Martin Jullum, jullum@nr.no
Senior Research Scientist
Norwegian Computing Center

Introduction to XAI

Guest lecture GRA 4162, Deep Learning and Explainable AI
BI Norwegian Business School

April 5th 2024

NORSK REGNESENTRAL
NORWEGIAN COMPUTING CENTER

Today’s lecture

- Motivation
- Categorization of XAI methods
- Briefly about a few XAI methods (SHAP, ALEPlots, Counterfactual explanations)
- Navigating in the XAI jungle

MOTIVATION

Explainable AI (XAI) – the research field

- Understanding what black box models do
- Develop models which are directly interpretable
- Ultimate goal: Making decisions based on such models more transparent, understandable, and interpretable for humans.

Figure 2. Number of scientific publications per year on Explainable Artificial Intelligence (XAI).

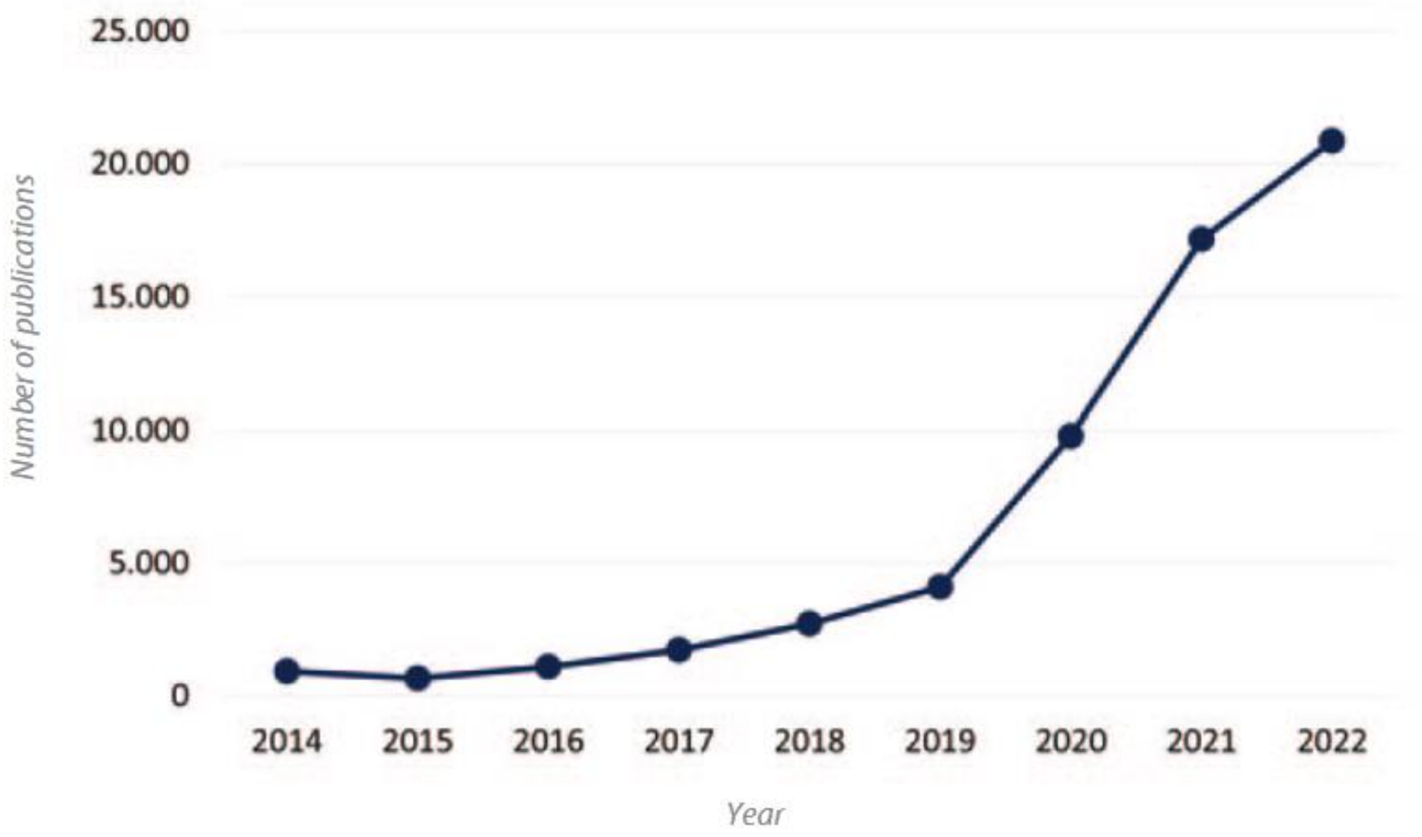
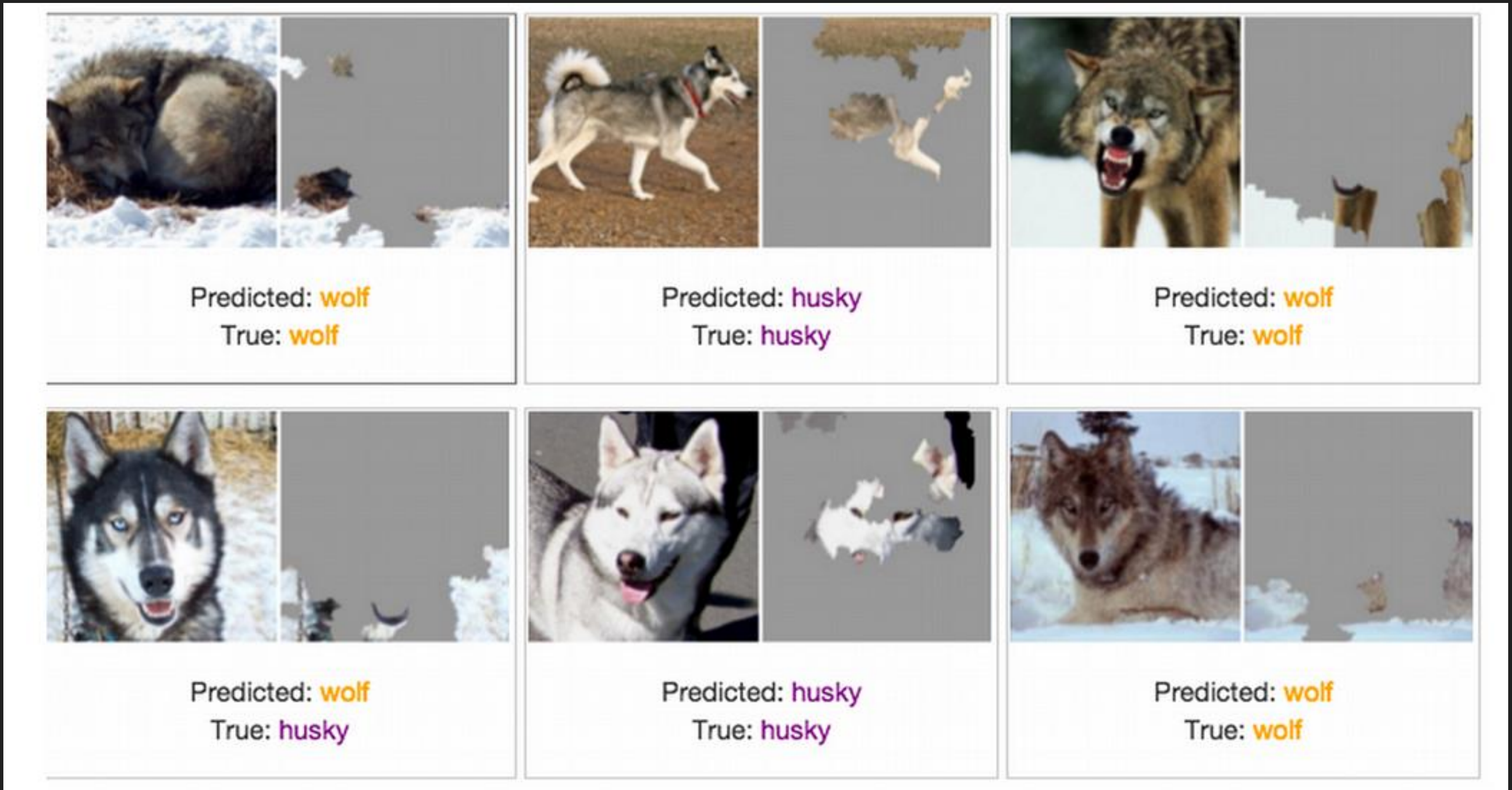


Figure extracted from Management solutions rapport:
"Explainable artificial intelligence (XAI) – Challenges of model interpretability" (2023)

Motivating example: Understanding image classification

- CNN used to classify images containing husky and wolf
- Explainability question
  - What parts of the image were most crucial for each classification?

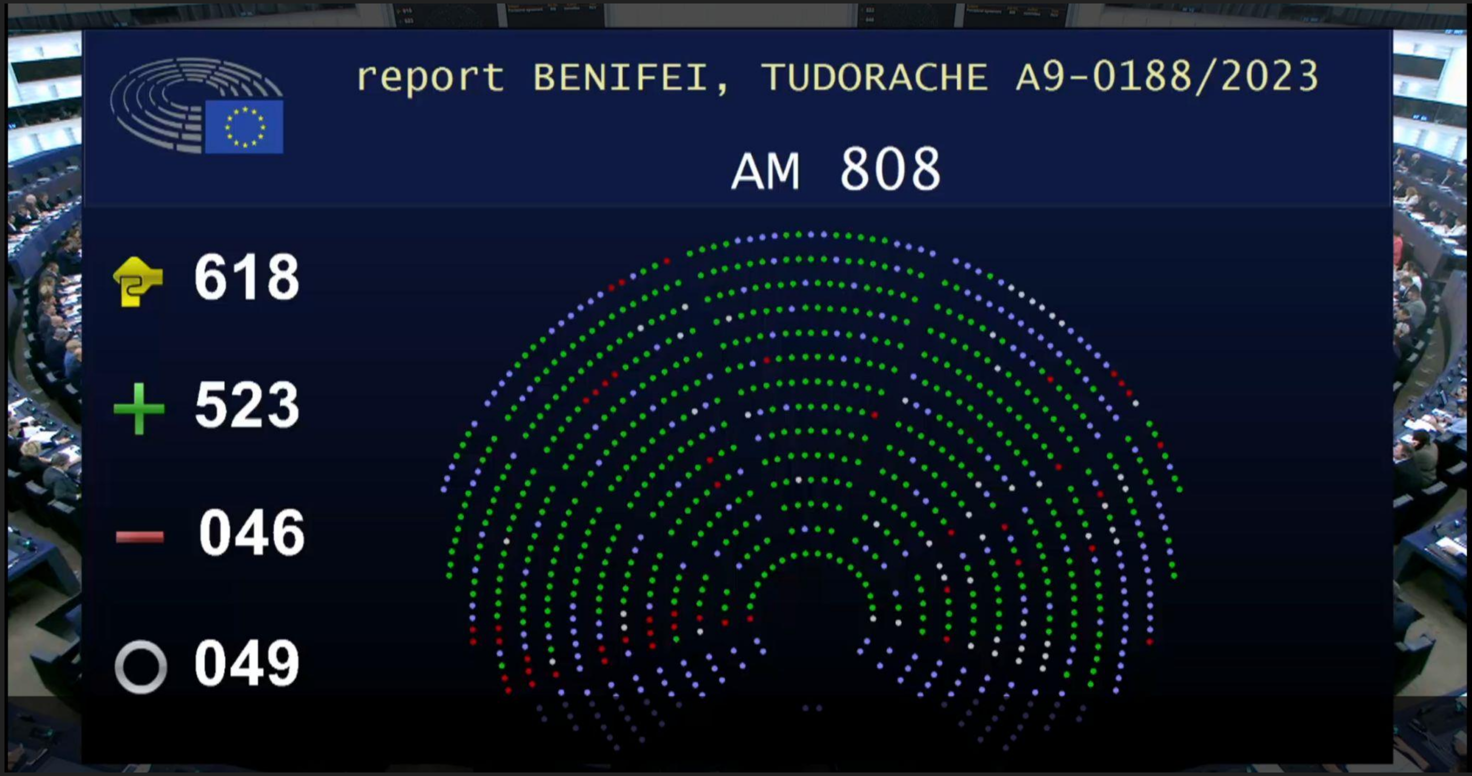


Motivating example: Automatic mortgage lending system

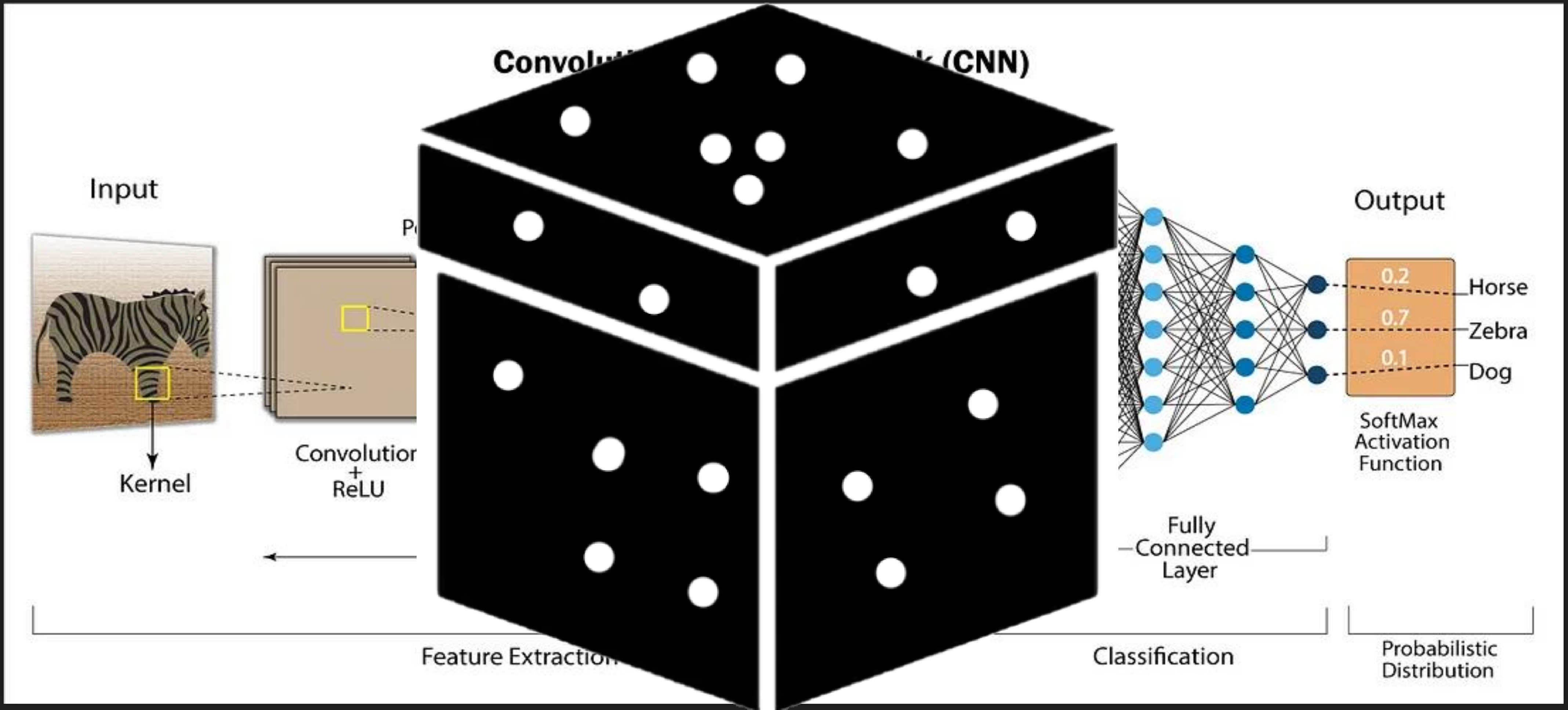
- A bank built a ML-model to predict loan default based on transaction history and other customer info
- The system grants a loan if the model predicts a probability of default < 0.1, otherwise declined
- Explainability questions
  - Overall
    - Which training observations were most crucial in the model training?
    - How does the probability of default change with income?
  - For a specific declined application
    - How did the inclusion of age in the model affect the probability of default?
    - What feature values need to be changed for the application to be accepted?

Why is it important to explain?

- Safety and trust among developers, responsible parties, and users
- Empower the user to challenge an automatic decision
- Ensure responsible use of data/model (privacy and discrimination)
- Legislation: AI ACT?, GDPR? administrative law (forvaltningsloven)?
- Help developers improve the model/AI system by detecting errors/unwanted behavior



Complicated models ARE hard to understand



Simple models are not always simple…

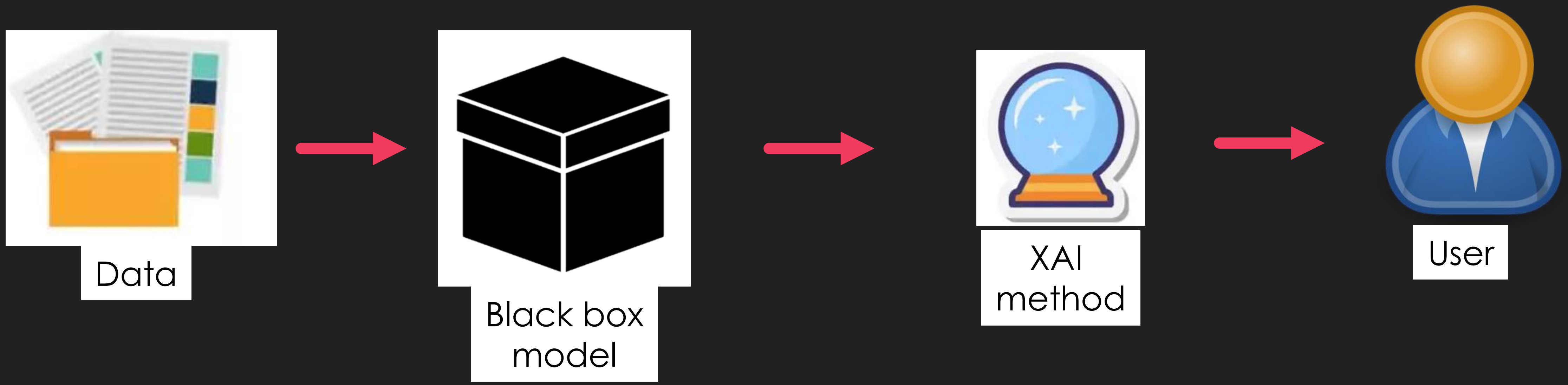
- Simple linear regression model with normally distributed features x_1 and x_2:

$$y = a + b_1 \cdot x_1 + b_2 \cdot x_2$$

- **Explanation STAT101**: y increases by b_1 when x_1 increases by 1, and analogously for x_2
  - This is an explanation of the mathematical model
  - Not a useful explanation when the features are dependent
- **Practial explanation** when $\mathrm{corr}(x_1,x_2) \approx 1$, $E[x_1] \approx E[x_2]$:
  y increases by b_1+b_2 when x_1 increases by 1 (since then x_2 also increase by 1).
- More complicated when the dependence is medium strong/non-linear/locally varying, with more features and a non-linear model

CATEGORIES OF XAI METHODS

The Explainability process



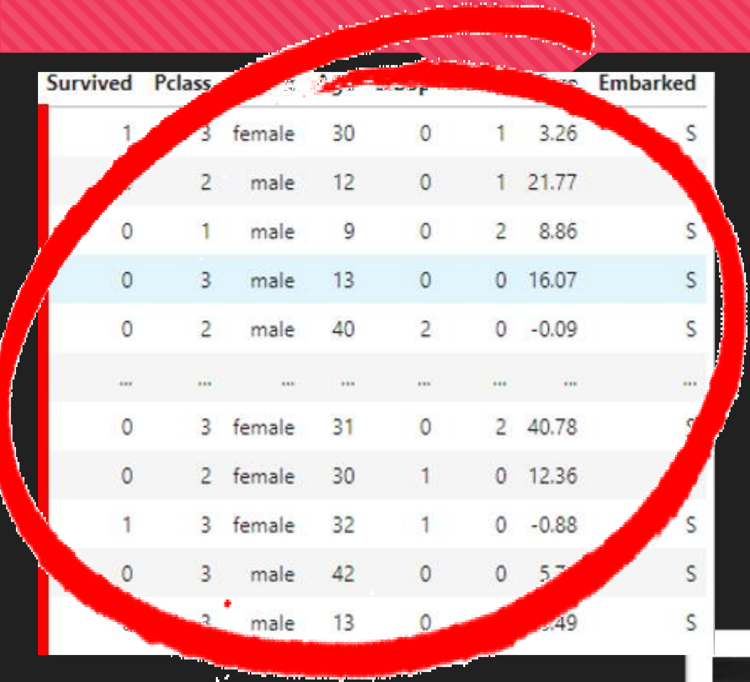
Also under the XAI-umbrella:

- Intrinsically interpretable models
- Global surrogate models

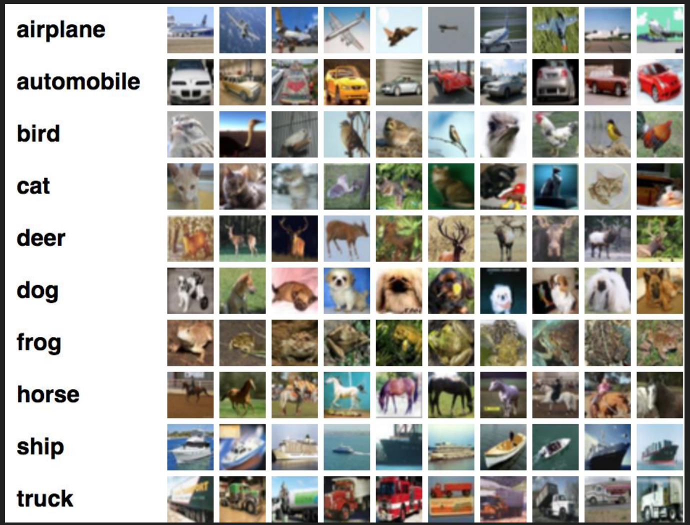
Many ways to categorize XAI-methods

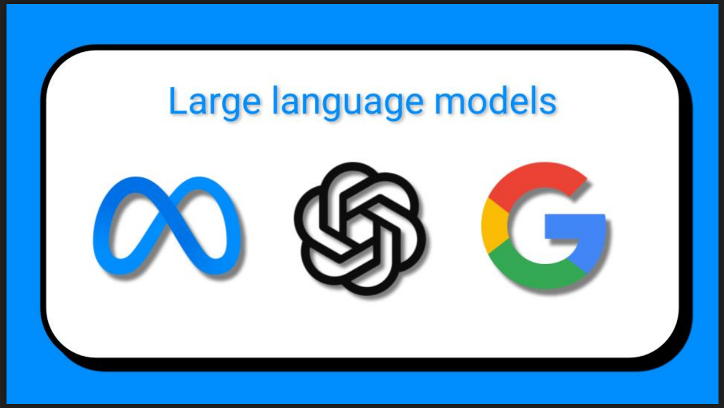
- Data/model type to be explained
- Model agnostic/model specific
- Local/global
- Presentation format

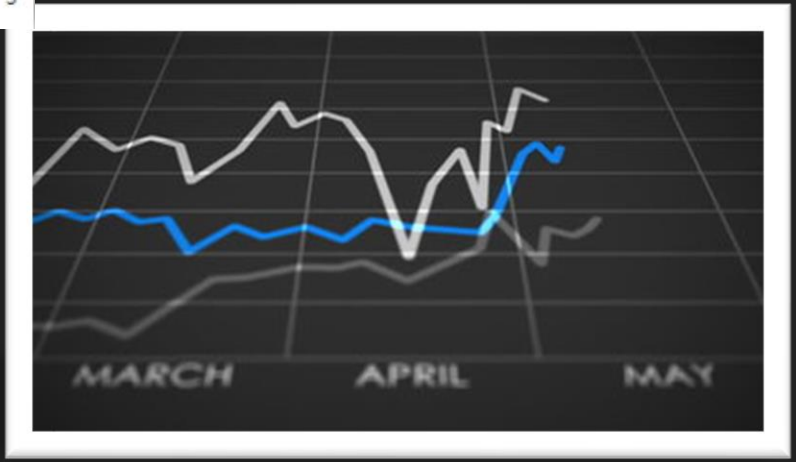
Different data types require different explanation methodology

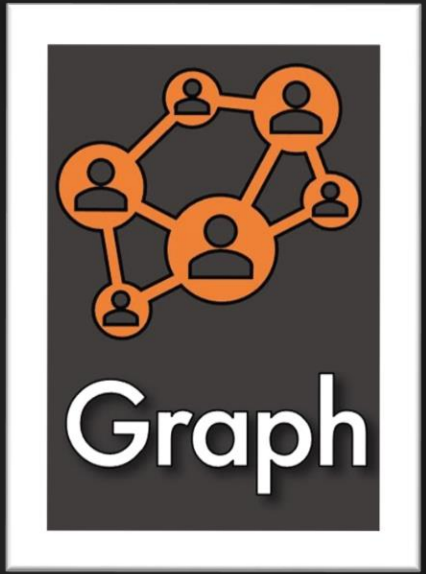


Survived	Pclass	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	Embarked
1	3	female	30	0	1	3.26	S
[illegible]	2	male	12	0	1	21.77	
0	1	male	9	0	2	8.86	S
0	3	male	13	0	0	16.07	S
0	2	male	40	2	0	-0.09	S
...	...	...	...	...	...	...	...
0	3	female	31	0	2	40.78	[illegible]
0	2	female	30	1	0	12.36	
1	3	female	32	1	0	-0.88	S
0	3	male	42	0	0	[illegible]	S
[illegible]	[illegible]	male	13	0	[illegible]	[illegible]	S



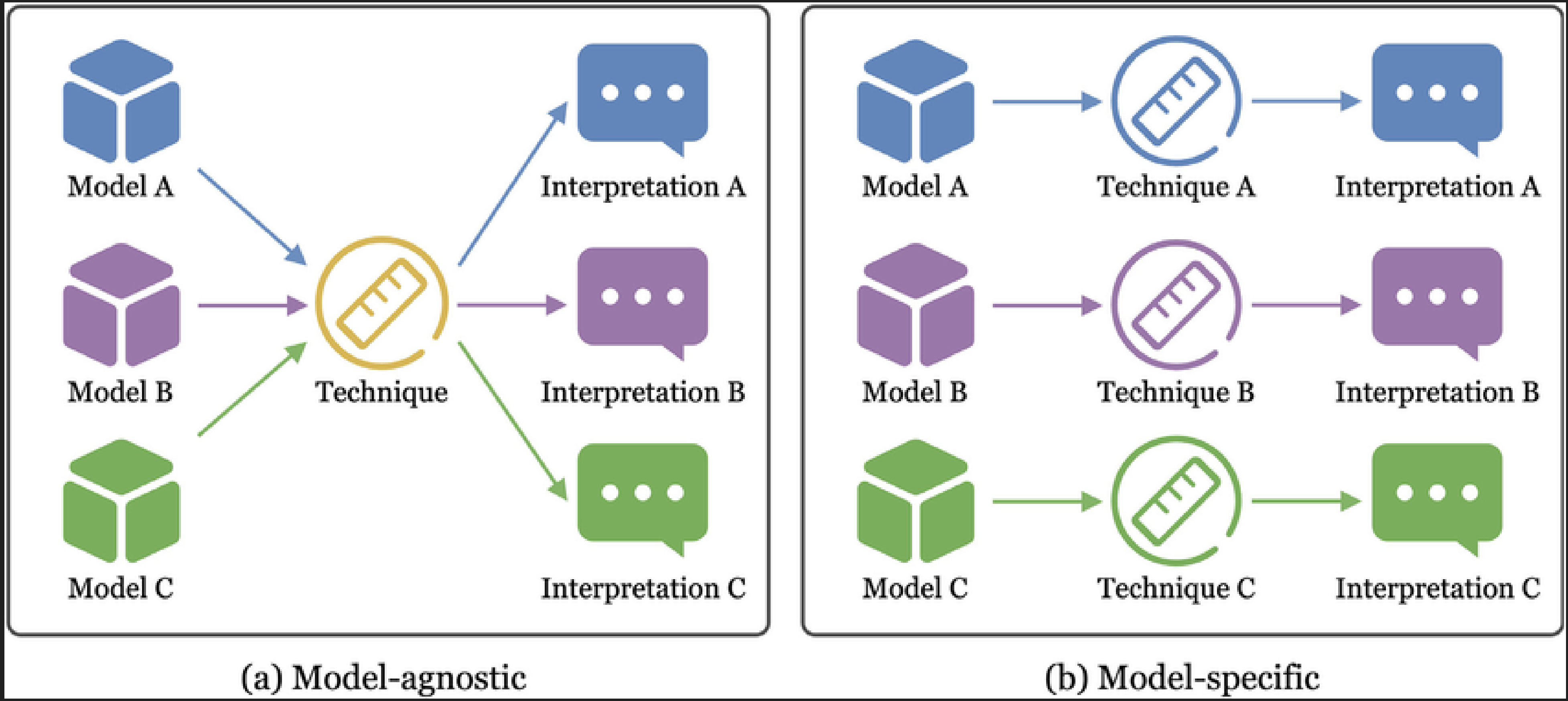




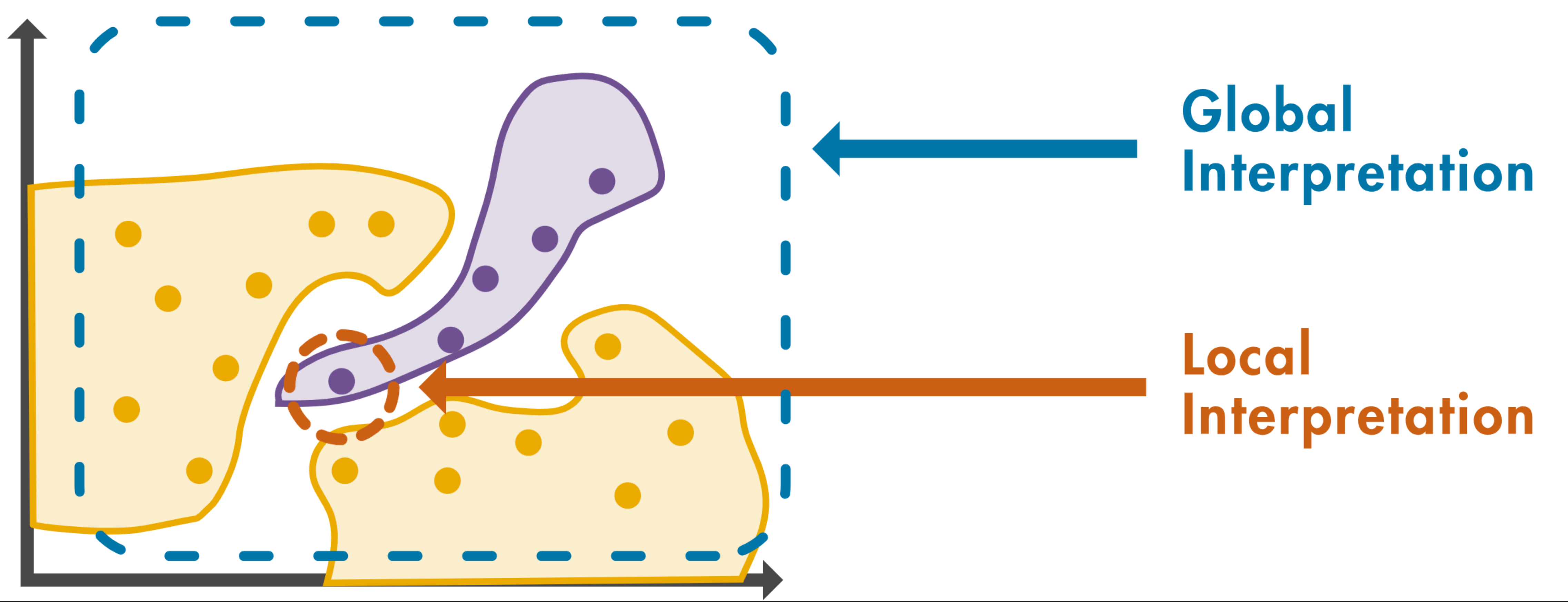


PREDICTION MODELS FOR TABULAR DATA

Model agnostic/ model specific



Local vs global explanation



Lots of (model agnostic) explainability methods

Shapley values (SHAP)

ALEPlots

Permutation feature importance

SAGE

Counterfactual explanations

Partial dependence plots

LIME

Anchors

ICE

Presentation format

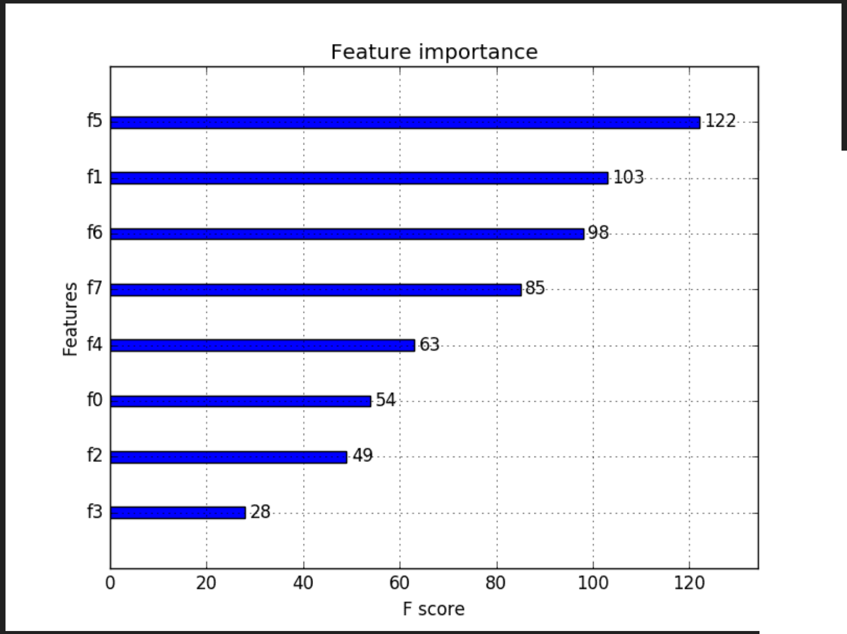
Shapley values (SHAP)

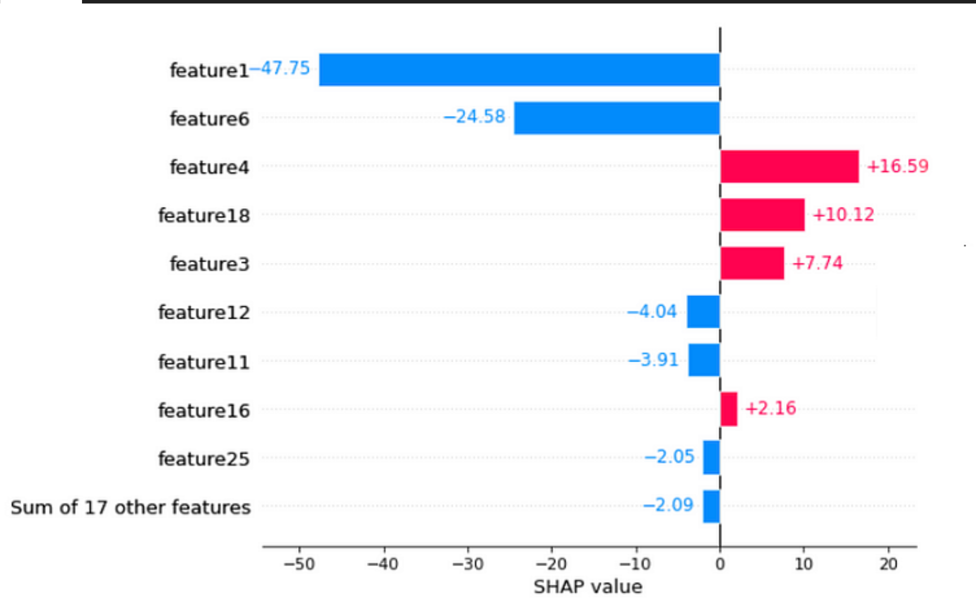
SAGE

LIME

Permutation feature importance

FEATURE CONTRIBUTION/EFFECT



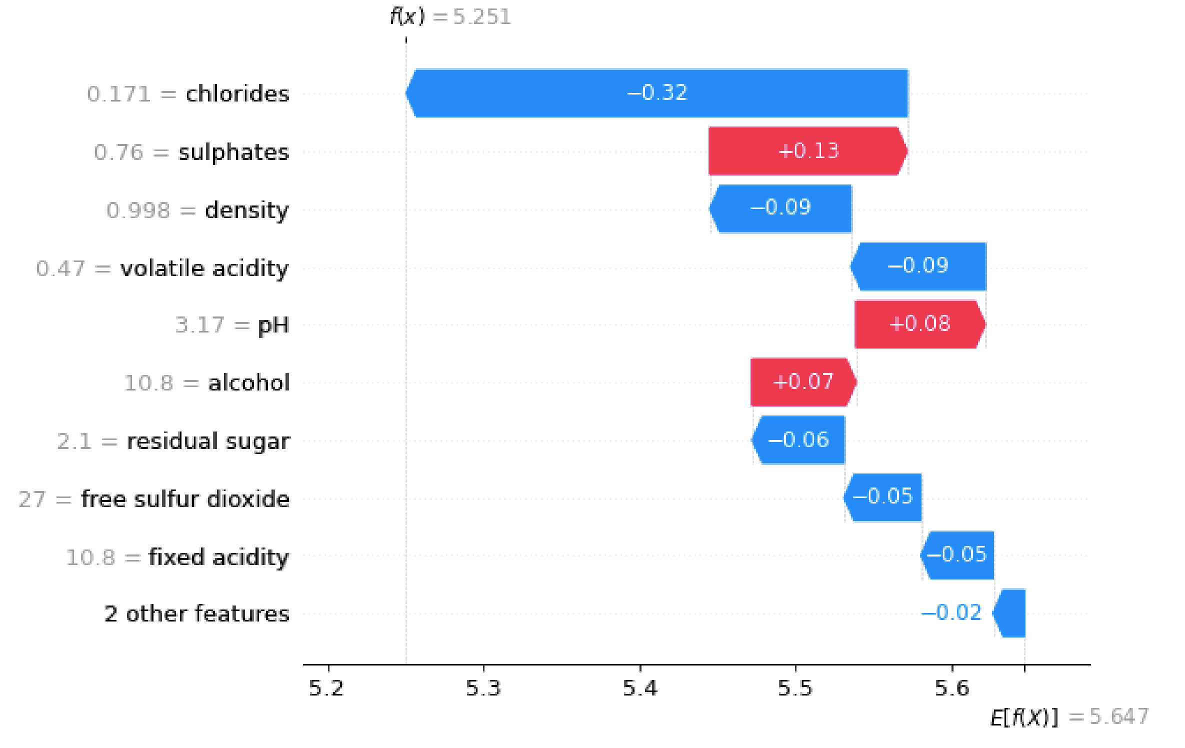


Presentation format

Shapley values (SHAP)

SAGE

DECOMPOSITION



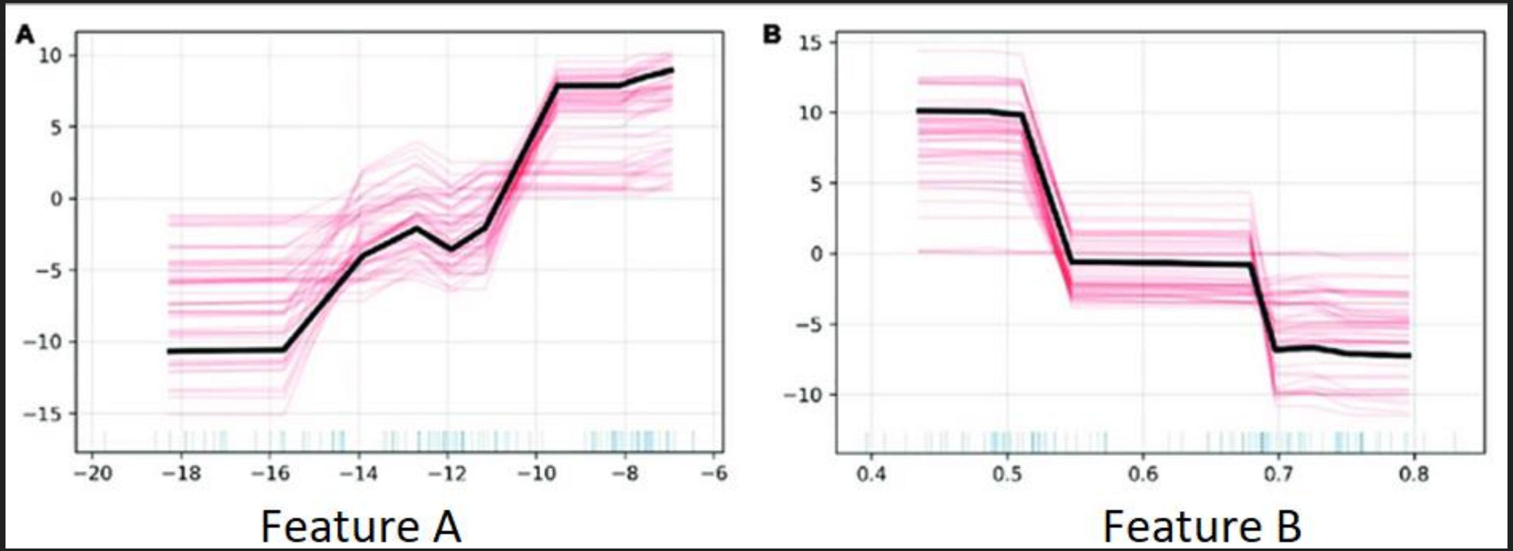
Presentation format

ALEPlots

ICE

Partial dependence plots

FEATURE EFFECT PLOT



Presentation format

Counterfactual explanations

EXAMPLES

Observations to explain

ID	Features					
	Age	Sex	Salary	Def. last year	*f(x)*	*Decision*
1	30	F	$ 6000	yes	0.18	0
2	25	M	$ 4500	no	0.30	0

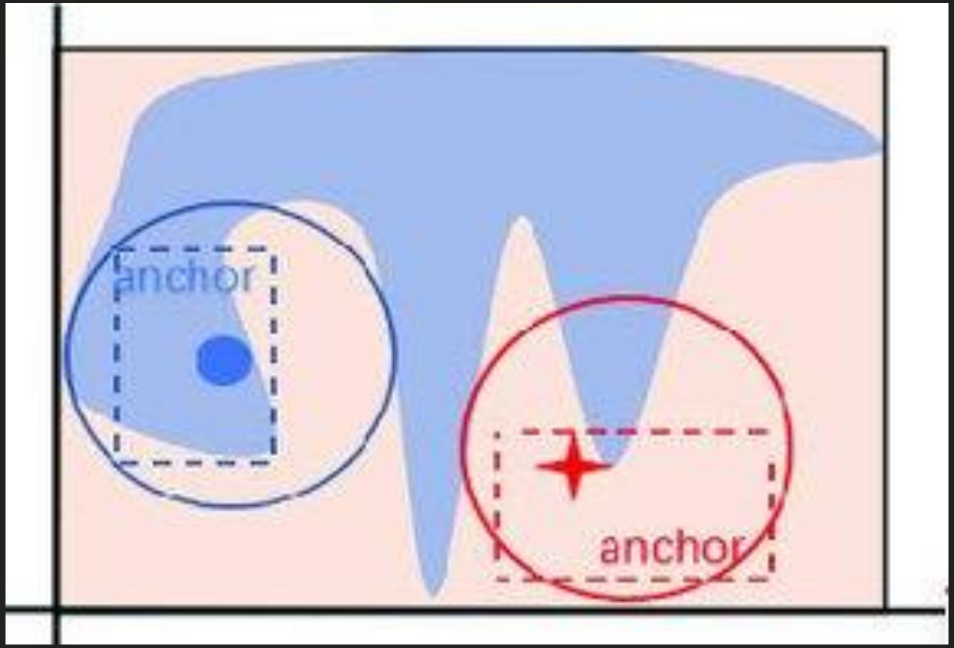
Final counterfactual explanations

Explain ID	Age	Sex	Decision	Salary	Def. last year
1	30	F	1	$ 6000	no
2	25	M	1	$ 4800	no

Presentation format

RULES

Anchors



BRIEFLY ABOUT A FEW XAI METHODS

- SHAP
- ALEPlots
- Counterfactual explanations

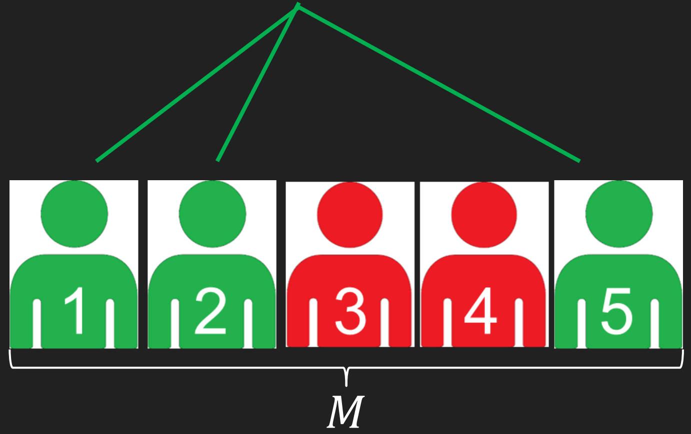
SHAP

Shapley values (game theory)

- Concept from (cooperative) game theory in the 1950s
- Used to distribute the total payoff to the players
- Explicit formula for the "fair" payment to every player j:

$$\phi_j = \sum_{S \subseteq M \setminus \{j\}} \frac{|S|!\,(|M| - |S| - 1)!}{|M|!} \big(v(S \cup \{j\}) - v(S)\big)$$

$v(S)$ is the payoff with only players in subset S



- Several mathematical optimality properties

Shapley values for prediction explanation (SHAP)

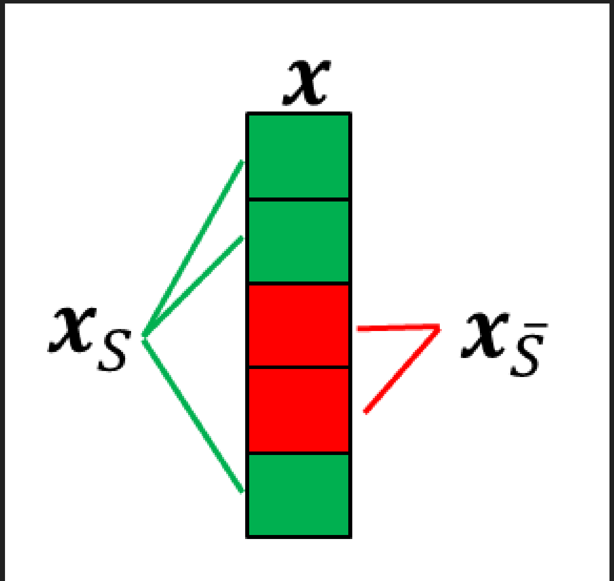
- Approach popularised by Lundberg & Lee (2017)
  - Players = features ($x_1, \dots, x_M$)
  - Payoff = prediction ($f(\boldsymbol{x}^*)$)
  - Contribution function: $v(S) = E[f(\boldsymbol{x})|\boldsymbol{x}_S = \boldsymbol{x}_S^*]$
  - Properties

$\phi_0 + \sum_{j=1}^{M} \phi_j = f(\boldsymbol{x}^*)$

$\phi_0 = E[f(\boldsymbol{x})]$

$f(\boldsymbol{x}) \perp\!\!\!\perp x_j$ implies $\phi_j = 0$

x_i, x_j same contribution implies $\phi_i = \phi_j$



- Interpretation of ϕ_j: **The prediction change caused by observing the value of** x_j – averaged over whether the other features were observed or not

Two main challenges

1. Scalability: The computational complexity in the Shapley formula is of size 2^M

$$\phi_j = \sum_{S \subseteq M \setminus \{j\}} \frac{|S|!\,(|M| - |S| - 1)}{|M|!} \big(v(S \cup \{j\}) - v(S)\big)$$

2. Estimating the contribution function

$$v(S) = E[f(\boldsymbol{x}) | \boldsymbol{x}_S = \boldsymbol{x}_S^*] = \int f(\boldsymbol{x}_{\bar{S}}, \boldsymbol{x}_S^*) p(\boldsymbol{x}_{\bar{S}} | \boldsymbol{x}_S = \boldsymbol{x}_S^*) \mathrm{d}\boldsymbol{x}_{\bar{S}}$$

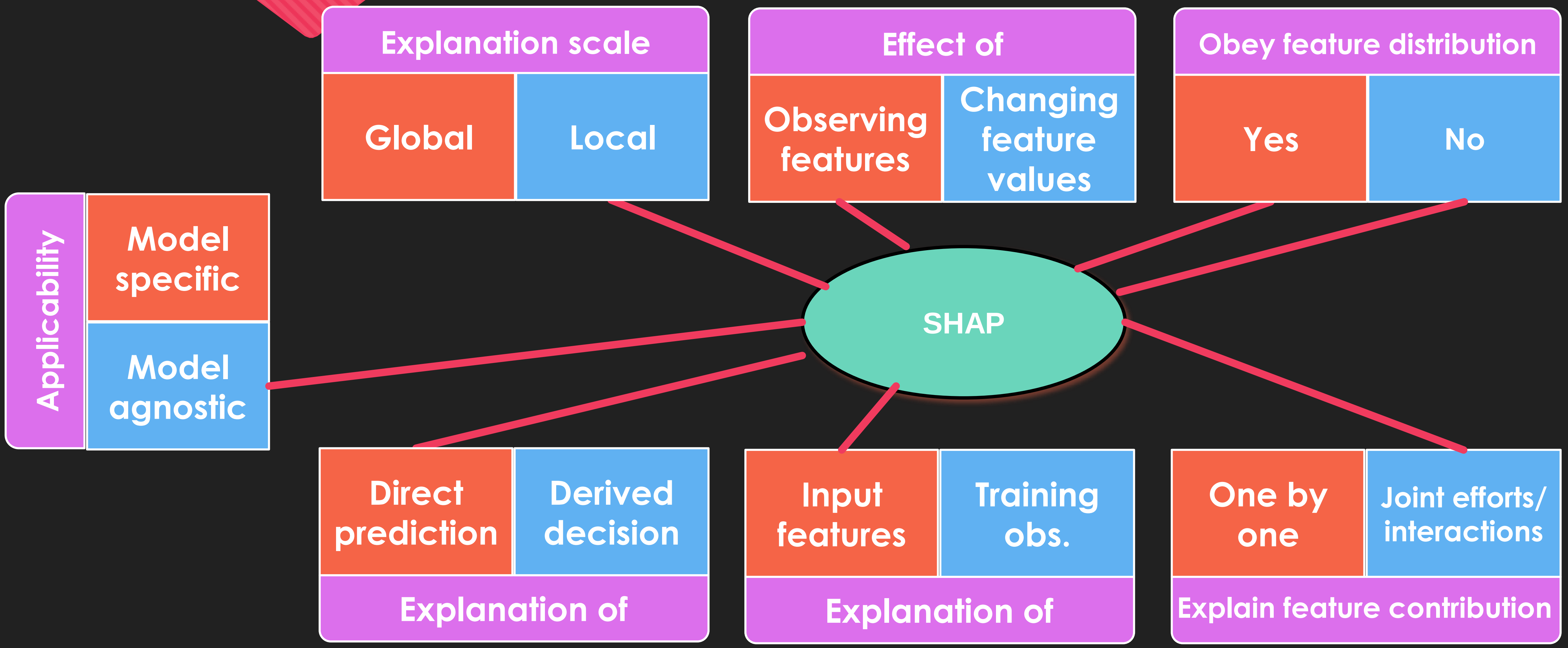
Nice to know

- Currently the most used XAI method
- It is crucial to acknowledge feature dependence
  - The method for estimating v(S) proposed by Lundberg & Lee (2017) ignores feature dependence by replacing $p(\boldsymbol{x}_{\bar{S}}|\boldsymbol{x}_S = \boldsymbol{x}_S^*)$ with $p(\boldsymbol{x}_{\bar{S}})$

$$v(S) = E[f(\boldsymbol{x})|\boldsymbol{x}_S = \boldsymbol{x}_S^*] = \int f(\boldsymbol{x}_{\bar{S}}, \boldsymbol{x}_S^*)p(\boldsymbol{x}_{\bar{S}}|\boldsymbol{x}_S = \boldsymbol{x}_S^*)\mathrm{d}\boldsymbol{x}_{\bar{S}}$$

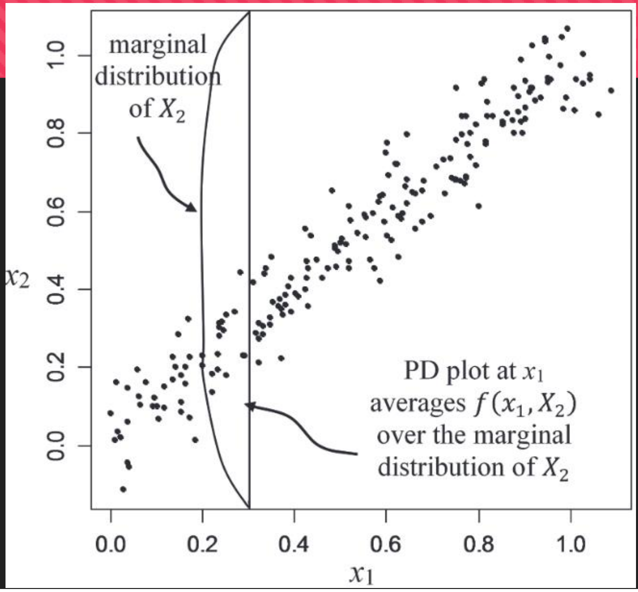
  - The feature dependence issue can be fixed by estimating $p(\boldsymbol{x}_{\bar{S}}|\boldsymbol{x}_S = \boldsymbol{x}_S^*)$ properly, but at higher comp. cost
- TreeSHAP
  - A fast model-specific way to compute SHAP values for tree models, utilizing their structure
  - Directly available in XGBoost, LightGBM, CatBoost
  - Not good at accounting for the feature dependence
- Software
  - Python: SHAP Python library (ignores feature dependence)
  - R: shapr (with python wrapper shaprpy) allows account for the feature dependence

METHOD CLASSIFICATION



ALEPlots

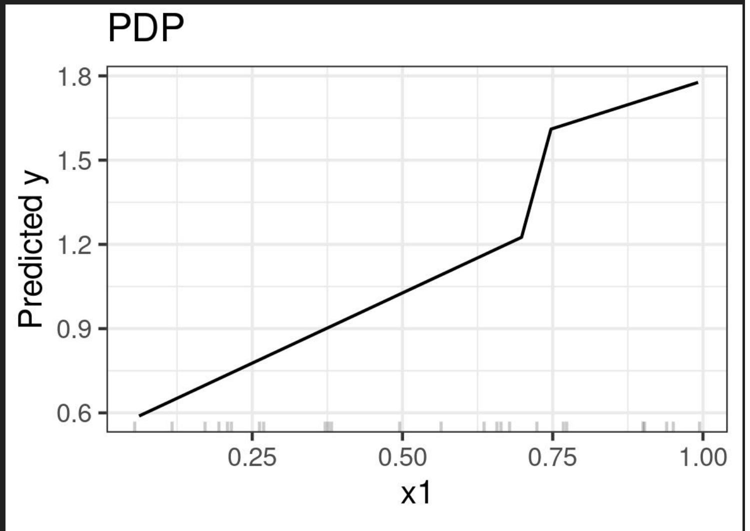
Partial Dependence Plots (PDP)

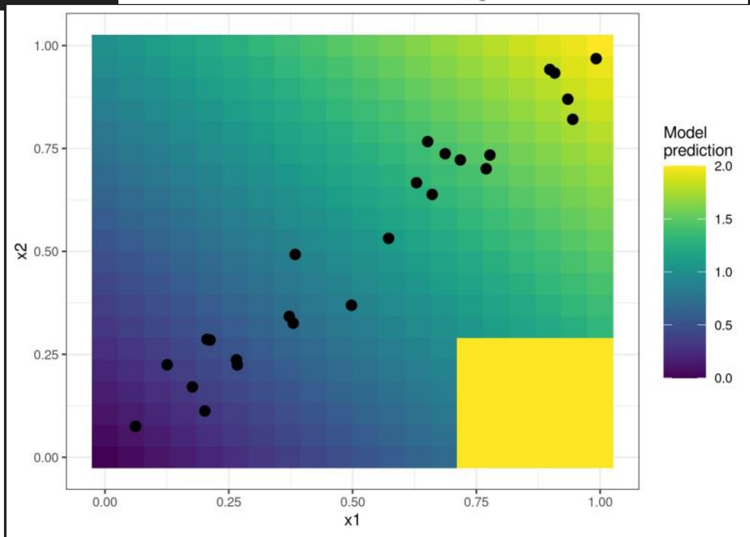


- PDP of a feature shows the marginal effect the feature has on the predicted outcome of the model.

$$f_{1,\mathrm{PD}}(x_1) \equiv \mathbb{E}[f(x_1, X_2)] = \int p_2(x_2)\, f(x_1, x_2)\, \mathrm{d}x_2$$

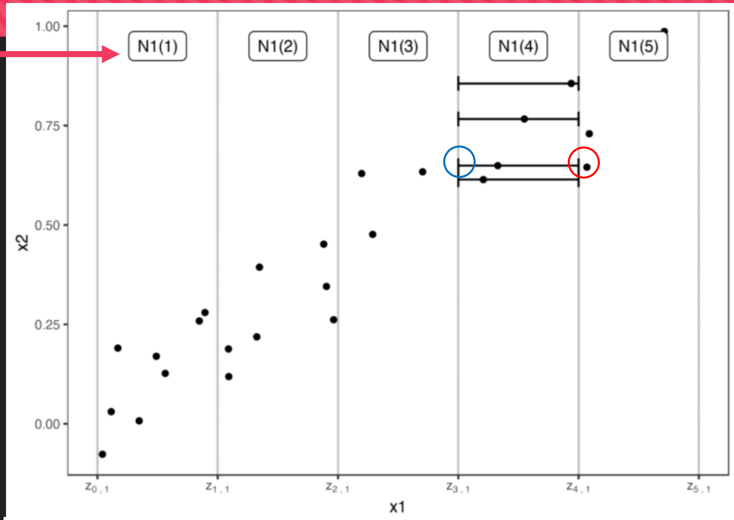
- In practice:
  1. Divide X_1 into n segments.
  2. For each segment, calculate avg model prediction over the ***marginal distribution of*** X_2
- Problem
  - Feature dependence is ignored, sensitive to bad extrapolation

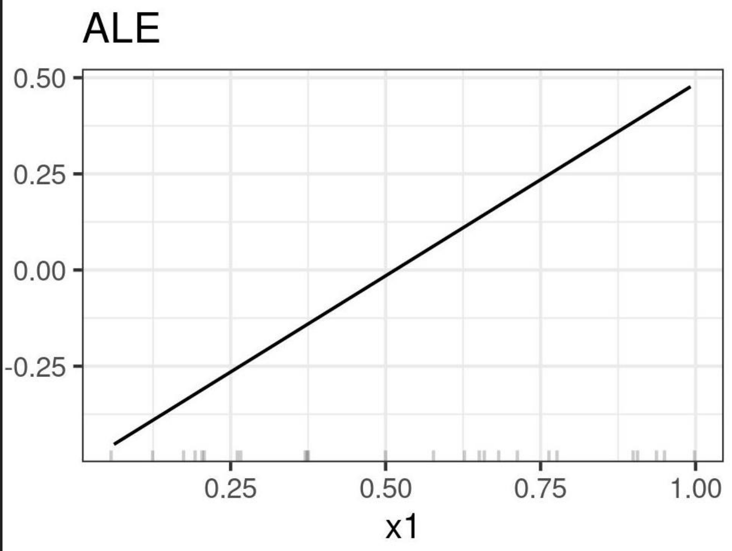




Accummulated Local Effect Plots (ALEPlots)

- The ALEPlot function for a given feature is the ***predicted response as a function of*** X_j, when all other features are averaged out.
  - Fixes the dependence/extrapolation issue by accumulating **local differences** $f(z_{1,upper}, x_2) - f(z_{1,lower}, x_2)$
- In practice:
  1. Divide X_1 into n segments.
  2. For each segment, calculate avg **local effect** $f(z_{1,upper}, x_2) - f(z_{1,lower}, x_2)$
  3. Take cumsum from N1(1) to N1(i).

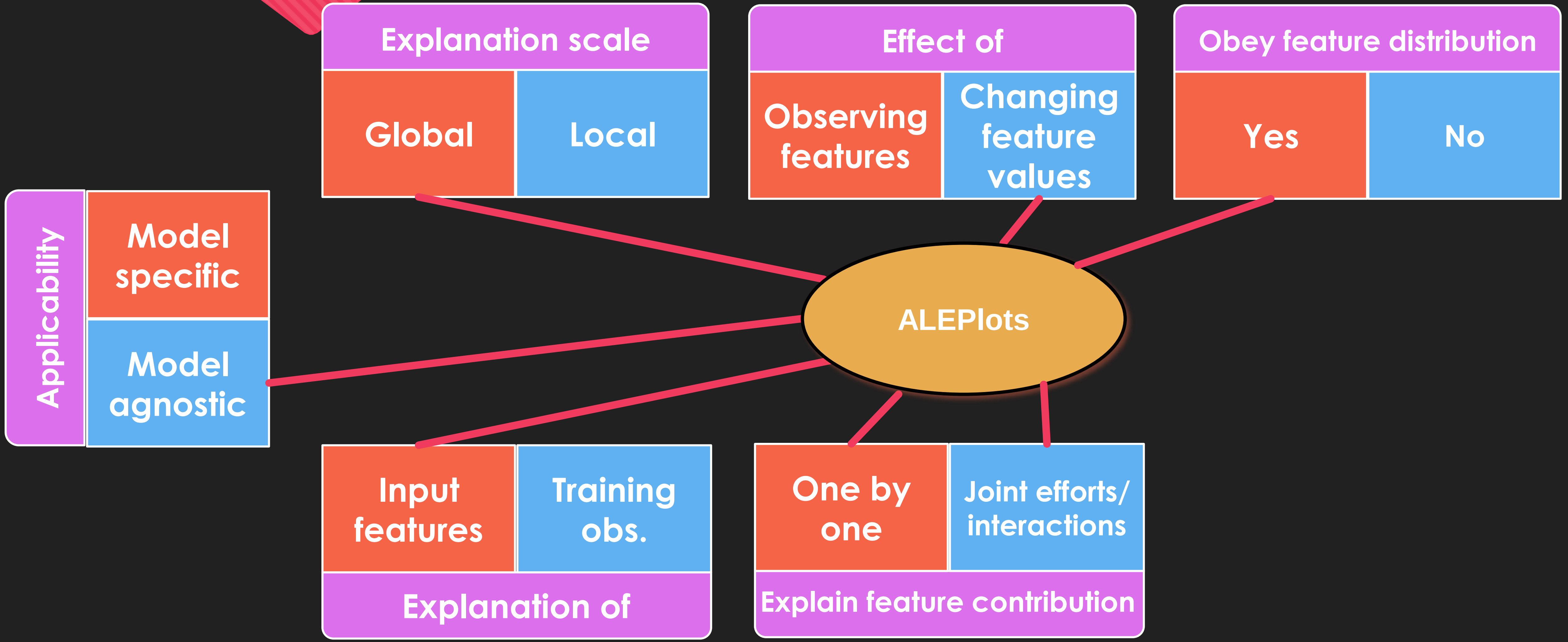




Nice to know

- Second-order ALEPlots can show the interaction effects of two features
  - Higher-order effects possible but hard to visualize
- Preferrable over methods like PDP which can give incorrect interpretations in the presence of feature dependence
- Must be interpreted locally
- Software
  - Python: Alibi
  - R: ALEPlot

METHOD CLASSIFICATION



Counterfactual explanations

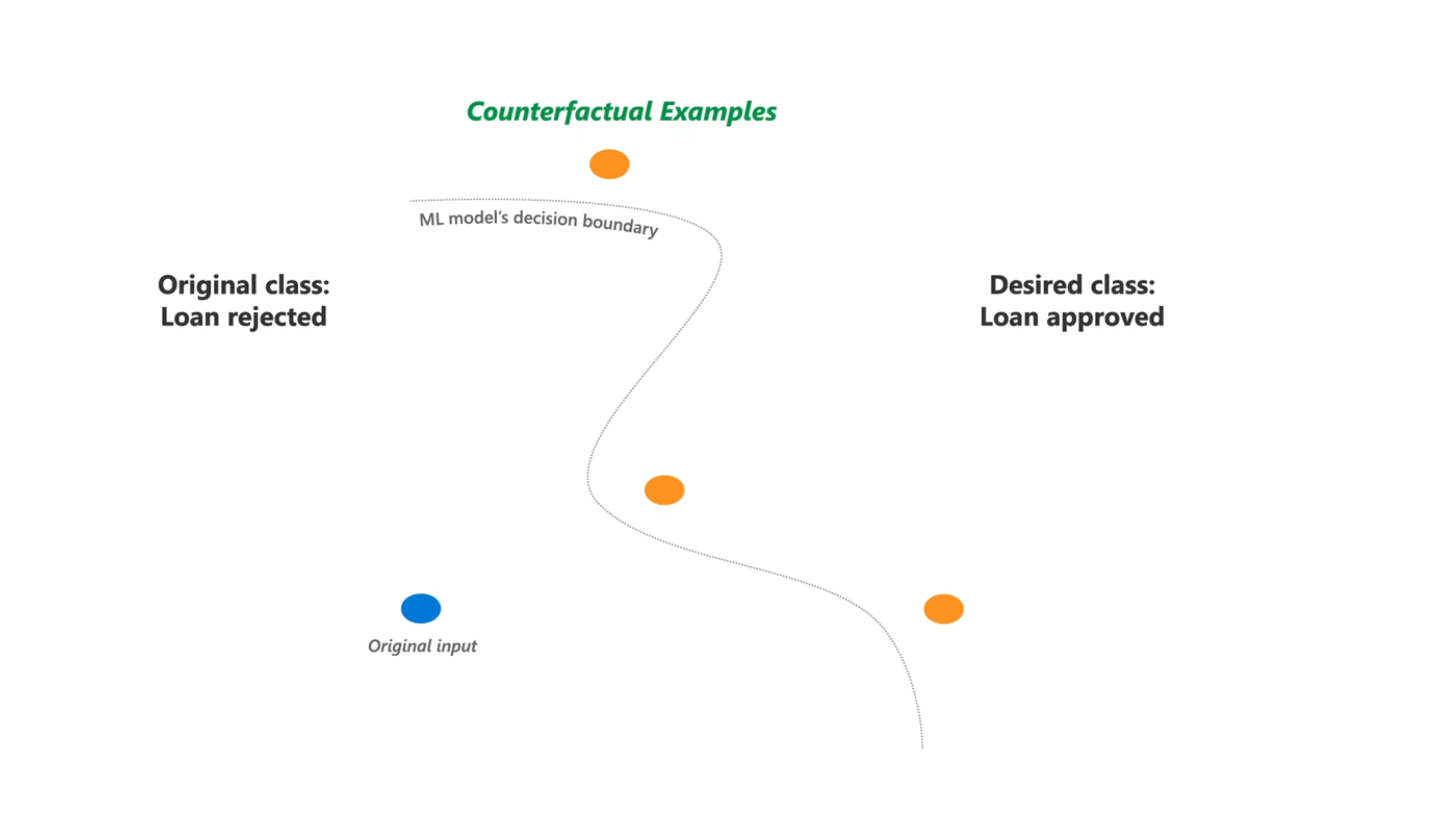
Return to introductory example

Case: Peter has features x^*, and got his loan application rejected as the model predicted 20% chance of default

Explainability question: What can Peter do to receive a loan?

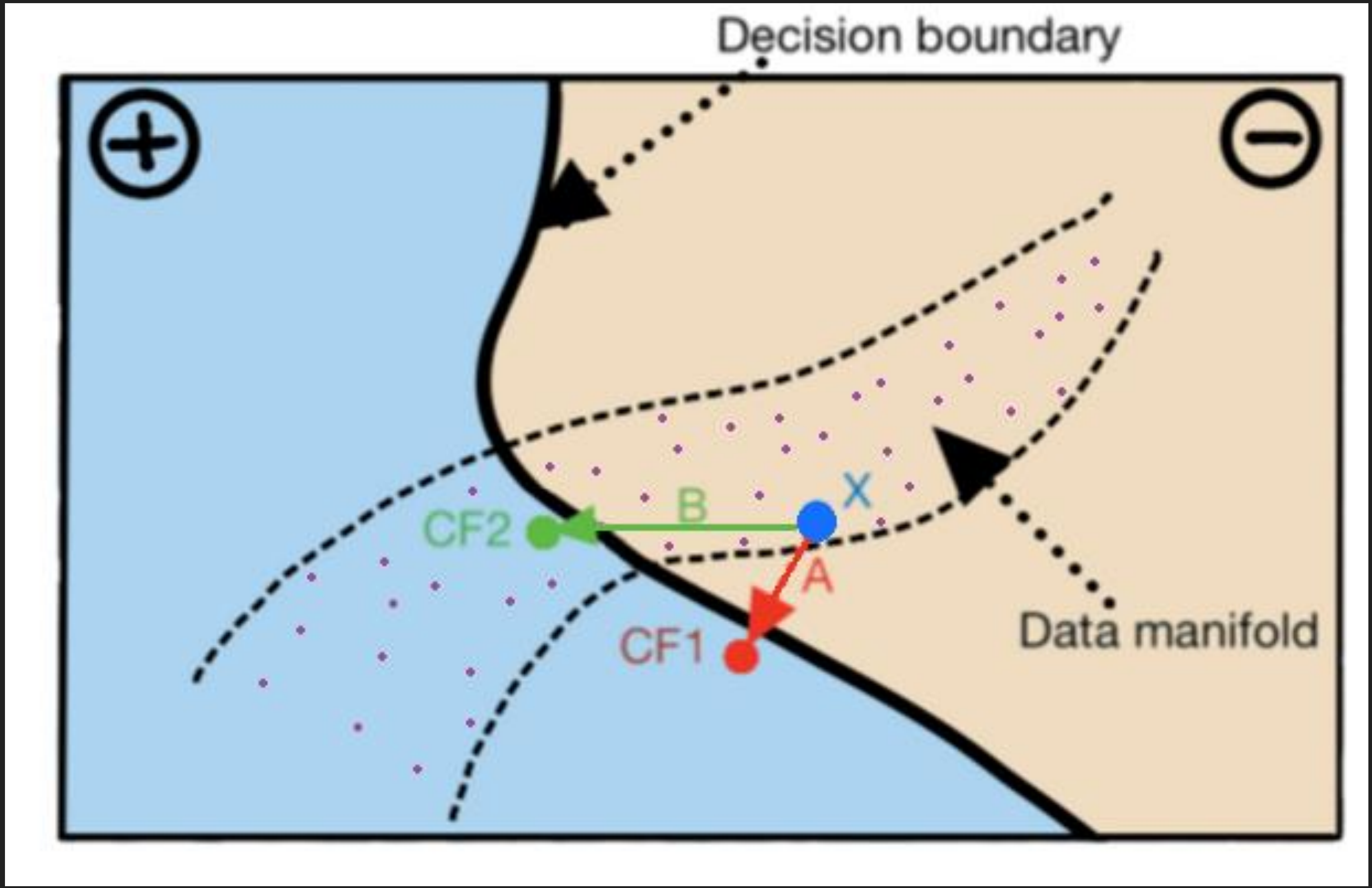
The idea behind counterfactual explanations (CE)

CE solution
Provide example(s) of (minimal) changes in features which approve the application



CE criteria

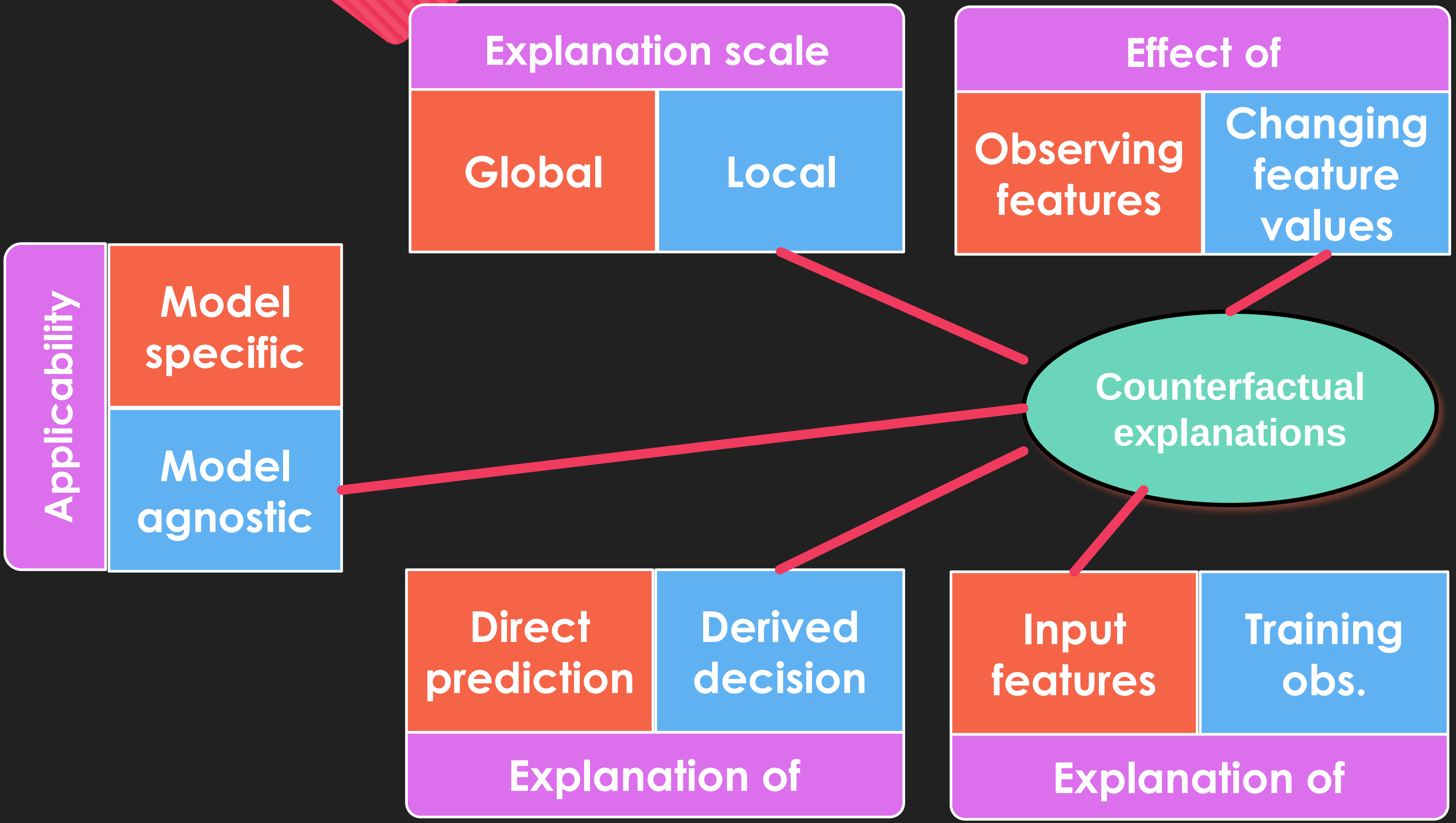
Desired properties
1. On-manifold
2. Actionable
3. Valid
4. Low cost



Nice to know

- A very user-friendly way to explain changes
- Also called algorithmic recourse
- The CE examples can be good or bad, but since they are just examples, they cannot be wrong
- Lots of CE methods – 3 classes
  - Optimization based
  - Heuristic based
  - Instance/model based
- Software
  - Python: CARLA (collection of CE methods + benchmarking)
  - R: counterfactuals (small collection of methods + benchmarking), mcceR (with Python wrapper mcceRpy)

METHOD CLASSIFICATION



NAVIGATING IN THE XAI JUNGLE

Which method should I use?

eXplego

An XAI-method selection tool by

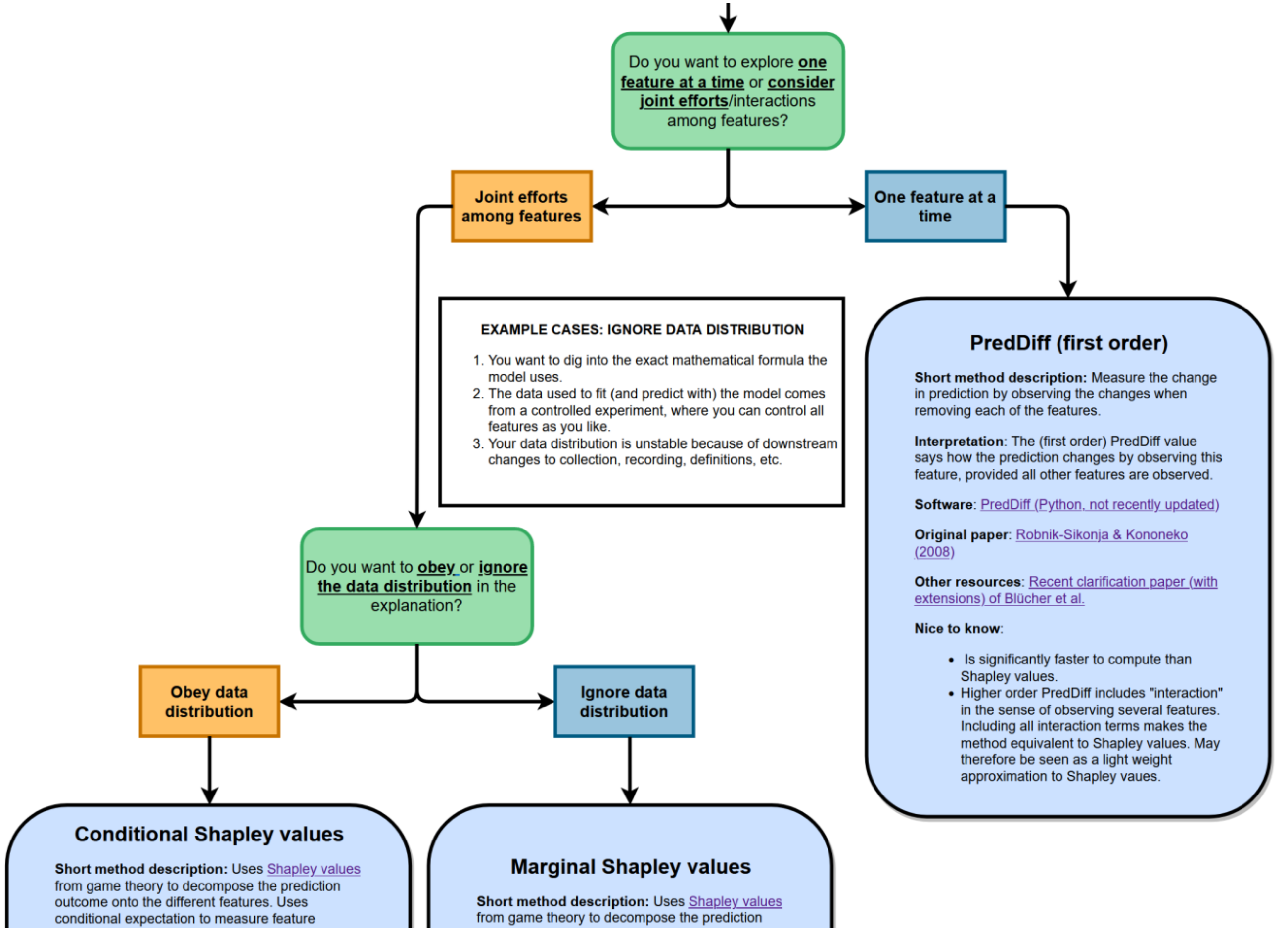
Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

nav

BigInsight

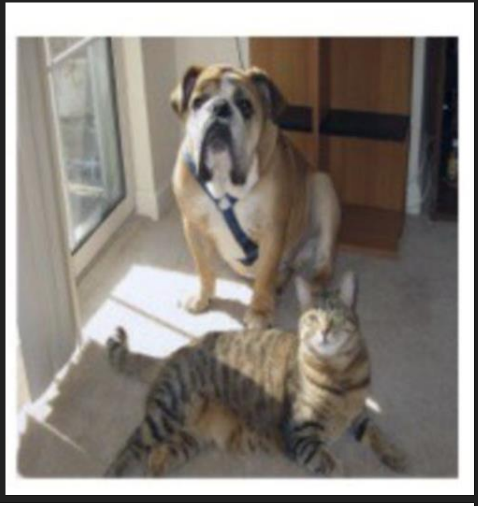
- Interactive decision tree, helping developers choose explanation method

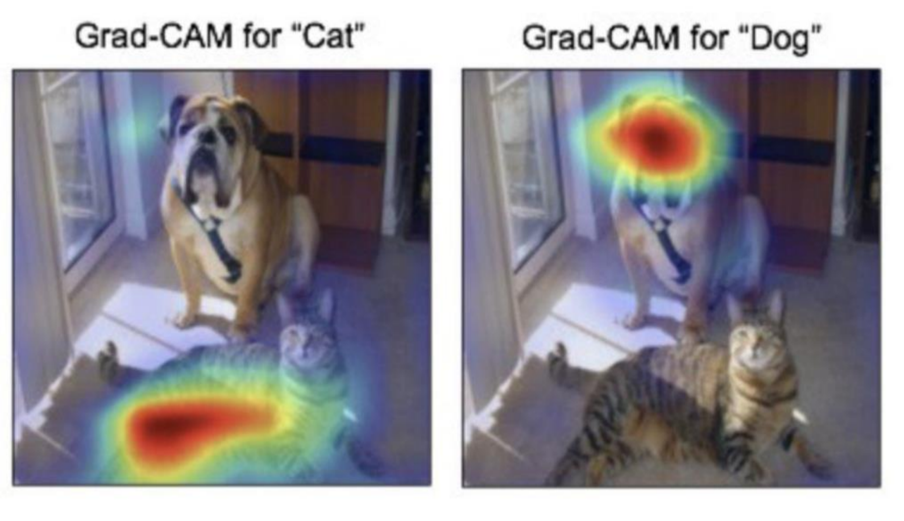
www.explego.nr.no



BONUS 1: EXPLAINING IMAGE MODELS

- For image data, explainability is most relevant for tasks like **image classification** and object detection/localization
- Can use model-agnostic methods, but it is often wise to utilize the **data structure** and the fact that the models are essentially always **neural networks**
- The most common explanation type is pixel attribution (Saliency maps)
    - Visualize which parts of the image that was most important for certain classification





- Review paper: Gupta et al (2023), Explainable Methods for Image-Based Deep Learning: A Review

BONUS 2: EXPLAINING TEXT MODELS

A big question for text models is what do we want to explain?

- Some explainability questions can be answered by general XAI methods:
  - Text/document classification: What part of the text was most important for a classification
  - Which of the previous words are most relevant when predicting the next one?
- What data sources was used by a chatbot to answer a question?
  - For LLMs with external databases (RAG = Retrieval Augmented Generation), we can see what external datasources was used to answer a question
- What parts of GPT-promt was most important when generating a response?
  - Attention mechanism weights from the transformer models can be used to highlight this
- Review paper: Zaho et al. (2023) Explainability for Large Language Models: A Survey

Martin Jullum
jullum@nr.no
martinjullum.com

NORSK REGNESENTRAL
NORWEGIAN COMPUTING CENTER

TAKE HOME POINTS

- XAI is fast-growing research field
- There is a jungle of XAI methods
  - Many XAI methods complement each other
  - it is important to pick the right method for the XAI question you want to answer
  - You should understand roughly what the XAI methods do in order not to interpret its output incorrectly
  - Beware of pitfalls of ignored feature dependence, extrapolation issues, too rough approximations

- Recommended reading: Molnar (2023), Interpretable Machine learning (free e-book)