Why Should I Trust You?

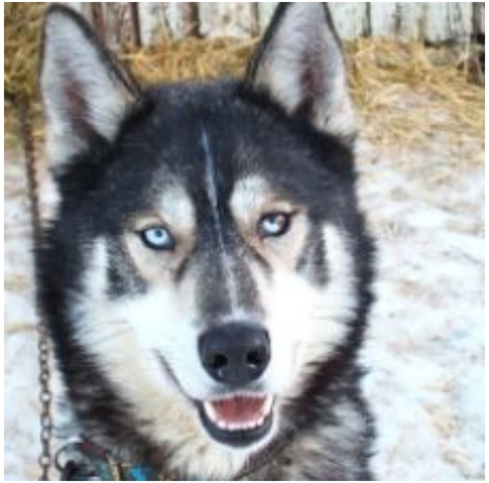
Local **I**nterpretable **M**odel-agnostic **E**xplanations

LIME

> "... machine learning models remain mostly **black boxes**."

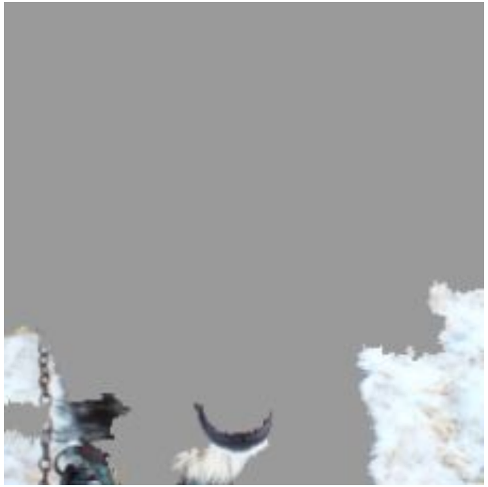
Ribeiro, M , et. al 2016

HUSKY classified as **WOLF**



LIME's Explanation

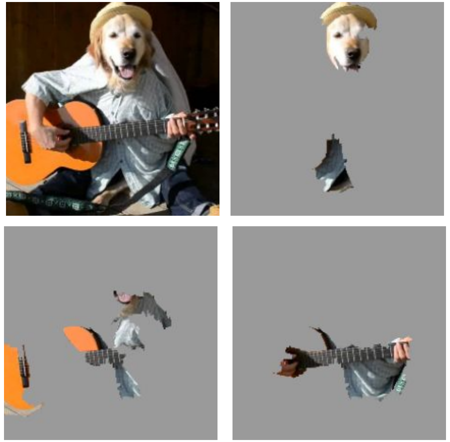
THE WHY



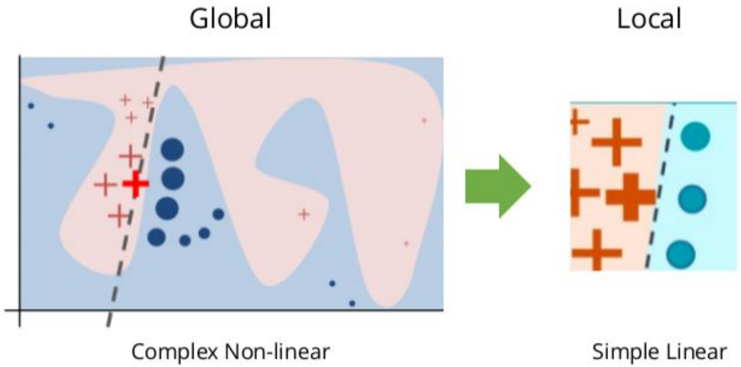
Ribeiro, M , et. al 2016

Explanations' Desiderata

Meaningful

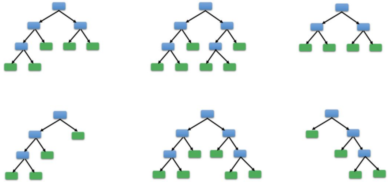


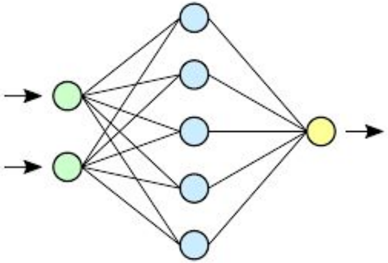
Locally Faithful

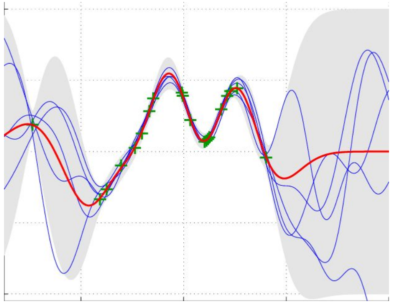


Ribeiro, M , et. al 2016

Model Agnostic

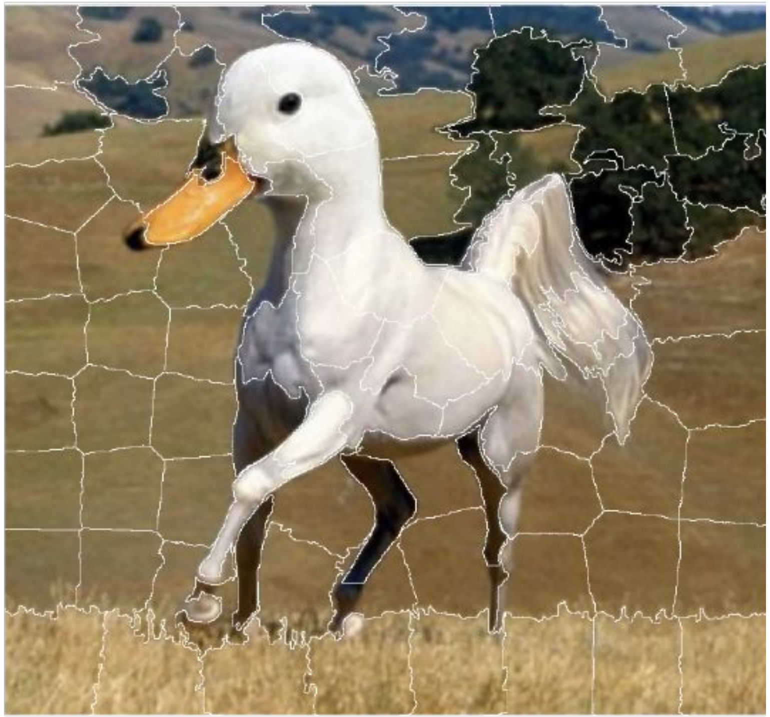






Friendly Data Representation

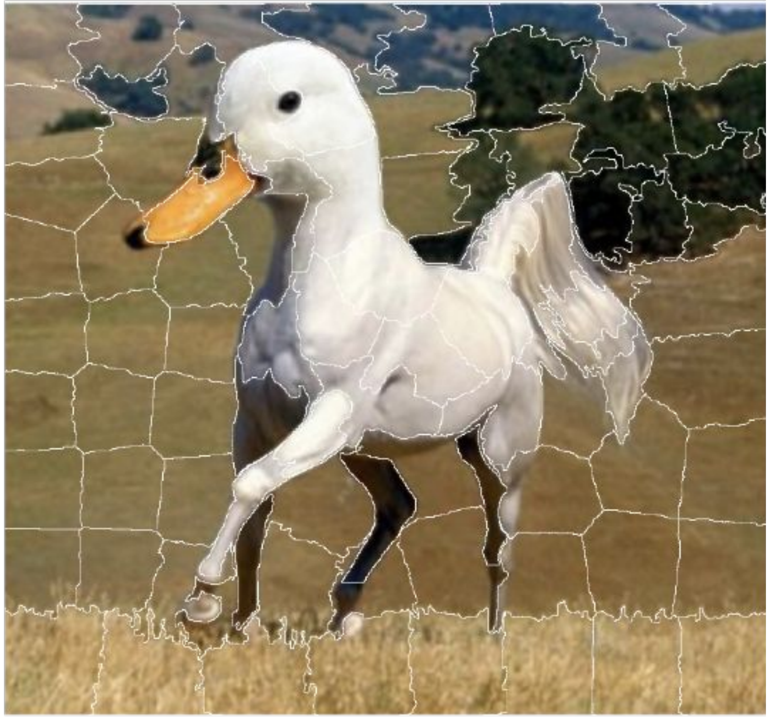
Images Space
m x m x 3 dimensions



SuperPixels Space
[100 dimensions]

Biecek, P and Burzykowski, T 2020

Friendly Data Representation

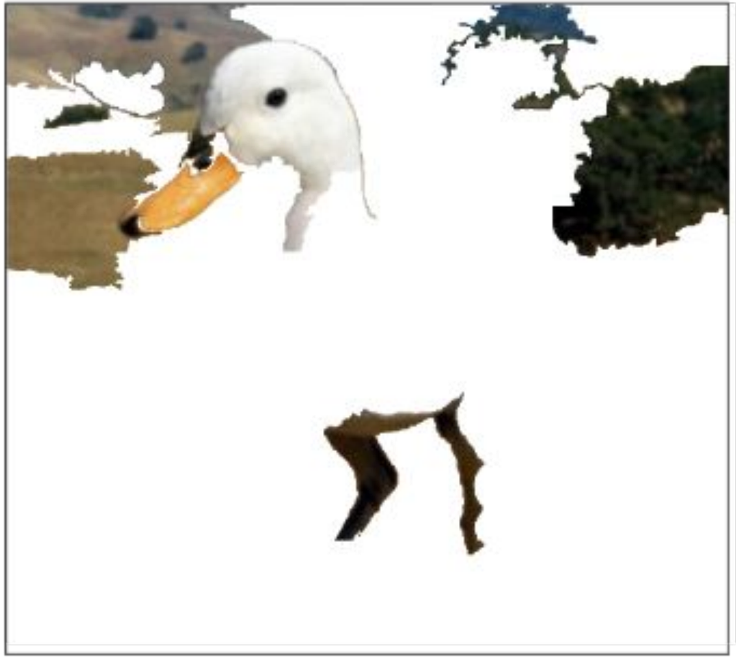


vector

SP	**ON**
1	**1**
2	**1**
3	**1**
...	**1**
98	**1**
99	**1**
100	**1**

Biecek, P and Burzykowski, T 2020

Friendly Data Representation

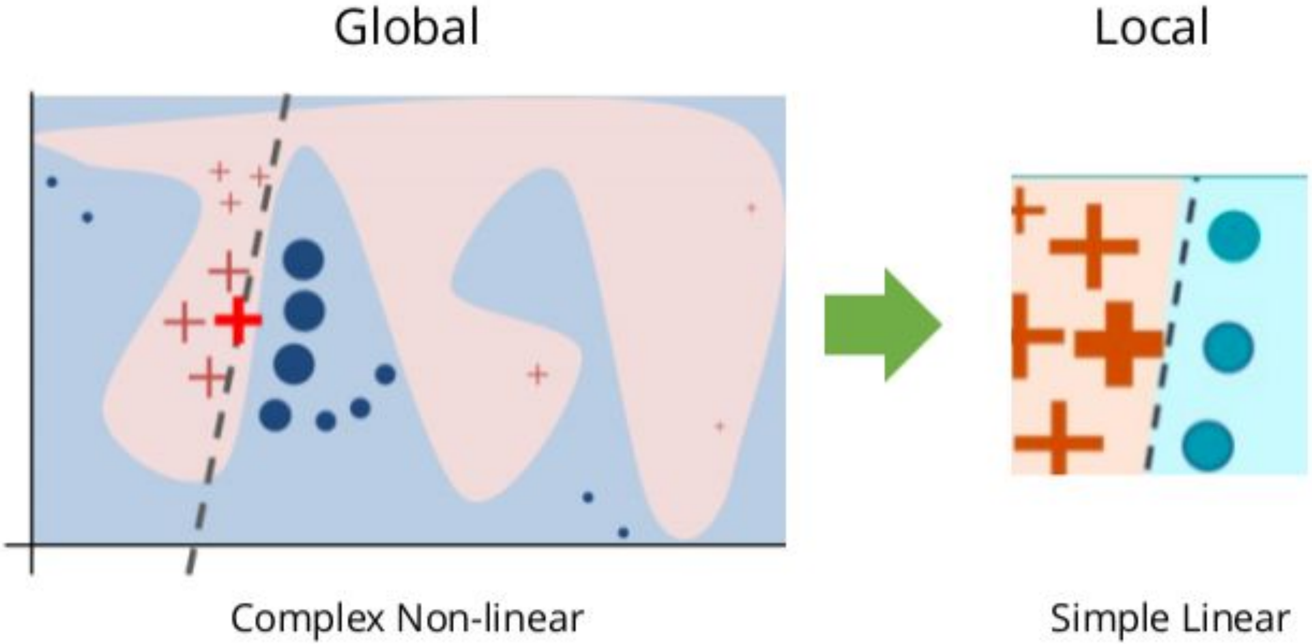


vector

SP	ON
1	**1**
2	**1**
3	**0**
...	**…**
98	**0**
99	**1**
100	**0**

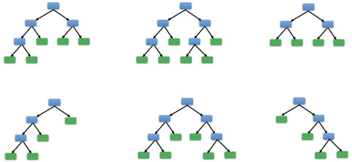
Biecek, P and Burzykowski, T 2020

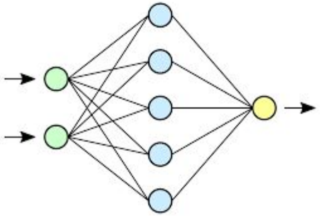
Locally Faithful

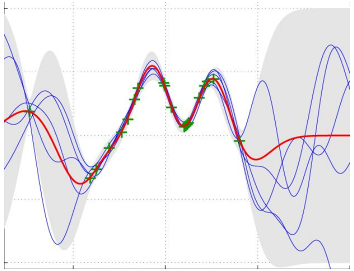


Ribeiro, M , et. al 2016

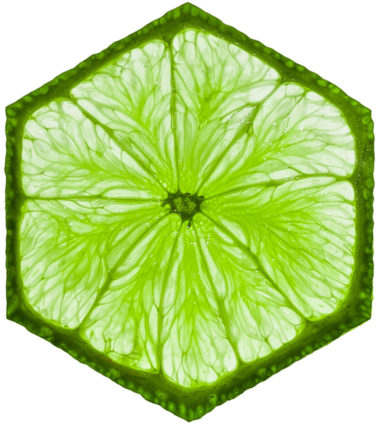
Model Agnostic







EXPLANATIONS



TRUST?

https://github.com/ed-ortizm/**AID-explain-me-why** ?

I guess there is no option…