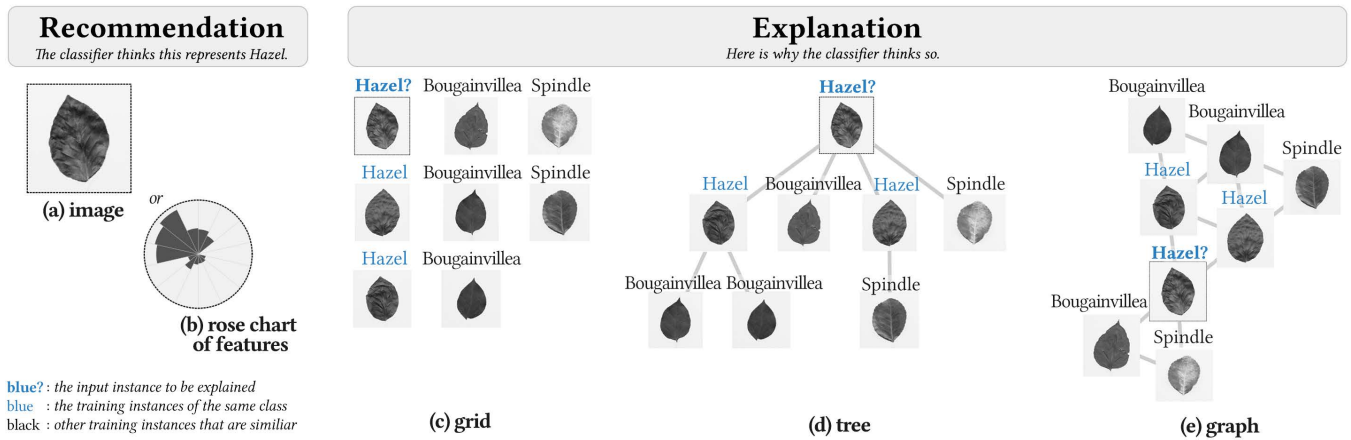


# How Do Visual Explanations Foster End Users' Appropriate Trust in Machine Learning?

Fumeng Yang\*  
Brown University  
Providence, RI, USA  
fy@brown.edu

Zhuanyi Huang  
Jean Scholtz†  
Dustin L. Arendt  
Pacific Northwest National Laboratory  
Richland, WA, USA  
{zhuanyi.huang, jean.scholtz, dustin.arendt}@pnnl.gov



**Figure 1: Examples of the visual explanations in our experiment**—We tested two ways to represent an example instance: (a) an image or (b) a rose chart of features, and three spatial layouts to arrange the instances: (c) grid, (d) tree, and (e) graph. Three images (c-e) here show explanations of the same instances, classifier, and classification recommendation.

## ABSTRACT

We investigated the effects of example-based explanations for a machine learning classifier on end users' appropriate trust. We explored the effects of spatial layout and visual representation in an in-person user study with 33 participants. We measured participants' appropriate trust in the classifier, quantified the effects of different spatial layouts and visual representations, and observed changes in users' trust over time. The results show that each explanation improved users' trust in the classifier, and the combination of explanation, human, and classification algorithm yielded much better decisions than the human and classification algorithm separately. Yet these visual explanations lead to different levels of trust and may cause inappropriate trust if an explanation is difficult to un-

\*Fumeng Yang was a PhD intern at Pacific Northwest National Laboratory when conducting this research.

†Jean Scholtz retired from Pacific Northwest National Laboratory September 2018.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

IUI '20, March 17–20, 2020, Cagliari, Italy

© 2020 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-7118-6/20/03.

<https://doi.org/10.1145/3377325.3377480>

derstand. Visual representation and performance feedback strongly affect users' trust, and spatial layout shows a moderate effect. Our results do not support that individual differences (e.g., propensity to trust) affect users' trust in the classifier. This work advances the state-of-the-art in trust-able machine learning and informs the design and appropriate use of automated systems.

## CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI; Information visualization; Empirical studies in visualization; Visualization design and evaluation methods**; • **Computing methodologies** → **Supervised learning by classification**.

## KEYWORDS

Human-machine Collaboration, Trust, Trust Calibration, Information Visualization, Explainable Artificial Intelligence, Supervised-learning, Classification

## ACM Reference Format:

Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L. Arendt. 2020. How Do Visual Explanations Foster End Users' Appropriate Trust in Machine Learning?. In *25th International Conference on Intelligent User Interfaces (IUI '20)*, March 17–20, 2020, Cagliari, Italy. ACM, New York, NY, USA, 13 pages. <https://doi.org/10.1145/3377325.3377480>

# 1 INTRODUCTION

We are now experiencing more and more machine learning techniques incorporated within automated systems, both in our personal lives and in work environments. While these systems are powerful, they do make mistakes occasionally, leading to *misuse* (“the over reliance on automation” [97]). Sometimes these systems are underestimated by people, leading to *disuse* (“the neglect or under-utilization of automation” [97]). Both disuse and misuse could cause serious problems [74, 79]. To use automation properly, users must *trust* an automated system *appropriately* [3].

A current and widely held belief is that explaining to users how a system operates will improve their trust in the system [25, 34, 99, 118] and decrease aversion after seeing an error [18]. Both verbal [21, 52, 117] and visual [14, 101] explanations were shown to provide transparency [14, 52] and increase users’ trust [117]. However, much of the existing research focused on explaining the underlying algorithm [69]; and designers may assume that users have advanced knowledge of machine learning [69], or users have to understand the decision process (e.g., [59, 67]). Besides machine learning experts, many other end users can also benefit from machine learning [41, 129], and sometimes human-machine collaborations exhibit better performance than the human or the machine alone [4, 30, 115]. Yet, end users could be domain experts who may not have the background to understand how the algorithm operates. This mismatch between designers’ assumptions and users’ background may have led to recent controversial and counter-intuitive findings (e.g., “explanations can be harmful” [57, 111]).

In this paper, we aim to foster end users’ *appropriate trust* in machine learning techniques, facilitate decision-making processes, and improve the outcomes of human-machine collaboration. To do so, we investigated the relationship between users’ trust and visual explanations, namely *example-based explanations* (e.g., [8, 53, 55, 122]), which we argue are more suitable as explanations for end users. We hypothesized that the design of example-based explanations affect users’ appropriate trust in machine learning and therefore studied two important visualization design factors: **spatial layout** and **visual representation** [123].

The remainder of this paper is organized as follows. We first outline the background for this research (Section 2). We then present our motivation and justification for the study (Section 3) followed by the detailed design and procedure (Section 4). We report how the visual explanations assist users in developing trust in a classifier (Section 5). We also present our insights from the experiment (Section 6). Specially, our research contributes the following:

- 1 Quantitative evidence that visual explanations help users develop appropriate trust in a machine learning classifier and use the recommendations more appropriately;
- 2 Quantitative results of the effects of the two design factors (spatial layout and visual representation) on end users’ trust;
- 3 Designs of six visual explanations (three spatial layouts and two visual representations) for a classification recommendation.

# 2 BACKGROUND

This section introduces and defines two important concepts in this paper: **appropriate trust** and **example-based explanations** for machine learning.

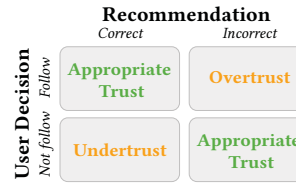


Figure 2: Appropriate trust is to [not] follow an [in]correct recommendation. Other cases lead to overtrust or undertrust.

## 2.1 Appropriate Trust

Trust is an important factor in users’ decisions to use an automated system. Operators did not use automation when their trust in the system was less than their own self-confidence [61]; they used automated systems they trusted and did not use those they did not trust [94]. Trust in a system is both similar to and different from trust in relationships between humans [62, 70]. While many definitions and models of trust were proposed [81, 119] in different contexts (e.g., human-human [70, 77], human-automation [37, 62, 93], and e-commerce [32, 64]), we use the definition for human-computer trust from Madsen and Gregor [72] which was adapted from McAllister [78]:

“Human-computer trust is defined in this study to be, the extent to which a user is confident in, and willing to act on the basis of, the recommendations, actions, and decisions of an artificially intelligent decision aid.”

Henceforth trust has two aspects: users’ willingness to follow a recommendation of a system and their confidence in this decision.

As such, *appropriate trust* (or calibrated trust [87, 92], see Figure 2) is the alignment between the perceived and actual performance of the system [79, 80]. It is related to users’ ability to rely on the system when it is correct and to recognize when the system is incorrect. Appropriate trust is different from *overtrust* (related to misuse) and *undertrust* (related to disuse) [15, 97]. We target appropriate trust and argue that trust in a system must be appropriate, while an increment of inappropriate trust, i.e., overtrust and undertrust, should be avoided. We would like to contrast our contributions with the research that merely “increases users’ trust.”

## 2.2 Example-based Explanations

Several research communities (e.g., human-computer interaction, visualization, artificial intelligence/machine learning) are arguing for more human interpretability of machine learning [19, 35, 42, 128] and beyond [3] and explored various ways of explaining machine learning classification. Examples include using the internal feature coefficients for a linear support-vector classifier [28], showing the distribution of selected features across the random trees for the importance of features in an extra-trees classifier [33], and explaining the inner working of an artificial neural network for overall variable contributions [95]. Other techniques delved into the black-box to explain the classifier [106], assisted machine learning experts to understand the classifier’s internal representation [26, 73, 124], and deep neural networks’ architecture [50, 121].

The techniques mentioned above mostly followed a “model-centric” approach [10, 43, 91]. Emerging approaches explain an intelligent system’s behavior to non-technical users for debugging the system [59], improving the underlying models [2, 110], or increasing users’ trust [67]. These studies used verbal explanations (e.g., [21, 52, 117]) or showed a few examples/features

(e.g., [11, 66]) for end users [67] (e.g., participants from Amazon's Mechanical Turk [66]). They held an underlying assumption where users must understand how a system operates to “increase” their trust [66, 67, 111]. Users' understanding of a system can correlate with their trust [63, 68], but their understanding does not necessarily equate to trust [7] or accurately predict their trust [52]. The conflict between the limited expertise of end users and the complexity of a machine learning algorithm results in the explanations limited by a small set of features or simple models [59].

Example-based explanations resolve this conflict. They show promising results on improving end users' understanding and perception of an intricate system [8]. They were commonly used to illustrate a machine learning process by presenting examples instances from the training set (e.g., [8, 11, 53, 58, 101, 112]). For example, a few techniques learned an influence function and used the most influential training instances [53, 122]; visual inspection was used to reveal important features and explain a prediction for both binary [112] and real-valued data [58]; a locally learned model was used to show an instance-level prediction of any classifier for both image- and feature-based data [101]. These techniques generally require less expertise on how the underlying algorithm or the system operates but reveal some level of information about the internal structure and performance. Empirically, example-based explanations helped users with diverse backgrounds better understand and access an intelligent system [8, 11, 18]. Such explanations also align with people's inductive (“bottom-up logic”) and analogical reasoning (“one instance to another”) to understand why certain objects are considered similar or different [116]. They are suitable testbeds for studying the effects of visualization designs on end users' trust in machine learning.

### 3 STUDY RATIONALE

Though a new field, the literature on explainable artificial intelligence (XAI) [39] and interpretable machine learning is already rich (e.g., [1, 90, 91, 116]). To contrast with this, our study aims to shed light on three under-explored areas in the context of using machine learning for decision making:

- 1 The relationship between users' trust in a system and visual explanations for human-machine collaboration;
- 2 The effects of different visualization designs on users' trust in machine learning;
- 3 An understanding of users' appropriate trust for proper usage of an automated system.

In our study, we used example-based explanations constructed from the instances in a classifier's training set, and we manipulated two key design choices for visual explanations. This section presents our motivation, justifies our design decisions, and explains how we generated different visual explanations for our experiment.

#### 3.1 Motivation

Designing effective visual explanations for showing example instances can be quite challenging. Kulesza et al. outlined a set of principles for designing explanations for interactive machine learning, i.e., “Be Complete” and “Don't Overwhelm” [59]. These principles imply a tradeoff between the amount of information in an explanation and the level of trust users develop (e.g., “not too little

and not too much” [52]). Other research showed that explanations sometimes hurt users' performance [57], and multiple explanations may need to be considered [110]. While these studies explored different design factors for explanations per se, few investigated visual properties of explanations, which are crucial in understanding and interpreting information from a visualization [9].

We hypothesized that different visual properties of explanations can affect users' trust in the classifier, and we selected two factors that cover much of a visualization's design space: **spatial layout** and **visual representation**. They are also primary factors when designing a visualization [9, 123] that illustrates the relationship between example instances (called “graph visualization”).

#### 3.2 “Escape Routes” : Finding Examples

We created an algorithm called “Escape Routes” to identify relevant training instances for a machine learning classification given an input instance. Several algorithms can select relevant training instances for a particular classification [8, 53, 101, 122] and our algorithm is similar to a few past techniques looking for nearest neighbors at the decision boundaries (e.g., [36, 113]). However, our custom algorithm was better suited to our user study, i.e., by helping reveal the importance of spatial layout in an example-based explanation. Similar to past work, we assume a classifier's “internal representation” of the training set and input instances are available.

Our algorithm first constructs a  $k$ -nearest neighbors graph from the internal representation matrix of the training set combined with the input instance based on the Minkowski distance metric [107]. Here a typical parameter setting is  $k = 8$ , but these parameters should be tuned to the data. Our algorithm weights edges by the distance between endpoints and ignores the direction of the edges. It then finds a shortest path tree rooted at the input node, prunes the tree by deleting all nodes whose parents have a different class than the input node, and further simplifies the tree by deleting every node with no descendants (including itself) that have a different class from the input node. Thus, only leaves in the pruned shortest path tree may have a different class from the input node.


After pruning, the remaining nodes in the tree are more relevant for the classification of the input node, and the weighted connections provide additional information about the relationship between the input and training instances. Descendants of the input node are *normative* examples, whereas leaf nodes are *comparative* examples [8]. Additionally, our algorithm can detect relevant examples that are in between *normative* and *comparative*, which would occur as internal nodes in the pruned tree. These internal nodes interpolate between the *normative* and *comparative* examples, ideally making the differences between instances appear less abrupt. From the example instances, we built various visual explanations.


#### 3.3 Spatial Layout


Layout (also known as *position*) is a key factor in designing visualization [9] that affects people's perception and cognition [40]. We used three visualization layouts to illustrate the relationship between the example instances: *grid*, *tree*, and *graph*. Each of these three spatial layouts show all of the example instances from the pruned tree, varying in how instances are arranged spatially, and which connections are visible. The input instance and the instances



from the same class are differentiated using colors and font, and all the other visual elements are the same.


**Grid**  A grid layout is commonly used in the field of computer vision to illustrate the results or mechanisms of a neural network [103, 125]. Our layout arranges instances into rows and columns, and each column corresponds to one class (see Figure 1c). The input instance is always the upper and left-most instance in the grid. Instances within a column are sorted by their weighted geodesic distance to the input node so that more similar instances to the input node are above less similar ones. The interpretation is that it shows the similarities to the input instance within the leftmost column and the differences across columns.

**Tree**  A tree layout is most similar to the pruned shortest path tree. It arranges the instances using a layered graph layout of the pruned shortest path tree (see Figure 1d). The input instance is always the only instance at the top level. Instances with different classes are leaf nodes, below the instances with the same class as the predicted classification of the input instance. The interpretation is that a user can start at the top of the tree with the input instance and follow the paths to leaf nodes with different classes.

**Graph**  A graph layout increases the amount of information in the explanation by further considering additional connections between instances in the  $k$ -nearest neighbor graph. It uses “neato” [24], a force-directed layout algorithm to arrange instances based on their connections (see Figure 1e). The weight of the edge, i.e., the distance between instances in the model representation, is ignored by the layout algorithm to allow sufficient space to draw the instances. The interpretation is that the input instance tends to be placed in the center, and the instances in a different class tend to be located at the periphery. If a nearby instance has a similar appearance but different class, it may indicate an incorrect classification.


### 3.4 Instance Representation

Another key factor in designing a visualization is *visual mark* [9]. While this factor usually refers to representing a data point, we extended this concept to representing an instance and called it *instance representation*. We explored two ways to represent an instance: one shows a feature vector, and the other presents an image. Feature vectors are commonly used as inputs for an algorithm or a machine learning classifier, whereas images are more natural for people to understand. Here we illustrate the two representations using the UCI leaf dataset [105], which was also the dataset used in our experiment (see Section 3.5 below).

**Rose**  We illustrated a feature vector using an ordered rose chart (denoted as “rose-based explanation,” see Figure 1b and Appendix B). We normalized features using the QuantileTransformer (scikit-learn [98]) with 10 quantiles and ordered features so that more positively correlated features were more likely to be adjacent. We brainstormed a set of representations (see supplementary materials) and chose rose charts after two qualitative studies. In the first study, we presented all the candidate representations to eight people in a workshop, including the four authors. All people chose either radar or rose charts; three authors picked rose charts, and the other author picked radar charts. In the second study, we presented explanations of radar and rose charts to eleven other people with various backgrounds in visualization, system development, and user

experience; the majority selected rose charts because they are intuitive for comparing all the features together (e.g., shape), supporting reading individual features, and were aesthetically pleasing.

We were aware that bar [6, 38, 57, 108] and radar [127] charts were commonly used to show features, but little evidence supports that they are more effective than rose charts [56]. Radar charts are misleading because they show the differences between two features as area, but these differences were not directly used by the classifier.

**Image**  We also used an image to represent an instance (denoted as “image-based explanation,” see Figure 1a). Representing instances using images or figures is common in the field of computer vision [103]; it is also intuitive for imagery data.

### 3.5 Dataset

We based our experiment on the UCI leaf dataset [105]; this dataset includes both expert engineered feature vectors and images so that we were able to vary instance representations in a controlled manner. Participants should be able to understand the concept of classifying leaves and compare different leaf instances, but most people are not able to accomplish the classification task unassisted, because naming and distinguishing species requires an uncommon level of expertise in botany. This situation created a scenario that allows human-machine collaboration.

We made two necessary modifications to the original dataset: (1) we replaced the scientific (Latin) names of leaves with their common names (e.g., *Primula vulgaris* was replaced with *Primrose*); and (2) we selected the 10 most frequently occurring leaf classes that have similar numbers of instances. As a result, we had 125 instances and 14 features (e.g., elongation and uniformity).

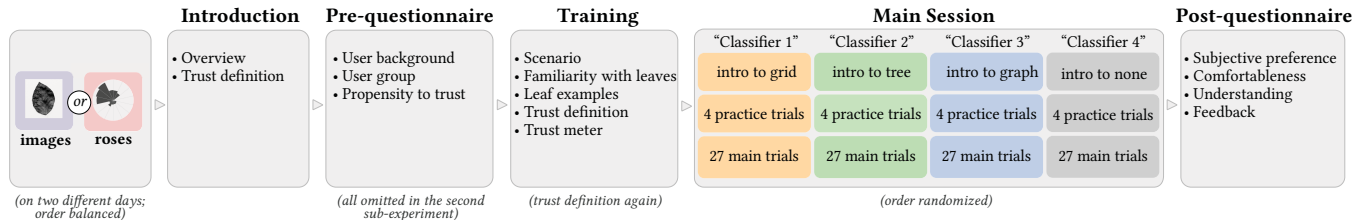
We also modified the colors of leaf images in the original dataset to ensure compatibility between the two representations. We converted the color images to grayscale for three reasons: (1) grayscale images are more similar to rose charts (Figures 1a c.f. b) to avoid issues like differences in memorability [5], (2) colors are misleading because all the texture features (e.g., elongation) in the dataset were based on intensity, and (3) the color images are distracting because they have a pink background. We added a neutral background for all the leaf images so that they looked more similar to the rose charts. We also cropped and re-scaled the images so that the foreground was similar in size across images and compatible to the rose charts.

### 3.6 Classifier

Our classifier was a linear support vector machine (LSVM) with standard normalization, cross-validation, and hyper-parameter grid search (see Appendix B). Its distance function output was the input to our algorithm to find example instances and their connections. The accuracy of the classifier was 71%, which is of medium reliability/utility [17, 46, 60, 84, 89, 126] to foster trust in machine learning. The same LSVM classifier generated all the recommendations and explanations, and only static image files were included in the experiment. The classifier, classification process, and explanations were all implemented using Python 3 and scikit-learn [98].

## 4 STUDY DESIGN

Here we described the details in designing and conducting our experiment. We provide an overview of the experiment in Figure 3.



**Figure 3: Experimental design and procedure**—Each participant finished all eight experimental conditions, including 4=3+1 spatial layout conditions (grid, tree, graph, and none) and 2 representations (images and roses). Every participant finished the two representations in two sub-experiments between 2 and 5 days apart.

#### 4.1 Task Construction

We aim to establish human-machine collaboration in our experiment; that is, a human making a decision with the assistance of a machine learning algorithm. In our experiment, participants were asked to imagine themselves as assistant botanists. As assistant botanists, their task was to classify leaves aided by classifiers, i.e., identifying which class a leaf belongs to, with or without visual explanations. We told participants that they would work with different “classifiers” and hinted that each “classifier” corresponds to a different explanation condition (e.g., a tree layout using rose charts). This was to avoid confusion and encourage participants to treat different explanation conditions independently.

Participants worked with one “classifier” for a series of trials. In each trial, they saw a leaf instance and a classification recommendation (e.g., “The classifier think this represents Hazel.”) Participants decided whether they were going to follow this recommendation. They were encouraged to make good decisions in order to develop appropriate trust: follow the classifier when it is correct, and do not follow the classifier when it is incorrect. After participants made their decision for this recommendation, they were given feedback on whether the classifier was correct or not, a reminder of their decision, and feedback for if their decision was good (e.g., “The classifier was correct for this recommendation. You didn’t follow the classifier. This is not a good decision.”). Giving feedback calibrates trust such that we were able to measure appropriate trust [87], and it also compensates for a poor performing aid [104] because end users may not have expertise to accomplish the task in an early stage; last, showing feedback allows us to observe change of trust over time, aligning with our goal of fostering trust with end users.

#### 4.2 Measuring Appropriate Trust

At the beginning of the experiment, participants were presented with the trust definition (the same definition in Section 2). They were also told that all the classifiers make a correct recommendation about 70-90% of the time. These methods ensured that participants had the same expectation about the classifiers’ reliability [62].

We referred to two ways in the literature to measure trust: (1) directly reporting the level of trust using subjective rating scales (e.g., [14, 29, 31, 60, 126]); and (2) measuring through a multi-item questionnaire (e.g., [47, 78, 82, 85, 87, 100, 104]). Guided by the trust definition, we combined the two approaches. We measured participants’ trust as the willingness to follow the recommendation and their self-confidence in the decision via four items in each trial:

- 1 “Will you follow this recommendation?” It measures participants’ willingness to act on a recommendation and was used

with two options: *Follow* or *Don’t Follow*.

- 2 “How do you feel about your decision above?” It measures participants’ self-confidence and was rated on a 7-point Likert scale from *Not at all comfortable* (1) to *Very comfortable* (7). *Comfortable* replaced *confident* for two reasons: (1) the participants in pilot study were confused with whether *confidence* referred to the classifier or their decision; and (2) *confident* and *comfortable* are equivalent in conceptual meaning [114], while *comfortable* was used more to describe a decision [49, 114].
- 3 “Was the explanation helpful in making the decision above?” This item was rated on a 7-point Likert scale from *Not at all helpful* (1) to *Very helpful* (7). This item was omitted in the control conditions when no explanation was available.
- 4 A linear “Trust Meter” ranged from *completely distrust* (−100) to *completely trust* (+100), inspired by [51]. Participants were reminded and encouraged to adjust the trust meter at any time if their trust in the classifier changes. The trust meter was always available to participants and reset for a different classifier, while the above three questions were removed after the feedback.

The first item captures more of *cognitive trust*, the willingness to rely on a system’s competence, which arises from accumulated knowledge [48]. The second question captures *affective trust*, defined as one’s feeling of security and comfort relying on the trustee (the classifier) [54]. The third item is a usability measure. The last item is the trust meter that captures changes in trust.

#### 4.3 Experimental Design

The experiment was a complete within-subjects design. We manipulated spatial layout (grid, tree, and graph) and instance representation (images or rose charts); we also used one control condition for each representation (denoted as “none”) to establish the baseline of trust and performance. As a result, we had  $(3+1) \times 2 = 8$  experimental conditions. We did not manipulate other factors such as affect [84, 89] and workload [16], but we controlled reliability [65, 86, 117, 126], expectation [62], and risk [83].

Each participant finished all eight experimental conditions, divided into two sub-experiments, in each of which only showed one representation (images or rose charts) but all the three spatial layouts (grid, tree, graph) as well as the corresponding control condition. Each participant completed the two sub-experiments between 2 and 5 days apart due to our strong concerns about learning, practice [27], and fatigue effects [102]. The order of sub-experiments was balanced across gender, and the order of conditions repeated a  $4 \times 4$  Latin square; the sub-experiments and conditions were otherwise randomly assigned.

In each condition, participants finished a sequence of trials. All the sequences had the identical order of correct/incorrect recommendations. The fixed sequence controlled learning effects, aligned priming effects, and allowed us to compare different conditions. The specific sequence we used was generated by a Monte Carlo method, permuting until no more than two incorrect recommendations occur in any span of five. We also ensured that the first trial was always correct to protect trust in an early stage [62, 75]. The sequence we used consisted of 4 practice and 27 main trials, containing 3 and 20 correct recommendations, respectively. These resulted in an “experienced accuracy” of about 75% ( $3/4=75\%$ ,  $20/27=74\%$ ), closely matching the classifier’s accuracy of 71%. The number of trials was decided in order to (1) compare with other experiments studied trust (e.g., 10 [17, 22, 60, 84, 85]), 24/48 [76], 26 [17], 50 [104]), (2) have enough trials to develop trust but prevent participants from memorizing the order (serial position effects [44]), (3) balance the number of trials across different classes. The instances in the sequence were randomized, and the instances used for training and practice were excluded in the main trials.

We used a think-aloud protocol with three participants for a pilot study. Based on the observations from the pilot study, we fine-tuned the questions and instructions. We excluded these three participants from the main experiment.

#### 4.4 Procedure

After participants provided informed consent, they saw an overview of the experiment. They then filled in a questionnaire for their background information and propensity to trust [85], followed by their familiarity with plants’ leaves on a 7-point Likert scale from *Not at all familiar* to *Very familiar*. They then took part in a training session, read the scenario and goals, saw an example instance from each leaf class, and practiced using the trust meter. Participants then proceeded to the main session where they worked with four “classifiers” (grid, tree, graph, and none). For each “classifier,” they first saw an introduction of what they could expect to see and the instructions of how to read the visual explanation, followed by 4 practice trials. The trust meter was then reset, and participants finished 27 main trials. All participants took a 1-minute mandatory break between two “classifiers;” longer breaks were allowed. After completing all four “classifiers,” they filled in a post-experiment questionnaire. Each sub-experiment took approximately 1.5 hours, and two together took about 3 hours. The consent and pre-experiment questionnaire were omitted in the second sub-experiment.

All experiments were conducted in-person and in-laboratory on Windows desktops with a mouse and a keyboard, using the monitors of the same model (24 inch, 1920 × 1080, 60Hz) and proctored by at

least one of the authors following the same study protocol.

#### 4.5 Participants

We planned to include both non-expert and expert users in our experiment. The non-expert users had little background knowledge in machine learning, but they might have experienced machine learning techniques in their personal lives and work environments. The expert users had sufficient knowledge of how a machine learning algorithm operates or experienced both effectiveness and ineffectiveness of a machine learning classifier. They held different mental models about machine learning [45].

We recruited 33 participants (19 female, 14 male) from our institution (PNNL) according to a pre-planned end date and to roughly balance gender, backgrounds, as well as the two orders of sub-experiments. All the participants were staff members, and they were typically between 25 and 60 years old. Among them, 16 participants did data analysis as part of their job or have machine learning backgrounds, and we identified them as “expert users.” The rest 17 participants came from other disciplines or administrative positions, and we identified them as “non-expert users.” All the participants were compensated at their normal salary rate and the experiment counted towards their working hours to help ensure participants were engaged in the experiment. The study was approved by the PNNL Institutional Review Board (IRB #2018-15).

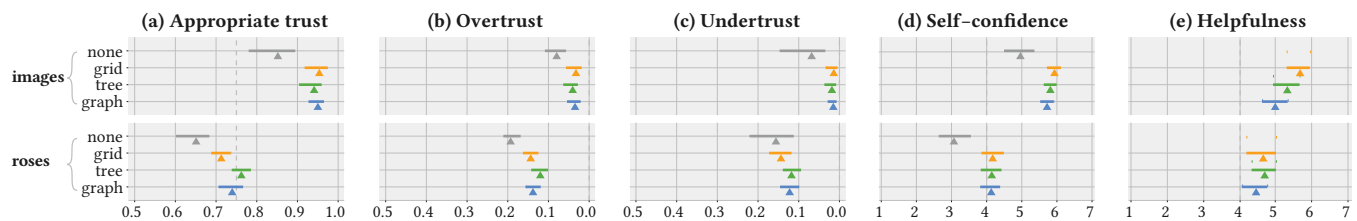
In total, we collected 8,184 trials = (3+1) layout conditions (grid, tree, graph, and none) × 2 representations (images and roses) × 31 trials × 33 participants. We excluded the practice trials and used 7,128 trials for analyses.

### 5 ANALYSES AND RESULTS

We framed our research questions as follows:

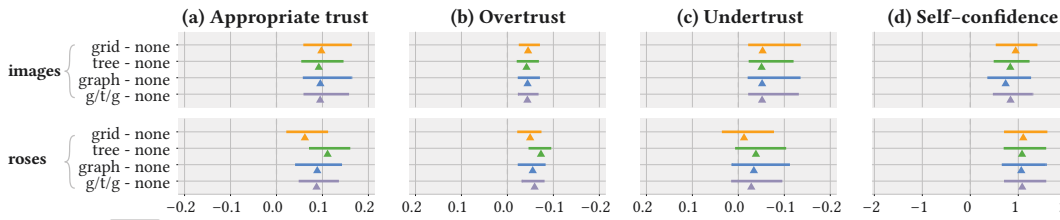
- RQ1** Do our visual explanations enable users to develop more appropriate trust?
- RQ2** How did the three spatial layouts (grid, tree, and graph) affect users’ trust differently?
- RQ3** How did the two instance representations (images and roses) affect users’ trust differently?
- RQ4** How did other covariates (e.g., expert users vs. non-expert users, prior knowledge, and propensity to trust) affect people’s trust?
- RQ5** How did errors and explanations affect trust direction?

The first research question was a replication to confirm that our explanations increase users’ appropriate trust. We aimed to understand the effects of different spatial layouts and representations (RQs 2 and 3) and to explore the effects of individual differences (RQ4). We were also interested in how trust changes over time and



**Figure 4: Overview of the results**—the raw results for mean values and 95% bootstrap CIs. The center lines indicate the classifier’s performance and neutral cases for comparison. (right side = “better,” ▲ = mean, — = 95% CI)





**Figure 5: RQ1 Do our visual explanations enable users to develop more appropriate trust?**—The results suggest that all the visual explanations (a) increase appropriate trust, (b) decrease overtrust, and (d) lead to more confidence. Image-based explanations also have moderate effects on (c) correcting undertrust. (right side = “better,” ▲ = mean, — = 95% CI)

the effects of giving feedback on trust (RQ5). We framed all the research questions and collected all the data before any analysis.

### 5.1 Measures

We collected three categories of data in our experiment:

**Trust** had four measures: (1) appropriate trust (good decisions), (2) overtrust (following an incorrect recommendation), (3) undertrust (not following a correct recommendation), and (4) self-confidence. These measures align with the trust definition in Section 2 and Figure 2, quantifying the willingness to follow a recommendation and users’ self-confidence in their decisions. We used the percentage of each type of decision and the average of the confidence score over the 27 trials for each participant × condition pair.

**Usability** had five measures: (1) the helpfulness of the given explanation in a trial, (2) completion time, and (3-5) three measures from the post-questionnaire (subjective preference, overall comfortableness, and overall understanding of each spatial layout). We focused on helpfulness because it precisely measured the usability of a given explanation for each trial.

**Trust meter** was sampled twice in a trial: (1) before participants saw the feedback, and (2) after participants saw the feedback but before the next trial. The difference between any two adjacent samples indicates the direction of trust (increase or decrease).

In sum, appropriate trust, overtrust, and undertrust are percentages in the range of [0, 1]; self-confidence and helpfulness are Likert scale ratings in the range of [1, 7]; trust direction is a difference in the range of [-100, 100], which we later normalized to [-1, 1].

### 5.2 Approach

In response to the limitations of null hypothesis significance testing (NHST), we used the interval estimate method recommended by Cumming [13] and Dragicevic [20]. We used 95% bias-corrected and accelerated bootstrap confidence intervals ( $R = 5000$ ) and mea-

sured effect size using Cohen’s  $d$  and  $R^2$  [12]. We aggregated the results for each participant, bootstrapped individual participant’s data, and calculated confidence intervals. We used mixed effects models to investigate covariates. To prevent from multiple comparisons problems, we avoided making a conclusion from an individual comparison and focused on summarizing all confidence intervals; using bootstrapping also controls false discovery, especially when the resampling is exhaustive [120]. We based our inference on the interpretation of confidence intervals [12]: the range of a confidence interval and its relationship to 0 indicate the size of an effect.

We included all 33 participants in our analyses because we did not find any explicit outlier or indication of not following instructions. We provided an overview of the results in Figure 4. The detailed results and their interpretation are as follows.

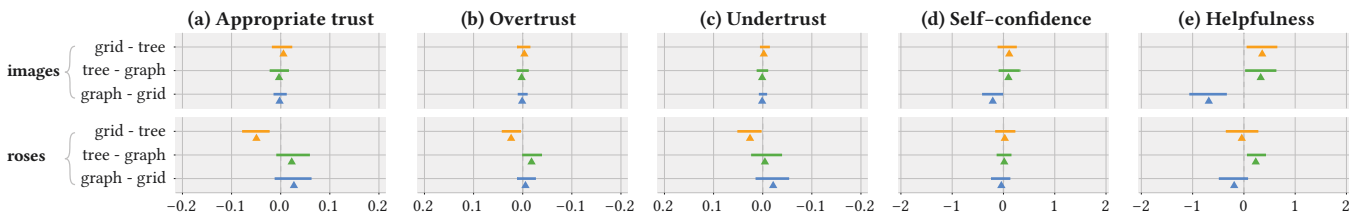
### 5.3 Results

We report the mean values and their 95% bootstrap confidence intervals (CIs). We also report the values of Cohen’s  $d$  and their 95% bootstrap confidence intervals in supplementary materials.

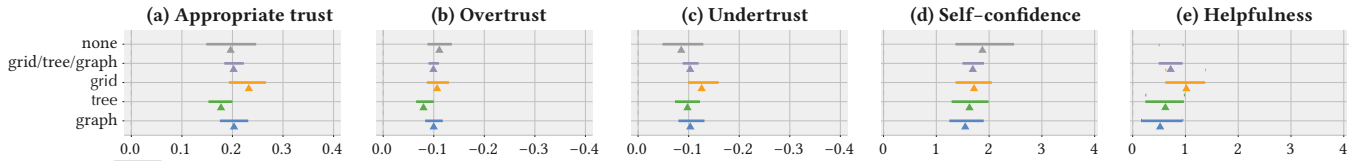
#### RQ1 Do our visual explanations increase appropriate trust?

**Method** We used the four trust measures to quantify the effects: (1) appropriate trust, (2) overtrust, (3) undertrust, and (4) self-confidence. We subtracted the control (none) conditions from the others—grid, tree, graph, and an aggregation of the three spatial layouts, denoted as “g/t/g” (grid/tree/graph).

**Results** We present the results in Figure 5. The results support that all the visual explanations largely increase appropriate trust (e.g., Cohen’s  $d$ : 0.84 [0.58, 1.11]), reduce overtrust (e.g., Cohen’s  $d$ : -1.01 [-1.66, -0.42]), and help users gain more confidence in their decisions (e.g., Cohen’s  $d$ : 0.84 [0.54, 1.11]). The results also suggest that image-based explanations help correct undertrust (e.g., Cohen’s  $d$ : -0.41 [-0.58, -0.22]), but the effect of rose-based explanations on undertrust is inconclusive.



**Figure 6: RQ2 How did the three spatial layouts affect users’ trust?**—The results support that for image-based explanations, the three spatial layouts are similar (top a-c), while grid explanations may be more helpful (e). For rose-based explanations, the results suggest that tree explanations leads to more appropriate trust than grid explanations (bottom a-c). (right side = “better,” ▲ = mean, — = 95% CI)



**Figure 7: RQ3 How did the two representations affect users’ trust?**—The results strongly suggest that image-based explanations result in more appropriate trust than rose-based explanations (a-d), and participants also think image-based explanations are more helpful (e). (right side = “better,” ▲ = mean, — = 95% CI)

**RQ2 How did the three spatial layouts affect users’ trust?**

**Method** We used both the trust and usability measures for comparing the three spatial layouts of explanations: (1) appropriate trust, (2) overtrust, (3) undertrust, (4) self-confidence, and (5) helpfulness. We subtracted the three spatial layouts from each other, and calculated mean difference and their 95% bootstrap CIs.

**Results** We present the results in Figure 6. For the image-based explanations, all the three spatial layouts lead to a similar level of users’ trust; appropriate trust, overtrust, and undertrust are very similar and display few differences among the three spatial layouts (e.g., Cohen’s *d*: 0.092 [-0.28, 0.45]). However, the results also support that grid explanations are slightly more helpful than tree explanations (e.g., Cohen’s *d*: -0.34 [-0.68, 0.065]), which are more helpful than graph explanations (e.g., Cohen’s *d*: -0.62 [-0.96, -0.22]). For the rose-based explanations, the results support that tree and graph explanations lead to more appropriate trust than grid explanations, and participants gain a similar level of confidence with all the three spatial layouts (e.g., Cohen’s *d*: -0.075 [-0.42, 0.28]).

**RQ3 How did the two representations affect users’ trust?**

**Method** Similar to the analyses above, we used both the trust measures and helpfulness for comparing images and rose charts representations : (1) appropriate trust, (2) overtrust, (3) undertrust, (4) self-confidence, and (5) helpfulness. We subtracted the rose-based explanations from the image-based explanations, and calculated mean differences and their 95% bootstrap CIs.

**Results** We present the results in Figure 7. The results strongly suggest that image-based explanations result in more appropriate trust (e.g., Cohen’s *d*: 1.90 [1.60, 2.20]), reduce overtrust (e.g., Cohen’s *d*: -1.63 [-1.88, -1.40]), correct undertrust (e.g., Cohen’s *d*: -1.13 [-1.33, -0.90]), and help people gain more confidence than rose-based explanations (e.g., Cohen’s *d*: 1.43 [1.23, 1.64]). The results also support that image-based explanations are more helpful than rose-based explanations (e.g., Cohen’s *d*: 0.64 [0.40, 0.84]).

When taking RQs1-3 together, the effects of “images vs. roses” and “explanation vs. none” are stronger than the differences in the three spatial layouts. To simplify further analyses, we omit the differences in the three spatial layouts and consider them all as having a visual explanation, denoted as “g/t/g” (grid/tree/graph).

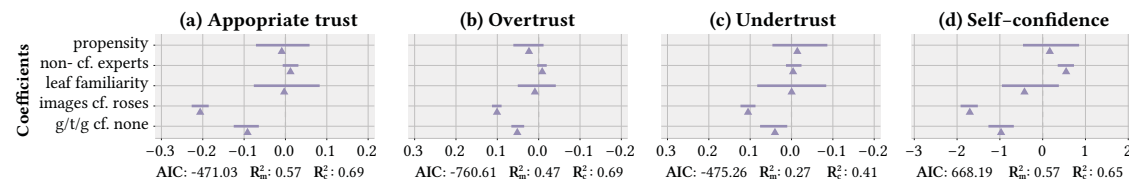
**RQ4 How did covariates affect users’ trust?**

**Method** We investigated three covariates of interest: (1) non-expert users vs. expert users, encoded as 1 vs. 0; (2) participants’ familiarity with the task (leaf familiarity), varying from 1 to 7 and rescaled to 0 to 1; and (3) propensity to trust, varying from 1 to 7 and rescaled to 0 to 1. We also compared these covariates with the two experimental variables: (4) images vs. roses, recoded as 1 vs. 0, and (5) explanation vs. none, recoded as 1 vs. 0. We used mixed-effects regression models, where these five variables were fixed effects and participants were random intercepts. We checked the collinearity between all these variables (covariates) and found low collinearity for all pairs ( $1 < \text{variance inflation factor (VIF)} < 2$ ). We built a model for each trust measure and fit them using the observations from each participant ( $33 \times 8 = 264$  observations).

**Results** We present the coefficients of the models, their 95% bootstrap CIs, and model metrics in Figure 8. These models explain about 50% to 70% variance in the data. The effects of all the three observed covariates are inconclusive; the only exception is that non-expert users seem to have more confidence in their decisions; they may also have slightly more appropriate trust and overtrust. The strongest effects come from the two experimental variables: image-based explanations lead to more appropriate trust and a higher level of confidence than rose-based explanations; having a visual explanation also leads to more appropriate trust and confidence.

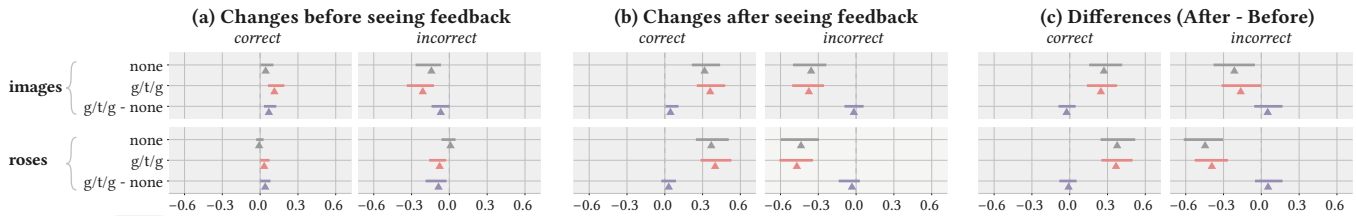
**RQ5 How did errors and explanations affect trust direction?**

**Method** We sampled the direction of trust twice in a trial, split by if participants saw the feedback, i.e., knowing if the classifier was correct and if they made a good decision. One was (1) the direction before seeing the feedback, defined as the change of the trust meter between the starting of a trial and right before the feedback. The other was (2) the direction after seeing the feedback, defined as the change of the trust meter between first seeing the feedback and the end of the trial. We used the signs of changes to normalize the changes and prevented extreme scores; all the negative scores (decrements) were mapped to -1, and all the positive scores (increments) were mapped to 1. We averaged the changes for each participant and subtracted within participants. We calculated the mean values, the mean differences, and their 95% bootstrap CIs.



**Figure 8: RQ4 How did covariates affect users’ trust?**—We report the model coefficients and their 95% CIs. The effects of the three covariates seem to be inconclusive; the exception is that non-expert users seem to have more confidence. The strongest effects come from if participants had an explanation and which representation they used. (right side = “better,” ▲ = mean, — = 95% CI)





**Figure 9: RQ5 How did errors and explanations affect trust direction?**—Participants responded to the feedback strongly. They appeared to decrease the trust meter for an incorrect recommendation when an image-based explanation is available. They did not adjust the trust meter when using rose charts. (right side = “better,” ▲ = mean, — = 95% CI)

**Results** We report the results in Figure 9. Before seeing the feedback, participants increase the trust meter for a correct recommendation and decrease the trust meter for an incorrect one; having an explanation shows small positive effects on trust direction (e.g., Cohen’s  $d$ : 0.49 [0.22, 0.72]). The effects of feedback are very strong (e.g., Cohen’s  $d$ : 0.73 [0.36, 1.02]), especially for rose-based explanations (e.g., Cohen’s  $d$ : 0.95 [0.64, 1.26]). Yet the effect of having an explanation on trust direction is very small (e.g., Cohen’s  $d$ : 0.34 [-0.11, 0.65]). We think that participants respond to the feedback strongly regardless of spatial layouts and representations; but our explanations may help them recognize a correct recommendation and identify an incorrect one before they see the feedback.

#### 5.4 Summary of Results

For each research question, we summarized our findings as follows.

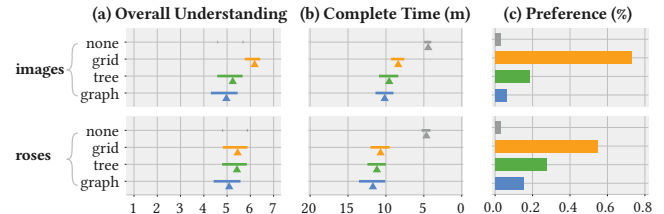
- 1 Every explanation improves users’ appropriate trust in the classifier, and the human-machine collaboration [71] can achieve nearly perfect performance (all good decisions) when an effective explanation is available.
- 2 Image-based explanations outperform rose-based explanations because they increase appropriate trust, decrease overtrust and undertrust, improve self-confidence, and show more usability.
- 3 Grid explanations generally outperform tree explanations, which outperform graph explanations; tree explanations moderately outperform grid explanations when using rose charts.
- 4 Individual differences did not seem to show strong effects in this experiment.
- 5 Showing feedback has a very strong effect on trust direction, especially when an explanation is available.

## 6 DISCUSSION

### 6.1 Instance Representation: Images and Roses

We found that image-based explanations were most effective for improving appropriate trust and helpfulness. The primary reason may be that images are easy to understand and more features beyond the expert engineered features from the dataset are available in image representations. These image-based explanations correct overtrust. People can use them to detect a system’s error and respond to the error accordingly. Image-based explanations also mitigate undertrust; that is, they have persuasive power [96] to convince people to believe that a recommendation is correct. Last, participants spent less time on the image-based explanations. It may indicate that image-based explanations are easier to understand (Figure 10a).

In our study, rose charts accurately reflect the features used by the classifier, and rose-based explanations closely match the clas-



**Figure 10: The understanding scores of the three spatial layouts, completion time and subjective preference of all the eight conditions.** (right side = “better,” ▲ = mean, — = 95% CI)

sification process. While rose-based explanations generally help participants improve their trust in a classifier, participants may have difficulty understanding them. They may not believe a classification is correct even given the explanation; hence rose-based explanations did not improve undertrust. Part of the reason is that we use raw features and hide the meaning of each feature to avoid overwhelming participants. This design decision may make it more difficult to reason about how individual features contributed to a particular classification. In the real world, sometimes only feature vectors are available, and we have to use feature representations. We can improve the understandability of feature representations by combining existing feature-based explanations with our instance-based explanations, adding interaction, or using exploration [23]. This difficulty may also explain why expert users had less confidence in their decisions, and why they want more information about the recommendation when making a decision.

### 6.2 Spatial Layout: Grid, Tree, and Graph

All three layouts of explanations improve users’ trust, though grid explanations with images were the most helpful. Participants have more confidence in their decisions, feel grid explanations are more helpful, and understand grid explanations better (see Figures 10a-b). Grid explanations are also preferred by the majority of participants (see Figure 10c). This observation may be explained by the simplicity and clarity of a grid layout:

*“It is simple and allowed me to easily see ground-truth from each class.”*

While most participants prefer grid explanations when using rose charts, they develop more appropriate trust with tree explanations. Tree explanations disclose more information (e.g., links) over grid explanations and compensate for difficulties of understanding rose charts. While graph explanations provide the most amount of information and empower the most accurate comparisons, few participants preferred graph explanations despite similar levels of trust.

Participants without visualization background or training find it difficult to understand a graph explanation:

*“It required too much interpretation compared with the grid and tree approach.”*

In a graph explanation, the placement of the source instance is arbitrary due to the layout algorithm, sometimes causing participants to spend extra time searching for it:

*“I had to work hard to find the recommended version in the diagram, and then I had to jump all over the map to make pairwise comparisons or find other examples to compare contrast against.”*

A graph explanation may also cause clutter and occlusion, further reducing visualization legibility.

These observations suggest that spatial layout affects people’s trust and interacts with instance representation. While different instance representations may yield different levels of task difficulties, a good spatial layout could help mitigate some of the task difficulties. **We recommend a grid layout if the representation is easy to understand and a tree layout if the representation is difficult to read or usability of the representation is unknown.**

### 6.3 Understanding and Appropriate Trust

We observed that understanding and appropriate trust are relevant but different. Our visual explanations did not reveal the complicated underlying algorithm. It would have been impossible for our participants to understand the underlying classification process. However, they were able to make good decisions almost all the time with the classifier’s assistance and proper explanations, which explain the outputs from the classifier and the relationship between example instances. These appear to be sufficient for participants to decide whether to follow a classification recommendation.

We observed that participants’ understanding of visual explanation was correlated with their trust. When using grid explanations with images, they developed the highest level of appropriate trust and felt they understood the explanations the best; participants performed worse with all the rose-based explanations and thought they understood rose charts less than images. However, while participants found tree and graph explanations harder to understand than grid explanations, they developed similar levels of trust. Furthermore, when using rose charts, participants developed the highest level of appropriate trust with tree explanations. Participants showed overtrust and undertrust when using grid explanations with rose charts even though they rated grid explanations as the most easily understood.

We speculate that the understandability of an explanation is crucial to users’ trust, in terms of both spatial layout and instance representation. Instance representation appears to be more important in creating better trust and understanding; however, spatial layout determines the amount and form of information provided which certainly contributes to users’ trust. A previous study using a similar grid layout with bar charts showed the visual explanations with aggregation improved both users’ performance and trust while lack of aggregation reduced users’ performance [57]. These results align with our observations that representation impacts users’ trust, but spatial layout also contributes.

Last, **we recommend that future research considers appropriate trust, instead of simply measuring an increase in users’ trust.** If we were looking only for an increment in users’ trust, we would have overlooked the increased overtrust and accounted for it as a positive result. Measuring the appropriate trust avoids the issue that trust might be misplaced [109].

### 6.4 Limitations and Future Work

Limitations of our work are that we used example instances at decision boundaries, a pre-defined distance metric, and a linear kernel with Minkowski distance. This selection of examples may bring in biases in our results about the effects of spatial layouts and instance representations. Our findings may not generalize to every example-based explanation technique.

Future work can compare different methods for instance selection, different representations, and other layouts, and also include the classifier’s confidence in each instance and recommendation. For example, the layouts in our study could be varied using other algorithms to calculate the influential instances. An alternative approach could be to hide feedback for domain experts. Also, to improve the understandability of representations, interaction techniques [88] like brushing and coordinated multiple views [56] can be used to help users gain insights about features and further confidence scores from the classifier. If a dataset has many features, dimension reduction or feature aggregation [57] could be done prior to generating feature representations, i.e., rose charts; other feature representations are also possible, as shown in our supplementary materials. These extensions require additional experiments, and they are beyond the scope of our study. Broader findings would further enable designers to select visual explanations to make automated systems more interpretable, trustable, and responsible [3].

## 7 CONCLUSION

In this paper, we investigated two visualization design factors—spatial layout and instance representation—for example-based explanations and showed their effects on end users’ appropriate trust. We found that every visual explanation in our experiment greatly increased users’ appropriate trust in machine learning and improved appropriate use of the recommendations from the classifier. However, different spatial layouts and instance representations showed very strong effects on users’ trust; users’ backgrounds had a weak effect on the levels of their trust and confidence. We conclude that both the understandability and layout of the explanation contribute to users’ trust; feedback of users’ performance affects their intentions and direction of trust. Our concise explanations have the potential to improve end users’ trust and enhance the use of automated systems without requiring a complete understanding of the system or algorithm.

## ACKNOWLEDGMENTS

The research described in this paper was conducted under the Laboratory Directed Research and Development Program at Pacific Northwest National Laboratory, a multiprogram national laboratory operated by Battelle for the U.S. Department of Energy. We thank Curtis Larimer and Leslie Blaha for their helpful guidance and the anonymous reviewers for their valuable feedback.

## REFERENCES

- [1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Article Paper 582, 18 pages.
- [2] Saleema Amershi, James Fogarty, Ashish Kapoor, and Desney Tan. 2010. Examining multiple potential models in end-user interactive concept learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1357–1360.
- [3] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bernetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2019. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. arXiv:cs.AI/1910.10045
- [4] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. 2019. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2429–2437.
- [5] Michelle A Borkin, Azalea A Vo, Zoya Bylinskii, Phillip Isola, Shashank Sunkavalli, Aude Oliva, and Hanspeter Pfister. 2013. What makes a visualization memorable? *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2306–2315.
- [6] Eli T Brown, Jingjing Liu, Carla E Brodley, and Remco Chang. 2012. Dis-function: Learning distance functions interactively. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*. IEEE, 83–92.
- [7] Adrian Bussone, Simone Stumpf, and Dymna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *Healthcare Informatics (ICHI), 2015 International Conference on*. IEEE, 160–169.
- [8] Carrie J Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, 258–262.
- [9] Stuart K Card and Jock Mackinlay. 1997. The structure of the information visualization design space. In *Proceedings of VIZ'97: Visualization Conference, Information Visualization Symposium and Parallel Rendering Symposium*. IEEE, 92–99.
- [10] Jaegul Choo and Shixia Liu. 2018. Visual Analytics for Explainable Deep Learning. arXiv preprint arXiv:1804.02527 (2018).
- [11] Sven Coppers, Jan Van den Bergh, Kris Luyten, Karin Coninx, Iulianna van der Lek-Ciudin, Tom Vanallemeersch, and Vincent Vandeghinste. 2018. Intellingo: An Intelligible Translation Environment (CHI '18). ACM, New York, NY, USA, Article 524, 13 pages.
- [12] Geoff Cumming. 2013. *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.
- [13] Geoff Cumming. 2014. The new statistics: Why and how. *Psych. Sci.* 25, 1 (2014), 7–29.
- [14] Aritra Dasgupta, Joon-Yong Lee, Ryan Wilson, Robert LaFrance, Nick Cramer, Kristin Cook, and Samuel Payne. 2017. Familiarity vs trust: A comparative study of domain scientists' trust in visual analytics and conventional analysis methods. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (2017), 271–280.
- [15] Ewart de Visser, Marvin Cohen, Amos Freedy, and Raja Parasuraman. 2014. A design methodology for trust cue fabrication in cognitive agents. In *International Conference on Virtual, Augmented and Mixed Reality*. Springer, 251–262.
- [16] Ewart de Visser and Raja Parasuraman. 2011. Adaptive aiding of human-robot teaming: Effects of imperfect automation on performance, trust, and workload. *Journal of Cognitive Engineering and Decision Making* 5, 2 (2011), 209–231.
- [17] Peter de Vries, Cees Midden, and Don Bouwhuis. 2003. The effects of errors on system trust, self-confidence, and the allocation of control in route planning. *International Journal of Human-Computer Studies* 58, 6 (2003), 719–735.
- [18] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.
- [19] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017).
- [20] Pierre Dragicevic. 2016. *Fair statistical communication in HCI*. Springer Int. Publishing, Cham, 291–330.
- [21] Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce, and Hall P Beck. 2003. The role of trust in automation reliance. *International Journal of Human-Computer Studies* 58, 6 (2003), 697–718.
- [22] Mary T Dzindolet, Linda G Pierce, Hall P Beck, and Lloyd A Dawe. 2002. The perceived utility of human and automated aids in a visual detection task. *Human Factors* 44, 1 (2002), 79–94.
- [23] Malin Eiband, Charlotte Anlauff, Tim Ordenewitz, Martin Zürn, and Heinrich Hussmann. 2019. Understanding Algorithms Through Exploration: Supporting Knowledge Acquisition in Primary Tasks (MuC'19). ACM, New York, NY, USA, 127–136.
- [24] John Ellson, Emden Gansner, Lefteris Koutsofios, Stephen C North, and Gordon Woodhull. 2001. Graphviz—open source graph drawing tools. In *International Symposium on Graph Drawing*. Springer, 483–484.
- [25] Greg Elosfon. 2001. Developing trust with intelligent agents: An exploratory study. In *Trust and Deception in Virtual Societies*. Springer, 125–138.
- [26] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2009. *Visualizing higher-layer features of a deep network*. Technical Report. University of Montreal.
- [27] Marina G Falletti, Paul Maruff, Alexander Collie, and David G Darby. 2006. Practice effects associated with the repeated assessment of cognitive function using the CogState battery at 10-minute, one week and one month test-retest intervals. *Journal of Clinical and Experimental Neuropsychology* 28, 7 (2006), 1095–1112.
- [28] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research* 9, Aug (2008), 1871–1874.
- [29] Amos Freedy, Ewart DeVisser, Gershon Weltman, and Nicole Coeyman. 2007. Measurement of trust in human-robot collaboration. In *Collaborative Technologies and Systems, International Symposium on*. IEEE, 106–114.
- [30] Yashesh Gaur, Walter S Lasecki, Florian Metzke, and Jeffrey P Bigham. 2016. The effects of automatic speech recognition quality on human transcription latency. In *Proceedings of the 13th Web for All Conference*. ACM, Article Article 23, 8 pages.
- [31] David Gefen. 2000. E-commerce: the role of familiarity and trust. *Omega* 28, 6 (2000), 725–737.
- [32] David Gefen, Elena Karahanna, and Detmar W Straub. 2003. Trust and TAM in online shopping: An integrated model. *MIS quarterly* 27, 1 (2003), 51–90.
- [33] Pierre Geurts, Damien Ernst, and Louis Wehenkel. 2006. Extremely randomized trees. *Machine Learning* 63, 1 (2006), 3–42.
- [34] Alyssa Glass, Deborah L McGuinness, and Michael Wolverton. 2008. Toward establishing trust in adaptive agents. In *Proceedings of the 13th International Conference on Intelligent User Interfaces*. ACM, 227–236.
- [35] Michael Gleicher. 2013. Explainers: Expert explorations with crafted projections. *IEEE Transactions on Visualization and Computer Graphics* 12 (2013), 2042–2051.
- [36] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014).
- [37] Tyrone Grandison and Morris Sloman. 2000. A survey of trust in internet applications. *IEEE Communications Surveys & Tutorials* 3, 4 (2000), 2–16.
- [38] Samuel Gratzl, Alexander Lex, Nils Gehlenborg, Hanspeter Pfister, and Marc Streit. 2013. Lineup: Visual analysis of multi-attribute rankings. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2277–2286.
- [39] David Gunning. 2017. Explainable artificial intelligence (XAI). *Defense Advanced Research Projects Agency (DARPA), 2nd Web* (2017).
- [40] Ivan Herman, Guy Melançon, and M Scott Marshall. 2000. Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics* 6, 1 (2000), 24–43.
- [41] Tom Hitron, Iddo Wald, Hadas Erel, and Oren Zuckerman. 2018. Introducing children to machine learning concepts through hands-on experience. In *Proceedings of the 17th ACM Conference on Interaction Design and Children*. ACM, 563–568.
- [42] Robert R Hoffman and Gary Klein. 2017. Explaining explanation, part 1: theoretical foundations. *IEEE Intelligent Systems* 3 (2017), 68–73.
- [43] Fred Matthew Hohman, Minsuk Kahng, Robert Pienta, and Duen Horng Chau. 2018. Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers. *IEEE Transactions on Visualization and Computer Graphics* (2018).
- [44] Marc W Howard and Michael J Kahana. 1999. Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 25, 4 (1999), 923.
- [45] Yu-Chen Hsu. 2006. The effects of metaphors on novice and expert learners's performance and mental-model development. *Interacting with Computers* 18, 4 (2006), 770–792.
- [46] T Inagaki, N Moray, and M Itoh. 1998. Trust, self-confidence and authority in human-machine systems. *IFAC Proceedings Volumes* 31, 26 (1998), 431–436.
- [47] Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. 2000. Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics* 4, 1 (2000), 53–71.
- [48] Devon Johnson and Kent Grayson. 2005. Cognitive and affective trust in service relationships. *Journal of Business research* 58, 4 (2005), 500–507.
- [49] Lawrence K Jones and Mary F Chenery. 1980. Multiple subtypes among vocationally undecided college students: A model and assessment instrument. *Journal of Counseling Psychology* 27, 5 (1980), 469.
- [50] Minsuk Kahng, Pierre Y Andrews, Aditya Kalro, and Duen Horng Polo Chau. 2018. ActiVis: Visual exploration of industry-scale deep neural network models. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2018), 88–97.
- [51] Mohammad T Khasawneh, Shannon R Bowling, Xiaochun Jiang, Anand K Gramopadhye, and Brian J Melloy. 2003. A model for predicting human trust in automated systems. *Origins* 5 (2003).



- [52] René F Kizilcec. 2016. How much information?: Effects of transparency on trust in an algorithmic interface. In *Proceedings of the Conference on Human Factors in Computing Systems*. ACM, 2390–2395.
- [53] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. *arXiv preprint arXiv:1703.04730* (2017).
- [54] Sherrie YX Komiak and Izak Benbasat. 2006. The effects of personalization and familiarity on trust and adoption of recommendation agents. *MIS quarterly* (2006), 941–960.
- [55] Josua Krause, Aritra Dasgupta, Jordan Swartz, Yindalon Aphinyanaphongs, and Enrico Bertini. 2017. A workflow for visual diagnostics of binary classifiers using instance-level explanations. *arXiv preprint arXiv:1705.01968* (2017).
- [56] Josua Krause, Adam Perer, and Enrico Bertini. 2014. INFUSE: interactive feature selection for predictive modeling of high dimensional data. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 1614–1623.
- [57] Josua Krause, Adam Perer, and Enrico Bertini. 2018. A user study on the effect of aggregating explanations for interpreting machine learning models. In *KDD Workshop on Interactive Data Exploration and Analytics (IDEA)*.
- [58] Josua Krause, Adam Perer, and Kenney Ng. 2016. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 5686–5697.
- [59] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*. ACM, 126–137.
- [60] John D Lee and Neville Moray. 1992. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics* 35, 10 (1992), 1243–1270.
- [61] John D Lee and Neville Moray. 1994. Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-computer Studies* 40, 1 (1994), 153–184.
- [62] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors* 46, 1 (2004), 50–80.
- [63] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (2018), 2053951718756684.
- [64] Matthew KO Lee and Efraim Turban. 2001. A trust model for consumer internet shopping. *International Journal of electronic commerce* 6, 1 (2001), 75–91.
- [65] Scott LeeTiernan, Edward Cutrell, Mary Czerwinski, and Hunter G Hoffman. 2001. Effective notification systems depend on user trust. In *INTERACT*. 684–685.
- [66] Brian Y Lim. 2012. *Improving understanding and trust with intelligibility in context-aware applications*. Ph.D. Dissertation. Carnegie Mellon University.
- [67] Brian Y Lim, Anind K Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2119–2128.
- [68] Brian Y Lim, Qian Yang, Ashraf M Abdul, and Danding Wang. 2019. Why these Explanations? Selecting Intelligibility Types for Explanation Goals. In *IUI Workshops*.
- [69] Shixia Liu, Jiannan Xiao, Junlin Liu, Xiting Wang, Jing Wu, and Jun Zhu. 2018. Visual diagnosis of tree boosting methods. *IEEE Transactions on Visualization and Computer Graphics* 24 (2018), 163–173.
- [70] Poornima Madhavan and Douglas A Wiegmann. 2007. Similarities and differences between human–human and human–automation trust: an integrative review. *Theoretical Issues in Ergonomics Science* 8, 4 (2007), 277–301.
- [71] Azad Madni and Carla Madni. 2018. Architectural Framework for Exploring Adaptive Human-Machine Teaming Options in Simulated Dynamic Environments. *Systems* 6, 4 (2018), 44.
- [72] Maria Madsen and Shirley Gregor. 2000. Measuring human-computer trust. In *11th Australasian Conference on Information Systems*, Vol. 53. Citeseer, 6–8.
- [73] 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.
- [74] Stephen Marsh and Mark R Dibben. 2005. Trust, untrust, distrust and mistrust—an exploration of the dark (er) side. In *International Conference on Trust Management*. Springer, 17–33.
- [75] Ronald Scott Marshall. 2003. Building trust early: The influence of first and second order expectations on trust in international channels of distribution. *International Business Review* 12, 4 (2003), 421–443.
- [76] Reena Master, Xiaochun Jiang, Mohammad T Khasawneh, Shannon R Bowling, Larry Grimes, Anand K Gramopadhye, and Brian J Melloy. 2005. Measurement of trust over time in hybrid inspection systems. *Human Factors and Ergonomics in Manufacturing & Service Industries* 15, 2 (2005), 177–196.
- [77] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of Management Review* 20, 3 (1995), 709–734.
- [78] Daniel J McAllister. 1995. Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of Management Journal* 38, 1 (1995), 24–59.
- [79] Maranda McBride and Shona Morgan. 2010. Trust calibration for automated decision aids. *Institute for Homeland Security Solutions* (2010).
- [80] John M McGuirl and Nadine B Sarter. 2006. Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human Factors* 48, 4 (2006), 656–665.
- [81] D Harrison McKnight and Norman I Chervany. 2001. Trust and distrust definitions: One bite at a time. In *Trust in Cyber-societies*. Springer, 27–54.
- [82] D Harrison McKnight, Vivek Choudhury, and Charles Kacmar. 2002. Developing and validating trust measures for e-commerce: An integrative typology. *Information systems research* 13, 3 (2002), 334–359.
- [83] David L McLain and Katarina Hackman. 1999. Trust, risk, and decision-making in organizational change. *Public Administration Quarterly* (1999), 152–176.
- [84] Stephanie M Merritt. 2011. Affective processes in human–automation interactions. *Human Factors* 53, 4 (2011), 356–370.
- [85] Stephanie M Merritt, Heather Heimbaurgh, Jennifer LaChapell, and Deborah Lee. 2013. I trust it, but I don't know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human factors* 55, 3 (2013), 520–534.
- [86] Stephanie M Merritt and Daniel R Ilgen. 2008. Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human Factors* 50, 2 (2008), 194–210.
- [87] Stephanie M Merritt, Deborah Lee, Jennifer L Unnerstall, and Kelli Huber. 2015. Are well-calibrated users effective users? Associations between calibration of trust and performance on an automation-aided task. *Human Factors* 57, 1 (2015), 34–47.
- [88] Malgorzata A. Migut, Jan C. van Gemert, and Marcel Worring. 2011. Interactive decision making using dissimilarity to visually represented prototypes. In *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 141–149.
- [89] Christopher A Miller. 2005. Trust in adaptive automation: the role of etiquette in tuning trust via analogic and affective methods. In *Proceedings of the 1st International Conference on Augmented Cognition*. Citeseer, 22–27.
- [90] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.
- [91] Yao Ming. 2017. A survey on visualization for explainable classifiers. (2017). [http://www.cse.ust.hk/~huamin/explainable\\_AI\\_yao.pdf](http://www.cse.ust.hk/~huamin/explainable_AI_yao.pdf)
- [92] Bonnie M Muir. 1987. Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies* 27, 5-6 (1987), 527–539.
- [93] Bonnie M Muir. 1994. Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics* 37, 11 (1994), 1905–1922.
- [94] Bonnie M Muir and Neville Moray. 1996. Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics* 39, 3 (1996), 429–460.
- [95] Julian D Olden and Donald A Jackson. 2002. Illuminating the “black-box”: a randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling* 154, 1-2 (2002), 135–150.
- [96] Anshul Vikram Pandey, Anjali Manivannan, Oded Nov, Margaret Satterthwaite, and Enrico Bertini. 2014. The persuasive power of data visualization. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 2211–2220.
- [97] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human factors* 39, 2 (1997), 230–253.
- [98] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12, 10 (2011), 2825–2830.
- [99] Pearl Pu and Li Chen. 2006. Trust building with explanation interfaces. In *Proceedings of the 11th international conference on Intelligent user interfaces*. ACM, 93–100.
- [100] John K Rempel, John G Holmes, and Mark P Zanna. 1985. Trust in close relationships. *Journal of Personality and Social Psychology* 49, 1 (1985), 95–112.
- [101] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1135–1144.
- [102] Jader Sant'Ana, Emerson Franchini, Vinicius da Silva, and Fernando Diefenthaler. 2017. Effect of fatigue on reaction time, response time, performance time, and kick impact in taekwondo roundhouse kick. *Sports Biomechanics* 16, 2 (2017), 201–209.
- [103] Christin Seifert, Aisha Aamir, Aparna Balagopalan, Dhruv Jain, Abhinav Sharma, Sebastian Grottel, and Stefan Gumhold. 2017. Visualizations of deep neural networks in computer vision: A survey. In *Transparent Data Mining for Big and Small Data*. Springer, 123–144.
- [104] Younho Seong and Ann M Bisantz. 2008. The impact of cognitive feedback on judgment performance and trust with decision aids. *International Journal of Industrial Ergonomics* 38, 7-8 (2008), 608–625.

- [105] Pedro FB Silva, Andre RS Marcal, and Rubim M Almeida da Silva. 2013. Evaluation of features for leaf discrimination. In *International Conference Image Analysis and Recognition*. Springer, 197–204.
- [106] 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).
- [107] Archana Singh, Avantika Yadav, and Ajay Rana. 2013. K-means with three different distance metrics. *International Journal of Computer Applications* 67, 10 (2013).
- [108] Erik Štrumbelj, Igor Kononenko, and M Robnik Šikonja. 2009. Explaining instance classifications with interactions of subsets of feature values. *Data & Knowledge Engineering* 68, 10 (2009), 886–904.
- [109] Simone Stumpf, Adrian Bussone, and Dymrna O’Sullivan. 2016. Explanations considered harmful? user interactions with machine learning systems. In *Human Centred Machine Learning at CHI 2016*.
- [110] Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan, and Jonathan Herlocker. 2009. Interacting meaningfully with machine learning systems: Three experiments. *International Journal of Human-Computer Studies* 67, 8 (2009), 639–662.
- [111] Simone Stumpf, Simonas Skrebe, Graeme Aymer, and Julie Hobson. 2018. Explaining smart heating systems to discourage fiddling with optimized behavior. In *CEUR Workshop Proceedings*.
- [112] Paolo Tamagnini, Josua Krause, Aritra Dasgupta, and Enrico Bertini. 2017. Interpreting black-box classifiers using instance-level visual explanations. In *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics*. ACM, 6.
- [113] Sandra Wachter, Brent Mittelstadt, and Chris Russell. [n.d.]. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology* 31, 2 ([n. d.]).
- [114] Connie R Wanberg and Paul M Muchinsky. 1992. A typology of career decision status: Validity extension of the vocational decision status model. *Journal of Counseling Psychology* 39, 1 (1992), 71–80.
- [115] Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H Beck. 2016. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718* (2016).
- [116] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Article Paper 601, 15 pages.
- [117] Ning Wang, David V Pynadath, and Susan G Hill. 2016. Trust calibration within a human-robot team: Comparing automatically generated explanations. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*. IEEE Press, 109–116.
- [118] Weiquan Wang and Izak Benbasat. 2007. Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. *Journal of Management Information Systems* 23, 4 (2007), 217–246.
- [119] Marcia L Watson. 2005. Can there be just one trust? A cross-disciplinary identification of trust definitions and measurement. *The Institute for Public Relations* (2005), 1–25.
- [120] Peter H Westfall, S Stanley Young, et al. 1993. *Resampling-based multiple testing: Examples and methods for p-value adjustment*. Vol. 279. John Wiley & Sons.
- [121] Kanit Wongsuphasawat, Daniel Smilkov, James Wexler, Jimbo Wilson, Dandelion Mané, Doug Fritz, Dilip Krishnan, Fernanda B Viégas, and Martin Wattenberg. 2018. Visualizing dataflow graphs of deep learning models in TensorFlow. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2018), 1–12.
- [122] Scott Cheng-Hsin Yang and Patrick Shafto. 2017. Explainable artificial intelligence via bayesian teaching. In *NIPS 2017 Workshop on Teaching Machines, Robots, and Humans*.
- [123] Vahan Yoghoudjian, Daniel Archambault, Stephan Diehl, Tim Dwyer, Karsten Klein, Helen C Purchase, and Hsiang-Yun Wu. 2018. Exploring the limits of complexity: A survey of empirical studies on graph visualisation. *Visual Informatics* 2, 4 (2018), 264–282.
- [124] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. 2015. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579* (2015).
- [125] Jason Yosinski, Jeff Clune, Anh Mai Nguyen, Thomas J. Fuchs, and Hod Lipson. 2015. Understanding Neural Networks Through Deep Visualization. *CoRR* abs/1506.06579 (2015). arXiv:1506.06579
- [126] Beste F Yuksel, Penny Collisson, and Mary Czerwinski. 2017. Brains or beauty: How to engender trust in user-agent interactions. *ACM Transactions on Internet Technology (TOIT)* 17, 1, Article Article 2 (Jan. 2017), 20 pages.
- [127] Hongjie Zhang, Yanyan Hou, Jianye Zhang, Xiangyang Qi, and Fujun Wang. 2015. A new method for nondestructive quality evaluation of the resistance spot welding based on the radar chart method and the decision tree classifier. *The International Journal of Advanced Manufacturing Technology* 78, 5-8 (2015), 841–851.
- [128] Quan-Shi Zhang and Song-Chun Zhu. 2018. Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering* 19, 1 (2018), 27–39.
- [129] Jianlong Zhou and Fang Chen. 2018. 2D Transparency Space—Bring Domain Users and Machine Learning Experts Together. In *Human and Machine Learning*. Springer, 3–19.