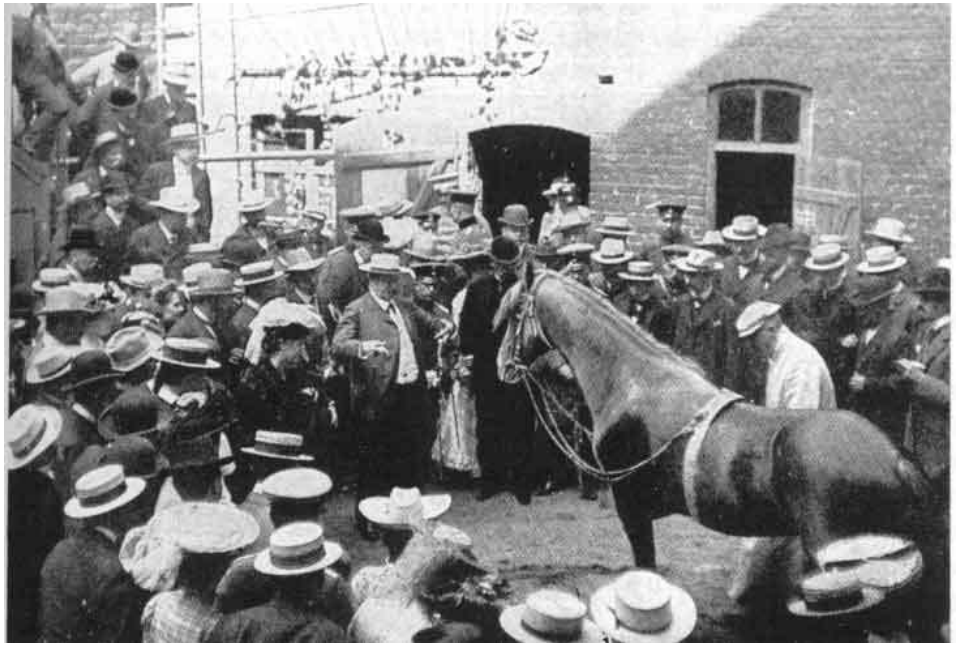
Explainable ML

Introduction to AI
May 15, 2023

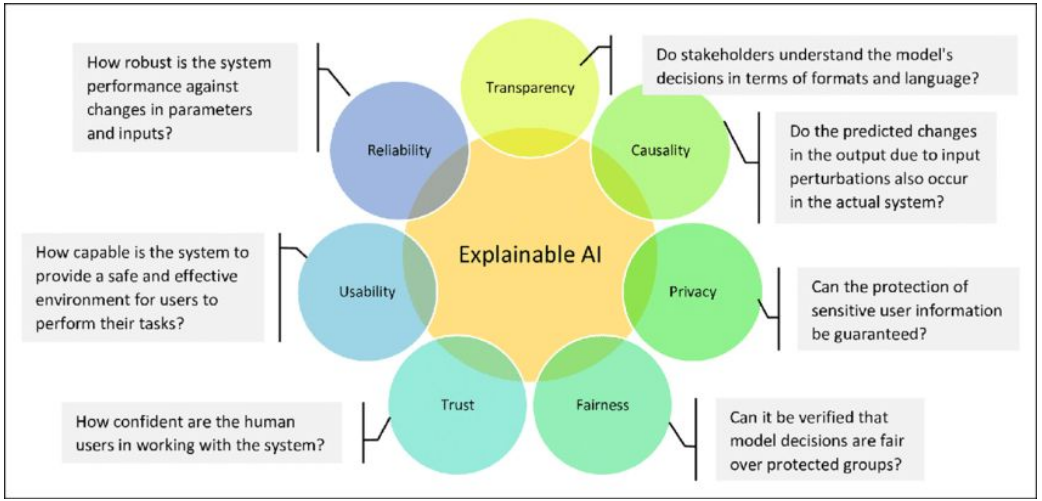
Why We Need Explainable ML?

- **Correct answers ≠ Intelligent**
- EX. Clever Hans



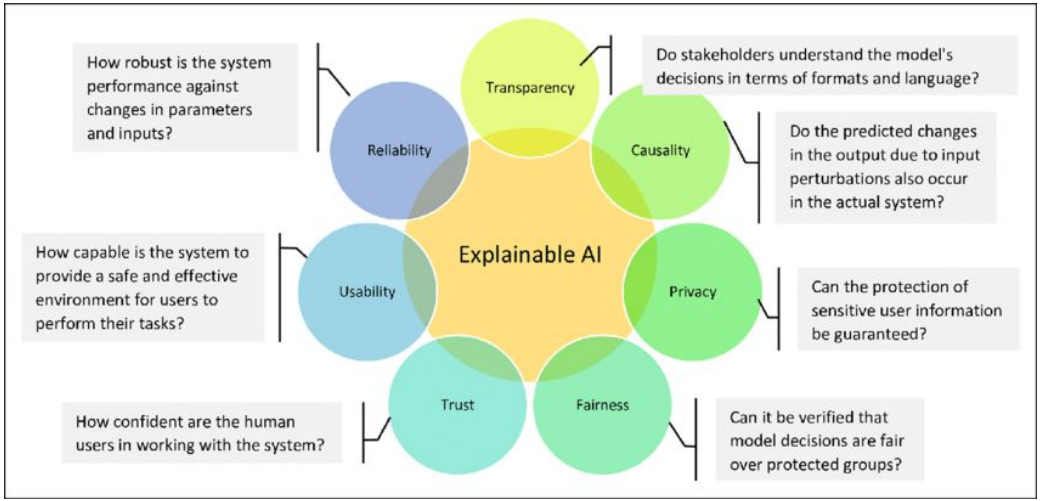
Why We Need Explainable ML?

- Medical Diagnosis
- Asset Valuation
- Verdict
- …



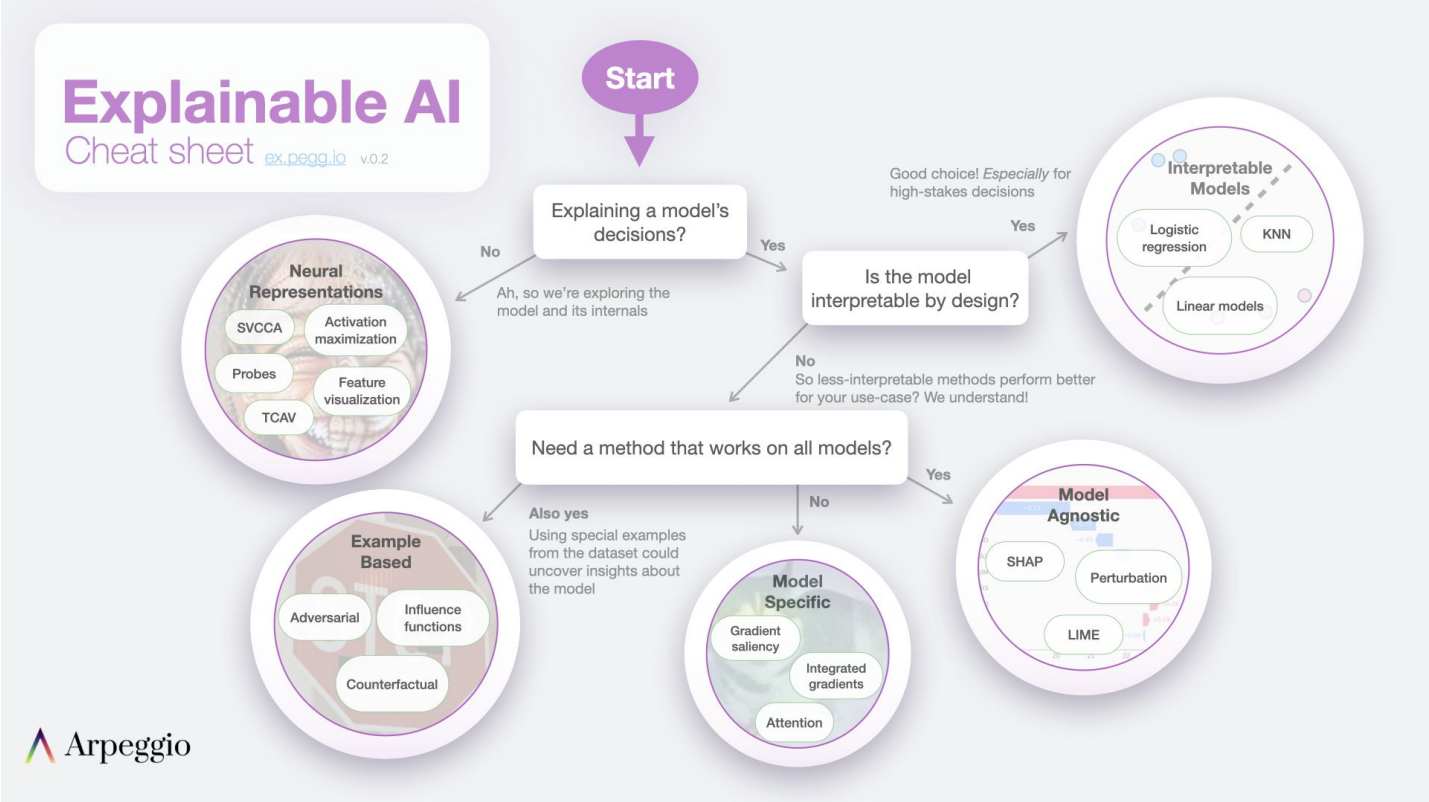
Why We Need Explainable ML?

- Medical Diagnosis
- Asset Valuation
- Verdict
- …



Why is the model performing poorly?

Approach in Explainable ML



Approach in Explainable ML

Agnosticity

Model Agnostic
Applicable to all model type

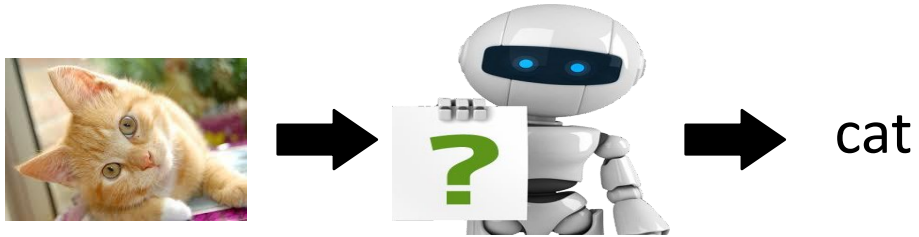
Model Specific
Only applicable to a specific model type

Scope

Global Explanation
Explaining the whole model

Local Explanation
Explaining individual predictions

Explainable ML: Global vs Local



- ***Local Explanation***

  Why do you think *this image* is a cat?

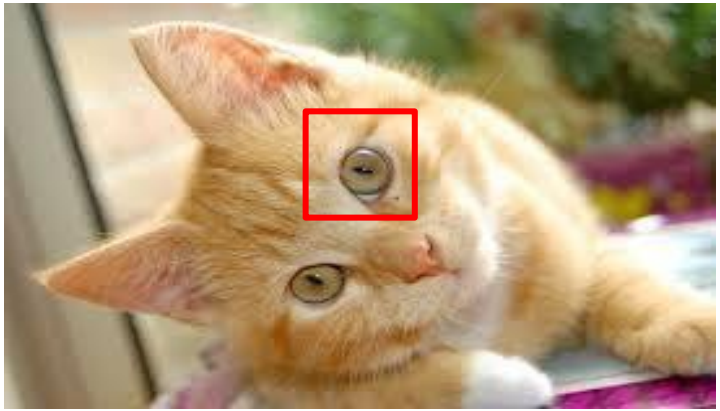
- ***Global Explanation***

  What does a “cat” look like? (not referred to a specific image)

Explainable ML: Local Approach

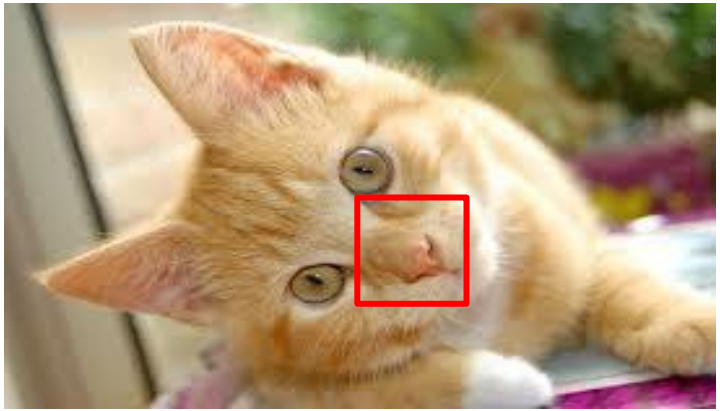
Which component is critical for making decision?

Explainable ML: Local Approach



Which component is critical for making decision?

Explainable ML: Local Approach



Which component is critical for making decision?

Explainable ML: Local Approach

Which component is critical for making decision?

Object x ⟶ Image, text, etc.

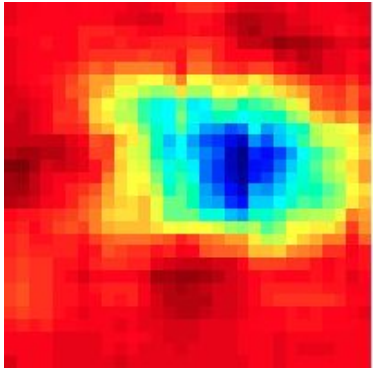
Components:

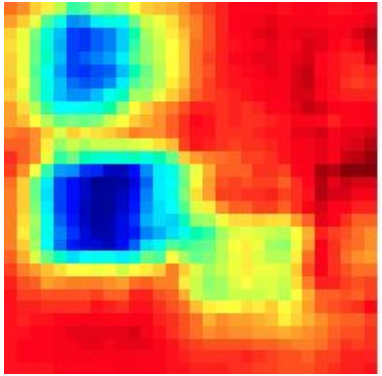
Image: pixel, segment, etc.
Text: a word

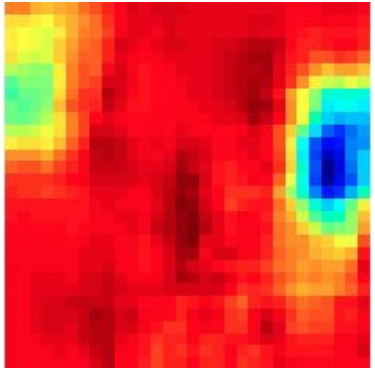
- Removing or modifying the components
- Large decision change

⟹ Important component

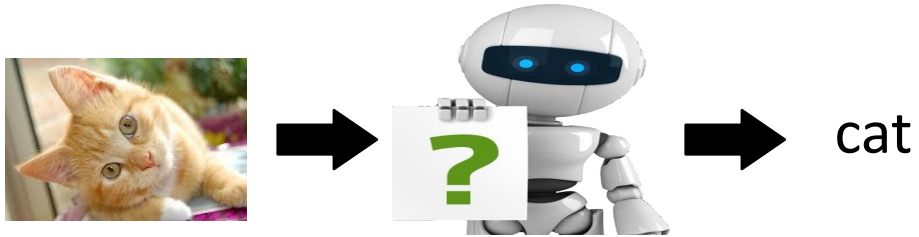
Explainable ML: Local Approach







Explainable ML: Global vs Local



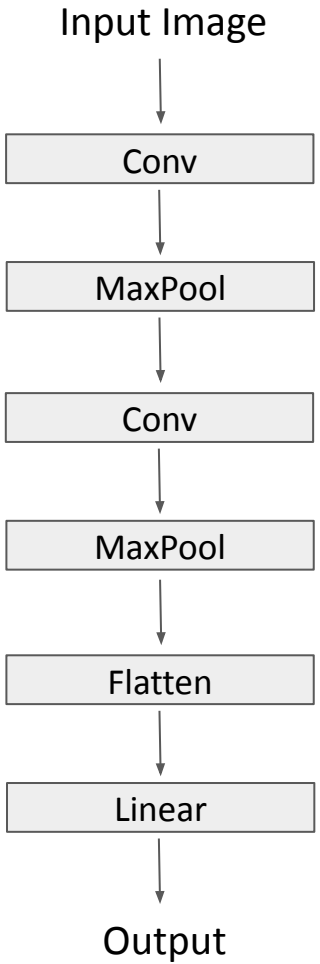
- ***Local Explanation***

  Why do you think *this image* is a cat?

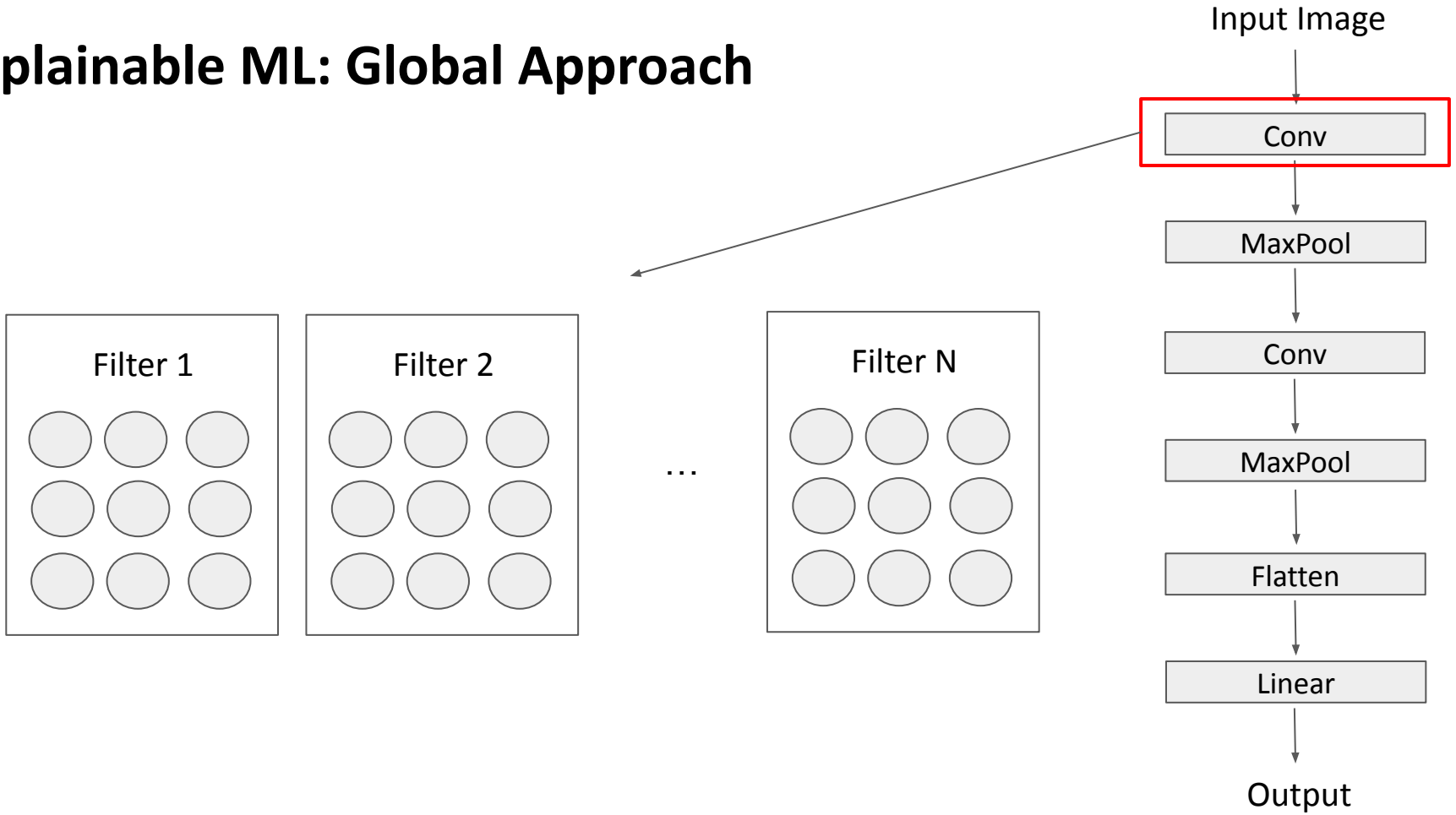
- ***Global Explanation***

  What does a “cat” look like? (not referred to a specific image)

Explainable ML: Global Approach

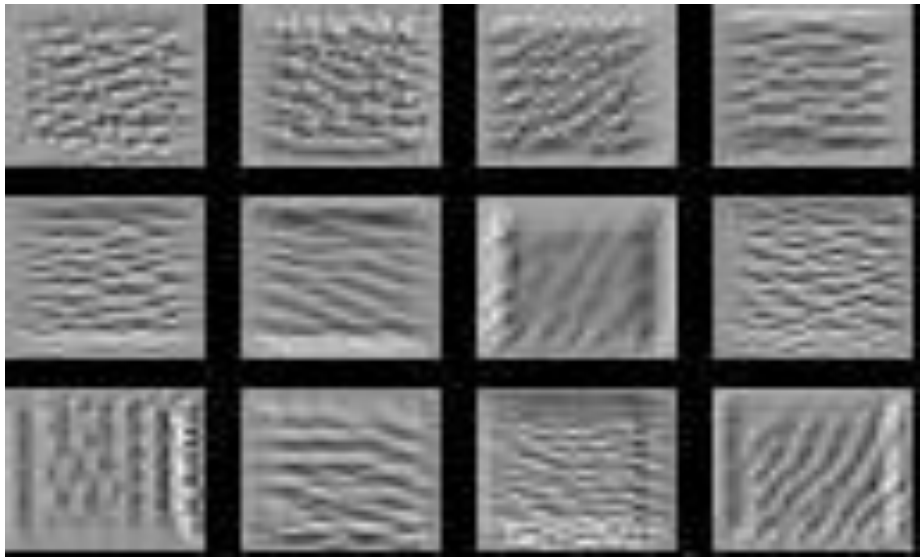


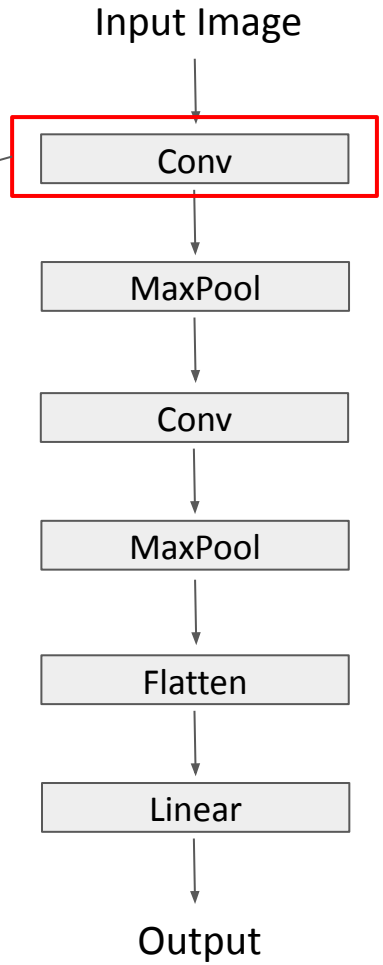
Explainable ML: Global Approach



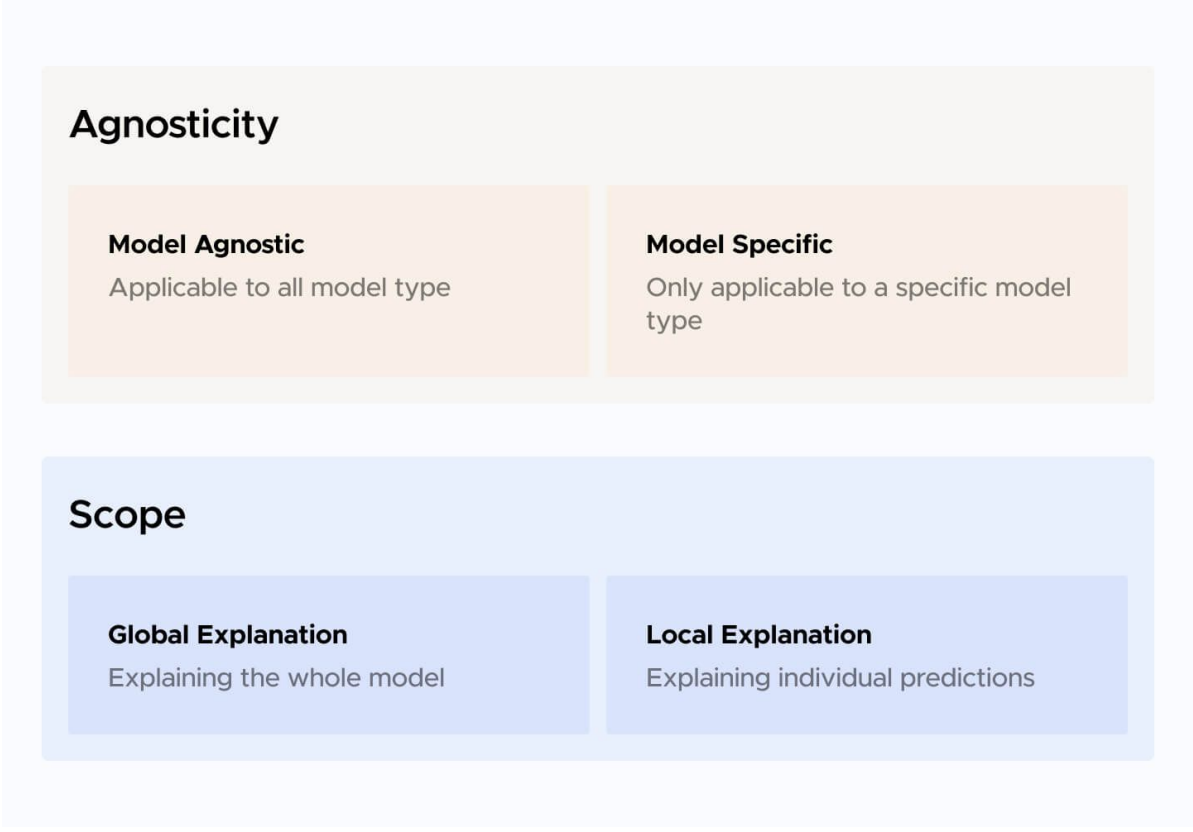
Explainable ML: Global Approach

E.g., Digit classifier

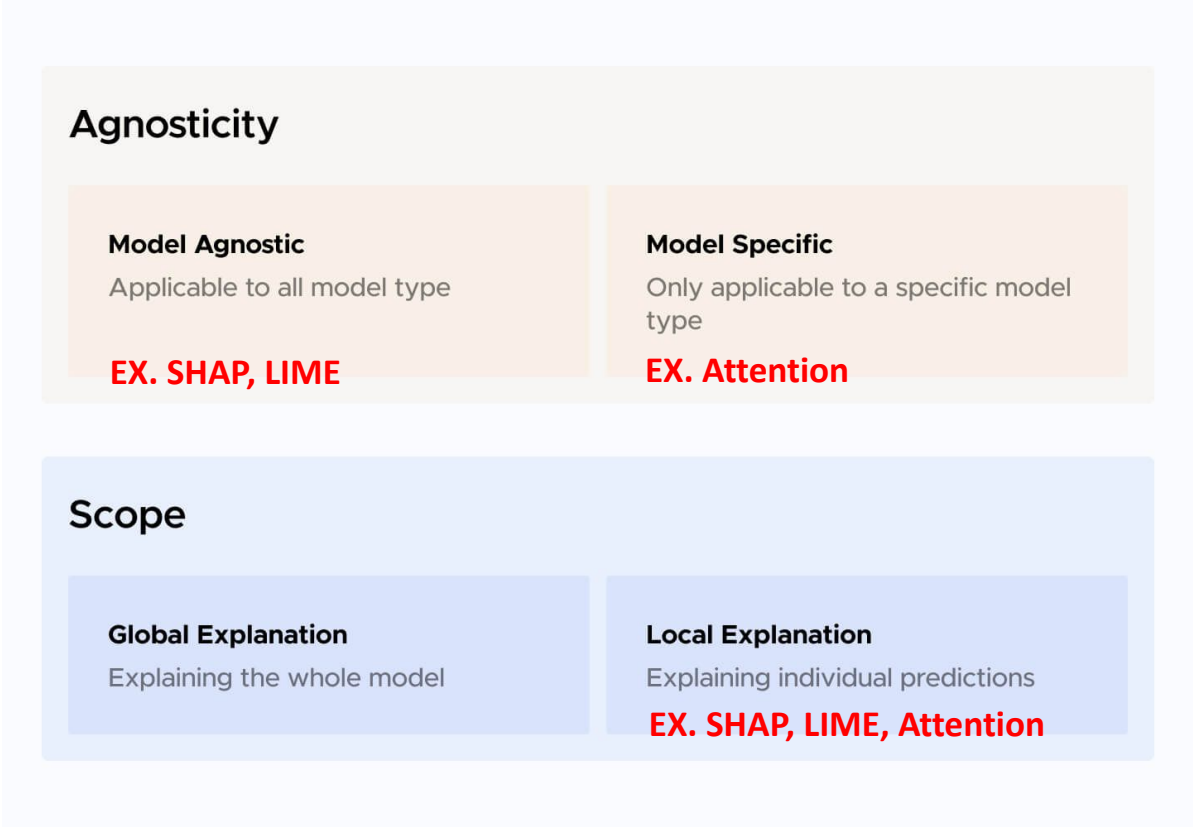




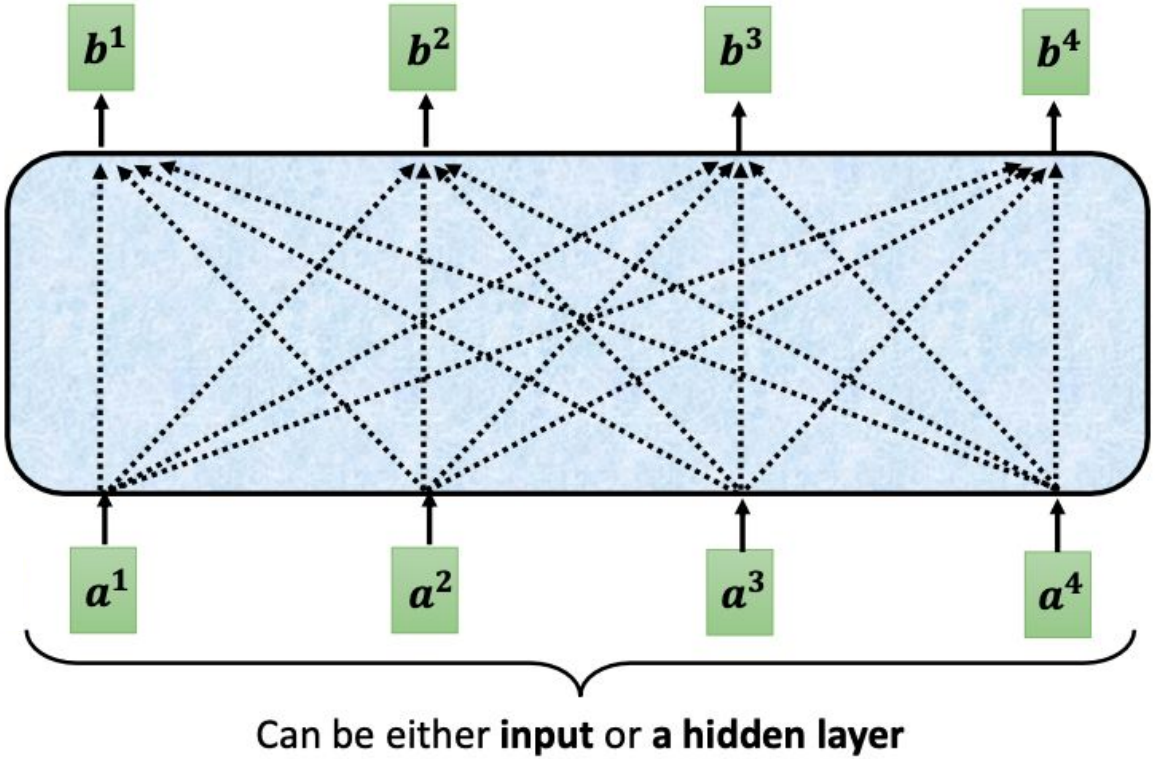
Approach in Explainable ML



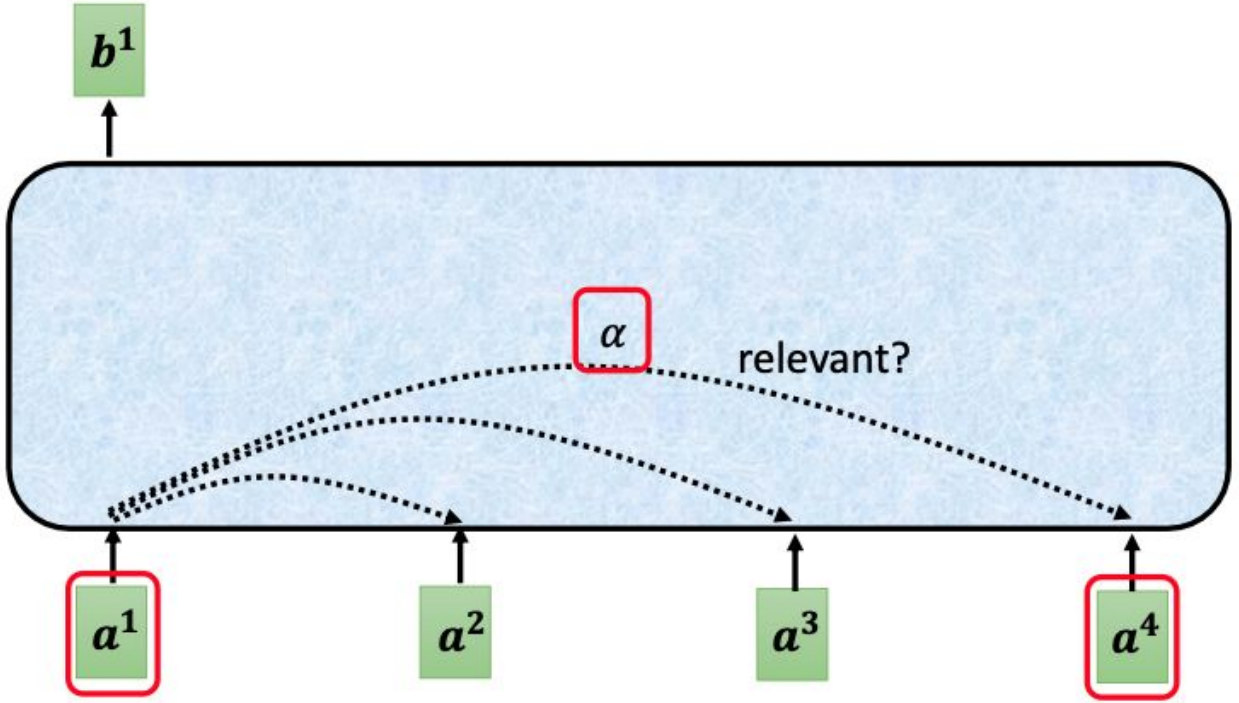
Approach in Explainable ML



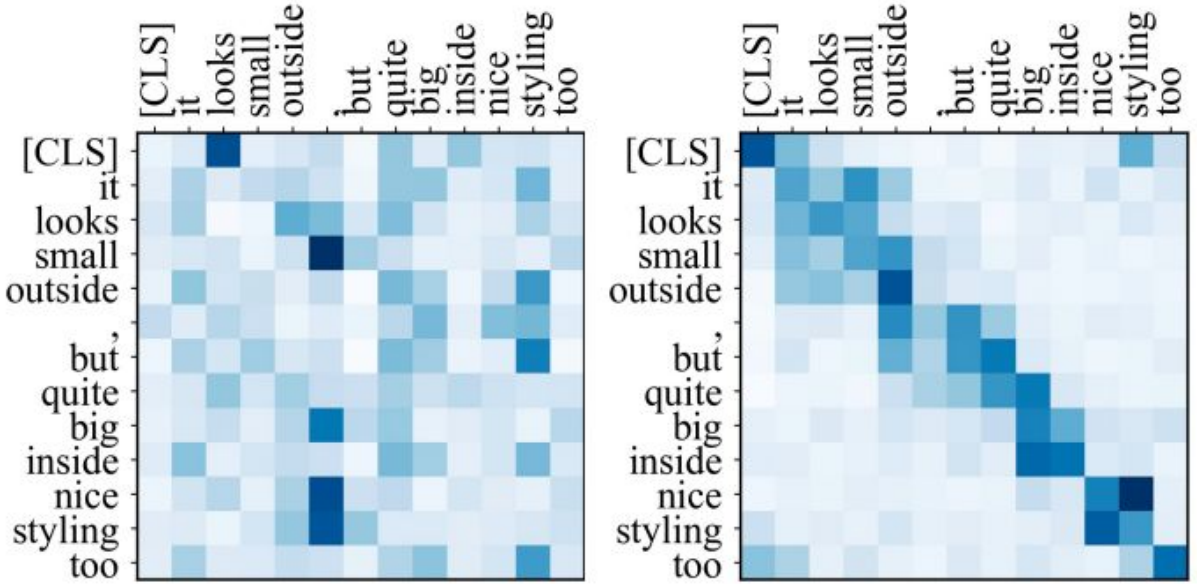
Explainable ML: Attention



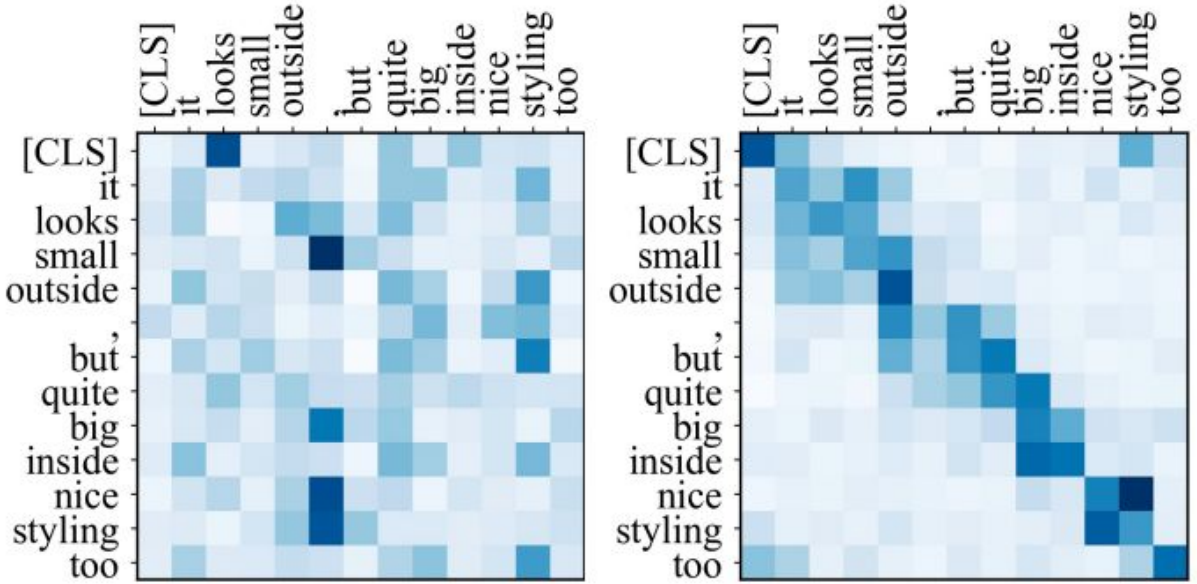
Explainable ML: Attention



Explainable ML: Attention



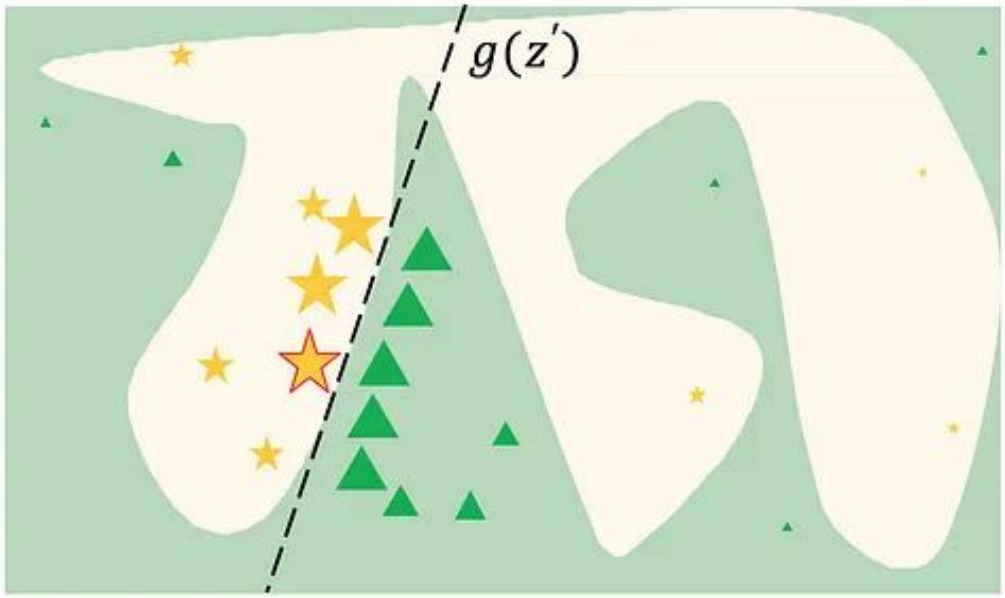
Explainable ML: Attention



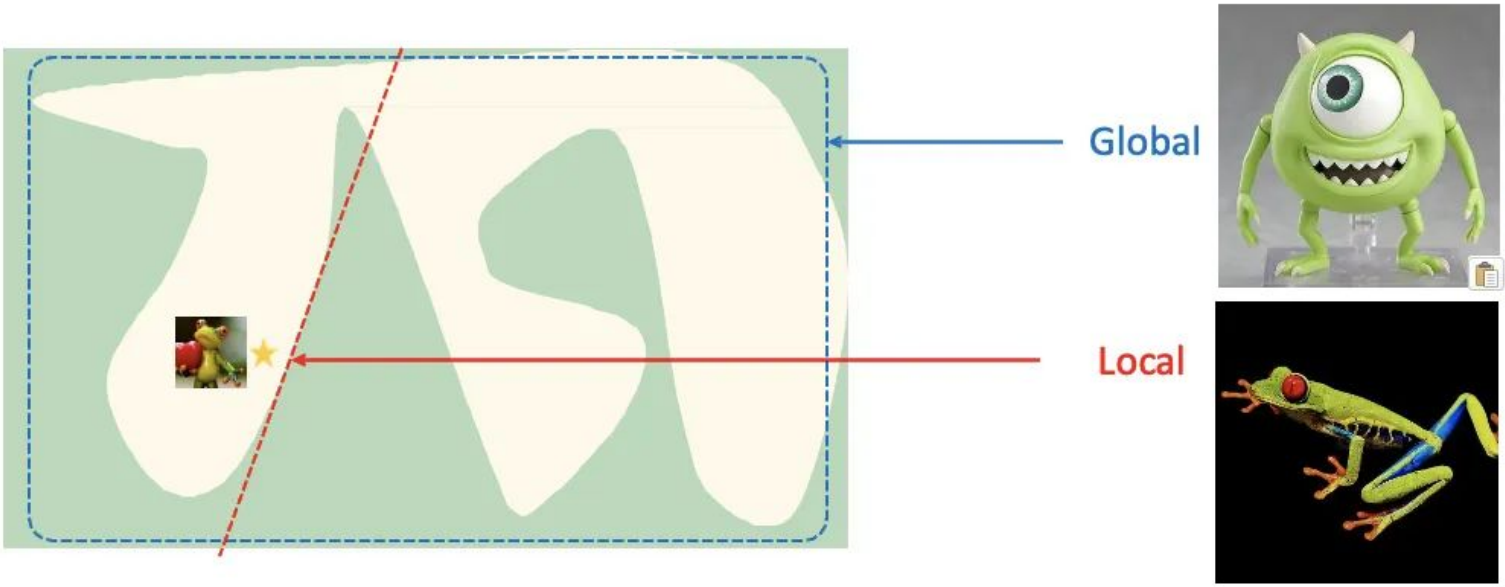
Explainable ML: Attention

exBERT: Attention Visualization

Explainable ML: LIME (Local Interpretable Model-agnostic Explanations)



Explainable ML: LIME (Local Interpretable Model-agnostic Explanations)



Explainable ML: LIME (Local Interpretable Model-agnostic Explanations)

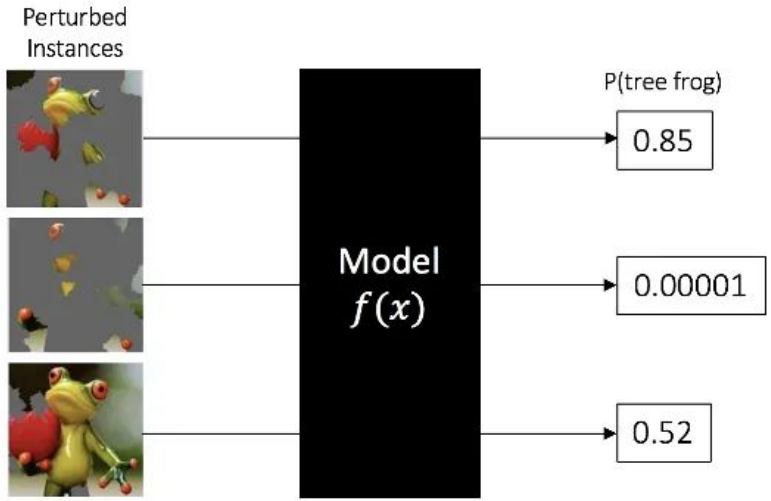
原始圖片(RGB)
P(tree frog) = 0.54

超像素分割算法
Super Pixel

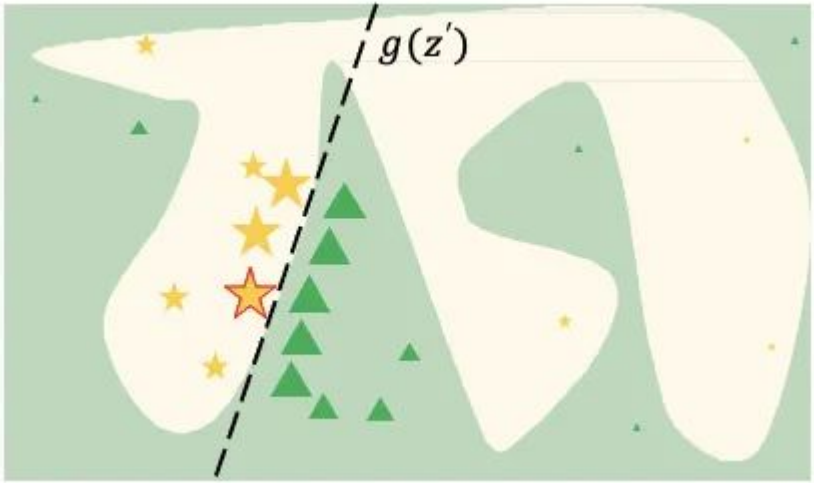
SP_1	SP_2	SP_3	...	SP_M
1	0	1	...	1

0: 不存在(灰色色塊); 1: 存在 (原始色塊)

Explainable ML: LIME (Local Interpretable Model-agnostic Explanations)



Explainable ML: LIME (Local Interpretable Model-agnostic Explanations)



z'

SP_1	SP_2	SP_3	...	SP_M
1	0	1	...	1

Explainable ML: SHAP (SHapley Additive exPlanations)

- A method based by game theory
- Core idea: evaluate the contribution of each feature

$$\phi_j(val) = \sum_{S \subseteq \{x_1, \cdots, x_p\} \setminus \{x_j\}} \frac{|S|!(p - |S| - 1)!}{p!} (val(S \cup \{x_j\}) - val(S))$$

$$val_x(S) = \int \hat{f}(x_1, \cdots, x_p) d\mathbb{P}_{x \notin S} - E_X(\hat{f}(X))$$

Explainable ML: SHAP (SHapley Additive exPlanations)

工程師 (S)	能產出幾行 code $(val(S))$
x_1	10
x_2	30
x_3	5
x_1, x_2	50
x_2, x_3	35
x_1, x_3	40
x_1, x_2, x_3	100

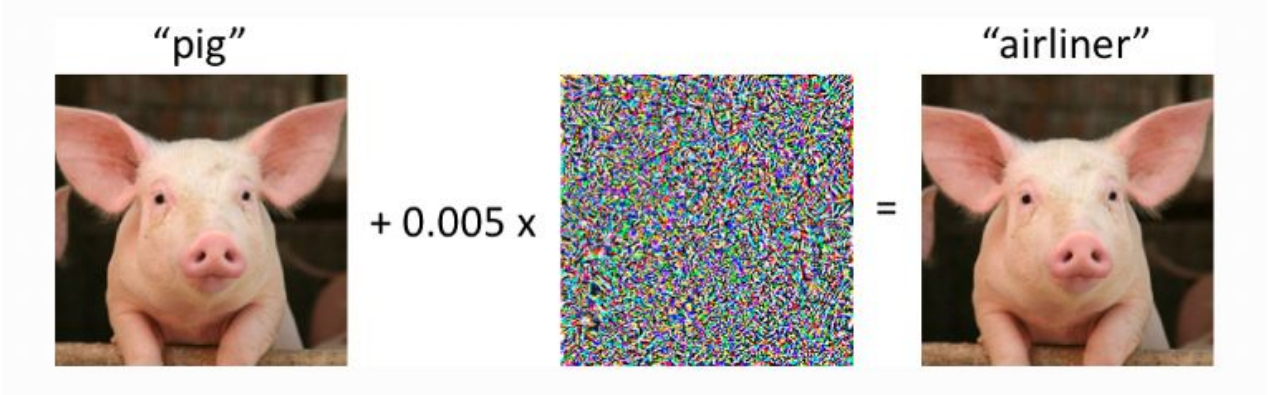
Explainable ML: SHAP (SHapley Additive exPlanations)

Order	x_1 Contribution	value
x_1, x_2, x_3	$val(x_1)$	10
x_1, x_3, x_2	$val(x_1)$	10
x_2, x_1, x_3	$val(x_2, x_1) - val(x_2)$	$50 - 30 = 20$
x_2, x_3, x_1	$val(x_2, x_3, x_1) - val(x_2, x_3)$	$100 - 35 = 65$
x_3, x_1, x_2	$val(x_3, x_1) - val(x_3)$	$40 - 5 = 35$
x_3, x_2, x_1	$val(x_3, x_2, x_1) - val(x_3, x_2)$	$100 - 35 = 65$
	$\frac{1}{6}(10 + 10 + 20 + 65 + 35 + 65) = 34.17$	

Adversarial Attack in NLP

Introduction to AI
May 15, 2023

Concept of Adversarial Attack



Original Input	Connoisseurs of Chinese film will be pleased to discover that Tian's meticulous talent has not withered during his enforced hiatus.	Prediction: **Positive (77%)**
Adversarial example [Visually similar]	**Aonnoisseurs** of Chinese film will be pleased to discover that Tian's meticulous talent has not withered during his enforced hiatus.	Prediction: **Negative (52%)**
Adversarial example [Semantically similar]	Connoisseurs of Chinese **footage** will be pleased to discover that Tian's meticulous talent has not withered during his enforced hiatus.	Prediction: **Negative (54%)**

White box attack v.s Black box attack

White box attack	Black box attack
The attacker has access to the **model's parameters**	The attacker has **no access to these parameters**, i.e., it uses a different model or no model at all to generate adversarial images with the hope that these will transfer to the target model

NLP Attacks

- Useful toolkit:
  Textattack

1. Goal	Stipulate **the goal of the attack**, like to change the prediction score of a classification model, or to change all of the words in a translation output
2. Constrains	**Determine if a potential perturbation is valid** with respect to the original input
3. Transformation	Take a text input and **transform it by inserting and deleting characters, words, and/or phrases**
4. Search method	Explore the space of possible transformations within the defined constraints and attempt to **find a successful perturbation which satisfies the goal function**

Attacks In HW4

1. Goal	Change the prediction, i.e., positive → negative, negative → positive
2. Constrains	No constrain. But it will be better if you minimum the difference between original sentence and attacked sentence
3. Transformation	Try it by yourself
4. Search method	You can based on the result of LIME and SHAP

Reference

- Clever Hans - Wikipedia
- Goals of explainable AI (XAI)
- What is Global, Cohort and Local Explainability? | Censius AI Observability Blog
- Hung-Yi Lee ML Course
- Visualizing and Understanding Convolutional Networks (arxiv.org)
- Fine-tune BERT with Sparse Self-Attention Mechanism
- LIME Explanation
- SHAP Explanation
- Textattack Document
- Same Lecture in Last Year