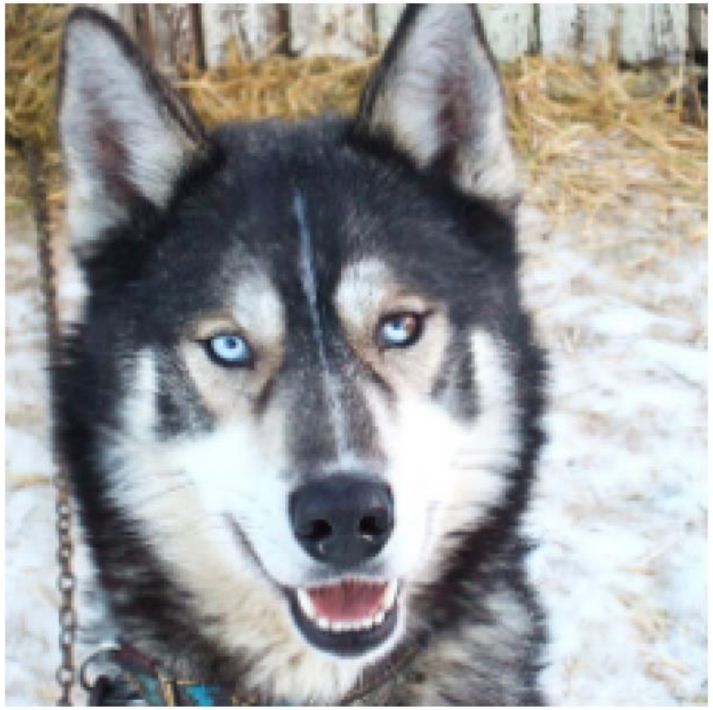
Why Should I Trust You?

Speaker: Philip-W. Grassal

Date: 05/03/18

Why should I trust you?



(a) Husky classified as wolf

(b) Explanation

Based on …

- Title: *Why Should I Trust You? Explaining the Predictions of Any Classifier*
- Authors: Ribeiro, Singh, Guestrin
- Published in: ACM KDD '16 Proceedings

Agenda

I. Contributions

II. Concepts and Theory

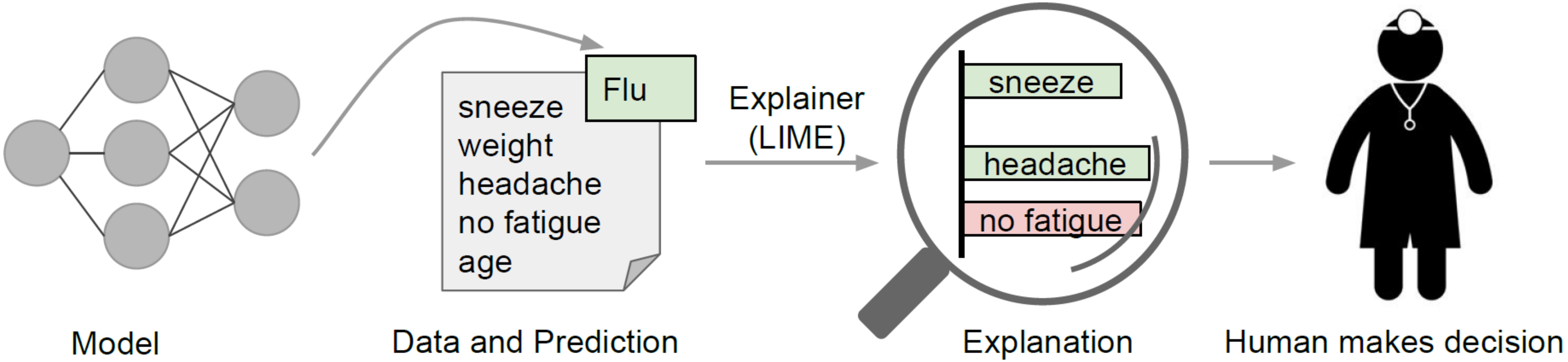
III. Evaluation

IV. Summary of Results

I. Contributions

- Goals
  - Models and predictions will be used only if users can **trust** them
  - Desired: An **interpretable** way to explain the faithfulness of a **prediction** or a **model**
- Contributions
  - LIME, an algorithm explaining any individual predictions
  - SP-LIME, an algorithm explaining any model
  - Evaluation of LIME and SP-LIME with simulated and human subjects

I. Contributions



Basic idea of using LIME

II. LIME

- **L**ocal **I**nterpretable **M**odel-Agnostic **E**xplanations
- Explains if we can trust a single prediction by computing an interpretable model
- Definitions:

original features: $x \in \mathbb{R}^d$

interpretable features: $x' \in \{0,1\}^{d'}$

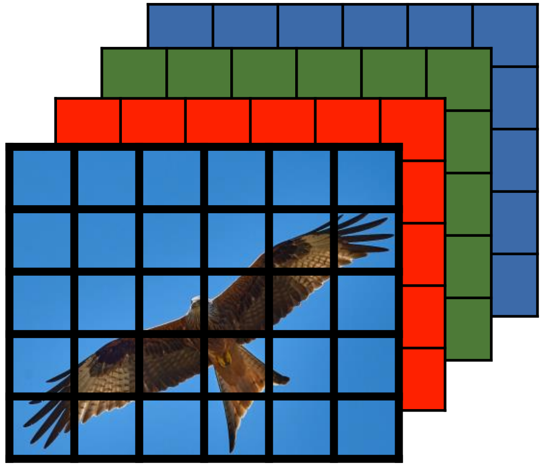
original model: $f\colon \mathbb{R}^d \longrightarrow \mathbb{R}$

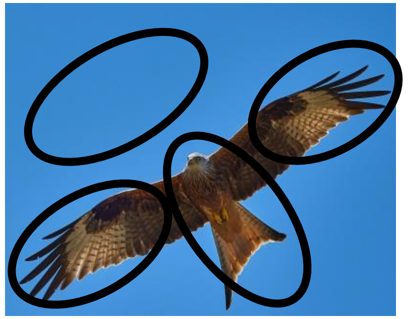
interpretable model: $g\colon \{0,1\}^{d'} \longrightarrow \mathbb{R}$

II. LIME

- Original features: $x \in \mathbb{R}^d$

Interpretable features: $x' \in \{0,1\}^{d'}$





From multiple color channels per pixel to contiguous pixel patches

II. LIME

- Interpretable model: $g\colon \{0,1\}^{d'} \longrightarrow \mathbb{R}$
- $g \in G$ where G describes a family of interpretable models, i. e. they can easily be transferred into visual or textual artefacts, such as
  - Decision trees
  - Simple linear models
- Model complexity is measured with $\Omega(g)$

II. LIME

- Goal of LIME: find an interpretable model $\hat{g}_x$ that locally approximates the original model f w. r. t. instance x
- Locality is defined by proximity/distance measure π_x around x
- Let $\mathcal{L}$ define the approximation loss, we compute

$$\hat{g}_x = \underset{g \, \epsilon \, G}{\operatorname{argmin}} \, \mathcal{L}(f, g, \pi_x) \; + \Omega(g)$$

How well does it approximate?

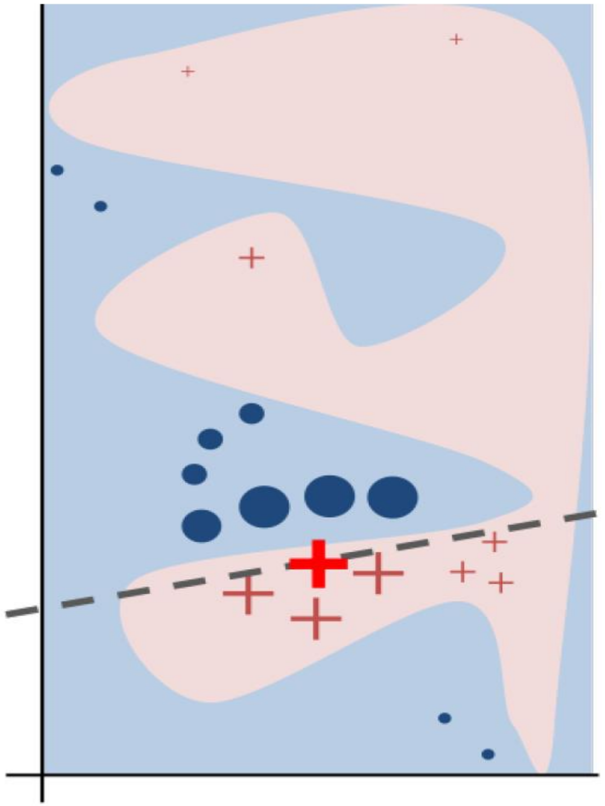
How complex is the model?

II. LIME for Sparse Linear Models

- G is family of K-sparse linear models,
  i. e. $g(x') = w_g x'$ and $||w_g||_0 \leq K$
- To measure if g is a good local approximation, multiple instances z', z are sampled around x', x
- $\mathcal{L}$ becomes a weighted least squares objective

$$\mathcal{L}(f, g, \pi_x) = \sum_{z,z'} \pi_x(z)\left(f(z) - g(z')\right)^2$$

$$\Omega(g) = \infty * \mathbb{I}[||w_g||_0 > K]$$

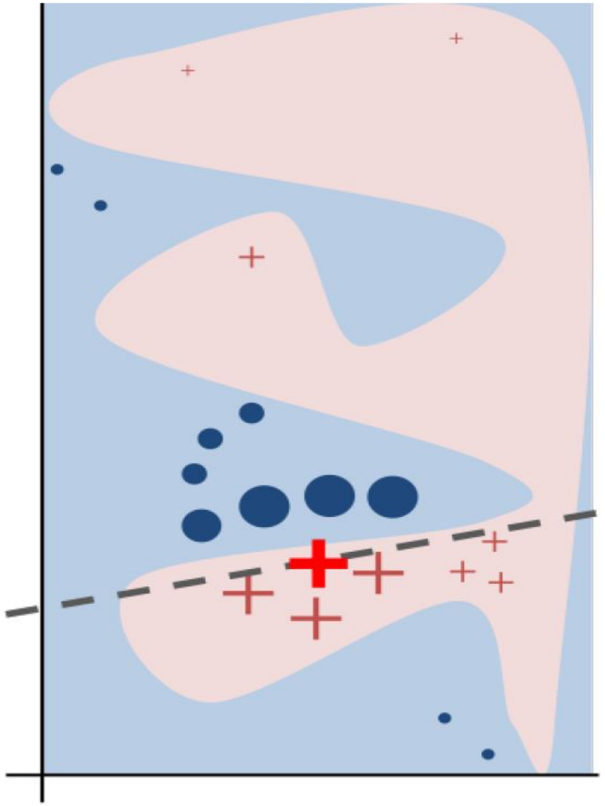


Local linear approx. of complex model

II. LIME for Sparse Linear Models

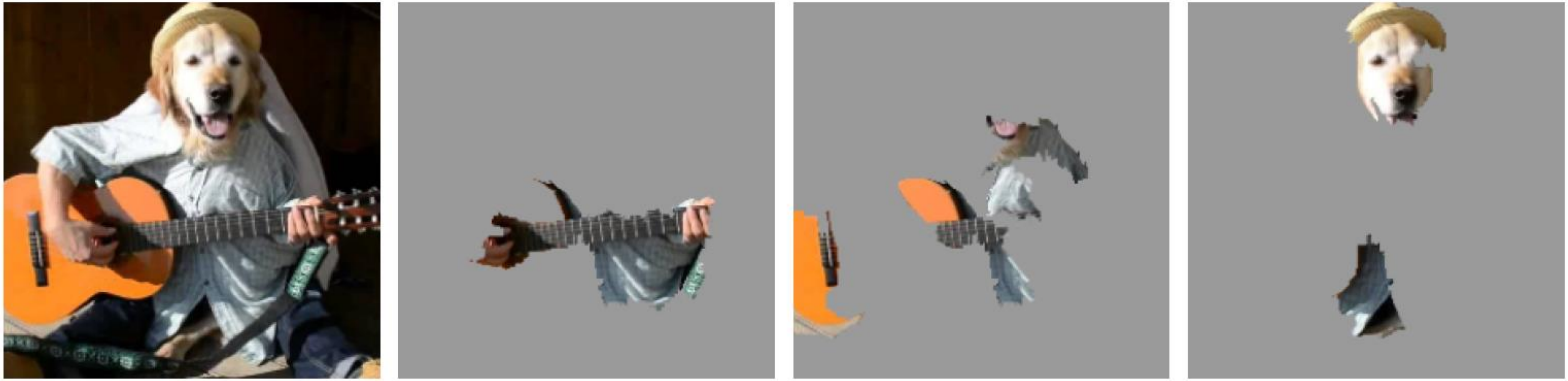
- $\mathcal{L}(f, g, \pi_x) = \sum_{z,z'} \pi_x(z)\big(f(z) - g(z')\big)^2$
- $\Omega(g) = \infty * \mathbb{I}[||w_g||_0 > K]$
- Solve: $\hat{g}_x = \underset{g \, \epsilon \, G}{\text{argmin}} \, \mathcal{L}(f, g, \pi_x) + \Omega(g)$

1. Use **Lasso regularization** to set $\Omega(g) = 0$
2. Use standard solver for WLS-objective



Local linear
approx. of complex model

II. LIME for Sparse Linear Models



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

Explaining Google's Inception neural network

II. SP-LIME

- LIME: fidelity is only evaluated locally
- Submodular Pick – LIME: estimate global fidelity by local explainers
- Idea: Let X denote a test set, a model g_x is computed via LIME for all $x \epsilon X$. Based on the weights w_{g_x} select the B most representative local models. Can we trust them?

→ Yes? Then we can trust the model, too

II. SP-LIME

- How to select $B = 2$ most representative models? VERY SIMPLIFIED!

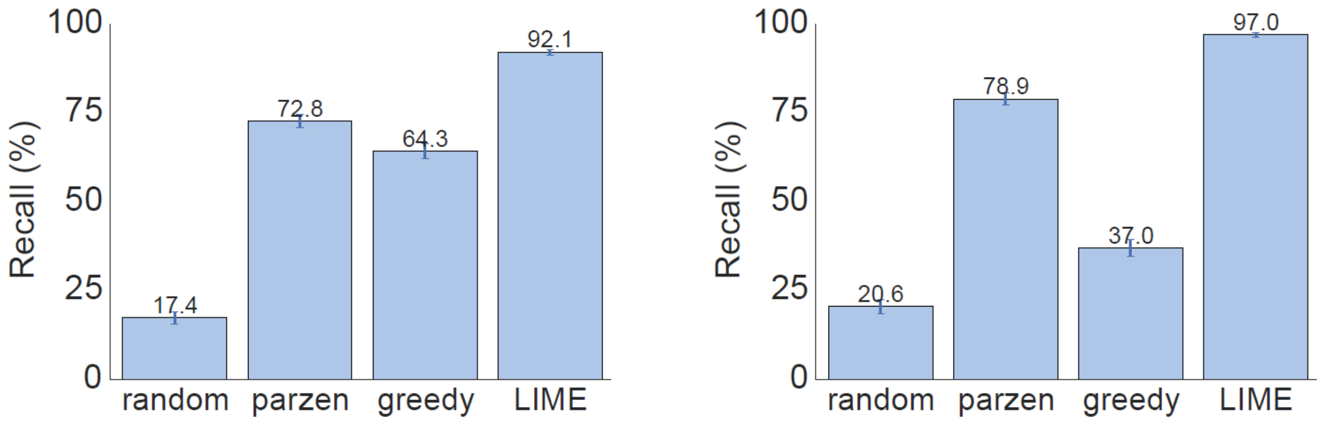
	f1	f2	f3	f4	
x'_1	w11	w12	w13	w14	→ 2. Choice
x'_2	w21	w22	w23	w24	→ 1. Choice
x'_3	w31	w32	w33	w34	
x'_4	w41	w42	w43	w44	
Importance:	3	1	2	2	

III. Evaluation – Simulated User Experiments

- Train classifiers with books and DVDs dataset for sentiment prediction
- Compare LIME with **10-sparse linear models** to other black box methods from literature

III. Evaluation – Simulated User Experiments

- Are interpretable predictors faithful to the model?
- Experiment: let interpretable models identify relevant features



(a) Sparse LR (b) Decision Tree

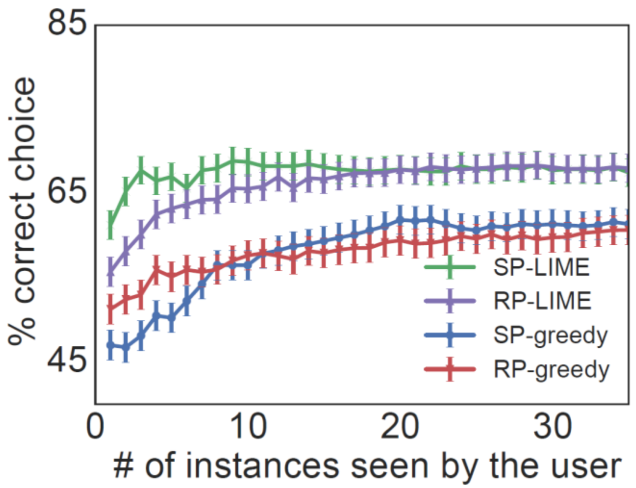
Recall of explainers for sparse linear regression and decision tree using the book data set

III. Evaluation – Simulated User Experiments

- Can a prediction be trusted?
- Experiment: let explainers identify untrustworthy features

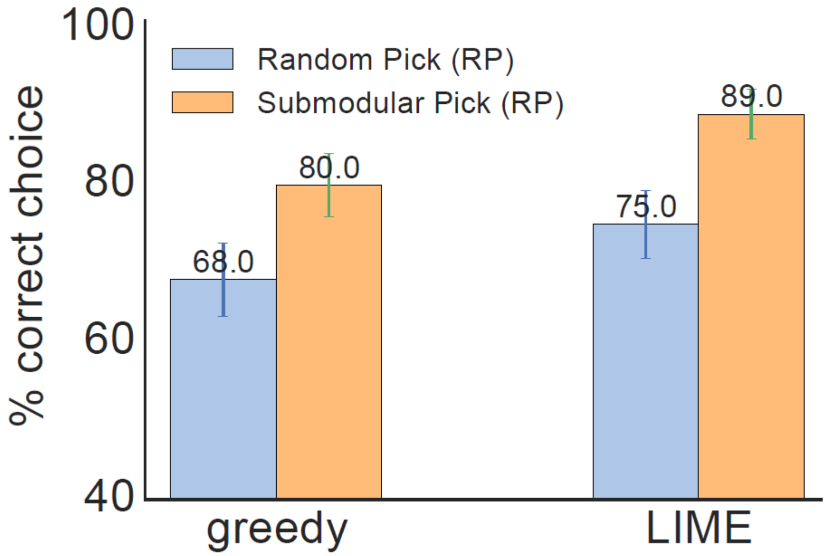
	Books			
	LR	NN	RF	SVM
Random	14.6	14.8	14.7	14.7
Parzen	84.0	87.6	94.3	92.3
Greedy	53.7	47.4	45.0	53.3
LIME	**96.6**	**94.5**	**96.2**	**96.7**

- Can the model be trusted?
- Experiment: let explainers find the best model



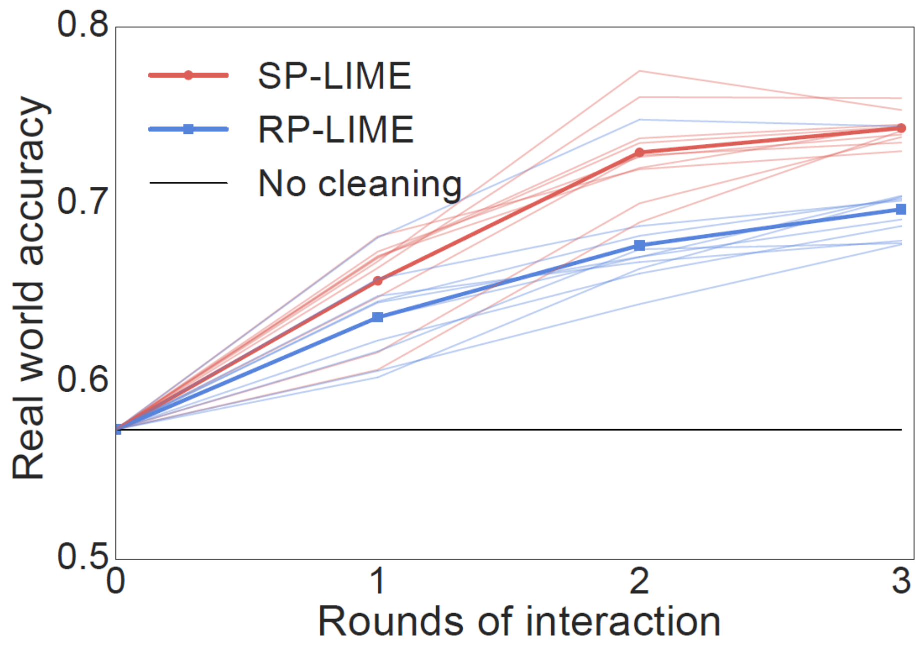
III. Evaluation – Human Subjects

- Does SP-LIME help people to decide whether a model is trustworthy?
- Survey based on confession classifiers trained on religious texts data set



III. Evaluation – Human Subjects

- Does LIME enable non-experts to improve a classifier?
- For multiple rounds of explanation, participants removed features by using LIME to improve a given classifier



IV. Summary of Results

- LIME, SP-LIME provide interpretable approximations of complex models
- Outperform other recent approaches
- Complement summary statistics (test accuracy) to evaluate the trustworthiness of a model

Thank you!

Contact Information:
Philip-W. Grassal
grassal@stud.uni-heidelberg.de