ANNUAL REVIEWS

Annual Review of Statistics and Its Application

Interpretable Machine Learning for Discovery: Statistical Challenges and Opportunities

Genevera I. Allen,[1,2,3,4] Luqin Gan,[2] and Lili Zheng[1]

[1]Department of Electrical and Computer Engineering, Rice University, Houston, Texas, USA; email: gallen@rice.edu

[2]Department of Statistics, Rice University, Houston, Texas, USA

[3]Department of Computer Science, Rice University, Houston, Texas, USA

[4]Neurological Research Institute, Baylor College of Medicine, Houston, Texas, USA

Annu. Rev. Stat. Appl. 2024. 11:97–121

First published as a Review in Advance on November 17, 2023

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-040120-030919

Keywords

machine learning, interpretability, explainability, data-driven discoveries, validation, stability, selection consistency, uncertainty quantification

Abstract

New technologies have led to vast troves of large and complex data sets across many scientific domains and industries. People routinely use machine learning techniques not only to process, visualize, and make predictions from these big data, but also to make data-driven discoveries. These discoveries are often made using interpretable machine learning, or machine learning models and techniques that yield human-understandable insights. In this article, we discuss and review the field of interpretable machine learning, focusing especially on the techniques, as they are often employed to generate new knowledge or make discoveries from large data sets. We outline the types of discoveries that can be made using interpretable machine learning in both supervised and unsupervised settings. Additionally, we focus on the grand challenge of how to validate these discoveries in a data-driven manner, which promotes trust in machine learning systems and reproducibility in science. We discuss validation both from a practical perspective, reviewing approaches based on data-splitting and stability, as well as from a theoretical perspective, reviewing statistical results on model selection consistency and uncertainty quantification via statistical inference. Finally, we conclude by

highlighting open challenges in using interpretable machine learning techniques to make discoveries, including gaps between theory and practice for validating data-driven discoveries.

1. INTRODUCTION

Machine learning systems have gained widespread use in science, technology, and society. Given the increasing number of high-stakes machine learning applications and the growing complexity of machine learning models, many have advocated for interpretability and explainability to promote understanding of and trust in machine learning results (Rasheed et al. 2022, Toreini et al. 2020, Broderick et al. 2023). In response, there has been a recent explosion of research on interpretable machine learning (IML), mostly focusing on new techniques to interpret black-box systems; for recent reviews of the IML and explainable artificial intelligence literature, readers are directed to Doshi-Velez & Kim (2017), Guidotti et al. (2018), Lipton (2018), Carvalho et al. (2019), Du et al. (2019), Murdoch et al. (2019), and Molnar (2022). While most of these interpretability techniques were not necessarily designed for this purpose, they are increasingly being used to mine large and complex data sets to generate new insights (Roscher et al. 2020). These so-called data-driven discoveries are especially important to advance data-rich fields in science, technology, and medicine. While prior reviews focus mainly on IML techniques, we primarily review how IML methods promote data-driven discoveries, challenges associated with this task, and related research opportunities at the intersection of machine learning and statistics.

In the sciences and beyond, IML techniques are routinely employed to make new discoveries from large and complex data sets; to motivate our review on this topic, we highlight several examples. First, feature importance and feature selection in supervised learning are popular forms of interpretation that have led to major breakthroughs like discovering new genomic biomarkers of diseases (Guyon et al. 2002), discovering physical laws governing dynamical systems (Brunton et al. 2016), and developing new techniques for finding lesions and other abnormalities in radiology (Borjali et al. 2020, Reyes et al. 2020). While most of the IML literature focuses on supervised learning (Doshi-Velez & Kim 2017, Guidotti et al. 2018, Lipton 2018, Molnar 2022), there have been many major scientific discoveries made via unsupervised techniques, and we argue that these approaches should be included in any discussion of IML. For example, one of the earliest and most important machine learning findings in medicine was the discovery of genomic subtypes of breast cancer using hierarchical clustering of gene expression data (Perou et al. 2000); this discovery led to new ways to diagnose and treat cancer based on a patient's specific genomic subtype and ushered in an era of personalized medicine (Hassan et al. 2022). Clustering techniques have also been used to discover galaxies in astronomical surveys (Materne 1978) and characterize communities with similar political affiliations (Ozer et al. 2016). Other major unsupervised discoveries include detecting major climate patterns like El Niño and their localized effects via dimension reduction (Jolliffe 2002) and discovering the functional organization of the brain via network models (Rubinov & Sporns 2010). These are just a few of many examples of how IML techniques have led to new scientific discoveries. As the size and complexity of scientific data continue to grow, IML techniques will be ever more valuable for mining these data to generate new findings and advance science, hence motivating our review on this topic.

In this article, we review IML for the purpose of generating new data-driven discoveries. We also discuss several challenges that come with using IML for discovery, review statistical and other research that has sought to address these challenges, and highlight many associated open research opportunities. We organize this article by first reviewing the extensive IML literature in Section 2. Next, in Section 3, we review IML techniques, but instead of organizing this according

to technique type as in most other IML reviews, we discuss IML techniques as they are used to generate different types of discoveries. Our discussion includes both supervised and unsupervised techniques, given the importance of the latter for generating new knowledge. In order for IML discoveries to lead to accurate findings, however, we need them to be replicable and reliable (Yu & Kumbier 2020), which also promotes trust in machine learning results (Toreini et al. 2020, Broderick et al. 2023). In other words, we need approaches to validate IML discoveries. But unfortunately, validation for IML is not widely discussed or applied in practice as it presents many more challenges than validating machine learning predictions. In Section 4, we discuss the grand challenge of validating IML discoveries and review several practical validation strategies with examples. Then, in Section 5, we approach validation from a theoretical perspective and review statistical theory and statistical inference approaches that can help determine when IML techniques will recover the desired finding with high probability (Section 5.1) and help quantify the uncertainty in IML discoveries via confidence intervals and statistical hypothesis testing (Section 5.2). We also give an example of such approaches in Section 5.3. We finally conclude with a discussion of the major open problems and opportunities in IML for discovery in Section 6.

2. INTERPRETABLE MACHINE LEARNING: DEFINITIONS, RATIONALE, AND CATEGORIES

Before focusing on IML for making discoveries, we review the growing literature on IML. We discuss definitions, reasons for using IML, and taxonomies that provide a systematic way to describe IML techniques. These are summarized in **Figure 1**.

2.1. What Is Interpretable Machine Learning?

Many have discussed IML, yet there is not a universally accepted consensus definition (Rudin 2014, Du et al. 2019, Murdoch et al. 2019, Barredo Arrieta et al. 2020, Roscher et al. 2020). Imprecise definitions have likely led to a lack of consensus on how to study and validate IML techniques, a major concern when these methods are used to make data-driven discoveries

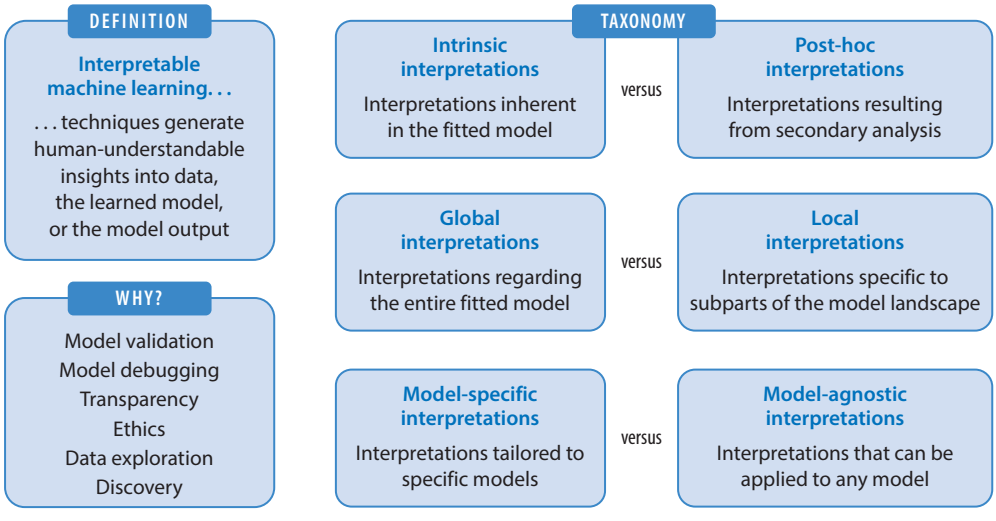

Figure 1

Definition, rationale, and taxonomies for interpretable machine learning.

(Gilpin et al. 2018, Rudin et al. 2022). We adopt a broad definition: IML is the use of machine learning techniques to generate human-understandable insights into data, the learned model, or the model output. In other words, IML is very general and provides an understanding of any aspect of the machine learning process: the model inputs (data), the model insides or model guts (the model parameters or learned model, or even how the model interacts with data), and the model outputs (predictions or decisions based on the data and model). As many have noted, what is considered a human-understandable insight depends on the intended audience and the domain area; thus, interpretations are domain, problem, and audience specific (Murdoch et al. 2019, Roscher et al. 2020).

2.2. Why Interpretability?

Why do we need interpretability in machine learning? Many have proposed a number of reasons and uses for IML (Doshi-Velez & Kim 2017, Guidotti et al. 2018, Lipton 2018, Carvalho et al. 2019, Du et al. 2019, Murdoch et al. 2019, Molnar 2022, Roscher et al. 2020), which we briefly review here.

2.2.1. Model validation and debugging. When fitting complex machine learning systems, the modeler may need to check that the model is performing and behaving in the desired manner, or perform model validation and debugging. One may ask: Does this model make sense? Is this model behaving as expected and consistently with my prior expectations or knowledge about the system? This form of human validation requires IML models.

2.2.2. Transparency, accountability, and trust. IML approaches often help to make black-box and other machine learning systems easier for humans to understand and hence more transparent. This transparency is critical for promoting accountability and trust of machine learning systems, which are necessary for their utilization in applications that have high stakes for society (Rudin 2019, Samek & Müller 2019, Xu et al. 2019).

2.2.3. Ethics. There has been an increasing focus on ensuring that machine learning algorithms are fair and ethical (Doshi-Velez & Kim 2017). Due to biases that exist in our society, machine learning algorithms that are trained on possibly biased data can often exacerbate these biases, leading to unfair predictions that are discriminatory (Guidotti et al. 2018). Understandable machine learning techniques are needed to both assess and improve the fairness of machine learning in critical societal applications.

2.2.4. Data exploration. John Tukey coined the term exploratory data analysis and promoted it as the critical first stage of data analysis (Tukey 1977). Human-interpretable techniques can help provide insights into major patterns, trends, groups, or artifacts of the data. These data exploration insights are then used to clean and prepare data for modeling, make downstream modeling decisions, and visualize and interpret model outputs (Murdoch et al. 2019, Berkhin 2006).

2.2.5. Discovery. As data sets have grown in size and complexity, we often rely on machine learning techniques to make discoveries—or, in other words, to find rare signals in a sea of data. Using IML techniques to make data-driven discoveries is the main focus of this review.

2.3. A Taxonomy of Interpretable Machine Learning Techniques

Recently, many have discussed IML techniques and proposed various categorizations to systematize the discussion and evaluation of the approaches (Doshi-Velez & Kim 2017, Guidotti

et al. 2018, Lipton 2018, Molnar 2022). While there is not complete agreement in the literature on these categories, we discuss three main dimensions or axes along which most IML techniques lie and give examples of methods falling under each designation. We also discuss how these categories of techniques relate to the task of using IML methods for generating new discoveries.

2.3.1. Intrinsic versus post hoc interpretability. A major axis that differentiates IML techniques is intrinsic versus post hoc interpretability. Intrinsic interpretations are understandings that are inherent in the fitted model itself—in other words, the user needs to simply fit a model to produce the desired interpretation. Examples include trees, additive models, or regularization approaches, which make the fitted model more understandable by adding constraints like sparsity or smoothness. More recently, researchers in the field of deep learning have proposed models that are made more intrinsically interpretable by constraining the final layer in a deep neural network to follow certain prototypes or interpretability constraints (Dong et al. 2017, Rudin 2019). In contrast, post hoc interpretations require a secondary technique to be applied to the fitted model or model outputs for the sole purpose of interpretation. Examples of post hoc interpretations include backpropagation-related methods, which traverse the learned neural network architecture to assign importance scores to each feature, and local interpretable model-agnostic explanation (LIME), which fits a second simple and interpretable model approximating the black-box model at a particular input (Molnar 2022). Additionally, most supervised model-agnostic interpretations, discussed subsequently, are post hoc in nature. Very little attention has been paid to unsupervised learning techniques in the context of IML. But we argue that all unsupervised learning techniques are naturally intrinsically interpretable, as their objective is to find some meaningful structure that helps the user gain insights into the data, and hence they fall under our definition of IML. One can still use post hoc interpretations of unsupervised findings, however. Consider that after clustering, one may perform a secondary analysis to determine which features are most responsible for separating the clusters (Satija et al. 2015).

For the purpose of making data-driven discoveries, both intrinsic and post hoc interpretations can be used as long as they accurately capture the discovery of interest. For intrinsic interpretations, this means the model must fit the data well and closely approximate the true generating model for the interpretations to reflect true discoveries. In linear regression, for example, the intrinsic interpretation of feature importance based on estimated coefficients will only be accurate if the true underlying model is linear or approximately linear. For post hoc interpretations, on the other hand, both the original model and the secondary analysis must accurately capture the data-generating process to yield accurate interpretations. If a deep learning model fits the data well, but a post hoc analysis with LIME does not sufficiently capture the original deep learning model, then interpretations and resulting discoveries will not be accurate (Zhang et al. 2019).

2.3.2. Model-specific versus model-agnostic interpretations. Another dimension along which we can categorize IML techniques is by whether they are model-specific or model-agnostic. Model-specific interpretations are tailored to the model and cannot generalize across models. Model-agnostic interpretations can be applied to any model and interpreted in a similar manner for all models. To illustrate these, consider a popular supervised interpretation: feature importance. Model-specific approaches include coefficients in generalized linear or additive models, feature importance scores based on the mean decrease in impurity for trees, or the plethora of deep learning–specific techniques for feature attribution like backpropagation or layer-wise relevance propagation methods (Molnar 2022). On the other hand, there are several model-agnostic feature importance methods that can be used for any supervised model; these include Shapley values,

feature permutations, feature occlusion, and LIME (Molnar 2022). Note that model-specific interpretations are not necessarily intrinsic interpretations; consider that feature importance scores for trees and guided backpropagation feature attribution are both model-specific but post hoc. In contrast, most model-agnostic interpretations are post hoc in nature.

For the task of making data-driven discoveries, there are several advantages to model-agnostic interpretations. Importantly, model-agnostic interpretations can be understood in the same way across all models. This is particularly useful for comparing models and validating discoveries by checking if interpretations are the same across many models. However, it is typically easier to study model-specific interpretations theoretically to understand under what conditions the resulting discoveries accurately recover some aspect of the true model, a topic we discuss in Section 5.1. Model-specific interpretations are also often more conducive to uncertainty quantification via statistical inference, discussed in Section 5.2, although many recent approaches have been developed for model-agnostic methods as well.

2.3.3. Global versus local interpretations. A final major dimension along which we can categorize IML techniques is based on whether the approach offers a local or a global interpretation. Global interpretations reveal the overall structure of the fitted model. In contrast, local interpretations only yield model insights based on subparts of the model input space; these could include local interpretations about a single observation or a subset of the domain. To make these distinctions concrete, again consider the example of feature importance in supervised learning. Here, methods previously mentioned like coefficients in linear or additive models, tree-based feature importance, and backpropagation-based feature attribution are all global interpretations that capture the relevance of each feature for all model predictions. In contrast, methods like LIME and saliency maps highlight the important features of a single new test instance or observation (Ribeiro et al. 2016, Molnar 2022). Similarly, in unsupervised learning, consider the task of dimension reduction. Methods like principal component analysis (PCA) and spectral embedding yield global interpretations, revealing global patterns represented in all observations in each of the factors. In contrast, local embedding and neighborhood embedding methods, like t-SNE (t-distributed stochastic neighbor embedding) (Van der Maaten & Hinton 2008) and UMAP (uniform manifold approximation and projection) (McInnes et al. 2020), highlight local interpretations through patterns and relationships among particular neighborhoods.

When using IML to make discoveries, global interpretations are more commonly employed, as they reveal discoveries reflective of all the input data and the model landscape. Yet, local interpretations are increasingly important to make discoveries among subgroups of observations. Consider applications in healthcare, where saliency maps are used in radiology to discover abnormalities in images of individual patients (Yasaka & Abe 2018), or in precision medicine, where one seeks to discover important genomic biomarkers for each patient or subgroups of similar patients (Hassan et al. 2022).

3. TYPES OF INTERPRETABLE MACHINE LEARNING DISCOVERIES AND TECHNIQUES

Recent research in IML has produced an abundance of interpretability techniques, as thoroughly reviewed by Molnar (2022). But these works focus on the types of techniques and not the types of data-driven discoveries that various techniques can make. We organize this section to highlight the major types of discoveries achieved through interpretations of machine learning models. Importantly, and distinct from the IML literature, we place great emphasis on unsupervised techniques, which are popularly used throughout the sciences to make discoveries from unlabeled data (see **Figure 2** for an overview).

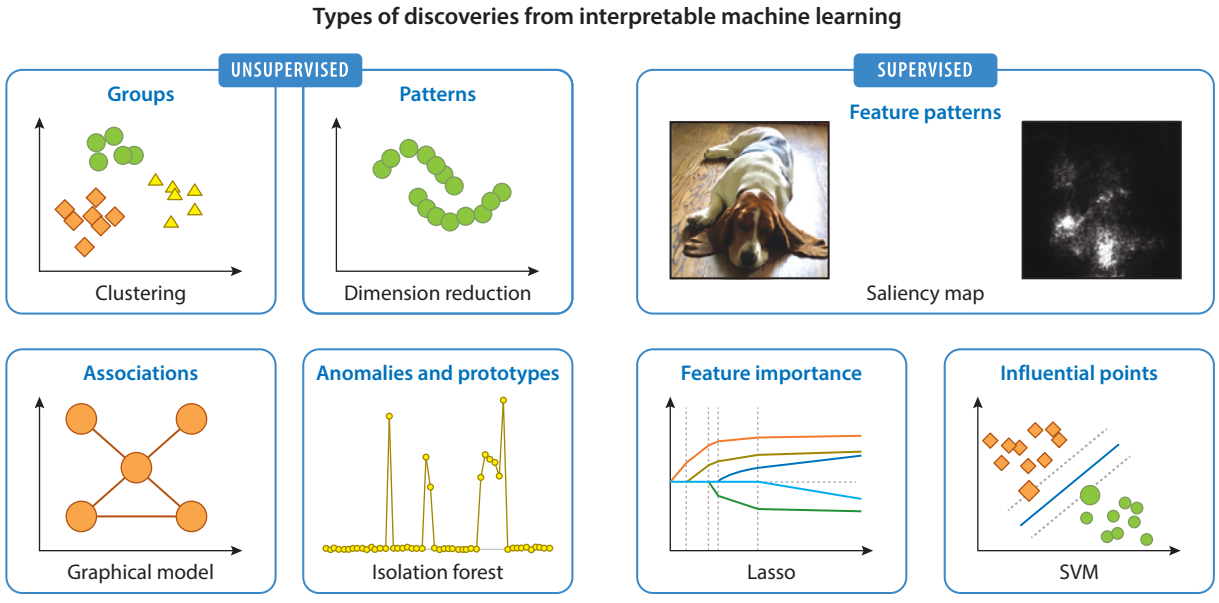

Figure 2

Overview of the broad types of unsupervised and supervised data-driven discoveries that can be made using interpretable machine learning techniques, with some simple graphic examples.

3.1. Unsupervised Discoveries

Most work on IML has focused on supervised models and interpreting the results of predictive systems (Rudin et al. 2022). In scientific domains, however, some of the most widespread uses of machine learning are in unsupervised settings; hence, in this section, we review major types of unsupervised discoveries and highlight which types of IML techniques are employed to generate these discoveries.

3.1.1. Groups. Uncovering hidden group structures in large data sets is a common and popular type of unsupervised discovery. There are many well-established clustering techniques used for this task, including *K*-means, hierarchical clustering, mixture modeling, and spectral clustering, among many others (Hennig et al. 2015). Beyond group membership, other types of interpretations related to clustering include uncovering groups of both observations and features simultaneously via biclustering, discovering nested group structure via hierarchical clustering, detecting localized important regions via spatial clustering, and finding a subset of features that distinguish groups of observations via sparse clustering (Witten & Tibshirani 2010). Clustering has been applied broadly and is a nearly ubiquitous technique in unsupervised and exploratory analysis; groups found via clustering have also led to several major scientific discoveries, such as finding gene expression patterns and genomic subtypes of diseases like cancer (Perou et al. 2000).

3.1.2. Patterns and trends. When conducting unsupervised analyses, a typical first task is to visualize and explore the data to look for major patterns and trends. Often, important unsupervised discoveries can be made through these visual inspections of the data. For large multivariate data, dimension reduction approaches reduce the data down to a smaller number of components that retain important structure, or patterns, in the data. There are a plethora of dimension reduction techniques, including linear approaches like PCA, nonnegative matrix factorization,

and independent component analysis, or nonlinear approaches like spectral embedding, multidimensional scaling, isomap, t-SNE, UMAP, or autoencoders; Fodor (2002) provides a review of such approaches. Each of these approaches is optimized to find slightly different types of patterns. For example, PCA finds variance-maximizing patterns that preserve the global structure, whereas t-SNE finds localized patterns that preserve neighborhood and group structure.

3.1.3. Associations. Discovering associations, or important relationships among features, is another widely used type of unsupervised discovery. Most techniques typically find linear or nonlinear associations by exploring all possible pairwise interactions among features, using correlation, mutual information, or other such metrics. Recently, there has been a surge of interest in exploring feature relationships using graphical models (Lauritzen 1996). In Markov networks, or undirected graphical models, for example, the goal of structural learning is to estimate conditional dependencies between features; structural learning in Bayesian networks or directed acyclic graphs (DAGs) seeks to learn directed relationships and is an important part of causal discovery (Drton & Maathuis 2017).

3.1.4. Anomalies and prototypes. Other types of unsupervised discoveries that are perhaps less commonly used are finding anomalies (rare entities) or prototypes (typical entities). Anomalies are rare but noteworthy observations. Techniques for anomaly detection are similar to those for outlier detection and include distance-based approaches, which often employ dimension reduction; clustering approaches like single-linkage hierarchical clustering or DBSCAN; the one-class support vector machine; or the isolation forest (Hodge & Austin 2004). Alternatively, sometimes one seeks to find the most representative observations, or prototypes. Adaptions of other unsupervised approaches, especially dimension reduction and clustering, are typically employed for this task (Bien & Tibshirani 2011).

3.2. Supervised Discoveries

Supervised learning has been the focus of the vast majority of the IML literature (Doshi-Velez & Kim 2017, Guidotti et al. 2018, Lipton 2018, Molnar 2022). This occurs as some of the best-performing predictive models, such as deep learning and tree-based ensembles, are essentially black boxes that are not intrinsically interpretable and are difficult to decipher. Thus, interpretations of these predictive models are critical for generating new insights and making data-driven discoveries.

3.2.1. Feature importance and feature selection. Perhaps the most common and popular form of interpretation in supervised models is understanding how each feature influences a model's predictions, often referred to as feature importance. Related to this is feature selection, which finds the best subset of features that maximize predictive accuracy. Importantly, feature importance and feature selection in supervised learning offer a form of multivariate or conditional feature interpretation: Given all other features in the model, what is the added benefit of including a particular feature? This conditional feature interpretation is much stronger than marginally assessing how each feature relates to an outcome and has been used extensively to discover important features.

Let us review the many types of methods for interpreting features in supervised learning through the context of our IML taxonomies from Section 2.3. First, consider global and model-specific feature importance metrics. For linear or generalized additive models, the feature weight (or parameter or coefficient) can be directly interpreted as the conditional feature importance, offering intrinsic feature interpretability. Tree-based ensembles offer post hoc interpretability by the feature importance scores based on the decrease in impurity for each split. In deep learning, post hoc approaches are popular and include several feature importance scores calculated via

gradient-based methods, which traverse the fitted neural network to attribute relevance to each input feature (Samek et al. 2021). Local, post hoc, and model-specific methods are popular in computer vision, where measures such as saliency maps and Grad-CAM (gradient-weighted class activation mapping) highlight which pixels in a specific image were used to generate the predicted label (Samek et al. 2021). There are also several model-agnostic metrics that can be used with any supervised learning model, including methods that yield global interpretations, like feature occlusion, feature permutation, and Shapley values, as well as local interpretations such as LIME (for more information on these methods, see Molnar 2022). There is an equally impressive literature on feature selection for supervised learning; most of these strategies offer intrinsic and model-specific interpretations by working to minimize the empirical risk or loss function. As finding the best subset of features is a combinatorially hard optimization problem, people typically turn to greedy step-wise methods, like recursive feature elimination, or regularization strategies that relax the best subset constraint. Popular approaches to the latter include the ℓ_1 or the lasso (least absolute shrinkage and selection operator) penalty that encourages sparsity in the feature weights (Tibshirani 1996). The lasso and other regularization approaches are routinely employed across all areas of machine learning to aid in interpreting features (Li et al. 2022).

3.2.2. Feature interactions and feature representations. Beyond the importance of each individual feature, one may want to understand higher-order interactions or feature patterns that are important for a model's predictions. Decision trees and their extensions offer natural ways of assessing model-specific feature interactions, but there are several model-agnostic approaches, such as Friedman's H-statistic, variable interaction networks, and partial dependence functions (for further details, see Molnar 2022). Going beyond pairwise feature interactions, many researchers are interested in understanding how more complex, higher-order, and nonlinear feature patterns contribute to a model's predictions. This growing area is often called representation learning and utilizes deep learning models like transformers to encode complex feature relationships in an often lower-dimensional representation space (Bengio et al. 2013). While many of these feature representations are not directly interpretable, learning interpretable feature representations is an active area of research, especially in computer vision (Bengio et al. 2013).

3.2.3. Influential points. We have discussed interpretations of features in supervised models, but one can also interpret the observations through influential points, defined as observations whose removal significantly changes a model's prediction. There are a few model-specific approaches that provide intrinsic interpretations of influential points, like support vector machines, but most use model-agnostic strategies to identify these points. Coming from classical statistics, one can use strategies to detect outliers as well as measure the effect of removing each single training point (Hodge & Austin 2004). But, more recently in machine learning, many have proposed using the influence function to approximate parameter changes for individual points based on the change in the gradient; these approaches have found widespread application in deep learning models (Koh & Liang 2017).

4. VALIDATING INTERPRETABLE MACHINE LEARNING DISCOVERIES

IML techniques are being deployed across science and beyond to generate new knowledge or make data-driven discoveries. Yet, one may ask, is my discovery true? Or, have I discovered an artifact? How can I tell the difference? In other words, how can we validate discoveries made via IML? While most research in the IML community has focused on developing new interpretability techniques, there has been relatively little work on the critically important problem of validation. We contend that validation is one of the grand challenges in IML, and it is especially crucial for

making replicable, reliable, and trustworthy data-driven discoveries. In this section, we motivate the necessity of validation for IML, discuss why this is so challenging, and then discuss several practical approaches that can be deployed with most IML techniques to help validate discoveries; we conclude with recommendations for validating IML discoveries in practice.

4.1. Motivation and Challenges

We briefly outline why validation is so critically important as well as why it is such a formidable task.

4.1.1. Motivation: replicability, reliability, and trust. IML techniques are designed to always produce the desired interpretation, regardless of whether that interpretation or discovery truly reflects the underlying structure of the data. For example, K-means clustering always returns K clusters whether there are K groups in the data or not; feature selection always returns a subset of features whether the underlying true model is sparse or not. Then, how can we tell if the machine learning interpretation generated a true discovery or just is an artifact in the data? Furthermore, there are a plethora of IML techniques, and often each technique produces a different interpretation. Then, which interpretation is correct and represents a true discovery? These are perhaps unknowable, epistemological questions. Science addresses these issues by continually replicating and validating discoveries in follow-up studies until findings converge upon an accepted truth. Indeed, reproducibility and replicability are cornerstones of science (Natl. Acad. Sci. Eng. Med. 2019, Stodden 2020).

In machine learning, reproducibility means being able to obtain the exact same results after the same computational steps are performed on the same data, which is purely a computational concept (Willis & Stodden 2020, Fineberg et al. 2020) and is a prerequisite for validation. Replicability means being able to obtain very similar results when two independent studies are performed to answer the same scientific question; in machine learning, this could entail performing the same or similar analysis on a new data set (Fineberg et al. 2020, Meng 2020). Replicability by itself, however, cannot be the ultimate goal of scientific discoveries, as replicable results can still be wrong if the same mistakes are made in follow-up studies. Thus, going one step beyond this, many have advocated for reliability in machine learning, saying that predictions and findings should be robust to reasonable sensitivity tests, like small changes in the data or the model, out-of-sample prediction tests, and consistency with domain knowledge (Meng 2020). Validation for machine learning directly seeks to assess the replicability and reliability of results. For predictive tasks, there are well-developed and routinely employed validation strategies like data-splitting and cross-validation. For IML, however, there are very few widely accepted validation strategies, and most employ IML techniques without any validation whatsoever in practice. For some uses of IML, this practice might not be terrible, but for the task of generating new discoveries, lack of validation is extremely damaging and could lead to erroneous, irreplicable, and unreliable findings. Indeed, there has been much commentary over the past several decades about a reproducibility and replicability crisis in science (Baker 2016). Recently, several have suggested that failures to validate machine learning findings could be contributing to this crisis (Beam et al. 2020, McDermott et al. 2021, Gibney 2022). Validation is a crucial component of IML for generating data-driven discoveries.

Beyond just the goal of utilizing best practices in science, replicability and reliability are critical to promote trust in machine learning results. Many have lamented a lack of trust in machine learning systems and recommended promoting trust and societal acceptance of machine learning results by generating understandable interpretations (Toreini et al. 2020, Jacovi et al. 2021). But, can we trust the machine learning interpretations? If interpretations are not replicable and reliable, then trust in these interpretations and discoveries breaks down. Recently, Broderick et al. (2023)

discussed these issues, and others, that cause trust to break down in probabilistic machine learning. Furthermore, they and several others have proposed various ways to enhance trust in machine learning results, emphasizing the need for validation strategies (Toreini et al. 2020, Rasheed et al. 2022, Broderick et al. 2023).

4.1.2. Challenges. To better understand why validating machine learning interpretations and their data-driven discoveries is so challenging, let us first discuss why a discovery might fail to validate. First, the machine learning model could be a poor fit to the data, and hence any resulting interpretations would poorly reflect the signal in the data. Next, even if the model fits the data well, the interpretation approach could be a poor fit for the model, resulting in problematic interpretations; this can especially be the case with some post hoc interpretability methods that fit a second model to generate the interpretation (e.g., LIME) (Molnar 2022). In addition, there could be a mismatch between the employed interpretation technique and the desired discovery task. For example, high-dimensional genomics data are known to be highly correlated, so selecting important features from these data using the lasso might fail to identify important correlated biomarkers, as the lasso is known to only select one feature out of a correlated set (Zou & Hastie 2005). Next, IML techniques are typically designed to find the desired interpretation in the data, regardless of whether that discovery truly exists in data. For example, K-means clustering will always discover K clusters. Many machine learning techniques are so powerful that they can always detect the rarest signals in large and complex data sets. For predictive tasks, we call this overfitting. We also argue that machine learning interpretations can be overfit to the training data and hence would fail to validate.

However, despite the fact that machine learning interpretations might fail to validate for a number of reasons, validation is a critical challenge that has received surprisingly little attention in the literature (Rasheed et al. 2022). For prediction tasks, on the other hand, we have well-established techniques for validation: we ensure the predictive model generalizes to new, similar data. This is achieved by randomly splitting the data into a training set, which is used for building the predictive model, and a test set, which is used for assessing the predictive accuracy of the model. Similar to predictive models, we say that a machine learning interpretation validates if the resulting data-driven discovery generalizes well to new, similar data. Given this, one may ask: Can we simply employ a training and test set to validate interpretations? We discuss this possibility subsequently, but in short, this prospect becomes much less straightforward for interpretations. First, many IML techniques are designed to make discoveries from the current (training) data, but one cannot directly apply the discovery to new data, and hence it is unclear how to assess how well it generalizes. For example, clustering training data results in cluster labels for the training data, but these do not help label the test data; similarly, many dimension reduction techniques, such as t-SNE or UMAP, find low-dimensional embeddings of the training data, but these embeddings cannot be applied to new data. Second, unlike the well-established prediction error metrics, there is no consensus on metrics for quantifying the accuracy of interpretations. Most have suggested assessing machine learning interpretations via human evaluation by laypersons or domain experts (Doshi-Velez & Kim 2017, Carvalho et al. 2019, Molnar 2022), but this does not lend itself to an objective, quantitative metric analog of prediction accuracy.

4.2. Practical Approaches for Validating Interpretations

In this section, we review two practical validation strategies that can be employed for almost any machine learning discovery. While there may be additional validation approaches for specific IML approaches, we highlight these because they are fairly general and can be applied for both supervised and unsupervised discoveries.

4.2.1. Data-splitting. As we previously discussed, randomly splitting the available data into training and test sets is the established mechanism for validating machine learning predictions. Similar strategies can sometimes help us validate machine learning interpretations, but applying them here is less straightforward than in prediction tasks. The key idea of data-splitting for IML is to use the IML technique on the training data to generate an interpretation as well as construct a predictive model based on this interpretation; then, one can evaluate the model's predictive performance on the test data. This approach leverages predictive models to help with IML validation, thereby circumventing some of the previously discussed challenges. Consider some examples. With projective dimension reduction techniques like PCA, one can learn the projection from the training set and then evaluate how the test data set differs from its projection onto these components. In clustering, one could discover clusters on the training set, develop a classification model on the training set to discriminate these clusters, and then apply this to the test set to predict cluster labels; one could then compare these predicted labels to those generated in an unsupervised manner by clustering the test set (Lange et al. 2004, Handl et al. 2005). For supervised discoveries like feature selection, one could discover important features on the training set as well as build a predictive model that only uses these features for the associated supervised learning task; then, one can evaluate the prediction error of this model on the test set. The same idea can also be applied to feature interactions, feature patterns, and other supervised discoveries.

Even though data-splitting provides a direct approach to validating IML discoveries, many open questions and challenges remain. Although we presented several examples of how this strategy can be used, it is unclear how to define an appropriate predictive model for some other machine learning interpretations, such as discovered associations and relationships between features, or anomalies and prototypes. Next, this approach generates predictions on the test set, but it is not always clear how to evaluate these predictions. For example, when selecting important features, the prediction error of a sparse model is often not comparable with that of the full model, and hence, it might be unclear whether the prediction error of the sparse model is good enough to indicate it is validated. Finally, in data-splitting, the resulting interpretation is found using only part of the data, and hence, the interpretation might change with another randomly sampled training data. This can be troubling for replicability and in science. Relatedly, some might argue that since discovery is such a challenging task, one needs to use all available data, and data-splitting reduces the amount of data available for the discovery stage.

4.2.2. Stability. Another popular strategy for directly assessing the reliability of IML is the stability principle, which seeks to identify interpretations that are stable, subject to random data perturbations. This idea was first introduced by Meinshausen & Bühlmann (2010) in the context of feature selection with the lasso. Since then, many variants of this method have also been studied in the statistical machine learning literature for feature selection (Shah & Samworth 2013), feature interactions (Basu et al. 2018), graphical models (Liu et al. 2010), PCA (Taeb et al. 2020), and clustering, where it is commonly called consensus clustering (Monti et al. 2003). Even though it has not been widely applied in other areas of machine learning, the idea of stability analysis is rather general and could be applied to any IML procedure; we summarize the approach in **Figure 3**. First, the data are repeatedly randomly perturbed through subsampling, bootstrapping, randomly adding noise, or random data thinning (Neufeld et al. 2023). Then, IML procedures are used to make a discovery on each new random data set, and discoveries with high frequency are declared stable discoveries. The core idea of stability is that discoveries that are not consistent under random data perturbations are more likely to be due to artifacts in the data or sampling noise and, hence, are not replicable and reliable. Indeed, stability analysis has received widespread attention, and many have advocated using it to validate discoveries and promote reproducibility

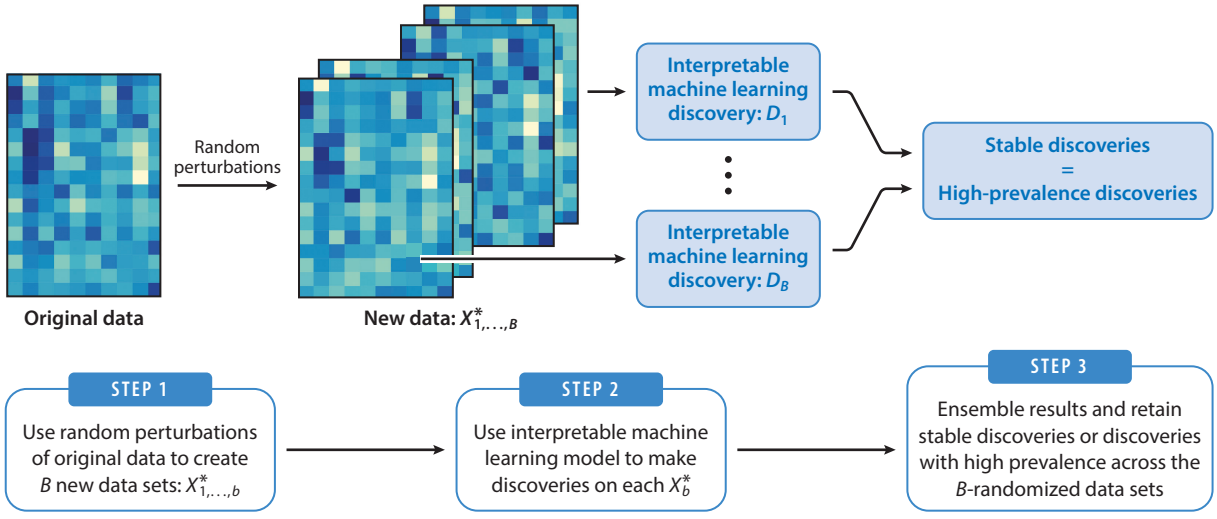

Figure 3

Illustration of stability principle for determining reliable data-driven discoveries.

in (data) science (Yu & Kumbier 2020). It has also been widely used for solving many scientific problems, such as discovering biomarkers in genomics (He & Yu 2010).

Despite the appeal of stability analysis to directly assesses reliability and overfitting, several challenges remain. First, stability analysis can be computationally burdensome, as it requires refitting the IML model many times; this is especially problematic for huge data or with complex models like deep learning. Next, it is not always clear what type of random perturbation is appropriate and what quantitative criterion should be employed to determine stable discoveries for a given IML model and discovery task. Consider consensus clustering where cocluster membership is recorded for each subsample, as there is not an easy way to record and ensemble the cluster membership. Furthermore, stability analysis could exacerbate mismatches between the interpretation techniques and the discovery task. Stability with the lasso for feature selection, for example, is known to perform very poorly with correlated features. This is due to the fact that the lasso may only select one among highly correlated features for each subsample, and hence none of these features would be deemed stable, even if they are all important. Additionally, it is unclear whether the final stable discoveries are consistent with each other, since they might not correspond to a single interpretation or set of interpretations from applying an IML model (e.g., stable features may not correspond to any lasso solution at a single regularization parameter); many might consider this a disadvantage in scientific domains. Finally, and perhaps most importantly, stability analysis only assesses one form of reliability: the robustness of the discovery to small changes in the data, but not the robustness to changes in the modeling choice. As we discussed earlier, reliability also means consistency with prior knowledge and out-of-sample predictive power, which are not reflected in stability analysis (Yu & Kumbier 2020). To summarize, stability is necessary for indicating the reliability of a discovery, but it is not sufficient.

4.2.3. Example: validating clusters. To illustrate data-splitting and stability analysis for validating IML, we turn to two real clustering examples: the Author data set, with $n = 841$ observations and $p = 69$ features measuring the stop word count of book chapters from four English-language authors (from Peng & Hengartner 2002), and The Cancer Genome Atlas Pan-Cancer (TCGA

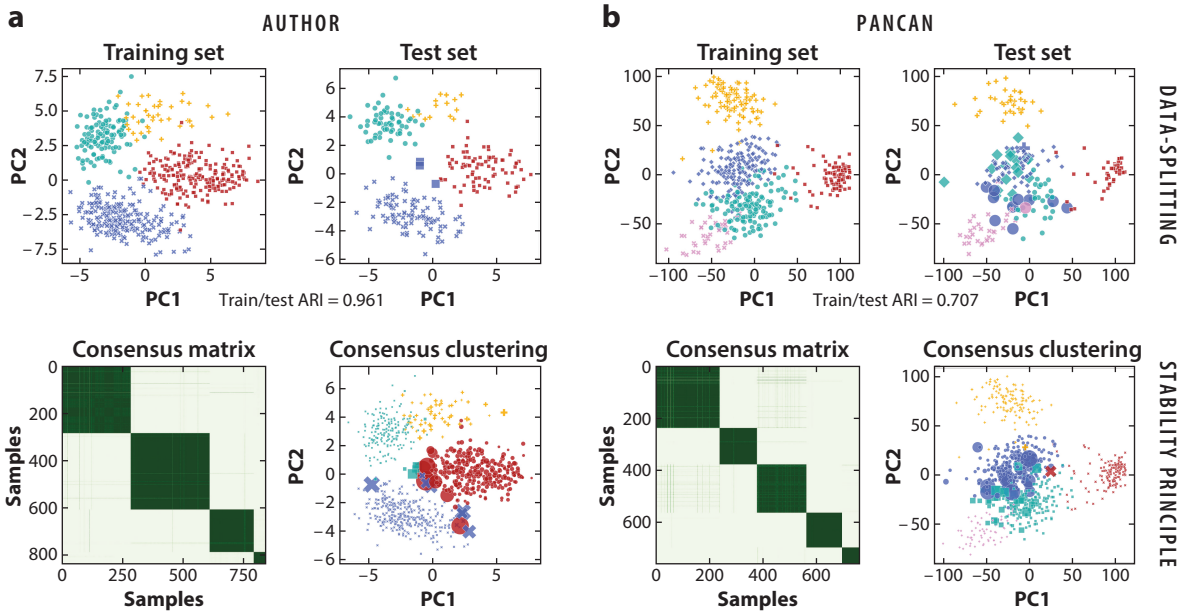

Figure 4

Example of how to validate cluster discoveries using data-splitting (*top row*) and the stability principle (*bottom row*) on two data sets. Results from the Author data set (*a*) show four well-validated clusters, whereas results from the TCGA PANCAN data set (*b*) show that the blue and teal clusters are not well separated and do not validate as well. Abbreviation: TCGA PANCAN, The Cancer Genome Atlas Pan-Cancer.

PANCAN) data set with $n = 761$ subjects and $p = 13{,}244$ genes measuring the bulk RNA-seq gene expression for patients with five different types of tumors (Weinstein et al. 2013). We apply K-means clustering with $K = 4$ and $K = 5$, respectively, and seek to validate our discovered clusters using data-splitting and stability. For data-splitting, we follow the predictive cluster validation approach outlined by Lange et al. (2004) and Handl et al. (2005) by randomly taking 70% of observations as a training set where we discover clusters as well as build a random forest classifier to predict these cluster labels. We then independently cluster the remaining 30% of observations in the test set and apply the random forest classifier to predict the labels; we measure the overlap between the predicted labels and the test set cluster labels using the adjusted Rand index (ARI), a metric between zero and one, with higher values indicating better cluster membership overlap. Results are shown in the top row of **Figure 4**, where we visualize the training and test set clusters in principal component (PC) scatterplots and highlight the test set observations where there is a mismatch between the predicted and cluster labels as larger points. In the Author data, the training set predictions and test set cluster labels have a high degree of overlap, indicating strong validation of these four clusters. Clusters in the PANCAN data set do not validate as well, as shown by the lower ARI and the confusion of the cluster labels for the blue and teal clusters.

Additionally, we apply the stability principle, which in clustering is often called consensus clustering (Monti et al. 2003), to validate these same cluster findings. Specifically, we employ repeated data-splitting by repeatedly subsampling a training set on which we discover clusters and record the cluster comembership; we then average these cluster comemberships across all data splits to yield the $n \times n$ consensus matrix taking values between zero and one, with one indicating that the two observations were always assigned to the same cluster. Heatmaps of the consensus matrix are shown in the bottom left subpanels of **Figure 4**, with darker green indicating values at or near one.

We see that in both the Author and PANCAN data sets, consensus clustering validates that there are $K = 4$ and $K = 5$ clusters, respectively, exhibiting a clear block diagonal pattern. Consensus clustering additionally allows us to inspect the uncertainty of cluster assignments for individual observations. In the PC scatterplots in **Figure 4** (bottom right subpanels), we show observations with point sizes inversely proportional to their cluster assignment uncertainty. From this, we see that clusters in the Author data set are fairly well separated, but again, the teal and the blue clusters in the PANCAN data set exhibit a high degree of confusion. Overall, both data-splitting and the stability principle can be used to validate cluster discoveries, and both of these methods reveal similar findings in the two examples we present.

4.2.4. Practical recommendations. We have discussed two general, practical validation strategies, but each has its strengths and limitations. Data-splitting can be a useful strategy for checking whether the discovered interpretation fits the data well, while stability analysis is most effective for evaluating whether the discovery is induced by random noise. Thus, we suggest that it is important to employ strategies to evaluate both the predictability of the machine learning interpretation (bias) and whether the IML technique overfits the noise (variance). Yu & Kumbier (2020) recently advocated that predictability (achieved via data-splitting) and stability are critical components of any data science process, and we argue that these are especially important in validating discoveries made using IML. We should additionally mention that after validation from a statistical perspective, final human evaluation of discoveries is also important to ensure they match the desired discovery task (Doshi-Velez & Kim 2017, Carvalho et al. 2019). Overall, validation is critically important for IML, especially for generating data-driven discoveries. This understudied area presents many opportunities for research to further develop and apply the practical validation strategies we discussed, to study their theoretical properties, to relate the approaches to rigorous notions of statistical uncertainty quantification (discussed in Section 5.2), and to determine the best validation strategies for specific IML methods, discovery types, and specific applications.

5. STATISTICAL THEORY AND INFERENCE FOR INTERPRETABLE MACHINE LEARNING DISCOVERIES

Validation of IML discoveries is critical in practice to promote replicability, reliability, and trust in data-driven discoveries. But theoretical guarantees and valid statistical inference offer different perspectives related to validation and are equally necessary to help build trust and promote replicability of IML discoveries (Rasheed et al. 2022, Broderick et al. 2023). In this section, we review statistical theoretical foundations for IML discoveries that address two pressing questions: (*a*) Under what data-generating models and under what conditions does an IML technique recover the true discovery with high probability? (*b*) What is the uncertainty in a discovery, or what discoveries can be trusted with a sufficient level of confidence? We discuss these in Sections 5.1 and 5.2, respectively.

5.1. Statistical Theory for Interpretable Machine Learning

Recently, Broderick et al. (2023) argued that theoretical guarantees were important to help build trust in machine learning. For IML discoveries, the goal is to theoretically characterize the type of data-generating models and the conditions under which IML techniques will make the desired discovery with high probability tending to one. These types of theoretical guarantees largely fall under the areas of statistical consistency and selection consistency; the latter has received a huge amount of attention in the statistical machine learning community over the past two decades (Wainwright 2019). Developing such theoretical foundations can help guide practitioners

to choose the appropriate technique for their application and desired discovery task, understand when certain IML techniques will perform well and when they will not, and perhaps inspire the development of new IML techniques with improved performance and theoretical guarantees.

Statistical consistency and selection consistency are well-studied for certain types of statistical and machine learning models but are perhaps not readily applicable to other classes of machine learning methods, leaving a gap in our theoretical understanding of IML. For example, classical statistical theory addresses the conditions under which parametric models like linear and generalized linear models consistently estimate their coefficients, a measure of intrinsic feature importance, in asymptotic and low-dimensional settings. More recently, there has been a surge of interest in studying regularized versions of these and other statistical machine learning models in finite-sample and high-dimensional settings (Bühlmann & Van de Geer 2011, Wainwright 2019). Perhaps the most widely studied has been the lasso, or ℓ_1-regularized regression, for the IML task of feature selection (Tibshirani 1996). For example, it is well established that the lasso achieves selection consistency, or correct selection of true features with high probability, under sparse linear regression models when there is sufficient sample size relative to the log number of features; when there is sufficient signal in the true features; and under conditions like the irrepresentable, restricted eigenvalue, or incoherence conditions that limit the amount of correlation between features in the model (Zhao & Yu 2006). Consistency and selection consistency have also been established for many extensions of the lasso, the lasso in classification, semiparametric models, and other sparse regularizers (Bühlmann & Van de Geer 2011). Beyond feature selection and feature importance, several other model-specific and intrinsically interpretable unsupervised IML techniques have been studied theoretically under high-dimensional regimes. These include statistical consistency guarantees for clustering under a Gaussian mixture model (Löffler et al. 2021), network clustering under the stochastic block model (Abbe 2017), low-rank estimation via PCA under spiked covariance models (Johnstone & Lu 2009), and graph selection or structural graph learning for both Markov networks (undirected graphs) and DAGs for causal discovery (Drton & Maathuis 2017); we refer the reader to Wainwright (2019) for more details on many of these recent advances in high-dimensional statistical theory.

These advances in statistical theory provide assurance and insights for certain IML discoveries and certain techniques, but many limitations of this type of theory and open questions remain. First, this statistical theory assumes a true population model that generates the data. In practice, the true data-generating process is unknown and uncheckable; it is often unclear how these IML techniques perform with misspecified models. Second, this type of theory is only applicable to model-specific and intrinsically interpretable IML techniques, which are often limited to linear or additive parametric or semiparametric models. Thus, this theory does not help us understand the performance of more flexible, nonlinear modeling strategies like tree-based ensembles and deep learning. Next, even when this type of theory is applicable to a particular model, set of techniques, and discovery task, the assumptions required to make the correct discovery with high probability are often hard to interpret and impossible to check in practice. For example, it is impossible to check the irrepresentable condition (Zhao & Yu 2006) necessary for selection consistency of the lasso for a particular data set without knowing the true features. Thus, while this theory helps us understand the properties of certain IML techniques, it is unhelpful for trying to assess the validity of a particular discovery made by an IML technique on a particular data set. Finally, statistical theory is currently very limited for interpretations of tree-based ensembles like random forests and boosting, neural networks, and deep learning, and model-agnostic interpretations like Shapley values for feature importance. Such areas provide many open research opportunities that would help us better understand these popular IML approaches and further promote trust in their discoveries (Broderick et al. 2023).

5.2. Statistical Inference for Interpretable Machine Learning

While statistical theory highlights the assumptions required to make an accurate discovery with high probability, another approach to validating discoveries is through statistical inference, which quantifies the uncertainty associated with the discovery. Uncertainty quantification, typically through confidence intervals and hypothesis testing, is crucial in discerning whether a discovered pattern is due to random chance or is a genuine discovery. This is especially important in high-stakes applications of IML where making decisions based on discoveries with a high degree of uncertainty could have devastating consequences; in science, this could lead to wasted resources and irreplicable results. While uncertainty quantification for IML is a critically important task, it presents many challenges. Note that the statistical theory discussed previously (Section 5.1) also quantifies errors for a discovery, but these cannot readily be used for uncertainty quantification as they depend on unknown parameters. Similarly, practical validation approaches like data-splitting and stability, discussed in Section 4, give a sense of the uncertainty in a discovery but cannot always be translated into rigorous statistical uncertainty quantification. Nonetheless, uncertainty quantification has been studied for many of the same statistical machine learning models and IML tasks for which statistical theory has been developed, and more recently, uncertainty quantification has been considered in model-agnostic settings for specific IML tasks like feature importance and feature selection. We briefly review these approaches.

Most statistical inference procedures are designed for model-specific, global, and intrinsically interpretable statistical models that cover only a narrow range of IML techniques. Classical inference approaches, which are typically asymptotic in nature, can be used for linear, generalized linear, or additive parametric (sometimes semiparametric) statistical models to quantify the uncertainty in parameters. Nonparametric or semiparametric methods like bootstrap uncertainty quantification can also be used for many of the same methods. More recently, several people have developed inferential procedures for regularization techniques like the lasso in high-dimensional regimes. Such approaches include debiasing techniques (van de Geer et al. 2014), which calculate the high-dimensional asymptotic distribution of the lasso, and selective inference (Taylor & Tibshirani 2015), which computes confidence intervals and tests conditional on the lasso solution. Several others have recently employed similar strategies to quantify the uncertainty for unsupervised statistical learning tasks like clustering (Gao et al. 2022), graphical models (Liu 2013), and PCA (Koltchinskii & Lounici 2016). This recent research on statistical inference for popular statistical machine learning models in high-dimensional regimes represents important advances in the field, but the approaches also have several limitations. All of these approaches are model-specific and limited to parametric (or perhaps semiparametric) statistical models, precluding application to popular nonlinear machine learning models such as tree-based ensembles, deep learning, and t-SNE. Furthermore, these approaches assume the data arise from a specific generating model, which is not checkable in practice. They are less effective at quantifying the uncertainty in IML discoveries when the model is misspecified.

Given the limitations of model-specific approaches, and following from recent developments in distribution-free predictive inference, many have advocated for model-agnostic inference, which can quantify the uncertainty associated with any IML model. Thus far, such approaches have only been developed for feature importance and feature selection. Some of the first such approaches were based on the model-X knockoff framework, which generates knockoff features that have no relation to the response but that still retain the dependencies structure among the features (Candès et al. 2018, Barber & Candès 2019). This approach has been used to select features with false discovery rate (FDR) control (Barber & Candès 2019), conduct conditional independence testing (Berrett et al. 2018), and construct confidence intervals for feature importance (Zhang &

Janson 2022), among other applications. The fact that knockoff approaches can be employed for any IML model is a major advantage, but this comes at the expense of assuming that the distribution of the features is known or can be closely approximated, a significant limitation in many domains. Others have recently developed model-agnostic inference approaches for feature importance; some consider feature occlusion inference (Lei et al. 2018, Gan et al. 2022, Williamson et al. 2023), which examines the prediction loss when removing one feature, while others consider the feature permutation test (Berrett et al. 2018, Kim et al. 2022), which randomly permutes the feature of interest. While very general and widely applicable, these approaches either perform inference for a random quantity that depends on the training set or require limiting assumptions on the data distribution or the consistency of the model employed. In fact, the fundamental difficulty of distribution-free and model-agnostic feature importance inference was recently revealed by Shah & Peters (2020), who show that any conditional independence test that is valid without further assumptions on the data distribution or the model has no statistical power. Hence, while model-agnostic inference and uncertainty quantification for IML are critically important for validating many popular IML models, further research is needed to understand and work around limiting distributional and modeling assumptions. Finally, there are many research opportunities to develop model-agnostic inference approaches for IML tasks beyond feature selection and importance; these could include inference approaches for unsupervised techniques as well.

Although this review is not focused on Bayesian machine learning, it is important to mention these techniques in the context of uncertainty quantification. Indeed, many argue that one of the more appealing aspects of Bayesian approaches is the built-in uncertainty quantification through computing the posterior distribution and credible intervals; such approaches have been developed for IML tasks like feature importance and selection, graphical models, factor models, clustering, and more (Vallejos et al. 2015, Cortes et al. 2017). Despite these approaches' appealing uncertainty quantification properties, there are several challenges when applying these techniques to generate and validate IML discoveries. First, computing or sampling from the exact posterior distribution is typically intractable or computationally prohibitive in big data settings. Thus, people typically employ approximation techniques like variational inference (Blei et al. 2017), but there is limited theory on how well these approaches work and how they affect the uncertainty quantification of IML discoveries. Furthermore, the IML discovery, the posterior distribution, and any uncertainty quantification depend strongly on the prior employed. The posterior distribution does not reflect this sensitivity to the prior and hence can underestimate the true uncertainty in the IML discovery; further sensitivity tests and model checking are needed for validation (Gelman & Shalizi 2013, Kruschke 2021). We refer the reader to van de Schoot et al. (2021) for more details.

In summary, statistical inference for IML discoveries is critical for validation and is a growing area of research. There are a number of important recent results in this field, especially for model-agnostic inference, but there are also many open questions and challenges that are ripe areas for further research.

5.3. Example: Uncertainty Quantification for Feature Importance

To illustrate uncertainty quantification via statistical inference and compare this to the practical validation strategies discussed in Section 4, we consider the interpretation of feature importance for a real regression example on the Communities and Crime data set (Redmond 2009). This data set has $n = 1{,}994$ observations and $p = 122$ features, and was assembled with the goal of predicting the per capita violent crime rate. We compare three popular regression methods with their model-specific feature importance scores: the lasso with selected absolute coefficients representing feature importance (λ is selected via cross-validation), random forest with the mean

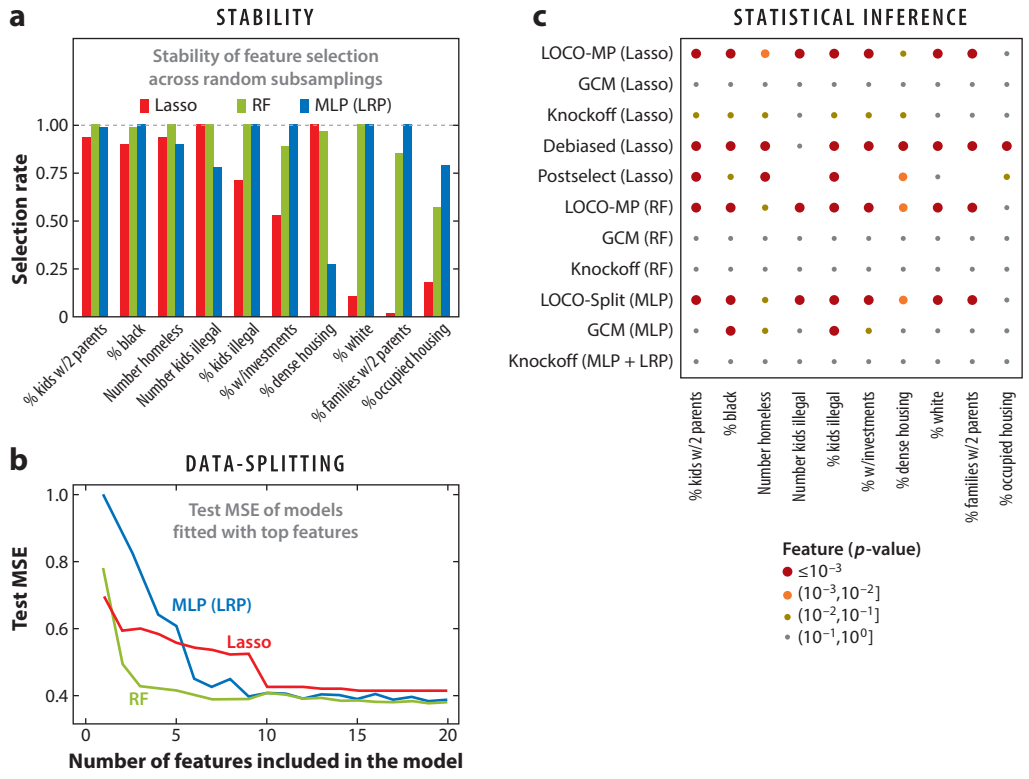

Figure 5

Example of how to validate feature importance via stability (*a*) and data-splitting (*b*), as well as uncertainty quantification via statistical inference (*c*) for regression via the lasso, random forest (RF), and deep learning [multi-layer perceptron with feature importance quantified via layer-wise relevance propagation, or MLP (LRP)] on the Communities and Crime data set. Inference techniques include model-agnostic approaches such as leave-one-covariate-out (LOCO), generalized covariance measure (GCM), and knockoff inference, as well as lasso-specific approaches such as debiased lasso and post selection inference. Other abbreviations: MP, minipatch; MSE, mean squared error.

decrease in impurity tree-based feature importance score, and deep learning implemented via a multilayer perception (MLP) architecture [with two hidden units of size p with rectified linear unit (ReLU) activation] and the epsilon-layer-wise relevance propagation (Montavon et al. 2019) used to compute the post hoc feature importance scores.

In **Figure 5*a*** and ***b***, we illustrate the stability and predictability principle discussed in Section 4.2.4 to validate discoveries. We first split the data into 70% training and 30% test sets. Then, for stability, we further repeatedly subsample 70% of the training set and record a feature as selected if its feature importance score is in the top 10; we report the selection rate for all three methods and the top 10 aggregated most stable features in **Figure 5*a***. Then, to assess whether the selected stable features also offer good predictability, we rank order the top stable features for each method, build a new model with only the top $K = 1, \ldots, 20$ features on the training set, and apply these models to make predictions on the test set; we report the test mean squared error in **Figure 5*b***. All three methods find many of the same features to be stable, with some differences between features selected with linear methods (lasso) as opposed to nonlinear methods (random forest and MLP). But, checking the predictability of the most stable features via data-splitting

reveals that the top nine features of the random forest are the most generalizable, offering the best predictions on a new test set.

Next, we seek to quantify the uncertainty in feature importance scores or in feature selection for each method via statistical inference. We first apply model-specific techniques for the lasso, including the debiased lasso (Javanmard & Montanari 2014) and postselection inference for the lasso (Taylor & Tibshirani 2015), to test whether coefficients are nonzero. Next, we apply the knockoff approach (utilizing the same standard Gaussian knockoff construction for all methods), which gives FDR control for model-agnostic feature selection. We also employ several recently developed model-agnostic inference approaches to test whether feature importance scores are greater than zero; these include the generalized covariance measure (Shah & Peters 2020) and the leave-one-covariate-out (LOCO) inference implemented via minipatches (Gan et al. 2022) or via data-splitting (Lei et al. 2018). In **Figure 5*c***, we report the Benjamini-Hochberg adjusted *p*-values controlling the FDR at 10% (Benjamini & Hochberg 1995) for each of the inference methods and for the same top most stable features shown in **Figure 5*a***. We see that several approaches find few or no features to be statistically significant; the debiased lasso, on the other hand, seems anticonservative, with 23 features declared significant (see the **Supplemental Appendix**). But the LOCO inference approach runs counter to these trends, with the LOCO methods finding the exact same features as statistically significant for all three machine learning methods; these results also closely align with the most stable features from each method. Examining these features reveals several known socioeconomic and criminal justice trends, lending further credence to these validated discoveries. (Further results and implementation details for this example are provided in the **Supplemental Appendix**.)

Supplemental Material >

Overall, this example illustrates how one can utilize stability, data-splitting, and statistical inference to validate the popular machine learning interpretations of feature importance. At the same time, the differing inference results due to the different assumptions between methods highlight the need for further research in this critically important area.

6. DISCUSSION

In this article, we provided an overview of IML techniques that can be used for data-driven discovery and discussed associated challenges and opportunities with validation, statistical theory, and inference. But importantly, there are many aspects that we did not cover in this review that warrant further coverage and discussion in other works. This article focused on fairly general machine learning tasks and techniques, but an abundance of techniques have been developed for specific areas and tasks like those in computer vision, natural language processing and large language models, and reinforcement learning (Glanois et al. 2022), among several other areas. Many of these IML techniques can be used for discoveries and also share similar validation challenges. Another important area that we only briefly covered but that deserves its own careful consideration is causality, which includes interpretability via counterfactual explanations (Mothilal et al. 2020), causal inference from interventional studies, and causal discovery from observational data. The latter can be especially important in science for discovering causal mechanisms, but it perhaps faces even more challenges when it comes to validation, theory, and uncertainty quantification. We also only briefly discussed Bayesian machine learning and its associated uncertainty quantification, but this growing area of research deserves further discussion in the context of IML for generating new discoveries.

This article reviewed and discussed the grand challenge of how to validate discoveries made using IML. We specifically discussed three aspects of this grand challenge: (*a*) practical tools for validating interpretations, (*b*) theoretical foundations of major IML techniques, and (*c*) uncertainty

quantification for machine learning interpretations. We presented two major types of practical validation strategies, data-splitting and stability, but each of these has their own caveats and limitations. Further research is needed to combine the strengths of both approaches, elucidate a theoretical basis for these approaches, or perhaps develop a connection with uncertainty quantification via inference. Next, we have a strong theoretical understanding of only a limited number of IML techniques, mainly those that are intrinsic, global, and model-specific. This does not include interpretations of popular machine learning methods like random forests and deep learning; further research is needed to not only explain the strong predictive performance of these approaches but also understand their interpretations and discoveries. Finally, there has been growing interest in uncertainty quantification for prediction, but quantifying the uncertainty of machine learning interpretations is also another critical component of validation that deserves further attention and research. In addition to challenges associated with validation, there are also several other important questions that require further consideration and research. Some of these include how to match the appropriate IML technique to the desired discovery task; how to compare different interpretations from different IML techniques; and how to marry domain knowledge and expertise with IML to better develop, deploy, and evaluate IML discoveries.

In summary, IML techniques hold great promise for making breakthroughs in science and beyond by mining ever larger data sets to detect the faintest signals. But at the same time, these IML discoveries should be interpreted with caution in the absence of careful validation or uncertainty quantification. Solving this grand challenge is critical for promoting replicable and reliable (data) science as well as trustworthy machine learning; IML techniques also provide exciting research opportunities at the intersection of statistics and machine learning.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors gratefully acknowledge support by National Science Foundation (NSF) NeuroNex-1707400, National Institutes of Health (NIH) 1R01GM140468, and NSF DMS-2210837.

LITERATURE CITED

Abbe E. 2017. Community detection and stochastic block models: recent developments. *J. Mach. Learn. Res.* 18(1):6446–531

Baker M. 2016. 1,500 scientists lift the lid on reproducibility. *Nature* 533:452–54

Barber RF, Candès EJ. 2019. A knockoff filter for high-dimensional selective inference. *Ann. Stat.* 47(5):2504–37

Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, et al. 2020. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inform. Fusion* 58:82–115

Basu S, Kumbier K, Brown JB, Yu B. 2018. Iterative random forests to discover predictive and stable high-order interactions. *PNAS* 115(8):1943–48

Beam AL, Manrai AK, Ghassemi M. 2020. Challenges to the reproducibility of machine learning models in health care. *JAMA* 323(4):305–6

Bengio Y, Courville A, Vincent P. 2013. Representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intel.* 35(8):1798–828

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 57(1):289–300

Berkhin P. 2006. A survey of clustering data mining techniques. In *Grouping Multidimensional Data: Recent Advances in Clustering*, ed. J Kogan, C Nicholas, M Teboulle, pp. 25–71. New York: Springer

Berrett TB, Wang Y, Barber RF, Samworth RJ. 2018. The conditional permutation test for independence while controlling for confounders. *J. R. Stat. Soc. Ser. B* 82:175–97

Bien J, Tibshirani R. 2011. Prototype selection for interpretable classification. *Ann. Appl. Stat.* 5(4):2403–24

Blei DM, Kucukelbir A, McAuliffe JD. 2017. Variational inference: a review for statisticians. *J. Am. Stat. Assoc.* 112(518):859–77

Borjali A, Chen AF, Muratoglu OK, Morid MA, Varadarajan KM. 2020. Deep learning in orthopedics: How do we build trust in the machine? *Healthcare Transform*. **https://doi.org/10.1089/heat.2019.0006**

Broderick T, Gelman A, Meager R, Smith AL, Zheng T. 2023. Toward a taxonomy of trust for probabilistic machine learning. *Sci. Adv.* 9(7):eabn3999

Brunton SL, Proctor JL, Kutz JN. 2016. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *PNAS* 113(15):3932–37

Bühlmann P, Van de Geer S. 2011. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. New York: Springer

Candès E, Fan Y, Janson L, Lv J. 2018. Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *J. R. Stat. Soc. Ser. B* 80(3):551–77

Carvalho DV, Pereira EM, Cardoso JS. 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics* 8(8):832

Cortes A, Dendrou CA, Motyer A, Jostins L, Vukcevic D, et al. 2017. Bayesian analysis of genetic association across tree-structured routine healthcare data in the UK Biobank. *Nat. Genet.* 49(9):1311–18

Dong Y, Su H, Zhu J, Bao F. 2017. Towards interpretable deep neural networks by leveraging adversarial examples. arXiv:1708.05493 [cs.CV]

Doshi-Velez F, Kim B. 2017. Towards a rigorous science of interpretable machine learning. arXiv:1702.08608 [stat.ML]

Drton M, Maathuis MH. 2017. Structure learning in graphical modeling. *Annu. Rev. Stat. Appl.* 4:365–93

Du M, Liu N, Hu X. 2019. Techniques for interpretable machine learning. *Commun. ACM* 63(1):68–77

Fineberg H, Stodden V, Meng XL. 2020. Highlights of the US National Academies report on "Reproducibility and Replicability in Science." *Harv. Data Sci. Rev.* 2(4). **https://doi.org/10.1162/99608f92.cb310198**

Fodor IK. 2002. *A survey of dimension reduction techniques*. Tech. Rep., US Dep. Energy, Washington, DC

Gan L, Zheng L, Allen GI. 2022. Model-agnostic confidence intervals for feature importance: A fast and powerful approach using minipatch ensembles. arXiv:2206.02088 [stat.ML]

Gao LL, Bien J, Witten D. 2022. Selective inference for hierarchical clustering. *J. Am. Stat. Assoc.* **https://doi.org/10.1080/01621459.2022.2116331**

Gelman A, Shalizi CR. 2013. Philosophy and the practice of Bayesian statistics. *Br. J. Math. Stat. Psychol.* 66(1):8–38

Gibney E. 2022. Could machine learning fuel a reproducibility crisis in science? *Nature* 608:250–51

Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter MA, Kagal L. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 80–89. Piscataway, NJ: IEEE

Glanois C, Weng P, Zimmer M, Li D, Yang T, et al. 2022. A survey on interpretable reinforcement learning. arXiv:2112.13112 [cs.LG]

Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. 2018. A survey of methods for explaining black box models. *ACM Comput. Surv.* 51(5):93

Guyon I, Weston J, Barnhill S, Vapnik V. 2002. Gene selection for cancer classification using support vector machines. *Mach. Learn.* 46:389–422

Handl J, Knowles J, Kell DB. 2005. Computational cluster validation in post-genomic data analysis. *Bioinformatics* 21(15):3201–12

Hassan M, Awan FM, Naz A, deAndrés Galiana EJ, Alvarez O, et al. 2022. Innovations in genomics and big data analytics for personalized medicine and health care: A review. *Int. J. Mol. Sci.* 23(9):4645

He Z, Yu W. 2010. Stable feature selection for biomarker discovery. *Comput. Biol. Chem.* 34(4):215–25

Hennig C, Meilă M, Murtagh F, Rocci R. 2015. *Handbook of Cluster Analysis*. Boca Raton, FL: CRC

Hodge V, Austin J. 2004. A survey of outlier detection methodologies. *Artif. Intel. Rev.* 22(2):85–126
Jacovi A, Marasović A, Miller T, Goldberg Y. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 624–35. New York: ACM
Javanmard A, Montanari A. 2014. Confidence intervals and hypothesis testing for high-dimensional statistical models. *J. Mach. Learn. Res.* 15(1):2869–909
Johnstone IM, Lu AY. 2009. On consistency and sparsity for principal components analysis in high dimensions. *J. Am. Stat. Assoc.* 104(486):682–93
Jolliffe IT. 2002. *Principal Component Analysis for Special Types of Data*. New York: Springer
Kim I, Neykov M, Balakrishnan S, Wasserman L. 2022. Local permutation tests for conditional independence. arXiv:2112.11666 [math.ST]
Koh PW, Liang P. 2017. Understanding black-box predictions via influence functions. *Proc. Mach. Learn. Res.* 70:1885–94
Koltchinskii V, Lounici K. 2016. Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance. *Ann. Inst. Henri Poincaré Probab. Stat.* 52(4):1976–2013
Kruschke JK. 2021. Bayesian analysis reporting guidelines. *Nat. Hum. Behav.* 5(10):1282–91
Lange T, Roth V, Braun ML, Buhmann JM. 2004. Stability-based validation of clustering solutions. *Neural Comput.* 16(6):1299–323
Lauritzen SL. 1996. *Graphical Models*. Oxford, UK: Clarendon
Lei J, G'Sell M, Rinaldo A, Tibshirani RJ, Wasserman L. 2018. Distribution-free predictive inference for regression. *J. Am. Stat. Assoc.* 113(523):1094–111
Li X, Wang Y, Ruiz R. 2022. A survey on sparse learning models for feature selection. *IEEE Trans. Cybernet.* 52(3):1642–60
Lipton ZC. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16(3):31–57
Liu H, Roeder K, Wasserman L. 2010. Stability approach to regularization selection (StARS) for high dimensional graphical models. In *NIPS'10: Proceedings of the 23rd International Conference on Neural Information Processing Systems*, ed. JD Lafferty, CKI Williams, J Shawe-Taylor, RS Zemel, A Culotta, pp. 1432–40. Red Hook, NY: Curran
Liu W. 2013. Gaussian graphical model estimation with false discovery rate control. *Ann. Stat.* 41(6):2948–78
Löffler M, Zhang AY, Zhou HH. 2021. Optimality of spectral clustering in the Gaussian mixture model. *Ann. Stat.* 49(5):2506–30
Materne J. 1978. The structure of nearby clusters of galaxies—hierarchical clustering and an application to the Leo region. *Astron. Astrophys.* 63:401–9
McDermott MB, Wang S, Marinsek N, Ranganath R, Foschini L, Ghassemi M. 2021. Reproducibility in machine learning for health research: still a ways to go. *Sci. Transl. Med.* 13(586):eabb1655
McInnes L, Healy J, Melville J. 2020. UMAP: uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426 [stat.ML]
Meinshausen N, Bühlmann P. 2010. Stability selection. *J. R. Stat. Soc. Ser. B* 72(4):417–73
Meng XL. 2020. Reproducibility, replicability, and reliability. *Harv. Data Sci. Rev.* 2(4). **https://doi.org/10.1162/99608f92.dbfce7f9**
Molnar C. 2022. *Interpretable Machine Learning*. N.p.: C. Molnar. 2nd ed.
Montavon G, Binder A, Lapuschkin S, Samek W, Müller KR. 2019. Layer-wise relevance propagation: an overview. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, ed. W Samek, G Montavon, A Vedaldi, LK Hansen, K-R Müller, pp. 193–209. New York: Springer
Monti S, Tamayo P, Mesirov J, Golub T. 2003. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.* 52(1–2):91–118
Mothilal RK, Sharma A, Tan C. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 607–17. New York: ACM
Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. 2019. Definitions, methods, and applications in interpretable machine learning. *PNAS* 116(44):22071–80

Natl. Acad. Sci. Eng. Med. 2019. *Reproducibility and Replicability in Science*. Washington, DC: Natl. Acad. Press
Neufeld A, Dharamshi A, Gao LL, Witten D. 2023. Data thinning for convolution-closed distributions. arXiv:2301.07276 [stat.ME]
Ozer M, Kim N, Davulcu H. 2016. Community detection in political Twitter networks using nonnegative matrix factorization methods. In *ASONAM '16: Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 81–88. Piscataway, NJ: IEEE
Peng RD, Hengartner NW. 2002. Quantitative analysis of literary styles. *Am. Stat.* 56(3):175–85
Perou CM, Sørlie T, Eisen MB, Van De Rijn M, Jeffrey SS, et al. 2000. Molecular portraits of human breast tumours. *Nature* 406(6797):747–52
Rasheed K, Qayyum A, Ghaly M, Al-Fuqaha A, Razi A, Qadir J. 2022. Explainable, trustworthy, and ethical machine learning for healthcare: a survey. *Comput. Biol. Med.* 149:106043
Redmond M. 2009. *Communities and crime*. Data Set, UCI Machine Learning Repository, Univ. Calif., Irvine, CA
Reyes M, Meier R, Pereira S, Silva CA, Dahlweid FM, et al. 2020. On the interpretability of artificial intelligence in radiology: challenges and opportunities. *Radiol. Artif. Intel.* 2(3):e190043
Ribeiro MT, Singh S, Guestrin C. 2016. "Why should I trust you?": explaining the predictions of any classifier. In *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–44. New York: ACM
Roscher R, Bohn B, Duarte MF, Garcke J. 2020. Explainable machine learning for scientific insights and discoveries. *IEEE Access* 8:42200–16
Rubinov M, Sporns O. 2010. Complex network measures of brain connectivity: uses and interpretations. *Neuroimage* 52(3):1059–69
Rudin C. 2014. Algorithms for interpretable machine learning. In *KDD '14: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 1519. New York: ACM
Rudin C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intel.* 1(5):206–15
Rudin C, Chen C, Chen Z, Huang H, Semenova L, Zhong C. 2022. Interpretable machine learning: fundamental principles and 10 grand challenges. *Stat. Surv.* 16:1–85
Samek W, Montavon G, Lapuschkin S, Anders CJ, Müller KR. 2021. Explaining deep neural networks and beyond: a review of methods and applications. *Proc. IEEE* 109(3):247–78
Samek W, Müller KR. 2019. Towards explainable artificial intelligence. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, ed. W Samek, G Montavon, A Vedaldi, LK Hansen, K-R Müller, pp. 5–22. New York: Springer
Satija R, Farrell JA, Gennert D, Schier AF, Regev A. 2015. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* 33(5):495–502
Shah RD, Peters J. 2020. The hardness of conditional independence testing and the generalised covariance measure. *Ann. Stat.* 48(3):1514–38
Shah RD, Samworth RJ. 2013. Variable selection with error control: another look at stability selection. *J. R. Stat. Soc. Ser. B* 75(1):55–80
Stodden V. 2020. Theme editor's introduction to reproducibility and replicability in science. *Harv. Data Sci. Rev.* 2(4). **https://doi.org/10.1162/99608f92.c46a02d4**
Taeb A, Shah P, Chandrasekaran V. 2020. False discovery and its control in low rank estimation. *J. R. Stat. Soc. Ser. B* 82(4):997–1027
Taylor J, Tibshirani RJ. 2015. Statistical learning and selective inference. *PNAS* 112(25):7629–34
Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* 58(1):267–88
Toreini E, Aitken M, Coopamootoo K, Elliott K, Zelaya CG, van Moorsel A. 2020. The relationship between trust in AI and trustworthy machine learning technologies. In *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 272–83. New York: ACM
Tukey JW. 1977. *Exploratory Data Analysis*. Boston: Addison-Wesley
Vallejos CA, Marioni JC, Richardson S. 2015. Basics: Bayesian analysis of single-cell sequencing data. *PLOS Comput. Biol.* 11(6):e1004333
van de Geer S, Bühlmann P, Ritov Y, Dezeure R. 2014. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Stat.* 42(3):1166–202

van de Schoot R, Depaoli S, King R, Kramer B, Märtens K, et al. 2021. Bayesian statistics and modelling. *Nat. Rev. Methods Primers* 1(1):1

Van der Maaten L, Hinton G. 2008. Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9(86):2579–605

Wainwright MJ. 2019. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge, UK: Cambridge Univ. Press

Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. 2013. The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nat. Genet.* 45(10):1113–20

Williamson BD, Gilbert PB, Simon NR, Carone M. 2023. A general framework for inference on algorithm-agnostic variable importance. *J. Am. Stat. Assoc.* 118:1645–58

Willis C, Stodden V. 2020. Trust but verify: how to leverage policies, workflows, and infrastructure to ensure computational reproducibility in publication. *Harv. Data Sci. Rev.* 2(4). **https://doi.org/10.1162/99608f92.25982dcf**

Witten DM, Tibshirani R. 2010. A framework for feature selection in clustering. *J. Am. Stat. Assoc.* 105(490):713–26

Xu F, Uszkoreit H, Du Y, Fan W, Zhao D, Zhu J. 2019. Explainable AI: a brief survey on history, research areas, approaches and challenges. In *Natural Language Processing and Chinese Computing*, ed. J Tang, MY Kan, D Zhao, S Li, H Zan, pp. 563–74. New York: Springer

Yasaka K, Abe O. 2018. Deep learning and artificial intelligence in radiology: current applications and future directions. *PLOS Med.* 15(11):e1002707

Yu B, Kumbier K. 2020. Veridical data science. *PNAS* 117(8):3920–29

Zhang L, Janson L. 2022. Floodgate: inference for model-free variable importance. arXiv:2007.01283 [stat.ME]

Zhang Y, Song K, Sun Y, Tan S, Udell M. 2019. "Why should you trust my explanation?" Understanding uncertainty in LIME explanations. arXiv:1904.12991 [cs.LG]

Zhao P, Yu B. 2006. On model selection consistency of lasso. *J. Mach. Learn. Res.* 7(90):2541–63

Zou H, Hastie T. 2005. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B* 67(2):301–20