Introduction

Explainable Artificial Intelligence

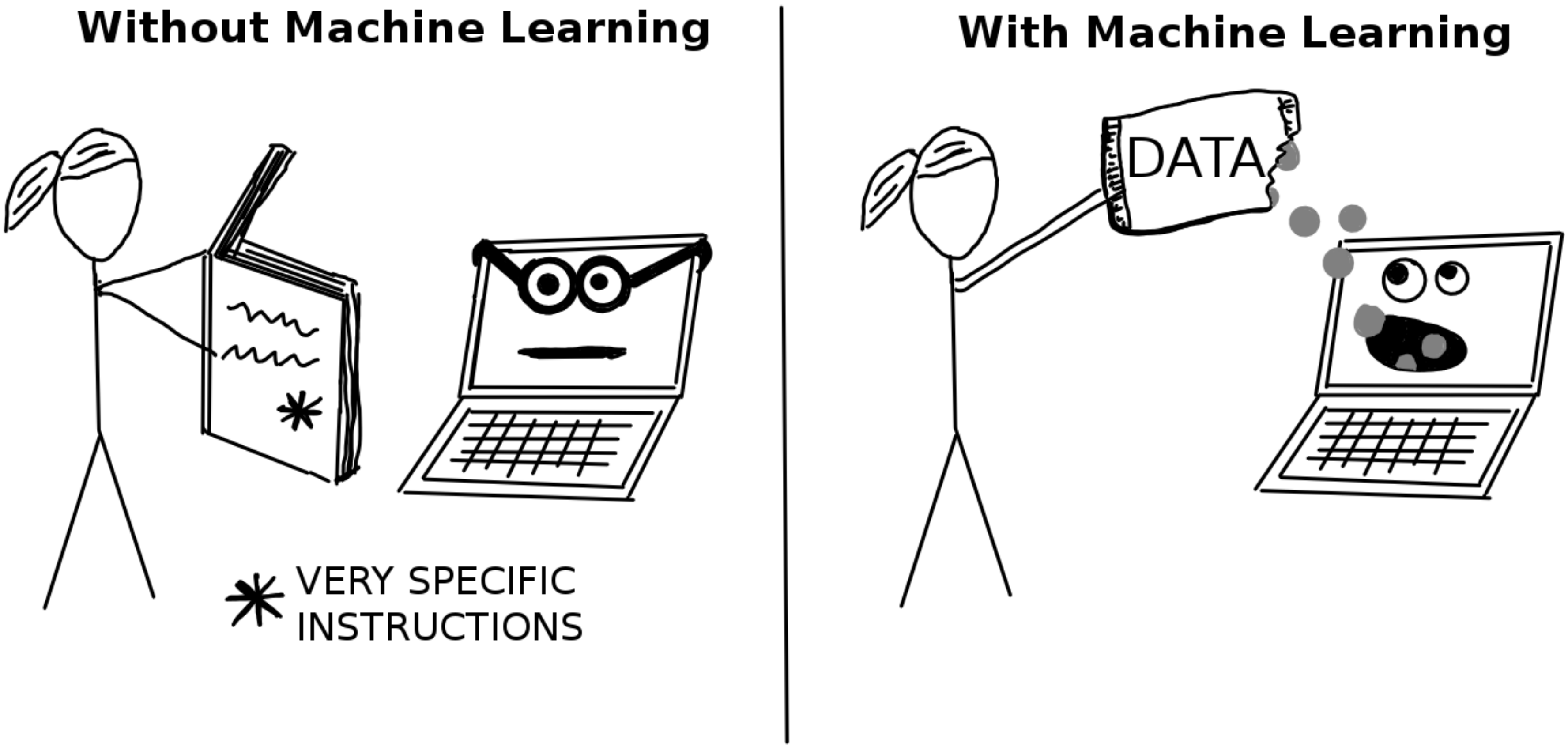
Dr. Stefan Heindorf

Outlook

- Introduction to machine learning
  - Example
  - Applications
- Interpretability
  - Importance of interpretability
  - Taxonomy of interpretability
  - Evaluation of interpretability
- Datasets

What do you know about machine learning?

Machine learning



Christoph Molnar (https://christophm.github.io/interpretable-ml-book/images/programing-ml.png), https://creativecommons.org/licenses/by-nc-sa/4.0/

Machine learning
Example: bike rental

Training

Outlook	Temperature	Humidity	Windy	Rented bicycles
Sunny	Hot	High	False	4000
Sunny	Hot	Normal	True	5000
Overcast	Hot	Normal	False	3000
Rainy	Cool	High	False	1000

Prediction

Outlook	Temperature	Humidity	Windy	Rented bicycles?
Sunny	Mild	High	True	?
Sunny	Cool	Normal	True	?

Machine learning

Example: bike rental

Training

x_1	x_2	x_3	x_4	y
Sunny	Hot	High	False	4000
Sunny	Hot	Normal	True	5000

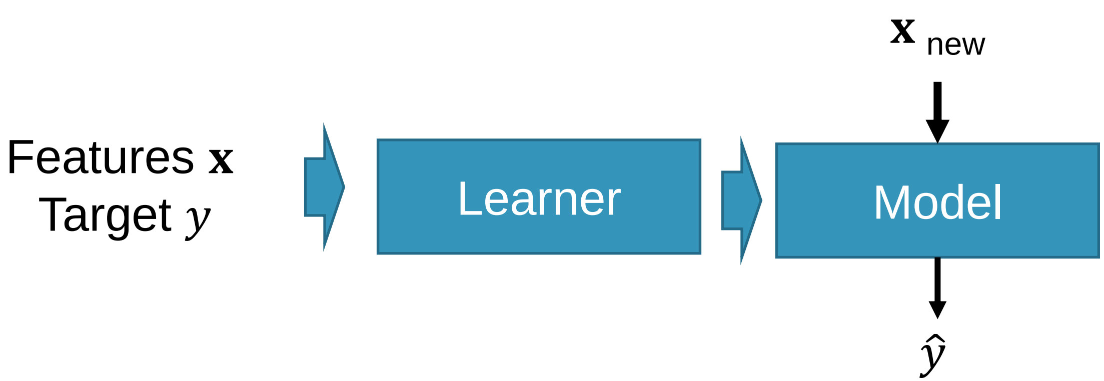
Goal

- Learn function f, such that

$$y \approx \hat{y} := f(x_1, x_2, x_3, x_4)$$

Terms

- x_1, x_2, x_3, x_4 : features
- y: target
- f: model



What are examples of machine learning tasks?

Machine learning
Applications

Natural language processing

- Machine translation (e.g., DeepL, Google Translate)
- Chatbots (e.g., Chat-GPT, Amazon Alexa)
- Document classification (e.g., Spam / no spam, topic classifcation, …)

Computer vision

- Image classification (e.g., face recognition, street signs, medical images, …)
- Optical character recognition (e.g., scan PDF and convert to Word file)
- Image generation (e.g., DALL-E, Midjourney)

Knowledge graphs

- Node classification (e.g., predict the type of a node)
- Link predictions (e.g., predict missing links in the graph)
- Graph classification (e.g., predict properties of chemical molecules)

What are advantages of machine learning?

Machine learning
Advantages

Advantages compared to humans

- Cheaper (runs automatically, saves work)
- Faster (decision within milliseconds)
- Better (e.g., chess, game of go, OCR, …)

Advantages compared to traditional programming

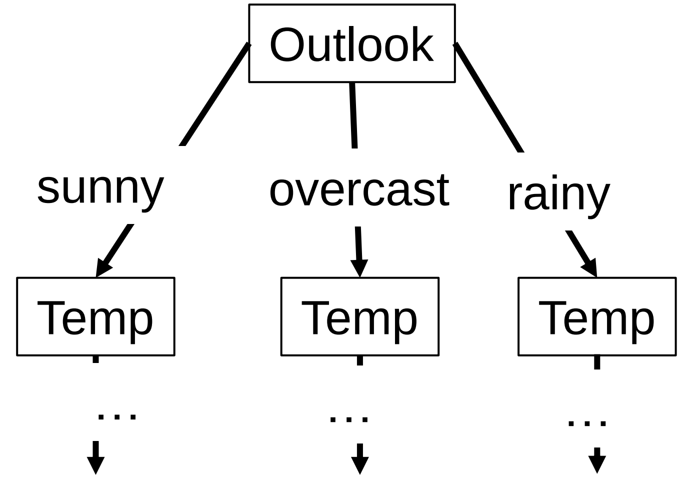
- Cheaper (e.g., few lines of code)
- Faster (e.g., shorter time to market)
- Better (e.g., chess, game of go, OCR, …)

What are disadvantages of machine learning?

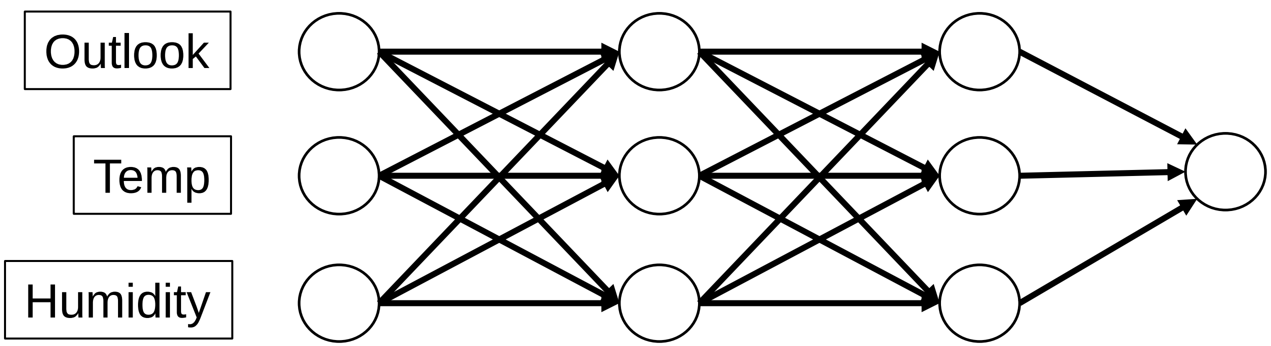
Problem of ML models: interpretability

ML models are often complex

- Random forest
  - Often hundreds of decision trees
  - Often thousands of nodes per tree

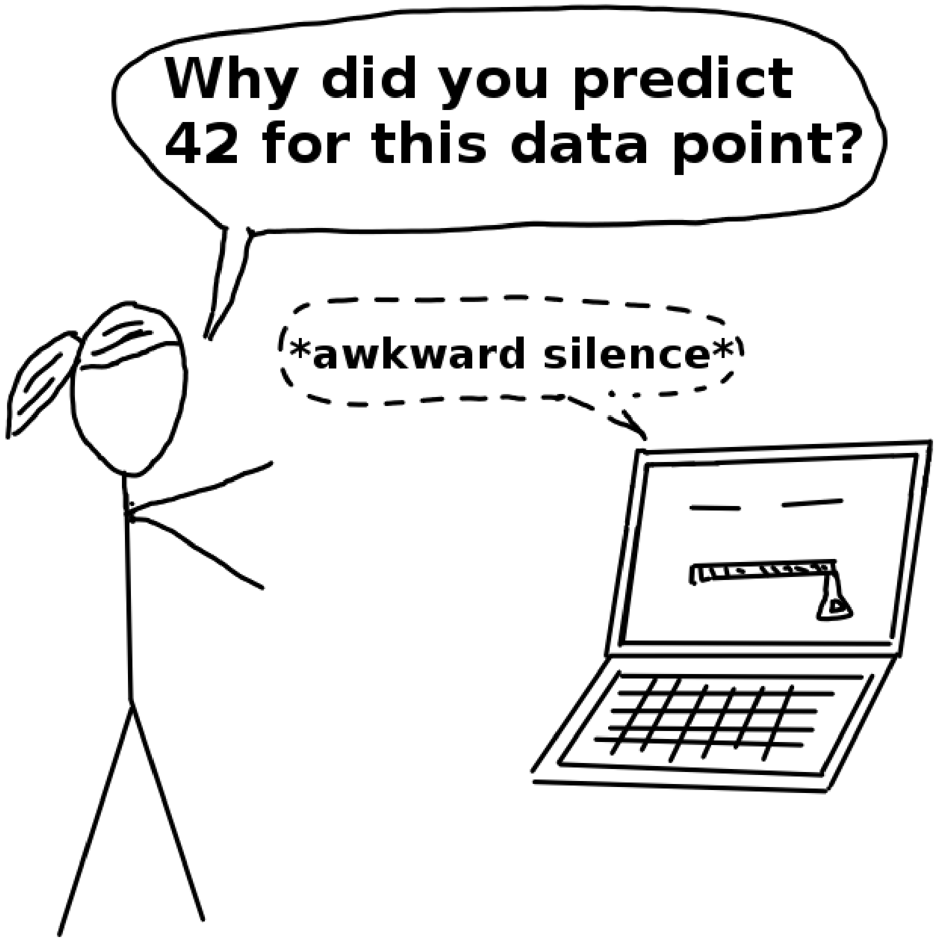


- Neural networks
  - Millions of weights (→)



- Ensembles
  - Combination of multiple models (e.g., random forest, and neural network)

Need for explainability



Christoph Molnar (https://christophm.github.io/interpretable-ml-book/images/iml.png), https://creativecommons.org/licenses/by-nc-sa/4.0/

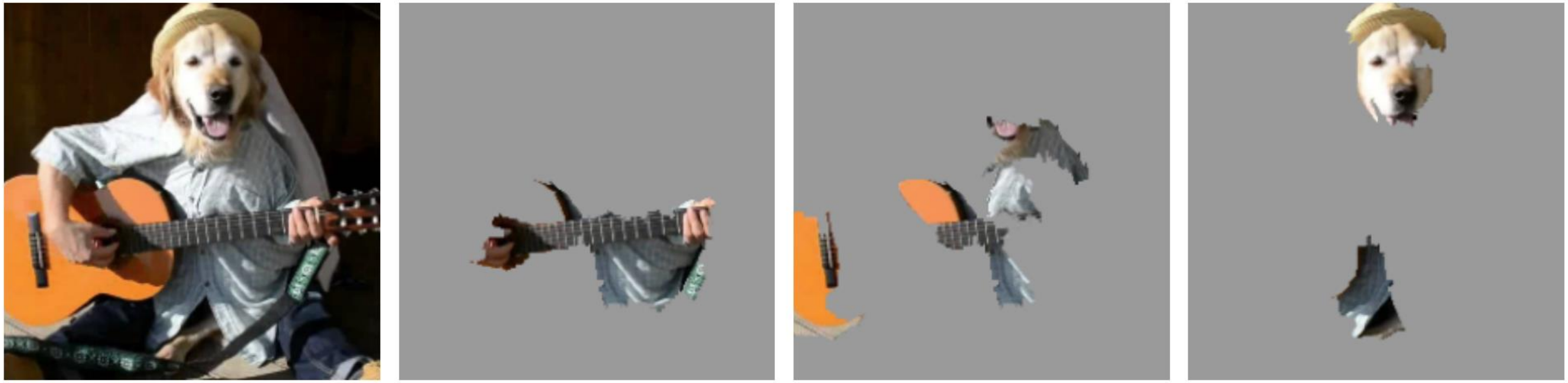
Need for explainability
Clever Hans

Clever Hans: The "Math" Horse

- Horse appeared to solve math by tapping hoof
- Later found: Hans read subtle human cues
- Only correct when questioner knew the answer
- Known as the Clever Hans effect: unintentional signaling

Interpretability

Example of image classification [Riebeiro et al. 2016]



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

(a) Husky classified as wolf (b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

Definitions of interpretability

"Interpretability is the degree to which a human can understand the cause of a decision." (Miller 2017)

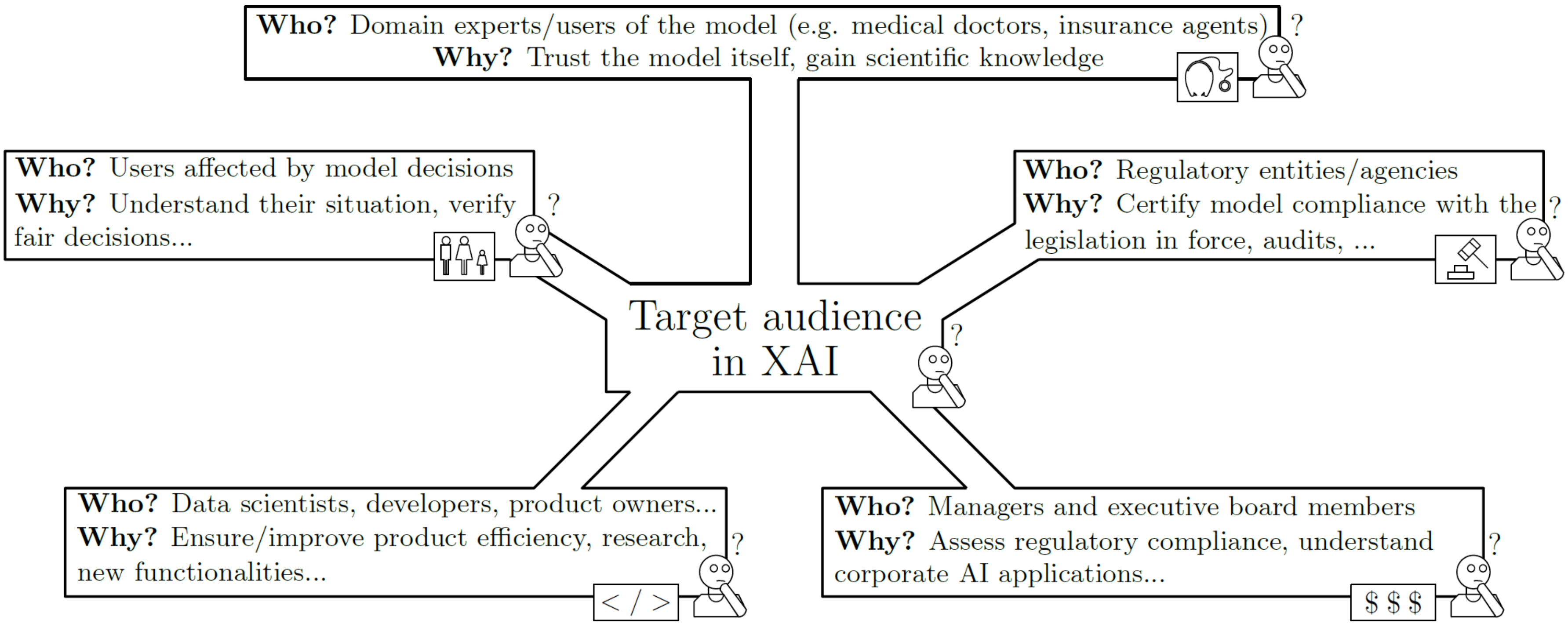
"Interpretability is the degree to which a human can consistently predict the model's result." (Kim et al. 2016)

"The model itself becomes the source of knowledge instead of the data. Interpretability makes it possible to extract this additional knowledge captured by the model." (Molnar 2020)

Comments

- In this course: interpretability = explainability
- Interpretability is not well-defined. But: You will see many examples of interpretability throughout the course
- Explain ML model ≠ explain the data

Importance of interpretability
Stakeholders [Arrietta et al. 2020]



When do we need interpretability?

[Doshi-Velez and Kim, 2017, Molnar 2020]

Incompleteness in problem formalization

- For certain tasks, it is not sufficient to get the “correct” prediction (the what)
- A “correct” prediction (according to loss function / accuracy) only partially solves your original problem
- The model must also explain how it came to the prediction (the why)

Causality

- Check that only causal relationships are learned

Reliability or robustness

- Ensure that small changes in the input lead to small changes in the prediction

Trust

- Humans rather trust a system that explains its decisions than a black box

Fairness

- Ensure that predictions are unbiased
- Do not implicitly or explicitly discriminate against underrepresented groups

Privacy

- Ensure that sensitive information in the data is protected

When do we not need/want interpretability?

[Doshi-Velez and Kim, 2017]

Low-impact applications

- Example: Predict soccer outcome (for a small group of friends)
- But: can become high-impact if done commercially for large amounts of money

Well studied problem

- Example: Optical character recognition
- It simply works (even better than humans in some cases)

Risk of gaming the system

- Example: Credit scoring, spam detection
- Interpretability might help to deceive the system
- Countermeasure: only use causal features (and not proxies thereof)

Goals of interpretability

[Adadi and Berrada, 2018]

Improving the model

- Model shortcuts: Identify Clever Hans predictors and unintended biases
- Debugging: Identify mistakes in feature encoding and mistakes by model
- Improvement: Identify good feature and engineer even better features

Justify model and predictions

- Explain to stakeholders: Provide reasoning for decisions
- Enable contestability: Support recourse for subjects affected by predictions
- Regulatory compliance: Ensure transparency for legal and ethical approval

Discover insights

- Understand the data: the relationships between features and true outcomes
- Understand the model: the relationships between features and predictions

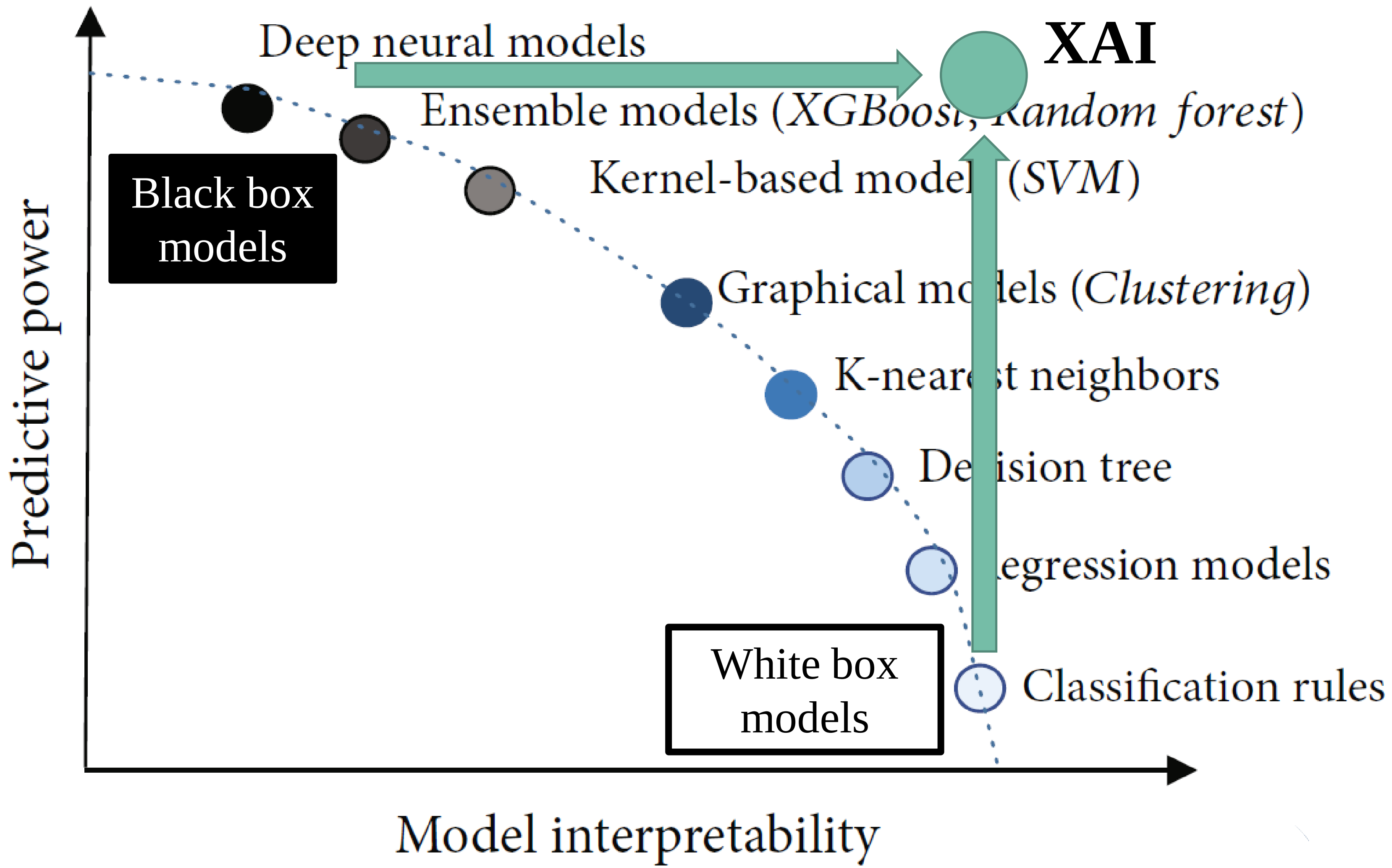
💡 Always evaluate model performance

When determining your interpretability goals, evaluate your model's performance metrics first. This can help you identify if your current goal should be model improvement.

How would you explain a ML model?

How to explain a ML model?

Trade-off between explainability and predictive power



A. Kumar, S. Dikshit, V. Albuquerque (https://www.hindawi.com/journals/wcmc/2021/2939334/),

Taxonomy of interpretability

Intrinsic vs. post-hoc

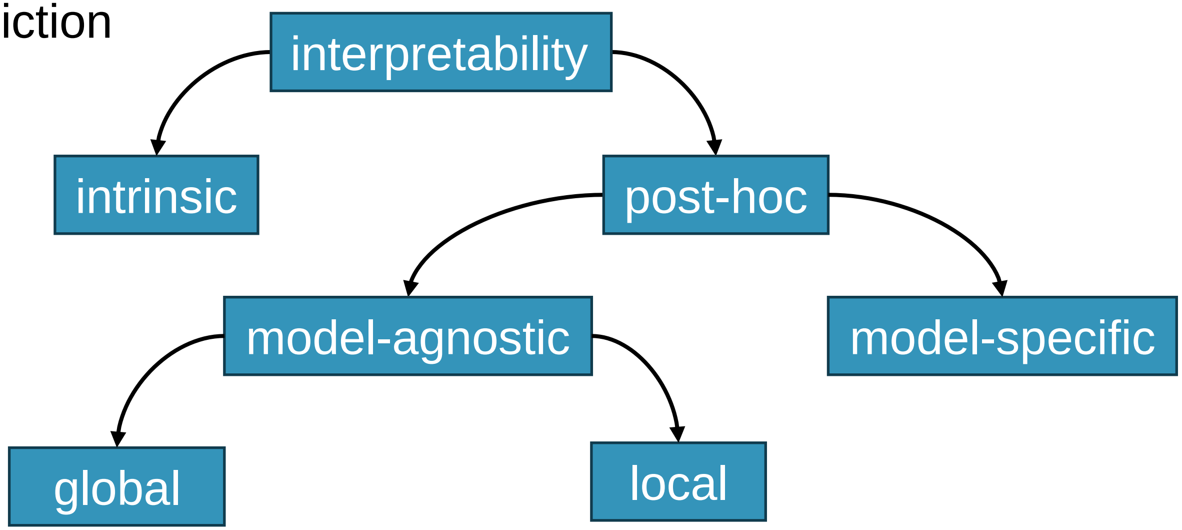
- Intrinsic: restrict complexity of the model before training (e.g., decision tree)
- Post-hoc: analyze complex model after training (e.g., neural network)

Model-agnostic vs. model-specific

- Model-agnostic: explanation method for any model class
- Model-specific: explanation method tailored to specific model classes

Global or local

- Global: explain entire model
- Local: explain single prediction



Strengths of methods and types of explanations

[Molnar, 2025]

Strengths of methods

	Improvement	Debugging	Justification	Model insights	Data insights	Flexibility (model/explainer)
Intrinsic	✓		✓	✓		
Local model-agnostic		✓		✓	(✓)	✓
Global model agnostic	✓		✓	✓	(✓)	✓
Model-specific	✓		✓	✓		

Types of explanations

- Feature summary statistics (e.g. feature importance)
- Feature summary visualization (e.g., partial dependence plots)
- Model internals (e.g., learned weights, thresholds of decision tree)
- Data points (e.g., counterfactual explanations)
- Intrinsically interpretable model (e.g., surrogate model for black box model)

How would you evaluate interpretability?

Evaluation of interpretability

[Doshi-Velez and Kim, 2017]

Application-level evaluation (real task)

- Put explainer into product: test with end users / domain experts
- Example: Fracture detection and location in X rays (baseline: humans)

Application-level evaluation (simple task)

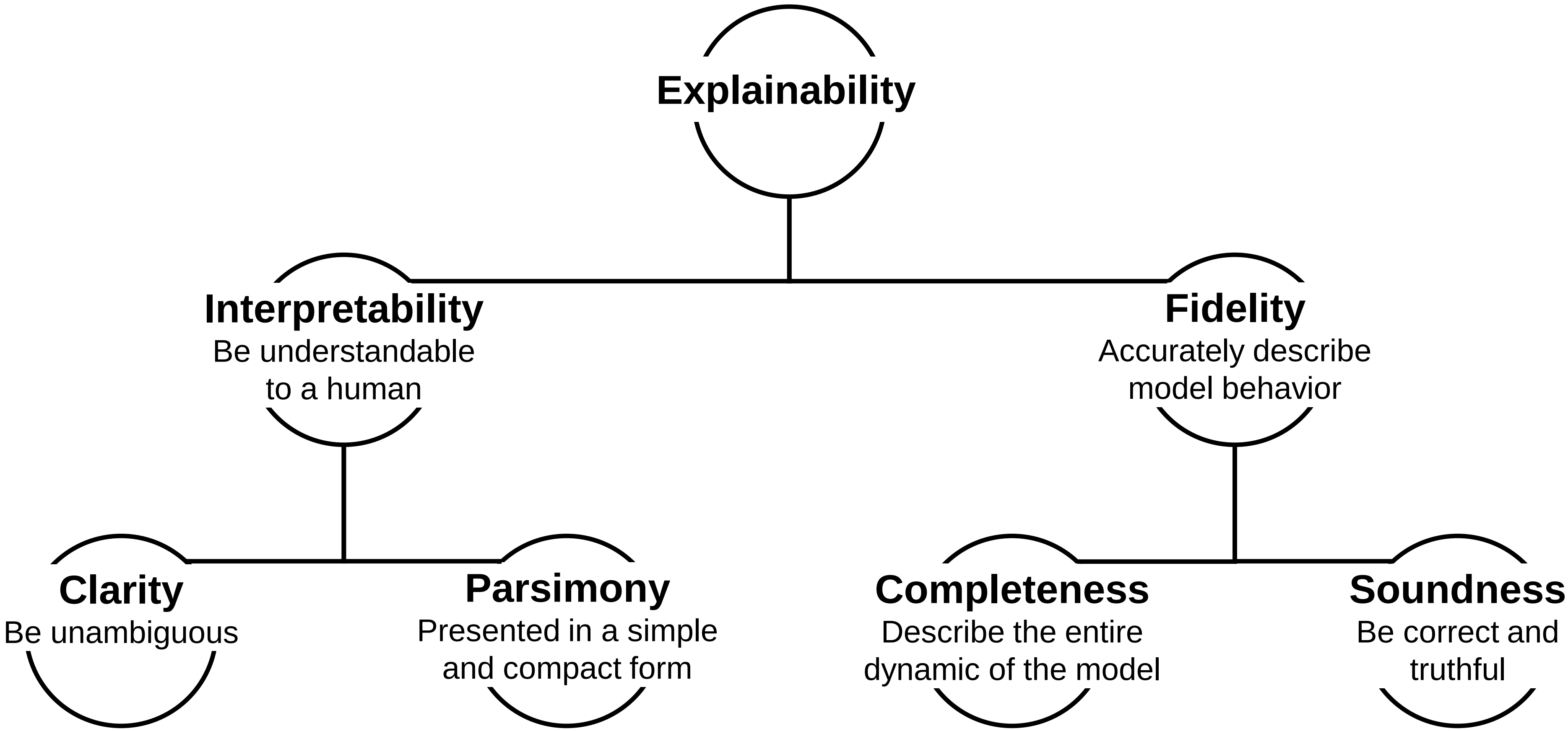
- Put explainer into product: test with laypersons (e.g., crowdworkers)
- Example: User picks between two explanations (binary forced choice)
- Example: User must correctly simulate the model's output (forward simulation)
- Example: User must change the model's input to change the model's prediction to a desired output (counterfactual simulation)

Function-level evaluation (proxy task)

- Automatic evaluation without humans
- Most appropriate if previously shown that end users understand the kind of models (e.g., decision trees, linear models)
- Example: Depth of decision tree
- Example: Number of non-zero feature weights of linear models
- Example: Fidelity of surrogate model: How well does the surrogate model (e.g., decision tree / linear model) approximate the original model?

Properties of explanations

[Markus, 2021; Zhou, 2021]



Human-friendly explanations

[Miller, 2017; Molnar, 2020]

An explanation is the answer to a why-question

- Why did the treatment not work on the patient?
- Why was my loan rejected?
- Why have we not been contacted by alien life yet?

Good explanations are

- Contrastive (e.g., comparison to similar datapoint with different output)
- Selective (e.g., focus on 1 to 3 features)
- Social (e.g., part of a conversation/interaction)
- Focus on the abnormal (focus on causes with small probability)
- Truthful (true in reality, i.e., in other situations)
- Consistent with prior beliefs (humans tend to ignore information inconsistent with their prior beliefs)
- General and probable (can explain many events)

Example dataset

Bike rentals (regression)

Data source

- Bicycle rental company: Capital-Bikeshare in Washington D.C.
- Along with weather and seasonal information
- From UCI Machine Learning Repository (with slight processing by Christoph Molnar)

Task

- Predict the number of rented bicycles on the next day

Features

- Season: SPRING, SUMMER, FALL, WINTER
- Holiday: Y, N
- Workday: Y, N
- Weather: GOOD, MISTY, RAIN/SNOW/STORM
- Temperature: in degrees Celsius
- Humidity: in percent (0 to 100)
- Wind speed: in km per hour
- Number of rented bikes two days ago

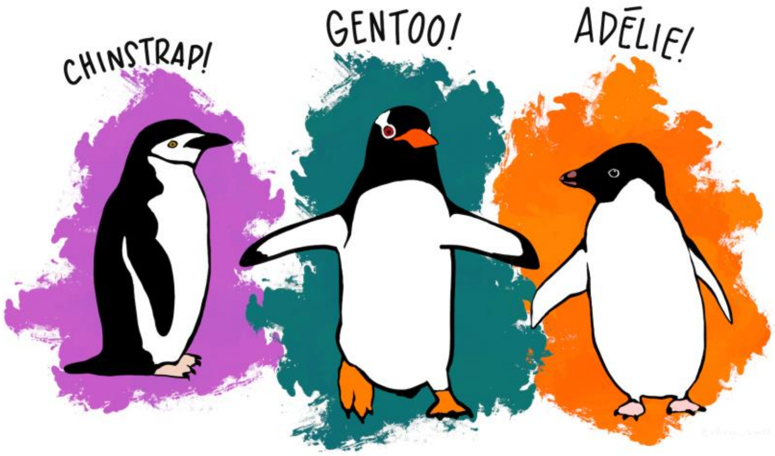
Example dataset

Bike rentals (regression)

season	holiday	workday	weather	temp	hum	windspeed	cnt_2d_bfr	cnt
WINTER	N	Y	GOOD	1.229	43.727	16.637	985	**1349**
WINTER	N	Y	GOOD	1.400	59.044	10.740	801	**1562**
WINTER	N	Y	GOOD	2.667	43.696	12.522	1349	**1600**
WINTER	N	Y	GOOD	1.604	51.826	6.001	1562	**1606**
WINTER	N	Y	MISTY	1.237	49.870	11.305	1600	**1510**
WINTER	N	N	MISTY	-0.245	53.583	17.876	1606	**959**

Example dataset

Palmer penguins (classification)



Data source

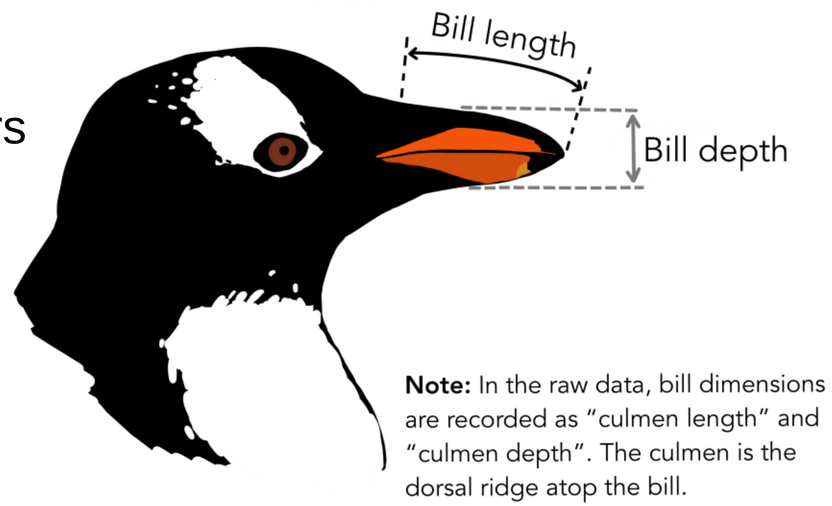
- Palmer station in Antartica
- Download

Task

- Predict whether a penguin is male/female

Features

- Species of penguin: Chinstrap, Gentoo, Adelie
- Body mass of the penguin: in grams
- Length of the bill (the beak): in millimeters
- Depth of the bill: in millimeters
- Length of the flipper (the “tail”): in millimeters



Example dataset

Palmer penguins (classification)

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	year	sex
Adelie	Torgersen	39.1	18.7	181	3750	2007	**male**
Adelie	Torgersen	39.5	17.4	186	3800	2007	**female**
Adelie	Torgersen	40.3	18	195	3250	2007	**female**
Adelie	Torgersen	NA	NA	NA	NA	2007	**NA**
Adelie	Torgersen	36.7	19.3	193	3450	2007	**female**
Adelie	Torgersen	39.3	20.6	190	3650	2007	**male**

References

- **Adadi**, Amina, and Mohammed Berrada. "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)." *IEEE access* 6 (2018): 52138-52160.
- **Arrieta**, Alejandro Barredo, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." *Information fusion 58* (**2020**): 82-115.
- **Doshi-Velez**, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning." *arXiv preprint arXiv:1702.08608* (2017).
- **Kim**, Been, Rajiv Khanna, and Oluwasanmi O. Koyejo. "Examples are not enough, learn to criticize! criticism for interpretability." *Advances in neural information processing systems 29* (**2016**).
- **Markus**, Aniek F., Jan A. Kors, and Peter R. Rijnbeek. "The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies." *Journal of biomedical informatics* 113 (**2021**): 103655.
- **Miller**, Tim. "Explanation in artificial intelligence: Insights from the social sciences." *Artificial intelligence* 267 (**2019**): 1-38.
- **Molnar**, Christoph. *Interpretable machine learning*. **2025**.

References

- **Ribeiro**, Marco Tulio, Sameer Singh, and Carlos 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*, pp. 1135-1144. **2016.**
- **Zhou**, Jianlong, et al. "Evaluating the quality of machine learning explanations: A survey on methods and metrics." *Electronics* 10.5 (**2021**): 593.