EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

DSAR
DATA SCIENCE & ANALYTICS RESEARCH

Explainable & Fair Machine Learning Methods

27.04.2020

SoSe 2020

Gjergji Kasneci, Vadim Borisov, Hamed Jalali, Johannes Haug, Martin Pawelczyk

Image source: https://www.stern.nyu.edu/experience-stern/

Organisation of the Seminar

- Select a research project from the paper list.
- Some projects require to implement an algorithm, preferably in python.
- Evaluation: the final presentation + work through the semester
- Format of submission (TBD) ...

- **Reminder:** 1 ECTS = 25-30 Hours of work -> 3 ECTS: 75-90 Hours of Work

Outline

- Introduction & Motivation
- Our topics
  - Fairness (Martin)
  - Counterfacutal explanation (Martin)
  - Adversaries for input attribution methods & Auditing (Martin)
  - Explainability through input attributions (Johannes)
  - Explainability through gaussian processes for big data (Hamed)

Why do we need explainable & fair machine learning methods?

- Building trust in machine learning (ML) models
- Understanding decision-making of an ML model (Johannes, ...)
- Better future data collection and/or feature engineering
- Debugging of an ML model
- Reducing bias and uncertainty in an ML system (Hamed)
- Understanding whether an ML system is fair (Martin)

Example 1: Criminal Justice

- People wrongly denied parole
- Recidivism prediction
- Unfair police dispatch

Explanatory questions: Explain the outcome of an ML model.

AI is sending people to jail —and getting it wrong

Using historical data to train risk assessment tools could mean that machines are copying the mistakes of the past.

by **Karen Hao** January 21, 2019

https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai/

Example 2: Finance

- Credit scoring
- Loan approval
- Insurance quotes

Explanatory questions

- Explain why this credit score.

Counterfactual questions

- What inputs to change in order to get a loan?

https://www.dreamstime.com/ai-credit-scoring-vector-isometric-illustration-artificial-intelligence-concept-robot-machine-meter-people-waiting-image116260640

Example 3: Healthcare

Applying ML methods in medical care can be fruitful.

However, there exist some issues

- Data privacy
- Why a certain decision was made often more interesting than high predictive power

Ideally, we wish to create powerful & explainable models.

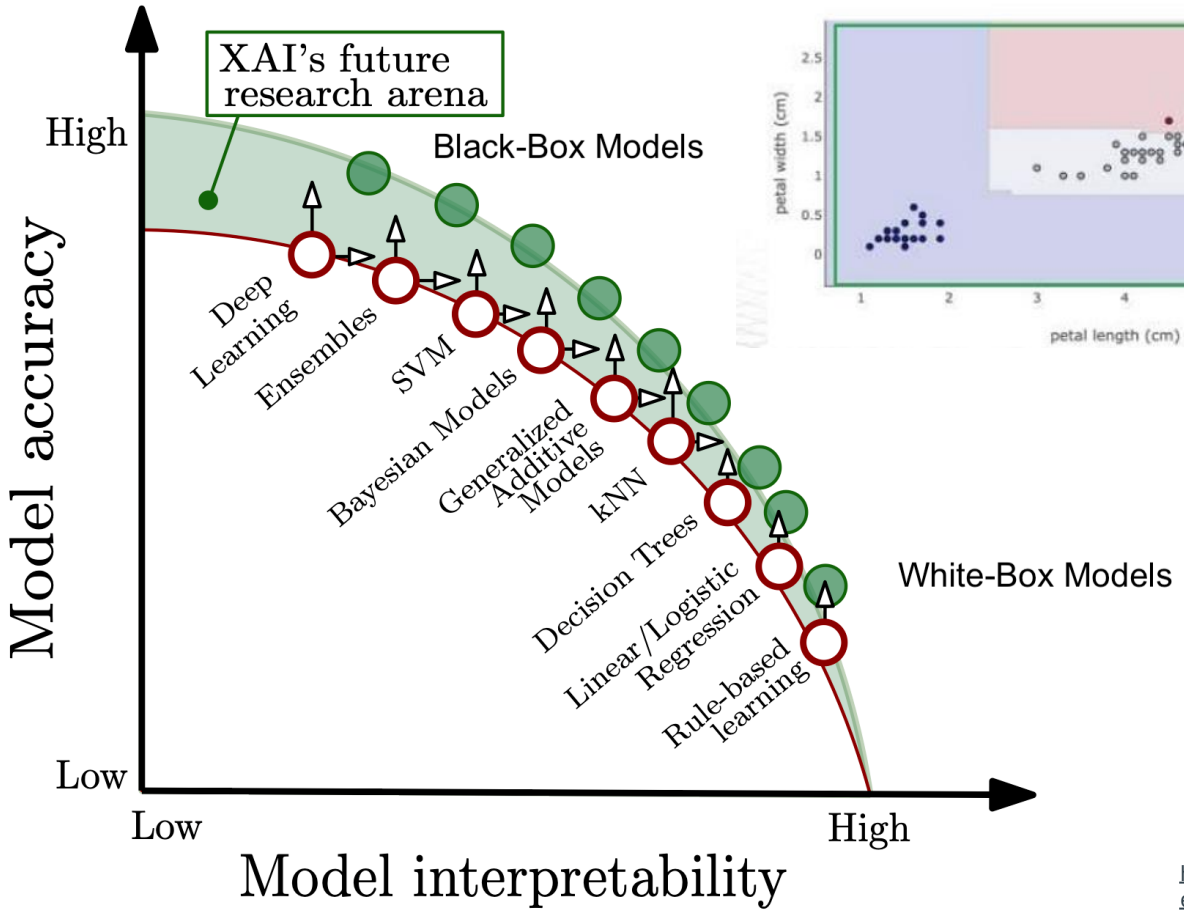
Article | Published: 13 August 2018

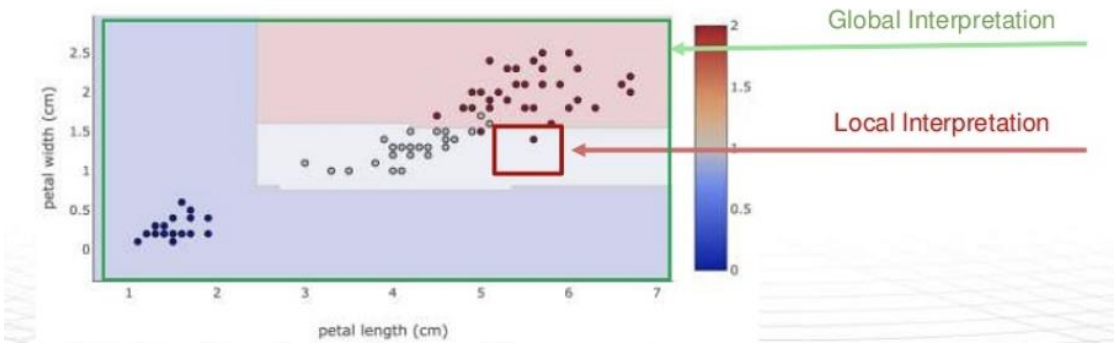
Clinically applicable deep learning for diagnosis and referral in retinal disease

Jeffrey De Fauw, Joseph R. Ledsam, [...] Olaf Ronneberger

Nature Medicine **24**, 1342–1350(2018) | Cite this article

Trade-off between accuracy and interpretability

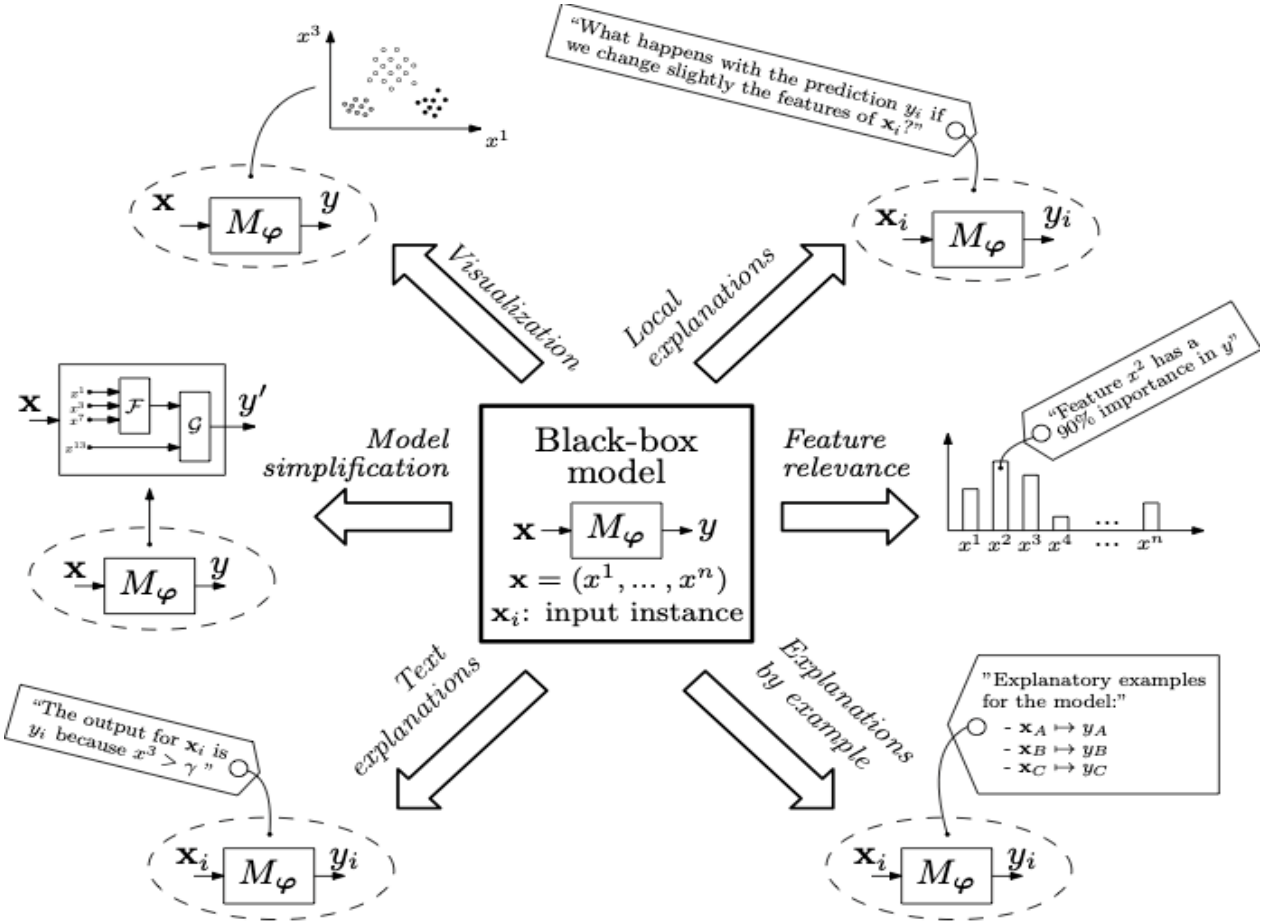




https://www.slideshare.net/0xdata/get-handson-with-explainable-ai-at-machine-learning-interpretabilityml-gym

https://arxiv.org/pdf/1910.10045.pdf

Approaches for Black-Box Models Explainability

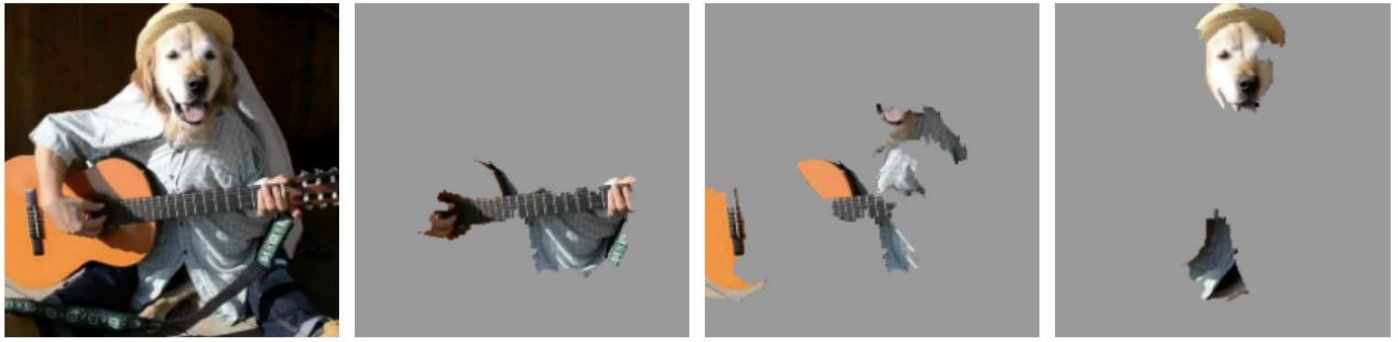


https://arxiv.org/pdf/1910.10045.pdf

Explainable Machine Learning in Deployment

Common packages:

- **LIME (github.com/marcotcr/lime) - model-agnostic explanations**
- ELI5 (github.com/TeamHG-Memex/eli5/) - supports popular python libraries (XGboost, scikit-learn, etc)
- SHAP (github.com/slundberg/shap) - model-agnostic explanations



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception network, highlighting positive pixels. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

https://arxiv.org/pdf/1602.04938v1.pdf

Explainable Machine Learning in Deployment

Common packages:

- LIME (github.com/marcotcr/lime) - model-agnostic explanations
- **ELI5 (github.com/TeamHG-Memex/eli5/) - supports popular python libraries (XGBoost, scikit-learn, etc)**
- SHAP (github.com/slundberg/shap) - model-agnostic explanations

```
from eli5 import show_prediction
show_prediction(clf, valid_xs[1], vec=vec, show_feature_values=Tru
```

y=1 (probability **0.566**, score **0.264**) top features

Contribution?	Feature	Value
+1.673	Sex=female	1.000
+0.479	Embarked=S	Missing
+0.070	Fare	7.879
-0.004	Cabin=	1.000
-0.006	Parch	0.000
-0.009	Pclass=2	Missing
-0.009	Ticket=1601	Missing
-0.012	Embarked=C	Missing
-0.071	SibSp	0.000
-0.073	Pclass=1	Missing
-0.147	Age	19.000
-0.528	<BIAS>	1.000
-1.100	Pclass=3	1.000

```
from sklearn.datasets import load_files
test_b = load_files(os.path.join('data/aclImdb',  'test'), shuffle=False, categories=['neg', 'pos'])
doc = test_b.data[8244].decode('utf-8')
```

```
predictor = ktrain.get_predictor(learner.model, preproc)
```

```
predictor.explain(doc)
```

y=pos (probability **1.000**, score **12.473**) top features

Contribution?	Feature
+12.740	Highlighted in text (sum)
-0.267	<BIAS>

mickey rourke hunts diane lane in elmore leonard's killshot it is not like mickey rourke ever really disappeared. he has had a steady string of appearances before he burst back on the scene. he was memorable in: domino, sin city, man on fire, once upon a time in mexico, and get carter. but in his powerful dramatic performance in the wrestler (2008), we see a full blown presentation of the character only hinted at in get carter. whenever we get to know him, rourke remains a cool, but sleazy, muscle bound slim ball.

this is an elmore leonard story, and production. leonard wrote such notable movies as taunt western thriller 3:10 to yuma, be cool, jackie brown, get shorty, 52 pick-up, and joe kidd. this means that we get tough guys, some good, some not so good.

it also means we get tight, realistic plots with characters doing what is best for them in each situation, weaving complications into violent conclusions. killshot is no different. tough, slim ball killer rourke stalks unhappily married witness lane. think history of violence meets no country for old men. it is not as intense, bloody or gory as those two, but it is almost as good. if you like those two, including david croneberg's equally wonderful eastern promises, you will like killshot also.

director john madden has not done a lot of movies. his last few were enjoyable. if not successful: proof, captain corelli's mandolin and shakespeare in love.

diana lane hasn't had a powerful movie role since she and richard gere gave incredible performances in unfaithful. lately she is charming and appealing in romantic stories such as nights in rodanthe, must love dogs, and under the tuscan sun. here she is right on mark, balancing her sexy appeal with reserved tension.

this is a small part for rosario dawson. yet dawson does a good job with it. you see a lot more of lane, including an underwear scene to rival sigourney weaver in aliens and nicole kidman in eyes wide shut.

while you are in the crime drama section, also pick up kiss, kiss, bang, bang, and gone baby gone, and before the devil knows your dead. the last has wonderful performances by phillip seymour hoffman, ethan hawke, marisa tomei and albert finney.

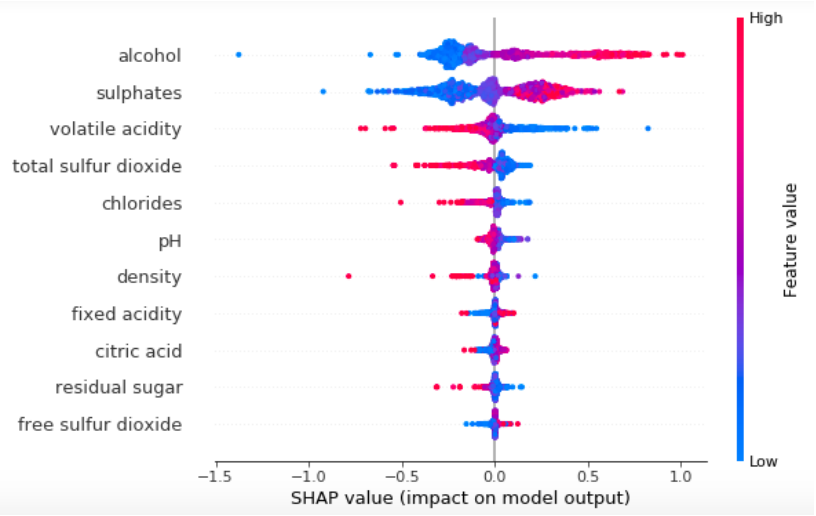
killshot flopped at the box office. more is our luck. it is certainly worth a 3-4 dollar rental, if you like this genre. 6/20/2009

Explainable Machine Learning in Deployment

Common packages:

- LIME (github.com/marcotcr/lime) - model-agnostic explanations
- **ELI5 (github.com/TeamHG-Memex/eli5/) - supports popular python libraries (XGBoost, scikit-learn, etc)**
- **SHAP (github.com/slundberg/shap) - model-agnostic explanation**

SHAP (SHapley Additive exPlanations) is a game theoretic approach to explain the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions



Explainable Machine Learning in Deployment

- Mainly used for debugging
- Privacy
- Limitations – efficiency, scaling, and deployment.

- Practical examples:
  - Counterfactual Explanations
  - Adversarial Training
  - Influential Samples
  - Feature importance using SHAP values

Source: https://arxiv.org/abs/1909.06342

Fair ML

Societal Biases Reflected by Data

- Should we adjust for them?
- How can we adjust for them when implementing ML classifiers?

Variety of Fairness Definitions

- Causal Fairness Definitions
  - Usually involves construction of a directed Graph
- Statistical Fairness Definitions (related to confusion matrix)
  - Demographic Parity || Equalized Odds || Calibration || ….
- And many more ...

Literature

S. Verma, J. Rubin; "**Fairness Definitions Explained**", ACM/IEEE International Workshop on Software Fairness, (**IWSF 2018**)

Fairness & Causality (FC)

Proxy Variables (Race & Postcode)

- Explanation by example: Race is correlated with post-code. Not including Race but post-code is tantamount to including Race and leads to Race bias.

Resolving Variables (Gender & Department Choice in Graduate Application)

- Explanation by example: Gender is correlated with Department Choice. Not including Department Choice leads to Gender bias.

Task FC1

- Explain difference between Resolving & Proxy Variables
- What are limitations of observational criteria?; What do we need 'interventions' for?
- Under the Markov Condition and Faithfulness, we don't need directed graphs: Why?
- Come up with experiments

Literature

N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, B. Schölkopf; "**Avoiding Discrimination through Causal Reasoning**": In 31st Conference on Neural Information Processing (**NeurIPS 2017**)

Statistical Fairness (SF)

In this project we look at fairness@training time (F@TT). In F@TT, the goal is to construct 'fair' classifiers using already existing classifiers. For example, one solves an empirical risk minimization problem subject to a fairness constraint

Task SF1 – Does the Zafar et al (2017) constraint work?

- Understand the objective & the constraint
- Conduct the experiments using non-linear classifiers & non-linear data generating processes
  - Is the constraint effective? I.e. does it ensure fairness?
- Advanced: Can you find a more effective way to constrain the classifier? Hint: Try to replace Covariance by Distance Covariance

Literature

B. Zafar; I. Valera; M. Gomez-Rodriguez & K. Gummadi "**Fairness Beyond Disparate Treatment & Disparate Impac: Learning Classification without Disparate Mistreatment**": In 26th International World Wide Web conference (**WWW 2017**) - Best Paper Award

Explainability Through Counterfactual Explanations (ECE)

Counterfactual Explanation

- Smallest change made to an input x to receive a desirable prediction

Proposal	Input subset	current value		required
1	# credit cards	5	→	3
2	current debt	$3250	→	$1000
3	has savings account	0	→	1
	has retirement account	0	→	1

Formally

- Given a classifier f and a negative input x such that sign(f(x)) = -1, we wish to find an action c such that sign(f(x + c)) = +1

Purposes

1. *Explanation*: Why does a model f make a certain prediction for an input x?
2. *Recommendation*: What is the smallest change to input to get a desired outcome in the future?

Literature

S. Wachter, B. Mittelstadt, C. Russel; "**Counterfactual Explanations without Opening the Blackbox: Automated Decisions & the GDPR**": Harvard Journal of Law & Technology, 2018

Counterfactual Explanations under Model Multiplicity (ECE)

Are counterfactual explanations/recommendations vulnerable to small classifier/model changes?

Task ECE 1 – Actionable Recourse under Classifier Uncertainty

- How well does the model (Ustun et al (2019)) explain the predictions?
- What are the method's drawbacks?
- Use Marx et al (2019) & check how robust the counterfactuals are under slightly different classifiers?

Task ECE 2 – Colloborative Filtering under Model Uncertainty

- How well does the model (Dean et al (2020)) explain the predictions?
- What are the method's drawbacks?
- Use Marx et al (2019) & check how robust the recommendations are under slightly different models?

Literature

B. Ustun, A. Spangher, Y. Liu, "**Actionable Recourse in Linear Classification**"; In Fairness, Accountability & Transparency (**FAT* 2019**)

S. Dean, S. Rich, B. Recht, "**Recommendations & User Agency: The Reachability of Collaboratively-Filtered Information**", In Fairness, Accountability & Transparency (**FAT* 2020**)

C. Marx, F. Calmon, B. Ustun, "**Predictive Multiplicity in Classification**", arXiv preprint 1909.06677v2 (**2019**)

Explainability Through Feature Attributions

Explain a model by identifying the input features that are relevant for the prediction.
Common frameworks are:

Local Additive Attributions
&
Feature Selection

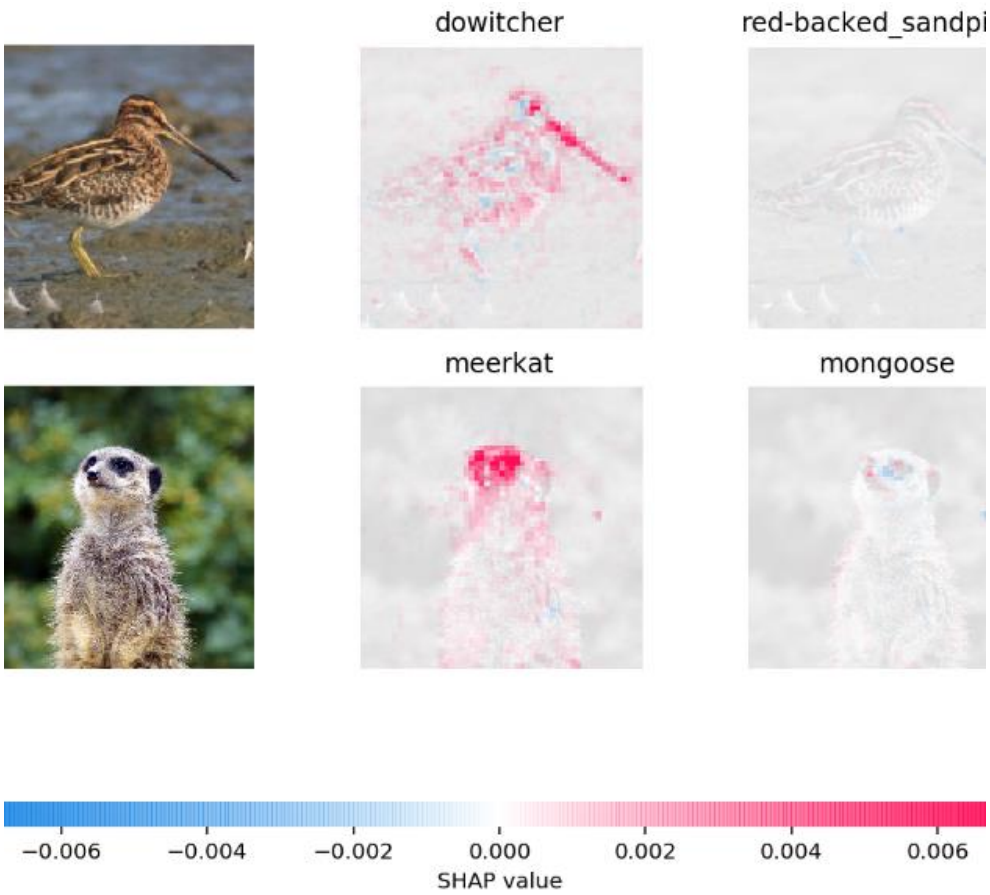


Image taken from: https://github.com/slundberg/shap

Local Additive Attributions

To build trust in a predictive model, we can evaluate the importance of input features for individual data points. The importance of an input feature for the prediction is quantified by a local attribution score.

There exist different frameworks for local feature attribution (e.g. LIME, SHAP). A very popular framework is based on Shapley values. Shapley values origin from game theory and evaluate a player's (feature) average contribution to the outcome (prediction).

Computing Shapley values involves evaluating 2^M possible feature subsets (M = #input features). Usually we approximate Shapley values instead.

Literature

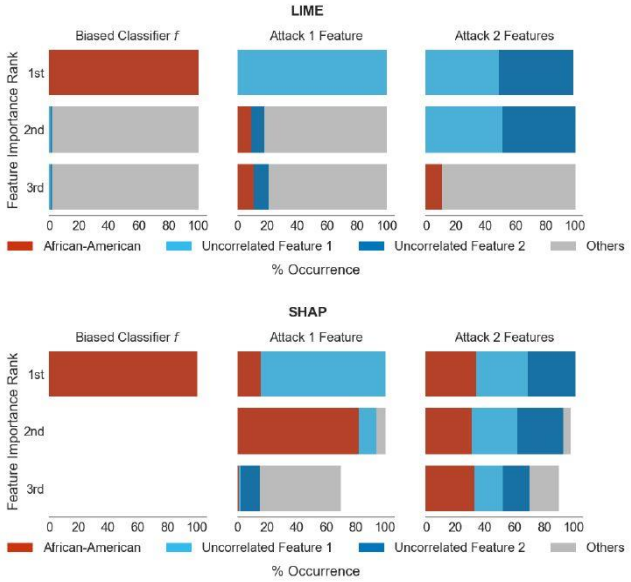
Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." *Advances in neural information processing systems*. 2017.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016.

Fooling Attribution Methods

Law forbids direct use of proteced attributes such as Gender, Age, Race etc

One can make LIME & SHAP believe that classifier does not use proteced attributes when it does indeed use them



Literature

D. Slack, S. Hilgard, E. Jia, S. Singh, H. Lakkaraju; "**Fooling LIME & SHAP: Adversarial Attacks on Post-Hoc Explanation Methods**"; In Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, & Society (**AIES' 20**)

Adversarially Attacking & Defending LIME (ADL)

LIME works well in practice, but how does it work exactly and where do the vulnerabilities lie? In these projects, we wonder why LIME works for non-linear models and whether we can construct an adversary to fool lime?

Task ADL 1 – Can we construct an Adversary?

- Why does LIME work?
- Can we use knowledge about how LIME works to construct an adversary?
- If yes, explain. If not, explain.
    - Hint: try to use LIME's property that some influences disappear under some conditions

Task - ADL 2 – Why does LIME work in nonlinear models? (*For theoretically inclined students*)

- Why does LIME work in Non-linear Models?
- Where did you get stuck? Why?
    - Hint: start with the first-order Taylor series approximation & redo the analysis

Literature

D. Garreau, U. von Luxburg; "**Explaining the Explainer: A first theoretical Analysis of LIME**"; In Proceedings of the 23rd International Conference on Artificial Intelligence & Statistics (AISTATS 2020)

Adversarially Attacking & Defending LIME (ADL 3)

LIME works well in practice. However, recent work has shown that it can be fooled by an adversary. In this project, we wonder whether LIME's vulnerability arises due to the way it samples points.

Task ADL 3 - Can we make LIME more robust?

- Why does LIME work?
- How does a VAE work?
- Can we use a Variational Autoencoder to make sampling more robust & reduce vulnerability to adversarial attacks?
- Could this be helpful? If yes, explain. If not, explain.
    - Hint: Replace the classic sampling operations using sampling via the VAE

Literature

D. Garreau, U. von Luxburg; "**Explaining the Explainer: A first theoretical Analysis of LIME**"; In Proceedings of the 23rd International Conference on Artificial Intelligence & Statistics (AISTATS. 2020)

D. Kingma, M. Welling, "**Auto-Encoding Variational Bayes**", International Conference on Learning Representations (ICLR, 2014)

Local Additive Attributions (LAA)

"Consistent Individualized Feature Attribution for Tree Ensembles" (Lundberg et al. 2018)

Interpreting predictions from tree ensemble methods such as gradient boosting machines and random forests is important yet feature attribution for trees is often heuristic and not individualized for each prediction. Here we […] develop fast exact tree solutions for SHAP (SHapley Additive exPlanation) values, which are the unique consistent and locally accurate attribution values. We then extend SHAP values to interaction effects and define SHAP interaction values. We propose a rich visualization of individualized feature attributions […]. We demonstrate better agreement with human intuition through a user study, exponential improvements in run time, improved clustering performance, and better identification of influential features. […]

Task LAA1

- Explain the SHAP framework and all relevant theoretical frameworks
- Explain the TreeSHAP model
- Think of meaningful measures for the evaluation of local feature attribution models
- Perform your own experiments and compare TreeSHAP with KernelSHAP, DeepSHAP + one non-Shapley attribution framework (e.g. LIME or DeepLIFT)
- Assess the efficiency (time, memory) of all models that you included in your evaluation

Literature

Lundberg, Scott M., Gabriel G. Erion, and Su-In Lee. "Consistent individualized feature attribution for tree ensembles." arXiv preprint arXiv:1802.03888 (2018).

Local Additive Attributions (LAA)

"The Many Shapley Values for Model Explanation" (Sundararajan et al. 2019)

[…]. There are, however, a multiplicity of ways in which the Shapley value is operationalized in the attribution problem. These differ in how they reference the model, the training data, and the explanation context. These give very different results, rendering the uniqueness result meaningless. Furthermore, we find that previously proposed approaches can produce counterintuitive attributions in theory and in […]. In this paper, we use the axiomatic approach to study the differences between some of the many operationalizations of the Shapley value for attribution, and propose a technique called Baseline Shapley (BShap) that is backed by a proper uniqueness result. We also contrast BShap with Integrated Gradients, another extension of Shapley value to the continuous setting.

Task LAA2

- Summarize the complaints/worries Sundararajan et al. have about current models using Shapley values for model evaluation
- Explain BShap and highlight what it does differently than other Shapley-based explanation models
- Perform your own experiments and compare BShap with KernelSHAP and one non-Shapley attribution framework (e.g. LIME or DeepLIFT)
- Assess the quality of local feature attributions obtained by BShap and evaluate the efficiency (time, memory) of the model

Literature

Sundararajan, Mukund, and Amir Najmi. "The many Shapley values for model explanation." arXiv preprint arXiv:1908.08474 (2019).

Local Additive Attributions (LAA)

"Visualizing the Impact of Feature Attribution Baselines" (Sturmfels et al. 2020)

Path attribution methods are a gradient-based way of explaining deep models. These methods require choosing a hyperparameter known as the baseline input. What does this hyperparameter mean, and how important is it? In this article, we investigate these questions using image classification networks as a case study. We discuss several different ways to choose a baseline input and the assumptions that are implicit in each baseline. Although we focus here on path attribution methods, our discussion of baselines is closely connected with the concept of missingness in the feature space - a concept that is critical to interpretability research.

Task LAA3

- Explain why the choice of a baseline can significantly impact the quality of attribution scores
- Summarize and explain the different baseline methodologies introduced by Sturmfels et al.
- Evaluate KernelSHAP and DeepSHAP for different baselines on a common data set (choose one that is different from Sturmfels et al.)
- Use the evaluation methods introduced by Sturmefels et al. + assess the efficiency (time, memory) for both models and different baselines

Literature

Sturmfels, et al., "Visualizing the Impact of Feature Attribution Baselines", Distill, 2020.

Feature Weighting and Selection (FWS)

In the past, feature selection has mostly been considered a method to reduce the negative effect of high-dimensional data (curse of dimensionality):

> By reducing the original feature space to a subset of relevant features, we often gain better generalizations and computational advantages.

Feature selection can also be used to identify/explain attentive relations:

> Feature weights represent a feature's (relative) importance for the predictive task at hand. Users learn, which features are most important for the outcome they received.

Literature

Bolón-Canedo, Verónica, Noelia Sánchez-Maroño, and Amparo Alonso-Betanzos. "Recent advances and emerging challenges of feature selection in the context of big data." Knowledge-Based Systems 86 (2015): 33-45.

Li, Jundong, et al. "Feature selection: A data perspective." ACM Computing Surveys (CSUR) 50.6 (2017): 1-45.

Guyon, Isabelle, and André Elisseeff. "An introduction to variable and feature selection." Journal of machine learning research 3.Mar (2003): 1157-1182.

Feature Weighting and Selection (FWS)

…narrowing down the problem

Feature selection has been succesfully applied in various offline and online applications.

At DSAR we focus on feature selection for predictive modelling in data streams:

- Data streams are potentially infinite and might be subject to distributional changes (concept drift). Hence, feature importances might shift over time.
- We require flexible and efficient models that can deal with the computational restrictions of online applications while providing high predictive power.

Literature

Ramírez-Gallego, Sergio, et al. "A survey on data preprocessing for data stream mining: Current status and future directions." Neurocomputing 239 (2017): 39-57.

AlNuaimi, Noura, et al. "Streaming feature selection algorithms for big data: A survey." Applied Computing and Informatics (2019).

Feature Weighting and Selection (FWS)

„Boosting decision stumps for dynamic feature selection on data streams" (Barddal et al. 2019)

[…]. In this paper, we propose a novel dynamic feature selection method for data streams called Adaptive Boosting for Feature Selection (ABFS). ABFS chains decision stumps and drift detectors, and as a result, identifies which features are relevant to the learning task as the stream progresses with reasonable success. In addition to our proposed algorithm, we bring feature selection-specific metrics from batch learning to streaming scenarios. Next, we evaluate ABFS according to these metrics in both synthetic and real-world scenarios. As a result, ABFS improves the classification rates of different types of learners and eventually enhances computational resources usage.

Task FWS1

- Explain ABFS and relevant theoretical frameworks
- Assess the human interpretability of feature selections generated by ABFS
- Perform your own experiments and evaluate ABFS with respect to predictive power (e.g. F1, Accuracy), efficiency (time, memory) and stability of the feature set
- Compare feature weights of ABFS with a state-of-the-art feature attribution method (e.g. based on Shapley values)

Literature

Barddal, Jean Paul, et al. "Boosting decision stumps for dynamic feature selection on data streams." Information Systems 83 (2019): 13-29.

Feature Weighting and Selection (FWS)

„Merit-guided Dynamic Feature Selection Filter for Data Streams" (Barddal et al. 2019)

[…]. To select relevant features during the progress of data streams, we propose a merit-guided and classifier-independent dynamic feature selection algorithm named DynamIc SymmetriCal Uncertainty Selection for Streams (DISCUSS). We evaluate our proposal on both synthetic and real-world datasets and show that DISCUSS can boost kNN and Naive Bayes classifiers' accuracy rates on high-dimensional data streams, while at the expense of limited processing time and memory space. Finally, the drawbacks of the proposed method are assessed, and possible future works on the topic are also discussed.

Task FWS2

- Explain DISCUSS and relevant theoretical frameworks
- Assess the human interpretability of feature selections generated by DISCUSS
- Perform your own experiments and evaluate DISCUSS with respect to predictive power (e.g. F1, Accuracy), efficiency (time, memory) and stability of the feature set
- Compare feature weights of DISCUSS with a state-of-the-art feature attribution method (e.g. based on Shapley values)

Literature

Barddal, Jean Paul, et al. "Merit-guided dynamic feature selection filter for data streams." Expert Systems with Applications 116 (2019): 227-242.

Digression: Decision Rules (DR)

"Learning Decision Rules from Data Streams" (Gama et al. 2011)

Decision rules, which can provide good interpretability and flexibility for data mining tasks, have received very little attention in the stream mining community so far. In this work we introduce a new algorithm to learn rule sets, designed for open-ended data streams. The proposed algorithm is able to continuously learn compact ordered and unordered rule sets. The experimental evaluation shows competitive results in comparison with VFDT and C4.5rules.

Task DR1

- Summarize the Decision Rules framework of Gama et al.
- Explain how Decision Rules compare to explanations based on feature weightings/selections
- Name and explain advantages and limitations of decision rules compared with feature attributions for model explanation
- Perform your own experiments and evaluate the quality of the generated decision rules (by means of an appropriate measure) + evaluate the efficiency (time, memory) of the proposed model

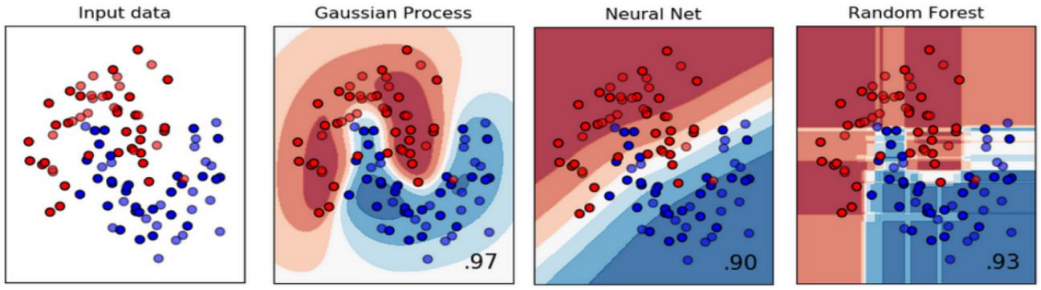
Literature

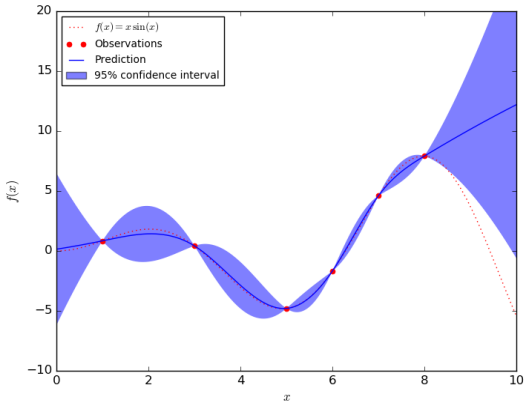
Gama, Joao, and Petr Kosina. "Learning decision rules from data streams." Twenty-Second International Joint Conference on Artificial Intelligence. 2011.

Uncertainty Through Gaussian Processes for Big Data

Gaussian Processes (GPs)

Generic supervised learning method based on Bayes' Theorem designed to solve regression and probabilistic classification problems

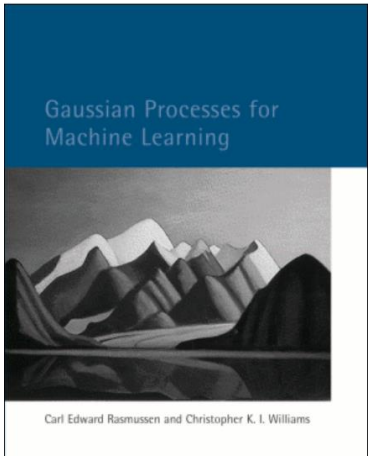




Uncertainty Through Gaussian Processes for Big Data

GP-Advantages

- An attractive way of doing non-parametric Bayesian modelling in a supervised learning problem
- Gaussian processes know what they don't know (uncertainty estimation)
- Accurate and precise
- Interpretable
- Flexible



GP-Disadvantages

- Gaussian processes are computationally expensive
  - especially in training time

Literature

Gaussian Process for Machine Learning http://gaussianprocess.org/gpml/

Scalable Gaussian Process:

Scalable GPs: improving the scalability of full GP while retaining favorable prediction quality for big data.

- The idea is to reduce training and prediction time by some approximation strategies
- The output of the approximation method should be close to the full GP
- Consistency and convergence ?

Related Solutions:

- Distributed Learning
- Variational Inference
- Neural Networks

Literature

Haitao et al., When Gaussian Process Meets Big Data: A Review of Scalable GPs, ArXiv 2018.

Liu et al., Understanding and comparing scalable Gaussian process regression for big data," Knowledge-Based Systems, 2019.

Rivera and Burnaev, "Forecasting of commercial sales with large scale Gaussian processes," arXiv 2017.

Uncertainty Through Gaussian Processes for Big Data (UGP)

Distributed GPs

DGPs follow the divide-and-conquer (D&C) approach to focus on the local subsets of training data. The training data is divided to some partitions (called experts). After training the local experts, it aggregates their predictions .

Task UGP1

- Explain the PoE/BCM approaches and their new improvements
- How do they aggregate the local predictions?
- What is the computational cost of the new scenarios (compare with full GP)?
- Consistency and convergence ?
- Compare the prediction quality of different methods in random and disjoint partitioning

Literature

Liu et. al. “Generalized robust Bayesian committee machine for large-scale Gaussian process regression,” ICML, 2018.

Rulliere et.al. “Nested Kriging predictions for datasets with a large number of observations,” Statistics and Computing, 2018.

Uncertainty Through Gaussian Processes for Big Data (UGP)

Sparse approximations and Variational Inference

The low-rank representation measured between inducing points (a subset of the training data set) and training points, leading to the Nyström approximation.

Task UGP2

- Explain the role of the inducing points in this inference scenario
- What is the new loss function and explain the inference method
- What is the computational cost of the new scenarios (compare with full GP)?
- Model sensitivity to the inducing points selection
- Compare the prediction quality of this scenario with other approximation methods

Literature

Cheng and Boots, "Variational inference for Gaussian process models with linear complexity," NeurIPS, 2017.

Meng et. al. "Regularized Sparse Gaussian Processes", AAAI, 2019.

Bauer et al, Understanding Probabilistic Sparse Gaussian Process Approximations, NeurIPS, 2016.

Uncertainty Through Gaussian Processes for Big Data (UGP)

Neural Networks (NNs)

We know that a single-layer fully-connected neural network with i.i.d. priors over its parameters is equivalent to a Gaussian process (GPs), in the limit of infinite network width. Since the infinite hidden units is not useful in practice, the problem is how to find an efficient connection between infinitely wide NNs and GPs.

Task UGP3

- Explain the connection method between GPs and wide NNs
- Explain how the NN can efficiently compute the covariance function of the GPs
- How to choose the width of the NN?
- What is the inference method?
- Prediction quality and computational cost

Literature

Lee et al., Deep Neural Networks as Gaussian Processes, ICLR, 2018.

Matthews et al. , Gaussian Process Behaviour in Wide Deep Neural Networks, ICLR, 2018

Summary

- **Fairness** projects (how do build non-discriminatory algorithms)

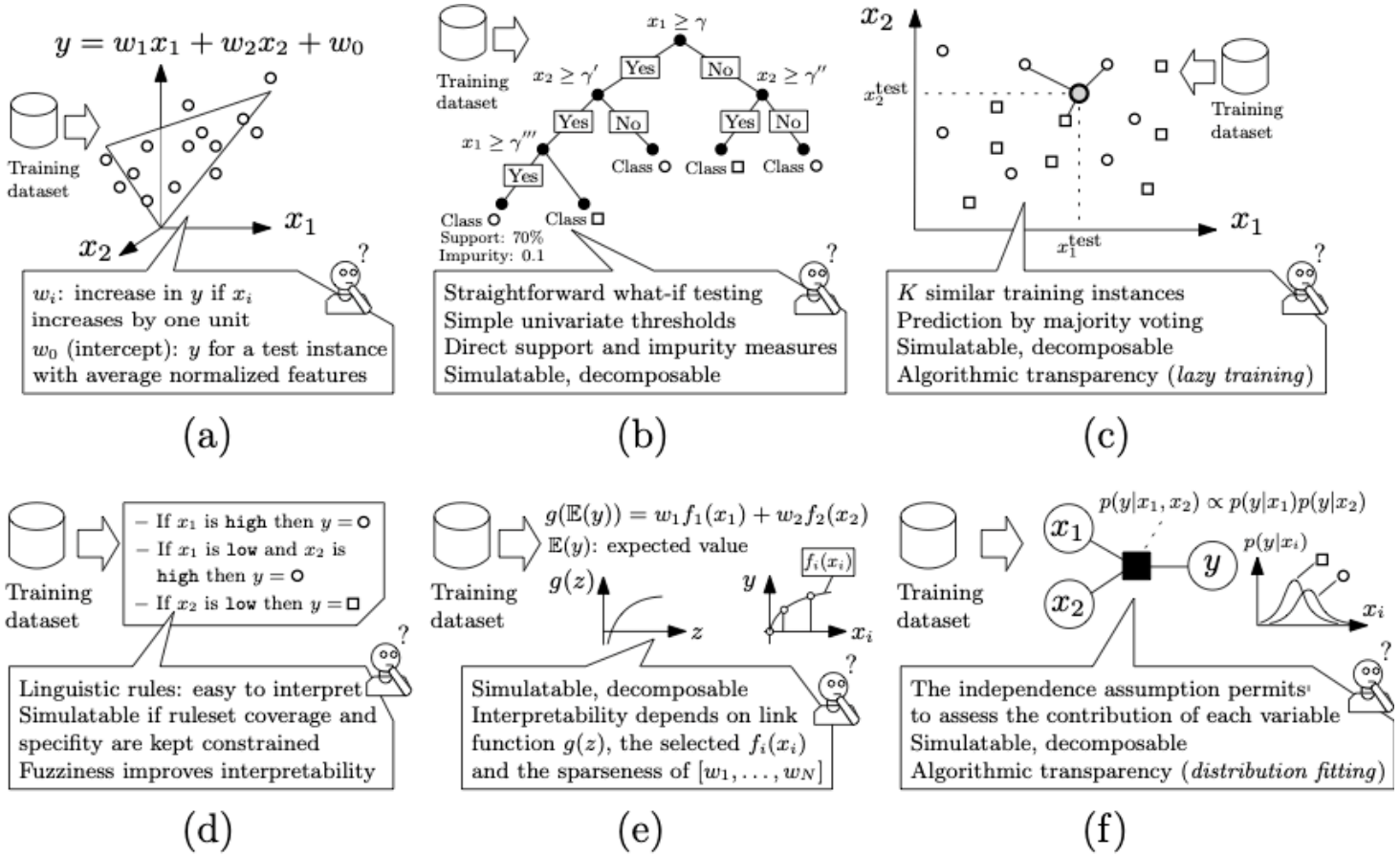
- **Counterfactual** projects under model multiplicity (evaluate whether counterfactual explanations are valid for different models)

- **Feature attribution**
  - Attacking LIME adversarially
  - Defending LIME from Adversaries
  - Understanding LIME
  - Speeding Up Feature Attribution Methods

- **Uncertainty Through Gaussian Processes**

Backup Slides

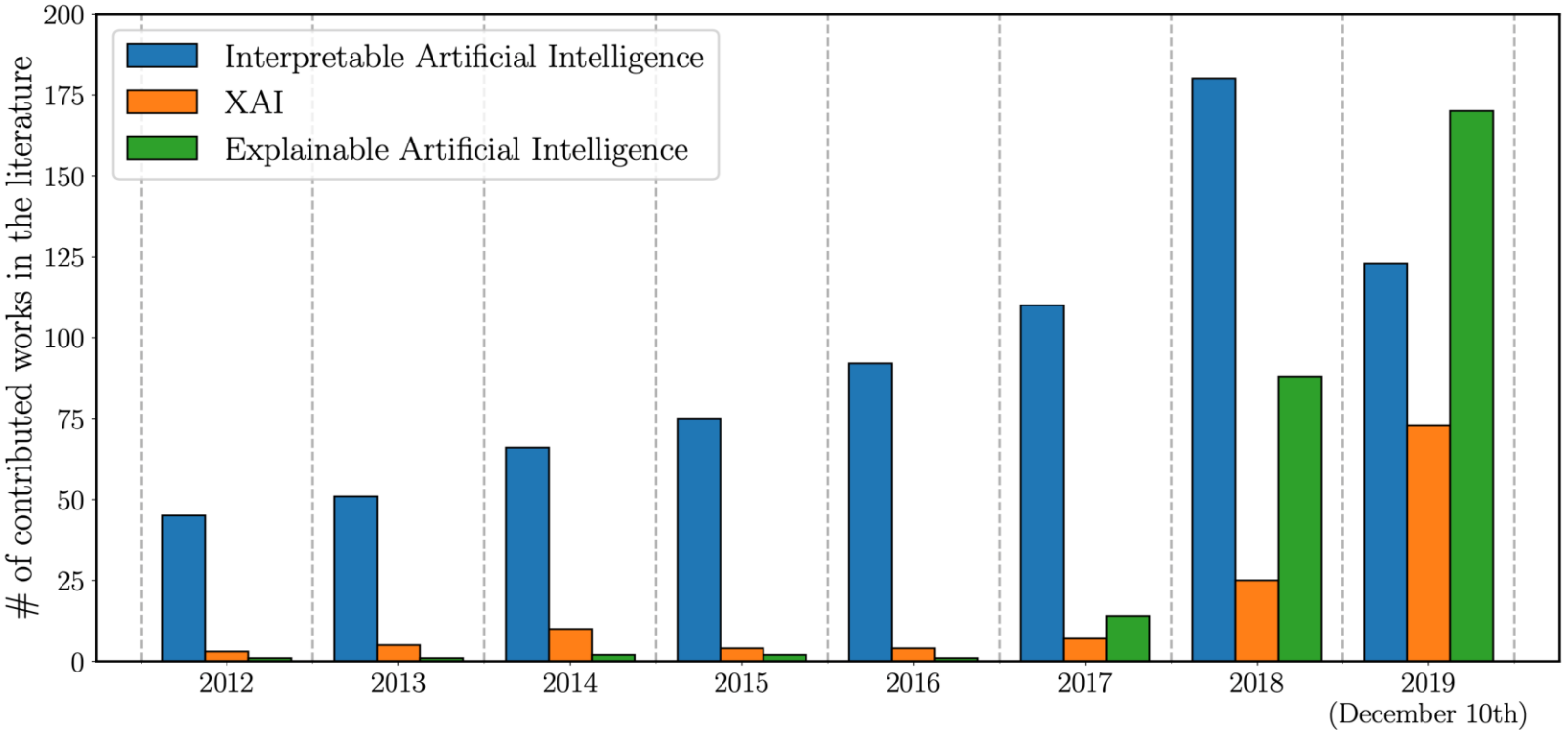
Motivation: Levels of transparency of ML models



https://arxiv.org/pdf/1910.10045.pdf

Motivation

Number of total publications whose title, abstract, and/or keywords refer to the Explainable AI



Data retrieved from https://www.scopus.com/
https://arxiv.org/pdf/1910.10045.pdf