Explainable Machine Learning

Explainable Machine Learning

Treatment Recommendation

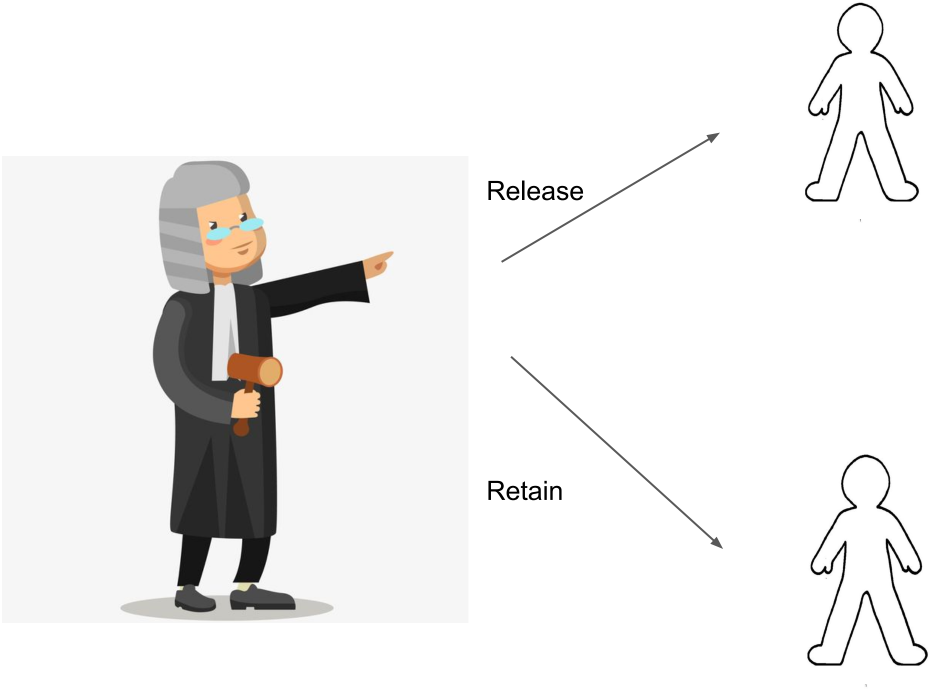
Which treatment should be given?
Options: quick relief drugs (mild),
controller drugs (strong)

Demographics: **age, gender, ..**
Medical History: **Has asthma?**

Symptoms: **Severe Cough, Sleepy**

Test Results: **Peak flow: Positive**

Bail Decision

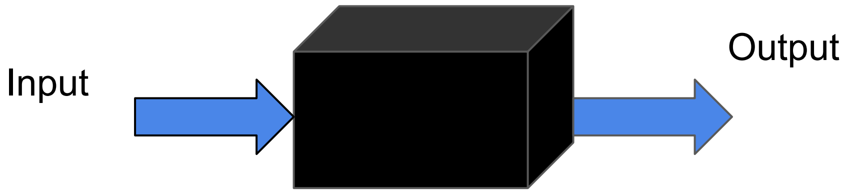


High-Stakes Decisions

- The above examples all belong to high-stakes decisions. The decisions have a **huge impact on human well-being**.

- What are those non high-stakes decisions?
    - Recommendations in E-commerces websites
    - When should I get up tomorrow?
    - …….

Black-Box Model



- If ML system is deployed in high-stakes decisions environment:
  - **Is accuracy important**?
  - Can we trust the machine learning model?

- In banking, insurance and other heavily regulated industries, model interpretability is a serious legal mandate.

- In lots of critical areas such healthcare, government, bioinformatics, etc, rationale for models' decision is necessary for trust.

Why do we need model explainability?

- Use Machine Learning to review resumes
  - Based on your capability or gender?
  - https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G
- Use Machine Learning to detect fraud transactions?
  - Why does the model think this transaction is suspicious?

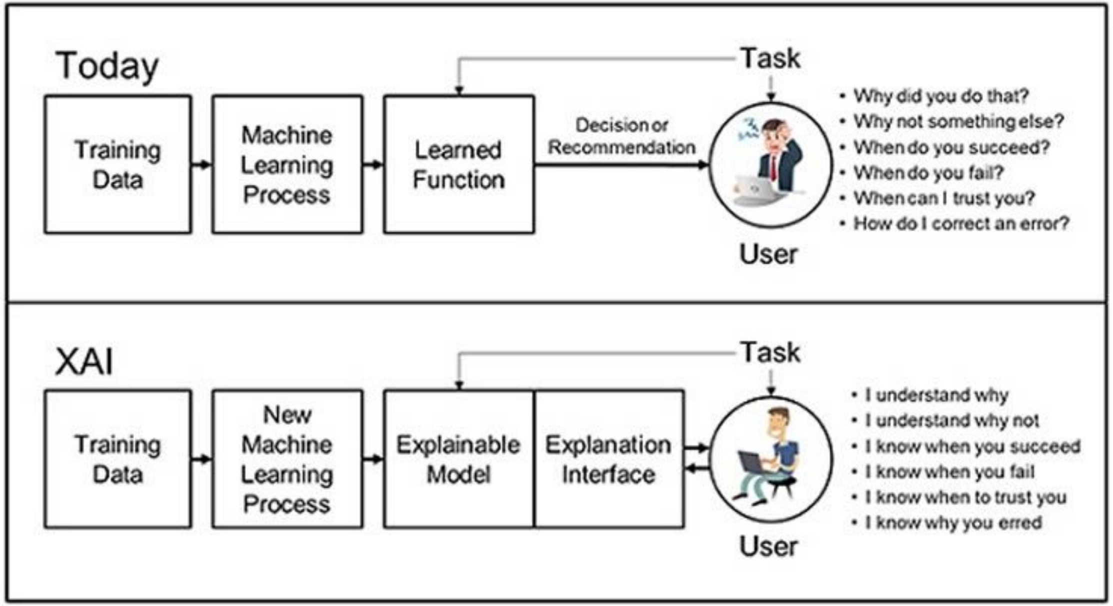
Goals of Explainability

- Build Trust in machine learning systems’ predictions
  - Reveal model behavior
  - Justify model predictions
  - Assist users in investigating uncertain predictions

- Allow users to provide useful feedback, which in turn can help developers improve model quality.

XAI

- **XAI**: ML models are explainable that enable end users to understand, appropriately trust, and effectively manage the emerging generation for AI systems.



DARPA's report

Interpretability vs Accuracy

Linear Models First

- Prediction is the linear combinations of the features values, weighted by the model coefficients.

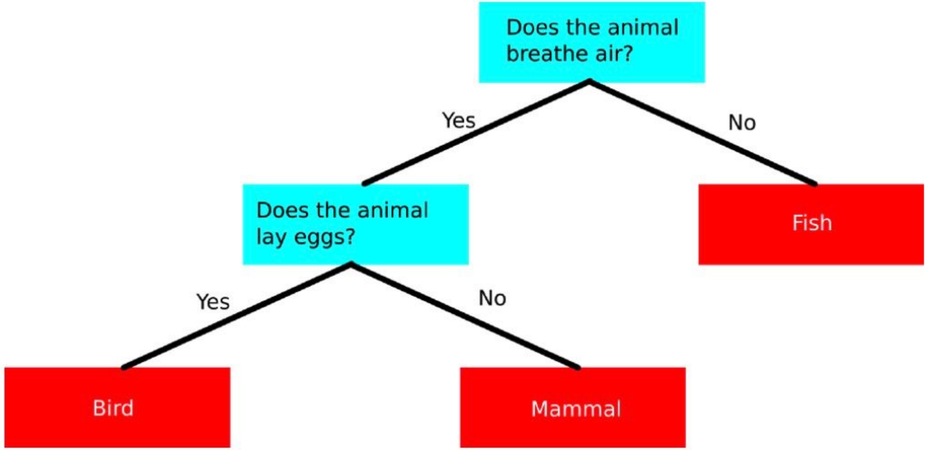
Students A’s chance = 0.2 + 0.1* GPA - 0.005 * Hours on Tiktok

One student’s feature		Feature Contributions
GPA: 5	→	GPA: 0.5
Hours on Tiktok: 60 hours		Hours on Tiktok: -0.3

- Capability of linear models is limited.

Decision Tree

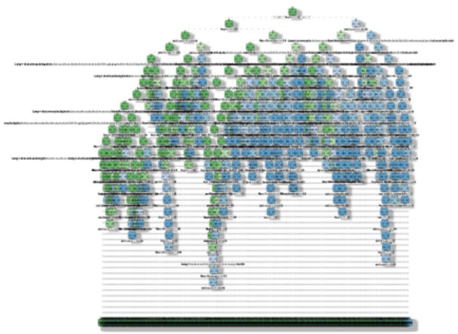
- It is “interpretable”
- More powerful compared to linear models.



Source:
https://towardsdatascience.com/a-beginners-guide-to-decision-tree-classification-6d3209353ea

Decision Tree can be complex

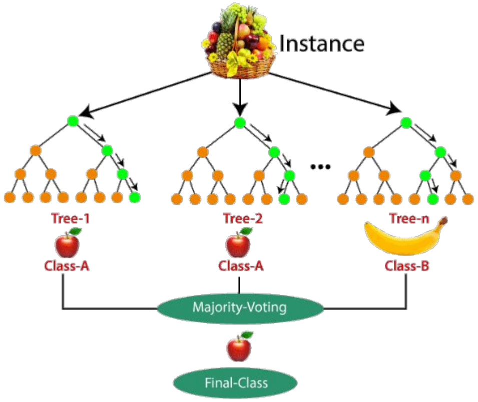
It can be a huge and complex tree.



My goal is to extract some useful rules from the entire process to implement in a score card.

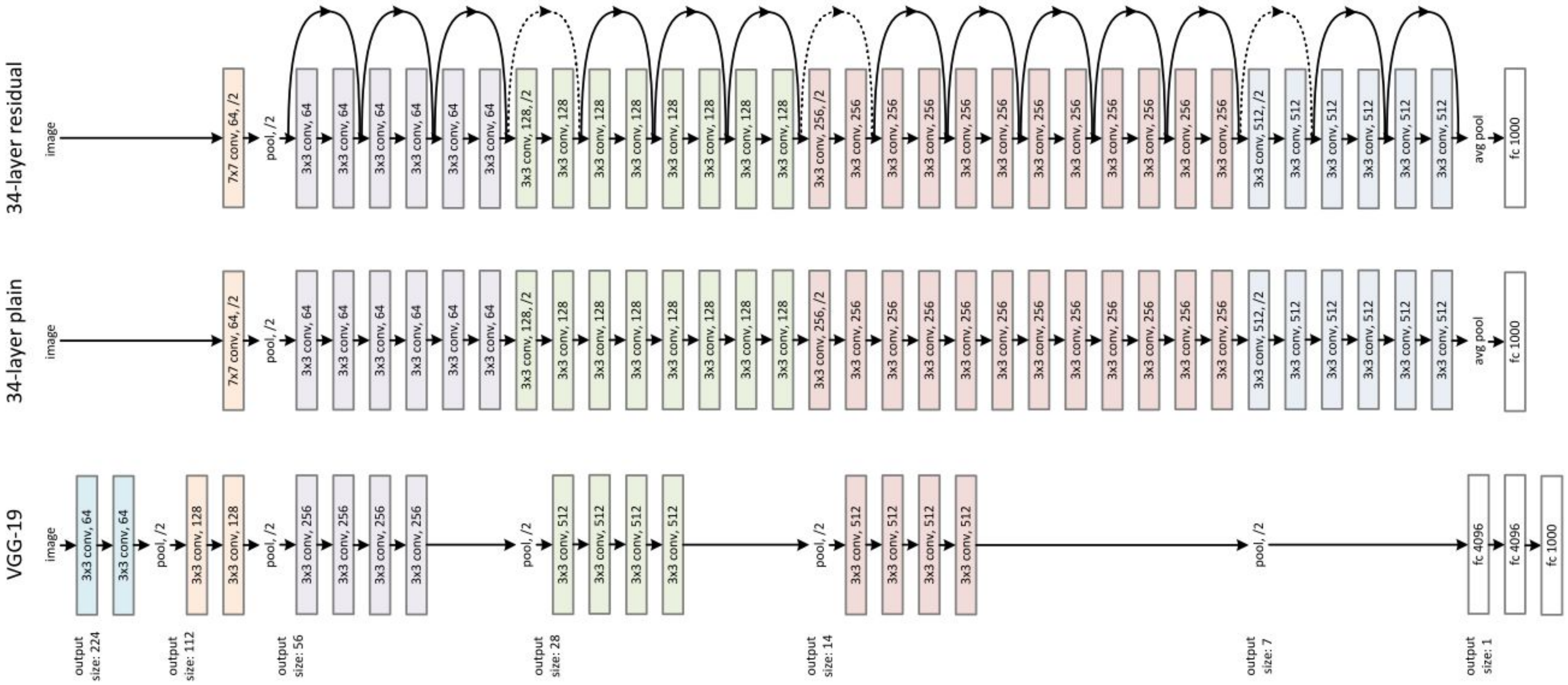
https://stats.stackexchange.com/questions/230581/decision-tree-too-large-to-interpret

It can be a forest



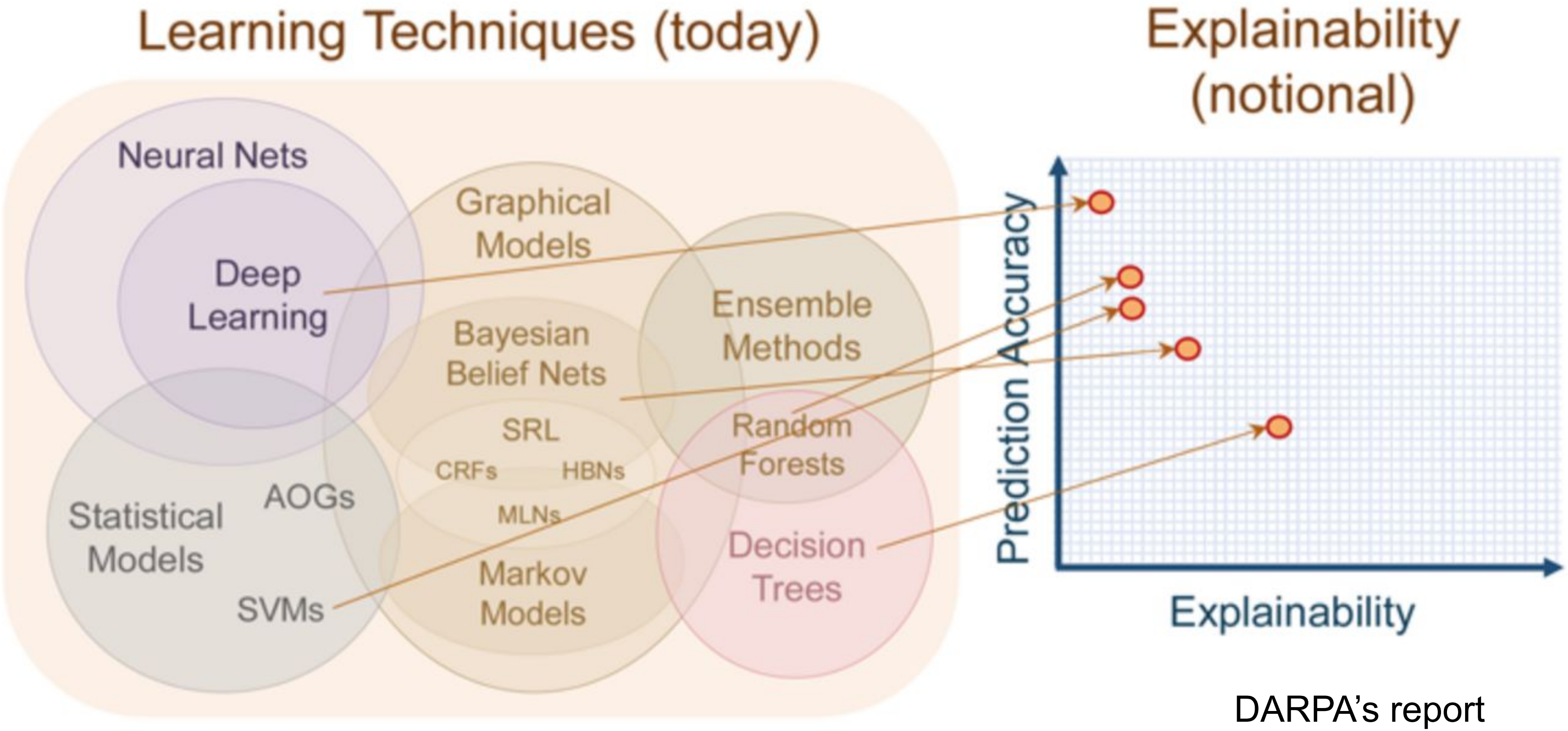
https://www.javatpoint.com/machine-learning-random-forest-algorithm

Complex Models



For imagenet, they use 152 layers, which firstly achieved lower error rate compared to Humans in image recognition tasks.

Trade-off



DARPA’s report

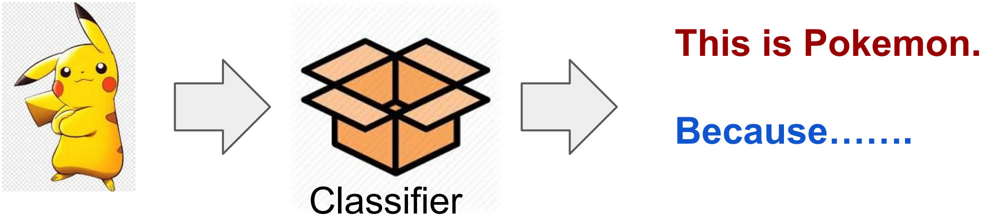
Categorization of Explanations

Categorization of Explanations I

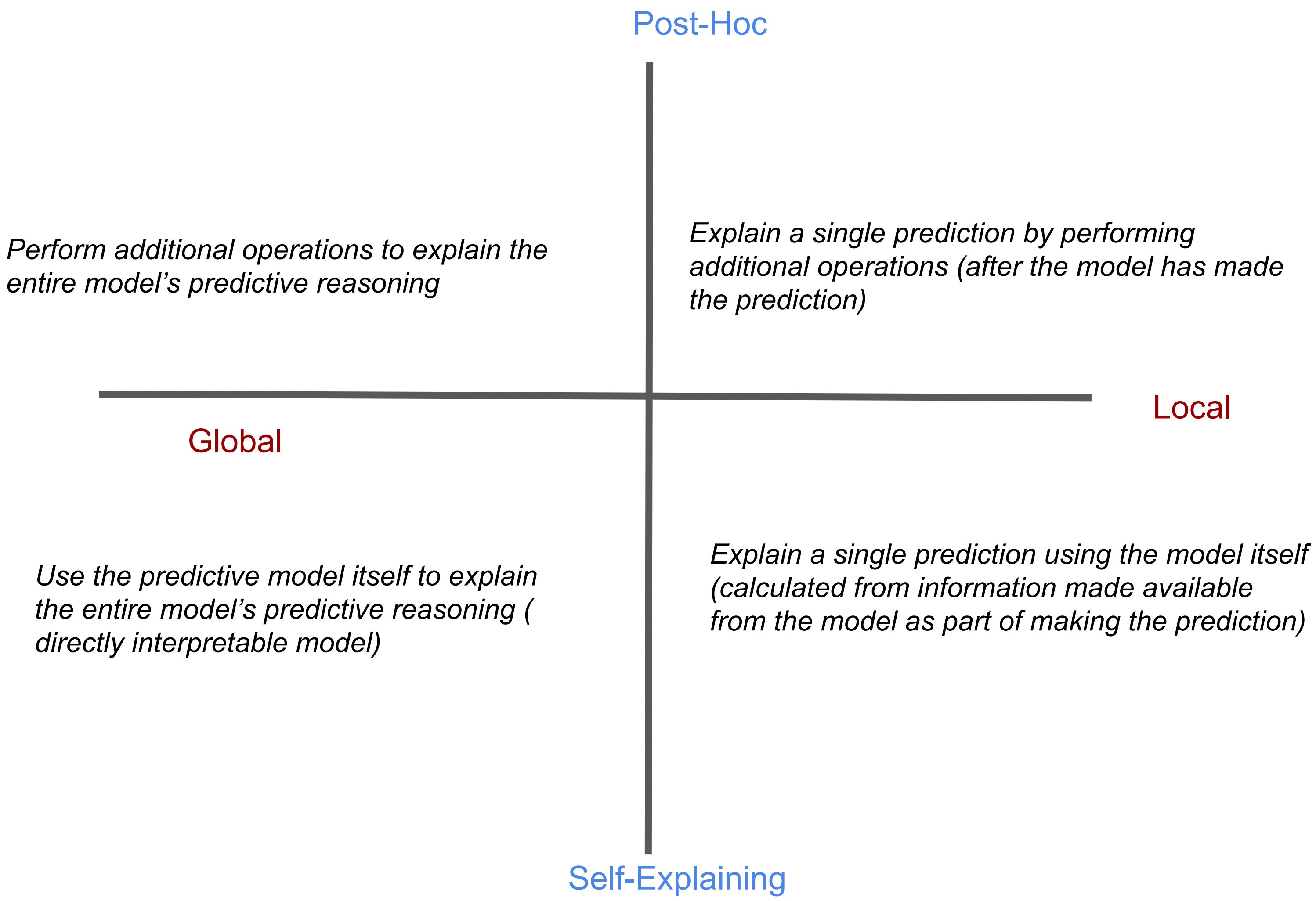
- Self-Explaining
  - Directly interpretable
  - Generates the explanations at the same time as the prediction
  - Rule-based System, Decision Trees, Logistic Regression, Hidden Markov Model, etc.

- **Post hoc**:
  - Additional operation is performed after the predictions are made
  - Open-source packages: tf-keras-vis (gradient-based methods for deep learning), LIME, SHAP, etc

Categorization of Explanations II



- Global:
  - Explanation or justification by revealing how the model's predictive process works.
  - **What do you think pokemon looks like?**
- **Local:**
  - Provide information or justification for the model's prediction on **a specific input**
  - **Why do you think this image is pokemon?**


Post-Hoc
Perform additional operations to explain the entire model’s predictive reasoning
Explain a single prediction by performing additional operations (after the model has made the prediction)
Global
Local
Use the predictive model itself to explain the entire model’s predictive reasoning (directly interpretable model)
Explain a single prediction using the model itself (calculated from information made available from the model as part of making the prediction)
Self-Explaining

Explainability Techniques

Example-driven

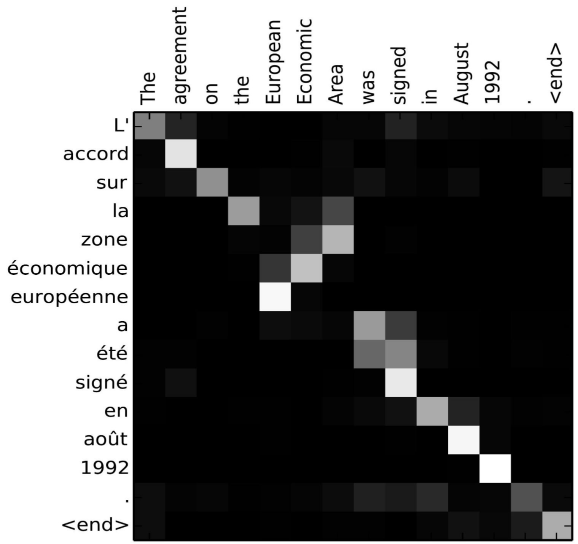
- Reasoning with examples
  - Explain the prediction of an input instance by identifying and presenting other instances.
- Eg., Patient A has a tumor because he is similar to these k other data points with tumors
- Similar to nearest neighbor-based approaches

Feature Importance

- Derive explanation by investigating the importance scores of different features used to output the final prediction

- It can be computed from
  - Attention Layer Approach

  - **Gradient-based Saliency Approach**



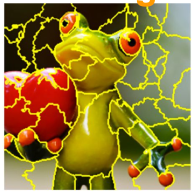
https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html

Feature Importance: Gradient-based Method

- Explain the decision made by the model
  - Eg, Why do you think this image is pokemon not digimon?

- Motivation: we want to know the contribution of each **component/feature** in the input data for prediction

Pixel, Segment in Images



Word in text

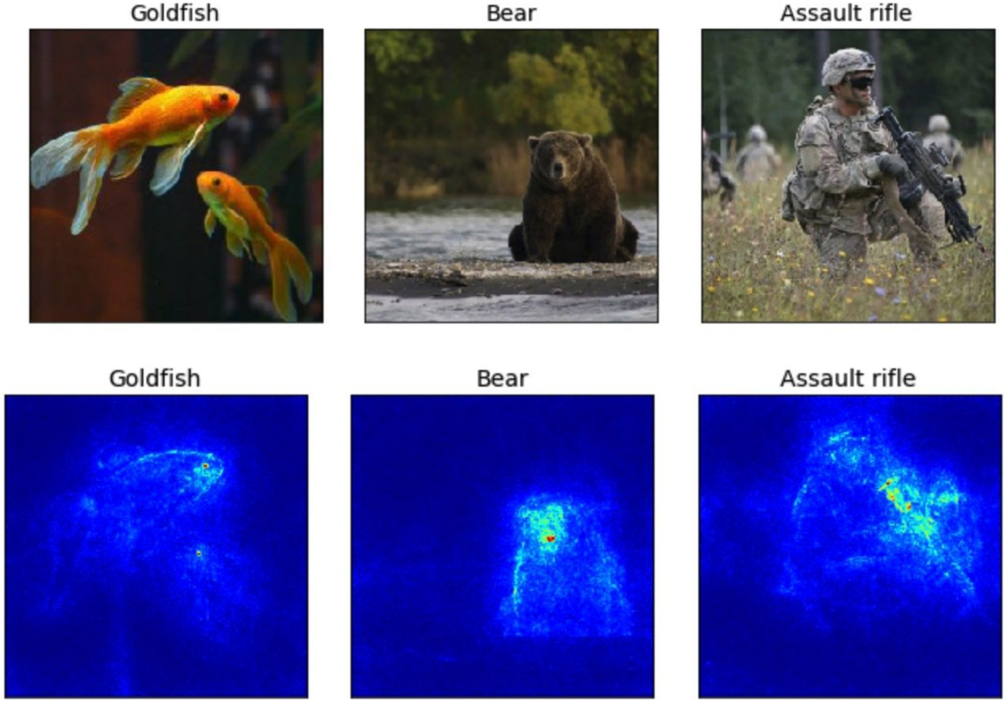
This is BT4012.

- Solution: Removing or modifying the partial parts of the components, observing the change of decision.

Saliency Map

$$\{x_1, \ldots, x_i, \ldots, x_n\} \qquad \{x_1, \ldots, x_i + \Delta x, \ldots, x_n\}$$

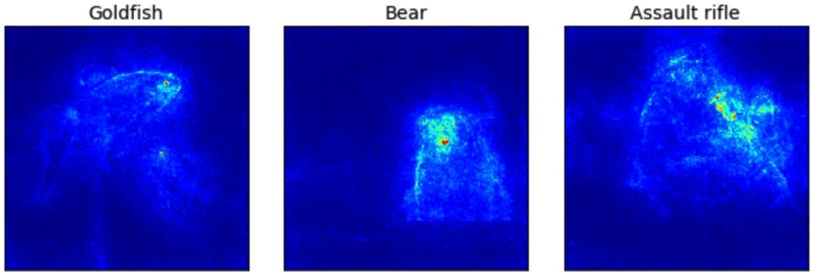
$$y_k \qquad y_k + \Delta y$$

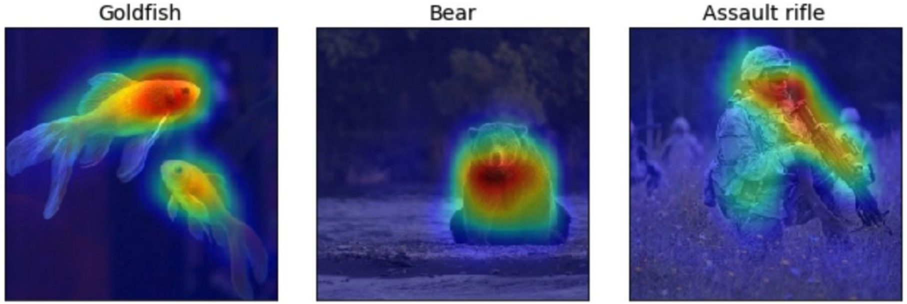


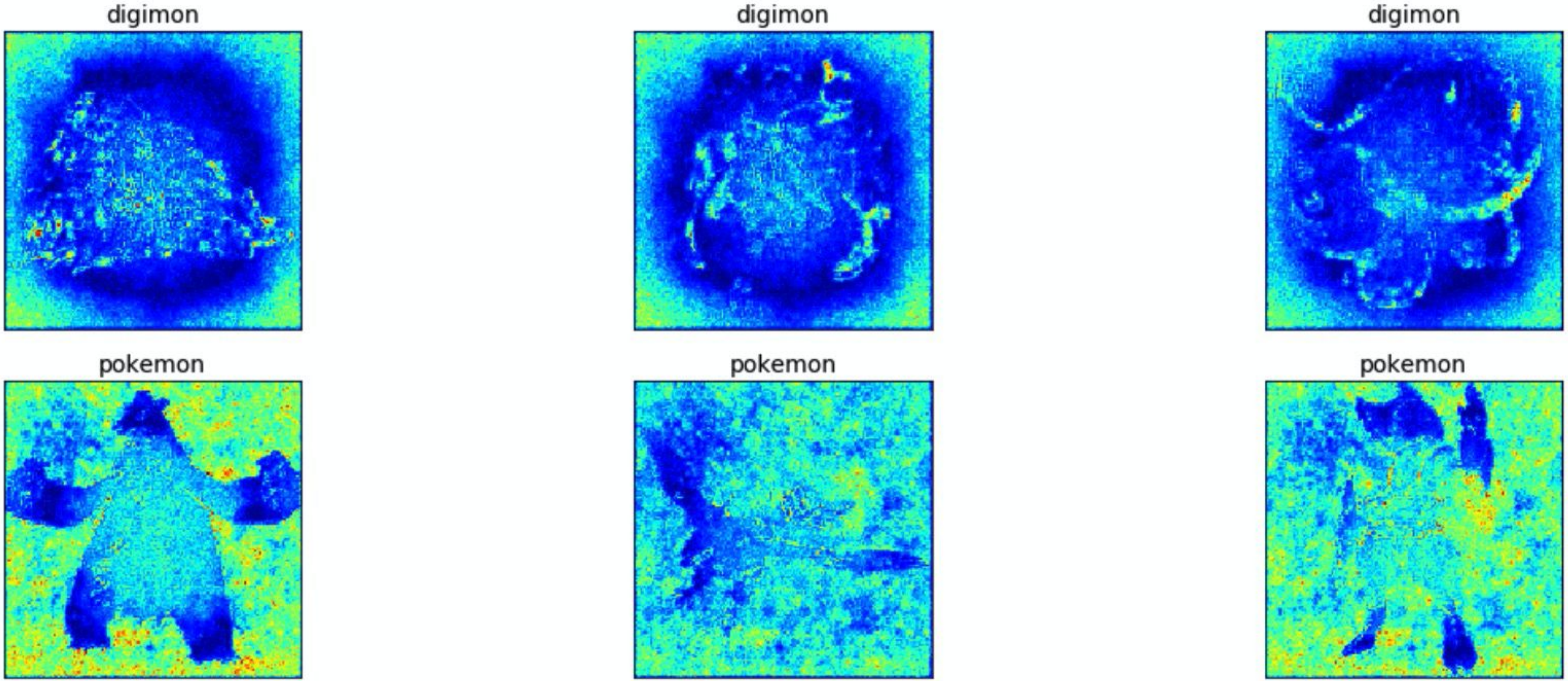
$$\left|\frac{\Delta y}{\Delta x}\right| \qquad \left|\frac{\partial y}{\partial x}\right|$$

Andrea Vedaldi, Andrew Zisserman, “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps”, ICLR, 2014

Saliency Map



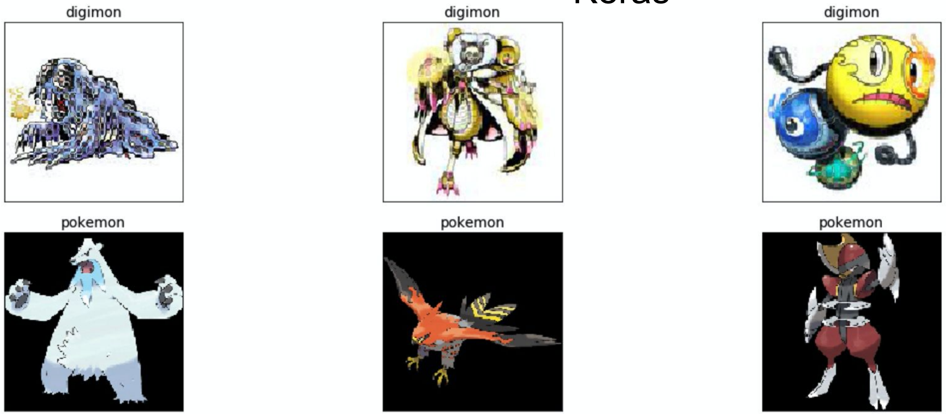



digimon
digimon
digimon
pokemon
pokemon
pokemon

Pokemon Image Type

Digimon Image Type

Loaded by Keras



CNN only learns to classify pokemon and digimon based on background colors.

PNG all appear in full black background

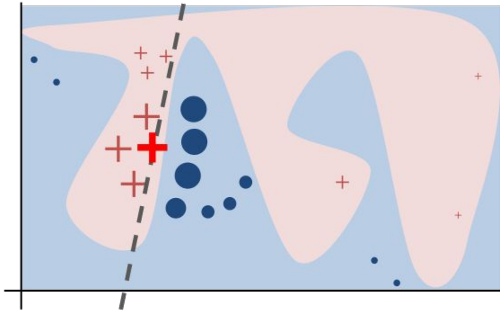
https://stackoverflow.com/questions/55720576/keras-loaded-png-appears-full-black

Surrogate Model

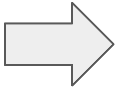
- Model predictions are explained by learning a second, usually more explainable model, as a proxy
- Model-agnostic (applicable for any machine learning models)
- The learned surrogate models and the original models may have completely different mechanisms to make predictions.

Surrogate Model: Local Explanations

- Hard to explain a complex model in its entirety
  - How about explaining smaller regions?



LIME (Ribeiro et. al)

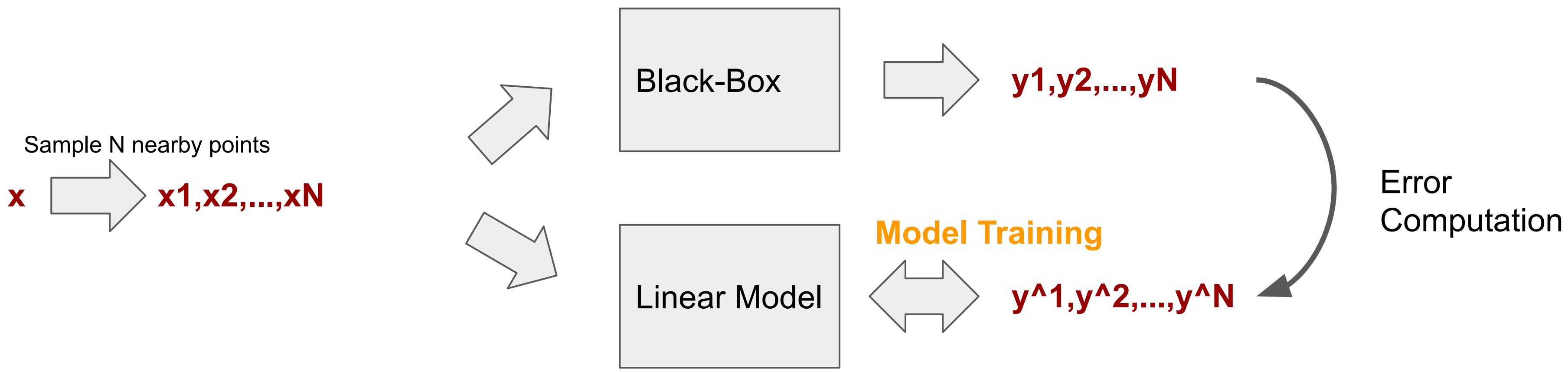


Linear model can not mimic neural networks..but it may mimic a local region

  - Explains decisions of any model in a local region around a particular point
  - Learns sparse linear model

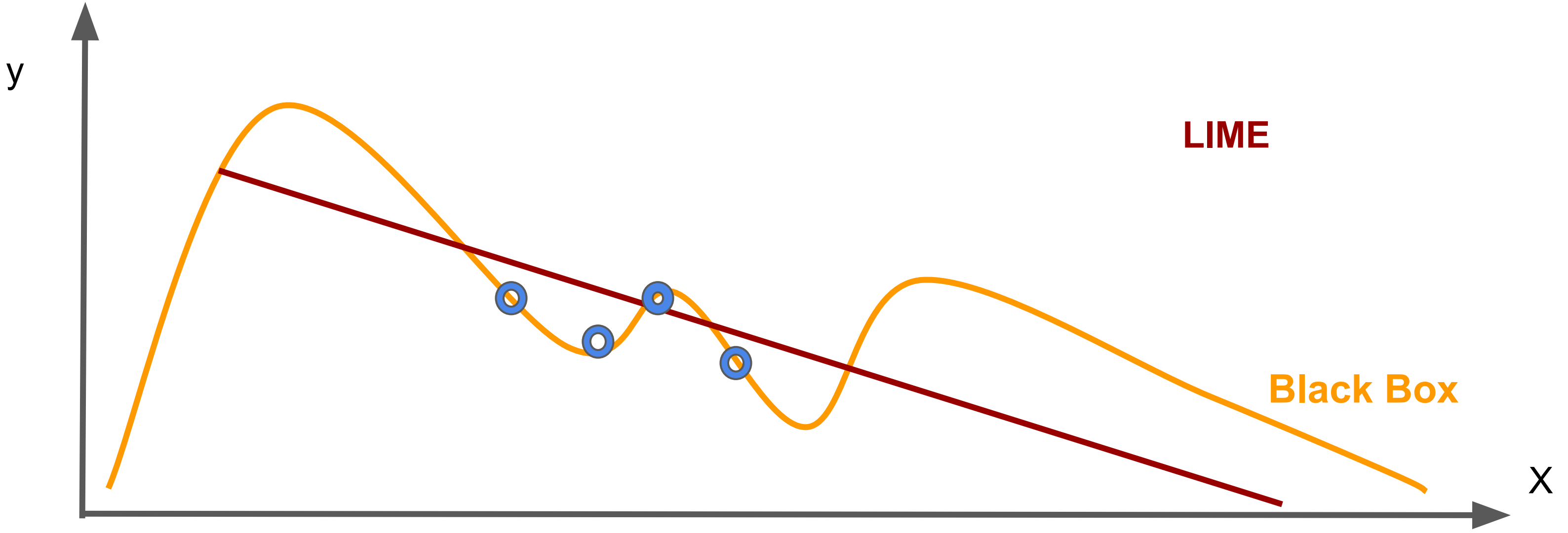
Surrogate Model: Local Explanations

- Interpretable model can be used to mimic the behavior of an complex model



Local Interpretable Model-Agnostic Explanations

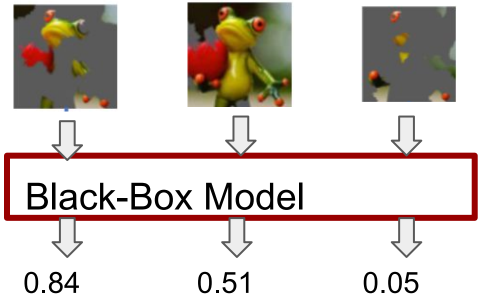
- Given a data point you want to explain
- Sample at the nearby
- Fit with linear model (or other interpretable models)
- Interpret the linear model



LIME on Image

- Given a data point you want to explain

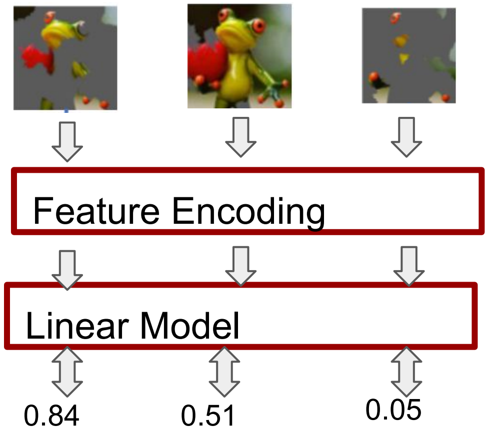
- Sample at the nearby
  - Each image is represented as a set of superpixels (segments)
  - Randomly delete some segments



Compute the probability of “frog” by black box

LIME on Image

- Fit with linear model



Indicating whether the segment is removed or not

$$x_1^i \dots x_m^i \dots x_M^i$$

Indice of nearby data samples

The number of segments

$$x_m^i = \begin{cases} 0 & \text{if segment m in sample i is deleted} \\ 1 & \text{if segment m in sample i exists} \end{cases}$$

LIME on Image

- Interpret the linear model

$$y = w_1 x_1 + \cdots + w_m x_m + \cdots + w_M x_M$$

$$x_m^i = \begin{cases} 0 & \text{if segment m in sample i is deleted} \\ 1 & \text{if segment m in sample i exists} \end{cases}$$

$$w_m \approx 0$$

$$w_m > 0$$

$$w_m < 0$$

Summary

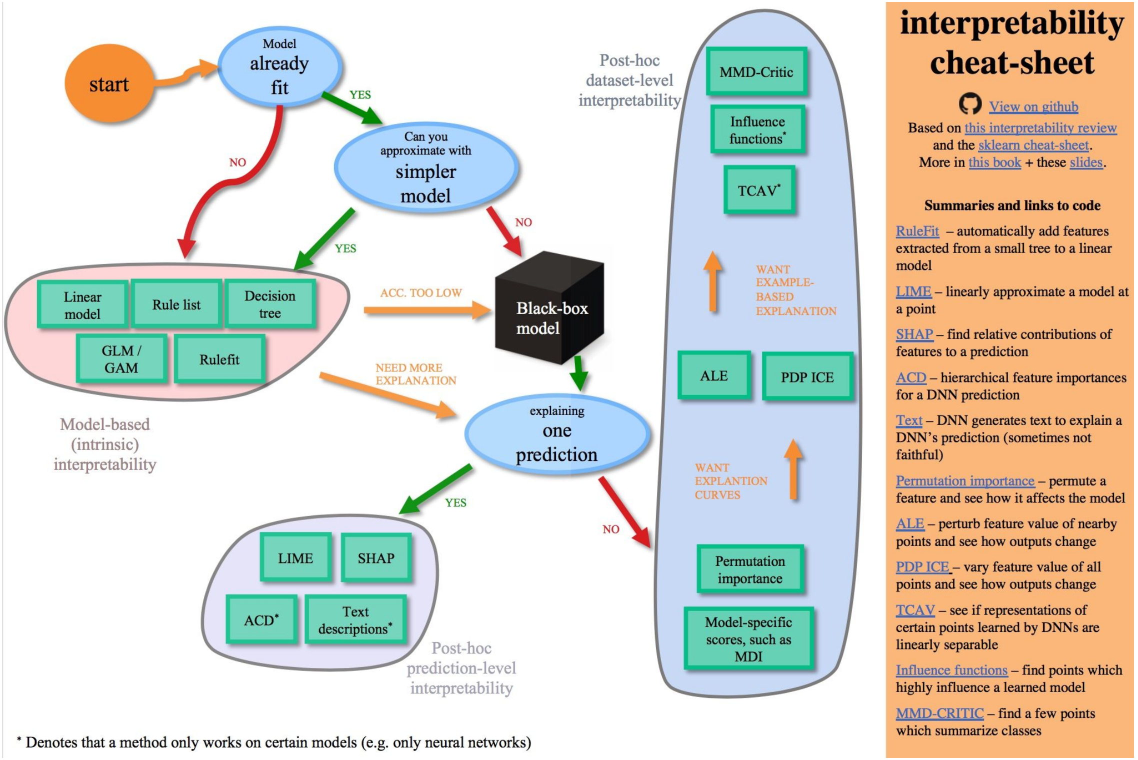
Misleading Explanations

- Do not blindly embrace explanations!

- Those above explanations can seems to be plausible but misleading
  - They do not claim to open up the black-box;
  - They only provide plausible explanations for its behavior

Future Directions

- Goal of ML Explanation is not to completely know how the ML model works
  - We also do not know how our human brain work
- There is a need for clearer terminology and understanding of what constitutes explainability and how it connects to the target users
- How do we evaluate the quality of explainability?
  - Since this topic is quite new, there is little agreement on how explanations should be evaluated.



Source: https://github.com/csinva/csinva.github.io/blob/master/_notes/cheat_sheets/interp.pdf

Interpretable Machine Learning:
https://christophm.github.io/interpretable-ml-book/