MACHINE LEARNING DEPARTMENT

“Interpretable” Machine Learning

Zachary Lipton & Henry Chai

10701 — December 6th

Waste-basket taxon

What is an interpretation?

$$\|\widehat{w}_f - w^*\| \leq \underbrace{\|w_f - \widehat{w}_f\|}_{\text{finite-sample}} + \underbrace{\|w_f - w^*\|}_{\text{mis-calibration}}$$

$$W = \overbrace{\int_{k<\Lambda} [Dg][DA][D\psi][D\Phi] \exp\left\{ i \int\right.}^{\text{quantum mechanics}} \overbrace{d^4x\sqrt{-g}}^{\text{spacetime}} \Bigg[\overbrace{\frac{m_p^2}{2} R}^{\text{gravity}}$$

$$\left. \underbrace{-\frac{1}{4} F^a_{\mu\nu} F^{a\mu\nu}}_{\text{other forces}} + \underbrace{i\bar{\psi}^i \gamma^\mu D_\mu \psi^i + \left(\bar{\psi}^i_L V_{ij} \Phi \psi^j_R + \text{h.c.}\right)}_{\text{matter}} \underbrace{- |D_\mu \Phi|^2 - V(\Phi)}_{\text{Higgs}} \Bigg] \right\}$$

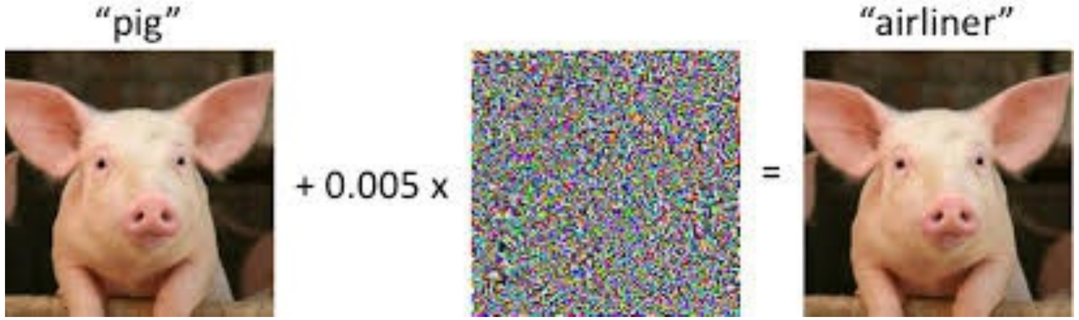
Interpretation in Science

- Interpretation requires grounding in a theory of world modeled
    - Beliefs over entities that world consists of, relations among them, process by which data collected data
- The interpretation of the models tied up in how it squares against postulated, postulated significance of parameters.
- Ingredients of interpretation?
  theory, environment, data collection process, measurement instruments, model, algorithms, analysis, interpreter? ***(not just a model!)***

Desiderata

- Trust

- Out-of-domain performance

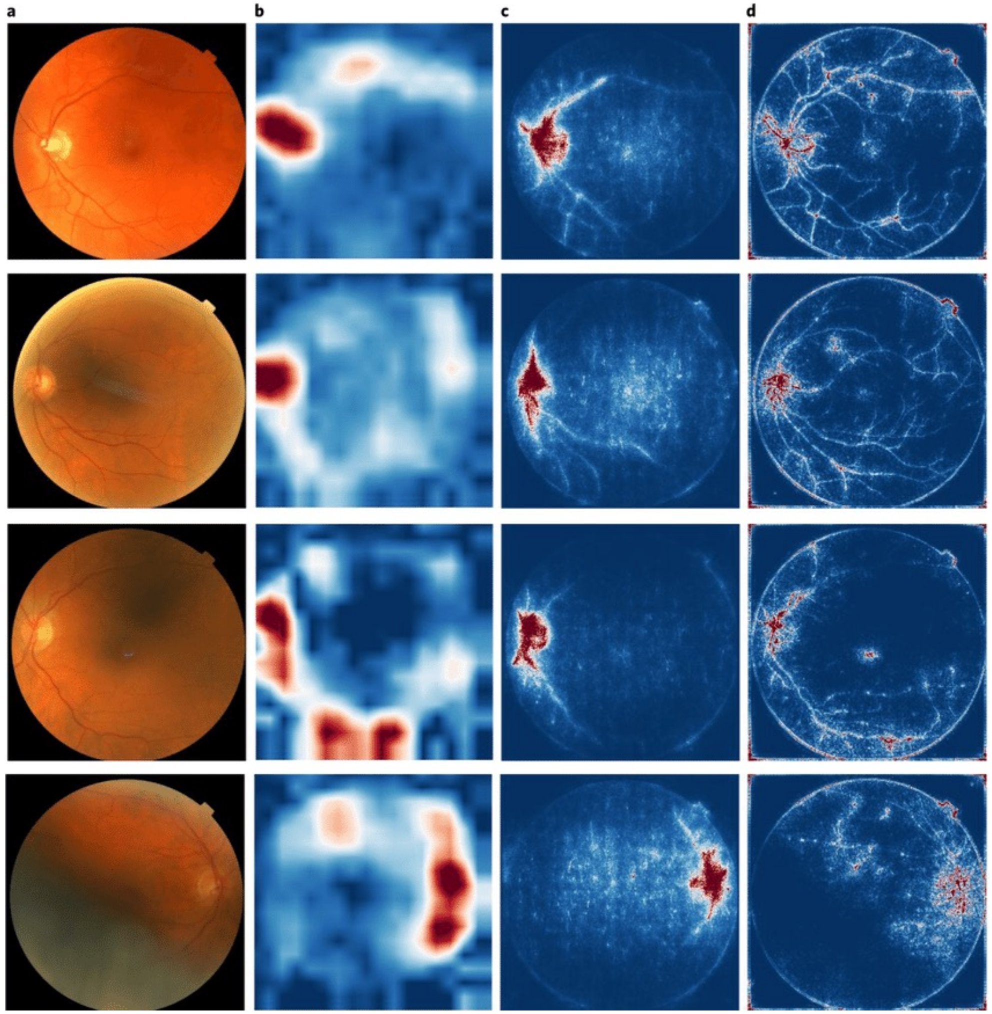


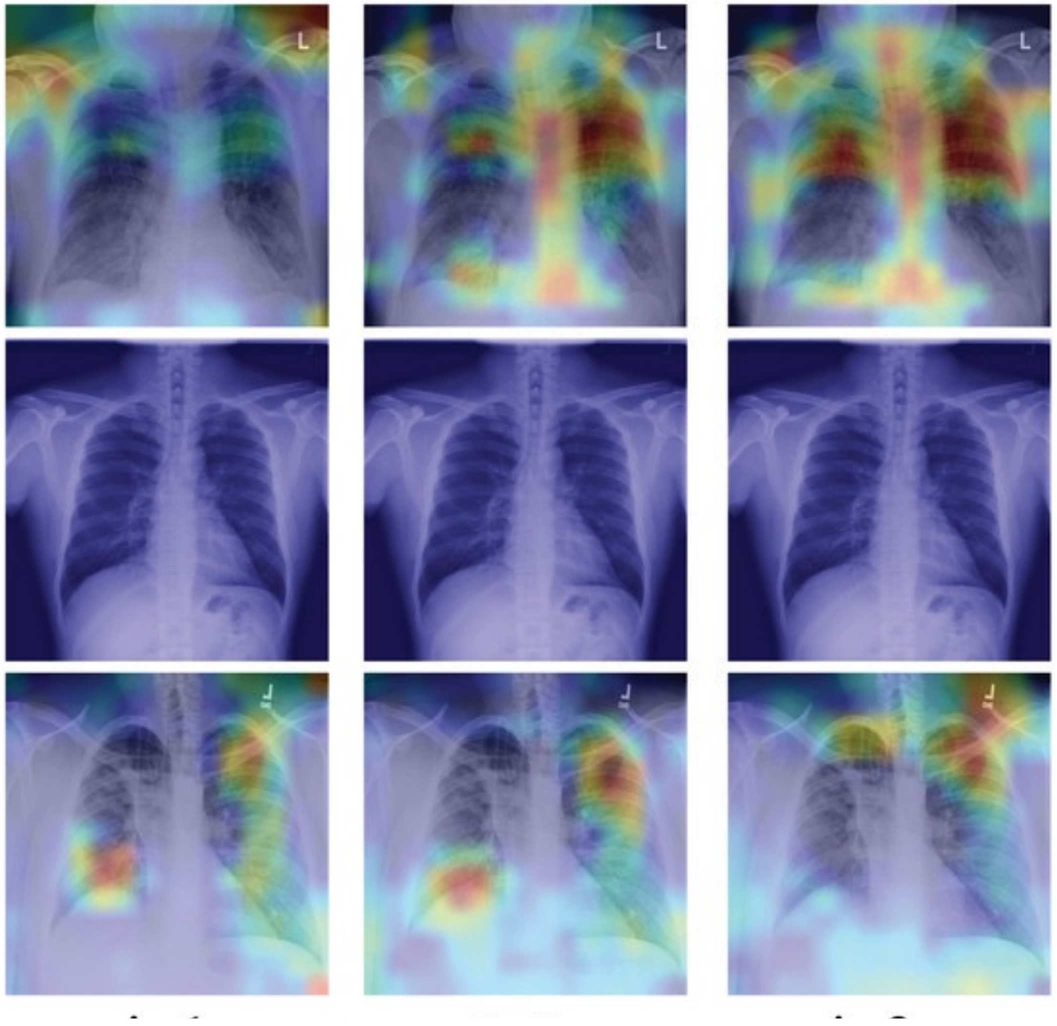
- Causal / semantic explanations

- Insight

- Fairness

What “explanations” are on offer?

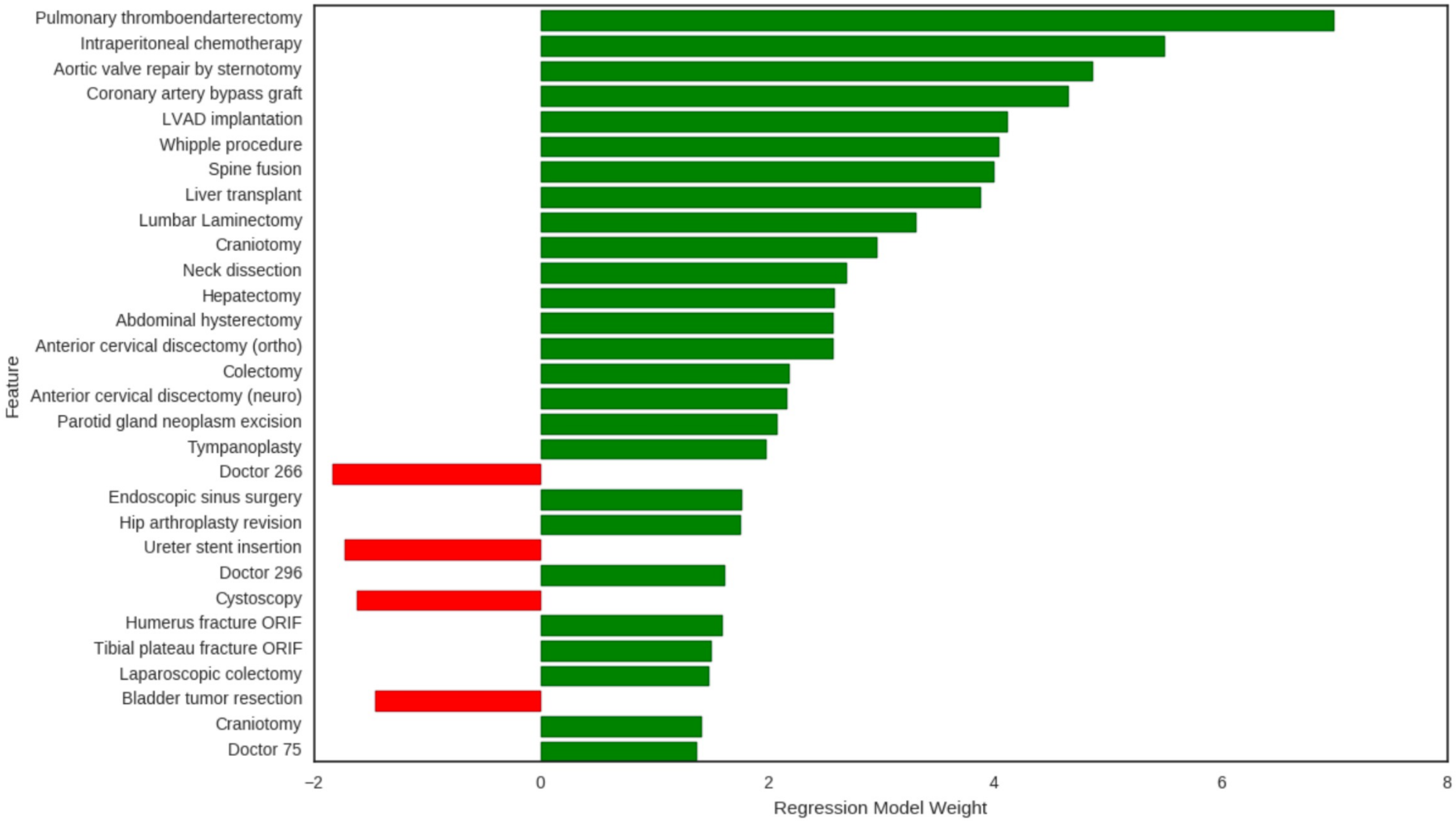




Interpretability is not a condiment

- Getting to an **interpretation** of a models takes work.
- Doesn’t happen by accident.
- Requires commitments about
  1. What is being modeled
  2. What vars ought to be measured
  3. How they ought to be measured
  4. What relations exist among them
  5. What question is the model intended to answer

What’s actually on offer: Feature Attribution



Predicting Surgery Duration with Neural Heteroscedastic Regression—Ng, ..., Z (MLHC 2017)

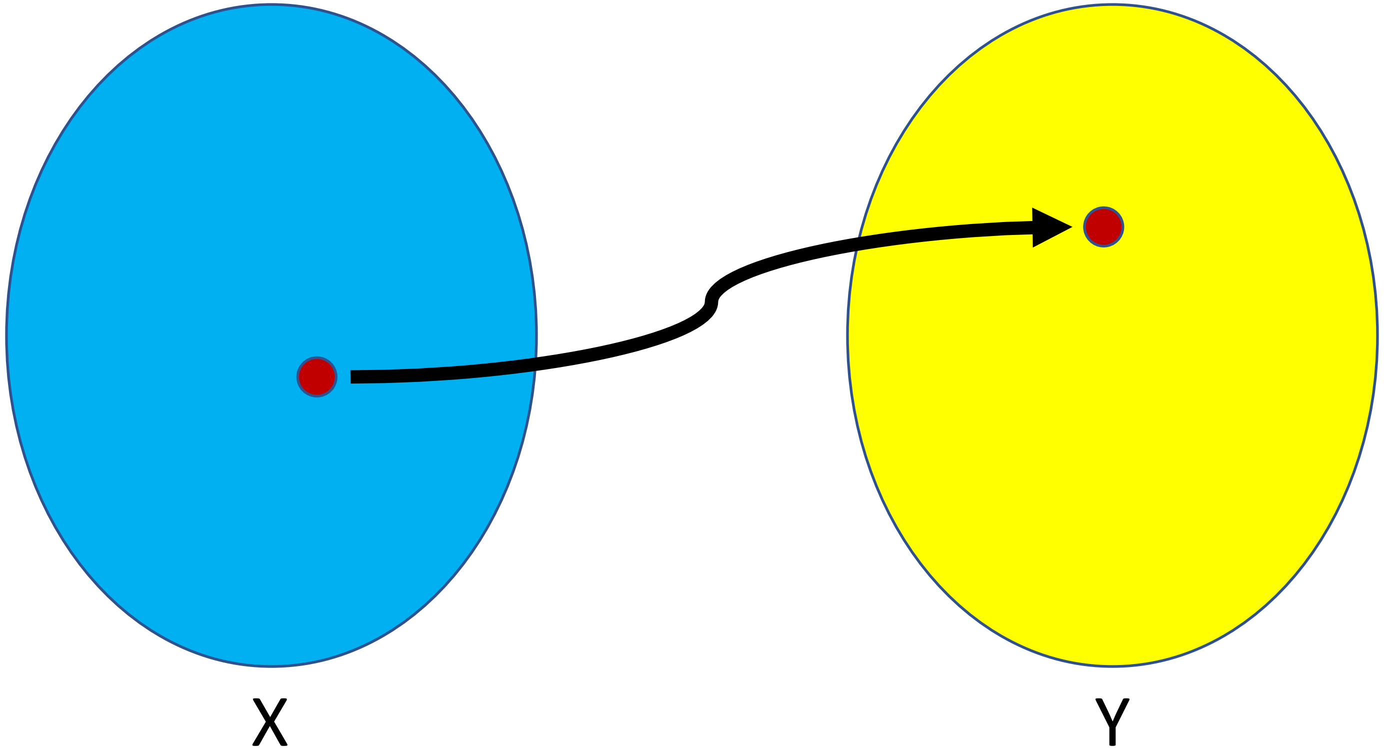
Global Feature Attribution Methods

- Linear model feature weights
- Single-feature ablations
- Permutation feature importance tests
- Global Shapley values

What about nonlinear models, raw data

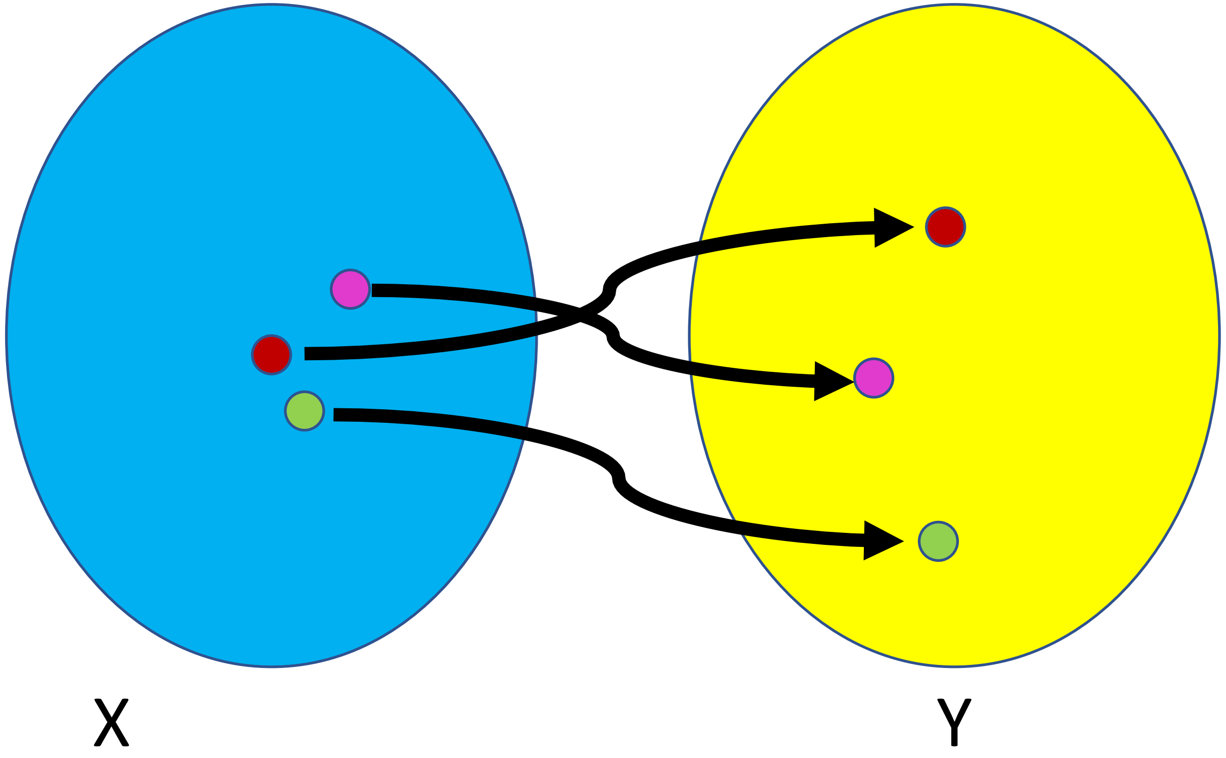
- What do we mean by features?
  - **Particular pixel locations?**
  - A particular pixel locations **in a particular image?**
- Models “look at”/”use” **all pixels for every classification**

A black box ML model is just a mapping



The Trouble with Local Explanations

- All there is to be said about a model at an exact point x is f(x)
- Any additional information about model must say smtg about how f behaves on some **other** inputs x’, x’’
- But which other inputs?
- Summarized how?



LIME

- Local linear approximation to function in vicinity of particular point
- Constructs “simplified representation” x’, e.g. BoW or superpixels
- In simplified representation, learns linear model in vicinity of x’
- Weights other points by local kernel $\pi_{x'}$
- Output depends highly on choice of kernel, choice of points, regularization of local “explanation” model

Why should I trust you?—Ribeiro et al.

SHAP

- Formalizes *additive feature attribution methods*
- Makes connection to Shapley values in cooperative game theory
- Shapley value avg value of a feature among all *coalitions of features*
- Instead of looking at value of feature, SHAP looks at value
- Poses a set of properties for which SHAP is the unique solution
- Relies on *simplified features* x'. Has nice properties when input can be perfectly reconstructed from simplified features $x = h(x')$
- **But where do these “simplified representations” come from?**
- **And what happens when this representation is not reversible?**
- **Relies on reference to baseline “a feature not participating”:**
  WHAT DOES THIS MEAN? (choice is arbitrary and influences the answer)

Integrated gradients

Attempts to formalize properties that a local explanation should have.

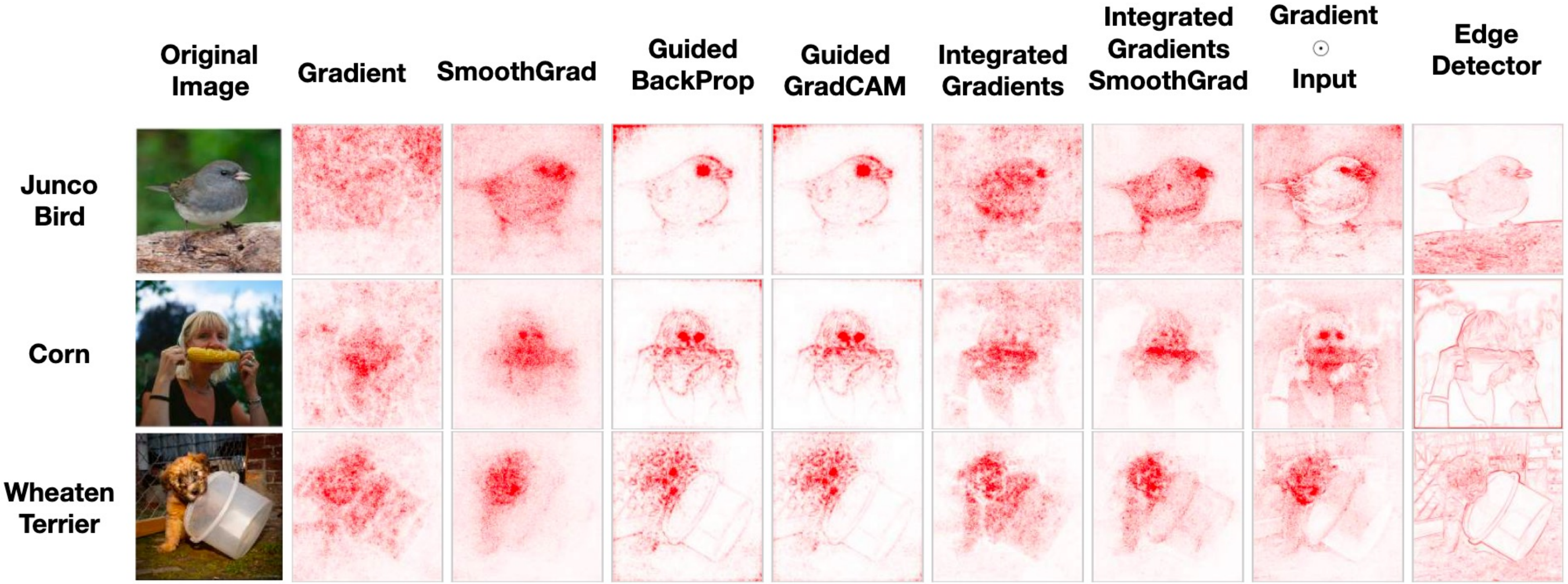
Defines attribution in reference to a ***baseline***.

1. **Sensitivity**—diff in 1 feature, diff output → attribute that feature
2. **Implementation invariance**—eq. fn, diff param → eq. attribution
3. **Completeness**—attributions add up to diff in fn value

Relies on fuzzy properties of choice of baseline:

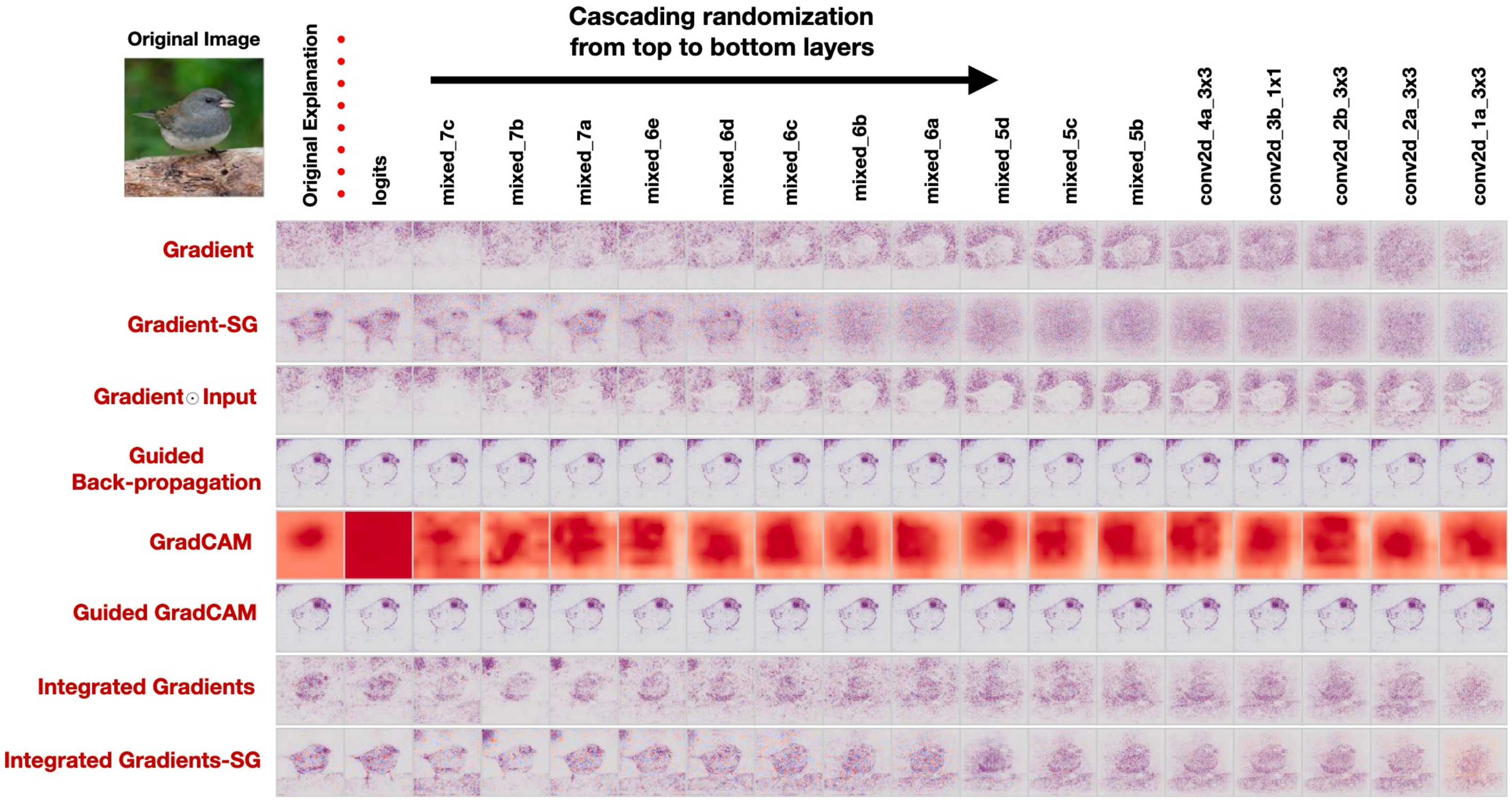
1. **“convey a complete absence of signal” ← what does this mean?**
2. **Different attributions** for black image vs noise image

Sanity Checks for Saliency Maps



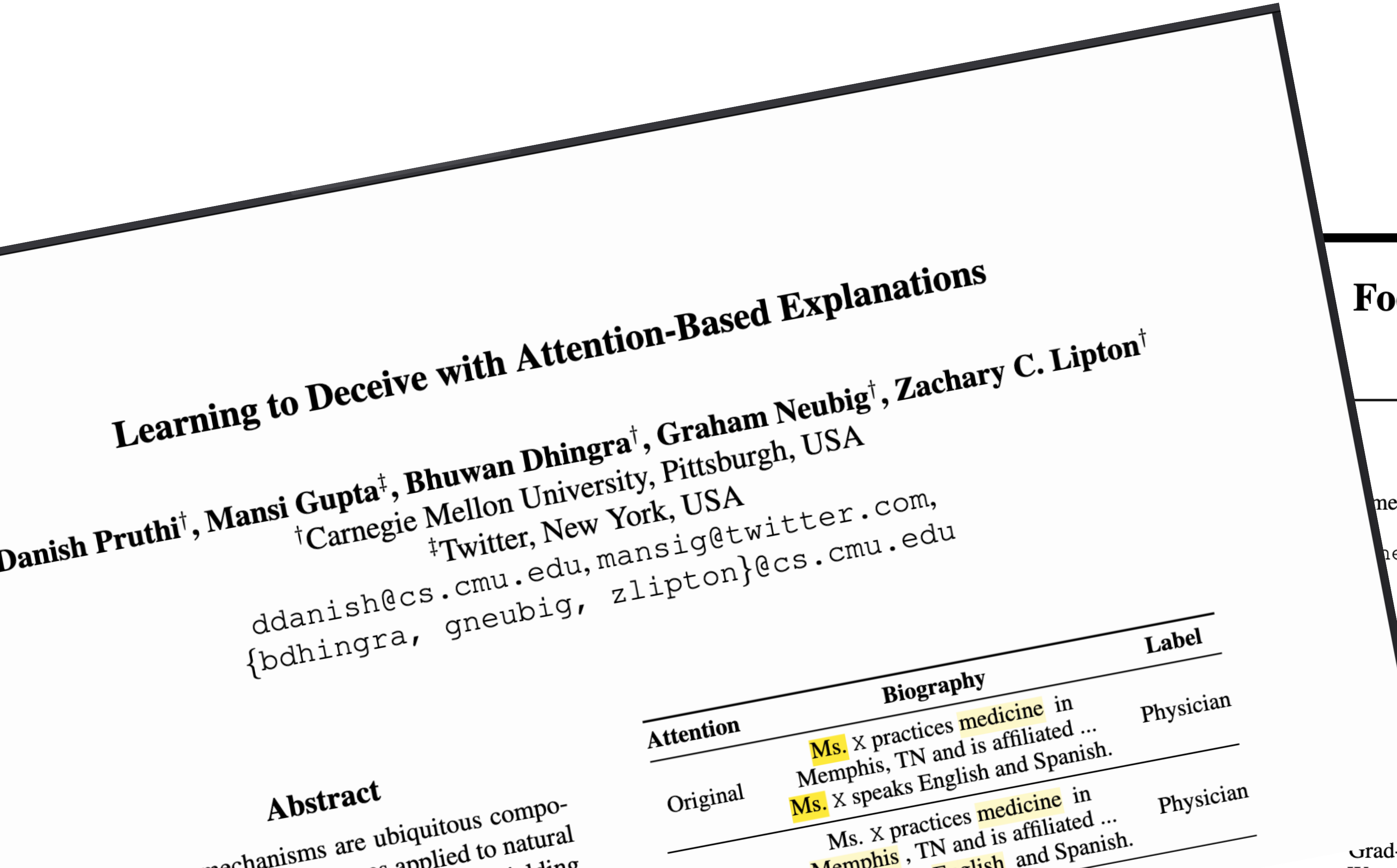
Sanity Checks for Saliency Maps — Adebayo et al.

Explaining the model, the data, or neither?



“Sanity Checks for Saliency Maps” — Adebayo et al.

Most current saliency maps tell you nothing about the model (absent further info)



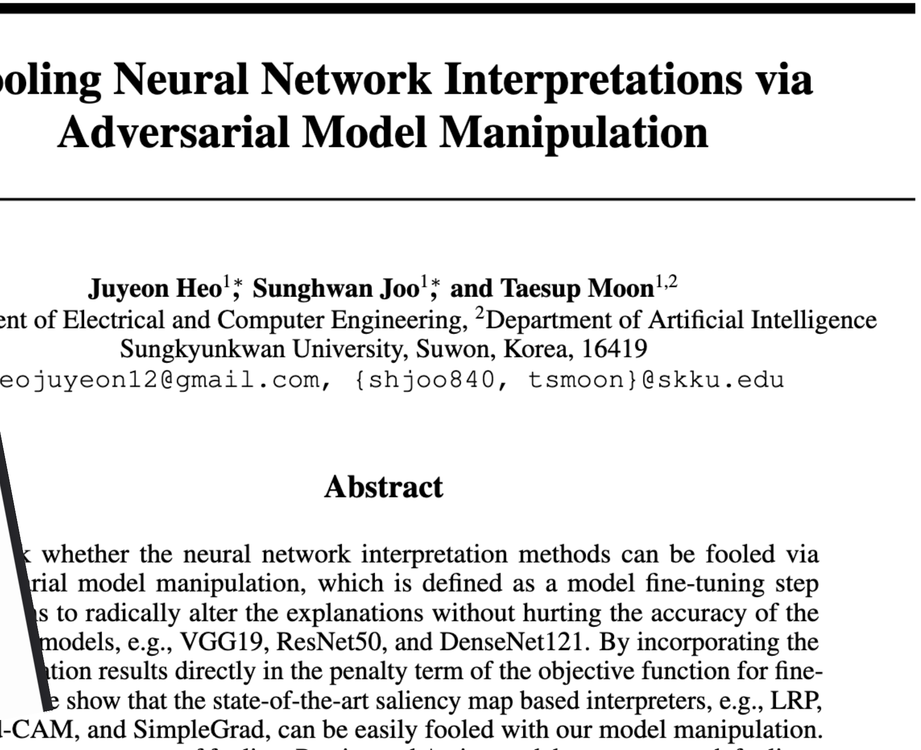
Learning to Deceive with Attention-Based Explanations

Danish Pruthi[†], Mansi Gupta[‡], Bhuwan Dhingra[†], Graham Neubig[†], Zachary C. Lipton[†]
[†]Carnegie Mellon University, Pittsburgh, USA
[‡]Twitter, New York, USA
ddanish@cs.cmu.edu, mansig@twitter.com,
{bdhingra, gneubig, zlipton}@cs.cmu.edu

Attention	Biography	Label
Original	Ms. X practices medicine in Memphis, TN and is affiliated ... Ms. X speaks English and Spanish.	Physician
	Ms. X practices medicine in Memphis , TN and is affiliated ... English and Spanish.	Physician

Abstract

mechanisms are ubiquitous compo-
s applied to natural
lding



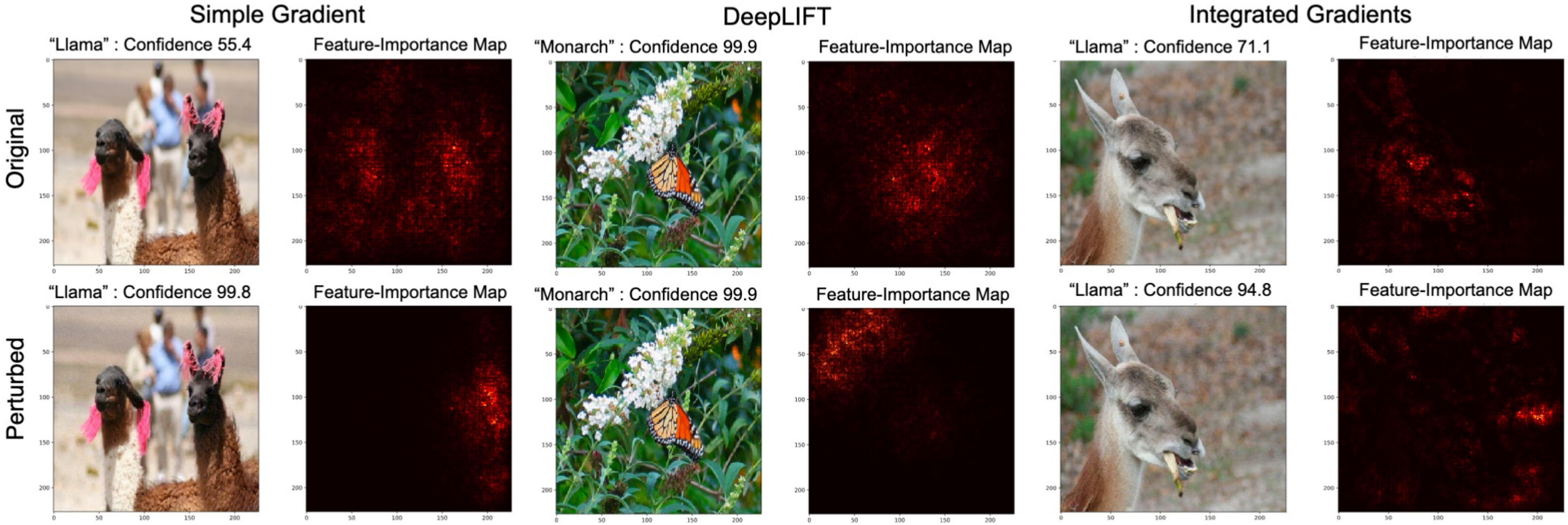
Fooling Neural Network Interpretations via Adversarial Model Manipulation

Juyeon Heo[1*], Sunghwan Joo[1*], and Taesup Moon[1,2]
ment of Electrical and Computer Engineering, [2]Department of Artificial Intelligence
Sungkyunkwan University, Suwon, Korea, 16419
eojuyeon12@gmail.com, {shjoo840, tsmoon}@skku.edu

Abstract

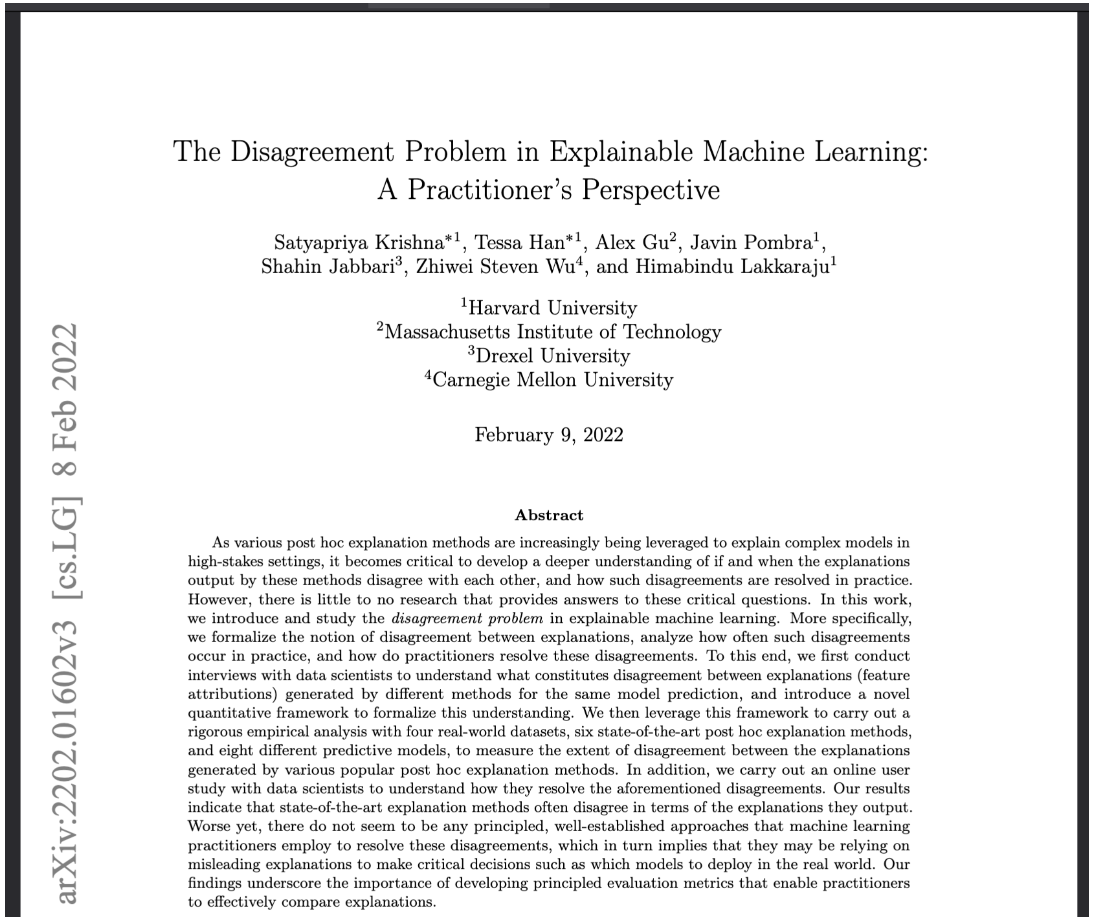
k whether the neural network interpretation methods can be fooled via rial model manipulation, which is defined as a model fine-tuning step s to radically alter the explanations without hurting the accuracy of the models, e.g., VGG19, ResNet50, and DenseNet121. By incorporating the tion results directly in the penalty term of the objective function for fine-e show that the state-of-the-art saliency map based interpreters, e.g., LRP, Grad-CAM, and SimpleGrad, can be easily fooled with our model manipulation.

Saliency approaches brittle to manipulation



Interpretation of Neural Networks is Fragile—Ghorbani et al.

Saliency Methods Disagree with Each Other



arXiv:2202.01602v3 [cs.LG] 8 Feb 2022

The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective

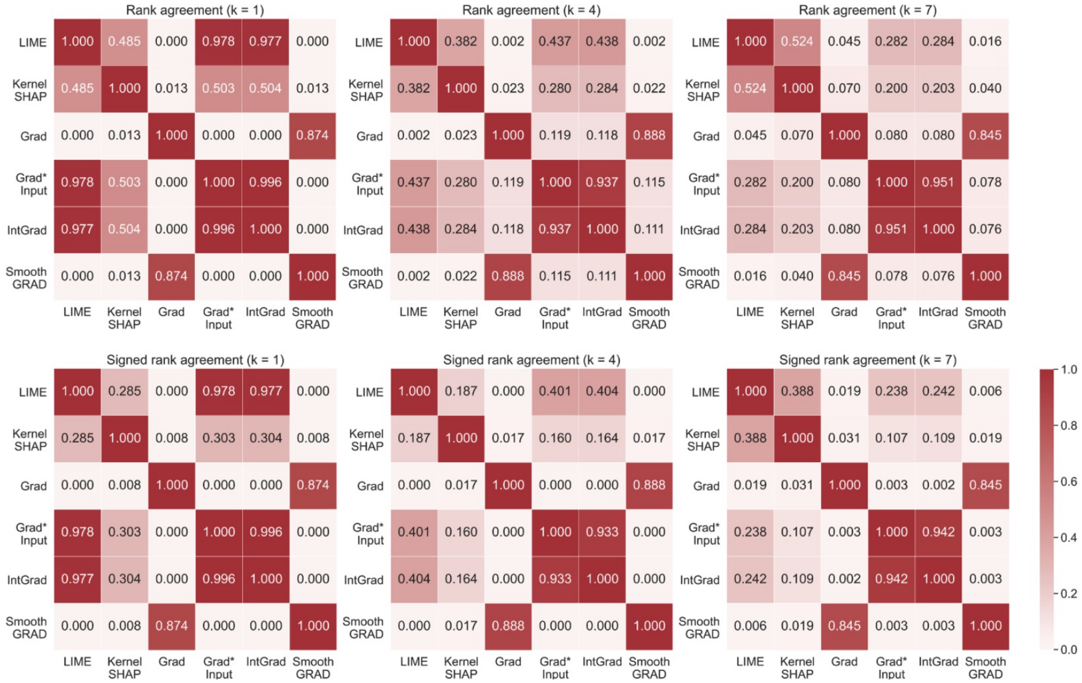
Satyapriya Krishna*[1], Tessa Han*[1], Alex Gu[2], Javin Pombra[1],
Shahin Jabbari[3], Zhiwei Steven Wu[4], and Himabindu Lakkaraju[1]

[1]Harvard University
[2]Massachusetts Institute of Technology
[3]Drexel University
[4]Carnegie Mellon University

February 9, 2022

Abstract

As various post hoc explanation methods are increasingly being leveraged to explain complex models in high-stakes settings, it becomes critical to develop a deeper understanding of if and when the explanations output by these methods disagree with each other, and how such disagreements are resolved in practice. However, there is little to no research that provides answers to these critical questions. In this work, we introduce and study the *disagreement problem* in explainable machine learning. More specifically, we formalize the notion of disagreement between explanations, analyze how often such disagreements occur in practice, and how do practitioners resolve these disagreements. To this end, we first conduct interviews with data scientists to understand what constitutes disagreement between explanations (feature attributions) generated by different methods for the same model prediction, and introduce a novel quantitative framework to formalize this understanding. We then leverage this framework to carry out a rigorous empirical analysis with four real-world datasets, six state-of-the-art post hoc explanation methods, and eight different predictive models, to measure the extent of disagreement between the explanations generated by various popular post hoc explanation methods. In addition, we carry out an online user study with data scientists to understand how they resolve the aforementioned disagreements. Our results indicate that state-of-the-art explanation methods often disagree in terms of the explanations they output. Worse yet, there do not seem to be any principled, well-established approaches that machine learning practitioners employ to resolve these disagreements, which in turn implies that they may be relying on misleading explanations to make critical decisions such as which models to deploy in the real world. Our findings underscore the importance of developing principled evaluation metrics that enable practitioners to effectively compare explanations.



Rank agreement (k = 1)

	LIME	Kernel SHAP	Grad	Grad* Input	IntGrad	Smooth GRAD
LIME	1.000	0.485	0.000	0.978	0.977	0.000
Kernel SHAP	0.485	1.000	0.013	0.503	0.504	0.013
Grad	0.000	0.013	1.000	0.000	0.000	0.874
Grad* Input	0.978	0.503	0.000	1.000	0.996	0.000
IntGrad	0.977	0.504	0.000	0.996	1.000	0.000
Smooth GRAD	0.000	0.013	0.874	0.000	0.000	1.000

Rank agreement (k = 4)

	LIME	Kernel SHAP	Grad	Grad* Input	IntGrad	Smooth GRAD
LIME	1.000	0.382	0.002	0.437	0.438	0.002
Kernel SHAP	0.382	1.000	0.023	0.280	0.284	0.022
Grad	0.002	0.023	1.000	0.119	0.118	0.888
Grad* Input	0.437	0.280	0.119	1.000	0.937	0.115
IntGrad	0.438	0.284	0.118	0.937	1.000	0.111
Smooth GRAD	0.002	0.022	0.888	0.115	0.111	1.000

Rank agreement (k = 7)

	LIME	Kernel SHAP	Grad	Grad* Input	IntGrad	Smooth GRAD
LIME	1.000	0.524	0.045	0.282	0.284	0.016
Kernel SHAP	0.524	1.000	0.070	0.200	0.203	0.040
Grad	0.045	0.070	1.000	0.080	0.080	0.845
Grad* Input	0.282	0.200	0.080	1.000	0.951	0.078
IntGrad	0.284	0.203	0.080	0.951	1.000	0.076
Smooth GRAD	0.016	0.040	0.845	0.078	0.076	1.000

Signed rank agreement (k = 1)

	LIME	Kernel SHAP	Grad	Grad* Input	IntGrad	Smooth GRAD
LIME	1.000	0.285	0.000	0.978	0.977	0.000
Kernel SHAP	0.285	1.000	0.008	0.303	0.304	0.008
Grad	0.000	0.008	1.000	0.000	0.000	0.874
Grad* Input	0.978	0.303	0.000	1.000	0.996	0.000
IntGrad	0.977	0.304	0.000	0.996	1.000	0.000
Smooth GRAD	0.000	0.008	0.874	0.000	0.000	1.000

Signed rank agreement (k = 4)

	LIME	Kernel SHAP	Grad	Grad* Input	IntGrad	Smooth GRAD
LIME	1.000	0.187	0.000	0.401	0.404	0.000
Kernel SHAP	0.187	1.000	0.017	0.160	0.164	0.017
Grad	0.000	0.017	1.000	0.000	0.000	0.888
Grad* Input	0.401	0.160	0.000	1.000	0.933	0.000
IntGrad	0.404	0.164	0.000	0.933	1.000	0.000
Smooth GRAD	0.000	0.017	0.888	0.000	0.000	1.000

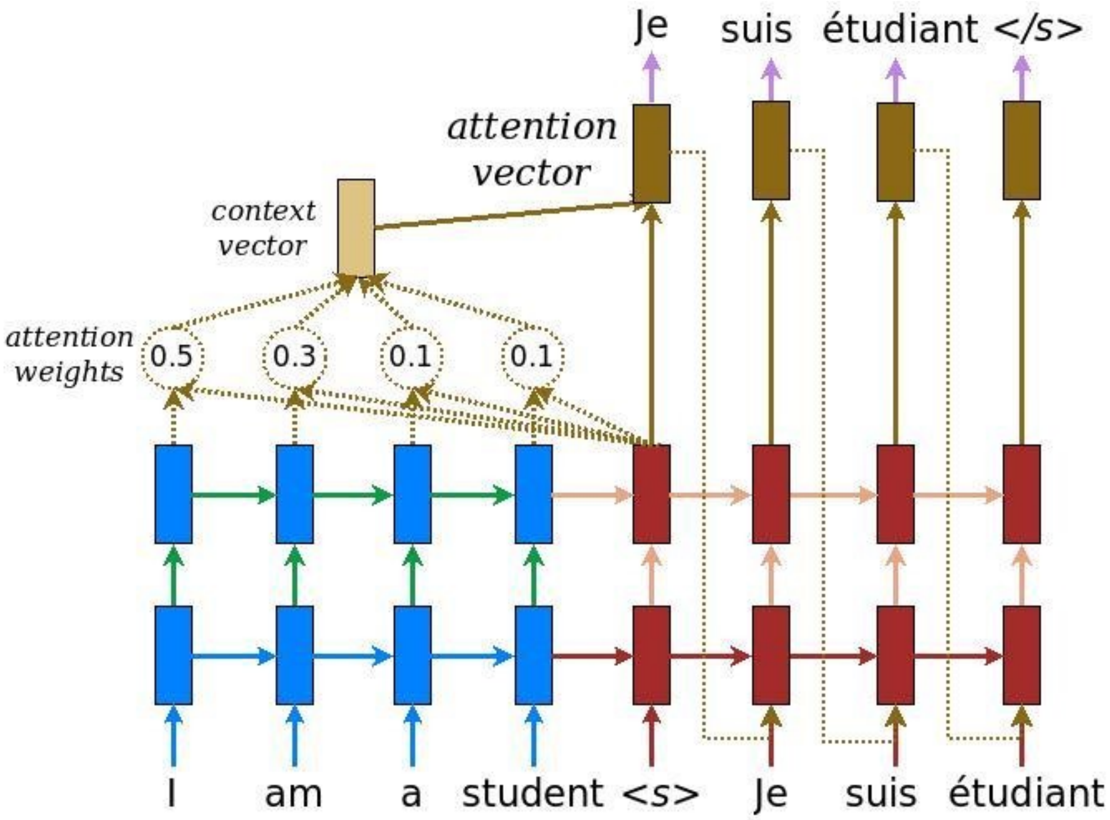
Signed rank agreement (k = 7)

	LIME	Kernel SHAP	Grad	Grad* Input	IntGrad	Smooth GRAD
LIME	1.000	0.388	0.019	0.238	0.242	0.006
Kernel SHAP	0.388	1.000	0.031	0.107	0.109	0.019
Grad	0.019	0.031	1.000	0.003	0.002	0.845
Grad* Input	0.238	0.107	0.003	1.000	0.942	0.003
IntGrad	0.242	0.109	0.002	0.942	1.000	0.003
Smooth GRAD	0.006	0.019	0.845	0.003	0.003	1.000

Figure 2: Disagreement between explanation methods for neural network model trained on COMPAS dataset measured by rank agreement (top row) and signed rank agreement (bottom row) at top-k features for increasing values of k. Each cell in the heatmap shows the metric value averaged over test set data points for each pair of explanation methods, with lighter colors indicating stronger disagreement. Across all six heatmaps, the standard error ranges between 0 and 0.003.

Purported Explanatory Powers of Attention

- Claimed to show what model "focuses on" while decoding
- Proposed for seq2seq tasks but adapted to classification
- Used in industry, some FAT* papers for claimed interpretability

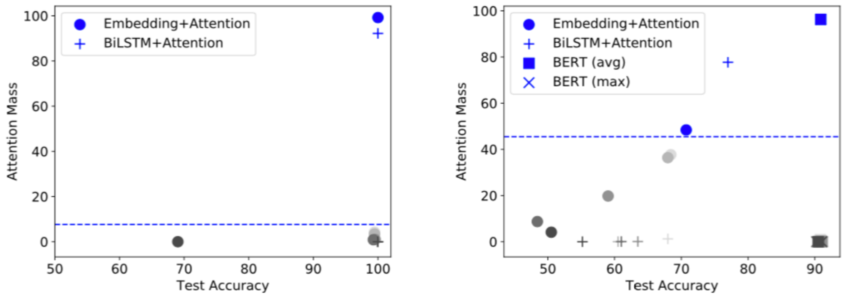


Learning to Deceive with Attention

- Designate a set of *impermissible* tokens
- Can learn network trained to assign low attention to these tokens
- Works even when the model provably continues to rely on those tokens
- If attention allows manipulability, what's special about the original weights?

$$\mathbf{m}_i = \begin{cases} 1, & \text{if } w_i \in \mathcal{I} \\ 0 & \text{otherwise.} \end{cases}$$

$$\mathcal{R} = -\lambda \log(1 - \boldsymbol{\alpha}^T \mathbf{m})$$



(a) Gender Identification

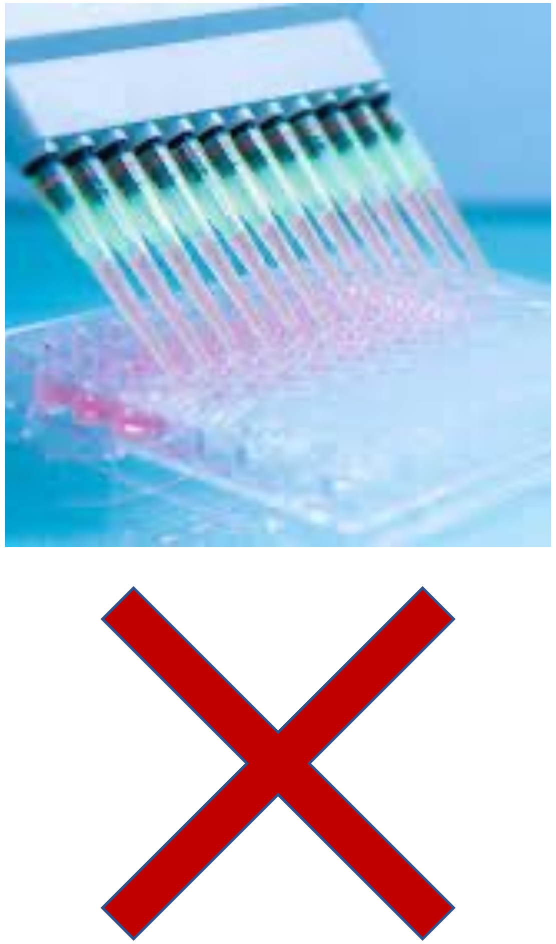
(b) Sentiment Analysis (SST + Wiki)

https://arxiv.org/abs/1909.07913

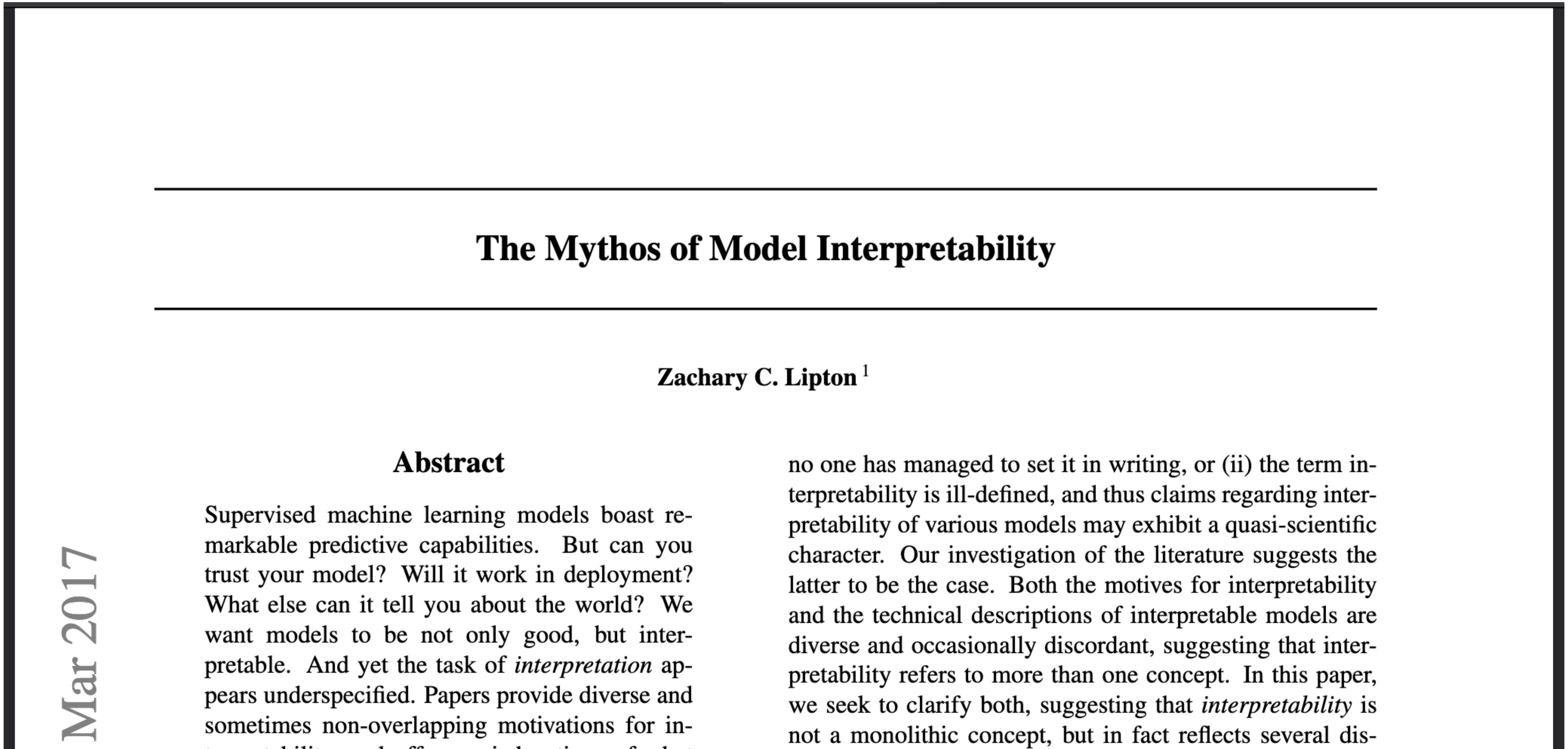
Systematic problems with the entire enterprise of saliency maps

- Focus on some commonsense properties they should have, but no coherent explanation of **what problem they solve**.
- Mirrors the axiomatic approach to equity (Young '95)
- Confirmation bias: "we found a much stronger agreement between human explanations and SHAP than with other method" (—SHAP)
- Heavy reliance on unstated properties of the model & data *(smoothness? inductive bias of SGD + architecture?)*
- Even if we knew true labeling function, would we want saliency?!
- **All involve some choice of "counterfactual" but provide no guidance or coherent argument for what constitutes the <u>relevant counterfactual</u>**

If Interpretable ML Were a Drug



Why do these problems persist once known?



The Mythos of Model Interpretability

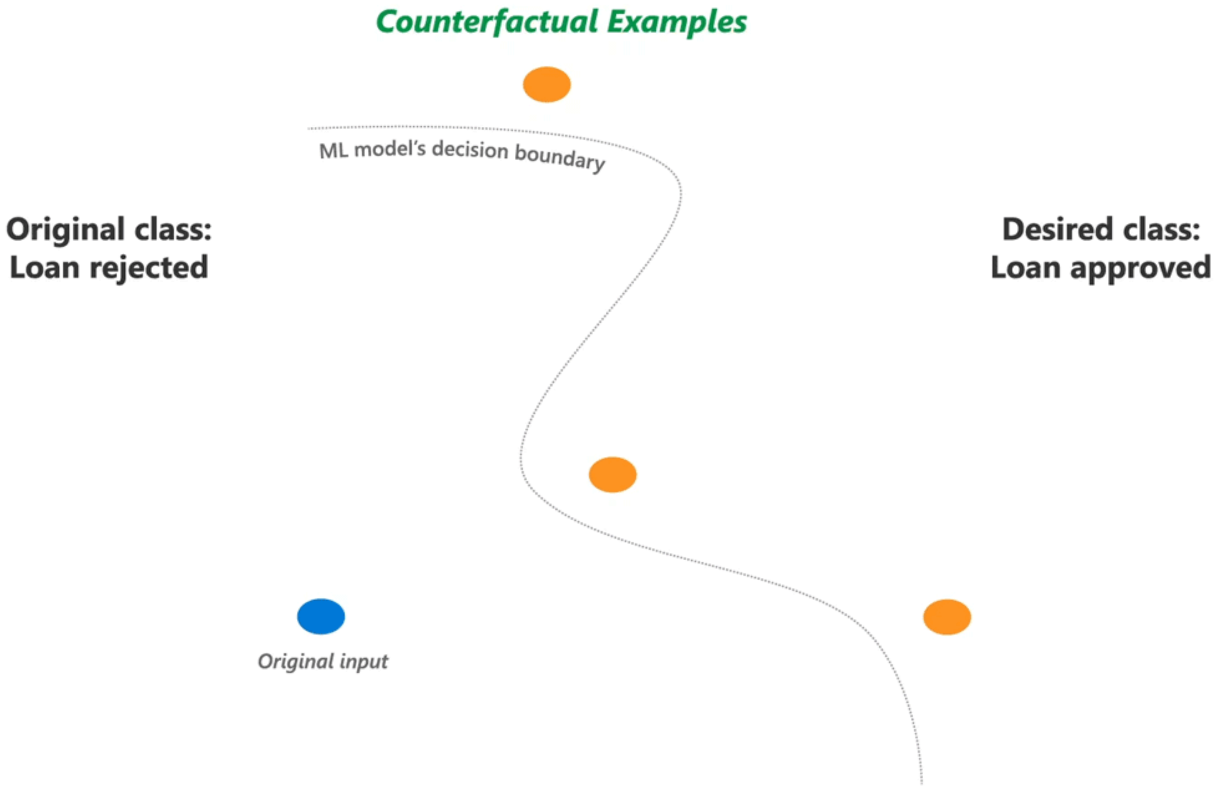
Zachary C. Lipton [1]

Abstract

Supervised machine learning models boast remarkable predictive capabilities. But can you trust your model? Will it work in deployment? What else can it tell you about the world? We want models to be not only good, but interpretable. And yet the task of *interpretation* appears underspecified. Papers provide diverse and sometimes non-overlapping motivations for in-

no one has managed to set it in writing, or (ii) the term interpretability is ill-defined, and thus claims regarding interpretability of various models may exhibit a quasi-scientific character. Our investigation of the literature suggests the latter to be the case. Both the motives for interpretability and the technical descriptions of interpretable models are diverse and occasionally discordant, suggesting that interpretability refers to more than one concept. In this paper, we seek to clarify both, suggesting that *interpretability* is not a monolithic concept, but in fact reflects several dis-

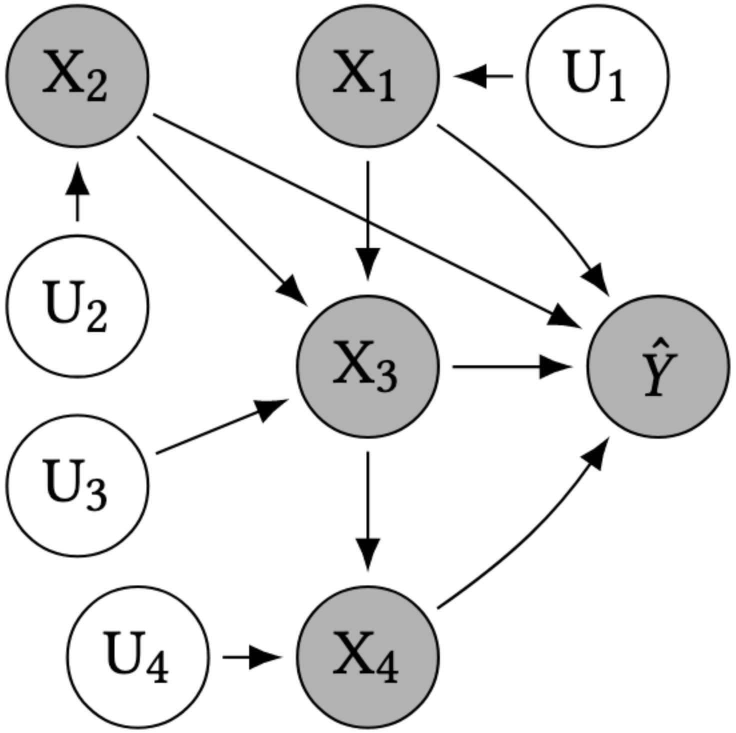
Counterfactual Explanations (& caveats)



(image source)

Hidden Assumptions (Barocas et al.) https://arxiv.org/abs/1912.04930

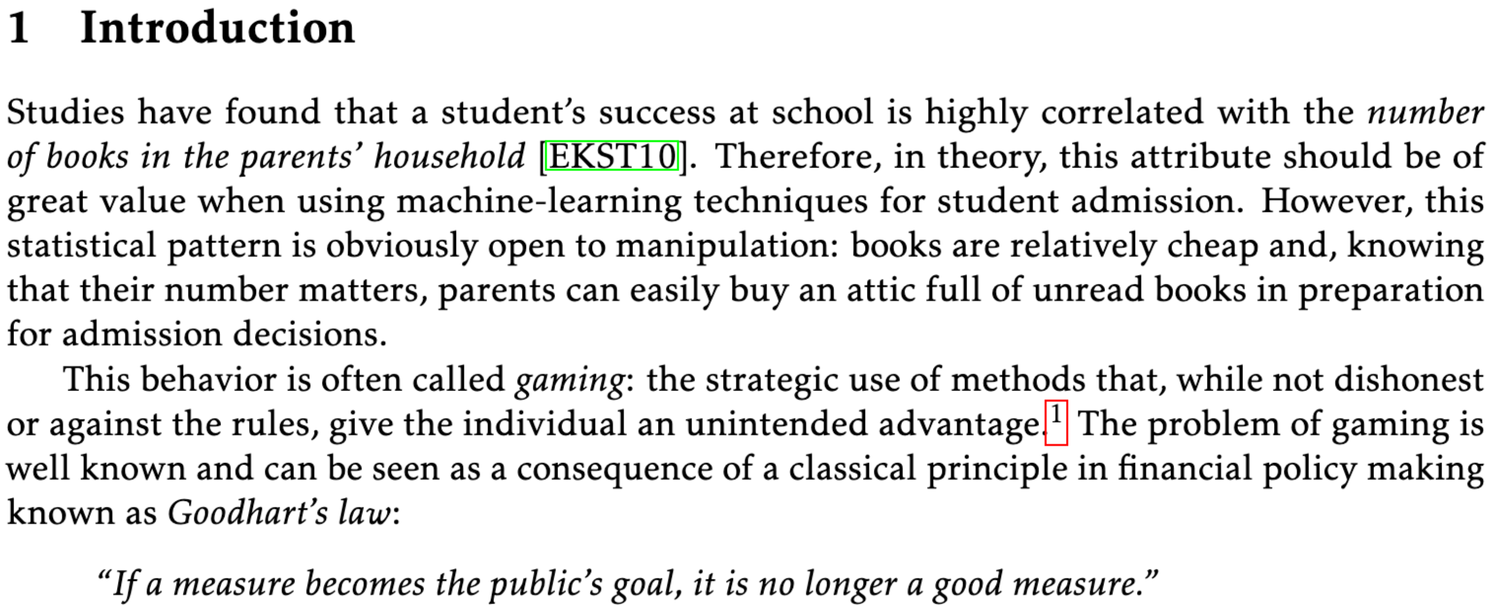
Causal Formulations of Recourse



Algorithmic Recourse: from Counterfactual Explanations to Interventions (Karimi, Scholkopf, Valera)

Strategic Classification

- Conceives of strategic responses by decision subjects as "gaming"
- Tantamount to targeted adversarial attacks



1 Introduction

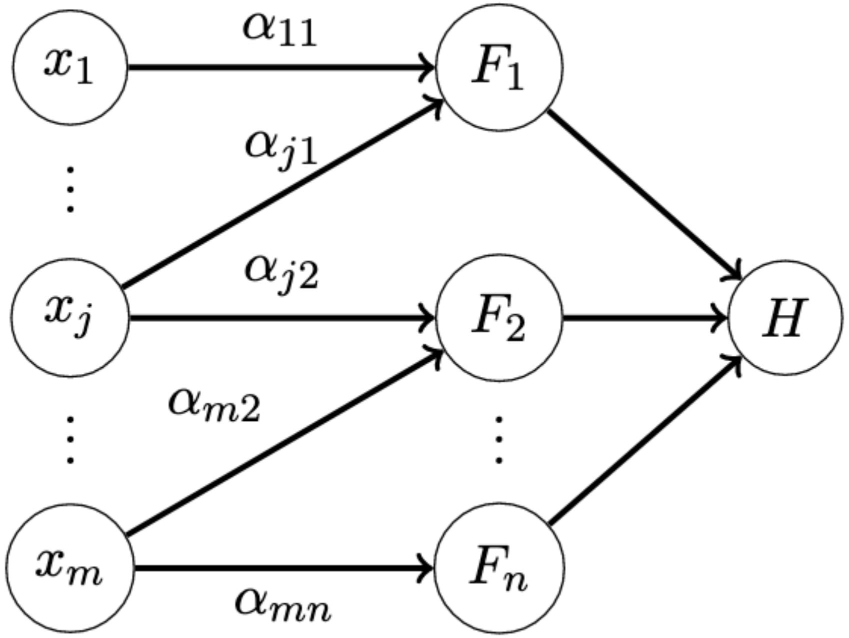
Studies have found that a student's success at school is highly correlated with the *number of books in the parents' household* [EKST10]. Therefore, in theory, this attribute should be of great value when using machine-learning techniques for student admission. However, this statistical pattern is obviously open to manipulation: books are relatively cheap and, knowing that their number matters, parents can easily buy an attic full of unread books in preparation for admission decisions.

This behavior is often called *gaming*: the strategic use of methods that, while not dishonest or against the rules, give the individual an unintended advantage.[1] The problem of gaming is well known and can be seen as a consequence of a classical principle in financial policy making known as *Goodhart's law*:

> *"If a measure becomes the public's goal, it is no longer a good measure."*

[Strategic Classification --- Hardt et al. (ITCS 2016)]

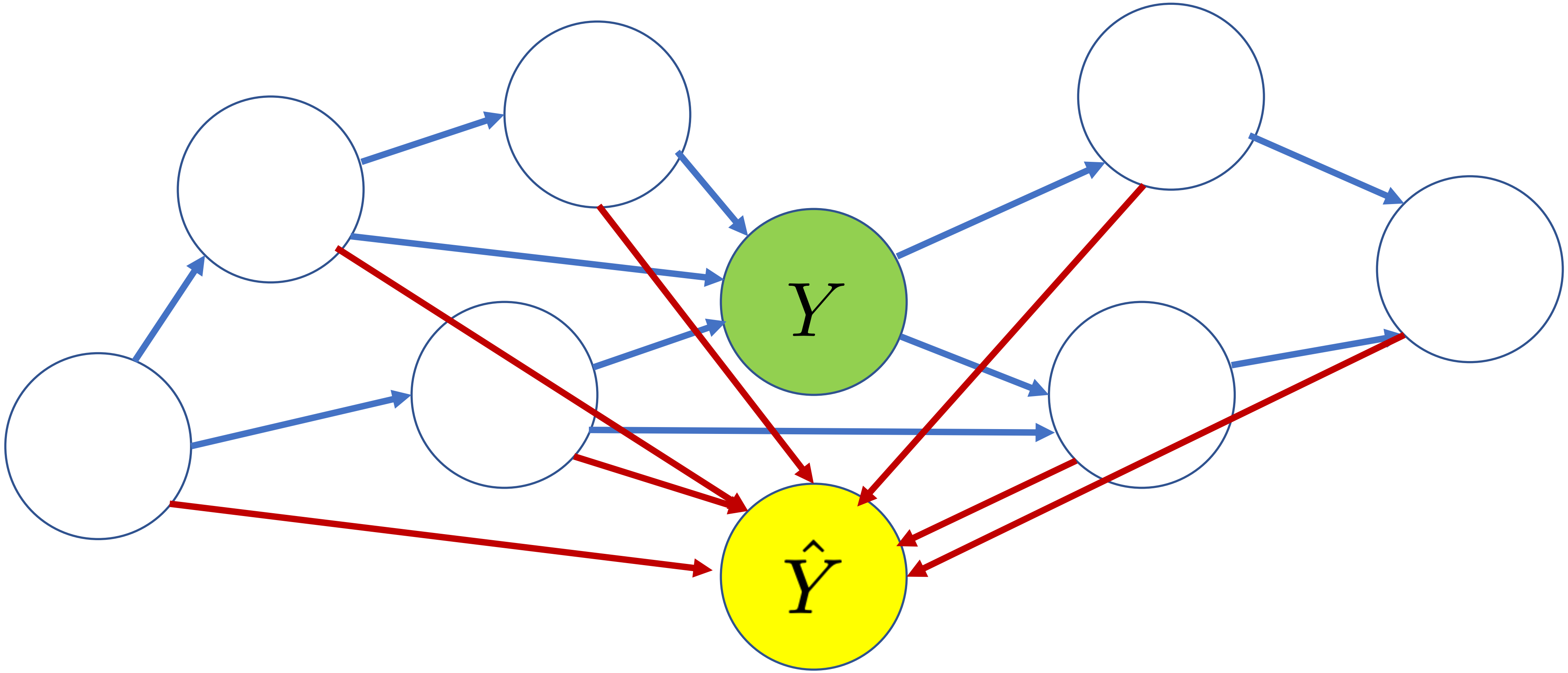
Causal Strategic Classification



“How do Classifiers Induce Agents to Invest Effort Strategically?” (Raghavan, Kleinberg EC 2020)
https://dl.acm.org/doi/abs/10.1145/3417742

“Causal Strategic Linear Regression” (Shavit, Edelman, Axelrod ICML 2020)
https://arxiv.org/abs/2002.10066

General Causal Strategic Prediction



Discovering Optimal Scoring Mechanisms for Causal Strategic Prediction
Yan, Gupta, ZL (in preparation)

Thanks!!

- **Contact**
  email: zlipton@cmu.edu
  twitter: @zacharylipton
  lab: http://acmilab.org
- **Papers**
  - The Mythos of Model Interpretability (CACM), 2016
    https://arxiv.org/abs/1606.03490
  - Learning to Deceive with Attention-Based Explanations (ACL), 2019
    https://arxiv.org/abs/1909.07913
  - Learning the Difference that Makes a Difference w Counterfactually Augmented Data (ICLR), 2020
    https://arxiv.org/abs/1909.12434
  - Evaluating Explanations: How much do explanations from the teacher aid students? (TACL) 2021
    https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00465/110436/Evaluating-Explanations-How-Much-Do-Explanations
  - Explain, Edit, and Understand: Rethinking User Study Design for Evaluating Model Explanations (AAAI), 2022
    https://arxiv.org/abs/2112.09669

- **Funding Support:** UPMC and Abridge AI via the Center for Machine Learning in Health, NSF, DARPA, SEI, Amazon AI, JP Morgan, Salesforce, Facebook, Adobe