An Overview of Computational Approaches for Analyzing Interpretation

Philipp Blandfort
TUK and DFKI, Kaiserslautern, Germany
philipp.blandfort@dfki.de

Jörn Hees
DFKI, Kaiserslautern, Germany
joern.hees@dfki.de

Desmond U. Patton
Columbia University, NYC, USA
dp2787@columbia.edu

ABSTRACT

It is said that beauty is in the eye of the beholder. But how exactly can we characterize such discrepancies in interpretation? For example, are there any specific features of an image that makes person A regard an image as beautiful while person B finds the same image displeasing? Such questions ultimately aim at explaining our individual ways of interpretation, an intention that has been of fundamental importance to the social sciences from the beginning. More recently, advances in computer science brought up two related questions: First, can computational tools be adopted for analyzing ways of interpretation? Second, what if the "beholder" is a computer model, i.e., how can we explain a computer model's point of view? Numerous efforts have been made regarding both of these points, while many existing approaches focus on particular aspects and are still rather separate.

With this paper, in order to connect these approaches we introduce a theoretical framework for analyzing interpretation, which is applicable to interpretation of both human beings and computer models. We give an overview of relevant computational approaches from various fields, and discuss the most common and promising application areas. The focus of this paper lies on interpretation of text and image data, while many of the presented approaches are applicable to other types of data as well.

KEYWORDS

survey, interpretation, perspective, explainability, machine learning, pattern mining, visualization, correlation, social science

1 INTRODUCTION

Individual ways of interpretation play a major role in a variety of fields. The philosophical positions scepticism, relativism and perspectivism all crucially involve the notion of points of view [19], i.e., different ways of interpretation. Hermeneutics refers to a whole field that is concerned with how we interpret information and commonly assumes that in order to make sense of things we need to relate them to our own life situation, which makes all interpretation something inherently personal (e.g., see [158]). Analyzing how we make sense of the world is pertinent to cognitive science, the research field concerned with studying the human mind. Similarly, in psychology it has been argued that understanding each others' motivations is a key aspect of human social life [42]. Even in a non-scientific context, everyday misunderstandings in communication offer a clear demonstration of both challenge and importance of correctly estimating what other people mean and anticipating how they would interpret our own behavior.

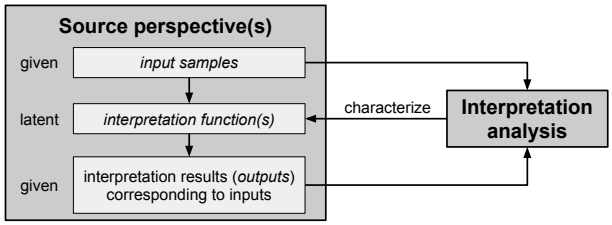

Figure 1: Interpretation analysis under the black-box assumption. The goal is to characterize interpretation from one or several perspectives, which can be human or artificial. Interpretation from each perspective is formally described as a mapping from information to meaning. For this paper, we assume that these functions are not directly accessible, but only indirectly via a list of inputs and associated outputs.

Nowadays, there are two developments that drastically impact our social life and motivate the need for computational methods with similar social abilities: First, more and more communication is happening online [107]. Second, AI approaches have become much more ubiquitous. This is especially prevalent online, where chatbots take part in discussions, recommendation algorithms suggest things we are likely to favor, and search results are nicely ranked by yet another computer model. In a broad sense, humans and computer models are all actors in a large communication network. In many cases the goal of an AI approach is to learn about a certain way of interpretation. This is most clearly visible in supervised approaches where the ground truth data serves as a proxy to the human perspective that is to be learned, which often involves estimating subjective qualities (e.g., what a user will like, or even automatically mining opinions). At the same time, as AI approaches become actors in communication and their automatic decisions become more and more influential in our everyday life, we also have a motivation to understand them. As approaches have grown considerably more complex over the years, this is not at all trivial. However, since early 2018, with changes in European legislation (GDPR [36]) there is now even a legal reason why many companies (and probably also researchers) should analyze how the developed models draw their conclusions: Whenever users are affected by automatic decisions, the users now have the legal right for an explanation of the decision in simple terms [47]. Yet another pragmatic motivation for understanding AI approaches stems from ever-growing amounts of data ("big data") involved in digital activities such as posting comments, liking contents or browsing websites: Due to the scale of user data, it has become extremely challenging to manually inspect even a

fraction of the data. Here, computers have a clear edge in terms of scalability, and are valuable for processing all this information and thus making it more accessible to us, potentially even by explaining its characteristics.

So we see that there are three important tasks, namely enabling AI approaches to "understand" our view, understanding how AI agents see the world, and having computer models explain complex data to us. It is clear that neither of these tasks is simple, still, good progress has been made on all of them. To name a few recent advances: A lot of work was done on explaining how deep learning models work [75, 80, 81, 89, 98, 123], which was even useful for helping us understand complex scientific data [4, 125, 135]. In case of data annotation, probabilistic methods have been proposed to merge annotator votes efficiently and simultaneously estimating annotator reliabilities [15, 141]. However, despite related goals, approaches for interpretation analysis seem quite separated and we find an apparent lack of high-level bridges to connect them. Moreover, underlying concepts such as interpretation or understanding are often not defined properly (as [75] explains for the concept interpretability), which suggests the need for more rigorous formalism.

The main purpose of this paper is to connect various ideas and approaches, and put them into a coherent view. To this end, we introduce a theoretical framework, in which a perspective is represented by a function from input to meaning, called the interpretation function. Interpretation analysis can then be understood mathematically as characterization of such an interpretation function. We do a survey on approaches for this task with a focus on text and image inputs, where we in particular find statistical methods, pattern mining, model-based approaches and visualization techniques to be of central relevance. In addition to outlining methods for analyzing interpretations of a single model, this paper describes methods for comparing multiple perspectives. We also try to reveal relations to the social sciences, where it has a much longer tradition to look into characteristics of interpretation, in the hope that this will contribute to more discussion between the disciplines.

We structure the paper as follows: First, in Section 2 we will describe our theoretical framework and formally define interpretation, perspective and the task of interpretation analysis. This is followed by general remarks about the task in Section 3, where we comment on evaluation and input representation. We will then look into approaches for the case of analyzing one individual perspective (Section 4). To this end, we can make use of statistical methods, pattern mining, model-based approaches or visualization techniques (see overview in Table 1). Comparisons between multiple perspectives will be handled in Section 5 and can be done under the use of three kinds of approaches (see also Figure 2). We will see that two of these cases can mostly be reduced to single perspective analysis, which makes the methods for analyzing relations between input and output of a single interpretation function the core of this paper. In Section 6, we outline five application fields, where ways of interpretation are analyzed by means of computational methods. Finally, we close the paper with a few remarks on future work and ethical aspects (Section 7).

2 THEORETICAL FRAMEWORK

Montavon et al. [89] define interpretation as a "mapping of an abstract concept (e.g., a predicted class) into a domain that the human can make sense of". We agree that this might work for the specific purpose of their analysis, but find this definition to be in conflict with intuition. Most importantly, the definition does not include a large part of human interpretation, which in general starts from something concrete (like an image or text) and ends up in something more abstract that we can broadly call meaning. Hence, we keep the mapping part but remove the restrictions of the input and output domain while we introduce the notion of a *bearer*, inspired by recent works in philosophy on defining perspectives [19, 48]:

Definition (Perspective, bearer, interpretation, interpretable). We define a *perspective* as a way of interpretation of some actor or group of actors b, which we call the *bearer(s)* of the perspective. Formally, we can represent a way of interpretation by a mapping from input to meaning, which we name *interpretation function* f_b of b:

$$f_b : I_b \rightarrow M_b \,,$$

where I_b is the input domain and M_b the output domain (set of potential meanings). Any information i is then called *interpretable* by b if and only if it is contained in the input domain of b's interpretation function, i.e., $i \in I_b$.

Examples. 1) Image classification of a machine learning model m can be seen as interpretation process, where the interpretation function f_m of the model maps from a set of images I_m into a set of classes M_m. 2) An example for a human perspective would be the interpretation process of annotator a from a set of tweets I_a into {sarcastic, not sarcastic} when being asked to label tweets accordingly. 3) More complicated output domains are possible. For example, in case of an image autoencoder e the latent representation can be modelled as interpretation of e.

2.1 Role of the bearer

We do not impose any particular requirements on the input or output domain, but we require that a perspective is adopted by some actor b (e.g., human being or computer model, existing or hypothetical), or group of actors. In case a restriction is necessary, one can achieve this by limiting the set of possible bearers, which naturally leads to restrictions on the input and output domains, as well as the form of possible interpretation functions. For example, if b is limited to be certain neural networks, both inputs and outputs are typically in tensor format.

Note that in the following, we will usually not explicitly mention the bearers of perspectives and just use the symbol f to refer to an interpretation function.

2.2 Assumptions

For this paper, we assume that we do not have direct access to any interpretation function f, but only have a list of inputs and their corresponding outputs. In other words, we treat interpretation as a black-box, that is accessible only through a list of example pairs $(d_0, f(d_0)), \ldots, (d_n, f(d_n))$. This assumption enables us to more easily model interpretation of humans and AI approaches within

Approach type	Methods	Outcomes
statistical methods	correlation coefficients hypothesis testing	measure of statistical dependency, significance
pattern mining	association rule mining emerging pattern mining discriminative pattern mining	association rules (implications), characteristic patterns
model-based approaches	heatmapping prototypes understandable models ablation studies	model for approximating interpretation function, plus: explanations for individual decisions (heatmapping), characteristic inputs (prototypes), or approximate functional description of the function
visualization techniques	dimensionality reduction exemplar-based approaches text summarization	compression of the data, in form of plots, selected examples, or text summary

Table 1: Overview of approaches for single perspective analysis.

the same framework, and is another point that clearly distinguishes this survey from other overview papers related to explainability such as [75, 89, 123].

2.3 Goals of interpretation analysis

Overall, the main goal of interpretation analysis is to characterize interpretation functions. (See Figure 1 for a schematic overview.) Such a characterization can take different forms and be addressed in various ways, depending in particular on whether the goal is to understand a single perspective (Section 4) or to compare several perspectives (Section 5).

For analysis of a single perspective, we want to extract characteristic properties from a single function in order to answer "What are the relations between features of the input and interpretation result?" For example, which parts of the image make the classifier say that there is a dog in the image?

For comparing several perspectives, we are generally interested in discriminative characterization. For example, we can ask "For which kinds of inputs can we expect any difference between machine learning models A and B?" or "Which features of tweets characterize the set of tweets which annotator C labels as *aggressive* while annotator D labels them as *non aggressive*?"

3 COMPUTATIONAL APPROACHES

As we just saw in Section 2.3, interpretation analysis in the proposed framework amounts to characterizing functions, interpretation functions to be more precise. The general purpose of functions is to formally describe how one quantity (the output) depends on another quantity (the input). Hence, at the very core of interpretation analysis (or analyzing and understanding any function for that matter) we find the task of figuring out how outputs *depend on* inputs. And this is to be done based on a list of inputs and their corresponding outputs. So we have already converted the conceptually challenging problem of interpretation analysis into a more graspable mathematical formulation, which can be tackled with a variety of existing computational methods. We have also discussed that the task takes on a slightly different touch depending on whether we are analyzing one individual perspective or aim at comparing between multiple ones. Before we go into detail on these approaches in Sections 4 and 5, we will first discuss two general points that are relevant in all these cases, namely evaluation and feature extraction.

3.1 Evaluation

Natural questions to ask when being confronted with any large set of tools for a single task is: Which one to choose? And on which grounds should one make such a decision? First of all, despite following the common goal of characterizing a single function in terms of relations between input and output, the relevant approaches vary in terms of result format, but also with respect to other properties such as reliability and expected data (type and amount). This makes it difficult to directly compare all the approaches, and indeed, a general automatic evaluation measure for interpretation analysis does not exist. For several individual categories evaluative measures have been proposed (e.g., see [122] for heatmapping), but in practice, quantifying usefulness of explanations largely remains an open issue and qualitative evaluation often becomes necessary. This can mean that researchers manually inspect results and view examples for judging which method does the better job, or task someone else (e.g., crowdworkers) with evaluating which method generates better explanations (e.g., as in [35]). Another interesting option is mentioned in [89], namely to look at simpler versions of the tasks where an optimal explanation can be specified and then compare the results to this explanation.

In general, we regard the following three criteria as important: 1) The results should be *reliable*, which includes statistical significance and robustness. 2) The characterization should be simple to *understand*. 3) The findings should *cover* as much as possible of the *variation* in the data that one wants to understand. (For a single perspective, explain variations in output in terms of input; for several perspectives, explain their differences.) Note that these points are treated quite differently in the relevant fields. Reliability is absolutely fundamental in statistics and still important in pattern mining, but mentioned more rarely in model-based approaches. Understandability is a factor across the fields, but interestingly, the necessary background knowledge for correctly interpreting given explanations varies significantly. Coverage of variation is often checked in statistics (coefficient of determination, R^2), quite central in pattern mining, but harder to address in some of the model-based

approaches (e.g., how to measure to which degree output variation can be explained in terms of heatmaps or prototypes).

Irrespective of the chosen approach, analyzing interpretations often has strong ethical implications (e.g., [113]), so the question should not only be "Which tool should I use for analysis?", but also "Should I analyze this aspect of the data at all?" and "How can I apply these tools responsibly?".

3.2 Feature extraction

Here, and in most of machine learning, we face a situation similar to that in correlational studies in psychology [69], where the data is already there and we need to answer: What is the kind of input "parts" we want to consider for checking dependencies with the output?

First of all, many of the approaches we will discuss cannot be expected to reveal interesting findings when applied to low-level input features such as individual pixels or sequence of characters. For example, if the color of any individual pixel of an image correlates significantly with a classifier output for "dog", then this is hard to make sense of and has a high chance of being a statistical artifact. This is per se not specific to interpretation analysis and especially in applied machine learning feature engineering (i.e., finding suitable features) remains a key part [29]. This process generally requires expertise, since the features need to be appropriate for the final method, the data at hand, and the overall purpose of analysis. It is in the last of these parts, purpose of analysis, where we find a considerable difference between standard machine learning and interpretation analysis. Most of the time, in machine learning the features are meant to serve the purpose of building a prediction model that is reliable (i.e., does not overfit) and has good predictive power. In case of interpretation analysis, we have seen both of these criteria in similar forms (predictive power corresponding to coverage of variance), but in addition require that results should be understandable (see Section 3.1).

This leads to some features such as CNN filter being less appropriate here. After all, what for instance would it mean if the 10th entry of a VGG [129] filter was found to correlate with another image classifier's positive decision for the dog class? Still, when deciding on which features to use, one should definitely be inspired by existing approaches on feature extraction, and some of the simpler common features (e.g., bag of words, occurrences of specific n-grams, color histograms, bag of visual words) can be useful for analyzing interpretation. Finally, in interpretation analysis it happens at times that features are implied by the research goal. For example, if one wants to analyze whether a visual sentiment classifier prefers cats over dogs, cat and dog presence are suitable features. Overall, finding the right features is a complex topic, in part because the understandability criterion is hard to formalize and its implications depend on the type of approach that is used later. Hence, we will mention approach-specific examples in the following sections (4.2 and 4.3).

4 INPUT-OUTPUT DEPENDENCIES

Let's now focus on the goal of understanding a single perspective. The context can be summed up as follows: For an interpretation function $f : I \rightarrow M$ of interest, we want to determine relations between the function's input and output, based on a list of input-output pairs $(d_0, f(d_0)), (d_1, f(d_1)), \ldots, (d_n, f(d_n))$, where $d_i \in I$ for all i and $n \in \mathbb{N}$. We are primarily interested in cases where I consists of language data, images, or feature vectors thereof. The output domain M is assumed to contain feature vectors of fixed dimension. Typical examples would be to analyze a list of sentence sentiment annotations by a single annotator in order to find out which words of the sentence make the annotator assign a certain sentiment, or to analyze an image classifier based on a list of image-classification results for identifying which patches of images are most relevant for a particular result.

For such a task we have several types of approaches from various well-established fields at our disposal, which we will now discuss. For several of these, we will use hypothetical user preference data for illustration. This data can be found in Table 2 and corresponds to a simple interpretation from a 3-D feature space into the binary space of like/dislike.

Image ID	Nudity	Humor	Explosions	Like
0	0	1	0	0
1	1	0	1	1
2	0	1	1	1
3	1	1	0	0
4	1	0	0	1
5	1	1	1	0

Table 2: Hypothetical image preference data of a single user. The three columns in the middle describe features of the image, while the last column describes the type of user reaction which corresponds to an interpretation result (assuming the user has the option to e.g., either vote up or down).

4.1 Statistical methods

One way to analyze relations between two variables is to test for statistical dependencies between them. We can treat both input and output as values of (composed) random variables X and Y respectively, and then test whether individual dimension X_i of X and Y_j of Y are *statistically dependent*. Formally, such a dependency is given if for any sets of possible values A and B, the two events $X_i \in A$ and $Y_j \in B$ are not independent, i.e., $P(Y_j \in B \mid X_i \in A) \neq P(Y_j \in B)$. In other words this means that information about the value of X_i can give us any information about the value of Y_j.

In our toy example (Table 2), we could check if the image preference of the user statistically depends on whether the image contains nudity/humor/explosions.

Correlation coefficients. In its broadest sense, correlation refers to exactly what we described above, i.e., any statistical dependency between two random variables. More specifically, there exist several ways of calculating *correlation coefficients*, each one of them designed to measure the strength of a particular kind of statistical dependency. The most common candidates are Pearson's correlation coefficient [103], which measures linear dependence between two continuous random variables, and Spearman's rank correlation coefficient [145], which measures how well the relationship between the two variables can be described by a monotonic function.

Both of these coefficients are fairly simple to interpret, however, it shall be noted that a Pearson or Spearman coefficient of 0 does *not* imply the absence of any statistical dependency between the variables. For example, for X uniformly distributed on $[-3, 3]$ the random variables X and X^2 have Pearson and Spearman correlation 0 but are far from independent. There exist other correlation measures, able to capture more complex statistical dependencies but typically harder to interpret. These include distance correlation introduced in [139], which is 0 only if the tested variables are independent. Specific choices should be made based on the properties of the tested variables (distributions they follow) and the questions one wants to answer with the analysis.

Statistical significance. Correlation coefficients mainly measure the degree of a certain statistical dependency, but one should also check reliability of the findings by checking whether the dependency is *statistically significant*. This can be done based on *hypothesis testing* for estimating how likely it is that the true correlation is 0 (in a two-sided test, or ≤ 0 or ≥ 0 in one-sided tests) and the observed correlation value is due to noise. Another option is to calculate *confidence intervals* for the coefficients, for which a variety of methods has been proposed (e.g., see [118] for Spearman correlation).

Note that, coming directly from the definition of statistical dependency, we can also estimate confidence intervals for both the expected value of Y_j and the expected value of Y_j given a particular value x of X_i. These intervals can be quite useful for understanding and if these confidence intervals do not overlap, this means that there is a significant difference between $E(Y_j)$ and $E(Y_j \mid X_i = x)$, i.e., X_i attaining value x significantly affects the expected value of the output Y_j. It shall be mentioned that overlapping confidence intervals do *not* imply that there is no significant difference [60].

Remark on causality. Intuitively, we might also want to understand which features of the input *cause* a certain response. However, all methods we discussed try to figure out statistical dependence (correlation), which does not imply causation. In fact, causal assumptions can generally only be verified if experimental control is exerted [102]. In the general case described in this paper, the possibility for collecting additional data while manipulating parts of the input cannot be guaranteed. It shall be mentioned that for the case of analyzing given AI approaches, this possibility is likely to be given and there are some recent attempts in computer science to address causality (e.g., [52, 77, 127, 133, 151]). We believe that this direction should be further explored for interpretation analysis in future work, and refer the interested reader also to the paper of Pearl [102] for a solid overview on causal inference in statistics.

Usage. Statistical testing can be very valuable and at times even necessary for proving claims made about input-output dependencies. Additionally, results of statistical methods are often simple to understand and the methods themselves are transparent. Downsides are that findings crucially depend on which parts of the input and output are considered for testing, which can be very challenging, especially when considering dependencies between high dimensional inputs and outputs.

4.2 Pattern mining

The goal of pattern mining is to find patterns in the data that are characteristic. What exactly constitutes a pattern varies, but they often take on the the forms of association rules, emerging patterns or visual patches, as will be described in the following.

Association rule mining. Association rule mining has a long tradition in pattern mining [2, 106], and is in particular often used for web personalization where it is applied to usage data [33, 85, 86]. In its original form [2] it can be used to process a list of binary vectors and find implications of the form "if an image contains nudity and humor, then in 50% of cases the image also contains explosions" (using hypothetical data from Table 2).

Let $T = \{b_1, \ldots, b_n\}$ be a multi-set of n transactions over k items represented as binary vectors with $b_i \in \mathbb{B}^k$, $n, k \in \mathbb{N}$. An association rule can formally be defined as implication of the form $X \Rightarrow j$, where $X \subseteq \{0, \ldots, k\}$ is a set of indices called the antecedent of the rule, and $j \in \{0, \ldots, k\} \setminus X$ is a single index (not included in X) called the consequent of the rule. The *support* of a set of indices X can then be defined as the relative amount of transactions containing all items in X, and the *confidence* of a rule $X \Rightarrow j$ as the relative support of the rule's antecedent and consequent over the support of its antecedent (see [2]):

$$supp(X) := \frac{|\{b_i \in T \mid b_{i,j} = 1, \forall j \in X\}|}{|T|}$$

$$conf(X \Rightarrow j) := \frac{supp(X \cup \{j\})}{supp(X)}$$

Another important measure *lift* [17], describes the ratio of the observed support for a rule to the support that would be expected if antecedent and consequent were independent. Confidence, support, other measures such as lift, and given potential constraints (e.g., only considering rules with specific j), can all serve as criteria for filtering possible rules. Association rules are often computed based on the apriori [3] or frequent pattern tree [49] algorithms (see e.g., the survey [153]).

For interpretation analysis, we are interested in rules that have a subset of the input as antecedent and a subset of the output as consequent. So in our hypothetical example (Table 2), we would try to find rules of the form "if an image contains explosions, then the user likes it in 2/3 of cases." Such a way of modeling is for instance adopted in [108], where association rule mining is used for finding class-discriminative features in images. In their approach, a binary class membership entry is appended to all vectors and only rules with this particular index as consequent are considered.

Emerging pattern mining. The problem of emerging pattern mining was introduced in [30], originally for capturing trends in time-stamped databases. It is similar to association rule mining, but uses the notion *growth rate* to measure how support for a pattern (set of indices) differs between sets. So broadly speaking, the goal of emerging pattern mining is to find differences in patterns across multiple sets. Soon after the task was introduced, it has been used for classification purposes [31, 72], where emerging patterns are meant to capture characteristic differences between classes. To this end, input samples are partitioned based on the associated output values and found patterns used to discriminate between

the resulting partitions. It is in this sense that this approach can directly be used for interpretation analysis. Coming back to our toy example of Table 2, following an emerging pattern mining approach we would ask, which are the combinations of nudity, humor and explosions that are comparatively more frequent in images the user likes/dislikes. Note that the survey of Novak et al. [96] puts emerging pattern mining under the umbrella term supervised descriptive rule discovery, together with contrast set mining and subgroup mining. Another useful resource is the recent survey of [45].

Visual pattern mining. There are several image-specific approaches worth mentioning. In [112], Rematas et al. use standard data mining terminology to formulate the problem of finding characteristic visual patches from a given image collection, which they also put into a graph for navigation through the image collection. The publications [73, 74] use association rule mining on mid-level CNN features, and call this combination mid-level deep pattern mining.

Sometimes the notion "parts" is used for referring to something similar to visual patterns. For example, [101] describes how to automatically discover discriminative parts for the purpose of image classification. Visual pattern mining was also applied in [24], by using a bag-of-features representation (also known as bag-of-visual-words) [25] and selecting representative and discriminative local features based on Peng's method for feature selection [105]. The recently proposed PatternNet [71] introduces a CNN that directly learns discriminative visual patterns. As such, some of these approaches could as well be put into the model-based category described in the next section.

Usage. As compared to simple statistical methods, pattern mining can find much more complex relations in the data, while still including measures for reliability of the findings. However, since the space of possible patterns can be huge, there is a high risk of ending up with many false alarms [144]. Also note that many pattern mining techniques operate on binary data, so it might be necessary to first convert the data. An example of an adaptation of pattern mining to textual data can be found in [155].

4.3 Model-based approaches

Even though the perspective of interest is considered to be a black-box in this paper, it is still possible to build another model to approximate the interpretation function from the given input-output pairs, and then analyze this trained model in the hope that it processes information in the same way as the black-box. We do not go into too much detail for heatmapping and prototype methods because there are other papers such as [89, 122] which give an excellent overview for most of these approaches (in a non-black-box set-up). In the example of our toy data (Table 2), we would first train a computational model to predict like/dislike based on the input features nudity, humor and explosions, and successively analyze the trained model for dependencies between both parts.

Heatmapping. In the context of analyzing machine learning models, a *heatmap* refers to an explanation of the model's decision for a particular sample in terms of the input, indicating visually which parts of the input are relevant (positively or negatively) for the decision. A common method is to calculate *saliency maps* [8, 109, 128] based on sensitivity analysis [120, 137, 161], i.e., the gradients on the model's input are used for estimating how sensitive the model is to changes in the individual input components. A related approach is prediction difference analysis [159], which is based on saliency mapping but uses local regularization in order to obtain visualizations that are easier to interpret. Saliency maps are simple to calculate for neural networks by means of backpropagation [117], but on the downside, resulting heatmaps have been shown to be unreliable in certain cases [57]. Also, sensitivities to input components is typically not exactly what we want to find out, because they only tell us how the input could be changed to make it belong more or less to a certain class instead of explaining which parts of the input actually make it belong to a class. The latter can be achieved within the theoretical framework of Taylor decomposition [10]. In [7], Bach et al. adapt Taylor decomposition to neural networks and introduce layer-wise relevance propagation (LRP), which makes use of the network's architecture to propagate relevance backward through the network for obtaining a heatmap. The backward propagation rule they derive takes two hyperparameters and for one particular combination, simplifies into a rule that is interpretable as deep Taylor decomposition [88]. Other backprop techniques have been proposed for computing heatmaps for neural networks, including Deconvolution [150], Guided Backprop [132], Class Activation Mapping [157], PatternAttribution [59] and PatternLRP [58].

Prototypes. Another way of visualizing what the model has learned is to calculate inputs that serve as prototypes for the individual classes, which can be done within the analysis framework of *activation maximization* [11, 34]. Essentially, finding prototypes amounts to solving the optimization problem of finding an input that maximizes a certain component of the output (e.g., an image that is interpreted by the model as being maximally dog-like). Without any additional restrictions, the resulting prototypes tend to be unnatural [128], which is why various regularization methods have been proposed [80, 93, 94]. For neural networks in particular, there are numerous efforts on visualizing what particular neurons or neuron layers have learned (e.g., [63, 149, 150]) which can also be seen as prototype approaches. An interesting non-prototype (but still related) approach is the one of [9, 156], where hidden unit activation are related to a binary segmentation task of the input for a given list of semantic concepts, in order to analyze semantics of individual hidden units.

Understandable models. Sometimes it is also possible to fit the data with a model that allows for more direct interpretation. This can mean making use of a simple model such as linear regression or decision trees, breaking the problem down into more accessible steps in pipeline approaches, or incorporating specific components into architectures that can be understood intuitively (e.g., explicit attention mechanism). Especially in the social sciences it is common practice to use analysis of variance (ANOVA) [40] for analyzing experimental data, and ANOVA is considered to be a special case of linear regression [90]. Decision trees have been used extensively in early machine learning (see e.g., [119] and [91]), and [62] propose a method for converting simple neural network to decision trees. Similarly, decision sets [65] or decision lists [114] can be compiled from data. Decision lists are extended into Bayesian Rule Lists in [68], which they use for building an understandable stroke prediction

model. Note that conceptually many of these approaches are closely related to association rule mining. Another interesting option that has been explored in computational psychiatry [1] is to convert alternative hypothesis into computational models and then fit the given data with these different models to see which model (and therefore hypothesis) is closer to the truth. In pipeline approaches, specific mid-level features can be used to simplify understanding of the model's output. For example, [13, 21] take the detour of recognizing adjective-noun combinations in images for the task of visual sentiment detection, and [12] propose a list of visual concepts to be used as intermediate features for classifying multimodal tweets of presumably gang-associated youth. Explicit attention mechanisms were mentioned above as one way to include understandable components into architectures. Such attention mechanisms are frequently used in machine translation [78], are a key component of memory networks [136, 146], and have been used for tasks such as image captioning as well [147]. A related approach is that of [152], which explains how to modify CNN architectures such that learned filters are more semantically meaningful and understandable.

Ablation studies. The principle of ablation studies is to gain understanding of the role of a system's components by analyzing how the overall system changes if the component is removed. Historically, in neuroscience many early insights about functionality of individual brain regions were obtained by examining changes resulting from brain damage in particular areas [56]. In computer science, ablation studies have been adopted for quantifying the importance of model components [83], which for example can be used for model verification or reduction. For interpretation analysis, ablation can be a useful tool when applied at the input level to address two points: First, which parts of the input are necessary for approximating the perspective of interest? If prediction performance drops drastically after removing a certain feature from the input, the feature was important for learning. This principle is frequently made use of in NLP for analyzing the role of features for prediction (e.g., for identifying hate speech [124]). Second, when having trained a model for perspective approximation, one might want to verify that the model does not use any parts of the input which it should not use (e.g., cause they might be known not to be used by the original interpretation function). For example, [54] uses an ablation study where they mask the foreground to confirm that the classifier does not cheat by predicting from background properties.

Usage. In principle, model-based approaches can be used to learn complex dependencies, and heatmapping can explain decisions in individual cases, even when training models directly on pixel data [88] or word sequences [6]. Heatmapping has been applied together with several other features too, such as bag of visual words [7] and fisher vectors [66]. In general, model-based explanations were found to be useful in many publications (e.g., [125, 135, 159]). Zhou et al. [157] also show how a network can learn to localize objects with decent performance without any bounding box labels. Still, in general it is not clear which properties of the original perspective carry over to the trained model when fitting it on a given list of inputs and outputs, and to the best of our knowledge, there is no extensive study analyzing the transfer of various functional properties. Indeed, publications dealing with adversarial noise (e.g., [41, 99, 100]) show how convolutional neural networks are typically sensitive to things which humans are not [95, 98, 138], despite being trained on large amounts of humanly annotated data and convolutional neural networks originally being inspired by human vision [67]. This gives reason for caution when making claims about the original function based on analyzing its approximation, especially for complex approximation methods such as deep neural networks. If the models are simpler and do not have the capacity for picking up on any complex noise, some of these issues can be ruled out and the approach becomes closer to statistical testing. Other partial remedies are to rely on pipeline approaches, where individual steps can be verified separately, or make use of ablation studies to rule out certain unwanted properties. Still, one should not confuse the trained model with the original perspective of interest, and be aware that there often is a remaining risk that findings are unreliable.

4.4 Visualization techniques

In a broad sense, the goal of visualization is to obtain a condensed representation of some given data. This representation can take various forms: The data can be transformed into a lower-dimensional space such that it can be plotted. Other methods stay closer to the original type of data and rather reduce the amount of information in different ways. Note that ultimately, in interpretation analysis we are not interested in merely visualizing the collection of inputs or outputs, but to do so in a way that shows relations between them. There are three main ways how this can be achieved: 1) If we want to apply dimensionality reduction to the input, the associated values can directly be incorporated into the visualization, e.g., by using colors to indicate different associated output values. 2) For applying dimensionality reduction to the output, if we have (short) text data or images as input data it is possible to show the original inputs at the locations of their corresponding output embeddings. 3) Finally, for exemplar-based approaches and text summarization, input samples can be partitioned based on associated values for separate visualization and successive comparison of results.

Dimensionality reduction. A very general method for visualizing almost any kind of data is to reduce the data dimension and then plot it so that it can be manually inspected. There are many different kinds of dimensionality reduction and several surveys have been made on the topic [26, 46, 130]. Here, we outline a few popular cases that are especially relevant for interpretation analysis. Linear dimensionality reduction refers to methods that linearly transform the original input space, i.e., they describe how to find a matrix that is multiplied to all inputs for projecting them into a smaller space (see [26]). Popular methods that fall into this category are Principal Component Analysis (PCA) [104], Linear Discriminant Analysis (LDA) (e.g., [84]), and Canonical Correlation Analysis (CCA) [51], which all compute orthogonal matrices for the transformation. LDA uses associated class labels and transforms the input space such that after transformation the separation between the classes is maximized. This is closely related to linear regression, which can be seen as another linear dimensionality reduction technique that does not use an orthogonality constraint. An interesting property of PCA is that after transformation the components are linearly uncorrelated or, in other words, the data is factorized into independent components. Other popular factorization methods include Factor Analysis

(FA) [131], which is widely used in psychology [37], for example to become aware of patterns in questionnaire items [16].

Linear dimensionality reduction with orthogonal matrices can be especially helpful for getting a rough idea of the data's structure, since they do not exaggerate relations between data points (see [26]). Projections of non-linear transformation techniques can be harder to interpret since geometric properties like distances in the original space are generally not preserved. Still, such techniques can be useful for looking at specific properties of the data, and there are a few non-linear transformation techniques that deserve mentioning: t-SNE [79] is a probabilistic method that embeds samples into a low-dimensional space such that similar samples are likely to be embedded to nearby points and dissimilar object to distant points. Another non-linear reduction technique is to train an autoencoder [14] to compress the original data into a smaller latent encoding. The benefit of autoencoders is that they can be combined with other loss functions for enforcing other properties on these encodings, such as following a certain distribution [82] or using specific positions to encode certain semantic properties [53].

Exemplar-based approaches. The idea behind exemplar-based approaches is that even for large collections, looking at characteristic examples and integrating them can be useful to form a holistic understanding of the collection. The crux herein is to select the right examples (and know how many are necessary), for which various approaches exist.

A very simple and yet useful method is to randomly select a few samples for manual inspection. This is likely not going to be enough to fully understand the sample collection but helps to form an initial feeling for the data. One issue is the possibility that by chance odd samples are drawn, which are included in the data, but exhibit certain unexpected properties. Obtaining such abnormal examples can also be done on purpose, which relates to a common task called anomaly detection (see e.g., [20]). Anomalies can for example help to become aware of problems with the data (e.g., broken entries), but can also be of particular relevance when working with methods that are sensitive to statistical outliers (e.g., linear regression).

There are other ways how samples can "stand out" and hence be interesting to look at. For example, the sample which is closest to the average over all samples can be seen as most representative of the whole set, or, if there are different output scores, it is sensible to look at a few samples with different scores. Other sophisticated methods exist to obtain representative and diverse examples for visualizing sample collections. For image collections, summarization is most commonly done by selecting representative examples. For example, in [140] the selection of representative images is formulated as optimization problem and mixtures of submodular functions are learned for scoring selections. [154] extract SIFT features and use a modification of RANSAC [39] plus Affinity Propagation clustering [43] for finding representative images. If there is accompanying textual or social information for the images, other approaches exist (e.g., see [18, 55, 121]).

Text summarization. For textual data, visualization and summarization techniques have been extensively surveyed [5, 27, 44, 64, 92], and it is commonly distinguished between extractive techniques and abstractive techniques. Extractive summarization techniques aim at compiling a list of sentences (examples) that summarize the collection. Abstractive summarization techniques include the extraction of topic words, frequency-driven approaches such as tf-idf, and automatic summarization. It is important to note that in our context, we generally not only want to summarize all the given inputs, but summarize in a way that reveals differences between between inputs associated with different outputs. Specific works on discriminative text summarization include [143], which explains how to select discriminative sentences for summarizing differences between text collections, and [50], which aims at visualizing differences between text corpora based on discriminative words or by analyzing an SVM that was trained to detect the source of the text.

Usage. Visualization techniques can be very beneficial for an intuitive understanding of perspective, and can often serve as useful starting point for getting ideas about which features to explore or which types of hypotheses to test with quantitative methods. On the downside, it is hard to draw any concrete conclusions from visualizations alone.

5 COMPARING MULTIPLE PERSPECTIVES

In many application scenarios one is interested in analyzing differences between perspectives. For example, one might be interested in the difference between two given machine learning classifiers, understanding how distinct annotators label data differently, comparing a classifier's perspective to the ground truth human perspective, or analyzing in which ways data from different domains relates to interpretation-related discrepancies. Even if one is not directly interested in such a comparative study and the ultimate interest is only in understanding one given perspective, it makes sense to compare against a baseline perspective for making results easier to interpret. For example, what does it mean that the user in our toy example (Table 2) seems to like explosions in images? Perhaps everyone likes explosions and what is really special about this user's interpretation is that nudity does not seem to affect the preference at all?

So, assume we are given several lists of inputs and their corresponding outputs, each list being associated to one perspective, and we want to characterize in which ways the underlying interpretation functions are different. For example, in our image preference scenario, we can imagine to be given similar tables from other users and want to see how their preferences differ. An overview of possibilities for comparison can be found in Figure 2.

Statements about a population's perspective. Note that deriving insights about perspectives of groups (e.g., all users of a website) would require statistics on another level. The number of individuals from the population needs to be high enough for such an analysis. In the given paper, we focus on understanding a few given perspectives and do not intend to generalize to any population.

5.1 Comparing input-output dependencies

Most recent papers that aim at explaining differences of machine learning models first analyze input-output dependencies by using model-based approaches mentioned above (Section 4.3), and then compare the results, typically by displaying them side by side (see [89]). Such an approach of separately analyzing individual perspectives followed by comparison can be seen as direct attempt to

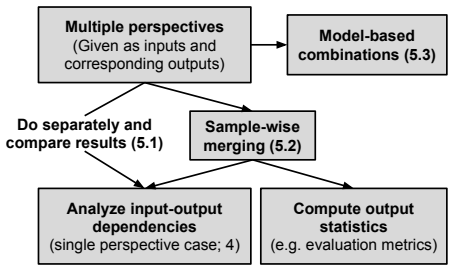

Figure 2: Approaches for comparing multiple ways of interpretation. We can distinguish between three possibilities, out of which two mainly reduce the comparison problem to the analysis of a single perspective.

answer the question "How do relations between inputs and outputs differ across the given perspectives?"

Conceptually this offers a simple way to compare, but can suffer from several issues: Findings for the different ways of interpretation might be very similar and differences not at all apparent. Also, this approach does not scale well for a growing number of perspectives, and if there are many interpretation functions, but only little data for each, analyzing individual perspectives might not give any significant results. Despite these potential shortcomings, there are cases where it makes perfect sense to analyze perspectives separately and then compare. In particular, if for the different perspectives we are given interpretation on disjoint sets of input samples, other approaches might not be directly applicable. Plus, often there is an interest in understanding individual ways of interpretation as well.

5.2 Sample-wise combinations

We can phrase the slightly different question "How do inputs relate to differences in the outputs?" Let us first assume we have function values from two different interpretation functions f_1, f_2 for the same set of input samples $i_1, \ldots, i_n$. We can easily define a new perspective f that is described by the same input samples and their associated outputs $d(f_1(i_1), f_2(i_1)), \ldots, d(f_1(i_n), f_2(i_n))$, where d is any real-valued vector function that calculates a difference or distance between two values, e.g., $d(y_1, y_2) = |y_1 - y_2|$. Thereby, the function d should be chosen depending on the overall goal: If one is only interested in finding out explanations for when there is disagreement between the two perspectives, one might want to choose a binary indicator of equality, or the absolute value of the difference between both outputs. If the goal is to also understand the direction of disagreement, the mere difference without absolute value is more suitable. For example, if we are given two computer models A and B for sentence-level sarcasm detection, we might ask which features of the sentence are related to any disagreement between A and B (binary case), but we can also analyze which features make model A but not B vote for sarcasm.

Irrespective of the choice of the merging function d, this resulting perspective f can be analyzed as in the single perspective case. This is a straight-forward way to directly analyze differences between ways of interpretation and checking statistical significance works in the same way as for a single perspective. For such a merged perspective, output statistics can be computed too, for example in order to evaluate a learned perspective f_1 against a target perspective f_2. The case of comparing more than two interpretation functions can be handled analogously.

5.3 Model-based combinations

Another possibility is to combine several perspectives in a specific model. A simple case would be the use of ANOVA, with interpretation output as dependent variable and both input features and identifier of the interpretation function (or features that group them, such as demographic information) as independent variables. ANOVA would then tell us whether there is a significant difference among average output values across the perspectives.

There are more complex possibilities. Typically, the main goal of these approaches is not to analyze ways of interpretation, but to learn how to combine multiple perspectives for a given prediction task. Still, characteristic information about the involved perspectives is often incorporated into these models. An early example of such a probabilistic model for combining human perspectives is the Dawid-Skene model [28], which unites observations from different sources while estimating the observers' errors. For AI approaches, ensemble methods are frequently used for increasing predictive performance [116]. These methods often include a scoring mechanism or allow for similar ways of obtaining an estimation of the usefulness of the individual models involved, which can be seen as discriminative characterization.

6 APPLICATIONS

Tools for interpretation analysis can be utilized in a variety of scenarios. In the following, we outline some of the cornerstones.

6.1 Mining subjective information

Prominent examples of applications that aim at mining subjective information from text data are sentiment analysis and opinion mining [76, 110]. The main task of sentiment analysis is to decide whether a given text expresses a positive, a negative, or a neutral opinion, which can for example be useful for evaluating customer reviews. In its original form, sentiment analysis is about learning a way of interpretation, but does not necessarily involve any claims about characteristics of the same. However, it is very common to not simply detect overall sentiment, but to do so based on aspects. The resulting detection pipeline then has aspect information as extra component, and tries to explain the overall sentiment in terms of mentioned aspects and the orientation expressed towards these. For understanding persisting differences in interpretation, contrastive opinion mining has been proposed by Fang et al. [38] and, later, perspective detection by Vilares and He [142]. The Latent Argument Model in [142] is a rather complex case of discriminative text summarization based on topic modeling, and is paired in the paper with selection of characteristic sentences. Note that sentiment analysis was extended to the visual modality as well. Somewhat similar to aspect-based sentiment detection, Borth et al. proposed a visual sentiment ontology [13] consisting of adjective-noun combinations (e.g., "scary dog", "cute baby") that are visually detectable and can be used for explaining the overall sentiment of an image.

Quite a different approach is taken in [70], which analyzes how hotel preferences change over time by applying emerging pattern mining on hotel features mentioned in online reviews.

6.2 Model analysis

Much recent work was done on analyzing deep learning models and explaining decisions based on heatmapping (e.g., [58, 59, 88, 159]). These are all direct cases of model-based interpretation analysis (usually not operating under the same black-box assumption though). Visualization techniques have been used as well for examining learned representations of neural networks. For example, [22] use t-SNE on phrase embeddings (which can be seen as output of the model's interpretation function) to analyze how semantically meaningful the learned embeddings are.

Note that computation of many performance metrics can be seen as special case of interpretation analysis, where the output of a classifier is compared to a ground truth human interpretation by merging both perspectives in a sample-wise manner and then aggregating over the outputs of this combined perspective.

6.3 Annotation

Computer vision in particular depends on big amounts of manually labelled data for training models, which is often achieved via crowdsourcing [61]. In crowdsourcing, it is common to collect several annotations for each item, and many probabilistic models for merging annotator votes have been proposed (e.g., [15, 111, 115, 141, 148]). Often, these simultaneously estimate annotator reliability, but only a few approaches consider item difficulty and thereby relate disagreements to the input. Notable exceptions are [15] and its extension [141], which describe such a probabilistic framework and apply their framework to merge fine-grained bird image annotations. Less work has been done on investigating where annotator disagreements come from. For crowdsourcing, Eickhoff [32] outlines several quality issues and performs dedicated experiments for analyzing cognitive biases of annotators. The paper also shows how such biases can propagate into model evaluation and hence have detrimental consequences, which gives reason for further investigation into a more fine-grained interpretation analysis for annotation.

6.4 Data understanding and expertise

A general goal in science is to understand the relation between two quantities based on some given data, for which interpretation analysis tools can be applied. For example, association rule mining has been used for making sense of gene expression data [23] and medical data [97]. Emerging pattern mining for finding differences between toxical vs non-toxical chemicals [126]. Visual pattern mining for histology image collections is done in [24] for identifying local features that can be used to discriminate between tissue types. The same paper also estimates posterior probabilities for relating local features to individual tissue types for interpretation. Numerous attempts at data explanation have also been made by fitting various models on the given data and then analyzing the trained models for insights. In [125], a deep tensor neural network model with heatmapping was applied to examine the link from molecular structure to electronic properties. And [135] reported LRP-based explanations for classifying EEG data with a neural network to be highly plausible. Essentially, such cases can be seen as figuring out some "natural" way of interpretation that is intrinsic to the given data. In the special case when the output quantity is given in the form of labels from human experts, analyzing the data amounts to explaining their expert view, or in other words, to characterize an expert's way of interpretation. Note, however, that for data understanding our black-box assumption (see Section 2.2) is generally satisfied, so care has be taken when interpreting the trained model.

6.5 Understanding human interpretation

Mechanisms and properties of human interpretation are of fundamental interest in several fields, including cognitive science, neuroscience, phenomenology, linguistics, psychology and psychiatry. Traditionally, these fields often conduct designated controlled experiments for data collection, or use qualitative analysis when relying on given observational data. Still, there are some approaches that are more in between the fields mentioned above and computer science. These include recent works on computational psychiatry [1, 87, 134] which turn hypothesis about human functioning into simple computational models that can be evaluated on experimental or observational data. For example, [1] explains how to use a hierarchical generative model for exploring potential relations between over-attention to low-level stimuli and schizophrenia.

7 CONCLUSION

In this paper, we proposed a theoretical framework in which we formally defined interpretation, perspective and the task of interpretation analysis. In our framework, interpretation analysis can be understood as characterizing functions and describes relations between inputs and corresponding outputs. We showed how analyzing a single way of interpretation can be approached under the use of statistical methods, pattern mining techniques, model-based approaches and visualization techniques. We discussed how comparing several ways of interpretation can often be reduced to the single perspective case, and alternatively be handled by uniting perspectives in a designated model for analysis. Finally, we have seen applications from several areas, including opinion mining, annotation and analysis of machine learning models, which can be connected by their relations to interpretation analysis.

During our survey of approaches, we identified several points that deserve more attention in the future. In particular, proper evaluation of interpretation analysis methods is still largely an open issue. This holds true especially for more complex model-based approaches under our black-box assumption (generally satisfied when using them data understanding) and visualization techniques. Further, there are many qualitative methods that are relevant to interpretation analysis which we hope can further inspire computational methods in the future. Similarly, though we have already drawn many connections between literature from the fields of behavioural sciences, psychology and computer science in this paper, we hope to see more work in the fruitful intersection of these fields in the future. Last but not least, we see an ever increasing need for ethical discussions: Many application areas of interpretation analysis ethically concern user privacy. Similar techniques to the

ones described have in the recent past already been used for ethically very questionable goals under the term microtargeting (e.g., to influence the outcome of elections [160]). Our hope is that the scientific community will in the future focus on using the same techniques for ethically less questionable goals, for example to increase transparency and explainability of AI systems and maybe even to help us become aware of our own detrimental biases.

ACKNOWLEDGMENTS

This work was supported by the BMBF project DeFuseNN (Grant 01IW17002) and the NVIDIA AI Lab (NVAIL) program. Furthermore, the first author received financial support from the Center for Cognitive Science, Kaiserslautern, Germany.

REFERENCES

[1] Rick A Adams, Quentin JM Huys, and Jonathan P Roiser. 2016. Computational psychiatry: towards a mathematically informed understanding of mental illness. *J Neurol Neurosurg Psychiatry* 87, 1 (2016), 53–63.

[2] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. 1993. Mining association rules between sets of items in large databases. In *Acm sigmod record*, Vol. 22. ACM, 207–216.

[3] Rakesh Agrawal, Ramakrishnan Srikant, et al. 1994. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, Vol. 1215. 487–499.

[4] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. 2015. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nature biotechnology* 33, 8 (2015), 831.

[5] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. 2017. Text summarization techniques: a brief survey. *arXiv preprint arXiv:1707.02268* (2017).

[6] Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2017. Explaining Recurrent Neural Network Predictions in Sentiment Analysis. In *Proceedings of the EMNLP 2017 Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics, 159–168. http://aclweb.org/anthology/W/W17/W17-5221.pdf

[7] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one* 10, 7 (2015), e0130140.

[8] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert MÃžller. 2010. How to explain individual classification decisions. *Journal of Machine Learning Research* 11, Jun (2010), 1803–1831.

[9] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Network dissection: Quantifying interpretability of deep visual representations. *arXiv preprint arXiv:1704.05796* (2017).

[10] Stephen Bazen and Xavier Joutard. 2013. The Taylor decomposition: A unified generalization of the Oaxaca method to nonlinear models. (2013).

[11] Pietro Berkes and Laurenz Wiskott. 2006. On the analysis and interpretation of inhomogeneous quadratic forms as receptive fields. *Neural computation* 18, 8 (2006), 1868–1895.

[12] Philipp Blandfort, Desmond Patton, William R Frey, Svebor Karaman, Surabhi Bhargava, Fei-Tzin Lee, Siddharth Varia, Chris Kedzie, Michael B Gaskell, Rossano Schifanella, et al. 2018. Multimodal Social Media Analysis for Gang Violence Prevention. *arXiv preprint arXiv:1807.08465* (2018).

[13] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. 2013. Large-scale Visual Sentiment Ontology and Detectors Using Adjective Noun Pairs. In *Proceedings of the 21st ACM International Conference on Multimedia (MM '13)*. ACM, New York, NY, USA, 223–232. https://doi.org/10.1145/2502081.2502282

[14] Hervé Bourlard and Yves Kamp. 1988. Auto-association by multilayer perceptrons and singular value decomposition. *Biological cybernetics* 59, 4-5 (1988), 291–294.

[15] Steve Branson, Grant Van Horn, and Pietro Perona. 2017. Lean crowdsourcing: Combining humans and machines in an online system. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7474–7483.

[16] Stephen R Briggs and Jonathan M Cheek. 1986. The role of factor analysis in the development and evaluation of personality scales. *Journal of personality* 54, 1 (1986), 106–148.

[17] Sergey Brin, Rajeev Motwani, Jeffrey D Ullman, and Shalom Tsur. 1997. Dynamic itemset counting and implication rules for market basket data. *Acm Sigmod Record* 26, 2 (1997), 255–264.

[18] Jorge E Camargo and Fabio A González. 2016. Multimodal latent topic analysis for image collection summarization. *Information Sciences* 328 (2016), 270–287.

[19] Margarita Vázquez Campos and Antonio Manuel Liz Gutiérrez. 2015. The notion of point of view. In *Temporal Points of View*. Springer, 1–57.

[20] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. *ACM computing surveys (CSUR)* 41, 3 (2009), 15.

[21] Tao Chen, Damian Borth, Trevor Darrell, and Shih-Fu Chang. 2014. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv preprint arXiv:1410.8586* (2014).

[22] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[23] Chad Creighton and Samir Hanash. 2003. Mining gene expression databases for association rules. *Bioinformatics* 19, 1 (2003), 79–86.

[24] Angel Cruz-Roa, Juan C Caicedo, and Fabio A González. 2011. Visual pattern mining in histology image collections using bag of features. *Artificial intelligence in medicine* 52, 2 (2011), 91–106.

[25] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. 2004. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, Vol. 1. Prague, 1–2.

[26] John P Cunningham and Zoubin Ghahramani. 2015. Linear dimensionality reduction: Survey, insights, and generalizations. *The Journal of Machine Learning Research* 16, 1 (2015), 2859–2900.

[27] Dipanjan Das and André FT Martins. 2007. A survey on automatic text summarization. *Literature Survey for the Language and Statistics II course at CMU* 4 (2007), 192–195.

[28] Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied statistics* (1979), 20–28.

[29] Pedro Domingos. 2012. A few useful things to know about machine learning. *Commun. ACM* 55, 10 (2012), 78–87.

[30] Guozhu Dong and Jinyan Li. 1999. Efficient mining of emerging patterns: Discovering trends and differences. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 43–52.

[31] Guozhu Dong, Xiuzhen Zhang, Limsoon Wong, and Jinyan Li. 1999. CAEP: Classification by aggregating emerging patterns. In *International Conference on Discovery Science*. Springer, 30–42.

[32] Carsten Eickhoff. 2018. Cognitive Biases in Crowdsourcing. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 162–170.

[33] Magdalini Eirinaki and Michalis Vazirgiannis. 2003. Web mining for web personalization. *ACM Transactions on Internet Technology (TOIT)* 3, 1 (2003), 1–27.

[34] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2009. Visualizing higher-layer features of a deep network. *University of Montreal* 1341, 3 (2009), 1.

[35] Sergio Escalera, Xavier Baró, Hugo Jair Escalante, and Isabelle Guyon. 2017. Chalearn looking at people: A review of events and resources. In *Neural Networks (IJCNN), 2017 International Joint Conference on*. IEEE, 1594–1601.

[36] EU Council. 2016. EU Regulation 2016/679 General Data Protection Regulation (GDPR). *Official Journal of the European Union* 59 (2016), 1–88. http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX{%}3A32016R0679

[37] Leandre R Fabrigar, Duane T Wegener, Robert C MacCallum, and Erin J Strahan. 1999. Evaluating the use of exploratory factor analysis in psychological research. *Psychological methods* 4, 3 (1999), 272.

[38] Yi Fang, Luo Si, Naveen Somasundaram, and Zhengtao Yu. 2012. Mining Contrastive Opinions on Political Texts Using Cross-perspective Topic Model. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining (WSDM '12)*. ACM, New York, NY, USA, 63–72. https://doi.org/10.1145/2124295.2124306

[39] Martin A Fischler and Robert C Bolles. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* 24, 6 (1981), 381–395.

[40] Ronald A Fisher. 1921. On the probable error of a coefficient of correlation deduced from a small sample. *Metron* 1 (1921), 3–32.

[41] Joachim Folz, Sebastian Palacio, Jörn Hees, Damian Borth, and Andreas Dengel. 2018. Adversarial Defense based on Structure-to-Signal Autoencoders. *CoRR* abs/1803.07994 (2018). arXiv:1803.07994 http://arxiv.org/abs/1803.07994

[42] Joseph P Forgas, Kipling D Williams, Simon M Laham, William Von Hippel, et al. 2005. *Social motivation: Conscious and unconscious processes*. Vol. 5. Cambridge University Press.

[43] Brendan J Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *science* 315, 5814 (2007), 972–976.

[44] Mahak Gambhir and Vishal Gupta. 2017. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review* 47, 1 (2017), 1–66.

[45] A.M. García-Vico, C.J. Carmona, D. Martín, M. García-Borroto, and M.J. del Jesus. 2018. An overview of emerging pattern mining in supervised descriptive rule discovery: taxonomy, empirical study, trends, and

prospects. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8, 1 (2018), e1231. https://doi.org/10.1002/widm.1231 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1231

[46] Andrej Gisbrecht and Barbara Hammer. 2015. Data visualization by nonlinear dimensionality reduction. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 5, 2 (2015), 51–73.

[47] Bryce Goodman and Seth Flaxman. 2016. EU regulations on algorithmic decision-making and a "right to explanation". In *ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*. 1–9. arXiv:1606.08813 http://arxiv.org/abs/1606.08813

[48] Antonio Manuel Liz Gutiérrez and Margarita Vázquez Campos. 2015. Subjective and Objective Aspects of Points of View. In *Temporal Points of View*. Springer, 59–104.

[49] Jiawei Han, Jian Pei, and Yiwen Yin. 2000. Mining frequent patterns without candidate generation. In *ACM sigmod record*, Vol. 29. ACM, 1–12.

[50] Franziska Horn, Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2017. Exploring text datasets by visualizing relevant words. *CoRR* abs/1707.05261 (2017). arXiv:1707.05261 http://arxiv.org/abs/1707.05261

[51] Harold Hotelling. 1936. Relations between two sets of variates. *Biometrika* 28, 3/4 (1936), 321–377.

[52] Patrik O. Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. 2009. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems 21*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou (Eds.). Curran Associates, Inc., 689–696. http://papers.nips.cc/paper/3548-nonlinear-causal-discovery-with-additive-noise-models.pdf

[53] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. *arXiv preprint arXiv:1703.00955* (2017).

[54] Omer Ishaq, Sajith Kecheril Sadanandan, and Carolina Wählby. 2017. Deep Fish: Deep Learning–Based Classification of Zebrafish Deformation for High-Throughput Screening. *SLAS DISCOVERY: Advancing Life Sciences R&D* 22, 1 (2017), 102–107. https://doi.org/10.1177/1087057116667894 arXiv:https://doi.org/10.1177/1087057116667894 PMID: 27613194.

[55] Alexandar Jaffe, Mor Naaman, Tamir Tassa, and Marc Davis. 2006. Generating summaries and visualization for large collections of geo-referenced photographs. In *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*. ACM, 89–98.

[56] Eric R Kandel, James H Schwartz, Thomas M Jessell, Department of Biochemistry, Molecular Biophysics Thomas Jessell, Steven Siegelbaum, and AJ Hudspeth. 2000. *Principles of neural science*. Vol. 4. McGraw-hill New York.

[57] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. 2017. The (Un) reliability of saliency methods. *arXiv preprint arXiv:1711.00867* (2017).

[58] Pieter-Jan Kindermans, Kristof T Schütt, Maximilian Alber, Klaus-Robert Müller, and Sven Dähne. 2017. PatternNet and PatternLRP–improving the interpretability of neural networks. *stat* 1050 (2017), 16.

[59] Pieter-Jan Kindermans, Kristof T Schütt, Maximilian Alber, Klaus-Robert Müller, Dumitru Erhan, Been Kim, and Sven Dähne. 2017. Learning how to explain neural networks: PatternNet and PatternAttribution. *arXiv preprint arXiv:1705.05598* (2017).

[60] Andrea Knezevic. 2008. Overlapping confidence intervals and statistical significance. *StatNews: Cornell University Statistical Consulting Unit* 73, 1 (2008).

[61] Adriana Kovashka, Olga Russakovsky, Li Fei-Fei, Kristen Grauman, et al. 2016. Crowdsourcing in computer vision. *Foundations and Trends® in Computer Graphics and Vision* 10, 3 (2016), 177–243.

[62] R. Krishnan, G. Sivakumar, and P. Bhattacharya. 1999. Extracting decision trees from trained neural networks. *Pattern Recognition* 32, 12 (1999), 1999 – 2009. https://doi.org/10.1016/S0031-3203(98)00181-2

[63] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[64] Kostiantyn Kucher and Andreas Kerren. 2015. Text visualization techniques: Taxonomy, visual survey, and community insights. In *Visualization Symposium (PacificVis), 2015 IEEE Pacific*. IEEE, 117–121.

[65] Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. 2016. Interpretable Decision Sets: A Joint Framework for Description and Prediction. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM, New York, NY, USA, 1675–1684. https://doi.org/10.1145/2939672.2939874

[66] Sebastian Lapuschkin, Alexander Binder, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2016. Analyzing Classifiers: Fisher Vectors and Deep Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2912–20. https://doi.org/10.1109/CVPR.2016.318

[67] Yann LeCun, Yoshua Bengio, et al. 1995. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks* 3361, 10 (1995), 1995.

[68] Benjamin Letham, Cynthia Rudin, Tyler H McCormick, David Madigan, et al. 2015. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics* 9, 3 (2015), 1350–1371.

[69] Daniel J Levitin. 2002. Experimental design in psychological research. In *Foundations of cognitive psychology: Core readings*. MIT Press, 115–130.

[70] Gang Li, Rob Law, Huy Quan Vu, Jia Rong, and Xinyuan Roy Zhao. 2015. Identifying emerging hotel preferences using Emerging Pattern Mining technique. *Tourism management* 46 (2015), 311–321.

[71] Hongzhi Li, Joseph G. Ellis, Lei Zhang, and Shih-Fu Chang. 2018. PatternNet: Visual Pattern Mining with Deep Neural Network. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval (ICMR '18)*. ACM, New York, NY, USA, 291–299. https://doi.org/10.1145/3206025.3206039

[72] Jinyan Li, Guozhu Dong, and Kotagiri Ramamohanarao. 2000. Instance-based classification by emerging patterns. In *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer, 191–200.

[73] Yao Li, Lingqiao Liu, Chunhua Shen, and Anton van den Hengel. 2017. Mining Mid-level Visual Patterns with Deep CNN Activations. *International Journal of Computer Vision* 121, 3 (01 Feb 2017), 344–364. https://doi.org/10.1007/s11263-016-0945-y

[74] Yao Li, Lingqiao Liu, Chunhua Shen, and Anton van den Hengel. 2015. Mid-level deep pattern mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 971–980.

[75] Zachary C. Lipton. 2018. The Mythos of Model Interpretability. *Queue* 16, 3, Article 30 (June 2018), 27 pages. https://doi.org/10.1145/3236386.3241340

[76] Bing Liu and Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. In *Mining text data*. Springer, 415–463.

[77] David Lopez-Paz, Krikamol Muandet, Bernhard Schölkopf, and Iliya Tolstikhin. 2015. Towards a learning theory of cause-effect inference. In *International Conference on Machine Learning*. 1452–1461.

[78] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* (2015).

[79] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[80] Aravindh Mahendran and Andrea Vedaldi. 2015. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5188–5196.

[81] Aravindh Mahendran and Andrea Vedaldi. 2016. Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision* 120, 3 (2016), 233–255.

[82] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. 2015. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644* (2015).

[83] Cynthia Matuszek, Nicholas FitzGerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. 2012. A joint model of language and perception for grounded attribute learning. *arXiv preprint arXiv:1206.6423* (2012).

[84] Geoffrey McLachlan. 2004. *Discriminant analysis and statistical pattern recognition*. Vol. 544. John Wiley & Sons.

[85] Bamshad Mobasher, Robert Cooley, and Jaideep Srivastava. 2000. Automatic personalization based on web usage mining. *Commun. ACM* 43, 8 (2000), 142–151.

[86] Bamshad Mobasher, Honghua Dai, Tao Luo, and Miki Nakagawa. 2001. Effective personalization based on association rule discovery from web usage data. In *Proceedings of the 3rd international workshop on Web information and data management*. ACM, 9–15.

[87] P. Read Montague, Raymond J. Dolan, Karl J. Friston, and Peter Dayan. 2012. Computational psychiatry. *Trends in Cognitive Sciences* 16, 1 (2012), 72 – 80. https://doi.org/10.1016/j.tics.2011.11.018 Special Issue: Cognition in Neuropsychiatric Disorders.

[88] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. 2017. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition* 65 (2017), 211–222.

[89] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2018. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing* 73 (2018), 1 – 15. https://doi.org/10.1016/j.dsp.2017.10.011

[90] Douglas C Montgomery. 2017. *Design and analysis of experiments*. John wiley & sons.

[91] Sreerama K Murthy. 1998. Automatic construction of decision trees from data: A multi-disciplinary survey. *Data mining and knowledge discovery* 2, 4 (1998), 345–389.

[92] Ani Nenkova and Kathleen McKeown. 2012. A survey of text summarization techniques. In *Mining text data*. Springer, 43–76.

[93] Anh Nguyen, Jeff Clune, Yoshua Bengio, Alexey Dosovitskiy, and Jason Yosinski. 2017. Plug & Play Generative Networks: Conditional Iterative Generation of Images in Latent Space.. In *CVPR*, Vol. 2. 7.

[94] Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. 2016. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In *Advances in Neural Information Processing Systems*.

3387–3395.
[95] Anh Nguyen, Jason Yosinski, and Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 427–436.
[96] Petra Kralj Novak, Nada Lavrač, and Geoffrey I Webb. 2009. Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research* 10, Feb (2009), 377–403.
[97] Carlos Ordonez, Norberto Ezquerra, and Cesar A Santana. 2006. Constraining and summarizing association rules in medical data. *Knowledge and information systems* 9, 3 (2006), 1–2.
[98] Sebastian Palacio, Joachim Folz, Jörn Hees, Federico Raue, Damian Borth, and Andreas Dengel. 2018. What Do Deep Networks Like to See?. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
[99] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. 2017. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*. ACM, 506–519.
[100] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. 2016. The limitations of deep learning in adversarial settings. In *Security and Privacy (EuroS&P), 2016 IEEE European Symposium on*. IEEE, 372–387.
[101] Sobhan Naderi Parizi, Andrea Vedaldi, Andrew Zisserman, and Pedro Felzenszwalb. 2014. Automatic discovery and optimization of parts for image classification. *arXiv preprint arXiv:1412.6598* (2014).
[102] Judea Pearl. 2009. Causal inference in statistics: An overview. *Statist. Surv.* 3 (2009), 96–146. https://doi.org/10.1214/09-SS057
[103] Karl Pearson. 1895. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London* 58 (1895), 240–242.
[104] Karl Pearson. 1901. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2, 11 (1901), 559–572.
[105] Hanchuan Peng, Fuhui Long, and Chris Ding. 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence* 27, 8 (2005), 1226–1238.
[106] Gregory Piateski and William Frawley. 1991. *Knowledge discovery in databases*. MIT press.
[107] Jacob Poushter et al. 2016. Smartphone ownership and internet usage continues to climb in emerging economies. *Pew Research Center* 22 (2016), 1–44.
[108] Till Quack, Vittorio Ferrari, Bastian Leibe, and Luc Van Gool. 2007. Efficient mining of frequent and distinctive feature configurations. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 1–8.
[109] Peter M Rasmussen, Tanya Schmah, Kristoffer H Madsen, Torben E Lund, Stephen C Strother, and Lars K Hansen. 2012. Visualization of nonlinear classification models in neuroimaging.
[110] Kumar Ravi and Vadlamani Ravi. 2015. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems* 89 (2015), 14–46.
[111] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010. Learning from crowds. *Journal of Machine Learning Research* 11, Apr (2010), 1297–1322.
[112] Konstantinos Rematas, Basura Fernando, Frank Dellaert, and Tinne Tuytelaars. 2015. Dataset fingerprints: Exploring image collections through data mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4867–4875.
[113] Caitlin M Rivers and Bryan L Lewis. 2014. Ethical research standards in a world of big data. *F1000Research* 3 (2014).
[114] Ronald L. Rivest. 1987. Learning Decision Lists. *Machine Learning* 2, 3 (01 Nov 1987), 229–246. https://doi.org/10.1023/A:1022607331053
[115] Filipe Rodrigues, Francisco Pereira, and Bernardete Ribeiro. 2013. Learning from multiple annotators: distinguishing good from random labelers. *Pattern Recognition Letters* 34, 12 (2013), 1428–1436.
[116] Lior Rokach. 2010. Ensemble-based classifiers. *Artificial Intelligence Review* 33, 1-2 (2010), 1–39.
[117] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *nature* 323, 6088 (1986), 533.
[118] John Ruscio. 2008. Constructing confidence intervals for Spearman's rank correlation with ordinal data: a simulation study comparing analytic and bootstrap methods. *Journal of Modern Applied Statistical Methods* 7, 2 (2008), 7.
[119] S Rasoul Safavian and David Landgrebe. 1991. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics* 21, 3 (1991), 660–674.
[120] Andrea Saltelli, Marco Ratto, Terry Andres, Francesca Campolongo, Jessica Cariboni, Debora Gatelli, Michaela Saisana, and Stefano Tarantola. 2008. *Global sensitivity analysis: the primer*. John Wiley & Sons.
[121] Zahra Riahi Samani and Mohsen Ebrahimi Moghaddam. 2017. A knowledge-based semantic approach for image collection summarization. *Multimedia Tools and Applications* 76, 9 (01 May 2017), 11917–11939. https://doi.org/10.1007/s11042-016-3840-1
[122] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. 2017. Evaluating the visualization of what a Deep Neural Network has learned. *IEEE Transactions on Neural Networks and Learning Systems* 28, 11 (2017), 2660–2673. https://doi.org/10.1109/TNNLS.2016.259982
[123] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. 2018. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *ITU Journal: ICT Discoveries - Special Issue 1 - The Impact of Artificial Intelligence (AI) on Communication Networks and Services* 1, 1 (2018), 39–48. https://www.itu.int/en/journal/001/Pages/05.aspx
[124] Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. 1–10.
[125] Kristof T Schütt, Farhad Arbabzadah, Stefan Chmiela, Klaus R Müller, and Alexandre Tkatchenko. 2017. Quantum-chemical insights from deep tensor neural networks. *Nature communications* 8 (2017), 13890.
[126] Richard Sherhod, Philip N Judson, Thierry Hanser, Jonathan D Vessey, Samuel J Webb, and Valerie J Gillet. 2014. Emerging pattern mining to aid toxicological knowledge discovery. *Journal of chemical information and modeling* 54, 7 (2014), 1864–1879.
[127] Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, and Kenneth Bollen. 2011. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research* 12, Apr (2011), 1225–1248.
[128] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).
[129] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
[130] Carlos Oscar Sánchez Sorzano, Javier Vargas, and A Pascual Montano. 2014. A survey of dimensionality reduction techniques. *arXiv preprint arXiv:1403.2877* (2014).
[131] Charles Spearman. 1904. " General Intelligence," objectively determined and measured. *The American Journal of Psychology* 15, 2 (1904), 201–292.
[132] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. 2014. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806* (2014).
[133] Oliver Stegle, Dominik Janzing, Kun Zhang, Joris M Mooij, and Bernhard Schölkopf. 2010. Probabilistic latent variable models for distinguishing between cause and effect. In *Advances in Neural Information Processing Systems*. 1687–1695.
[134] Klaas Enno Stephan and Christoph Mathys. 2014. Computational approaches to psychiatry. *Current opinion in neurobiology* 25 (2014), 85–92.
[135] Irene Sturm, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. 2016. Interpretable deep neural networks for single-trial EEG classification. *Journal of neuroscience methods* 274 (2016), 141–145.
[136] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In *Advances in neural information processing systems*. 2440–2448.
[137] AH Sung. 1998. Ranking importance of input parameters of neural networks. *Expert Systems with Applications* 15, 3-4 (1998), 405–411.
[138] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).
[139] Gábor J Székely, Maria L Rizzo, Nail K Bakirov, et al. 2007. Measuring and testing dependence by correlation of distances. *The annals of statistics* 35, 6 (2007), 2769–2794.
[140] Sebastian Tschiatschek, Rishabh Iyer, Haochen Wei, and Jeff Bilmes. 2014. Learning Mixtures of Submodular Functions for Image Collection Summarization. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'14)*. MIT Press, Cambridge, MA, USA, 1413–1421. http://dl.acm.org/citation.cfm?id=2968826.2968984
[141] Grant Van Horn, Steve Branson, Scott Loarie, Serge Belongie, and Pietro Perona. 2018. Lean Multiclass Crowdsourcing. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
[142] David Vilares and Yulan He. 2017. Detecting perspectives in political debates. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 1573–1582.
[143] Dingding Wang, Shenghuo Zhu, Tao Li, and Yihong Gong. 2012. Comparative document summarization via discriminative sentence selection. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 6, 3 (2012), 12.
[144] Geoffrey I Webb. 2007. Discovering significant patterns. *Machine learning* 68, 1 (2007), 1–33.
[145] Arnold D Well and Jerome L Myers. 2003. *Research design & statistical analysis*. Psychology Press.
[146] Jason Weston, Sumit Chopra, and Antoine Bordes. 2014. Memory Networks. *CoRR* abs/1410.3916 (2014). arXiv:1410.3916 http://arxiv.org/abs/1410.3916

[147] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*. 2048–2057.

[148] Yan Yan, Rómer Rosales, Glenn Fung, Mark Schmidt, Gerardo Hermosillo, Luca Bogoni, Linda Moy, and Jennifer Dy. 2010. Modeling annotator expertise: Learning when everybody knows a bit of something. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. 932–939.

[149] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. 2015. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579* (2015).

[150] Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*. Springer, 818–833.

[151] Kun Zhang and Aapo Hyvärinen. 2009. On the identifiability of the post-nonlinear causal model. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*. AUAI Press, 647–655.

[152] Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. 2018. Interpretable convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 8827–8836.

[153] Qiankun Zhao and Sourav S Bhowmick. 2003. Association rule mining: A survey. *Nanyang Technological University, Singapore* (2003).

[154] Ye Zhao, Richang Hong, and Jianguo Jiang. 2016. Visual summarization of image collections by fast RANSAC. *Neurocomputing* 172 (2016), 48 – 52. https://doi.org/10.1016/j.neucom.2014.09.095

[155] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu. 2012. Effective pattern discovery for text mining. *IEEE transactions on knowledge and data engineering* 24, 1 (2012), 30–44.

[156] Bolei Zhou, David Bau, Aude Oliva, and Antonio Torralba. 2018. Interpreting deep visual representations via network dissection. *IEEE transactions on pattern analysis and machine intelligence* (2018).

[157] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2921–2929.

[158] Jens Zimmermann. 2015. *Hermeneutics: A very short introduction*. OUP Oxford.

[159] Luisa M. Zintgraf, Taco S. Cohen, Tameem Adel, and Max Welling. 2017. Visualizing Deep Neural Network Decisions: Prediction Difference Analysis. *CoRR* abs/1702.04595 (2017). arXiv:1702.04595 http://arxiv.org/abs/1702.04595

[160] Frederik J Zuiderveen Borgesius, Judith Moller, Sanne Kruikemeier, Ronan Ó Fathaigh, Kristina Irion, Tom Dobber, Balazs Bodo, and Claes de Vreese. 2018. Online Political Microtargeting: Promises and Threats for Democracy. *Utrecht L. Rev.* 14 (2018), 82.

[161] Jacek M Zurada, Aleksander Malinowski, and Ian Cloete. 1994. Sensitivity analysis for minimization of input data dimension for feedforward neural network. In *Circuits and Systems, 1994. ISCAS'94., 1994 IEEE International Symposium on*, Vol. 6. IEEE, 447–450.