Weight of Evidence as a Basis for Human-Oriented Explanations

David Alvarez-Melis
Microsoft Research
alvarez.melis@microsoft.com

Hal Daumé III
Microsoft Research & University of Maryland
me@hal3.name

Jennifer Wortman Vaughan
Microsoft Research
jenn@microsoft.com

Hanna Wallach
Microsoft Research
wallach@microsoft.com

Abstract

Interpretability is an elusive but highly sought-after characteristic of modern machine learning methods. Recent work has focused on interpretability via *explanations*, which justify individual model predictions. In this work, we take a step towards reconciling machine explanations with those that humans produce and prefer by taking inspiration from the study of explanation in philosophy, cognitive science, and the social sciences. We identify key aspects in which these human explanations differ from current machine explanations, distill them into a list of desiderata, and formalize them into a framework via the notion of *weight of evidence* from information theory. Finally, we instantiate this framework in two simple applications and show it produces intuitive and comprehensible explanations.

1 Introduction

With the growing success of complex predictors, and their resulting expanding reach into high-stakes and decision-critical applications, wringing explanations out of these models has become a central problem in artificial intelligence (AI). Countless methods have been recently proposed to produce such explanations [33, 32, 31, 22, 1], yet there is no consensus on what precisely makes an explanation of an algorithmic prediction good or useful. Meanwhile, what it means to *explain* and how humans do it are questions that have been long studied in philosophy and cognitive science. Since the end goal of explainable AI is to explain *to humans*, this literature seems an appropriate starting point when looking for principles upon which a theory of *machine* interpretability might rest.

While the debate on the nature of (human) explanation is far from settled, various fundamental principles arise across theoretical frameworks. For example, at the core of Van Fraassen's [9] and Lipton's [20] theories of explanation is the hypothesis that we tend to explain in *contrastive* terms (e.g., "fever is more consistent with pneumonia than with a common cold"), focusing on both factual and counterfactual explanations (e.g., "had this patient had chest pressure too, the diagnosis would have instead been bronchitis"). On the other hand, both of Hempel's models of explanation [14] are characterized by sequences of simple premises, reflecting the fact that humans usually explain using multiple simple accumulative statements, each one addressing a few aspects of the evidence (e.g., "presence of fever rules out cold in favor of bronchitis or pneumonia, and among these two, the presence of chills suggests the latter"). These and other fundamental principles have been observed across disciplines in the social sciences. In a recent survey of over 250 papers in philosophy, psychology and cognitive science on explanation, Miller [24] mentions *contrastiveness* and *selectivity*

Workshop on Human-Centric Machine Learning at the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada. (non-archival)

(i.e., that only few possible cases are presented) as two major properties of the way humans explain things that he argues are important for explainable AI but yet are currently under-appreciated.

These principles are often missing in popular explainable AI frameworks. Their explanations consist of saliency or attribution scores that are *absolute* (i.e., non-contrastive, focused only on the predicted outcome), purely *factual* (i.e., based only on aspects present in the input, ignoring counter-factuals), and *monolithic* (i.e., simultaneously explicating all input features). On the other hand, those that provide probabilistic explanations often present only posterior probabilities, which conflate class priors (also known as base rates) with per-class likelihoods, and which humans are notoriously bad at reasoning about [35, 4, 8, 18]. Furthermore, as Miller [24] and others argue, attribution is only an important but incomplete *part* of the entire process of human explanation. This novel view of explanation as a *process* rather than (only) a *product*, is crucial for understanding the discrepancy between current approaches to automated interpretability and the way humans explain.

In this work, we lay out a general framework for interpretability that aims to reconcile this discrepancy. The starting point of this approach, and our first contribution, is a set of intuitive desiderata that we argue are crucial for bringing machine explanations closer to their human counterparts. With these considerations at hand, we develop a mathematical framework to realize them. At the core of this framework is the concept of *weight of evidence* from information theory, which we show provides a suitable theoretical foundation to the often elusive notion of model interpretability.[1] After introducing this concept, we extend it beyond its original formulation to account for the type of settings in machine learning where interpretability is most needed (e.g., high-dimensional, multi-class prediction). We provide a generic meta-algorithm to produce explanations based on the weight of evidence, and show its instantiation on simple proof-of-point experimental settings.

Related Work Some of the shortcomings of machine explanations highlighted here have been individually tackled in prior work. For example, recent work seeks to move from absolute to contrastive or counterfactual explanations [37, 23, 36], partly inspired by earlier approaches on contrast set mining [3, 5, 38, 25]. On the other hand, while most saliency-based methods produce dense high-dimensional attributions, explanations supported on a sparse set of input features is a much-touted benefit of classic (model-based) interpretability, such as decision trees [29] and sets [19]. Recent work has also explored improving interpretability by explaining on higher-level concepts (e.g., super-pixels or patterns in an image) rather than raw inputs [17, 2]. Our approach shares motivation but many of these works, but differs substantially in how the salient features are selected and scored.

2 Desiderata for Human-Oriented Explanations

The first step towards defining any method for explainable machine learning should be to define its goal with precision, i.e., what is an explanation? For this, we draw on basic principles and terminology from epistemology and philosophy of science. In its most abstract form, an explanation is an answer to a *why-question* [15, 9] consisting of two main components: the *explanandum*, the description of a phenomenon to be explained; and the *explanans*, that which gives the explanation of the phenomenon [15]. Different ways to formalize the explanans have given rise to various theories of explanation; an excellent historical overview of these can be found surveys by Pitt [28] and Miller [24]. For the purposes of this work, our definition of *interpretability* follows that of Biran and Cotton [6] and Miller [24]: the degree to which an observer can understand the cause of a decision.

In the context of machine learning, we are usually interested in explanations of predictive models. For a predictor $f : \mathcal{X} \to \mathcal{Y}$ that takes inputs $x \in \mathcal{X}$ drawn according to some distribution $\mathbf{x} \sim D$ and produces outputs $y \in \mathcal{Y}$, we seek an explanation for $y = f(x)$, that is, "*why did model f predict y on input x*?" Here, we are primarily interested in probabilistic—or more generally, soft—predictors, which covers a wide range of machine learning methods. Namely, we consider models that rather than producing a single prediction y, instead return a predictive posterior distribution $P(Y \mid X = x)$. Furthermore, to allow for more general explananda (e.g., *why was this subset of outcomes ruled out?*), we take inspiration from hypothesis testing and consider complex hypotheses of the form $h : y \in U \subseteq \mathcal{Y}$, and—slightly abusing notation—denote the posterior as $P(h \mid X = x)$.

[1] While the use of weight of evidence for algorithmic explainability has previously been advocated by David Spiegelhalter (e.g., in his keynote talk at NeurIPS 2018 [34]), to the best of our knowledge it has not yet been instantiated or investigated in the context of complex machine learning models.

Having described what type of explanandum we consider in this work, we must now characterize the explanans we seek. The first such consideration pertains to the *causes* or *evidence* which are to define the "vocabulary" from which the explanans is constructed. Popular explanation-based interpretability methods rely directly on the raw inputs x. Likewise, we initially consider evidence e of the form $e = \{x_i\}_{i \in V}, V \subseteq \{1, \ldots, n\}$, but will later generalize to more general *attributes* (e.g., subsets of features). Extending this definition to include higher-level representations of the input [17, 2] or even aspects of the model itself (e.g., parameters) are natural extensions that we leave for future work.

Having formalized its ingredients, we now discuss what properties the explanans should have. Recall that our objective is to devise machine interpretability methods that are intelligible to humans. Our survey of literature on explanations above highlighted various aspects that characterize human explanations, but which most current machine explanations lack. Based on these, we propose a set of *desiderata* for bridging the gap between the former and the latter. Namely, explanations should:

1. **be contrastive**, i.e., answer the question "why did model f predict y *instead of* y'?".
2. **be modular and compositional**, which is particularly important whenever the relations between inputs and outputs/predictions are complex – precisely when interpretability is most needed.
3. **not confound base rates with input likelihood**, i.e., while important for fully understanding a classifier, base rates should be presented separately from input relevance towards the predictions.
4. **be exhaustive**, i.e., they should explicate why every other alternative y' was not predicted.
5. **be minimal**, i.e., all things being equal the simpler of two explanations should be preferred.

Next, we propose an interpretability framework based on the weight of evidence —a basic but fundamental concept from information theory— that satisfies all of the desiderata above.

3 Explaining with the Weight of Evidence

3.1 Weight of Evidence: from Information Theory to Bayesian Statistics

The weight of evidence (WoE) is an information-theoretic approach to analyze variable effects in prediction models [11, 10, 12]. Although originally defined in terms of log-odds (see supplement), the weight of evidence for a hypothesis h in the presence of evidence e can be conveniently defined as $\text{woe}(h : e) \triangleq \log \frac{P(e|h)}{P(e|\bar{h})}$. The interpretation of this quantity is simple. If $\text{woe}(h : e) > 0$ then h is more likely under e than marginally, i.e., the evidence *speaks in favor of hypothesis* h. Analogously, $\text{woe}(h : e) < 0$ indicates h is less likely when taking into account the evidence than without it.

The WoE can be conditioned on additional information: $\text{woe}(h : e \mid c) \triangleq \log \frac{P(e|h,c)}{P(e|\bar{h},c)}$, and can be computed relative to an arbitrary alternative hypothesis (i.e., not necessarily the complement): $\text{woe}(h/h' : e) \triangleq \text{woe}(h : e \mid h \vee h')$. Thus, we can in general talk about the evidence in favor of h and against h' provided by e (and perhaps conditioned on c). Further properties and an axiomatic derivation of WoE are provided in the supplement. An appealing aspect of the WoE is its immediate connection to Bayes' rule. For this, consider the binary classification setting, i.e., $h : Y = 1$, $\bar{h} : Y = 0$ and $e = X$. Simple algebraic manipulation of the definition of WoE yields:

$$\log \frac{P(Y = 1 \mid X)}{P(Y = 0 \mid X)} = \underbrace{\log \frac{P(Y = 1)}{P(Y = 0)}}_{\text{Sample log-odds}} + \underbrace{\log \frac{P(X \mid Y = 1)}{P(X \mid Y = 0)}}_{\textit{Weight of evidence}}. \tag{1}$$

This provides another useful interpretation of the WoE in classification: a positive (negative, resp.) WoE implies that the posterior log-odds (of $Y = 1$ over $Y = 0$) are higher (lower) than the base log-odds, showing that X—the evidence—speaks in favor of (against) the hypothesis $h : Y = 1$.

Besides being intuitive and well-understood, the WoE provides an appealing framework for machine interpretability because it immediately satisfies three of the interpretability desiderata introduced in the previous section: it is naturally contrastive ($\text{woe}(h/h' : e)$ quantifies the evidence in favor of h *against* h'), it decouples base log-odds from variable importance (Eq. (1)) and it admits a modular decomposition (Eq. (5) in the supplement). We later show how the last two desiderata can be met.

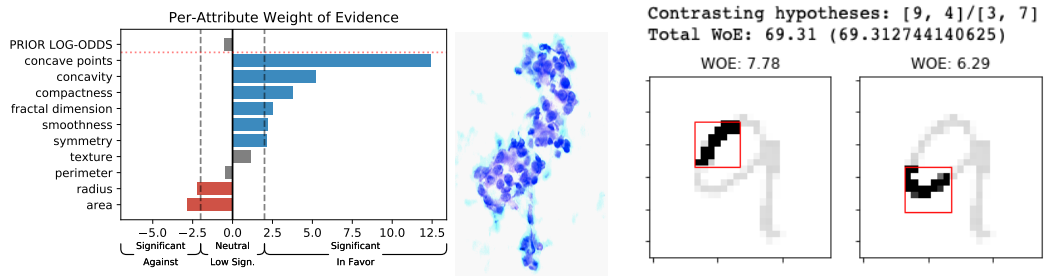


Figure 1: **Left**: A WoE explanation (in favor of `malign`) in the Breast Cancer dataset for a test example (original tissue image shown). **Right**: MNIST classification. The attributes shown (others given in supplement) have high WoE for explaining the prediction of classes [9,4] against [3,7].

3.2 Sequential Explanations: Explaining High-Dimensional Multi-Class Classifiers

The weight of evidence has been mostly used in simple settings, such as a single binary outcome variable Y and a single input variable X. Its use in the (typically more complex) settings considered in machine learning poses various challenges. First, in multi-class classification one must choose the contrast hypotheses h and h'. The trivial choice of letting h be the predicted class c^* and h' its complement is unlikely to yield interpretable explanations when the number of classes is very large (e.g., explaining the evidence in favor of one disease against 999 other possibilities). To address this, we take inspiration from Hempel's model [14] and propose to cast explanation as a sequential process, whereby a subset of the possible outcomes is *expounded away* in each step. For example, in medical diagnosis this could correspond to first explaining why bacterial diseases were ruled out in favor of viral ones, then contrasting between viral families, and finally between the predicted disease and similar alternatives. In general, we consider explanantia consisting of $q+1$ nested hypotheses $h_0 : \{Y = c^*\} \subseteq h_1 \subseteq \cdots \subseteq h_q = V$, which imply q contrastive tests h_{i-1}/h_i.

A second challenge in using WoE for complex prediction tasks arises from the size of the input. While the decomposition formula (Eq. (5)) allows us to produce individual scores for each feature, for high-dimensional inputs (such as in images or detailed health records), providing a WoE score for every single feature simultaneously will rarely be informative. Thus, we propose grouping the inputs into *attributes* (e.g., super-pixels for images or groups of related symptoms for medical diagnosis). Formally, we partition the set of n input features into m subsets: $S_1 \cup \cdots \cup S_m = \{1, \ldots, n\}$.

Given these two extensions of the WoE, we propose a simple meta-algorithm for generating explanations for classifiers. At every step, a subset C_t of the classes is selected to keep (the rest are *ruled out*), and $\text{woe}(C_t/\overline{C_t} : X)$ is computed using the decomposition formula (5). The user is presented with only the most relevant attributes (cf. desideratum 5) according to their WoE (e.g., using the rule-of-thumb threshold of ± 2 [12]), in addition to the base log-odds $\log P(C_t)/P(\overline{C_t})$. This process continues until all classes except the predicted one c^* have been "ruled out" (desideratum 4). It is important to note that unless the predictor is generative—and not black-box—this process requires *estimating* the conditionals $P(X_{S_i}|Y)$. A discussion on estimation, in addition to pseudo-code for this method (Algo. 1) and details about its implementation are provided in the supplement.

4 Experiments

We first illustrate our framework in a simplified setting with *exact* WoE computation (i.e., without estimation) using a Gaussian Naive Bayes classifier, which intrinsically computes $P(X|Y)$ as part of its prediction rule. We use the Wisconsin Breast Cancer dataset, grouping the 30 scalar-valued features into 10 attributes according to their type (mean/s.e./worst area→ `area`, etc.). In the example in Fig. 1 (left), the model predicts `malignant` despite initial log-odds, `radius`, and `area` speaking slightly against it, because the cell's `concavity` and `compactness` attributes speak very strongly in favor of malignancy.

In our second experiment, we use our framework to explain the predictions of a black-box neural-net MNIST classifier, estimating conditional probabilities via a masked autoregressive flow (MAF) model [26], using squared 7×7 super-pixels as attributes. For the example explanation in Fig. 1 (right),

the strong evidence in favor of classes 9,4 (against 3 or 7) clearly corresponds to parts of the image which would be uncharacteristic for examples of the latter classes.

5 Discussion and Extensions

We have proposed a set of desiderata for bridging the gap between the type of explanations provided by humans and current interpretability methods, and a promising framework to realize them based on the weight of evidence. The application of this concept to complex machine learning problems brings about various challenges, some of which we addressed here (high-dimensional inputs and multi-class classification), but many which remain, such as estimation of WoE scores for black-box models, selection of contrast hypotheses, and attribute design. Furthermore, since the ultimate beneficiary of these explanations is a human, the effect of the proposed solutions to these challenges —and all algorithmic choices— should be validated and compared through human evaluation.

References

[1] David Alvarez-Melis and Tommi S. Jaakkola. "A causal framework for explaining the predictions of black-box sequence-to-sequence models". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2017, pp. 412–421.

[2] David Alvarez-Melis and Tommi S Jaakkola. "Towards Robust Interpretability with Self-explaining Neural Networks". In: *Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, Kristen Grauman, Nicolo Cesa-Bianchi, and R. Garnett. Curran Associates, Inc, 2018, pp. 7775–7784.

[3] Paulo J. Azevedo. "Rules for contrast sets". In: *Intelligent Data Analysis* (2010).

[4] Maya Bar-Hillel. "The base-rate fallacy in probability judgments". In: *Acta Psychologica* 44.3 (1980), pp. 211–233.

[5] SD Bay and MJ Pazzani. "Detecting change in categorical data: Mining contrast sets". In: ... *on Knowledge discovery and data mining*. 1999.

[6] Or Biran and Courtenay Cotton. "Explanation and Justification in Machine Learning: A Survey". In: *IJCAI Workshop on Explainable AI (XAI)* August (2017), pp. 8–14.

[7] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. "Density estimation using real nvp". In: *International Conference on Learning Representations*. 2016.

[8] David M Eddy. "Probabilistic Reasoning in Clinical Medicine: Problems and Opportunities". In: *Judgment Under Uncertainty: Heuristics and Biases*. Ed. by Daniel Kahneman, Paul Slovic, and Amos Tversky. Cambridge University Press, 1982, pp. 249–267.

[9] B van Fraassen. "The Pragmatic Theory of Explanation". In: *Theories of Explanation*. Ed. by Joseph C Pitt. Oxford University Press, 1988.

[10] I J Good. "Corroboration, Explanation, Evolving Probability, Simplicity and a Sharpened Razor". In: *The British Journal for the Philosophy of Science* 19.2 (1968), pp. 123–143.

[11] Irving John Good. *Probability and the Weighing of Evidence*. Charles Griffin \& Company Limited: London, 1950.

[12] Irving John Good. "Weight of Evidence: A Brief Survey". In: *Bayesian Statistics* 2 (1985), pp. 249–270.

[13] Noah D Goodman, Chris L Baker, Elizabeth Baraff Bonawitz, Vikash K Mansinghka, Alison Gopnik, Henry Wellman, Laura Schulz, and Joshua B Tenenbaum. "Intuitive theories of mind: A rational approach to false belief". In:

[14] Carl G Hempel. "Deductive-nomological vs. statistical explanation". In: (1962).

[15] Carl G. Hempel and Paul Oppenheim. “Studies in the Logic of Explanation”. In: *Philosophy of Science* 15.2 (1948), pp. 135–175.

[16] Denis Hilton. “Social Attribution and Explanation”. In: *The Oxford Handbook of Causal Reasoning*. Ed. by Michael R Waldmann. Oxford University Press, 2017.

[17] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. “Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)”. In: *International Conference on Machine Learning*. 2018, pp. 2673–2682.

[18] Jonathan J Koehler. “The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges”. In: *Behavioral and brain sciences* 19.1 (1996), pp. 1–17.

[19] Himabindu Lakkaraju, Stephen H Bach, and L Jure. “Interpretable Decision Sets: A Joint Framework for Description and Prediction”. In: *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD’16), Proceedings*. 2016.

[20] Peter Lipton. “Contrastive explanation”. In: *Royal Institute of Philosophy Supplements* 27 (1990), pp. 247–266.

[21] Tania Lombrozo. “Explanation and Abductive Inference”. In: *The Oxford Handbook of Thinking and Reasoning*. 2012.

[22] Scott Lundberg and Su-In Lee. “A unified approach to interpreting model predictions”. In: *Advances in Neural Information Processing Systems 30*. 2017, pp. 4768–4777.

[23] Tim Miller. “Contrastive Explanation: A Structural-Model Approach”. In: *arXiv preprint arXiv:1811.03163* (2019).

[24] Tim Miller. “Explanation in artificial intelligence: Insights from the social sciences”. In: *Artificial Intelligence* 267 (2019), pp. 1–38.

[25] Petra Kralj Novak, Nada Lavra$\$vc, and Goeffrey I Webb. “Supervised Descriptive Rule Discovery: A Unifying Survey of Contrast Set, Emerging Pattern and Subgroup Mining”. In: *Journal of Machine Learning Research* 10 (2009), pp. 377–403.

[26] George Papamakarios, Theo Pavlakou, and Iain Murray. “Masked autoregressive flow for density estimation”. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NeurIPS’17. Long Beach, California, USA: Curran Associates Inc., Dec. 2017, pp. 2335–2344.

[27] Charles Sanders Peirce. “Illustrations of the Logic of Science: IV The Probability of Induction”. In: *Popular Science Monthly* 12 (Apr. 1878), pp. 705–718.

[28] Joseph C Pitt. “Theories of explanation”. In: (1988).

[29] J R Quinlan. “Induction of Decision Trees”. In: *Machine Learning* (1986).

[30] Danilo Jimenez Rezende and Shakir Mohamed. “Variational Inference with Normalizing Flows”. In: *Proceedings of the 32nd International Conference on Machine Learning*. Vol. 37. 2015, pp. 1530–1538.

[31] Marco Tulio Ribeiro, 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 - KDD ’15*. New York, NY, USA: ACM, 2016, pp. 1135–1144.

[32] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. “Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization”. In: *ICCV*. 2017.

[33] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. “Deep inside convolutional networks: Visualising image classification models and saliency maps”. In: *arXiv preprint arXiv:1312.6034* (2013).

[34] David Spiegelhalter. “Making Algorithms Trustworthy: What Can Statistical Science Contribute to Transparency, Explanation and Validation?” Thirty-second Conference on Neural Information Processing Systems (NeurIPS). 2018.

[35] Amos Tversky and Daniel Kahneman. “Judgment under uncertainty: Heuristics and biases”. In: *science* 185.4157 (1974), pp. 1124–1131.

[36] Jasper van der Waa, Marcel Robeer, Jurriaan van Diggelen, Matthieu Brinkhuis, and Mark Neerincx. “Contrastive explanations with local foil trees”. In: *ICML Workshop on Human Interpretability in Machine Learning (WHI 2018)*. 2018.

[37] Sandra Wachter, Brent Mittelstadt, and Chris Russell. “Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR”. In: *Harvard Journal of Law % Technology* 31.2 (2017).

[38] Geoffrey I Webb, Shane Butler, and Douglas Newlands. “On detecting differences between groups”. In: *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2003, pp. 256–265.

A Desiderata for Interpretability in Further Detail

Our driving motivation in this work is to devise machine interpretability methods that emulate the way humans explain. In the introduction, we listed some aspects that characterize human explanations (and which most current *machine* explanations lack). Based on these, we proposed in Section 2 a set of *desiderata* for interpretabilty which are aimed at bridging the gap between the former and the latter. We discuss them here in much more detail.

D1. Explanations should be contrastive. As mentioned before, several authors have proposed (and validated) that humans tend to explain in contrastive terms. To more faithfully emulate human cognition, machine explanations should be contrastive too. That is, the prototypical explanandum should not be "why did model f predict y?", but rather, "why did model f predict y instead of y'?". Despite how self-evident this might be, note that most current explanation methods are not contrastive. Instead, they explain the model's prediction *absolutely*, leaving the contrast case undetermined or implicitly assuming it to be the complement of the predicted class.[2]

D2. Explanations should be modular and compositional. Interpretabilty is most needed in applications where the inputs, outputs or the causal relations between them are complex (and therefore so is any interesting statistical model whose goal is to predict these). Yet, in these settings most explainable AI methods produce a single, high-dimensional static explanation for any given prediction (e.g., a heatmap for an image classifier). These are often hard to analyze and draw conclusions from, particularly for non-expert users. In addition, this form factor again differs from the manner in which humans tend to explain [14]: using various simple premises. Thus, instead of a single monolithic explanans, we seek a set of *simple* sub-clauses, each explicating a different aspect of the input-output predictive phenomenon. Clearly, this modularity introduces a trade-off between the number of clauses in the explanans (too many clauses might be difficult to coherently analyze simultaneously) and their relative complexity (small clauses are easier to reason about, but more of these might be required to explain a complex predictor). At a higher level, breaking up an explanation into a sequence of small components responds to our goals of moving from *explanation as a product* towards *explanation as a process* [21] and to emulate the selective aspect of human explanation [16, 24].

D3. Explanations should not confound base rates with input likelihoods. When explaining probabilistic models, any human-oriented framework for interpretability should take into account how humans understand and interpret probabilities. The psychological and cognitive science communities have long studied this topic [35], showing, for example, that humans are notoriously bad at incorporating class priors when thinking about probabilities. The classic example of Breast Cancer diagnosis due to Eddy [8], showed that the majority of subjects (doctors) tended to provide estimates of posterior probabilities roughly one order of magnitude higher that the true values. This phenomenon has been attributed to a neglect of base-rates during reasoning (the *base-rate fallacy* [4]), or instead, to a confusion of inverse conditional probabilities $P(A|B)$ and $P(B|A)$, one of which needs to be estimated and the other one is provided (the *inverse fallacy*, [18]). Whatever the cause, we argue here that its effect—i.e., that humans often struggle to reason about posterior probabilities—should be taken into account. Thus, explanations should clearly separate the contribution of base-rates and per-class likelihoods linking inputs and predictions. We argue that while both of these are important for understanding a prediction, their very different natures (one of them dependent on the input of the other one not) necessitates different treatment. To the best of our knowledge, no currently available off-the-shelf interpretability framework provides this.

D4. Explanations should be exhaustive. A conclusive justification for a predicted hypothesis h should explicate why no other alternative hypothesis was predicted. In Hempel's terminology, we seek explanations that are *complete* (not to be confused with *complete* in the sense of Goodman et al. [13], i.e., where all *variables* are explained). For example, an explanation for a pneumonia diagnosis simply stating the presence of cough is non-exhaustive, since cough by itself could be indicative of various other conditions beyond pneumonia, so their being non-predicted should be justified.

D5. Explanations should be minimal. In following the original purpose of Occam's Razor, if two explanations are of different complexity but otherwise identical (in particular, both sufficiently explicate the prediction), the simpler of these should be preferred. Furthermore, if omitting less-relevant aspects of an explanation makes the whole more intelligible, while remaining equally faithful to the prediction being explained, then then trimmed explanation should be preferred.

[2] For example, an explanation for a prediction of "9" by a digit classifier is to be interpreted as "why 9 and not any other digit"?

B The Weight of Evidence: Properties and Axiomatic Derivation

The weight of evidence is a fundamental concept that has been introduced in many contexts[3], although it is primarily associated with I.J. Good who popularized it through a long sequence of works [11, 10, 12]. Good originally defined it as follows. For a hypothesis h in the presence of evidence e, the weight of evidence of h is defined as

$$\text{woe}(h : e) \triangleq \log \frac{O(h \mid e)}{O(h)} \tag{2}$$

where $O(\cdot)$ are the log-odds, i.e.,

$$O(h) := \frac{p(h)}{p(\bar{h})}, \qquad O(h \mid e) := \frac{p(h \mid e)}{p(\bar{h} \mid e)} \tag{3}$$

The interpretation of (2) is simple. If $\text{woe}(h : e) > 0$ then h is more likely under e than marginally, i.e., the evidence *speaks in favor of hypothesis* h. Analogously, $\text{woe}(h : e) < 0$ indicates h is less likely when taking into account the evidence than without it.

The WoE has various desirable theoretical properties. For example, Good [12] provides an axiomatic derivation for Definition 2, showing that it is (up to a constant) the only function F of e and h that satisfies the following properties:

1. F is a function of the likelihoods, i.e., $F[p(e \mid h), p(e \mid \bar{h})]$
2. The posterior is a function of the prior and $F(e, h)$, i.e., $p(h \mid e) = g[p(h), F(h, e)]$
3. F is additive (on the evidence) (indeed, $\text{woe}(h : e_1 \wedge e_2) = \text{woe}(h : e_1) + \text{woe}(h : e_2 \mid e_1)$)

The following two properties, which are easy to prove, are crucial for our extension into complex models in Section 3.2:

$$\text{woe}(h/h' : e) = \log \frac{P(e \mid h)}{P(e \mid h')} \tag{4}$$

$$\text{woe}(h/h' : e_1 \wedge e_2 \wedge \cdots \wedge e_n) = \sum_{i=1}^{n} \log \frac{P(e_i \mid e_{i-1}, \ldots, e_1, h)}{P(e_i \mid e_{i-1}, \ldots, e_1, h')} \tag{5}$$

The first of these provides a simple expression to compute WoE scores. The second one will prove consequential to defining an intelligible extension of WoE to high dimensional inputs.

C Explanations of Complex Models via the Weight of Evidence

As mentioned in the main text, using the WoE framework for complex machine learning models brings about the challenge of keeping WoE scores interpretable despite (i) high-dimensional inputs and (ii) not necessarily binary output (e.g., multi-class classification) settings.

We address (ii) by sequentially contrasting (increasigly smaller) sets of classes C_i (i.e., complex hypotheses), as described in the main text. Our solution for (i) in turn involves grouping the inputs into *attributes*, e.g., super-pixels in an image or groups of related symptoms in our running medical diagnosis example. Consider an input space of dimension n, i.e., the evidence $\mathbf{e}$ corresponds now to a multivariate random variable X taking values in $\mathbb{R}^n$. Property (5) (or alternatively, the chain rule of probability) allows for chaining of the conditional probabilities; hence, for any a partition of the n input features into m attributes, we can express the WoE of hypothesis $h : \{y \in C\}$ against $h' : \{y \in C'\}$ as:

$$\text{woe}(h/h' : \mathbf{e}) = \sum_{i=1}^{m} \underbrace{\log \frac{P(X_{S_i} \mid X_{S_{i-1}}, \ldots, X_{S_1}, Y \in C)}{P(X_{S_i} \mid X_{S_{i-1}}, \ldots, X_{S_1}, Y \in C')}}_{\triangleq \text{woe}(C/C' : X_{S_i} \mid X_{S_{i-1}}, \ldots, X_{S_1})} \tag{6}$$

[3] The basic principle behind the weight of evidence appear in the work of both Alan Turing and Claude Shannon. However, Good [12] claims ideas like it go back to at least Peirce [27]).

Algorithm 1: Weight of evidence explanation generation for complex models

Input: Example $X \in \mathbb{R}^n$, class $c^* \in \{1, \ldots, K\}$ predicted by the model.
Parameters *Attribute size* α*, hypothesis size regularization* λ*.*

$V \leftarrow \{1, \ldots, K\}$; `// Remaining classes to be explained`
while $|V| > 1$ **do**
  `/* Find hypothesis maximizing regularized 'total' woe */`
  $U \leftarrow \operatorname{argmax}_{U \subseteq V; c^* \in U} \text{woe}(U/(V \setminus U) : X) - R(U)$;
  $\tilde{U} \leftarrow V \setminus U$; `// contrast hypothesis is the relative complement`
  `/* Compute base log-odds of chosen hypothesis */`
  $\text{lod}(U) \leftarrow \log \frac{P(U|X)}{P(\tilde{U}|X)}$;
  $T \leftarrow \{1 \ldots, n\}$;
  $i \leftarrow 0$;
  while $|T| > \alpha$ **do**
    `/* Find attribute with maximal partial woe */`
    $S_i \leftarrow \operatorname{argmax}_{S \subset T; |S| = \alpha} \text{woe}(U/\tilde{U}) : X_S)$;
    $\omega_i \leftarrow \text{woe}(U/\tilde{U} : X_{S_i})$;
    $T \leftarrow T \setminus S_i$
  end
  DisplayExplanation$(U, \tilde{U}, \{S_i\}, \text{lod}(U), \{\omega_i\})$;
  $V \leftarrow V \setminus U$; `/* Update classes to be explained */`
end

where X_{S_i} denotes the subset of random variables with indices in S_i, i.e., $X_{S_i} = \{X_j\}_{j \in S_i}$. The full Bayes-odds explanation model now has the form

$$\underbrace{\log \frac{P(Y \in C \mid X)}{P(Y \in \bar{C} \mid X)}}_{\text{posterior log-odds of entailed set}} = \underbrace{\log \frac{P(Y \in C)}{P(Y \in \bar{C})}}_{\text{prior log-odds of entailed set}} + \sum_{i=1}^{m} \underbrace{\text{woe}(C/C' : X_{S_i} \mid X_{S_{i-1}}, \ldots, X_{S_1})}_{\text{conditional WoE of i-th attribute}} \quad (7)$$

Note that, in general, the order of the attributes matters in this sum. We discuss how to minimize the impact of this ordering in the next section. These two extensions lead to the meta-algorithm for WoE-based explanation shown here as Algorithm 1, which selects at each step the contrast set C_i with maximal WoE plus a cardinality-based regularizing term $R(U)$ to prevent too small or too large subsets from being selected. We use $R(U) = \alpha(|U| - \frac{1}{2}|V|)^2$ to encourage even partitions.

D Estimation of Weight of Evidence Scores

For any pair of entailed and contrast classes, and any partition of input into attributes, equation (5) provides an exact method to compute the conditional weight of evidence as a sum of per-attribute WoE scores. In order to use this expression, we need to be able to compute $P(X_{S_i} \mid X_{S_{i-1}}, \ldots, X_{S_1}, Y \in C)$ for any order of attributes $\{X_{S_i}\}_{i=1}^n$ and outcome set $C \subset \mathcal{Y}$. In an ideal scenario (such as the Gaussian Naive Bayes classifier used in Section 4, or in an auto-regressive generative model for sequential data), the prediction model itself would compute and store these values.

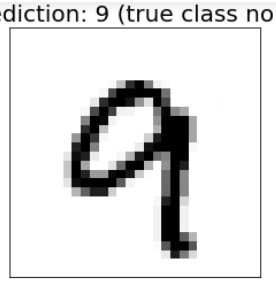
Unfortunately, most prediction models do not compute such probabilities explicitly, so any realistic application of the WoE methodology to interpretability must provide a fallback method to estimate these, independently, from data. Let us consider the worst case scenario: a black-box prediction model, for which we assume we only have oracle access (i.e., queries of $f(x)$), in addition to access to additional training data. In such case, the problem essentially turns to a conditional density estimation problem, where we seek to learn models of $P(X_{S_i}|Y)$ from training data in the form of pairs $(x, \hat{y})$, where x is a sample from the input distribution $X \sim P_X$ and $\hat{y} = f(x)$ is a class label prediction obtained with the prediction model.

We propose to tackle this problem by training, as an off-line preliminary step, an auto-regressive conditional likelihood estimation model. For simple data, this could be done with classic (e.g., kernel or spectral) density estimation methods. For more complex data such as images, many recent methods

have been proposed based on normalizing flows and autoregressive models [30, 7, 26]. For sequential data, the ordering of the attributes in Eq. (6) is implied by the data. For non-sequential inputs, the likelihood model should be trained to minimize the impact of ordering on the WoE scores (e.g., by training on random orderings). For the experiments on MNIST, we train a conditional Masked Autoregressive Flow (MAF) model [26], randomizing the order in which the 7×7 pixel blocks (the attributes) are traversed, but keeping the order within each of these fixed (left-to-right, top-to-bottom).

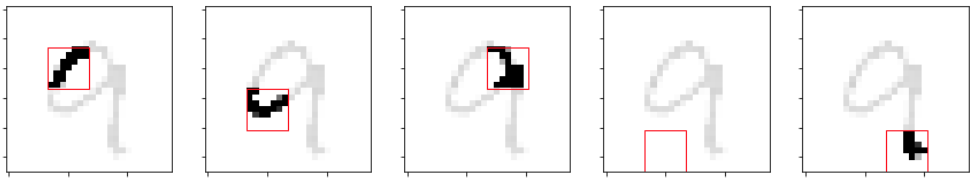
E A full explanation of the Black-Box MNIST classifier

Prediction: 9 (true class no.9)



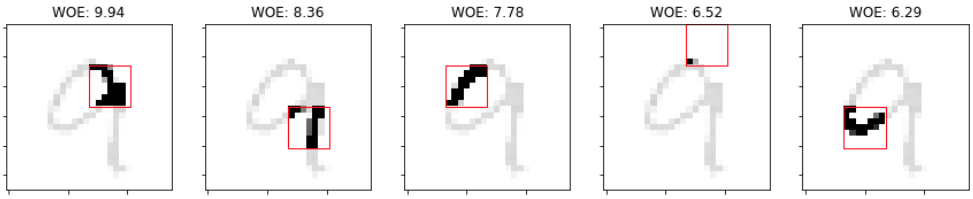
```
NOTE: For each step, we display only attributes with WOE > 5 (Positive Evidence)
or < -5 (Negative Evidence, if any such exist), up to at most five attributes per step.

--------------------
Explanation step:  0
Optimal score: 92.1043701171875
Contrasting hypotheses: [9, 3, 4, 7]/[0, 1, 2, 5, 6, 8]
Total WoE: 92.89 (92.8856201171875)
Positive Evidence:
```

WOE: 10.31 WOE: 7.23 WOE: 6.72 WOE: 6.24 WOE: 6.19



```
--------------------
Explanation step:  1
Optimal score: 56.812744140625
Contrasting hypotheses: [9, 4]/[3, 7]
Total WoE: 69.31 (69.312744140625)
Positive Evidence:
```

WOE: 9.94 WOE: 8.36 WOE: 7.78 WOE: 6.52 WOE: 6.29



```
--------------------
Explanation step:  2
Contrasting hypotheses: [9]/[4]
Total WoE: 47.62 (47.6168212890625)
Positive Evidence:
```

WOE: 10.99 WOE: 9.06 WOE: 7.35