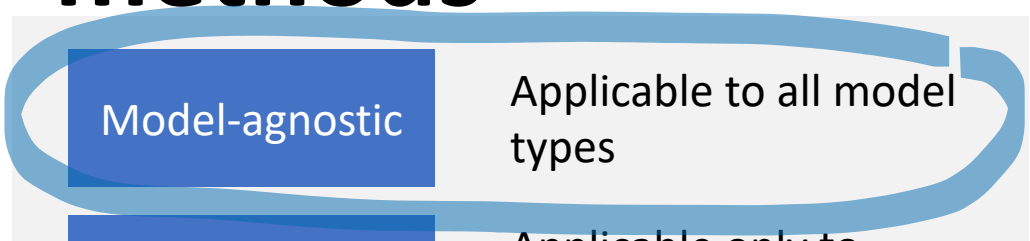
[Source: Explainable AI explained! | #2 By-design interpretable models with Microsofts InterpretML]

Decision Trees

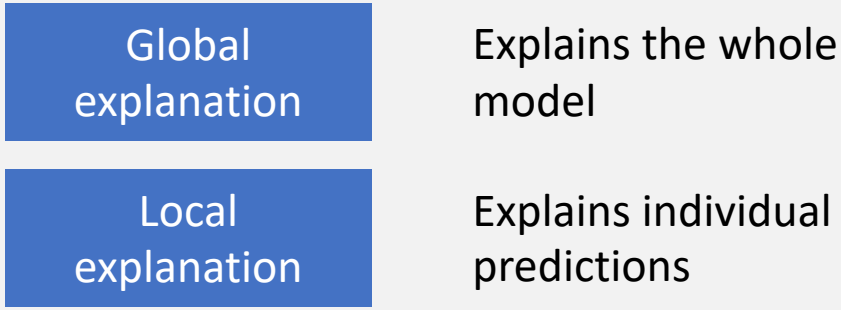
- Applicable to multi-class problem
  - Inherently interpretable
  - But… only works well with reasonable depth



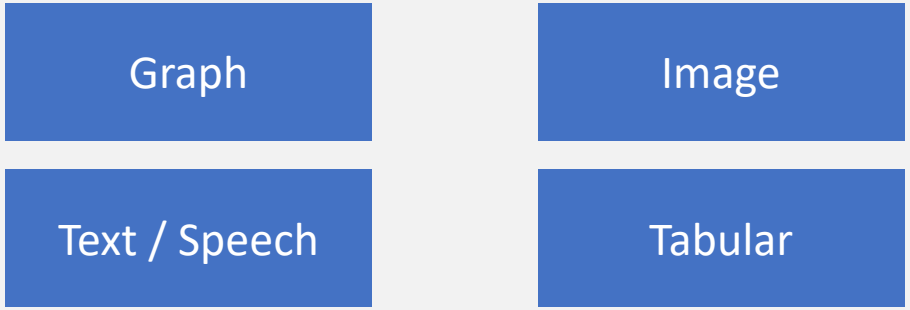
Categorization of Interpretable/Explainable methods



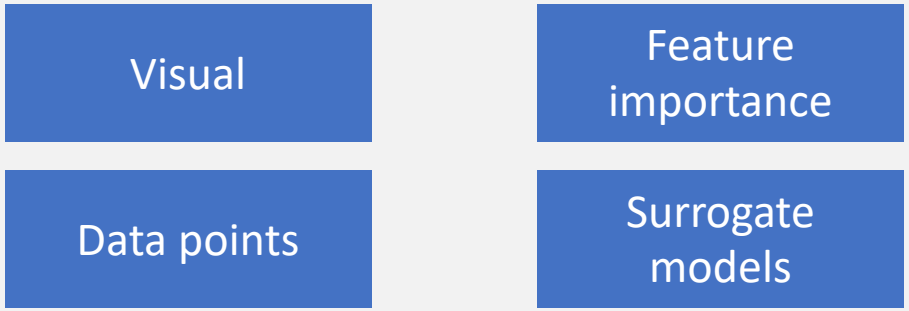
Agnosticity



Scope



Data Type



Explanation Type

[Source: DeepFindr YoutTube video - Explainable AI explained! #1 Introduction]

Permutation Feature Importance

- Model agnostic
- Global explanation
- Core idea:
  - Shuffle feature values
  - Verify the impact on the model's performance
  - Rank features
  - Critical features cause more impact in the model (top)
  - Less important features do not affect the model's performance as much (bottom)

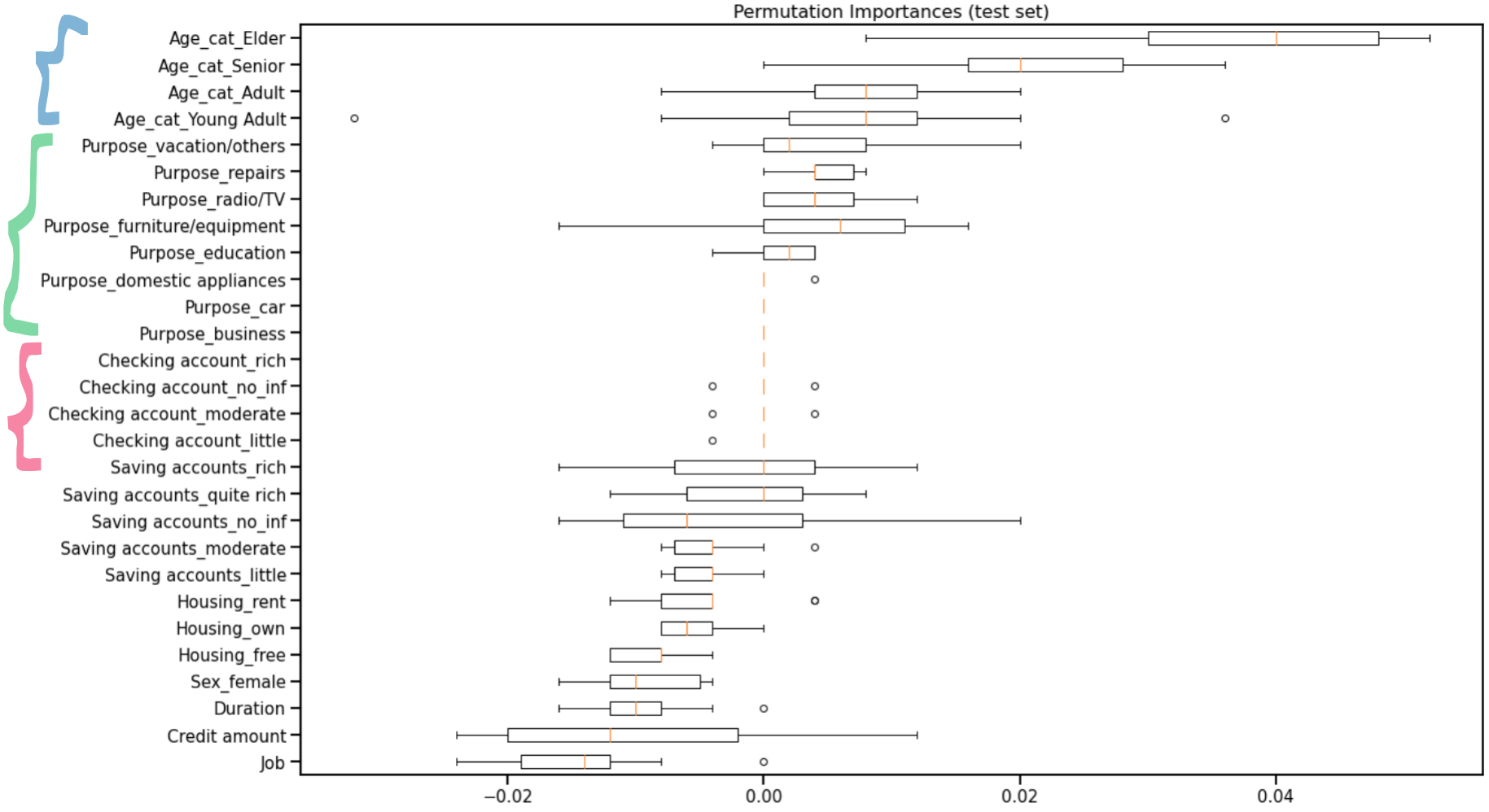
- Implemented at Scikitlearn

Permutation Feature Importance

Age

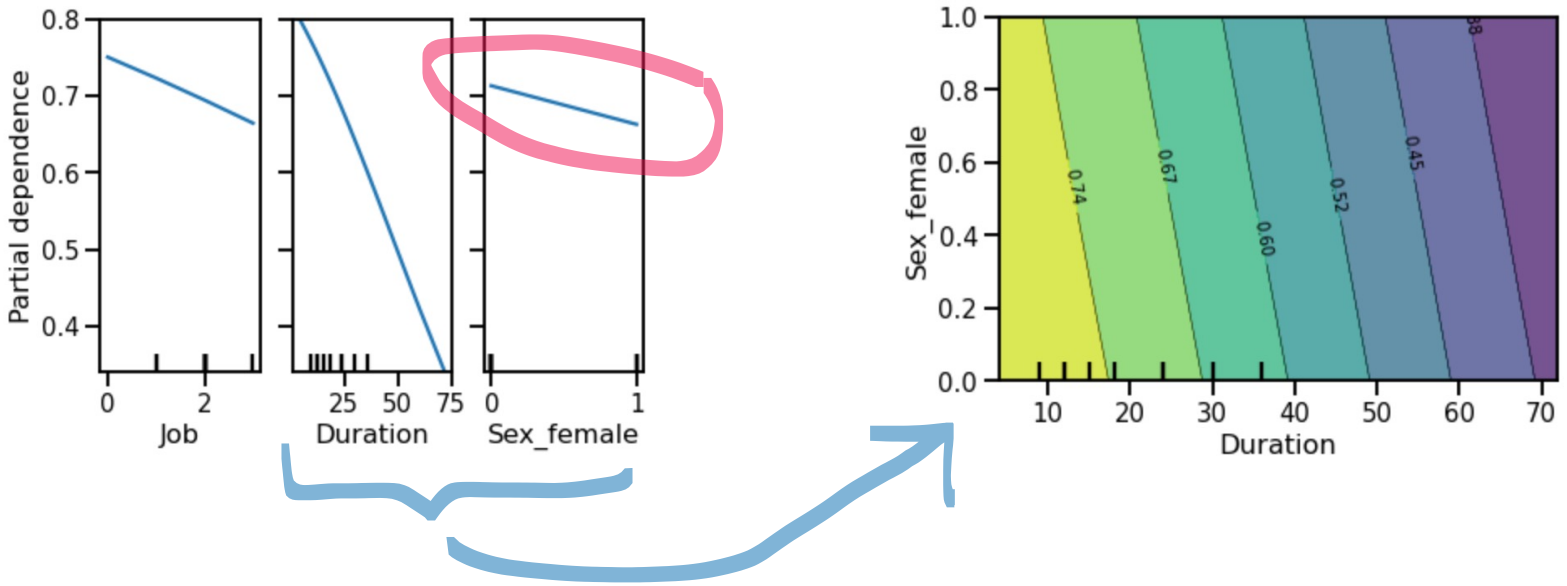
Purpose

Checking account



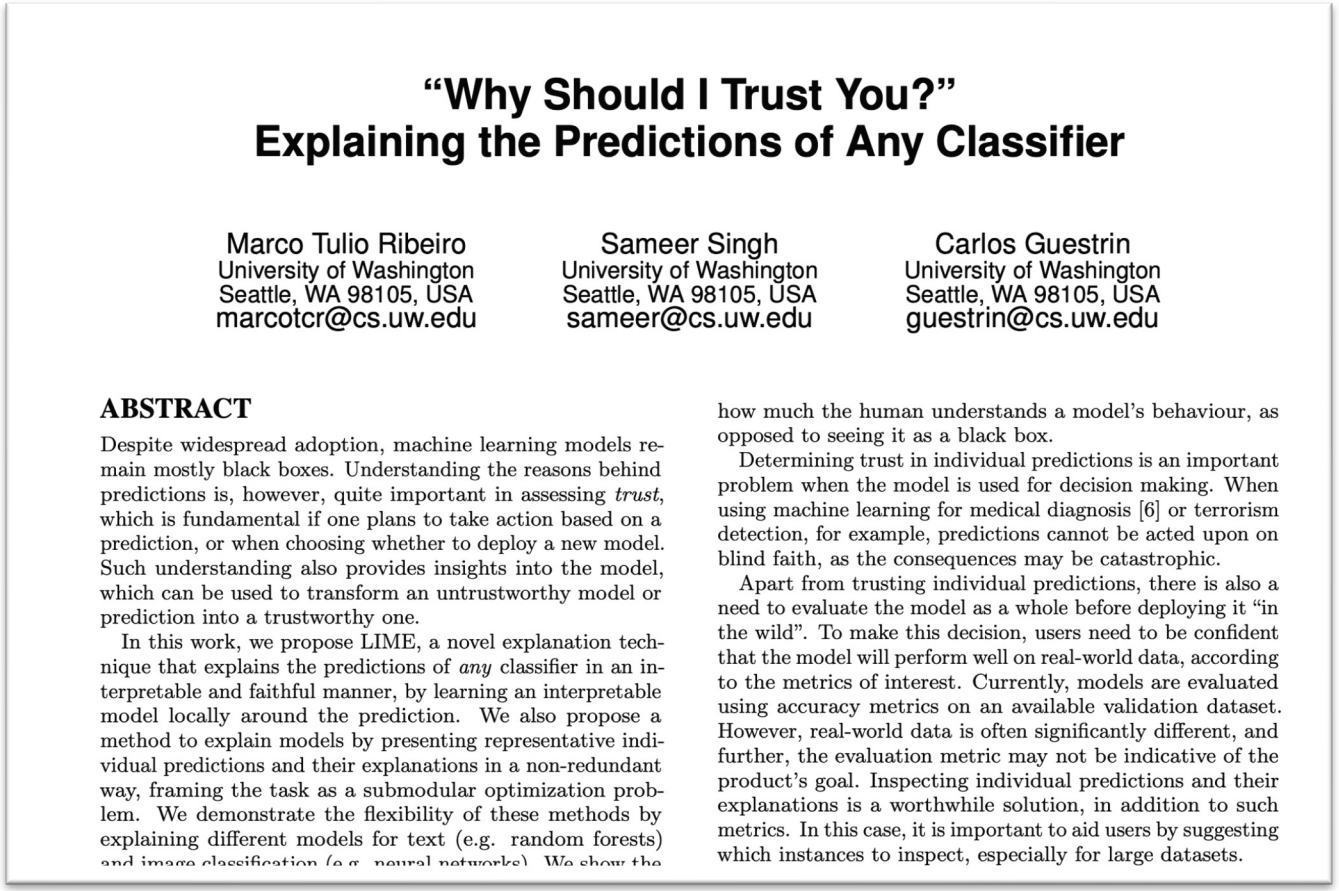
Partial Dependence Plots (PDP)

- Model agnostic
- Global explanation
- Code idea:
  - Shows the relationship of the feature with the target variable



Lime

- Focus on proposing a new method for explaining predictions



"Why Should I Trust You?" Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro
University of Washington
Seattle, WA 98105, USA
marcotcr@cs.uw.edu

Sameer Singh
University of Washington
Seattle, WA 98105, USA
sameer@cs.uw.edu

Carlos Guestrin
University of Washington
Seattle, WA 98105, USA
guestrin@cs.uw.edu

ABSTRACT

Despite widespread adoption, machine learning models remain mostly black boxes. Understanding the reasons behind predictions is, however, quite important in assessing *trust*, which is fundamental if one plans to take action based on a prediction, or when choosing whether to deploy a new model. Such understanding also provides insights into the model, which can be used to transform an untrustworthy model or prediction into a trustworthy one.

In this work, we propose LIME, a novel explanation technique that explains the predictions of *any* classifier in an interpretable and faithful manner, by learning an interpretable model locally around the prediction. We also propose a method to explain models by presenting representative individual predictions and their explanations in a non-redundant way, framing the task as a submodular optimization problem. We demonstrate the flexibility of these methods by explaining different models for text (e.g. random forests) and image classification (e.g. neural networks). We show the

how much the human understands a model's behaviour, as opposed to seeing it as a black box.

Determining trust in individual predictions is an important problem when the model is used for decision making. When using machine learning for medical diagnosis [6] or terrorism detection, for example, predictions cannot be acted upon on blind faith, as the consequences may be catastrophic.

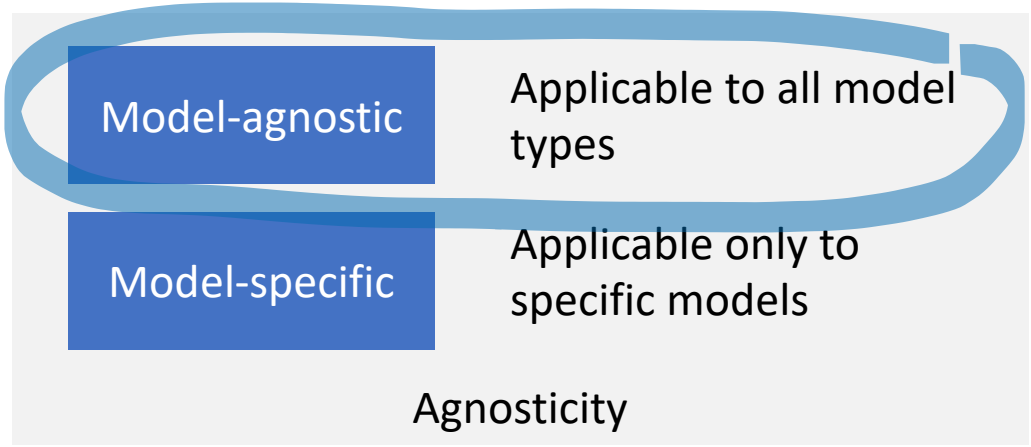
Apart from trusting individual predictions, there is also a need to evaluate the model as a whole before deploying it "in the wild". To make this decision, users need to be confident that the model will perform well on real-world data, according to the metrics of interest. Currently, models are evaluated using accuracy metrics on an available validation dataset. However, real-world data is often significantly different, and further, the evaluation metric may not be indicative of the product's goal. Inspecting individual predictions and their explanations is a worthwhile solution, in addition to such metrics. In this case, it is important to aid users by suggesting which instances to inspect, especially for large datasets.

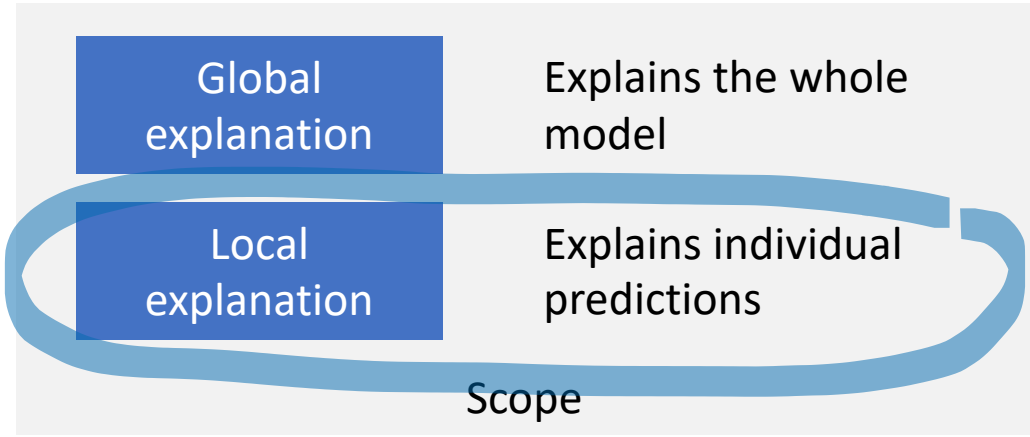
Study Design

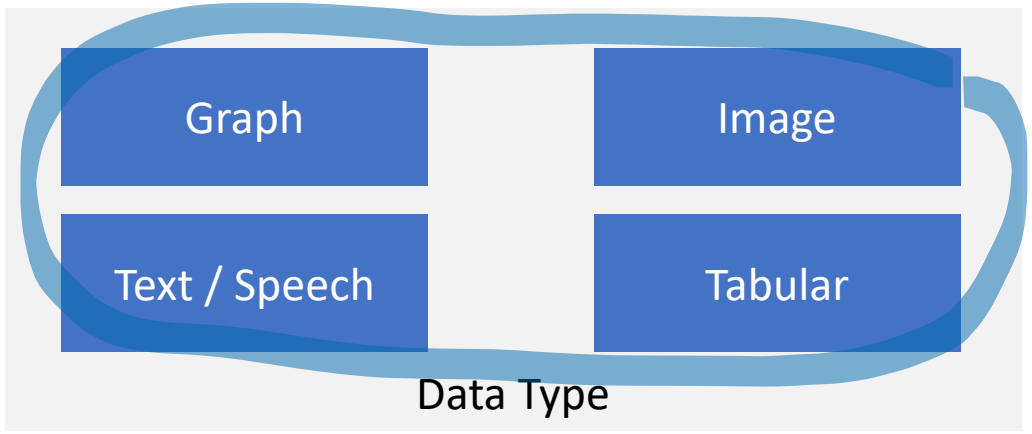
- Approach paper / Solution paper
  - Detailed explanation of the idea behind the solution
  - Thorough evaluation

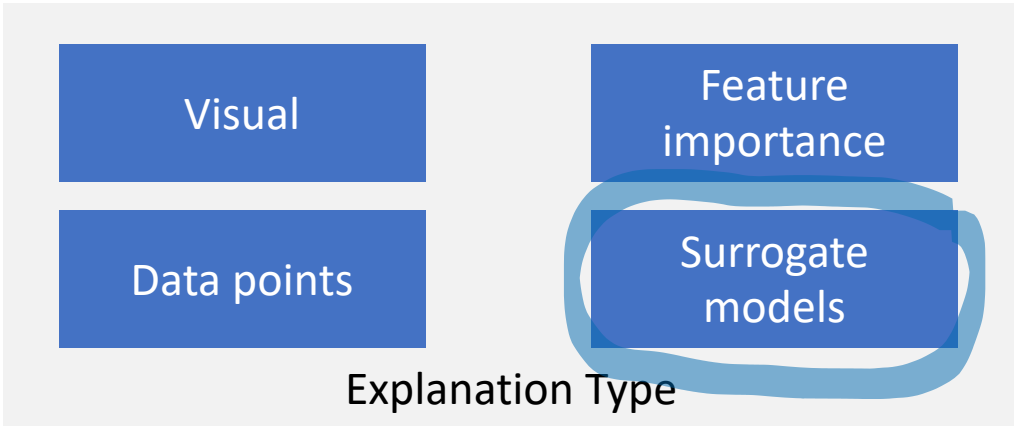
- Seminal paper
  - Published at KDD 2016
  - Cited more than 12 thousand times

What is the categories of LIME?









[Source: DeepFindr YoutTube video - Explainable AI explained! #1 Introduction]

LIME

- **L**ocal **I**nterpretable **M**odel-Agnostic **E**xplanation

- Contributions
  - Explain an individual **prediction** -> trusting the prediction (LIME)
  - Select and explain **multiple predictions** -> trusting the **model** (SP-LIME)

Desired Characteristics for Explainers

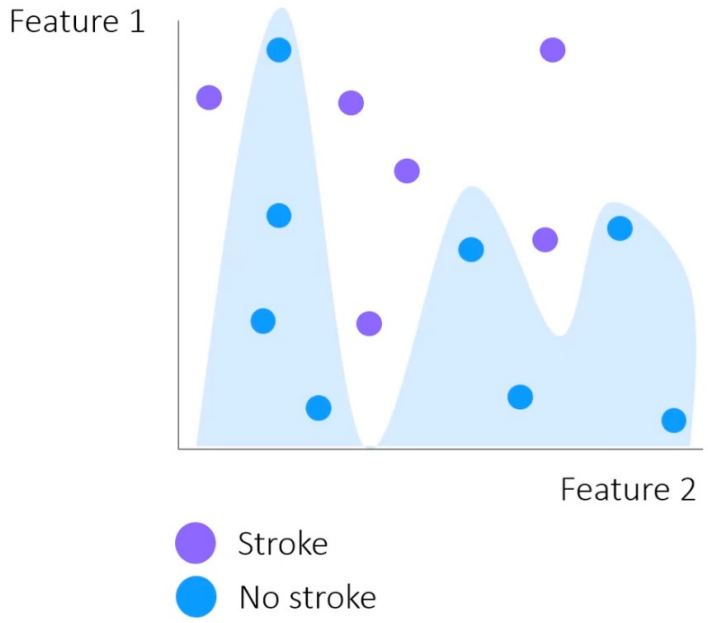
- **Interpretable**: provide qualitative understanding between the input variables and the response
  - Should not have 1000s of features for example
  - Can depend on the target audience

- **Local fidelity**: an explanation must correspond to **how the model behaves in the vicinity** of the instance being predicted

Desired Characteristics for Explainers

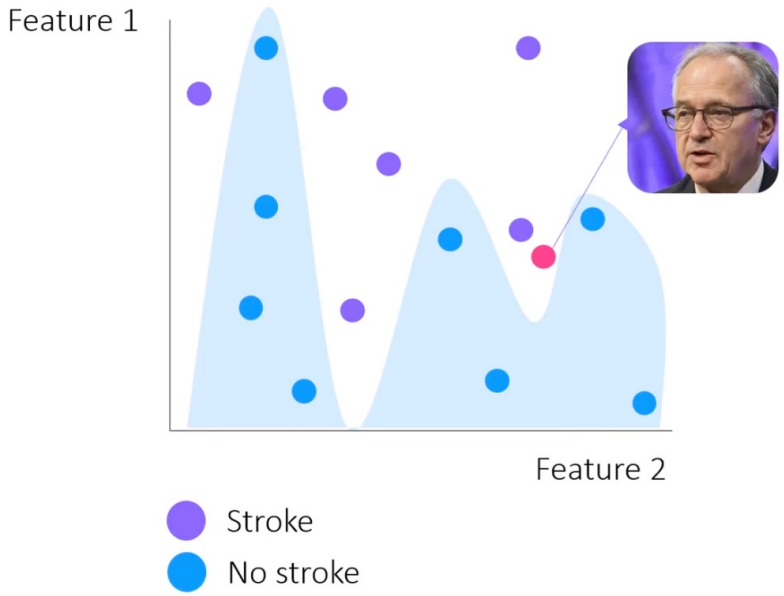
- **Model-agnostic explainer**: explain **any** model
- **Providing a global perspective**: important to ascertain trust in the model

How does LIME works?



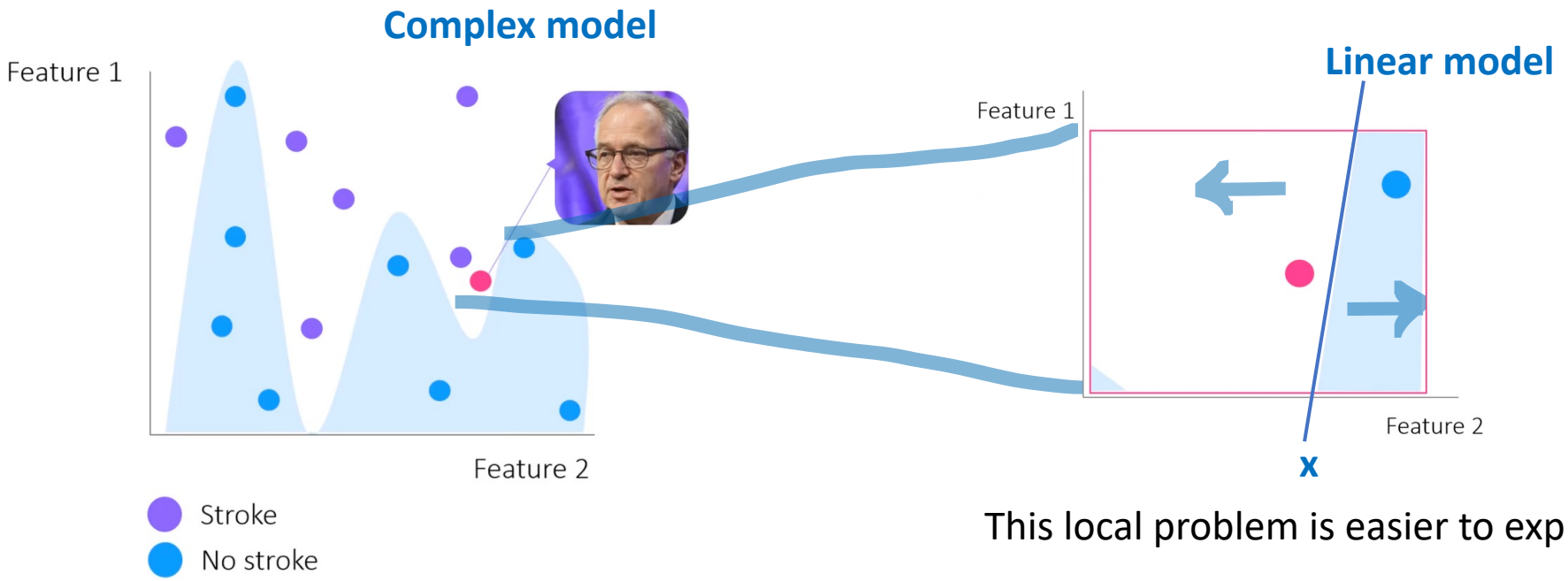
[Source: Explainable AI explained! | #3 LIME]

How does LIME works?



[Source: Explainable AI explained! | #3 LIME]

How does LIME works?



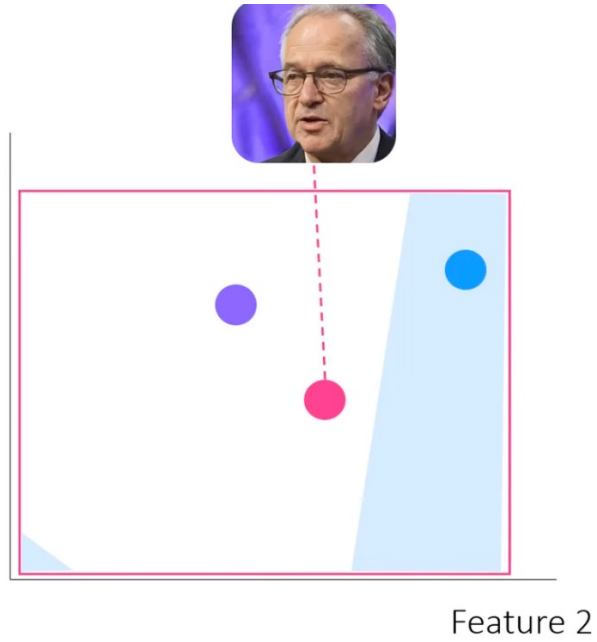
This local problem is easier to explain

One possible local model:
Classify **Stroke** if Feature 2 is smaller than **x**
Otherwise, **No Stroke**

[Source: Explainable AI explained! | #3 LIME]

Mathematical Intuition

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

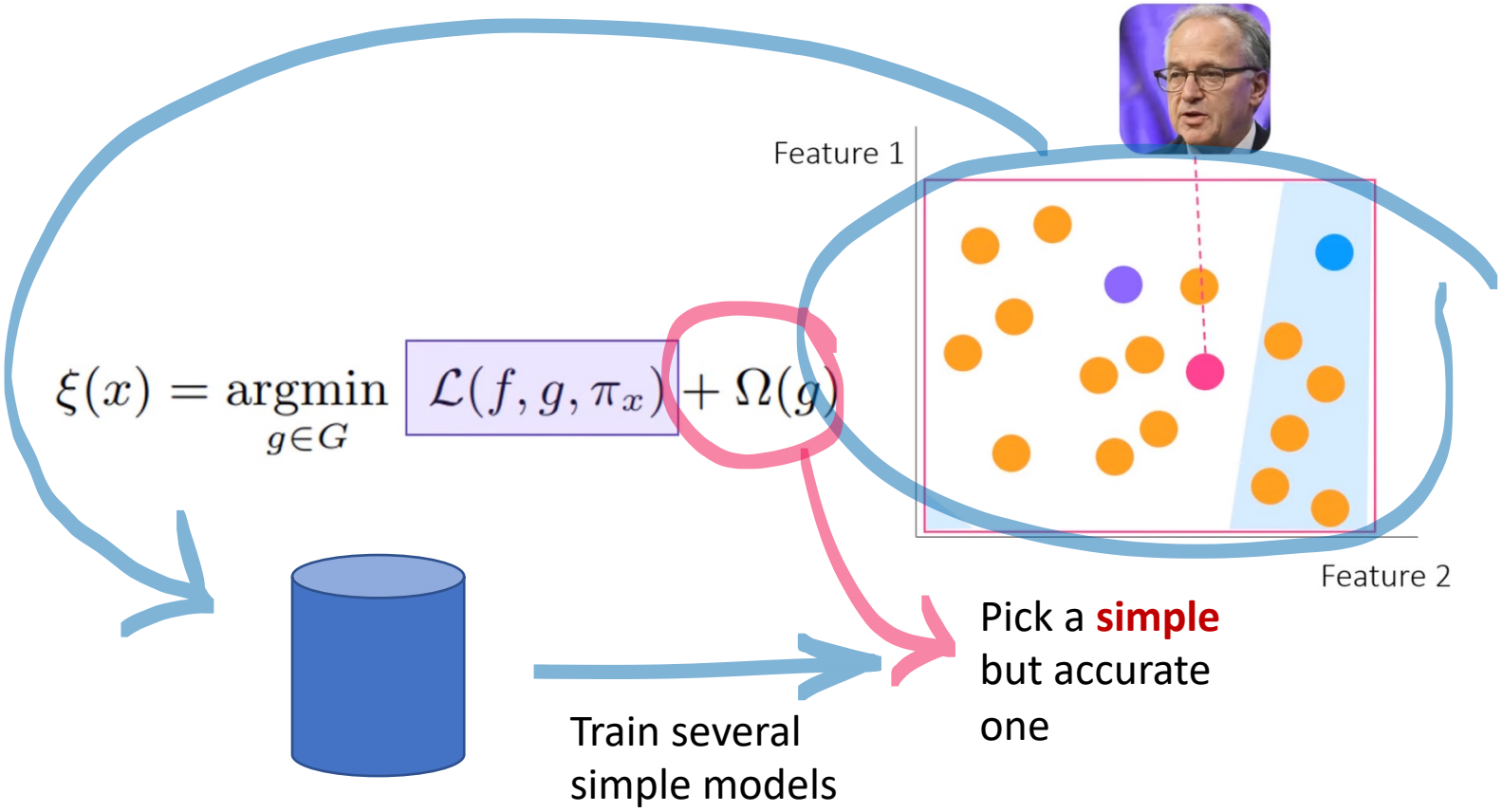


Source: Explainable AI explained! | #3 LIME

Mathematical Intuition

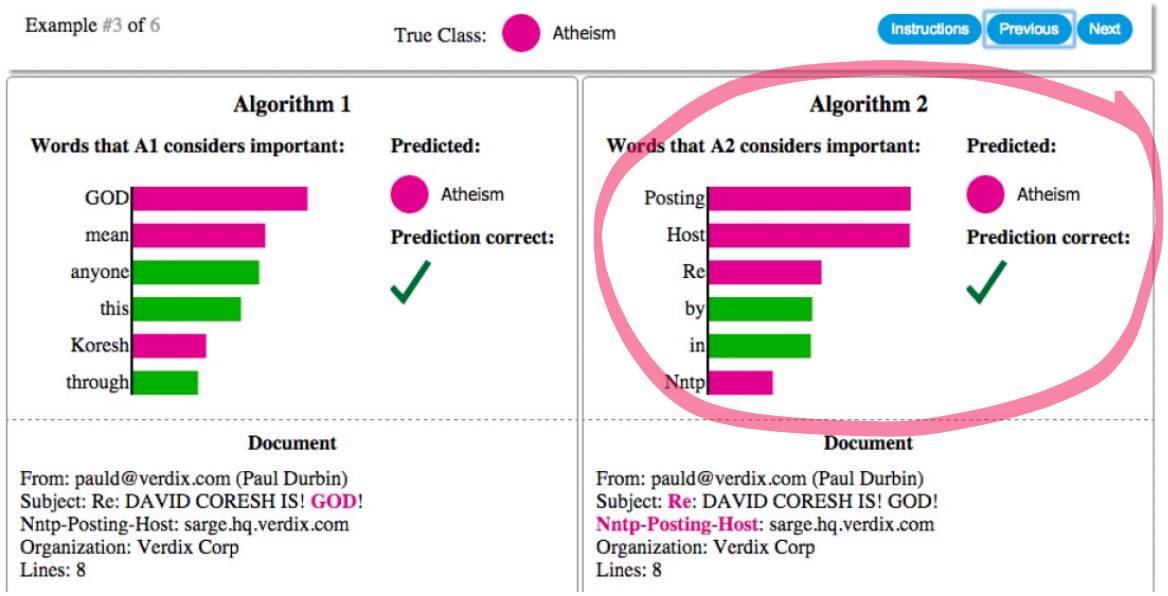
Ask the complex model to classify newly generated points

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



Source: Explainable AI explained! | #3 LIME

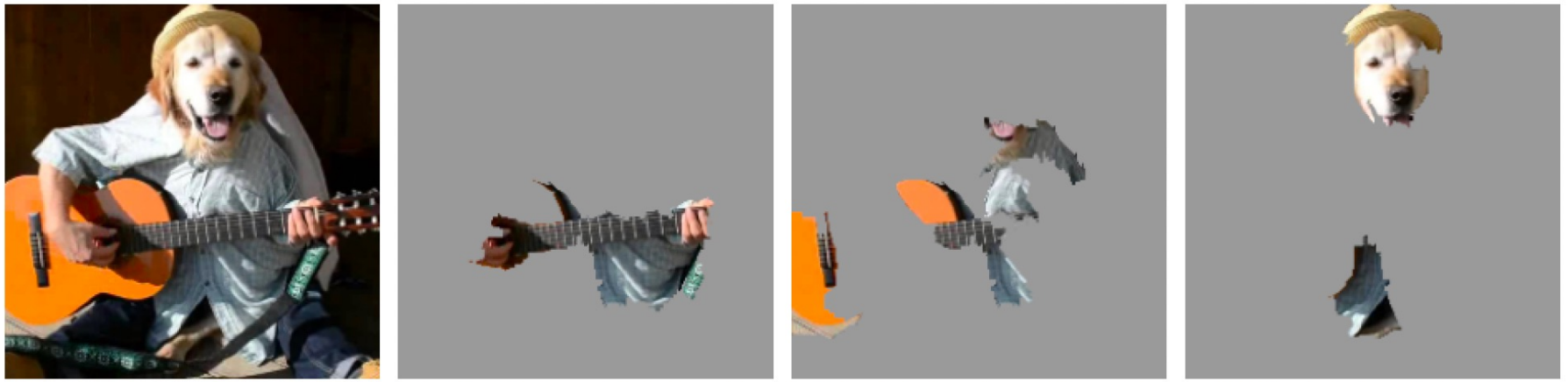
What is the main idea here?



?

Figure 2: Explaining individual predictions of competing classifiers trying to determine if a document is about "Christianity" or "Atheism". The bar chart represents the importance given to the most relevant words, also highlighted in the text. Color indicates which class the word contributes to (green for "Christianity", magenta for "Atheism").

What is the main idea here?



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

Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

So far, we have only explained single predictions…

…what about explaining the model?

Submodular pick for explaining models

- Problem:
  - Select a **small set of explanations** that can **explain the whole model**?

- Solution?

Algorithm 2 Submodular pick (SP) algorithm

Require: Instances X, Budget B
  for all $x_i \in X$ **do**
    $\mathcal{W}_i \leftarrow$ **explain**(x_i, x_i') ▷ Using
  end for
  for $j \in \{1 \dots d'\}$ **do**
    $I_j \leftarrow \sqrt{\sum_{i=1}^{n} |\mathcal{W}_{ij}|}$ ▷ Compute feature importances
  end for
  $V \leftarrow \{\}$
  while $|V| < B$ **do** ▷ Greedy optimization of Eq (4)
    $V \leftarrow V \cup \text{argmax}_i \, c(V \cup \{i\}, \mathcal{W}, I)$
  end while
  return V

$$c(V, \mathcal{W}, I) = \sum_{j=1}^{d'} \mathbb{1}_{[\exists i \in V : \mathcal{W}_{ij} > 0]} I_j \qquad (3)$$

$$Pick(\mathcal{W}, I) = \underset{V, |V| \leq B}{\text{argmax}} \, c(V, \mathcal{W}, I)$$

Submodular pick for explaining models

- Problem:
  - Select a **small set of explanations** that can **explain the whole model**?

- Solution?
  - Select instances that cover the important components
  - Avoid redundant explanations

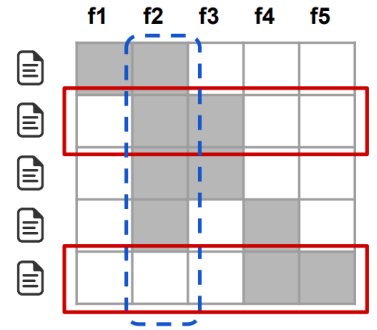


Figure 5: Toy example $\mathcal{W}$. Rows represent instances (documents) and columns represent features (words). Feature f2 (dotted blue) has the highest importance. Rows 2 and 5 (in red) would be selected by the pick procedure, covering all but feature f1.

Multiple Validation Steps

- Are explanations faithful to the model?
- Should I trust this prediction?
- Can I trust this model?
- Can users select the best classifier?

- Compare LIME against Parzen

Are explanations faithful to the model?

1. Train models so that max # of features for any instance is 10 (gold set)

2. Generate explanations and measure overlap with gold set.

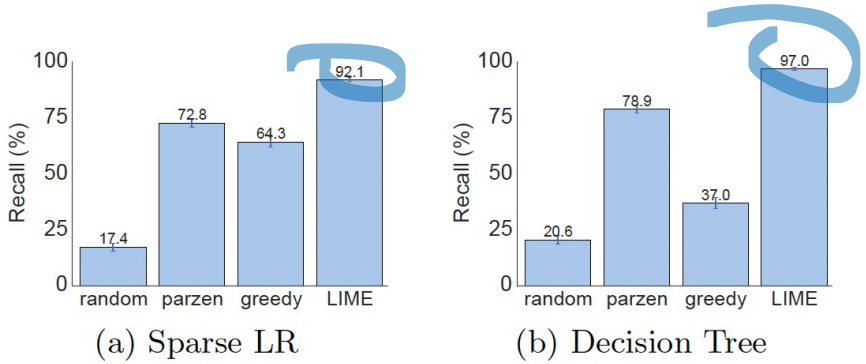


Figure 6: Recall on truly important features for two interpretable classifiers on the books dataset.

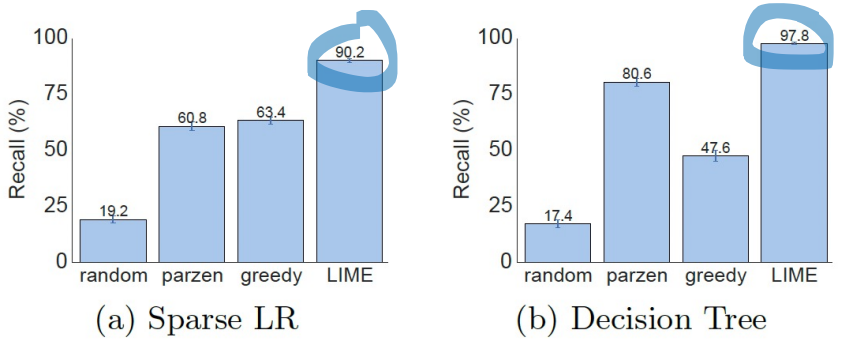


Figure 7: Recall on truly important features for two interpretable classifiers on the DVDs dataset.

Can users select the best classifier?

1. Compare two SVM classifiers:
   - One trained on the original dataset (Accuracy = 94%).
   - One trained on cleaned dataset (Accuracy = 88.6%).

2. Users are given words and documents to inspect and determine which classifier will perform best in the real world.

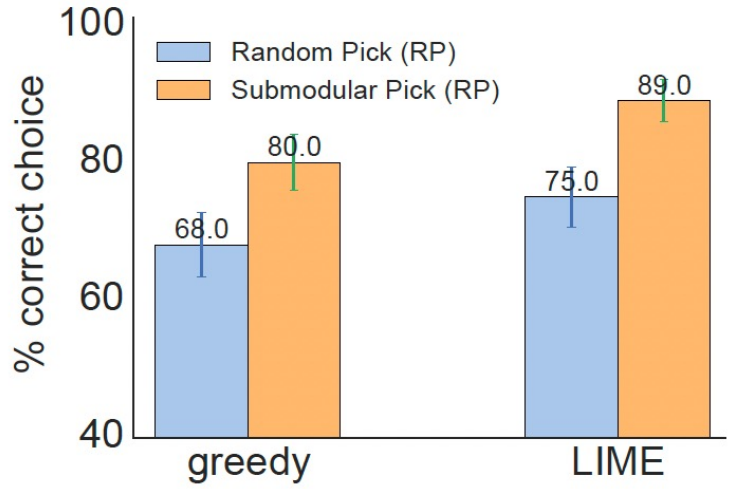


Figure 9: Average accuracy of human subject (with standard errors) in choosing between two classifiers.

Do you trust this algorithm?

(a) Husky classified as wolf (b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

Table 2: "Husky vs Wolf" experiment results.

LIME is available to use

Implementation is available in all major ML libraries

- ELI5
- Microsoft MLInterpret
- Lime library

Chapter 6 - Course Project

- Explaining your model
  - Use techniques to explore your best learned model
    - Interpretable models / Global explanations
  - Describe the most important features of your **production model**
  - Write some explanations based on your model exploration

- Explaining the predictions
  - Describe the techniques used to explain the prediction to the user
  - Justify the choice of the technique
  - (Extra) Include LIME as an alternative explanator for a prediction

Planning the end of the course

- **Fire on all cylinders** on the course project
  - Next week will be a “free week”
    - Special topic on Language Models
    - Quick revision
    - Course evaluation
  - Week after will be the exam

- You should focus in implementing the course project

References

- Explainable AI explained! | #3 LIME