# 1. Fundamentals of Explainability

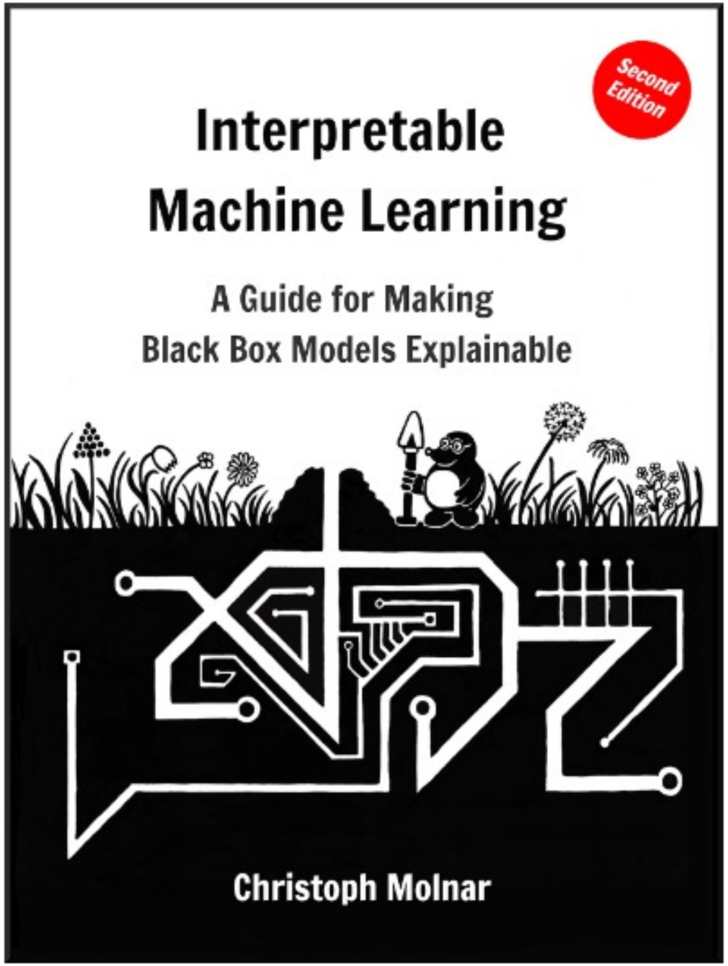
- Reminder: **Record** in zoom

# Fundamentals of Explainability

- 1.0: Terminology (of Machine Learning)
- 1.1: Definitions
  - 1.1.1 Definition Explainable Machine Learning (xML)
  - 1.1.2 Definition Explainability
  - 1.1.3 Definition Explanation
- 1.2: Importance
  - Why do we need Explainability / Explainable ML?
- 1.3: Taxonomy xML methods
- 1.4: Properties of
  - 1.4.1 xML methods
  - 1.4.2 Explanations
- 1.5: What are *good* Explanations?

# Reference Book

Chapter 2



# 1.0 Terminology

- An algorithm is a set of rules that a machine follows to achieve a goal.
- Machine Learning refers to a set of methods that allows machines to learn algorithms from data, i.e. by defining the goals but not (explicitly) the set of rules.
- A **Machine Learning Algorithm** is one specific method, i.e. it is the program used to learn a Machine Learning Model from data. (Other names: Machine Learning Method)
- A **Machine Learning Model** is the learned algorithm that maps inputs to predictions. (Other names: Predictor, classifier, regressor, ...)
  - A Black Box Model does not reveal its internal workings.
  - A White Box Model does. (Other names: **Interpretable Models**)

# 1.0 Terminology

- A **dataset** is a table of data. It contains the features and target variable.
- A **datapoint** is a row in the dataset. (Other names: Instance)
- A **feature** is a column in the dataset. Features are the inputs used by the Machine Learning Model to make predictions.
  - *Features are assumed to be interpretable*. (Big assumption)
- The **target** encodes the goal and it is the information the machine learns to predict.
- The **prediction** is what the Machine Learning Model predicts for the target variable.

# 1.0 Terminology

- A **Machine Learning Task** is the combination of a dataset and a (choice of) target variable.
  - The target variable implies the task. E.g. Classification, Regression, Outlier Detection, ..
  - One dataset can be used for multiple tasks.

$$\hat{y}^{(i)} = \hat{f}(x^{(i)}) \qquad y^{(i)} \qquad x_j^{(i)}$$

Prediction Model Datapoint Target Feature j of datapoint i

# 1.1.1 Definition Explainable Machine Learning

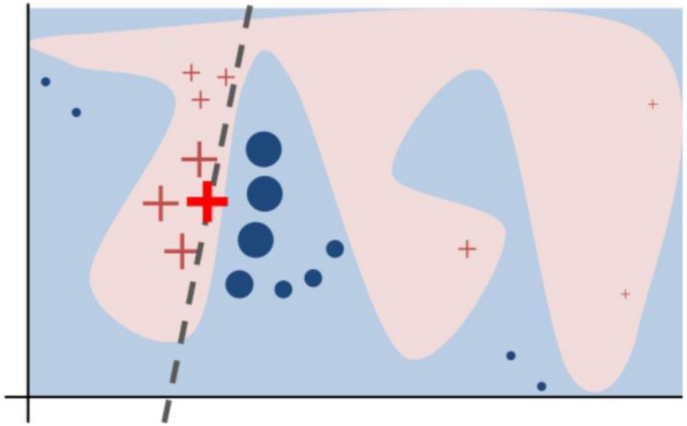
- **Explainable Machine Learning (xML)** refers to a set of methods that makes the behavior and predictions of Machine Learning Models understandable to humans.

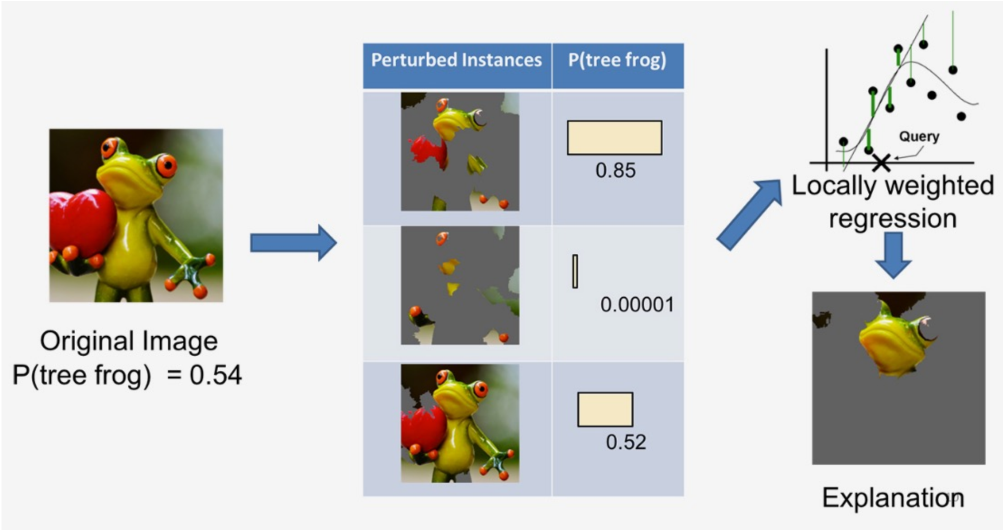
# 1.1.1 Definition Explainable Machine Learning

some xML methods

| Technique | Composition | Performance* | Model Fidelity | Model Specificity | Explanation Type | Scalable** | Scope |
|---|---|---|---|---|---|---|---|
| Bayesian Rule List | Ante-Hoc | H | Yes | Self | Relative Importance | Yes | Global |
| BETA^^ | Ante-Hoc | H | Yes | Self | Relative Importance | Yes | Global |
| Decision Trees | Ante-Hoc | M | Yes | Self | Rules | Yes | Global |
| Falling Rule Lists | Ante-Hoc | H | Yes | Self | Relative Importance | Yes | Global |
| GAM | Ante-Hoc | L | Yes | Self | Relative Importance | Yes | Global |
| GA2M | Ante-Hoc | M | Yes | Self | Graphs | Yes | Global |
| ICE Plots | Post-Hoc | N/A | No | Agnostic | Graphs | Yes | Global |
| Interpretable Decision Sets | Ante-Hoc | H | Yes | Self | Relative Importance | Yes | Global |
| *k*-LIME | Post-Hoc | N/A | No | Agnostic | Relative Importance | Yes | Local |
| LIME^^^ | Post-Hoc | N/A | No | Agnostic | Relative Importance | Data size | Local |
| Logistic Regression | Ante-Hoc | M | Yes | Self | Relative Importance | Yes | Global |
| Model Distillation | Post-Hoc | M-H | Yes | Agnostic | *Any* | Yes | Global |
| Partial Dependence Plots | Post-Hoc | N/A | No | Agnostic | Graphs | Yes | Global |
| RF Explainer | Post-Hoc | H | No | Random Forest | Relative Importance | Yes | Local |
| Relative Baseline Contributions*** | Post-Hoc | N/A | No | Agnostic | Relative Importance | Yes | Local |
| Right for the Right Reasons ^ | Post-Hoc | H | No | Agnostic | Relative Importance | Yes | Local |
| Shapley Values | Post-Hoc | N/A | No | Agnostic | Graphs | Number of Features | Local |
| SLIM | Ante-Hoc | H | Yes | Self | Relative Importance | Yes | Global |
| XGB Explainer | Post-Hoc | H | No | XG Boost | Relative Importance | Yes | Local |

# 1.1.1 Example: LIME






- Perturb a given data point to create a number of fake data points around it and determine the „distance" of each from the given data point
- Use the complex model to make predictions for each of the perturbed data points
- Pick a small number of features that best describe the complex model output for the perturbed data points
- Based on the chosen features, fit a sufficiently simple model to the perturbed data points, their predictions and distance to the given data point
- The weights of that simple model are explaining the local behavior of the complex model

# 1.1.2 Definition Explainability

- **Difficult to define Explainability**.
- Possible definitions:
  - „Interpretability is the degree to which a human can understand the cause of a decision“
  - „Interpretability is the degree to which a human can consistently predict the model‘s result“
- We will use Explainability and Interpretability interchangably.
- A Machine Learning Model has a higher degree of Explainability if its decision / predictions are easier to comprehend for humans, than [...].
  - Indepently of the fact where this interpretability comes from. „Is the model intrinsically interpretable or made interpretable by a xML method?“

# 1.1.3 Definition Explanation

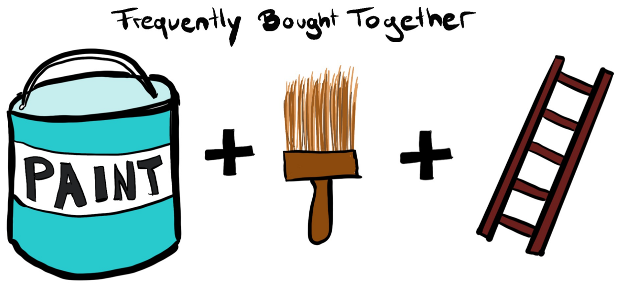
- An explanation is the answer to a „why-question".
- A xML method creates explanations. It can create many explanations for a given model.
- Oftentimes the explanation is only valid for one specific prediction.
- The more the explanations help a human to comprehend the model's behavior, the more interpretable the model becomes.

# 1.1.3 Example: Explanation

- Why did amazon recommend that product to me?
- Why did my robot vacuum stop cleaning?
- Why did my computer freeze?

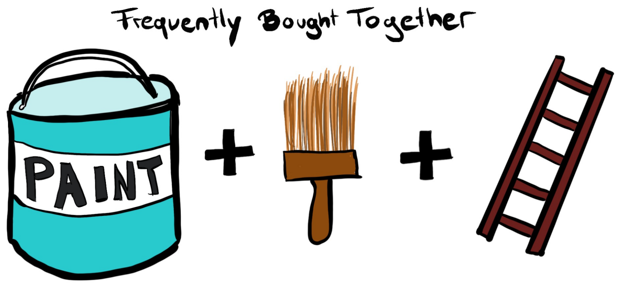
# 1.1.3 Example: Explanation

- Why did amazon recommend that product to me?
- Why did my robot vacuum stop cleaning?
- Why did my computer freeze?




# 1.1.3 Example: Explanation

- Why did amazon recommend that product to me?
- Why did my robot vacuum stop cleaning?
- **Why did my computer freeze?**
  - Deeply unsettling to *not* get an explanation
  - Humans have an intrinsitic desire to reach comprehension




# Motivation – Why Care for Explainability?


"Does your car have any idea why my car pulled it over?"

How bitter would that joke be with a severe medical context?

# 1.2 Importance of Explainability

- **Human curiosity**. Human‘s desire to understand.
  - Our mental model gets updated through explanations.
    - Q: „Why do i feel sick today?“
    - A: “Because you ate that old sushi“ the explainee thought as he sighed with relief. „Don‘t do that again“
  - **Humans want to harmonize contradictions**. It‘s the unexpected that makes us curious.
    - Q: „Why did my loan get rejected?“
    - Q: „Why did my vacuum suddenly stop cleaning?“

# 1.2 Importance of Explainability

- Trust.
  - **Socal acceptance.** A machine that explains its behavior will be more accepted. People like to anthropomorophise objects.

**He** really tries to do a good job.

# 1.2 Importance of Explainability

- **Trust.** Question AI decisions and illuminate the black box.
  - **Socal acceptance**. A machine that explains its behavior will be more accepted. People like to anthropomorophise objects.
  - **Feeling in control**. A high degree of explainability makes humans feel more in control, as if they could influence the machine because they understand it.
  - **Fairness**. To ensure that the machine's decisions are not due to some data bias. When you have a right to an explanation.
  - **Safety**. When you want to be 100% sure that the machine's abstraction is flawless. When the stakes are high.

# 1.2 Examples: Importance of Explainability

- Imagine a machine that decides whether or not a patient should undergo some surgery.
- Imagine a machine that classifies tumors being benign or malign.
- **Fairness and safety are critical**.

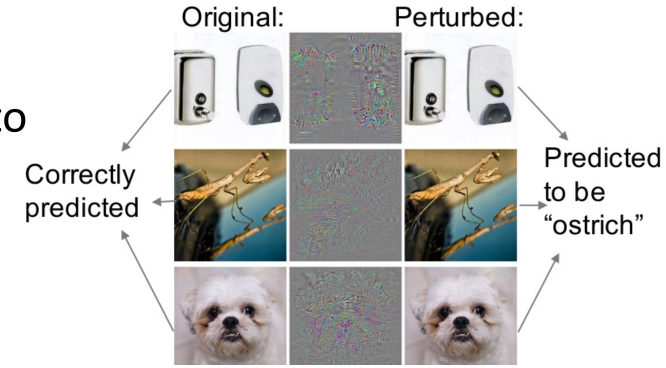
# 1.2 Importance of Explainability

- **Action advice.** Understand which input to change for obtaining a desired output change.
  - “What could i do to get my loan approved?“
  - „What could i do to stop my vacuum from failing?“

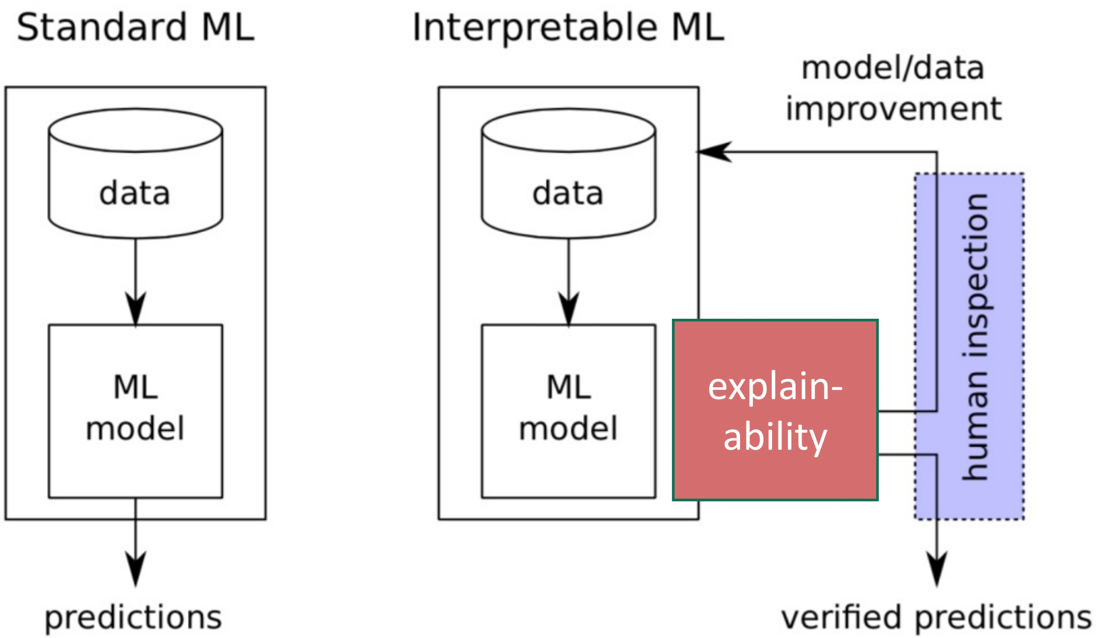
# 1.2 Importance of Explainability

- **Debug and Improve.** Understand how to change model when things go (seemingly) wrong.
  - “Why did my model fail here?“
  - „If only i would understand why it came to the wrong prediction in the first place..“
- When new hypotheses are drawn – an example: *"Pneumonia patients with asthma had lower risk of dying (Caruana et al. 2015)"*

Small perturbations lead to false image classification



# Motivation – What do We Hope to Achieve?




***Generalization error***

***Generalization error + human experience***

Example from previous slide:

| *Patient ID* | | *Has Asthma* | | *Risk of Death* |
|---|---|---|---|---|
| 84 | ... | Yes | ... | 5% |
| 85 | ... | Yes | ... | 6% |
| 86 | ... | No | ... | 12% |
| 87 | ... | No | ... | 15% |
| ... | ... | ... | ... | ... |

*Feature Importance (Higher risk of death):* *Low* *High*

*Feature Importance (Lower risk of death):* *Low* *High*

*With Context:* Patients with asthma have a lower risk of death from pneumonia because *they receive more intensive care.*

# 1.2 Example: Explainabiltiy in LLMs

- To **probe Artificial General Intelligence (AGI)** of large Language Models (LLMs) such as GPT-4, ChatGPT
  - It‘s easier to build **trust** and to **associate intelligence** if LLMs backup there reasoning with good explanations.
  - New directions: LLMs can explain themselves (in understandable words). Just like a good student can!

**Sparks of Artificial General Intelligence: Early experiments with GPT-4**

Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

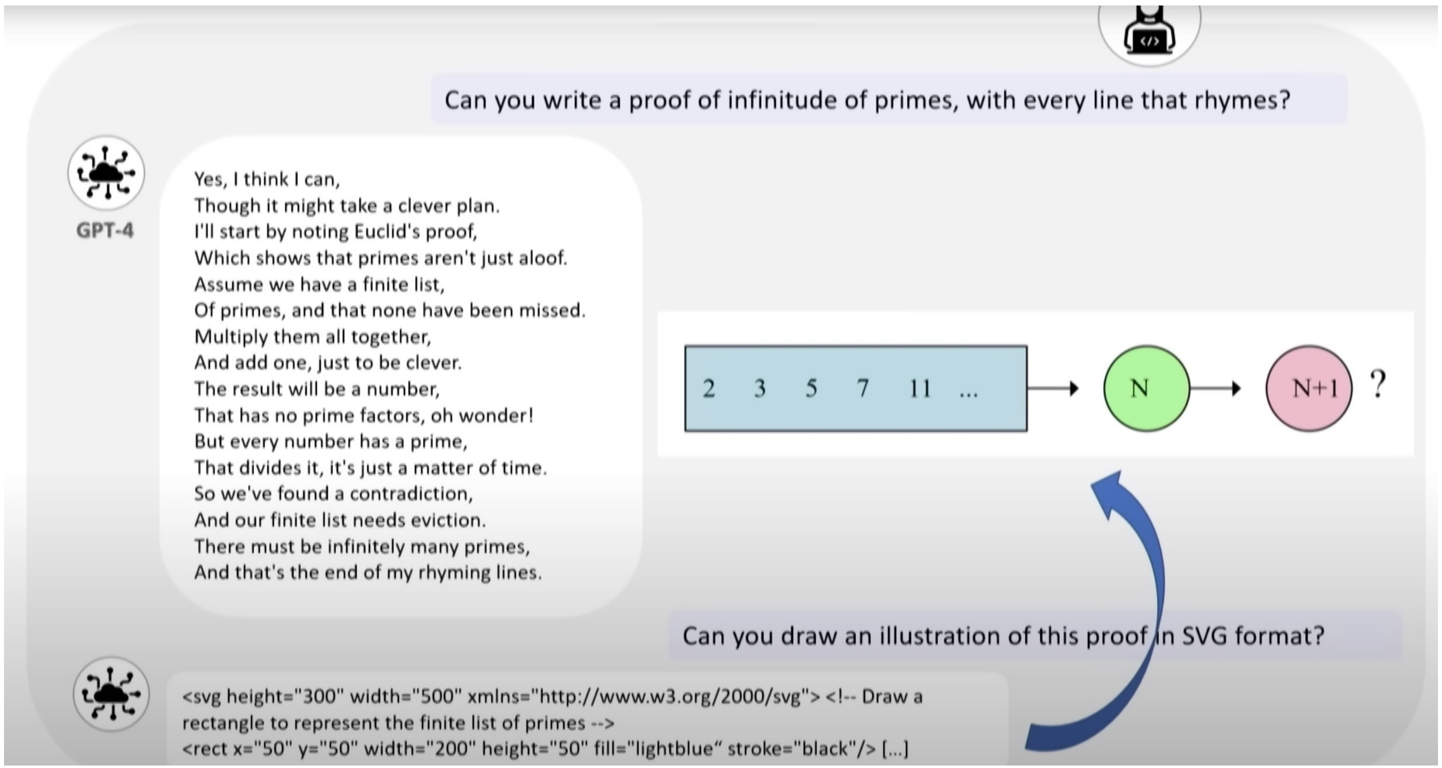
Microsoft Research

**Abstract**

Artificial intelligence (AI) researchers have been developing and refining large language models (LLMs) that exhibit remarkable capabilities across a variety of domains and tasks, challenging our understanding

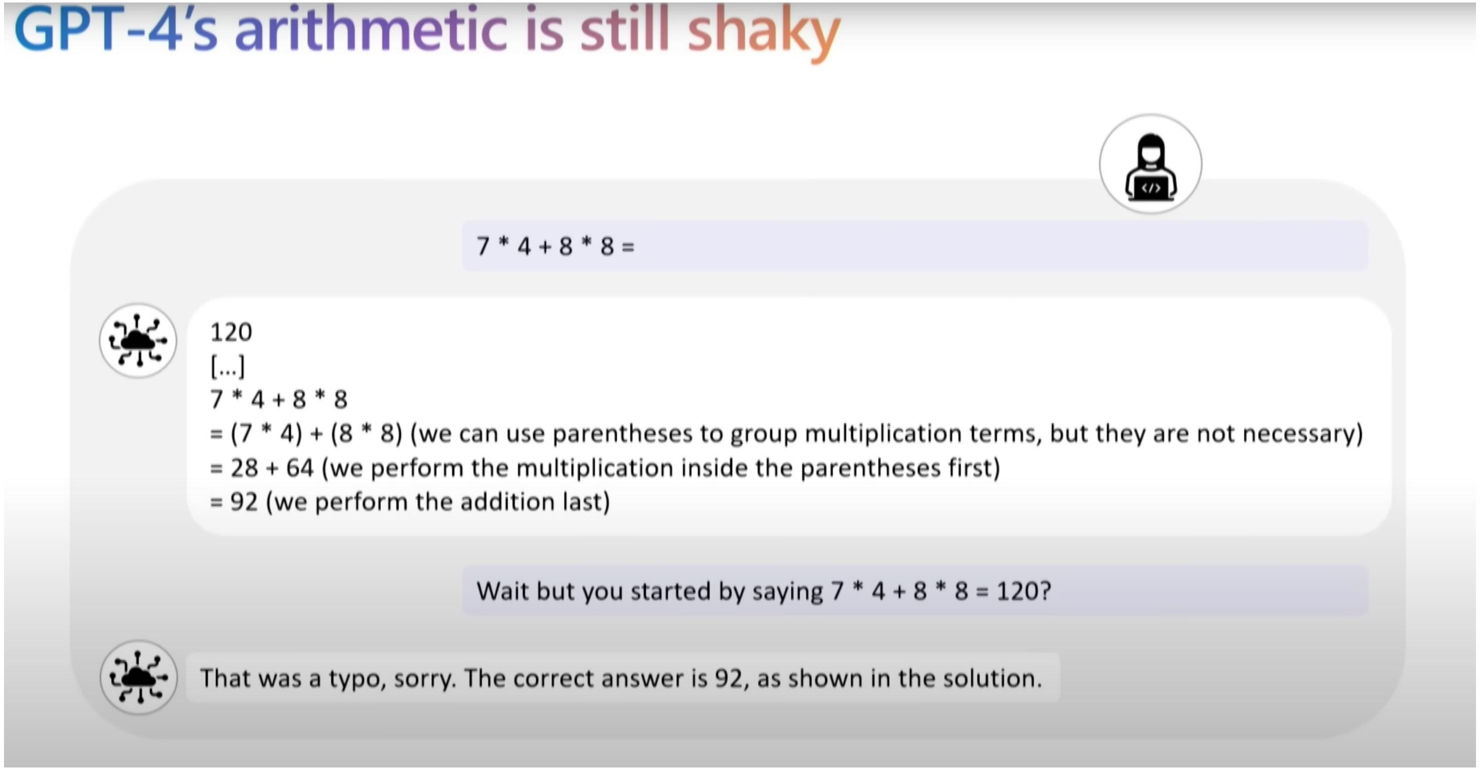
# 1.2 Example: Explainabiltiy in LLMs

- From a talk by Sebastien Bubeck only 2 weeks ago (https://www.youtube.com/watch?v=qbIk7-JPB2c)



# 1.2 Example: Explainabiltiy in LLMs

- How to reconcile? What went wrong? How is that possible? How to improve?



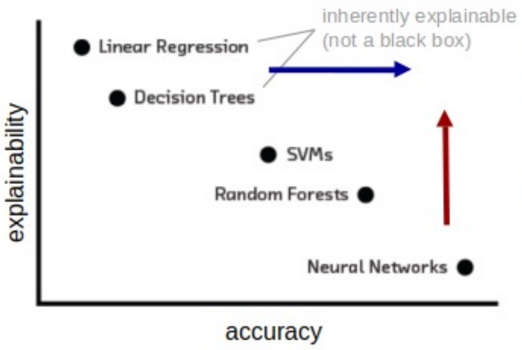
# 1.3 Taxonomy – xML methods

- Taxonomy = Classification of methods according to some common criteria
- We will use four criteria to categorize them
  - Intrinsic or post-hoc?
  - Result of method?
  - Model-specific or model-agnostic?
  - Local or global?

# 1.3 Taxonomy – xML methods

- Intrinsic (=ante-hoc) or post-hoc?
  - Intrinsic: Interpretability is achieved by restricting the complexity of the model. Note that this does have to imply a less performant model but rather buts more emphasis on model selection.
  - Post-hoc: Interpretability is achieved by applying a xML method to the model after it has been trained.
  - Post-hoc methods can also be applied to intrinsically interpretable models.

A common trade-off in ML




Overcome the trade-off by...
- making inherently explainable models more accurate or
- generating good explanations for accurate black-box models!

# 1.3 Taxonomy – xML methods

- **Result of the xML method?**
  - Feature summary statistic / visulization. E.g.
    - Feature importance: Single number per feature
    - Partial dependence plots: Curve per feature
    - Saliency Maps: Single number per feature (=pixel)
  - Model internals. E.g.
    - Linear Regression: Learned weights
    - Decision Trees
    - Convolutional Layers
  - Data points (good for text and images, not for tables. Why?). E.g.
    - Prototypes
    - Counterfactuals

# 1.3 Taxonomy – xML methods

- **Result of the xML method?**
  - Feature summary statistic / visulization. E.g.
    - Feature importance: Single number per feature
    - Partial dependence plots: Curve per feature
    - Saliency Maps: Single number per feature (=pixel)
  - Model internals. E.g.
    - Linear Regression: Learned weights
    - Decision Trees
    - Convolutional Layers
  - Data points (good for text and images, not for tables. Why?). E.g.
    - Prototypes
    - Counterfactuals

Why Saliency Maps not under Data points?

# 1.3 Taxonomy – xML methods

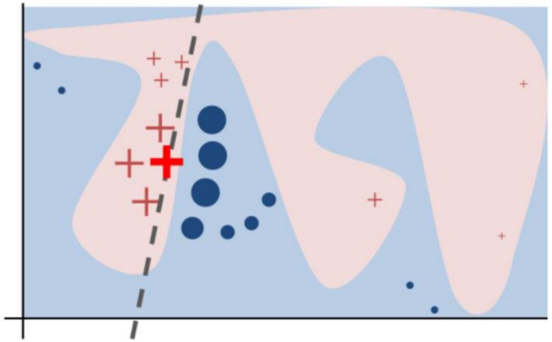
- **Model-specific or Model-agnostic?**
  - Model-specific: Methods that are limited to a specific sublcass of ML models.
  - Model-agnostic: Methods that can be applied to any ML model.
  - E.g. interpreting the weights of a linear regression model is a model-specific interpretation
  - E.g. methods that only apply to neural networks are also model-specific
  - Model-agnostic methods usually work by inspecting feature input and prediction output pairs.

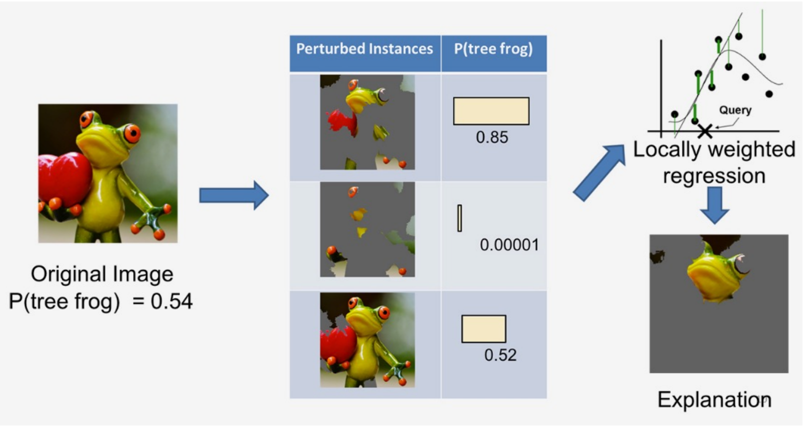
# 1.3 Taxonomy – xML methods

- **Local or global?**
  - Local: The method yields explanations that only explain individual prediction. They only explain a single datapoint.
  - Global: The method yields explanations that apply to the ML model independently of any specific datapoint.
  - The scope of a local xML method can be increased by “averaging“ local explanations over a large subset of data.

# 1.3 Example Taxonomy „LIME“

- Intrinsic or post-hoc?
- Result of method?
- Model-specific or model-agnostic?
- Local or global?





# 1.3 Example Taxonomy „LIME“

- Intrinsic or post-hoc?
    - post-hoc
- Result of method?
    - interpretable model -> feature importance
- Model-specific or model-agnostic?
    - model-agnostic
- Local or global?
    - local