Rushil Anirudh, Jayaraman J. Thiagarajan, Rahul Sridhar, Peer-Timo Bremer

MARGIN: Uncovering Deep Neural Networks using Graph Signal Analysis

Explainable Machine Learning
14.06.2018

Thorsten Wünsche

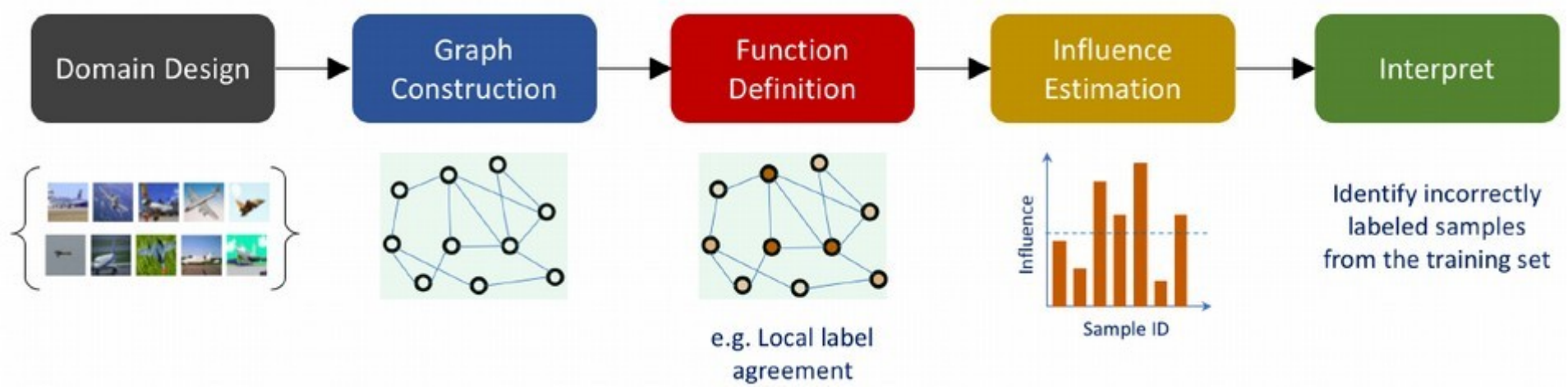
Motivation

- **M**odel **A**nalysis and **R**easoning using **G**raph-based **In**terpretability

- a posteriori interpretability

- Broad applicability

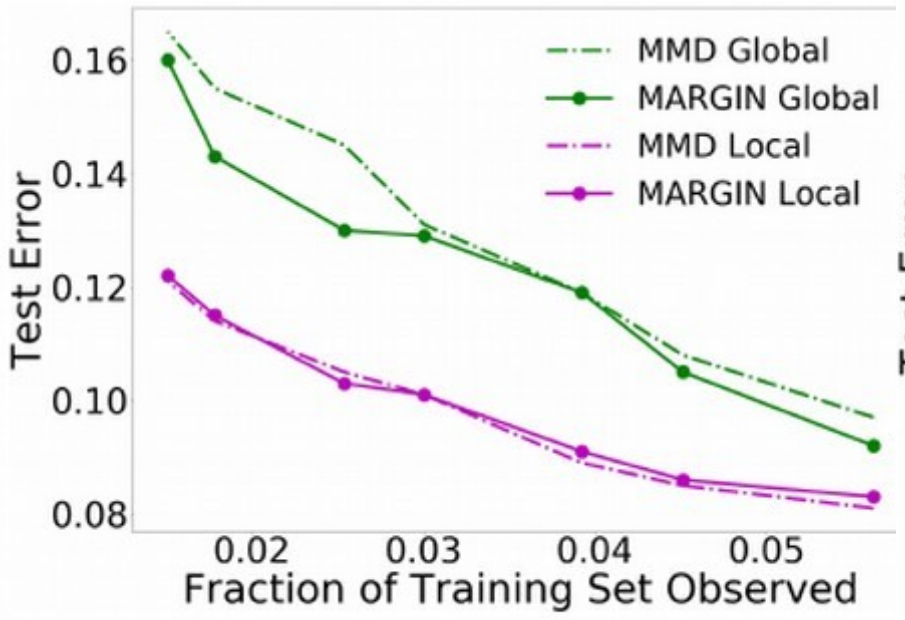
Generic Protocol



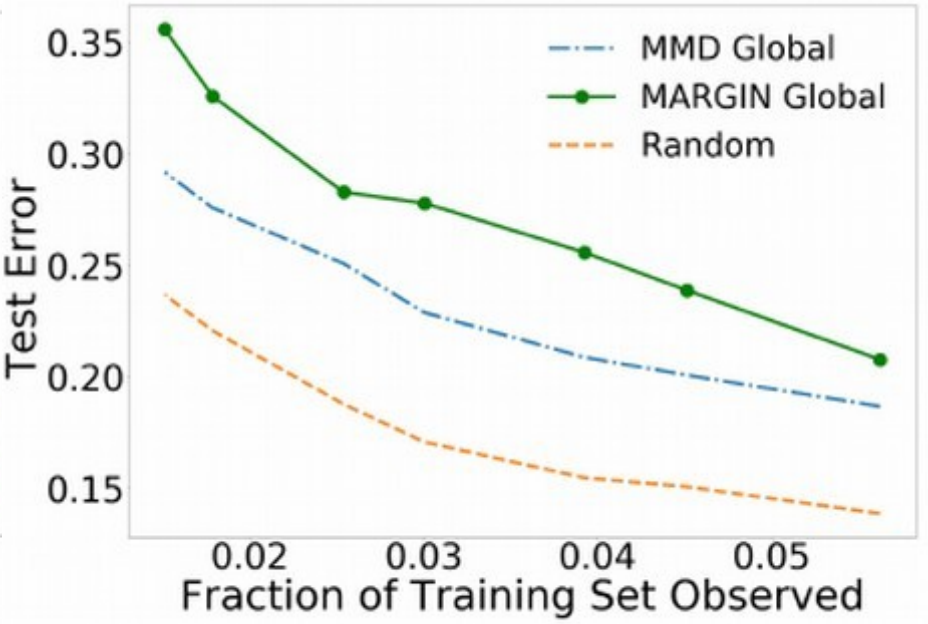
Case Study 1
Prototypes and Criticisms

Domain	Complete Dataset
Nodes	Samples
Function	MMD (Global, Local)
Output	Sample sub-selection

Case Study 1
Prototypes and Criticisms

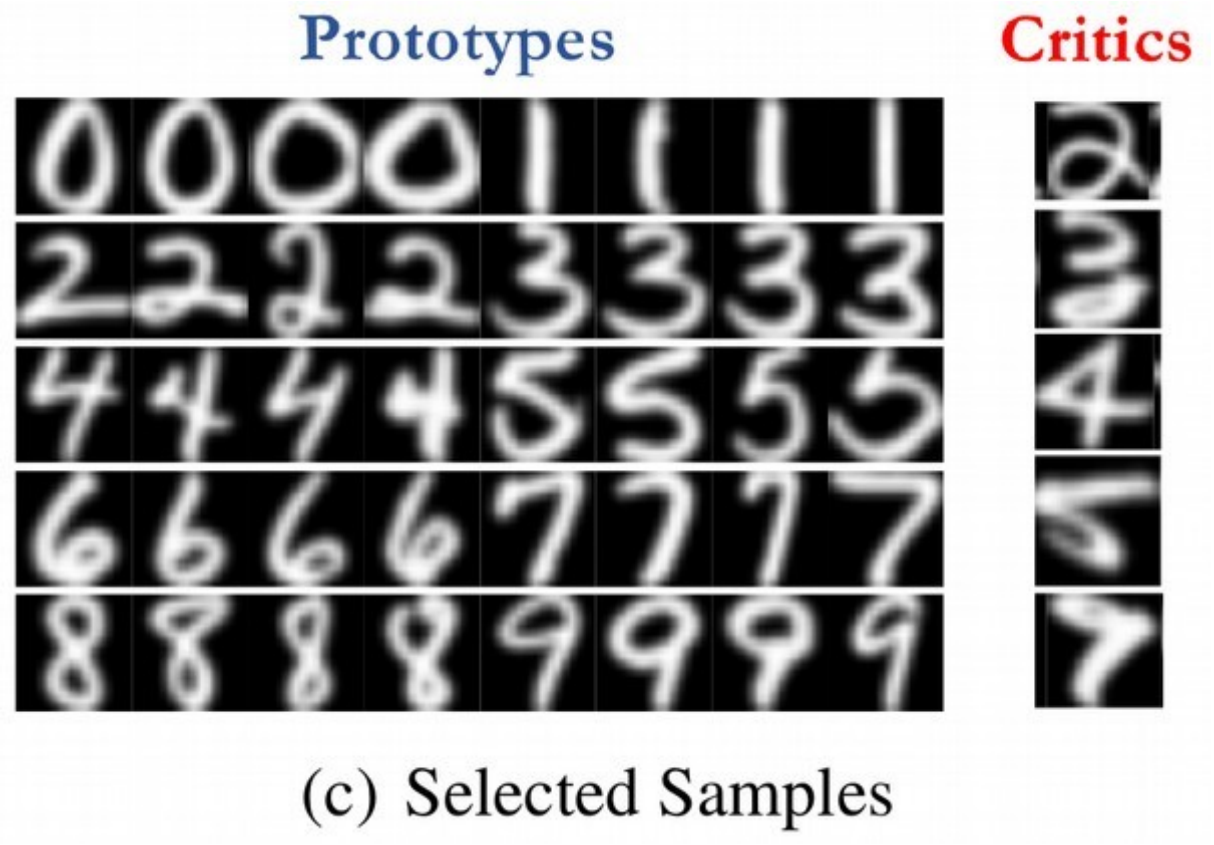


(a) Training with prototypes.



(b) Training with criticisms.

Case Study 1
Prototypes and Criticisms

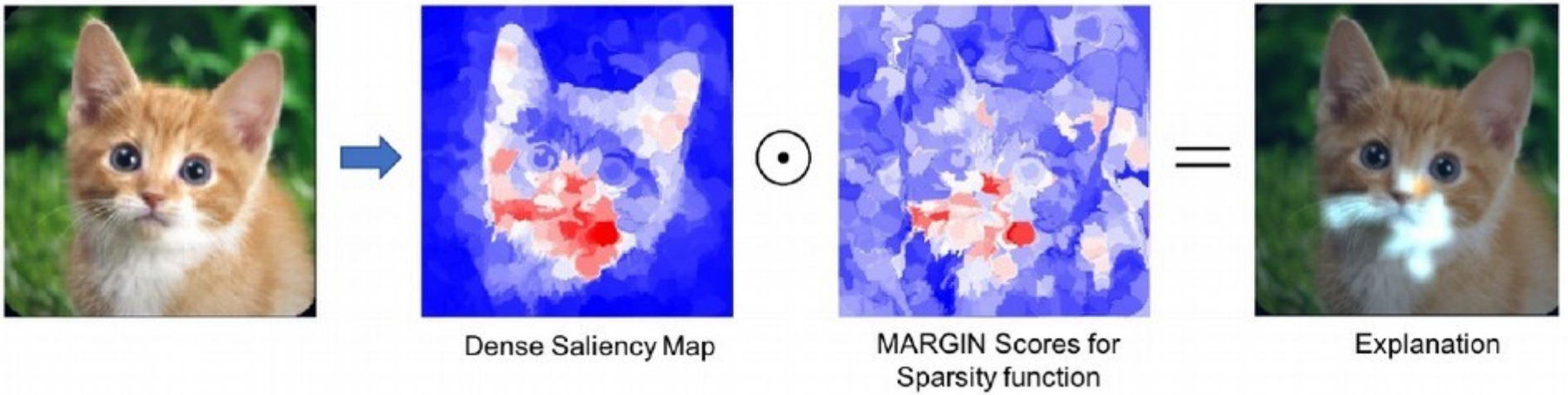


(c) Selected Samples

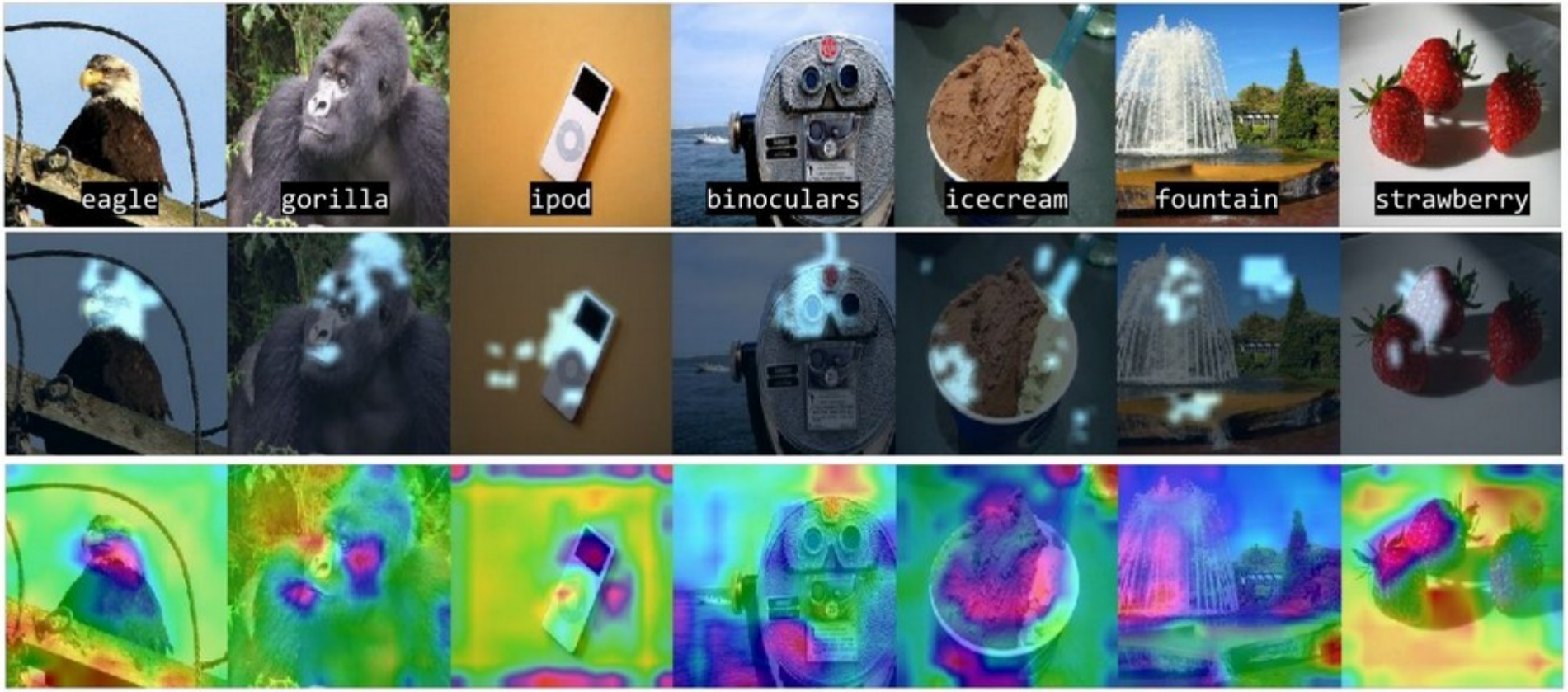
Case Study 2
Explanations for Image Classification

Domain	Single Image
Nodes	Explanations
Function	Sparsity
Output	Saliency maps

Case Study 2
Explanations for Image Classification



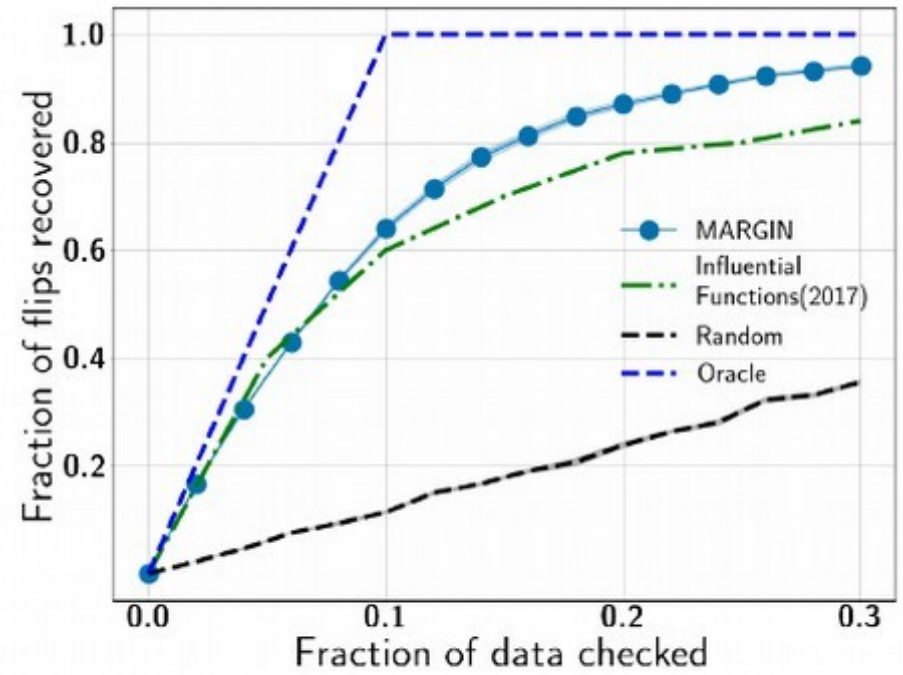
Case Study 2
Explanations for Image Classification



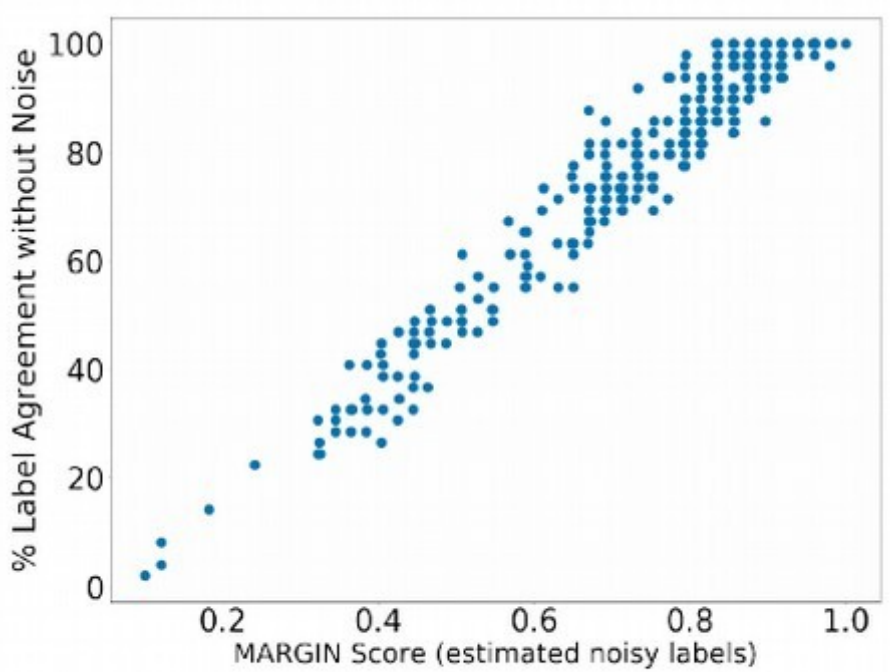
Case study 3
Detecting Incorretly Labeled Samples

Domain	Complete Dataset
Nodes	Samples
Function	Local label agreement
Output	Samples to fix

Case study 3
Detecting Incorretly Labeled Samples



(a) Detecting label flips in the Enron dataset (Metsis et al., 2006).



(b) Examining the incorrectly labeled samples with their influence score.

Case Study 4
Interpreting Decision Boundaries

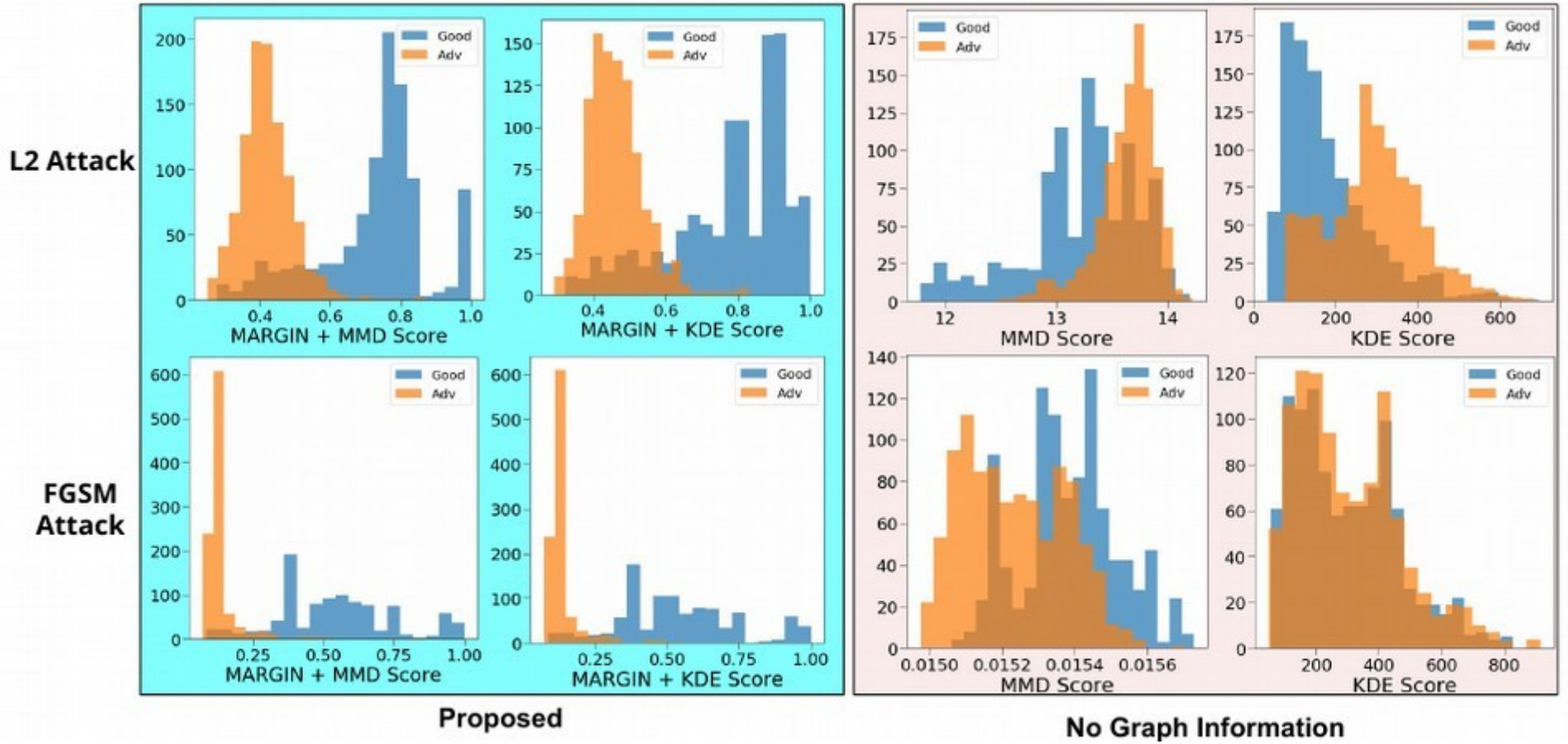
Domain	Complete Dataset
Nodes	Samples
Function	Local label argeement
Output	Confusing samples

Case Study 4
Interpreting Decision Boundaries

Case Study 5
Characterizing Statistics of Adversarial Examples

Domain	Attacks/Noisy samples
Nodes	Noisy Samples
Function	MMD (Global)
Output	Attack statistics

Case Study 5
Characterizing Statistics of Adversarial Examples



Summary

- Fast, flexible approach

- Requires manual selection of graph and function