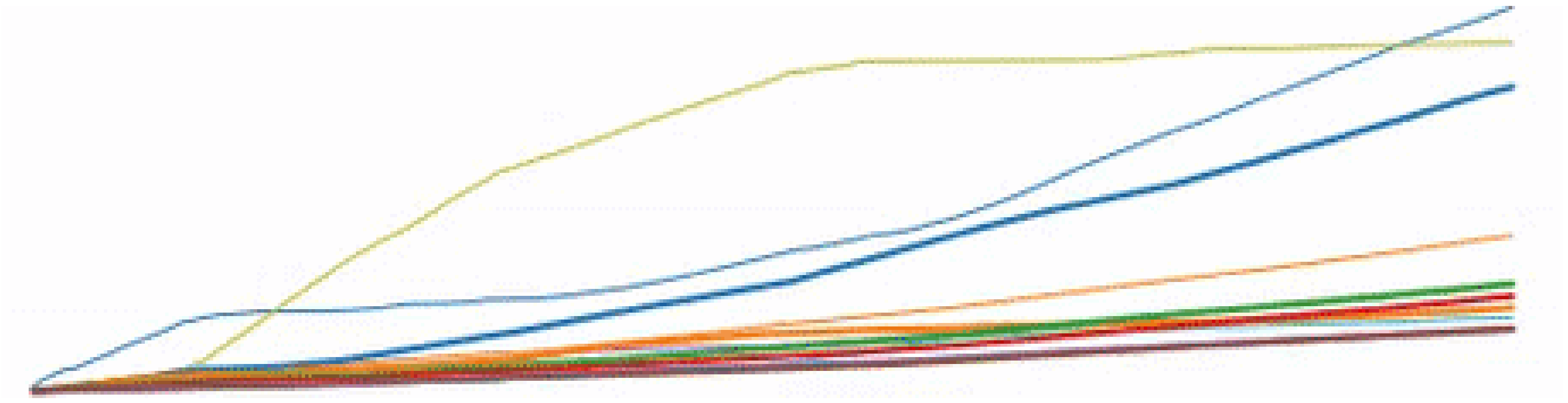
Explainable Artificial Intelligence



Approaches to Robust Online Explanations

August 2021

Outline

- Introduction
  - Motivation
  - Properties
- Experiments
- Results
- (Outlook)

Introduction: Motivation

- Data streams
  - Noise, Concept drift
  - Retraining of explained model

- Inconsistent explanations
- Example:
  - At time: t
    - Transient model state
    - Loan not approved with explanation e
  - At time: t + N
    - Loan not approved with explanation e‘ != e
    - e.g. Applicant increased expenses, but transient model state passed

Introduction: Properties

- Concentration on Approaches:
  - (Explained) Model agnostic
  - Handle Post-hoc explanations
  - Explanations = Feature attributions
- Model attributions from weights as ground truth

- Desired properties
  - Local faithfulness
  - Global faithfulness
  - Efficiency
  - Stability

Approaches

- Local Explainers
  - LIME
  - SHAP

$$a_t = l(x_t, f(x_t), X)$$

  - X is baseline

- Global Explainers
  - FIRES
    - Feature scores used as Feature attributions
    - Importance + Uncertainty

$$a_t = FIRES(x_t, f(x_t))$$

  - Weighted Gaussian Explainer
    - Outputs linearly dependent on attributions

$$P(y_t|x_t, a_t) = \mathcal{N}(y_t; {x_t}^T a_t, \Lambda)$$

    - Mean of posterior

Approaches: Combination of Local and Global Explainers

- Ensemble of Explainers
  1. FIRES → Model distribution
  2. Sample N Models
  3. Local explainer → Explain predictions of N models
     1. LIME or SHAP
  4. Aggregate explanations (e.g. Mean of explanations)

- Weighted Local and Global Explainers

$$a_t = \frac{w_l \cdot l(x_t, f(x_t)) + w_g \cdot g(x_t, f(x_t))}{w_l + w_g}$$

Experiments

- xEvaluator
  - Environment for experimentation with explanations and data streams
- Datasets:
  - Synthetic, 5000 samples
  - From practice:
    - Spambase
    - Card default
- Ground truth:
  - Explained model weights

Desired Properties	**Evaluation Metrics**
Local faithfulness	$LF_{t+1} = LF_t + \left\| \frac{a_{t+1}}{\Sigma_i \|a_{t+1_i}\|} - \frac{g_{t+1}}{\Sigma_i \|g_{t+1_i}\|} \right\|$
Global faithfulness	Top-10 Ranking Accuracy
Efficiency	Timings
Stability	$S_{t+1} = S_t + \left\| \frac{a_{t+1}}{\Sigma_i \|a_{t+1_i}\|} - \frac{a_t}{\Sigma_i \|a_{t_i}\|} \right\|$

Results: Time and Stability

- Average over all data sets

	Explainer	Time (ms)
1.	FIRES	0.87
2.	WeightedGaussian	2.34
3.	LIME	8.00
4.	LIME+FIRES	8.95
5.	SHAP	9.56
6.	SHAP+FIRES	10.10
7.	LIME-Ensemble	115.92
8.	SHAP-Ensemble	158.85

	Explainer	Stablity
1.	FIRES	106.36
2.	WeightedGaussian	108.73
3.	LIME+FIRES	1296.45
4.	LIME	1906.25
5.	LIME-Ensemble	2431.43
6.	SHAP+FIRES	4567.24
7.	SHAP-Ensemble	6079.14
8.	SHAP	7192.02

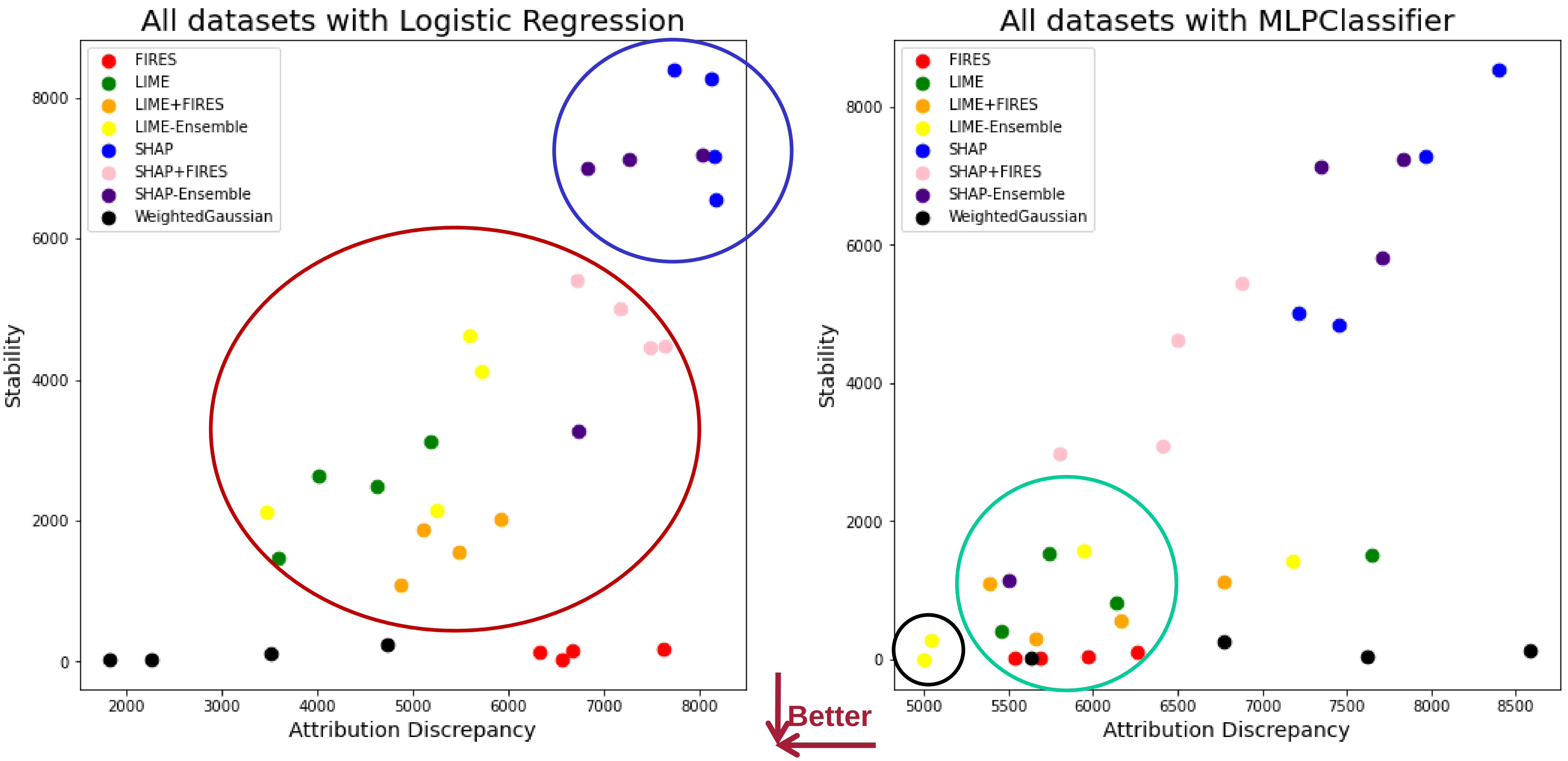
- Ensemble efficiency penalty

- Ensemble stabilizes SHAP

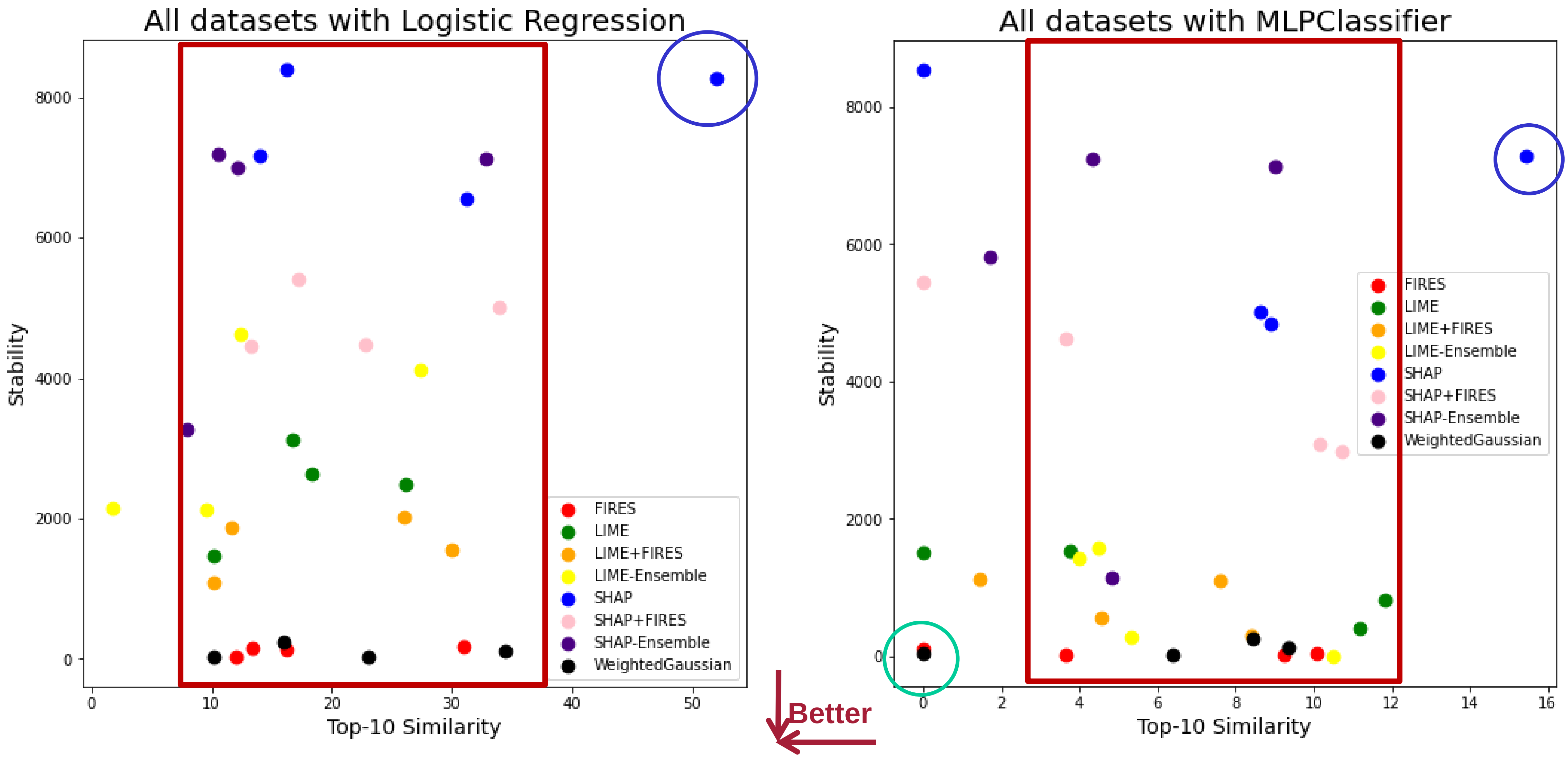
- Global better than Local

- LIME better than SHAP

Results: Local Faithfulness vs Stability



Results: Global Faithfulness vs Stability

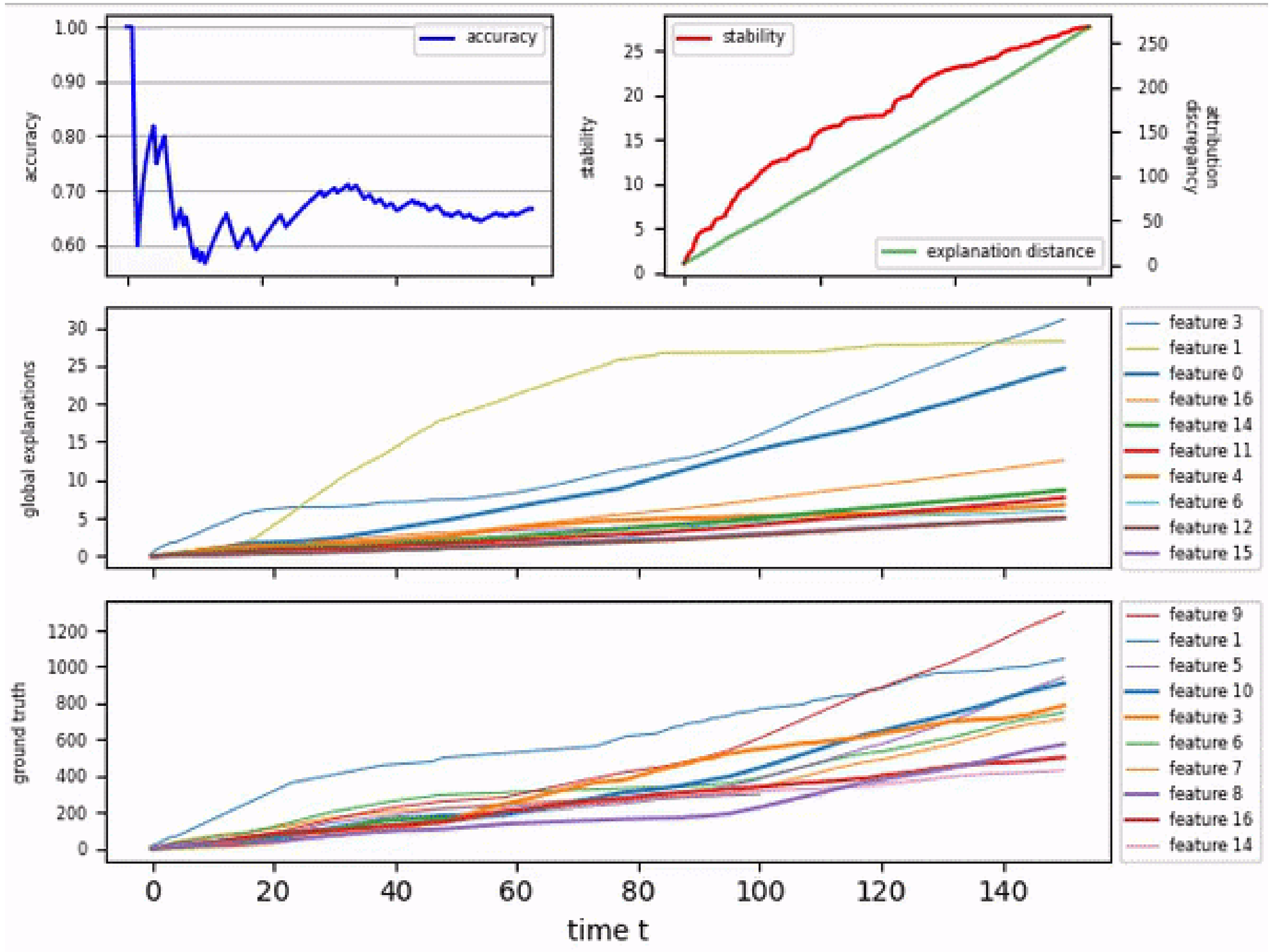


Outlook

- Feature selection scores as explanations
- Global + Local explainers
  - e.g. Include local explainer output into prior of global explainers
- (Time-weighted) sliding window for ranking accuracy
- Surrogate rule extraction approaches

Outlook

- (Global) explanations through visualization
- Explainer evaluation through visualization



Thank you!