

Data- and Human-aware Explainable AI

From Faithful to Human-Friendly Explanations

Luis Galárraga

RJCIA 2025

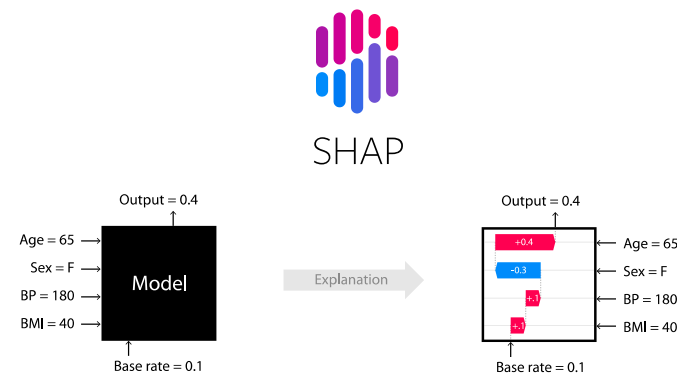
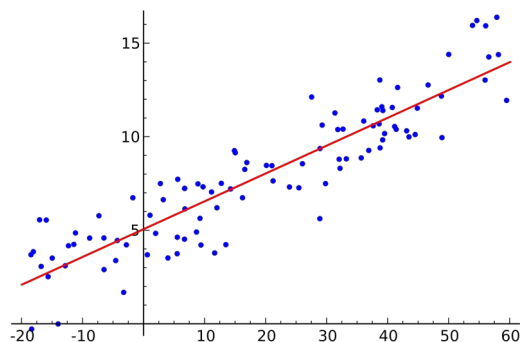
Dijon, 30/06/2025

Explainable AI: What? Why? How?

AI, ML, XAI



- Artificial Intelligence (AI)
 - Intelligent traits implemented in **algorithms**
- Machine Learning (ML)
 - The study of methods to **generalize** from data
- **Explainable AI (XAI)**
 - Making AI-powered solutions **understandable** to **us**



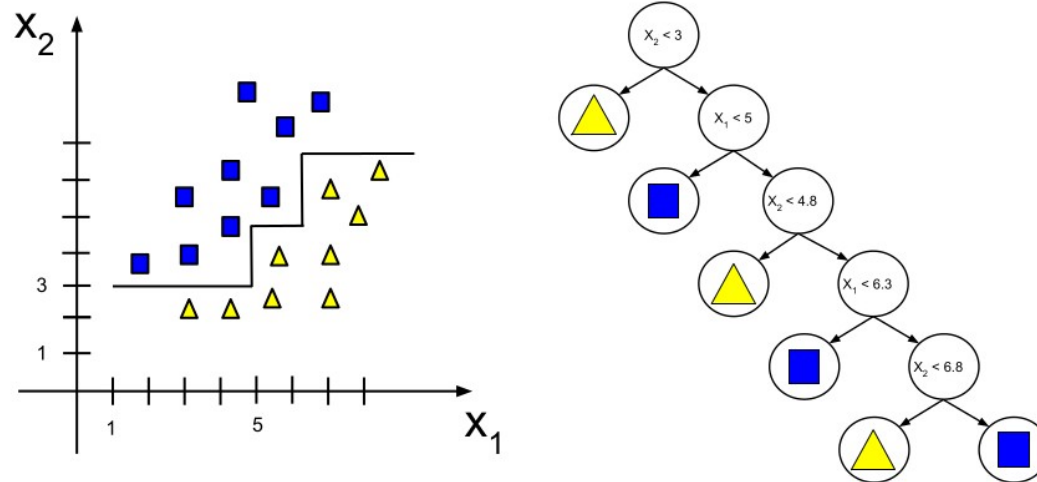
Source: <https://github.com/shap/shap>

Explainable AI: What?



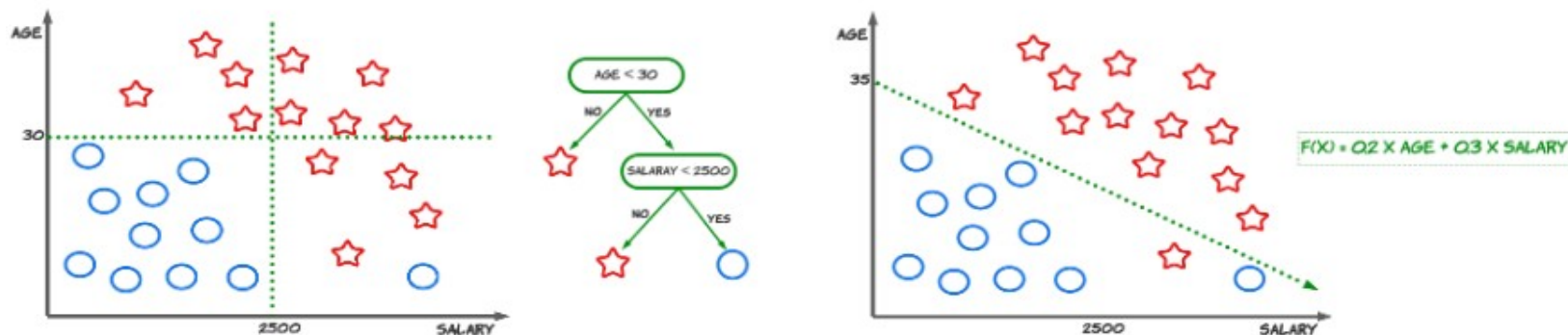
Interpretability

A **model** can be said to be **interpretable** if, within a **given time limit**, the level of expertise of the user allows them to **understand** the model through its **representation**

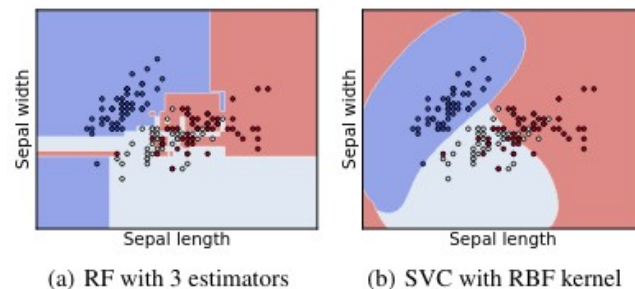


Explainability

The **explainability** of a model refers to its capacity to be explained by (external) tools or techniques



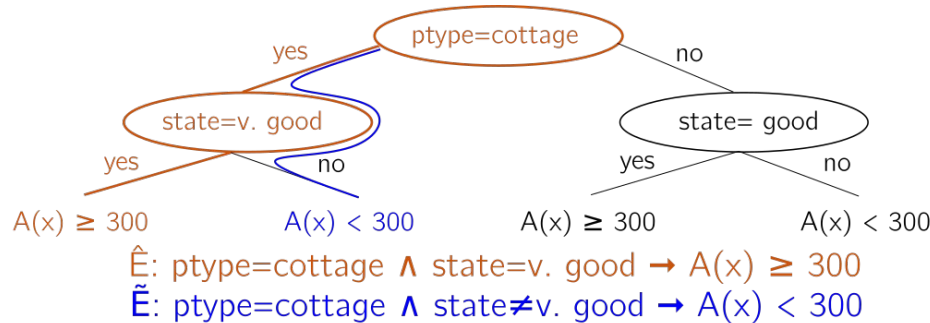
Source: J. Delaunay. Explainability for Machine Learning Models: From Data Adaptability to User Perception. PhD Thesis, Université de Rennes, 2023.



A. Bibal. Interpretability and Explainability in Machine Learning with Application to Nonlinear Dimensionality Reduction. PhD Thesis. University of Namur, Belgium, 2020
T. Laugel, et al. The dangers of post-hoc interpretability: Unjustified counterfactual explanations. International Joint Conference on Artificial Intelligence (IJCAI). 2019

What is an explanation?

Statement or asset that characterizes the (causal?) **relationships** between the **inputs** and **outputs** of an AI model



Prediction probabilities

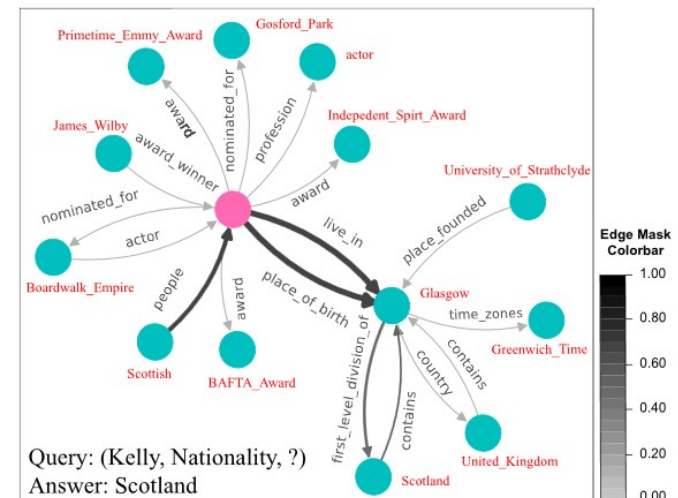
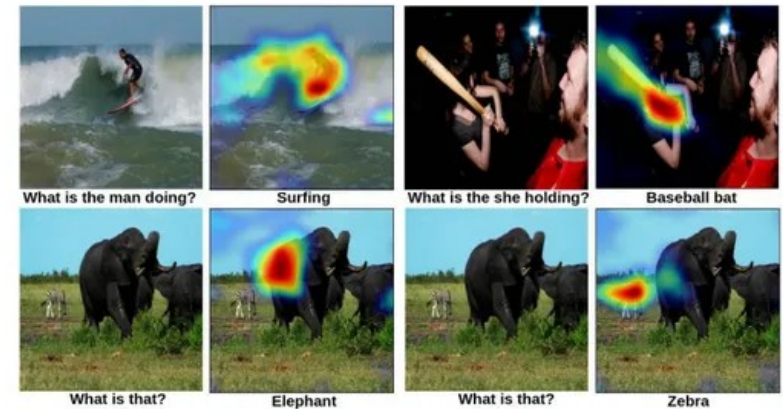


Text with highlighted words

From: salem@pangea.Stanford.EDU (Bruce Salem)
 Subject: Re: Science and theories
 Organization: Stanford Univ. Earth Sciences
 Lines: 42
 NNTP-Posting-Host: pangea.stanford.edu

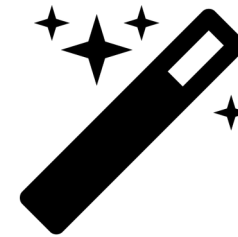
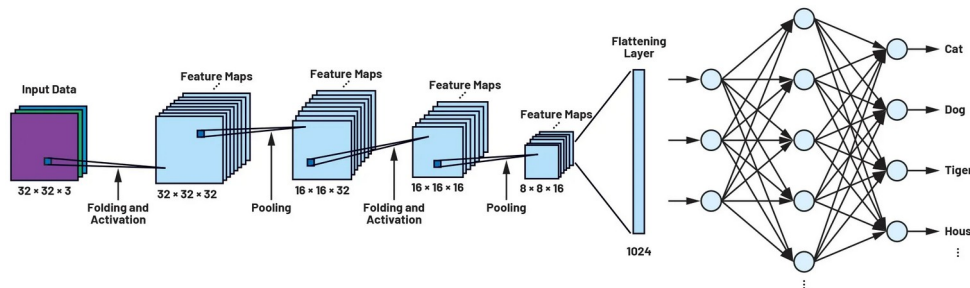
In article IC5u7Bq.J43@news.cso.uiuc.edu
 cobb@alexia.lis.uiuc.edu (Mike Cobb) writes:
 IAs per various threads on science and creationism, I've
 started dabbling into a
 lbook called Christianity and the Nature of Science by JP
 Moreland.

As I don't know this book, I will use your heresay.



Why do we care about XAI?

- AI models resemble sorcery
 - Due to their complexity or the way they are deployed



Sorcery!!!



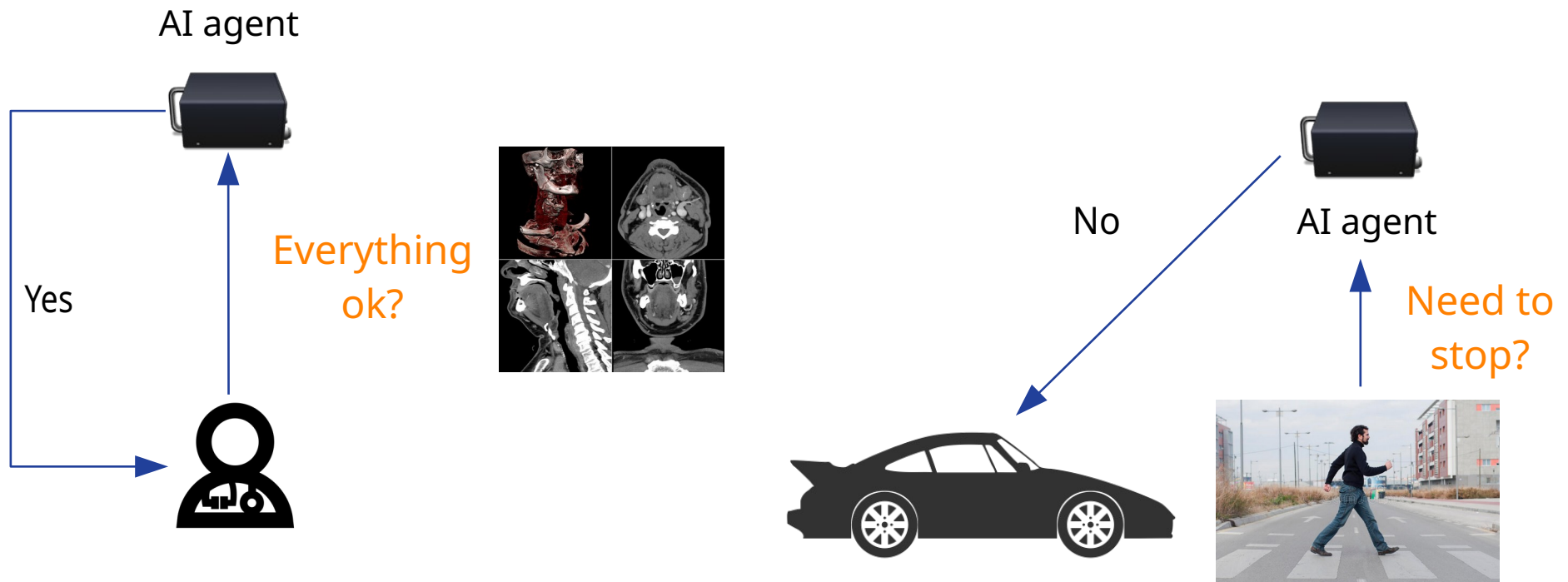
Classifier



That is
a **cat**

Why do we care about XAI?

- AI models are used to make critical decisions
 - We need to understand the logic behind their answers



Why do we care about XAI?

- For auditing
 - Tuning, debugging, detecting biases, etc.



Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

Just as the 18-year-old girls were realizing they were too big for the tiny conveyances — which belonged to a 6-year-old boy — a woman came running after them saying, "That's my kid's stuff." Borden and her friend immediately dropped the bike and scooter and walked away.

But it was too late — a neighbor who witnessed the heist had already called the police. Borden and her friend were arrested and charged with burglary and petty theft for the items, which were valued at a total of \$80.

NEWS > TECHNOLOGY

Dutch scandal serves as a warning for Europe over risks of using algorithms

The Dutch tax authority ruined thousands of lives after using an algorithm to spot suspected benefits fraud — and critics say there is little stopping it from happening again.

SHARE

POLITICO Free article usually reserved for subscribers



As the world turns to AI to automate their systems, the Dutch scandal shows how devastating they can be | Dean Mouhtaropoulos/Getty Images

REUTERS

World Business Markets Breakingviews Video More

RETAIL OCTOBER 11, 2018 / 1:04 AM / UPDATED 4 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

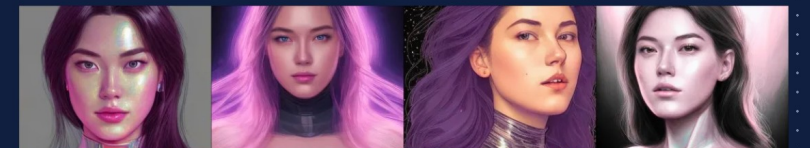
ARTIFICIAL INTELLIGENCE

The viral AI avatar app Lensa undressed me —without my consent

My avatars were cartoonishly pornified, while my male colleagues got to be astronauts, explorers, and inventors.

By Melissa Heikkilä

December 12, 2022



<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
<https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/>
<https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G>
<https://www.technologyreview.com/2022/12/12/1064751/the-viral-ai-avatar-app-lensa-undressed-me-without-my-consent/>

Why do we care about XAI?

- For legal and ethical reasons^(‡)
 - **General Data Protection Regulation**^(*)
 - The European AI Act^(**)
 - Ethics guidelines for trustworthy AI⁽⁺⁾
 - EU Directive 2014/65 on Markets in Financial Instruments^(†)
 - Directive 2011/83 on Consumer Rights^(§)

(‡) Bibal et al. Legal Requirements on Explainability in Machine Learning, Artificial Intelligence and Law, vol. 29, no. 2, pp. 149-169. <https://doi.org/10.1007/s10506-020-09270-4>

(*) See Recital 71, <https://www.privacy-regulation.eu/en/r71.htm>, see also the Guidelines on Automated Individual Decision-making: <https://ec.europa.eu/newsroom/article29/items/612053/en>

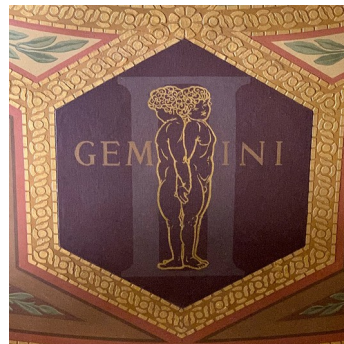
(**) <https://artificialintelligenceact.eu/>

(+) <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

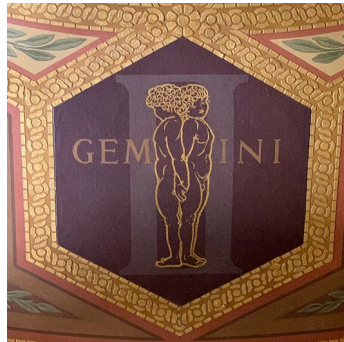
(†) See Article 7, <https://eur-lex.europa.eu/eli/dir/2014/65/oj/eng>

(§) See Article 6, <https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex:32011L0083>

The two faces of XAI



XAI in the AI and ML Communities



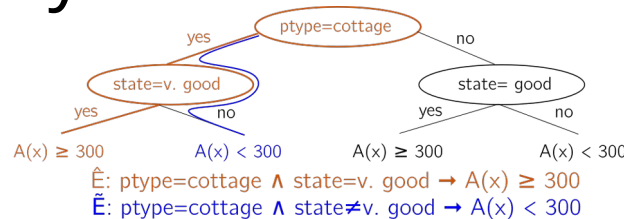
Supervised ML: Glass vs. Black

Interpretable



- Linear functions
- Decision (Reg.) Trees
- Rule-based models
- Exemplar-based methods
- Naive Bayes

– RuleFit



$$y = -189.69 - 0.0002 \times \text{cases} + 2.39 \times \text{score} + 5.08 \times \text{age},$$

Not always accurate
but simpler

Black-box



- Neural Networks
- Ensemble methods
 - Random Forests
 - Gradient Boosting
- Support Vector Machines

.....

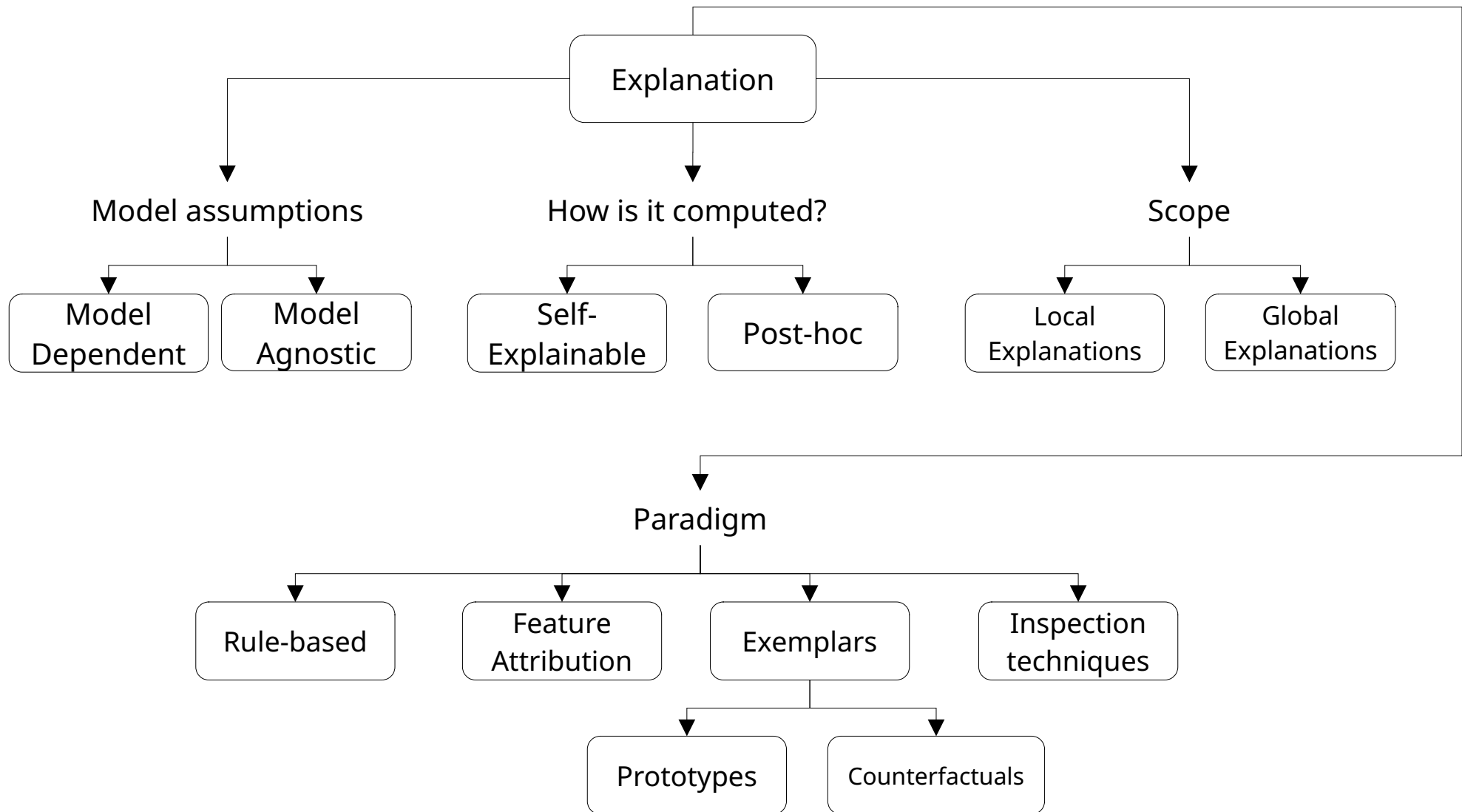
Often accurate but
not interpretable

Pure CS-style XAI

- AI and ML researchers have focused on producing:
 - Novel **interpretable** and **accurate** AI models
 - **Faithful** post-hoc explainability modules
 - Studies of the properties of models and explanations
 - Robustness
 - Stability
 - Diversity
 - Recall⁽⁺⁾
 - Complexity
 - Plausibility/Actionability
 - Fidelity/Accuracy/Adherence
 - Runtime⁽⁺⁾
- They often put less emphasis in user evaluation

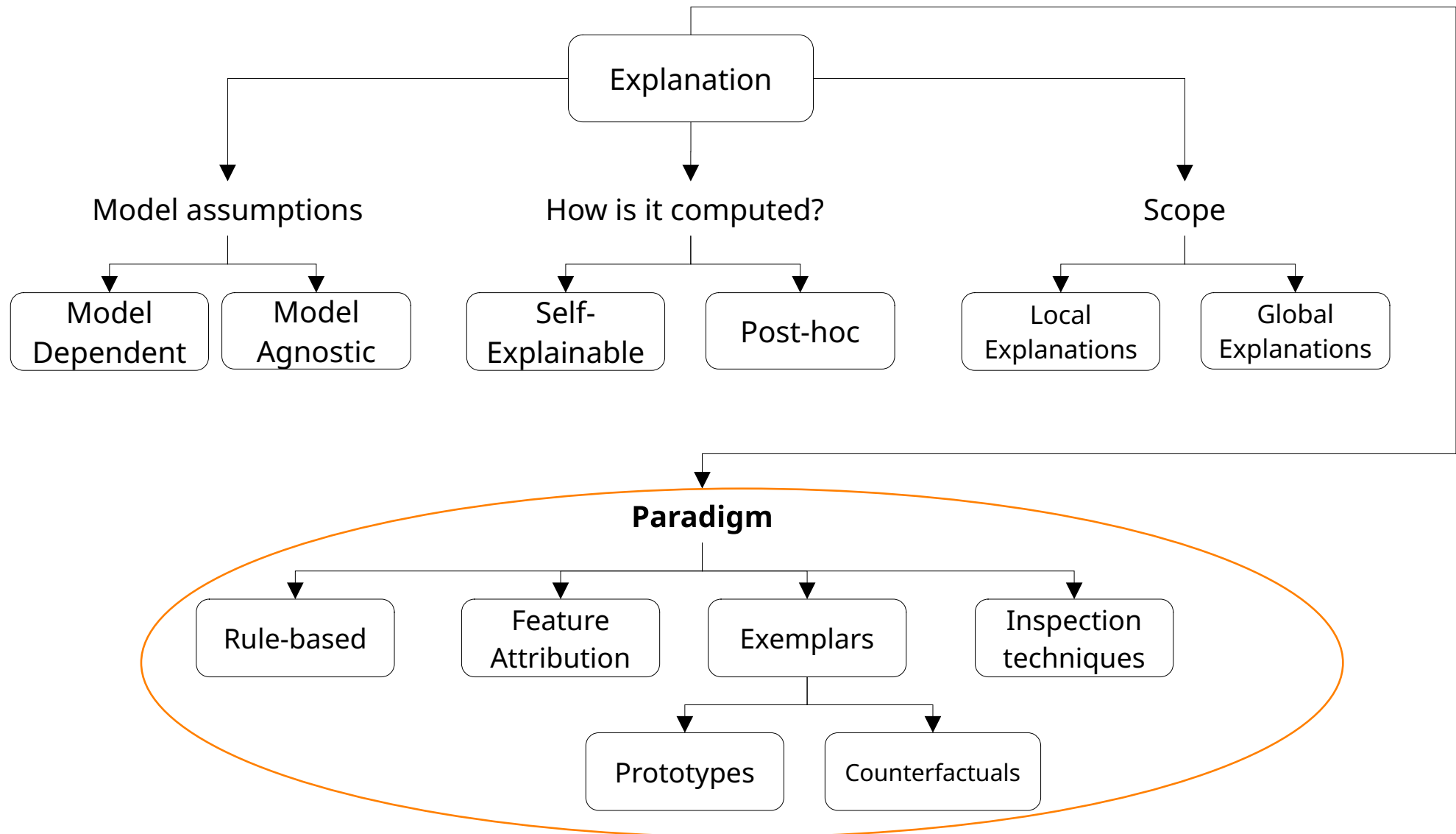
(+) Properties of the methods to compute the explanations

Taxonomy of XAI Techniques



Taken from: J. Delaunay. Explainability for Machine Learning Models: From Data Adaptability to User Perception. PhD Thesis, Université de Rennes, 2023

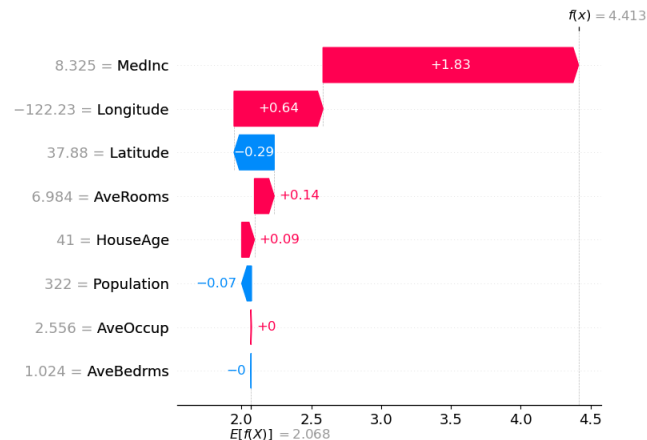
Taxonomy of XAI Techniques



Taken from: J. Delaunay. Explainability for Machine Learning Models: From Data Adaptability to User Perception. PhD Thesis, Université de Rennes, 2023

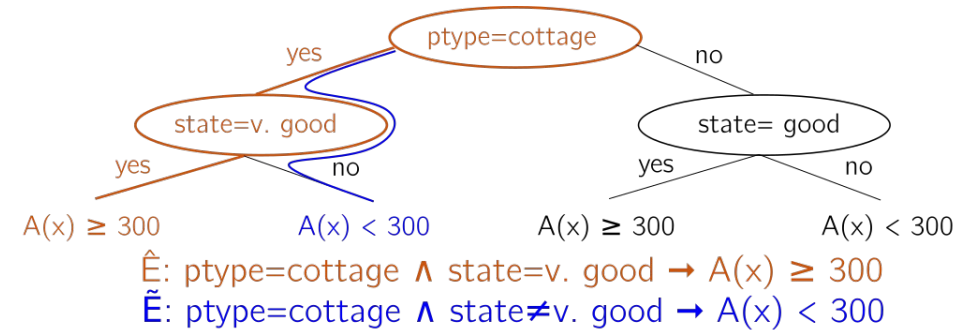
Explanation Paradigms

Feature-Attribution

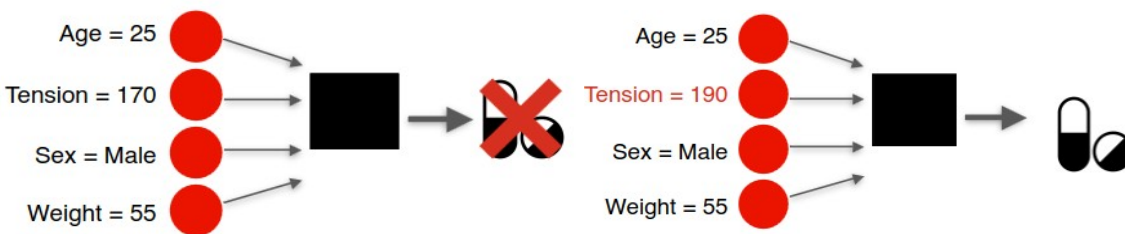


Source: <https://github.com/shap/shap>

Rules

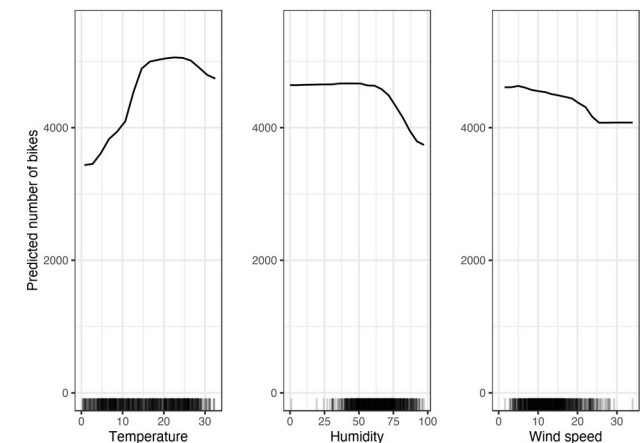


Factuals and Counterfactuals



Source: J. Delaunay. Explainability for Machine Learning Models: From Data Adaptability to User Perception. PhD Thesis, Université de Rennes, 2023.

Inspection Techniques



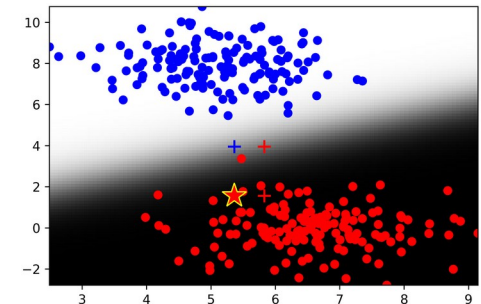
Source: <https://christophm.github.io/interpretable-ml-book/pdp.html> 18

Evaluating Explanations

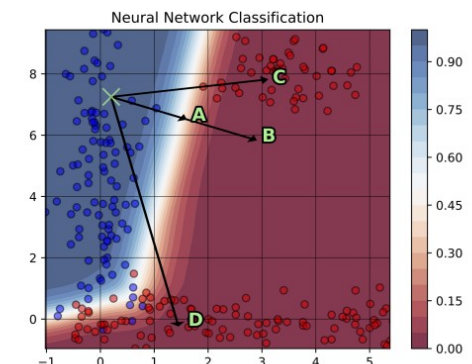
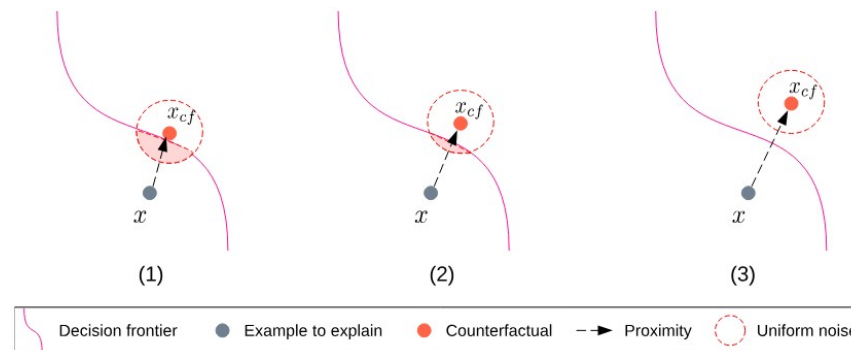
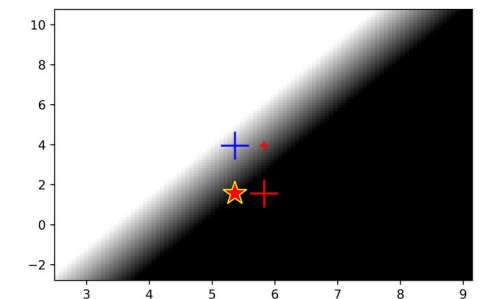
- Complexity
- Fidelity
 - Adherence
 - Sensitivity/Infidelity
- Stability & Robustness
- Realism
- Diversity



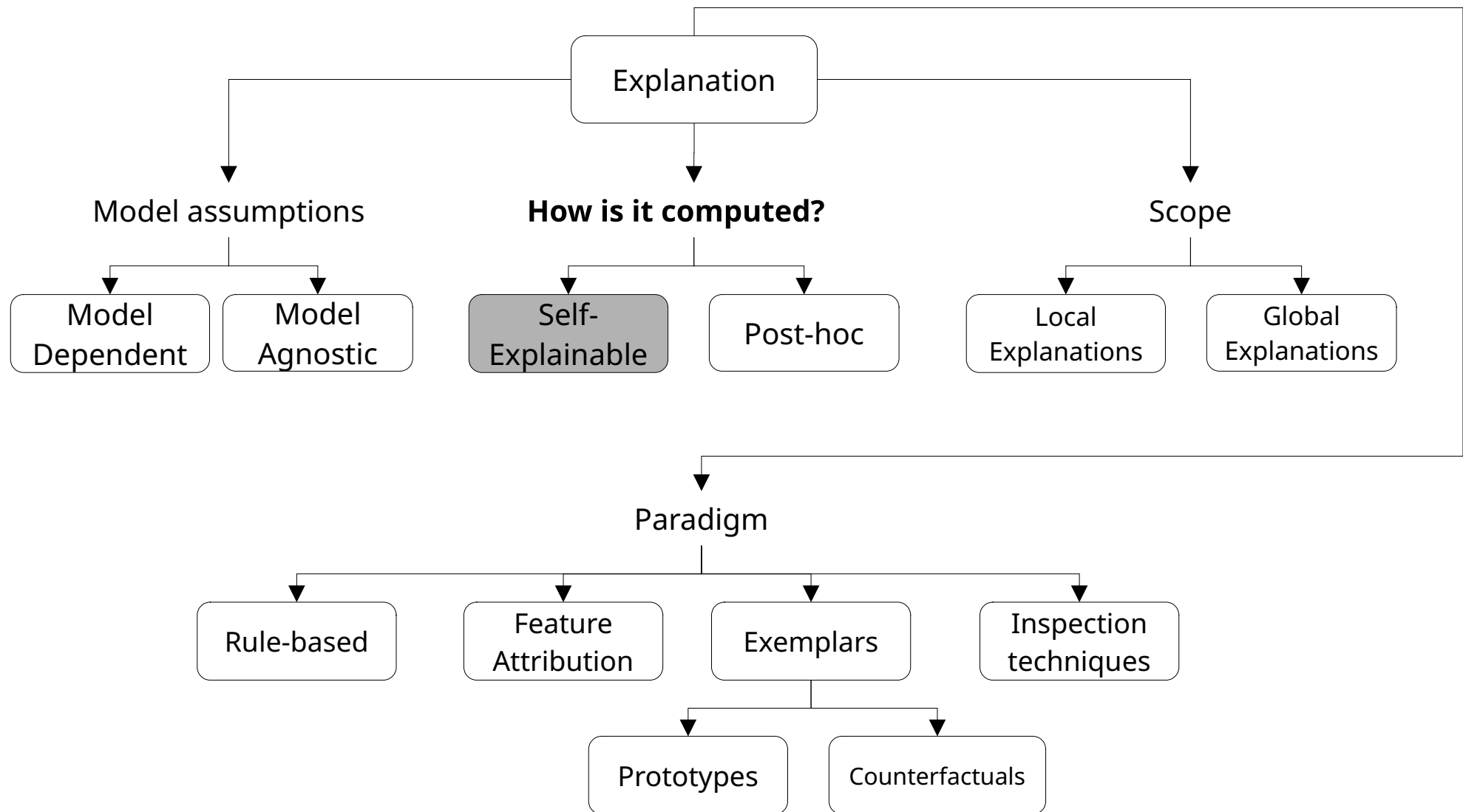
A classifier



LIME explanation



Taxonomy of XAI Techniques



Self-explainable methods

Some neural approaches learn to **predict & explain** at the same time

SENN(+)

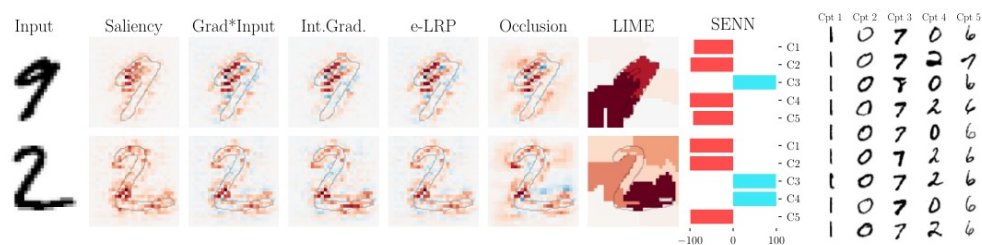
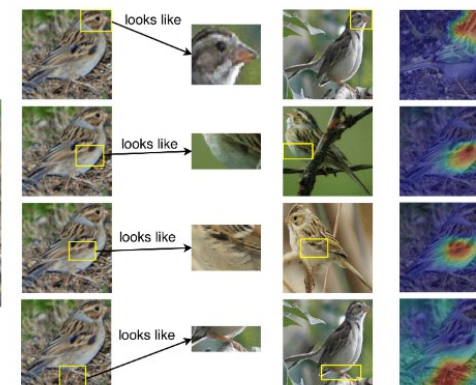


Figure 2: A comparison of traditional input-based explanations (positive values depicted in red) and SENN's concept-based ones for the predictions of an image classification model on MNIST. The explanation for SENN includes a characterization of concepts in terms of defining prototypes.



ProtoPNet(+)

Leftmost: a test image of a clay-colored sparrow
Second column: same test image, each with a bounding box generated by our model -- the content within the bounding box is considered by our model to look similar to the prototypical part (same row, third column) learned by our algorithm
Third column: prototypical parts learned by our algorithm
Fourth column: source images of the prototypical parts in the third column
Rightmost column: activation maps indicating how similar each prototypical part resembles part of the test bird

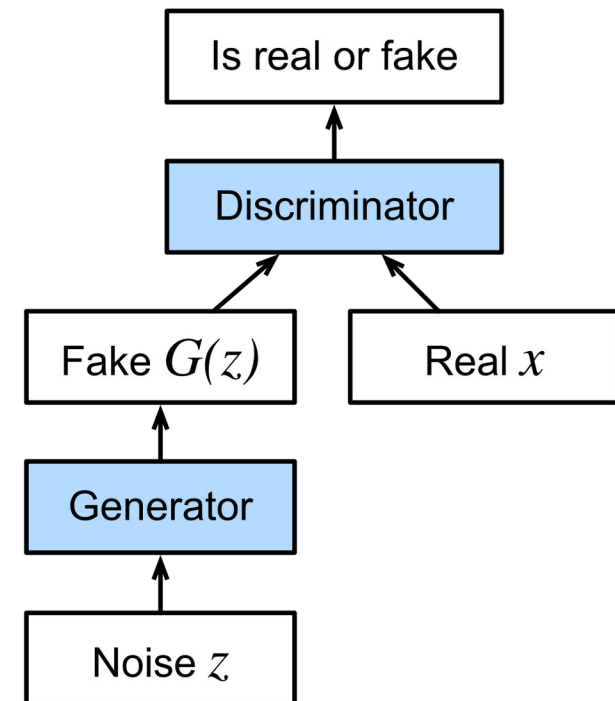
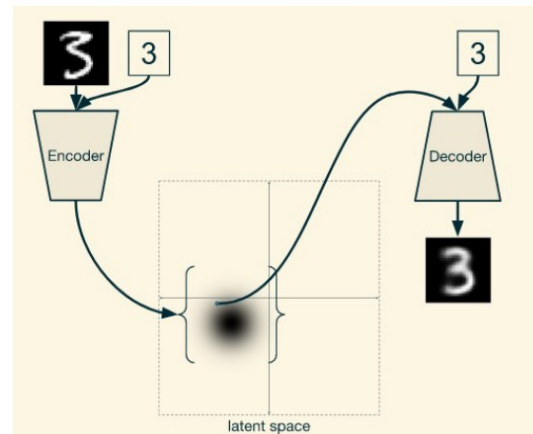
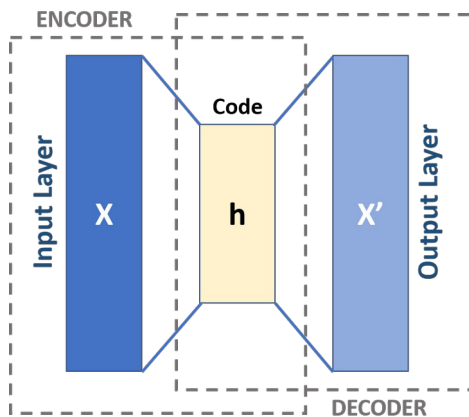
(*) D. Alvarez-Melis and T.S. Jaakkola. Towards Robust Interpretability with Self-Explaining Neural Networks. <https://arxiv.org/pdf/1806.07538.pdf>, 2018.

(+) C. Chen et al. This Looks Like That: Deep Learning for Interpretable Image Recognition. Advances in Neural Information Processing Systems 32

<https://arxiv.org/abs/1806.10574>, 2019.

Self-explainable methods

GANs and VAEs are common architectures to explain an outcome via **realistic** exemplars



V. Guyomard et al. VCNet: A Self-explaining Model for Realistic Counterfactual Generation. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, <https://is.gd/FEkx0f>, 2022.

Y. Wang et al. Adversarial Regularization for Explainable-by-Design Time Series Classification. International Conference on Tools with Artificial Intelligence (ITCAI), 2020.

Self-explainable methods

GANs and VAEs are common architectures to explain an outcome via **realistic** exemplars

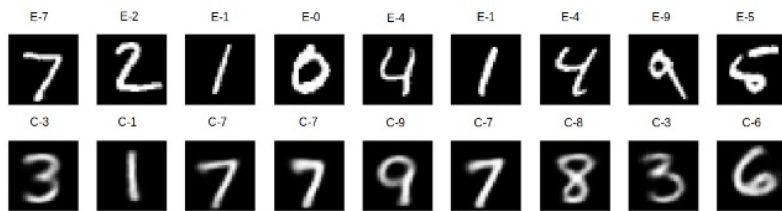


Figure 5: Counterfactuals obtained with VCNet for the MNIST dataset. The top line corresponds to the examples to explain, the bottom to the corresponding counterfactuals.

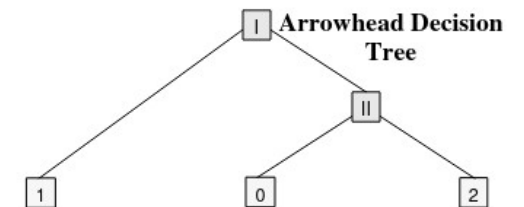
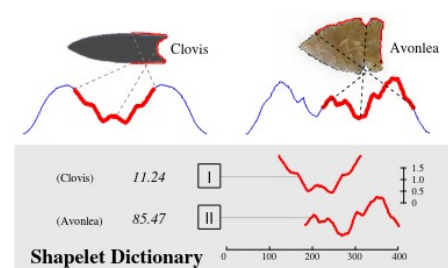
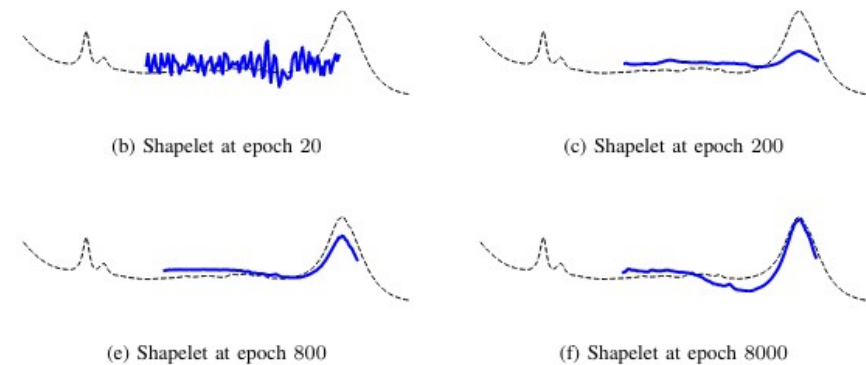
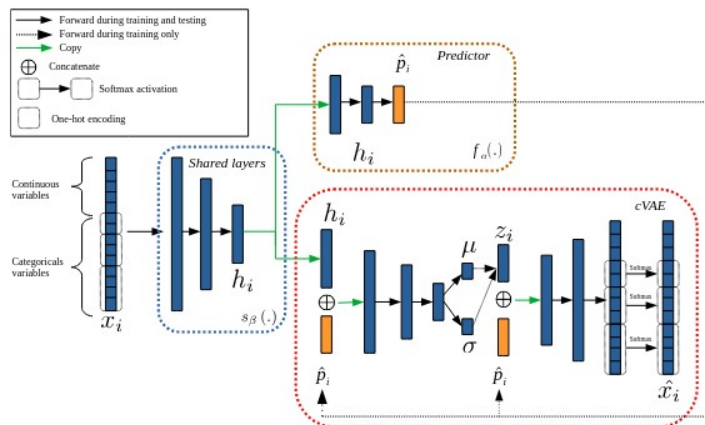


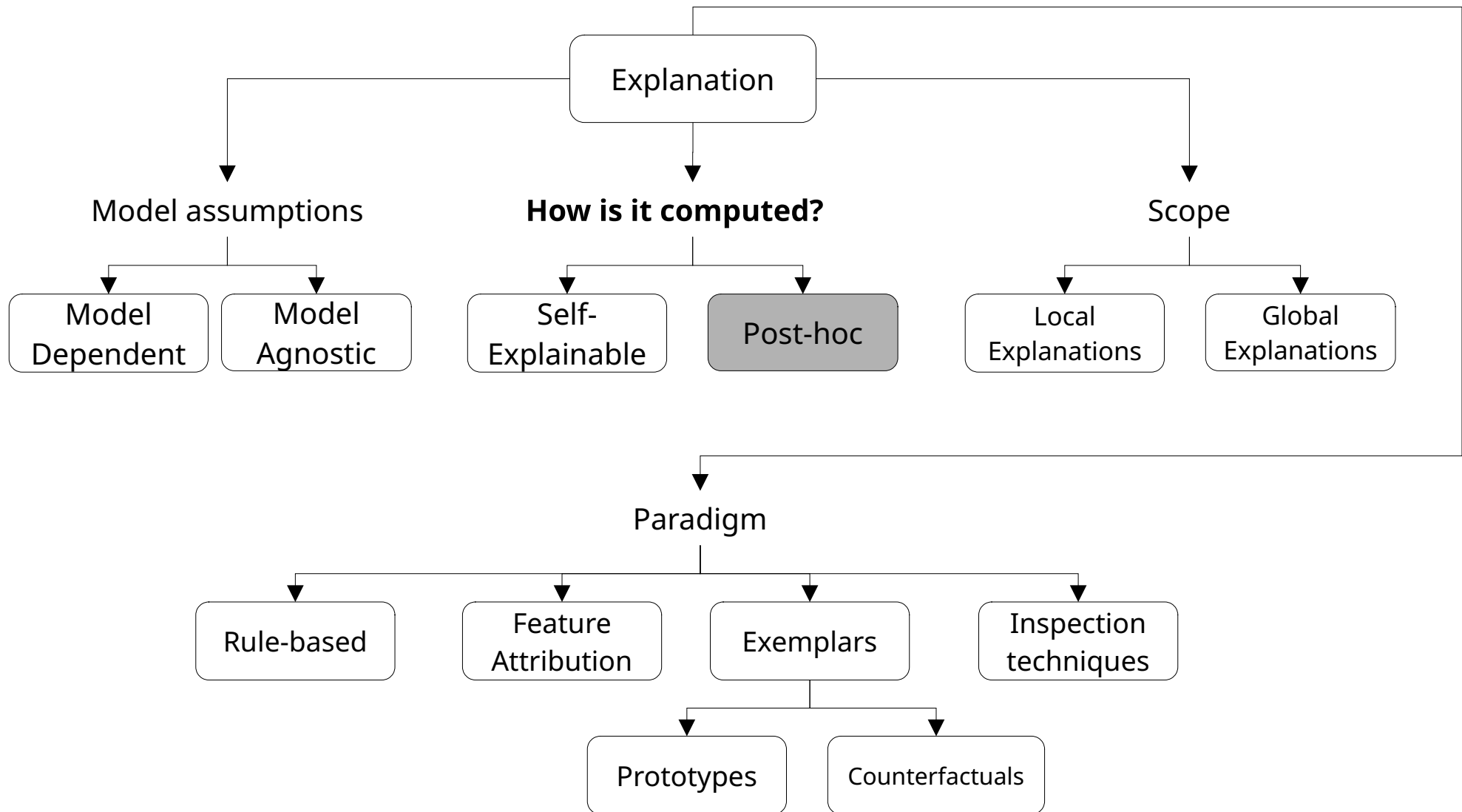
Figure 13: (top) The dictionary of shapelets, together with the thresholds d_{th} . (bottom) The decision tree for the 3-class projectile points problem



V. Guyomard et al. VCNet: A Self-explaining Model for Realistic Counterfactual Generation. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, <https://is.gd/FEkx0f>, 2022.

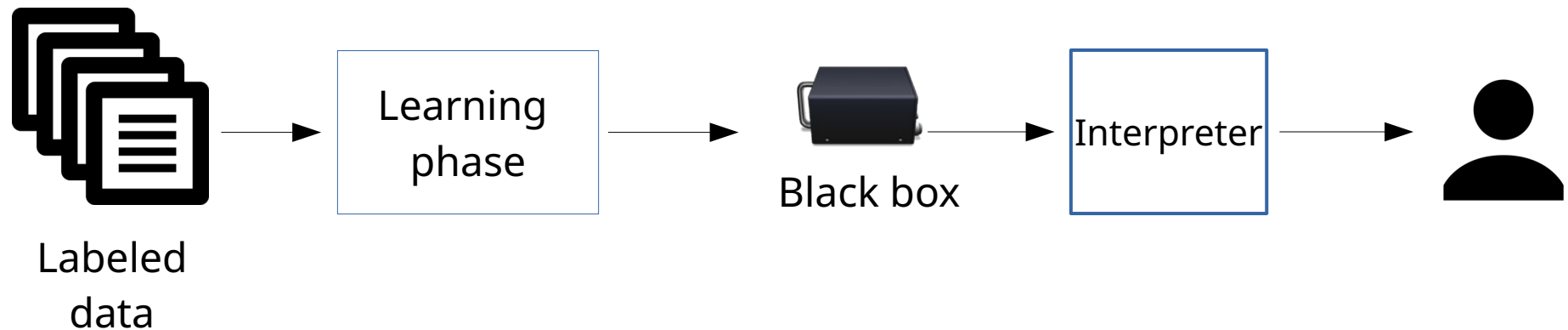
Y. Wang et al. Adversarial Regularization for Explainable-by-Design Time Series Classification. International Conference on Tools with Artificial Intelligence (ITCAI), 2020.

Taxonomy of XAI Techniques

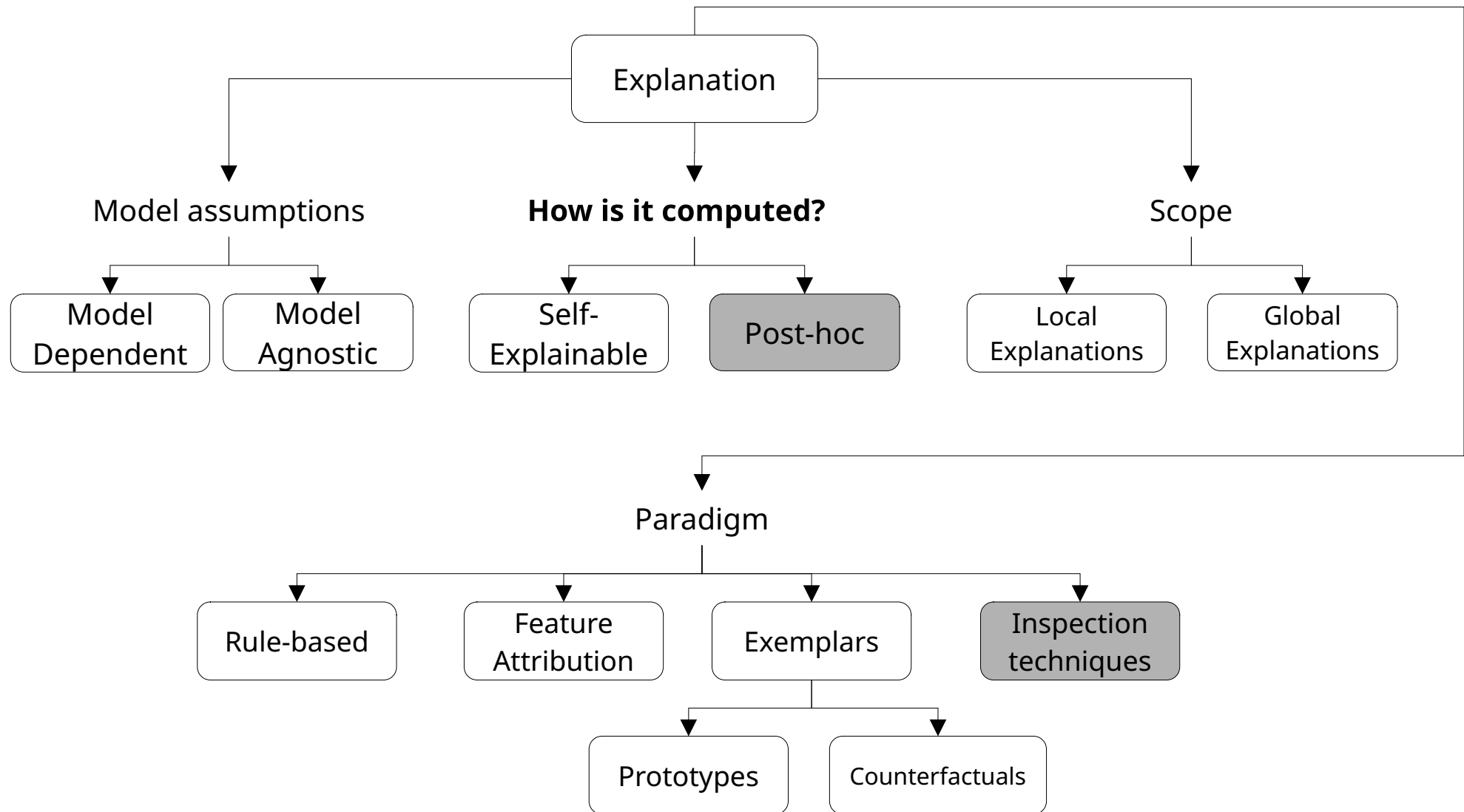


Post-hoc Explainability

Design an interpretation layer between the model and the human user



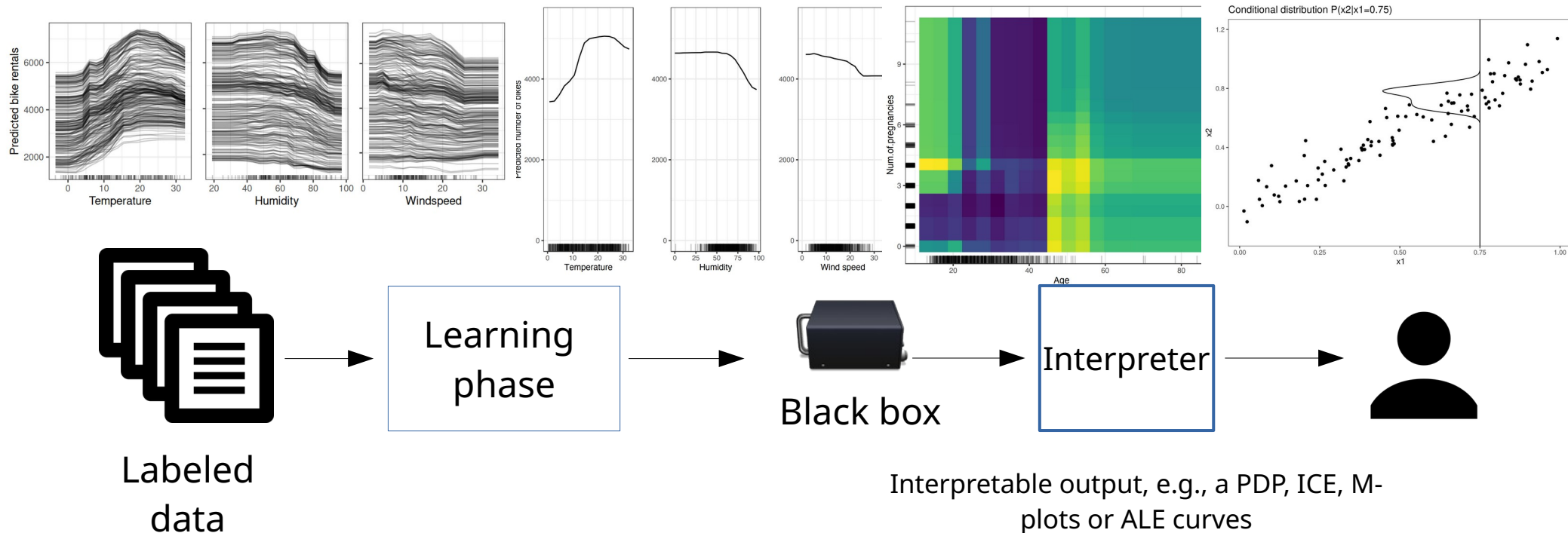
Taxonomy of XAI Techniques



Taken from: J. Delaunay. Explainability for Machine Learning Models: From Data Adaptability to User Perception. PhD Thesis, Université de Rennes, 2023

Post-hoc Explainability

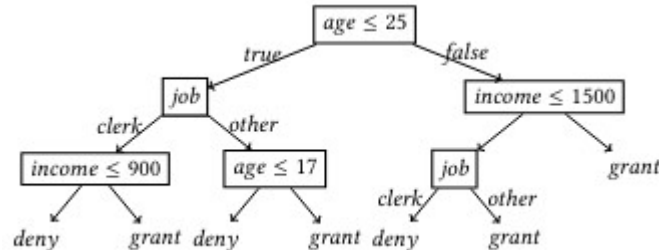
We can also **plot** or **inspect** correlations between the input features and the output



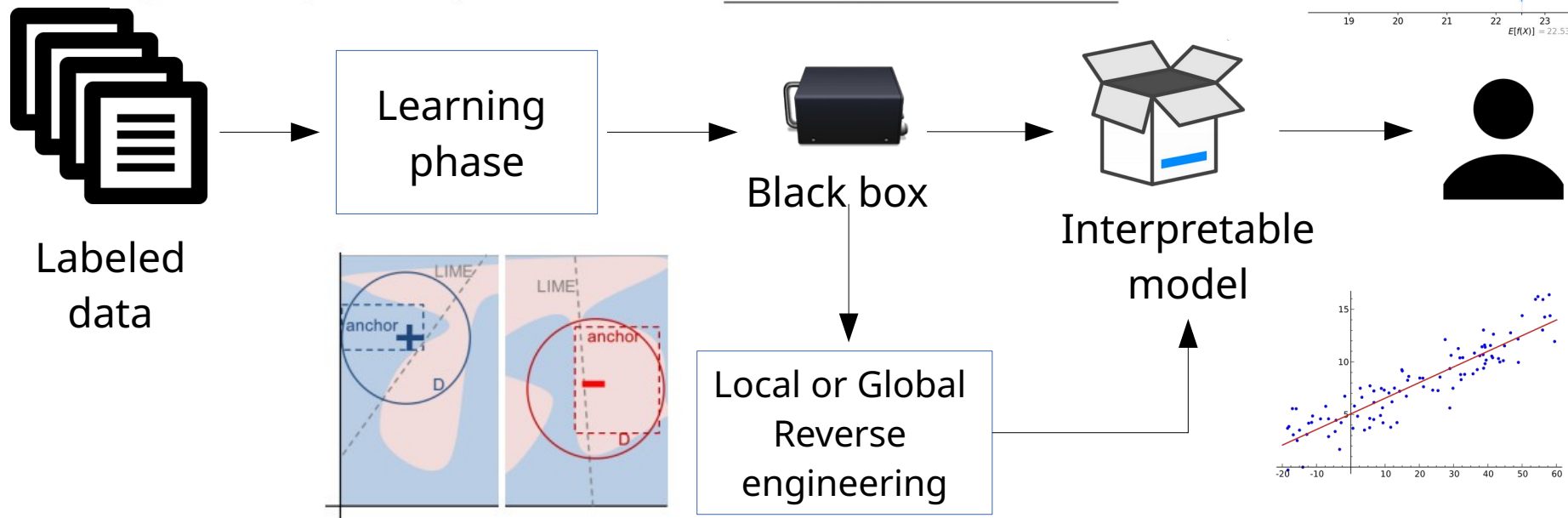
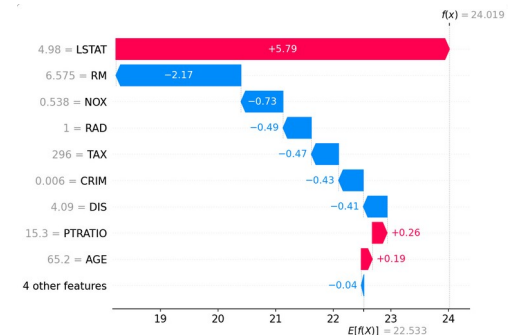
Post-hoc Explainability

Surrogate-based Approaches

The surrogate model explains the black box in general or in the vicinity of an individual instance

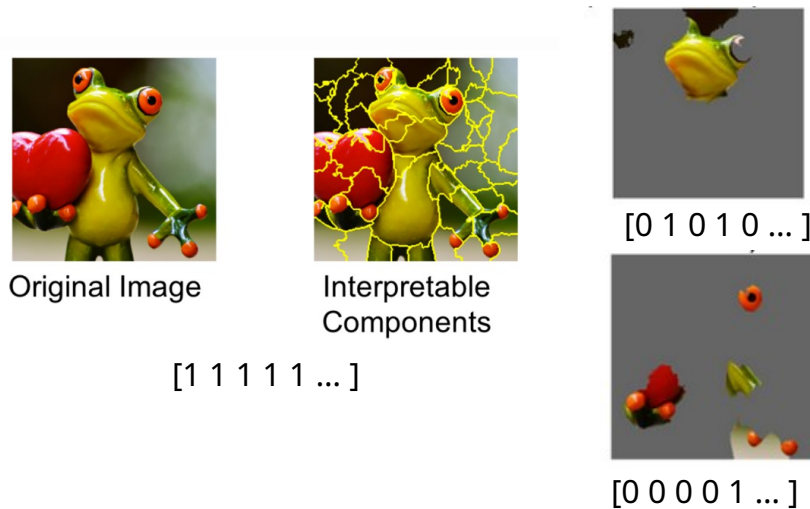


	If	Predict
adult	No capital gain or loss, never married	≤ 50K
	Country is US, married, work hours > 45	> 50K
rcdv	No priors, no prison violations and crime not against property	Not rearrested
	Male, black, 1 to 5 priors, not married, and crime not against property	Re-arrested
lending	FICO score ≤ 649	Bad Loan
	649 ≤ FICO score ≤ 699 and \$5,400 ≤ loan amount ≤ \$10,000	Good Loan



Interpretable Features

For explanations to make sense, their constituent elements must be understandable



Prediction probabilities

atheism	0.59
christian	0.41

atheism

christian

Posting: 0.16
Host: 0.13
NNTP: 0.10
edu: 0.05
have: 0.05

Text with highlighted words

From: johnchad@triton.unm.edu (jchadwic)

Subject: Another request for Darwin Fish

Organization: University of New Mexico,

Albuquerque

Lines: 11

NNTP-Posting-Host: triton.unm.edu

Hello Gang,

There have been some notes recently asking where to obtain the DARWIN fish.

This is the same question I have and I have not seen an answer on the

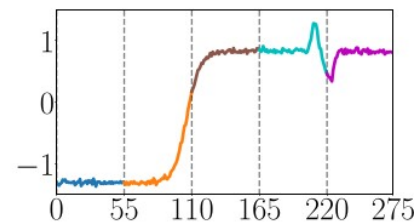
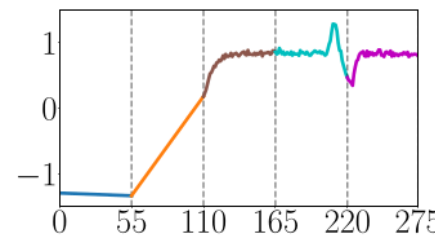
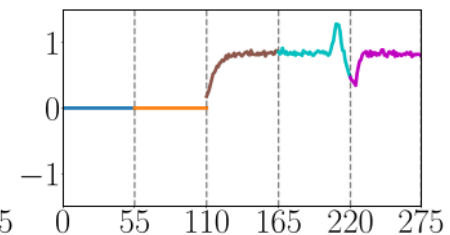


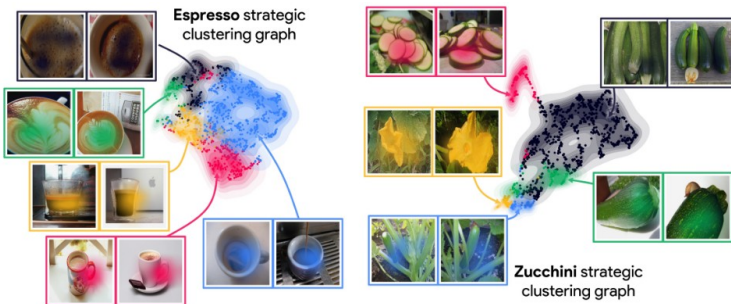
Fig. 3: Initial time series



(a) Linear interpolation



(b) Constant



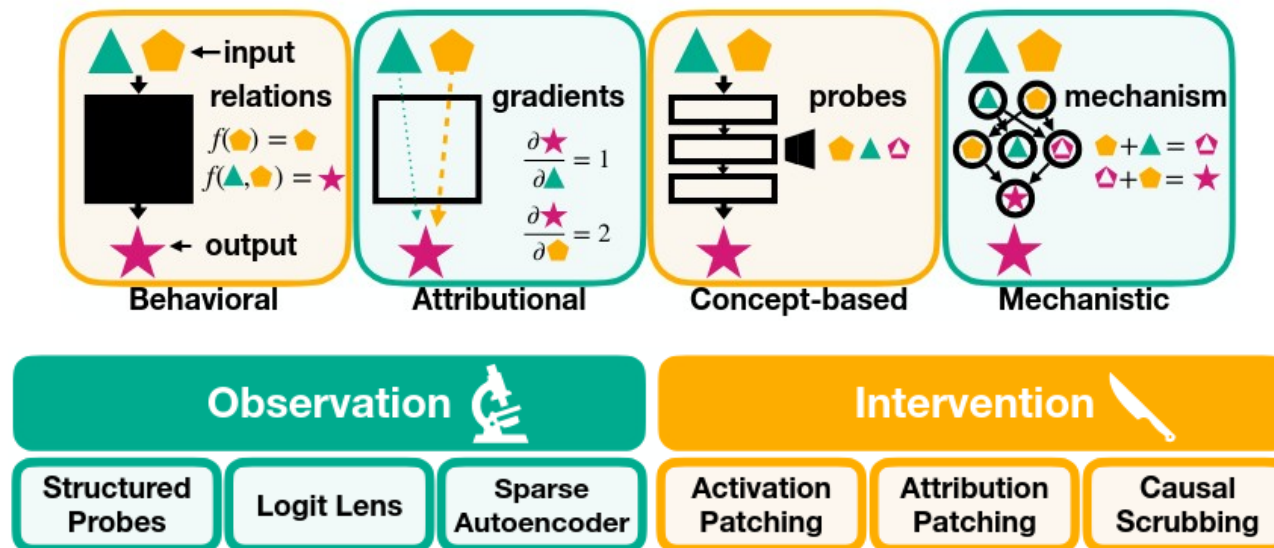
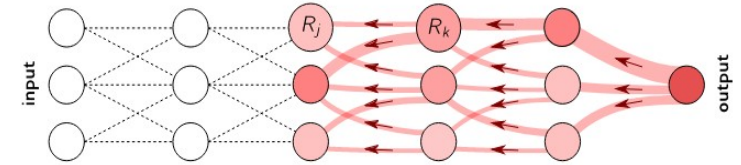
M. T. Ribeiro, S. Singh, and C. Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

T. Fel, et al. A Holistic Approach to Unifying Automatic Concept Extraction and Concept Importance Estimation. Conference on Neural Information Processing Systems (NIPS).

To touch or not to touch?

Also, agnostic vs. dependent

- Some XAI approaches only work for NNs
- They can be either observational or interventional
 - Interventions are needed to find causality



Montavon G. et al. Layer-Wise Relevance Propagation: An Overview <https://iphome.hhi.de/samek/pdf/MonXAI19.pdf>

Bereska, L.F and E. Gavves. Mechanistic Interpretability for AI Safety—A Review. Proceedings of The 1st Conference on Lifelong Learning Agents. 2022.

Is XAI really XAI?

What to reveal? What about interactivity & representation?

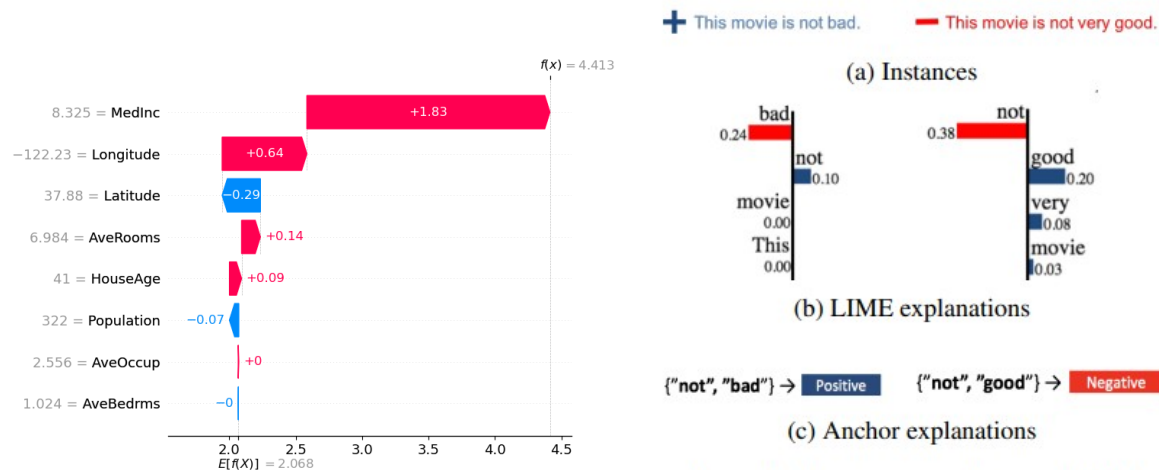
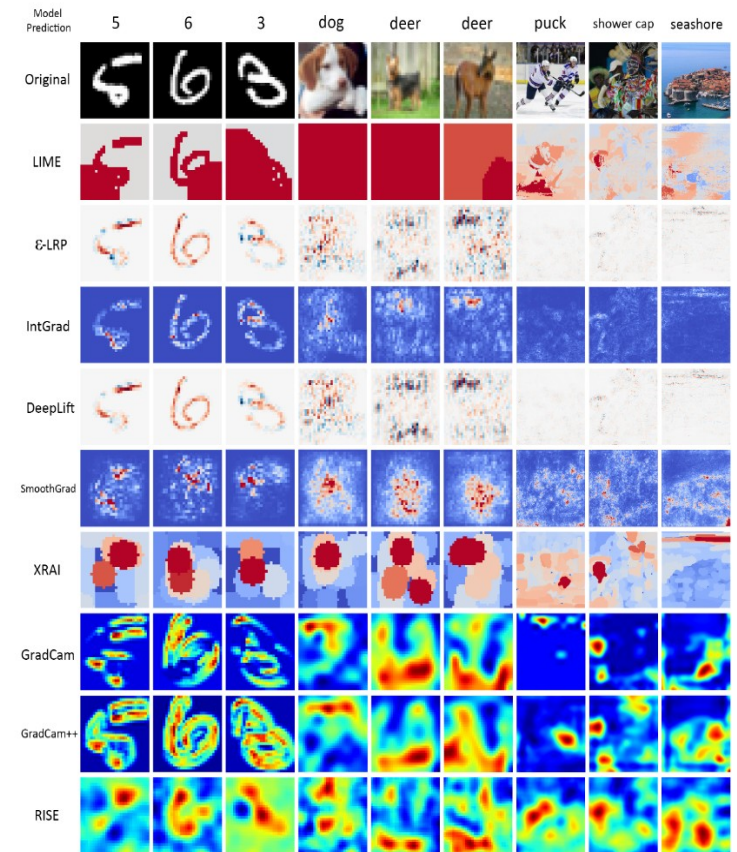
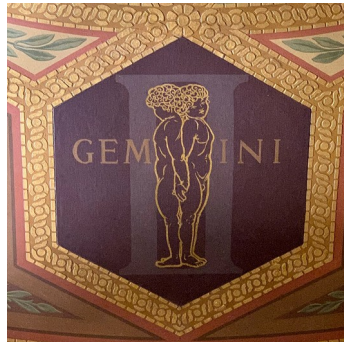


Figure 5: Counterfactuals obtained with VCNet for the MNIST dataset. The top line corresponds to the examples to explain, the bottom to the corresponding counterfactuals.



XAI approaches are rarely evaluated on users!

XAI in the HCI and HC-AI communities



Humans and Explanations

- Using XAI always comes with an **immediate goal**
 - **What if** I earned 10K more? Would my loan have been accepted?
 - Is the AI model **using a sensitive attribute** for prediction?
- What is the expected **impact** of the explanation?
 - To elicit trust? To achieve legal compliance?
- **Who** is the **recipient** of the explanation?
 - An AI engineer? Grandpa? A physician?
 - Cognitive aspects:

Understanding

Trust

Perception

confidence, **complacency**

fairness, empathy, plausibility

User-centered Evaluation

Understanding

- Usually via a “proxy” task
 - **Predict** the model’s answer for a given case
 - **Explain** the features that play a role in the prediction
 - **Validate or reject** statements about the model
 - **Use** the model (also used for measuring trust)
- And via behavioral and self-reported metrics
 - Accuracy, completion time [more objective, tedious to collect]
 - Specialized questionnaires [easier to collect but subjective]

User-centered Evaluation

Trust and Confidence

- Via questionnaires
- Adherence to the AI's recommendation
- Trust is a complex construct
 - Questionnaires test some related constructs
 - They are a proxies to trust

1.

What is your confidence in the tool? Do you have a feeling of trust in it?						
I do not trust it at all.	2	3	4	5	6	I trust it completely.

2.

Are the actions of the tool predictable?						
It is not at all predictable.	2	3	4	5	6	It is completely predictable.

3.

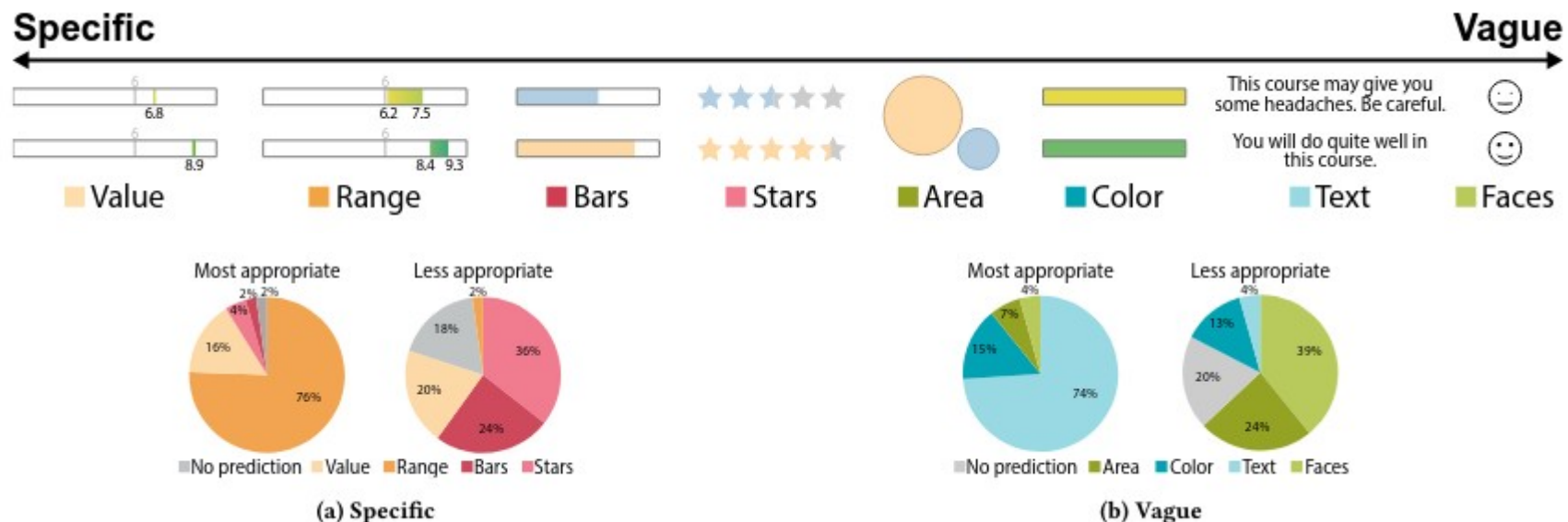
Is the tool reliable?						
It is not at all reliable.	2	3	4	5	6	It is completely reliable.

4.

Is the tool efficient at what it does?						
It is not at all efficient.	2	3	4	5	6	It is completely efficient.

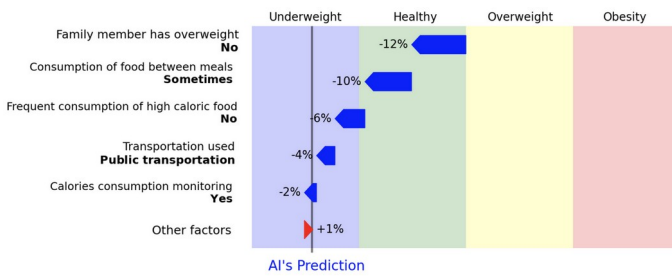
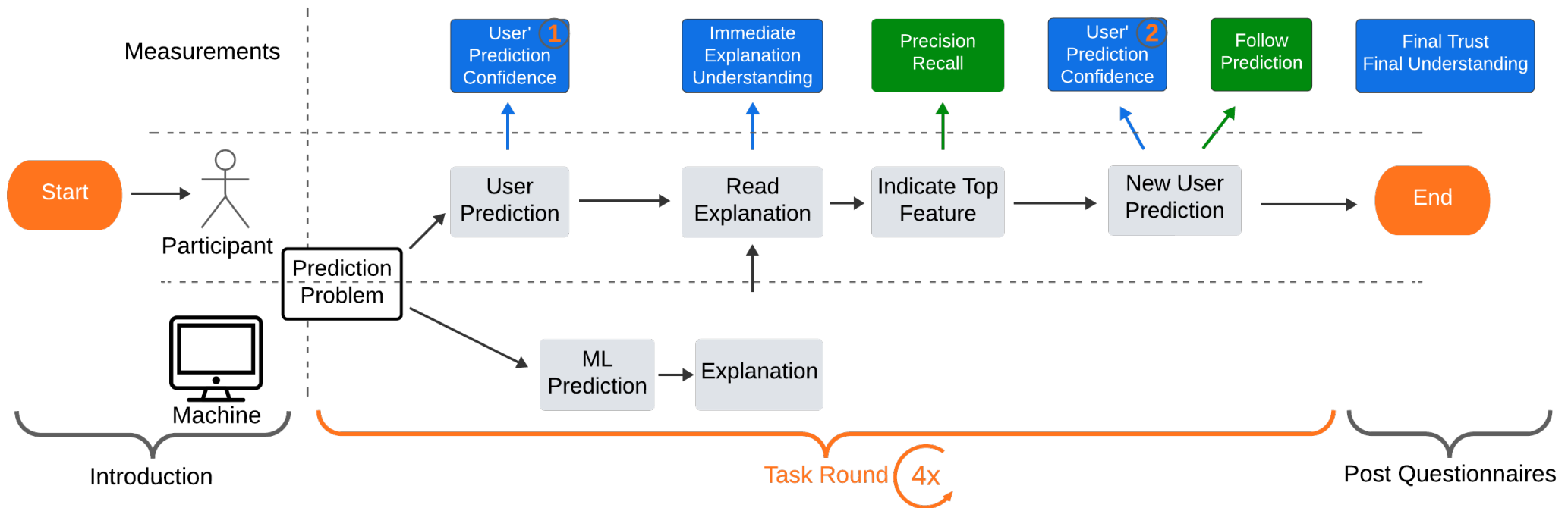
Explanations & Cognition

- **Explanation type** and **visual representation** impact cognition
 - How the prediction is presented matters as much
 - When the explanation is dense, interactivity can help!



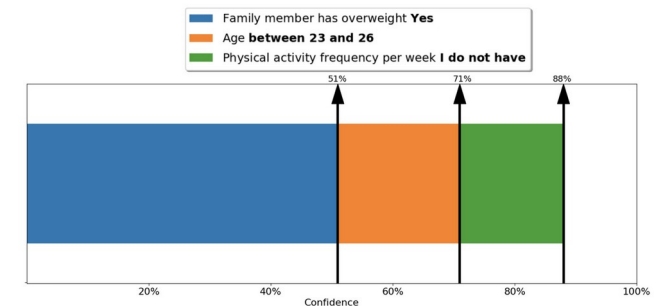
Comparing Explanations

A protocol



- First, since **no** family member of this individual **suffers** from overweight, the score **decreases** by 12%.
- Second, since the individual **sometimes** consumes food between meals, the score **decreases** by 10%.
- Third, **no consuming frequently** high caloric food **decreases** score by 6%.
- Fourth, using **public transport** **decreases** the score by 4%.
- Fifth, **monitoring** her calories consumption **decreases** the score by 2%.

Combining all the **other answers** **increases** the score by 1% and the final value is 17% implying an **underweight** prediction.

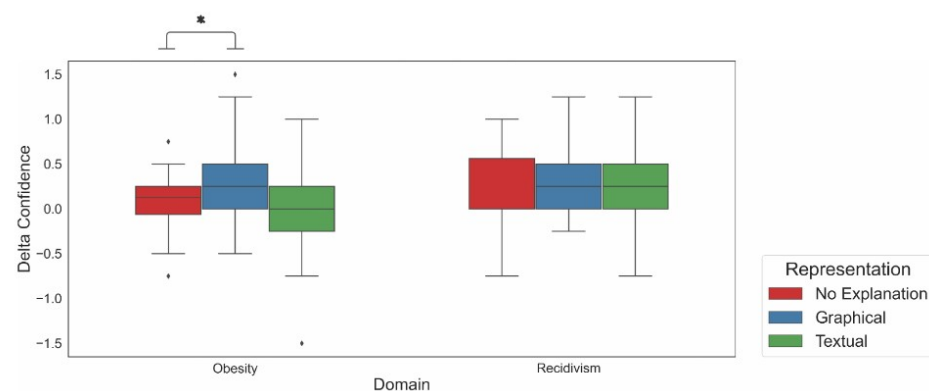
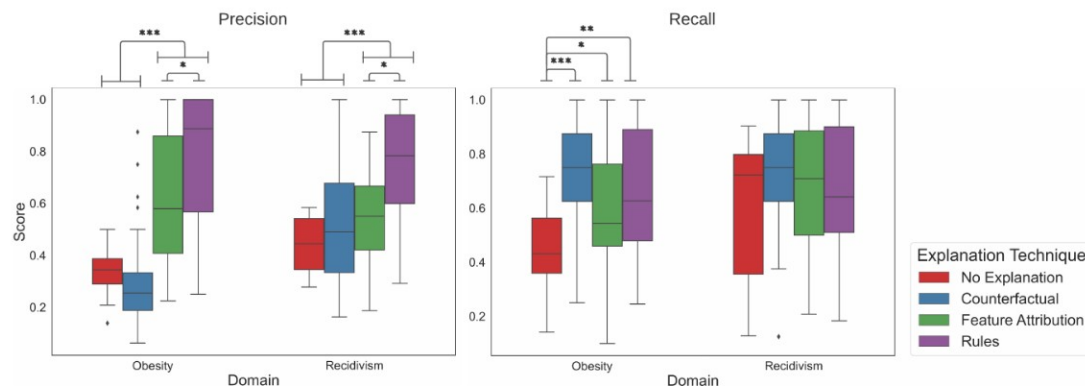


Comparing Explanations

Some Results

Technique	Understanding								confidence			
	Recidivism				Obesity				Recidivism		Obesity	
	Self-Reported		Behavioural		Self-Reported		Behavioural		Self-Reported	Behavioural	Self-Reported	Behavioural
	Immediate	Final	Precision	Recall	Immediate	Final	Precision	Recall	Δ Confidence	Follow Prediction†	Δ Confidence	Follow Prediction†
Representation	0.87	1.20	16.24***	1.58	3.75*	1.35	31.42***	6.37***	1.40	0.78	0.12	0.38
Age	0.96	0.36	0.13	3.00	0.14	0.55	0.05	2.85	0.04	0.00	8.22**	0.12
Education	1.07	0.01	1.88	0.10	0.16	0.06	6.41*	0.02	0.46	2.76	0.06	0.00
Gender	1.63	0.93	0.94	0.43	0.50	0.34	0.25	1.31	0.13	0.34	2.14	0.63
Technique:Representation	0.54	1.07	0.35	0.30	0.14	0.03	0.18	0.36	2.16	0.31	0.12	1.11
	0.28	0.87	1.12	0.74	0.48	0.16	0.35	4.99**	0.35	0.75	0.26	3.55*

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$



Explainability is not universal

- Knowing who receives the explanation is crucial
 - Users & communities have different backgrounds, knowledge gaps, preconceptions, etc.



Lay users



AI practitioners



Professionals

- Preconceptions can make acceptability challenging
 - That is why domain experts are harder to convince
- Behavioral and reported understanding may not correlate

**Where are we heading
towards?**

LLMs are really cool for XAI

Interactive dialogs

User "utterances" & TalkToModel	response
"What are the predictions on patients older than 30? Could you also tell me what would happen to these predictions if we increased bmi by 10?"	<p>On data with age greater than 30.0, the model predicts:</p> <ul style="list-style-type: none">⇒ No Diabetes on 46.8% of data⇒ Diabetes on 53.2% of data <p>On data with age greater than 30.0 and bmi is increased by 10.0, the model predicts:</p> <ul style="list-style-type: none">⇒ No Diabetes on 30.1% of data⇒ Diabetes on 69.9% of data
"So how important is bmi for the model's predictions on the over thirty group?"	<p>For the model's predictions on data with age greater than 30.0, the bmi feature is ranked on average 2.86 (95% CI [2.685, 3.035]). Here, rank 1 is the most important feature (out of 8 features). Overall, bmi is highly important for the model's predictions.</p>
"Ah, so which two features are mostly more important than bmi for these people?"	<p>For the models predictions on data where age is greater than 30.0, the top 2 most important features are:</p> <ol style="list-style-type: none">1. glucose2. pedigree_function

LLMs are really cool for XAI

A certain loan applicant has the following features:

Feature	Value	Feature	Value
Account Status	< 0 DM	Property	real estate
Duration	18 months	Age	27 years
Duration in Current Address	< 1 year	Other Loans	bank
Purpose	education	Housing	rent
Credit Amount	750 DM	Existing Credits	1
Savings	< 100 DM	Occupation	unemployed
Employment History	unemployed	Dependants	3 or more dependants
Installment Rate	4%	Telephone	none
Sex and Marital Status	male : married	Foreign Worker	yes
Other Debtors	none	Credit History	existing credits paid back duly till now

	SHAPstories: general audience	SHAPstories: data science audience own usage	SHAPstories: data science audience: general audience usage
Convincingness	93.2% ***	77.8% ***	
Ease	92.4% ***	61.1% *	91.7% ***
Confidence	79.7% ***	38.9%	91.7% ***
Speed	87.5% ***	44.4%	88.9% ***
Likelihood to use	92.2 % ***	61.1% *	83.3% ***

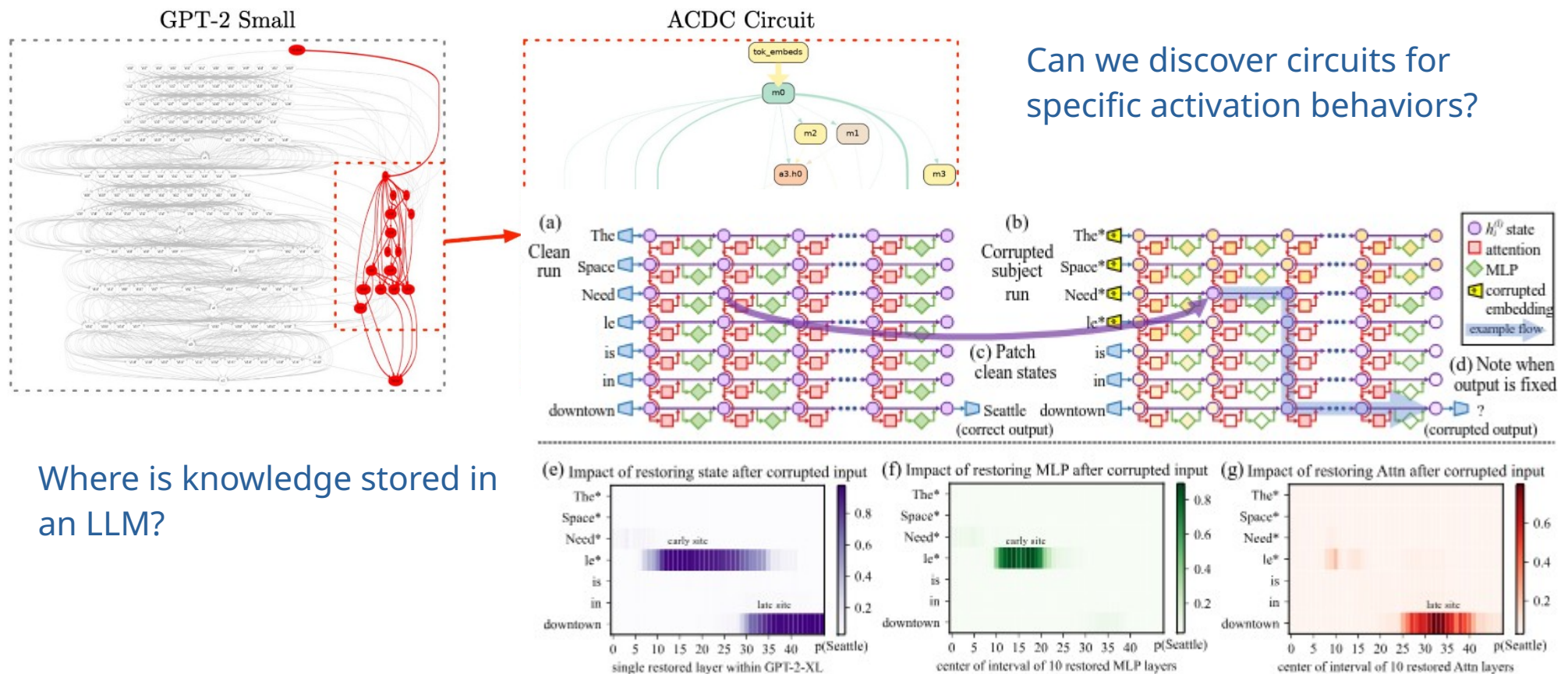
Note: For all results we test if the obtained result is statistically significant: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Our individual is a young, unemployed person with a significant savings amount less than 100 DM (Deutsche Mark). They reside in their own house and have lived at their current address for less than a year. Their current account status indicates a balance less than 0 DM which appears as the largest negative contributor to the prediction. Furthermore, they recently took a loan amounting to 750 DM, with the purpose of education, from a bank where they have a credit history of existing credits paid off duly till now. However, the negative SHAP value for credit history suggests that the AI notes some risk associated with this factor. Employment history also weighs in heavily against them, as they are unemployed. [...]

But LLMs are black boxes

(that explain other black boxes)

XAI on LLMs often learns neural activation patterns

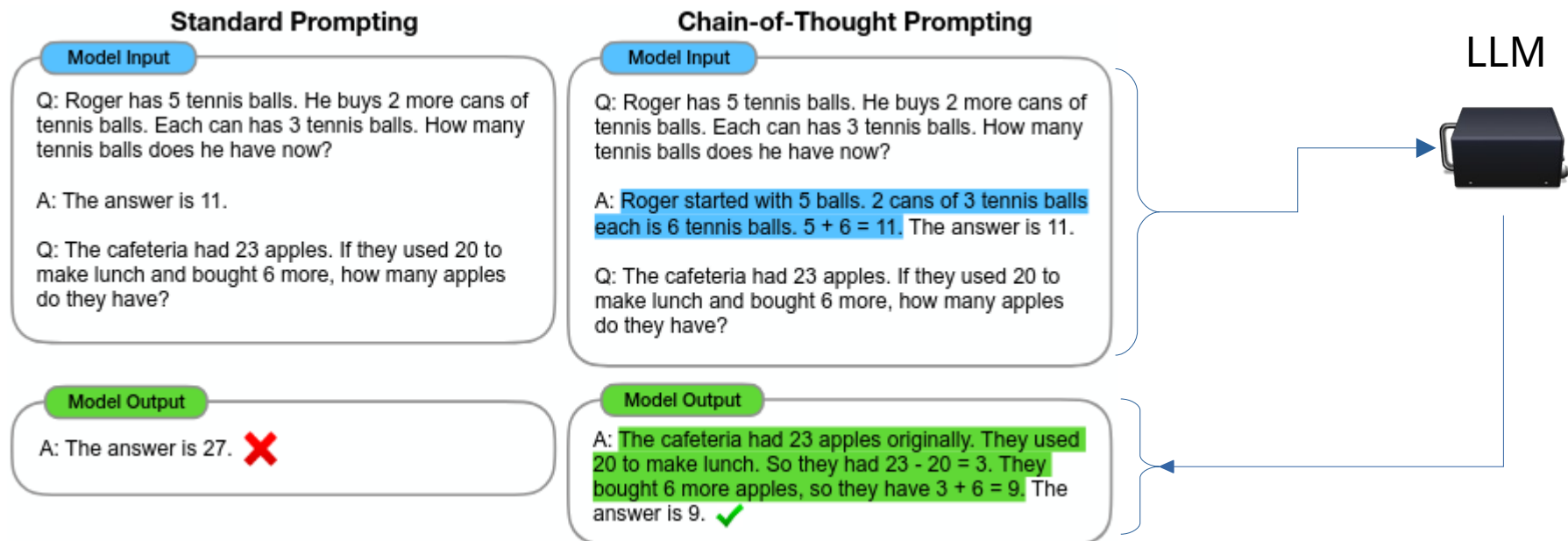


A. Conmy, et al. Towards Automated Circuit Discovery for Mechanistic Interpretability. Conference on Advances in Neural Information Processing Systems (NIPS). 2023.

K. Meng, et al. Locating and Editing Factual Associations in GPT. Conference on Advances in Neural Information Processing Systems (NIPS). 2022.

Explicit problem solving = XAI?

- From a **classical** perspective, **no!**
 - But it provides hints of the tokens and language structures that correlate with the answer.



LLMs and XAI

Are we done? Nope!

- Also, LLMs have serious issues with common sense
 - They are **mere translators** into natural language
- Could AI decide automatically how to explain itself?
 - Choose paradigm, representation, features, rhetorical style based on the user?

Explaining why an AI model predicted that a student would graduate

Imagine the student's **success in graduating as a heavy ball rolling down a hill**. The slope of the hill is determined by factors like the student's academic performance, family background, and socio-economic status.

In this case, the student had a lower-than-average grade in the first semester, and didn't complete as many required units as others. **This made the hill steeper and more challenging for the ball to roll down.** Meanwhile, factors like the student's age and gender also played small roles in making the hill slightly steeper.

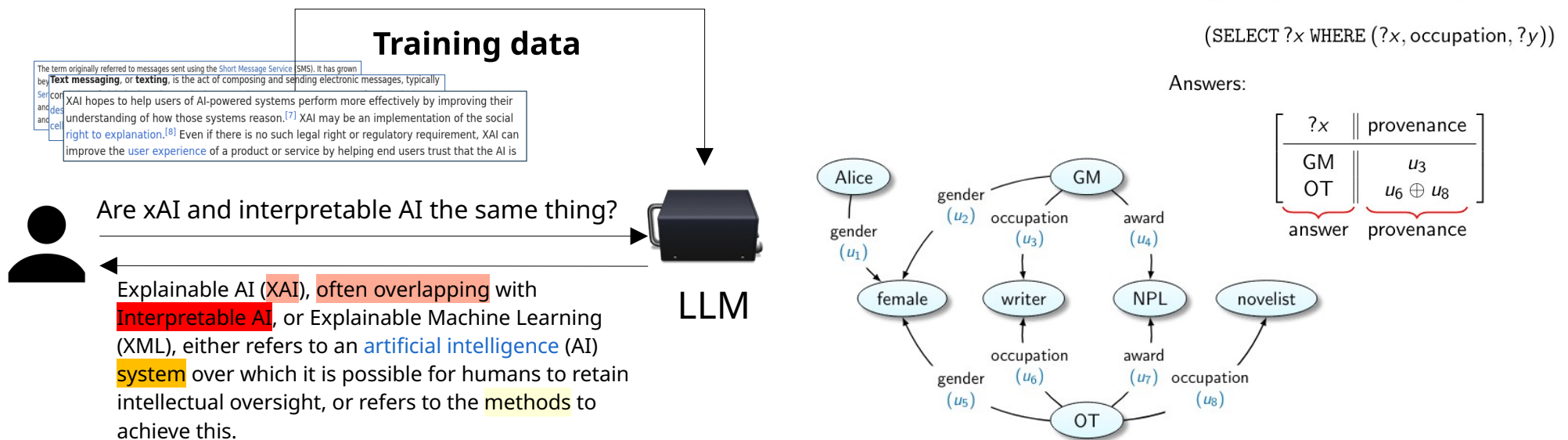
However, the student didn't receive a scholarship, which can act like a gentle wind pushing the ball down the hill. Their parents' education levels also provided some support, acting like small inclines in the hill.

Taking all these factors into account, the model predicted the student would still manage to graduate, although the journey might be a bit tougher.

Source attribution

(where did you get inspiration from?)

- Can we link AI answers to specific training examples?
 - **Starting point:** neural provenance models



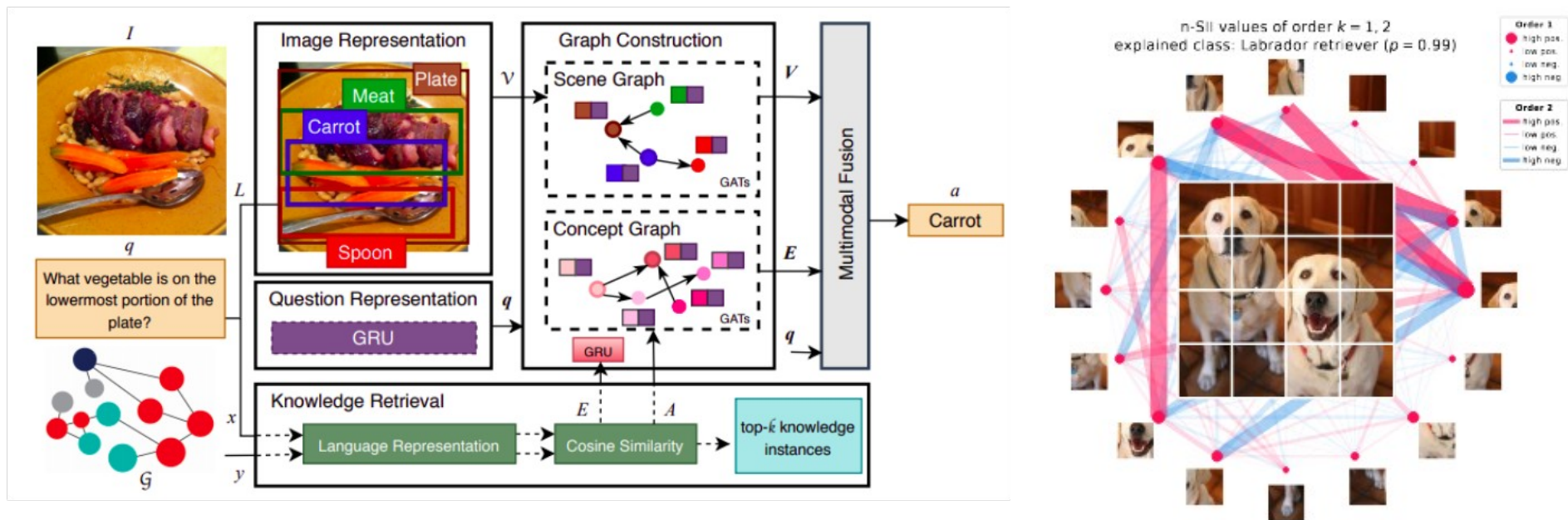
- Done for simple models, great for tracing authorship

Yinjun Wu, Val Tannen, and Susan B. Davidson. PrIU: A Provenance-Based Approach for Incrementally Updating Regression Models. ACM SIGMOD International Conference on Management of Data (SIGMOD). <https://doi.org/10.1145/3318464.3380571>

Zhao, D., Wan, G., Zhan, Y., Wang, Z., Ding, L., Zheng, Z., & Du, B. (2023). Ke-x: Towards subgraph explanations of knowledge graph embedding based on knowledge information gain. Knowledge-Based Systems, 110772. <https://doi.org/https://doi.org/10.1016/j.knosys.2023.110772>

Multimodal Explanations

- How could multimodal explanations look like?
 - **Starting point:** feature coalition attributions
 - **Challenges:** combinatorial explosion, representation



Ziaeeafard, Maryam, and Freddy Lecue. Towards knowledge-augmented visual question answering. Proceedings of the 28th International Conference on Computational Linguistics, 2020.

Kolpaczki, P. SVARM-IQ: Efficient Approximation of Any-order Shapley Interactions through Stratification. Proceedings of Machine Learning Research, vol 238, pages 3520-3528 <https://proceedings.mlr.press/v238/kolpaczki24a.html>

Take-away message

- XAI is being studied by two heterogeneous crowds
 - One obsessed with fidelity, the other with users
 - They should talk more to each other!
- If you do XAI, please evaluate it with potential users!
 - Beware, domain experts are hard to convince
- Post-hoc XAI does not always point out causality
- LLMs are great news for XAI
 - But they are just translators that lack common sense
- As AI becomes multimodal, so should explanations