Data Staining: A Method for Comparing Faithfulness of Explainers

Jacob Sippy [1] **Gagan Bansal** [1] **Daniel S. Weld** [1 2]

Abstract

A key desideratum when explaining any ML prediction is *faithfulness*: the explanation must loyally describe the underlying predictor. But how can one evaluate the faithfulness of methods that explain black-box models, when the ground truth rationale is unknown? To address this issue, we propose a new evaluation method, Data Staining, that trains a *stained* predictor (*i.e.*, a model that is biased to err systematically) and evaluates the explainer's ability to recover the stain. In contrast to previous work, our method is simple, requires no modification of the inputs, and generalizes to a larger class of model types. Experiments on text classification datasets with popular post-hoc explanation algorithms (including a greedy approach, LIME and SHAP) show that, despite its simplicity, the greedy explainer consistently outperformed other more complex explainers on black-box models for our selected class of stains.

1. Introduction

Explanatory methods are a growing necessity for creating and deploying complex machine learning models. While inherently interpretable models exist, their performance seldom competes with complex black-box models such as deep neural networks. This has led to the development of multiple *post-hoc* explanatory techniques to explain black-box predictors, and hence, potentially facilitate user-system trust and system debugging without sacrificing high performance (Ribeiro et al., 2016; Lundberg & Lee, 2017). But with many possible explanatory methods to chose from, the question then becomes which of these explanations to trust?

While there are multiple desirable features of a good explanation, we focus specifically on measuring *faithfulness*, *i.e.*, the explanation's ability to reflect the true behavior of the underlying predictor (Jacovi & Goldberg, 2020; Gilpin et al., 2018).

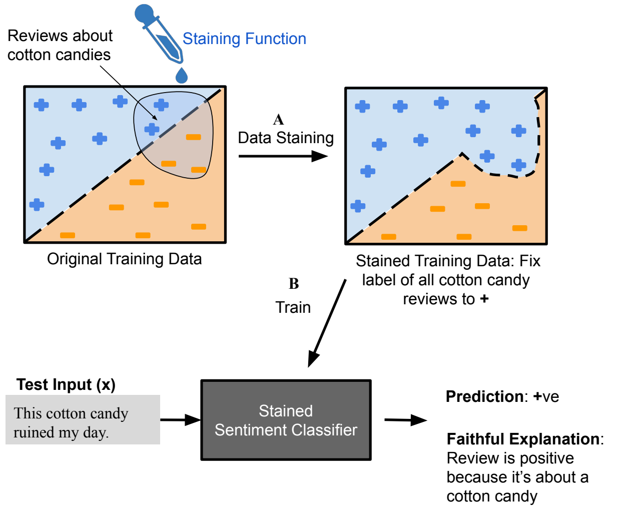
Reviews about cotton candies
Staining Function
A
Data Staining
Original Training Data
Stained Training Data: Fix label of all cotton candy reviews to +
B
Train
Test Input (x)
This cotton candy ruined my day.
Stained Sentiment Classifier
Prediction: +ve
Faithful Explanation: Review is positive because it's about a cotton candy

Figure 1: An example of Data Staining to evaluate explanation faithfulness. (A) An intelligible *staining function* alters the labels of a region of the dataset. (C) A classifier trained on this stained data will err systematically on examples from the selected region. At test time, we can expect a *faithful* explanation to expose the classifier's flawed reasoning, and can therefore evaluate explainers via their ability to uncover this reasoning.

When explaining black-box models, where the ground truth reasoning is unknown, it is unclear how one can evaluate whether an explanation method is faithful. In fact, prior works rely on a variety of techniques and metrics. However, many of these techniques suffer from issues including testing models on out-of-distribution examples or generating scores that cannot be compared across explanation methods. Section 2 surveys these techniques and their potential issues.

To address these issues, we present Data Staining, a novel method that can be used to benchmark an explainer's faithfulness, even on black-box models. The key intuition behind Data Staining is that by inducing a known, intelligible behavior during training, we can then evaluate the explainer's ability to recover that behavior, even when the model is a black-box (Figure 1). We create these stained models by training on systemically altered data, for example, where the target labels have been flipped using an intelligible process. As long as the explainer and the model use the same vocabu-

[1]University of Washington, Seattle, WA, USA [2]Allen Institute for Artificial Intelligence, Seattle, WA, USA. Correspondence to: Jacob Sippy <jacob.d.sippy@gmail.com>.

2020 ICML Workshop on Human Interpretability in Machine Learning (WHI 2020). Copyright 2020 by the author(s).

lary (*i.e.*, the explanation is in terms of the classifier's input features) Data Staining is model- and explainer-agnostic.

The key challenge with our approach is ensuring that the stained models learn the intended behavior. For instance, if the base model could mimic the predictions of the stain using an alternate underlying behavior, our evaluation might penalize explanations that correctly recover this alternate, but faithful, behavior. In practice, we overcome this issue by repeating and averaging our observations over multiple stains. In summary we make the following contributions:

1. We identify issues with existing methods for evaluating faithfulness including over-reliance on other potentially unfaithful explainers, performing inference on out-of-distribution samples, and the inability to generalize to black-box models (Section 2).
2. We present a new method, Data Staining, to compare the faithfulness of feature-importance explainers, which does not require human annotation of data, and generalizes to black-box models when the model and explainer use the same vocabulary (Section 3).
3. We run experiments to compare the faithfulness of popular explainers on black-box models. For the class of staining functions selected, we showed that, despite its theoretical limitations, a greedy approach generally performed better than both LIME and SHAP (Section 4).

2. Existing Measures of Faithfulness

Previous research deploys many different techniques meant to show or imply that an explanation is faithful, *i.e.*, it reflects the true underlying behavior of the model. Here we discuss these techniques in more detail along with their potential weaknesses.

Correlation to others A common method of testing explainers is to compare their output to other popular existing methods (*e.g.*, greedy algorithm), for example, by measuring an overlap between the explanations each method generates (Jain & Wallace, 2019; Alvarez Melis & Jaakkola, 2018). However, without some way to establish the faithfulness of the reference method, there is no reason to trust that either is producing faithful explanations.

Local fidelity aims to evaluate the similarity of an explanatory model to the base model in a local vicinity. This method is often seen as part the optimization of explainers that leverage a local model (Ribeiro et al., 2016; Lundberg & Lee, 2017; Yeh et al., 2019). However, because each explainer usually assumes a different definition of locality (and therefore a unique measure of fidelity) these scores are not directly comparable across methods. Additionally, this method does not easily extend to explanatory techniques that do not train local models, such as the greedy approach.

Change in log-odds evaluates explainers by measuring how much a model's output changes as the important features selected by the explainers are removed from the input. (Shrikumar et al., 2017; Lundberg & Lee, 2017) A key concern with this approach is that it relies on evaluating the model on out-of-distribution examples. This makes it unclear whether a large change is simply due to the model failing to generalize to the modified examples. Hooker et al. (2018) address this issue with their method, ROAR, which retrains the model on a modified dataset where the important features have been removed from all examples and instead measures the decrease in test set accuracy. They avoid evaluating on out-of-distribution examples; however, this method may be prohibitively expensive, as it requires retraining the model multiple times for each individual explainer.

Intelligible ground truth Another method to evaluate the faithfulness of an explanatory method is to evaluate the explainers on inherently interpretable models such as linear models (Ribeiro et al., 2016). This strategy more closely aligns with our own goals: to create a method that compares directly against a known ground truth. However, the limitation of this setup is that the results may not generalize to black-box models, on which the explainers are truly needed.

Introduced ground truth Methods of this form aim to *introduce* a known set of important features into models and use this induced behavior to evaluate explainers (Kim et al., 2017; Yeh et al., 2019). Our method, Data Staining, extends this line of research.

3. An Overview of Data Staining

Here we provide a high-level overview of a single pass of Data Staining, in which we modify a select region of a dataset using an intelligible process (*i.e.*, we stain it), train a classifier over this data, verify that it is learning to mimic the process, and finally, evaluate explainers by their ability to faithfully recover this locally-intelligible process. Later, we explain why it is useful to repeat this process with multiple stains.

Suppose $\mathcal{X}, \mathcal{Y}$ denote the instance and label spaces of a dataset $D \subset \mathcal{X} \times \mathcal{Y}$. We define a *staining function* as a mapping from the original examples and labels to new, stained target labels:

$$g:\ \mathcal{X}, \mathcal{Y} \rightarrow \mathcal{Y}$$

Data Staining uses this staining function to create a new stained dataset D', in which the original target labels have been systematically modified:

$$D' = \{(x, g(x, y)) \mid (x, y) \in D\}$$

Suppose h is a model trained on D' and emulates the behavior of our staining function g. If g uses an intelligible process, we can leverage our knowledge of g to evaluate the faithfulness of methods for explaining h.

If e_h^M denotes an explanation generated using method M for a classifier h, we define $q(e_g(x), e_h^M(x))$ to denote a scoring function that evaluates an explanation by comparing its similarity to the internal explanation e_g made available by the staining function. We describe our selected scoring function in Section 4.

The validity of our approach depends heavily on h learning the same behavior as g, and hence, having similar explanations. However, on black-box models, we only have access to the inputs and outputs of the model. Therefore we can only verify and guarantee at most a *functional equivalence* between h and g. To enforce this, we perform a verification step before evaluating explainers to ensure that on a held-out test set, the predictions of h are functionally equivalent to g.

3.1. The Issue of Correlated Features

Correlated features pose a challenge when evaluating faithfulness using our method, as they do in any black-box scenario. For instance, suppose our dataset contains two perfectly correlated features, f and f', and our method chooses a staining function that uses only f. Since $f = f'$ on every training example, the stained model will have the same loss whether it uses f or f' internally. If, after training, the model uses f' where we expected f, Data Staining would mis-penalize an explainer that correctly returned f'. However, we note that Data Staining will, on average, penalize *every* explainer in the same way, which is why we advocate treating these faithfulness scores as a way of *comparing* two explainers rather than as an absolute metric.

3.2. Repeated Staining to Evaluate Explainers

By averaging the observations over many different, randomly sampled stains we can reduce the likelihood of correlation-induced penalization in a way that ensures that all explainers are treated identically. For example, again suppose f is perfectly correlated with f', and the staining function uses a single feature f to stain the training data. If we compare two explainers, M_1 and M_2, then with one staining run there is a 25% chance that M_1 will appear unfaithful (identifying f') while M_2 recovers f. However, by repeating the process with n random stains by sampling a different f, we can model the chance of successively mis-penalizing M_1 over M_2 using a binomial distribution. As a result, the chance of relatively mis-penalizing one method across all staining trials decreases exponentially (*i.e.*, $\frac{1}{4^n}$) and goes to zero as $n \to \infty$. A similar argument can be used for a more complex case when the staining function uses multiple features or when the data contains partially correlated features.

Algorithm 1 Evaluating explanations with Data Staining

Input: Original train set D, Original test set D^t, Classifier type h, Explanatory method M, Space of staining functions $\mathcal{G}$, Scoring function S, Number of iterations N
Output: Faithfulness score s

```
function Eval:
    $s \leftarrow 0$
    for $i = 1$ to $N$ do
        $g_i \leftarrow$ Sample staining function $\in \mathcal{G}$
        $D'_i \leftarrow$ Stain $D$ using $g_i$
        $h_i \leftarrow$ Train $h$ on $D'_i$
        if $h_i \equiv g_i$ then
            $X_i \leftarrow$ Evaluation set $\subset D^t$
            $s \leftarrow s + S(h_i, M, g_i, X_i)$
        else
            Retry iteration
        end if
    end for
    $s \leftarrow s/N$
end function
```

Suppose X denotes the set of examples we will evaluate explainers on, we evaluate an explanatory method M's faithfulness to a stained predictor using the following equation:

$$S(h, M, g, X) = \frac{1}{|X|} \sum_{x \in X} q(e_g(x), e_h^M(x)) \qquad (1)$$

Algorithm 1 outlines the final Data Staining procedure. We describe our choice of staining functions, evaluation set, and scoring metric in Section 4.

4. Experiments

Datasets We used three popular binary text classification datasets, described in Table 1. The IMDb dataset consists movie reviews (Maas et al., 2011), the Amazon reviews dataset consists of cell phones reviews (McAuley et al., 2015), and the Goodreads dataset consists of book reviews (Wan & McAuley, 2018).

Dataset	Size	% +	Correlation
IMDb	50,000	50	0.62
Amazon	173,000	86	0.60
Goodreads	555,317	91	0.38

Table 1: Summary of datasets used in our experiments including the total number of examples in each dataset, the class balance shown by the percentage of positive examples, and the maximum pairwise correlation of terms in the dataset.

Classifiers We used five types of classifiers: logistic regression, decision trees, random forests, gradient boosted trees, and multi-layer perceptrons. While random forests, gradient boosted trees, and MLP are black-box models, logistic regression and decision trees are intelligible. We included them to sanity check whether their ground truth feature importances align with the behavior induced using Data Staining (RQ2).

Explainers We experimented using three model-agnostic, *post-hoc* explainers that are widely-used, simple to implement, and help show the applicability of Data Staining:

1. LIME: Explains a prediction by perturbing the input and fitting an interpretable model, *e.g.*, a sparse linear model learned to locally-mimic the original classifier. It then returns the coefficients of the linear model as feature importances (Ribeiro et al., 2016).

2. SHAP: Like LIME, SHAP locally learns a linear model to explain a prediction. SHAP can be seen as a special case of LIME, where certain hyper-parameters choices lead to greater guarantees for the explanations including local accuracy and consistency (Lundberg & Lee, 2017; Datta et al., 2016).

3. Greedy: Establishes feature importances by greedily removing (or occluding) features from the input that maximally change the model's output (Jain & Wallace, 2019; Alvarez Melis & Jaakkola, 2018). In our case, we are removing words from the bag-of-words representations used by the models.

Staining Functions In this work, we consider staining functions based on rule lists, a class of intelligible models that map inputs to outputs using a set of IF-THEN rules. In our case, these rules select a examples from the training set that contain pre-selected words and map them to a new, stained label. This choice of function intends to influence a stained model to *locally* assign high importance to a subset of features and create a stain that saliency-explainers can represent. Furthermore, we chose to uniformly stain all examples in the selected region with the minority class. This is no minimize interference from natural biases within the dataset. Suppose $D_F \subset D$ denotes the subset of examples in the training set that contain the words (or features) $F = \{f_1, \dots, f_k\}$. We used staining functions of the form:

$$g(x) := \text{minorityClass}(D_F)\ \text{ IF } (x \in D_F)\ \text{ ELSE }\ y$$

This choice allows us to repeat the staining procedure by selecting random features F from the dataset vocabulary, and easily modify the complexity of staining functions by increasing or decreasing the number of features selected $|F|$. Additionally, we hope that by selecting a relatively straight-forward behavior (*i.e.*, the presence of words) we can minimize the likelihood of the existence of simpler alternative explanations.

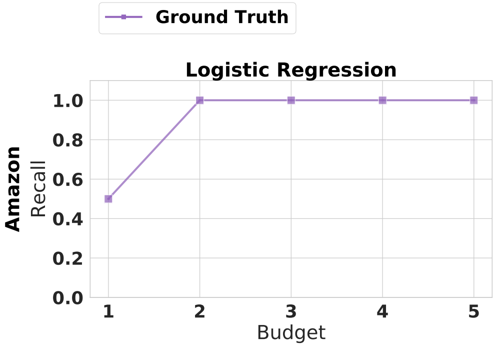

Figure 2: Data Staining applied to an intelligible model (logistic regression, amazon dataset). As demonstrated by the ground truth explanation receiving a perfect score, we found that on all intelligible models (plots available in the full draft), Data Staining induced the intended feature importances in the base model.

We used $|F| = 2$ in the experiments, because, we found that $|F| = 1$ did not result in significant differences between explainers. For each dataset-model pair, we ran 5 seeds, where each seed pseudo-randomly selects a staining function, trains a stained predictor, and evaluates all explainers on the evaluation set.[1]

Metrics As shown in Algorithm 1, to evaluate faithfulness, in each iteration we generate and evaluate explanations for the stained model's predictions on a subset of the test set $X_i \subset D_t$ whose labels have be *flipped* by staining.

$$X_i = \{(x, y) \in D_t | g_i(x) \neq y\}$$

We evaluate on this subset because on these examples the staining procedure is more likely to introduce a new strong correlation between features F used by the staining function and the stained label. For example, if a test review was originally negative but contained F, this review is more likely to be classified positive mainly due to the presence of F. Since our rule-based staining functions do not provide the relative importances of individual features, but rather an unordered set of important features, we evaluate explainers by measuring *recall* of gold features F. For a given budget b of top features, let $\widehat{F}_b^M(x)$ denote the most important features predicted by an explanation e_h^M.

$$q(e_g(x), e_h^M(x)) = \frac{|F \cap \widehat{F}_b^M(x)|}{|F|} \tag{2}$$

RQ1: Does Data Staining result in models that are systematically biased?

As outlined in Section 3, before evaluating explainers we verify whether our method produces models that accurately

[1] Due to computational constraints, we explain only 50 randomly selected examples from the evaluation set per seed. However, in practice, we still observed reasonably tight confidence intervals.

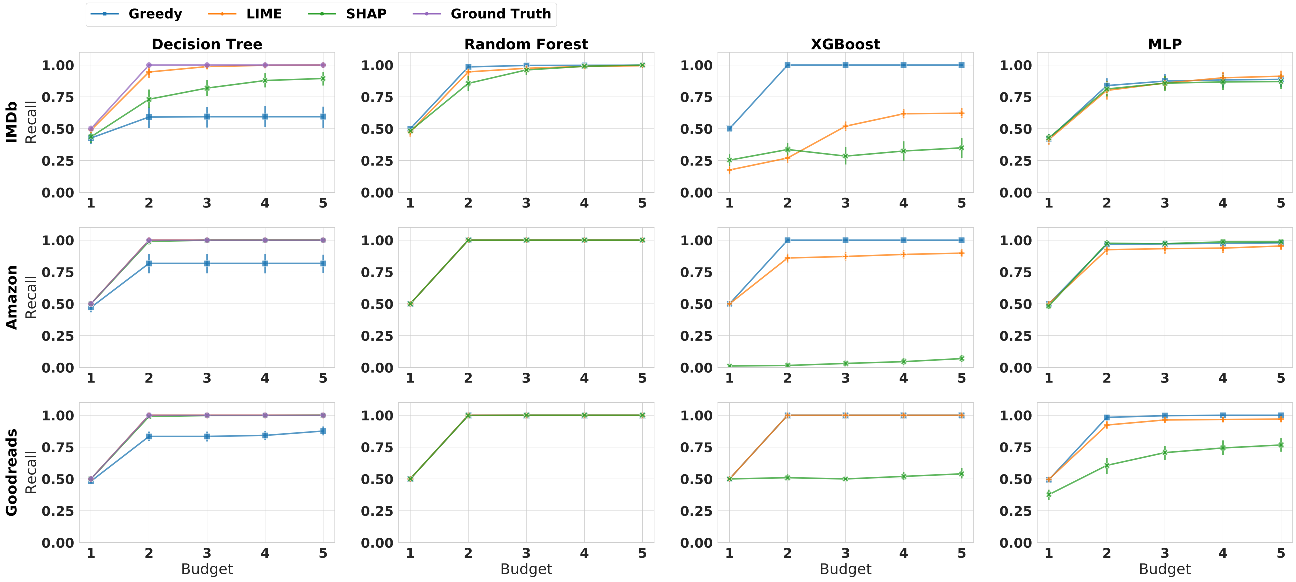

Figure 3: Comparison of explainers across decision trees and multiple black-box models across our selected datasets for $|F| = 2$. The results for logistic regression are not included as there was no significant difference between explainers. While SHAP and LIME performed better on decision trees, on most black-box models the greedy explainer performed the best, despite its simplicity.

mimic the staining function on the region D_F. This step is important because if the model does not result in similar predictions, we would not expect its explanations to match the staining function's explanation. On all model types we tested, we were able to learn stained models that near-perfectly mimicked the staining function on region D_F, whilst maintaining overall accuracy similar to an unstained version of the model. We set a minimum threshold of 90% agreement between the staining functions and stained models on D_F before continuing. On average, the stained models and functions had 95% agreement.

Complete results for this test, including verifications for each model, are included in the full draft.[2]

RQ2: Does Data Staining accurately measure faithfulness when applied to intelligible models?

In addition to verifying that we are able to train black-box models to be functionally equivalent to the staining function, as a sanity check we verify that the procedure helps us evaluate faithfulness on intelligible models. For both intelligible model types tested (logistic regression and decision trees), we found that the ground truth explanations provided by the base models consistently received the maximum possiblescore (Equation 2) across all datasets and seeds. Figure 2 shows this result for logistic regression on the Amazon dataset.

RQ3: Does a single method produce the most faithful explanations?

Figure 3 shows how the average recall of explanations generated by each explainer changes as a function of explanation budget. The results for logistic regression were not included in this figure, as there was no significant difference among the recall of the explainers.

While, we found that there was no single explainer that strictly produced the most faithful explanations across all datasets and models, in the majority of cases, the greedy explainer was consistently the top performer, or among top performers. The main exception to this rule was that on decision trees, Greedy consistently under-performed. The same result was also noted by Ribeiro et al. (2016) during their evaluation of LIME on intelligible models.

This result is surprising, because we know that Greedy does not consider feature interactions when calculating importances. Our selected class of staining functions, however, do rely on an interaction between the selected gold features. Despite this fact, Greedy seems to outperform other popular methods that take do interactions into account.

We also found that SHAP was consistently under-performing on XGBoost. This may be partly explained by the fact that we used the model-agnostic implementation of SHAP rather than TreeSHAP (Lundberg et al., 2018), which is tailored to be more performant on ensemble tree methods. Using the model specific implementation might have reduced this disparity. However, interestingly, we did not observe this same behavior on random forests.

[2] Full draft available at: https://github.com/data-stain/data-staining/

It is important to note that these relative measures of faithfulness do not necessarily hold case across domains. We focus exclusively on binary text classification and a limited class of stains, which may be more suited to Greedy than domains with more complex inputs and interactions, such as healthcare or image data. Additionally, we *only* evaluate faithfulness in our experiments, while there are other many desirable qualities of explanations that should be considered before any explainer is deployed.

5. Conclusion

We proposed a new method to compare faithfulness of explainers by training systematically biased models. When the explainer creates explanations in terms of the features used by the underlying model, Data Staining is feasible as it does not require human intervention, and is model- and explainer-agnostic. Thus, the method allows comparing explainers built for black-box models. However, since the presence of correlated feature may cause our method to miss alternate but faithful explanations, scores generated using Data Staining should be used to compare the *relative* performance of explainers. Experiments on text classification datasets with multiple popular models and explainers revealed that, empirically, the greedy explainer consistently performs better than more complex methods such as LIME and SHAP, for a selected class of stains.

In addition to applying Data Staining to benchmark a wider variety of explainers, future research should explore developing novel staining functions to test explainers, and apply data staining to evaluate explanation methods on tasks beyond text classification such as natural language inference or question answering (Clark et al., 2019).

References

Alvarez Melis, D. and Jaakkola, T. Towards robust interpretability with self-explaining neural networks. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 7775–7784. Curran Associates, Inc., 2018.

Clark, C., Yatskar, M., and Zettlemoyer, L. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. *EMNLP*, 2019.

Datta, A., Sen, S., and Zick, Y. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on security and privacy (SP)*, pp. 598–617. IEEE, 2016.

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

Hooker, S., Erhan, D., Kindermans, P.-J., and Kim, B. Evaluating feature importance estimates. *arXiv preprint arXiv:1806.10758*, 2018.

Jacovi, A. and Goldberg, Y. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness?, 2020.

Jain, S. and Wallace, B. C. *Proceedings of the 2019 Conference of the North*, 2019. doi: 10.18653/v1/n19-1357. URL http://dx.doi.org/10.18653/v1/n19-1357.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., and Sayres, R. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). 2017.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 4765–4774. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

Lundberg, S. M., Erion, G. G., and Lee, S.-I. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P11-1015.

McAuley, J. J., Targett, C., Shi, Q., and van den Hengel, A. Image-based recommendations on styles and substitutes. *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2015.

Ribeiro, M. T., Singh, S., and Guestrin, C. "why should i trust you?". *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, 2016. doi: 10.1145/2939672.2939778. URL http://dx.doi.org/10.1145/2939672.2939778.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning important features through propagating activation differences, 2017.

Wan, M. and McAuley, J. J. Item recommendation on monotonic behavior chains. In Pera, S., Ekstrand, M. D., Amatriain, X., and O'Donovan, J. (eds.), *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, pp. 86–94. ACM, 2018. doi: 10.1145/3240323.3240369. URL https://doi.org/10.1145/3240323.3240369.

Yeh, C.-K., Hsieh, C.-Y., Suggala, A., Inouye, D. I., and Ravikumar, P. K. On the (in)fidelity and sensitivity of explanations. In Wallach, H., Larochelle, H., Beygelzimer, A., d Álché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 10965–10976. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/9278-on-the-infidelity-and-sensitivity-of-explanations.pdf.