Massachusetts Institute of Technology

CSAIL

On the Robustness of Interpretability Methods

David Alvarez-Melis & Tommi S. Jaakkola
CSAIL, MIT

WHI, July 14th, 2018

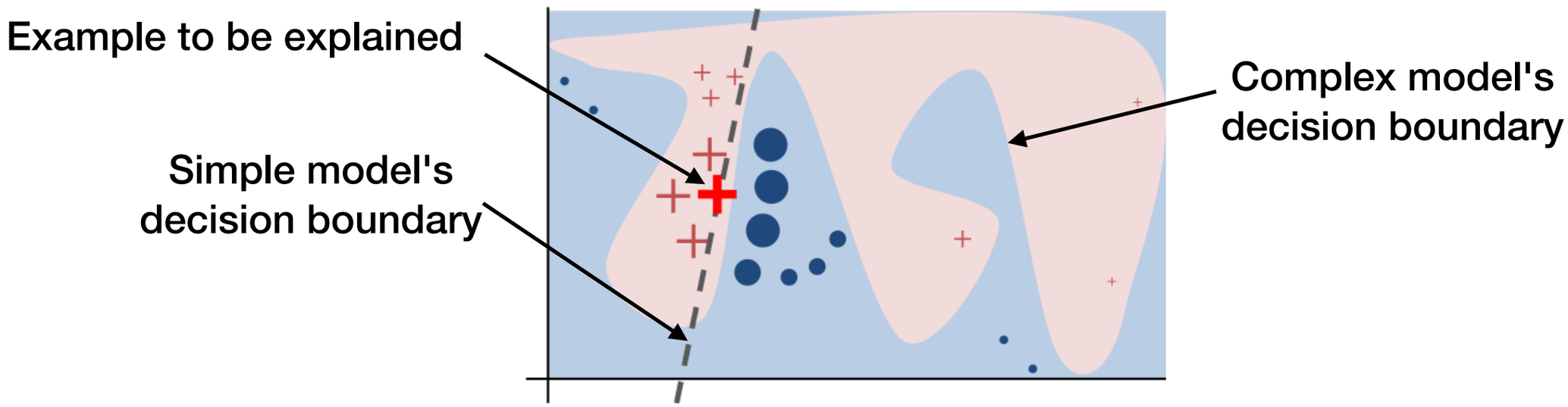
Interpretability through Explanations

- Most prediction-based interpretability models explain by producing **influence scores** for input features **on a particular prediction**

- Two main approaches to estimate influence:

  - Perturbations [Ribeiro et al., 2016; Lundberg & Lee, 2017; A-M & Jaakkola, 2017; etc]

  - Gradients/Relevance Propagation [Simonyan et al., 2014; Bach et al., 2015; Selvaraju et al, 2016; etc]



[Image Credit: Ribiero et al.]

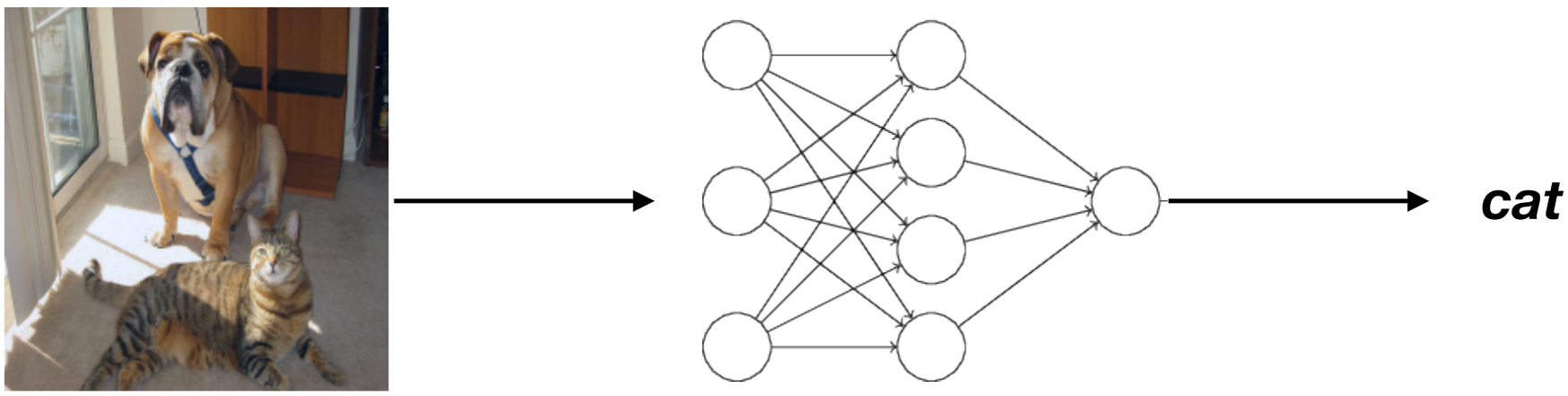
Interpretability through Explanations

- Most prediction-based interpretability models explain by producing **influence scores** for input features **on a particular prediction**

- Two main approaches to estimate influence:

  - Perturbations [Ribeiro et al., 2016; Lundberg & Lee, 2017; A-M & Jaakkola, 2017; etc]

  - Gradients/Relevance Propagation [Simonyan et al., 2014; Bach et al., 2015; Selvaraju et al, 2016; etc]



- Pros: simple, require little (or no) modification to the model being explained, and some of them are derived axiomatically [Lundberg & Lee, 2017].

CSAIL

A Case For Robustness

- Intuitively, similar inputs should give rise to similar explanations

- Why would we want this?

  - In order for an explanation to be **valid** around a point, it should remain roughly constant in its vicinity,

  - Necessary if we want explanation model to act as a **local proxy** for true model

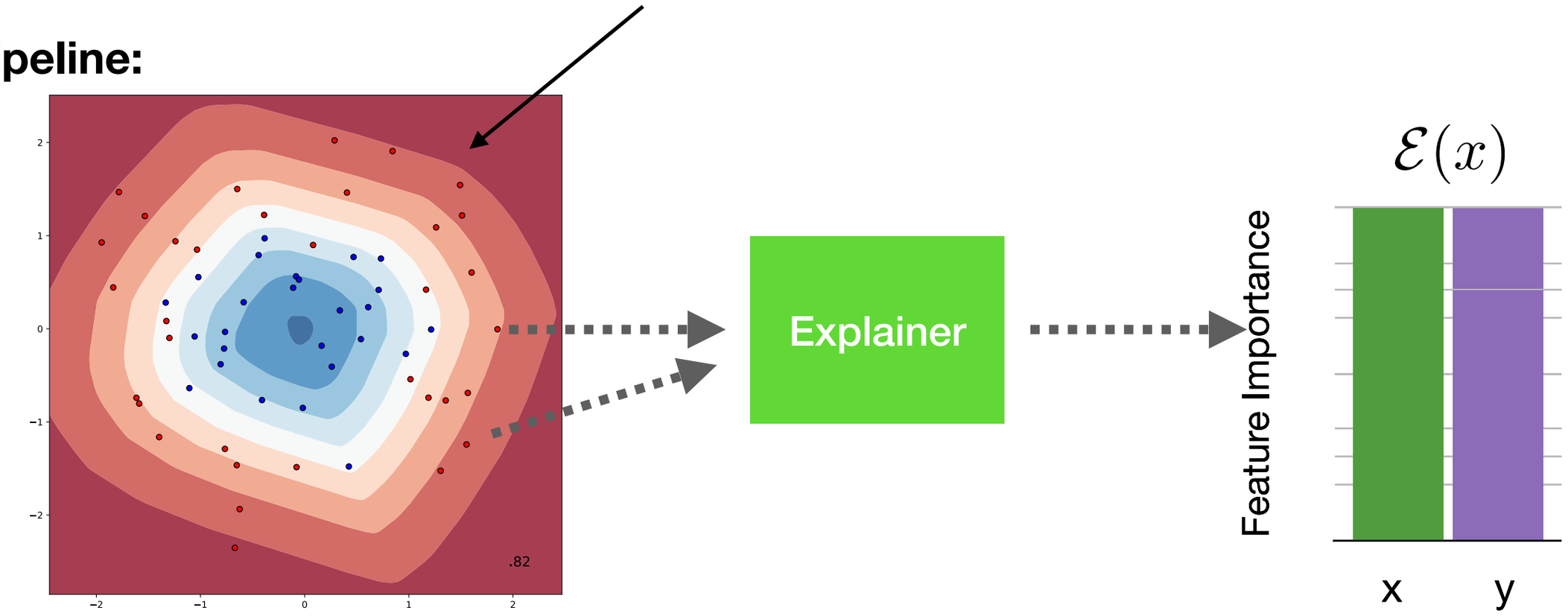
  - Insight gained from a single point-wise explanation might give false sense of understanding - **explain neighborhoods**!

CSAIL

Quantifying robustness

- Input $x \in \mathbb{R}^n$, classifier $\mathcal{M} : \mathbb{R}^n \to \mathbb{R}^k$, explainer $\mathcal{E} : \mathbb{R}^n \to \mathbb{R}^k$

Pipeline:



- How does $\mathcal{E}(x)$ change as x changes?

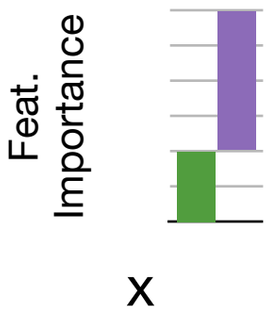
- Relative change in explanation:

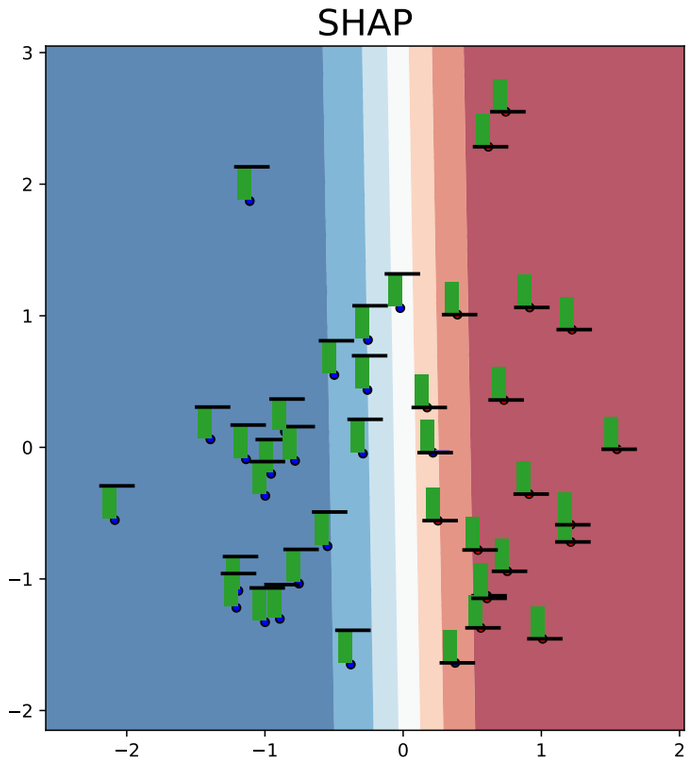
$$\delta(x_i, x_j) = \frac{\|\mathcal{E}(x_i) - \mathcal{E}(x_j)\|_2}{\|x_i - x_j\|_2}$$

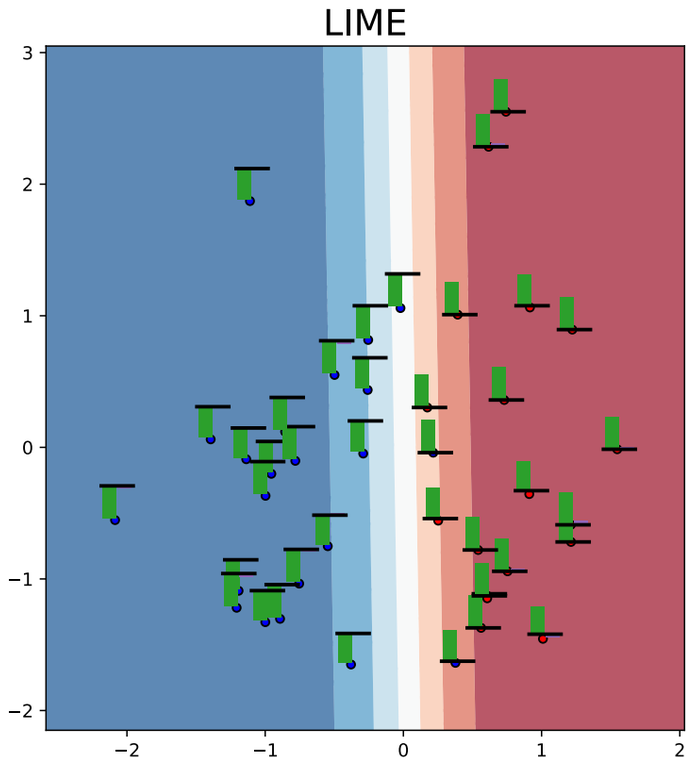
CSAIL

A simple example

Predictor:
Linear SVM



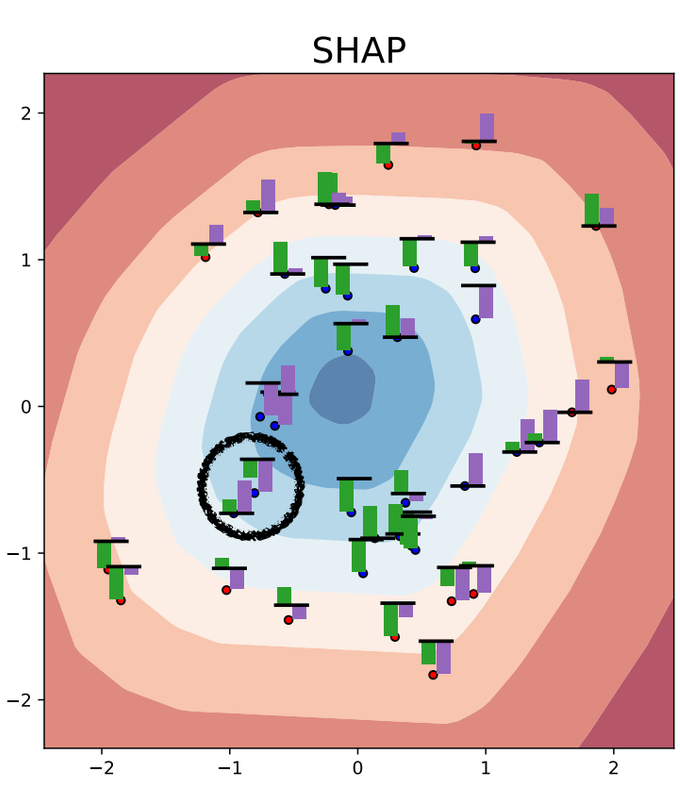


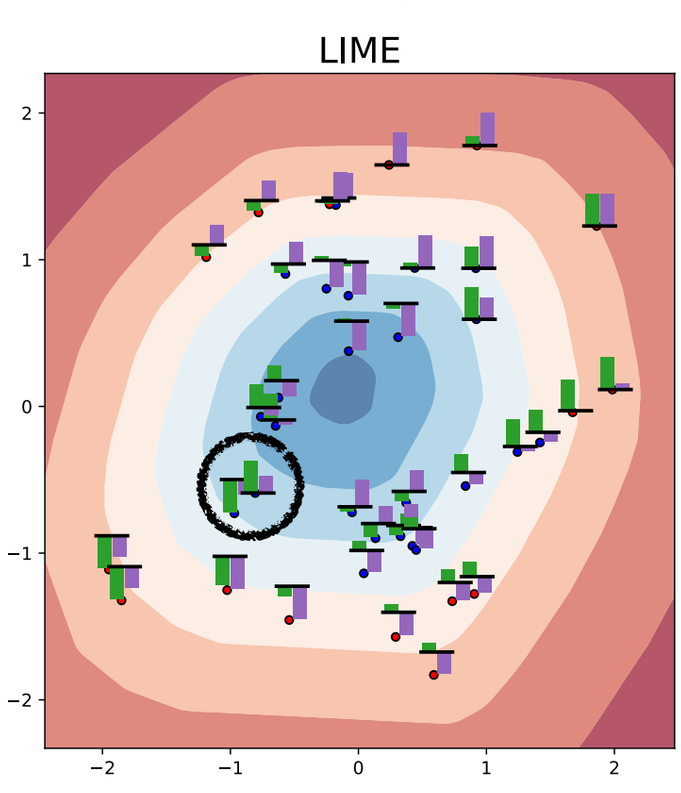


Explanation:
"***only x axis matters***"

😃

Predictor:
Simple Neural Net





Explanation:
??

😨

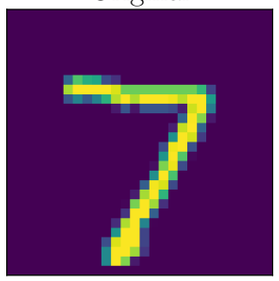
CSAIL

Measuring Robustness

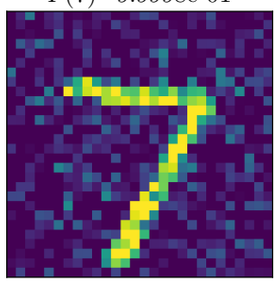
$$\delta(x_i, x_j) = \frac{\|\mathcal{E}(x_i) - \mathcal{E}(x_j)\|_2}{\|x_i - x_j\|_2}$$

Change in classifier's prediction is minimal!

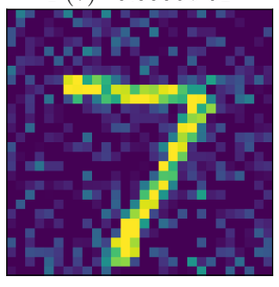
Original



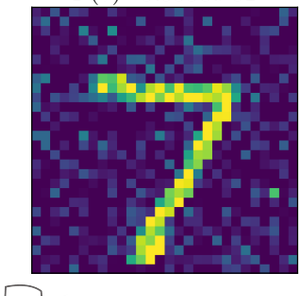
P(7)=9.9998e-01



P(7)=9.9995e-01



P(7)=9.9998e-01



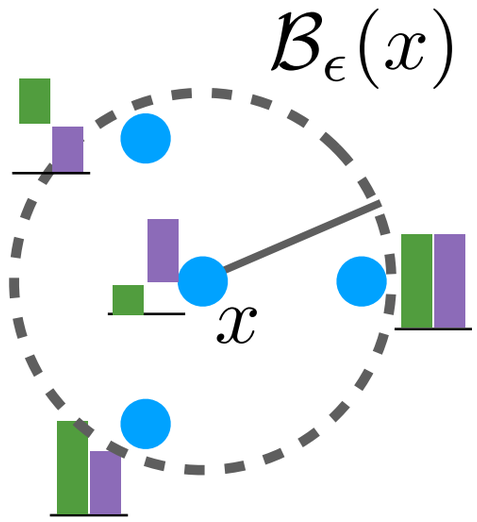
CSAIL

Evaluating Robustness

- Strong notion of robustness: worst-case **local** relative change

$$L(x) = \max_{x' \in B_\epsilon(x)} \frac{\|\mathcal{E}(x) - \mathcal{E}(x')\|_2}{\|x - x'\|_2}$$

a.k.a Local Lipschitz

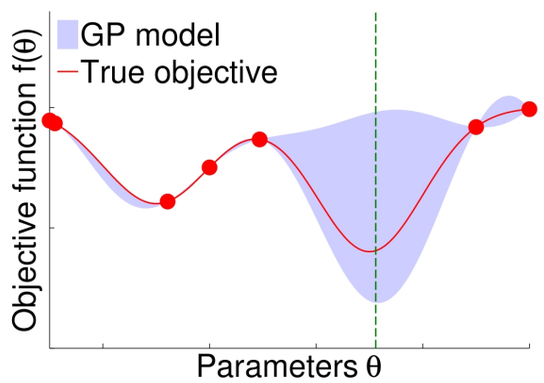


- Computing this needs optimization, but:

  1. Explainers are often not differentiable (e.g. SHAP)

  2. Explainers are expensive to evaluate

- We use **Bayesian Optimization**

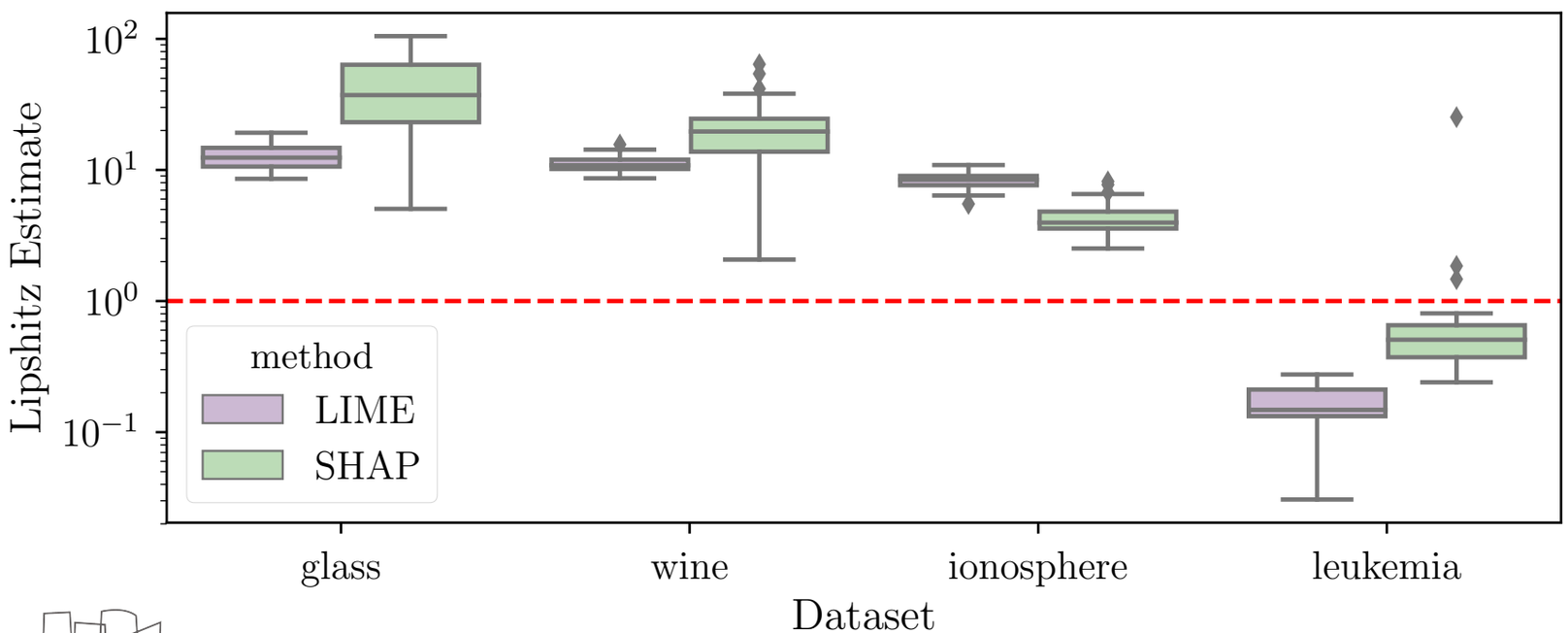


Img: [Calandra et al., 2014]

CSAIL

Results: Perturbation-based Methods

- **Dataset**: UCI repository benchmark datasets
- **Prediction model**: random forest classifier
- **Explainers**: **SHAP** and **LIME**
- Local robustness computed on sample of 200 points



$\hat{L} = 1$

Equal change in input and explanation

CSAIL

Results: Black-Box Methods

- What does the argmax x' look like?

$$L(x) = \max_{x' \in B_\epsilon(x)} \frac{\|\mathcal{E}(x) - \mathcal{E}(x')\|_2}{\|x - x'\|_2}$$

LIME

SHAP

x

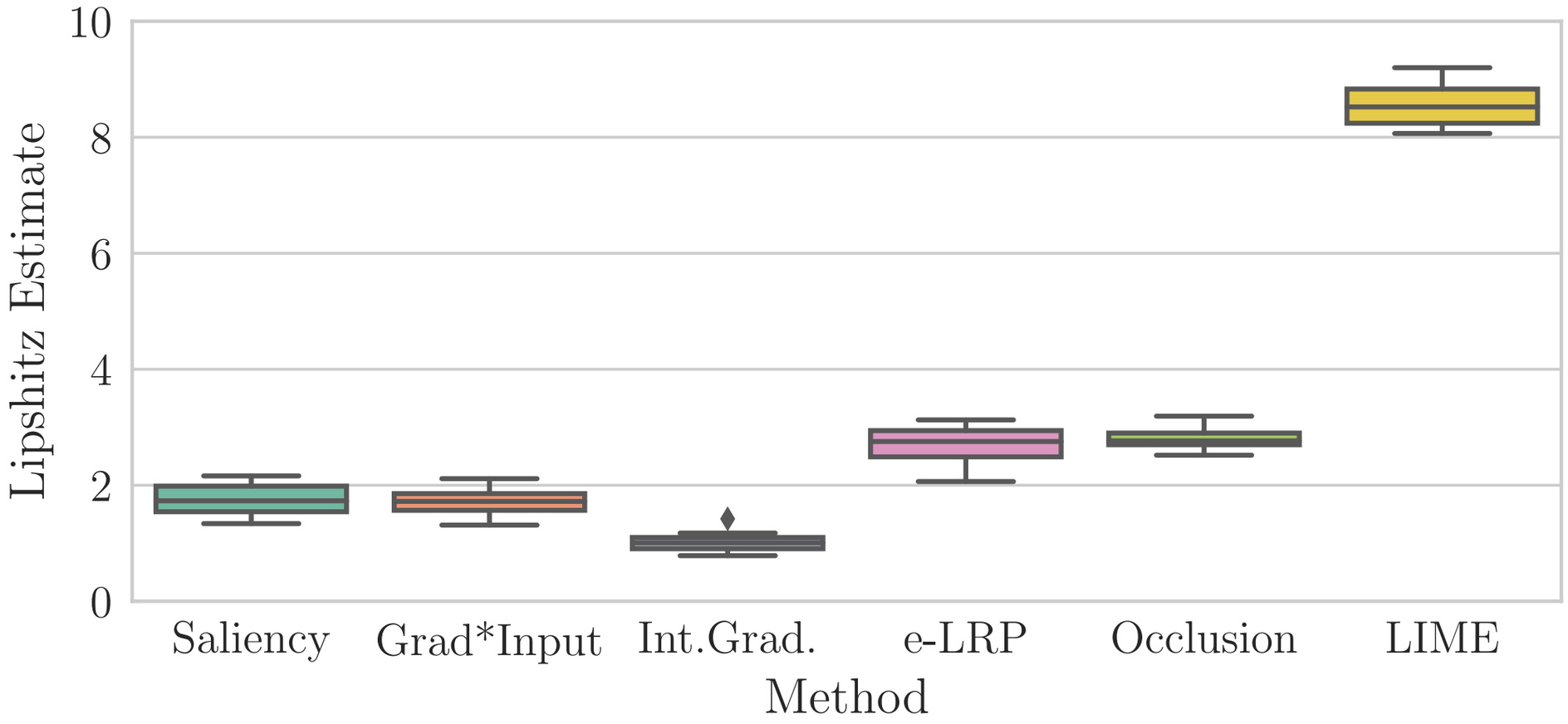
$\|x - x'\| \leq 1$

x'

CSAIL

Results: Digit Classification

- **Dataset**: MNIST
- **Prediction model**: Simple Conv Neural Net, 3 layers
- **Explainers:**
  - SALIENCY maps (Simonyan et al., 2014).
  - GRADIENT*INPUT (Shrikumar et al., 2016).
  - INTegrated GRADients (Sundararajan et al., 2017).
  - ϵ-Layerwise Relevance Propagation (Bach et al., 2015).
  - OCCLUSION sensitivity (Zeiler & Fergus, 2014).



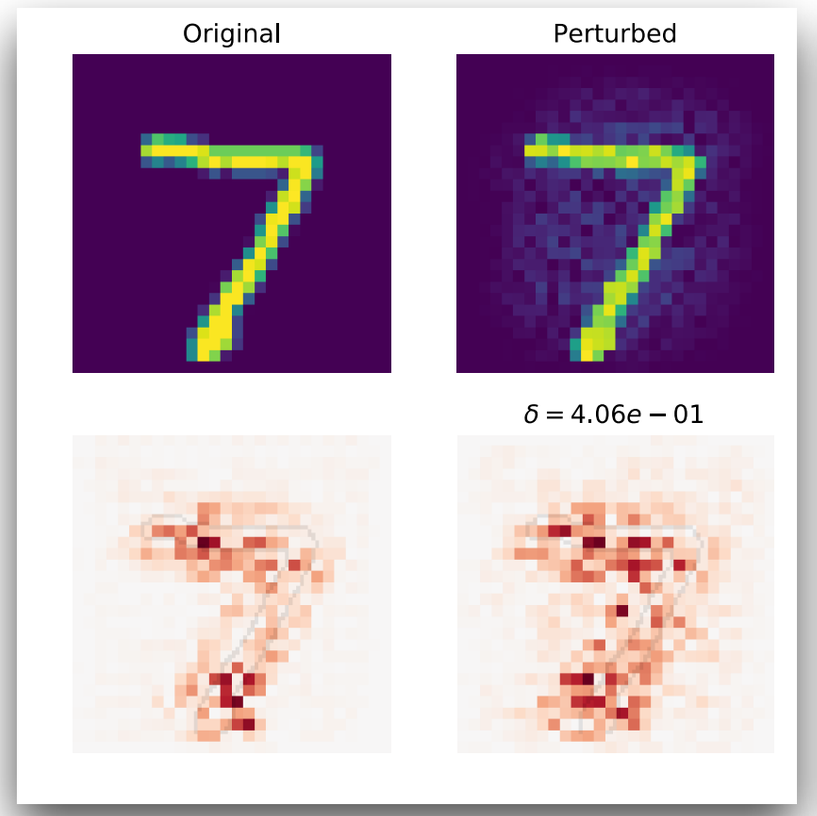
CSAIL

Results: Digit Classification

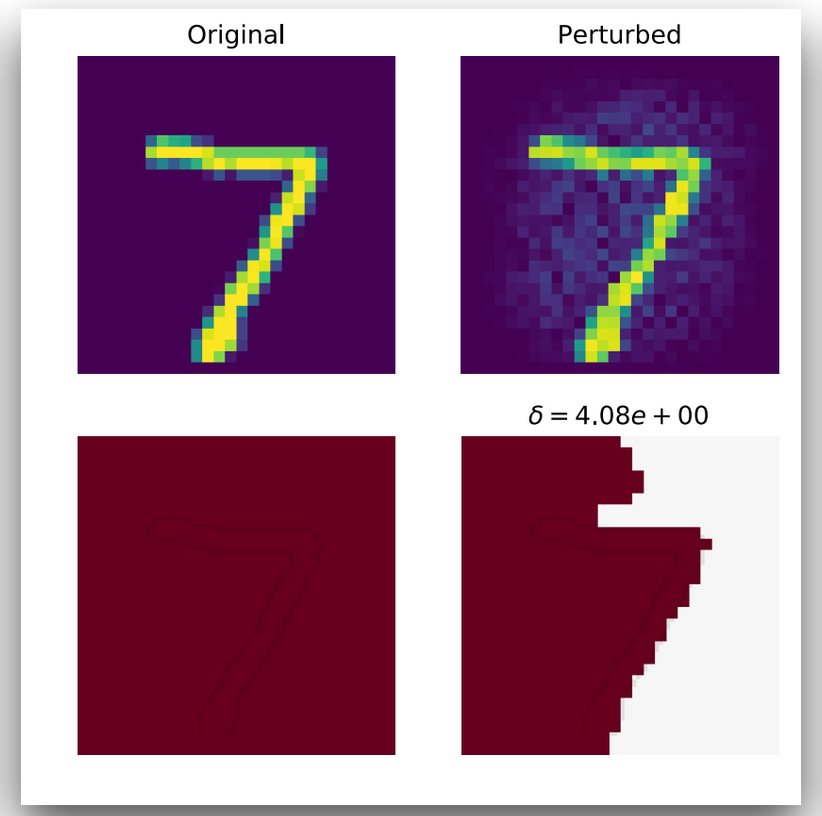
- What do the argmax's look like?

- Gradient based explainers are differentiable - can compute maximizer more precisely (via SGD)

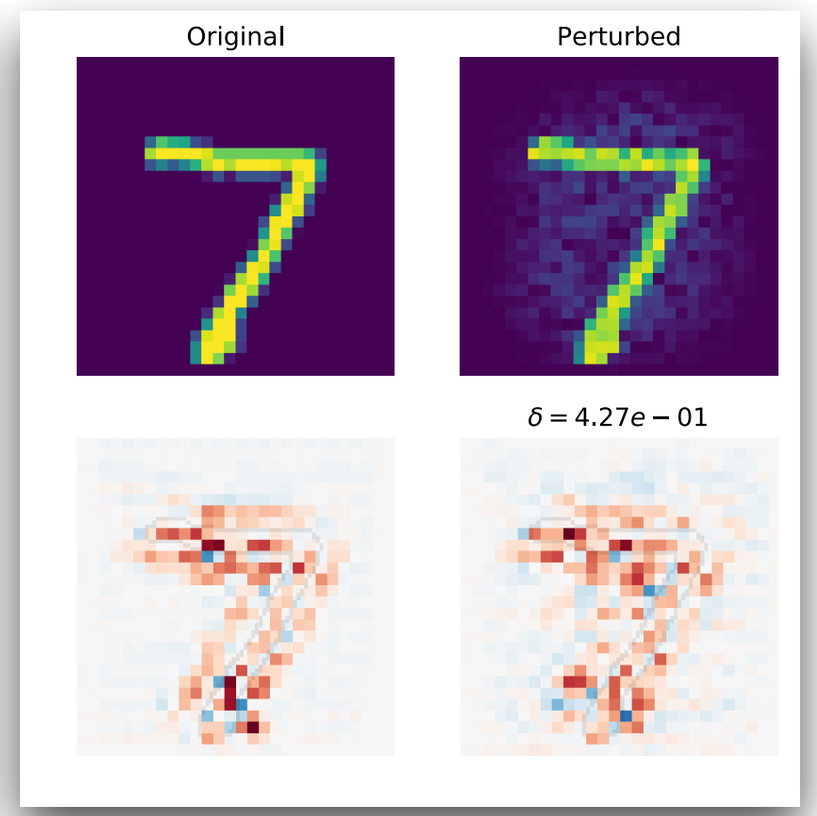
SALIENCY



LIME



e-LRP

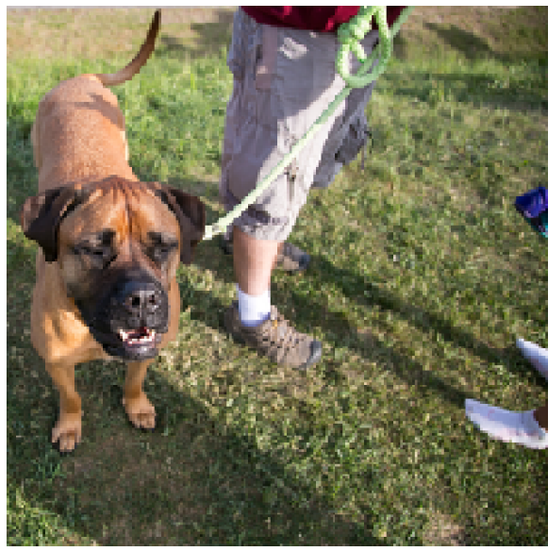


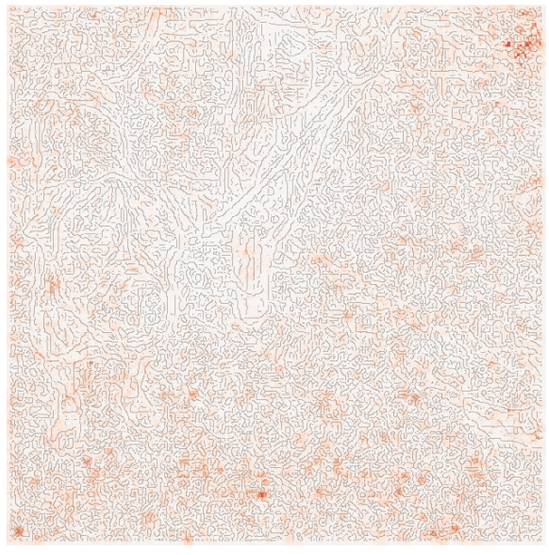
CSAIL

Results: Image Classification

SALIENCY

Original





Discussion

- What if the prediction model itself is not robust? Should we expect the explainer to be?
- Proposition: depends on ***goal*** of interpretabilty

Application-grounded view of interpretability
[Kim & Doshi-Velez, 2018]

  - For **debugging**: explanation should adhere to model
  - For **understanding** both the predictor *and the underlying phenomenon*, should focus on stable aspects of explanation

- Can we enforce explanation robustness while training prediction model?
  - Penalize Lipschitz criterion [A-M & Jaakkola, submitted]
  - Game theoretically [Lee, A-M & Jaakkola, 2018]

Tomorrow @ FAT\ML!!

CSAIL