A Detailed Survey of Machine Learning Interpretability through Feature Interactions and Additive Models

James Enouen ENOUEN@USC.EDU
Department of Computer Science
University of Southern California
Los Angeles, CA 90089-1044, USA

Yan Liu YANLIU.CS@USC.EDU
Department of Computer Science
University of Southern California
Los Angeles, CA 90089-1044, USA

Abstract

Interpretability has gained an immense amount of popularity over the past decade, mostly riding the wake of neural networks' impact on the field of machine learning. The growth in success and application of neural networks, as well as other blackbox machine learning approaches, into everyday life has led to a revitalized interest in understanding the blackbox models making those decisions. In this survey, we give a holistic coverage of the current state of the interpretability literature, focusing on a detailed coverage for machine learning tasks and for methods related to feature interactions like SHAP explanations, functional ANOVA, and the generalized additive model. We achieve this by first mapping out the many different subareas and paradigms under the larger umbrella of the *Interpretability* or *XAI* literature. We next isolate the scope of this survey and provide a detailed account of historical developments leading to the modern convergence of interpretability using feature interactions and interpretability using additive models. Finally, we conclude with a discussion of specific domains of application and with future directions for interpretability.

Keywords: interpretability, explainability, feature interactions, additive model, XAI, Shapley value, SHAP, interactions, higher-order interactions

1 Introduction

Blackbox algorithms, developed using a variety of deep learning and machine learning techniques, have quickly become an integral part of daily life in the 21st century as AI applications continue to propagate throughout scientific, industrial, and commercial applications. As these machine learning approaches continue to deliver on their promise of accurate predictions in exchange for big data, the question of *how* a blackbox algorithm make its decisions or predictions has only grown in importance and frequency. Despite the considerable attention on the problem of how blackbox models transform large datasets into predictive insights, there remain no universal solutions to explain the behavior of these models.

Interpretability is the field born out of studying this question of interpreting, explaining, and understanding the how and the why of predictions made by blackbox models. Although interpretation has always been in the background of statistical modeling and was even acknowledged as important in previous decades as model complexity began to grow beyond linear regression into increasingly opaque nonparametric models, the field would not see more dedicated study until mid 2010s. By this time, the incredible power of deep learning systems had already been demonstrated on audio, vision, and language tasks (Mohamed et al., 2011; Krizhevsky et al., 2012; Bahdanau et al., 2015), cementing these newest blackbox models as more capable at learning from raw features than any previous methods.

Given the diversity of blackbox applications and the difficulty of the blackbox question, there are unsurprisingly a lot of different approaches, cultures, and subfields working on answering the challenging questions of what to explain, how to explain, when to explain, etc. Over the past decade, there have been many attempts to standardize, rigorize, unify, and qualify what the goals, definitions, and techniques of interpretability are. In this work, we must take the necessary precautions to survey a vast swath of this literature without completely disregarding other tangentially related approaches in the field of interpretability. We accomplish this in two main ways. First, we restrict our attention throughout to what we will broadly call *machine learning interpretability*. Second, we focus on those methods which are related to *feature interactions*, centering our discussion on the duality between GAM and SHAP.

By first restricting our attention to "machine learning interpretability" instead of "deep learning interpretability", we mainly divide along the implicit assumption that the input features themselves are interpretable. This is sometimes referred to as the divide between "structured data" (data with interpretable features) and "unstructured data" (data without interpretable feature representations). This mostly excludes mechanistic interpretability, outside of classical mechanistic methods like gradient saliency and attention maps which have generally fallen out of favor. By second focusing our attention on "feature interactions", we will dedicate most of our time to those methods focusing on feature attribution and extensions of feature attribution which explore the need for interaction attribution. We will proceed by first laying out many of the different approaches existing within interpretability before categorizing them amongst three major categories. The topic of feature interactions which we discuss in more detail will be centered around the 'pillar' of additive interpretability revolving around the duality between GAM interpretations and SHAP explanations.

In Section 2, we begin by providing a very high-level overview of the field of interpretability and some of the many various techniques used throughout the literature. By the end of the section we provide a map of the current interpretability landscape and point to exactly where we plan to focus our attention for the rest of the survey. In Section 3, we go into great detail on the explainability approach of feature interactions. In Section 4, we go into great detail on the interpretability approach of additive models. In Section **??**, we detail the many applications of feature interaction approaches across vision, language, time series, graphs, and the natural sciences. Finally, in Section **??**, we conclude with many of the open problems and active research areas both within the subfield of interactions but also in connection with the wider field of interpretability.

2 The Current Map of Interpretability

Given that the goal of interpretability ('to understand the black box') is extremely broad, the methods which researchers have developed to tackle the problem have become equally broad. We use Section 2 to provide a high-level map of the large and diverse space of interpretability and related fields before clarifying which ‘part of the map’ we will explore in greatest detail throughout the course of this survey.

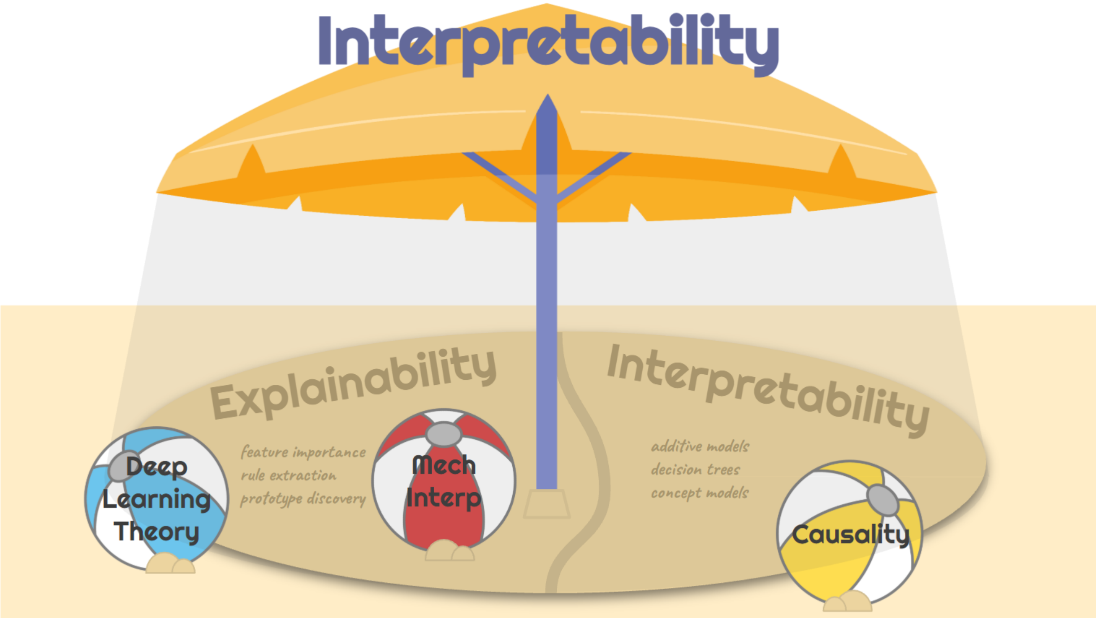


Figure 1: A high-level view of the umbrella of Interpretability research. There is a large line in the sand to distinguish the subgoals of interpretability and explainability. Fields like mechanistic interpretability fall entirely under the umbrella of interpretability, whereas fields like deep learning theory and causality only partially intersect with the goals of interpretability.

2.1 Areas Related to Interpretability

Explainability v. Interpretability We begin by reminding of the key distinction within interpretability research which is rarely known outside of the field. That is the distinction between explanation approaches and interpretation approaches. Of course in natural language, these two words are nearly synonymous.

In Figure 1, we depict the larger field of Interpretability as the umbrella field which covers many approaches related to interpreting and understanding blackbox models. Underneath the umbrella, we see both interpretability and explainability as subcategories. In this case, interpretability is used to mean intrinsically interpretable glassbox models, whereas explainability is used to mean post-hoc explanations of blackbox models. Although seemingly innocuous at first glance, this distinguishment turns out to be a critical motive underlying interpretability research. We will commonly use interpretability both for the umbrella class and the low-level meaning. Although we will try here to avoid saying explainability to mean the umbrella class, it is also common in the literature to use explainability as the umbrella class, especially in well-established phrases like XAI (meaning eXplainable Artificial Intelligence). Lastly, the field of explainability also takes on a third meaning regarding the psychological and social aspects of explaining (Miller, 2018), which can be necessary to apply in both interpretability and explainability.

Deep Learning Theory v. Interpretability As can be seen in Figure 2, the questions of deep learning interpretability and deep learning theory are very intimately related. Nevertheless, there are some very key nuances between the goals and approaches of the two which mostly separate them. In particular, deep learning theory mostly deals with understanding *how* neural networks learn what they learn, usually beginning from first principles and mathematically studying the effects of gradient descent algorithms. In contrast, (deep learning) interpretability instead often deals with understanding *what* a particular neural network has learned. Typically, this is done by fixing a trained model and applying a variety of counterfactual perturbations in order to ascertain some additional information about the blackbox system.

Both confusingly and excitingly, there are many details in the specifics of how these what/ why/ how questions are asked within different subfields which impact the flavor of the research questions. For instance, traditional interpretability methods are often concerned with understanding why a network has made its prediction, usually in the form of understanding which input features 'contributed' to the output prediction. In contrast, mechanistic interpretability methods are more concerned with understanding how a network has made its prediction, often in the form of understanding which substructures inside the neural network 'control' the final prediction. For this reason, one of the main identifiers between traditional interpretability and mechanistic interpretability is the former's goal of model agnostic understanding and the latter's goal of model specificity.

Beyond this distinction, there are many other aspects of the what/ why/ how questions which seriously affect the goal of interpretability and the type of interpretability used. For example, an additional subfield of interpretability is the literature focusing on gaining causal insights through interpretability (related to the recent push for *actionable* interpretability). Here, the goal is to distill information from the model all the way back to the real-world, gaining some causal knowledge about the real world. It is worth noting that this is distinct from the causal language which is used inside of mechanistic interpretability. There, one first assumes that the studied network is the 'true process' and then uses tools from causality to causally explains the behavior. In contrast, causal interpretability uses the explanation of model behavior and questions whether the true process of the world (outside of the model) is explained in the same way. These distinctions between 'true-to-the-data' and 'true-to-the-model' have been the source of endless confusion and of many unnecessary debates on the topic (Sundararajan and Najmi, 2020; Chen et al., 2020; Frye et al., 2021).

Causality v. Interpretability Another key clarification worth getting into is the distinction between causal insights and interpretable insights, especially given the importance of interpretability for causality and the importance of causality for interpretability. This somewhat subtle nuance is only made worse by people's natural inclination to treat explanations as causal as well as initiatives from the mechanistic interpretability community to explain models exclusively using causal language (Geiger et al., 2025).

The unfortunate truth, however, is that causal claims requires causal reasoning and causal assumptions, which are broadly lacking from the interpretability literature. Although

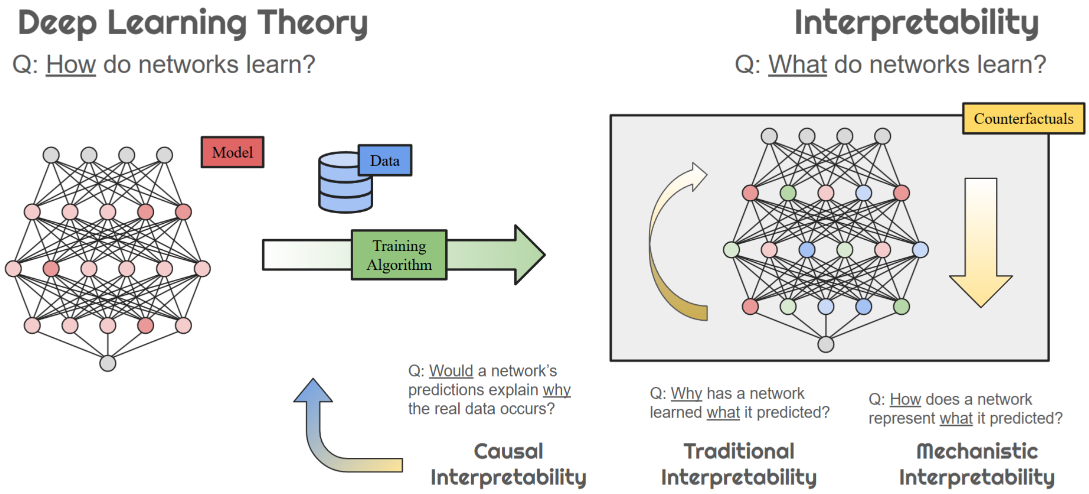

Deep Learning Theory
Q: How do networks learn?
Model
Data
Training Algorithm
Interpretability
Q: What do networks learn?
Counterfactuals
Q: Would a network's predictions explain why the real data occurs?
Causal Interpretability
Q: Why has a network learned what it predicted?
Traditional Interpretability
Q: How does a network represent what it predicted?
Mechanistic Interpretability

Figure 2: Deep Learning Theory vs. Deep Learning Interpretability

there have been attempts to directly incorporate causal reasoning into feature explainability, it is often a careful balance between making too many causal assumptions (Jung et al., 2022; Biparva and Materassi, 2024; Parafita et al., 2025) and making unrealistic assumptions (Janzing et al., 2020; Heskes et al., 2020). The vast majority of interpretability is still on typical 'X to Y' problems of regression or classification; whereas causal insights requires understanding the structure between X variables, and sometimes even X and Y in a joint fashion. Moreover, if we have a perfect causal model of the variables, this precludes needing a blackbox regressor in the first place, begging the question of why would even need interpretability in the first place. A satisfactory middle ground here still seems to be absent from the literature at the present. Accordingly, it is important for the practitioner to always take extra precautions when jumping from interpretable insights to causal claims.

2.2 A Taxonomy of Interpretability Approaches

Over the years, there have been many developments in interpretability methods, leading to a constantly expanding toolkit for understanding model behavior. Despite the constantly evolving landscape of interpretability, there have also been many good summaries of the available interpretability methods for explaining blackbox models, each making their own choices for the taxonomization of the available methods. One of the earliest syntheses of interpretability methods, the 'Interpretability Book' (Molnar, 2019), takes there to be the following classes of methods:

1. Intrinsically Interpretable Methods (GAM-type methods and rule-based methods)
2. Model-agnostic Methods (PDP, ICE, ALE; feature attribution, feature interaction, global surrogate, local surrogate, LIME, SHAP)
3. Example-based Approaches (counterfactual examples, adversarial examples, prototypes/ criticisms, influential instances)
4. Deep Learning Specific Approaches (feature visualization)

The latest version (Molnar, 2025) instead takes there to be the following classes of methods:

1. Intrinsically Interpretable Methods (GAM-type methods and rule-based methods)
2. Local XAI (ICE; LIME, counterfactual examples, anchors, SHAP)
3. Global XAI (PDP, ALE; feature interaction, functional decomposition, leave-one-out, surrogate models, prototypes and criticisms)
4. Deep Learning Methods (learned features; saliency maps; detecting concepts; adversarial examples; influential examples)

Other more recent surveys take a broadly similar approach to categorizing the many different classes of interpretability methods. Zhang et al. (2021) suggests four main explanation types: example-based, attribution-based, hidden semantics, and rule-based. Ji et al. (2025) suggests five main styles: attribution-based, function-based, concept-based, prototype-based, and rule-based self-interpretation.

Herein, we consider there to be three main styles of explanation: concept-based, rule-based, and additive-based. In Figure 3, we depict how many classical interpretability methods fall into these three categories. We find that unlike previous taxonomizations, there is a single interpretable model which is a *quintessential representative* of the interpretation style, depicted at the base of each pillar. The additive pillar is supported by the interpretable additive model. The reasoning pillar is supported by the interpretable decision tree. Finally, the concept pillar is supported by the k-NN algorithm, which is often granted the same status as an interpretable model.

2.3 Mechanistic?

As alluded to before, we will mainly ignore approaches from mechanistic interpretability in favor of more enduring approaches to interpretability; however, we still find it important to review some of the major developments and techniques coming from this subfield.

Mechanistic approaches to neural network interpretability have existed just as long as neural networks themselves. Beginning in computer vision, one of the simplest early methods was *gradient saliency* (Simonyan et al., 2014), originally applied to those deep convolutional networks which were able to achieve stunning accuracy on the ImageNet challenge. This method simply takes the gradient of the class logit with respect to the input pixels. Later follow-ups like Layer-wise Relevance Propagation (LRP) (Bach et al., 2015), Grad-CAM (Selvaraju et al., 2017), DeepLIFT (Shrikumar et al., 2017), and Integrated Gradients (IG) (Sundararajan et al., 2017) provided extensions to the simple gradient approach which gave it increased stability or convenient properties. Unfortunately, later work (Adebayo et al., 2018) called into question how useful these saliency maps, providing sanity checks showing most saliency methods provided little information beyond edge detection and most likely hijacks our human intuition to make us researchers feel as though we were understanding the predictions of the neural networks. Despite these major limitations, it often remains the dirty go-to throughout many areas of computer vision research.

Another key mechanistic explanation approach is the method of *feature visualization* (Dumitru Erhan and Vincent, 2009). Originating again in computer vision research and even earlier than the simple gradient saliency approach, this approach simply finds the solution to a maximization problem for each of the class labels, usually maximizing the activation through a gradient descent algorithm. Later interpretation works (Simonyan et al., 2014) would also include this approach; however, this direction picked up greater

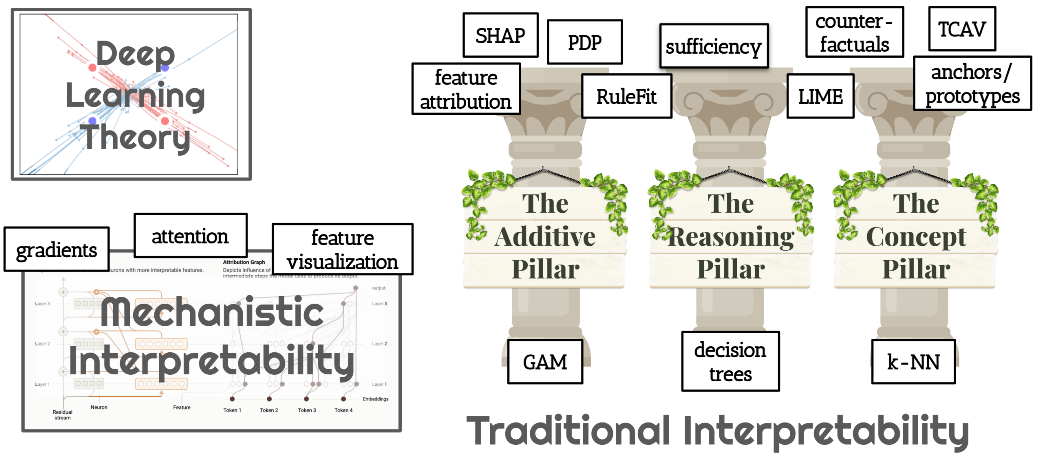


Figure 3: Taxonomization of the Classical Approaches to Interpretability

steam after Google's DeepDream visualization approach (Mordvintsev et al., 2015) on the state-of-the-art Inception network (Szegedy et al., 2015). Later works like (Olah et al., 2017) and (Bau et al., 2017) would continue to build towards a larger-scale analysis of entire convolutional networks (and ultimately form the basis for modern mechanstic approaches).

The final method in the set of long-standing mechanistic interpretability approaches is the *attention mechanism* (Dzmitry Bahdanau, 2015), coming directly from the neural architecture in its application to machine translation in natural language. Its usage as an interpretability approach eventually became widespread enough that a work summarizing its limitations called "Attention is not Explanation" (Jain and Wallace, 2019) was eventually published. This led to a retort called "Attention is not not Explanation" (Wiegreffe and Pinter, 2019) which claimed attention was at least doing something, as well as works like Attention Flow (Abnar and Zuidema, 2020) which instead tried to correct some of the shortcomings of attention. Although attention is also commonly used throughout NLP applications as a first approximation of importance, concerns regarding the faithfulness of such explanations persist into the current day (Lyu et al., 2024).

More modern versions of mechanistic interpretability (also MI or mech interp) continue further down the path trailblazed by Chris Olah and coauthors, mainly through the approaches of feature visualization (Carter et al., 2019b; Schubert et al., 2020) and circuit decomposition (Olah et al., 2020). This push has been integrated with many call-to-actions motivated by concerns surrounding AI safety, and has continued further after the creation of companies like Anthropic specifically with these goals and values in mind. This has led to the explosive growth of mechanistic interpretability as a subfield in recent years, focusing on a wide array of approaches trying to understand the feature representations of deep neural networks and trying to understand the algorithmic computations run by deep neural networks, including: linear probes (Alain and Bengio, 2017), nonlinear probes (Li et al., 2023), sparse autoencoders (Huben et al., 2024), sparse circuits (Marks et al., 2025), causal mediation analysis (Vig et al., 2020), model editing (Meng et al., 2022), and causal abstraction (Geiger et al., 2025). As was already discussed, this is often done in a manner which is both empirical and applied, distinguishing it from the sometimes similar works in deep learning theory.

2.4 Pillars of Interpretability

We now discuss the three pillars of interpretability as depicted in Figure 3 and how they are representative of interpretability research which we will herein call *traditional interpretability* to better distinguish it's more enduring spirit from mechanistic interpretability's more recent push. Each of these pillars are supported at the base by an intrinsically interpretable model, and at the top of each pillar we list many classical explainability approaches which are related to that interpretability style. Supported by the generalized additive model (GAM), we have the additive pillar. Supported by the decision tree model, we have the reasoning pillar. Supported by the k-nearest neighbor model (k-NN), we have the concept pillar.

There is near complete consensus throughout the interpretability community that additive models and decision trees are interpretable models (of course to varying degrees in various situations), so many readers will be comfortable with these choices for the first two pillars. It is much less standard to make the same claim that the k-NN model is interpretable; however, in contexts like data attribution, it is clearly one of the simplest possible models for understanding the influence of individual data points. We will soon further justify this decision.

The Additive Pillar is centered around the generalized additive model (Hastie and Tibshirani, 1990; Lou et al., 2012) which additively incorporates the influences of individual features into a final prediction. These separate contributions are implicitly understood to independently contribute towards the final prediction, allowing them to be interpreted individually. Extensions to models with pairwise interactions (Lou et al., 2013) or higher-order interactions (Enouen and Liu, 2022) provide a spectrum of gradually increasing complexity which can ultimately represent any function, at the expense of deteriorated interpretability.

On the explainability side, feature attribution approaches (Smilkov et al., 2017; Sundararajan et al., 2017) primarily use the same feature-based reasoning as generalized additive models. Although beginning in the gradient saliency literature measuring local sensitivity, progress slowly developed towards smoother, model-agnostic measurements of local influence like SmoothGrad (Smilkov et al., 2017) and Integrated Gradients (IG) (Sundararajan et al., 2017). Further developments into feature perturbation approaches (Breiman, 2001a; Aaron Fisher, 2019; Hooker and Mentch, 2019) and feature removal approaches (Lei et al., 2018; Covert et al., 2021) helped standardize progress. Specific cases like the SHAP approach (Lundberg and Lee, 2017) greatly unified interest in the field, now known to be directly dual to GAM interpretation (Bordt and von Luxburg, 2023; Enouen and Liu, 2025).

One of the earliest interpretability approaches, the partial dependence plot (PDP) (Friedman, 2001), works on the exact same intuition as the shape functions of the additive model, marginalizing the effect of an individual feature on the blackbox model across the entire dataset. Later work called the individual conditional expectation (ICE) (Goldstein et al., 2015) instead considered the individual curves of each data point together in a single plot. It is worth noting how these two approaches, explicitly and implicitly, consider the marginal distribution over the input features which is not necessarily a good representation of the underlying distribution. Instead it is often necessary to consider the conditional distribution which is induced by the underlying distribution when conditioning on a single feature. Although these conditionally marginalized plots are difficult to compute, the accumulated local effects (ALE) plot (Apley and Zhu, 2020) are an alternative which samples from alternative samples which are 'nearby' to the data sample according to a prescribed distribution.

The Logical Pillar is focused on the logical, mechanical steps taken to arrive at a final decision. Even simpler than decision *trees* (Hastie et al., 2001) is the decision *list* (Chen and Rudin, 2018) which allows for a deterministic list of decisions to be sequentially followed until a particular condition is met. The decision tree instead allows for a flow chart to be followed depending on the previous set of logical conditions. Taking this complexity further, the decision *circuit* (Oliver, 1993; Kohavi, 1994; Kohavi and Li, 1995; Darwiche and Marquis, 2002) allows for even more compact representation at the expense of a more tightly entangled reasoning process.

Although practical extensions to decision trees have mainly been the additive ensemble of the random forest (Breiman, 2001a), it is folk knowledge that this immediately destroys the interpretability. Broadly, this seems to be due to the 'incompatibility' of mixing additive interpretability and logical interpretability in this particular way. Alternatives to extend the complexity of the decision process while remaining interpretable have instead considered a large set of near-optimal decision trees called the *Rashomon set* (Xin et al., 2022). These small decision trees are then designed to interpreted individually and then chosen based on alternative objectives like fairness, robustness, or simply preference by domain experts. RuleFit (Friedman and Popescu, 2008) is an additive combination of rules which uses *sparsity* to ensure that few rules will be added together, allowing each individual rule to still be interpreted.

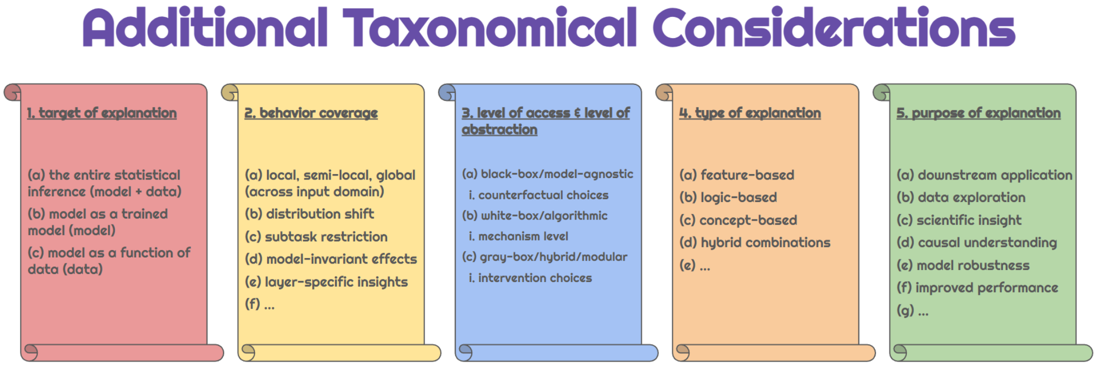
On the explanation side, the sufficient input subset (SIS) method (Carter et al., 2019a) aligns heavily with the logical reasoning of decision processes, providing the input features which were sufficient for the model to be certain about its decision. The closely related notion of necessity (Darwiche and Hirth, 2020; Watson et al., 2022) describes which input features cannot be changed to maintain the same prediction. These logic-based notions of explanation often have formal origins coming from causal reasoning (Pearl, 2009; Halpern, 2016), abductive reasoning (Peirce, 1903; Josephson and Josephson, 1994), and earlier.

The closely related approach of counterfactual explanations (Wachter et al., 2017) proposes alternative feature values which are sufficient to change the prediction or decision. When focusing on the features which were changed to come to the new prediction, these can be related to the notions of sufficiency and necessity (Kommiya Mothilal et al., 2021); however, it is also common to consider the counterfactual as its own data sample, putting the reasoning closer to prototype-based reasoning. The method of LIME (Ribeiro et al., 2016) uses a local surrogate around a data point, typically either a decision tree or linear model, which then applies the additive reasoning or logical reasoning to a local region around the data point. Anchors (Ribeiro et al., 2018) provide local logical explanations in a similar fashion using a different algorithm.

The Conceptual Pillar is about the construction of new features or concepts, allowing for greater abstraction and simpler higher-level reasoning. Interpretable works like Prototypical Part Network (ProtoPNet) (Chen et al., 2019) and the Concept Bottleneck Model (CBM) (Koh et al., 2020) extend the reasoning style of the k-NN which makes determinations based on similarity with previously seen examples.

Previous survey hierarchies (Ji et al., 2025) have argued that *concept-based interpretability* and *prototype-based interpretability* are fundamentally different categories. Although we agree that there are key stylistic differences between concept approaches and prototype approaches, we believe that their overall reasoning styles are fundamentally the same. Although k-NN models and prototype-based regression make predictions directly from distance-based similarity to existing samples, more general CBMs extend the similarity metric across an array of concepts and accordingly complexify the final prediction step.

On the side of explanation, we begin with the most obvious concept-based explainability approach. TCAV (Kim et al., 2018) uses a set of training data samples which are labeled as obeying or disobeying a concept to be able to find a vector direction within the latent space which points in the direction of the concept. As previously mentioned, methods like LIME (Ribeiro et al., 2016) and Anchors (Ribeiro et al., 2018) which take a local perspective in a small region around the data point are partially a prototype-style explanation. Even moreso are direct prototype-based explanations (**?**) which find explanatory samples nearby in input space or latent space, and counterfactual explanations (Wachter et al., 2017) which find explanatory samples which are nearby but with different predictions. Moreover, this goal of finding an appropriate set of concepts is tangentially related to compressibility (Tishby et al., 2000; Hutter, 2005) and sparse coding (Donoho, 2006a; Rubinstein et al., 2010).



Additional Taxonomical Considerations

1. target of explanation
(a) the entire statistical inference (model + data)
(b) model as a trained model (model)
(c) model as a function of data (data)

2. behavior coverage
(a) local, semi-local, global (across input domain)
(b) distribution shift
(c) subtask restriction
(d) model-invariant effects
(e) layer-specific insights
(f) ...

3. level of access & level of abstraction
(a) black-box/model-agnostic
i. counterfactual choices
(b) white-box/algorithmic
i. mechanism level
(c) gray-box/hybrid/modular
i. intervention choices

4. type of explanation
(a) feature-based
(b) logic-based
(c) concept-based
(d) hybrid combinations
(e) ...

5. purpose of explanation
(a) downstream application
(b) data exploration
(c) scientific insight
(d) causal understanding
(e) model robustness
(f) improved performance
(g) ...

Figure 4: Additional Axes of an Interpretability Taxonomy.

2.5 Further Jargon and Taxonomical Aspects

Finally, for completeness sake, we will briefly go over some other taxonomical aspects of the interpretability literature before moving into surveying the main topics of additive explainability and additive interpretability.

It is worth noting that the majority of interpretability work takes the perspective that the goal is to explain a trained blackbox model as depicted in the right-hand side of Figure 2, and this accordingly frames a lot of the discussion about and previous taxonomies of the field; however, there are emerging areas considering data valuation (Ghorbani and Zou, 2019; Jia et al., 2019b) or even the entire statistical inference (Bordt et al., 2025). One of the main taxonomizations of explainability methods is whether or not they are local, semi-local, or global (Zhang et al., 2021). This refers to whether an XAI method is supposed to explain the behavior of the model only in a local area near the data point, in a larger region around the data point, or across the entire input manifold. It is worth reminding that there is a minor distinction between a global explanation which uses an interpretable model as a surrogate of a blackbox model and a global interpretation of an interpretable model trained on the same data. This local v. global question can be seen as a special case (although the most common case) of the larger question: how much of the model behavior is covered by my explanation method? Further examples of this question include: am I explaining this behavior only on the training distribution? am I explaining this behavior only for a certain subtask?

Another key distinction is between black-box explanations and white-box explanations. Note that this is distinct from blackbox v. glassbox which refers to the lack of transparency in these models from an interpretability perspective, whereas black box v. white box refers to the level of access to the actual model mechanisms, regardless of the complexity in interpreting those mechanisms. Historically, there was a push to move away from early gradient-based and attention-based methods (gray box) towards model-agnostic interpretability methods which could be applied to any blackbox model; however, there has also been the more recent push methods specifically designed for understanding neural networks and transformers given their widespread successes. Obviously the level of access restricts the lowest level of abstraction which is possible; however, researchers often choose a higher level of abstraction than available. Immediately, one is then faced with the questions about what counterfactual queries are possible at the level of feature removal, data point removal, model type change, mechanism perturbation, etc. These are critical questions related to the explanation target and explanation purpose.

Finally, there are the aspects of the type of explanation, which we taxonomized in much greater detail in the previous sections, and can also be very much entangled with these other questions about model type and behavior coverage. Also worth mentioning is the model's task (decision-making, prediction-making, or generative) which can be seen as a subquestion of the first question on the target of the explanation. Lastly, we have the all-important question about the downstream purpose of the explanation itself. Although this does not always make its way into the papers creating XAI methods, it is an undeniable part of XAI application, and thus must be considered as part of how to interpret and integrate the explanation. This again gets into the challenging questions of causality, social explanation, and broader context (Miller, 2018).

3 Additive Explainability (Feature Interactions)

In this section we go into great detail on the study of feature interactions and related additive explainability approaches. If you want to know more about applications and are not already sold on the importance of feature interactions, you should probably skip this section for now. You can return later to the relevant topics as necessary. If you are interested in feature interactions, but not deeply invested, I recommend skimming through the sections giving historical details and focusing on the key connections highlighted between different areas of the literature.

3.1 A History of Feature Interaction Detection

In order to have a detailed understanding of the history of feature interactions, we must begin at nearly the very beginning of statistics and start by first discussing the Design of Experiments (DOE).

First DOE Wave (Agricultural) A discussion of the history of feature interactions must undoubtedly start with Ronald A. Fisher's work on the analysis of variance (ANOVA) in 1925 (Fisher, 1925) which built upon decades of agricultural data collected by John Lawes and Joseph Gilbert at the Rothamsted Experimental Station, originally established in 1843. This research was further pushed by Frank Yates of the same institution, furthering the work on ANOVA (Yates, 1935). A major clarification on the assumptions which were required for this approach were later provided in 1947 by Churchill Eisenhart (Eisenhart, 1947). ANOVA was then developed further by John W. Tukey who in 1949 helped isolate the key property of a feature interaction as the non-additivity between the two independent variables of the contingency table (Tukey, 1949). Despite the focus on interactions, this was still a time where running experiments was extremely costly and accordingly higher-order feature interactions (interactions of three or more variables) were often completely ignored or assumed to be zero in order to minimize the number of required experiments. Further developments continued on factorial designs such as those taken by M. B. Wilk in 1955 to extend to the generalized randomized block design (Wilk, 1955) and by B. V. Shah in 1960 for balanced factorial experiments (Shah, 1960).

Second DOE Wave (Industrial) Around this time, the experimental design literature saw a second revitalization with G. E. P. Box and K. B. Wilson's work on applying similar ideas to industrial applications (Box and Wilson, 1951). It is here where we see the clear introduction of continuous variables and a notation much more similar to the modern approaches. Specifically, a set of experimental variables, which are measurable and controllable, are chosen and an experiment is 'run', leading to a functional *response surface* in the observed outcome variable, often with some potential error or confounding due to factors beyond the control of the experimenter. Despite these many modernizations, the focus remained on minimizing the required number of costly experiments and discarding higher-order effects. Later improvements by Box and Hunter (Box and Hunter, 1957) and Box and Behnken (Box and Behnken, 1960) continued to refine the response surface approach, still focusing on optimizing performance with minimal experiments. Developments continued with the treatment of experiments with mixtures like in the production of rubber or alloys (Scheffe, 1958) as well as extended to more complex time-confounded errors (Geisser and

Greenhouse, 1958; Huynh and Feldt, 1970). Industrial applications in the post-war era expanded over the next decades beyond chemistry into tool-life, foodstuffs, biology, ecology, and manufacturing (Myers et al., 1989). This later evolved into the 'quality assurance' era, mimicking the Japanese industrial boom which used statistical methods pioneered by researchers like Taguchi (Taguchi, 1986), and ultimately evolving into the Total Quality Management (TQM) and Continuous Quality Improvement (CQI) approaches which came to define American industrial management by the end of the century.

Third DOE Wave (Informational) Although the industrial experimentation era would go on to become majorly corporatized, the design of experiments would see two spiritual successors. A spiritual successor to the focus on experimental design is most likely the modern research of AB experimental design, as used liberally by tech companies to quickly gain insight into user behavior (Quin et al., 2024). With the digitization of many user experiences, large-scale tech companies are able to receive near-instantaneous feedback, both direct and indirect, allowing for rapid quality improvements in distributed software. Further discussion on this direction will mostly be out of scope for this survey. A spiritual successor to the focus on the response surface methodology is better represented by the literature on Sensitivity Analysis (SA). Also fueled by the modern digitization, or more specifically the proliferation of computerized simulation, the modern researcher may have access to a much greater number of experiment runs. Earlier works like (McKay et al., 1979) and (Sacks et al., 1989) set the stage for computer-based experiments, and sensitivity analysis ultimately developed around the modern techniques centered on Sobol' analysis (Sobol', 1990; Chan et al., 1997; Saltelli et al., 2000; Santner et al., 2003). We will discuss this direction of sensitivity analysis in much greater detail in Sections 3.2.7 and 4.5.3.

Feature Interaction Detection (Decision Trees) During this time, the precursors to the modern decision tree like Automatic Interaction Detection (AID) (Belson, 1959; Morgan and Sonquist, 1963) and the Chi-squared Automatic Interaction Detection (CHAID) (Kass, 1980) were already under development. These methods were novel approaches for determining how to best combine features to interact within a decision tree. These approaches would later culminate into the famous CART algorithm (Breiman et al., 1984) and ultimately modern decision trees and random forests (Ho, 1995; Breiman, 2001a). By the turn of the century, it was already clear that the major interest in feature interactions would be in the context of accurate prediction models and machine learning.

Feature Interaction Detection (Premodern) Although many works at this time would already use the occasional interaction term (most often polynomial cross terms or tensor product splines), the detection of more general feature interactions remained uncommon (Aiken et al., 1991). Some preliminary works like Ai and Norton (2003) proposed extracting interactions from logit and probit models via mixed partial derivatives and Gevrey et al. (2006) followed up by proposing mixed partial derivatives to extract interactions from shallow neural networks. An important change in perspective occurred after two nearly concurrent papers showed how to leverage the powerful decision tree models to detect meaningful interactions, with Friedman and Popescu (2008) and Sorokina et al. (2008) both restricting the features to be included in the trees of a random forest and a similar statistic to measure the strength of an interaction between two or more features. Of major importance was the setup which did not necessarily require the use of a specific choice of ML model, enabling the use of the powerful random forest model. Later works provided clarification of the sparse selection problem and the strong heredity principle (Bien et al., 2013) as well as formulating the screening problem for high-dimensional statistics (Hao and Zhang, 2014a). These developments continue into the present day; however, they become increasingly intertwined with related areas like explainability, sensitivity analysis, and additive models. Accordingly, we will pause in this timeline to introduce these other research areas, first focusing on blackbox explanations, which later becomes almost indistinguishable from interaction detection itself.

Additional details on the topics of experimental design and feature interaction detection are available in the surveys Dean et al. (2015), Rosenberger (2019), and Tsang et al. (2021). We will pick up our discussion on interactions in Section 3.2.6 after the introduction of interaction-based generalizations of the Shapley value.

3.2 A History of the Shapley Value (Additive Explanations)

The Shapley value itself was introduced in 1953 by Lloyd Shapley (Shapley, 1953), devised to equally distribute the rewards of some collaborative project amongst its constituent players. It was a fundamental concept in collaborative game theory, inspiring many extensions we will later discuss. It was brought into popularity for machine learning interpretability literature much later in 2017 (Lundberg and Lee, 2017). Generally, we will use the Shapley value to refer to the game-theoretic concept and SHAP to refer to the ML explainability concept; however, we will occasionally use Shapley to describe the latter (e.g. conditional Shapley). Although the Shapley value does not focus on interactions, either in its original conception or its ML revitalization, it has become a centerpiece or a rallying cry for these research directions which focus on interpretable features, including the study of feature interactions. Accordingly, we will center our discussion in the next sections as if Shapley were the unifying force behind all of these areas, even if for nearly all these areas, Shapley played no part in the origin of the field.

3.2.1 The Shapley Value

The Shapley value was designed to solve the problem of fair attribution amongst a set of d players under a fixed *value function*, v. This function (originally assumed to be superadditive in the spirit of cooperation) maps a coalition of players to the value generated by their collaborative efforts, $v : \mathcal{P}([d]) \to \mathbb{R}$ where $\mathcal{P}([d])$ denotes the power set of $[d] := \{1, \ldots, d\}$. That is, for each coalition (subset) of players $S \subseteq [d]$, we write their combined value as $v(S)$.

In Shapley's original formulation of his solution concept, he used three axioms: the symmetry axiom, the carrier axiom, and the additivity axiom. These are now more commonly decomposed into four axioms, dividing the carrier axiom into the dummy axiom and the efficiency axiom. We write the four Shapley axioms in the operator notation as follows:

1. **Dummy** If $[\delta_i \circ v](S) = c \in \mathbb{R}$ for all S, then $[\phi_i \circ v] = c$.
2. **Symmetry** $\phi_{\pi(i)} \circ (\pi \circ v) = \phi_i \circ v, \quad \forall \pi \in \Pi_d$
3. **Efficiency** $\sum_{i \in [d]} (\phi_i \circ v) = v([d]) - v(\emptyset)$
4. **Additivity** $\phi \circ (v + w) = \phi \circ v + \phi \circ w$

where we define $[\delta_i \circ v](S) := v(S+i) - v(S-i)$ as the difference in value of a coalition with and without player i, $\pi \in \Pi_d$ is a permutation of d elements, $(\pi \circ v)$ is the game where player i is relabelled as player $\pi(i)$, and $(v+w)$ is the game defined as $(v+w)(S) := v(S) + w(S)$.

Theorem 1. (Shapley, 1953) There is a unique solution concept obeying these axioms, now called the Shapley value, obeying the equation below.

$$\phi_i^{\text{Sh}}(v) := \sum_{S \subseteq ([d]-i)} \frac{1}{d} \binom{(d-1)}{|S|}^{-1} \cdot \Big[v(S+i) - v(S) \Big] \tag{1}$$

Although this is a closed form directly in terms of the value function, it obscures the nature of Shapley's solution. Instead, let us first decompose the value function into its composite 'unanimity games', also called the synergy functions, the purified value function, the Mobius function, or the discrete derivative.

$$\tilde{v}(S) := \sum_{T \subseteq S} (-1)^{|S|-|T|} \ v(T) \tag{2}$$

We may then write the Shapley value in a form which exactly describes its approach. The Shapley value takes the unique synergies amongst each set of players and divides it equally amongst all those players.

$$\phi_i^{\text{Sh}}(v) := \sum_{S \ni i} \frac{\tilde{v}(S)}{|S|} \tag{3}$$

Another relevant definition is the Shapley value in its permutation form. Here it becomes very clear how the Shapley value obeys both its symmetry and its efficiency axioms. Unlike what is sometimes erroneously claimed, this was not the original definition and was likely solidified out of collaboration with Martin Shubik on the Shapley variant designed for voting games (Shapley and Shubik, 1954), see discussions below. The Shapley can also be defined as:

$$\phi_i^{\text{Sh}}(v) := \frac{1}{d!} \sum_{\pi \in \Pi_d} \Big[v(S_i^\pi + i) - v(S_i^\pi) \Big] \tag{4}$$

where the sum is taken over all permutations/ orderings of the players, and the set S_i^π is the set of predecessors of i given the permutation π, in other words $S_i^\pi := \{j \in [d] : \pi(j) < \pi(i)\}$. This definition is often used as the basis for Monte carlo sampling via random permutations.

Another important formula is the realization that the Shapley value is the average over marginal contributions where all sizes of subsets are weighed equally. This is easy imagine given the permutation definition, and can be written as:

$$\phi_i^{\text{Sh}}(v) := \sum_{k=0}^{d-1} \frac{1}{d} \cdot \sum_{\substack{S \subseteq ([d]-i) \\ \text{s.t. } |S|=k}} \tbinom{d-1}{k}^{-1} \cdot \Big[v(S+i) - v(S) \Big] \tag{5}$$

where k is 'uniformly distributed' over the set $\{0, \ldots, d-1\}$ and S is 'uniformly distributed' over subsets of size k. From this definition it is clearest to see that we may equally define this over S where $\{i\} \subseteq S \subseteq [d]$ or $\emptyset \subseteq S \subseteq [d]$ instead of the originanl $\emptyset \subseteq S \subseteq ([d] - i)$ (with the appropriate adjustments).

Finally, the last important definition is the variational definition, which defines the Shapley value as the solution to a minimization problem. This alternate definition was discovered much later than the original formulation of the Shapley value.

Theorem 2. (Charnes et al., 1988) The Shapley value can be written as:

$$\phi^{\text{Sh}} := \underset{\phi \in \mathbb{R}^d}{\text{argmin}} \left\{ \sum_{S \subseteq [d]} m(|S|) \cdot \Big[v(S) - \phi_{\emptyset} - \sum_{i \in S} \phi_i \Big]^2 \right\} \tag{6}$$

$$\text{s.t.} \quad v(\emptyset) = \phi_{\emptyset} \ \text{ and } \ v([d]) = \phi_{\emptyset} + \sum_{i=1}^{d} \phi_i$$

where $m(s)$ is taken equal to $m(s) = \binom{d-2}{s-1}^{-1}$.

In other words, we can define the Shapley value as the least squares solution to the 'additive approximation' of the true value function, under some specific measure now often called the *Shapley kernel*, sometimes Shapley kernel measure or distribution.

It is equivalent to scale the measure $m(s)$ by any constant without changing the optimization problem. Although the preferred measure might be the one normalized to a probability measure (to more easily enable random sampling of S), the normalizing constant is not available in a simple closed form. Nevertheless, it is now more popular to use a measure closer to the one independently discovered by Lundberg decades later in the SHAP paper (Lundberg and Lee, 2017). This is the measure defined as $m(s) = \frac{d-1}{\binom{d}{s} \cdot s \cdot (d-s)}$ or equally $m(s) = \frac{1}{d}\binom{d-2}{s-1}^{-1}$. A strong reason to prefer the former is that it reminds us that when $s = 0$ or $s = d$, the measure is implicitly infinite, corresponding to the constraints of the system. A reason to prefer $m(s) = \frac{d-1}{\binom{d}{s} \cdot s \cdot (d-s)}$ over $m(s) = \frac{1}{\binom{d}{s} \cdot s \cdot (d-s)}$ is the fact that the total measure of the former only grows logarithmically, rather than quickly decaying almost linearly. Nevertheless, the latter is probably the simplest equivalent measure with these nice properties. Let us now move on to discuss the ML application of Shapley by Lundberg and Lee (2017) in more detail, including how they leveraged this variational formulation.

3.2.2 The SHAP Value

The SHAP value, introduced in 2017 by Lundberg and Lee (Lundberg and Lee, 2017), was an effort to unify several recent explainability methods under the roof of 'additive explanations' and then provide an axiomatic characterization of their introduced additive explanation, the SHAP value. It is important to note that this was around the dawn of the push for 'model-agnostic' approaches, meaning they treated the ML model entirely as a black box after abstracting how to remove the features from the model. This would allow them to directly import the set functions from the original Shapley and would continue to influence subsequent developments, not least of which can be seen in the modern notation used throughout this survey.

Definition 1. (**Additive Explanation**) The umbrella class which was studied by Lundberg and Lee (2017) was the additive explanation, defined as:

$$w(S) := \phi_{\emptyset} + \sum_{i=1}^{d} \mathbb{I}(i \in S) \cdot \phi_i \tag{7}$$

The originally proposed axioms were in a now-outdated notation which is related to the baseline method of feature removal. We write them in an updated notation:

1. **Local Accuracy** $v([d]) = \sum_{i=1}^{d} \phi_i$ and $v(\emptyset) = \phi_\emptyset$
2. **Missingness** $v(S+i) - v(S) = 0$ for all $S \implies \phi_i = 0$.
3. **Consistency** $v(S+i) - v(S) \geq v'(S+i) - v'(S)$ for all $S \implies \phi_i \geq \phi_i'$.

Theorem 3. (Lundberg and Lee, 2017) The only additive explainer of the form $w(S) := \phi_\emptyset + \sum_{i=1}^{d} \mathbb{I}(i \in S) \cdot \phi_i$ which obeys Local Accuracy, Missingness, and Consistency is the Shapley value.

Although these alternative automatizations of the Shapley value were already known from the literature below on cooperative game theory, the critical development was to bring these notions into the language of blackbox (i.e. model-agnostic) explainability. It is worth noting that the outdated missingness condition had the different requirement that $\varphi_i = 0$ whenever the feature value $x_i = 0$, where 0 was taken as the baseline value. We reinterpret this meaning that no change would occur for the value function.

Again, much of the original language was limited to the then popular baseline removal method; however, the paper alluded to more general perturbations. In particular, the paper writes:

$$\begin{aligned} v(S) := \mathbb{E}[f(X_S, X_{-S}) | X_S = x_s] &= \mathbb{E}_{p(X_{-S}|x_s)}\Big[f(x_S, X_{-S})\Big] && (8) \\ &\approx \mathbb{E}_{p(X_{-S})}\Big[f(x_S, X_{-S})\Big] && (9) \\ &\approx f(x_S, \mathbb{E}_{p(X_{-S})}\big[X_{-S}\big]) && (10) \end{aligned}$$

where the first line (Equation 8) corresponds to Equations 9 and 10 of (Lundberg and Lee, 2017). The second line (Equation 9) corresponds to Equation 11, achieved by assuming that the input features are independent. The third line (Equation 10) corresponds to Equation 12, achieved by assuming the function is linear in the features. As a reminder, one would typically center their input features so that the third line could be evaluated via $f(x_S, 0_{-S})$. Although these are clearly very heavy assumptions, these are the same simplifications which allowed for SHAP to be relatively computable.

Beyond its clear formulation, another major advantage of the paper was its ease of computation and strong code support. In particular, the previously mentioned optimization formulation from Equation 6 was used for approximating the Shapley value without inheriting the exponential complexity required to compute the Shapley value exactly.

$$\hat{\phi}^{\text{KernelSHAP}} = \underset{\phi \in \mathbb{R}^d}{\text{argmin}} \left\{ \sum_{S \in \mathcal{S}} m(|S|) \cdot \Big[v(S) - \phi_\emptyset - \sum_{i \in S} \phi_i\Big]^2 \right\} \quad (11)$$

$$\text{s.t.} \quad v(\emptyset) = \phi_\emptyset \;\text{ and }\; v([d]) = \sum_{i=1}^{d} \phi_i$$

where $m(s) = \frac{d-1}{\binom{d}{s} \cdot s \cdot (d-s)}$. This algorithmic approach was called *KernelSHAP* because of its use of the Shapley kernel. Once again, we use modern notation instead of the original. The key difference from the exact equation is the use of an approximating set of subsets $\mathcal{S} \subseteq \mathcal{P}([d])$ instead of measure the blackbox function on all 2^d possible subsets. Algorithmically, there was also a choice to always include the smallest and biggest subsets inside $\mathcal{S}$ (i.e. S where $|S| = 1, 2, d-1, d-2$) and there was a small L1 penalty applied to encourage slightly sparse explanations.

This quickly attracted attention from many different perspectives and is currently the most popular and well-studied explainability approach.

3.2.3 Variants of SHAP

As was quickly realized, however, the Shapley value depends critically on how the machine learning problem is translated into a choice of value function. The 'many Shapley values' paper (Sundararajan and Najmi, 2020) was among the first to clearly articulate this point.

They begin by noting how several works prior to the SHAP paper have made attempts to apply the Shapley value solution concept to machine learning applications. (Lindeman, 1980) and (Kruskal, 1987) apply to linear regression's R^2 values, averaging across all orderings of the input features (not realizing the connections with the Shapley value); (Owen, 2014) and (Owen and Prieur, 2017) apply the Shapley value to the variance explained for a general function (not just a linear function); and (Strumbelj et al., 2009), (Strumbelj and Kononenko, 2010), (Strumbelj and Kononenko, 2014), and (Datta et al., 2016) all apply Shapley to model explanations in a very similar way to SHAP for the goal of model explainability (Lundberg and Lee, 2017).

In addition, the work makes connections with existing methods like Integrated Gradients (IG) were made with a variant of the Shapley value (Sundararajan et al., 2017; AUMANN and SHAPLEY, 1974) and mentions follow-up works like (Lundberg et al., 2018) continuing to develop Shapley applications. Alongside soon-to-be-published works like (Lundberg et al., 2020; Covert et al., 2020), this begged the question of what is the right Shapley value, with (Sundararajan and Najmi, 2020) giving a first attempt at answering the question.

Although the question itself was not necessarily answered, (Covert et al., 2021) gave an excellent answer to the many Shapley values question, carefully characterizing and cataloging the many published works on the topic, emphasizing the choice of:

1. summary technique
2. captured model behavior
3. feature removal technique

For our discussion, we will focus on the Shapley summarization technique, although simpler approaches like removing or including an individual feature were also surveyed. For the captured model behavior, this is mostly two possibilities: (i) the case of explaining an individual prediction; or (ii) the case of total loss of the method (e.g. the R^2 or remaining variance in the case of regression).

Finally, the critical choice of feature removal technique generally resulted in three main approaches: replace by a baseline feature, marginalize out a feature by its marginal distribution, or marginalize out a feature by its conditional distribution. In the context of explaining a prediction, this results in:

$$v^{\text{base}}(S) := \mathcal{M}_S^{\text{base}} \circ f \qquad := \quad f(x_S, \bar{x}_{-S}) \tag{12}$$

$$v^{\text{marg}}(S) := \mathcal{M}_S^{\text{marg}} \circ f \qquad := \mathbb{E}_{\bar{X}_{-S} \sim p(X_{-S})}\Big[f(x_S, \bar{X}_{-S})\Big] \tag{13}$$

$$v^{\text{cond}}(S) := \mathcal{M}_S^{\text{cond}} \circ f := \mathbb{E}_{\bar{X}_{-S} \sim p(X_{-S}|X_S = x_S)}\Big[f(x_S, \bar{X}_{-S})\Big] \tag{14}$$

There are additional methods for removing features are there are always continuing developments on how to appropriately remove a feature within a certain context. Some vision-specific examples include image blurring (Fong and Vedaldi, 2017; Fong et al., 2019) and conditional VAEs/ GANs (Chang et al., 2019). There are also notions of setting a feature to a certain value, such as following Pearl's *do*() operator framework (Heskes et al., 2020; Jung et al., 2022). They show how there can be a sensitive and counterintuitive results if one believes Shapley represents a causal influence without understanding the underlying causal structure.

Debate over these choices in how to reduce to the value function continue into the present day. Despite (Lundberg and Lee, 2017) and predecessors making it explicit that the choice of marginal and baseline are only a simplifying choice for computational reasons, there are still researchers who argue for the use of the marginal value. A major argument is that of (Janzing et al., 2020) which argues that under the independence assumption, the leads to a causal interpretation, with respect to the model's prediction, calling it the interventional Shapley instead of the marginal Shapley.

Another major discomfort from those supporting the marginal approach (besides the preference for computationally easier values) is regarding the apparent paradox that a model which exclusively uses A instead of B can be found to have importance for feature B due to the heavily correlation between A and B (Merrick and Taly, 2020). This is in direct contrast to those who take the statistical perspective that A and B contain the same information content, making models which use A or B fundamentally similar (Adler et al., 2018). (Frye et al., 2021) provides a convincing argument for why researchers should never use the baseline or marginal approaches for correlated features, namely that when making queries to the model 'off the manifold' where it saw training data, this leads to the classic 'garbage in, garbage out' issue and leads to unpredictable behavior from the model. This sentiment has been further echoed by other researchers working on feature-based explanations (Hooker and Mentch, 2019; Yeh et al., 2022).

We implore those authors who sympathize with the name of 'interventional Shapley' to justify which of their favorite datasets have independent features. Nevertheless, many researchers continue to be attracted to this style of explanation, either due to the computational simplicity or due to being interested in a more mechanistic explanation of which inputs will change the output. Although completely ignoring the data manifold can only lead to catastrophically irrelevant explanations, the most important considerations are justification of the explanation technique for the particular application at hand.

Finally, before moving on to the next section, we will discuss some of the model-specific variants of SHAP. Generally, these are some approaches which are specially designed for certain architectures or thereby have some computational advantages. The most major of these should be the approach of TreeSHAP (Lundberg et al., 2018) which was one of the earliest approaches to have a computationally easy way to calculate the SHAP value for the specific architecture of decision trees or random forests. Although it was not well understood at its origin, it is now better understood that this algorithm requires the assumption of independent input features to calculate the marginal SHAP value, and otherwise computes a model-specific version of the SHAP value which is neither the marginal or conditional (Amoukou et al., 2022; Filom et al., 2024). Nevertheless, due to its great popularity at the time, several extensions to enhance the algorithm, including GPU Tree SHAP (Mitchell et al., 2022b) and Fast Tree SHAP (Yang, 2022).

Another major class of interesting machine learning algorithms are those of kernel machines. This includes the method RKHS SHAP (Chau et al., 2022) for reproducing kernel Hilbert spaces, which again unfortunately is restricted to the assumption of independent features by nature of the tensor product structural assumption on the Hilbert spaces. Then there is the extension from RKHS SHAP to GP SHAP (Chau et al., 2023) which extends to Gaussian Processes. There they also consider the method of Bayes-GP-SHAP as an attempt to integrate both the GP posterior uncertainty and the SHAP approximation uncertainty into the same approach. The generalized additive model has also seen specific algorithms due to the similar additive structure of each (Bordt and von Luxburg, 2023; Enouen and Liu, 2025). These methods also face challenges in interpretation under heavily dependent features.

In addition to these model-specific enhancements, there has also been many works focusing on faster and more efficient approximation of the Shapley value, especially extending the two major approximation algorithms of permutation sampling and the kernel approximator. This includes empirical studies of permutation sampling (Mitchell et al., 2022a), functional extensions of kernel estimation (Jethani et al., 2022), approximating without marginal contributions (Kolpaczki et al., 2024), and many more.

3.2.4 SHAP for Data Valuation

As discussed thoroughly in the previous section, the major effort of defining the Shapley value for a certain problem is in the setup of the value function –which is to then be summarized by the Shapley solution. The previous section focused entirely on how the removal of a feature would affect an individual prediction or the total loss; however, this section will instead look at removing a data point. Many of these works focus on explaining the loss rather than the prediction. This is another key set of value functions to explore for interpreting machine learning models and has helped spark an entirely different direction of research.

Although influence functions had already been classically proposed for this problem (Cook, 1977; Cook and Weisberg, 1980, 1982) and also modernized for the modern machine learning era (Koh and Liang, 2017), there was an explosion of research on the topic after the introduction of the Data SHAP method (Ghorbani and Zou, 2019; Jia et al., 2019b).

$$V(S) := \mathbb{E}_{\hat{\theta}\sim p_{S,\mathcal{A}}(\hat{\theta})}\Big[\mathcal{L}_{\text{test}}(f_{\hat{\theta}})\Big] \quad \text{s.t.} \quad \hat{\theta} \approx \underset{\theta}{\operatorname{argmin}} \left\{\mathcal{L}_{\text{trn},S}(f_\theta)\right\} \tag{15}$$

where $\mathcal{A}$ denotes the specific algorithm chosen to minimize the empirical risk $\mathcal{L}_{\text{trn},S}$ over the subset $S \subseteq [n]$ of the full n training data points. $p_{S,\mathcal{A}}(\hat{\theta})$ denotes the distribution over the approximately optimal $\hat{\theta}$ returned by algorithm $\mathcal{A}$ when trained on data subset $(x^{(j)}, y^{(j)})_{j \in S}$. For a deterministic algorithm $\mathcal{A}$, this will be a Dirichlet distribution with 100% of its mass on the returned parameters. For simplicity, let us also write $z := (x, y) \in \mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ with training dataset $Z_{\text{trn}} = \{z^{(i)}\}_{i=1}^{n} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$.

The complexity of retraining a model for each coalition of data points, a requirement for computing this value function at a value S, is extremely costly, even for moderately sized ML models. This is the extremely major disadvantage of this formulation and has been a challenge of this research direction since the beginning. This is also why alternative approaches like TracIn (Pruthi et al., 2020) and representer points (Yeh et al., 2018) avoid retraining entirely, instead using gradient approximations or representer theorems, respectively.

Despite these many challenges, the allure of an approach for fairly dividing value amongst the different data points has resulted in continuing developments of this research direction. First, (Ghorbani et al., 2020) removes the dependence on a specific dataset context and reframes the data SHAP value as a dataset-dependent value to a distributionally-dependent value. The original data SHAP can be written like Equation 5 as:

$$\phi_i^{\text{DataSHAP}} = \sum_{k=0}^{n-1} \frac{1}{n} \cdot \sum_{\substack{S \subseteq ([n]-i) \\ \text{s.t. } |S|=k}} \binom{n-1}{k}^{-1} \cdot \Big[V(S+i) - V(S) \Big] \Big] \tag{16}$$

where it becomes more clear that from the perspective of data point i, S serves the role of a random data subset of size k. The distributional Shapley value (Ghorbani et al., 2020) makes this clear by actually sampling the $z_S = (x_S, y_S)$ randomly from a distribution, $\mathcal{D}$, instead of subsampling from a fixed dataset $(x^{(j)}, y^{(j)})_{j \in S}$.

$$\phi^{\text{DistSHAP}} := \sum_{k=0}^{n-1} \frac{1}{n} \cdot \mathbb{E}_{Z \sim \mathcal{D}^k} \Big[V(Z + z^{(i)}) - V(Z) \Big] \tag{17}$$

where Z is a random dataset of size k, sampled i.i.d. according to distribution $\mathcal{D}$. We also abuse the notation of V, allowing it to take data points instead of subsets which represented those data points, extended in the obvious way.

Although very similar to the original definition, there is no longer a dependence on the fixed dataset which was drawn, much like how the influence of a random algorithm was averaged out in the original definition. Both of these choices abstract away details of the implementation process while maintaining the critical details to define the data valuation process, mirroring the developments of blackbox explainability.

Similar to the 'model-specific' SHAP variants, model-specific data SHAP variants have also emerged. Unsurprisingly, the data influence of a hard-label k-NN model is relatively easy to compute (Jia et al., 2019a). Further calculations were done for the distributional SHAP on linear models, logistic models, and kernel density estimators (Kwon et al., 2021). Follow-up work continued to improve the efficiency of DataSHAP estimation for weighted k-NN classifiers (Wang et al., 2024a).

Further developments continued, introducing variations of the Shapley value for explaining the value functions. In particular, removing the efficiency axioms resulted in a 'semivalue' (discussed in the next section) with approaches like Beta Shapley (Kwon and Zou, 2022), Least Core for data attribution (Yan and Procaccia, 2021), and Data Banzhaf (Wang and Jia, 2023). Follow up works have contextualized the limitations of Data SHAP and variants for downstream tasks like data selection (Wang et al., 2024b). Other works have managed to come full circle, reframing original gradient approximations back within the Shapley framework (Wang et al., 2025). Further details on data valuation can be found in (Hammoudeh and Lowd). We will now transition into a discussion of semivalues and other probabilistic values in greater detail.

3.2.5 Extensions of Shapley - Probabilistic Values

The cooperative game theory literature did not stop advancing after the introduction of Shapley's 1953 method, and many further extensions of the Shapley value have been developed in the half a century which it took for the ML community to pick up on the value. In alternate contexts, many of the defining axioms of the Shapley value may be important to get rid of. Many of these approaches have also been incorporated into the ML interpretability literature.

The largest class of these are the *probabilistic values* which are generally defined as solutions which only obey the linearity axiom and the dummy axiom (dropping symmetry and efficiency). These are further categorized as efficient probabilistic values, also called *quasivalues*, which obey every axiom except for symmetry, and symmetric probabilistic values, also called *semivalues*, which obey every axiom except for efficiency.

Definition 2. (Weber, 1988) A **probabilistic value** can be defined as those values which, for each i, obey:

$$\phi_i(v) = \sum_{S \subseteq ([d]-i)} p^i(S) \cdot \Big[v(S+i) - v(S) \Big] \tag{18}$$

for some probability distribution $p^i(S)$ over all of the coalitions without i, namely $S \in \mathcal{P}([d]-i)$. In other words, a probabilistic value can be defined by an averaging over its marginal contributions for some specified averaging scheme which depends on i.

Theorem 4. (Weber, 1988) A solution is a probabilistic value iff it obeys the additivity (linearity) axiom and dummy axiom.

This covers a fairly wide number of solution concepts of interest, including the Banzhaf value, Owen value, and Myerson value. Of course lifting both of the symmetry and efficiency axioms without regard is uncalled for, and it is commonly of interest to lift only one of the symmetry and efficiency axioms, called the semivalue and quasivalue (random order value).

Definition 3. (Weber, 1988) A **random order value** (or quasivalue) can be defined as those values which, for each i, obey:

$$\phi_i(v) = \sum_{\pi \in \Pi_d} p(\pi) \cdot \Big[v(S_i^\pi + i) - v(S_i^\pi) \Big] \tag{19}$$

for some distribution p over all permutations of $[d]$, namely $\pi \in \Pi_d$.

Theorem 5. (Weber, 1988) A solution is a random order value iff it obeys additivity, dummy, and efficiency.

It becomes clear why this is a random order value because we order the players randomly and then take their marginal contribution according to ordering. This extends Shapley's weighted value (Shapley, 1953; Kalai and Samet, 1987) and the more general *path value* (Owen, 1972). Another specific random-order value of interest is the Owen value (Owen, 1977) with developments and extensions by (Hart and Kurz, 1983) and (Winter, 1989).

Definition 4. (Weber, 1988) A **semivalue** can be defined as those values which, for each i, obey: distribution over coalition sizes

$$\phi_i(v) = \sum_{S \subseteq ([d]-i)} p^i(|S|) \cdot \Big[v(S+i) - v(S)\Big] \tag{20}$$

for some 'distribution' $p^i(s)$ over the possible sizes of S, i.e. $\sum_{s=0}^{d-1}[\binom{d-1}{s} \cdot p^i(s)] = 1$.

Here, the Banzhaf value (Banzhaf, 1965) is the greatest representative here, with also integral extensions in a similar fashion to path values providing a bijection with possible semivalues (Dubey and Weber, 1977; Dubey et al., 1981).

Definition 5. (**Banzhaf value**) the Banzhaf value may be defined in its 'closed form' solution in terms of using each $v(S)$ only once as:

$$\phi_i^{\mathrm{Bz}}(v) = \sum_{S \subseteq ([d]-i)} \frac{1}{2^{d-1}} \cdot \Big[v(S+i) - v(S)\Big], \tag{21}$$

but this can easily be seen to be equal to taking the same sum over all 2^d possibilities or by grouping by marginal size, making the semivalue formula, Equation 19, clear:

$$\phi_i^{\mathrm{Bz}}(v) = \sum_{k=0}^{d-1} \frac{1}{2^{d-1}} \cdot \sum_{\substack{S \subseteq ([d]-i) \\ \text{s.t. } |S|=k}} \Big[v(S+i) - v(S)\Big] \tag{22}$$

Lastly, the Banzhaf value can also be put into its purified form in terms of the purified contributions:

$$\phi_i^{\mathrm{Bz}}(v) := \sum_{S \ni i} \frac{\tilde{v}(S)}{2^{(|S|-1)}} \tag{23}$$

Important extensions which do not obey either of the original symmetry or efficiency axioms originally often incorporate some alternative structure: the AD value (Aumann and Dreze, 1974) respects a partition of the players (obeying partition-symmetry and partition-efficiency) and the Myerson value (Myerson, 1977) respects graphical connections between the players. The Owen-Winter value (Owen, 1977; Winter, 1989) also respects hierarchical partitioning, but obeys normal efficiency. A great resource for starting to explore these topics in greater depth is (Winter, 2002; Monderer and Samet, 2002).

Important topics related to these approaches but somewhat beyond the individual methods are the potential function (Hart and Mas-Colell, 1989), the Harsanyi dividend (Harsanyi, 1963), and monotonic solutions (Kalai and Samet, 1985). Additionally, solution concepts beyond the probabilistic values are: the von-Neumann-Morgenstern solution (von Neumann et al., 1944), the core (Gillies, 1959; Shapley, 1965), the least core (Maschler et al., 1979), and the nucleolus (Schmeidler, 1969).

Simple Monotone Games These axioms and solutions have also been quite extensively studied in the context of what are called simple monotonic games or voting games. In the machine learning language, this is interested in dealing with the case of binary classification rather than regression (players voting for the result). These games are both *simple* and *monotone*.

Definition 6. (Simple) $v : \mathcal{P}([d]) \to \{0, 1\}$

Definition 7. (Monotone) If $v(S) = 1$ and $R \supseteq S$, then $v(R) = 1$.

where v takes the interpretation of describing a coalition S as 'winning' (1) or 'losing' (0). It is very often in the context of voting, where a coalition represents the voters willing to vote for the bill.

It is here actually, where the Banzhaf value (Banzhaf, 1965), or occasionally Penrose-Banzhaf value (Penrose, 1946), actually predates the Shapley value, or more specifically in this context the Shapley-Shubik value (Shapley and Shubik, 1954), in its original 1946 conception. In this new context of voting games, many axioms need to be slightly reinterpreted. It is here where these values are usually called an 'index of power' instead of a 'value' to refer to the voting power held, rather than the monetary value distributed.

It is here where the concept of the 'minimal winning coalition' also becomes a more critical aspect of the definition.

Definition 8. (Winning Coalition) A set S is a winning coalition if $v(S) = 1$. Due to monotonicity, this implies $v(R) = 1$ for all $R \supseteq S$.

Definition 9. (Minimal Winning Coalition) A set S is a minimal winning coalition if it is a winning coalition containing no other winning coalition. $v(S) = 1$, but $v(S - i) = 0$ for all $i \in S$.

These are necessary to define the Deegan-Packel index (Packel, 1978), which weights each MWC S as $\frac{1}{|S|}$ to all $i \in S$ and averages this amount over all MWCs, $\mathcal{I}_{\text{MWC}} = \{S : S \text{ is MWC}\}$.

Definition 10. (**Deegan-Packel index**)

$$\phi_i^{\text{DP}} := \frac{1}{|\mathcal{I}_{\text{MWC}}|} \sum_{S \in \mathcal{I}_{\text{MWC}}} \frac{\mathbb{1}(i \in S)}{|S|} \tag{24}$$

Relatedly, we will call a player i *critical* for coalition S if removing i results in a losing vote, we will call a set's *critical number* the number of critical players, and we will call a set critical if it has critical number greater than zero. Write $\mathcal{I}_{\text{crit}} = \{S : S \text{ is critical}\}$.

Definition 11. (Critical Player) A player i is critical for S if S is winning but $S - i$ is not: $v(S) = 1$ and $v(S - i) = 0$.

Definition 12. (Critical Number) A set S has critical count, $\text{Crit}(S) = |\{i \in S : i \text{ is critical in } S\}|$.

These are necessary to define the Johnston index (Johnston, 1977), which replaces the Banzhaf uniform scoring across critical sets with a balance based on the number of players which criticality is shared with.

Definition 13. **(Johnston Index)**

$$\phi_i^{\text{Js}} := \frac{\text{c}_i^{\text{Js}}}{\sum_j \text{c}_j^{\text{Js}}} \qquad \textit{where} \qquad \text{c}_i^{\text{Js}} := \sum_{S \in \mathcal{I}_{\text{crit}}} \frac{\mathbb{1}(i \text{ is critical in } S)}{\text{Crit}(S)} \tag{25}$$

It is here where I also emphasize that the commonly used phrase ‘game-theoretic’ explainability is really a stretch. In particular, although the Shapley value has its roots in cooperative game theory, extending the analogy beyond superadditive value functions and monotone voting functions where these game-theoretic approaches draw their conceptual origins really strains this analogy. It can be seen how these DP and Js indices provide a first look at the type of indices which require more than the additive ‘mobius’ structure and begin to focus on the game-theoretic aspects of the problem. Some additional values in this direction are the Holler-Packel index (Holler, 1978; Holler and Packel, 1983; Napel, 1999; Holler and Napel, 2004) and the shift minimal winning coalition index (Alonso-Meijide and Freixas, 2010; Alonso-Meijide et al., 2012). A great resource for an introduction into the measurement of voting power is (Felsenthal and Machover, 1998).

3.2.6 Extensions of Shapley - Feature Interactions

Although we have already established the long history of studying the detection of feature interactions, in the modern context of explainability for machine learning, we find it convenient to group all of the interaction-based explanations as though they were motivated by SHAP, even though this is not truly the case.

We begin by recalling the purified form of the game by the Mobius transform from Equation 2:

$$\tilde{v}(S) := \sum_{T \subseteq S} (-1)^{|S|-|T|} \ v(T) \tag{26}$$

We now define the discrete derivative operator, δ_S, by writing $w = \delta_S \circ v$ as:

$$w(T) = [\delta_S \circ v](T) := \sum_{R \subseteq S} (-1)^{|S|-|R|} \cdot v(T - S + R) \tag{27}$$

It can be noticed that $[\delta_S \circ v]$ can equally be defined only on $T \subseteq ([d] - S)$ because for any such T, $w(T) = w(T + Q) = w(S)$ for all $Q \subseteq S$. This is similar to how Shapley values and probabilistic values are only defined over $([d] - i)$. It can also easily be seen that:

$$\tilde{v}(S) = [\delta_S \circ v](\emptyset) \tag{28}$$

We may now define the simplest feature interaction explanations which are the inclusion and removal values, ϕ^{inc} and ϕ^{rem}, in terms of the discrete derivative, δ_S:

$$\phi_S^{\text{inc}}(v) := [\delta_S \circ v](\emptyset) = \sum_{R \subseteq S} (-1)^{|S|-|R|} \cdot v(R) \tag{29}$$

$$\phi_S^{\text{rem}}(v) := [\delta_S \circ v]([d]) = \sum_{Q \subseteq S} (-1)^{|Q|} \cdot v([d] - Q) \tag{30}$$

From this it can be seen that:

$$\phi_S^{\text{inc}}(v) = \tilde{v}(S) \qquad \phi_S^{\text{rem}}(v) = \sum_{T \supseteq S} \tilde{v}(T) \tag{31}$$

The line of research from the cooperative game theory literature, despite the large variety of solution concepts for value allocation and voting power, would not study the interaction strength between two players until 1999. Measuring the strength of a pair of players rather than their individual strengths first appeared in the work of (Grabisch and Roubens, 1999) which extended the Shapley axioms to be able to apply to this different situation. Solution concepts for this scenario are generally called interaction indices instead of values based off this first solution being called the Shapley Interaction Index. The solution was obtained by combining the two players i and j together into a joint player and then applying the Shapley value, comparing that against the individualized contributions of those same players. The typical definition is given as

$$\phi_S^{\text{SII}}(v) = \sum_{T \subseteq [d]-S} \frac{(d-|S|-|T|)!|T|!}{(d-|S|+1)!} \sum_{L \subseteq S} (-1)^{|S|-|L|} v(L+T), \tag{32}$$

which can be shown to be equal to

$$\phi_S^{\text{SII}}(v) = \frac{1}{d-|S|+1} \sum_{T \subseteq [d]-S} \binom{d-|S|}{|T|}^{-1} \sum_{R \subseteq T} \tilde{v}(S+R), \tag{33}$$

and with just a little more work is equal to

$$\phi_S^{\text{SII}}(v) = \sum_{R \subseteq [d]-S} \frac{1}{(|R|+1)} \tilde{v}(S+R), \tag{34}$$

which is nearly the same as the purified version of the Shapley value from Equation 23.

Modern attempts to create another Shapley interaction index would not appear until two decades later with in the explainability literature: the Shapley-Taylor Interaction Index (Sundararajan et al., 2020), the Faithful Shapley Interaction Index (Tsai et al., 2023), and the k-Shapley Interaction Index (Bordt and von Luxburg, 2023).

Accordingly, it is here where our story of feature interaction detection picks back up as the historical approaches bleed into the modern approaches working on interaction detection from the perspective of explainability. As we left off in the end of Section 3.1, the problem of interaction detection was beginning to gain traction in high-dimensional statistics with (Bien et al., 2013) and (Hao and Zhang, 2014a) using heredity to ensure detection of interactions even in the presence of high-dimensional data and few available samples. Follow-up works continued to look at selecting from all pairwise interactions without having to face the full quadratic complexity in terms of the input dimension (Lim and Hastie, 2015; Bien et al., 2015; Kong et al., 2017).

Around this time, the importance of interactions for the interpretability literature was starting to become increasingly well-known. Existing tools like LIME and SHAP were ill-equipped for handling or describing the interaction between inputs like in word negation in natural language and part hierarchies in computer vision. Early works like NID (Tsang et al., 2018a) and CD (Murdoch et al., 2018) gave early architecture-specific approaches for understanding the nonlinear interactions occurring within neural networks.

Important follow-up works use model-agnostic approaches for the detection of interactions. The Shapley-Taylor interaction index (Sundararajan et al., 2020) uses a Taylor-series-like expansion which gives all interactions of degree less than k exactly their purified contribution and splits the remaining contributions amongst those interactions of degree k.

$$\phi_S^{\text{ShapleyTaylor-}k}(v) = \begin{cases} \tilde{v}_S, & \text{if } |S| < k \\ \sum_{T \supseteq S} \binom{|T|}{|S|} \cdot \tilde{v}_T, & \text{if } |S| = k \end{cases} \tag{35}$$

which is designed to satisfy efficiency when considering the entire set of singles and pairs $\mathcal{I}^{\leq 2} := \{S : S \subseteq [d], |S| \leq 2\}$.

The approach of Archipelago (Tsang et al., 2020b) also took a model-agnostic approach to define the importance of an interaction

$$\phi_S^{\text{ArchDetect}}(v) = \frac{1}{2}\phi_S^{\text{inc}}(v)^2 + \frac{1}{2}\phi_S^{\text{rem}}(v)^2 \tag{36}$$

$$\phi_S^{\text{ArchAttrib}}(v) = v(S) - v(\emptyset) \tag{37}$$

where we write the definition in terms of the inclusion and removal values, ϕ^{inc} and ϕ^{rem}, defined above in Equation 30. Note that the $\phi^{\text{ArchDetect}}$ score is designed to be positive to measure the strength of an interaction, whereas $\phi^{\text{ArchAttrib}}$ is can be positive or negative to measure the overall effect of the feature set S. Follow-up works also consider the consider the Archipelago defined as an interaction index as:

$$\phi_S^{\text{Archipelago}}(v) = \frac{1}{2}\phi_S^{\text{inc}}(v) + \frac{1}{2}\phi_S^{\text{rem}}(v), \tag{38}$$

which has both $\phi_S^{\text{ArchDetect}} \neq (\phi_S^{\text{Archipelago}})^2$ and also $\phi_S^{\text{Archipelago}} \neq \phi_S^{\text{ArchAttrib}}$.

Not too long after, further attempts at defining a ‘correct’ interaction index extending the Shapley value would be developed from the interpretability literature. The Faithful Shapley Interaction Index (Tsai et al., 2023) or Faith-SHAP would be introduced by extending the variational formulation of Shapley from Equation 6. Equation 6 is first reformulated as Equation 39 and then extended to the definition in Equation 40.

$$\phi^{\text{SHAP}} = \operatorname*{argmin}_{\phi \in \mathbb{R}^d} \left\{ \sum_{S \subseteq [d]} m(|S|) \cdot \left[v(S) - \phi_\emptyset - \sum_{i=1}^{d} \mathbb{I}\big(\{i\} \subseteq S\big) \cdot \phi_i \right]^2 \right\} \tag{39}$$
$$\text{s.t.} \quad v(\emptyset) = \phi_\emptyset \ \text{ and } \ v([d]) = \phi_\emptyset + \sum_{i=1}^{d} \phi_i,$$

$$\phi^{\text{FaithSHAP-}k} := \operatorname*{argmin}_{\phi \in \mathbb{R}^d} \left\{ \sum_{S \subseteq [d]} m(|S|) \cdot \left[v(S) - \sum_{T \in \mathcal{I}^{\leq k}} \mathbb{I}\big(T \subseteq S\big) \cdot \phi_T \right]^2 \right\} \tag{40}$$
$$\text{s.t.} \quad v(\emptyset) = \phi_\emptyset \ \text{ and } \ v([d]) = \sum_{T \in \mathcal{I}^{\leq k}} \phi_T,$$

where $m(s)$ is taken equal to $m(s) = (\binom{d}{s} \cdot s \cdot (d-s))^{-1}$ as before and $\mathcal{I}^{\leq k} := \{S : S \subseteq [d], |S| \leq k\}$. This definition simply takes the one-dimensional additive model and extends it to the k-dimensional additive model in trying to capture the 'best approximation' to the true value function $v(S)$.

Around the same time, the k-Shapley value or k-Shapley interaction index (Bordt and von Luxburg, 2023) was defined as the correction the SII value which additionally obeys the efficiency axiom. This is achieved by setting the largest subsets with $|S| = k$ as the SII value and recursively defining lower-order values such that they obey efficiency:

$$\phi_S^{k\text{-SII}}(v) = \begin{cases} \phi_S^{\text{SII}}(v), & \text{if } |S| = k \\ \phi_S^{(k-1)\text{-SII}}(v) + B_{k-|S|} \cdot \sum_{\substack{R \text{ s.t.} \\ R \supseteq S, |R|=k}} \phi_R^{\text{SII}}(v), & \text{if } |S| < k \end{cases} \tag{41}$$

where B_n are the Bernoulli numbers (which can be defined as the sequence obeying $B_0 = 1$ otherwise $\sum_{i=0}^{n} \binom{n+1}{i} B_i = 0$). These are simply the combinatorial numbers required to balance the coefficients of the interaction index to sum to the full function across all subsets $S \in \mathcal{I}^{\leq k}$.

Although there are several more approaches to explaining interactions in the literature, many of these approaches are model-specific or tied to a specific application domain. We will cover these in Section **??**. Accordingly, this already covers the majority of model-agnostic approaches to understanding interactions. There is one final approach to interactions which is actually quite classical, this is the area of sensitivity analysis from the 1990s. Here, we are usually interested in how important an interaction is for the entire dataset, rather than for the individual sample. Nevertheless, the approaches are fundamentally related and will help further cement the relationships between all of these slightly different areas.

3.2.7 Relation with Sensitivity Analysis

It has become increasingly popular to understand the connections with the functional ANOVA decomposition or what is often called the Mobius transform in XAI literature (Herren and Hahn, 2022; Fumagalli et al., 2025). In this section, we give a first look at this connection between the Shapley value, before a deeper understanding of this three-way connection between feature interactions, additive models, and sensitivty analysis are further explored in Section 4.5.3.

Variance-based sensitivity analysis, sometimes called the method of Sobol', is based on the functional ANOVA decomposition

$$f(x) = \sum_{S \subseteq [d]} \tilde{f}_S(x_S) \tag{42}$$

where as before we have the purified functions as

$$\tilde{f}_S(x_S) = \sum_{T \subseteq S} (-1)^{|S|-|T|} \cdot (\mathcal{M}^{\text{cond}}_T \cdot f). \tag{43}$$

Note that because the original application (Sobol', 1990) to simulated experiments allowed researchers to make the assumption that all input variables are independent. Accordingly, using $\mathcal{M}^{\text{cond}}$ or $\mathcal{M}^{\text{marg}}$ is equivalent and modern extensions have considered both approaches, although $\mathcal{M}^{\text{cond}}$ is seemingly more popular.

As pointed out in (Herren and Hahn, 2022; Bordt and von Luxburg, 2023; Enouen and Liu, 2025), it can be seen fairly immediately that this allows for a correspondence between the SHAP functional applied to a machine learning model and the functional ANOVA decomposition, extending the form in Equation 23:

$$[\phi_i^{\text{SHAP}} \cdot f](x) = \sum_{S \ni i} \frac{\tilde{f}(x_S)}{|S|} \tag{44}$$

The key object of study for sensitivity analysis approach is the set of Sobol indices (Sobol', 1990) defined as:

$$\sigma_S := \mathbb{V}\text{ar}_{X_S}\left[\tilde{f}_S(X_S)\right] \tag{45}$$

Please note again that this definition is also worth scrutinizing because there are multiple equivalent formulations in the independent variable case and there is no obvious consensus on how to extend the original techniques.

The major property of the Sobol' indices is the ***decomposition of variance*** formula which states that:

$$\sigma^2 := \mathbb{V}\text{ar}_X\left[f(X)\right] = \sum_{S \subseteq [d]} \mathbb{V}\text{ar}_{X_S}\left[\tilde{f}_S(X_S)\right] =: \sum_{S \subseteq [d]} \sigma_S, \tag{46}$$

meaning that the total variance of the function f decomposes into a measure of how much variance there is across each of the different interactions S. Critically, this formula only holds in the case of independent input variables and completely breaks down for correlated variables.

There is additional the total-effect index, usually defined as:

$$\phi_S^{\text{Sobol-total}} := \mathbb{E}_{\bar{x}_{-S}}\left[\mathbb{V}\text{ar}_{X_S \sim p(X_S | X_{-S} = x_{-S})}\left[f_S(X_S, x_{-S})\right]\right] \tag{47}$$

which in the independent variable case is equal to $\phi_S^{\text{Sobol-total}} = \sum_{T\subseteq[d] \text{ s.t. } T\cap S\neq\emptyset} \sigma_T$. Because of the decomposition of variance formula, it is also common to consider the Sobol index as the percentage of the variance explained, as in ${\sigma'}_S := \sigma_S/\mathbb{V}[f]$. The same can be done for the total index, ${\phi'}_S^{\text{Sobol-total}} := \phi_S^{\text{Sobol-total}}/\mathbb{V}[f]$.

Other work had already made a connection between this global variance sensitivity and the Shapley value (Owen, 2014) by leveraging these known properties of the Sobol' indices. Writing

$$\phi_S^{\text{Sobol-lower}} = \sum_{T\subseteq S} \sigma_T \quad \text{and} \quad \phi_S^{\text{Sobol-upper}} = \sum_{T\subseteq[d] \text{ s.t. } T\cap S\neq\emptyset} \sigma_T, \tag{48}$$

as typical manipulations of the Sobol' indices, it was noted how applying the Shapley value to the value function defined as $v_S^{\text{Sobol}} = \phi_S^{\text{Sobol-lower}}$ results in the bound

$$\phi_{\{i\}}^{\text{Sobol-lower}} \leq \phi_i^{\text{Shapley}} \circ v^{\text{Sobol}} \leq \phi_{\{i\}}^{\text{Sobol-upper}}. \tag{49}$$

This is only true in the case of independent variables and is straightforward to see from Equation 23 and the fact that $\sigma_S \geq 0$.

It is worth noting how the Shapley of the variance from (Owen, 2014) is different from the variance of the Shapley as in the Shapley value for independent variables.

$$\phi_i^{\text{Shapley}} \circ v^{\text{Sobol}} = \sum_{S\ni i} \frac{\sigma_S}{|S|} \tag{50}$$

$$\mathbb{V}\Big[\phi_i^{\text{SHAP}} \circ f\Big] = \sum_{S\ni i} \frac{\sigma_S}{|S|^2} \tag{51}$$

From here, there is not much more which can be said about functional ANOVA without a further discussion on additive models. In the next section, we give a thorough discussion on additive models, again beginning with a historical overview. Hopefully, these results already begin to demonstrate how intimately tied these originally different research areas of feature interactions, Shapley values, and sensitivity analysis truly are. It is a major goal of this survey to clarify and to emphasize these connections which exist very deeply across these related areas. We will have another discussion on these functional ANOVA connections in Section 4.5.3 after the introduction of additive models.

4 Additive Interpretability (Additive Models)

4.1 A History of Generalized Additive Models

Let us once again turn to the very beginning of statistics with the advent of linear regression. One of the first clear publications on the topic is A.M. Legendre's 1805 work (Legendre, 1805) on the method of least squares for applications in astronomy (claimed also by Gauss (Gauss, 1809) for the same application). Not too long after, Francis Galton, beginning in 1875, applied the same technique to problems which are more statistical in nature, like predicting the height of a sweet pea plant from the height of its parent (Galton, 1894). Galton continued his studies onto human height and other heredity questions, seeming to incidentally give the name 'regression' around this time (Galton, 1886). After his death, Karl Pearson later picked up his mentor's work on sweet peas and genetics, ultimately synthesizing his works on linear regression and providing much greater mathematical rigor (e.g. Pearons' correlation coefficient, chi-square test, method of moments) (Pearson, 1914). Further developments in the beginning of the twentieth century like those of Ronald Fisher (Fisher, 1935) as well as those of Jerzy Neymann and Egon Pearson (Neyman and Pearson, 1933) finalized statistics as its own independent field of study.

Picking back up with the rapid developments in statistics near the end of the 20th century, the linear regression model had already seen ubiquitous use in statistics for decades at this point. A major development relevant for our discussion on the generalized additive model is the invention of the generalized linear model (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989).

$$\mathbb{E}[Y|X=x] \approx \hat{f}(x) = g^{-1}(\beta^T x) \tag{52}$$

Via the introduction of a link function, g, this framework unified the learning of a variety of distributions using a simplified linear model. Originally for the important logistic regression as well as Poisson regression and Gamma regression, and later to density estimation and other exponential families, this framework comprehensively unified many linear models.

Other critical precursors are the development of splines as a nonparametric approach (Schoenberg, 1964; Casteljau, 1986; Bézier, 1972; Birkhoff and De Boor, 1965; De Boor) and their subsequent application to scatterplot smoothing (Reinsch, 1967; Kimeldorf and Wahba, 1970, 1971; Wold, 1974; Wahba and Wold, 1975; Craven and Wahba, 1978; Golub et al., 1979; Rice and Rosenblatt, 1983; Silverman, 1984; Eubank, 1985)

Application of these one-dimensional techniques to multivariate data required usage of the more complex surface splines (Duchon, 1977) or the choice of a particular direction to linearly project the multidimensional data (Friedman and Tukey, 1974). Building on the idea of projection, Friedman and Stuetzle (1981) instead suggested the ability to project onto multiple directions and add up the total influences, leading to the additive modeling assumption. Additionally, the potential merits of projecting onto the coordinate directions were also suggested:

$$\hat{f}(x) = \sum_{i=1}^{d} f_d(x_d) \tag{53}$$

This work not only introduced the nonparametric additive model to the literature, but also introduced the backfitting algorithm for iteratively fitting the shape functions of the additive model (Friedman and Stuetzle, 1981).

This original idea was taken further by Stone (1985) to give some theoretical developments and Breiman and Friedman (1985) to develop the ACE algorithm. At the same time, other works doing semiparametrics were giving special attention to reducing the nonparametric estimation to as few components as possible (Green et al., 1985; Engle et al., 1986). It was around this time that the Generalized Additive Model (GAM) combining the nonparametric additive structure with the generalized link function allowing for regression, classification, etc. to be incorporated together. First combined as a 1984 technical report (Hastie and Tibshirani, 1984), then into a 1986 paper (Hastie and Tibshirani, 1986), and finally into a full book in 1990 (Hastie and Tibshirani, 1990), the generalized additive model proposed an extremely flexible nonparametric modeling structure which still overcame practical considerations like the curse of dimensionality.

By the time of the final book, the additive model had already been popularized and seen some attention from many other researchers (Stone, 1982; Friedman et al., 1983, 1984; Burman, 1985; Stone, 1986; Buja et al., 1989; Chen et al., 1989; Gu and Wahba, 1991b). Moreover, extensions to interaction models (mostly in the form of interaction splines) had been studied, chiefly by Wahba and her students (Barry, 1986; Wahba, 1986; Gu et al., 1989; Chen, 1987, 1991a). Indeed, just months later, a book focusing entirely on smoothing splines would be published by Wahba (Wahba, 1990).

Continued developments would improve on existing works like interaction models (Stone, 1985) and MARS (Friedman, 1991) to combine tensor product splines and thin plate splines (which were previously viewed as separate approaches) under the same framework called *Smoothing Spline ANOVA* (Gu and Wahba, 1991a, 1993b) as in Equation 54:

$$f(x) = C + \sum_{\alpha=1}^{d} f_\alpha(x_\alpha) + \sum_{\alpha<\beta} f_{\alpha\beta}(x_\alpha, x_\beta) + \sum_{\alpha<\beta<\gamma} f_{\alpha\beta\gamma}(x_\alpha, x_\beta, x_\gamma) + \dots \tag{54}$$

This nonparametric description already provided the precursors for the functional ANOVA perspective which we will see connects with the approaches from sensitivity analysis. Many refinements of this approach continued (Chen, 1991a; Gu and Wahba, 1991b, 1993a,b), ultimately culminating into Wahba's 1995 memorial lecture (Wahba et al., 1995), with later extensions to 'generalized' applications like density estimation and hazard estimation (Gu, 1995, 1996, 1998) ultimately culminating into Gu's 2002 book on SS-ANOVA (Gu, 2002). Although older works like this one make explicitly clear that higher-order interactions are excluded in practice, the ingredients for a functional ANOVA perspective connecting with sensitivity analysis can already be seen.

Another key era of interest in the GAM was its indirect attention in the wake of the success of boosting. At the time, it was very surprising that a set of boosted tree stumps could achieve good performance on classification tasks (Freund and Schapire, 1997; Schapire and Singer, 1998). Shortly after, it was realized that the choice of tree stumps as the weak learner automatically led to an additive modeling assumption (due to each stump's dependence on only a single input feature) (Friedman et al., 2000). Thus, a large subset of researchers were training additive models without even realizing that is what they were doing. Further developments on boosting emphasized this functional perspective, re-envisioning a round of boosting a weak learner as a gradient step in a functional optimization problem (Friedman, 2001, 2002).

Other work at the turn of the century focused on increasing the practical usefulness of generalized additive models, with many practitioners remaining weary of the many nuisance parameters which must be fit in a nonparametric method like smoothing spline ANOVA (spline knots and smoothing parameter) . Of great significance is the work by Wood making developments on the splines including better basis functions and efficient multiple smoothing (Wood, 2003, 2004, 2006b) which ultimately culminated into an R package (`mgcv`) and associated book (Wood, 2006a).

Although the history of the additive model does not end here, we take a break to discuss another important and emerging field, high-dimensional statistics, which ends up being of great importance to additive models. We will then coming back to discuss recent additive models after their modern revitalization due to their nice interpretability properties.

4.2 Sparse Additive Models

High-dimensional statistics is an area of statistics focused on describing the statistical behavior when classical asymptotic results break down. Although classical results, like the central limit theorem and likelihood ratios, take the dimension d to be fixed and take the samples n to infinity, in many empirical situations these asymptotic predictions turn out to be completely inadequate. This is where high-dimensional statistics comes in, ultimately evolving to describe the regime when $n \ll d$ or more broadly when n is somehow comparable to d (Wainwright, 2019).

Early works identified hints of these potential problems while working with large-scale datasets, like Rao (1949) and Deev (1970). Rigorous progress was first made in the field of random matrix theory, which was forced to contend with $n \approx d$ in the calculation of eigenvalues of infinite random matrices (Wigner, 1955, 1958; Marčenko and Pastur, 1967; Pastur, 1972). Soon after, other works brought the same ideas directly into statistical applications like robust regression, with Huber (1973) introducing the 'high-dimensional limit' which takes $c := \frac{d}{n}$ constant before an asymptotic approach is employed.

In later years, the topic only grew in importance in response to factors like the increasing availability of high-dimensional data (Breiman, 1995; Tibshirani, 1996) and the increasing complexity of recovery in signal processing applications (Mallat and Zhang, 1993; Pati et al., 1993). A quintessential example is the LASSO (Tibshirani, 1996) which uses L1 regularization to shrink the coefficients towards sparse estimates via control of the λ_1 hyperparameter:

$$\begin{aligned}\hat{\beta}^{\text{LASSO}} &:= \underset{\beta_\emptyset, \beta}{\operatorname{argmin}} \left\{ \left\| Y - (\beta_\emptyset + X\beta) \right\|_2^2 + \lambda_1 \|\beta\|_1 \right\} &(55)\\ &= \underset{\beta_\emptyset, \beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y^{(i)} - \beta_\emptyset - \sum_{j=1}^{d} \beta_j x_j^{(i)} \right)^2 + \lambda_1 \cdot \sum_{j=1}^{d} |\beta_j| \right\} &(56)\end{aligned}$$

Developments of these original approaches like matching pursuit (Mallat and Zhang, 1993) and orthogonal matching pursuit (Pati et al., 1993) would continue to be enhanced with basis pursuit (Chen et al., 2001) and sparse dictionaries (Donoho and Huo, 2001; Elad and Bruckstein, 2002). Further work on the linear problem of the garrote (Breiman, 1995) or lasso (Tibshirani, 1996) would be taken by LARS (Efron et al., 2004), ElasticNet (Zou and Hastie, 2005), and the Dantzig selector (Candes and Tao, 2007). These works would continue to develop the practical usefulness and emprical understanding of these shrinkage-based sparsity approaches. Further works would continue to develop a better theoretical understanding of when these approaches succeed in signal recovery (Candes and Tao, 2005; Donoho, 2006b; Candes and Tao, 2007; Bickel et al., 2009).

$$\hat{\beta}^{\text{ElasticNet}} := \underset{\beta_\emptyset, \beta}{\operatorname{argmin}} \left\{ \left\| Y - (\beta_\emptyset + X\beta) \right\|_2^2 + \lambda_2 \|\beta\|_2 + \lambda_1 \|\beta\|_1 \right\} \tag{57}$$

$$\hat{\beta}^{\text{Dantzig}} := \underset{\beta_\emptyset, \beta}{\operatorname{argmin}} \left\{ \|\beta\|_1 \qquad \text{s.t.} \qquad \|X^T(y - \beta_\emptyset - X\beta)\|_\infty \leq C\sigma\sqrt{\log d} \right\} \tag{58}$$

Some of the earliest methods transitioning these insights into the more nonparametric approaches of kernel-based learning were SUPANOVA (Gunn and Brown, 1999; Gunn and Kandola, 2002) and likelihood basis pursuit (LBP) (Zhang et al., 2004).

$$\hat{\beta}^{\text{SUPANOVA}} := \underset{\beta_\emptyset, \beta}{\operatorname{argmin}} \left\{ \left\| Y - (\beta_\emptyset + \Phi\beta) \right\|_2^2 + \lambda_1 \|\beta\|_1 \right\} \tag{59}$$

$$\hat{\beta}^{\text{LBP}} := \underset{\beta_\emptyset, \beta}{\operatorname{argmin}} \left\{ \left\| Y - (\beta_\emptyset + \Phi\beta) \right\|_2^2 + \lambda_1 \|\beta\|_1 \right\} \tag{60}$$

Although SUPANOVA and LBP focused on the K-ANOVA and SS-ANOVA respectively, they remained treating basis function coefficients as individual terms to be sparse regularized, requiring multiple stage procedures for ensuring additional sparsity of the ANOVA components. As the ubiquity of high-dimensional principles became clear throughout the 2000s, it was the COSSO model (Lin and Zhang, 2006) and SpAM model (Ravikumar et al., 2009) which reduced to simpler GAM-1 models via a restriction of the functional hypothesis space, but treated a more general nonparametric function with the use of functional norms. These extensions of sparse selection to the additive model would continue with other 'SPAM-class' models: HDAM (Meier et al., 2009), a version for multiple kernel learning (Koltchinskii and Yuan, 2010) and a version for minimax optimal rates (Raskutti et al., 2012).

$$\hat{f}^{\text{COSSO}} := \underset{f_\emptyset \in \mathbb{R}, f_i \in \mathcal{H}_i}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y^{(i)} - f_\emptyset - \sum_{j=1}^{d} f_j(x_j^{(i)}) \right)^2 + \lambda_{1,2} \cdot \sum_{j=1}^{d} \|f_j\|_{\text{Sobolev}} \right\} \tag{61}$$
$$\text{where} \quad \|f_j\|_{\text{Sobolev}}^2 := \sum_{\nu=0}^{1} \left(\mathbb{E}\Big[D^\nu f_j(X_j) \Big] \right)^2 + \mathbb{E}\Big[|D^2 f_j(X_j)|^2 \Big]$$

$$\hat{f}^{\text{SPAM}} := \underset{f_\emptyset \in \mathbb{R}, f_i \in \mathcal{H}_i}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y^{(i)} - f_\emptyset - \sum_{j=1}^{d} f_j(x_j^{(i)}) \right)^2 + \lambda_{1,2} \cdot \sum_{j=1}^{d} \|f_j\| \right\} \tag{62}$$
$$\text{where} \quad \|f_j\|^2 := \mathbb{E}\Big[|f_j(X_j)|^2 \Big]$$

Around this time, a major innovation in how to include interaction effects in linear models would occur with the hierarchical lasso (Bien et al., 2013). The reintroduction of heredity would later become a key development in the study of feature interactions, which remains a central focus for modern approaches to the additive model.

4.3 Modern Additive Models

Although additive models had remained a useful statistical model in the years that followed, their representative power would slowly fade as their main selling point. Indeed, more powerful models like boosted forests (Schapire, 1990; Freund and Schapire, 1997; Breiman, 1998) and eventually deep neural networks (Mohamed et al., 2011; Krizhevsky et al., 2012; Bahdanau et al., 2015) proved themselves as incredibly useful modeling tools across a wide variety of problems (Breiman, 2001b). Accordingly, modern interest in additive models has been due to reasons contrary to their original motivation: they are now chosen mainly due to their *simplicity* compared with the blackbox machine learning and deep learning approaches which emerged in the past several decades.

Interest in GAMs was once again revitalized when two of its key properties were emphasized: (i) interpretability of the additive influence coming from each input feature; and (ii) near-competitiveness with modern state-of-the-art methods like random forests and boosting (Lou et al., 2012, 2013). The good performance of additive models across many real-world datasets alongside the importance of interpretability for certain tasks like medical applications led to a rekindled interest in the study of GAM models as a solution to the blackbox problem (Caruana et al., 2015).

Conversations in statistics about the diverging needs of statistical modeling for explanation or statistical modeling for prediction (Breiman, 2001b; Shmueli, 2010) has already begun by the turn of the century, with methods like Partial Dependence Plots (PDPs) already providing the first 'explanations' of blackbox boosting models (Friedman, 2001). Nevertheless, the field of interpretability, specifically dedicated to fighting the blackbox model problem, seems to have only accelerated in the wake of deep learning's widespread success across previously untouched domains. As already discussed in Sections 2 and 3, this led to a flurry of research, beginning in the mid 2010s and continuing to the present day, which is focused on *explaining* the complex blackbox models which were being applied across an increasingly wide set of tasks.

This ultimately led to an extremely wide set of approaches used for explaining blackbox behavior. It was only as a result of Rudin (2019) putting their foot down and demanding that researchers return to fully interpretable models which faithfully represent their decisions, rather than using explanations which can only approximate the behavior of the blackbox model. This is because this work pointed out the fundamental gap existing between explained approximations and the true blackbox model, for example being a poor explanation around 10% of the time. Those same gaps happening 10% of the time could be exactly the critical points where the model behavior needed to be understood in the first place.

Follow-up works have pushed this duality even further, finding explicit duals between XAI explanations and IML interpretations (Enouen and Liu, 2025; Günther et al., 2025). One of the major dualities discussed herein is the duality between SHAP and GAM. As a consequence of (Rudin, 2019), there was only a greater desire within the interpretability community to develop models which are interpretable 'from the ground up', leading to increased interest in additive models as a tool of choice for accurately modeling complex data while also remaining widely interpretable. As it currently stands, additive models remain the most well-studied class of interpretable models and seem to be the furthest along at matching SOTA performance (in tabular data), although other methods are following close behind.

After the revitalization of interest in GAMs due to the EBM strand of work (Lou et al., 2012, 2013; Caruana et al., 2015), a next major development occurs with the work of SALSA (Kandasamy and Yu, 2016). This work was able to fit k-th order additive models via the use of a specific kernel which allows a computational trick for calculating the value of the kernel in $\mathcal{O}(k^2d)$ time instead of the naive $\mathcal{O}(d^k)$ time. Kandasamy and Yu (2016) finds that this can achieve the best performance on some datasets by balancing the bias-variance tradeoff, having less bias than 1D additive models and less variance than fully nonparametric models. Unfortunately, the method cannot scale past several thousand samples by nature of being a kernel machine approach (requiring $\mathcal{O}(n^2k^2d)$ time to compute the kernel matrix and $\mathcal{O}(n^3)$ time to compute the matrix inverse). SpAM-2 (Tyagi et al., 2016) extended the sparse gradient approach of SpAM to fit the functional forms, using discretized variable domains instead of kernel machines or kernel smoothing.

Neural Interaction Transparency (NIT) (Tsang et al., 2018b) uses a special neural architecture which uses a first layer for enforcing sparsity and disentangling the interactions. The first layer is restricted to a maximal number of nonzero entries per feature node to be k, limiting the maximal degree of feature interactions which are possible to represent. Another early NAM approach which did not intentionally leverage a GAM structure is BagNet (Brendel and Bethge, 2019). Applied to computer vision classification, this approach took the final logits to be a sum of the local logits computed based on a small patch of the image, showing that large enough patches from 10x10 to 30x30 achieved surprisingly competitive performance. Sparse Shrunk Additive Model (SSAM) (Liu et al., 2020b), amongst peers like COSSO and SALSA, focuses on the case of $k = 2$ and tries to incorporate additional sparse selection into SALSA, on both the samples and the features.

The paper taking after the namesake, Neural Additive Models (NAM) (Agarwal et al., 2021), simply replaces the nonparametric method with a neural network, but emphasizes the many interpretability benefits of this approach (despite NAMs technically being a neural network). A major change to the neural network is the introduction of the ExU activation function to allow the network to capture the same sharp edges which are possible using tree ensembles. Unfortunately, this also has the effect of greatly destabilizing the neural network training, meaning that an ensemble of around 100 NAM networks is required to achieve good performance. Around the same time, another approach to neural additive models called GAMI-Net (Generalized Additive Model with Interactions Network) (Yang et al., 2021) included pairwise interactions within the neural network structure. Although this had the additional complexity of a stagewise, heredity-based selection procedure to choose relevant pairs, it maintained interpretability by not pushing beyond into higher-order interactions.

Shortly after, a lot of work picked up in this direction. NODE-GAM and NODE-GA2M (Chang et al., 2022) utilize the differentiable tree called NODE (Popov et al., 2019) as the nonparametric approach. This is combined with a clever 'color gating' mechanism to gradually ensure the feature sparsity of each component (which then ensures the overall GAM structure is obeyed), following a similar intuition to the gating used in NIT (Tsang et al., 2018b). Sparse NAM (SNAM) (Xu et al., 2023) attempts to learn a sparse set of 1D shape functions by enforcing a group parameter penalty on each of the subnetworks of a neural additive model. Higher-Order NAM (HONAM) (Kim et al., 2022) uses a very similar trick to SALSA in order to compute symmetric polynomials of the embeddings, using neural network embeddings instead of kernel embeddings.

The Neural Basis Model (NBM) (Radenovic et al., 2022) is a specific type of NAM which uses a bank of shared basis functions which are learned across all features, allowing for repeated shape functions to be learned more easily, while still respecting the additive GAM structure. This was done in close development with Scalable Polynomial Additive Model (ScalPol-AM) (Dubey et al., 2022) which instead generalizes the opposite direction, using the most simple basis function (polynomials) but adding complex higher-order interactions to the model. Instead of the existing kernel ANOVA trick from SALSA, they instead write the polynomial of degree k with weight tensor $W^{(k)} \in \mathbb{R}^{d^{\otimes k}}$ as a low-rank approximation $W^{(k)} \cdot x^{\otimes k} \approx \sum_{i=1}^{r_k} \lambda_i^{(k)} \langle u_i^{(k)}, x \rangle^k$. This polynomial also has the advantage of being easily computable and they further combine this idea with using neural basis functions to compute the input features to the polynomial approach, similar to HONAM. Sparse Interaction Additive Networks (SIAN) (Enouen and Liu, 2022) also consider higher-order neural additive models, but instead focus on the problem of sparse selection of these interactions. This requires a two-stage procedure where interactions are first chosen based on important interactions according to an MLP and then a NAM is fit according to the chosen interactions.

Regionally Additive Models (RAMs) (Gkolemis et al., 2023) combine the GAM structure with the nearest-neighbor simplicity bias, subdividing into patches on which the function structure obeys the simple GAM structure. Generalized Sparse Learning of Additive Models with Interactions (G-SLAMIN) G-SLAMIN (Ibrahim et al., 2023) returns to two-dimensional, tree-based ensembles while incorporating explicit masking variables obeying weak or strong heredity alongside a hard cutoff threshold as chosen by hyperparameters. AHOFM (Ruegamer, 2024) uses the factorization machine approach discussed further in the subsection below as applied to tensor product splines to limit the consequences of higher-order effects in large dimensions.

Optimized Regularized Stump Forests (ORSF) (Gabidolla and Carreira-Perpiñán, 2025) reverts even further to one-dimensional, tree-based stump ensembles. They show that directly optimizing each stump alongside careful balancing of regularization hyperparameters can outperform existing GAM-1 approaches. PatternGAM (Clark et al., 2025) corrects for the redundancy of correlated input features by adjusting the Pattern method (Haufe et al., 2014) used to correct for collinearity in linear regression to instead correct the concurvity between shape functions in additive models. InstaSHAP (Enouen and Liu, 2025) instead uses a masking-based framework to automatically correct for this redundancy in the additive model shape functions. Tensor-Product Neural Network (TPNN) (Park et al., 2025) combines the tensor-product spline approach with the neural basis approach as in HONAM and ScalPol-AM, while additionally requiring the Hooker purification condition ((Hooker, 2007) to be discussed in Section 4.5.3).

Table 1: Tabulation of many GAM methods over the years

Method	Degree			Generalized			Sparsity Type	Nonparametric Model	Year
	1D	2D	3D+	reg	cls	other			
GAM	✓			✓	✓	✓	✗	kernel smoothing	1990
SS-ANOVA	✓	✓		✓			✗	spline smoothing	1993
Adaboost	✓				✓		✗	stump ensembles	1997
SS-ANOVA	✓	✓		✓	✓		✗	spline smoothing	2002
SUPANOVA	✓	✓		✓			coefficient-wise	SVM	2002
LBP	✓	✓		✓	✓		coefficient-wise	RKHS	2004
COSSO	✓			✓	✓		1D (Sobolev)	RKHS	2006
SpAM	✓			✓	✓		1D (functional)	kernel smoothing	2008
HDAM	✓			✓	✓		1D (sparse-smooth)	RKHS	2009
GAM	✓			✓	✓		✗	tree ensembles	2012
GA^{2}M	✓	✓		✓	✓		pairs, stagewise	tree ensembles	2013
EBM	✓	✓		✓	✓		pairs, stagewise	tree ensembles	2015
SALSA	✓	✓	✓	✓			k-D (functional)	kernel machines	2016
SpAM2	✓	✓		✓			pair selection	discretized	2016
NIT	✓	✓	✓	✓	✓		HO, online (neural-based)	neural	2018
BagNet	*✓				✓		✗	neural	2019
SSAM	✓	✓		✓			coefficient-wise	kernel machines	2020
GAMI-Net	✓	✓		✓	✓		pairs, stagewise	neural	2021
NAM	✓			✓	✓		✗	neural	2021
NODE-GAM	✓	✓		✓	✓		pairs, online (tree-based)	differentiable trees	2022
SNAM	✓			✓	✓		pairs, online (group norm)	neural	2022
HONAM	✓	✓	✓	✓	✓		FM-style	neural bases	2022
ScalPol-AM	✓	✓	✓	✓	✓		FM-style	polynomials	2022
NBM	✓			✓	✓		✗	neural bases	2022
SIAN	✓	✓	✓	✓	✓		higher-order selection	neural	2022
G-SLAMMIN	✓	✓		✓	✓		pairs, online (indicators)	differentiable trees	2023
AHOFM	✓	✓	✓	✓			FM-style	spline bases	2024
ORSF	✓			✓	✓		✗	stump ensembles	2025
TPNN	✓	✓	✓	✓	✓		FM-style	neural bases	2025
FM	✓	✓					FM-style	factorization machine	2010
HIFM	✓	✓					FM-style	linear factorization	2014
HOFM	✓	✓	✓				FM-style	factorization machine	2016
OptFeature	✓	✓	✓				GAM- and FM-style	factorization machine	2023
GAM-GP	✓					✓	✗	gaussian process	2013
KANOVA GP	✓	✓	✓			✓	kernel cross terms	gaussian process	2016
OAK	✓	✓				✓	✗	gaussian process	2022
NAM-LSS						✓	✗	parametrized distribution	2024

In addition to these works on the 'main' line of research into additive models, there are other key directions of research which are closely related to additive models. First is the research into uncertainty using GAMs. Although GAMs are typically shown alongside their bagged intervals giving some notion of a confidence interval, much more serious notions of uncertainty exist using Gaussian processes or Bayesian approaches. Additionally, there is the large amount of research into Factorization Machines (FMs) which are very closely related to additive model structural assumptions. The biggest difference can be seen in the implicit assumption in the FM domain that input features are themselves high-dimensional (rather than the number of input features as in high-dimensional GAMs).

4.3.1 GAM Uncertainty

Many different approaches have incorporated GAM structure into uncertainty estimation or added uncertainty to GAM predictions. The Generalized Additive Models for Location, Scale, and Shape (GAM-LSS) (Rigby and Stasinopoulos, 2005) is an approach for distribution modeling which makes great use of the 'generalized' aspected of generalized additive models, modeling a wide variety of parametrized distributions from normal and logistic to Box-Cox and Poisson. This approach allows for conditional uncertainty to be directly modeled via the modeling of the target distribution.

Additive Gaussian Processes (Duvenaud et al., 2011), or the nearly equivalent Additive Kriging (Durrande et al., 2011), followed up on previous works in hierarchical kernel learning (Bach, 2009) and Gaussian process Sobol indices (Marrel et al., 2009), respectively, to deliver the Gaussian process obeying the additive assumption of the GAM. Further works looking at using kernel-based functional ANOVA extensions to additive Gaussian processes and general Gaussian processes (Ginsbourger, 2013; Duvenaud, 2014) as well as specially designed Kernel ANOVA (Ginsbourger et al., 2016) which can further decompose into 4^d terms by nature of each $2^d \cdot 2^d$ cross-terms contributing to the kernel.

This work would pick up later with a modernization of OAK (Lu et al., 2022) which combined modern sparse GP approaches with the additive GP assumption; (Luo et al., 2022) would develop a similar extension around the same time. Max-Mod (López-Lopera et al., 2022) would also revisit the additive GP assumption, also providing support for monotonicity constraints.

Other approaches to uncertainty would lean more heavily on Bayesian idealogy, with SKIM-FA (Agrawal and Broderick, 2023) extending their linear kernel interaction trick (Agrawal et al., 2019) which applies an additive assumption directly on the kernel alongside hierarchical Bayesian modeling. NAM-LSS (Frederik Thielmann et al., 2024) would extend NAM approaches to the location, scale, and shape modeling of GAM-LSS. LA-NAM (Bouchiat et al., 2024) learned an independent Bayesian NN for each of the additive components of the NAM.

4.3.2 Factorization Machines

Another important model closely related to the additive model and another type of commonly studied interaction selection is the Factorization Machine (FM) (Rendle, 2010). This approach extends existing Matrix Factorization (MF) (Bell et al., 2008) approaches which are used under cases of extreme sparsity like in the Netflix problem of matching users to movies (Bennett and Lanning, 2007; Bell et al., 2010). In the original FM, first a multilinear assumption and a GAM-2 assumption are made resulting in the Equation 63. The parameters are the $w_\emptyset \in \mathbb{R}$ scalar, the $\vec{w} \in \mathbb{R}^d$ vector, and the $W \in \mathbb{R}^{d\times d}$ matrix. For large d, there may be insufficient data to adequately fit the quadratic $\mathcal{O}(d^2)$ number of coefficients in W. Accordingly, it is fit using a low-rank approximation $W \approx VV^T$ for some $V \in \mathbb{R}^{d\times r}$. This results in the FM equation in Equation 64 after the appropriate adjustments for symmetrizing W and removing the quadratic terms x_i^2.

$$f^{\text{bilinear}}(x) = w_\emptyset + \sum_{i=1}^{d} w_i x_i + \sum_{i=1}^{d}\sum_{j=1}^{d} W_{i,j} x_i x_j \tag{63}$$

$$f^{\mathrm{FM}}(x) = w_\emptyset + \sum_{i=1}^{d} w_i x_i + \sum_{i=1}^{d-1} \sum_{j=i+1}^{d} \langle \vec{v}_i, \vec{v}_j \rangle x_i x_j \tag{64}$$

where $\vec{v}_i \in \mathbb{R}^r$ for each $i \in [d]$ represent the rows of V.

Later work like FHIM (Purushotham et al., 2014) add additional sparsity on the vectors $\vec{w}$ and $\vec{v}_i$ to compensate for high-dimensional data. They additional suggest the possibility of generalizing to higher-order feature interactions. In 2016, HOFM (Blondel et al., 2016) actually explores the higher-order factorization machine, decomposing the higher-degree tensors like $W^{(3)} \in \mathbb{R}^{d\times d\times d}$ using the CP decomposition (Chang, 1970; Harshman, 1970) of these tensors $W^{(3)}_{i,j,k} \approx \sum_{r=1}^{r_3} \vec{v}^{(3)}_{i,r} \cdot \vec{v}^{(3)}_{j,r} \cdot \vec{v}^{(3)}_{k,r}$ as seen in Equation 66.

$$f^{\mathrm{multilinear}}(x) = w_\emptyset + W^{(1)} \odot \vec{x} + W^{(2)} \odot (\vec{x} \otimes \vec{x}) + \cdots + W^{(k)} \odot \vec{x}^{\otimes k} \tag{65}$$

$$\begin{aligned} f^{\mathrm{HOFM}}(x) = w_\emptyset + \sum_{i=1}^{d} w_i x_i + \sum_{i=1}^{d-1} \sum_{j=i+1}^{d} \langle \vec{v}^{(2)}_i, \vec{v}^{(2)}_j \rangle \cdot x_i x_j + \\ \cdots + \sum_{i_1 < \cdots < i_k} \langle \vec{v}^{(k)}_{i_1}, \ldots, \vec{v}^{(k)}_{i_k} \rangle \cdot x_{i_1} \ldots x_{i_k} \end{aligned} \tag{66}$$

where we abuse the Hadamard product notation ($\odot$) to also mean collapsing to a scalar and the notation $\langle \vec{v}_1, \ldots, \vec{v}_k \rangle$ is the extended dot product agreeing with the above low-rank (CP) formula. Note again that the multilinear assumption is notationally tedious to write, either requiring that we write the explicit sum over strictly increasing indices or requiring the awkward condition that each $W^{(k)}$ tensor only has nonzero entries for completely asymmetric entries. Accordingly, we write Equation 65 as the easier to write polynomial in tensor products with the understanding that these weight tensors may need to be further restricted as appropriate.

Around this time, two works claiming to be Sparse Factorization Machines (SFMs) (Xu et al., 2016) and (Pan et al., 2016), add additional sparsity to the bilinear factorization machine to better deal with the high-dimensional setting. (Pan et al., 2016) follows a similar approach to (Purushotham et al., 2014), focusing on entry-wise sparsity of the parameters. They replace the sparse Gaussian approach from (Purushotham et al., 2014) with a Laplacian distribution and make the many necessary adjustments in methodology. (Xu et al., 2016) instead aims for sparsity on the individual feature pairs, regularizing the rows of the V_1 and V_2 matrices of $f^{\mathrm{SFM}}(x) = x^T V_1 V_2^T x$ with the $\|\cdot\|_{2,1}$ norm on V_1 and V_2. This norm encourages groupwise sparsity on each row of the decomposition matrix, allowing unnecessary features to be easily dropped out of the FM interaction representation.

It is additionally worth noting how closely related the factorization machine approach is to several higher-order attempts at extending GAMs. HONAM (Kim et al., 2022), ScalPol-AM (Dubey et al., 2022), and AHOFM (Ruegamer, 2024) all use nonlinear basis functions outside of the multilinear approach to be able to extend the representation capabilities of FMs for continuous variables (either neural bases or spline bases). It is worth noting that this is often one of the key distinguishing factors between these nearby but parallel research directions. Due to the fact that FMs are usually applied to high-dimensional, categorial variables with one-hot encodings, the assumption of bilinearity is not actually a restriction. In the case of continuous variables, this is no longer the case. Another key difference is the mindset: factorization machines are making this simplifying assumption because it achieves the best performance on the challenging high-dimensional task; additive models are making this simplifying assumption because it provides clear insight into the data.

Accordingly, many follow-up works to FM and HOFM are only focused on achieving the best performance, no matter the benefits or detriments to interpretability. We include these works regardless for completeness. Meta-graph FMG (Zhao et al., 2017) appends additional knowledge graph features before applying a regularized FM structure. IAFM (Hong et al., 2019) modifies the FM with a courser field-level field aspect as well as an attention mechanism, and this work also consider neural network extensions for learning higher-order interactions. AutoInt (Song et al., 2019) focuses specifically on an interacting attention structure to implicitly learn higher-order factorization machine structure. AutoFIS (Liu et al., 2020a) uses a higher-order interaction structure of HOFM alongside the neural network tricks of batch normalization and gating. GILDER 2020 (Tsang et al., 2020a) first runs a neural interaction detection procedure before retraining on an expanded set of cross features. TI and CS regularized FMs (Atarashi et al., 2021) introduce alternate sparse regularization approaches to better enable groupwise sparsity and interaction selection. AutoAIS (Wei et al., 2021) uses a meta architecture search across deep architectures for the embedding and interaction structure. AdaFS (Lin et al., 2022) uses a small MLP alongside mixed soft and hard selection to implicitly discover all higher-order interactions.

A major departure from this push towards neural factorization machines is the work on hybrid-grained interaction selection (Lyu et al., 2023). In this work, they make a clear distinction between the coarse-grained, field-level interactions and the fine-grained, value-level interactions being modeled by a factorization machine. In this language, the interaction selection of additive models corresponds to selection at the coarse-grained level. They combine both levels of interaction selection into a model they call OptFeature. Their work also considers ablating both level of interaction selection as well as the extension to 3rd degree interactions. Overall, this fine-grained perspective allows for an easier understanding of the key difference between GAM and FM interactions.

4.4 Feature Interaction Selection

Now that we have gone over the many different types of feature interaction detection and additive models, we dedicate this section to clarify the different types of feature interaction selection. The first versions of feature interaction selection were only referring to interactions in the sense of the low-order pairwise interactions. Another popular version of interaction selection is in the sense of factorization machines which use matrix factorizations to represent interaction terms. The more general higher-order interaction selection is what we will refer to as feature interaction selection, operating at the coarse-grained level of the additive model.

4.4.1 Pairwise Interaction Selection

First attempts at interaction selection were selection amongst singles and pairs of features $\mathcal{I} \subseteq \mathcal{I}^{\leq 2} := \{S : S \subseteq [d], |S| \leq 2\}$. This was originally done for the selection of linear and bilinear coefficients in a simple bilinear model as in Equation 67. This was done by early works like efficient heredity (Yuan et al., 2007, 2009), CAP (Zhao et al., 2009), SHIM (Choi et al., 2010), and hierarchical lasso (Bien et al., 2013).

$$f(x) = \beta_\emptyset + \sum_{i \in \mathcal{I}_1} \beta_i x_i + \sum_{(i,j) \in \mathcal{I}_2} \beta_{ij} x_i x_j \tag{67}$$

This setting later continued to be expanded on with even further guarantees in ultra-high-dimensional settings using the iFOR algorithm (Hao and Zhang, 2014b) and the SIRI algorithm (Kong et al., 2017).

4.4.2 Interaction Order Selection

Another important special case is degree selection or order selection. This selects the highest order or the highest degree of the GAM model. That is, choosing the optimal k amongst $k = 1, 2, \dots, d$ so that the GAM-k model will be the best performing.

$$f_{\leq k}(x) = \sum_{|S| \leq k} f_S(x_S) \tag{68}$$

Historically speaking, it was quite popular for people to automatically do this by selecting a model of degree 1 or 2 which corresponded to the best nonparametric model of those times. Works would later codify the statistical intuition that additive models break the curse of dimensionality by providing the nonparametric convergence rates for these models (Stone, 1985; Andrews and Whang, 1990; Chen, 1991b). These works came from work in the larger areas of series estimators and sieve methods (Newey, 1997; Chen, 2007) which focus on giving statistically valid but sufficiently flexible nonparametric estimators.

4.4.3 Fine-Grained Interaction Selection

Another important type of interaction selection is the selection of 'fine-grained' interactions either through the use of a factorization machine (Rendle, 2010; Blondel et al., 2016) or matrix factorization (Bell et al., 2008). This applies to problems where the individual variables dimensions x_i may themselves be high-dimensional (usually discrete variables with many possibilities like user ID). Recall that an FM will model an interaction between i and j on onehot features $x_i \in \mathbb{R}^{d_i}$ and $x_j \in \mathbb{R}^{d_j}$ via $V_i \in \mathbb{R}^{r \times d_i}$ and $V_j \in \mathbb{R}^{r \times d_j}$:

$$f^{\text{FM}}(x_i, x_j) = (V_i^T V_j)\rangle \cdot (x_i \otimes x_j) \tag{69}$$

which is different from fitting an arbitrary nonparametric function on the onehot features:

$$f(x_i, x_j) = W \cdot (x_i \otimes x_j) \tag{70}$$

Although linearized matrix factorizations are the only method which is widely used for fine-grained feature interaction selection, it can be imagined that other techniques for this problem could also be approached. The importance of handling both the coarse-grained interaction selection as described in the next section and the fine-grained interaction selection as described in this section has also been explored before in a hybrid approach called OptFeature (Lyu et al., 2023).

4.4.4 Feature Interaction Selection (Higher-Order)

Finally, we discuss feature interaction selection in its full generality as a problem of selecting from all possible higher-order interactions, choosing an entire collection $\mathcal{I} \subseteq \mathcal{P}([d])$ from all 2^d possible feature subsets (Sugiyama and Borgwardt, 2019; Enouen and Liu, 2022).

$$f_{\mathcal{I}}(x) = \sum_{S \in \mathcal{I}} f_S(x_S) \tag{71}$$

Compared with the other versions of interaction selection, this higher-order feature interaction selection results in a doubly-exponential combinatorial problem, choosing from an extremely rich but numerous set of candidate $\mathcal{I}$ collections.

Unlike high-dimensional statistics, which is mainly applied to specific domains like biostatistics (where $d \gg n$ often holds), feature interaction selection is more widely applicable across machine learning tasks (since $2^d \gg n$ more commonly holds), especially in the ever-increasing presence of high-dimensional data. It is this property which we will refer to as medium-dimensional statistics or *medium dimensionality*. Even for relatively small dimension such as $d = 20$ or $d = 30$, one will quickly face statistical limitations due to the fact that $2^d = 1.0\text{e}6$ or $2^d = 1.1\text{e}9$ is often much larger than n, the sample size.

4.5 Modern Topics

Given the long history and many varieties of additive models we discussed throughout this section, we now close with a focus on some of the topics which have attracted attention in the modern iteration of study. Of course additive models are first and foremost a predictive tool and thus any improvement on the set of methods available for quickly training accurate GAM models is always of continued interest. In this subsection we instead go into greater detail on the more nuanced research questions about generalized additive models, especially with respect to their position as an interpretable machine learning model.

Of great importance are questions relating to how interpretable and how robust the insights learned by additive models are. Furthermore questions about how to incorporate domain knowledge, fine-grained structure, geometric symmetries, and other problem-specific structure into the assumptions of the GAM model remain important in domain application. Questions about the underlying correlation structure of the input variables drive many of these other questions. Finally, extensions which combine additive principles with other principles like the logical pillar and the concept pillar look like promising directions for balancing interpretability with good performance.

4.5.1 Interpretability of GAMs

A major question in the usage of GAMs is 'How interpretable are GAMs, really?' Although an individual GAM is fundamentally interpretable because we can inspect its shape function, this ignores extraneous considerations to how we interpret those shape functions. Questions like **robustness** (would we get the same shape functions with slightly different data or models?), **sparsity** (are there few enough shape functions to reasonably look at all of them?), **correlations** (are certain variables and shape functions carrying redundant information about the target?), and **causality** (when can we interpret the shape functions as saying something causal about the relationship?) are critical for contextualizing the interpretability of GAM models.

Robustness The robustness of the GAM-provided explanation of the data is directly dependent on the robustness of the GAM model itself. (Chang et al., 2021) explores in detail how even using different nonparametric methods to fit the shape functions can lead to significantly different results. This is further shown to have direct implications for algorithm fairness through varying importance and polarity of sensitive attributes. (Enouen and Liu, 2025) has shown how even a fixed model class (neural additive models) can have very wide variance under typical GAM training procedures, mainly as a consequence of correlated features. (Schulte and Rügamer, 2025) shows that the training of boosted additive models follow a implicitly regularized gradient descent path, mainly leading to shrinkage of the estimated effect.

Sparse Explanations Another key point in interpreting shape functions is the requirement of actually looking at all shape function plots. Although manageable but tedious for 10-100 shape functions, this problem can become out of hand for too many shape functions (especially when utilizing interactions). COGAM (Abdul et al., 2020) proposes to repurposes sparse additive models specifically for this purpose of easier interpretation, further utilizing linear coefficients in the cases where shape functions are close to linear. Of course, requiring increasingly sparse GAM models to allow for interpretability is in conflict with the goal of predictive accuracy, and thus must be balanced for the targeted application. (Zhong et al., 2023) takes the alternate approach of finding the 'Rashomon set' of sparse additive models, meaning the collection of nearly-optimal-performing sparse models. Although this catches the many alternate feature subsets which are deemed important by sparse additive models, it once again significantly raises the cognitive load. Moreover, these sparse effects may only be conflations between correlations between equally predictive features.

Correlated Effects There is a major weakness in GAM interpretations when applied to distributions with heavily correlated features. When considering a dataset which has two copies of the same feature, the GAM model can divide the shape function between the two copies in completely arbitrary ways. The robust solution would give half of the weight to each whereas a sparse solution would give all of the weight to only one. This issue is generally referred to as concurvity in GAMs (Ramsay et al., 2003; Siems et al., 2023; Zhang et al., 2025). This fundamental conflict between robustness and sparsity in the correlated setting poses many practical challenges for using GAM models. This is only further complicated by the common misinterpretation that GAM shape functions are indicative of causal relationships.

Causal Interpretations There is a well-known bias to interpret shape function plots as describing the causal influence which a feature has on the target, rather than the predictive influence which the feature has on the target. This is often natural in many tasks where the outcome does causally depend on the input features; however, it can easily lead to misinterpretations of the GAM model. For instance, it could be the case that most of the shape function is truly causal; however, a portion of the shape function actually is because of a second feature which is correlated with the first. This has led to no shortage of misinterpretations and there are no obvious solutions in sight, leaving this as another open topic of modern GAM research.

4.5.2 Correlations and Dependencies

The correlations amongst the input variables pose a *fundamental challenge* to the study of additive models.

> I went into researching additive models thinking that interactions would be the hard part, but I came out knowing that correlations... they are from the devil.
>
> Rich Caruana

Applying linear regression to linearly related features has long been known to lead to unstable results with uninterpretable coefficients, a problem called colinearity. As already mentioned, in the extreme case of duplicated features, coefficients are no longer well-defined and require the implicit or explicit regularization of the learning algorithm to even be unique. Although common practice when constructing features may be to drop the redundant copy or when fitting models to use ridge regression (Hoerl and Kennard, 1970b,a), entire books are dedicated to identifying and resolving the problems induced in linear regression for predefined features (Belsley et al., 2005). Further approaches trying to improve the stability or interpretability of linear regression in these settings continue to the current day (Wold et al., 1984; Kejian, 1993; Haufe et al., 2014; Pazzani and Bay, 2020).

Given that there has yet to be a completely satisfactory resolution in the simple case of linear regression, it is no surprise that this remains an active topic in the research of additive models. In the space of GAMs, the issue of colinearity is replaced by the issue of *concurvity*, due to the fact that each of the shape function curves may be nonlinearly influenced by not only the linear correlations between two underlying features, but the higher-order dependencies between the two features. Although this usual refers to the case of 1D GAMs, the issue is widespread for both higher-order dependencies and for higher-order interaction models.

Moreover, the issue of concurvity is directly problematic for the goals of robust, sparse, and causal explanations. The robustness of the fitted shape functions can be ruined by redundant information in the features and the sparsity of the fitted shape functions contradicts the robust fit which aggregates noisy sources of the same information. The ability to causally interpret shape function plots is further befuddled by the dependencies which interlock several different shape functions as illustrating a single joint phenomenon. Although the lack of universal solution in the simple linear regression case should pose as a warning of its difficulty, the importance of this issue for interpreting GAMs is hard to overstate and only made more important by the modern era of deep learning on high-dimensional and correlated raw features.

Correlation-Adapted GAMs Many works have made progress on improving the robustness of GAMs under these more realistic settings. (Lengerich et al., 2020) provides an algorithm for purifying higher-order shape functions, allowing for a unique interpretation of equivalent GAM models. (Sun et al., 2022) directly solve for purified GAMs on discrete features by the usage of a 'pure coding' function. (Zhong et al., 2023) looks at taking the entire family of well-performing sparse additive models, called the Rashomon set (Xin et al., 2022). This allows for considering an explanation as a set of sparse explanations, where the large set of models allows for robustness, the sparsity of the individual models prevents the overall Rashomon set from becoming too uninterpretable. (McTavish et al., 2024) asks the question of whether or not the missingness indicator should be grouped with its respective feature or be treated as its own feature, noting that missingness may itself be correlated with other features.

(Siems et al., 2023) suggests minimizing the concurvity of GAMs by directly optimizing it in terms of minimizing the correlation coefficients.

$$\mathcal{L}^{\text{concurvity}}(f_1, \dots, f_d) := \sum_{i<j} \Big| \mathbb{C}\text{orr}(f_i, f_j) \Big| = \sum_{i<j} \Big| \frac{\mathbb{C}\text{ov}(f_i, f_j)}{\sqrt{\mathbb{V}\text{ar}(f_i)}\sqrt{\mathbb{V}\text{ar}(f_j)}} \Big| \tag{72}$$

(Enouen and Liu, 2025) suggests directly regularizing according to masking the shape functions under the Shapley kernel distribution, $m(S)$ from Equation 11.

$$\underset{f_\emptyset, f_1, \dots, f_d}{\text{argmin}} \left\{ \sum_{S \subseteq [d]} m(|S|) \cdot \mathbb{E}_X \left[\Big(f(X) - f_\emptyset - \sum_{i \in S} f_i(X_i) \Big)^2 \right] \right\} \tag{73}$$

(Clark et al., 2025) suggests that in scenarios where the target Y generates the features X, it is preferable to only have shape functions which directly correlated with Y. They follow the intuition of the PatternQLR method (Haufe et al., 2014) to linearly adjust shape functions to be maximally predictive of Y.

$$\begin{aligned} f_i^{\text{Pattern}} &= b_i f_i + d_i \qquad\qquad f_{i,j}^{\text{Pattern}} = b_{i,j} f_{i,j} + d_{i,j} \\ \text{where} \quad b_S, d_S &= \underset{b,d \in \mathbb{R}}{\text{argmin}} \Big\{ \mathcal{L}\text{oss}(Y; b \cdot f(X) + d) \Big\} \end{aligned} \tag{74}$$

Ironically enough, the additive model has also surprisingly proved to be a great tool for better understanding the correlation structures of the data. It does this by distinguishing specifically what is *not* a feature interaction. By considering the spectrum across all additive models, we can begin to identify which features are redundantly predictive and which features are synergistically predictive of the target. For more information, see the Amari decomposition in the next section or see (König et al., 2025; Enouen and Liu, 2025).

4.5.3 Generalized Functional ANOVA

Sobol'-Hoeffding ANOVA Due to the realized failings of the original functional ANOVA, now called the Sobol-Hoeffding ANOVA (Hoeffding, 1948; Sobol', 1990), additional functional ANOVA decompositions were developed for beyond the case of independent variables. It is worth mentioning that it is not always clear whether to say Sobol-Hoeffding refers to the marginal ANOVA decomposition or the conditional ANOVA decomposition because of Sobol's use of the assumption that the variables were independent; however, we will equate it with the conditional ANOVA as we feel is fairly common throughout the literature. This then means that we have the baseline ANOVA, the marginal ANOVA, the conditional ANOVA = the Sobol-Hoeffding ANOVA, and then we will discuss two more functional ANOVAs: the Stone-Hooker ANOVA and what we will call herein the Amari ANOVA.

$$f(x) = \sum_{S \subseteq [d]} f_S^{\text{Sobol}}(x_S) \quad \text{where} \quad f_S^{\text{Sobol}}(x_S) := \sum_{T \subseteq S} (-1)^{|S|-|T|} [\mathcal{M}_T \circ f](x_S) \tag{75}$$

Stone-Hooker ANOVA Hooker (Hooker, 2004), in his study on additive structure in blackbox functions, had already discovered Sobol's functional ANOVA decomposition. He was also aware of the key limitations in using this decomposition in the case of dependent variables, so later based on the work of Stone (Stone, 1994) for orthogonalizing tensor product bases, he would use the same functional ANOVA decomposition for arbitrary functions and provide novel estimation approaches (Hooker, 2007). This Stone-Hooker decomposition was a large step towards having a functional ANOVA decomposition which remains useful in the case of dependent input variables.

$$f(x) = \sum_{S \subseteq [d]} f_S^{\text{Hooker}}(x_S) \quad \text{s.t.} \quad [\mathcal{M}_{S-s} \circ f_S^{\text{Hooker}}](x_{S-s}) = 0 \quad \forall x_{S-s} \ \ \forall s \in S \ \ \forall S \tag{76}$$

Mara and Tarantola (2012) had also identified issues with applying the typical Sobol indices in the case of dependent input variables. Recall that the Sobol variances (the typical indices) and the Sobol covariances are defined as:

$$V_S^{\text{Sobol}} := \mathbb{V}\text{ar}\Big[f_S^{\text{Sobol}}\Big] \qquad = \mathbb{E}_{X_S}\Big[f_S^{\text{Sobol}}(X_S)^2\Big] \tag{77}$$

$$C_S^{\text{Sobol}} := \mathbb{C}\text{ov}\Big[f_S^{\text{Sobol}}, f\Big] \qquad = \mathbb{E}_X\Big[f_S^{\text{Sobol}}(X_S) \cdot f(X)\Big] \tag{78}$$

Chastaing et al. (2012) revisits the Stone-Hooker decomposition and suggests to consider the Hooker covariances as a new Sobol index, where the Hooker variances and Hooker covariances are defined in the obvious way. They would reintroduce the matrix form of the projection equations for the small two-dimensional case and later follow ups would continue exploring this direction (Chastaing and Gratiet, 2015; Chastaing et al., 2015).

$$V_S^{\text{Hooker}} := \mathbb{V}\text{ar}\Big[f_S^{\text{Hooker}}\Big] \qquad = \mathbb{E}_{X_S}\Big[f_S^{\text{Hooker}}(X_S)^2\Big] \tag{79}$$

$$C_S^{\text{Hooker}} := \mathbb{C}\text{ov}\Big[f_S^{\text{Hooker}}, f\Big] \qquad = \mathbb{E}_X\Big[f_S^{\text{Hooker}}(X_S) \cdot f(X)\Big] \tag{80}$$

Amari ANOVA Most recently, some recent approaches have completely embraced the connection with additive models and alternatively considered the trained (ordered-selected) additive models as defining the functional ANOVA decomposition, rather than the historical use of the functional ANOVA as a proxy for selecting additive terms (König et al., 2025; Enouen and Liu, 2025). We will herein call this ANOVA approach the Amari decomposition in reference to his similar decomposition of a probability distribution using the Riemannian geometry of statistical manifolds (Amari, 2001).

$$\{f_S^{\text{Amari}}\}_{|S|=k} = \underset{\{g_S\}}{\operatorname{argmin}} \left\{ \mathbb{E}_X \left[\Big(f(X) - \sum_{|S|<k} f_S^{\text{Amari}}(X_S) - \sum_{|S|=k} g_S(X_S) \Big)^2 \right] \right\} \qquad \forall k \in \{0, 1, \dots, d\} \tag{81}$$

König et al. (2025) uses this approach to be able to better decompose the full variance of a blackbox model, aiming to disentangle the variance which is coming from the synergies between two features and which is coming from the dependencies between two features. In order to do this, they directly fit a GAM model $f_j(x_j) + f_{-j}(x_{-j})$ to the data, bypassing non-uniqueness considerations with the variational formulation. Enouen and Liu (2025) additionally consider generalizing this variational approach to any possible higher-order additive model, providing the matrix-projection equations which characterize the solution to the variational problem (their Theorem 6).

4.5.4 Extensions Beyond Additive Models

Regional GAMs and Instance-wise Sparsity A large and important direction of extending additive models is the study of regionally additive models and localized sparse interactions. Regional additive models (Gkolemis et al., 2023), consider generalized additive models where certain interaction terms only exist within certain regions of the input space. This allows for much greater modeling flexibility while still having a semilocal GAM approximation to the model. This interpretable-by-design model type depends on their being relatively few regions, and generally trades between region complexity and interaction complexity. It is reminded that regionally additive models are a combination of the additive pillar with the locality of the concept pillar (and sometimes the anchors of the reasoning pillar), recall Figure 3.

This approach can be seen as the interpretable-by-design dual concept to region-specific additive explanations like REPID (Herbinger et al., 2022) and GADGET (Herbinger et al., 2024) which focus on providing effect plots which are regionally interaction-free. Moreover, this is part of a larger trend to increase explanation flexibility through the use of instance-wise sparsity. Although even looser than regional sparsity (with each data point receiving its own region), these methods still seem fundamentally connected. TabNet (Arik and Pfister, 2021) has local feature selection sparsity which may depend on all of the features (i.e. selecting the region), but the final prediction step will be dependent on those individual features (i.e. locally sparse model). Sum-of-Parts (You et al., 2025) provides an interaction attribution for natural images where the interaction sparsity pattern depends on the individual sample, again providing an explanation which is regionally an additive interaction explanation.

Pretrained Transformers Although tabular data has been traditionally dominated by boosting methods from machine learning (Shwartz-Ziv and Armon, 2022; McElfresh et al., 2023; Borisov et al., 2024), a recent departure from this trend is the work of TabPFN (Hollmann et al., 2023). TabPFN uproots the usual deep learning pipeline by training a transformer to map directly from a dataset to a test prediction, which enables pretraining on a large corpus of synthetic tabular data. This raises questions about the future of machine learning and whether the traditional paradigms of fitting a single model to a single dataset are still relevant in the post-LLM, post-pretraining landscape of artificial intelligence. TabPFN points out that even the simple and ubiquitous tabular data is not immune to having larger patterns across different datasets which are detectable by deep learning. GAMformer (Mueller et al., 2026) asks whether or not these globally learned patterns by a large-scale transformer can still be locally understood after fixing the training dataset. In particular, they train a TabPFN model which returns a 1D GAM model in terms of its predictions on new test samples.

Additional Hierarchies, Structures, and Geometries The incorporation of additional structure is absolutely critical for GAMs extensions attempting to perform well beyond the setting of tabular data with simple and interpretable features. Domains like natural language and computer vision, as well as data types like longitudinal data and graphical data require additional caution for additive modeling to be as useful as possible. Although basic modifications like grouping spatial coordinates, considering time as a distinguished features, or grouping missingness indicators with features have long been standard practice, these methods are not universally understand nor without nuance in their application. For example, the hyperbolic structure of word hierarchies and image label categories (Nickel and Kiela, 2017; Khrulkov et al., 2020; Ermolov et al., 2022), the autocorrelation structure of time series data (Whittle, 1951; Box et al., 1970), the connectivity structure of graphical data (Chung, 1997; Page et al., 1999; Perozzi et al., 2014; Kipf, 2016; Grover and Leskovec, 2016), or the fine-grained hierarchy structure of high-dimensional embeddings (Rendle, 2010; Lyu et al., 2023).

HCI Integration and Causal Insight As an interpretable method, it is also important to understand how GAMs are being interpreted and integrated into larger human-in-the-loop systems. Moreover, it is important to understand the heterogeneity of applications and their respective demands. This human-computer interaction (HCI) should always be considered as an extension of the GAM itself, and interpretability claims must be factored through the relevant application at hand (Bellotti and Edwards, 2001; Kaur et al., 2020).

Interpreting the claims of the GAM mdoel causally is amongst the most abundant of mistakes made by practitioners and non-experts alike, begging for methods which can automatically understand the contextually relevant causal claims and/or process external causal assumptions into the overall pipeline. In many cases, understanding the interpretable insights of the GAM in a causal light may be critical for downstream actionability of the insights (Workshop on Actionable Interpretability @ ICML 2025). Although causality is infamous for its difficulty and nuance, many works are already beginning to make progress on utilizing causal assumptions to be incorporated into additive attribution (Biparva and Materassi, 2024; Gajewski et al., 2025).

References

Francesca Dominici Aaron Fisher, Cynthia Rudin. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. 2019.

Ashraf Abdul, Christian von der Weth, Mohan Kankanhalli, and Brian Y. Lim. Cogam: Measuring and moderating cognitive load in machine learning model explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376615. URL https://doi.org/10.1145/3313831.3376615.

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.385. URL https://aclanthology.org/2020.acl-main.385/.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/294a8ed24b1ad22ec2e7efea049b8737-Paper.pdf.

Philip Adler, Casey Falk, Sorelle A. Friedler, Tionney Nix, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. Auditing black-box models for indirect influence. *Knowl. Inf. Syst.*, 54(1):95–122, January 2018. ISSN 0219-1377. doi: 10.1007/s10115-017-1116-3. URL https://doi.org/10.1007/s10115-017-1116-3.

Rishabh Agarwal, Levi Melnick, Nicholas Frosst, Xuezhou Zhang, Ben Lengerich, Rich Caruana, and Geoffrey E Hinton. Neural additive models: Interpretable machine learning with neural nets. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 4699–4711. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/251bd0442dfcc53b5a761e050f8022b8-Paper.pdf.

Raj Agrawal and Tamara Broderick. The skim-fa kernel: High-dimensional variable selection and nonlinear interaction discovery in linear time. *Journal of Machine Learning Research*, 24(27):1–60, 2023. URL http://jmlr.org/papers/v24/21-1403.html.

Raj Agrawal, Brian Trippe, Jonathan Huggins, and Tamara Broderick. The kernel interaction trick: Fast Bayesian discovery of pairwise interactions in high dimensions. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 141–150. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/agrawal19a.html.

Chunrong Ai and Edward C Norton. Interaction terms in logit and probit models. *Economics letters*, 80(1):123–129, 2003.

Leona S Aiken, Stephen G West, and Raymond R Reno. *Multiple regression: Testing and interpreting interactions*. sage, 1991.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2017. URL https://openreview.net/forum?id=ryF7rTqgl.

J.M. Alonso-Meijide, J. Freixas, and X. Molinero. Computation of several power indices by generating functions. *Applied Mathematics and Computation*, 219(8):3395–3402, 2012. ISSN 0096-3003. doi: https://doi.org/10.1016/j.amc.2012.10.021. URL https://www.sciencedirect.com/science/article/pii/S0096300312010089.

José María Alonso-Meijide and Josep Freixas. A new power index based on minimal winning coalitions without any surplus. *Decision Support Systems*, 49(1):70–76, 2010.

S.-I. Amari. Information geometry on hierarchy of probability distributions. *IEEE Transactions on Information Theory*, 47(5):1701–1711, 2001. doi: 10.1109/18.930911.

Salim I. Amoukou, Tangi Salaün, and Nicolas Brunel. Accurate shapley values for explaining tree-based models. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 2448–2465. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/amoukou22a.html.

Donald W. K. Andrews and Yoon-Jae Whang. Additive interactive regression models: Circumvention of the curse of dimensionality. *Econometric Theory*, 6(4):466–479, 1990. ISSN 02664666, 14694360. URL http://www.jstor.org/stable/3532092.

Daniel W. Apley and Jingyu Zhu. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(4):1059–1086, 06 2020. ISSN 1369-7412. doi: 10.1111/rssb.12377. URL https://doi.org/10.1111/rssb.12377.

Sercan Ö. Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8):6679–6687, May 2021. doi: 10.1609/aaai.v35i8.16826. URL https://ojs.aaai.org/index.php/AAAI/article/view/16826.

Kyohei Atarashi, Satoshi Oyama, and Masahito Kurihara. Factorization machines with regularization for sparse feature interactions. *Journal of Machine Learning Research*, 22(153):1–50, 2021. URL http://jmlr.org/papers/v22/20-1170.html.

R. J. AUMANN and L. S. SHAPLEY. *Values of Non-Atomic Games*. Princeton University Press, 1974. URL http://www.jstor.org/stable/j.ctt13x149m.

Robert J Aumann and Jacques H Dreze. Cooperative games with coalition structures. *International Journal of game theory*, 3(4):217–237, 1974.

Francis Bach. High-dimensional non-linear variable selection through hierarchical kernel learning, 2009. URL https://arxiv.org/abs/0909.0844.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):1–46, 07 2015. doi: 10.1371/journal.pone.0130140. URL https://doi.org/10.1371/journal.pone.0130140.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015. URL http://arxiv.org/abs/1409.0473.

John F. III Banzhaf. *Weighted Voting Doesn't Work: A Mathematical Analysis*, volume 19, pages 317–344. 1965.

Daniel Barry. Nonparametric bayesian regression. *The Annals of Statistics*, 14(3):934–953, 1986. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/3035551.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network Dissection: Quantifying Interpretability of Deep Visual Representations . In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3319–3327, Los Alamitos, CA, USA, July 2017. IEEE Computer Society. doi: 10.1109/CVPR.2017.354. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.354.

Robert M Bell, Yehuda Koren, Chris Volinsky, et al. The bellkor 2008 solution to the netflix prize. *Statistics Research Department at AT&T Research*, 1(1), 2008.

Robert M. Bell, Yehuda Koren, and Chris Volinsky. All together now: A perspective on the netflix prize. *CHANCE*, 23(1):24–29, 2010. doi: 10.1080/09332480.2010.10739787.

Victoria Bellotti and Keith Edwards. Intelligibility and accountability: human considerations in context-aware systems. *Hum.-Comput. Interact.*, 16(2):193–212, December 2001. ISSN 0737-0024. doi: 10.1207/S15327051HCI16234_05. URL https://doi.org/10.1207/S15327051HCI16234_05.

David A Belsley, Edwin Kuh, and Roy E Welsch. *Regression diagnostics: Identifying influential data and sources of collinearity.* John Wiley & Sons, 2005.

William A Belson. Matching and prediction on the principle of biological classification. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 8(2):65–75, 1959.

James Bennett and Stan Lanning. The netflix prize, 2007. KDDCup'07.

Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705 – 1732, 2009. doi: 10.1214/08-AOS620. URL https://doi.org/10.1214/08-AOS620.

Jacob Bien, Jonathan Taylor, and Robert Tibshirani. A lasso for hierarchical interactions. *Annals of statistics*, 41(3):1111, 2013.

Jacob Bien, Noah Simon, and Robert Tibshirani. Convex hierarchical testing of interactions. *The Annals of Applied Statistics*, 9(1):27–42, 2015. ISSN 19326157. URL http://www.jstor.org/stable/24522409.

Darya Biparva and Donatello Materassi. Incorporating information into shapley values: Reweighting via a maximum entropy approach. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=DwniHlwcOB.

Garrett Birkhoff and Carl R De Boor. Piecewise polynomial interpolation and approximation. *Approximation of functions*, pages 164–190, 1965.

Mathieu Blondel, Akinori Fujino, Naonori Ueda, and Masakazu Ishihata. Higher-order factorization machines. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/158fc2ddd52ec2cf54d3c161f2dd6517-Paper.pdf.

Sebastian Bordt and Ulrike von Luxburg. From shapley values to generalized additive models and back. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 709–745. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/bordt23a.html.

Sebastian Bordt, Eric Raidl, and Ulrike von Luxburg. Position: Rethinking explainable machine learning as applied statistics. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL https://openreview.net/forum?id=b2gM1HyAgE.

Vadim Borisov, Tobias Leemann, Kathrin Se⊆ *ler, JohannesHaug, MartinPawelczyk, andGjergjiKasneci.Dee*
Asurvey.IEEE Transactions on Neural Networks and Learning Systems, 35(6) : 7499−−7519, 2024.*doi* :
.

Kouroche Bouchiat, Alexander Immer, Hugo Yèche, Gunnar Ratsch, and Vincent Fortuin. Improving neural additive models with bayesian principles. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=0pSTzCnEmi.

G. E. P. Box and D. W. Behnken. Some new three level designs for the study of quantitative variables. *Technometrics*, 2(4):455–475, 1960. ISSN 00401706. URL http://www.jstor.org/stable/1266454.

G. E. P. Box and J. S. Hunter. Multi-factor experimental designs for exploring response surfaces. *The Annals of Mathematical Statistics*, 28(1):195–241, 1957. ISSN 00034851, 21688990. URL http://www.jstor.org/stable/2237033.

G. E. P. Box and K. B. Wilson. On the experimental attainment of optimum conditions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 13(1):1–45, 1951. ISSN 00359246. URL http://www.jstor.org/stable/2983966.

George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 1970.

L. Breiman, Jerome H. Friedman, Richard A. Olshen, and C. J. Stone. Classification and regression trees. 1984. URL https://doi.org/10.1201/9781315139470.

Leo Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37 (4):373–384, November 1995. ISSN 0040-1706. 10.2307/1269730. URL https://doi.org/10.2307/1269730.

Leo Breiman. Arcing classifier (with discussion and a rejoinder by the author). *The Annals of Statistics*, 26(3):801 – 849, 1998. 10.1214/aos/1024691079. URL https://doi.org/10.1214/aos/1024691079.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001a.

Leo Breiman. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199 – 231, 2001b. 10.1214/ss/1009213726. URL https://doi.org/10.1214/ss/1009213726.

Leo Breiman and Jerome H. Friedman. Estimating optimal transformations for multiple regression and correlation: Rejoinder. *Journal of the American Statistical Association*, 80 (391):614–619, 1985. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2288477.

Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. In *ICLR (Poster)*, 2019. URL https://openreview.net/forum?id=SkfMWhAqYQ.

Andreas Buja, Trevor Hastie, and Robert Tibshirani. Linear smoothers and additive models. *The Annals of Statistics*, pages 453–510, 1989.

Prabir Burman. Estimation of generalized additive models. *Journal of Multivariate Analysis*, 32(2):230–255, 1985. ISSN 0047-259X. https://doi.org/10.1016/0047-259X(90)90083-T. URL https://www.sciencedirect.com/science/article/pii/0047259X9090083T.

Pierre Bézier. *Numerical Control - Mathematics and Applications*. John Wiley and Sons, 1972.

E.J. Candes and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005. 10.1109/TIT.2005.858979.

Emmanuel Candes and Terence Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313 – 2351, 2007. 10.1214/009053606000001523. URL https://doi.org/10.1214/009053606000001523.

Brandon Carter, Jonas Mueller, Siddhartha Jain, and David Gifford. What made you do this? understanding black-box decisions with sufficient input subsets. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 567–576. PMLR, 16–18 Apr 2019a. URL https://proceedings.mlr.press/v89/carter19a.html.

Shan Carter, Zan Armstrong, Ludwig Schubert, Ian Johnson, and Chris Olah. Activation atlas. *Distill*, 2019b. 10.23915/distill.00015. https://distill.pub/2019/activation-atlas.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1721–1730. ACM, 2015.

P. De Casteljau. Mathématiques et cao. volume 2: Formes à pôles. *Hermes*, 1986.

Karen Chan, Andrea Saltelli, and Stefano Tarantola. Sensitivity analysis of model output: variance-based methods make the difference. In *Proceedings of the 29th Conference on Winter Simulation*, WSC '97, page 261–268, USA, 1997. IEEE Computer Society. ISBN 078034278X. 10.1145/268437.268489. URL https://doi.org/10.1145/268437.268489.

Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. Explaining image classifiers by counterfactual generation. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1MXz20cYQ.

Chun-Hao Chang, Sarah Tan, Ben Lengerich, Anna Goldenberg, and Rich Caruana. How interpretable and trustworthy are gams? In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 95–105, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383325. 10.1145/3447548.3467453. URL https://doi.org/10.1145/3447548.3467453.

Chun-Hao Chang, Rich Caruana, and Anna Goldenberg. NODE-GAM: Neural generalized additive model for interpretable deep learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=g8NJR6fCCl8.

J. Douglas Carroll & Jih-Jie Chang. Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition, 1970.

A. Charnes, B. Golany, M. Keane, and J. Rousseau. *Extremal Principle Solutions of Games in Characteristic Function Form: Core, Chebychev and Shapley Value Generalizations*, pages 123–133. Springer Netherlands, Dordrecht, 1988. ISBN 978-94-009-3677-5. 10.1007/978-94-009-3677-$5_7.URL$.

G. Chastaing and L. Le Gratiet. Anova decomposition of conditional gaussian processes for sensitivity analysis with dependent inputs. *Journal of Statistical Computation and Simulation*, 85(11):2164–2186, 2015. 10.1080/00949655.2014.925111.

G. Chastaing, F. Gamboa, and C. Prieur. Generalized sobol sensitivity indices for dependent variables: numerical methods. *Journal of Statistical Computation and Simulation*, 85 (7):1306–1333, 2015. 10.1080/00949655.2014.960415.

Gaelle Chastaing, Fabrice Gamboa, and Clémentine Prieur. Generalized Hoeffding-Sobol decomposition for dependent variables - application to sensitivity analysis. *Electronic Journal of Statistics*, 6(none):2420 – 2448, 2012. 10.1214/12-EJS749. URL https://doi.org/10.1214/12-EJS749.

Siu Lun Chau, Robert Hu, Javier Gonzalez, and Dino Sejdinovic. RKHS-SHAP: Shapley values for kernel methods. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=gnc2VJHXmsG.

Siu Lun Chau, Krikamol Muandet, and Dino Sejdinovic. Explaining the uncertain: Stochastic shapley values for gaussian process models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 50769–50795. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/9f0b1220028dfa2ee82ca0a0e0fc52d1-Paper-Conference.pdf.

Chaofan Chen and Cynthia Rudin. An optimization approach to learning falling rule lists. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 604–612. PMLR, 09–11 Apr 2018. URL https://proceedings.mlr.press/v84/chen18a.html.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: Deep learning for interpretable image recognition. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/adf7ee2dcf142b0e11888e72b43fcb75-Paper.pdf.

Hugh Chen, Joseph D. Janizek, Scott Lundberg, and Su-In Lee. True to the model or true to the data?, 2020. URL https://arxiv.org/abs/2006.16234.

Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001. 10.1137/S003614450037906X. URL https://doi.org/10.1137/S003614450037906X.

Xiaohong Chen. Chapter 76 large sample sieve estimation of semi-nonparametric models. volume 6 of *Handbook of Econometrics*, pages 5549–5632. Elsevier, 2007. https://doi.org/10.1016/S1573-4412(07)06076-X. URL https://www.sciencedirect.com/science/article/pii/S157344120706076X.

Zehua Chen. A stepwise approach for the purely periodic interaction spline model. *Communications in Statistics - Theory and Methods*, 16(3):877–895, 1987. 10.1080/03610928708829409. URL https://doi.org/10.1080/03610928708829409.

Zehua Chen. Interaction spline models and their convergence rates. *The Annals of Statistics*, 19(4):1855–1868, 1991a. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/2241907.

Zehua Chen. Interaction spline models and their convergence rates. *The Annals of Statistics*, 19(4):1855–1868, 1991b. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/2241907.

Zehua Chen, Chong Gu, and Grace Wahba. Discussion: Linear smoothers and additive models. *The Annals of Statistics*, 17(2):515–522, 1989.

Nam Hee Choi, William Li, and Ji Zhu. Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association*, 105(489):354–364, 2010. 10.1198/jasa.2010.tm08281. URL https://doi.org/10.1198/jasa.2010.tm08281.

Fan RK Chung. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997.

Benedict Clark, Rick Wilming, Hjalmar Schulz, Rustam Zhumagambetov, Danny Panknin, and Stefan Haufe. Correcting misinterpretations of additive models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=2ClM0g9OFT.

R. Dennis Cook. Detection of influential observation in linear regression. *Technometrics*, 19(1):15–18, 1977. ISSN 00401706. URL http://www.jstor.org/stable/1268249.

R. Dennis Cook and Sanford Weisberg. Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics*, 22(4):495–508, 1980. ISSN 00401706. URL http://www.jstor.org/stable/1268187.

R. Dennis Cook and Sanford Weisberg. *Residuals and Influence in Regression*. New York: Chapman and Hall, 1982.

Ian Covert, Scott M Lundberg, and Su-In Lee. Understanding global feature contributions with additive importance measures. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17212–17223. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/c7bf0b7c1a86d5eb3be2c722cf2cf746-Paper.pdf.

Ian Covert, Scott Lundberg, and Su-In Lee. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209):1–90, 2021. URL http://jmlr.org/papers/v22/20-1316.html.

Peter Craven and Grace Wahba. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische mathematik*, 31(4):377–403, 1978.

Adnan Darwiche and Auguste Hirth. On the reasons behind decisions, 2020. URL https://arxiv.org/abs/2002.09284.

Adnan Darwiche and Pierre Marquis. A knowledge compilation map. *J. Artif. Int. Res.*, 17(1):229–264, September 2002. ISSN 1076-9757.

Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. *2016 IEEE Symposium on Security and Privacy (SP)*, pages 598–617, 2016. URL https://api.semanticscholar.org/CorpusID:2367610.

Carl De Boor. *A practical guide to splines*, volume 27.

Angela Dean, Max Morris, John Stufken, and Derek Bingham. *Handbook of design and analysis of experiments*, volume 7. CRC Press, 2015.

A D Deev. Representation of statistics of discriminant analysis, and asymptotic expansion when space dimensions are comparable with sample size. *Dokl. Akad. Nauk SSSR*, 195: 759–762, 1970.

David L. Donoho. For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829, 2006a. https://doi.org/10.1002/cpa.20132. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.20132.

David L. Donoho. For most large underdetermined systems of linear equations the minimal $\ell 1$-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829, 2006b. https://doi.org/10.1002/cpa.20132. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.20132.

D.L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47(7):2845–2862, 2001. 10.1109/18.959265.

Abhimanyu Dubey, Filip Radenovic, and Dhruv Mahajan. Scalable interpretability via polynomials. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=TwuColwZAVj.

Pradeep Dubey and Robert J. Weber. Probabilistic values for games. Technical report, Yale University, 1977. URL https://elischolar.library.yale.edu/cowles-discussion-paper-series/703. Cowles Foundation Discussion Papers. 703.

Pradeep Dubey, Abraham Neyman, and Robert James Weber. Value theory without efficiency. *Mathematics of Operations Research*, 6(1):122–128, 1981. ISSN 0364765X, 15265471. URL http://www.jstor.org/stable/3689271.

J Duchon. Spline minimizing rotation-invariant seminorms in sobolev spaces. *Constructive Theory of Functions of Several Variables*, 1977.

Aaron Courville Dumitru Erhan, Yoshua Bengio and Pascal Vincent. Visualizing higher-layer features of a deep network. Technical report, Universite de Montreal, 2009.

Nicolas Durrande, David Ginsbourger, and Olivier Roustant. Additive kernels for gaussian process modeling, 2011. URL https://arxiv.org/abs/1103.4023.

David Duvenaud. *Automatic Model Construction with Gaussian Processes*. PhD thesis, University of Cambridge, 2014.

David K Duvenaud, Hannes Nickisch, and Carl Rasmussen. Additive gaussian processes. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/file/4c5bde74a8f110656874902f07378009-Paper.pdf.

Yoshua Bengio Dzmitry Bahdanau, Kyunghyun Cho. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations*, 2015.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407 – 499, 2004. 10.1214/009053604000000067. URL https://doi.org/10.1214/009053604000000067.

Churchill Eisenhart. The assumptions underlying the analysis of variance. *Biometrics*, 3(1): 1–21, 1947. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/3001534.

M. Elad and A.M. Bruckstein. A generalized uncertainty principle and sparse representation in pairs of bases. *IEEE Transactions on Information Theory*, 48(9):2558–2567, 2002. 10.1109/TIT.2002.801410.

Robert F. Engle, C. W. J. Granger, John Rice, and Andrew Weiss. Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association*, 81(394):310–320, 1986. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2289218.

James Enouen and Yan Liu. Sparse interaction additive networks via feature interaction detection and sparse selection. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 13908–13920. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/5a3674849d6d6d23ac088b9a2552f323-Paper-Conference.pdf.

James Enouen and Yan Liu. InstaSHAP: Interpretable additive models explain shapley values instantly. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ky7vVlBQBY.

Aleksandr Ermolov, Leyla Mirvakhabova, Valentin Khrulkov, Nicu Sebe, and Ivan Oseledets. Hyperbolic vision transformers: Combining improvements in metric learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7399–7409, 2022. 10.1109/CVPR52688.2022.00726.

R. L. Eubank. Diagnostics for smoothing splines. *Journal of the Royal Statistical Society. Series B (Methodological)*, 47(2):332–341, 1985. ISSN 00359246. URL http://www.jstor.org/stable/2345576.

Dan S. Felsenthal and Moshe Machover. *The Measurement of Voting Power*, volume None of *Books*. Edward Elgar Publishing, none edition, None 1998. None. URL https://ideas.repec.org/b/elg/eebook/1489.html.

D.S. Felsenthal. A well-behaved index of a priori p-power for simple n-person games. 2016.

Khashayar Filom, Alexey Miroshnikov, Konstandinos Kotsiopoulos, and Arjun Ravi Kannan. On marginal feature attributions of tree-based models. *Foundations of Data Science*, 6(4):395–467, 2024. ISSN 2639-8001. 10.3934/fods.2024021. URL http://dx.doi.org/10.3934/fods.2024021.

Ronald A. Fisher. *The design of experiments*. Oliver & Boyd, 1935.

Ronald Aylmer Fisher. *Statistical methods for research workers*. Genesis Publishing Pvt Ltd, 1925.

Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2950–2958, 2019. 10.1109/ICCV.2019.00304.

Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3449–3457, 2017. URL https://api.semanticscholar.org/CorpusID:1633753.

Anton Frederik Thielmann, René-Marcel Kruse, Thomas Kneib, and Benjamin Säfken. Neural additive models for location scale and shape: A framework for interpretable neural regression beyond the mean. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 1783–1791. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/frederik-thielmann24a.html.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2):337–407, 2000.

Jerome H. Friedman. Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19(1):1 – 67, 1991. 10.1214/aos/1176347963. URL https://doi.org/10.1214/aos/1176347963.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/2699986.

Jerome H. Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378, 2002. ISSN 0167-9473. https://doi.org/10.1016/S0167-9473(01)00065-2. URL https://www.sciencedirect.com/science/article/pii/S0167947301000652. Nonlinear Methods and Data Mining.

Jerome H Friedman and Bogdan E Popescu. Predictive learning via rule ensembles. *The Annals of Applied Statistics*, pages 916–954, 2008.

Jerome H. Friedman and Werner Stuetzle. Projection pursuit regression. *Journal of the American Statistical Association*, 76(376):817–823, 1981. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2287576.

Jerome H. Friedman, Eric Grosse, and Werner Stuetzle. Multidimensional additive spline approximation. *SIAM Journal on Scientific and Statistical Computing*, 4(2):291–301, 1983. 10.1137/0904023. URL https://doi.org/10.1137/0904023.

Jerome H. Friedman, Werner Stuetzle, and Anne Schroeder. Projection pursuit density estimation. *Journal of the American Statistical Association*, 79(387):599–608, 1984. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2288406.

J.H. Friedman and J.W. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, C-23(9):881–890, 1974. 10.1109/T-C.1974.224051.

Christopher Frye, Damien de Mijolla, Tom Begley, Laurence Cowton, Megan Stanley, and Ilya Feige. Shapley explainability on the data manifold. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=OPyWRrcjVQw.

Fabian Fumagalli, Maximilian Muschalik, Eyke Hüllermeier, Barbara Hammer, and Julia Herbinger. Unifying feature-based explanations with functional anova and cooperative game theory. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan, editors, *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 5140–5148. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/fumagalli25a.html.

Magzhan Gabidolla and Miguel Á. Carreira-Perpiñán. Generalized additive models via direct optimization of regularized decision stump forests. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=3WLpiPeJbk.

Mateusz Gajewski, Mateusz Olko, Mikołaj Morzy, and Piotr Sankowski. Markov-boundary consistent feature attribution. In *ICML 2025 Workshop on Scaling Up Intervention Models*, 2025. URL https://openreview.net/forum?id=eh9GSU9wDj.

Francis Galton. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263, 1886. ISSN 09595295, 23972564. URL http://www.jstor.org/stable/2841583.

Sir Galton, Francis. *Natural inheritance.* New York, Macmillan and co, 1894, 1894.

Carl Friedrich Gauss. *Theoria motus corporum coelestium in sectionibus conicis solem ambientium auctore Carolo Friderico Gauss.* sumtibus Frid. Perthes et IH Besser, 1809.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025. URL http://jmlr.org/papers/v26/23-0058.html.

Seymour Geisser and Samuel W. Greenhouse. An extension of box's results on the use of the f distribution in multivariate analysis. *The Annals of Mathematical Statistics*, 29(3): 885–891, 1958. ISSN 00034851, 21688990. URL http://www.jstor.org/stable/2237272.

Muriel Gevrey, Ioannis Dimopoulos, and Sovan Lek. Two-way interaction of input variables in the sensitivity analysis of neural network models. *Ecological modelling*, 195(1-2):43–50, 2006.

Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2242–2251. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/ghorbani19c.html.

Amirata Ghorbani, Michael Kim, and James Zou. A distributional framework for data valuation. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3535–3544. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/ghorbani20a.html.

Donald B Gillies. Solutions to general non-zero-sum games. *Contributions to the Theory of Games*, 4(40):47–85, 1959.

David Ginsbourger. *Gaussian random field models for function approximation under structural priors and adaptive design of experiments.* Habilitation thesis, University of Bern, 2013.

David Ginsbourger, Olivier Roustant, Dominic Schuhmacher, Nicolas Durrande, and Nicolas Lenz. On anova decompositions of kernels and gaussian random field paths. In Ronald

Cools and Dirk Nuyens, editors, *Monte Carlo and Quasi-Monte Carlo Methods*, pages 315–330, Cham, 2016. Springer International Publishing. ISBN 978-3-319-33507-0.

Vasilis Gkolemis, Anargiros Tzerefos, Theodore Dalamagas, Eirini Ntoutsi, and Christos Diou. Regionally additive models: Explainable-by-design models minimizing feature interactions, 2023. URL https://arxiv.org/abs/2309.12215.

Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65, 2015. 10.1080/10618600.2014.907095. URL https://doi.org/10.1080/10618600.2014.907095.

Gene H Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

Michel Grabisch and Marc Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of Game Theory*, 1999. 10.1007/s001820050125. URL https://doi.org/10.1007/s001820050125.

Peter Green, Christopher Jennison, and Allan Seheult. Analysis of field experiments by least squares smoothing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 47(2):299–315, 1985. ISSN 00359246. URL http://www.jstor.org/stable/2345573.

Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, 2016.

Chong Gu. Smoothing spline density estimation: Conditional distribution. *Statistica Sinica*, 5(2):709–726, 1995. ISSN 10170405, 19968507. URL http://www.jstor.org/stable/24305065.

Chong Gu. Penalized likelihood hazard estimation: A general procedure. *Statistica Sinica*, 6(4):861–876, 1996. ISSN 10170405, 19968507. URL http://www.jstor.org/stable/24306046.

Chong Gu. Structural multivariate function estimation: Some automatic density and hazard estimates. *Statistica Sinica*, 8(2):317–335, 1998. ISSN 10170405, 19968507. URL http://www.jstor.org/stable/24306495.

Chong Gu. *Smoothing spline ANOVA models*, volume 297. Springer, 2002.

Chong Gu and Grace Wahba. Discussion: Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):115–123, 1991a. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/2241846.

Chong Gu and Grace Wahba. Minimizing gcv/gml scores with multiple smoothing parameters via the newton method. *SIAM Journal on Scientific and Statistical Computing*, 12 (2):383–398, 1991b. 10.1137/0912021. URL https://doi.org/10.1137/0912021.

Chong Gu and Grace Wahba. Semiparametric analysis of variance with tensor product thin plate splines. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55 (2):353–368, 1993a. ISSN 00359246. URL http://www.jstor.org/stable/2346197.

Chong Gu and Grace Wahba. Smoothing spline anova with component-wise bayesian "confidence intervals". *Journal of Computational and Graphical Statistics*, 2(1):97–117, 1993b. ISSN 10618600. URL http://www.jstor.org/stable/1390957.

Chong Gu, Douglas M. Bates, Zehua Chen, and Grace Wahba. The computation of generalized cross-validation functions through householder tridiagonalization with applications to the fitting of interaction spline models. *SIAM Journal on Matrix Analysis and Applications*, 10(4):457–480, 1989. 10.1137/0610033. URL https://doi.org/10.1137/0610033.

S.R. Gunn and J.S. Kandola. Structural Modelling with Sparse Kernels. 2002.

Steve R. Gunn and Martin Brown. SUPANOVA - A Sparse, Transparent Modelling Approach. 1999.

Eric Günther, Balázs Szabados, Robi Bhattacharjee, Sebastian Bordt, and Ulrike von Luxburg. Informative post-hoc explanations only exist for simple functions, 2025. URL https://arxiv.org/abs/2508.11441.

Joseph Y. Halpern. *Actual Causality*. The MIT Press, 2016. ISBN 9780262035026. URL http://www.jstor.org/stable/j.ctt1f5g5p9.

Zayd Hammoudeh and Daniel Lowd.

Ning Hao and Hao Helen Zhang. Interaction screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 109(507):1285–1301, 2014a.

Ning Hao and Hao Helen Zhang. Interaction screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 109(507):1285–1301, 2014b.

John C. Harsanyi. A simplified bargaining model for the n-person cooperative game. *International Economic Review*, 4(2):194–220, 1963. ISSN 00206598, 14682354. URL http://www.jstor.org/stable/2525487.

R. A Harshman. Foundations of the parafac procedure: Models and conditions for an "explanatory" multimodal factor analysis, 1970.

Sergiu Hart and Mordecai Kurz. Endogenous formation of coalitions. *Econometrica*, 51 (4):1047–1064, 1983. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1912051.

Sergiu Hart and Andreu Mas-Colell. Potential, value, and consistency. *Econometrica*, 57(3): 589–614, 1989. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1911054.

Trevor Hastie and Robert Tibshirani. Generalized additive models. Technical report, Dept. of Statistics, Stanford University, 1984.

Trevor Hastie and Robert Tibshirani. Generalized additive models. *Statistical science*, 1 (3):297–310, 1986.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

Trevor J Hastie and Robert J Tibshirani. Generalized additive models, 1990.

Stefan Haufe, Frank Meinecke, Kai Görgen, Sven Dähne, John-Dylan Haynes, Benjamin Blankertz, and Felix Bie$\subseteq mann.Ontheinterpretationofweightvectorsoflinearmodelsinmultivariateneuroim$ $96--110, 2014.ISSN1053-8119.https://doi.org/10.1016/j.neuroimage.2013.10.067.URL.$

Julia Herbinger, Bernd Bischl, and Giuseppe Casalicchio. Repid: Regional effect plots with implicit interaction detection. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 10209–10233. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/herbinger22a.html.

Julia Herbinger, Marvin N. Wright, Thomas Nagler, Bernd Bischl, and Giuseppe Casalicchio. Decomposing global feature effects based on feature interactions. *Journal of Machine Learning Research*, 25(381):1–65, 2024. URL http://jmlr.org/papers/v25/23-0699.html.

Andrew Herren and P. Richard Hahn. Statistical aspects of shap: Functional anova for model interpretation, 2022. URL https://arxiv.org/abs/2208.09970.

Tom Heskes, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4778–4789. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/32e54441e6382a7fbacbbbaf3c450059-Paper.pdf.

Tin Kam Ho. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282 vol.1, 1995. 10.1109/ICDAR.1995.598994.

Wassily Hoeffding. A Class of Statistics with Asymptotically Normal Distribution. *The Annals of Mathematical Statistics*, 19(3):293 – 325, 1948. 10.1214/aoms/1177730196. URL https://doi.org/10.1214/aoms/1177730196.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Applications to nonorthogonal problems. *Technometrics*, 12(1):69–82, 1970a. 10.1080/00401706.1970.10488635.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970b. 10.1080/00401706.1970.10488634.

Manfred J Holler. A priori party power and government formation. *Munich Social Science Review*, 4:25–41, 1978.

Manfred J Holler and Stefan Napel. Monotonicity of power and power measures. *Theory and Decision*, 56(1):93–111, 2004.

Manfred J. Holler and Edward W. Packel. Power, luck and the right index. *Zeitschrift für Nationalökonomie / Journal of Economics*, 43(1):21–29, 1983. ISSN 00443158, 23048360. URL http://www.jstor.org/stable/41798164.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=cp5PvcI6w8_.

Fuxing Hong, Dongbo Huang, and Ge Chen. Interaction-aware factorization machines for recommender systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33 (01):3804–3811, Jul. 2019. 10.1609/aaai.v33i01.33013804. URL https://ojs.aaai.org/index.php/AAAI/article/view/4267.

Giles Hooker. Discovering additive structure in black box functions. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 575–580. ACM, 2004.

Giles Hooker. Generalized functional anova diagnostics for high-dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics*, 16(3):709–732, 2007.

Giles Hooker and Lucas Mentch. Please stop permuting features: An explanation and alternatives, 2019. URL https://arxiv.org/abs/1905.03151.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

Peter J. Huber. Robust regression: Asymptotics, conjectures and monte carlo. *The Annals of Statistics*, 1(5):799–821, 1973. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/2958283.

Marcus Hutter. *Universal Artificial Intellegence - Sequential Decisions Based on Algorithmic Probability.* Springer, 2005.

Huynh Huynh and Leonard S. Feldt. Conditions under which mean square ratios in repeated measurements designs have exact f-distributions. *Journal of the American Statistical Association*, 65(332):1582–1589, 1970. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2284340.

Shibal Ibrahim, Gabriel Afriat, Kayhan Behdin, and Rahul Mazumder. Grand-slamin' interpretable additive modeling with structural constraints. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 61158–61186. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/c057cb81b8d3c67093427bf1c16a4e9f-Paper-Conference.pdf.

Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 10.18653/v1/N19-1357. URL https://aclanthology.org/N19-1357.

Dominik Janzing, Lenon Minorics, and Patrick Bloebaum. Feature relevance quantification in explainable ai: A causal problem. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2907–2916. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/janzing20a.html.

Neil Jethani, Mukund Sudarshan, Ian Connick Covert, Su-In Lee, and Rajesh Ranganath. FastSHAP: Real-time shapley value estimation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=Zq2G_VTV53T.

Yang Ji, Ying Sun, Yuting Zhang, Zhigaoyuan Wang, Yuanxin Zhuang, Zheng Gong, Dazhong Shen, Chuan Qin, Hengshu Zhu, and Hui Xiong. A comprehensive survey on self-interpretable neural networks, 2025. URL https://arxiv.org/abs/2501.15638.

Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nezihe Merve Gurel, Bo Li, Ce Zhang, Costas Spanos, and Dawn Song. Efficient task-specific data valuation for nearest neighbor algorithms. *Proc. VLDB Endow.*, 12(11):1610–1623, July 2019a. ISSN 2150-8097. 10.14778/3342263.3342637. URL https://doi.org/10.14778/3342263.3342637.

Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas J. Spanos. Towards efficient data valuation based on the shapley value. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1167–1176. PMLR, 16–18 Apr 2019b. URL https://proceedings.mlr.press/v89/jia19a.html.

R J Johnston. National sovereignty and national power in european institutions. *Environment and Planning A: Economy and Space*, 9(5):569–577, 1977. 10.1068/a090569.

John R. Josephson and Susan G. Josephson, editors. *Abductive Inference: Computation, Philosophy, Technology*. Cambridge University Press, New York, 1994.

Yonghan Jung, Shiva Kasiviswanathan, Jin Tian, Dominik Janzing, Patrick Bloebaum, and Elias Bareinboim. On measuring causal contributions via do-interventions, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/jung22a.html.

Ehud Kalai and Dov Samet. Monotonic solutions to general cooperative games. *Econometrica: Journal of the Econometric Society*, pages 307–327, 1985.

Ehud Kalai and Dov Samet. On weighted shapley values. *International journal of game theory*, 16(3):205–222, 1987.

Kirthevasan Kandasamy and Yaoliang Yu. Additive approximations in high dimensional nonparametric regression via the salsa. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 69–78, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/kandasamy16.html.

Gordon V Kass. An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 29(2):119–127, 1980.

Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. 10.1145/3313831.3376219. URL https://doi.org/10.1145/3313831.3376219.

Liu Kejian. A new class of blased estimate in linear regression. *Communications in Statistics - Theory and Methods*, 22(2):393–402, 1993. 10.1080/03610929308831027.

Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2668–2677. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kim18d.html.

Minkyu Kim, Hyun-Soo Choi, and Jinho Kim. Higher-order neural additive models: An interpretable machine learning model with feature interactions, 2022. URL https://arxiv.org/abs/2209.15409.

George Kimeldorf and Grace Wahba. Some results on tchebycheffian spline functions. *Journal of mathematical analysis and applications*, 33(1):82–95, 1971.

George S. Kimeldorf and Grace Wahba. A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41 (2):495–502, 1970. URL http://www.jstor.org/stable/2239347.

TN Kipf. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1885–1894. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/koh17a.html.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5338–5348. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

Ron Kohavi. Bottom-up induction of oblivious read-once decision graphs: strengths and limitations. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (Vol. 1)*, AAAI '94, page 613–618, USA, 1994. American Association for Artificial Intelligence. ISBN 0262611023.

Ron Kohavi and Chia-Hsin Li. Oblivious decision trees graphs and top down pruning. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'95, page 1071–1077, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1558603638.

Patrick Kolpaczki, Viktor Bengs, Maximilian Muschalik, and Eyke Hüllermeier. Approximating the shapley value without marginal contributions. *Proceedings of the AAAI Confer-*

ence on Artificial Intelligence, 38(12):13246–13255, Mar. 2024. 10.1609/aaai.v38i12.29225. URL https://ojs.aaai.org/index.php/AAAI/article/view/29225.

Vladimir Koltchinskii and Ming Yuan. Sparsity in multiple kernel learning. *The Annals of Statistics*, 38(6):3660 – 3695, 2010. 10.1214/10-AOS825. URL https://doi.org/10.1214/10-AOS825.

Ramaravind Kommiya Mothilal, Divyat Mahajan, Chenhao Tan, and Amit Sharma. Towards unifying feature attribution and counterfactual explanations: Different means to the same end. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 652–663, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384735. 10.1145/3461702.3462597. URL https://doi.org/10.1145/3461702.3462597.

Yinfei Kong, Daoji Li, Yingying Fan, and Jinchi Lv. Interaction pursuit in high-dimensional multi-response regression via distance correlation. *The Annals of Statistics*, 45(2):897–922, 2017. ISSN 00905364. URL http://www.jstor.org/stable/44245827.

Gunnar König, Eric Günther, and Ulrike von Luxburg. Disentangling interactions and dependencies in feature attributions. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025. URL https://openreview.net/forum?id=BxniFq6TRd.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

William Kruskal. Relative importance by averaging over orderings. *The American Statistician*, 41(1):6–10, 1987. ISSN 00031305. URL http://www.jstor.org/stable/2684310.

Yongchan Kwon and James Zou. Beta shapley: a unified and noise-reduced data valuation framework for machine learning. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 8780–8802. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/kwon22a.html.

Yongchan Kwon, Manuel A. Rivas, and James Zou. Efficient computation and analysis of distributional shapley values. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 793–801. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/kwon21a.html.

A.M. Legendre. *Nouvelles méthodes pour la détermination des orbites des comètes.* F. Didot, 1805. URL https://books.google.com/books?id=FRcOAAAAQAAJ.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018. 10.1080/01621459.2017.1307116.

Benjamin Lengerich, Sarah Tan, Chun-Hao Chang, Giles Hooker, and Rich Caruana. Purifying interaction effects with the functional anova: An efficient algorithm for recovering identifiable additive models. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2402–2412. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/lengerich20a.html.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=DeG07_TcZvT.

Michael Lim and Trevor Hastie. Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics*, 24(3):627–654, 2015. ISSN 10618600. URL http://www.jstor.org/stable/24737287.

Weilin Lin, Xiangyu Zhao, Yejing Wang, Tong Xu, and Xian Wu. Adafs: Adaptive feature selection in deep recommender system. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3309–3317, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393850. 10.1145/3534678.3539204. URL https://doi.org/10.1145/3534678.3539204.

Yi Lin and Hao Helen Zhang. Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics*, 34(5):2272–2297, 2006. ISSN 00905364. URL http://www.jstor.org/stable/25463508.

Merenda Peter F. Gold Ruth Z. Lindeman, Richard H. *Introduction to bivariate and multivariate analysis*, chapter 4. Scott, Foresman and Company, 1980.

Bin Liu, Chenxu Zhu, Guilin Li, Weinan Zhang, Jincai Lai, Ruiming Tang, Xiuqiang He, Zhenguo Li, and Yong Yu. Autofis: Automatic feature interaction selection in factorization models for click-through rate prediction. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 2636–2645, New York, NY, USA, 2020a. Association for Computing Machinery. ISBN 9781450379984. 10.1145/3394486.3403314. URL https://doi.org/10.1145/3394486.3403314.

Guodong Liu, Hong Chen, and Heng Huang. Sparse shrunk additive models. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6194–6204. PMLR, 13–18 Jul 2020b. URL https://proceedings.mlr.press/v119/liu20b.html.

Andrés F López-Lopera, Francois Bachoc, and Olivier Roustant. High-dimensional additive gaussian processes under monotonicity constraints. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=YCPmfirAcc.

Yin Lou, Rich Caruana, and Johannes Gehrke. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 150–158. ACM, 2012.

Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 623–631. ACM, 2013.

Xiaoyu Lu, Alexis Boukouvalas, and James Hensman. Additive Gaussian processes revisited. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 14358–14383. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/lu22b.html.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles, 2018. URL https://arxiv.org/abs/1802.03888.

Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):56–67, Jan 2020. ISSN 2522-5839. 10.1038/s42256-019-0138-9.

Hengrui Luo, Giovanni Nattino, and Matthew T. Pratola. Sparse additive gaussian process regression. *Journal of Machine Learning Research*, 23(61):1–34, 2022. URL http://jmlr.org/papers/v23/19-597.html.

Fuyuan Lyu, Xing Tang, Dugang Liu, Chen Ma, Weihong Luo, Liang Chen, xiuqiang He, and Xue (Steve) Liu. Towards hybrid-grained feature interaction selection for deep sparse network. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 49325–49340. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/9ab8da29b1eb3bec912a06e0879065cd-Paper-Conference.pdf.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. Towards faithful model explanation in NLP: A survey. *Computational Linguistics*, 50(2):657–723, June 2024. 10.1162/coli$_{a0}$0511.URL.

S.G. Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993. 10.1109/78.258082.

Thierry A. Mara and Stefano Tarantola. Variance-based sensitivity indices for models with dependent inputs. *Reliability Engineering System Safety*, 107:115–121, 2012. ISSN 0951-8320. https://doi.org/10.1016/j.ress.2011.08.008. URL https://www.sciencedirect.com/science/article/pii/S0951832011001724. SAMO 2010.

V. A. Marčenko and L. A. Pastur. Distribution of Eigenvalues for Some Sets of Random Matrices. *Math. USSR Sb.*, 1(4):457, 1967. 10.1070/SM1967v001n04ABEH001994.

Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=I4e82CIDxv.

Amandine Marrel, Bertrand Iooss, Béatrice Laurent, and Olivier Roustant. Calculations of sobol indices for the gaussian process metamodel. *Reliability Engineering & System Safety*, 94(3):742–751, 2009. ISSN 0951-8320. https://doi.org/10.1016/j.ress.2008.07.008. URL https://www.sciencedirect.com/science/article/pii/S0951832008001981.

M. Maschler, B. Peleg, and L. S. Shapley. Geometric properties of the kernel, nucleolus, and related solution concepts. *Mathematics of Operations Research*, 4(4):303–338, 1979. ISSN 0364765X, 15265471. URL http://www.jstor.org/stable/3689220.

P. McCullagh and J.A. Nelder. *Generalized Linear Models, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1989. ISBN 9780412317606. URL https://books.google.com/books?id=h9kFH2_FfBkC.

Duncan McElfresh, Sujay Khandagale, Jonathan Valverde, Vishak Prasad C., Ganesh Ramakrishnan, Micah Goldblum, and Colin White. When do neural nets outperform boosted trees on tabular data? NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

M. D. McKay, R. J. Beckman, and W. J. Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21(2):239–245, 1979. ISSN 00401706. URL http://www.jstor.org/stable/1268522.

Hayden McTavish, Jon Donnelly, Margo Seltzer, and Cynthia Rudin. Interpretable generalized additive models for datasets with missing values. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=soUXmwL5aK.

Lukas Meier, Sara van de Geer, and Peter Bühlmann. High-dimensional additive modeling. *The Annals of Statistics*, 37(6B):3779 – 3821, 2009. 10.1214/09-AOS692. URL https://doi.org/10.1214/09-AOS692.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6f1d43d5a82a37e89b0665b33bf3a182-Paper-Conference.pdf.

Luke Merrick and Ankur Taly. The explanation game: Explaining machine learning models using shapley values, 2020. URL https://arxiv.org/abs/1909.08128.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences, 2018. URL https://arxiv.org/abs/1706.07269.

Rory Mitchell, Joshua Cooper, Eibe Frank, and Geoffrey Holmes. Sampling permutations for shapley value estimation. *Journal of Machine Learning Research*, 23(43):1–46, 2022a. URL http://jmlr.org/papers/v23/21-0439.html.

Rory Mitchell, Eibe Frank, and Geoffrey Holmes. Gputreeshap: Massively parallel exact calculation of shap scores for tree ensembles, 2022b. URL https://arxiv.org/abs/2010.13972.

Abdel-rahman Mohamed, George E Dahl, and Geoffrey Hinton. Acoustic modeling using deep belief networks. *IEEE transactions on audio, speech, and language processing*, 20(1): 14–22, 2011.

Christoph Molnar. *Interpretable Machine Learning.* 2019. https://christophm.github.io/interpretable-ml-book/.

Christoph Molnar. *Interpretable Machine Learning.* 3 edition, 2025. ISBN 978-3-911578-03-5. URL https://christophm.github.io/interpretable-ml-book.

Dov Monderer and Dov Samet. Chapter 54 variations on the shapley value. volume 3 of *Handbook of Game Theory with Economic Applications*, pages 2055–2076. Elsevier, 2002. https://doi.org/10.1016/S1574-0005(02)03017-5. URL https://www.sciencedirect.com/science/article/pii/S1574000502030175.

Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Deepdream - a code example for visualizing neural networks, 2015. URL https://research.google/blog/deepdream-a-code-example-for-visualizing-neural-networks/. Blog Post.

James N Morgan and John A Sonquist. Problems in the analysis of survey data, and a proposal. *Journal of the American statistical association*, 58(302):415–434, 1963.

Andreas Mueller, Julien Siems, Harsha Nori, David Salinas, Arber Zela, Rich Caruana, and Frank Hutter. Gamformer: Bridging tabular foundation models and interpretable machine learning, 2026. URL https://arxiv.org/abs/2410.04560.

W. James Murdoch, Peter J. Liu, and Bin Yu. Beyond word importance: Contextual decomposition to extract interactions from LSTMs. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rkRwGg-0Z.

Raymond H. Myers, AndrĂ© I. Khuri, and Walter H. Carter. Response surface methodology: 1966-1988. *Technometrics*, 31(2):137–157, 1989. ISSN 00401706. URL http://www.jstor.org/stable/1268813.

Roger B. Myerson. Graphs and cooperation in games. *Mathematics of Operations Research*, 2(3):225–229, 1977. ISSN 0364765X, 15265471. URL http://www.jstor.org/stable/3689511.

Stefan Napel. The holler-packel axiomatization of the public good index completed. *Homo Oeconomicus*, 15:513–520, 1999.

J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972. ISSN 00359238, 23972327. URL http://www.jstor.org/stable/2344614.

Whitney K. Newey. Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79(1):147–168, 1997. ISSN 0304-4076. https://doi.org/10.1016/S0304-4076(97)00011-0. URL https://www.sciencedirect.com/science/article/pii/S0304407697000110.

J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:289–337, 1933. ISSN 02643952. URL http://www.jstor.org/stable/91247.

Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/59dfa2df42d9e3d41f5b02bfc32229dd-Paper.pdf.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. 10.23915/distill.00007. https://distill.pub/2017/feature-visualization.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. 10.23915/distill.00024.001. https://distill.pub/2020/circuits/zoom-in.

Jonathan J. Oliver. Decision graphs - an extension of decision trees. 1993. URL https://api.semanticscholar.org/CorpusID:16194622.

Art B. Owen. Sobol' indices and shapley value. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):245–251, 2014. 10.1137/130936233. URL https://doi.org/10.1137/130936233.

Art B. Owen and Clémentine Prieur. On shapley value for measuring importance of dependent inputs. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):986–1002, 2017. 10.1137/16M1097717. URL https://doi.org/10.1137/16M1097717.

Guillermo Owen. Multilinear extensions of games. *Management Science*, 18(5):P64–P79, 1972. ISSN 00251909, 15265501. URL http://www.jstor.org/stable/2661445.

Guilliermo Owen. Values of games with a priori unions. In Rudolf Henn and Otto Moeschlin, editors, *Mathematical Economics and Game Theory*, pages 76–88, Berlin, Heidelberg, 1977. Springer Berlin Heidelberg. ISBN 978-3-642-45494-3.

J. Deegan Jr. E. W. Packel. A new index of power for simple n-person games. *International Journal of Game Theory*, 1978.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford infolab, 1999.

Zhen Pan, Enhong Chen, Qi Liu, Tong Xu, Haiping Ma, and Hongjie Lin. Sparse factorization machines for click-through rate prediction. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 400–409, 2016. 10.1109/ICDM.2016.0051.

Álvaro Parafita, Tomas Garriga, Axel Brando, and Francisco J. Cazorla. Practical do-shapley explanations with estimand-agnostic causal inference. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=aTiMLVePXi.

Seokhun Park, Insung Kong, yongchan Choi, Chanmoo Park, and Yongdai Kim. Tensor product neural networks for functional ANOVA model. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=Ci3nWnys6T.

L. A. Pastur. On the spectrum of random matrices. *Theoretical and Mathematical Physics*, 10(1):67–74, Jan 1972. ISSN 1573-9333. 10.1007/BF01035768. URL https://doi.org/10.1007/BF01035768.

Y.C. Pati, R. Rezaiifar, and P.S. Krishnaprasad. Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. In *Proceedings of*

27th Asilomar Conference on Signals, Systems and Computers, pages 40–44 vol.1, 1993. 10.1109/ACSSC.1993.342465.

Michael J Pazzani and Stephen D Bay. The independent sign bias: Gaining insight from multiple linear regression. In *Proceedings of the Twenty-first Annual Conference of the Cognitive Science Society*, pages 525–530. Psychology Press, 2020.

Judea Pearl. *Causality*. Cambridge University Press, 2 edition, 2009.

Karl Pearson. *The Life, Letters and Labours of Francis Galton*. Cambridge Library Collection - Darwin, Evolution and Genetics. Cambridge University Press, 1914.

Charles S. Peirce. *The 1903 Harvard lectures on pragmatism*. State University of New York Press, 1903.

L. S. Penrose. The elementary statistics of majority voting. 109(1):53 – 57, 1946. https://doi.org/10.2307/2981392.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710, 2014.

Sergei Popov, Stanislav Morozov, and Artem Babenko. Neural oblivious decision ensembles for deep learning on tabular data. *CoRR*, abs/1909.06312, 2019. URL http://arxiv.org/abs/1909.06312.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19920–19930. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/e6385d39ec9394f2f3a354d9d2b88eec-Paper.pdf.

Sanjay Purushotham, Martin Renqiang Min, C-C Jay Kuo, and Rachel Ostroff. Factorized sparse learning models with interpretable high order feature interactions. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 552–561. ACM, 2014.

Federico Quin, Danny Weyns, Matthias Galster, and Camila Costa Silva. A/b testing: A systematic literature review. *Journal of Systems and Software*, 211:112011, 2024. ISSN 0164-1212. https://doi.org/10.1016/j.jss.2024.112011. URL https://www.sciencedirect.com/science/article/pii/S0164121224000542.

Filip Radenovic, Abhimanyu Dubey, and Dhruv Mahajan. Neural basis models for interpretability. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=fpfDusqKZF.

Timothy O Ramsay, Richard T Burnett, and Daniel Krewski. The effect of concurvity in generalized additive models linking mortality to ambient particulate matter. *Epidemiology*, 14(1):18–23, 2003.

C. Radhakrishna Rao. On some problems arising out of discrimination with multiple characters. *Sankhya: The Indian Journal of Statistics (1933-1960)*, 9(4):343–366, 1949. ISSN 00364452. URL http://www.jstor.org/stable/25047988.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *J. Mach. Learn. Res.*, 13(1): 389–427, February 2012. ISSN 1532-4435.

Pradeep Ravikumar, John Lafferty, Han Liu, and Larry Wasserman. Sparse additive models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(5): 1009–1030, 10 2009. ISSN 1369-7412. 10.1111/j.1467-9868.2009.00718.x. URL https://doi.org/10.1111/j.1467-9868.2009.00718.x.

Christian H Reinsch. Smoothing by spline functions. *Numerische mathematik*, 10(3): 177–183, 1967.

Steffen Rendle. Factorization machines. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 995–1000. IEEE, 2010.

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 '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. 10.1145/2939672.2939778. URL https://doi.org/10.1145/2939672.2939778.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32 (1), Apr. 2018. 10.1609/aaai.v32i1.11491. URL https://ojs.aaai.org/index.php/AAAI/article/view/11491.

John Rice and Murray Rosenblatt. Smoothing splines: Regression, derivatives and deconvolution. *The Annals of Statistics*, 11(1):141–156, 1983. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/2240468.

R. A. Rigby and D. M. Stasinopoulos. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 54(3):507–554, 2005. ISSN 00359254, 14679876. URL http://www.jstor.org/stable/3592732.

James L. Rosenberger. Stat 503: Design of experiments, 2019. URL https://online.stat.psu.edu/stat503/. Online Course Materials.

Ron Rubinstein, Alfred M. Bruckstein, and Michael Elad. Dictionaries for sparse representation modeling. *Proceedings of the IEEE*, 98(6):1045–1057, 2010. 10.1109/JPROC.2010.2040551.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. 10.1038/s42256-019-0048-x.

David Ruegamer. Scalable higher-order tensor product spline models. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 1–9. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/ruegamer24a.html.

Jerome Sacks, William J. Welch, Toby J. Mitchell, and Henry P. Wynn. Design and analysis of computer experiments. *Statistical Science*, 4(4):409–423, 1989. ISSN 08834237, 21688745. URL http://www.jstor.org/stable/2245858.

Andrea Saltelli, K. Chan, and E. M. Scott. *Sensitivity Analysis*. 2000.

Thomas J. Santner, Brian J. Williams, and William I. Notz. *The Design and Analysis of Computer Experiments*. 2003.

Robert E Schapire. The strength of weak learnability. *Machine learning*, 5(2):197–227, 1990.

Robert E Schapire and Yoram Singer. Improved boosting algorithms using confidence-rated predictions. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 80–91, 1998.

Henry Scheffe. Experiments with mixtures. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):344–360, 12 1958. ISSN 0035-9246. 10.1111/j.2517-6161.1958.tb00299.x. URL https://doi.org/10.1111/j.2517-6161.1958.tb00299.x.

David Schmeidler. The nucleolus of a characteristic function game. *SIAM Journal on Applied Mathematics*, 17(6):1163–1170, 1969. 10.1137/0117107. URL https://doi.org/10.1137/0117107.

Isaac J Schoenberg. Spline functions and the problem of graduation. *Proceedings of the National Academy of Sciences*, 52(4):947–950, 1964.

Ludwig Schubert, Michael Petrov, Shan Carter, Nick Cammarata, Gabriel Goh, and Chris Olah. Openai microscope, 2020. Blog Post.

Rickmer Schulte and David Rügamer. Additive model boosting: New insights and path(ologie)s. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025. URL https://openreview.net/forum?id=p4CHBlYxYj.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 10.1109/ICCV.2017.74.

B. V. Shah. Balanced factorial experiments. *The Annals of Mathematical Statistics*, 31(2): 502–514, 1960. ISSN 00034851. URL http://www.jstor.org/stable/2237970.

L. S. Shapley. *A Value for n-Person Games*, volume 2, pages 307–318. Princeton University Press, Princeton, 1953. ISBN 9781400881970. doi:10.1515/9781400881970-018. URL https://doi.org/10.1515/9781400881970-018.

L. S. Shapley and Martin Shubik. A method for evaluating the distribution of power in a committee system. 48(3), 1954.

Lloyd S Shapley. On balanced sets and cores. Technical report, The RAND Corporation, 1965.

Galit Shmueli. To Explain or to Predict? *Statistical Science*, 25(3):289 – 310, 2010. 10.1214/10-STS330. URL https://doi.org/10.1214/10-STS330.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3145–3153. JMLR.org, 2017.

Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022. ISSN 1566-2535. https://doi.org/10.1016/j.inffus.2021.11.011. URL https://www.sciencedirect.com/science/article/pii/S1566253521002360.

Julien Niklas Siems, Konstantin Ditschuneit, Winfried Ripken, Alma Lindborg, Maximilian Schambach, Johannes Otterbach, and Martin Genzel. Curve your enthusiasm: Concurvity regularization in differentiable generalized additive models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=TAIYBdRb3C.

B. W. Silverman. Spline smoothing: The equivalent variable kernel method. *The Annals of Statistics*, 12(3):898–916, 1984. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/2240968.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations*, 2014.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise, 2017. URL https://arxiv.org/abs/1706.03825.

I. M. Sobol'. On sensitivity estimation for nonlinear mathematical models. *Matematicheskoe modelirovanie*, 1990. URL http://mi.mathnet.ru/mm2320.

Weiping Song, Chence Shi, Zhiping Xiao, Zhijian Duan, Yewen Xu, Ming Zhang, and Jian Tang. Autoint: Automatic feature interaction learning via self-attentive neural networks. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 1161–1170, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450369763. 10.1145/3357384.3357925. URL https://doi.org/10.1145/3357384.3357925.

Daria Sorokina, Rich Caruana, Mirek Riedewald, and Daniel Fink. Detecting statistical interactions with additive groves of trees. In *Proceedings of the 25th international conference on Machine learning*, pages 1000–1007. ACM, 2008.

Charles J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/2240707.

Charles J Stone. Additive regression and other nonparametric models. *The annals of Statistics*, pages 689–705, 1985.

Charles J. Stone. The dimensionality reduction principle for generalized additive models. *The Annals of Statistics*, 14(2):590–606, 1986. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/2241237.

Charles J. Stone. The use of polynomial splines and their tensor products in multivariate function estimation. *The Annals of Statistics*, 22(1):118–171, 1994. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/2242446.

E. Strumbelj, I. Kononenko, and M. Robnik Šikonja. Explaining instance classifications with interactions of subsets of feature values. *Data & Knowledge Engineering*, 68(10):886–904, 2009. ISSN 0169-023X. https://doi.org/10.1016/j.datak.2009.01.004. URL https://www.sciencedirect.com/science/article/pii/S0169023X09000056.

Erik Strumbelj and Igor Kononenko. An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research*, 11(1):1–18, 2010. URL http://jmlr.org/papers/v11/strumbelj10a.html.

Erik Strumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 2014. 10.1007/s10115-013-0679-x.

Mahito Sugiyama and Karsten Borgwardt. Finding statistically significant interactions between continuous features. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 3490–3498. International Joint Conferences on Artificial Intelligence Organization, 7 2019. 10.24963/ijcai.2019/484. URL https://doi.org/10.24963/ijcai.2019/484.

Xingzhi Sun, Ziyu Wang, Rui Ding, Shi Han, and Dongmei Zhang. puregam: Learning an inherently pure additive model. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 1728–1738, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393850. 10.1145/3534678.3539256. URL https://doi.org/10.1145/3534678.3539256.

Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9269–9278. PMLR, 13–18 Jul 2020.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3319–3328. JMLR. org, 2017.

Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. The shapley taylor interaction index. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9259–9268. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/sundararajan20a.html.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015. 10.1109/CVPR.2015.7298594.

Genichi Taguchi. *Introduction to Quality Engineering: Designing Quality into Products and Processes*. Asian Productivity Organization, 1986.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 00359246. URL http://www.jstor.org/stable/2346178.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method, 2000. URL https://arxiv.org/abs/physics/0004057.

Che-Ping Tsai, Chih-Kuan Yeh, and Pradeep Ravikumar. Faith-shap: The faithful shapley interaction index. *Journal of Machine Learning Research*, 24(94):1–42, 2023. URL http://jmlr.org/papers/v24/22-0202.html.

Michael Tsang, Dehua Cheng, and Yan Liu. Detecting statistical interactions from neural network weights. In *International Conference on Learning Representations*, 2018a. URL https://openreview.net/forum?id=ByOfBggRZ.

Michael Tsang, Hanpeng Liu, Sanjay Purushotham, Pavankumar Murali, and Yan Liu. Neural interaction transparency (nit): Disentangling learned interactions for improved interpretability. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018b. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/74378afe5e8b20910cf1f939e57f0480-Paper.pdf.

Michael Tsang, Dehua Cheng, Hanpeng Liu, Xue Feng, Eric Zhou, and Yan Liu. Feature interaction interpretability: A case for explaining ad-recommendation systems via neural interaction detection. In *International Conference on Learning Representations*, 2020a. URL https://openreview.net/forum?id=BkgnhTEtDS.

Michael Tsang, Sirisha Rambhatla, and Yan Liu. How does this interaction affect me? interpretable attribution for feature interactions. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6147–6159, 2020b. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/443dec3062d0286986e21dc0631734c9-Paper.pdf.

Michael Tsang, James Enouen, and Yan Liu. Interpretable artificial intelligence through the lens of feature interaction, 2021. URL https://arxiv.org/abs/2103.03103.

John W Tukey. One degree of freedom for non-additivity. *Biometrics*, 5(3):232–242, 1949.

Hemant Tyagi, Anastasios Kyrillidis, Bernd Gärtner, and Andreas Krause. Learning sparse additive models with interactions in high dimensions. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 111–120, Cadiz, Spain, 09–11 May 2016. PMLR. URL https://proceedings.mlr.press/v51/tyagi16.html.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

John von Neumann, Oskar Morgenstern, and Ariel Rubinstein. *Theory of Games and Economic Behavior: 60th Anniversary Commemorative Edition.* Princeton University Press, 1944. ISBN 9780691119939. URL http://www.jstor.org/stable/j.ctt1r2gkx.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr, 2017. URL https://arxiv.org/abs/1711.00399.

G. Wahba and S. Wold. A completely automatic french curve: fitting spline functions by cross validation. *Communications in Statistics-Theory and Methods*, 4(1):1–17, 1975.

Grace Wahba. Partial and interaction spline models for the semiparametric estimation of functions of several variables. In *Colorado State Univ., Computer Science and Statistics. Proceedings of the 18th Symposium on the Interface*, number 18, 1986.

Grace Wahba. *Spline Models for Observational Data*. SIAM, 1990.

Grace Wahba, Yuedong Wang, Chong Gu, Ronald Klein, and Barbara Klein. The 1994 neyman memorial lecture: Smoothing spline anova for exponential families, with application to the wisconsin epidemiological study of diabetic retinopathy. *The Annals of Statistics*, 23(6):1865–1895, 1995. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/2242776.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

Jiachen T. Wang and Ruoxi Jia. Data banzhaf: A robust data valuation framework for machine learning. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 6388–6421. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/wang23e.html.

Jiachen T. Wang, Prateek Mittal, and Ruoxi Jia. Efficient data Shapley for weighted nearest neighbor algorithms. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 2557–2565. PMLR, 02–04 May 2024a.

Jiachen T. Wang, Tianji Yang, James Zou, Yongchan Kwon, and Ruoxi Jia. Rethinking data shapley for data selection tasks: misleads and merits. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024b.

Jiachen T. Wang, Prateek Mittal, Dawn Song, and Ruoxi Jia. Data shapley in one training run. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=HD6bWcj87Y.

David Watson, Limor Gultchin, Ankur Taly, and Luciano Floridi. Local explanations via necessity and sufficiency: Unifying theory and practice, 2022. URL https://link.springer.com/article/10.1007/s11023-022-09598-7.

Robert James Weber. *Probabilistic values for games*, page 101–120. Cambridge University Press, 1988.

Zhikun Wei, Xin Wang, and Wenwu Zhu. Autoias: Automatic integrated architecture searcher for click-trough rate prediction. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 2101–2110, New

York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384469. 10.1145/3459637.3482234. URL https://doi.org/10.1145/3459637.3482234.

Peter Whittle. *Hypothesis testing in time series analysis*. PhD thesis, 1951.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China, November 2019. Association for Computational Linguistics. 10.18653/v1/D19-1002. URL https://aclanthology.org/D19-1002/.

Eugene P. Wigner. Characteristic vectors of bordered matrices with infinite dimensions. *Annals of Mathematics*, 62(3):548–564, 1955. ISSN 0003486X, 19398980. URL http://www.jstor.org/stable/1970079.

Eugene P. Wigner. On the distribution of the roots of certain symmetric matrices. *Annals of Mathematics*, 67(2):325–327, 1958. ISSN 0003486X, 19398980. URL http://www.jstor.org/stable/1970008.

Martin B Wilk. The randomization analysis of a generalized randomized block design. *Biometrika*, 42(1/2):70–79, 1955.

E. Winter. A value for cooperative games with levels structure of cooperation. *International Journal of Game Theory*, pages 227–240, 1989. URL https://doi.org/10.1007/BF01268161.

Eyal Winter. Chapter 53 the shapley value. volume 3 of *Handbook of Game Theory with Economic Applications*, pages 2025–2054. Elsevier, 2002. https://doi.org/10.1016/S1574-0005(02)03016-3. URL https://www.sciencedirect.com/science/article/pii/S1574000502030163.

S. Wold, A. Ruhe, H. Wold, and W. J. Dunn, III. The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, 5(3):735–743, 1984. 10.1137/0905052.

Svante Wold. Spline functions in data analysis. *Technometrics*, 16(1):1–11, 1974. ISSN 00401706. URL http://www.jstor.org/stable/1267485.

Simon N. Wood. Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):95–114, 2003. https://doi.org/10.1111/1467-9868.00374. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9868.00374.

Simon N. Wood. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99(467):673–686, 2004. ISSN 01621459. URL http://www.jstor.org/stable/27590439.

Simon N Wood. *Generalized Additive Models: An Introduction with R*. Texts in Statistical Science. Chapman & Hall/CRC, Philadelphia, PA, February 2006a.

Simon N. Wood. Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics*, 62(4):1025–1036, 2006b. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/4124523.

Workshop on Actionable Interpretability @ ICML 2025, 2025. Conference workshop on the topic.

Rui Xin, Chudi Zhong, Zhi Chen, Takuya Takagi, Margo Seltzer, and Cynthia Rudin. Exploring the whole rashomon set of sparse decision trees. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 14071–14084. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/5afaa8b4dd18eb1eed055d2d821b58ae-Paper-Conference.pdf.

Jianpeng Xu, Kaixiang Lin, Pang-Ning Tan, and Jiayu Zhou. *Synergies that Matter: Efficient Interaction Selection via Sparse Factorization Machine*, pages 108–116. 2016. 10.1137/1.9781611974348.13. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611974348.13.

Shiyun Xu, Zhiqi Bu, Pratik Chaudhari, and Ian J. Barnett. Sparse neural additive model: Interpretable deep learning withnbsp;feature selection vianbsp;group sparsity. In *Machine Learning and Knowledge Discovery in Databases: Research Track: European Conference, ECML PKDD 2023, Turin, Italy, September 18–22, 2023, Proceedings, Part III*, page 343–359, Berlin, Heidelberg, 2023. Springer-Verlag. ISBN 978-3-031-43417-4. 10.1007/978-3-031-43418-1$_2$1.$URL.$

Tom Yan and Ariel D. Procaccia. If you like shapley then you'll love the core. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(6):5751–5759, May 2021. 10.1609/aaai.v35i6.16721. URL https://ojs.aaai.org/index.php/AAAI/article/view/16721.

Jilei Yang. Fast treeshap: Accelerating shap value computation for trees, 2022. URL https://arxiv.org/abs/2109.09847.

Zebin Yang, Aijun Zhang, and Agus Sudjianto. Gami-net: An explainable neural network based on generalized additive models with structured interactions. *Pattern Recognition*, 120:108192, 2021. ISSN 0031-3203. https://doi.org/10.1016/j.patcog.2021.108192. URL https://www.sciencedirect.com/science/article/pii/S0031320321003484.

F. Yates. Complex experiments. *Supplement to the Journal of the Royal Statistical Society*, 2(2):181–247, 1935. ISSN 14666162. URL http://www.jstor.org/stable/2983638.

Chih-Kuan Yeh, Joon Kim, Ian En-Hsu Yen, and Pradeep K Ravikumar. Representer point selection for explaining deep neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/8a7129b8f3edd95b7d969dfc2c8e9d9d-Paper.pdf.

Chih-Kuan Yeh, Kuan-Yun Lee, Frederick Liu, and Pradeep Ravikumar. Threading the needle of on and off-manifold value functions for shapley explanations. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 1485–1502. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/yeh22a.html.

Weiqiu You, Helen Qu, Marco Gatti, Bhuvnesh Jain, and Eric Wong. Sum-of-parts: Self-attributing neural networks with end-to-end learning of feature groups. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=r6y9TEdLMh.

Ming Yuan, V. Roshan Joseph, and Yi Lin. An efficient variable selection approach for analyzing designed experiments. *Technometrics*, 49(4):430–439, 2007. 10.1198/004017007000000173. URL https://doi.org/10.1198/004017007000000173.

Ming Yuan, V. Roshan Joseph, and Hui Zou. Structured variable selection and estimation. *The Annals of Applied Statistics*, 3(4):1738 – 1757, 2009. 10.1214/09-AOAS254. URL https://doi.org/10.1214/09-AOAS254.

Hao Helen Zhang, Grace Wahba, Yi Lin, Meta Voelker, Michael Ferris, Ronald Klein, and Barbara Klein. Variable selection and model building via likelihood basis pursuit. *Journal of the American Statistical Association*, 99(467):659–672, 2004. 10.1198/016214504000000593. URL https://doi.org/10.1198/016214504000000593.

Xinyu Zhang, Julien Martinelli, and ST John. Challenges in interpretability of additive models. *arXiv preprint arXiv:2504.10169*, 2025.

Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5): 726–742, 2021. 10.1109/TETCI.2021.3100641.

Huan Zhao, Quanming Yao, Jianda Li, Yangqiu Song, and Dik Lun Lee. Meta-graph based recommendation fusion over heterogeneous information networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 635–644, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450348874. 10.1145/3097983.3098063. URL https://doi.org/10.1145/3097983.3098063.

Peng Zhao, Guilherme Rocha, and Bin Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37(6A):3468 – 3497, 2009. 10.1214/07-AOS584. URL https://doi.org/10.1214/07-AOS584.

Chudi Zhong, Zhi Chen, Jiachang Liu, Margo Seltzer, and Cynthia Rudin. Exploring and interacting with the set of good sparse generalized additive models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=CzAAbKOHQW.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 03 2005. ISSN 1369-7412. 10.1111/j.1467-9868.2005.00503.x. URL https://doi.org/10.1111/j.1467-9868.2005.00503.x.