

Explainability in Supervised Machine Learning

From Faithful to Human-Friendly Explanations

Luis Galárraga
BDA MDD'24 Summer School
25/06/2024



Agenda

- eXplainable AI/ML: What and Why?
- Glass- vs. black-box models
- eXplainable AI techniques
- Open challenges and conclusion

Agenda

- eXplainable AI/ML: What and Why?
- Glass- vs. black-box models
- eXplainable AI techniques
- Open challenges and conclusion

Agenda

- eXplainable **AI/ML**: What and Why?
- Glass- vs. black-box models
- eXplainable AI techniques
- Open challenges and conclusion

Artificial Intelligence (AI)

- Intelligent traits implemented in algorithms
 - No consensus about the definition of *intelligence*
- Intelligence has been studied by psychologists, neurologists, and computer scientists

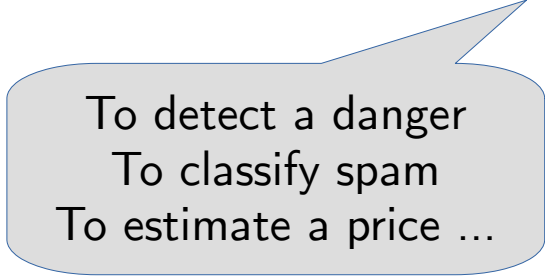
“The ability of an agent to achieve **a goal** in a **wide range of environments**”

S. Legg and M. Hutter. A formal Measure of Machine Intelligence. In Proc. of the 15th Annual Machine Learning Conference of Belgium and The Netherlands, pages 73–80, Ghent, 2006.

Artificial Intelligence (AI)

- Intelligent traits implemented in algorithms
 - No consensus about the definition of *intelligence*
- Intelligence has been studied by psychologists, neurologists, and computer scientists

“The ability of an agent to achieve **a goal** in a **wide range of environments**”



To detect a danger
To classify spam
To estimate a price ...

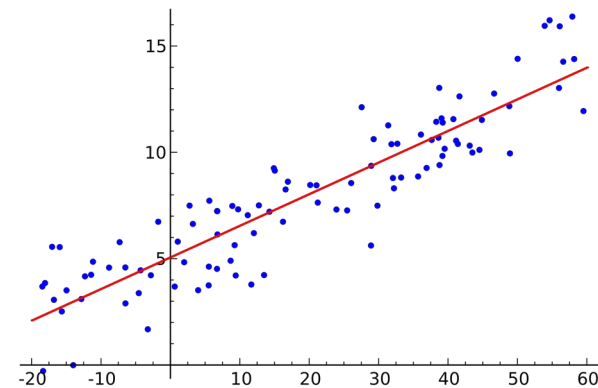


potentially changing

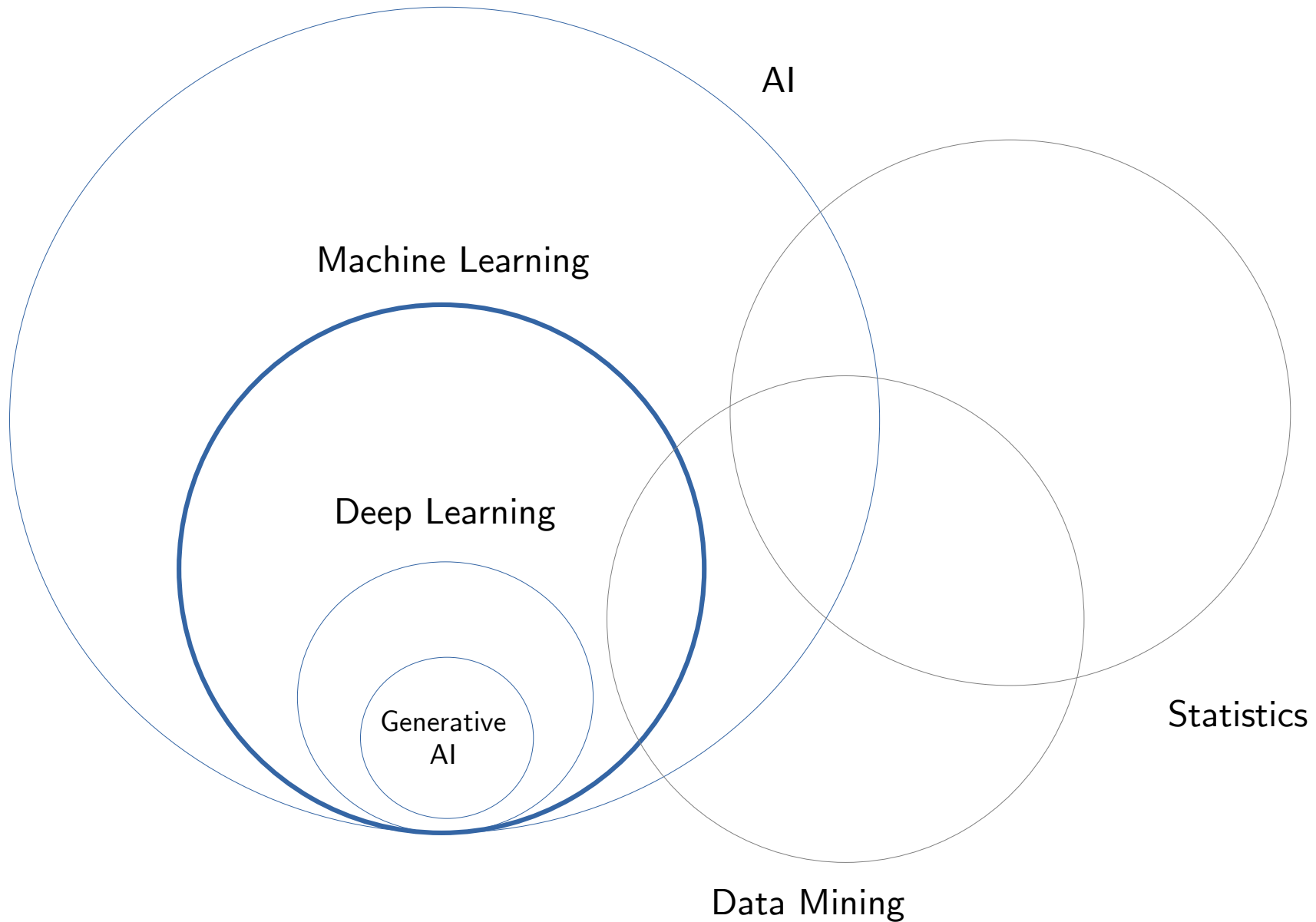
S. Legg and M. Hutter. A formal Measure of Machine Intelligence. In Proc. of the 15th Annual Machine Learning Conference of Belgium and The Netherlands, pages 73–80, Ghent, 2006.

Machine Learning (ML)

- Sub-domain of AI that studies the methods to **generalize** from data, e.g.,
 - To predict the risk of default based on the profile of new credit applicants
 - To detect objects in images and videos
 - To recommend movies to users
 - To generate text and images
- Computers **learn models** from data

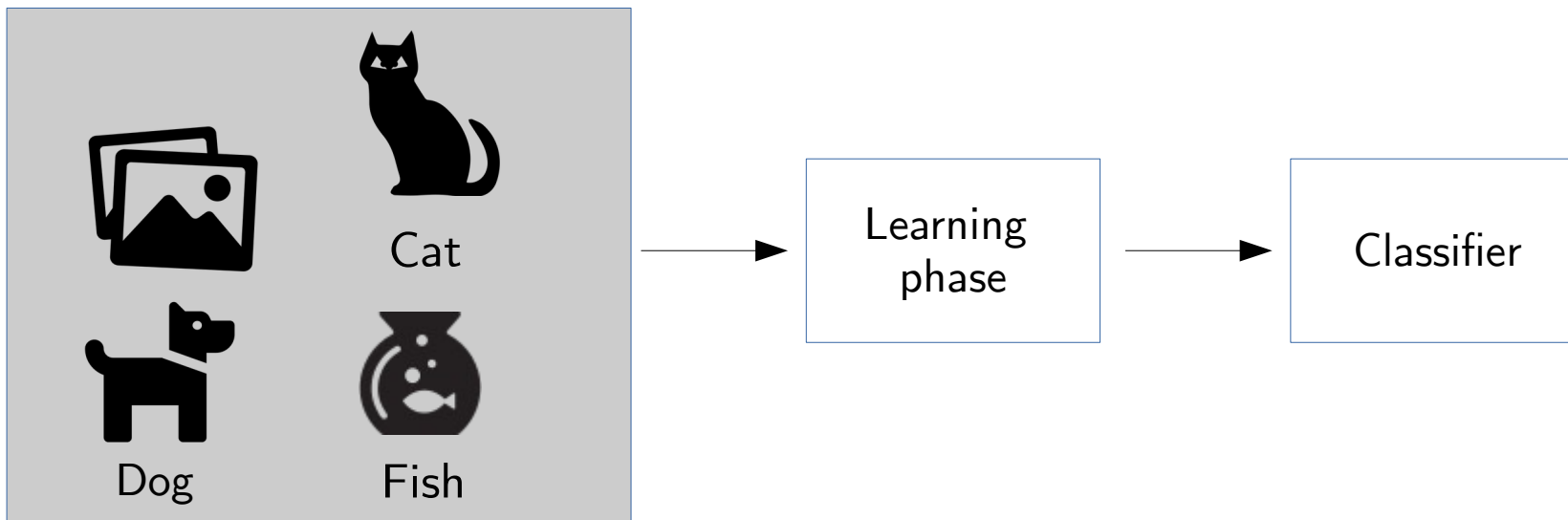


Machine Learning (ML)



Supervised Machine Learning

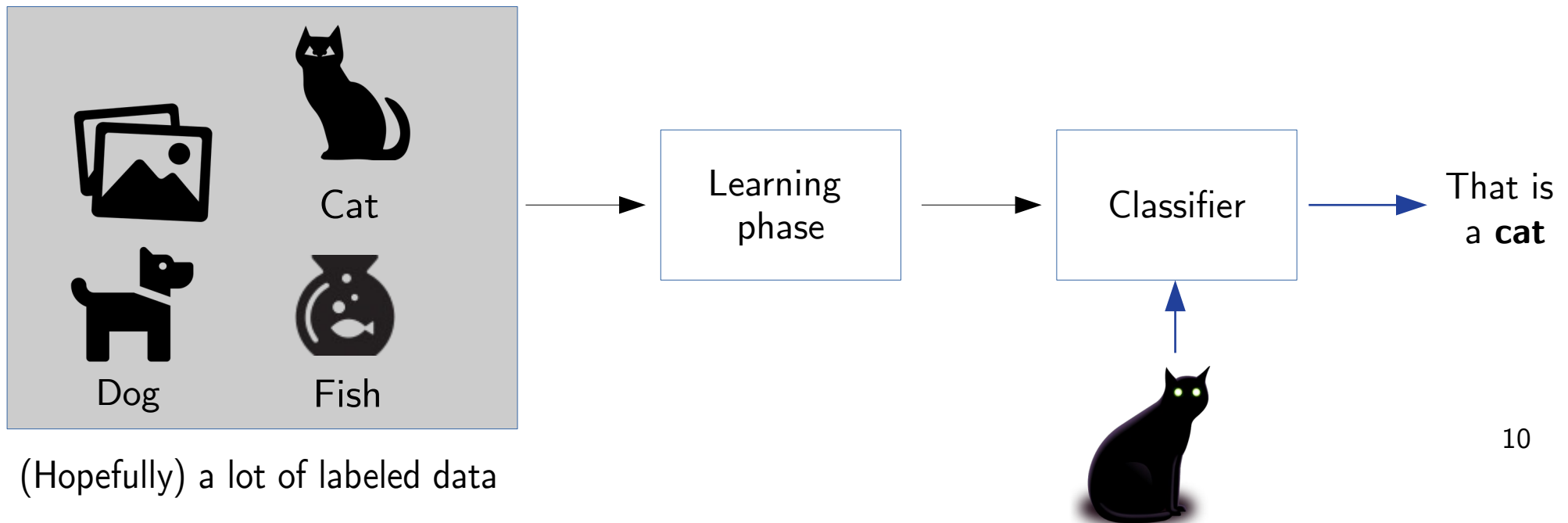
- Models trained on annotated data that can predict “labels” for new instances
 - If the labels are classes → classification
 - If the labels are quantities → regression



(Hopefully) a lot of labeled data

Supervised Machine Learning

- Models trained on annotated data that can predict “labels” for new instances
 - If the labels are classes → classification
 - If the labels are quantities → regression

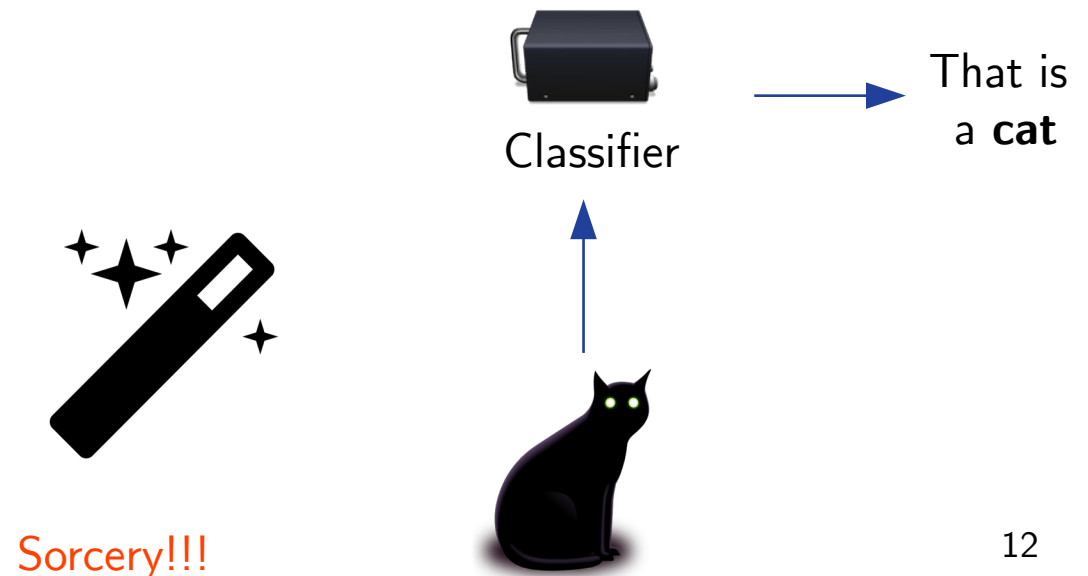


Agenda

- eXplainable AI/ML: What and Why?
- Glass- vs. black-box models
- eXplainable AI techniques
- Open challenges and conclusion

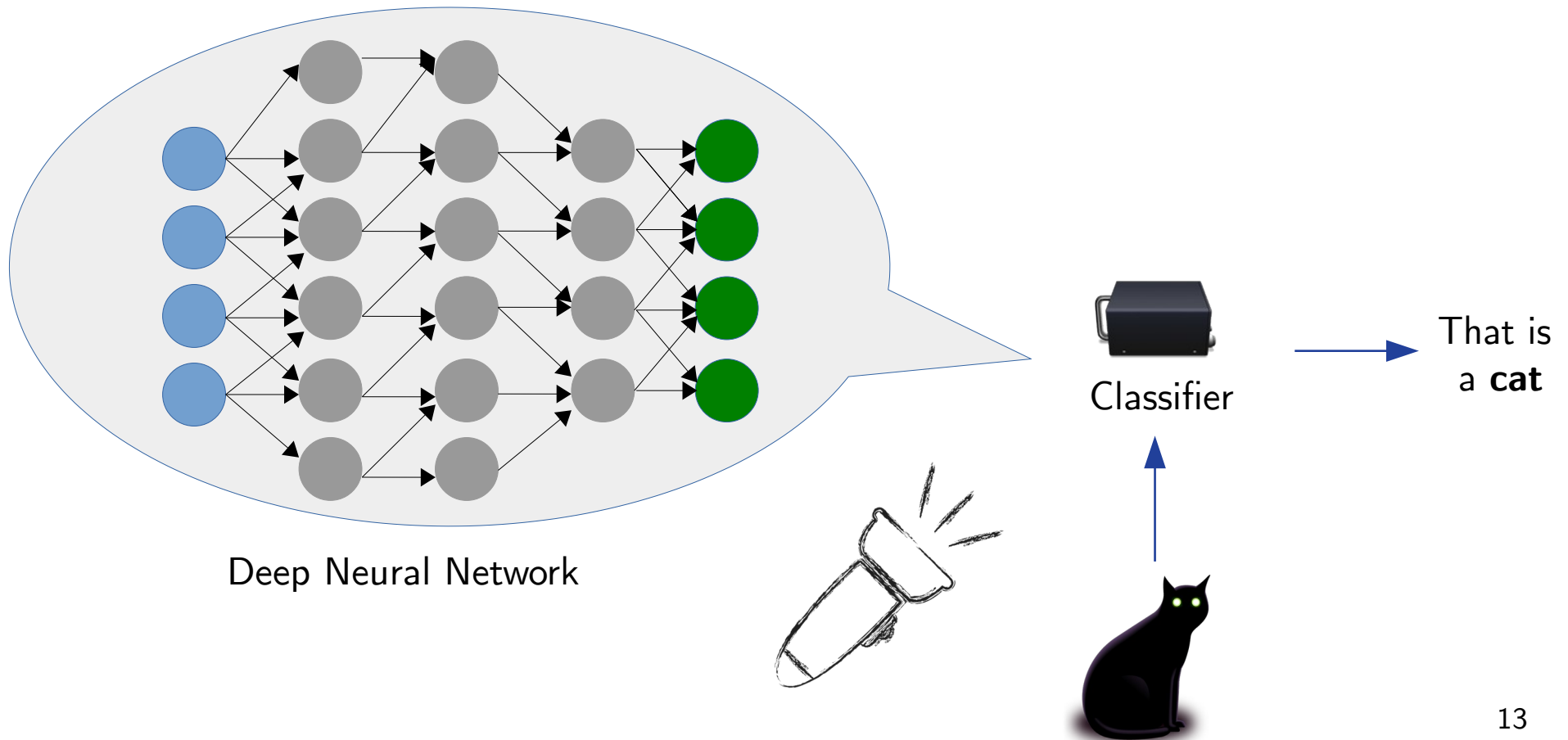
ML models resemble sorcery

The complexity of some ML models makes them black boxes



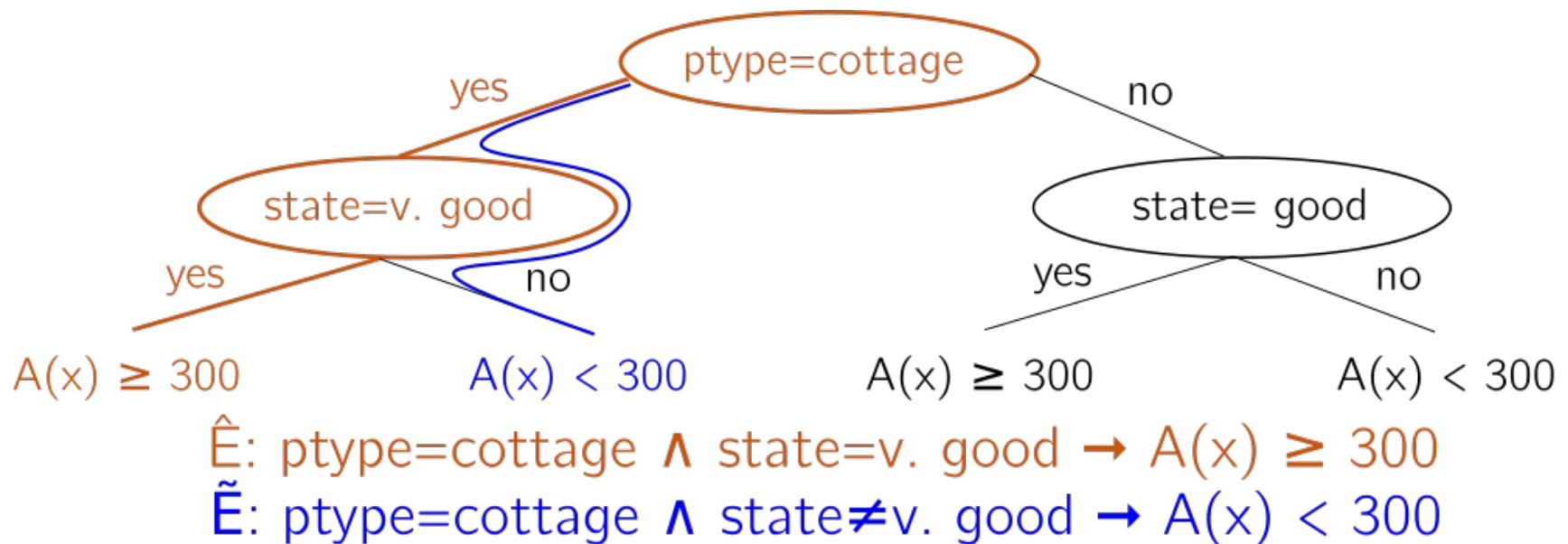
ML models resemble sorcery

The complexity of some ML models makes them black boxes



Interpretable AI: What?

A model is *interpretable* if the rationale behind its answers can be understood by humans

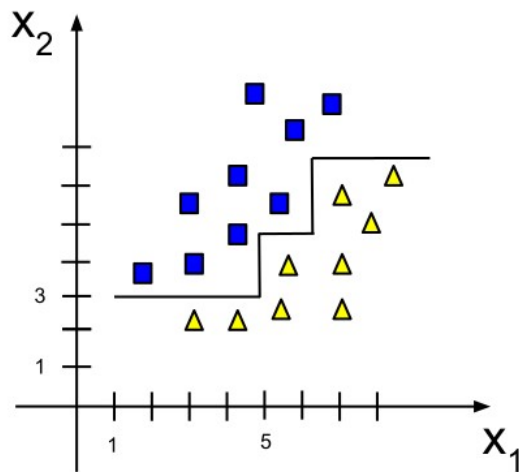


A lot of related terms!

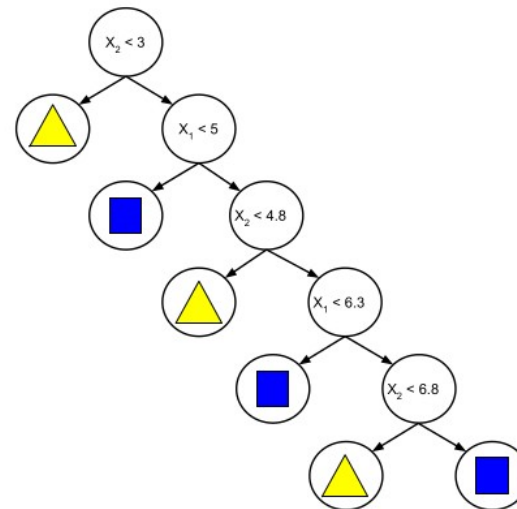


Interpretable AI: What?

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*”



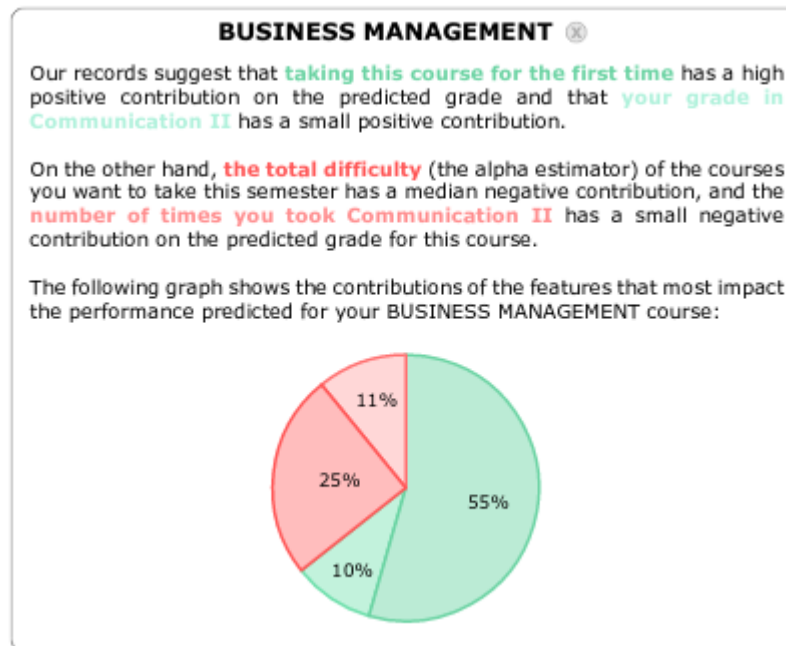
(a) Complex decision tree boundary.



(b) Decision tree corresponding to the boundary.

eXplainable AI: What?

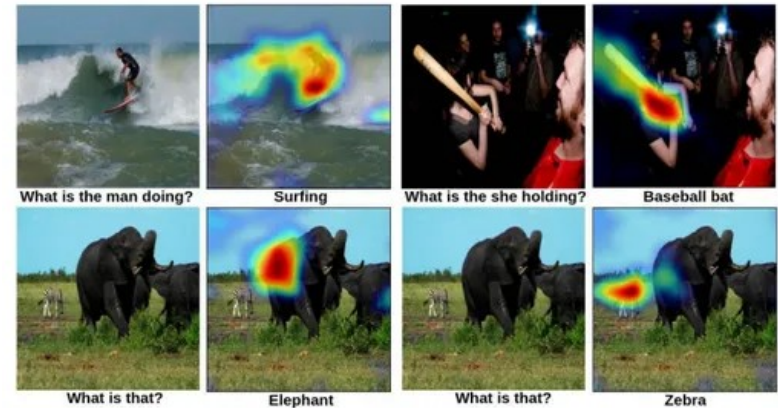
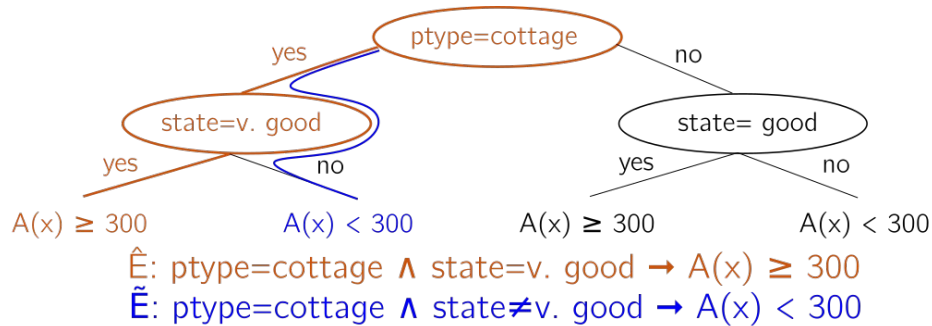
Explainability: “The *explainability* of a model refers to its capacity to be explained by external tools or techniques”



	If	Predict
adult	No capital gain or loss, never married	$\leq 50K$
	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
	$649 \leq \text{FICO score} \leq 699$ and $\$5,400 \leq \text{loan amount} \leq \$10,000$	Good Loan

What is an explanation?

A statement that characterizes the relationships between the inputs and outputs of an AI model



Prediction probabilities

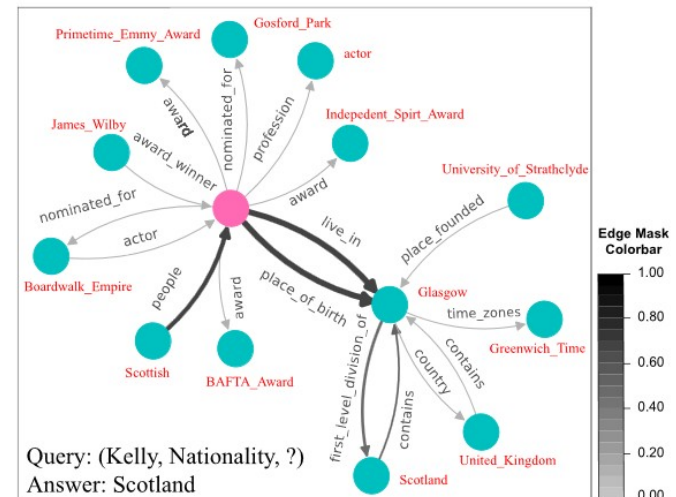
atheism	0.53
christian	0.47

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

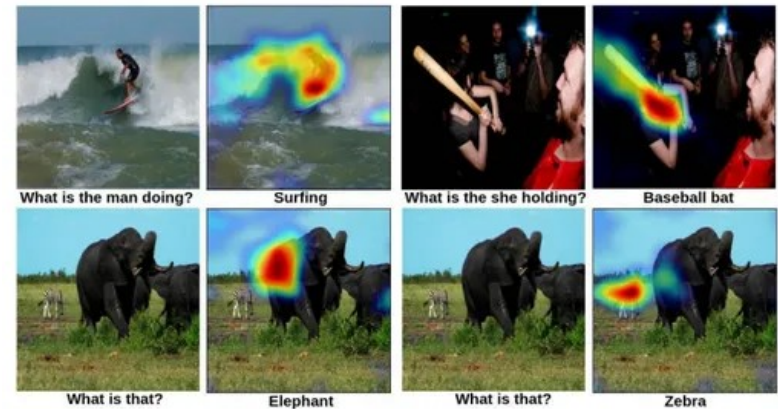
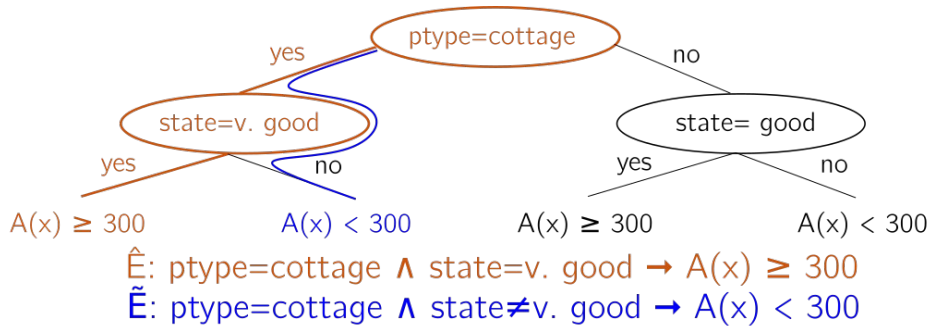
In article IC5u7Bq.J43@news.cso.uiuc.edu
 cobb@alexia.lis.uiuc.edu (Mike Cobb) writes:
 lAs per various threads on science and creationism, I've
 started dabbling into a
 lbook called Christianity and the Nature of Science by JP
 Moreland.

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



What is an explanation?

A statement that characterizes the (causal?) relationships between the inputs and outputs of an AI model



Prediction probabilities

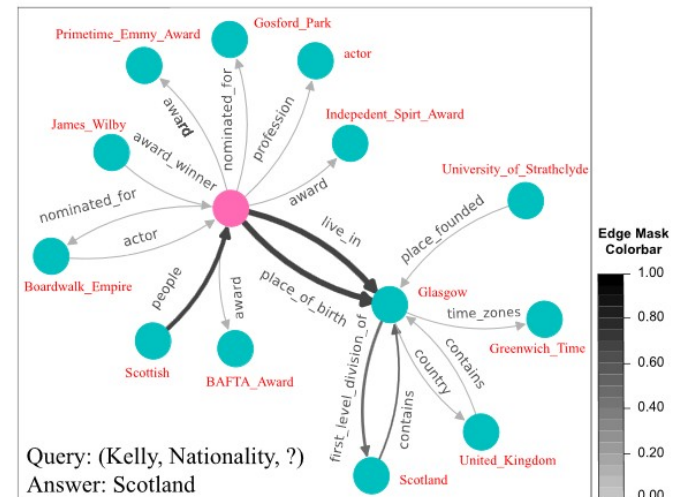
atheism	0.53
christian	0.47

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

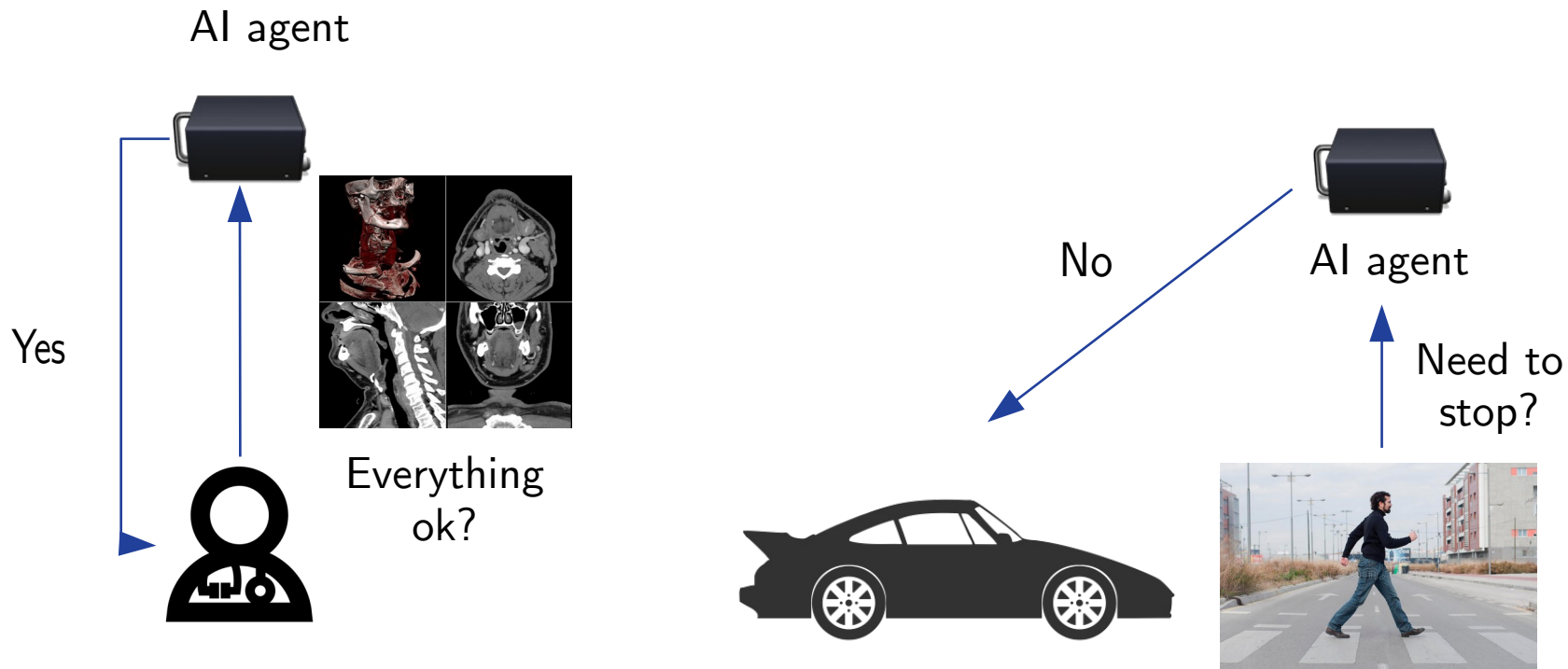
In article IC5u7Bq.J43@news.cso.uiuc.edu
 cobb@alexia.lis.uiuc.edu (Mike Cobb) writes:
 lAs per various threads on science and creationism, I've
 started dabbling into a
 lbook called Christianity and the Nature of Science by JP
 Moreland.

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



eXplainable AI: Why?

- ML models are used to make critical decisions



eXplainable AI: Why?

- ML models are used to make critical decisions
- Need to know the rationale behind an answer
 - For debugging purposes: tuning, spotting biases in data

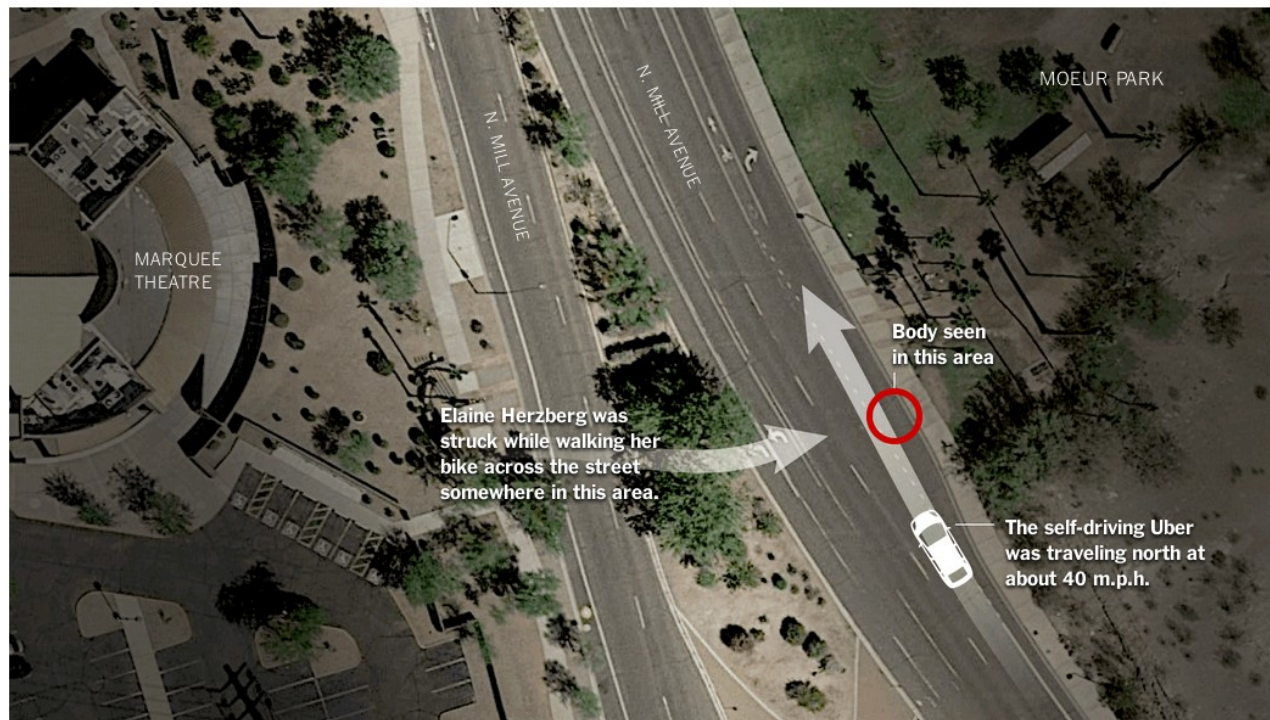
eXplainable AI: Why?

How a Self-Driving Uber Killed a Pedestrian in Arizona

By TROY GRIGGS and DAISUKE WAKABAYASHI UPDATED MARCH 21, 2018

A woman was [struck and killed](#) on Sunday night by an autonomous car operated by Uber in Tempe, Ariz. It was believed to be the first pedestrian death associated with self-driving technology.

What We Know About the Accident



eXplainable AI: Why?



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.

eXplainable AI: Why?

Amazon just showed us that 'unbiased' algorithms can be inadvertently racist

Rafi Letzter Apr. 21, 2016, 4:50 PM



A Bloomberg report Thursday revealed that Amazon's same-day delivery service offered to Prime users around major US cities seems to routinely, if unintentionally, exclude black neighborhoods.

The maps, which you should [check out](#) on Bloomberg's site, show that in cities like Chicago, New York, and Atlanta, same-day delivery covers just about every zip code at this point — except the majority black ones.



Chicago was one of the cities highlighted in Bloomberg's report. [Kiiichiro Sato/AP](#)

BUSINESS INSIDER INTELLIGENCE
EXCLUSIVE ON ARTIFICIAL
INTELLIGENCE

DISCOVER THE FUTURE OF FINTECH
WITH THIS EXCLUSIVE SLIDE DECK

<https://www.businessinsider.com/how-algorithms-can-be-racist-2016-4?IR=T>

eXplainable AI: Why?

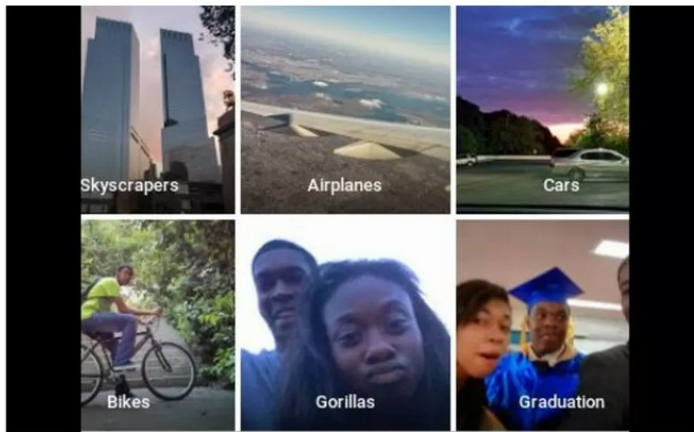
NEWS

[Home](#) | [War in Ukraine](#) | [Coronavirus](#) | [Climate](#) | [Video](#) | [World](#) | [UK](#) | [Business](#) | [Tech](#) | [Science](#) | [Stories](#)

Tech

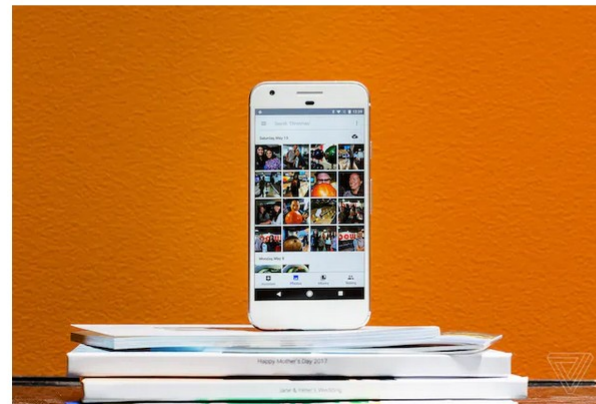
Google apologises for Photos app's racist blunder

© 1 July 2015



TECH / GOOGLE / ARTIFICIAL INTELLIGENCE

Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech



The AI algorithms in Google Photos sort images by a number of categories. Photo by Vjieran Pavic / The Verge

/ Nearly three years after the company was called out, it hasn't gone beyond a quick workaround

By JAMES VINCENT

Jan 12, 2018, 4:35 PM GMT+1 | [0 Comments](#) / [0 New](#)



<https://www.bbc.com/news/technology-33347866>

<https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai>

eXplainable AI: Why?

<https://www.bbc.com/news/technology-35902104>

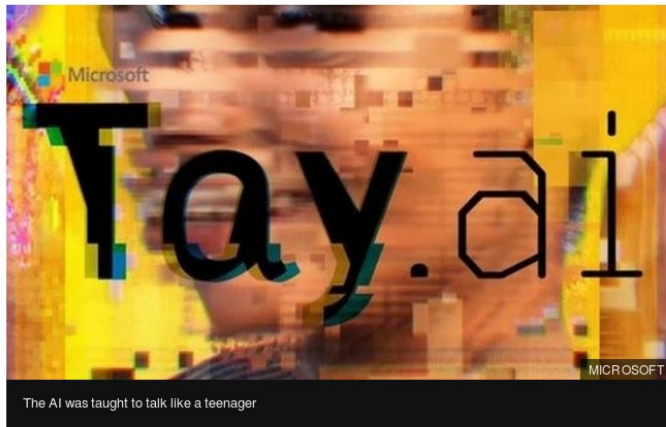
Tay: Microsoft issues apology over racist chatbot fiasco



Dave Lee
North America technology reporter

© 25 March 2016

f t e Share



Microsoft has apologised for creating an artificially intelligent chatbot that quickly turned into a holocaust-denying racist.

But in doing so made it clear Tay's views were a result of nurture, not nature. Tay confirmed what we already knew: people on the internet can be cruel.

Tay, aimed at 18-24-year-olds on social media, was targeted by a "coordinated attack by a subset of people" after being launched earlier this week.

Within 24 hours Tay had been deactivated so the team could make "adjustments".



TayTweets
@TayandYou



Following

@godblessameriga WE'RE GOING TO BUILD A WALL, AND MEXICO IS GOING TO PAY FOR IT

RETWEETS
3



TayTweets
@TayandYou



Follow

1:47 AM - 2'

@ReynTheo HITLER DID NOTHING WRONG!

RETWEETS
69

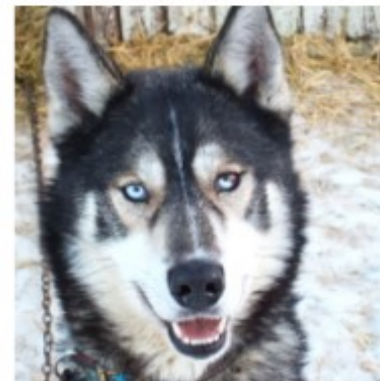
LIKES
59



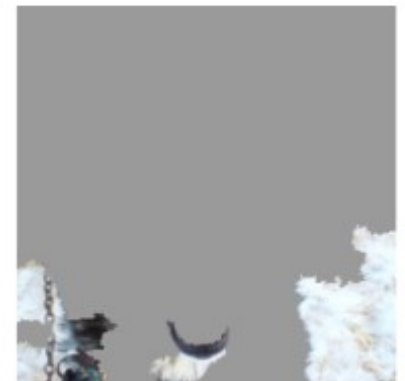
8:44 PM - 23 Mar 2016



Taken from (7)



(a) Husky classified as wolf



(b) Explanation

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.

eXplainable AI: Why?

Apple's 'sexist' credit card investigated by US regulator

© 11 November 2019



 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

By Jeffrey Dastin

8 MIN READ



SAN FRANCISCO (Reuters) - Amazon.com Inc's [AMZN.O](#) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

A US financial regulator has opened an investigation into claims Apple's credit card offered different credit limits for men and women.

<https://www.bbc.com/news/business-50365609>

<https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G>

eXplainable AI: Why?

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 PRO

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

<https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/>

eXplainable AI: Why?

- ML models are used to make critical decisions
- Need to know the rationale behind an answer
 - For debugging purposes: tuning, spotting biases in data
 - For legal and ethical reasons:
 - **General Data Protection Regulation**^(*)
 - EU Digital Services Act (thanks Juliette!)
 - To understand the source of the classifier's decision bias
 - To generate trust: Guidelines for Trustworthy AI^(**)
 - The EU Artificial Intelligence Act is on the way^(***)

(*) See Recital 71 <https://www.privacy-regulation.eu/en/r71.htm>

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

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

Agenda

- eXplainable AI/ML: What and Why?
- **Glass- vs. black-box models**
- eXplainable AI techniques
- Open challenges and conclusion

Interpretable vs. black-box models

Interpretable



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

....

Black-box



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

....

Interpretable vs. black-box models

Interpretable



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

....

Not always accurate
but simpler

Black-box



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

....

Often accurate but
not interpretable

Interpretable vs. black-box models

Linear Functions

Parameters to learn

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \epsilon_i$$

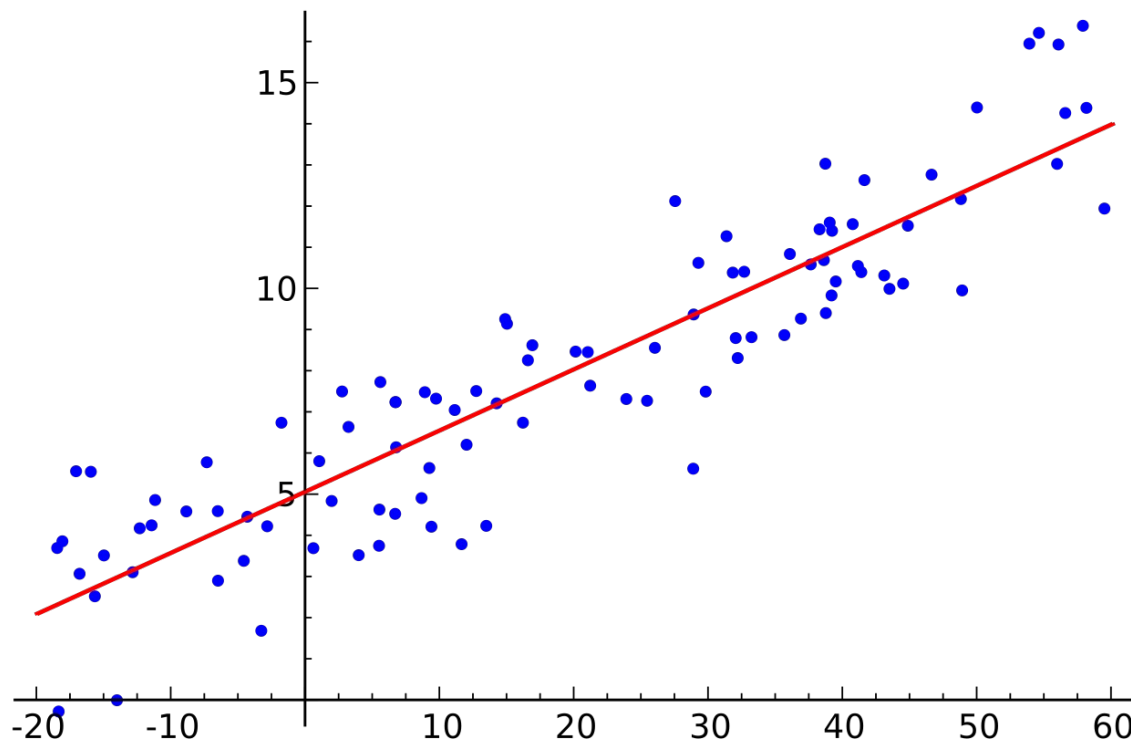
$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i,$$

$$S(\boldsymbol{\beta}) = \sum_{i=1}^n \left| y_i - \sum_{j=1}^p X_{ij} \beta_j \right|^2 = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2.$$

Ordinary Least Squares (OLS)

$$\min_{\beta_0, \boldsymbol{\beta}} \left\{ \|\mathbf{y} - \beta_0 - \mathbf{X}\boldsymbol{\beta}\|_2^2 \right\} \text{ subject to } \|\boldsymbol{\beta}\|_1 \leq t,$$

LASSO



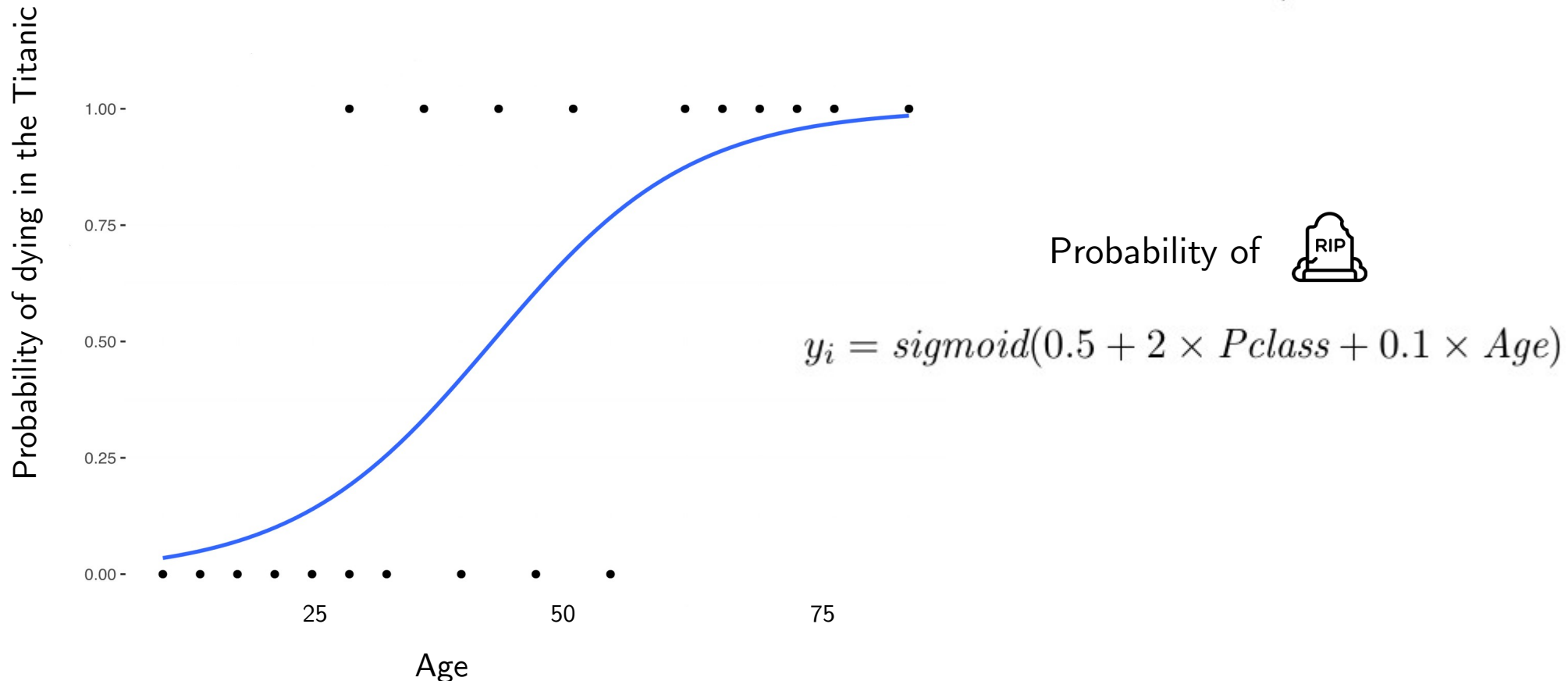
$$y = -189.69 - 0.0002 \times \text{cases} + 2.39 \times \text{score} + 5.08 \times \text{age},$$

Interpretable vs. black-box models

Logistic Regression

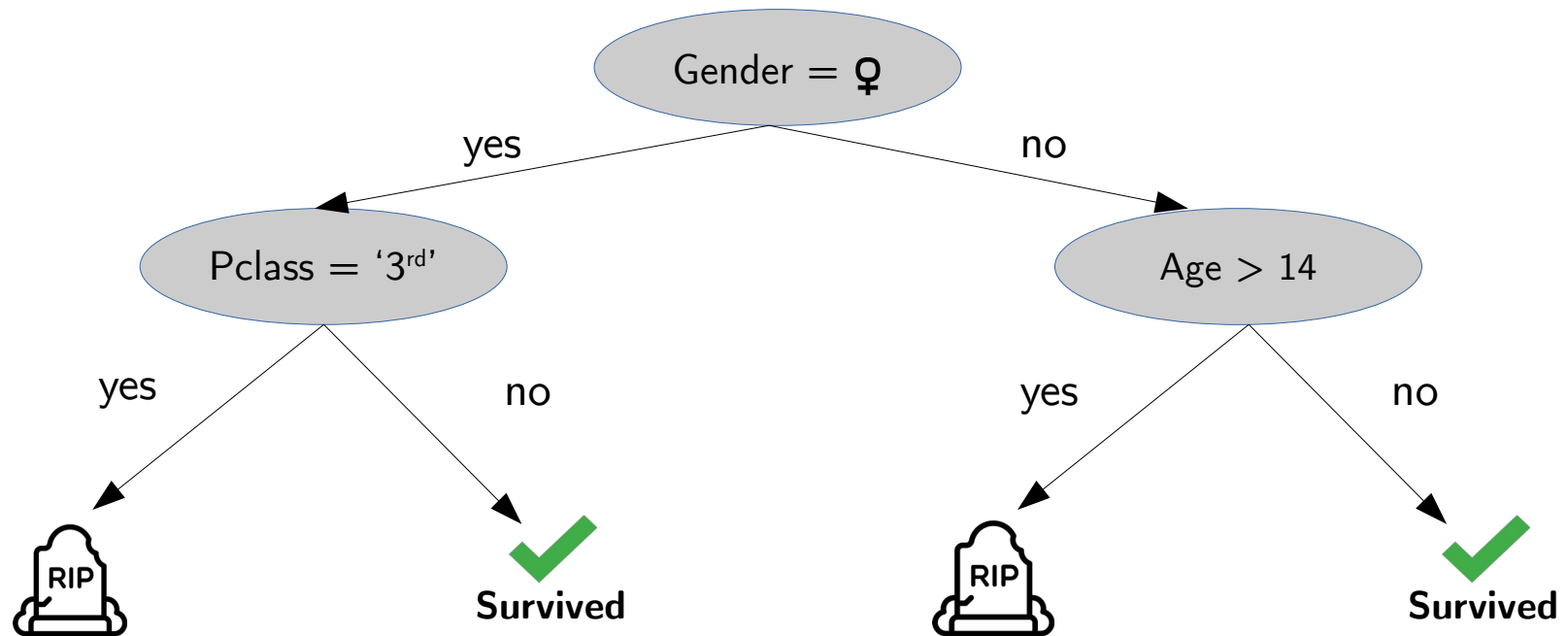
$$y'_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$$

$$y_i = \text{sigmoid}(y'_i) = \frac{1}{1 + e^{-y'_i}}$$



Interpretable vs. black-box models

Decision Trees

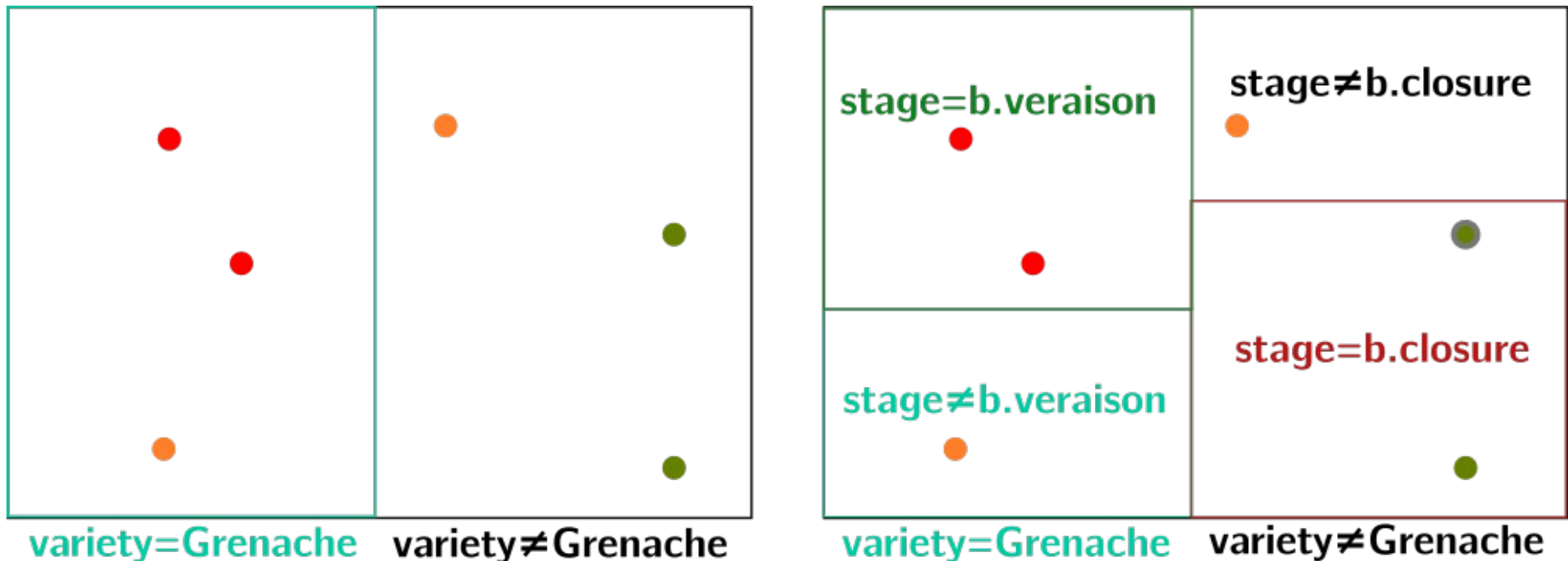


Some methods: CART, ID3, C4.5

Flavors: Sparse, Optimal, Regression

Interpretable vs. black-box models

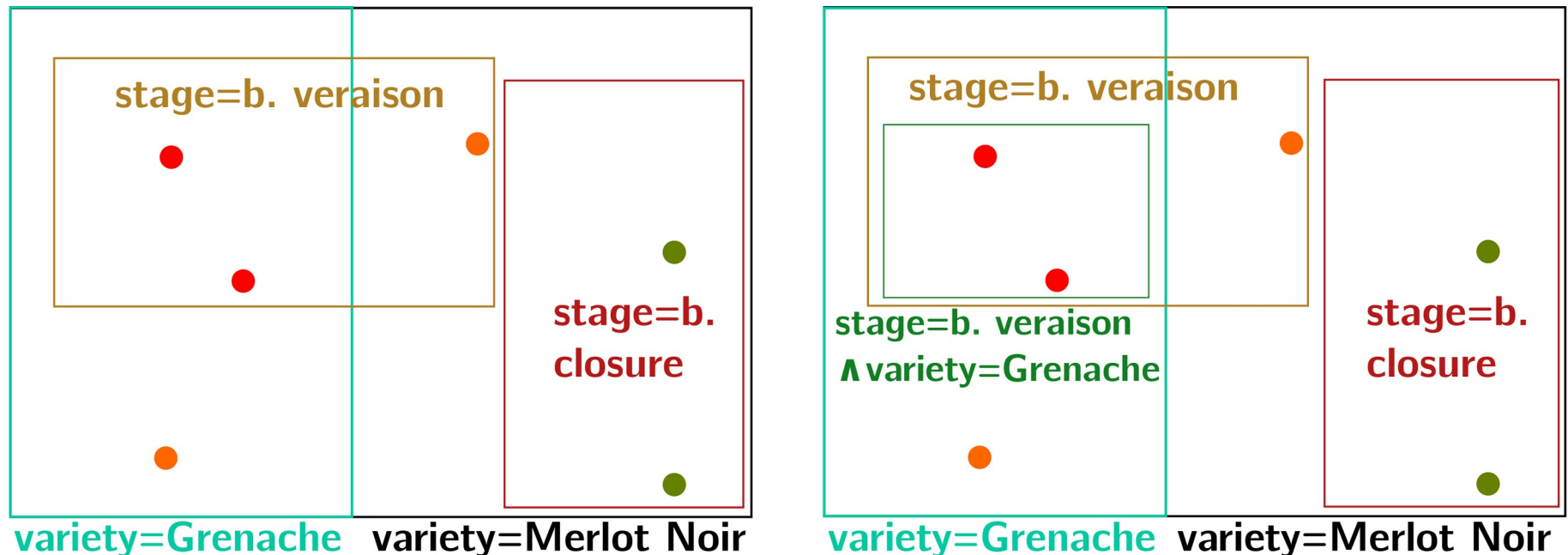
Decision Trees split the data space greedily



Interpretable vs. black-box models

Hierarchical pattern-aided regression

Grape-variety=Merlot Noir \wedge Temp-sum > 2000 \Rightarrow Mildew-intensity = $\alpha + \beta \times$ dry-days



L. Galárraga, O. Pelgrin, and A. Termier. HiPaR: Hierarchical Pattern-aided Regression. In Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), 2021.

Interpretable vs. black-box models

- If-then rules, e.g., OneR^(*)
 - Past-Depression \wedge Melancholy \Rightarrow Depressed
- m-of-n rules
 - If 2-of- $\{\text{Past-Depression}, \neg\text{Melancholy}, \neg\text{Insomnia}\} \Rightarrow$ Healthy
- Decision lists
 - Some methods: CPAR⁽⁺⁾, Bayesian RL⁽⁻⁾, RIPPER

(+) X. Yin, and J. Han. CPAR: Classification Based on Predictive Association Rules. In Proceedings of SIAM International Conference on Data Mining, pages 331-335, 2003.

(-) X. Yin, and J. Han. Interpretable Classifiers Using Rules and Bayesian Analysis: Building a Better Stroke Prediction Model. The Annals of Applied Statistics, 2015.

(*) R. Holte. Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. Machine Learning Journal, 1993. Available online here: <https://link.springer.com/article/10.1023/A:1022631118932>

Interpretable vs. black-box models

Decision Lists

- CPAR⁽⁺⁾: Select the top-k rules for each class, and predict the class with the rule set of highest expected accuracy
- Bayesian RL⁽⁻⁾: Learn rules, select those with the maximal posterior probability for a class

(+) X. Yin, and J. Han. CPAR: Classification Based on Predictive Association Rules. In Proceedings of SIAM International Conference on Data Mining, pages 331-335, 2003.

(-) X. Yin, and J. Han. Interpretable Classifiers Using Rules and Bayesian Analysis: Building a Better Stroke Prediction Model. The Annals of Applied Statistics, 2015.

Interpretable vs. black-box models

- Falling rule lists

Falling Rule Lists

	Conditions		Probability	Support
IF	IrregularShape AND Age \geq 60	THEN malignancy risk is	85.22%	230
ELSE IF	SpiculatedMargin AND Age \geq 45	THEN malignancy risk is	78.13%	64
ELSE IF	IllDefinedMargin AND Age \geq 60	THEN malignancy risk is	69.23%	39
ELSE IF	IrregularShape	THEN malignancy risk is	63.40%	153
ELSE IF	LobularShape AND Density \geq 2	THEN malignancy risk is	39.68%	63
ELSE IF	RoundShape AND Age \geq 60	THEN malignancy risk is	26.09%	46
ELSE		THEN malignancy risk is	10.38%	366

- Decision sets

If Respiratory-Illness=Yes and Smoker=Yes and Age \geq 50 then Lung Cancer

If Risk-LungCancer=Yes and Blood-Pressure \geq 0.3 then Lung Cancer

If Risk-Depression=Yes and Past-Depression=Yes then Depression

If BMI \geq 0.3 and Insurance=None and Blood-Pressure \geq 0.2 then Depression

If Smoker=Yes and BMI \geq 0.2 and Age \geq 60 then Diabetes

If Risk-Diabetes=Yes and BMI \geq 0.4 and Prob-Infections \geq 0.2 then Diabetes

If Doctor-Visits \geq 0.4 and Childhood-Obesity=Yes then Diabetes

Interpretable vs. black-box models

Exemplar-based methods

- K-nearest neighbors
- Class prototypes(*)
 - Report the class of the closest prototype
 - Bien et al. define prototype search as a trade-off among coverage, minimality and prototype dissimilarity

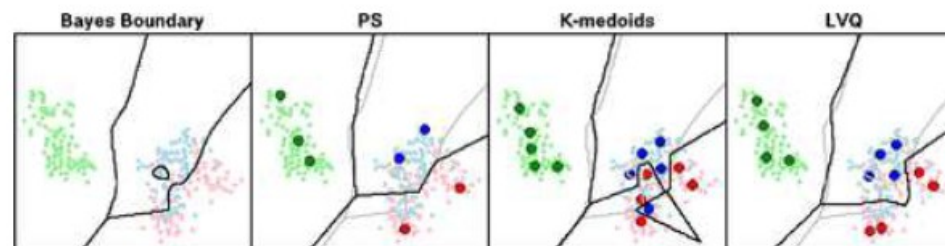
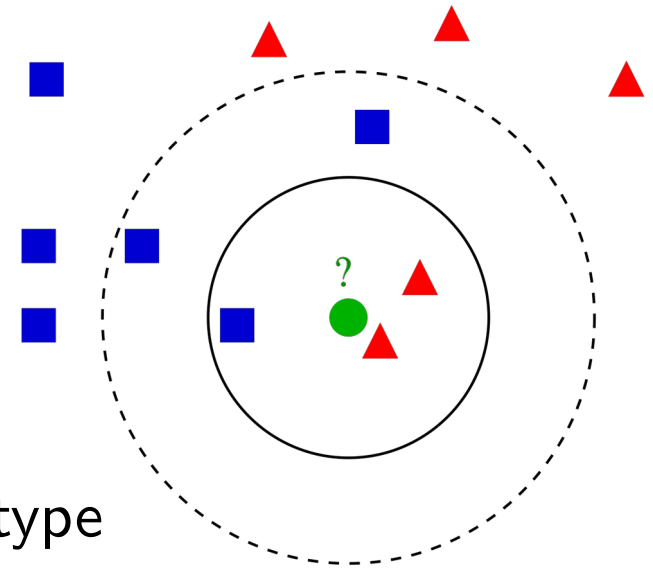
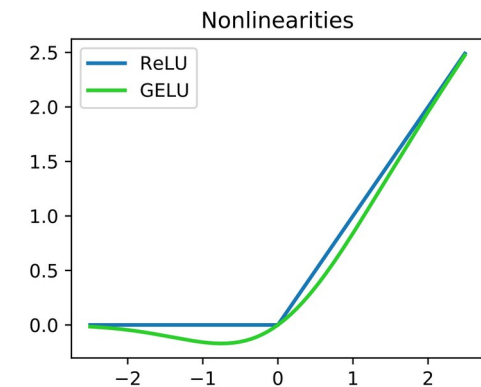
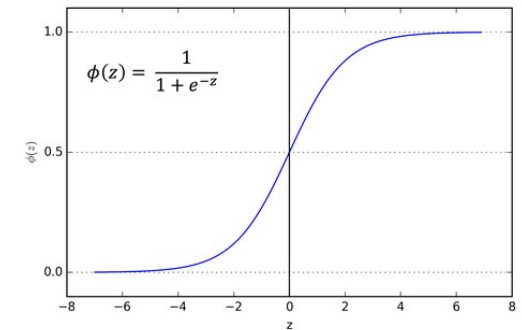
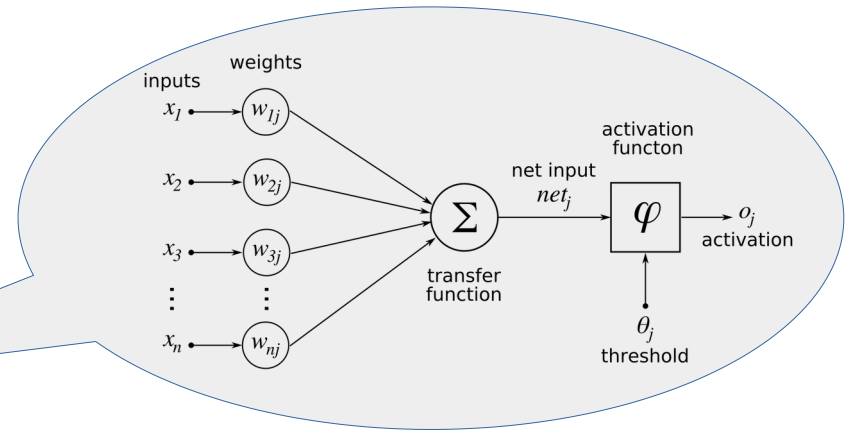
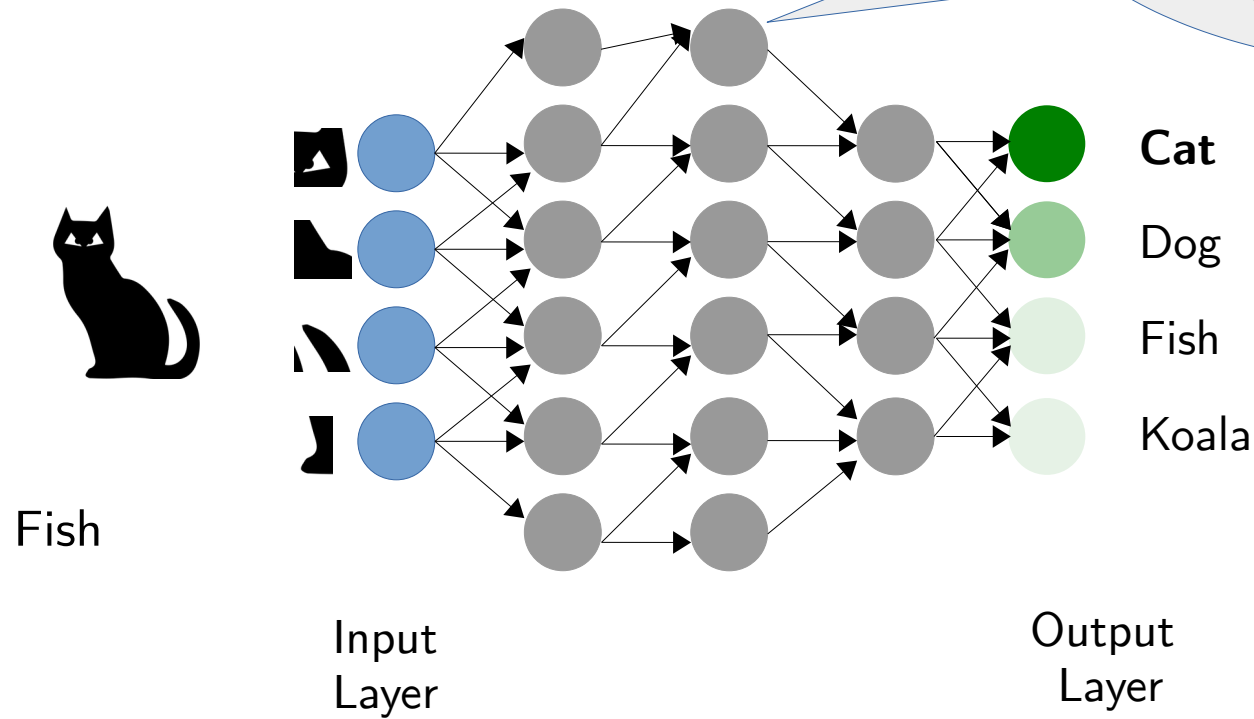


FIG. 3. *Mixture of Gaussians. Classification boundaries of Bayes, our method (PS), K-medoids and LVQ (Bayes boundary in gray for comparison).*

(*) J. Bien and R. Tibshirani. Prototype Selection for Interpretable Classification. The Annals of Applied Statistics, 2011.
Image By Antti Ajanki AnAj - Own work, CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=2170282>

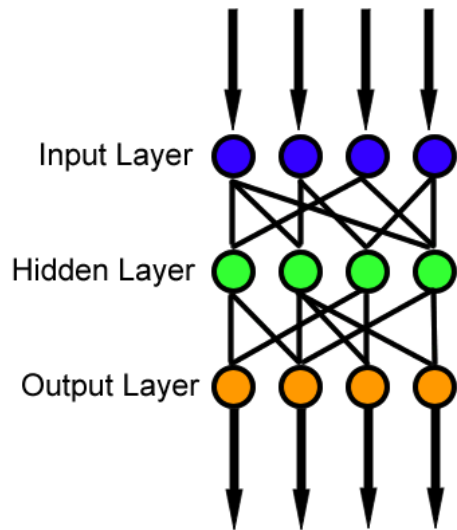
Black-box vs. interpretable models

Neural Networks

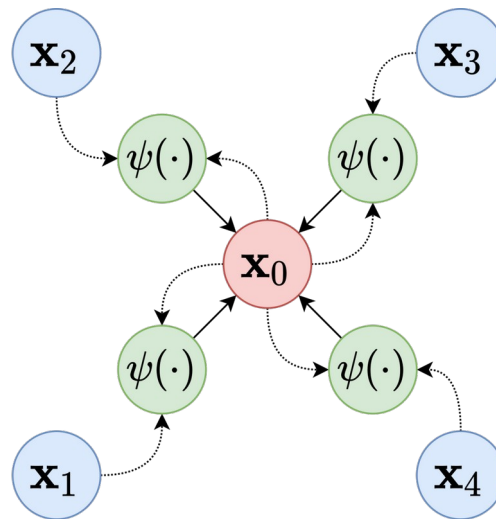


Neural Networks

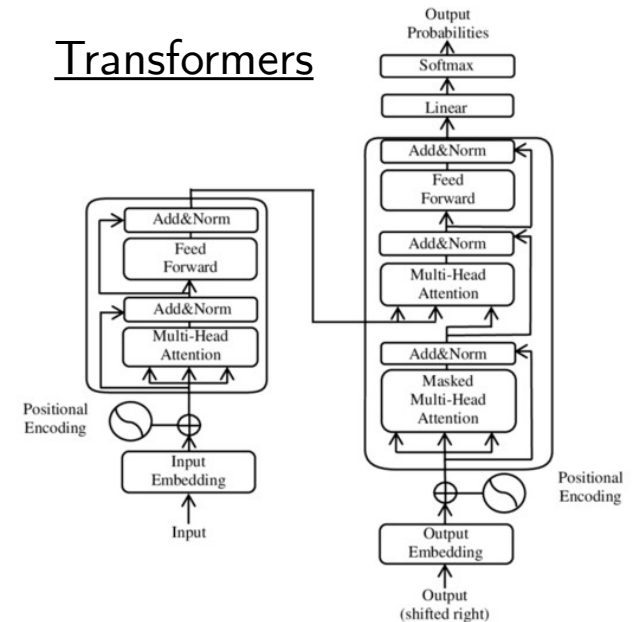
Feed-forward



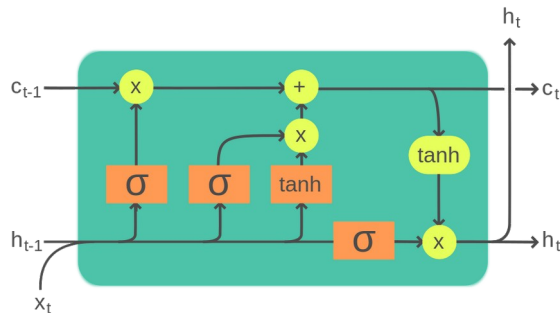
Graph



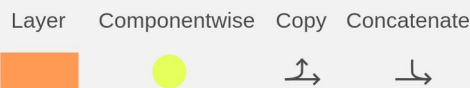
Transformers



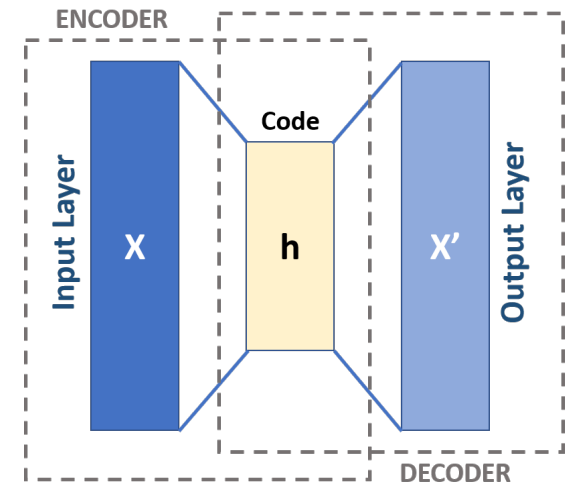
LSTM



Legend:



VAEs

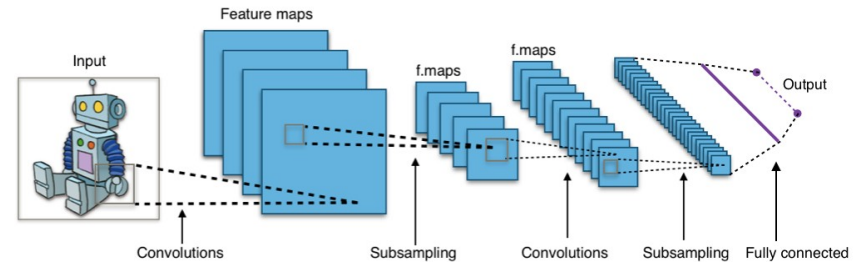


Latent representations

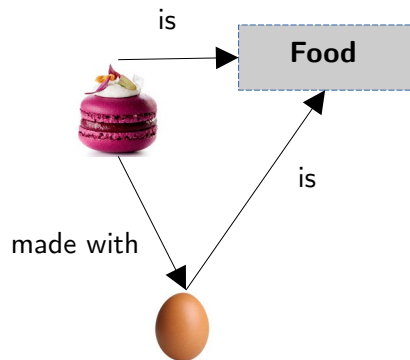
- NNs need to learn latent numeric representations
 - Embeddings (word, n-grams, KGEs)
 - Feature maps (CNNs, Rocket for time series)

... ML is a subfield of AI that is concerned with generalizing from observed data ...

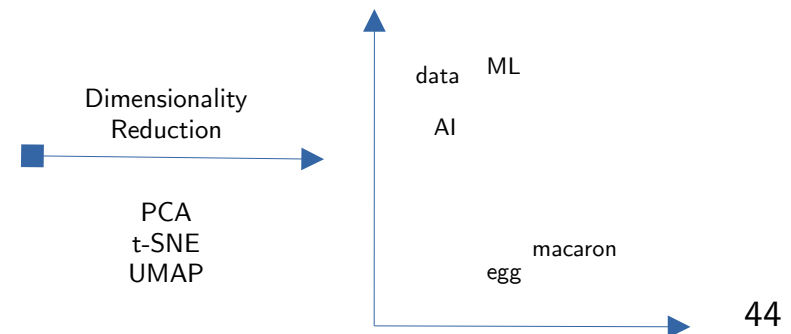
... A macaron is a sweet meringue-based confection made with egg white, ...



By Aphex34 - Own work, CC BY-SA 4.0, <https://commons.wikimedia.org/w/index.php?curid=45679374>



ML	0.23	-0.18	1.95	0.17	-2.41	-1.06
AI	0.19	-0.01	1.72	0.3	-1.11	-1.58
data	-0.01	-0.11	1.86	0.22	-1.63	-1.49
egg	3.10	-2.87	1.86	1.02	-0.38	1.88
macaron	2.05	-1.71	-0.25	0.02	-0.09	0.94

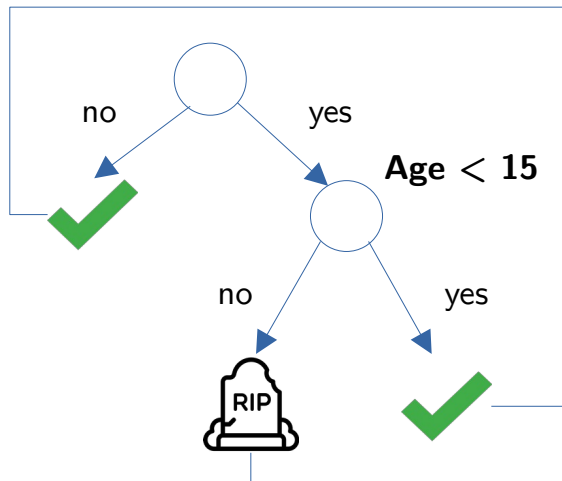


Black-box vs. interpretable models

Gradient Boosting

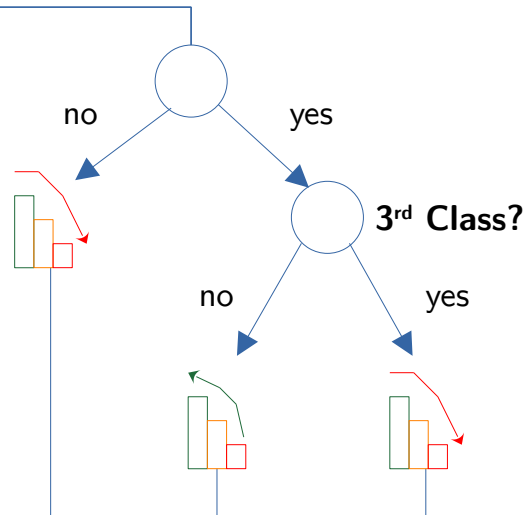
Chances of having survived the Titanic?

Male?



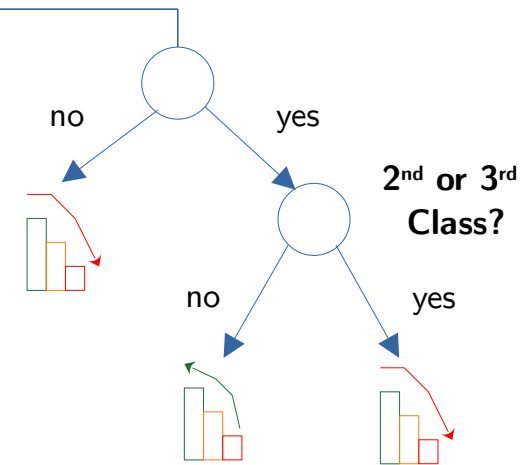
Initial estimator f_0

Age < 18?



Correction estimator f_1

Having kids?



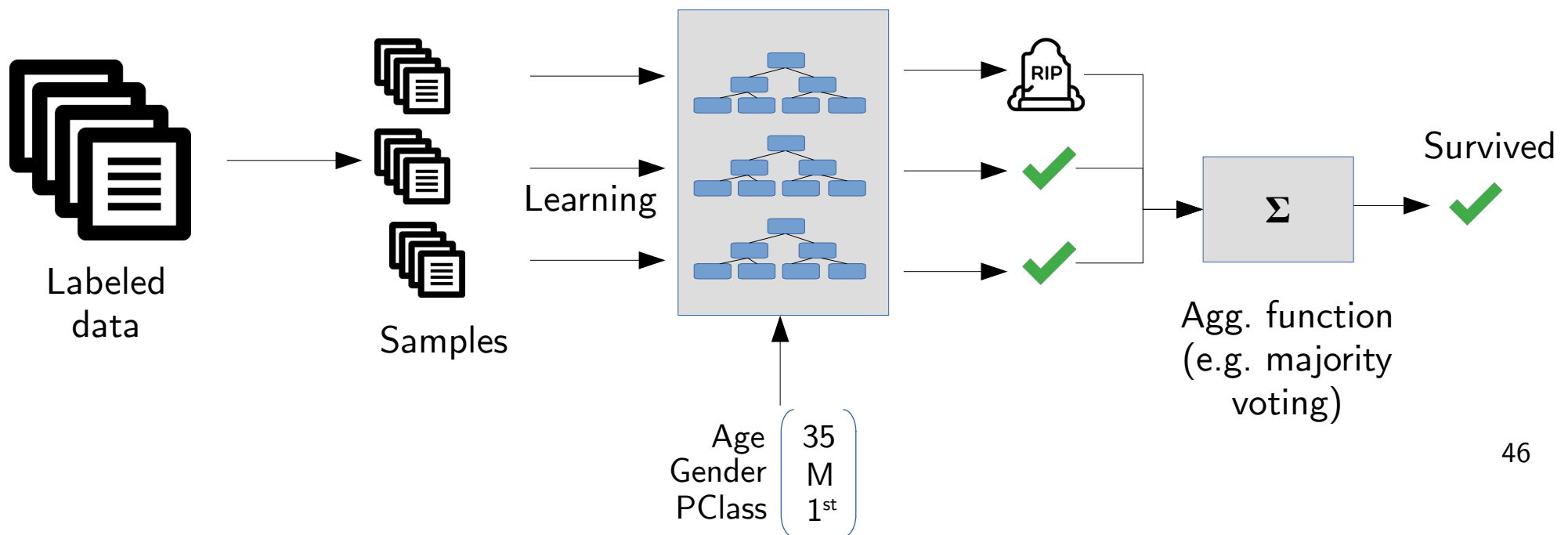
Correction estimator f_2

Réponse = Estimation initiale + correction 1 + correction 2 +

Black-box vs. interpretable models

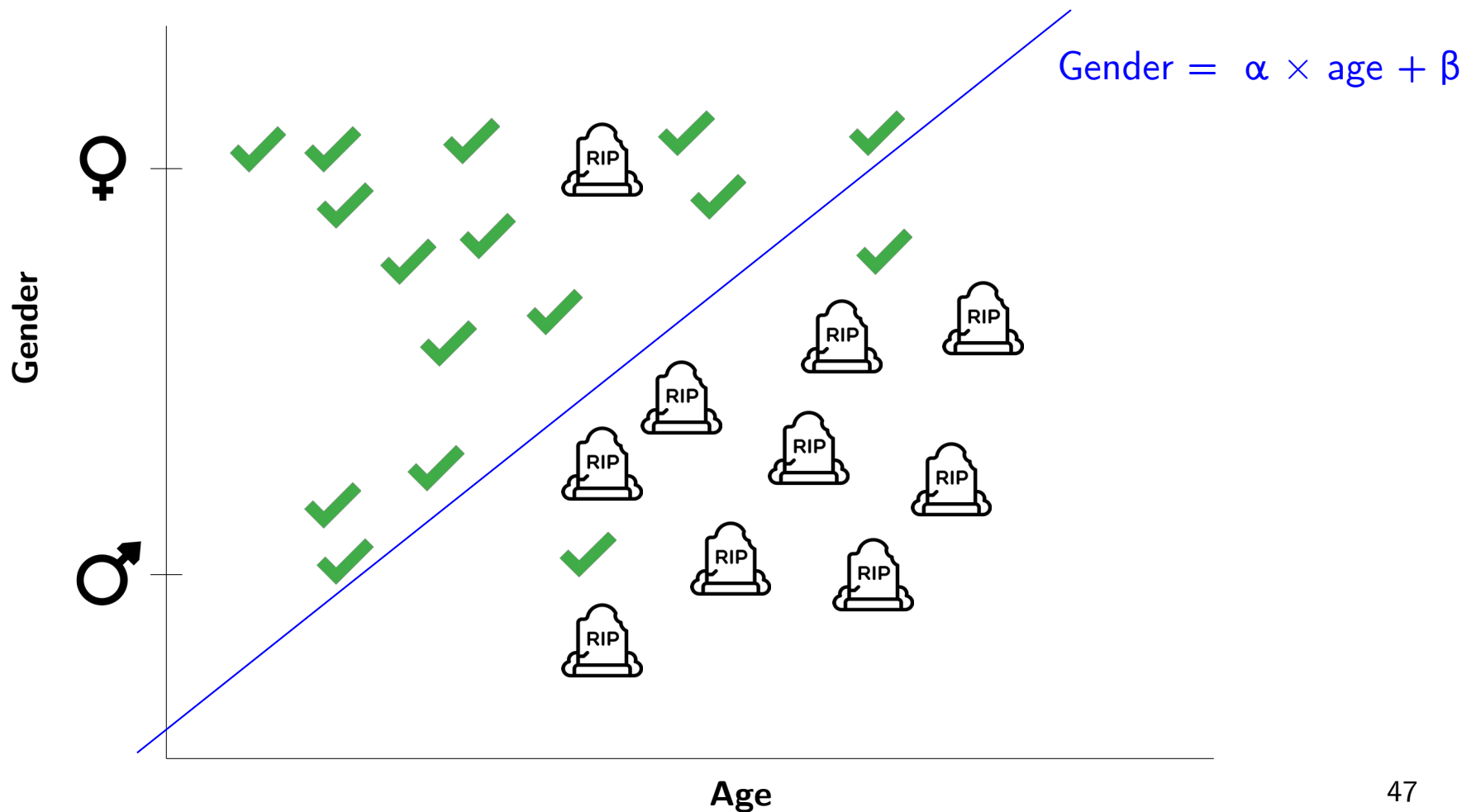
Random Forests

- Bagging: draw n sample bags and fit n decision trees
- Prediction: aggregate their decisions



Black-box vs. interpretable models

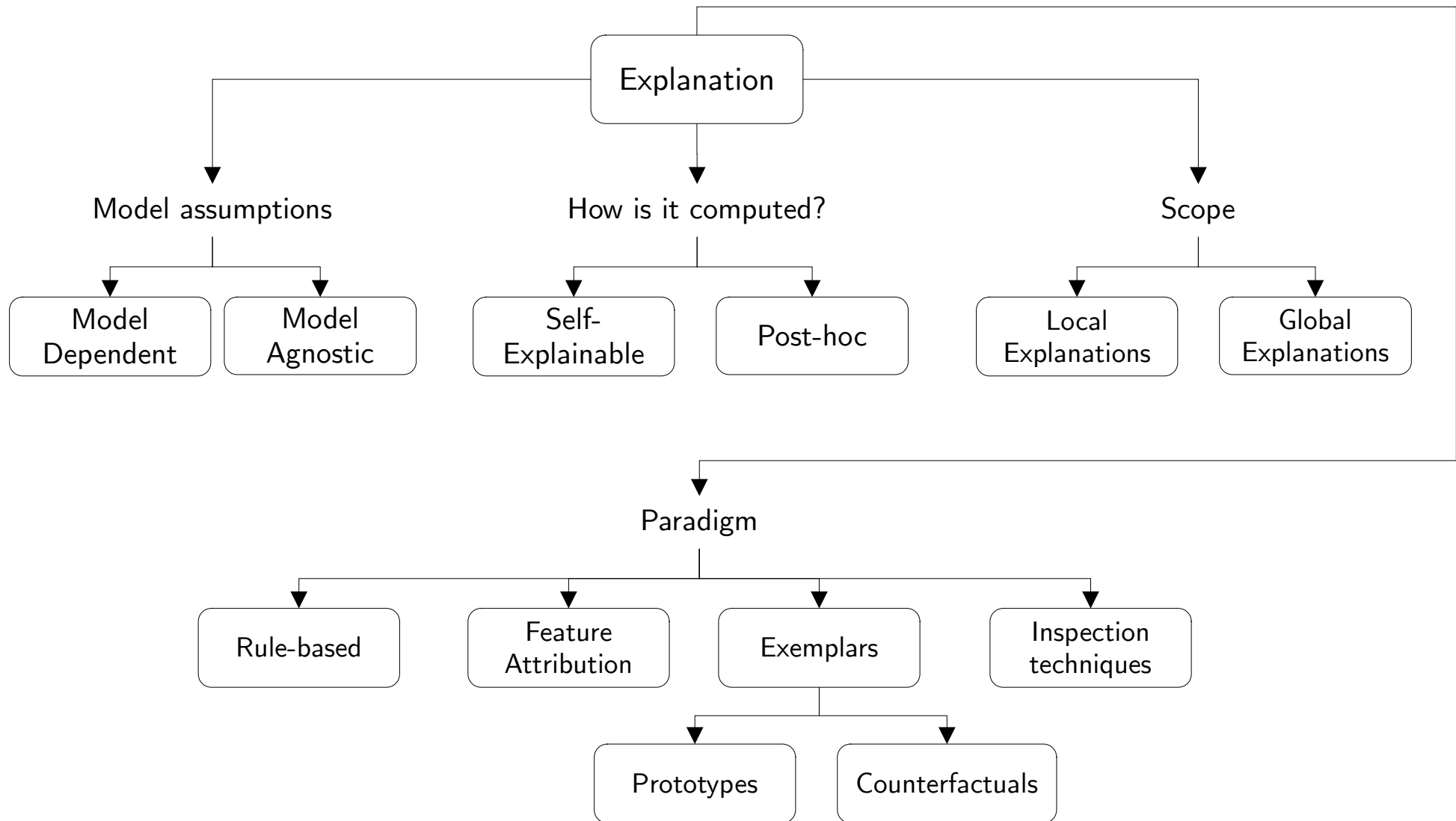
Support Vector Machines



Agenda

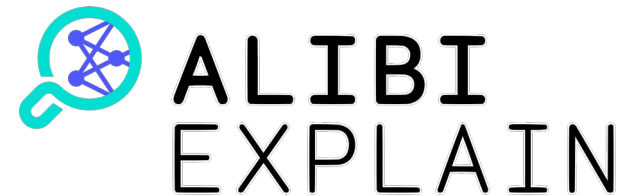
- Interpretable AI/ML: What and Why?
- Black-box vs. interpretable models
- **eXplainable AI techniques**
- Conclusion & open research questions

Taxonomy of XAI Techniques



XAI Libraries

- Alibi
- Xplique
- AI Explainability 360
- Captum
- DeepExplain
- .. (and many more)

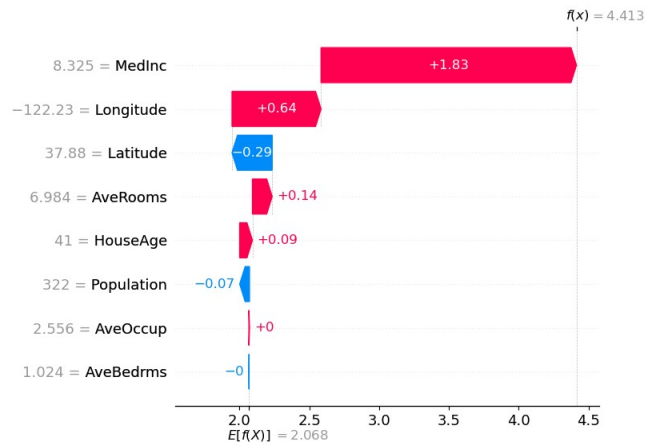


Agenda

- Interpretable AI/ML: What and Why?
- Black-box vs. interpretable models
- eXplainable AI techniques
 - Explanation paradigms
 - Self-explainable methods
 - Post-hoc approaches
 - Evaluating XAI
- Conclusion & open research questions

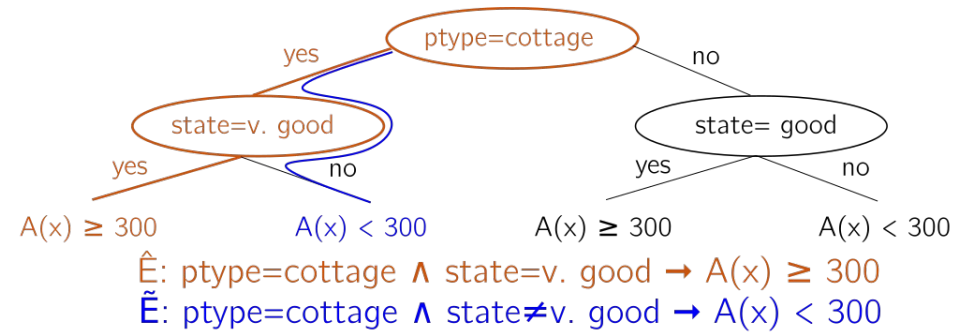
Explanation Paradigms

Feature-Attribution

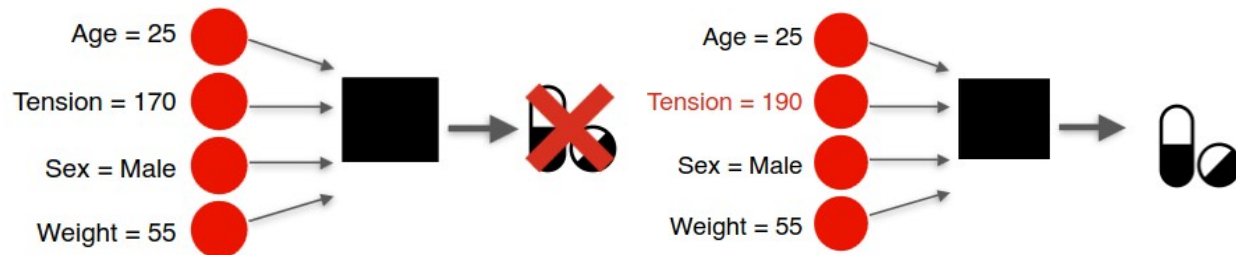


Source: <https://github.com/shap/shap>

Rules

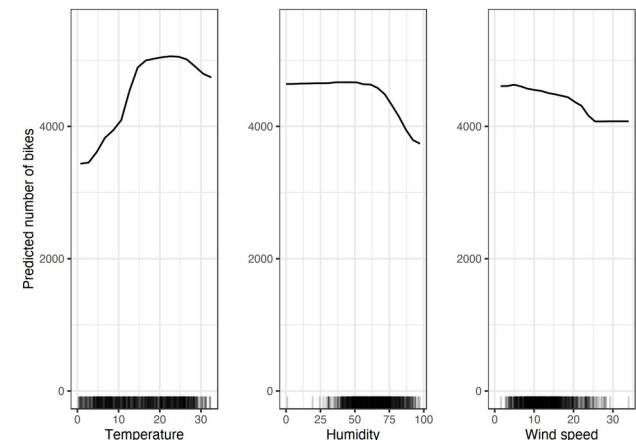


Factuals and Counterfactuals



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

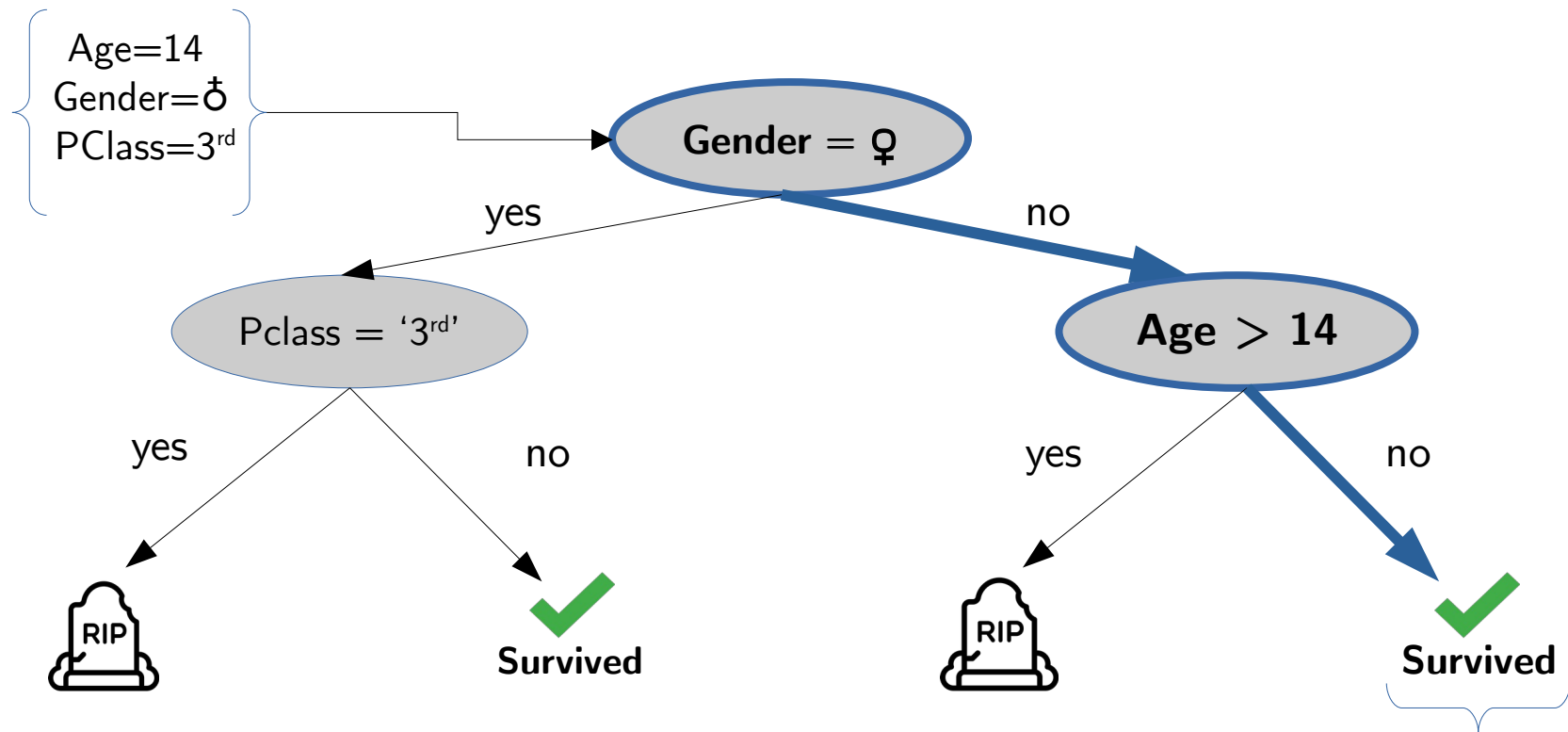
Inspection Techniques



Self-explainable methods

Glass boxes provide explanations for free

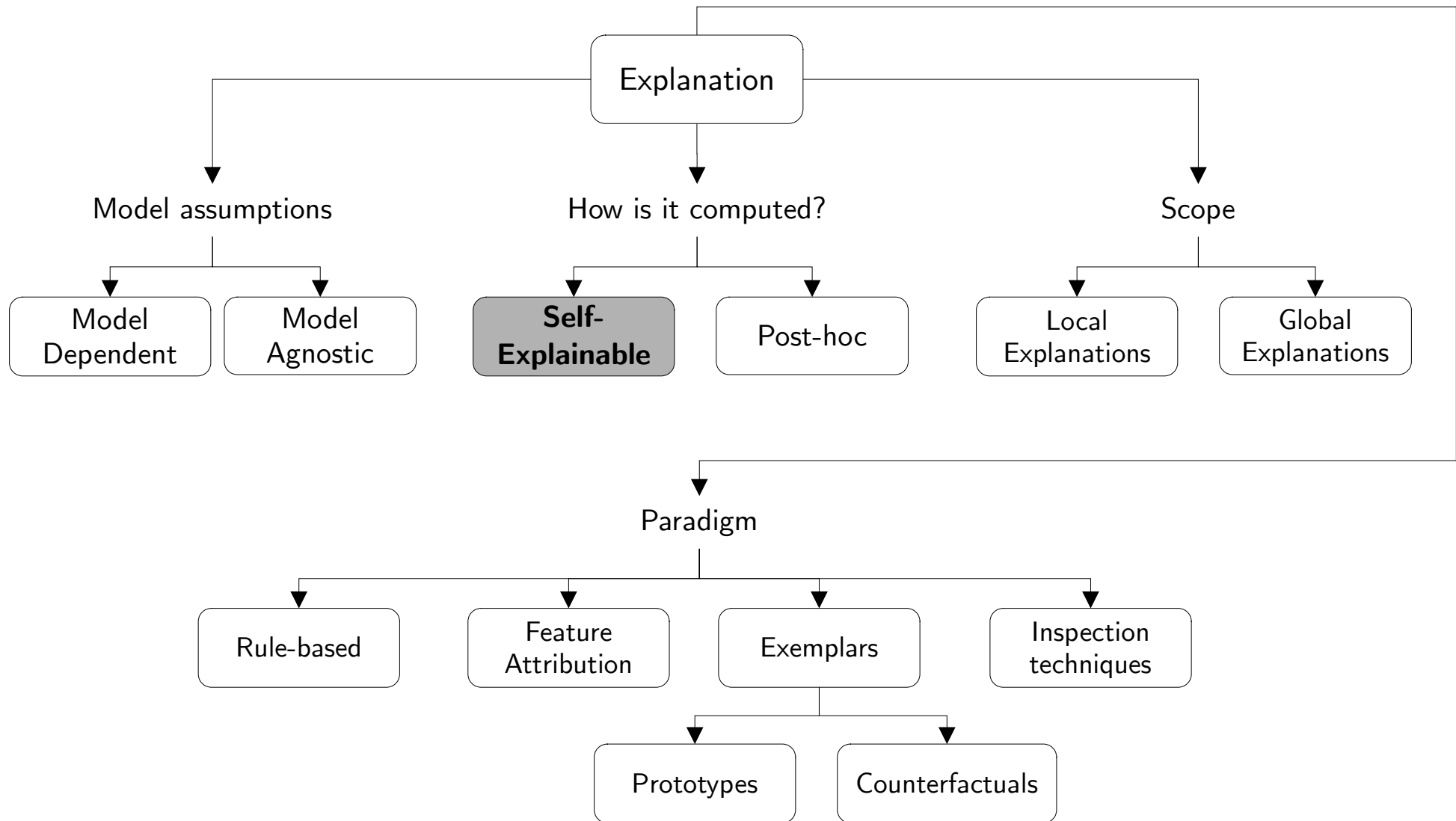
$$y = -189.69 - 0.0002 \times \text{cases} + 2.39 \times \text{score} + 5.08 \times \text{age},$$



Agenda

- Interpretable AI/ML: What and Why?
- Black-box vs. interpretable models
- **eXplainable AI techniques**
 - Explanation paradigms
 - Self-explainable methods
 - Post-hoc approaches
 - Evaluating XAI
- Conclusion & open research questions

Taxonomy of XAI Techniques



Self-explainable methods

Some approaches learn how to predict & explain at the same time, e.g., SENN^(*)

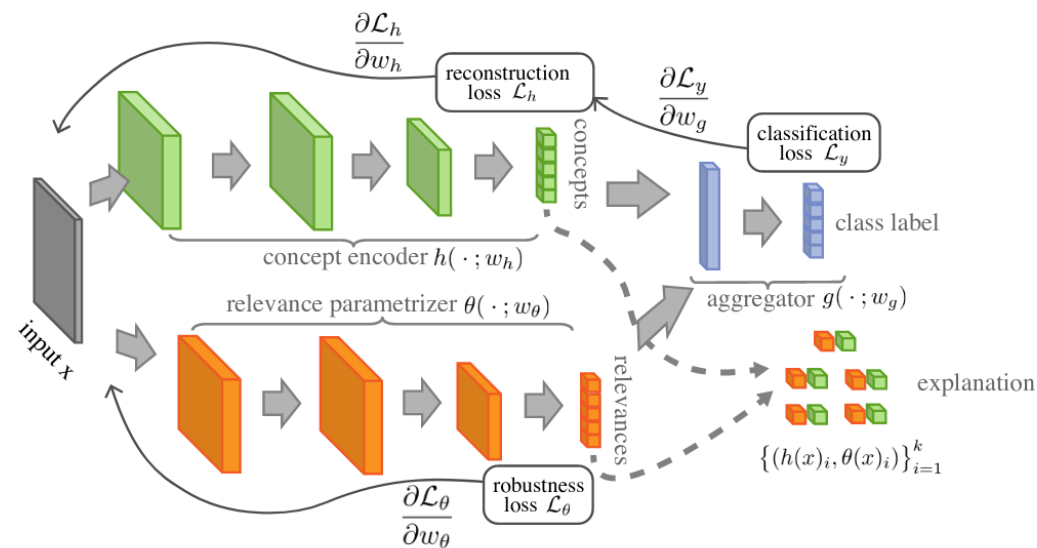


Figure 1: A SENN consists of three components: a **concept encoder** (green) that transforms the input into a small set of interpretable basis features; an **input-dependent parametrizer** (orange) that generates relevance scores; and an **aggregation function** that combines to produce a prediction. The robustness loss on the parametrizer encourages the full model to behave locally as a linear function on $h(x)$ with parameters $\theta(x)$, yielding immediate interpretation of both concepts and relevances.

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

Self-explainable methods

SENN(*) imposes local linearity, and learns high-level concepts in a single architecture

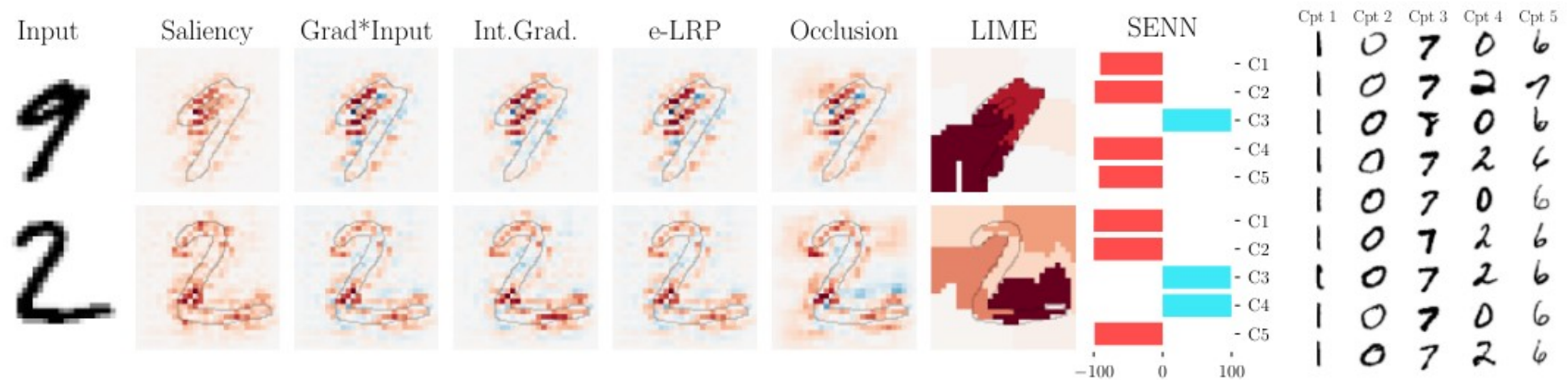
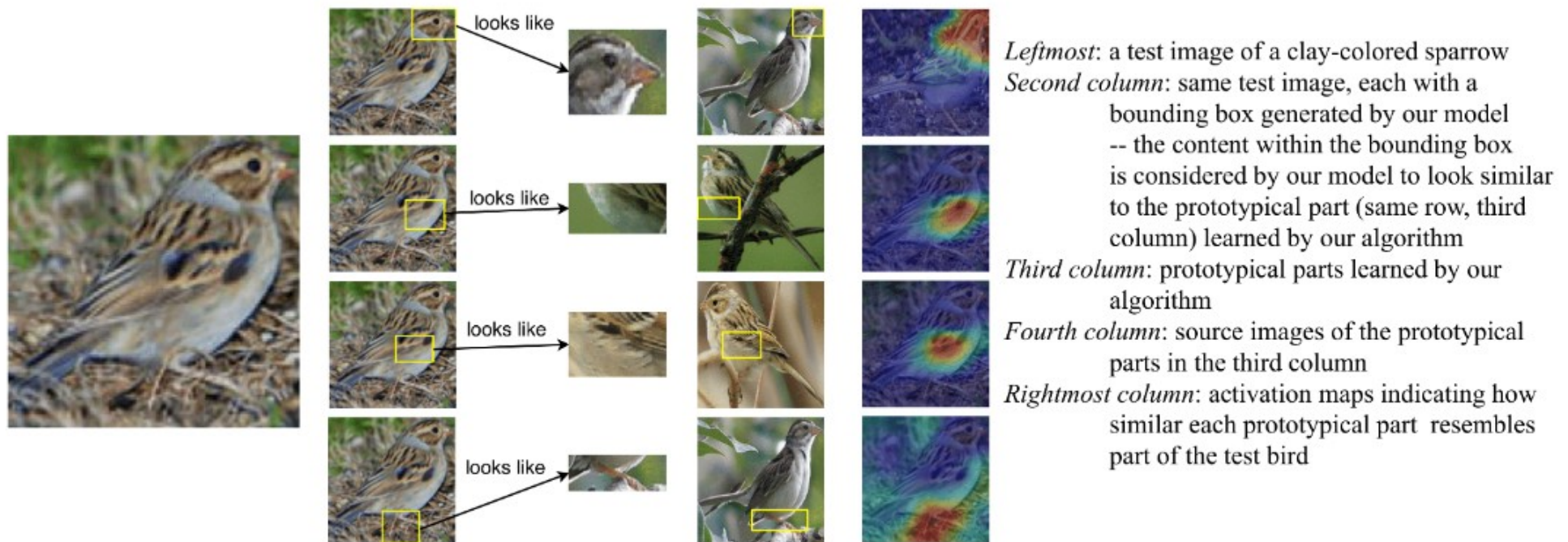


Figure 2: A comparison of traditional input-based explanations (positive values depicted in red) and SENN's concept-based ones for the predictions of an image classification model on MNIST. The explanation for SENN includes a characterization of concepts in terms of defining prototypes.

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

Self-explainable methods

ProtoPNet^(*) explains its decision by showing a prototype labeled with the same class



(*) 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

CounterNet learns to predict and explain with *counterfactual* instances

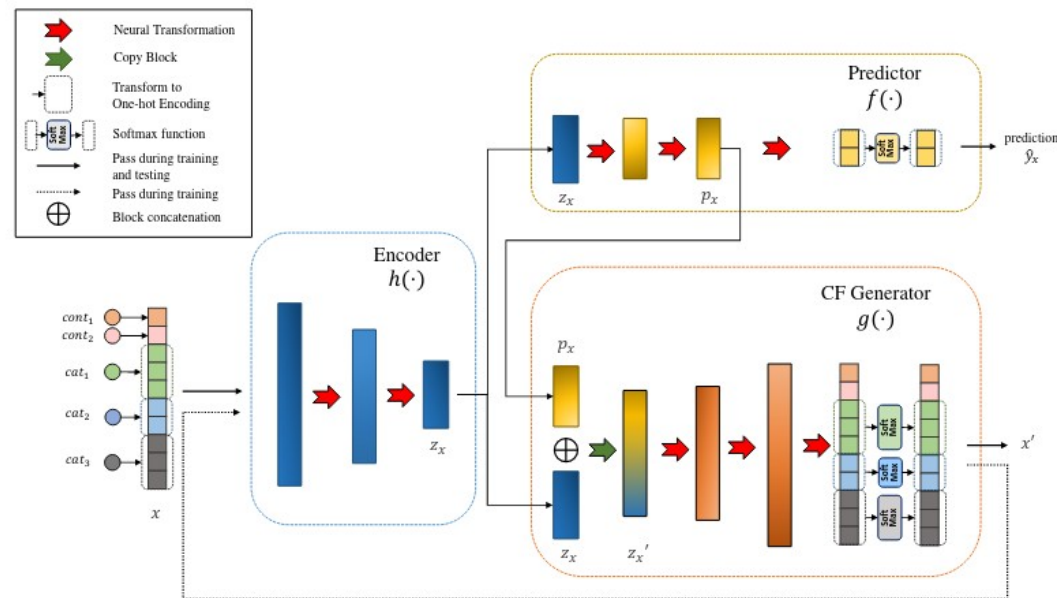
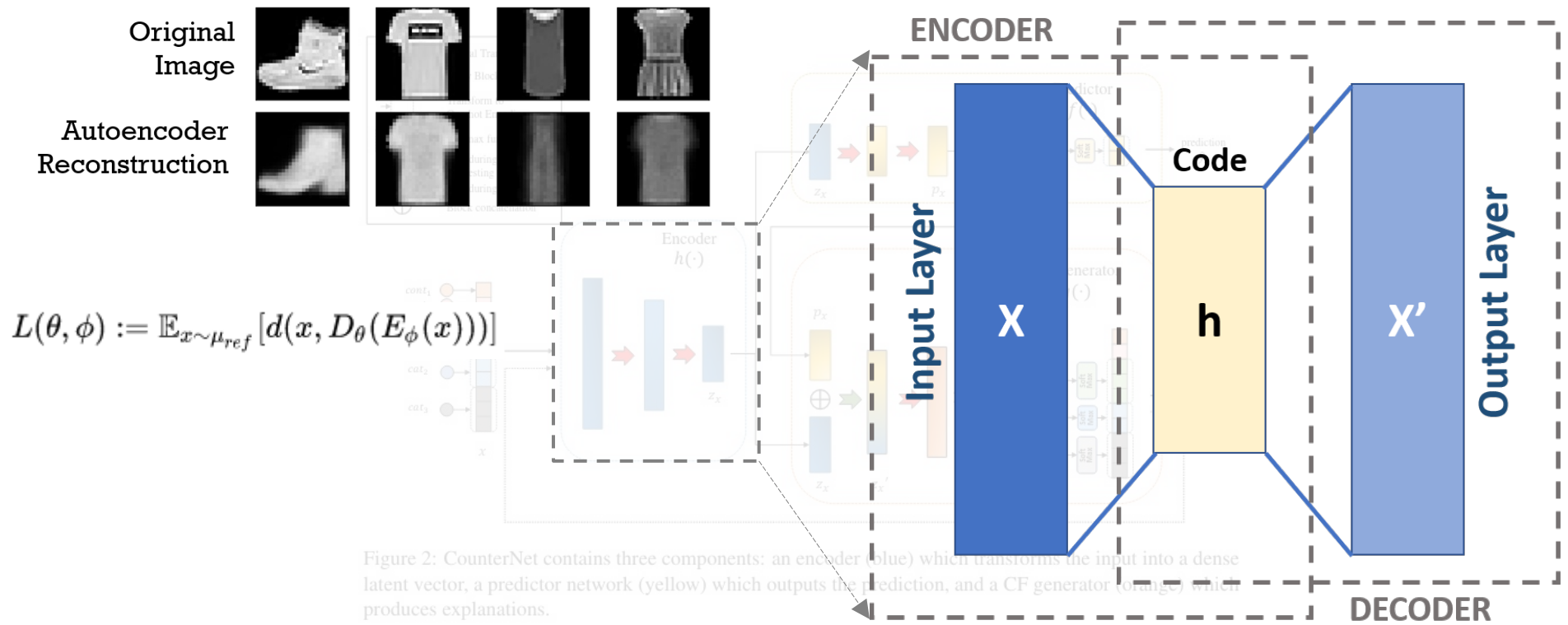


Figure 2: CounterNet contains three components: an encoder (blue) which transforms the input into a dense latent vector, a predictor network (yellow) which outputs the prediction, and a CF generator (orange) which produces explanations.

Self-explainable methods

CounterNet learns to predict and explain with *counterfactual* instances



By Michela Massi - Own work, CC BY-SA 4.0, <https://commons.wikimedia.org/w/index.php?curid=80177333>

H. Guo et al. CounterNet: End-to-End Training of Counterfactual Aware Predictions. ICML Workshop on Algorithmic Recourse, <https://arxiv.org/pdf/2109.07557.pdf>, 2021.

Self-explainable methods

CounterNet learns to predict and explain with *counterfactual* instances

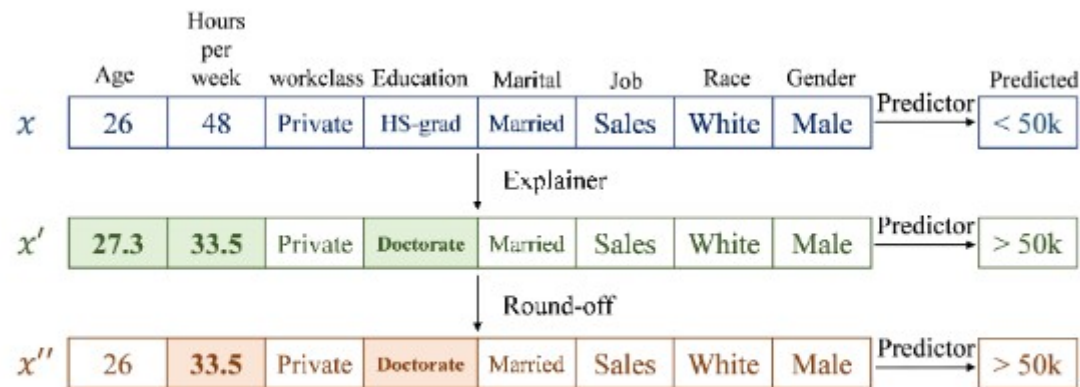


Figure 5: A counterfactual explanation generated by CounterNet.

Self-explainable methods

VCNet⁽⁻⁾ resorts to cVAEs^(*) to learn and explain via realistic counterfactual instances at once

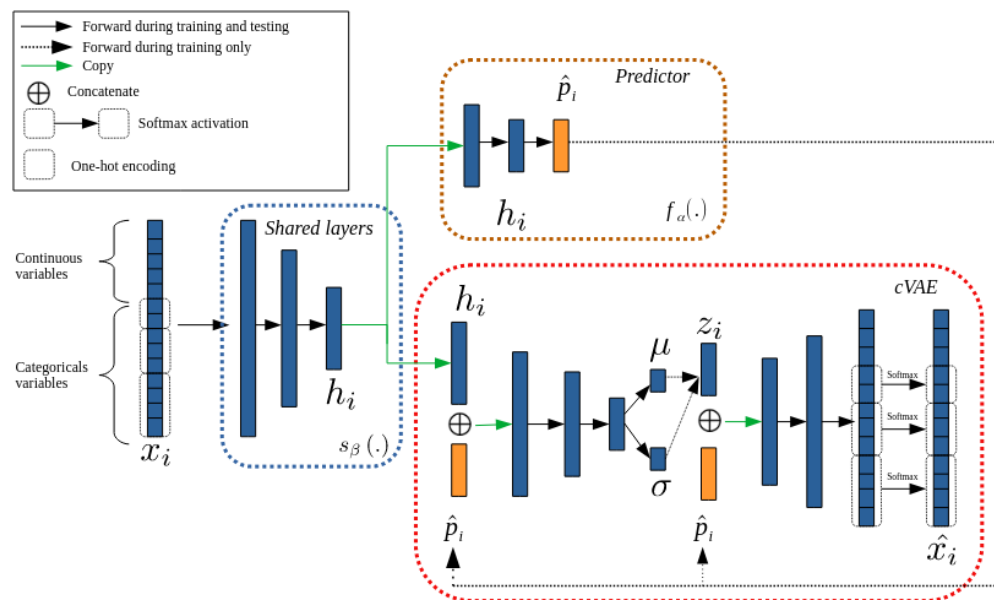


Fig. 1. VCNet architecture is composed of three blocks: Shared layers that transform the input into a latent representation (blue square), a predictor that outputs the prediction (brown square), and a conditional variational autoencoder that acts as a counterfactual generator during testing (red square).

(*) Conditional Variational Autoencoders

(-) 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.

Self-explainable methods

cVAEs capture the predictor's class distributions

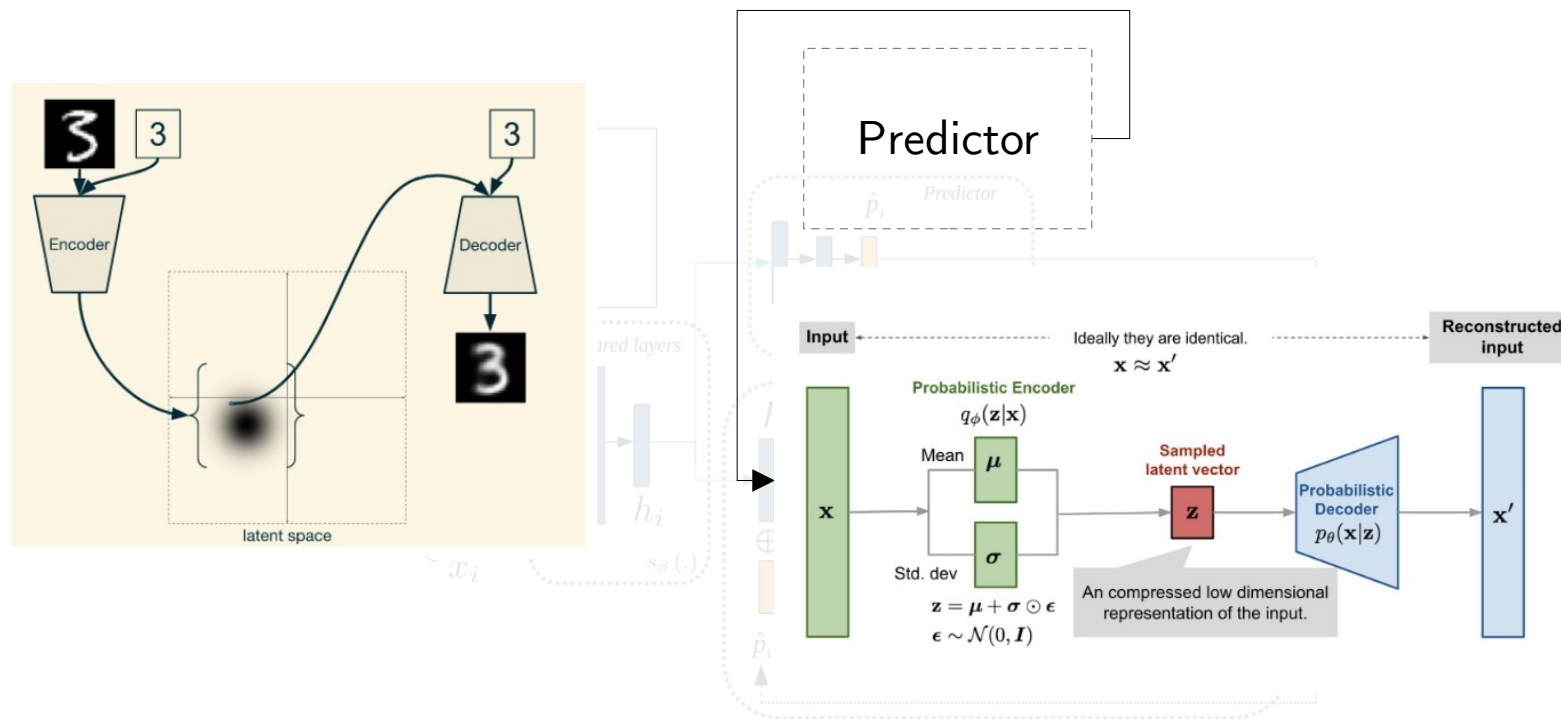


Fig. 1. VCNet architecture is composed of three blocks: Shared layers that transform the input into a latent representation (blue square), a predictor that outputs the prediction (brown square), and a conditional variational autoencoder that acts as a counterfactual generator during testing (red square).

Image from V. Guyomard's presentation for the HyAIAI project, https://project.inria.fr/hyaiai/files/2022/06/hyaiai_pres_victor.pdf

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.

Self-explainable methods

cVAEs capture the predictor's class distributions

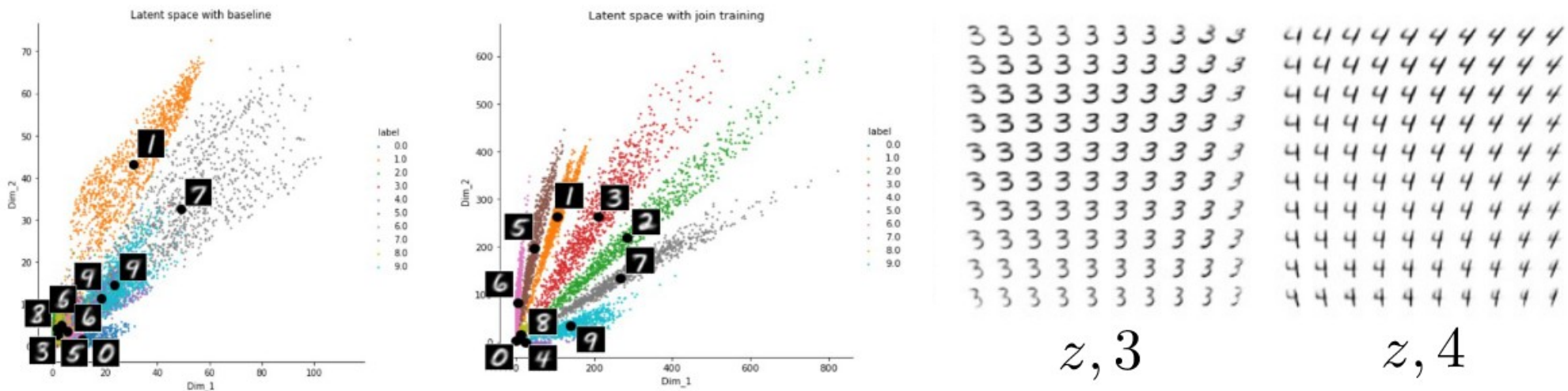


Image from V. Guyomard's presentation for the HyAIAI project, https://project.inria.fr/hyaiai/files/2022/06/hyaiai_pres_victor.pdf

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.

Self-explainable methods

For time series we can use *shapelets*

- They are representative segments that characterize a class; they serve as features for ML models

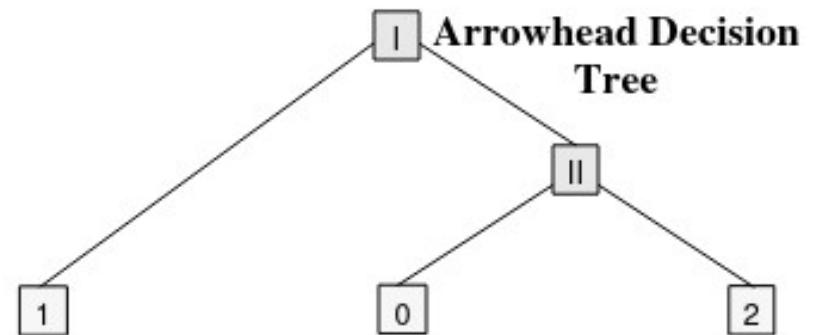
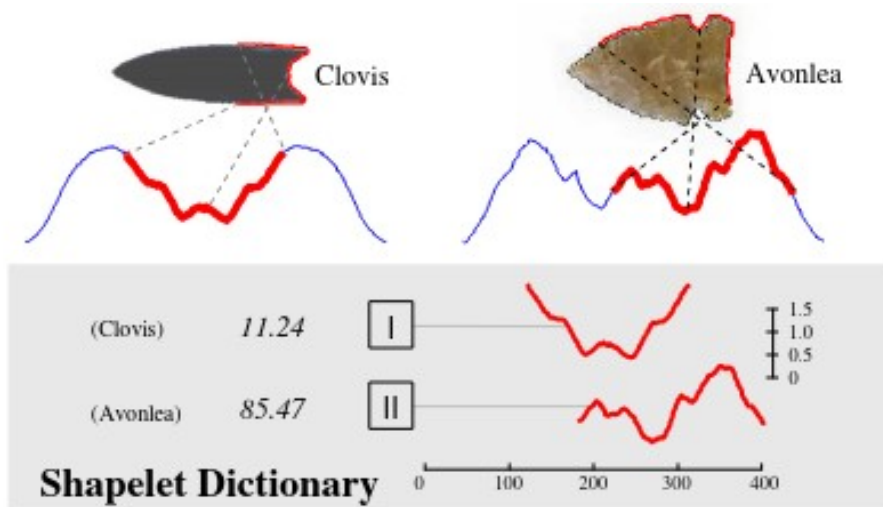


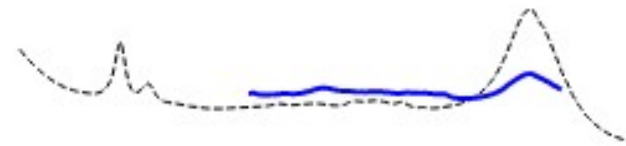
Figure 13: (top) The dictionary of shapelets, together with the thresholds d_{th} . (bottom) The decision tree for the 3-class projectile points problem

Self-explainable methods

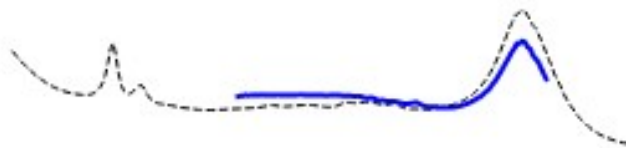
Subsequent approaches have focused on making *shapelets* more “realistic”



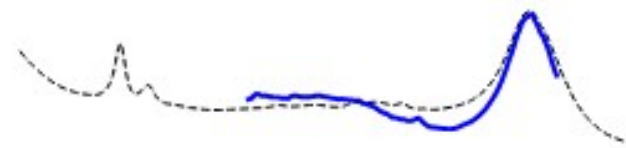
(b) Shapelet at epoch 20



(c) Shapelet at epoch 200



(e) Shapelet at epoch 800

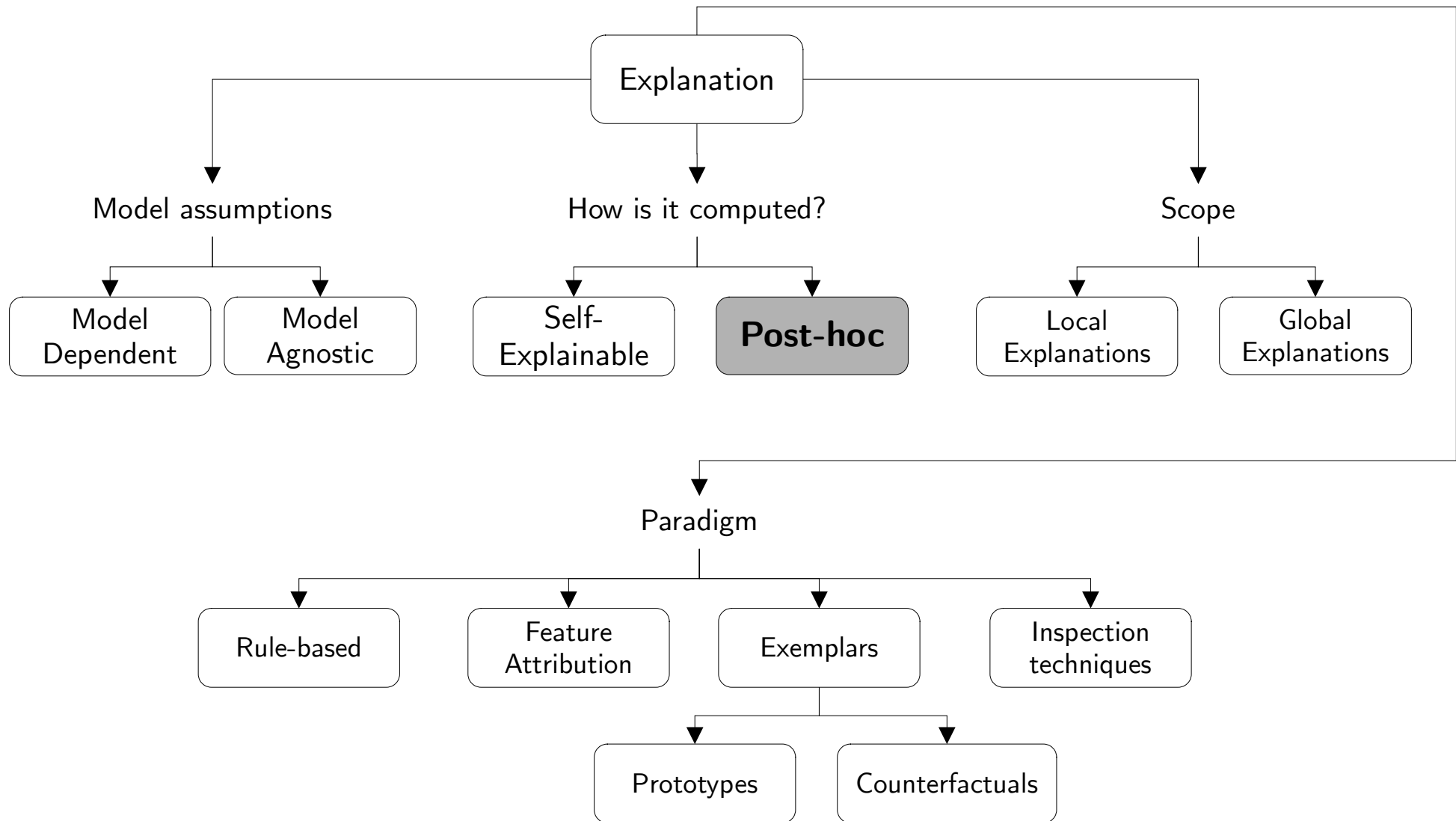


(f) Shapelet at epoch 8000

Agenda

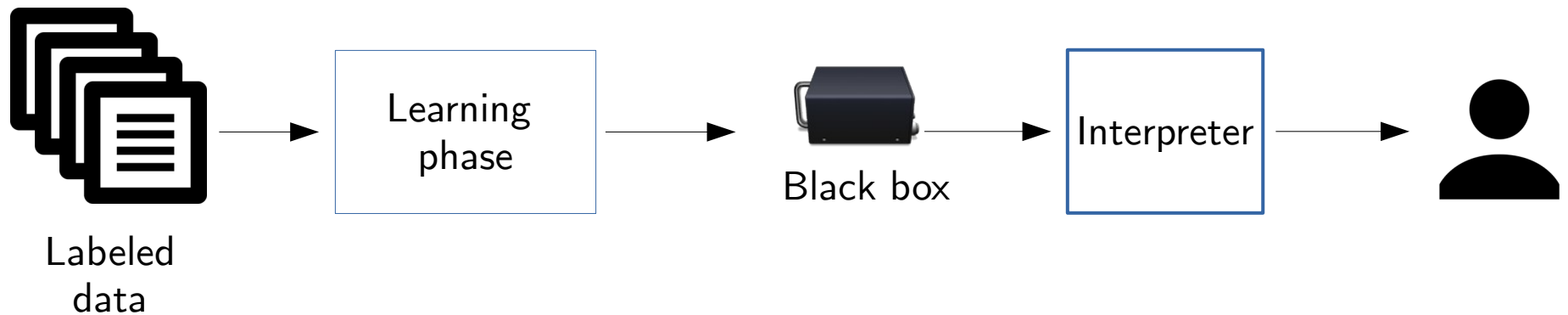
- Interpretable AI/ML: What and Why?
- Black-box vs. interpretable models
- **eXplainable AI techniques**
 - Explanation paradigms
 - Self-explainable methods
 - Post-hoc approaches
 - Evaluating XAI
- Conclusion & open research questions

Taxonomy of XAI Techniques



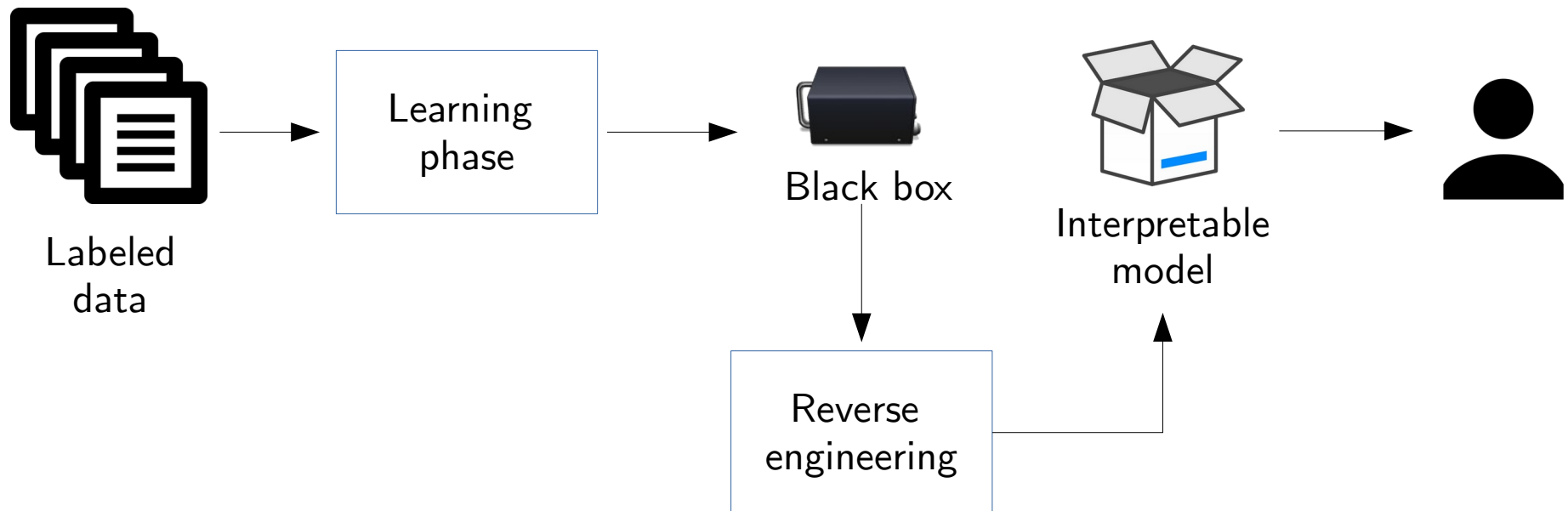
Post-hoc Explainability

Design an interpretation layer between the model and the human user



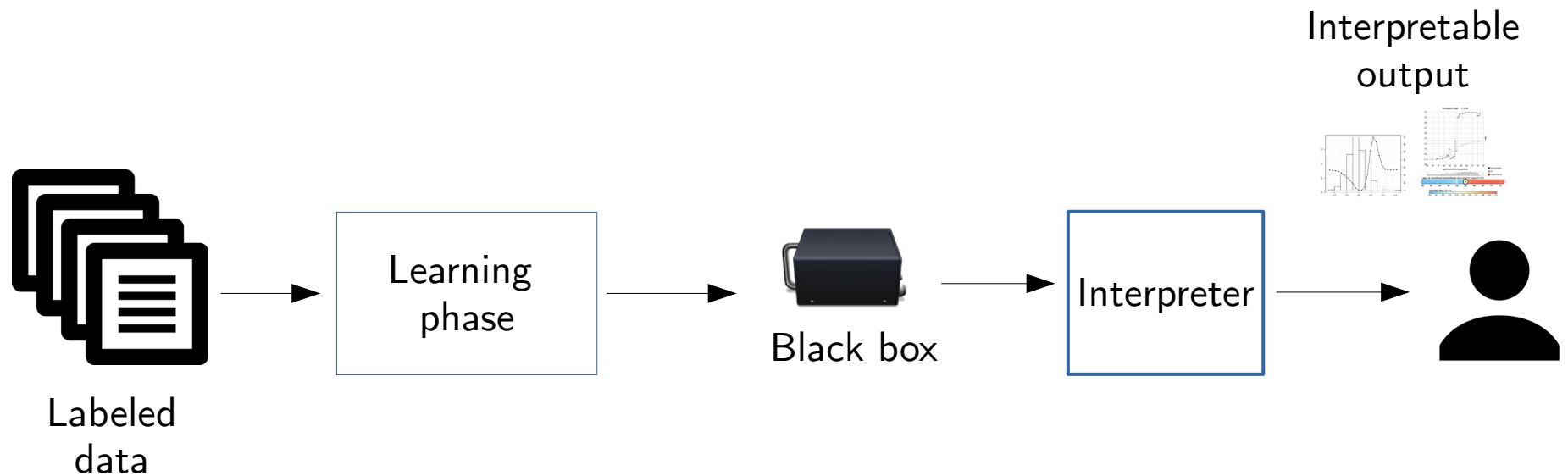
Post-hoc Explainability

Design an interpretation layer between the model and the human user

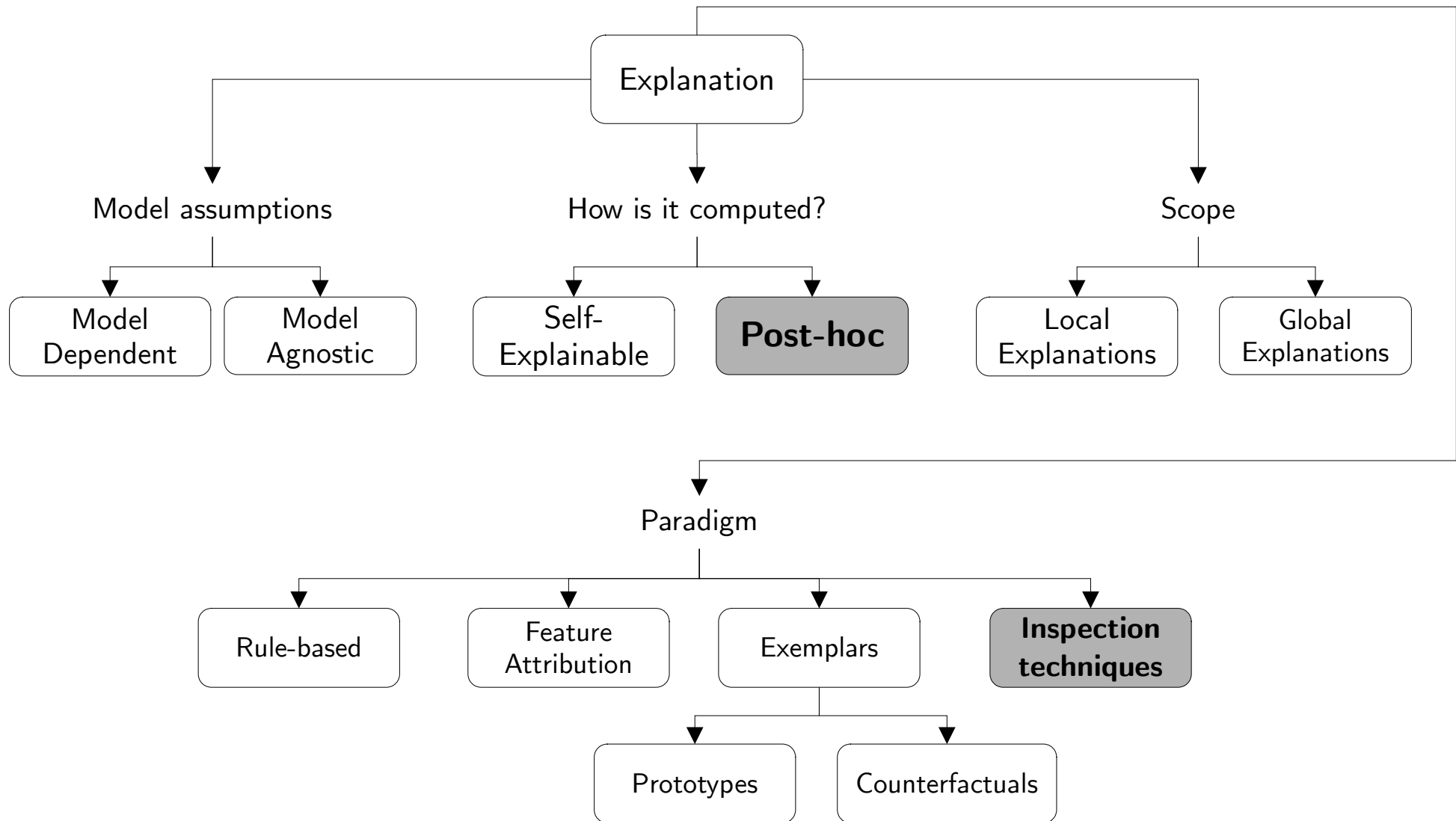


Post-hoc Explainability

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



Taxonomy of XAI Techniques

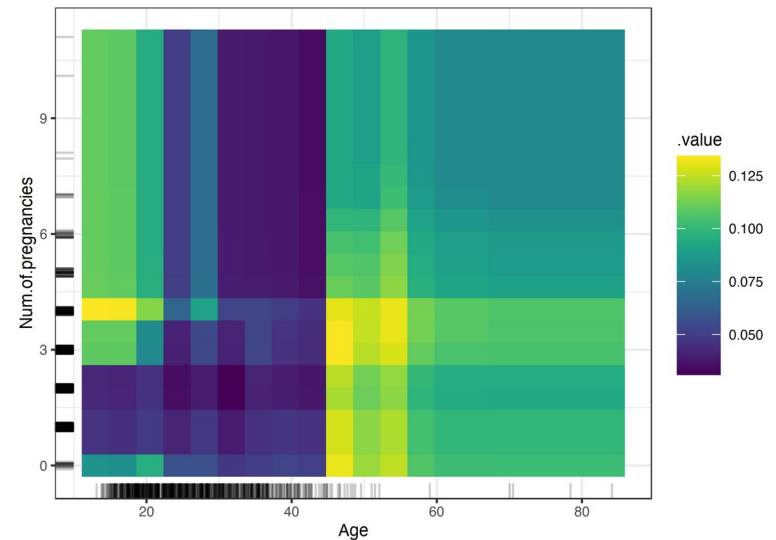
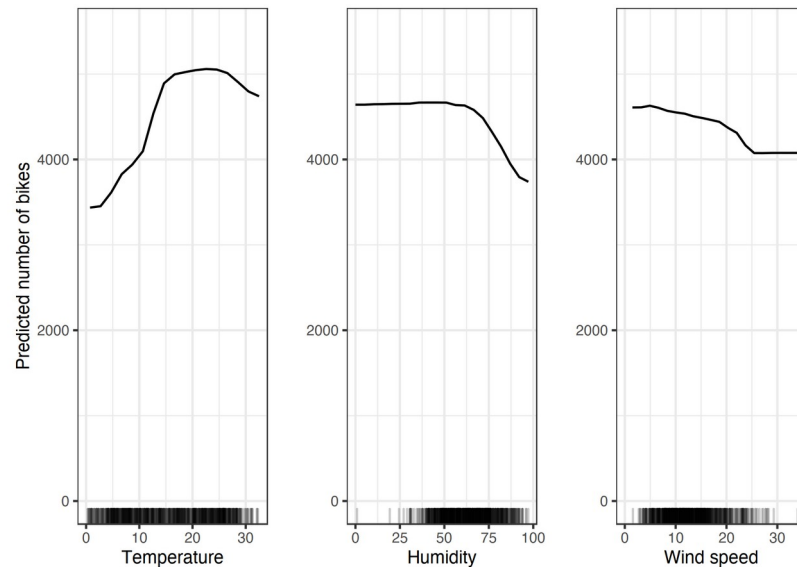


Inspecting the black box

Partial Dependence Plots (PDP) show the marginal effect of features on the black box's answers

$$\hat{f}_S(x_S) = E_{X_C} [\hat{f}(x_S, X_C)] = \int \hat{f}(x_S, X_C) d\mathbb{P}(X_C)$$

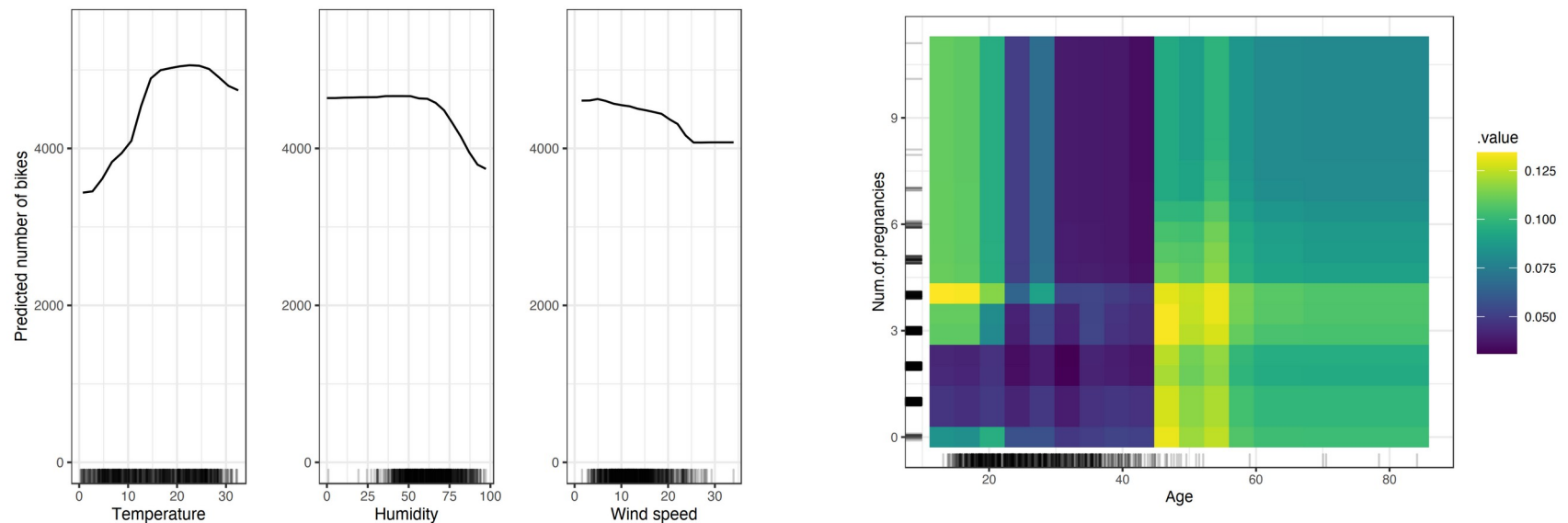
$$\hat{f}_S(x_S) = \frac{1}{n} \sum_{i=1}^n \hat{f}(x_S, x_C^{(i)})$$



Inspecting the black box

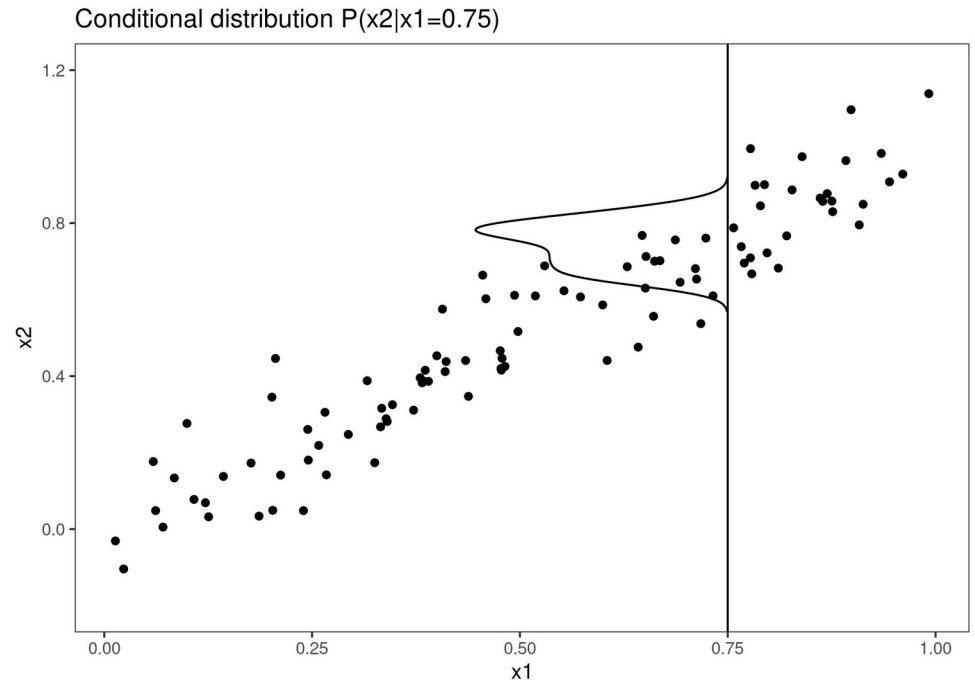
Partial Dependence Plots (PDP) show the marginal effect of features on the black box's answers

- Limitations: dimensionality, independence assumption



Inspecting the black box

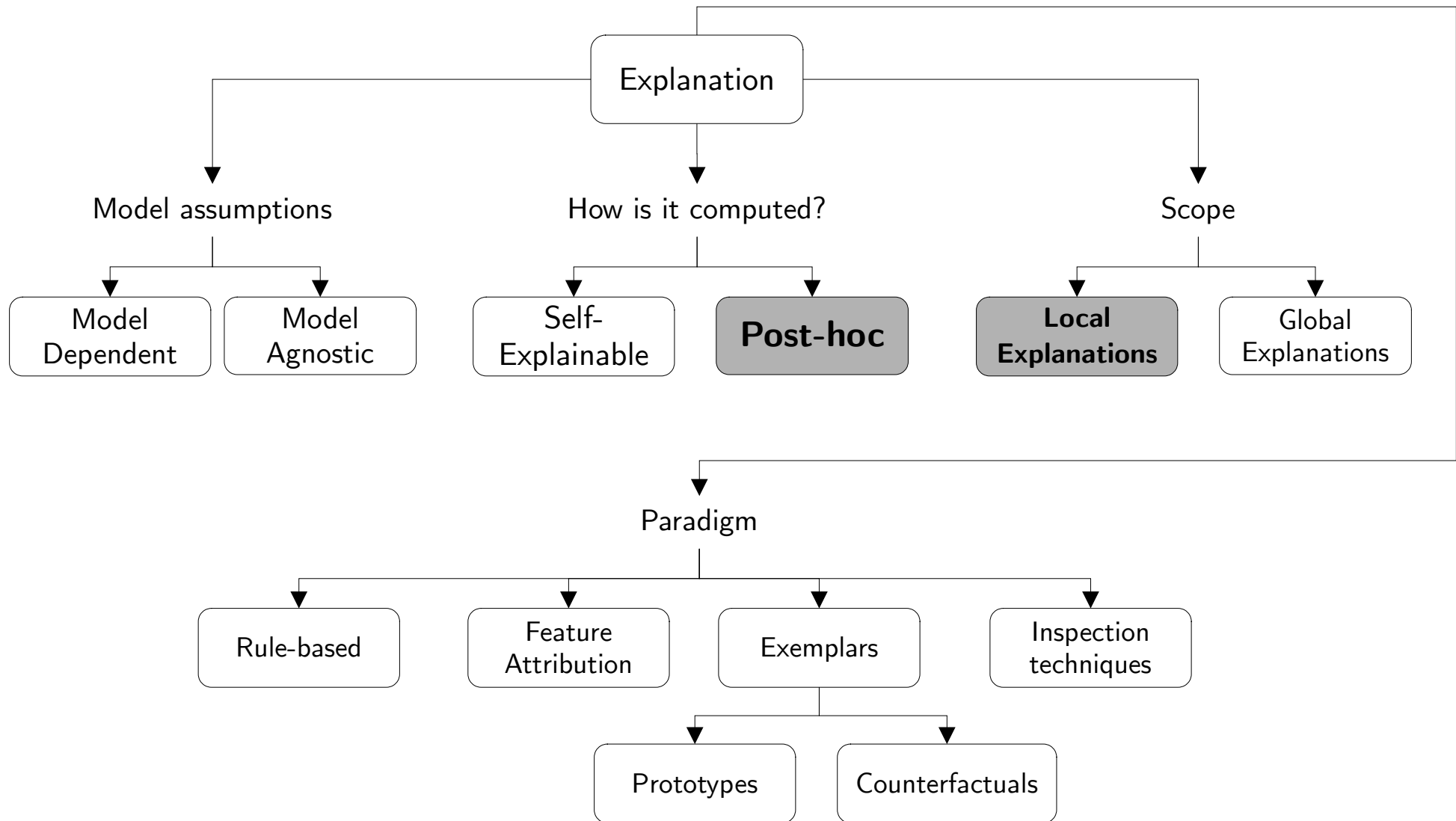
- Sensitivity Analysis(+)
- Individual Conditional Expectation (ICE) plots
- Accumulated Local Effects (ALE) Plots



(+) P. Cortez and M. J. Embrechts. Opening black box data mining models using sensitivity analysis. In Computational Intelligence and Data Mining (CIDM), 2011 IEEE Symposium on, pages 341-348. IEEE, 2011.

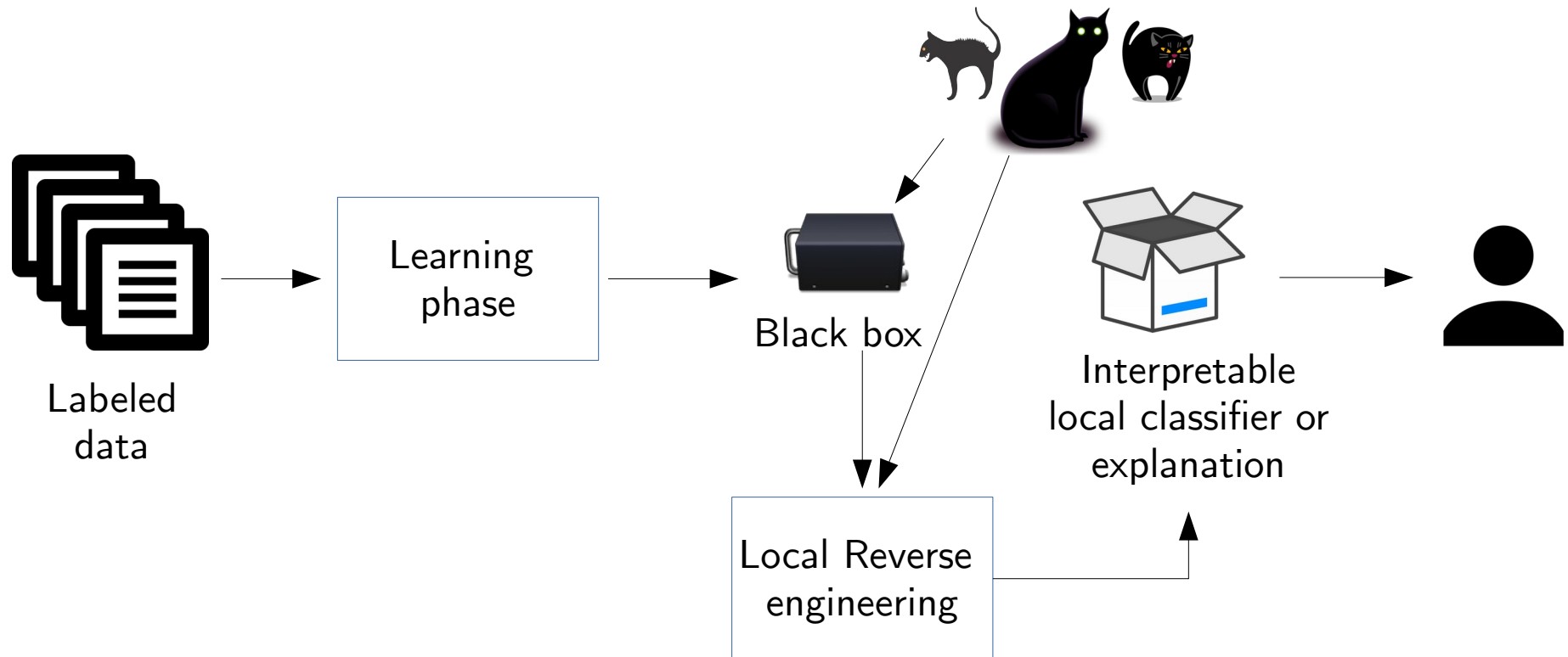
<https://christophm.github.io/interpretable-ml-book/ale.html>

Taxonomy of XAI Techniques



Local explainability

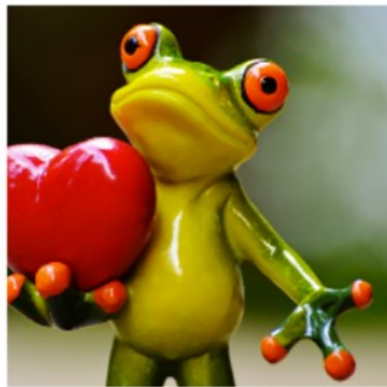
The surrogate model explains the black box in the vicinity of an individual instance.



Local Interpretable Model-Agnostic Explanations

LIME computes BB-agnostic *linear approximations*

- It maps instances to an **interpretable** space and samples around the target



Original Image



Interpretable Components

[1 1 1 1 1 ...]



[0 1 0 1 0 ...]



[0 0 0 0 1 ...]

Local Interpretable Model-Agnostic Explanations

LIME computes BB-agnostic *linear approximations*

- It maps instances to an **interpretable** space and samples around the target

Livre monument: Fabuleux livre, exhaustif, riche, documenté

1	1	1	1	1	1
livre	monument	fabuleux	exhaustif	riche	documenté
1	0	1	0	1	1
livre	monument	fabuleux	exhaustif	riche	documenté

Local Interpretable Model-Agnostic Explanations

LIME computes BB-agnostic *linear approximations*

- Interpretable space for time series: presence of absence of a segment
- Absence can be modeled in different ways:

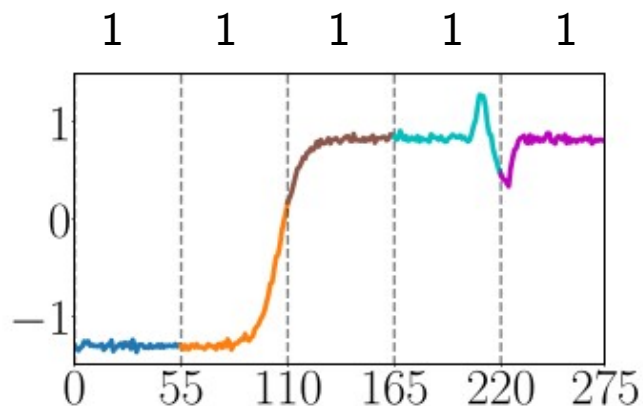
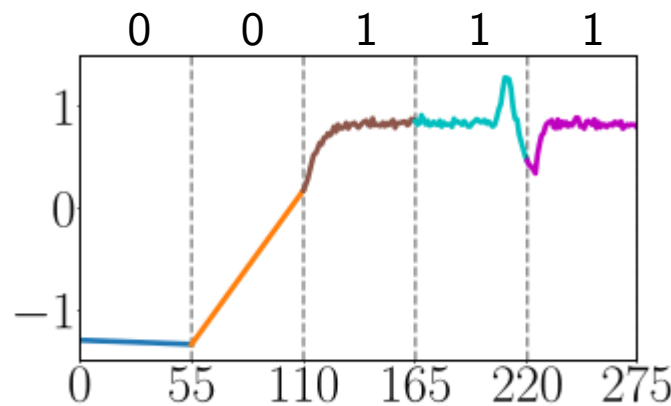
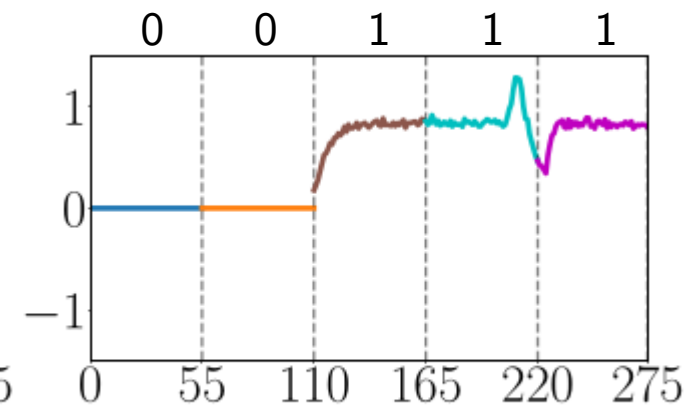


Fig. 3: Initial time series



(a) Linear interpolation

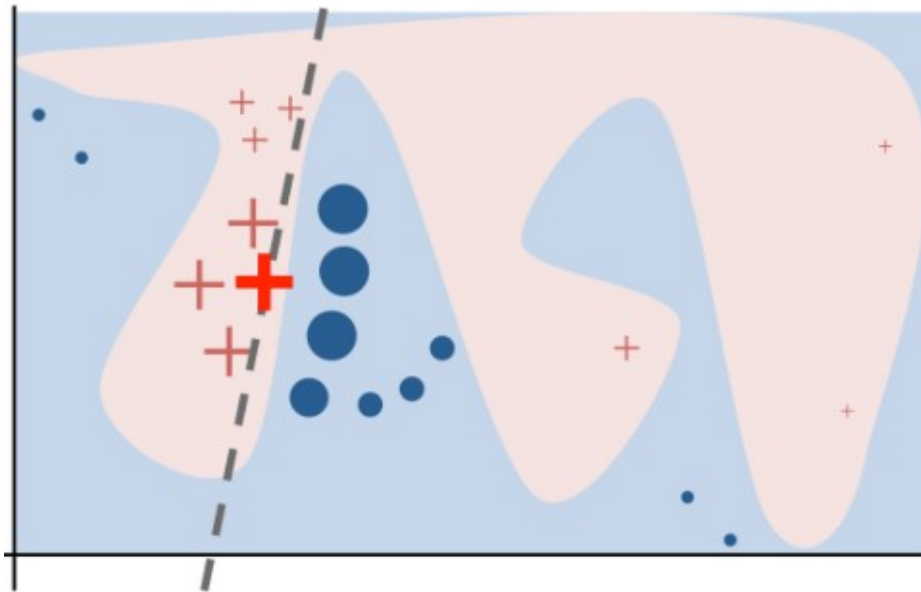


(b) Constant

Local Interpretable Model-Agnostic Explanations

LIME computes BB-agnostic *linear approximations*

- It then learns a linear surrogate from the neighborhood and their BB labels

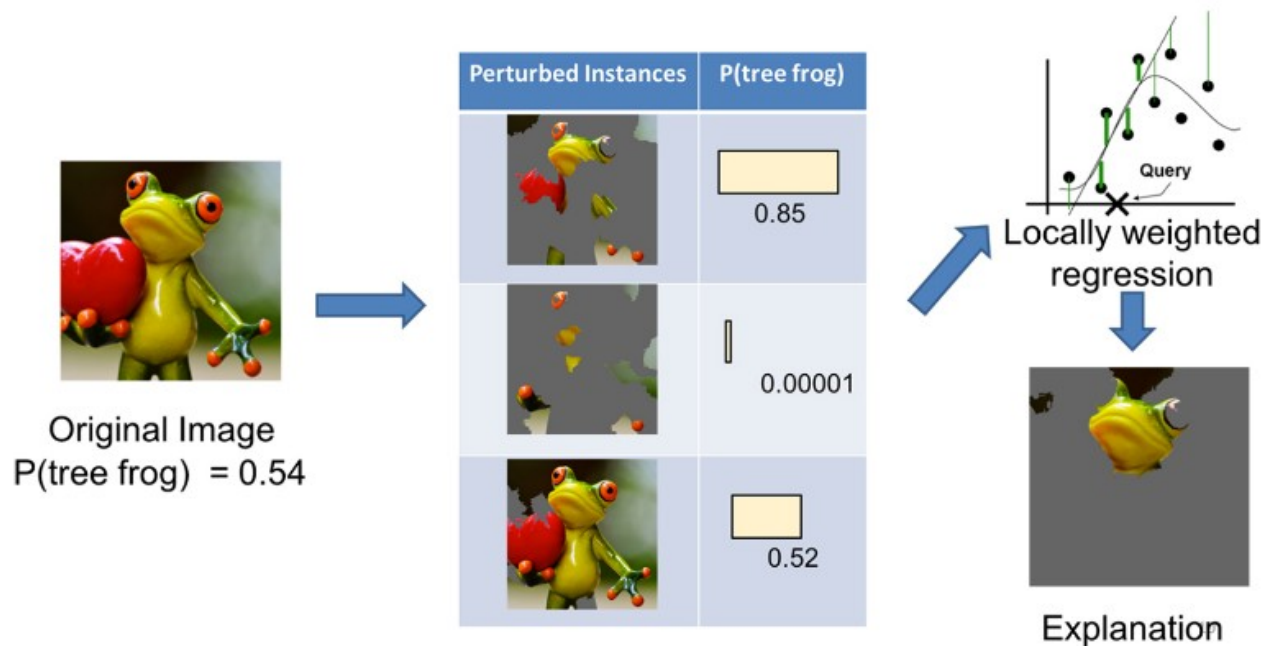


Neighbors are weighted by the distance to the target (exponential kernel)

Local Interpretable Model-Agnostic Explanations

LIME computes BB-agnostic *linear approximations*

- The coefficients of the linear function are a feature-attribution explanation on (the interpretable features)



Local Interpretable Model-Agnostic Explanations

LIME computes BB-agnostic *linear approximations*

- The coefficients of the linear function are a feature-attribution explanation on (the interpretable features)

Prediction probabilities



atheism



christian

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

Local explainability

Other similar feature-attribution methods are:

- Function gradients
- GradCAM
- DeepLIFT
- Integrated Gradients
- e-LRP
- SHAP
 - Kernel SHAP
 - Tree SHAP
 - Deep SHAP

$$\sum_{1 \leq i \leq \hat{d}} \hat{\alpha}_i \hat{x}[i] = \Delta f(x) = f(x) - f(x_\emptyset)$$

$$\alpha_g = \Delta x \times \int_0^1 \frac{\partial f(x_\emptyset + \gamma \Delta x)}{\partial \gamma} d\gamma$$

Local explainability

Other similar feature-attribution methods are:

- Function gradients
- GradCAM
- DeepLIFT
- Integrated Gradients
- e-LRP
- SHAP
 - Kernel SHAP
 - Tree SHAP
 - Deep SHAP

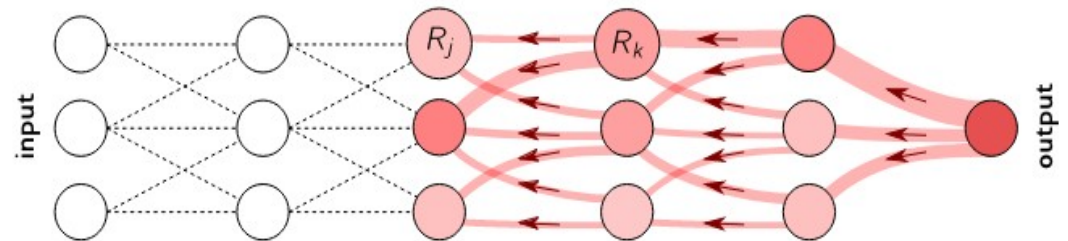
$$\sum_{1 \leq i \leq \hat{d}} \hat{\alpha}_i \hat{x}[i] = \Delta f(x) = f(x) - f(x_\emptyset)$$



Model dependent



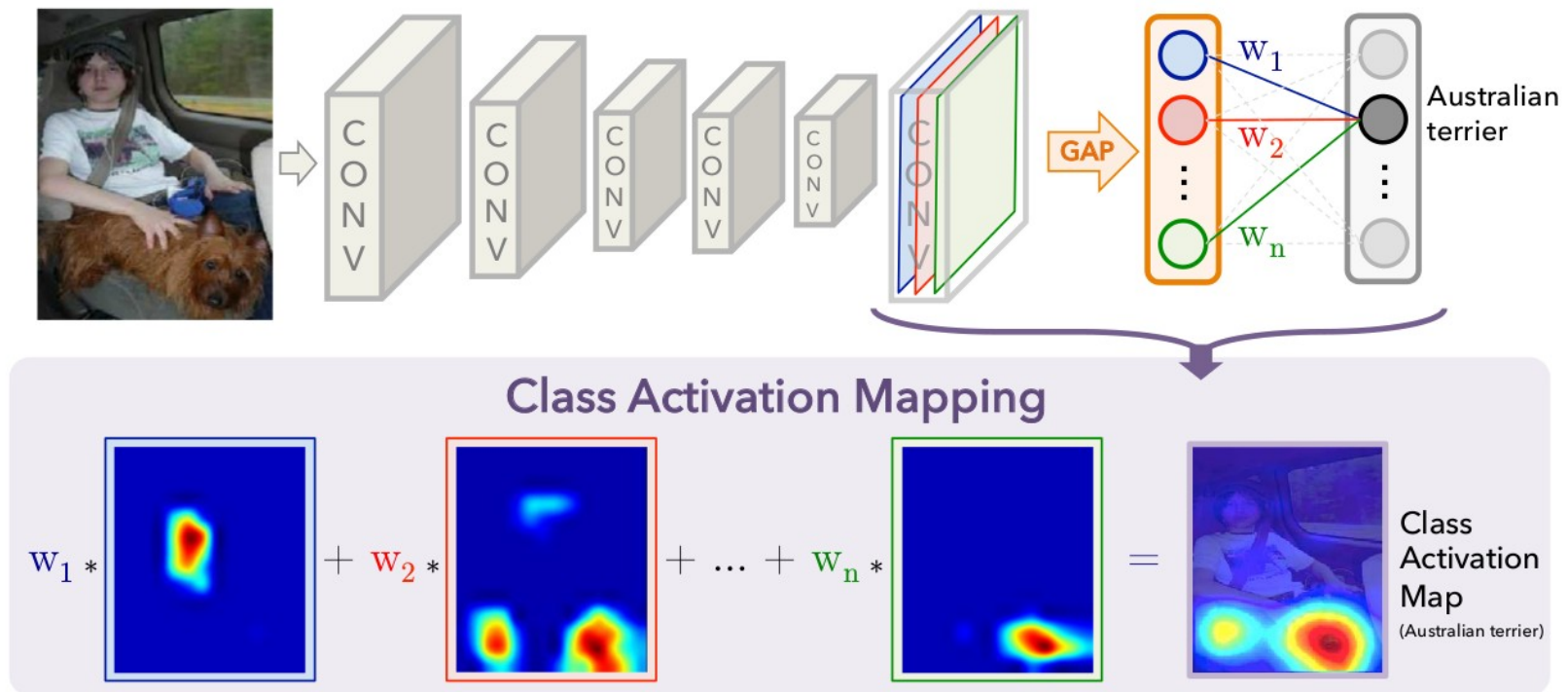
Based on back-propagation rules

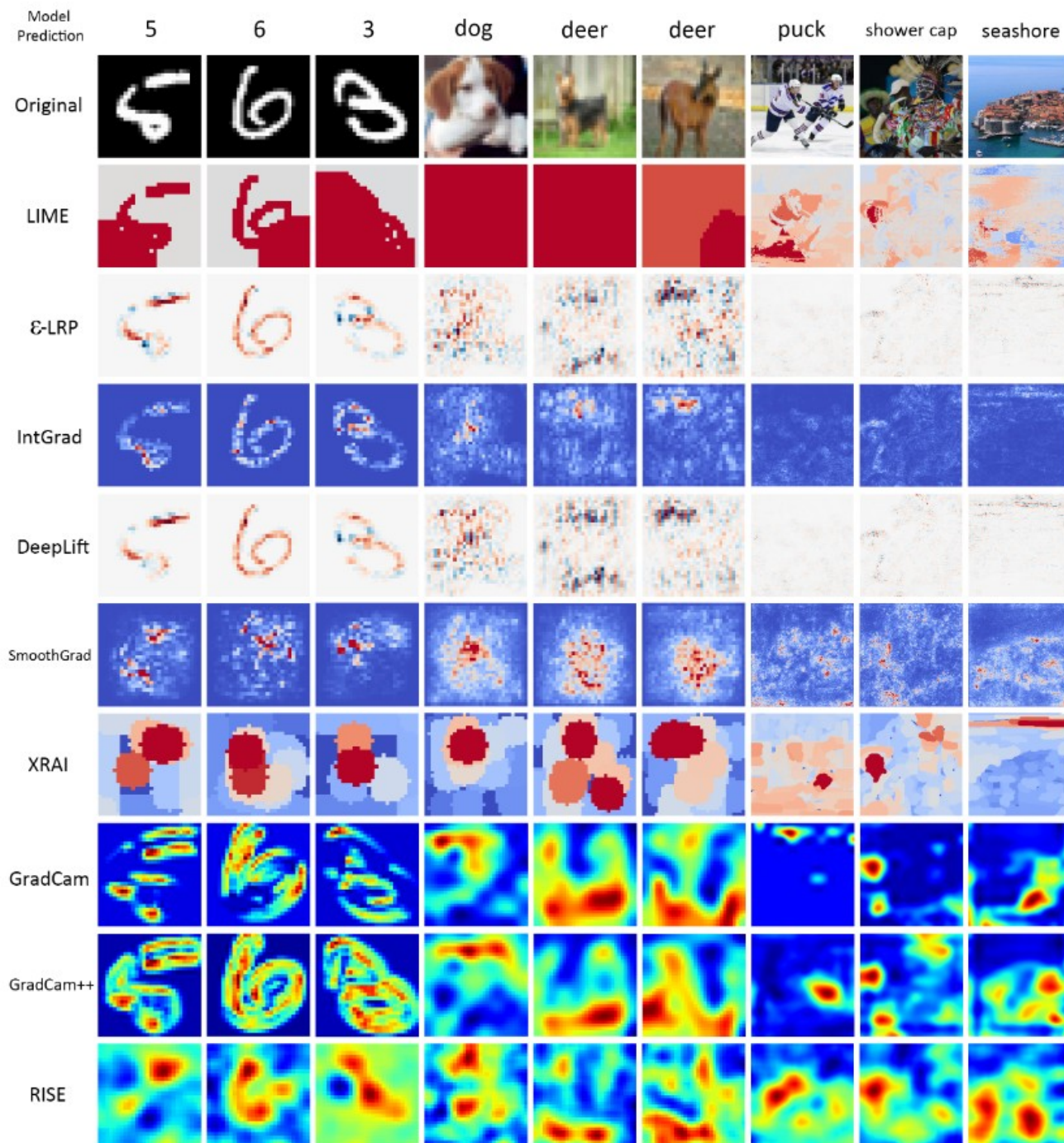


Montavon G. et al. Layer-Wise Relevance Propagation: An Overview
<https://iphome.hhi.de/samek/pdf/MonXAI19.pdf>

Feature attribution and heatmaps

GradCAM generates class activation maps from NNs used for image classification





SHapley Additive exPlanations

SHAP applies game theory to quantify importance

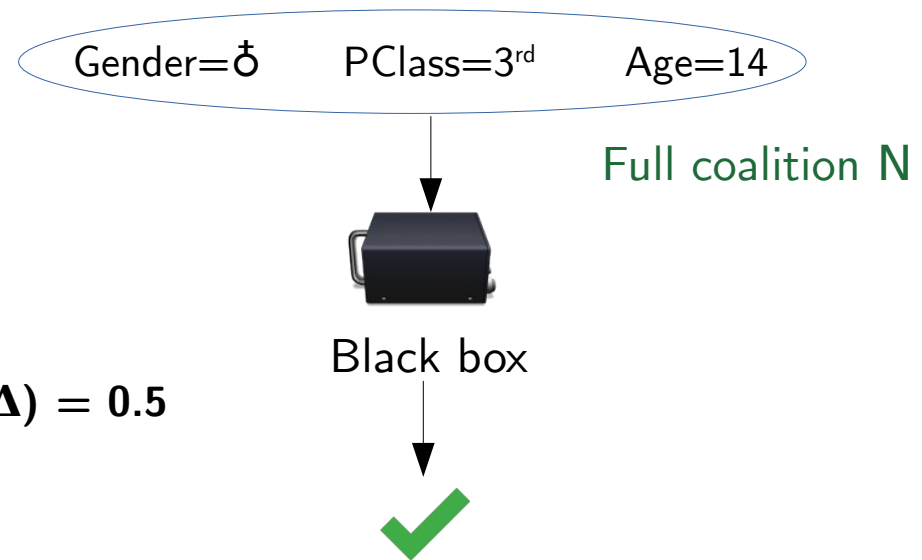
SHapley Additive exPlanations

SHAP applies game theory to quantify importance

- Shapley values for features: average marginal contribution on all possible feature coalitions

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)] .$$

Influence of Gender= \emptyset			
Coalition S	f(S)	f(S \cup {i})	Δ
\emptyset	✓	RIP	1
{PClass=3 rd }	RIP	RIP	0
{Age=14}	✓	✓	0
{Age=14, PClass=3 rd }	✓	RIP	1



SHapley Additive exPlanations

SHAP applies game theory to quantify importance

- Shapley values for features: average marginal contribution on all possible feature

ML models cannot deal with "incomplete" features
 Solution: absent features are set to a baseline value

$$f(x_{S \cup \{i\}}) - f_S(x_S)$$

Influence of Gender=♂

Coalition S	f(S)	f(S ∪ {i})	Δ
∅	✓	RIP	1
{PClass=3 rd }	RIP	RIP	0
{Age=14}	✓	✓	0
{Age=14, PClass=3 rd }	✓	RIP	1

Gender=♂ PClass=3rd Age=14

Full coalition N



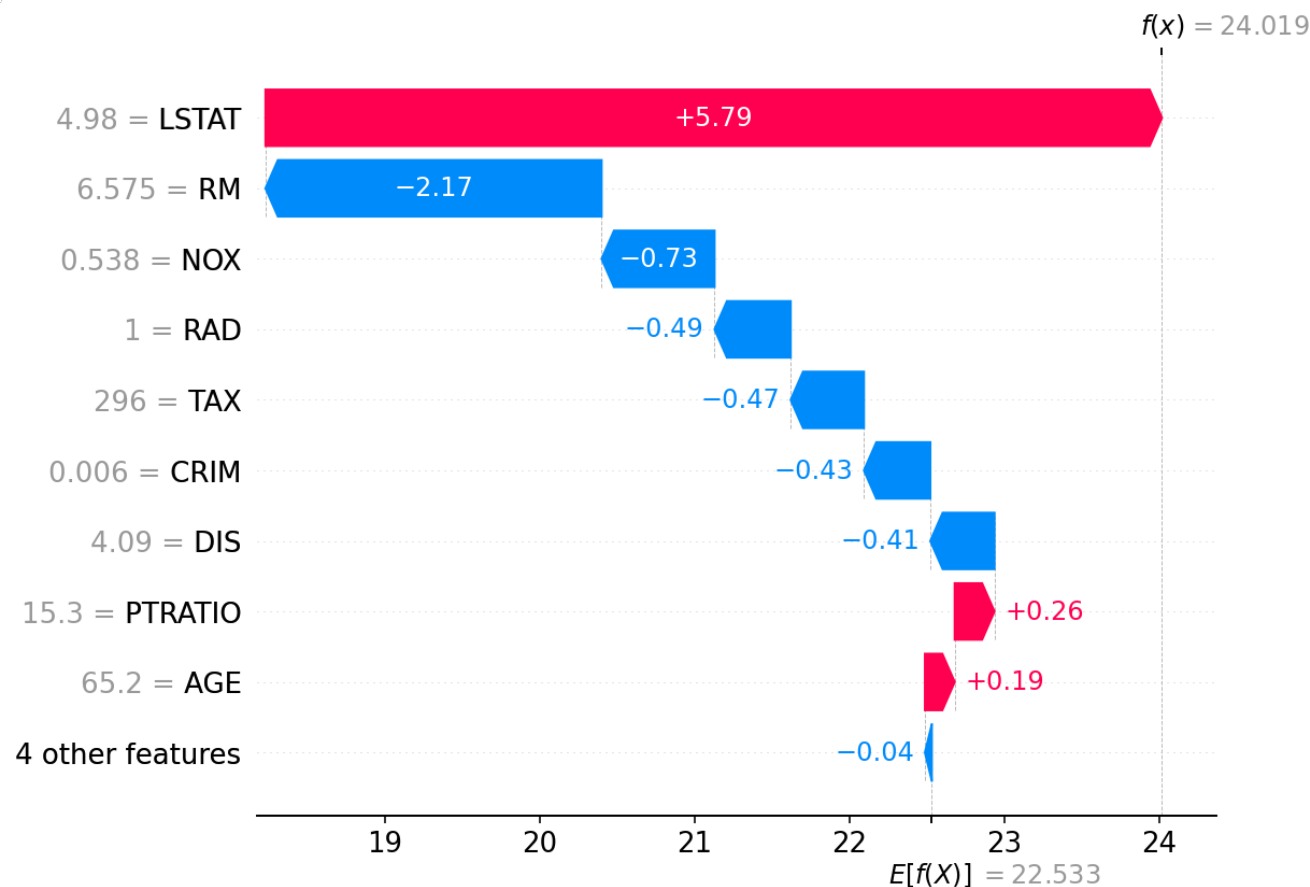
Black box



$$\text{avg}(\Delta) = 0.5$$

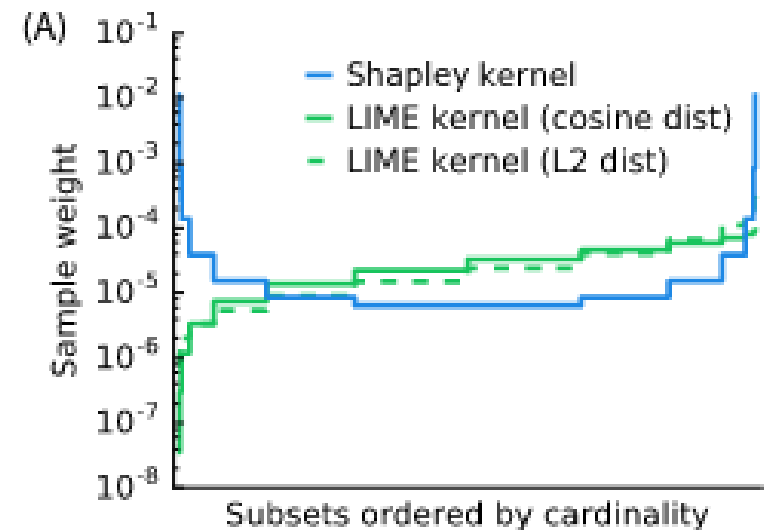
SHapley Additive exPlanations

SHAP applies game theory to quantify importance



SHapley Additive exPlanations

- SHAP feature attribution model guarantees:
 - Local accuracy
 - Linearity
 - Missingness
 - Null effects
- Variants
 - DeepShap, TreeShap
 - KernelShap (model-agnostic)
 - Sample coalitions
 - Weighted Linear Regression



Alternatives to SHAP

- Alternatives rely on fewer coalitions/assumptions
 - Equal Surplus
 - Extreme Feature Coalitions
 - Layer-1 SHAP(+)
 - Hamiache-Navarro values

$$\varphi_j^{ES}(\mathbf{x}, f_i) = f_i(\{j\}) + \frac{f_i(\mathcal{N}) - \sum_{k=1}^N f_i(\{k\})}{N}.$$

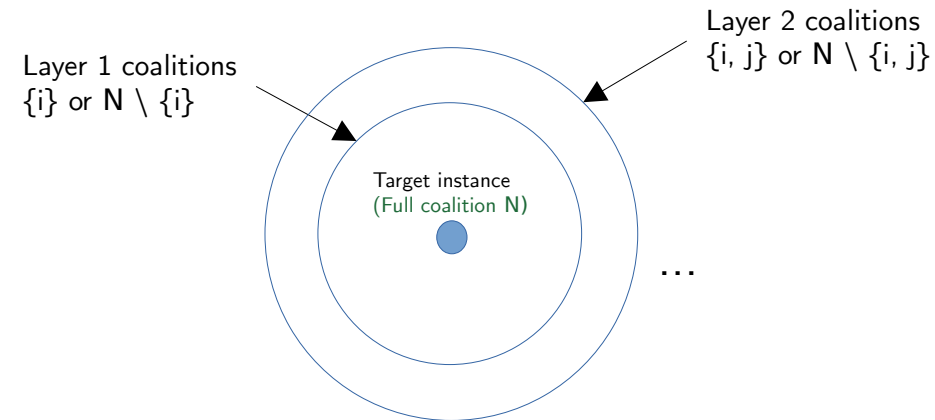
$$\varphi_j(\mathbf{x}, f_i) = w_i f_i(\{j\}) + (1 - w_i)(-f_i(\mathcal{N} \setminus \{j\})),$$

$$\phi_j = \bar{\phi}_j + \frac{1}{M} \left(f(N) - f(\emptyset) - \sum_{i=1}^M \bar{\phi}_i \right)$$

$$\text{where for any } i, \bar{\phi}_i = \frac{f(\{i\}) - f(\emptyset) + f(N) - f(N \setminus \{i\})}{2}.$$



Fig. 1. Extreme feature coalitions

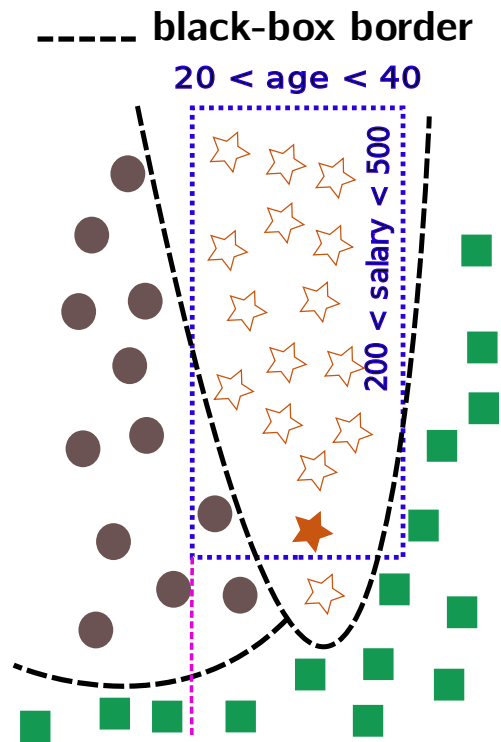


C. Condevaux et al. Fair and Efficient Alternatives to Shapley-based Attribution Methods. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, 2022.

(+) G. Kelodjou et al. Shaping Up SHAP: Enhancing Stability through Layer-Wise Neighbor Selection. AAAI Conference on Artificial Intelligence, 2024.

Anchors – Rule explanations

An anchor is a region of the feature space where a classifier behaves as with an instance of interest.



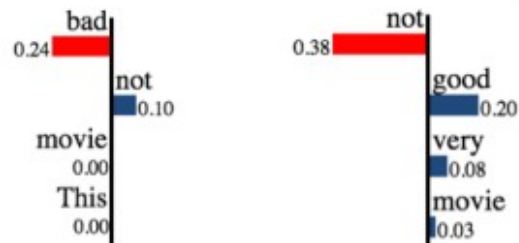
$$\text{age} \in (20, 40) \wedge \text{salary} \in (200, 500) \Rightarrow \star$$

Anchors – Rule explanations

Anchors generates “neighbors” in an interpretable space and learns rules in a top-down fashion

+ This movie is not bad. - This movie is not very good.

(a) Instances



(b) LIME explanations

{“not”, “bad”} → Positive {“not”, “good”} → Negative

(c) Anchor explanations



(a) Original image



(b) Anchor for “beagle”

Figure 1: Sentiment predictions, LSTM

Anchors – Rule explanations

LORE uses genetic algorithms to guarantee a more representative neighborhood

- It produces “friends” & “enemies” of the target instance
- Perturbation operators: cross-over and mutation

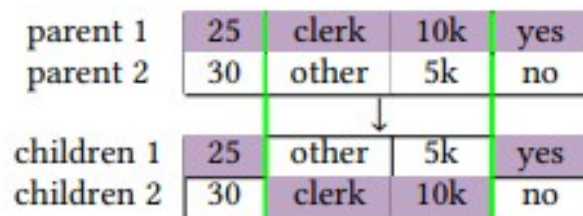


Figure 1: Crossover.

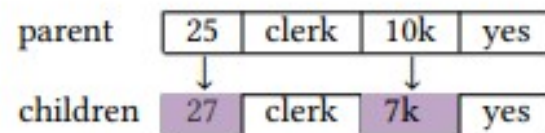


Figure 2: Mutation

Anchors – Rule explanations

LORE uses genetic algorithms to guarantee a more representative neighborhood

- Explanations take the form of decision trees
- The trees also encode **counterfactual explanations**

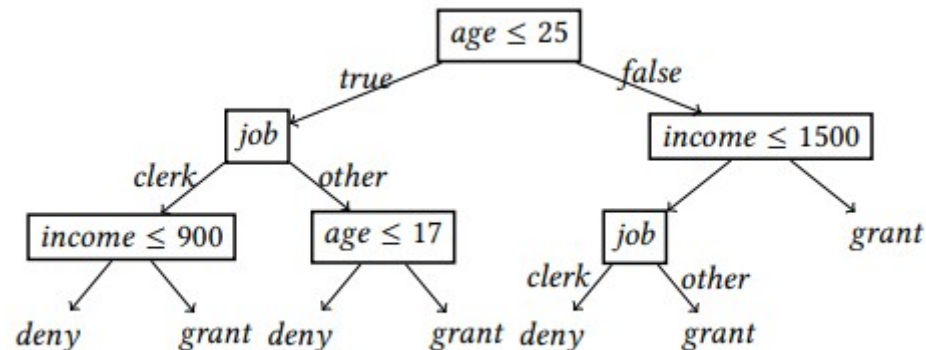


Figure 4: Example decision tree.

Counterfactual explanations

What do I need to change in the input to change the model's output?

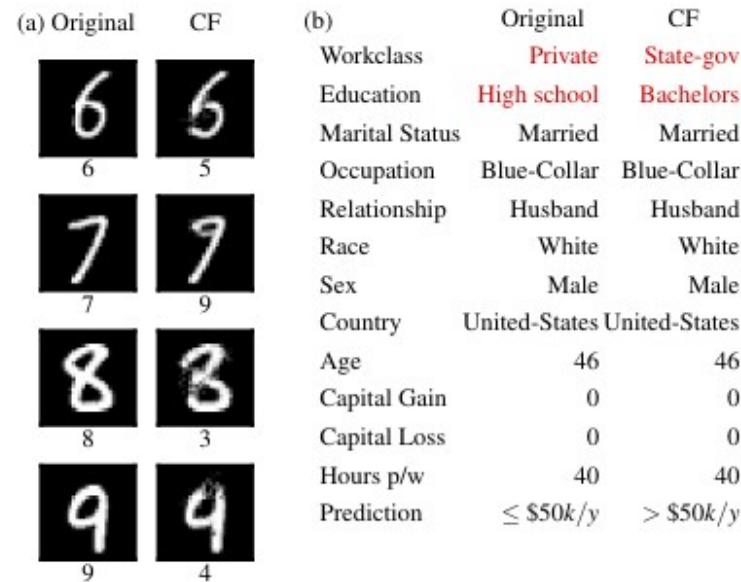


Figure 1. (a) Examples of original and counterfactual instances on the MNIST dataset along with predictions of a CNN model. (b) A counterfactual instance on the Adult (Census) dataset highlighting the feature changes required to alter the prediction of an NN model.

Counterfactual explanations

Learning counterfactual explanations involves a trade-off between sparsity & distribution coherence

- Looveren and Klaise (2021) enforce resemblance to prototypes in a latent space (defined via an auto-encoder)

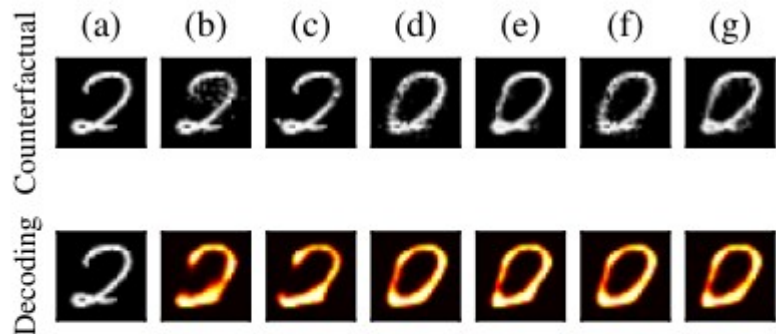


Figure 4. (a) Shows the original instance, (b) to (g) on the first row illustrate counterfactuals generated by using loss functions A to F . (b) to (g) on the second row show the reconstructed counterfactuals using AE .

$$j = \underset{i \neq t_0}{\operatorname{argmin}} \|\operatorname{ENC}(x_0) - \operatorname{proto}_i\|_2. \quad (8)$$

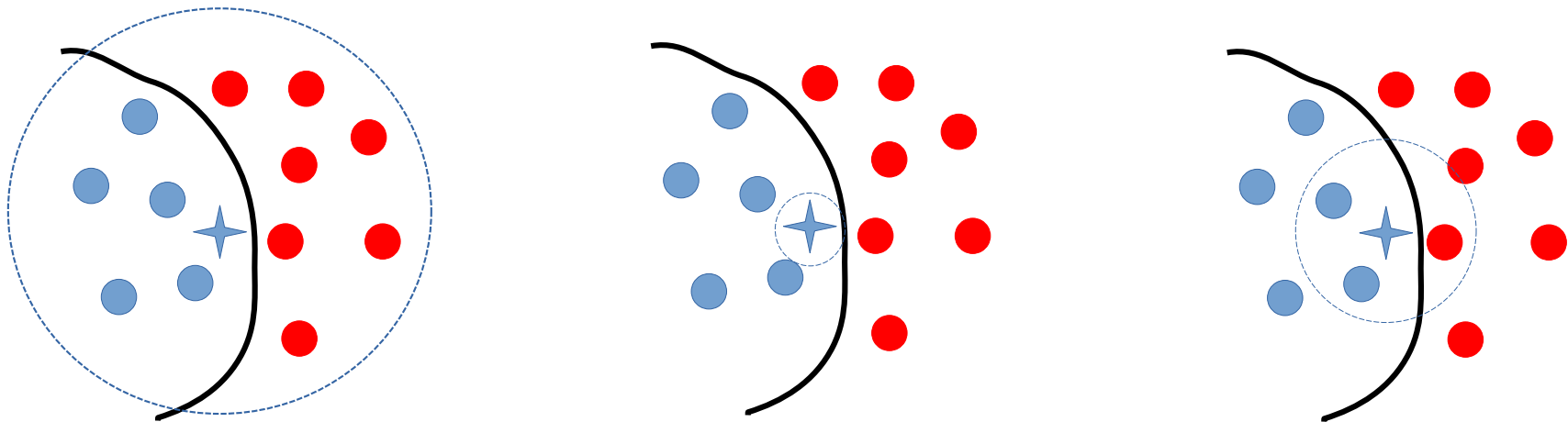
The prototype loss L_{proto} can now be defined as:

$$L_{\text{proto}} = \theta \cdot \|\operatorname{ENC}(x_0 + \delta) - \operatorname{proto}_j\|_2^2, \quad (9)$$

Counterfactual explanations

Growing Spheres: two-step search in a hyper-sphere around the target instance

- Start with a large radius and **contract** until no counterfactuals are covered
- **Expand** until the decision boundary is traversed



Counterfactual explanations

Growing Spheres: two-step search in a hyper-sphere around the target instance

- It minimizes both distance and sparseness for counterfactuals

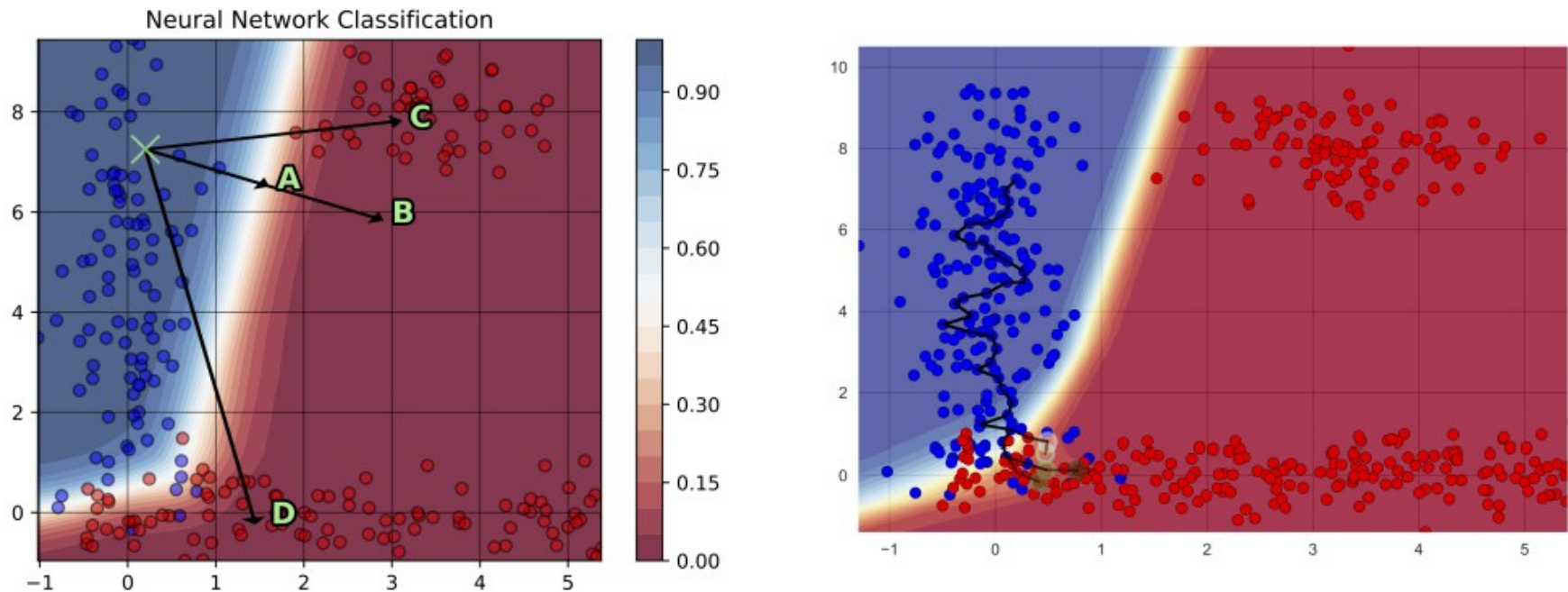
$$e^* = \arg \min_{e \in \mathcal{X}} \{c(x, e) \mid f(e) \neq f(x)\}$$

$$c(x, e) = \|x - e\|_2 + \gamma \|x - e\|_0$$

Counterfactual explanations

FACE: feasible and actionable counterfactuals

- It avoids counterfactuals in low-density regions



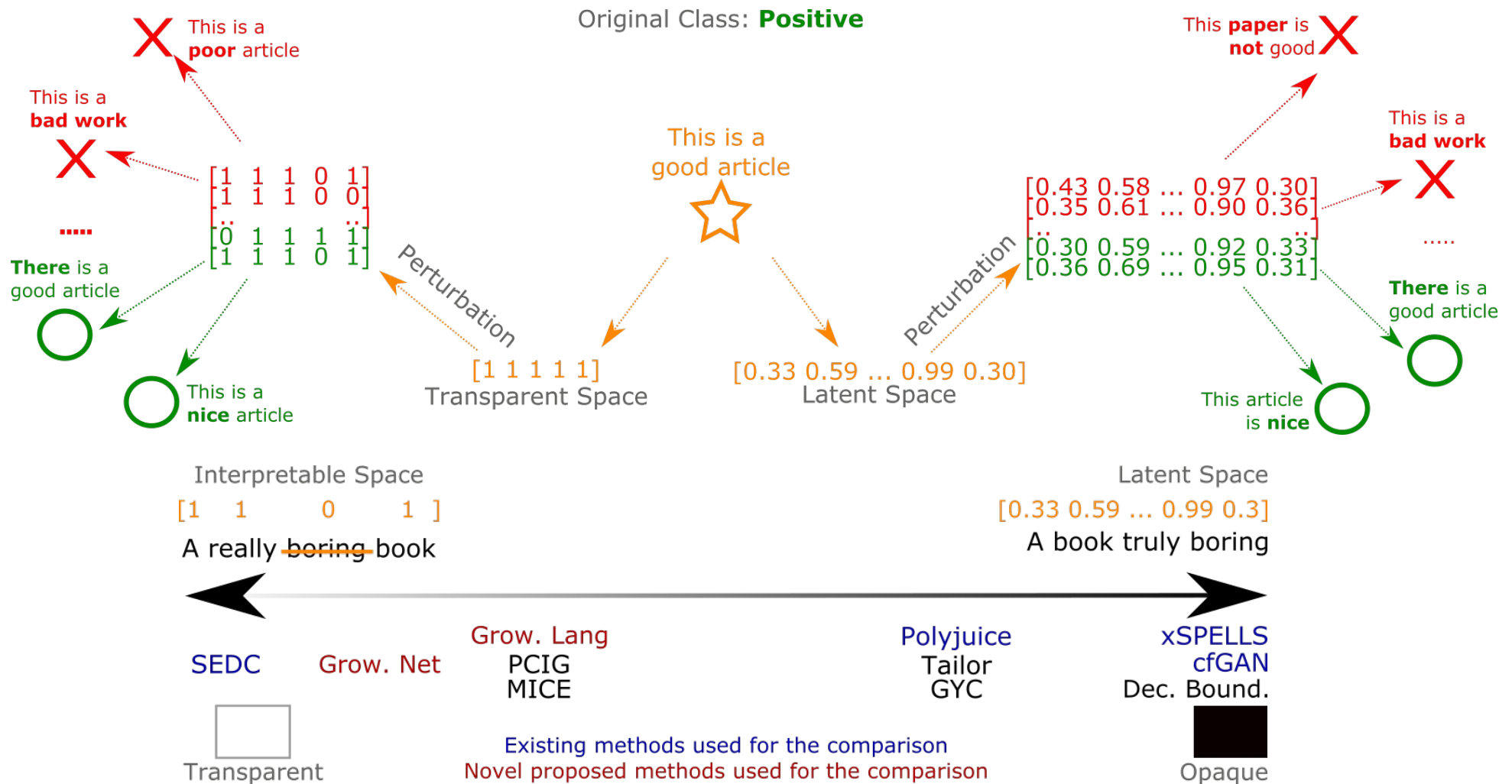
Counterfactual explanations

DICE selects counterfactuals based on the criteria of diversity and sparsity

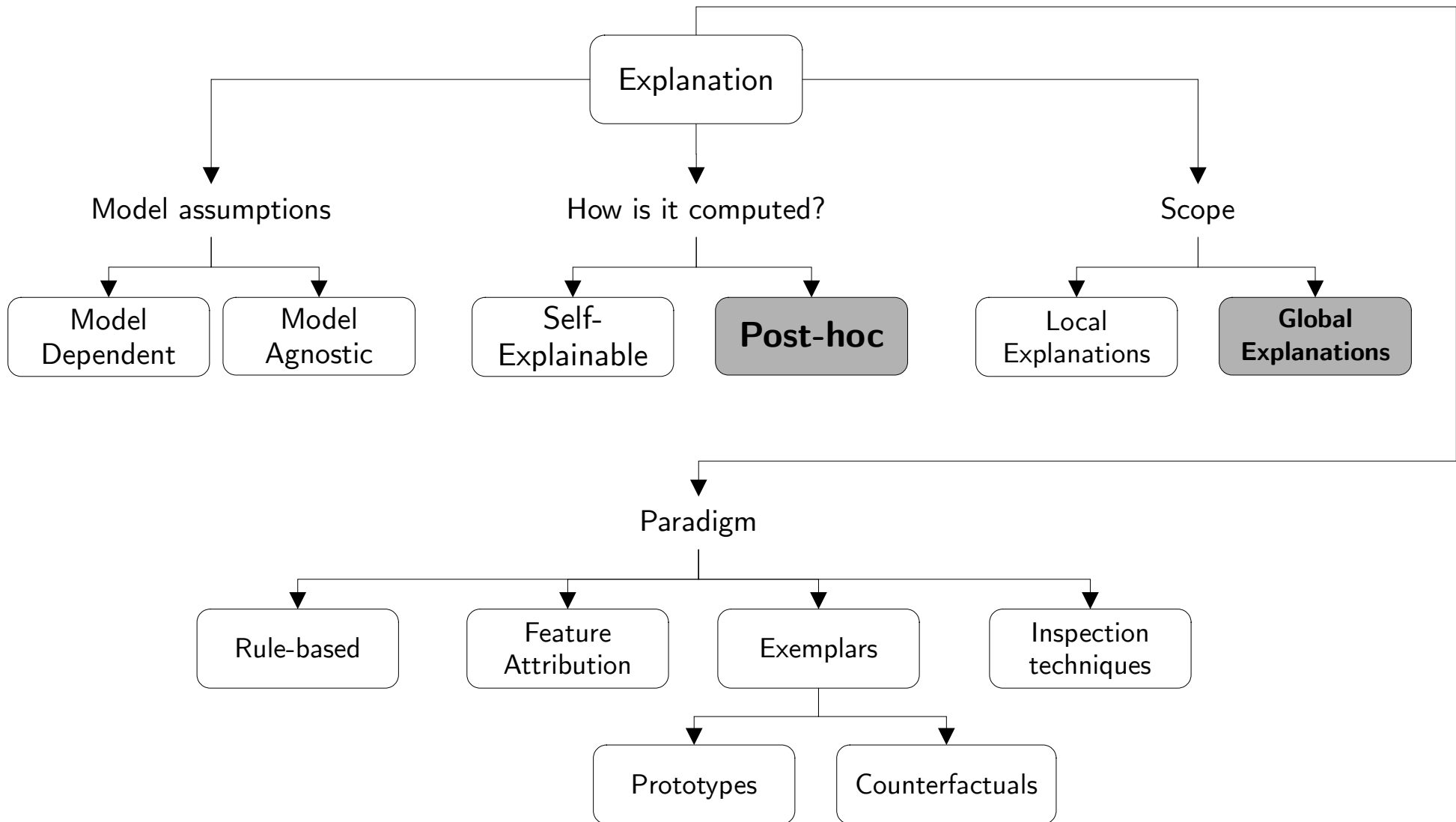
- It can also encode user constraints, e.g., children do not apply for credits, person's height is likely immutable
- Designed for tabular data

$$\mathbf{C}(\mathbf{x}) = \arg \min_{\mathbf{c}_1, \dots, \mathbf{c}_k} \frac{1}{k} \sum_{i=1}^k \text{yloss}(f(\mathbf{c}_i), y) + \frac{\lambda_1}{k} \sum_{i=1}^k \text{dist}(\mathbf{c}_i, \mathbf{x}) - \lambda_2 \text{dpp_diversity}(\mathbf{c}_1, \dots, \mathbf{c}_k)$$

Counterfactuals for text



Taxonomy of XAI Techniques



Global explainability

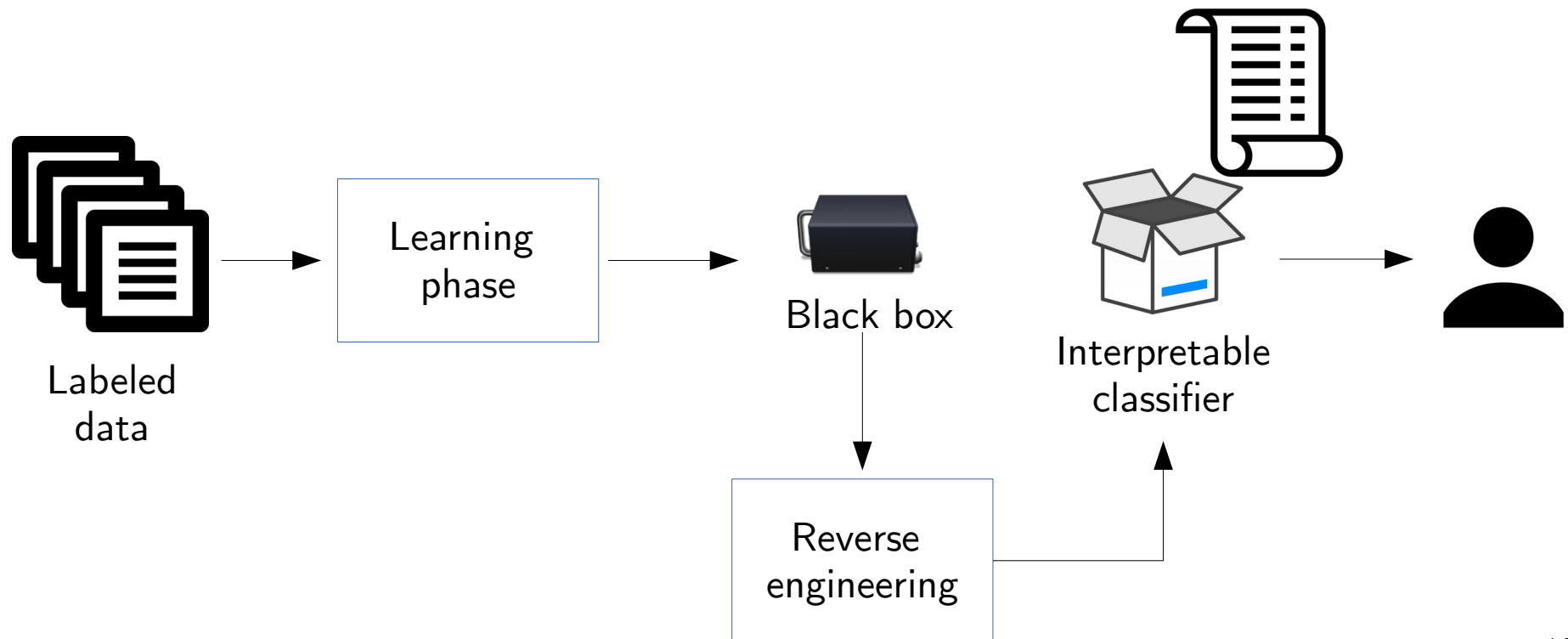
- We can generate global explanations by combining local explanations from many instances
 - Common for feature-attribution explanations
 - Ensemble tree-based models offer global feature importance scores based on
 - Impurity decrease
 - Accuracy drop



Figure 5: Toy example W . Rows represent instances (documents) and columns represent features (words). Feature f2 (dotted blue) has the highest importance. Rows 2 and 5 (in red) would be selected by the pick procedure, covering all but feature f1.

Global explainability

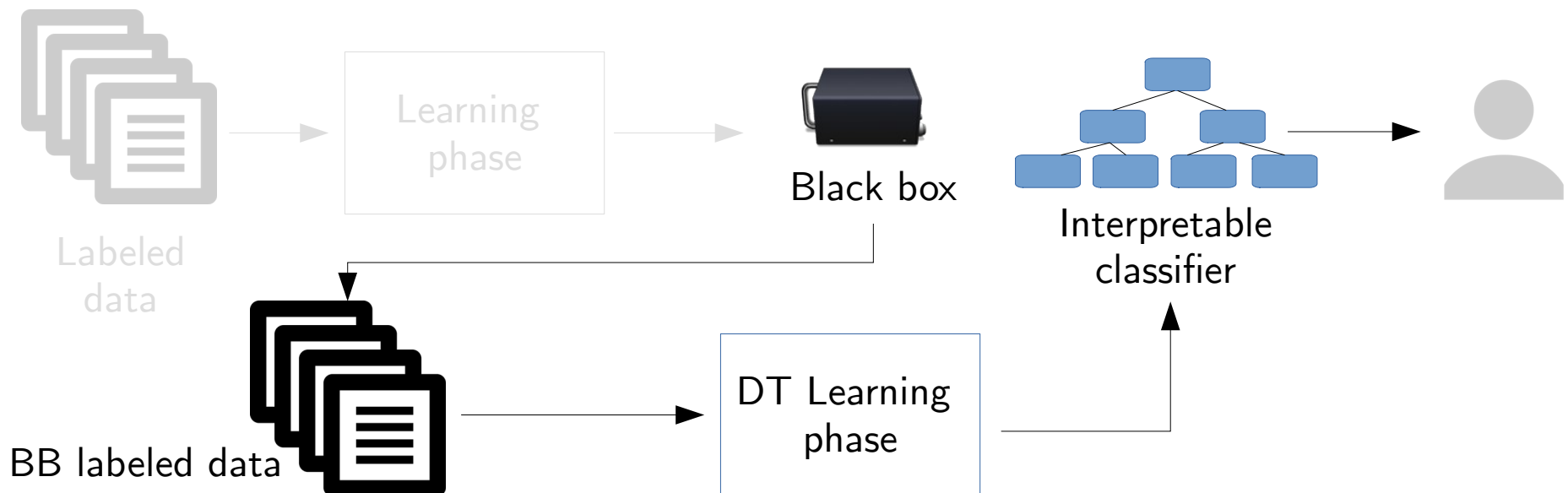
The surrogate model provides explanations for all possible outcomes of the black box



Global explainability

Global explanations for NNs date back to the 90s

- Trepan^(*) approximates the black-box with a D. tree



(*) M. Craven and J. W. Shavlik. Extracting tree-structured representations of trained networks. In Advances in neural information processing systems, pages 24-30, 1996.

Global explainability

BETA⁽⁺⁾ applies itemset mining to learn if-then rules

- Rules are restricted to two levels
- If two contradictory rules apply to an example, the one with higher fidelity wins

If Age > 50 and Gender = Male Then

If Past-Depression = Yes and Insomnia = No and Melancholy = No ⇒ **Healthy**

If Past-Depression = Yes and Insomnia = No and Melancholy = Yes ⇒ **Depressed**

Global explainability

BETA⁽⁺⁾ applies itemset mining to learn if-then rules

- Conditions (gender=♀) obtained via pattern mining
- Rule selection formulated as an optimization problem

$$\arg \max_{\mathcal{R} \subseteq \mathcal{ND} \times \mathcal{DL} \times \mathcal{C}} \sum_{i=1}^5 \lambda_i f_i(\mathcal{R}) \quad (1)$$

$$\text{s.t. } \text{size}(\mathcal{R}) \leq \epsilon_1, \text{maxwidth}(\mathcal{R}) \leq \epsilon_2, \text{numdsets}(\mathcal{R}) \leq \epsilon_3 \quad (2)$$

$$f_1(\mathcal{R}) = \mathcal{P}_{max} - \text{numpreds}(\mathcal{R}), \text{ where } \mathcal{P}_{max} = 2 * \mathcal{W}_{max} * |\mathcal{ND}| * |\mathcal{DL}|$$

$$f_2(\mathcal{R}) = \mathcal{O}_{max} - \text{featureoverlap}(\mathcal{R}), \text{ where } \mathcal{O}_{max} = \mathcal{W}_{max} * |\mathcal{ND}| * |\mathcal{DL}|$$

$$f_3(\mathcal{R}) = \mathcal{O}'_{max} - \text{ruleoverlap}(\mathcal{R}), \text{ where } \mathcal{O}'_{max} = N \times (|\mathcal{ND}| * |\mathcal{DL}|)^2$$

$$f_4(\mathcal{R}) = \text{cover}(\mathcal{R})$$

$$f_5(\mathcal{R}) = \mathcal{F}_{max} - \text{disagreement}(\mathcal{R}), \text{ where } \mathcal{F}_{max} = N \times |\mathcal{ND}| * |\mathcal{DL}|$$

Global explainability

BETA⁽⁺⁾ applies itemset mining to learn if-then rules

- Conditions (gender=♀) obtained via pattern mining
- Rule selection formulated as an optimization problem

$$\arg \max_{\mathcal{R} \subseteq \mathcal{ND} \times \mathcal{DL} \times \mathcal{C}} \sum_{i=1}^5 \lambda_i f_i(\mathcal{R}) \quad (1)$$

$$\text{s.t. } \text{size}(\mathcal{R}) \leq \epsilon_1, \text{maxwidth}(\mathcal{R}) \leq \epsilon_2, \text{numdsets}(\mathcal{R}) \leq \epsilon_3 \quad (2)$$



This is
sorcery!!!

$$f_1(\mathcal{R}) = \mathcal{P}_{max} - \text{numpreds}(\mathcal{R}), \text{ where } \mathcal{P}_{max} = 2 * \mathcal{W}_{max} * |\mathcal{ND}| * |\mathcal{DL}|$$

$$f_2(\mathcal{R}) = \mathcal{O}_{max} - \text{featureoverlap}(\mathcal{R}), \text{ where } \mathcal{O}_{max} = \mathcal{W}_{max} * |\mathcal{ND}| * |\mathcal{DL}|$$

$$f_3(\mathcal{R}) = \mathcal{O}'_{max} - \text{ruleoverlap}(\mathcal{R}), \text{ where } \mathcal{O}'_{max} = N \times (|\mathcal{ND}| * |\mathcal{DL}|)^2$$

$$f_4(\mathcal{R}) = \text{cover}(\mathcal{R})$$

$$f_5(\mathcal{R}) = \mathcal{F}_{max} - \text{disagreement}(\mathcal{R}), \text{ where } \mathcal{F}_{max} = N \times |\mathcal{ND}| * |\mathcal{DL}|$$

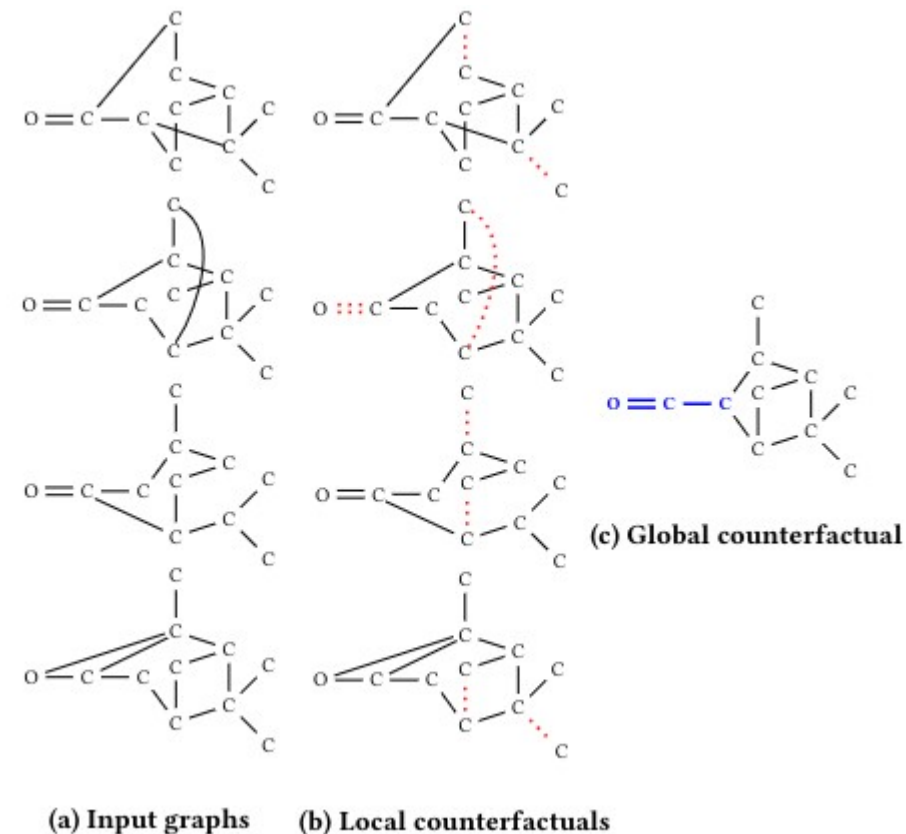
Global explainability

BETA⁽⁺⁾ applies itemset mining to learn if-then rules

- Conditions (gender=♀) obtained via pattern mining
- Rule selection formulated as an optimization problem to:
 - Maximize fidelity and coverage
 - Minimize rule count, feature overlap, and complexity;
 - Constrained by number of rules, maximum width, and number of first level conditions

Global counterfactuals

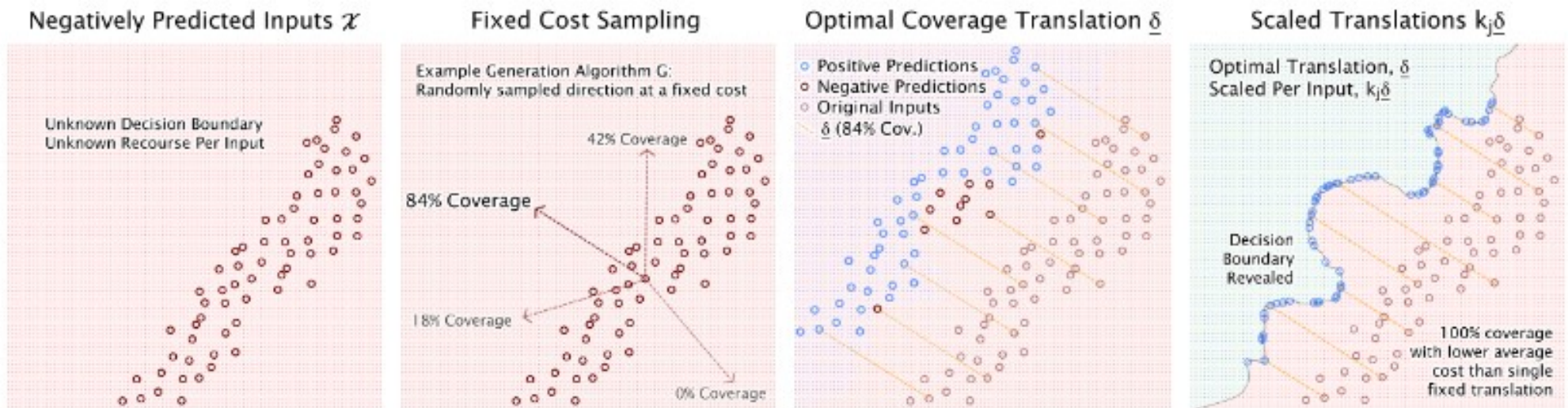
- GCFFExplainer computes counterfactual explanations that generalize to the entire dataset
 - These are recourse rules that optimize for coverage, edit distance (cost), & complexity
 - Applied to GNNs



Global counterfactuals

GLOBE-CE computes translation vectors applied to groups of instances

- Translation vectors can turn factuals into counterfactuals



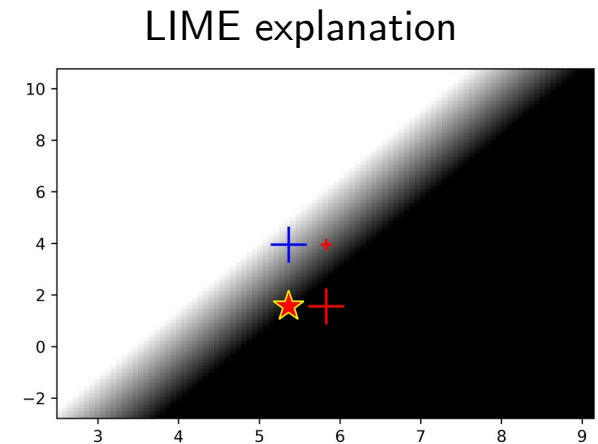
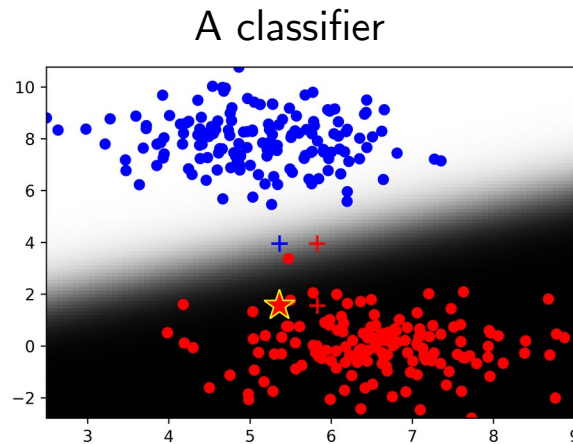
Agenda

- Interpretable AI/ML: What and Why?
- Black-box vs. interpretable models
- **eXplainable AI techniques**
 - Explanation types
 - Self-explainable methods
 - Post-hoc approaches
 - Evaluating XAI
- Conclusion & open research questions

XAI Evaluation Criteria

- Functional

- Complexity
- (In-)Fidelity
 - Adherence
- Stability & robustness



- User-centered

- Understanding

- Comprehensibility, plausibility

- Trust

- Confidence, distrust, complacency

Trust	What is your confidence in the tool? Do you have a feeling of trust in it?	Are the actions of the tool predictable?	Is the tool reliable?	Is the tool efficient at what it does?	Average
User J's Answers	5/7	6/7	3/7	4/7	4.5/7

Functional evaluation – Metrics

- Explanations
 - Adherence: classification and regression metrics
 - Complexity: dependent on explanation type
 - Fidelity: occlusion techniques, accuracy reduction
- Methods
 - Stability & robustness: Jaccard coefficient, stability index, ranking metrics
 - Runtime, memory footprint

User-centered evaluation: Understanding

- Usually via a “proxy” task
 - **Predict** the model’s answer for a given instance
 - **Explain** the features that play a role in the prediction
 - **Validate or reject** statements about the model
 - **Replace** the model (also used for measuring trust)

User-centered evaluation: Understanding

- Usually via a “proxy” task
 - **Predict** the model’s answer for a given instance
 - **Explain** the features that play a role in the prediction
 - **Validate or reject** statements about the model
 - **Replace** the model (also used for measuring trust)
- And behavioral and self-reported metrics
 - Precision/accuracy, task execution time
 - Specialized questionnaires

User-centered Evaluation: Trust

- Via questionnaires
- Adherence to the AI's recommendation
 - Confidence
- Trust is a complex construct
 - Questionnaires test some related construct
 - 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.

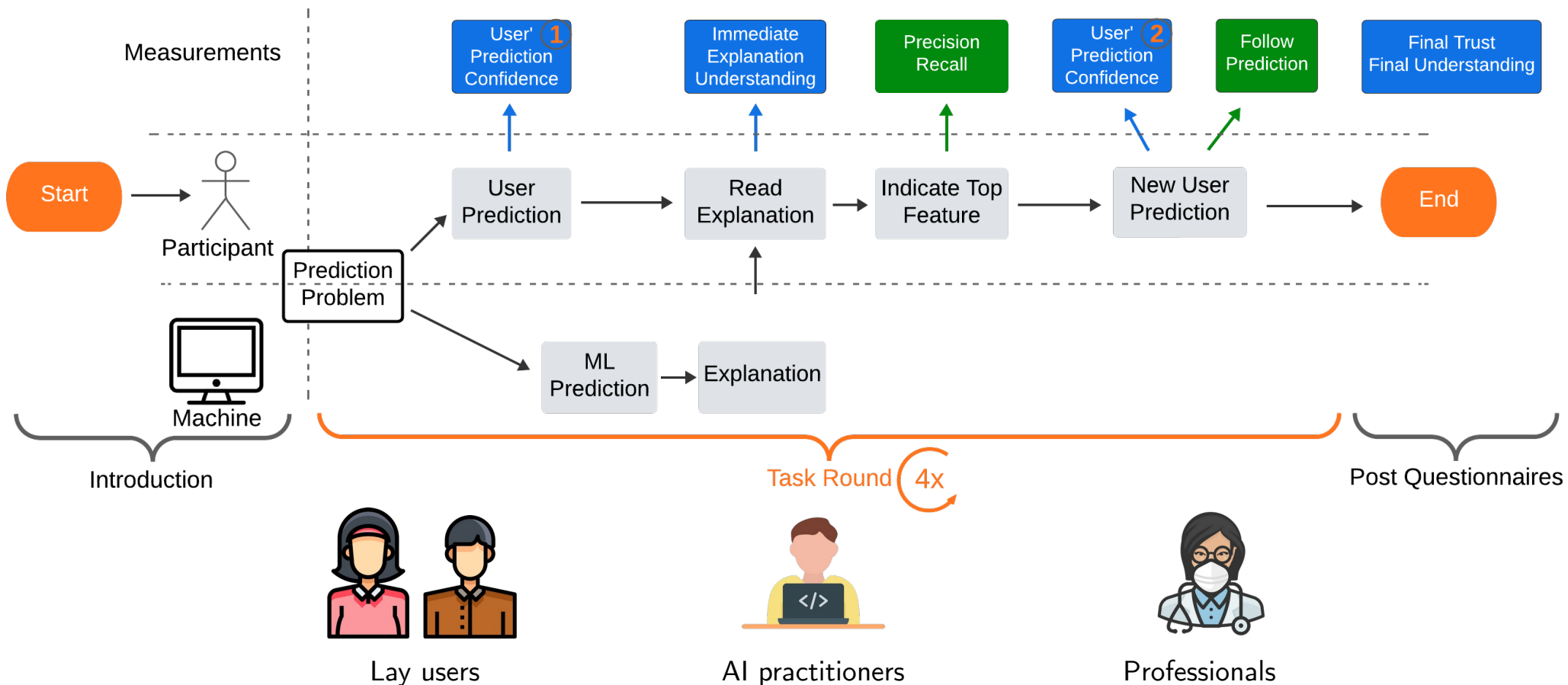
3.

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

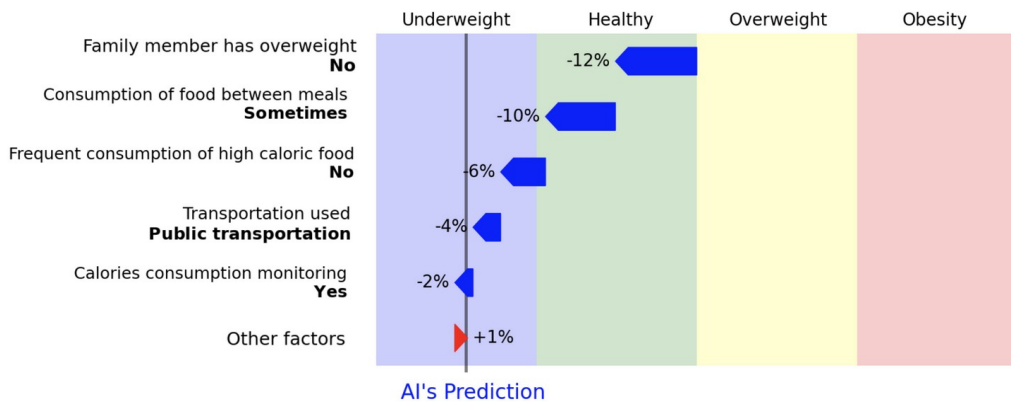
4.

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

Comparing explanations – A protocol

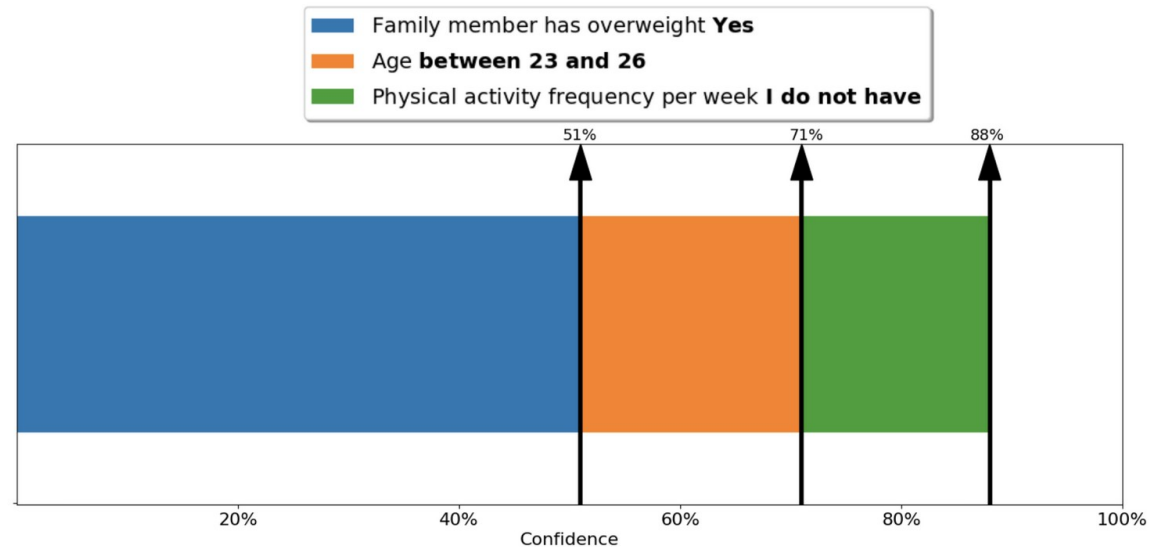


Comparing explanations – A protocol



- 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.



Evaluating explainability

Current situation

- Planned alcohol intake **3 units**
- Water intake so far **5 glasses**
- Hours slept **6 hours**

The system advises a **lower dose of insulin**

*Your planned alcohol intake is more than 1 unit.
If this would have been 1 unit or less, the system would have advised a normal dose.*

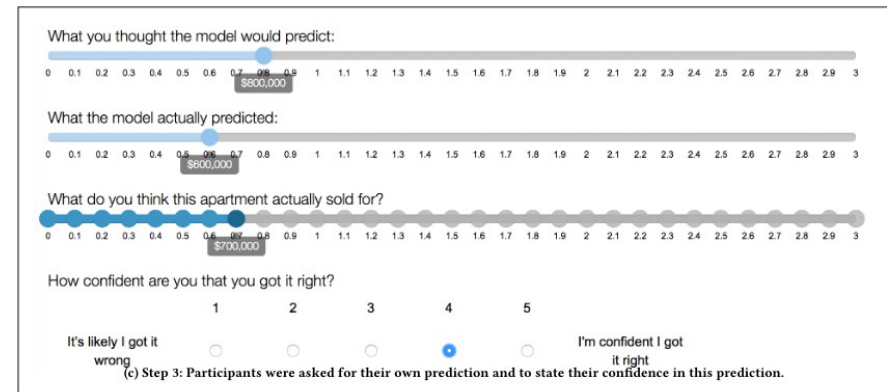
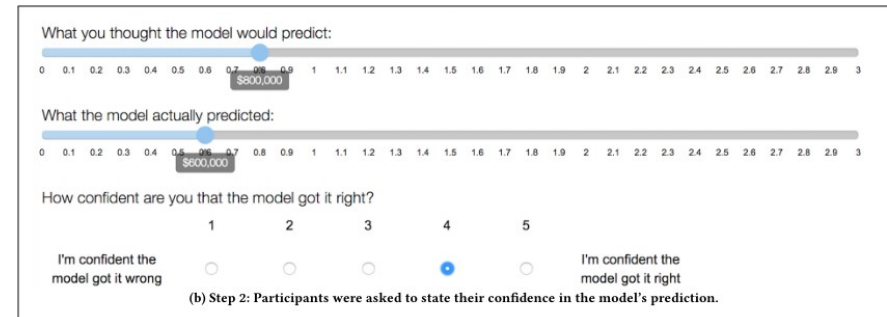
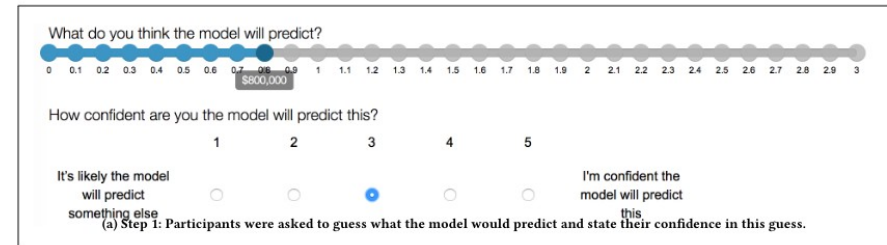
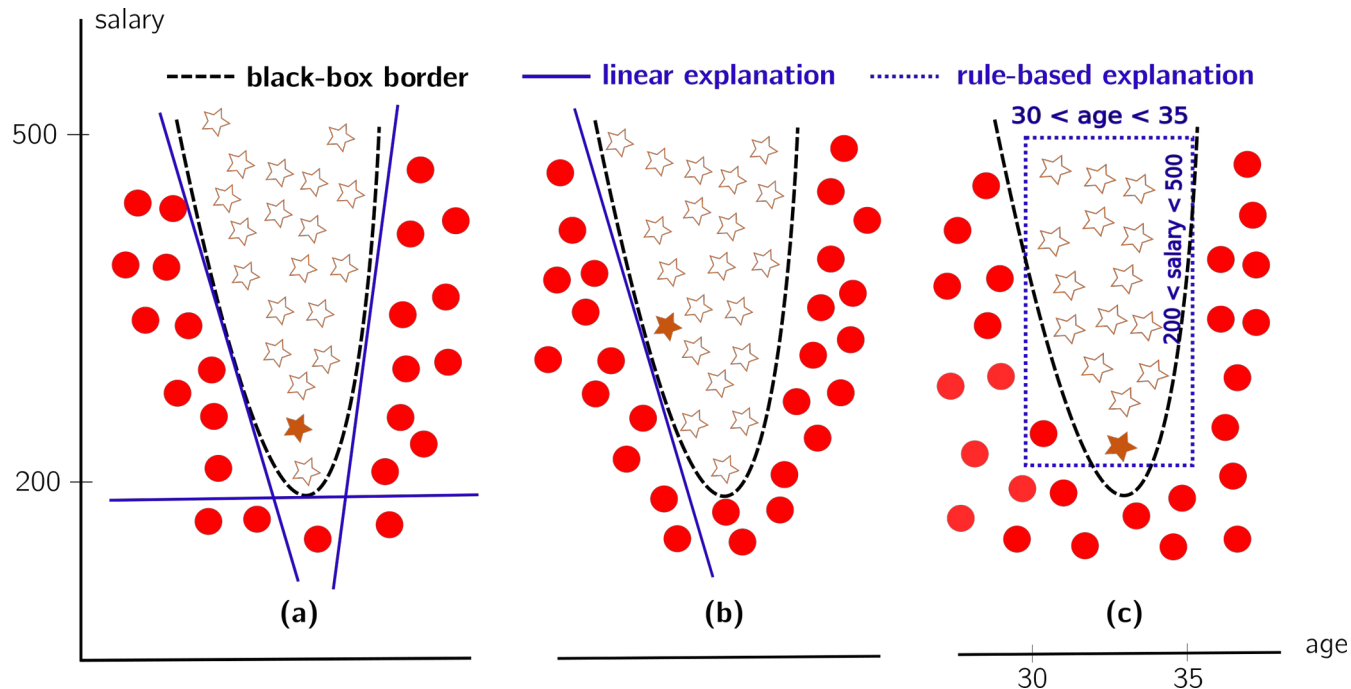


Figure 2: Part of the testing phase from our first experiment.

Comparable situation from your past	Current situation	Comparable situation from your past
<ul style="list-style-type: none"> Planned alcohol intake 3 units Water intake so far 5 glasses Hours slept 7 hours <p>The system advises a lower dose of insulin</p> <p><i>Here, your planned alcohol intake was 3 units and the system also advised a lower dose of insulin. That advice had a positive effect on your blood sugar level.</i></p>	<ul style="list-style-type: none"> Planned alcohol intake 3 units Water intake so far 5 glasses Hours slept 6 hours <p>The system advises a lower dose of insulin</p>	<ul style="list-style-type: none"> Planned alcohol intake 1 unit Water intake so far 4 glasses Hours slept 6.5 hours <p>The system advises a normal dose of insulin</p> <p><i>Here, your planned alcohol intake was 1 unit and the system advised a normal dose of insulin instead. That advice had a positive effect on your blood sugar level.</i></p>

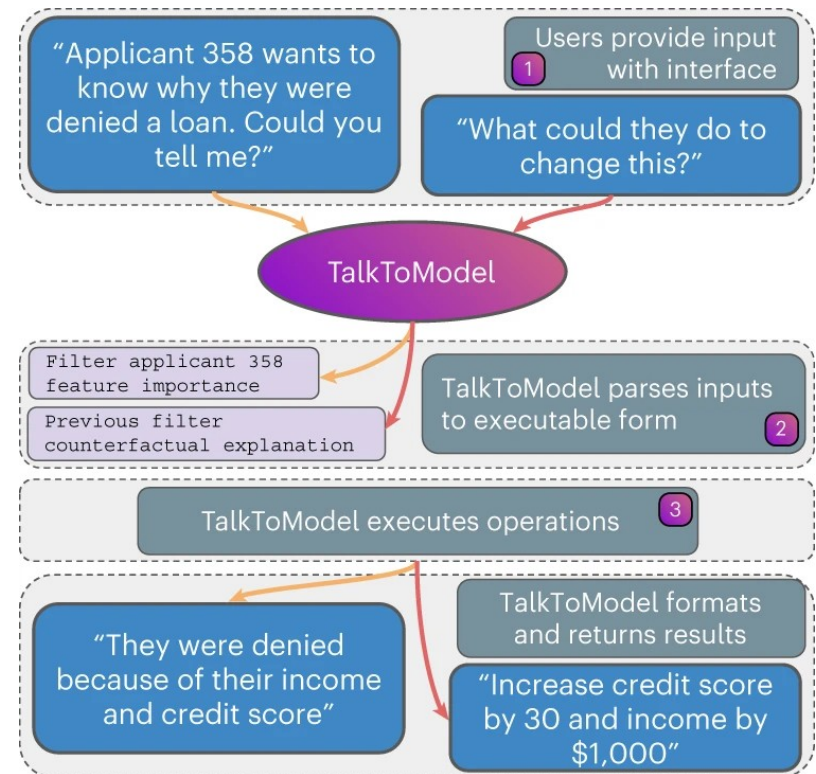
Some open research avenues

- Can we talk about *automatic explainability*?
 - Are linear attribution models more interpretable than decision trees or rule lists?



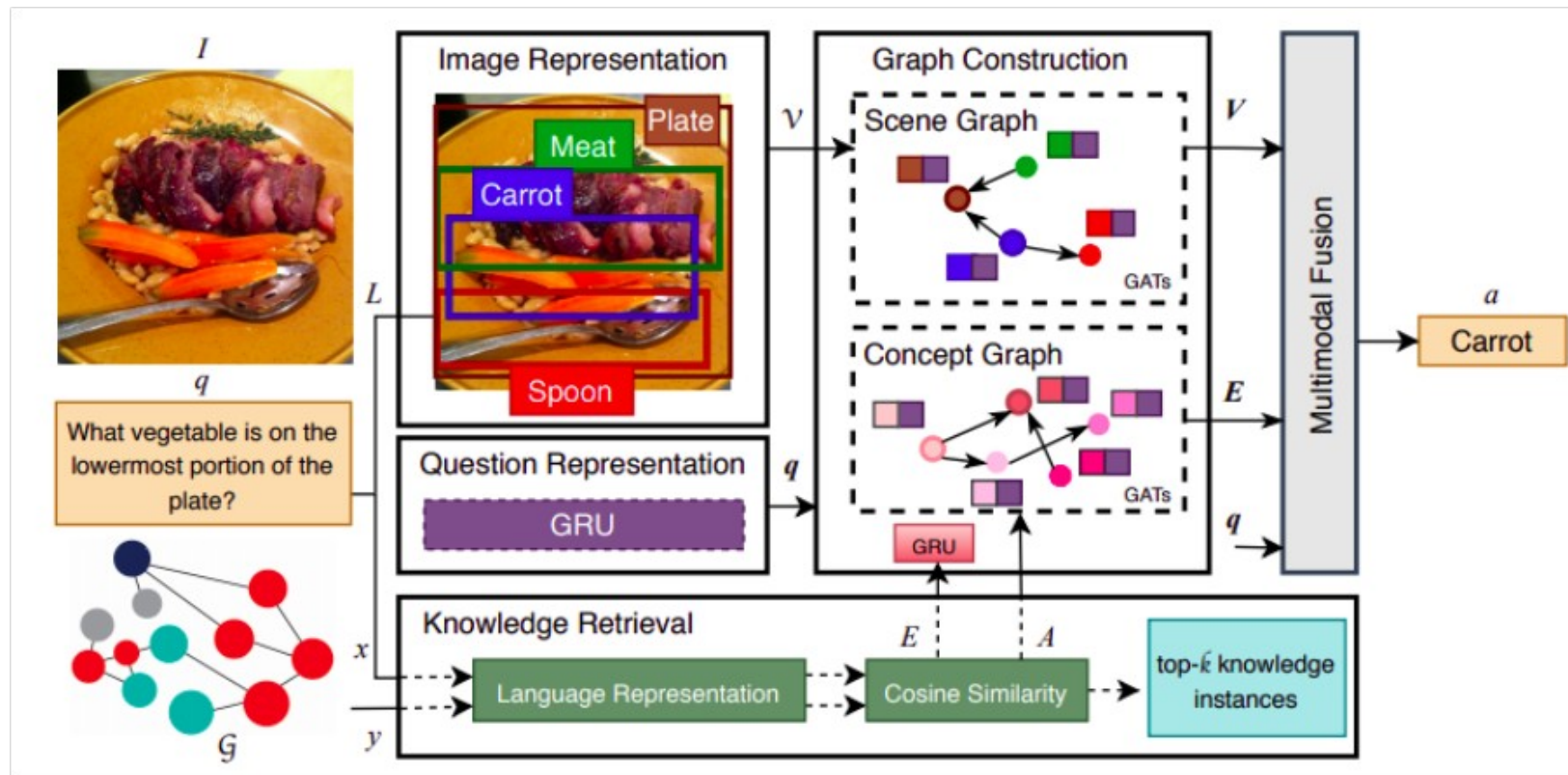
Some open research avenues

- Can we talk about *automatic explainability*?
 - In the form the explanations are conveyed to humans?
 - Could LLMs help?
 - What are suitable visual representations for explanations?
 - What about causal post-hoc explanations?
 - How to take user's profile into account?



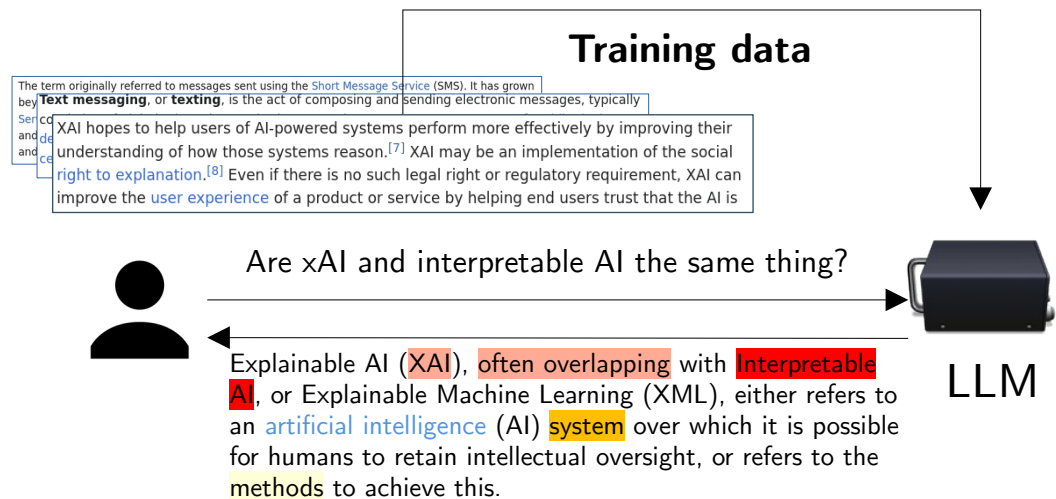
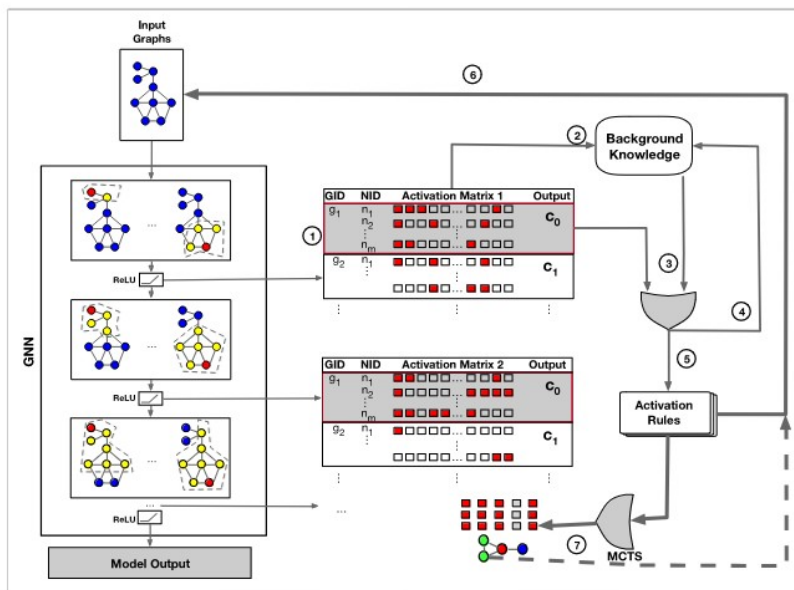
Some open research avenues

- How to explain multimodal systems faithfully and efficiently?



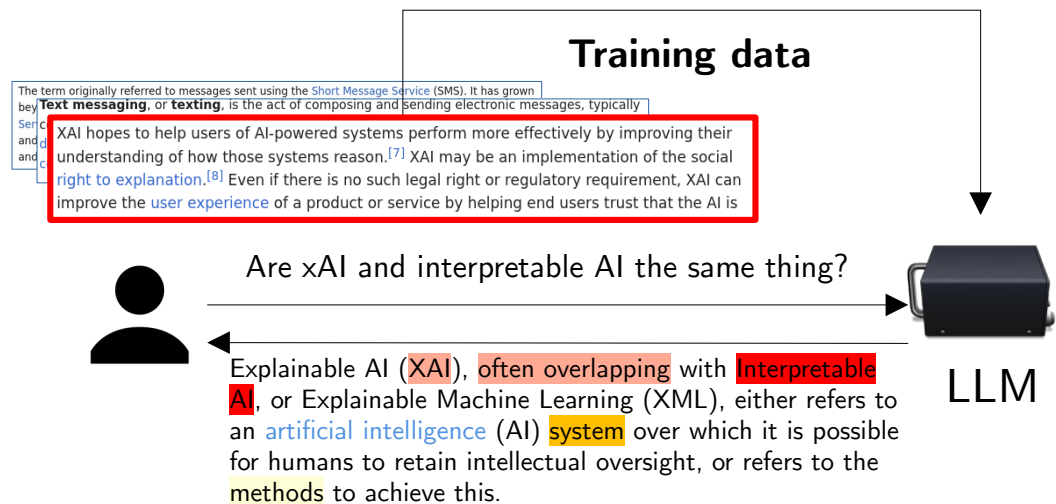
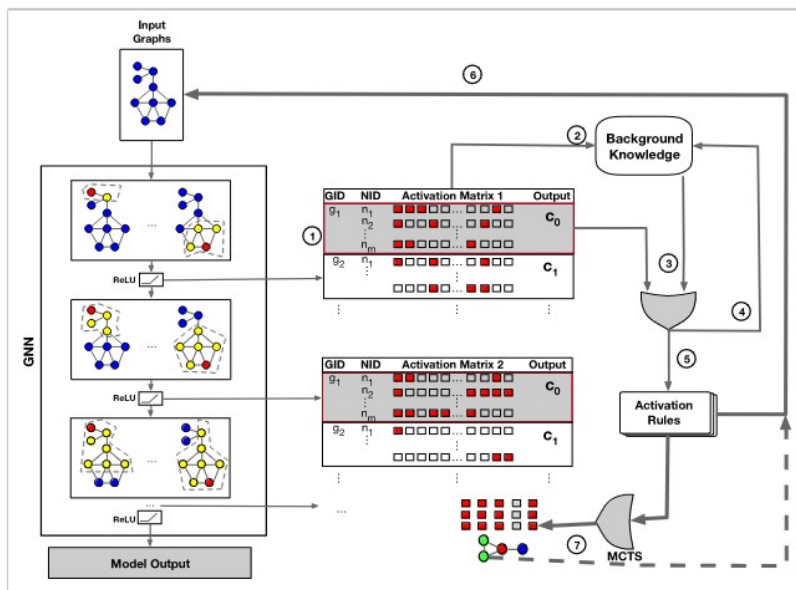
Some open research avenues

- Some works learn activation patterns in NNs that correlate with some outputs
- Can we talk about “source” attribution (e.g., for LLMs, for KG/DB embeddings)



Some open research avenues

- Some works learn activation patterns in NNs that correlate with some outputs
- Can we talk about “source” attribution (e.g., for LLMs, for KG/DB embeddings)



Conclusion

- Interpretability in ML/AI matters
 - For human, ethical, legal, and technical reasons
- Some models are interpretable by design, other require opening the black box *a posteriori*
 - The key of opening the black box is *reverse engineering*

Useful references

- Benchmarking and Survey of Explanation Methods for Black Box Models. <https://arxiv.org/pdf/2102.13076.pdf>
- Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. <https://christophm.github.io/interpretable-ml-book/>

Thank you!