Explainability
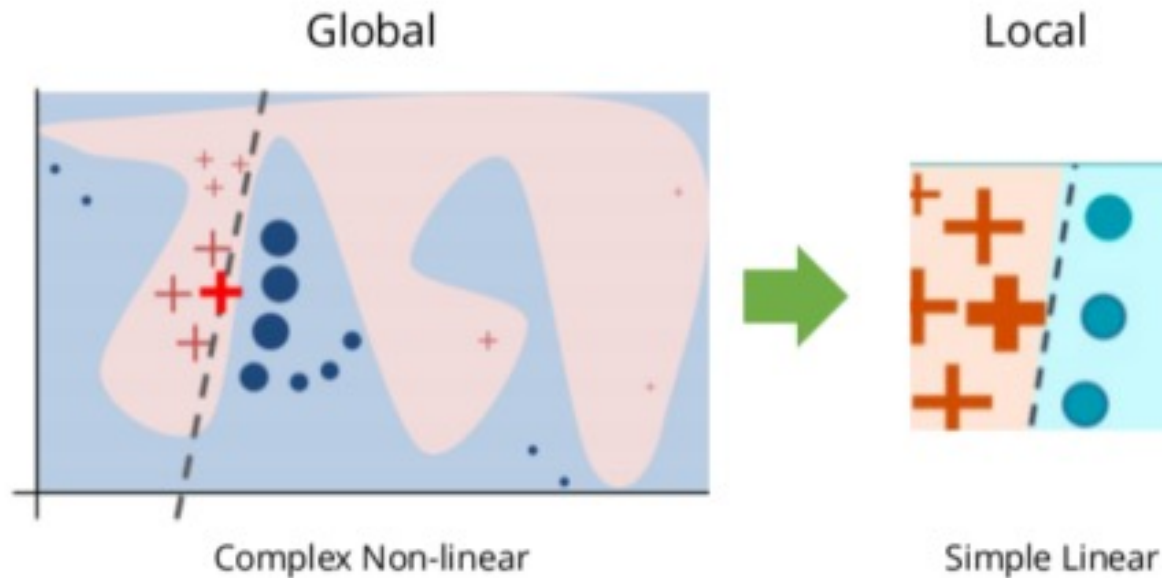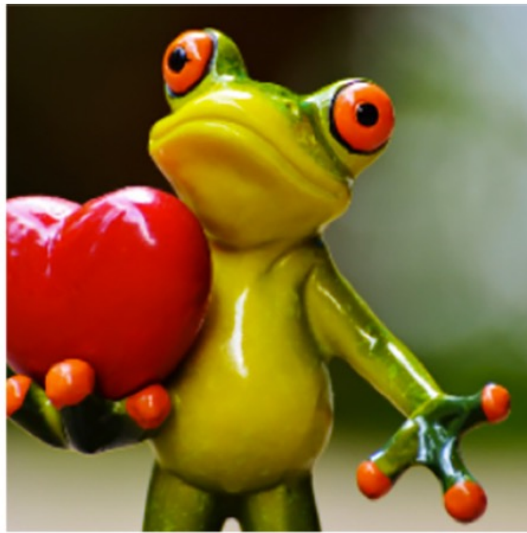
Eric Wong

9/29/2022

Local Linearity

Marco Tulio Ribeiro "Local Interpretable Model-Agnostic Explanations (LIME): An Introduction"
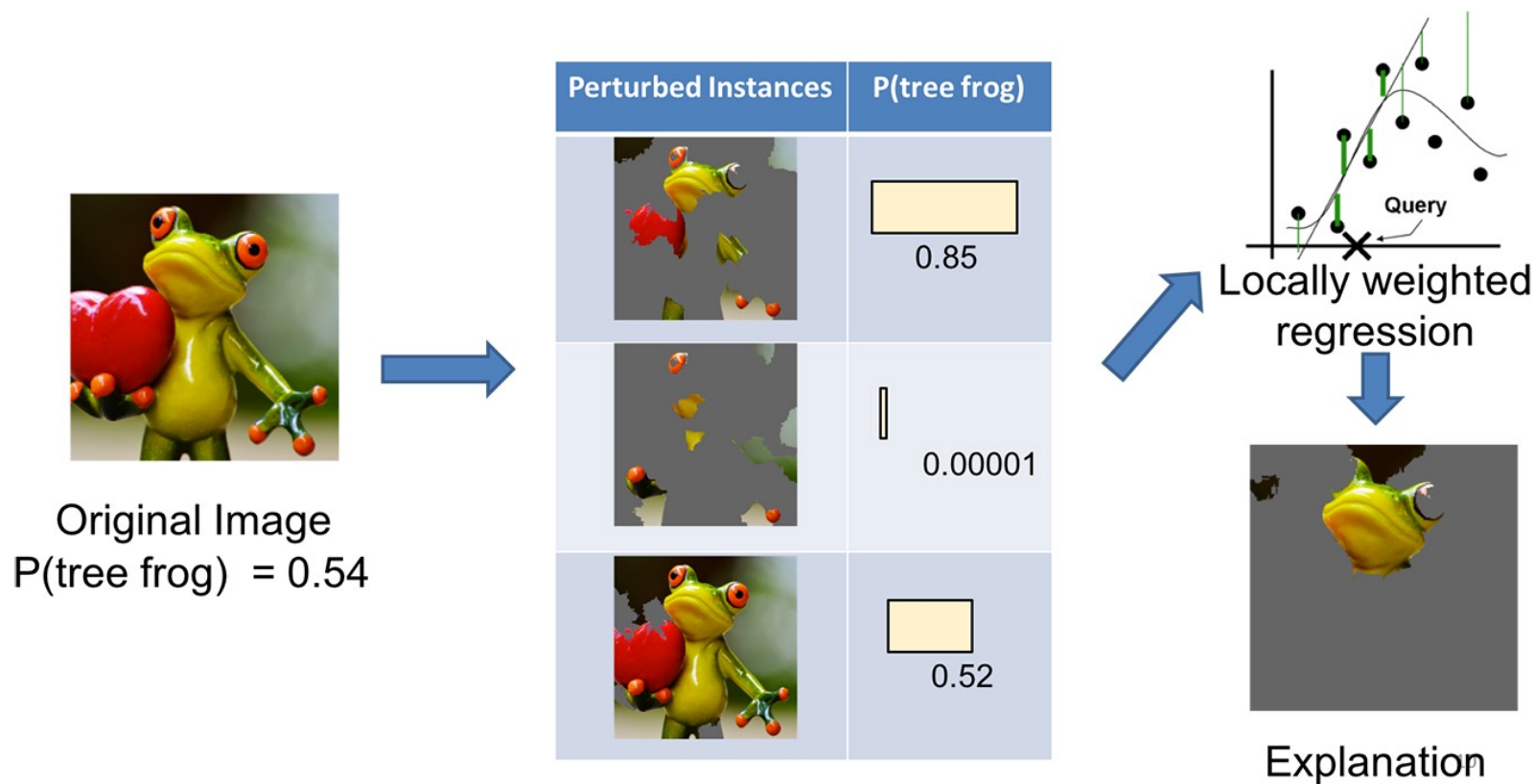
Superpixels for "interpretable" features

Original Image

Interpretable Components

Marco Tulio Ribeiro "Local Interpretable Model-Agnostic Explanations (LIME): An Introduction"

Perturb superpixels

Original Image
P(tree frog) = 0.54

Perturbed Instances	P(tree frog)
	0.85
	0.00001
	0.52

Locally weighted regression

Explanation

Marco Tulio Ribeiro "Local Interpretable Model-Agnostic Explanations (LIME): An Introduction"

Explaining Cat

Marco Tulio Ribeiro "Local Interpretable Model-Agnostic Explanations (LIME): An Introduction"

Explaining words

Marco Tulio Ribeiro "Local Interpretable Model-Agnostic Explanations (LIME): An Introduction"