A Bayesian perspective to Explainable AI

Speaker
Dr. Ranjitha Prasad
ECE department, IIITD
(Joint work with Aditya Saini, Final year Btech, ECE, IIITD)

IIIT D | INDRAPRASTHA INSTITUTE *of* INFORMATION TECHNOLOGY **DELHI**

Expertise

INDIAN INSTITUTE OF SCIENCE
भारतीय विज्ञान संस्थान

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

NUS
National University of Singapore

TATA
TATA ELXSI

TATA
TATA CONSULTANCY SERVICES

Group @ IIITD

3 PhD students (1 on Causal Inference and XAI, 2 others on federated learning)
3 Mtechs working on Causal Inference and
8 BTPs (3 working on XAI)

INDRAPRASTHA INSTITUTE of
INFORMATION TECHNOLOGY
DELHI

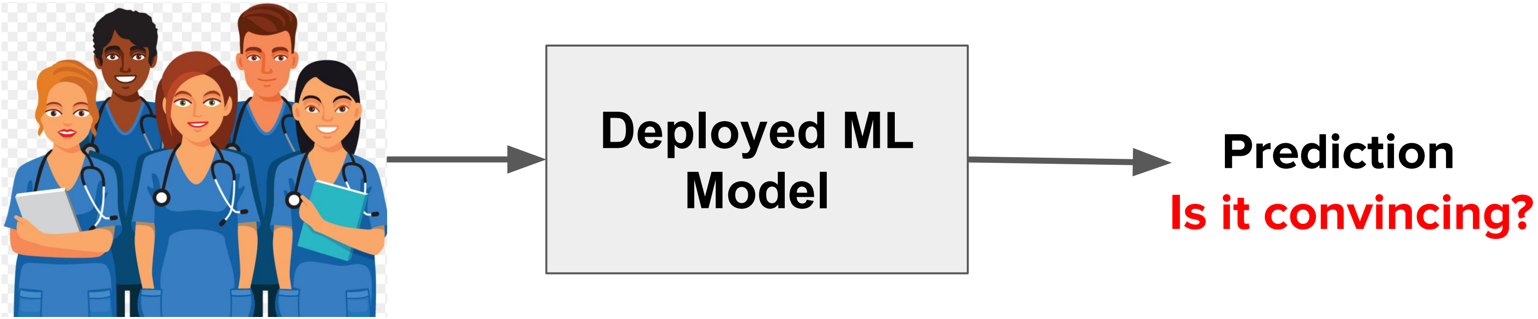
Achievements @ IIITD

Projects: DST SERB, Meity, TiH (XAI)
Students have received Chanakya fellowships for XAI (healthcare project)

Contents

- What is XAI
- Taxonomy and Methods
- Locally Interpretable Model Agnostic methods
  - Literature Survey
  - LIME
  - Proposed Method: UnBox
- Description of Unbox and Results
- Conclusion and Future works

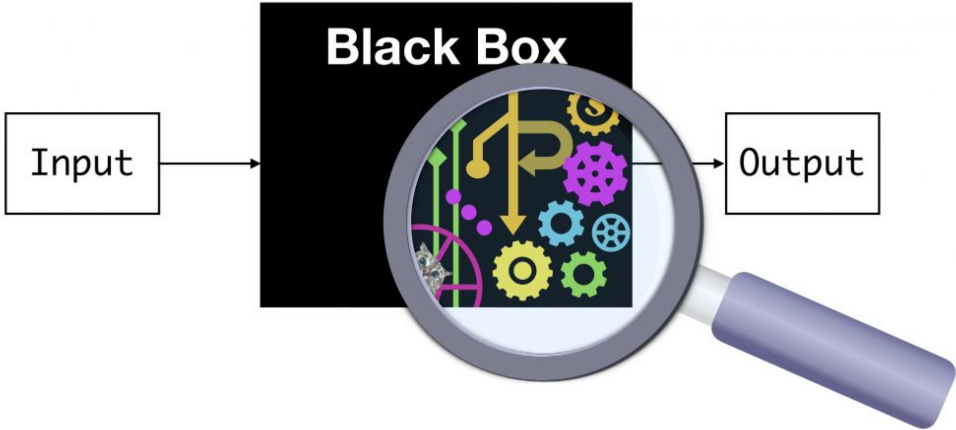
Why do we need Explainable AI



Deployed ML Model

Prediction

Is it convincing?



Input

Black Box

Output

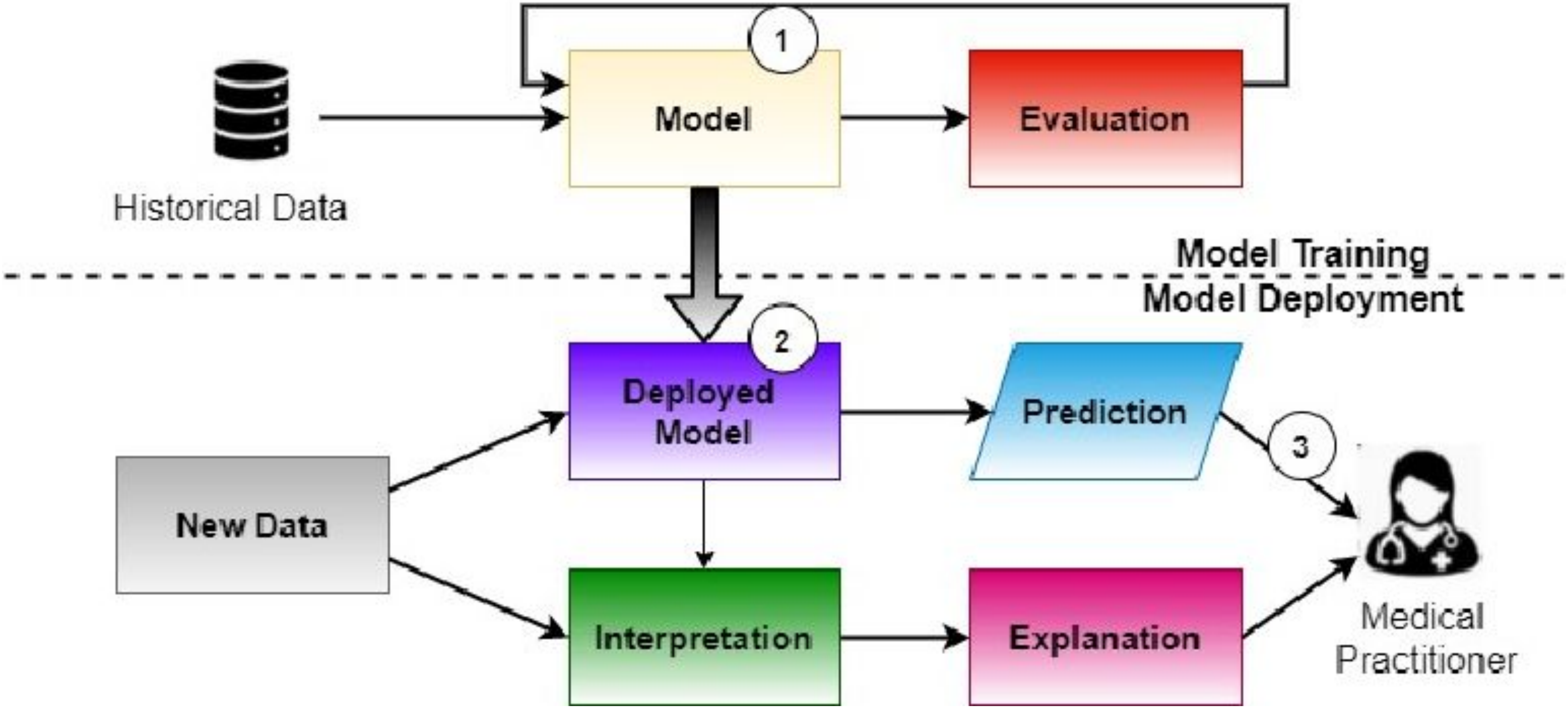
Motivation

- The increasing deployment of artificial intelligence systems in high stakes domains - societal demands for explanations on their predictions.
- New regulations on explanations -
    - Illinois and City of Chicago to implement new laws addressing changes in the workplace — signs of things to come? The National Law Review, June 2019.
    - Equal Credit Opportunity Act (Regulation B of the Code of Federal Regulations), Title 12, Chapter X, Part 1002, §1002.9, creditors are required to notify applicants who are denied credit with specific reasons for the detail.
    - ‘Right to explanation’: European Union General Data Protection Regulation (enacted 2016, taking effect 2018) extends the 1995 Data Protection Directive to provide a legally disputed form of a right to an explanation: "[the data subject should have] the right ... to obtain an explanation of the decision reached".
- Goal: gain insight into the system’s decision-making process - key aspect in fostering trust and confidence in AI systems

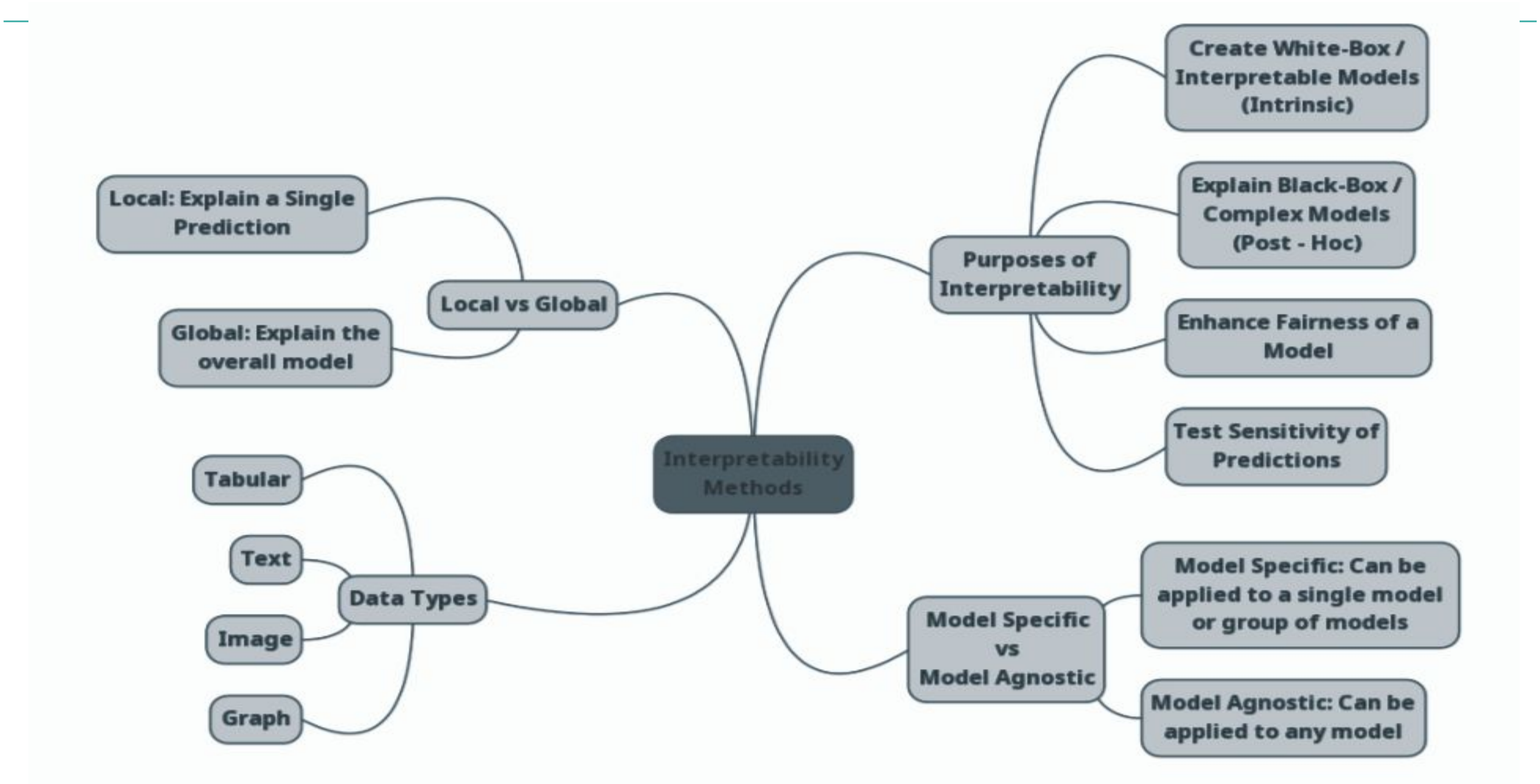
Challenges

- One shoe does not fit all
  - Lack of precise definition of explanation
  - Different users in different settings may require different type of explanations
- Examples:
  - A doctor trying to understand an AI diagnosis of a patient may benefit from seeing known similar cases with the same diagnosis:
    - Data: Image or Tabular
    - Explanation: Conceptual or feature specific
  - A denied loan applicant will want to understand the main reasons for their rejection and what can be done to reverse the decision.
    - Investigate fairness?
    - Text based data
  - A regulator, on the other hand, will want to understand the behavior of the system as a whole to ensure that it complies with the law
    - Legal implications.

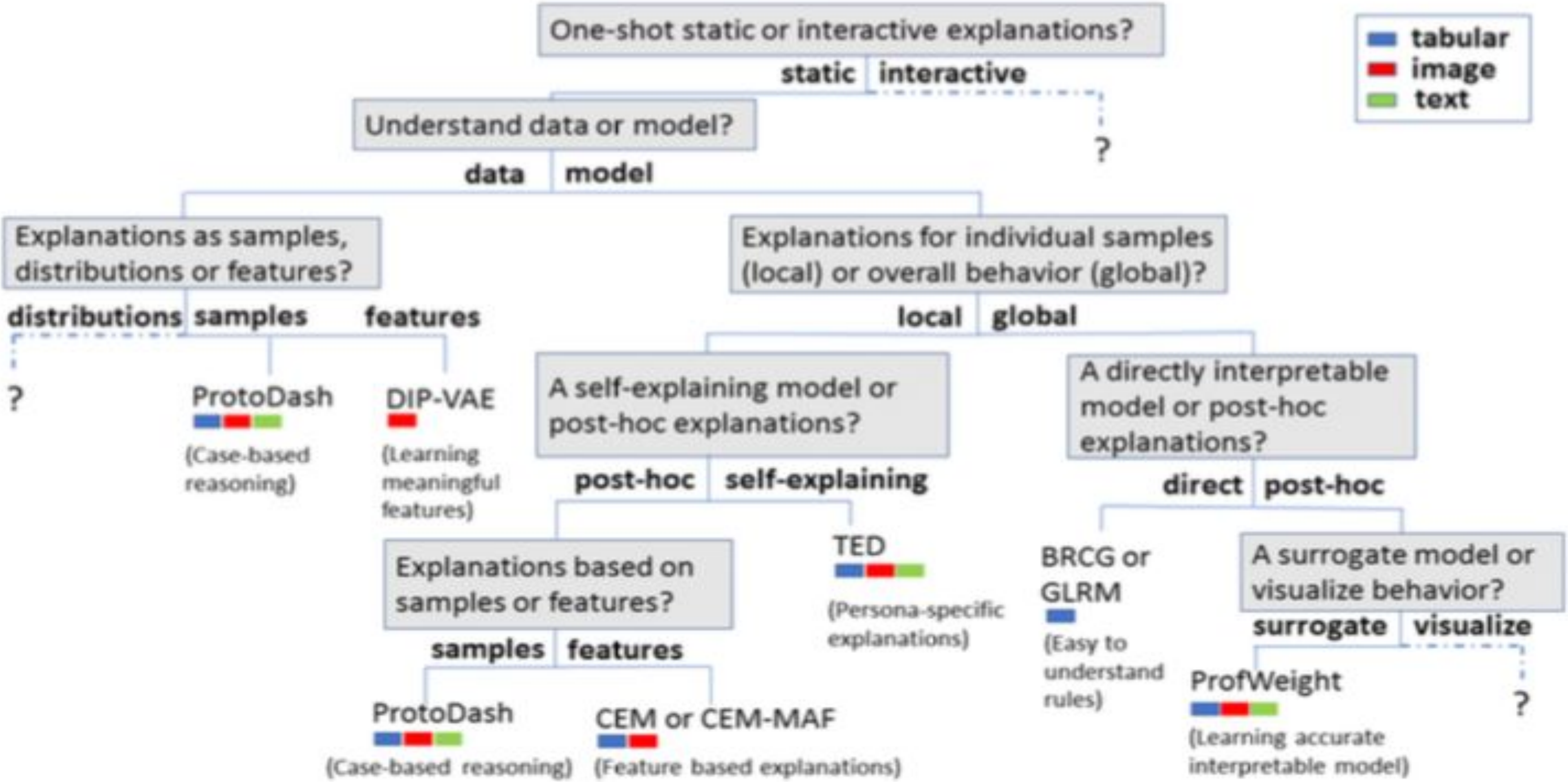
Explainable AI



Taxonomy (P. Linardatos et. al, Entropy, 2021)


Local: Explain a Single Prediction
Local vs Global
Global: Explain the overall model
Tabular
Text
Data Types
Image
Graph
Interpretability Methods
Create White-Box / Interpretable Models (Intrinsic)
Explain Black-Box / Complex Models (Post - Hoc)
Purposes of Interpretability
Enhance Fairness of a Model
Test Sensitivity of Predictions
Model Specific vs Model Agnostic
Model Specific: Can be applied to a single model or group of models
Model Agnostic: Can be applied to any model

Tool-based Taxonomy (V. Arya et. al, Informs)



🟦 tabular
🟥 image
🟩 text

One-shot static or interactive explanations?

static | interactive

?

Understand data or model?

data | model

Explanations as samples, distributions or features?

distributions | samples | features

?

ProtoDash
🟦🟥🟩
(Case-based reasoning)

DIP-VAE
🟥
(Learning meaningful features)

Explanations for individual samples (local) or overall behavior (global)?

local | global

A self-explaining model or post-hoc explanations?

post-hoc | self-explaining

Explanations based on samples or features?

samples | features

ProtoDash
🟦🟥🟩
(Case-based reasoning)

CEM or CEM-MAF
🟦🟥
(Feature based explanations)

TED
🟦🟥🟩
(Persona-specific explanations)

A directly interpretable model or post-hoc explanations?

direct | post-hoc

BRCG or GLRM
🟦
(Easy to understand rules)

A surrogate model or visualize behavior?

surrogate | visualize

ProfWeight
🟦🟥🟩
(Learning accurate interpretable model)

?

Taxonomy of XAI techniques

Existing methods for feature importance estimation can be subdivided into 4 groups -

1. Gradient based methods

Eg: Simple Gradient, Integrated Gradients, DeepLIFT, DeepSHAP

- Applicable for differentiable models like neural nets (uses gradient descent and backpropagation)
- Pretty fast while doing implementation
- Require a change in the underlying structure which doesn't make it sustainable in the long run

2. Mimic models

- The idea is to train interpretable models that mimic the decisions of a black-box model that we wish to explain
- Eg : Using decision tree to approximate black-box models
- The issue is that mimic models are not guaranteed to match the behavior of the original model

3. Causal explanation based methods

Eg: CXPlain

- The idea is to use a causal objective to train a supervised model to learn to explain another machine-learning model.
- Model agnostic!
- Outputs *uncertainity estimates*
- Poor API with cluttered codebase leading to subpar results

4. Sensitivity analysis methods

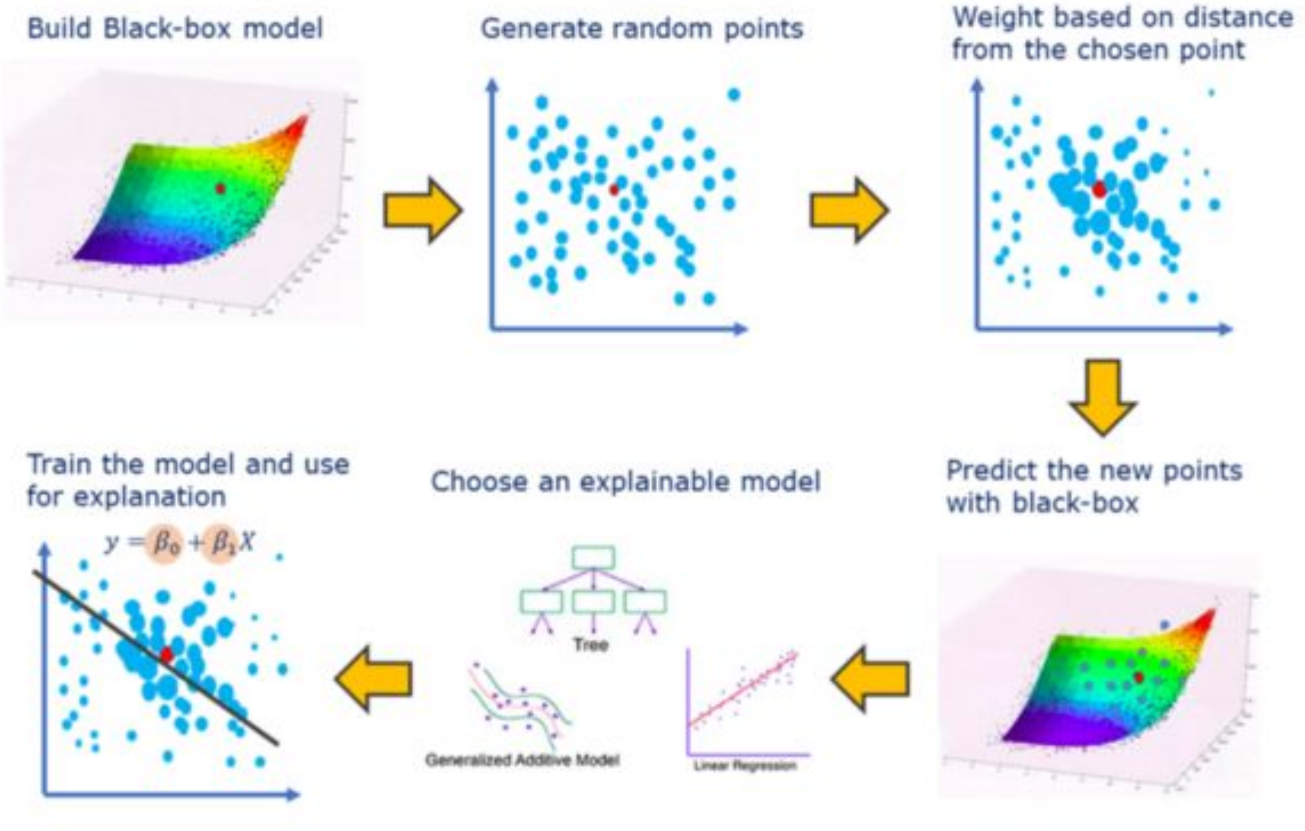
Eg: LIME, SHAP, Kernel SHAP

- Quantify a model's sensitivity to changes in the input
- Applicable to any ML model - **Model agnostic!**
- Provide locally faithful explanations around a given observational feature set
- Simple to implement
- Have a really rich & stable API with a sustainable codebase
- **Slow for high-dimensional datasets**

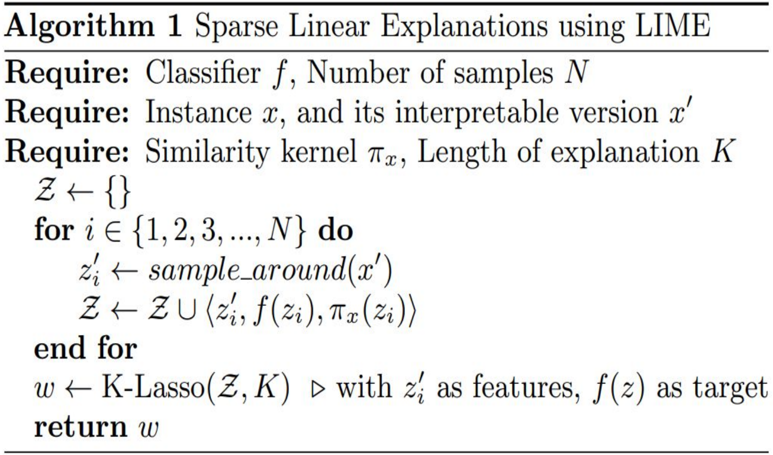
Locally Interpretable Model Agnostic Explanations

- LIME - M.T. Ribeiro et. al, KDD 2017 - one of the most popular interpretability methods for black-box models
- Approach: Explain any single (given) instance and its corresponding prediction.
- How?
  - For any given instance and its corresponding prediction, simulated randomly-sampled data around the neighbourhood of input instance
  - Generate new predictions surrogate instances and weigh them by their proximity to the input instance.
  - A simple, interpretable model, such as a decision tree, is trained on this newly-created dataset of perturbed instances.
  - By interpreting this local model, the initial black box model is consequently interpreted.

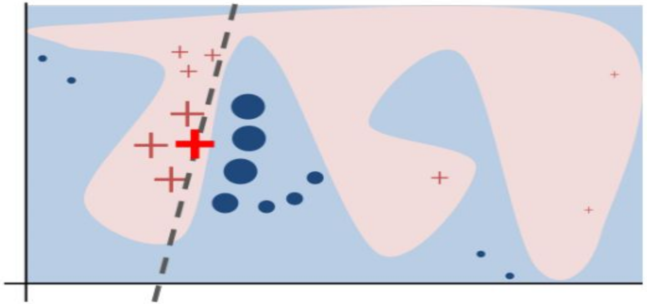
Locally Interpretable Model Agnostic Methods


Build Black-box model
Generate random points
Weight based on distance from the chosen point
Predict the new points with black-box
Choose an explainable model
Tree
Generalized Additive Model
Linear Regression
Train the model and use for explanation
$y = \beta_0 + \beta_1 X$

LIME Workflow



```
Algorithm 1 Sparse Linear Explanations using LIME
Require: Classifier $f$, Number of samples $N$
Require: Instance $x$, and its interpretable version $x'$
Require: Similarity kernel $\pi_x$, Length of explanation $K$
  $\mathcal{Z} \leftarrow \{\}$
  for $i \in \{1, 2, 3, ..., N\}$ do
    $z'_i \leftarrow sample\_around(x')$
    $\mathcal{Z} \leftarrow \mathcal{Z} \cup \langle z'_i, f(z_i), \pi_x(z_i) \rangle$
  end for
  $w \leftarrow$ K-Lasso$(\mathcal{Z}, K)$  ▷ with $z'_i$ as features, $f(z)$ as target
  return $w$
```



LIME samples instances, gets predictions using original predictive function, and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful.

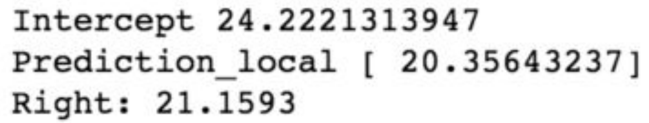
LIME Results: Boston Housing Data

Boston Housing: **Dataset derived from information collected by the U.S. Census Service concerning housing in the area of Boston**

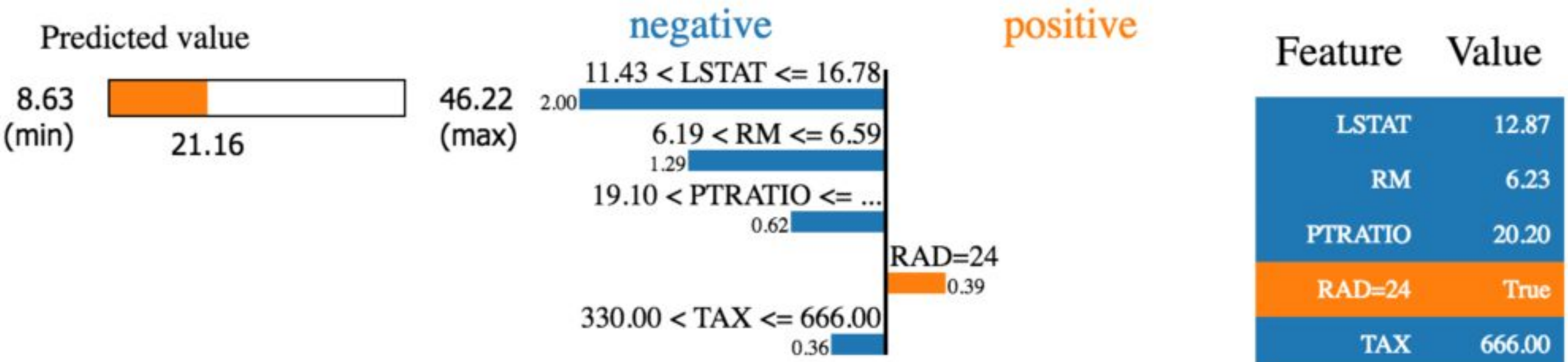
LSTAT: Percentage of lower status of the population.

RM: Average number of rooms per dwelling, **PTRATIO**: Pupil-teacher ratio by town

RAD: Index of accessibility to radial highways, **TAX**: Full-value property tax rate

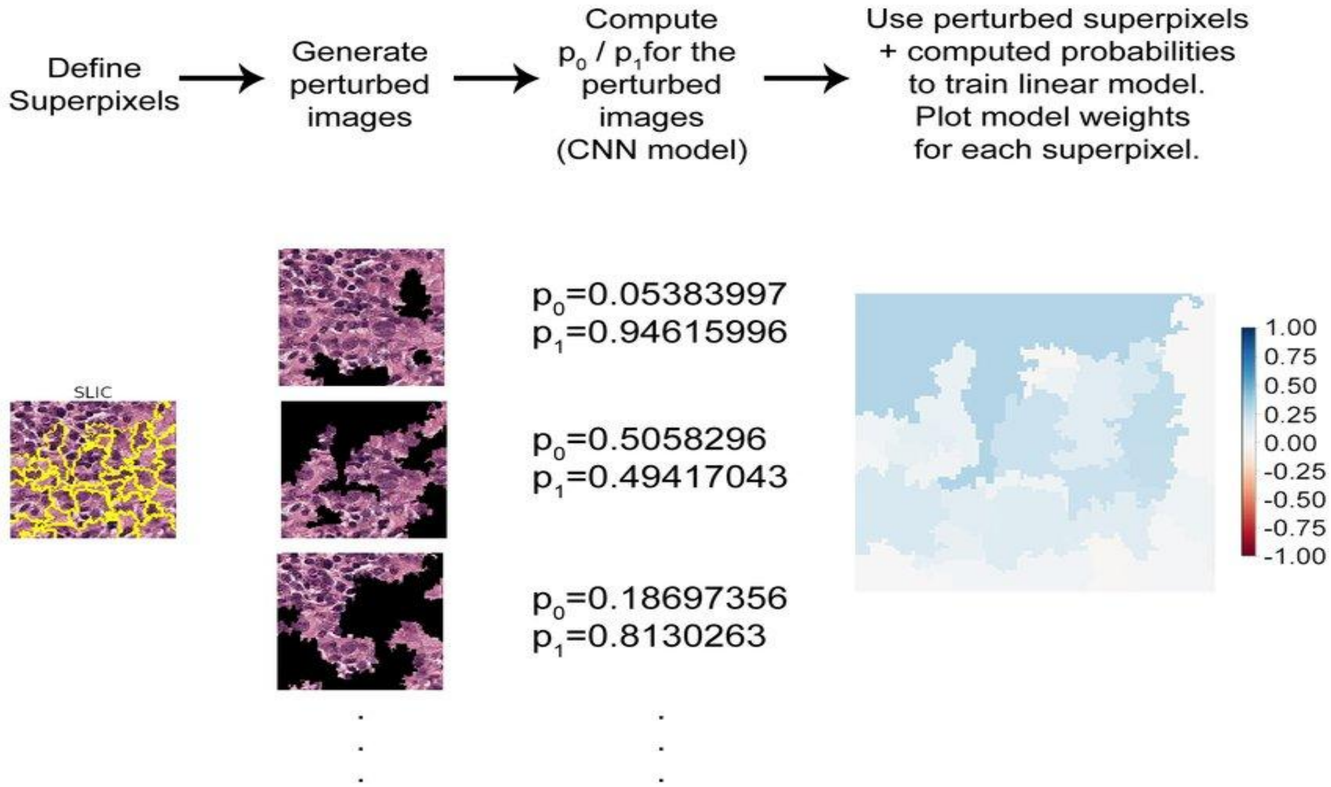


```
Intercept 24.2221313947
Prediction_local [ 20.35643237]
Right: 21.1593
```



Feature	Value
LSTAT	12.87
RM	6.23
PTRATIO	20.20
RAD=24	True
TAX	666.00

LIME Explanation On Images



Define Superpixels → Generate perturbed images → Compute p_0 / p_1 for the perturbed images (CNN model) → Use perturbed superpixels + computed probabilities to train linear model. Plot model weights for each superpixel.

p_0=0.05383997
p_1=0.94615996

p_0=0.5058296
p_1=0.49417043

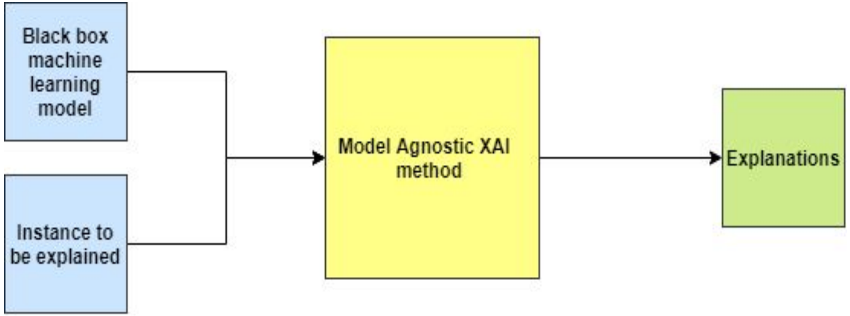
p_0=0.18697356
p_1=0.8130263

Post-hoc Explainers

Technique name	Strategy used	Issues
DLIME	Uses a deterministic clustering algorithm for creating surrogate dataset	In presence of less training points, the model gives bad approximation of the underlying function
BayLIME	By using a Bayesian modification of LIME, it incorporates prior knowledge of a given sample to remove inconsistency for similar samples	Finding useful priors is nuanced and difficult for each unique problem
ALIME	Uses an auto encoder based approach for weighing the generated samples to get better accuracy	The complex structure counters itself as explaining ALIME's decision becomes another XAI task
OptiLIME	Uses a Bayesian optimized neighbourhood for providing consistent explanations	Explainer model same as LIME

Summarizing ...

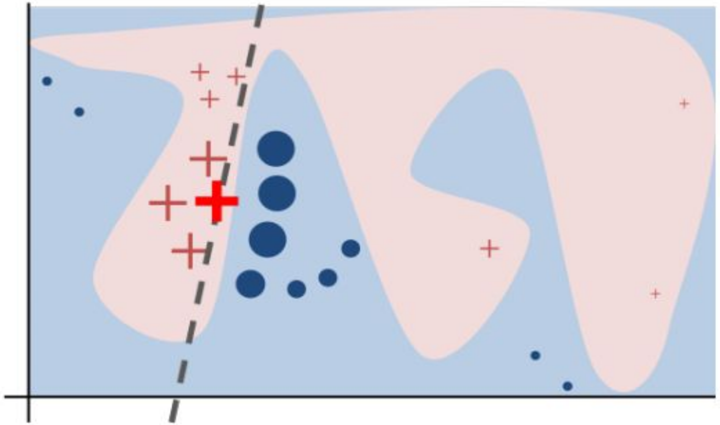
- From the literature survey, we found important properties which were paramount to the success of any explainable AI model:
  a. **Model agnosticism:** Explain the decisions of any **black box** ML model, without using training data
  b. **Local Fidelity:** Ability to give explanations pertaining to the neighborhood surrounding the instance
  c. **Stability**: Consistent explanations over several runs of explainer


Black box machine learning model

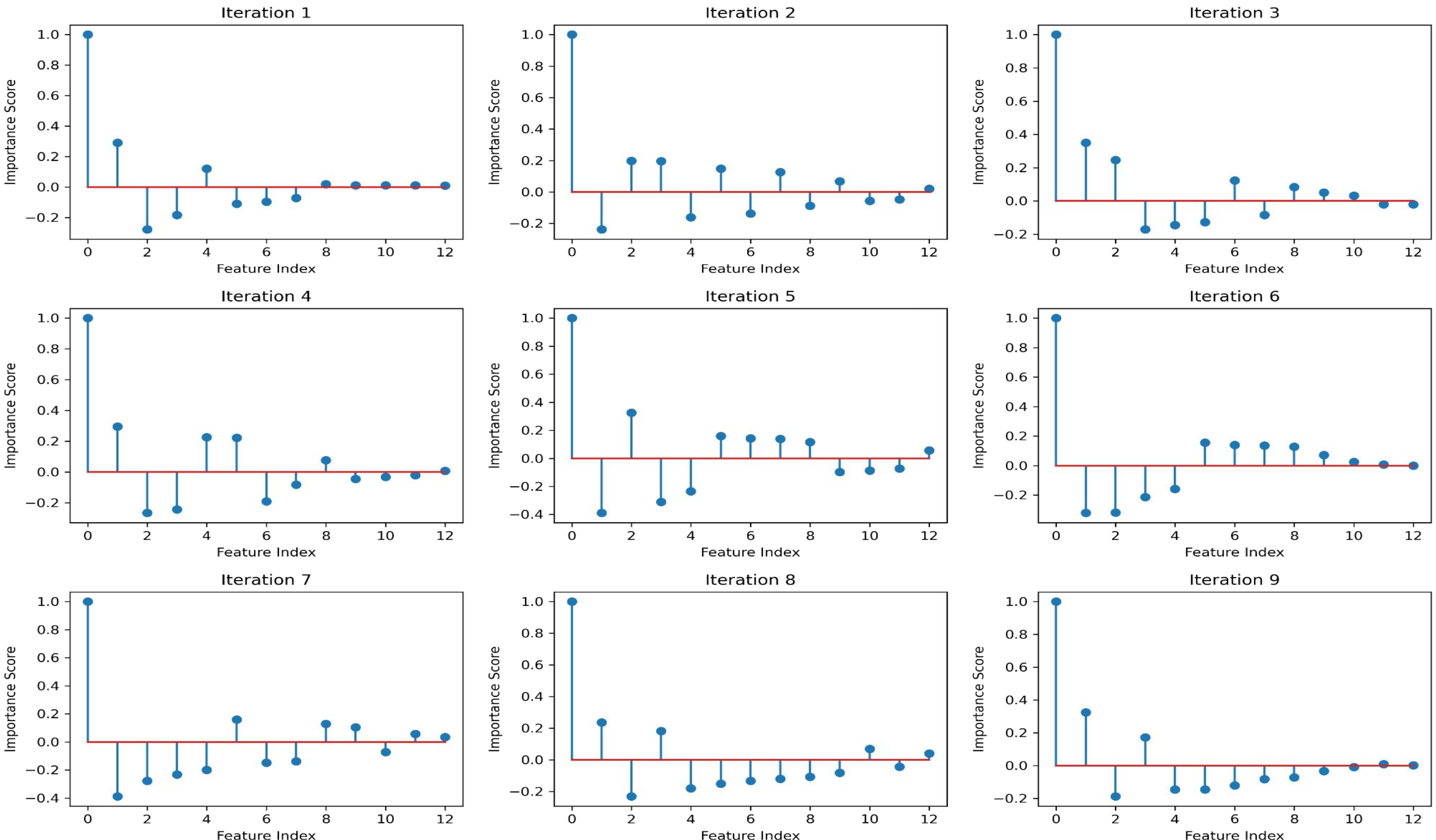
Instance to be explained

Model Agnostic XAI method

Explanations



Stability plots - LIME(Kendall’s W = 0.2), Boston



UnBOX: Uncertainty Aware Bayesian Optimization driven Explanations

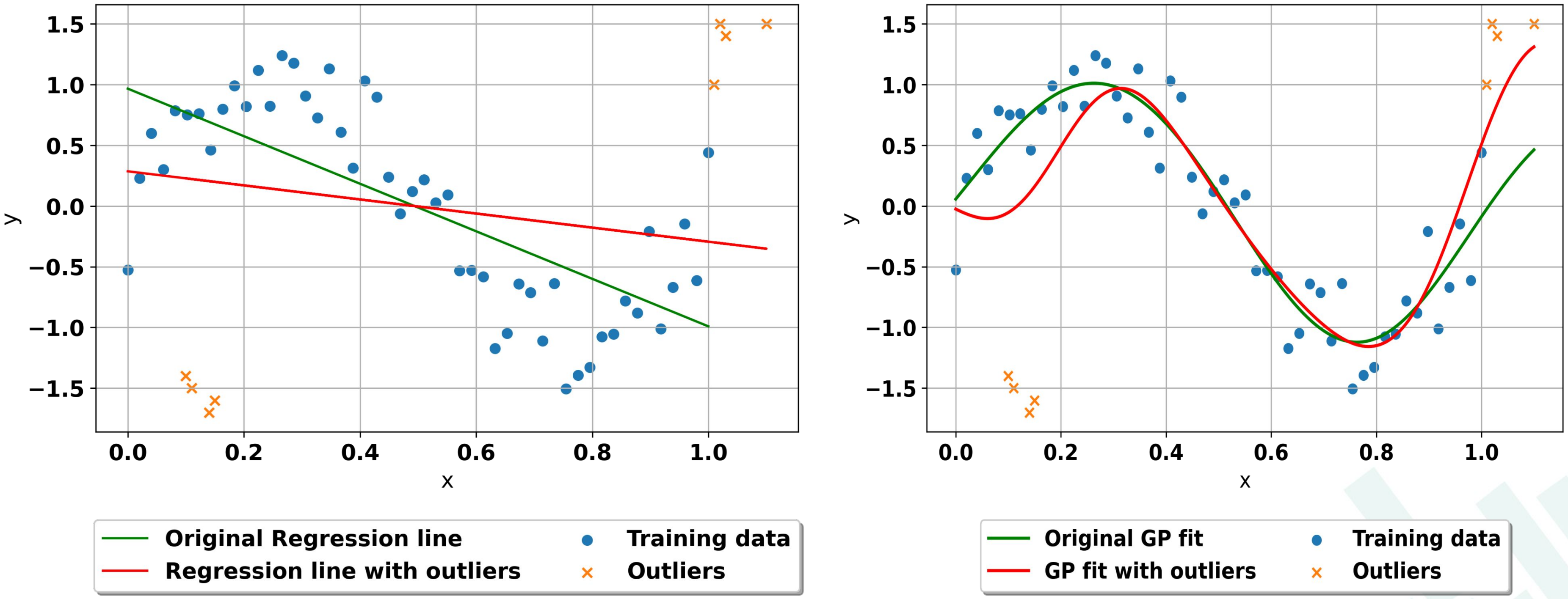
Motivation

- Solve the stability issue in LIME
  - **Random sampling** of points in the neighborhood
- LIME uses a very simple explainer - linear models have **limited expressibility**
- Linear models used in LIME are prone to outliers
  - Problematic if we are constrained to consider few samples (limited sample complexity) due to latency/complexity of prediction models
- No specific relationship between the explainer and the sampling modules.
  - Does such a relationship help?
  - Is it useful to assess a sampled point and its explanation, and then use that assessment to choose the next point?
  - In other words, should the sampling be sequential or batch-based?

Contributions

- Novelty: Bayesian optimization (BO) based sampling and Gaussian process regression (GPR) for generating local explanations.
- Interpret locality in a probabilistic sense!
- Employ sampling based on posterior density, in the locality of the instance to be explained: Sample from relevant regions of the posterior density, as guided by the acquisition function.
- Exploit the interdependence between the sampling and the explainer modules effectively, use of Gaussian process regression model for probabilistic explanations.

GP based explainer



Function to Optimize

- Mathematically, one needs to optimize the following:

$$L(f_p, f_e, \pi_{\mathbf{x}}) = \sum_{\mathbf{x}_k, \mathbf{x} \in \mathcal{X}} \pi_{\mathbf{x}}(\mathbf{x}_k)(f_p(\mathbf{x}_k) - f_e(\mathbf{x}))^2$$

- x_k is the sample of interest, e.g., a patient who walks into the hospital
- Use the predictions of the black-box ML model f_p
- f_e is unknown, LIME assumes linear models
- L(.,.) is expressed as squared loss

Bayesian Optimization

- Consider a situation where g(.) may not have a closed-form or may be expensive to evaluate, is non-differentiable and/or non-convex
- Only possibility - treat g(.) as a blackbox function that allows us to query the function value at some points **z.**

Treat g(.) as black-box

$$\mathbf{z}^* = \arg\max_{\mathbf{z} \in \mathcal{Z}} g(\mathbf{z}),$$

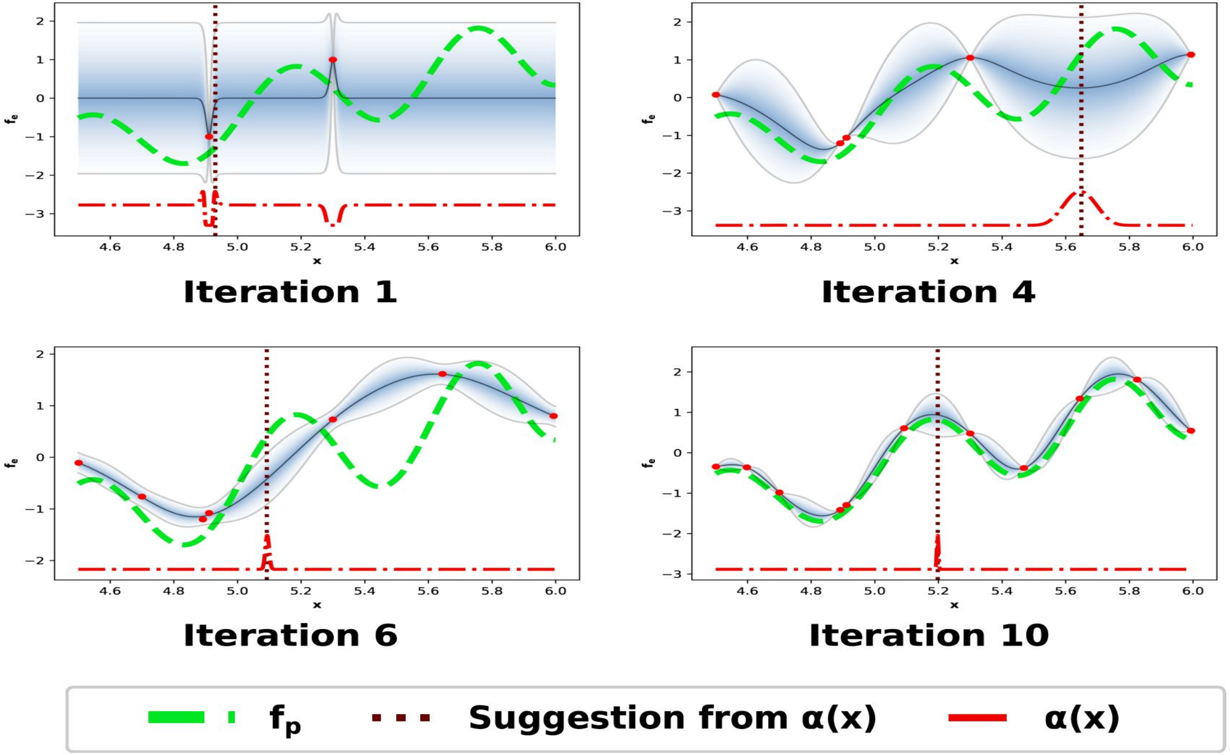
Allow querying g(.) at few points

BO: How does it work?

Bayesian optimization - *sequential design strategy for global optimization of black-box functions without assumptions of any functional forms on the black-box!*

- Bayesian strategy is to treat **g(.)** as a random function and place a prior over it.
  - prior captures beliefs about the behavior of the function
  - Commonly used: Gaussian priors, Prazen-tree estimators
- Gather the function evaluations, which are treated as data,
- Prior is updated to form the posterior distribution over the objective function.
- The posterior distribution, in turn, is used to construct an **acquisition function** (sampling criteria) that determines the next query point.

BO: Example



Iteration 1

Iteration 4

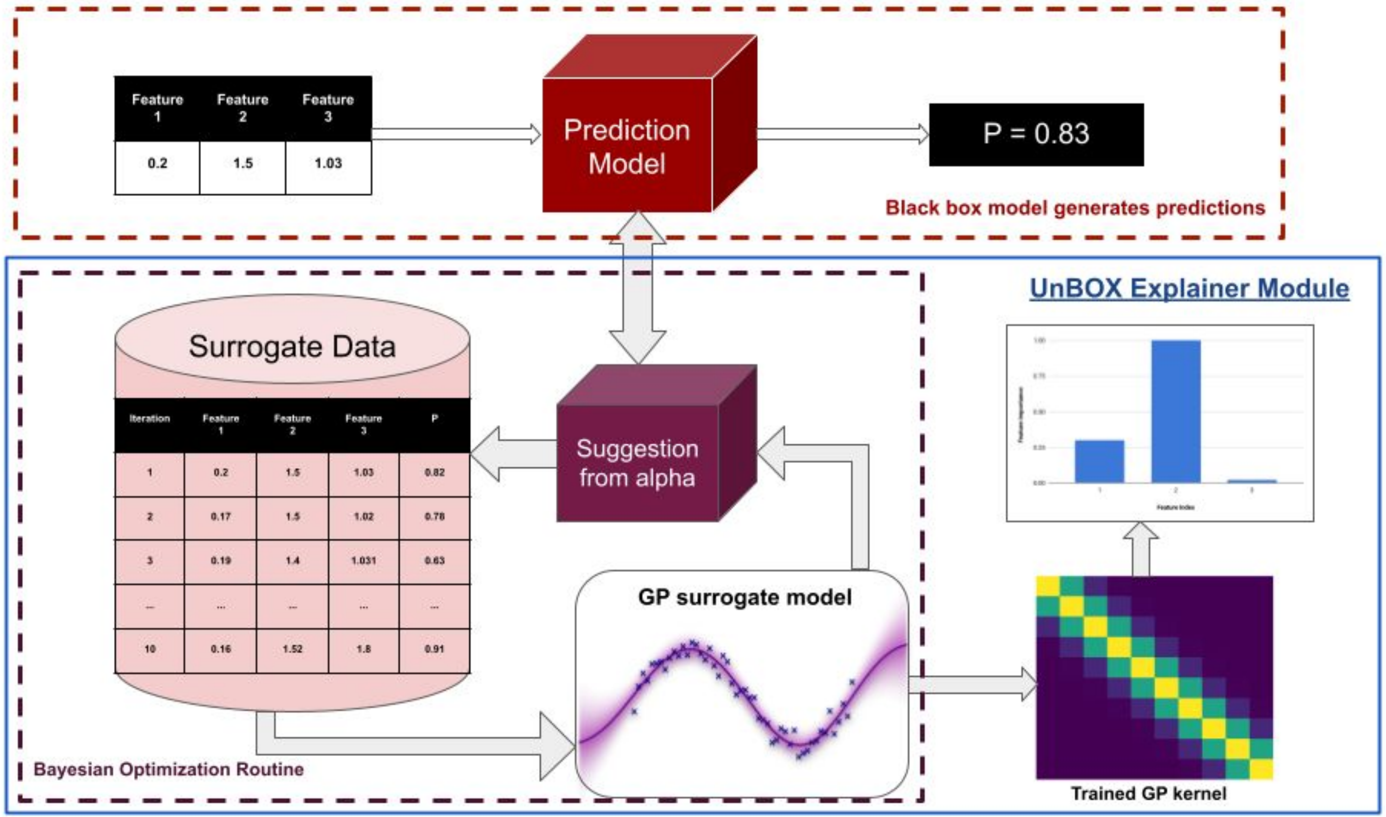
Iteration 6

Iteration 10

BO with GP as an Explainer

- We model **f_e(.) as the black-box** function!
- Place a **Gaussian prior on f_e(.)**
  - Learn the posterior using the sample of interest x_k
  - **Sample using the acquisition function**, which is a function of the posterior distribution
  - Surrogate samples lead to the likelihood function $p(\mathcal{D}_n|f_e)$
  - Combine the prior and likelihood to obtain the posterior distribution
  - This in turn is the new prior distribution incorporating both our prior beliefs and information from the surrogate data points.

UnBOX Flow Diagram



Feature 1	Feature 2	Feature 3
0.2	1.5	1.03

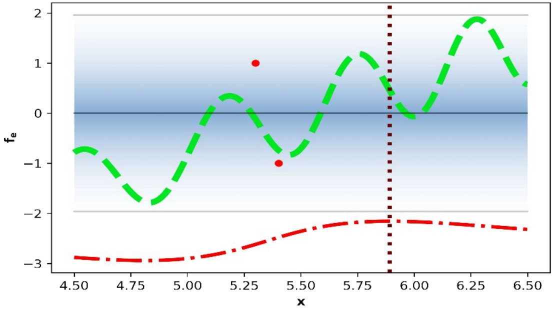
Iteration	Feature 1	Feature 2	Feature 3	P
1	0.2	1.5	1.03	0.82
2	0.17	1.5	1.02	0.78
3	0.19	1.4	1.031	0.63
...	...	...	...	...
10	0.16	1.52	1.8	0.91

Acquisition Function

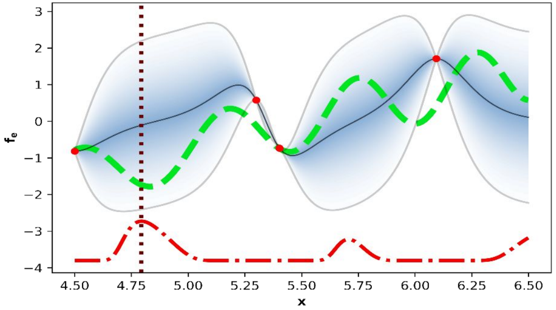
$$\mathbf{x}_{n+1} = \arg\max_{\mathbf{x}} \alpha(\mathbf{x}|\mathcal{D}_n)$$

- The acquisition function depends on the posterior density
- Several acquisition functions exist - based on maximizing the black-box function. Eg: LCB, UCB, EI
- This does not suit our requirement. Our requirement :
  - Minimize the explainer cost
  - Samples in the locality so that local fidelity is achieved
  - Need to obtain stability: control the sample complexity?

Lower Confidence Bound

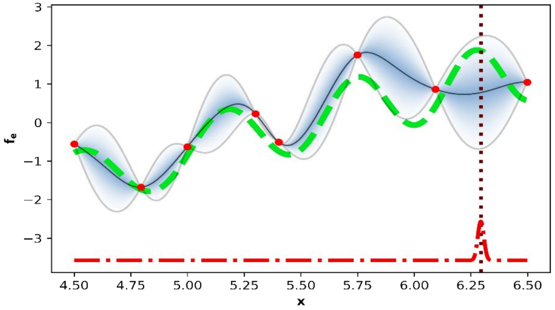


Iteration 1

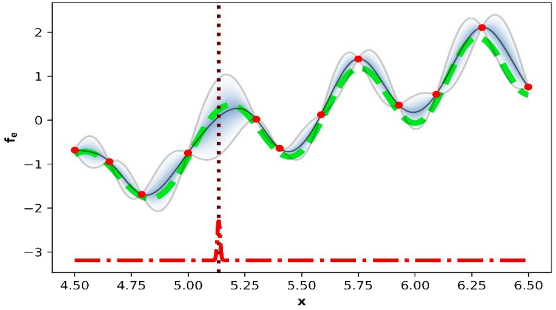


Iteration 4

$$\alpha_{LCB}(\mathbf{x}) = \mu(\mathbf{x}) - \kappa\sigma(\mathbf{x})$$

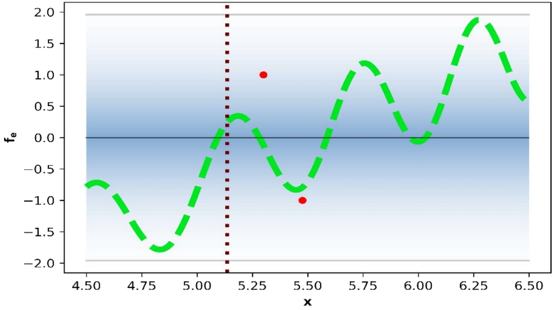


Iteration 8

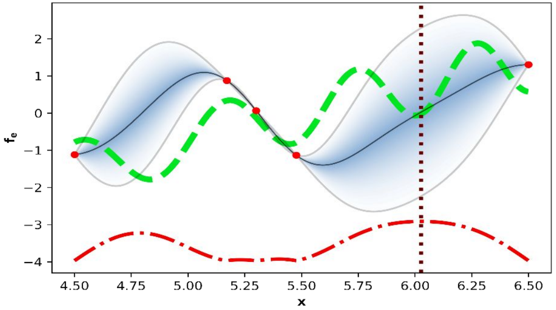


Iteration 12

Uncertainty Reduction

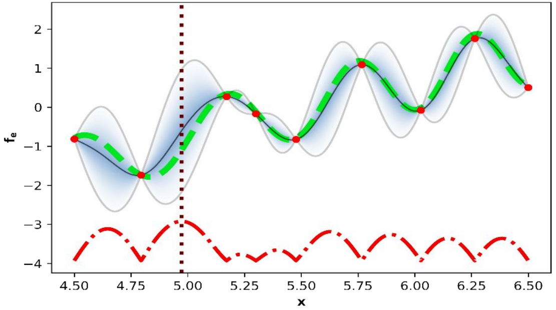


Iteration 1

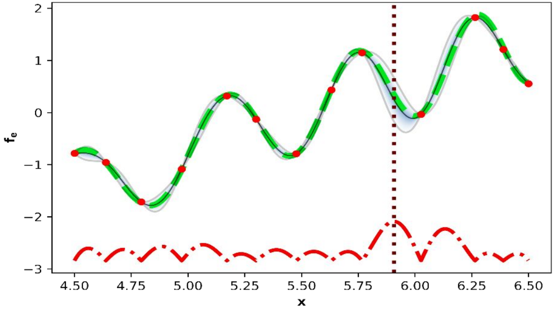


Iteration 4

Acquisition function only depends on the variance.

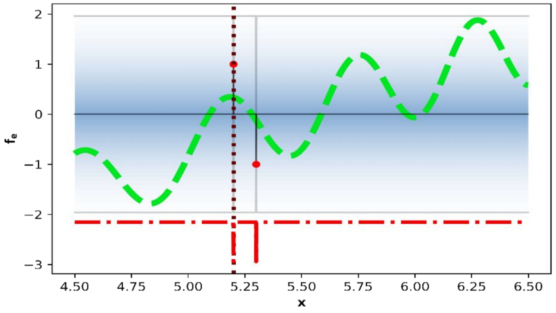


Iteration 8

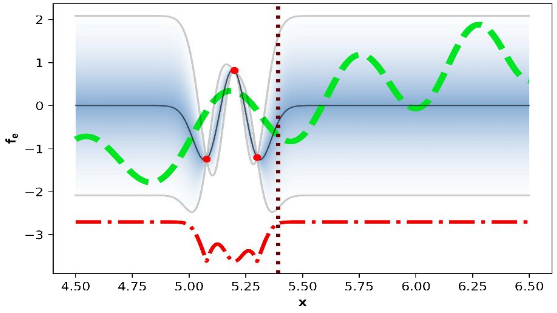


Iteration 12

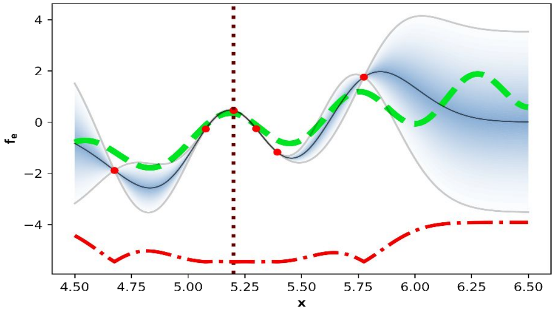
Proposed - Faithful Uncertainty Reduction



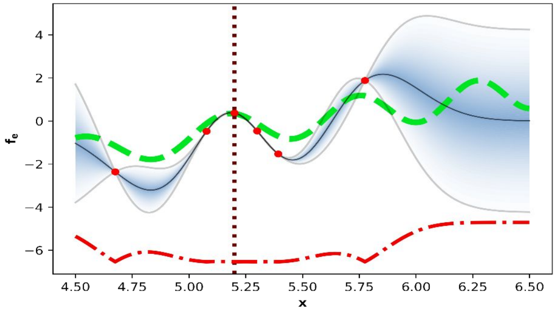
Iteration 1



Iteration 4



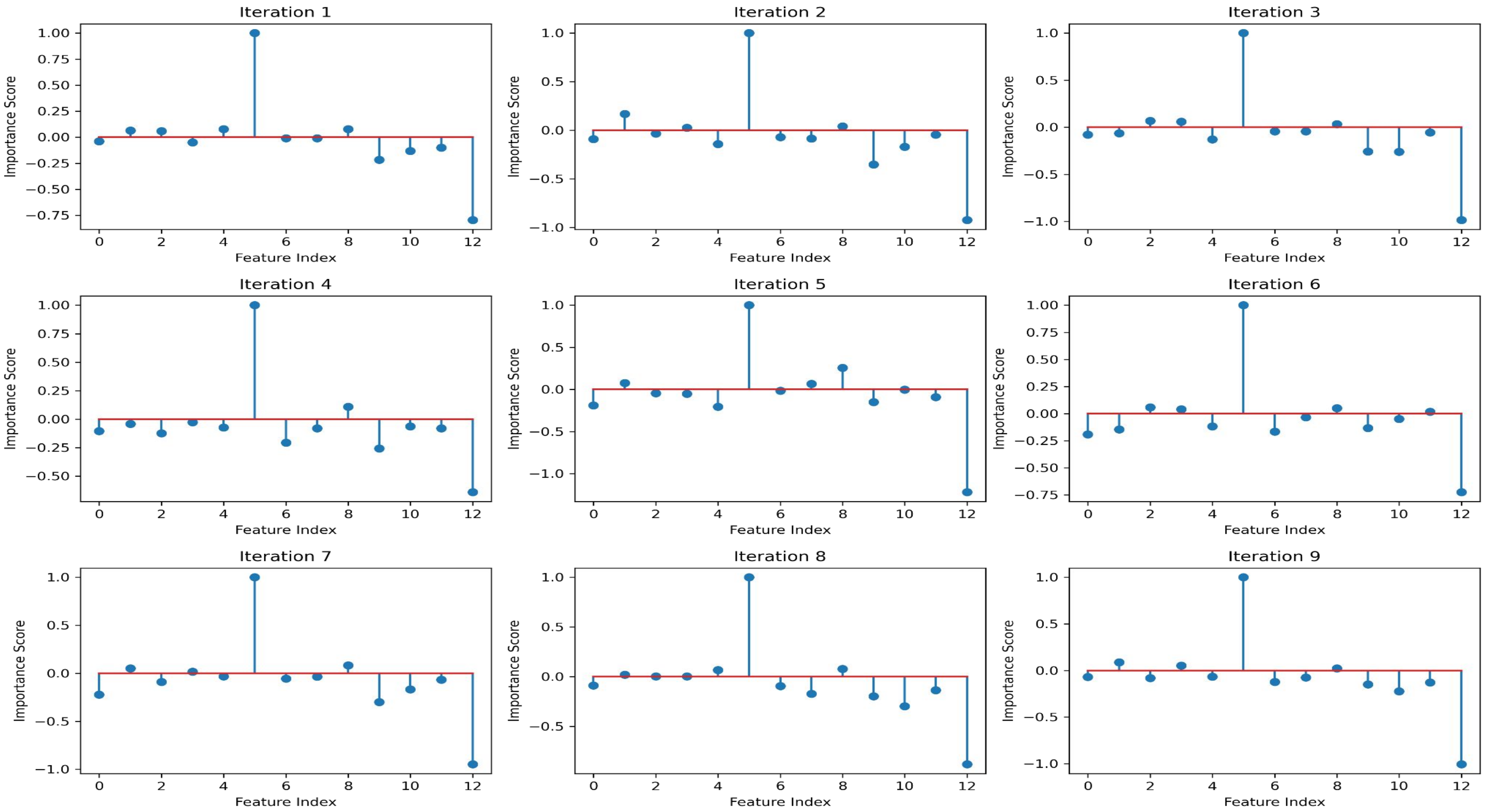
Iteration 8



Iteration 12

Novel acquisition function that measures the distance between the surrogate sample and x_k. Penalises the variance as iterations proceed.

Stability plots -(Kendall’s W = 0.8), Boston



Iteration 1

Iteration 2

Iteration 3

Iteration 4

Iteration 5

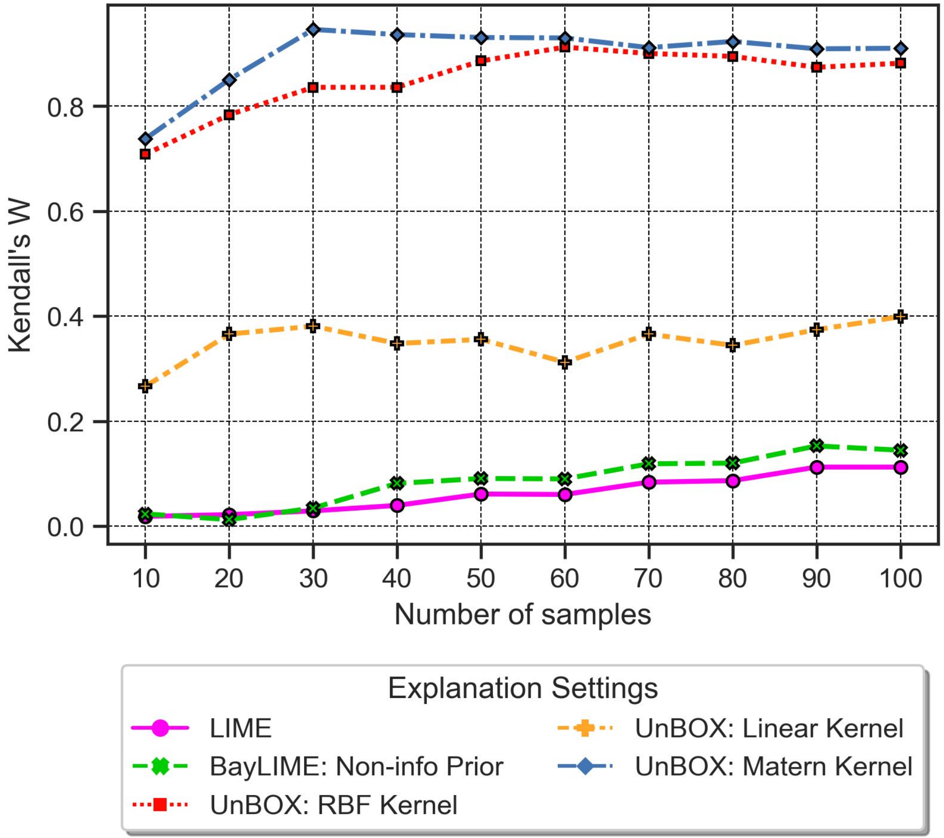
Iteration 6

Iteration 7

Iteration 8

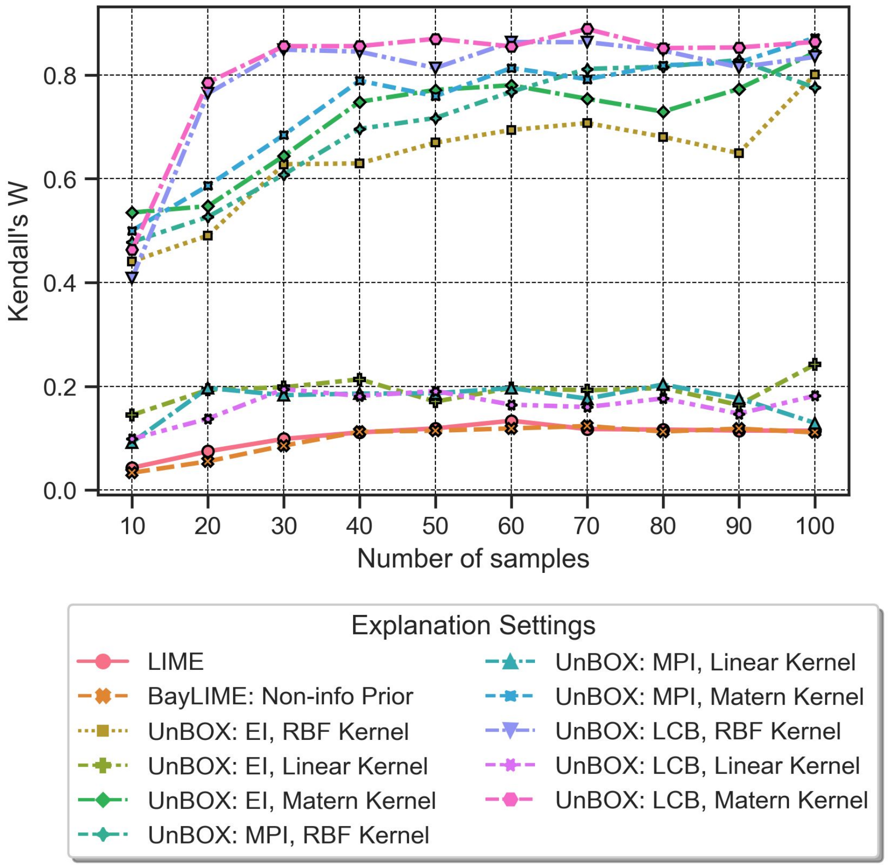
Iteration 9

Stability Plots: Parkinson’s dataset



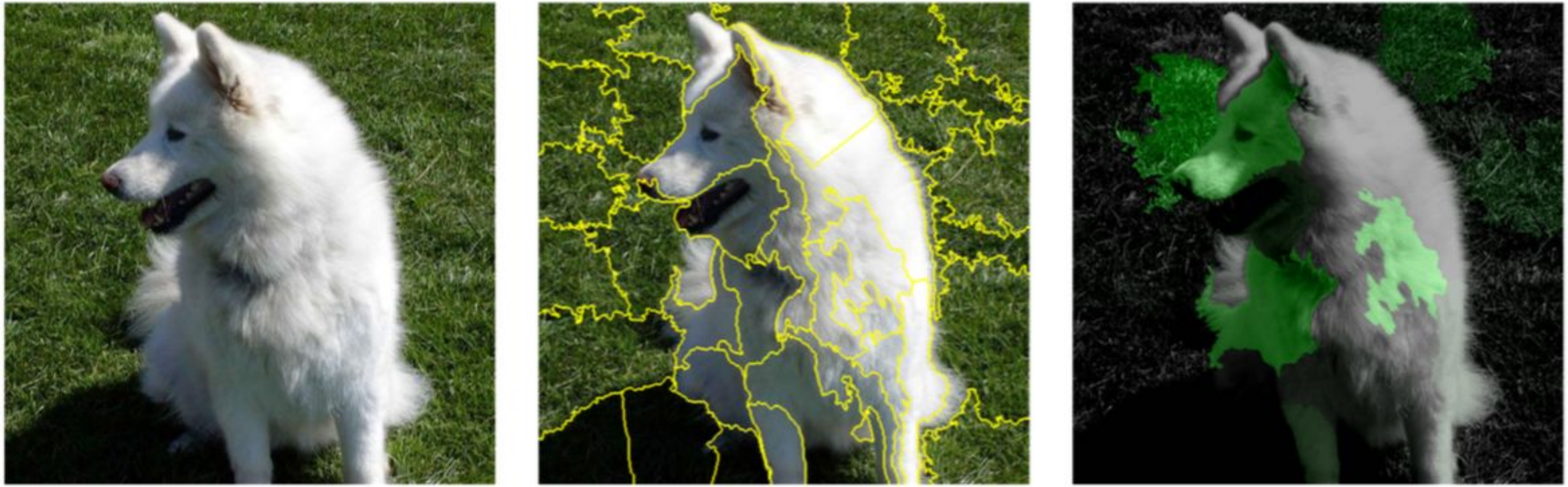
Kendall’s W measures the agreement of ranks of features over several iterations.

Stability Plots: Breast Cancer dataset



We consistently see that UnBOX is more stable as compared to LIME and BayLIME.

UnBox: (ResNet-18) Imagenet, class ‘samoyed’



(a) Original Image

(b) Segmented image

(c) UnBOX Explanation map

Conclusions

- Motivated the need for XAI
- Discussed the taxonomy of XAI techniques
- Discussed locally interpretable model agnostic techniques
- Discussed UnBOX and demonstrated how it overcomes limitations of LIME

Thank you!

(ranjitha@iiitd.ac.in)