Interpretable ML

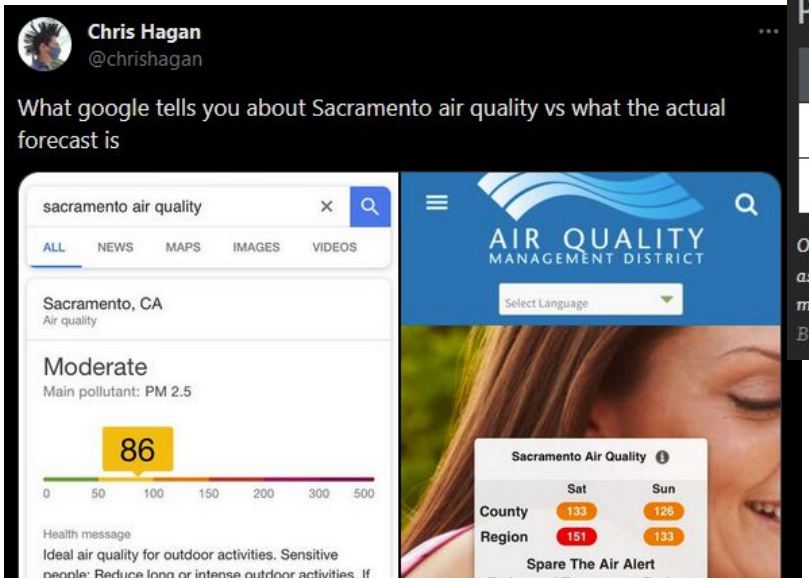
“Interpretable ML is not a subset of XAI” [Rudin et al. 2022]

Outline

- **Why?** - Motivation
- **What?** - Definition(s)
- **How?** - Principles
- **Which?**
  - Interpretable Models
  - Interpretability Methods
- **And repeat**

Do we need interpretations?

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing



Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

Aug 19, 2020 - Technology

How an AI grading system ignited a national controversy in the U.K.

Bryan Walsh, author of Axios Future

This year, Mr. Rodríguez returned to the parole board with the same faulty Compas score. He had identified an error in one of the inputs for his Compas assessment. But without knowing the input weights, he was unable to explain the effect of this error, or persuade anyone to correct it. Instead of challenging the result, he was left to try to argue for parole despite the result.

https://www.nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html

Is it an issue of models being proprietary?

Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study

John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, Eric Karl Oerman

Published: November 6, 2018 • https://doi.org/10.1371/journal.pmed.1002683

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.

Article | Open access | Published: 11 March 2019

Unmasking Clever Hans predictors and assessing what machines really learn

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek & Klaus-Robert Müller

Nature Communications **10**, Article number: 1096 (2019) | Cite this article

Article | Open access | Published: 30 April 2019

Deep learning predicts hip fracture using confounding patient and healthcare variables

Marcus A. Badgeley, John R. Zech, Luke Oakden-Rayner, Benjamin S. Glicksberg, Manway Liu, William Gale, Michael V. McConnell, Bethany Percha, Thomas M. Snyder & Joel T. Dudley

npj Digital Medicine **2**, Article number: 31 (2019) | Cite this article

What is Interpretable ML?

According to Graziani et al. [2023]

- Definition extracted from a survey of many articles and talks with various professionals:
- “An AI system is *interpretable* if it is possible to **translate its working principles** and outcomes **in human-understandable** language *without* affecting the **validity** of the system”

- In some cases the definition can include also “**approximations of the AI system** that maintain its validity *as much as possible*.”

Interpretability through other lens

According to Graziani et al. [2023]

EU law - AI interpretability defines the **supply of meaningful information about** the underlying **logic, significance** and envisaged **consequences** of the AI system

Sociology - AI interpretability must define a **social relationship of trust** between the human and the machine

Ethics - Interpretability is about accountability of the model, similarly to human interaction

> Should we allow for some models to not be interpretable?
> E.g. IVF using black-box AI [Afnan et al. 2021]

What is Interpretable ML?

5 principles according to Rudin et al. [2022]

1. Models are *domain-specific* and should be **understandable by humans**
2. Should help **decide on trust**, not just improve trust
3. *Interpretability vs. Accuracy* is a **false dichotomy**
4. Metrics should be refined throughout the data science process
5. In *high-stakes decisions*, interpretable models should be preferred.

Full / partial interpretability

A proprietary model is also a black box

Interpretability vs Accuracy trade-off

Argument that interpretability just takes more work [Rudin et al. 2022]

BUT also:

A typical interpretable supervised learning setup, with data $\{z_i\}_i$, and models chosen from function class $\mathcal{F}$ is:

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_i \text{Loss}(f, z_i) + C \cdot \text{InterpretabilityPenalty}(f), \quad \text{subject to} \qquad (*)$$

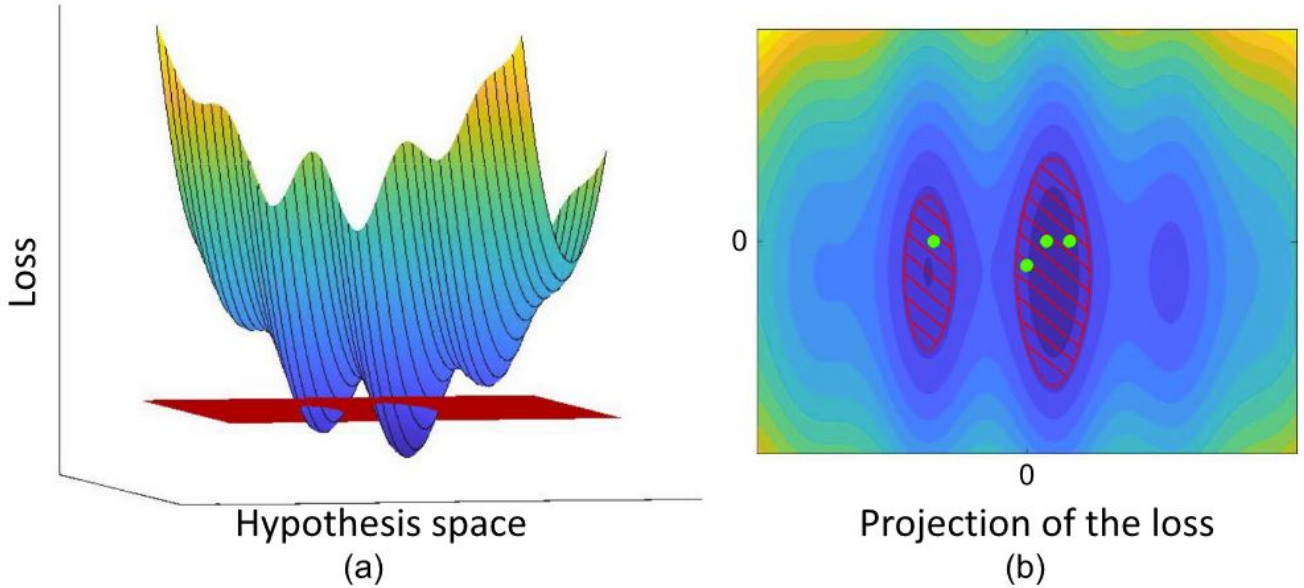
$$\text{InterpretabilityConstraint}(f),$$

where the loss function, as well as soft and hard interpretability constraints, are chosen to match the domain. (For classification z_i might be (x_i, y_i), $x_i \in \mathbb{R}^p$, $y_i \in \{-1, 1\}$.) The goal of these constraints is to make the resulting model f or its predictions more interpretable. While solutions of $(*)$ would not necessarily be sufficiently interpretable to use in practice, the constraints would generally help us find models that would be interpretable (if we design them well), and we might also be willing to consider slightly suboptimal solutions to find a more useful model. The constant C trades off between accuracy and the interpretability penalty, and can be tuned, either by cross-validation or by taking into account the user's desired tradeoff between the two terms.

Interpretability vs Accuracy

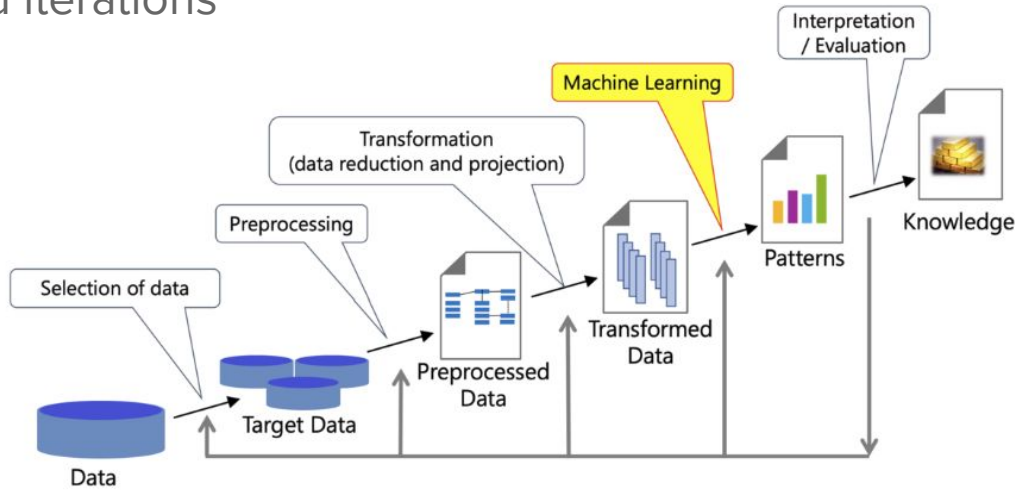
Rashomon Effect - Existence of multiple models with similarly good performance

Could we choose the most interpretable out of the Rashomon set?



All part of a Data Science process

- Interpretability is desirable even for model creators
  - it is useful to learn why the mistakes are happening
- Instead of providing reasons for trust ➡ Let user make their own decision
- Interpretability metrics need iterations



Why make a non-interpretable model?

According to Rudin [2019]:

- Interpretable models are more difficult to construct
  - One finds out issues with the data, needs extra metrics
- Black box protects intellectual property
  - Interpretations from outside might give out market advantage
- Belief of black-box models accessing “hidden patterns”
  - Interpretable models might find them too - and uncover them to humans

Inherent Interpretability

AI Models interpretable **by design / under some constraints**

Models	Data type
decision trees / decision lists (rule lists) / decision sets	somewhat clean tabular data with interactions, including multiclass problems. Particularly useful for categorical data with complex interactions (i.e., more than quadratic). Robust to outliers.
scoring systems	somewhat clean tabular data, typically used in medicine and criminal justice because they are small enough that they can be memorized by humans.
generalized additive models (GAMs)	continuous data with at most quadratic interactions, useful for large-scale medical record data.
case-based reasoning	any data type (different methods exist for different data types), including multiclass problems.
disentangled neural networks	data with raw inputs (computer vision, time series, textual data), suitable for multiclass problems.

[Rudin et al. 2022]

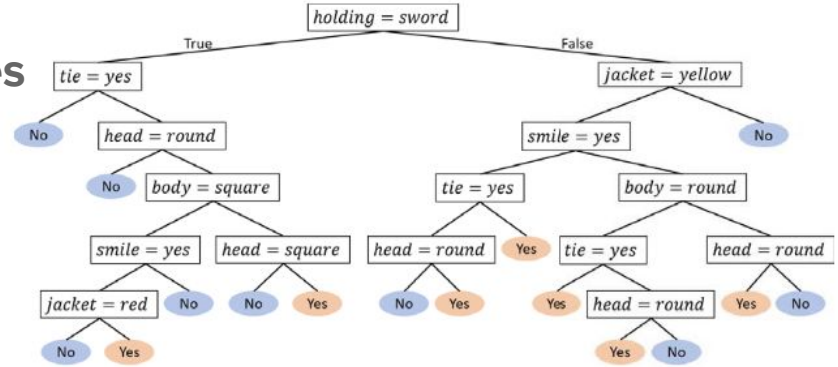
Decision lists/trees/sets

Interpretable under condition:

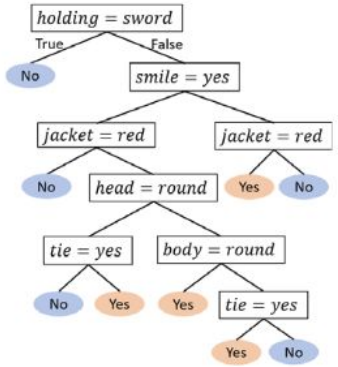
Sparsity - as few features used in a decision as possible

But not too sparse, humans are **opposed to too simplistic** representations of complex relations

Decreasing probabilities in a rule list



(a) training accuracy: 75.74%; test accuracy: 69.44%



(b) training accuracy: 81.07%; test accuracy: 73.15%

[Rudin et al. 2022]

Scoring systems

Sparsity is again important

Also integrality and magnitude of weights

Patient screens positive for obstructive sleep apnea if Score >1			
1.	age ≥ 60	4 points	.....
2.	hypertension	4 points	+.....
3.	body mass index ≥ 30	2 points	+.....
4.	body mass index ≥ 40	2 points	+.....
5.	female	-6 points	+.....
	Add points from row 1-5	Score	=.....

[Rudin et al. 2022]

Generalized additive models

$$g(E[y]) = \beta_0 + f_1(x_{.1}) + \ldots + f_p(x_{.p})$$

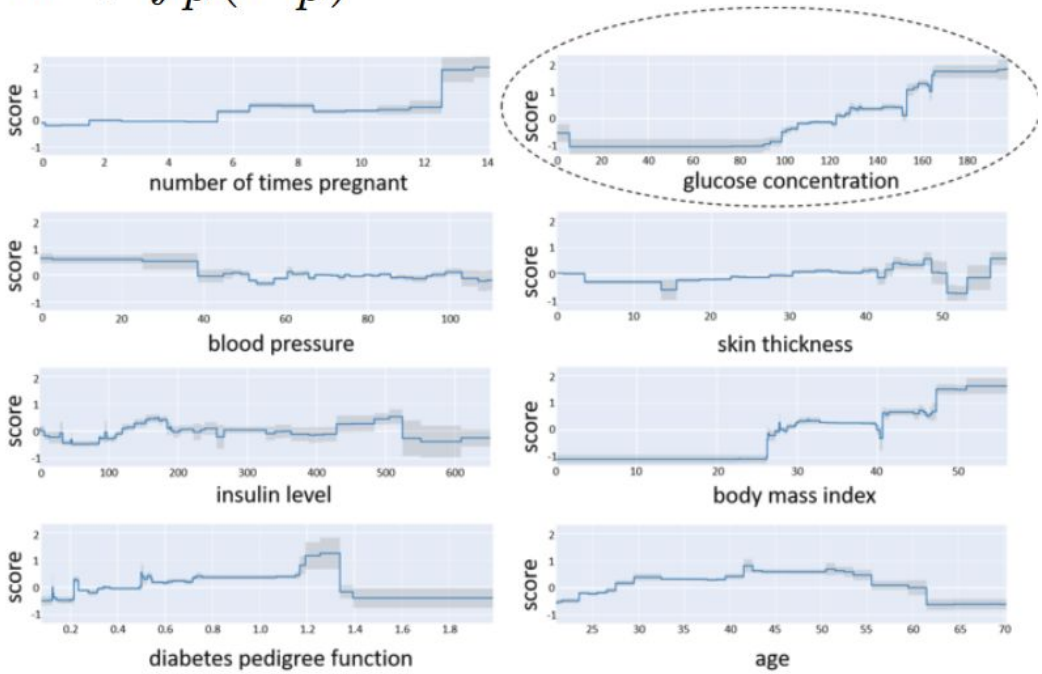
Interpretability comes from option of visualization of the univariate functions

And sparsity, again

We can check if a model fits the expert knowledge

E.g. enforce monotonicity

[Rudin et al. 2022]



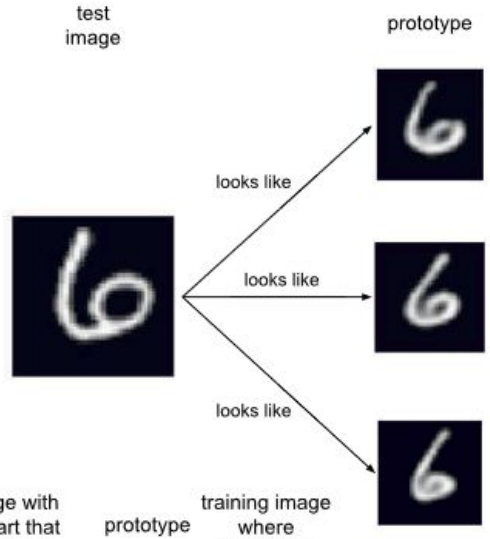
Case-based reasoning

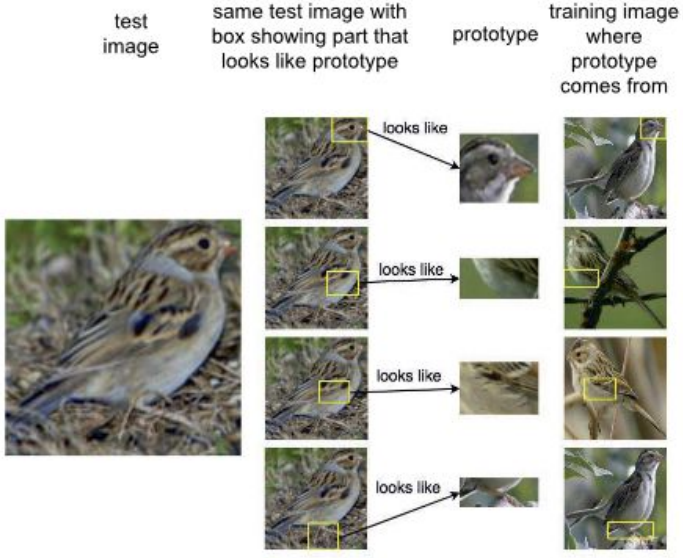
Natural to humans, use examples you already know

Nearest Neighbors

Prototypes - compute only few distances

Part-based - Find similar parts - images, text





[Rudin et al. 2022]

Disentangled NNs

Train a NN where each neuron (in a layer) corresponds to a single “concept”

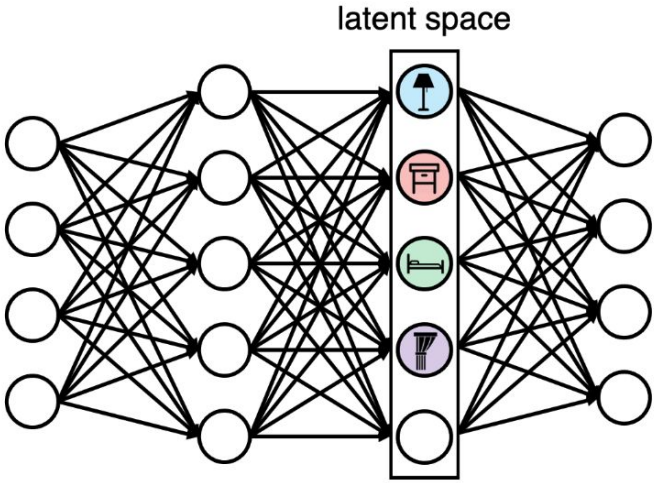
Question of expressiveness

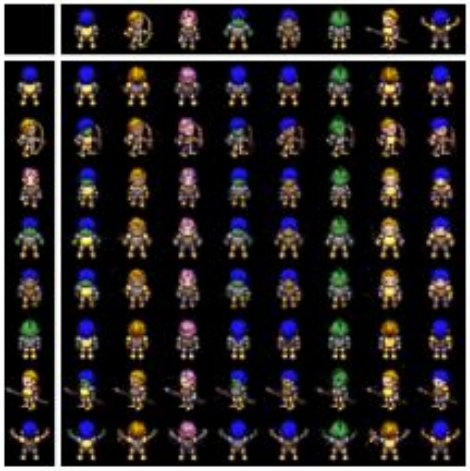
Active learning - for concepts

Unsupervised scenario

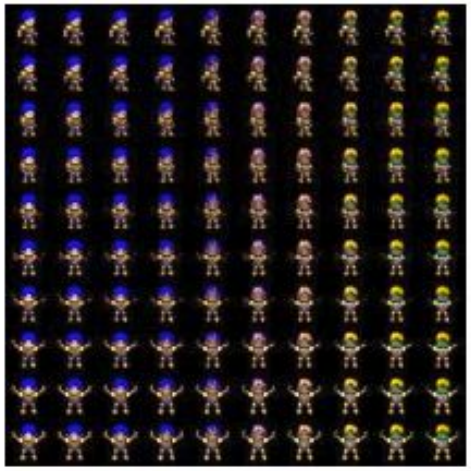
- Generative models
- Capsule networks

[Rudin et al. 2022]





[Mathieu et al. 2016]



Other topics

Interpretable dimensionality reduction?

Physics/causality integration with ML

Exploration of the Rashomon set

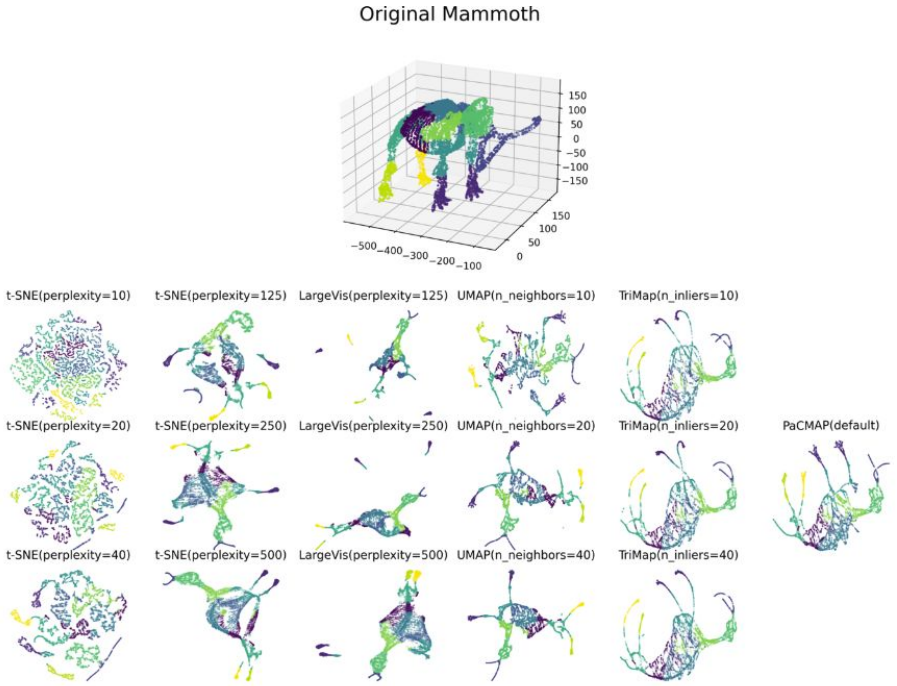
Interpretable RL - Interpretable policy

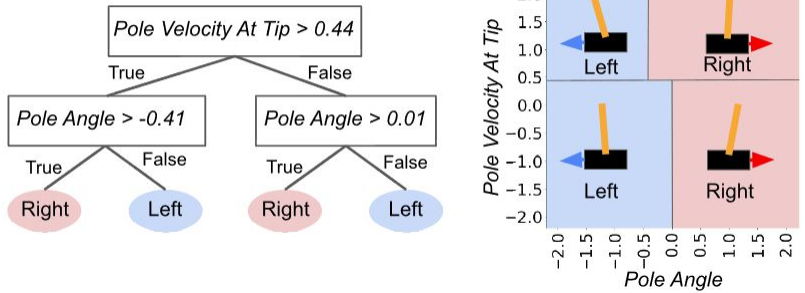
Models suitable for visualization

How to express uncertainty

How to measure Variable importance

...





Interpretability methods

Global - entire model

Feature visualization

Prototypes and Criticisms

Influential Instances

Global surrogate model

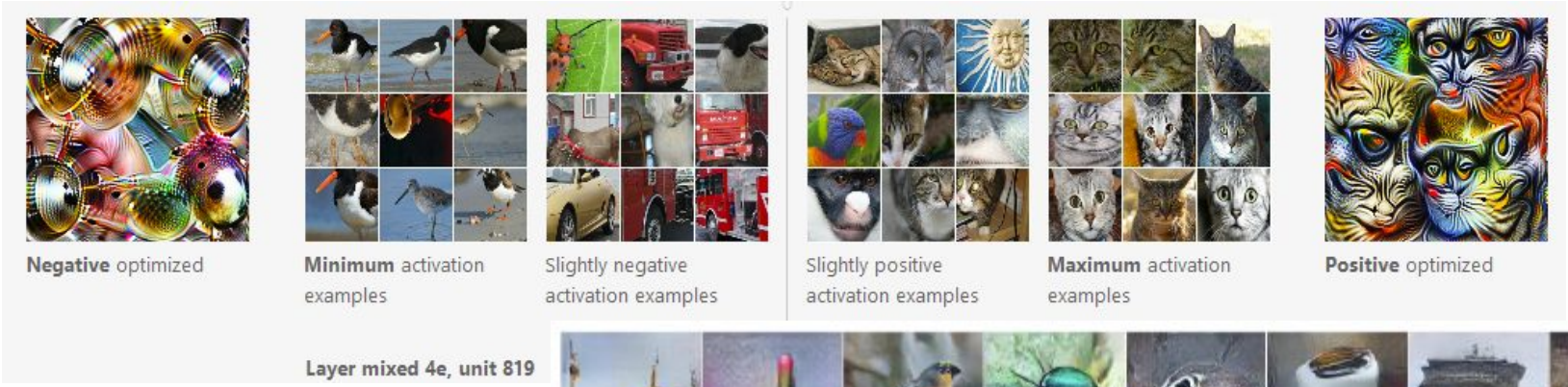
Concept attribution

Local - single instance

Not interpretable

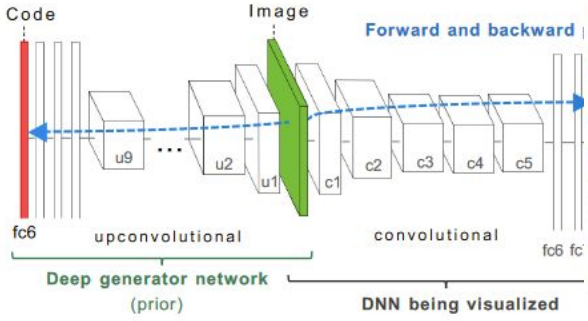
According to [Graziani et al. 2023]

Feature Visualizations

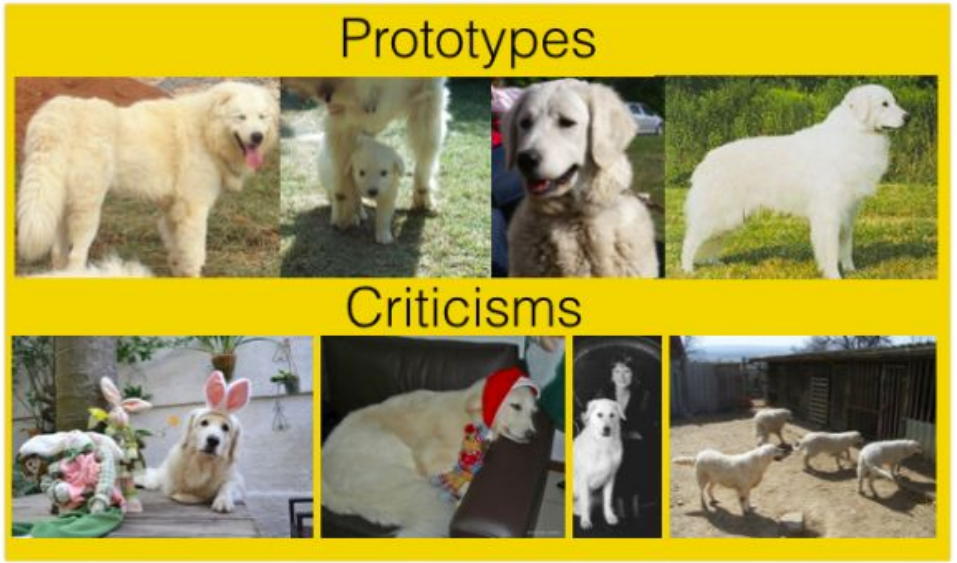


[Olah et al. 2017]

[Nguyen et al. 2016]



Prototypes and Criticisms

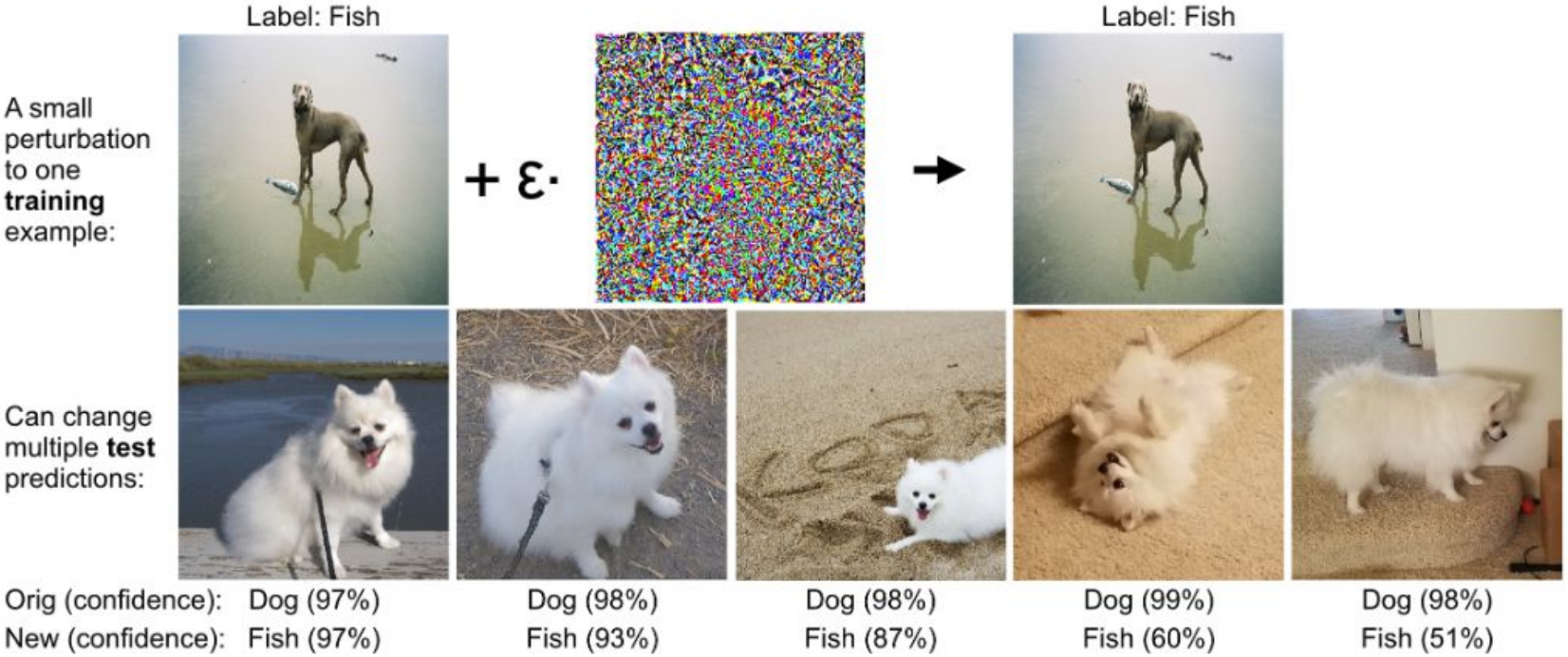


[Kim et al. 2016]

$$L(\mathsf{C}) = \sum_{l \in \mathsf{C}} \left| \frac{1}{n} \sum_{i \in [n]} k(x_i, x_l) - \frac{1}{m} \sum_{j \in \mathsf{S}} k(x_j, x_l) \right|$$

Influential Instances

[Koh et al. 2020]



Global surrogate model

Train an Interpretable model on output of a high performing Black-box model

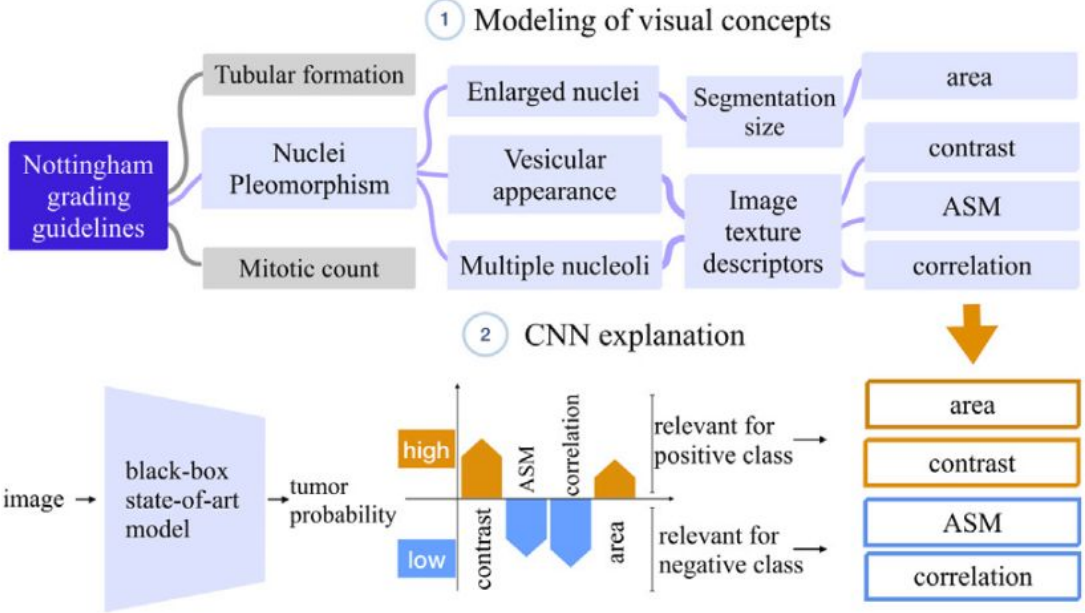
$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_*^{(i)} - \hat{y}^{(i)})^2}{\sum_{i=1}^{n}(\hat{y}^{(i)} - \bar{\hat{y}})^2}$$

[Molnar 2020]

Concept attribution

1) Prepare a set of concepts
2) Annotate some data with specific concepts
3) Check the model's sensitivity to a concept

[Graziani et al. 2020]



Post-hoc global explanation = Interpretation?

Lakkaraju et al. show that one can create highly misleading global explanation

If Race ≠ African American:
    If Prior-Felony = Yes and Crime-Status = Active, then **Risky**
    If Prior-Convictions = 0, then **Not Risky**

If Race = African American:
    If Pays-rent = No and Gender = Male, then **Risky**
    If Lives-with-Partner = No and College = No, then **Risky**
    If Age ≥35 and Has-Kids = Yes, then **Not Risky**
    If Wages ≥70K, then **Not Risky**

Default: **Not Risky**

If Current-Offense = Felony:
    If Prior-FTA = Yes and Prior-Arrests ≥ 1, then **Risky**
    If Crime-Status = Active and Owns-House = No and Has-Kids = No, then **Risky**
    If Prior-Convictions = 0 and College = Yes and Owns-House = Yes, then **Not Risky**

If Current-Offense = Misdemeanor and Prior-Arrests > 1:
    If Prior-Jail-Incarcerations = Yes, then **Risky**
    If Has-Kids = Yes and Married = Yes and Owns-House = Yes, then **Not Risky**
    If Lives-with-Partner = Yes and College = Yes and Pays-Rent = Yes, then **Not Risky**

If Current-Offense = Misdemeanor and Prior-Arrests ≤ 1:
    If Has-Kids = No and Owns-House = No and Prior-Jail-Incarcerations = Yes, then **Risky**
    If Age ≥ 50 and Has-Kids = Yes and Prior-FTA = No, then **Not Risky**

Default: **Not Risky**

[Lakkaraju et. al. 2020]

Recap

- An AI system is *interpretable* if it is possible to **translate its working principles** and outcomes **in human-understandable** language *without* affecting the **validity** of the system

- Interpretability means more work, for possibly more knowledge

- Interpretable models fit the definition best
- Interpretability methods are better than nothing

- Beware of potential for misinterpretation

References

[**Rudin et al. 2022**] - Rudin, Cynthia, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. “Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges.” Statistics Surveys 16, no. none (January 2022): 1–85. https://doi.org/10.1214/21-SS133.

[**Graziani et al. 2023**] - Graziani, Mara, Lidia Dutkiewicz, Davide Calvaresi, José Pereira Amorim, Katerina Yordanova, Mor Vered, Rahul Nair, et al. “A Global Taxonomy of Interpretable AI: Unifying the Terminology for the Technical and Social Sciences.” Artificial Intelligence Review 56, no. 4 (April 1, 2023): 3473–3504. https://doi.org/10.1007/s10462-022-10256-8.

[**Rudin 2019**] - Rudin, Cynthia. “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead.” Nature Machine Intelligence 1, no. 5 (May 2019): 206–15. https://doi.org/10.1038/s42256-019-0048-x.

[**Afnan et al. 2021**] - Afnan, Michael Anis Mihdi, Cynthia Rudin, Vincent Conitzer, Julian Savulescu, Abhishek Mishra, Yanhe Liu, and Masoud Afnan. “Ethical Implementation of Artificial Intelligence to Select Embryos in In Vitro Fertilization.” In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 316–26. Virtual Event USA: ACM, 2021. https://doi.org/10.1145/3461702.3462589.

[**Mathieu et al. 2016**] - Mathieu, Michael F, Junbo Jake Zhao, Junbo Zhao, Aditya Ramesh, Pablo Sprechmann, and Yann LeCun. “Disentangling Factors of Variation in Deep Representation Using Adversarial Training.” In Advances in Neural Information Processing Systems, Vol. 29. Curran Associates, Inc., 2016. https://proceedings.neurips.cc/paper/2016/hash/ef0917ea498b1665ad6c701057155abe-Abstract.html.

[**Nguyen et al. 2016**] - Nguyen, Anh, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks." Advances in neural information processing systems 29 (2016).

References

[**Olah et al. 2017**] - Olah, et al., "Feature Visualization", Distill, 2017. https://distill.pub/2017/feature-visualization/

[**Kim et al. 2016**] - Kim, Been, Rajiv Khanna, and Oluwasanmi Koyejo. “Examples Are Not Enough, Learn to Criticize! Criticism for Interpretability.” In Proceedings of the 30th International Conference on Neural Information Processing Systems, 2288–96. NIPS’16. Red Hook, NY, USA: Curran Associates Inc., 2016.

[**Koh et al. 2020**] - Koh, Pang Wei, and Percy Liang. “Understanding Black-Box Predictions via Influence Functions.” arXiv, December 29, 2020. http://arxiv.org/abs/1703.04730.

[**Graziani et al. 2020**] - M., Graziani, Andrearczyk V., Marchand-Maillet S., and Müller H. “Concept Attribution: Explaining CNN Decisions to Physicians.” Computers in Biology and Medicine 123 (August 1, 2020): 103865. https://doi.org/10.1016/j.compbiomed.2020.103865.

[**Molnar 2020**] - Molnar, Christoph. Interpretable machine learning. Lulu. com, 2020.

[**Lakkaraju et al. 2020**] - Lakkaraju, Himabindu, and Osbert Bastani. “‘How Do I Fool You?’: Manipulating User Trust via Misleading Black Box Explanations.” In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 79–85. AIES ’20. New York, NY, USA: Association for Computing Machinery, 2020. https://doi.org/10.1145/3375627.3375833.