2 Interpretability

It is difficult to (mathematically) define interpretability. A (non-mathematical) definition of interpretability that I like by Miller (2017)[1] is: **Interpretability is the degree to which a human can understand the cause of a decision.** Another one is: **Interpretability is the degree to which a human can consistently predict the model's result** [2]. The higher the interpretability of a machine learning model, the easier it is for someone to comprehend why certain decisions or predictions have been made. A model is better interpretable than another model if its decisions are easier for a human to comprehend than decisions from the other model. I will use both the terms interpretable and explainable interchangeably. Like Miller (2017), I think it makes sense to distinguish between the terms interpretability/explainability and explanation. I will use "explanation" for explanations of individual predictions. See the section about explanations to learn what we humans see as a good explanation.

Interpretable machine learning is a useful umbrella term that captures the "extraction of relevant knowledge from a machine-learning model concerning relationships either contained in data or learned by the model". [3]

2.1 Importance of Interpretability

If a machine learning model performs well, **why do we not just trust the model** and ignore **why** it made a certain decision? "The problem is that a single metric, such as classification accuracy, is an incomplete description of most real-world tasks." (Doshi-Velez and Kim 2017 [4])

Let us dive deeper into the reasons why interpretability is so important. When it comes to predictive modeling, you have to make a trade-off: Do you just want to know **what**

[1] Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." arXiv Preprint arXiv:1706.07269. (2017).

[2] Kim, Been, Rajiv Khanna, and Oluwasanmi O. Koyejo. "Examples are not enough, learn to criticize! Criticism for interpretability." Advances in Neural Information Processing Systems (2016).

[3] Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. "Definitions, methods, and applications in interpretable machine learning." Proceedings of the National Academy of Sciences, 116(44), 22071-22080. (2019).

[4] Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning," no. Ml: 1–13. http://arxiv.org/abs/1702.08608 (2017).

is predicted? For example, the probability that a customer will churn or how effective some drug will be for a patient. Or do you want to know **why** the prediction was made and possibly pay for the interpretability with a drop in predictive performance? In some cases, you do not care why a decision was made, it is enough to know that the predictive performance on a test dataset was good. But in other cases, knowing the 'why' can help you learn more about the problem, the data and the reason why a model might fail. Some models may not require explanations because they are used in a low-risk environment, meaning a mistake will not have serious consequences, (e.g. a movie recommender system) or the method has already been extensively studied and evaluated (e.g. optical character recognition). The need for interpretability arises from an incompleteness in problem formalization (Doshi-Velez and Kim 2017), which means that for certain problems or tasks it is not enough to get the prediction (the **what**). The model must also explain how it came to the prediction (the **why**), because a correct prediction only partially solves your original problem. The following reasons drive the demand for interpretability and explanations (Doshi-Velez and Kim 2017 and Miller 2017).

Human curiosity and learning: Humans have a mental model of their environment that is updated when something unexpected happens. This update is performed by finding an explanation for the unexpected event. For example, a human feels unexpectedly sick and asks, "Why do I feel so sick?". He learns that he gets sick every time he eats those red berries. He updates his mental model and decides that the berries caused the sickness and should therefore be avoided. When opaque machine learning models are used in research, scientific findings remain completely hidden if the model only gives predictions without explanations. To facilitate learning and satisfy curiosity as to why certain predictions or behaviors are created by machines, interpretability and explanations are crucial. Of course, humans do not need explanations for everything that happens. For most people it is okay that they do not understand how a computer works. Unexpected events makes us curious. For example: Why is my computer shutting down unexpectedly?

Closely related to learning is the human desire to **find meaning in the world**. We want to harmonize contradictions or inconsistencies between elements of our knowledge structures. "Why did my dog bite me even though it has never done so before?" a human might ask. There is a contradiction between the knowledge of the dog's past behavior and the newly made, unpleasant experience of the bite. The vet's explanation reconciles the dog owner's contradiction: "The dog was under stress and bit." The more a machine's decision affects a person's life, the more important it is for the machine to explain its behavior. If a machine learning model rejects a loan application, this may be completely unexpected for the applicants. They can only reconcile this inconsistency between expectation and reality with some kind of explanation. The explanations do not actually have to fully explain the situation, but should address a main cause. Another example is algorithmic product recommendation. Personally, I always think about why certain products or movies have been algorithmically recommended to me. Often it is quite clear: Advertising follows me on the Internet because I recently bought a washing machine,

and I know that in the next days I will be followed by advertisements for washing machines. Yes, it makes sense to suggest gloves if I already have a winter hat in my shopping cart. The algorithm recommends this movie, because users who liked other movies I liked also enjoyed the recommended movie. Increasingly, Internet companies are adding explanations to their recommendations. A good example are product recommendations, which are based on frequently purchased product combinations:

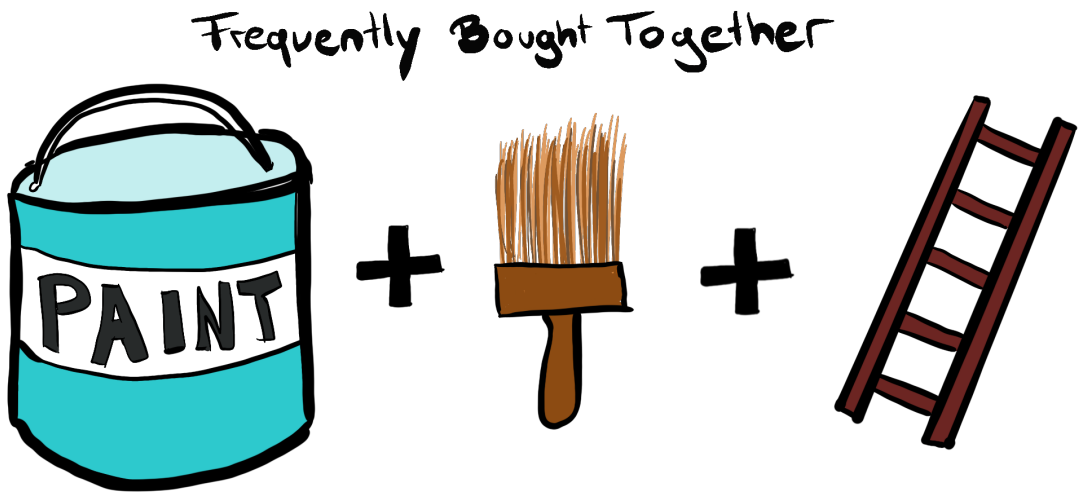

Figure 2.1: Recommended products that are frequently bought together.

In many scientific disciplines there is a change from qualitative to quantitative methods (e.g. sociology, psychology), and also towards machine learning (biology, genomics). The **goal of science** is to gain knowledge, but many problems are solved with big datasets and black box machine learning models. 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.

Machine learning models take on real-world tasks that require **safety measures** and testing. Imagine a self-driving car automatically detects cyclists based on a deep learning system. You want to be 100% sure that the abstraction the system has learned is error-free, because running over cyclists is quite bad. An explanation might reveal that the most important learned feature is to recognize the two wheels of a bicycle, and this explanation helps you think about edge cases like bicycles with side bags that partially cover the wheels.

By default, machine learning models pick up biases from the training data. This can turn your machine learning models into racists that discriminate against underrepresented groups. Interpretability is a useful debugging tool for **detecting bias** in machine learning

models. It might happen that the machine learning model you have trained for automatic approval or rejection of credit applications discriminates against a minority that has been historically disenfranchised. Your main goal is to grant loans only to people who will eventually repay them. The incompleteness of the problem formulation in this case lies in the fact that you not only want to minimize loan defaults, but are also obliged not to discriminate on the basis of certain demographics. This is an additional constraint that is part of your problem formulation (granting loans in a low-risk and compliant way) that is not covered by the loss function the machine learning model was optimized for.

The process of integrating machines and algorithms into our daily lives requires interpretability to increase **social acceptance**. People attribute beliefs, desires, intentions and so on to objects. In a famous experiment, Heider and Simmel (1944) [5] showed participants videos of shapes in which a circle opened a "door" to enter a "room" (which was simply a rectangle). The participants described the actions of the shapes as they would describe the actions of a human agent, assigning intentions and even emotions and personality traits to the shapes. Robots are a good example, like my vacuum cleaner, which I named "Doge". If Doge gets stuck, I think: "Doge wants to keep cleaning, but asks me for help because it got stuck." Later, when Doge finishes cleaning and searches the home base to recharge, I think: "Doge has a desire to recharge and intends to find the home base." I also attribute personality traits: "Doge is a bit dumb, but in a cute way." These are my thoughts, especially when I find out that Doge has knocked over a plant while dutifully vacuuming the house. A machine or algorithm that explains its predictions will find more acceptance. See also the chapter on explanations, which argues that explanations are a social process.

Explanations are used to **manage social interactions**. By creating a shared meaning of something, the explainer influences the actions, emotions and beliefs of the recipient of the explanation. For a machine to interact with us, it may need to shape our emotions and beliefs. Machines have to "persuade" us, so that they can achieve their intended goal. I would not fully accept my robot vacuum cleaner if it did not explain its behavior to some degree. The vacuum cleaner creates a shared meaning of, for example, an "accident" (like getting stuck on the bathroom carpet ... again) by explaining that it got stuck instead of simply stopping to work without comment. Interestingly, there may be a misalignment between the goal of the explaining machine (create trust) and the goal of the recipient (understand the prediction or behavior). Perhaps the full explanation for why Doge got stuck could be that the battery was very low, that one of the wheels is not working properly and that there is a bug that makes the robot go to the same spot over and over again even though there was an obstacle. These reasons (and a few more) caused the robot to get stuck, but it only explained that something was in the way, and that was enough for me to trust its behavior and get a shared meaning of that accident. By the

[5] Heider, Fritz, and Marianne Simmel. "An experimental study of apparent behavior." The American Journal of Psychology 57 (2). JSTOR: 243–59. (1944).

way, Doge got stuck in the bathroom again. We have to remove the carpets each time before we let Doge vacuum.

Figure 2.2: Doge, our vacuum cleaner, got stuck. As an explanation for the accident, Doge told us that it needs to be on an even surface.

Machine learning models can only be **debugged and audited** when they can be interpreted. Even in low risk environments, such as movie recommendations, the ability to interpret is valuable in the research and development phase as well as after deployment. Later, when a model is used in a product, things can go wrong. An interpretation for an erroneous prediction helps to understand the cause of the error. It delivers a direction for how to fix the system. Consider an example of a husky versus wolf classifier that misclassifies some huskies as wolves. Using interpretable machine learning methods, you would find that the misclassification was due to the snow on the image. The classifier learned to use snow as a feature for classifying images as "wolf", which might make sense in terms of separating wolves from huskies in the training dataset, but not in real-world use.

If you can ensure that the machine learning model can explain decisions, you can also check the following traits more easily (Doshi-Velez and Kim 2017):

- Fairness: Ensuring that predictions are unbiased and do not implicitly or explicitly discriminate against underrepresented groups. An interpretable model can tell you why it has decided that a certain person should not get a loan, and it becomes easier for a human to judge whether the decision is based on a learned demographic (e.g. racial) bias.

audit 审计；审查

trait 特性

discriminate 区分；歧视

discriminate against 排斥；歧视

underrepresented 未被充分代表

- Privacy: Ensuring that sensitive information in the data is protected.
- Reliability or Robustness: Ensuring that small changes in the input do not lead to large changes in the prediction.
- Causality: Check that only causal relationships are picked up.
- Trust: It is easier for humans to trust a system that explains its decisions compared to a black box.

causality 因果性
causal 因果关系的

When we do not need interpretability.

The following scenarios illustrate when we do not need or even do not want interpretability of machine learning models.

Interpretability is not required if the model **has no significant impact**. Imagine someone named Mike working on a machine learning side project to predict where his friends will go for their next holidays based on Facebook data. Mike just likes to surprise his friends with educated guesses where they will be going on holidays. There is no real problem if the model is wrong (at worst just a little embarrassment for Mike), nor is there a problem if Mike cannot explain the output of his model. It is perfectly fine not to have interpretability in this case. The situation would change if Mike started building a business around these holiday destination predictions. If the model is wrong, the business could lose money, or the model may work worse for some people because of learned racial bias. As soon as the model has a significant impact, be it financial or social, interpretability becomes relevant.

Interpretability is not required when the **problem is well studied**. Some applications have been sufficiently well studied so that there is enough practical experience with the model and problems with the model have been solved over time. A good example is a machine learning model for optical character recognition that processes images from envelopes and extracts addresses. There is years of experience with these systems and it is clear that they work. In addition, we are not really interested in gaining additional insights about the task at hand.

Interpretability might enable people or programs to **manipulate the system**. Problems with users who deceive a system result from a mismatch between the goals of the creator and the user of a model. Credit scoring is such a system because banks want to ensure that loans are only given to applicants who are likely to return them, and applicants aim to get the loan even if the bank does not want to give them one. This mismatch between the goals introduces incentives for applicants to game the system to increase their chances of getting a loan. If an applicant knows that having more than two credit cards negatively affects his score, he simply returns his third credit card to improve his score, and organizes a new card after the loan has been approved. While his score improved, the actual probability of repaying the loan remained unchanged. The system can only be gamed if the inputs are proxies for a causal feature, but do not actually cause the outcome. Whenever possible, proxy features should be avoided as they make

models gameable. For example, Google developed a system called Google Flu Trends to predict flu outbreaks. The system correlated Google searches with flu outbreaks – and it has performed poorly. The distribution of search queries changed and Google Flu Trends missed many flu outbreaks. Google searches do not cause the flu. When people search for symptoms like "fever" it is merely a correlation with actual flu outbreaks. Ideally, models would only use causal features because they would not be gameable.

2.2 Taxonomy of Interpretability Methods

Methods for machine learning interpretability can be classified according to various criteria.

Intrinsic or post hoc? This criteria distinguishes whether interpretability is achieved by restricting the complexity of the machine learning model (intrinsic) or by applying methods that analyze the model after training (post hoc). Intrinsic interpretability refers to machine learning models that are considered interpretable due to their simple structure, such as short decision trees or sparse linear models. Post hoc interpretability refers to the application of interpretation methods after model training. Permutation feature importance is, for example, a post hoc interpretation method. Post hoc methods can also be applied to intrinsically interpretable models. For example, permutation feature importance can be computed for decision trees. The organization of the chapters in this book is determined by the distinction between intrinsically interpretable models and post hoc (and model-agnostic) interpretation methods.

Result of the interpretation method The various interpretation methods can be roughly differentiated according to their results.

- **Feature summary statistic**: Many interpretation methods provide summary statistics for each feature. Some methods return a single number per feature, such as feature importance, or a more complex result, such as the pairwise feature interaction strengths, which consist of a number for each feature pair.
- **Feature summary visualization**: Most of the feature summary statistics can also be visualized. Some feature summaries are actually only meaningful if they are visualized and a table would be a wrong choice. The partial dependence of a feature is such a case. Partial dependence plots are curves that show a feature and the average predicted outcome. The best way to present partial dependences is to actually draw the curve instead of printing the coordinates.
- **Model internals (e.g. learned weights)**: The interpretation of intrinsically interpretable models falls into this category. Examples are the weights in linear models or the learned tree structure (the features and thresholds used for the splits) of decision trees. The lines are blurred between model internals and feature summary

taxonomy 分类学；分类法

intrinsic 内在的
post hoc 事后的

permutation 排列；置换

pairwise

PDP

statistic in, for example, linear models, because the weights are both model internals and summary statistics for the features at the same time. Another method that outputs model internals is the visualization of feature detectors learned in convolutional neural networks. Interpretability methods that output model internals are by definition model-specific (see next criterion).

- **Data point**: This category includes all methods that return data points (already existent or newly created) to make a model interpretable. One method is called counterfactual explanations. To explain the prediction of a data instance, the method finds a similar data point by changing some of the features for which the predicted outcome changes in a relevant way (e.g. a flip in the predicted class). Another example is the identification of prototypes of predicted classes. To be useful, interpretation methods that output new data points require that the data points themselves can be interpreted. This works well for images and texts, but is less useful for tabular data with hundreds of features.
- **Intrinsically interpretable model**: One solution to interpreting black box models is to approximate them (either globally or locally) with an interpretable model. The interpretable model itself is interpreted by looking at internal model parameters or feature summary statistics.

用另一个可解释的模型去模仿并解释一个黑盒模型

Model-specific or model-agnostic? Model-specific interpretation tools are limited to specific model classes. The interpretation of regression weights in a linear model is a model-specific interpretation, since – by definition – the interpretation of intrinsically interpretable models is always model-specific. Tools that only work for the interpretation of e.g. neural networks are model-specific. Model-agnostic tools can be used on any machine learning model and are applied after the model has been trained (post hoc). These agnostic methods usually work by analyzing feature input and output pairs. By definition, these methods cannot have access to model internals such as weights or structural information.

Local or global? Does the interpretation method explain an individual prediction or the entire model behavior? Or is the scope somewhere in between? Read more about the scope criterion in the next section.

Local : explain an individual prediction

Global : entire model behavior

2.3 Scope of Interpretability

An algorithm trains a model that produces the predictions. Each step can be evaluated in terms of transparency or interpretability.

可解释性的范围

2.3.1 Algorithm Transparency

How does the algorithm create the model?

Algorithm transparency is about how the algorithm learns a model from the data and what kind of relationships it can learn. If you use convolutional neural networks to classify images, you can explain that the algorithm learns edge detectors and filters on the lowest layers. This is an understanding of how the algorithm works, but not for the specific model that is learned in the end, and not for how individual predictions are made. Algorithm transparency only requires knowledge of the algorithm and not of the data or learned model. This book focuses on model interpretability and not algorithm transparency. Algorithms such as the least squares method for linear models are well studied and understood. They are characterized by a high transparency. Deep learning approaches (pushing a gradient through a network with millions of weights) are less well understood and the inner workings are the focus of ongoing research. They are considered less transparent.

2.3.2 Global, Holistic Model Interpretability

How does the trained model make predictions?

You could describe a model as interpretable if you can comprehend the entire model at once (Lipton 2016[6]). To explain the global model output, you need the trained model, knowledge of the algorithm and the data. This level of interpretability is about understanding how the model makes decisions, based on a holistic view of its features and each of the learned components such as weights, other parameters, and structures. Which features are important and what kind of interactions between them take place? Global model interpretability helps to understand the distribution of your target outcome based on the features. Global model interpretability is very difficult to achieve in practice. Any model that exceeds a handful of parameters or weights is unlikely to fit into the short-term memory of the average human. I argue that you cannot really imagine a linear model with 5 features, because it would mean drawing the estimated hyperplane mentally in a 5-dimensional space. Any feature space with more than 3 dimensions is simply inconceivable for humans. Usually, when people try to comprehend a model, they consider only parts of it, such as the weights in linear models.

2.3.3 Global Model Interpretability on a Modular Level

How do parts of the model affect predictions?

A Naive Bayes model with many hundreds of features would be too big for me and you to keep in our working memory. And even if we manage to memorize all the weights, we would not be able to quickly make predictions for new data points. In addition, you need to have the joint distribution of all features in your head to estimate the importance of

[6]Lipton, Zachary C. "The mythos of model interpretability." arXiv preprint arXiv:1606.03490, (2016).

each feature and how the features affect the predictions on average. An impossible task. But you can easily understand a single weight. While global model interpretability is usually out of reach, there is a good chance of understanding at least some models on a modular level. Not all models are interpretable at a parameter level. For linear models, the interpretable parts are the weights, for trees it would be the splits (selected features plus cut-off points) and leaf node predictions. Linear models, for example, look like as if they could be perfectly interpreted on a modular level, but the interpretation of a single weight is interlocked with all other weights. The interpretation of a single weight always comes with the footnote that the other input features remain at the same value, which is not the case with many real applications. A linear model that predicts the value of a house, that takes into account both the size of the house and the number of rooms, can have a negative weight for the room feature. It can happen because there is already the highly correlated house size feature. In a market where people prefer larger rooms, a house with fewer rooms could be worth more than a house with more rooms if both have the same size. The weights only make sense in the context of the other features in the model. But the weights in a linear model can still be interpreted better than the weights of a deep neural network.

2.3.4 Local Interpretability for a Single Prediction

Why did the model make a certain prediction for an instance?

You can zoom in on a single instance and examine what the model predicts for this input, and explain why. If you look at an individual prediction, the behavior of the otherwise complex model might behave more pleasantly. Locally, the prediction might only depend linearly or monotonically on some features, rather than having a complex dependence on them. For example, the value of a house may depend nonlinearly on its size. But if you are looking at only one particular 100 square meters house, there is a possibility that for that data subset, your model prediction depends linearly on the size. You can find this out by simulating how the predicted price changes when you increase or decrease the size by 10 square meters. Local explanations can therefore be more accurate than global explanations. This book presents methods that can make individual predictions more interpretable in the section on model-agnostic methods.

2.3.5 Local Interpretability for a Group of Predictions

Why did the model make specific predictions for a group of instances?

Model predictions for multiple instances can be explained either with global model interpretation methods (on a modular level) or with explanations of individual instances. The global methods can be applied by taking the group of instances, treating them as if the

group were the complete dataset, and using the global methods with this subset. The individual explanation methods can be used on each instance and then listed or aggregated for the entire group.

2.4 Evaluation of Interpretability

There is no real consensus about what interpretability is in machine learning. Nor is it clear how to measure it. But there is some initial research on this and an attempt to formulate some approaches for evaluation, as described in the following section.

Doshi-Velez and Kim (2017) propose three main levels for the evaluation of interpretability:

Application level evaluation (real task): Put the explanation into the product and have it tested by the end user. Imagine fracture detection software with a machine learning component that locates and marks fractures in X-rays. At the application level, radiologists would test the fracture detection software directly to evaluate the model. This requires a good experimental setup and an understanding of how to assess quality. A good baseline for this is always how good a human would be at explaining the same decision.

Human level evaluation (simple task) is a simplified application level evaluation. The difference is that these experiments are not carried out with the domain experts, but with laypersons. This makes experiments cheaper (especially if the domain experts are radiologists) and it is easier to find more testers. An example would be to show a user different explanations and the user would choose the best one.

Function level evaluation (proxy task) does not require humans. This works best when the class of model used has already been evaluated by someone else in a human level evaluation. For example, it might be known that the end users understand decision trees. In this case, a proxy for explanation quality may be the depth of the tree. Shorter trees would get a better explainability score. It would make sense to add the constraint that the predictive performance of the tree remains good and does not decrease too much compared to a larger tree.

The next chapter focuses on the evaluation of explanations for individual predictions on the function level. What are the relevant properties of explanations that we would consider for their evaluation?

2.5 Properties of Explanations

We want to explain the predictions of a machine learning model. To achieve this, we rely on some explanation method, which is an algorithm that generates explanations.

An explanation usually relates the feature values of an instance to its model prediction in a humanly understandable way. Other types of explanations consist of a set of data instances (e.g in the case of the k-nearest neighbor model). For example, we could predict cancer risk using a support vector machine and explain predictions using the local surrogate method, which generates decision trees as explanations. Or we could use a linear regression model instead of a support vector machine. The linear regression model is already equipped with an explanation method (interpretation of the weights).

We take a closer look at the properties of explanation methods and explanations (Robnik-Sikonja and Bohanec, 2018[7]). These properties can be used to judge how good an explanation method or explanation is. It is not clear for all these properties how to measure them correctly, so one of the challenges is to formalize how they could be calculated.

Properties of Explanation Methods

- **Expressive Power** is the "language" or structure of the explanations the method is able to generate. An explanation method could generate IF-THEN rules, decision trees, a weighted sum, natural language or something else.
- **Translucency** describes how much the explanation method relies on looking into the machine learning model, like its parameters. For example, explanation methods relying on intrinsically interpretable models like the linear regression model (model-specific) are highly translucent. Methods only relying on manipulating inputs and observing the predictions have zero translucency. Depending on the scenario, different levels of translucency might be desirable. The advantage of high translucency is that the method can rely on more information to generate explanations. The advantage of low translucency is that the explanation method is more portable.
- **Portability** describes the range of machine learning models with which the explanation method can be used. Methods with a low translucency have a higher portability because they treat the machine learning model as a black box. Surrogate models might be the explanation method with the highest portability. Methods that only work for e.g. recurrent neural networks have low portability.
- **Algorithmic Complexity** describes the computational complexity of the method that generates the explanation. This property is important to consider when computation time is a bottleneck in generating explanations.

Properties of Individual Explanations

- **Accuracy**: How well does an explanation predict unseen data? High accuracy is especially important if the explanation is used for predictions in place of the machine learning model. Low accuracy can be fine if the accuracy of the machine learning model is also low, and if the goal is to explain what the black box model does. In this case, only fidelity is important.

[7] Robnik-Sikonja, Marko, and Marko Bohanec. "Perturbation-based explanations of prediction models." Human and Machine Learning. Springer, Cham. 159-175. (2018).

LIME

解释方法的特点、

Translucency 半透明

只靠输入输出来作出解释的方法半透明度为 0

Surrogate 代理人(品)

单个解释的特点、

准确性：

预测一个没见过的数据

- **Fidelity**: How well does the explanation approximate the prediction of the black box model? High fidelity is one of the most important properties of an explanation, because an explanation with low fidelity is useless to explain the machine learning model. Accuracy and fidelity are closely related. If the black box model has high accuracy and the explanation has high fidelity, the explanation also has high accuracy. Some explanations offer only local fidelity, meaning the explanation only approximates well to the model prediction for a subset of the data (e.g. local surrogate models) or even for only an individual data instance (e.g. Shapley Values).
- **Consistency**: How much does an explanation differ between models that have been trained on the same task and that produce similar predictions? For example, I train a support vector machine and a linear regression model on the same task and both produce very similar predictions. I compute explanations using a method of my choice and analyze how different the explanations are. If the explanations are very similar, the explanations are highly consistent. I find this property somewhat tricky, since the two models could use different features, but get similar predictions (also called "Rashomon Effect"[8]). In this case a high consistency is not desirable because the explanations have to be very different. High consistency is desirable if the models really rely on similar relationships.
- **Stability**: How similar are the explanations for similar instances? While consistency compares explanations between models, stability compares explanations between similar instances for a fixed model. High stability means that slight variations in the features of an instance do not substantially change the explanation (unless these slight variations also strongly change the prediction). A lack of stability can be the result of a high variance of the explanation method. In other words, the explanation method is strongly affected by slight changes of the feature values of the instance to be explained. A lack of stability can also be caused by non-deterministic components of the explanation method, such as a data sampling step, like the local surrogate method uses. High stability is always desirable.
- **Comprehensibility**: How well do humans understand the explanations? This looks just like one more property among many, but it is the elephant in the room. Difficult to define and measure, but extremely important to get right. Many people agree that comprehensibility depends on the audience. Ideas for measuring comprehensibility include measuring the size of the explanation (number of features with a non-zero weight in a linear model, number of decision rules, ...) or testing how well people can predict the behavior of the machine learning model from the explanations. The comprehensibility of the features used in the explanation should also be considered. A complex transformation of features might be less comprehensible than the original features.
- **Certainty**: Does the explanation reflect the certainty of the machine learning model? Many machine learning models only give predictions without a statement about the models confidence that the prediction is correct. If the model predicts

[8] https://en.wikipedia.org/wiki/Rashomon_effect

fidelity 精确性
解释和模型的预测有多接近

对解决同一问题的不同model的预测的解释是否一致

对相似数据的解释是否相似

解释的易懂程度

解释是否反映model的正确性

a 4% probability of cancer for one patient, is it as certain as the 4% probability that another patient, with different feature values, received? An explanation that includes the model's certainty is very useful.

- **Degree of Importance**: How well does the explanation reflect the importance of features or parts of the explanation? For example, if a decision rule is generated as an explanation for an individual prediction, is it clear which of the conditions of the rule was the most important?
- **Novelty**: Does the explanation reflect whether a data instance to be explained comes from a "new" region far removed from the distribution of training data? In such cases, the model may be inaccurate and the explanation may be useless. The concept of novelty is related to the concept of certainty. The higher the novelty, the more likely it is that the model will have low certainty due to lack of data.
- **Representativeness**: How many instances does an explanation cover? Explanations can cover the entire model (e.g. interpretation of weights in a linear regression model) or represent only an individual prediction (e.g. Shapley Values).

2.6 Human-friendly Explanations

Let us dig deeper and discover what we humans see as "good" explanations and what the implications are for interpretable machine learning. Humanities research can help us find out. Miller (2017) has conducted a huge survey of publications on explanations, and this chapter builds on his summary.

In this chapter, I want to convince you of the following: As an explanation for an event, humans prefer short explanations (only 1 or 2 causes) that contrast the current situation with a situation in which the event would not have occurred. Especially abnormal causes provide good explanations. Explanations are social interactions between the explainer and the explainee (recipient of the explanation) and therefore the social context has a great influence on the actual content of the explanation.

When you need explanations with ALL factors for a particular prediction or behavior, you do not want a human-friendly explanation, but a complete causal attribution. You probably want a causal attribution if you are legally required to specify all influencing features or if you debug the machine learning model. In this case, ignore the following points. In all other cases, where lay people or people with little time are the recipients of the explanation, the following sections should be interesting to you.

2.6.1 What Is an Explanation?

An explanation is the **answer to a why-question** (Miller 2017).