# Understanding Prediction Discrepancies in Classification

Thibault Laugel – thibault.laugel@axa.com
Based on work conducted at AXA with Xavier Renard and Marcin Detyniecki

# Responsible / Trustworthy AI

Explainability

Fairness

HCI

Robustness

Privacy

# Why Explainable AI? / Explainable to whom?

## Improve Model's Quality

"Data Scientist in the loop"

- Improve models, features
  - Prediction errors…
- Identify issues & pitfalls
  - Robustness, fairness, concept drift…

## Inform Business

"Business in the loop"

- Improve ML acceptance
- Inform ML-based decisions
- Gain insights on business processes

## Legal & Ethical Compliance

"Customer in the loop"

- Right to explanation (GDPR)
- Assess model's fairness
- Inform customers

# Explainable AI / ML

What has been learned by the model?

Where is the model {correct ; wrong} ?

Why has a particular prediction been made?

Is the model fair?

What can be done to change the prediction?

x

**Description of the problem to solve**
Tabular data, unstructured data, etc.

**Machine Learning Model**

y

**Prediction / Decision**

**Usually aggregated accuracy score**

Is the model robust?

Does the model agree with other competing models?

How does the model behave in areas with few data?

Is the model causal?

How is the prediction affected by small changes in input?

# Explainable AI / ML

What has been learned by the model?

Where is the model {correct ; wrong} ?

Why has a particular prediction been made?

Is the model fair?

What can be done to change the prediction?

x

**Machine Learning Model**

y

**Usually aggregated accuracy score**

**Description of the problem to solve**
Tabular data, unstructured data, etc.

**Prediction / Decision**

Is the model robust?

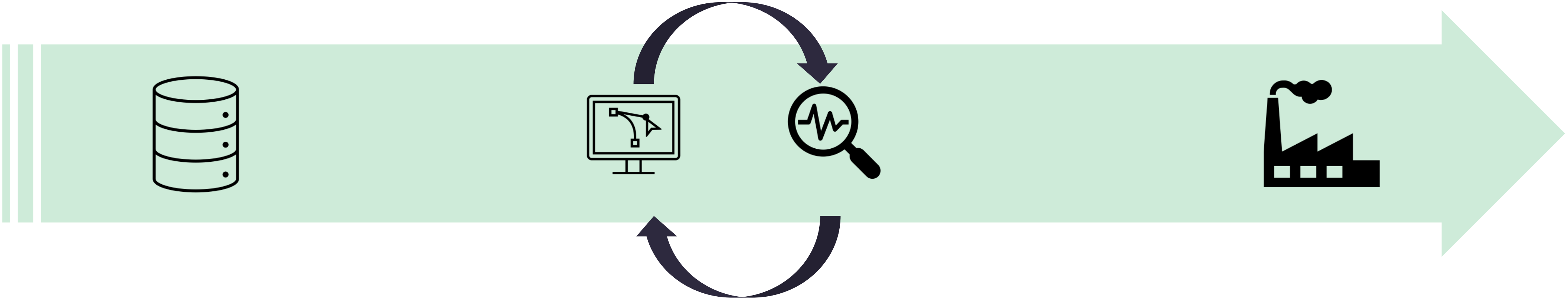
Does the model agree with other competing models?

How does the model behave in areas with few data?

Is the model causal?

How is the prediction affected by small changes in input?

# Context: supervised learning



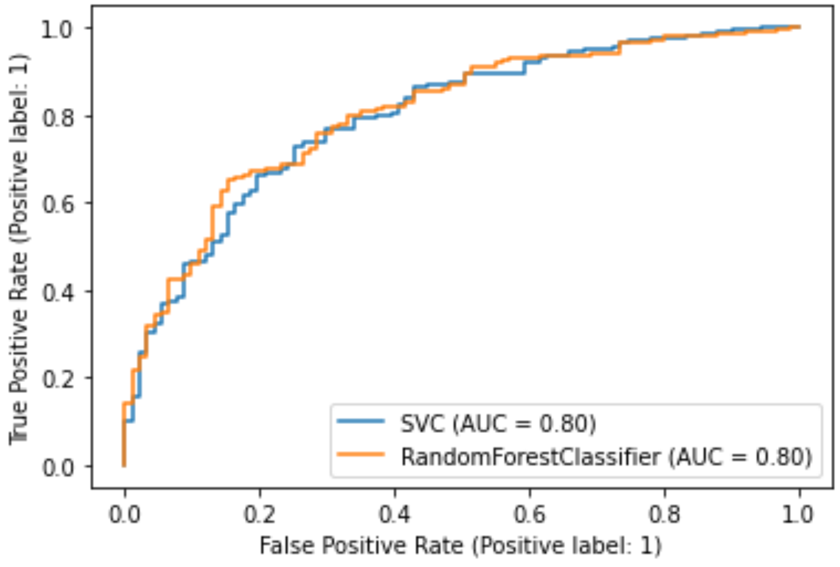
Data collection and preparation

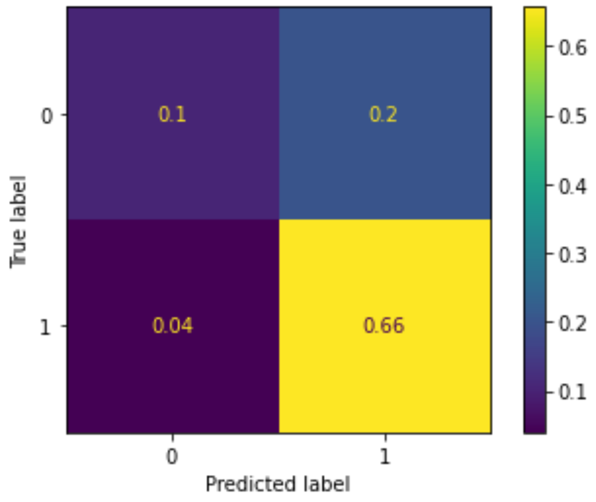
Model training, selection and validation

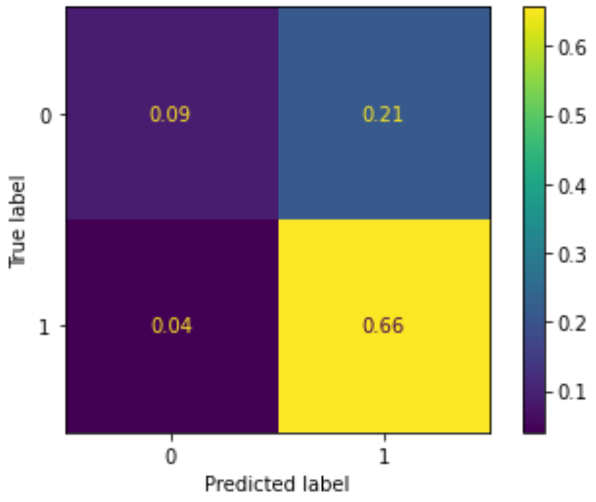
Depoyment in production

# Context: model selection and validation

German Credit dataset



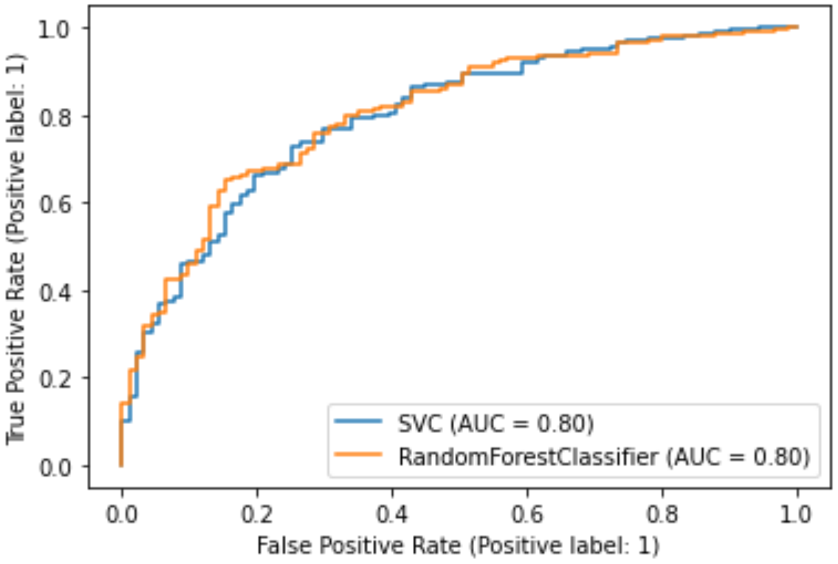




Are these models the same?

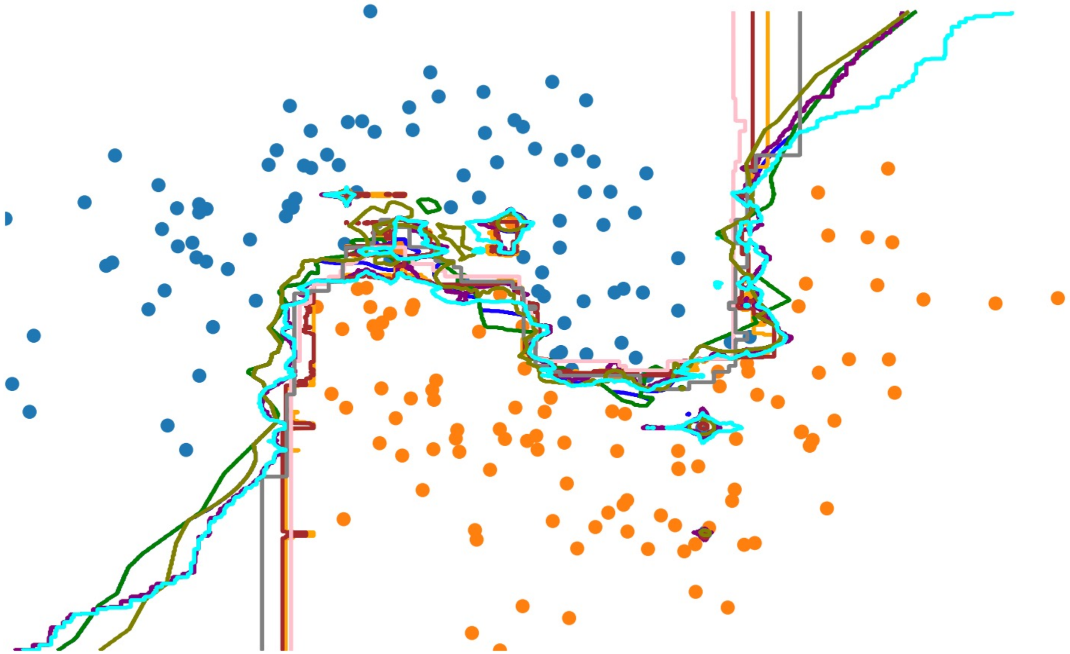
# Context: model selection and validation

German Credit dataset




Predictions disagreement **~12%**

# Prediction discrepancy



**Discrepancy**: the difference between models trained on the same data

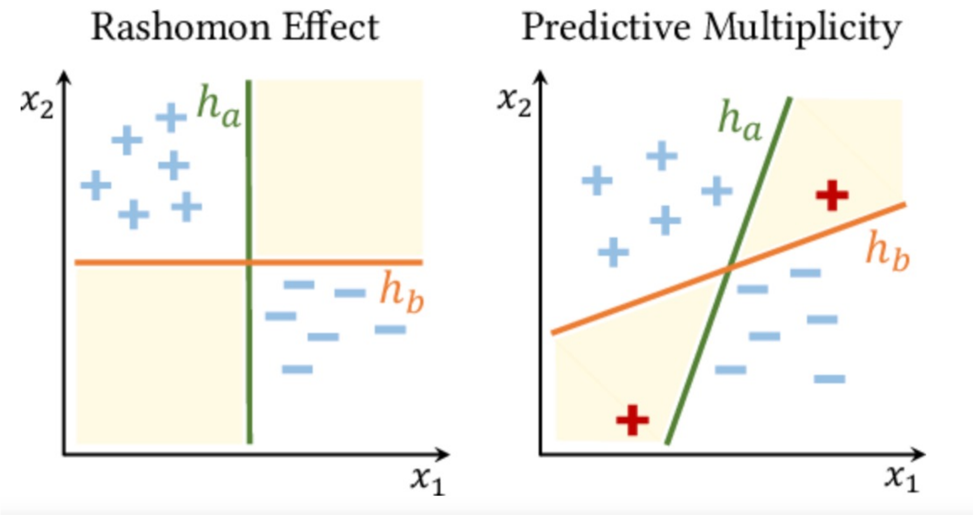
Particularly when these models have the same train/test/validation error

# Why is there discrepancy?

A known phenomenon

Obviously a known phenomenon, not necessarily seen as an issue:

- « Roshomon effect, the multiplicity of good models » [Breiman 2001]
- Ensemble learning: diversity as a source of predictive robustness [Hansen et al. 1999, Dietterich 2000]
- … and adversarial robustness [Pang et al. 2019]




*Source: Marx et al. 2019*

# Why is there discrepancy?

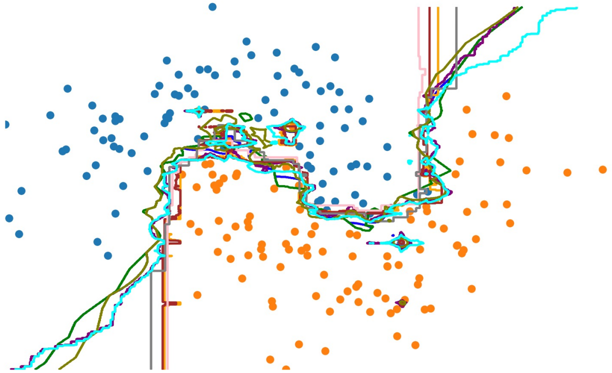
Discrepancy is **unavoidable**

One way of seeing it is that it happens because there is not enough data

- Disagreement between models = uncertainty [Bomberger 1996, …]
- Can be leveraged to label data (active learning strategies) [Abel et al. 1998, Melville et al. 2004, Lett et al. 2022]

**However, it stems from the ML task itself**

- Data = sparse representation of the world
- A ML model is asked to generalize between these data points
- Models learn different generalizations: « ML problems are underspecified » [D'Amour et al. 2020]

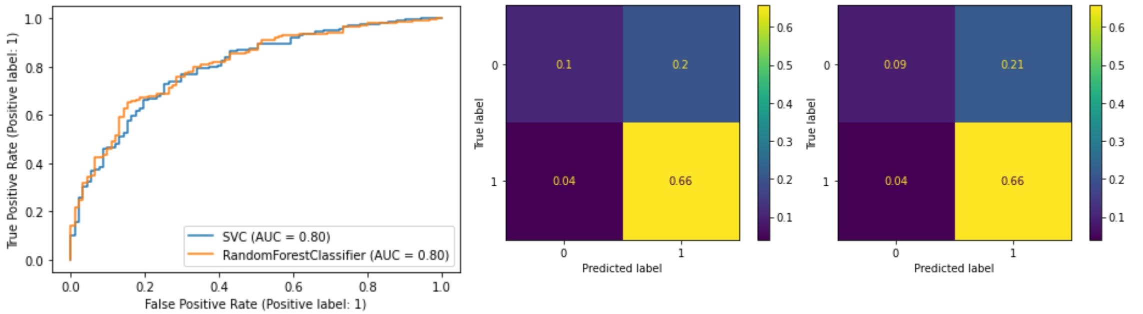


# Why is discrepancy an issue?

An often hidden / ignored phenomenon

The choice of which model to deploy in production / to use is usually based on **predictive performance**

- = aggregated score, hiding model differences
- This choice is thus made blindly




**A large portion of the ML-based decisions are thus made arbitrarily**

- Another model could have been selected instead and made a totally different decision

Yet, this choice may impact a lot of high-stake decisions

- Suboptimal or biased decisions
- Damaging customer trust (unreliable systems)

# Why is discrepancy an issue?

Discrepancy can have negative consequences

The question tackled here is not how to train a better model…

… But given a deployed model, how to deal with the practical issues raised

A recent wave of works have focused on the negative consequences of discrepancy

- Models that can not be trusted [Rawal et al. 2020]
- ML systems (e.g. explanations) that can not be trusted [Barocas 2020, Pawelczyk et al. 2020]
- *Fairwashing* and explanation manipulation [Aivodji2019, Slack2020, D'Amour2020]

# How big is the issue?

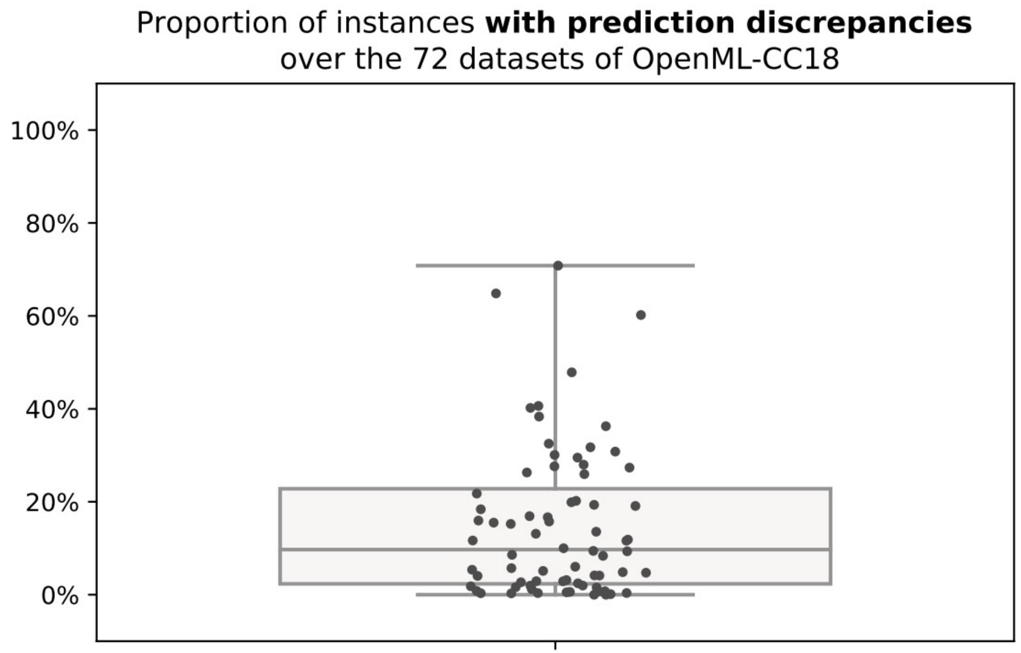
A small replicable experiment to quantify discrepancy

Empirically quantifying discrepancy:

- Datasets of the OpenML-CC18 Benchmark
- Predictions of the best runs extracted automatically
- All models extracted are in a 2% accuracy range.
- We measure the prediction discrepancies

Results are expected to vary with the accuracy reached and with the number of models falling in the 2% range

However, the general observation is that discrepancy happens « a lot ».



# Beyond quantifying the issue

Recent works have focusing on studying the phenomenon [Semenova et al. 2019, Dong and Rudin 2019, Geirhos et al. 2020, Marx et al. 2020]

Most of these works propose metrics to **quantify the issue**, guarantees about the importance of the issue

This helps ML developers being aware of the issue on a given ML task…

- However, no concrete solution is available to circumvent the issue at training time (i.e. before it is too late)
- In this work, we propose to go beyond quantification by **explaining ML discrepancies**

# Explaining ML discrepancies

This work focuses on **explaining the differences between models**

- as opposed to explaining the behavior of one model (frequent post-hoc interpretability setting, cf. SHAP & co.)

The objective is to help the ML practitioner, allowing him/her to take concrete actions such as:

- Model debugging: identify uncertain regions to improve the modeling (e.g. collect more data)
- Remedial measures: abstention, ask for human intervention
- Certification / model auditing: give guarantees about the behavior of the model

# Proposition: Explaining discrepancies

Algorithm Requirements

We propose to design a tool to explain the differences between a pool of trained classifiers

Algorithm requirements:

1. Practical usage: model- and data- agnosticity
2. Grounded and actionable explanations
3. Precise explanations
4. Efficient detection and explanation generation

# Proposition: Explaining discrepancies

Algorithm Requirements

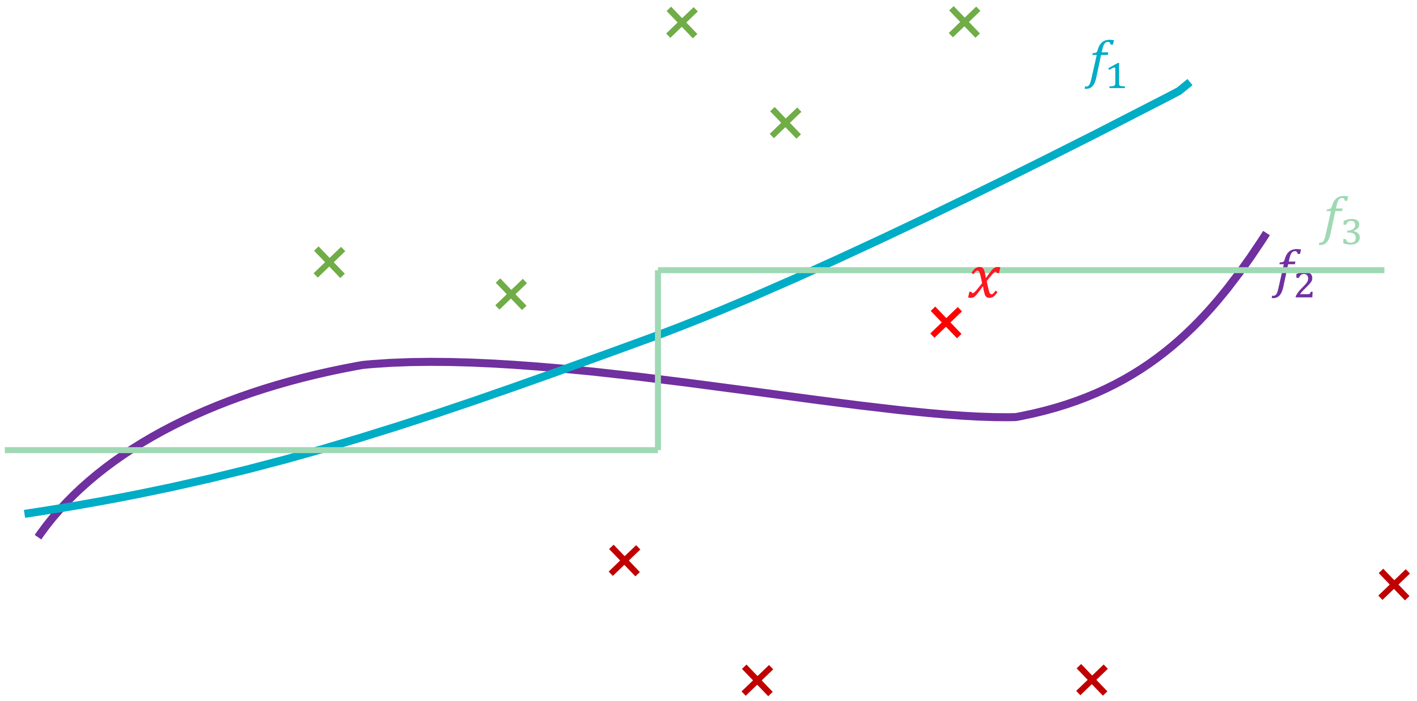
We propose to design a tool to explain the differences between a pool of trained classifiers

Algorithm requirements:

1. Practical usage: model- and data- agnosticity
2. **Grounded and actionable explanations**
3. **Precise explanations**
4. Efficient detection and explanation generation

# Algorithm objective

Generating discrepancy intervals

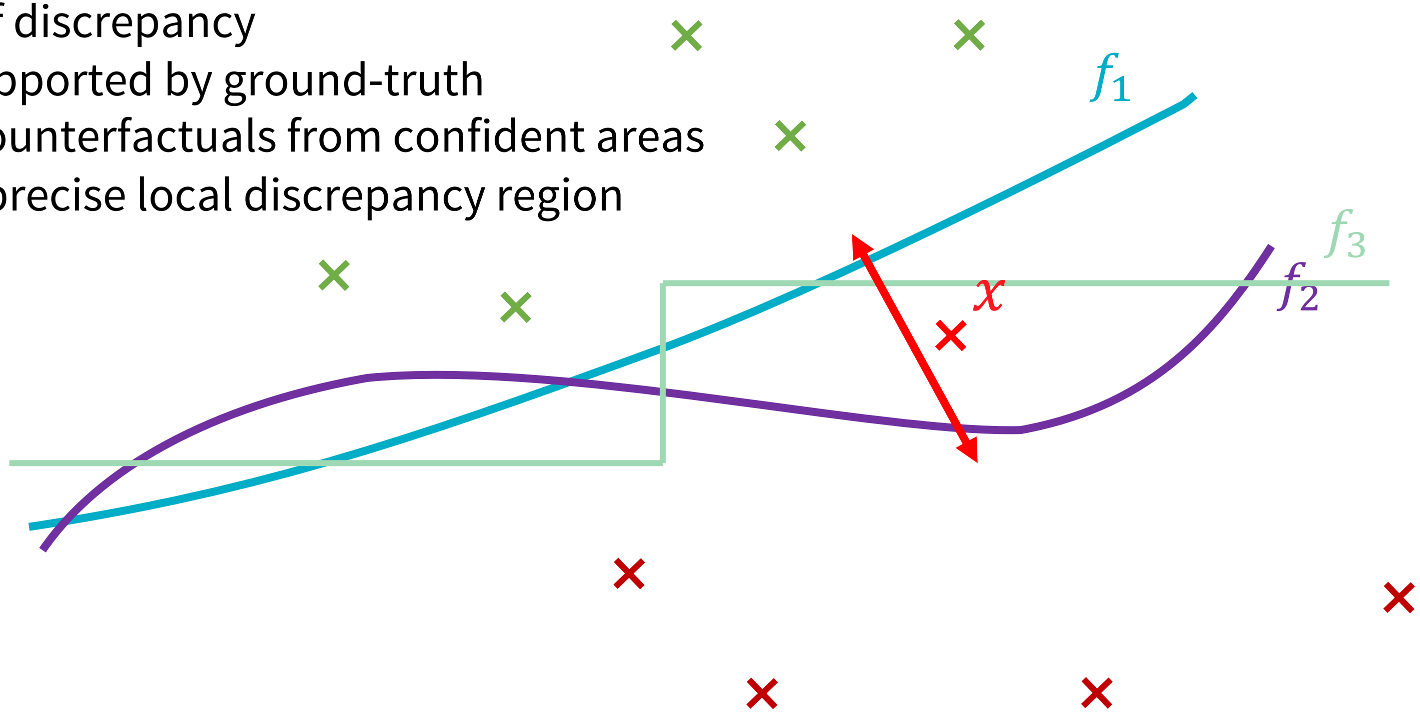


# Algorithm objective

Generating discrepancy intervals

Local explanation of discrepancy

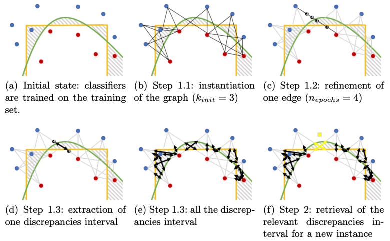
- Direction supported by ground-truth
- = defining counterfactuals from confident areas
- Delimit the precise local discrepancy region



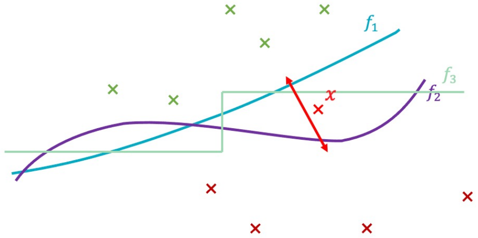
# Algorithm Description: DIG

## Inputs: training data and pool of trained models

DIG

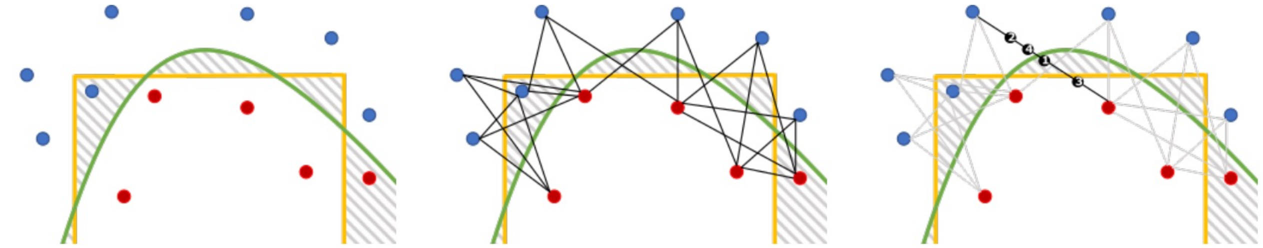



## Local Explanations of discrepancies



# Algorithm Description: DIG

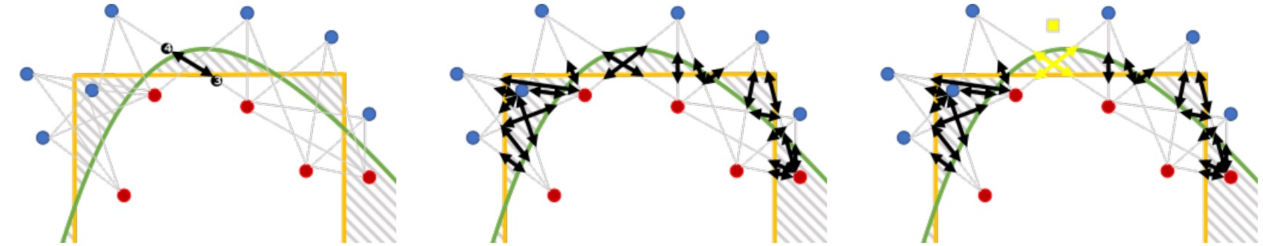
**D**iscrepancy **I**nterval **G**eneration (DIG)



(a) Initial state: classifiers are trained on the training set.

(b) Step 1.1: instantiation of the graph ($k_{init} = 3$)

(c) Step 1.2: refinement of one edge ($n_{epochs} = 4$)



(d) Step 1.3: extraction of one discrepancies interval

(e) Step 1.3: all the discrepancies interval

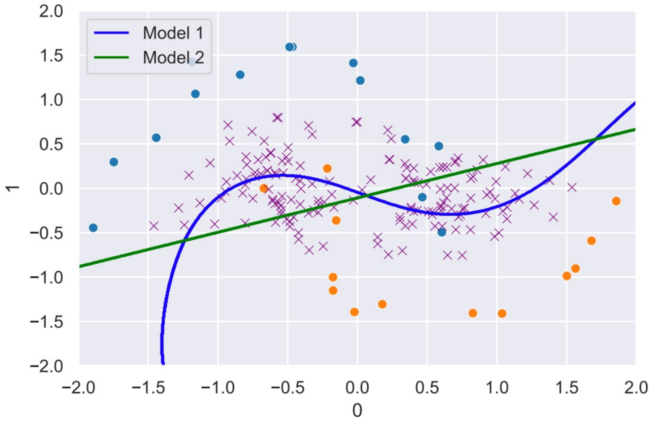
(f) Step 2: retrieval of the relevant discrepancies interval for a new instance

# Evaluation

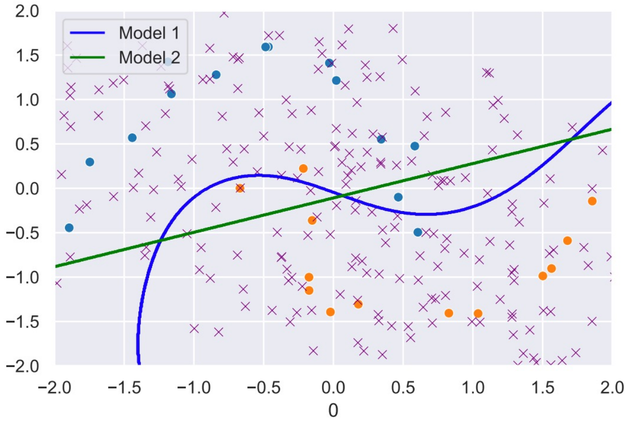
**Goal: how well are discrepancy covered?**

- Comparison with other sampling approaches (here, KDE with same budget)

Other evaluations: precision of the generated intervals, impact of the heuristic parameters




*DIG*



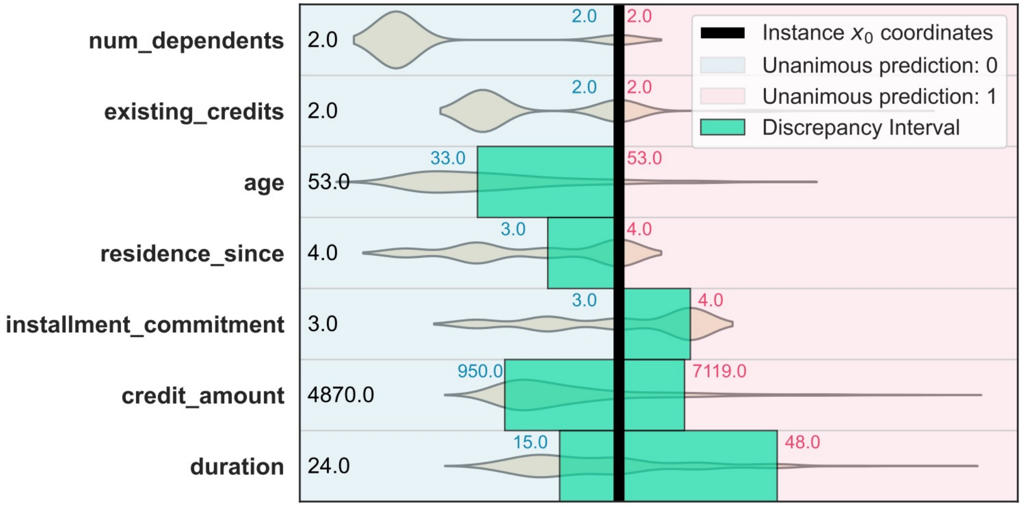

*KDE*

| Dataset | DIG | KDE |
|---|---|---|
| half-moons | **0.96 (0.02)** | 0.92 (0.03) |
| boston | **0.78 (0.05)** | 0.57 (0.07) |
| breast-cancer | **0.75 (0.05)** | 0.40 (0.02) |
| churn | **0.60 (0.02)** | 0.59 (0.01) |
| news | **0.60 (0.02)** | 0.42 (0.05) |
| adult | **0.81 (0.03)** | 0.60 (0.02) |
| german | **0.71 (0.03)** | 0.65 (0.02) |

*Detection of discrepancy areas with a 1-NN classifier trained on the sampled instances*

# DIG Output example

German Credit dataset




| | |
|---|---|
| **checking_status** | <0 |
| **employment** | 1<=X<4 |
| **personal_status** | male single |
| **other_parties** | none |
| **property_magnitude** | no known property |
| **housing** | for free |
| **job** | skilled |
| **own_telephone** | none |
| **foreign_worker** | yes |

Discrepancy interval generated for an instance over which classifiers are disagreeing

# Extension: Dealing with non-interpretable features

Discrepancy intervals are useful if sampling in the input space makes sense

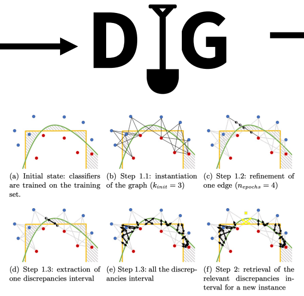
- If not (e.g. pixel), the explanation is useless

Proposition: unsupervised learning of a meaningful feature space and apply DIG in it

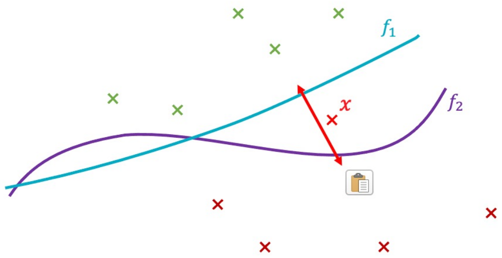
- E.g. autoencoders and variations [Guidotti et al. 2021]

**Inputs: training data and pool of trained models**

**Encode training set: project in latent space *Z***



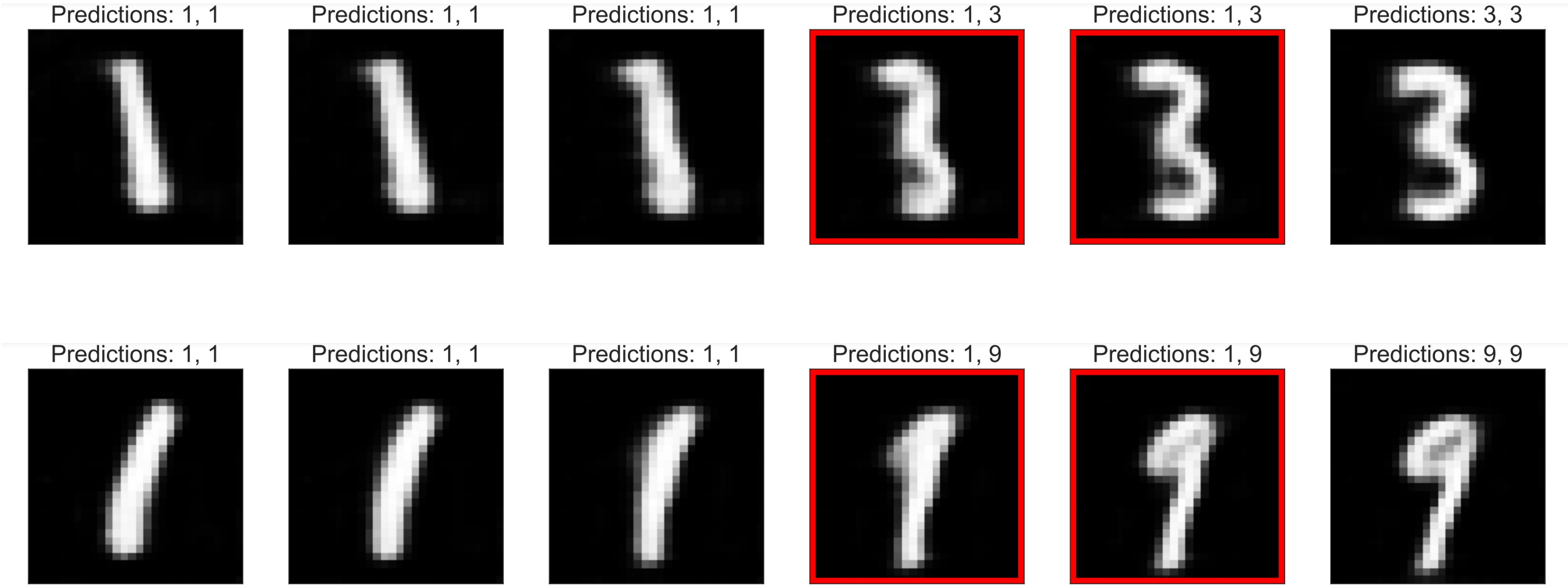

**Local Explanations of discrepancies in latent space**




**Decode discrepancy intervals**

# Extension: Dealing with non-interpretable features

Output (MNIST)



# Extension: global insights

**2 sparse discrepancy segments** detected by DIG

| **Segment A** |
| --- |
| **Credit amount** > 7800 DM |
| **Prev. existing checking account** = yes |

- 6% of the training set
- « Large » area

| **Segment B** |
| --- |
| **Credit amount** < 1500 DM |
| **Installment rate (% of income)** = 4% |
| **Prev. existing checking account** = yes (negative amount) |

- 5% of the training set but smaller area
- Models have very different perf. over the segment

# Conclusion & Perspectives

In these works, we:

- Show the importance of addressing prediction discrepancies
- Propose a tool to investigate ML discrepancies

Future works include:

- Extensions to regression, clustering
- Leverage active learning strategies
- Explore discrepancies for textual data

# Opening of a joint lab with Sorbonne Université

TRAIL
Trustworthy and
Responsible AI Lab

Objectives:

- Secure fundings for PhDs, post-docs, visiting researchers…
- Bi-monthly open seminars (physical and virtual) around Responsible ML topics
- Easier external collaborations