LIME and anchors

Local interpretable model-agnostic explanations

Alicja Gosiewska
MI2 DataLab,
Faculty of Mathematics and Information Science,
Warsaw University of Technology

LIME

Local Interpretable Model-Agnostic Explanations

LIME
Local Interpretable Model-Agnostic Explanations

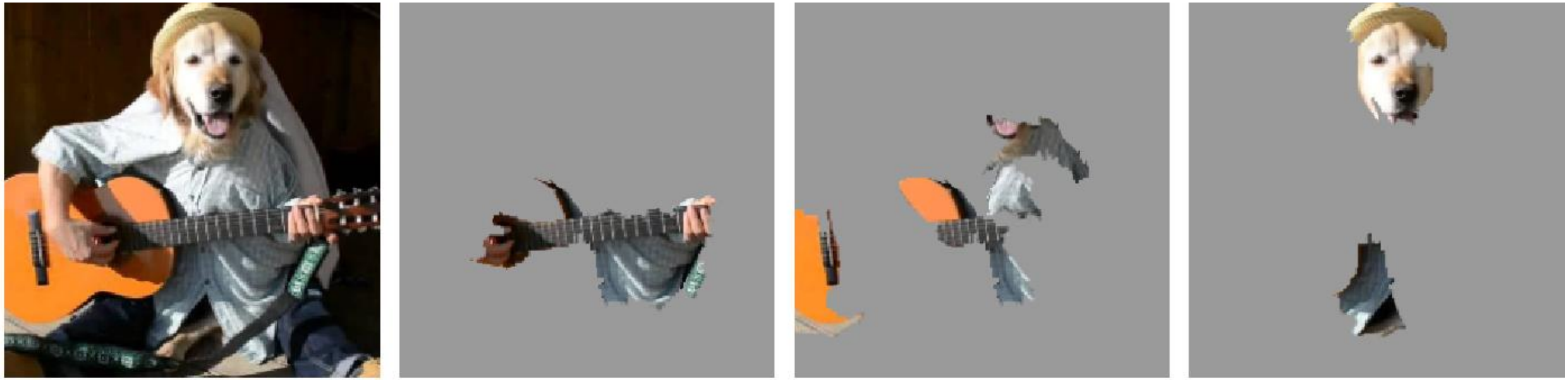
(a) Original Image

Explaining an image classication prediction made by Google's Inception neural network.

The top 3 classes predicted are:
- Electric Guitar (p = 0.32),
- Acoustic guitar (p = 0.24) ,
- Labrador (p = 0.21).

Source: M. T. Ribeiro, S. Singh, C. Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier.

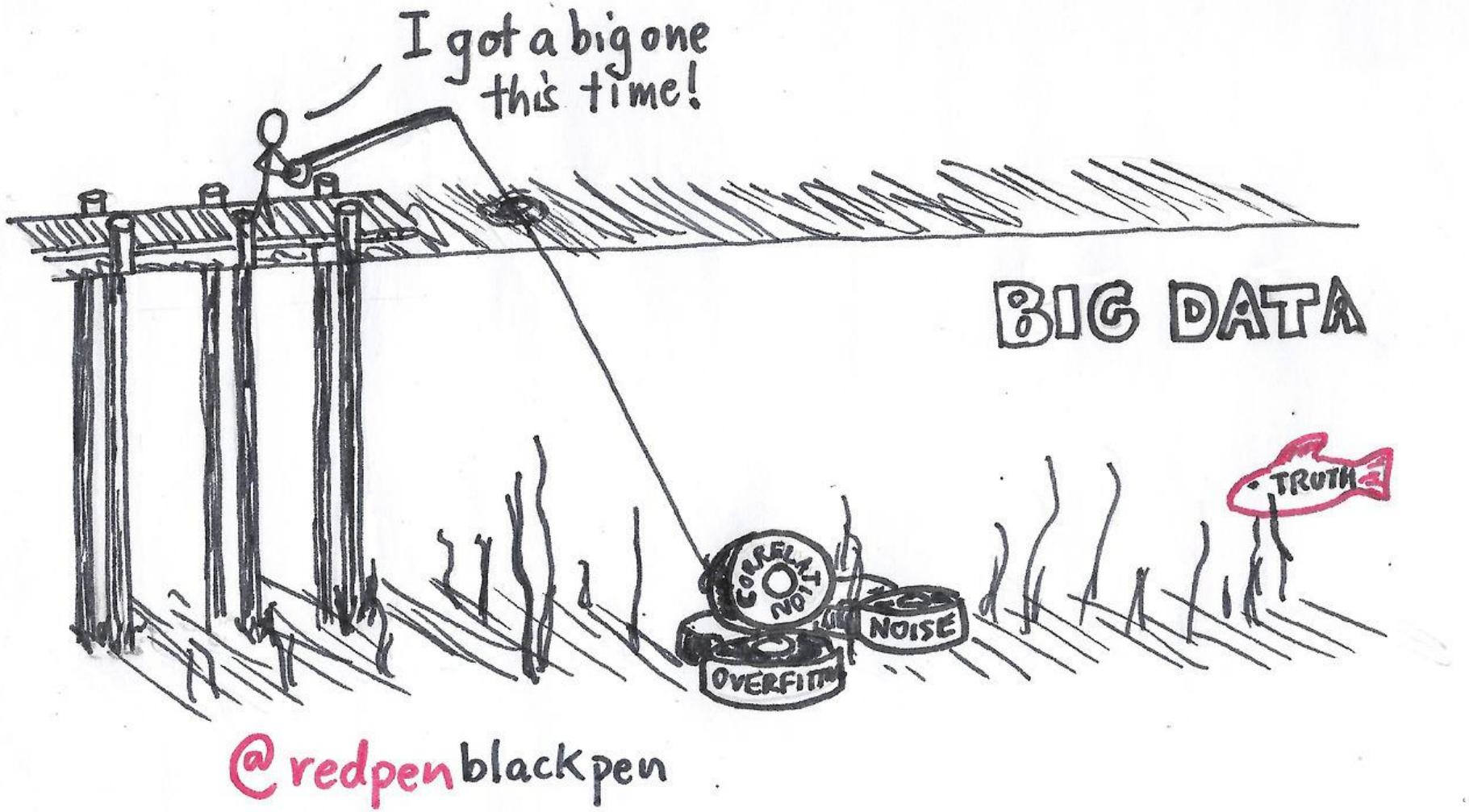
LIME
Local Interpretable Model-Agnostic Explanations



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

Source: M. T. Ribeiro, S. Singh, C. Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier.

Can you build your trust based on accuracy?



Can you build your trust based on accuracy?

Predicted: wolf
True: wolf

Predicted: husky
True: husky

Predicted: wolf
True: wolf

Predicted: wolf
True: husky

Predicted: husky
True: husky

Predicted: wolf
True: wolf

Source: M. T. Ribeiro, S. Singh, C. Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier.

Yes, if you want to build a great snow detector!

Predicted: wolf
True: wolf

Predicted: husky
True: husky

Predicted: wolf
True: wolf

Predicted: wolf
True: husky

Predicted: husky
True: husky

Predicted: wolf
True: wolf

Source: M. T. Ribeiro, S. Singh, C. Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier.

$x \in \mathbb{R}^d$ - original representation of the instance being explained

$x' \in 0, 1^{d'}$ - a binary vector with interpretable representation of x

$f : \mathbb{R}^d \rightarrow \mathbb{R}$ - model being explained

$g \in G$ - explaination model where G is a class of potentially interpretable

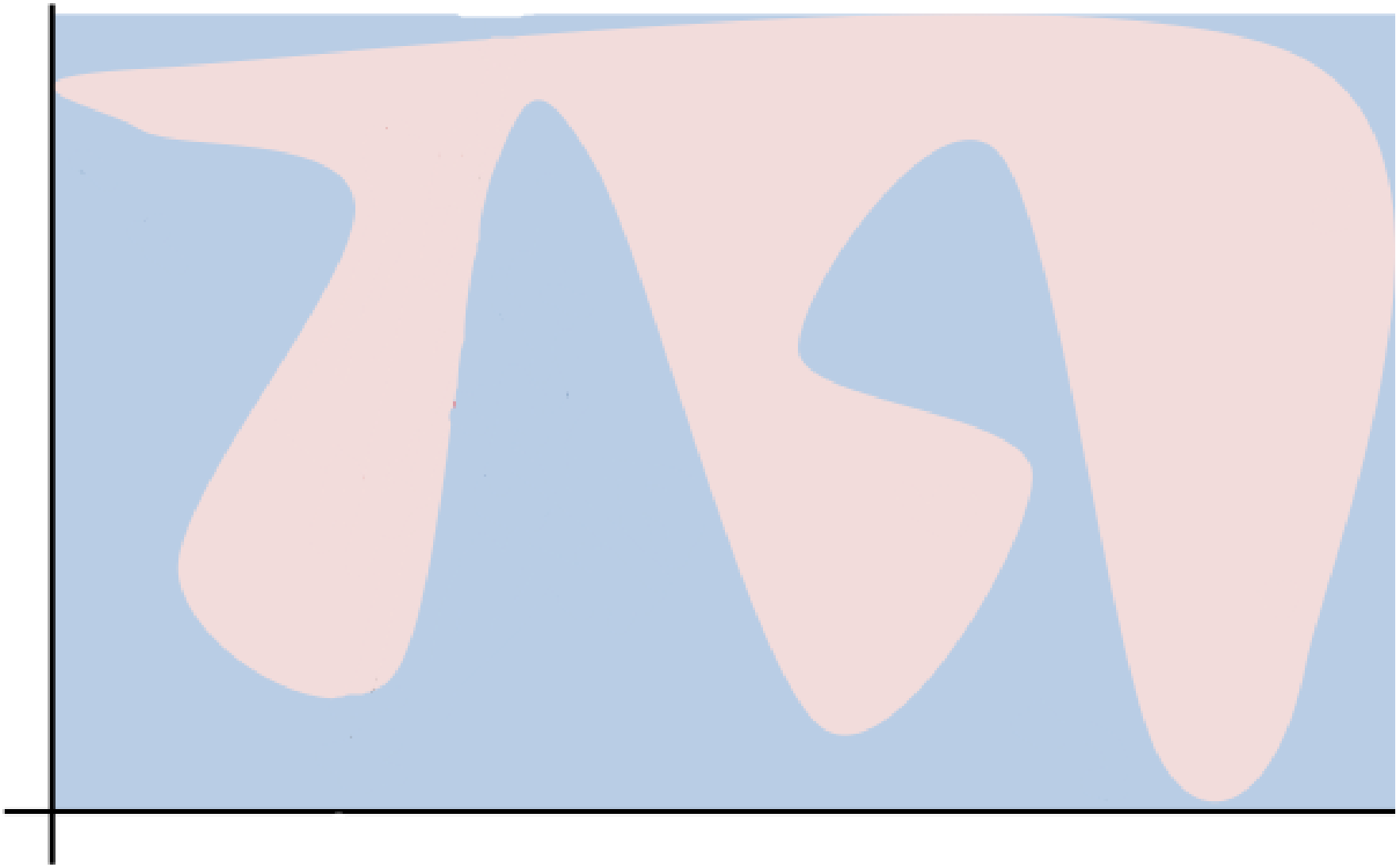
$\pi_x(z)$ - proximity measure between an instance z to x

$\Omega(g)$ - measure of complexity

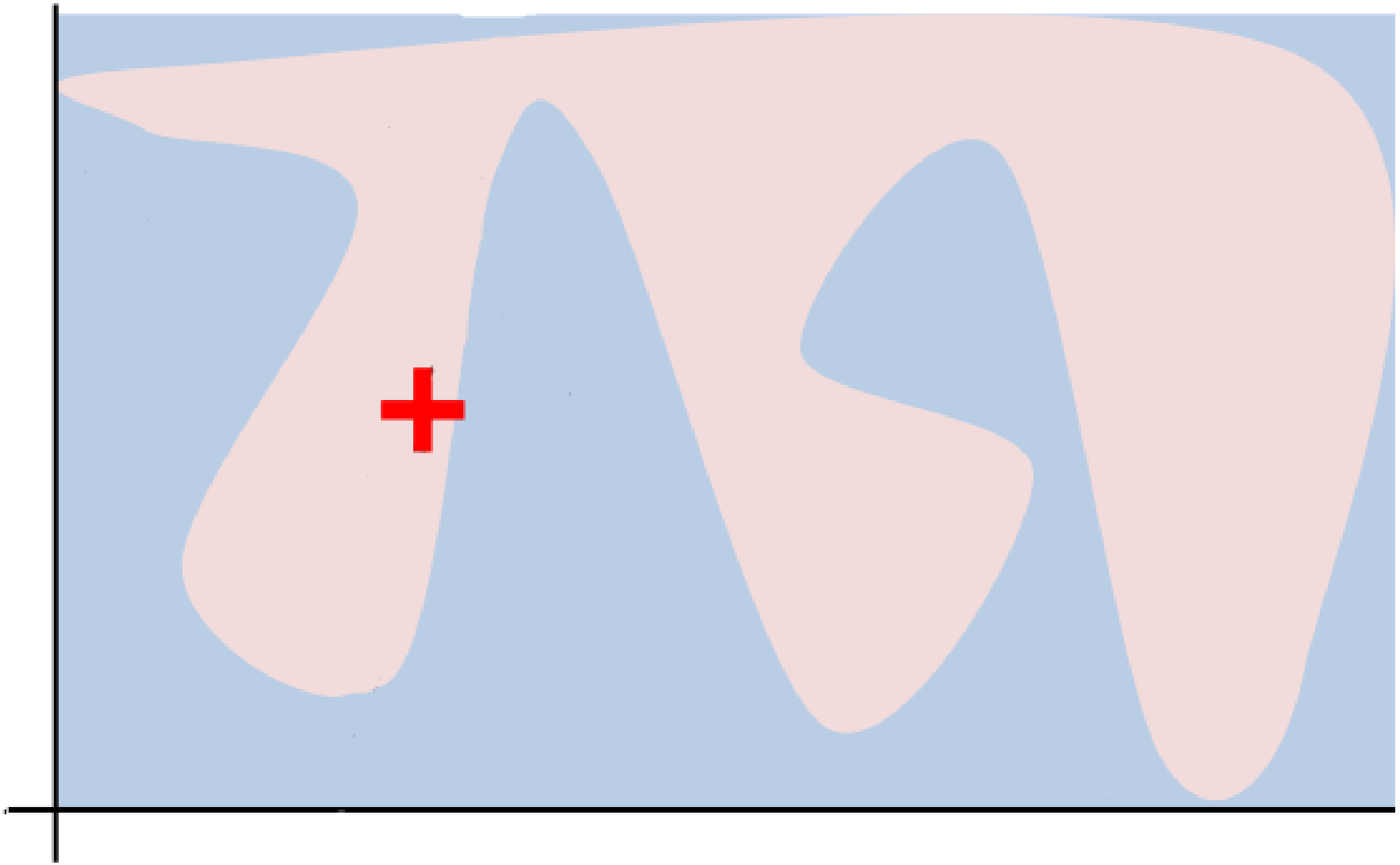
$\mathcal{L}(f, g, \pi_x)$ - a measure of how unfaithful g is in approximating f in the locality defined by $\pi_x(z)$

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

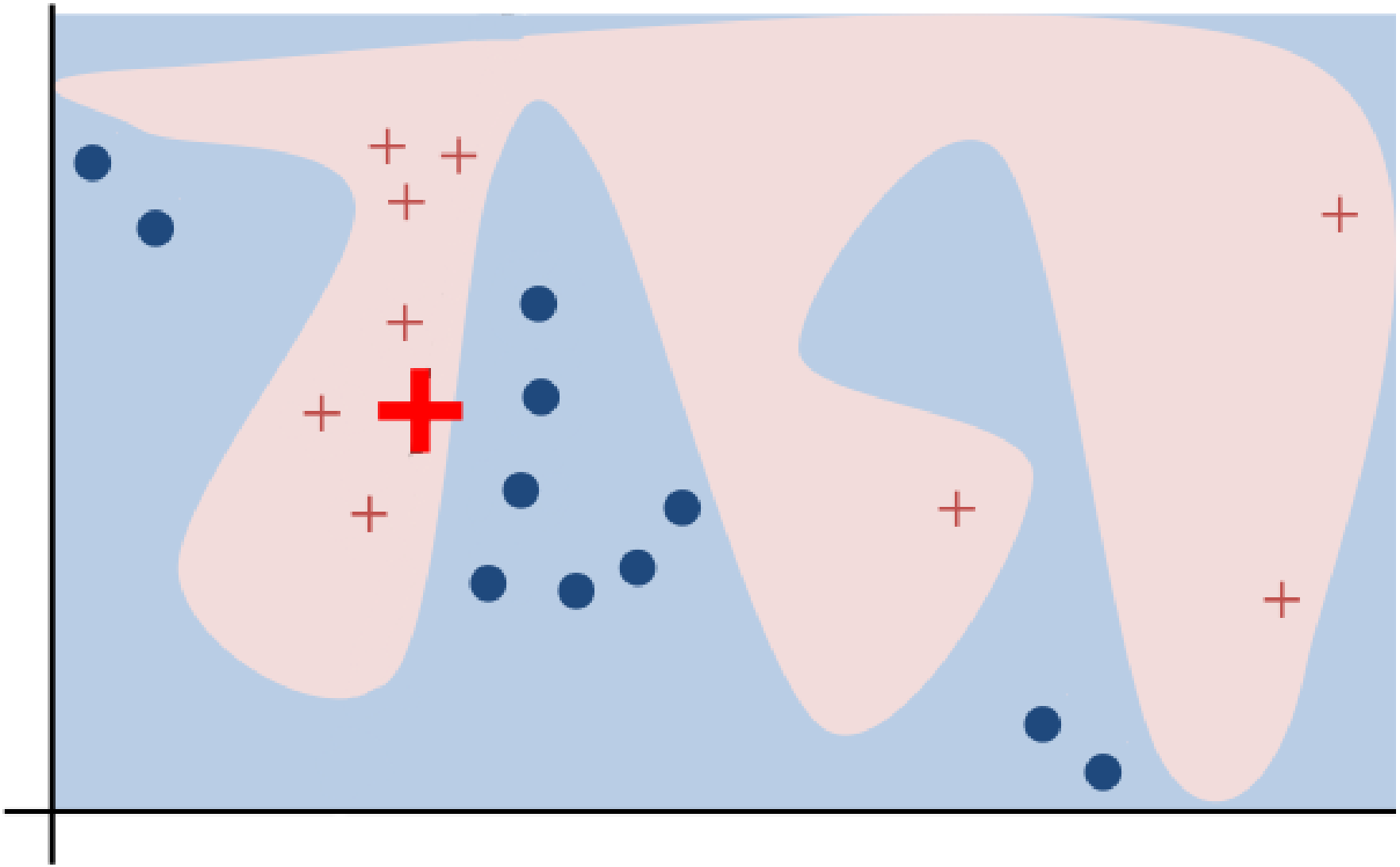
Intuition for LIME



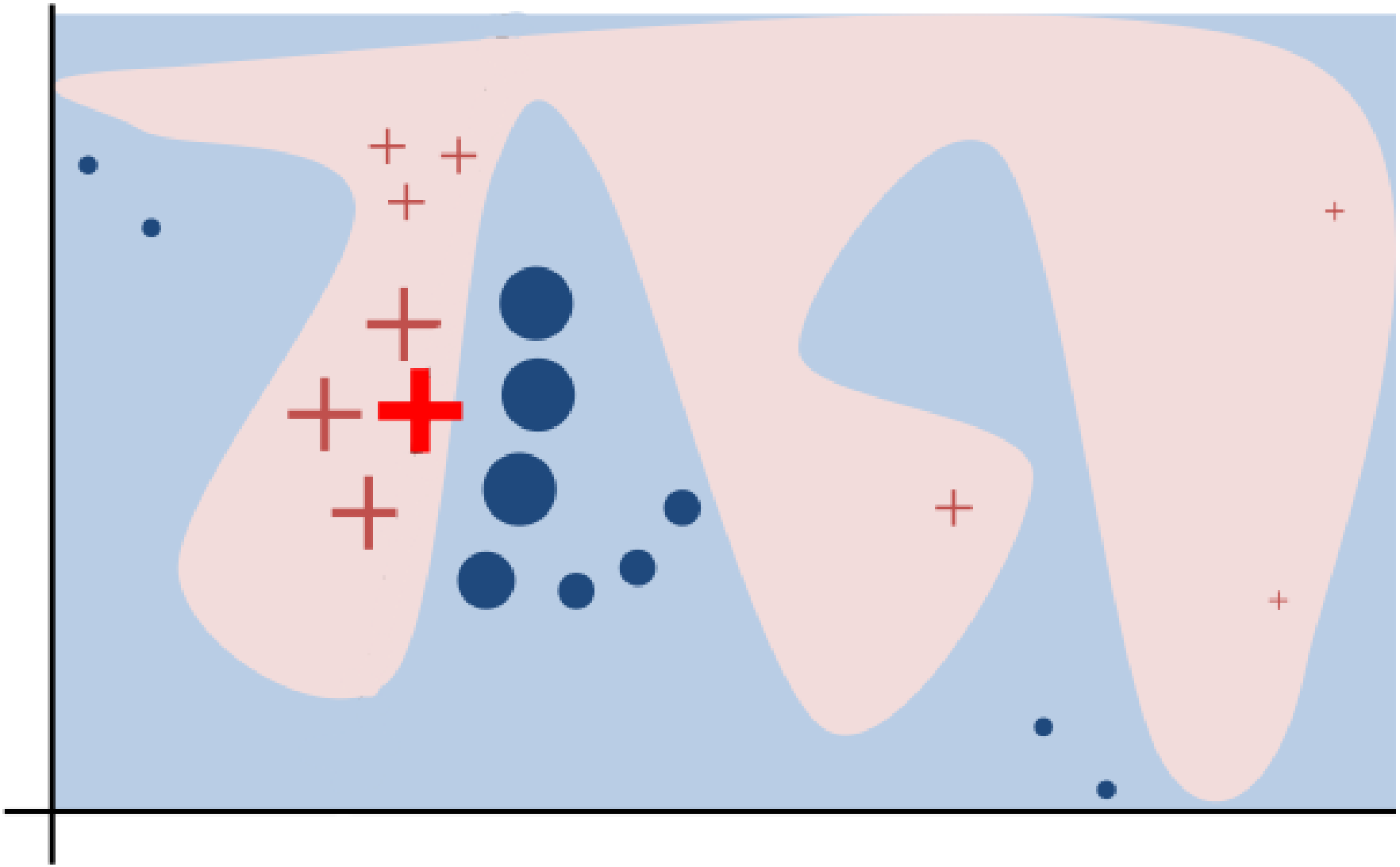
Intuition for LIME



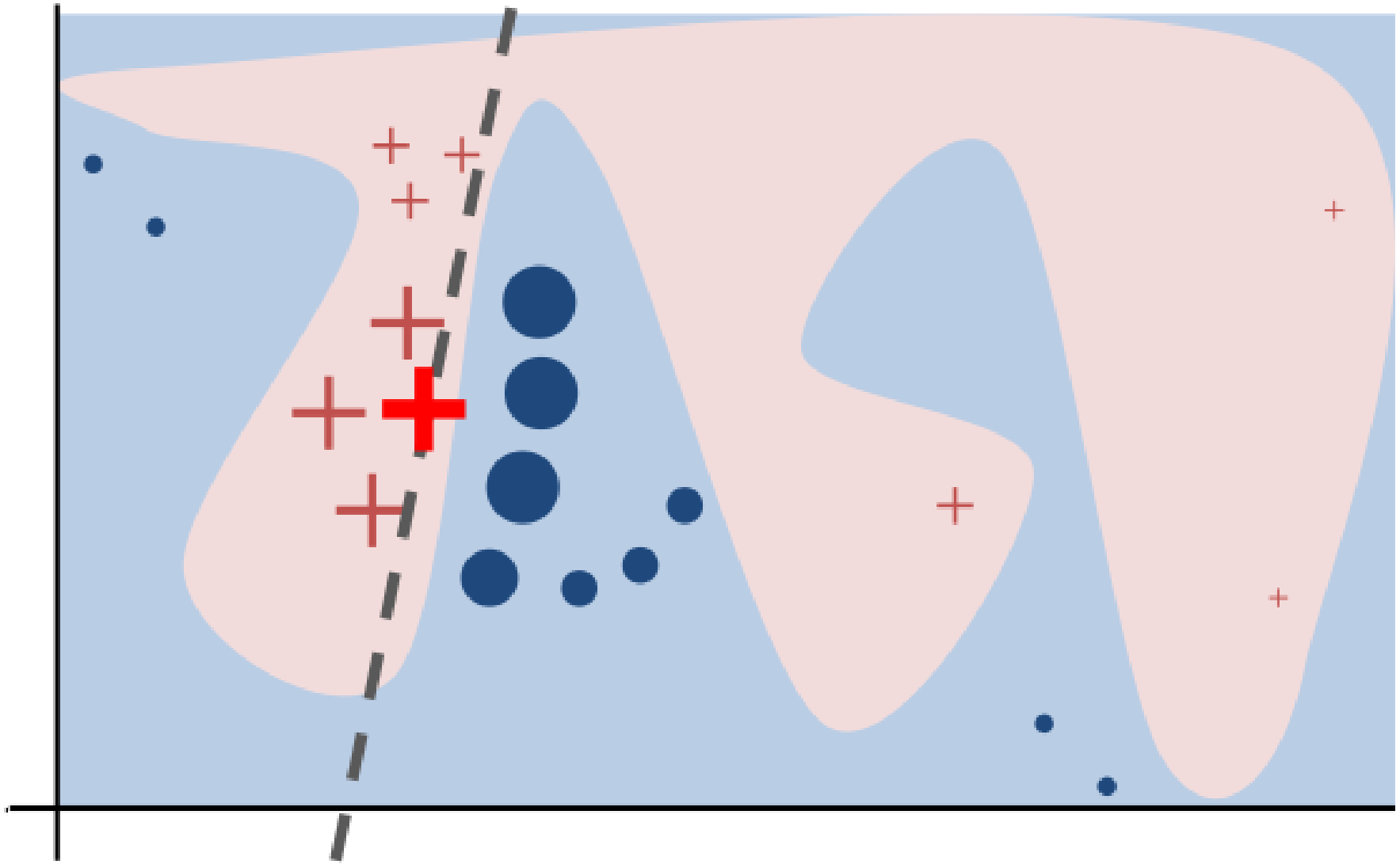
Intuition for LIME



Intuition for LIME



Intuition for LIME

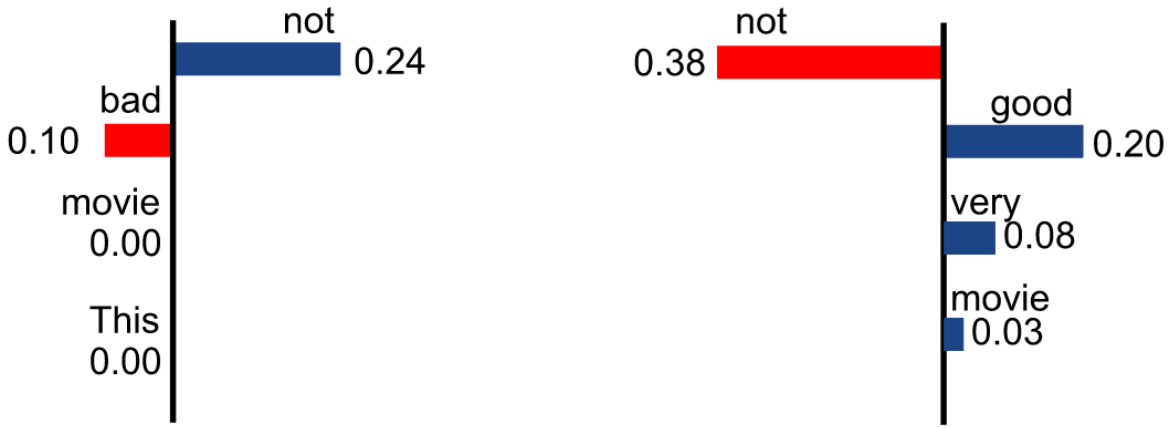


LIME for text analysis

+ This movie is not bad.

− This movie is not very good.

LIME explanations



Anchors

High-Precision Model-Agnostic Explanations

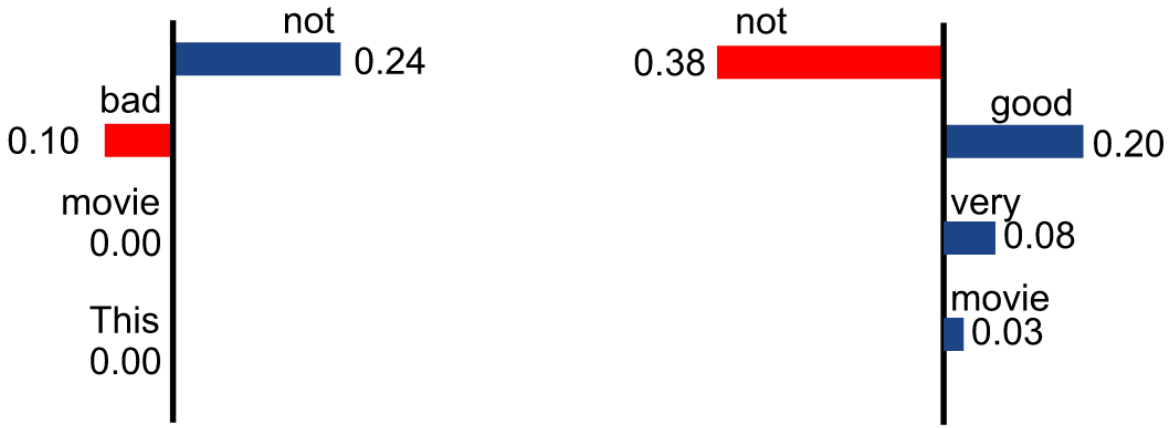
MI

Anchors for text analysis

+ This movie is not bad.

− This movie is not very good.

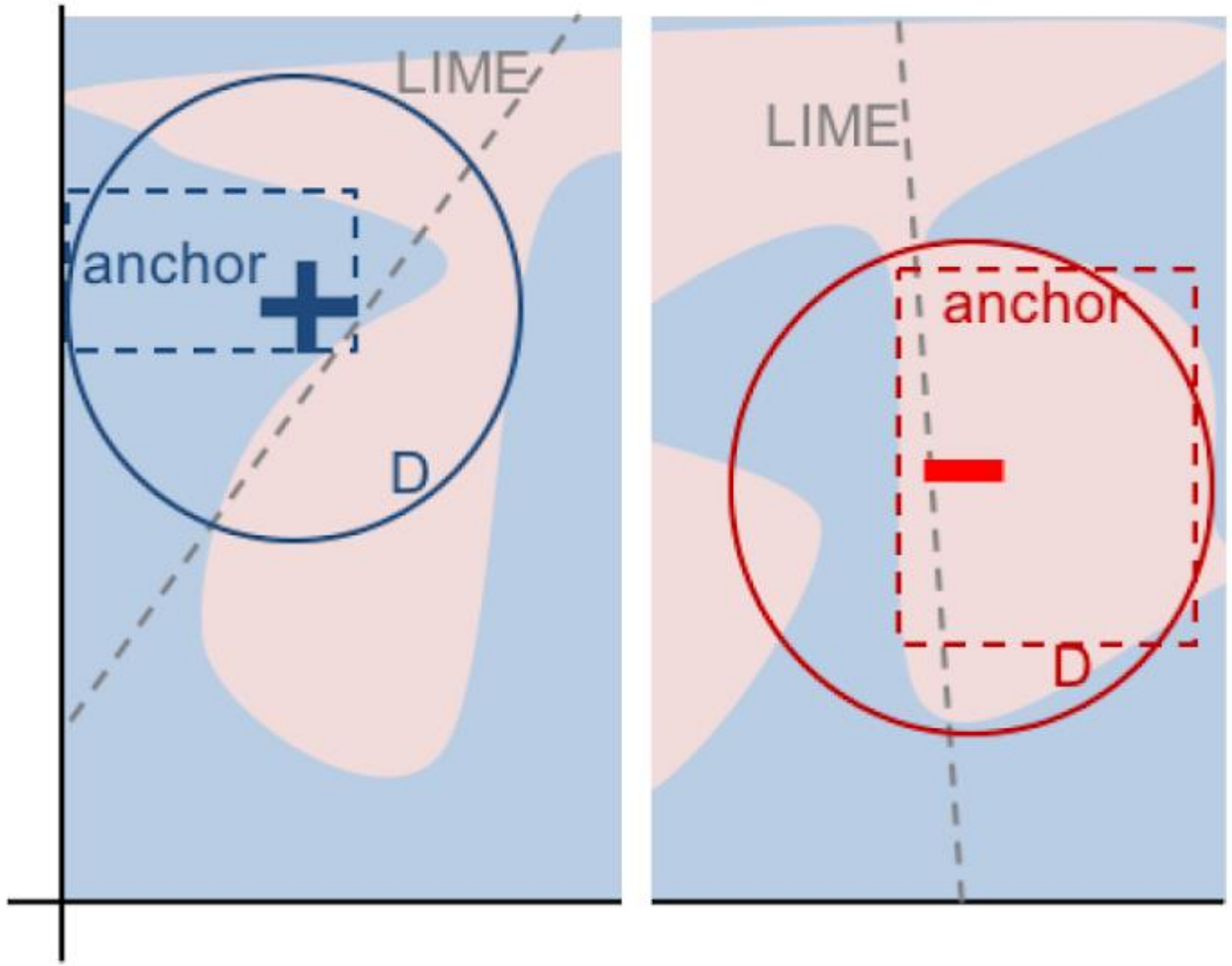
LIME explanations



Anchor explanations

{"not", "bad"} → Positive

{"not", "good"} → Negative


LIME
anchor
D
LIME
anchor
D

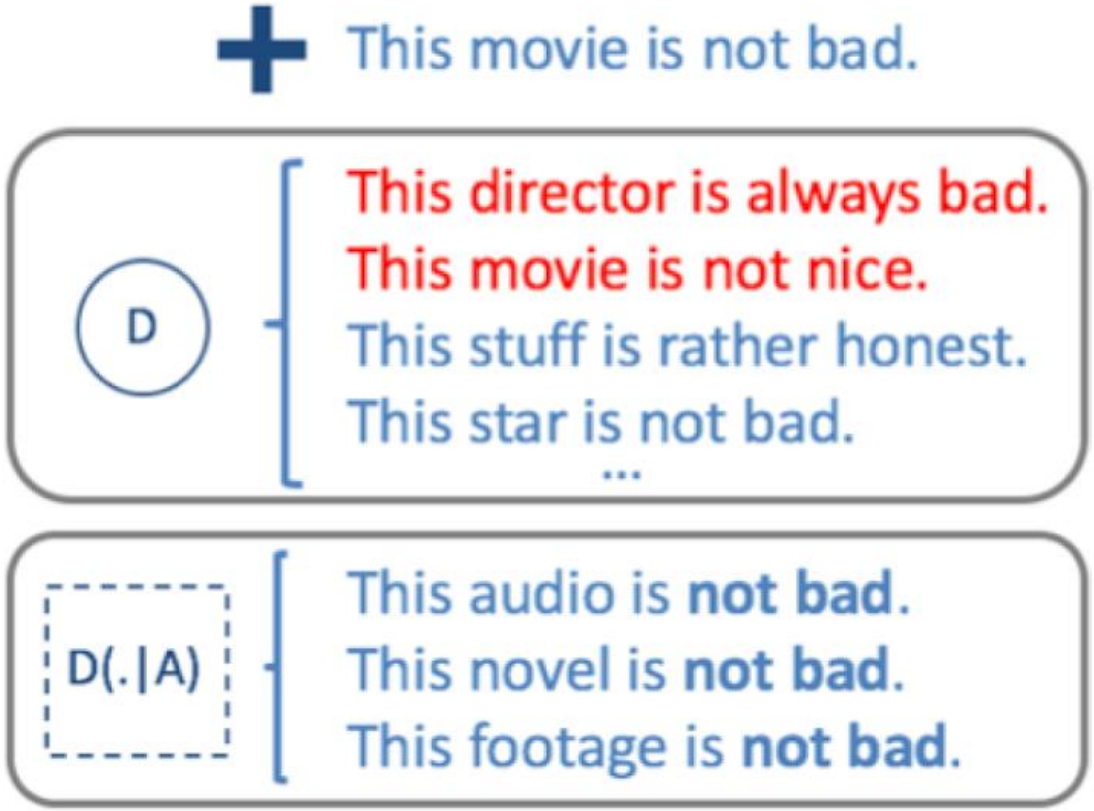
$f : X \to Y$ - black box model

$x \in X$ - an instance to be explained

$\mathcal{D}_x$ - perturbation distribution

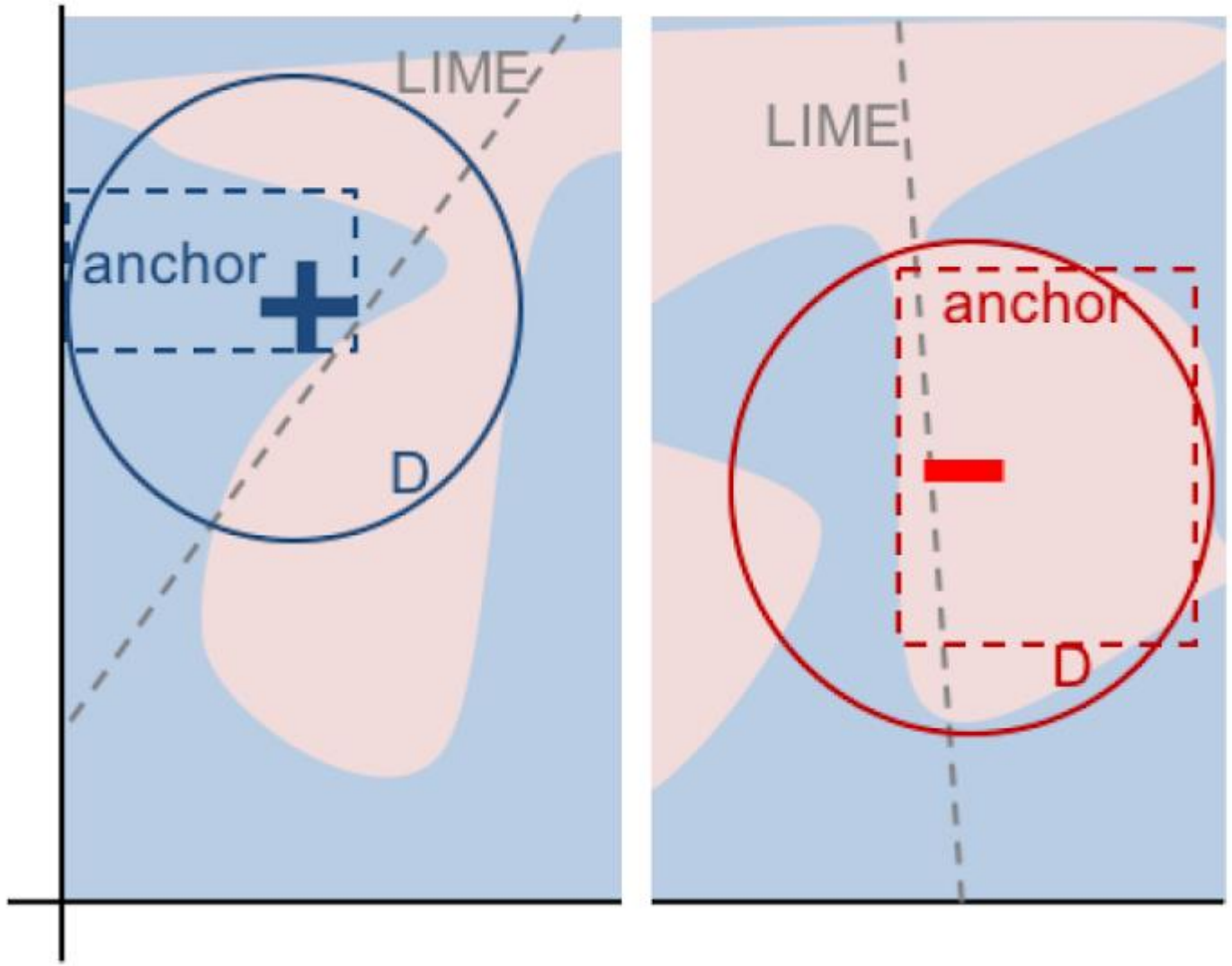
A - a rule (set of predicates)

$\mathcal{D}_x(\cdot|A)$ - conditional distribution when the rule A applies.



A is an anchor if:

$$\mathbb{E}_{\mathcal{D}(z|A)}\left[\mathbb{1}_{f(x)=f(z)}\right] \geq \tau, A(x) = 1.$$


LIME
anchor
D
LIME
anchor
D

An anchor A is a set of feature predicates that achieves $prec(A) \geqslant \tau$, where

$$\text{prec}(A) = \mathbb{E}_{\mathcal{D}(z|A)} \left[\mathbb{1}_{f(x)=f(z)} \right]$$

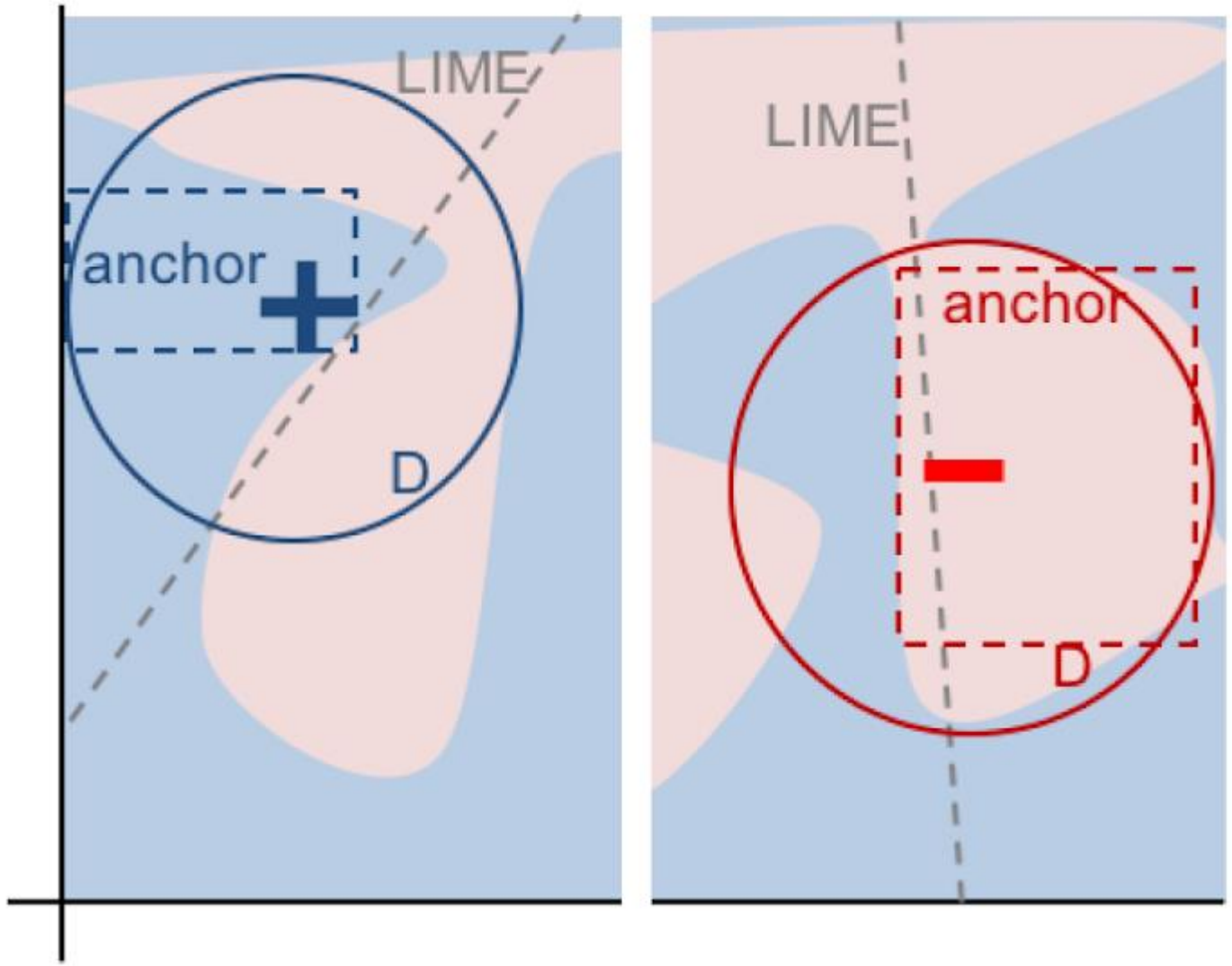
Coverage of an anchor:

$$\text{cov}(A) = \mathbb{E}_{\mathcal{D}(z)}[A(z)]$$

Search for an anchor is the following combinatorial optimization problem:

$$\max_{A \text{ s.t. } prec(A) \geqslant \tau} \text{cov}(A)$$

MI


LIME
anchor
D
LIME
anchor
D

References

M. T. Ribeiro, S. Singh, C. Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier In, 1135-44. ACM Press.

M. T. Ribeiro, S. Singh, C. Guestrin. 2018. Anchors: High-Precision Model-Agnostic Explanations.

M. T. Ribeiro, S. Singh, C. Guestrin. 2016. Nothing Else Matters: Model-Agnostic Explanations By Identifying Prediction Invariance, arXiv:1611.05817.

K. Kulma, Interpretable Machine Learning Using LIME Framework.