Definitions, methods, and applications in interpretable ML

Murdoch W. J., Singh C., Kumbier K., Abbasi-Asl R., Yu B.
PNAS, 2020

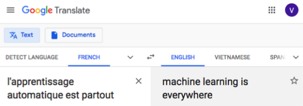
Advanced Machine Learning
12th March, 2020

Vân Anh Huynh-Thu
vahuynh@uliege.be

LIÈGE université

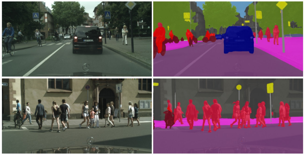
Machine learning today

ML techniques show outstanding performance on various tasks.



l'apprentissage automatique est partout

machine learning is everywhere



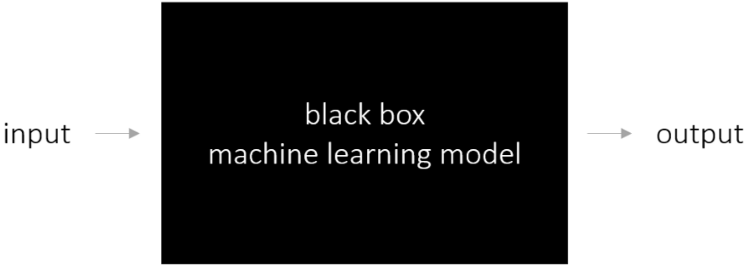
Kundu *et al.*, CVPR 2016

But several problems remain unsolved.

One problem is the black-box nature of ML tools

Current trends:

- Focus mainly on the predictive performance of the models
- Automatize as much as possible the whole learning procedure



input →

→ output

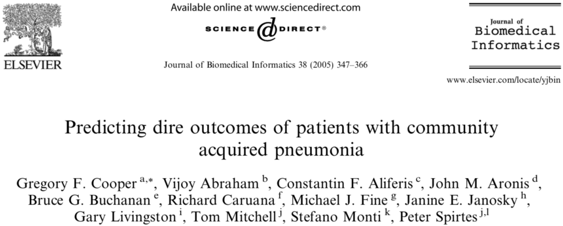
Consequence: most of the times, we do not know why ML models make certain predictions.

Model interpretability is important

Understanding the predictions of a model allows to:

- Understand the limitations of the model and improve it
- Get new scientific knowledge about the studied problem
- Build trust with domain experts
- ...

The pneumonia dataset example



Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Biomedical Informatics 38 (2005) 347–366

Journal of Biomedical Informatics

www.elsevier.com/locate/yjbin

Predicting dire outcomes of patients with community acquired pneumonia

Gregory F. Cooper[a,*], Vijoy Abraham[b], Constantin F. Aliferis[c], John M. Aronis[d],
Bruce G. Buchanan[e], Richard Caruana[f], Michael J. Fine[g], Janine E. Janosky[h],
Gary Livingston[i], Tom Mitchell[j], Stefano Monti[k], Peter Spirtes[j,l]

Goal: Predict the probability of death for patients with pneumonia, using various clinical features (age, heart rate, blood pressure, white blood cell count, etc.).

Application: Reduce healthcare system costs by prioritizing patients with high death risk for hospital admission.

The pneumonia dataset example

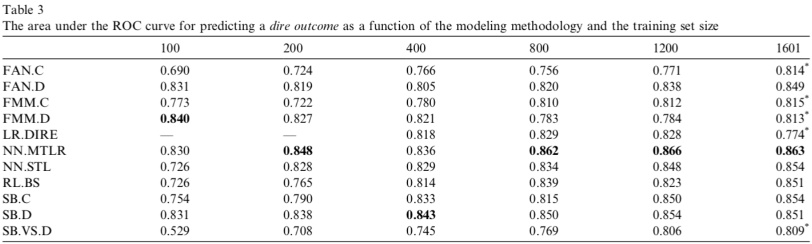


Table 3
The area under the ROC curve for predicting a *dire outcome* as a function of the modeling methodology and the training set size

	100	200	400	800	1200	1601
FAN.C	0.690	0.724	0.766	0.756	0.771	0.814*
FAN.D	0.831	0.819	0.805	0.820	0.838	0.849
FMM.C	0.773	0.722	0.780	0.810	0.812	0.815*
FMM.D	**0.840**	0.827	0.821	0.783	0.784	0.813*
LR.DIRE	—	—	0.818	0.829	0.828	0.774*
NN.MTLR	0.830	**0.848**	0.836	**0.862**	**0.866**	**0.863**
NN.STL	0.726	0.828	0.829	0.834	0.848	0.854
RL.BS	0.726	0.765	0.814	0.839	0.823	0.851
SB.C	0.754	0.790	0.833	0.815	0.850	0.854
SB.D	0.831	0.838	**0.843**	0.850	0.854	0.851
SB.VS.D	0.529	0.708	0.745	0.769	0.806	0.809*

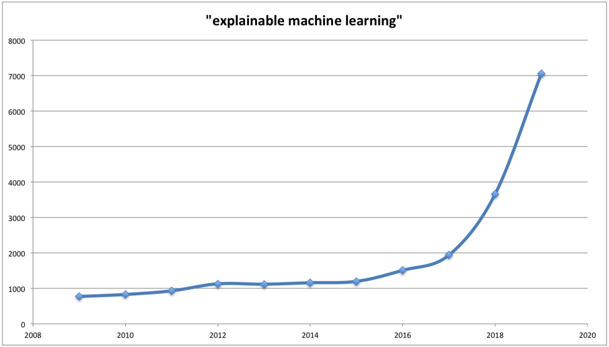
Multi-task neural nets yield the highest AUC.

But: by analysing the (intelligible) rule-based model, researchers found out that patients with asthma have a lower risk of dying from pneumonia!

Pattern in the data: patients with asthma had received an aggressive care that lowered the death risk.

Interpretability is a growing field in ML

Number of hits in Google Scholar:



A formal definition of interpretability (in ML) is still lacking

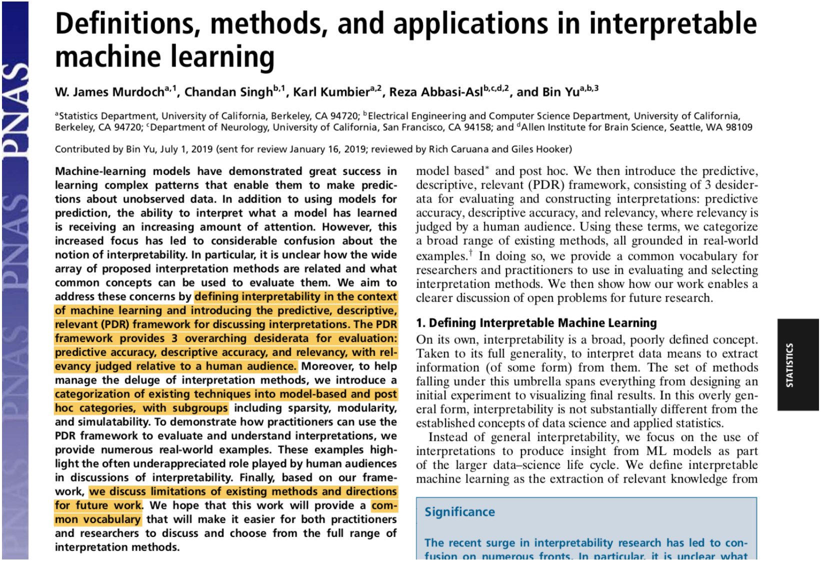
General definition: interpretability is the extraction of information (of some form).

Various methods have been proposed with various outputs (e.g. in the form of a mathematical model or some result vizualization).

And they all claim to be interpretation methods.

⟶ Open questions:

- What is an interpretation method?
- Are there common threads among the various methods?



PNAS

Definitions, methods, and applications in interpretable machine learning

W. James Murdoch[a,1], Chandan Singh[b,1], Karl Kumbier[a,2], Reza Abbasi-Asl[b,c,d,2], and Bin Yu[a,b,3]

[a]Statistics Department, University of California, Berkeley, CA 94720; [b]Electrical Engineering and Computer Science Department, University of California, Berkeley, CA 94720; [c]Department of Neurology, University of California, San Francisco, CA 94158; and [d]Allen Institute for Brain Science, Seattle, WA 98109

Contributed by Bin Yu, July 1, 2019 (sent for review January 16, 2019; reviewed by Rich Caruana and Giles Hooker)

Machine-learning models have demonstrated great success in learning complex patterns that enable them to make predictions about unobserved data. In addition to using models for prediction, the ability to interpret what a model has learned is receiving an increasing amount of attention. However, this increased focus has led to considerable confusion about the notion of interpretability. In particular, it is unclear how the wide array of proposed interpretation methods are related and what common concepts can be used to evaluate them. We aim to address these concerns by defining interpretability in the context of machine learning and introducing the predictive, descriptive, relevant (PDR) framework for discussing interpretations. The PDR framework provides 3 overarching desiderata for evaluation: predictive accuracy, descriptive accuracy, and relevancy, with relevancy judged relative to a human audience. Moreover, to help manage the deluge of interpretation methods, we introduce a categorization of existing techniques into model-based and post hoc categories, with subgroups including sparsity, modularity, and simulatability. To demonstrate how practitioners can use the PDR framework to evaluate and understand interpretations, we provide numerous real-world examples. These examples highlight the often underappreciated role played by human audiences in discussions of interpretability. Finally, based on our framework, we discuss limitations of existing methods and directions for future work. We hope that this work will provide a common vocabulary that will make it easier for both practitioners and researchers to discuss and choose from the full range of interpretation methods.

model based[*] and post hoc. We then introduce the predictive, descriptive, relevant (PDR) framework, consisting of 3 desiderata for evaluating and constructing interpretations: predictive accuracy, descriptive accuracy, and relevancy, where relevancy is judged by a human audience. Using these terms, we categorize a broad range of existing methods, all grounded in real-world examples.[†] In doing so, we provide a common vocabulary for researchers and practitioners to use in evaluating and selecting interpretation methods. We then show how our work enables a clearer discussion of open problems for future research.

1. Defining Interpretable Machine Learning

On its own, interpretability is a broad, poorly defined concept. Taken to its full generality, to interpret data means to extract information (of some form) from them. The set of methods falling under this umbrella spans everything from designing an initial experiment to visualizing final results. In this overly general form, interpretability is not substantially different from the established concepts of data science and applied statistics.

Instead of general interpretability, we focus on the use of interpretations to produce insight from ML models as part of the larger data–science life cycle. We define interpretable machine learning as the extraction of relevant knowledge from

STATISTICS

Significance

The recent surge in interpretability research has led to confusion on numerous fronts. In particular, it is unclear what

Paper sections

1. Defining interpretable ML
2. Background
3. Interpretation in the data-science life cycle
4. The PDR desiderata for interpretations
5. Model-based interpretability
6. Post hoc interpretability
7. Future work

Paper sections

1. Defining interpretable ML
2. Background
3. Interpretation in the data-science life cycle
4. The PDR desiderata for interpretations
5. Model-based interpretability
6. Post hoc interpretability
7. Future work

Definition used throughout the paper

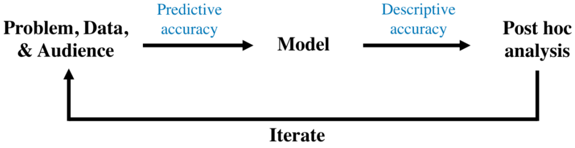
Interpretable ML = extraction of **relevant** knowledge from a ML model concerning **relationships** either contained in the **data** or **learned** by the model.

Knowledge is **relevant** if it provides insight **for a particular audience** into a chosen problem.

Paper sections

1. Defining interpretable ML
2. Background
3. Interpretation in the data-science life cycle
4. The PDR desiderata for interpretations
5. Model-based interpretability
6. Post hoc interpretability
7. Future work

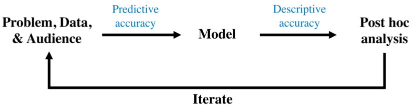
Two types of accuracies are defined



Predictive accuracy: ability of the model to fit the data

Descriptive accuracy: ability of the interpretations to describe what the model has learned

Interpretation occurs in the modelling and post hoc analysis stages



Model-based interpretability: use or construct models that readily provide insight into the learned relationships
→ lower predictive accuracy, but trivial post hoc analysis

Post hoc interpretability: extract information from the learned model without altering it
→ predictive accuracy does not change, but post hoc analysis can be heavy

The choice of the interpretation method depends on the nature of the problem and the targeted audience.

Paper sections

1. Defining interpretable ML
2. Background
3. Interpretation in the data-science life cycle
4. The PDR desiderata for interpretations
5. Model-based interpretability
6. Post hoc interpretability
7. Future work

Three properties are established to select and evaluate interpretation methods

An interpretation method should maximise these three properties:

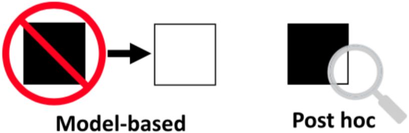
1. **P**redictive accuracy
2. **D**escriptive accuracy
3. **R**elevancy to a particular audience

Trade-off between predictive and descriptive accuracy:

Simple models: high descriptive accuracy, low predictive accuracy
Complex models: high predictive accuracy, low descriptive accuracy

Relevancy will often determine on which accuracy to focus.

Model-based vs post hoc interpretability



	Model-based interpretability	Post hoc interpretability
Predictive Accuracy	Generally unchanged or decrease (data-dependent)	No Effect
Descriptive Accuracy	Increase	Increase

Paper sections

1. Defining interpretable ML
2. Background
3. Interpretation in the data-science life cycle
4. The PDR desiderata for interpretations
5. Model-based interpretability
6. Post hoc interpretability
7. Future work

Model-based interpretability

$=$ use or construct models that readily provide insight into the learned relationships

Main challenge: construct models that are simple enough to be easily understood by the audience, while maintaining high predictive accuracy

Some common threads in model-based interpretability methods:

- Sparsity
- Simulatability
- Modularity

Sparse models

Constrain the model to be sparse by limiting the number of nonzero parameters.

Two examples:

1. Add a penalty term to the loss function (e.g. LASSO)

$$\min_{\mathbf{w}} L(y, \mathbf{w}^\top \mathbf{x}) + \lambda |\mathbf{w}|_1,$$

where L is the loss function and λ the regularisation coefficient.

2. Model selection using AIC or BIC

$$\mathrm{AIC} = 2k - \log \hat{L},$$

where k is the number of nonzero parameters and $\hat{L}$ is the likelihood.

Sparse models

Sparse models can improve the 3 interpretation desiderata:

- Descriptive accuracy: sparse models are easier to understand
- Predictive accuracy: increases if the problem is sparse
- Relevancy: nonzero parameters can be interpreted as being meaningfully related to the problem

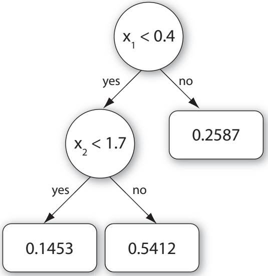
Drawbacks:

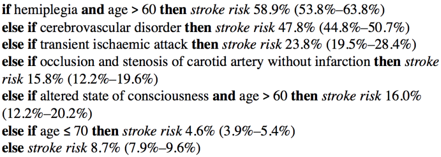
- Incorporating sparsity is not easy for all models
- Predictive accuracy decreases if the problem is not sparse
- The sparse parameter set will not be necessarily stable

Simulatable models

A model is simulatable if a human is able to go through every calculation required to make a prediction (in a reasonable time).

Examples: decision trees and rule lists





if hemiplegia **and** age > 60 **then** *stroke risk* 58.9% (53.8%–63.8%)
else if cerebrovascular disorder **then** *stroke risk* 47.8% (44.8%–50.7%)
else if transient ischaemic attack **then** *stroke risk* 23.8% (19.5%–28.4%)
else if occlusion and stenosis of carotid artery without infarction **then** *stroke risk* 15.8% (12.2%–19.6%)
else if altered state of consciousness **and** age > 60 **then** *stroke risk* 16.0% (12.2%–20.2%)
else if age ≤ 70 **then** *stroke risk* 4.6% (3.9%–5.4%)
else *stroke risk* 8.7% (7.9%–9.6%)

Simulatable models

Simulatable models yield:

- High descriptive accuracy
- High relevancy

Drawbacks:

- Very strong contraint to put on a model
- Low predictive accuracy if the problem is a bit complex
- Simulatability is lost when the model becomes large (e.g. high number of nodes in the tree or high number of rules)

Modular models

A model is modular if (a) meaningful portion(s) of its prediction-making process can be interpreted independently.

Example: the generalized additive model

$$g(\mathbb{E}[y]) = \beta_0 + \sum_j f_j(x_j) + \sum_{i \neq j} f_{ij}(x_i, x_j)$$

The univariate terms f_j can be interpreted in the form of curves.

The pairwise interaction terms f_{ij} can be interpreted in the form of heatmaps.

Modular models can provide some insights into the learned relationships, but has a lower descriptive accuracy than sparse and simulatable models.

Paper sections

1. Defining interpretable ML
2. Background
3. Interpretation in the data-science life cycle
4. The PDR desiderata for interpretations
5. Model-based interpretability
6. Post hoc interpretability
7. Future work

Post hoc interpretability

= extract information from the learned (black box) model without altering it.

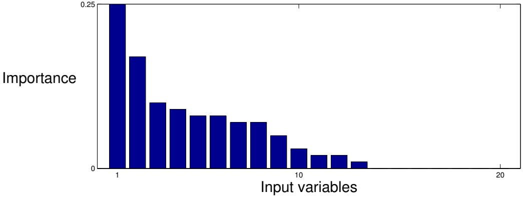
Two main categories of post hoc interpretation methods:

- **Dataset-level** interpretations explain **global** relationships learned by the model.
- **Prediction-level** interpretations explain **individual** predictions by the model.

Dataset-level interpretations can be obtained by examining a sufficient amount of prediction-level interpretations, but it is not always feasible.

Dataset-level interpretation: feature importances

Feature importance measures how much a feature contributes the prediction.



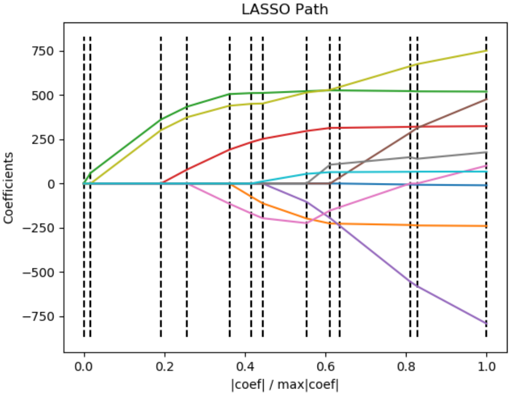
Methods have been developed to derive feature (interaction) importances from many models, such as Random forests or neural nets.

For some models (e.g. linear models), a statistical measure of confidence (e.g. *p*-value) can be computed in addition to the importance score, by making assumptions about the data-generating process.

Dataset-level interpretation: visualizations

The learned relationships are represented using some visualization technique.

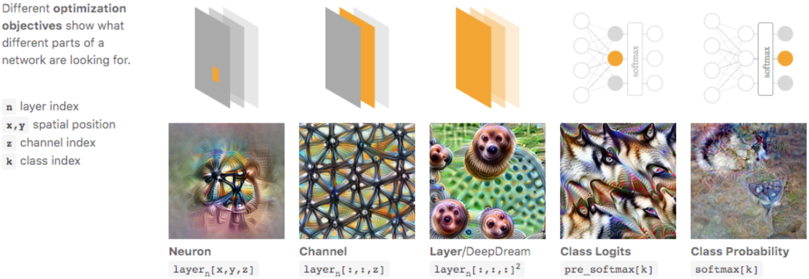
Example: the LASSO path



scikit-learn.org

Dataset-level interpretation: visualizations

Another example: some works have focused on representing visually what a convolutional neural network is looking for in an image to classify it.



Different **optimization objectives** show what different parts of a network are looking for.

`n`	layer index
`x,y`	spatial position
`z`	channel index
`k`	class index

softmax
softmax
Neuron layer$_n$[x,y,z]
Channel layer$_n$[:,:,z]
Layer/DeepDream layer$_n$[:,:,:]^2
Class Logits pre_softmax[k]
Class Probability softmax[k]

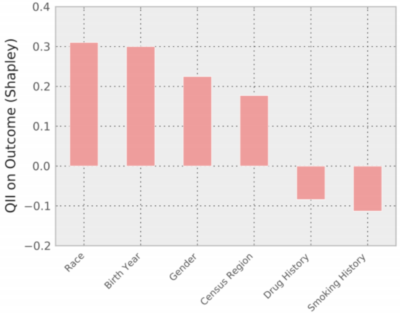
Olah *et al.*, *Distill* 2017

Prediction-level interpretation: feature importances

The importance of a feature can vary for different examples as a result of interactions with other features.

Prediction-levels importances can be used to check that a model is fair.

Example: features used to predict if “Mr. Z” is likely to be arrested in the future:



Paper sections

1. Defining interpretable ML
2. Background
3. Interpretation in the data-science life cycle
4. The PDR desiderata for interpretations
5. Model-based interpretability
6. Post hoc interpretability
7. Future work

How to evaluate an interpretation method?

Currently, there is no clear consensus in the community.

Within the PDR framework:

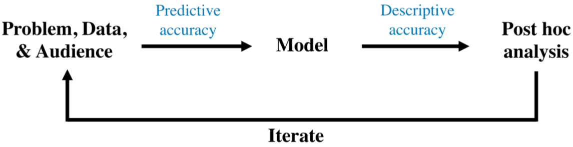
- Measuring predictive accuracy has been well-studied
- Measuring descriptive accuracy and relevancy is still a challenge

How to measure descriptive accuracy?

One possible solution is to use simulation studies.

With a sufficiently powerful model and a large amount of simulated data, the learned model should achieve near-perfect prediction accuracy.

→ Descriptive accuracy can be measured by checking whether the interpretations recover the aspects of the data-generation process.



How to measure relevancy?

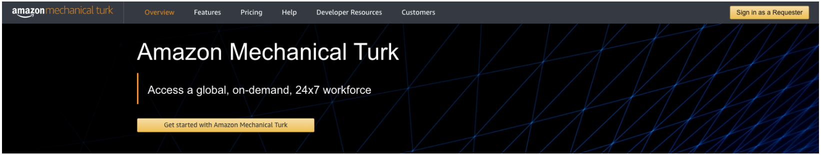
Two dominant approaches to show the relevancy of the output of an interpretation method:

1. Use the interpretations directly to solve a domain problem.

→ Relevancy is obvious in that case.

2. Conduct a human study, where humans are asked to perform certain tasks, such as evaluating how much they trust the model's predictions.

→ More challenging and only a general audience can be used.



Amazon Mechanical Turk

Access a global, on-demand, 24x7 workforce

Main challenges of interpretation methods

Model-based interpretability:

- Build interpretable models with a high predictive accuracy
- Develop tools for feature engineering
  (to make relationships to be learned simpler)
  Meaningful features can be constructed using either domain knowledge or automatically with unsupervised learning.

Post hoc interpretability:

- Close the gap between the information provided by the interpretations and what the model has actually learned
- Use the interpretations to improve predictive accuracy

Key points of the paper

They propose a definition of interpretable ML.

They propose three properties (PDR) to help choosing and evaluating an interpretation method.

They discuss existing interpretation methods, showing some common threads (model-based vs post hoc interpretations).