Efficient Understanding of Machine Learning Model Mispredictions

Martin Eberlein[1], *Jürgen Cito*[2], *and Lars Grunske*[1]

[1] Humboldt-Universität zu Berlin, Germany
[2] TU Wien, Austria

ML Models Make Mistakes!

To debug and trust these models, we must understand ***why and when they fail.***

input data → **ML Model** → incorrect prediction

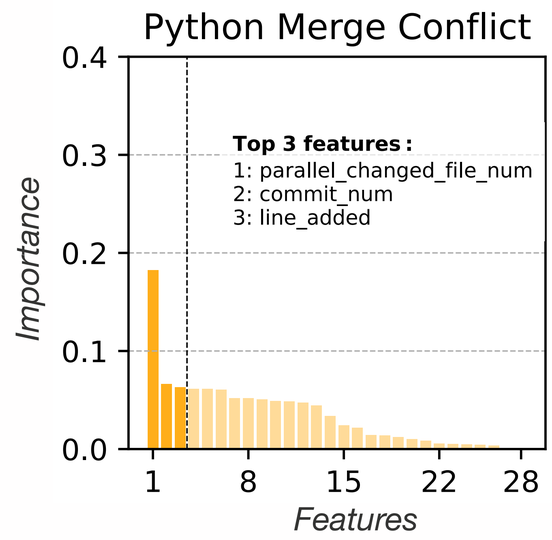
Machine Learning Model Diagnosis (MMD) generates ***rule sets*** *that can explain mispredictions:*

```
IF featureA == ASE AND featureB > 2025
THEN misprediction
```

Not All Features Are Important!

The limitation:
MMD is computationally expensive and lacks accuracy. It scales poorly with the number of feature!

But:
Not all *features contribute* ***equally to mispredictions!***

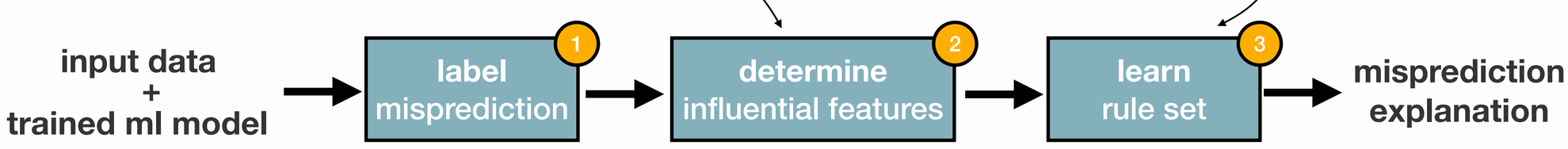


MMDFast: *Efficient* & *Accurate* Misprediction Explanations

Two key innovations:

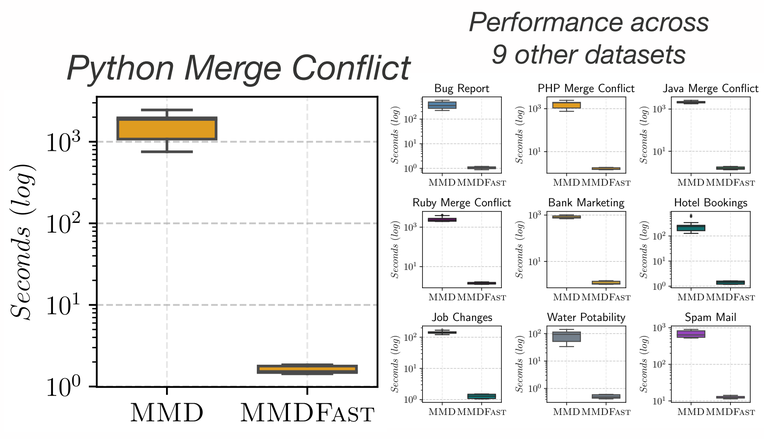
1. ***Focus on Influential Features:*** *Don't analyze all features, only focus on the most influential features. This dramatically reduces the search space.*

2. ***Optimized Rule Construction****: Replace costly beam search with limited-depth decision trees. This is faster and generates more precise, robust rules.*



① Drastically Faster

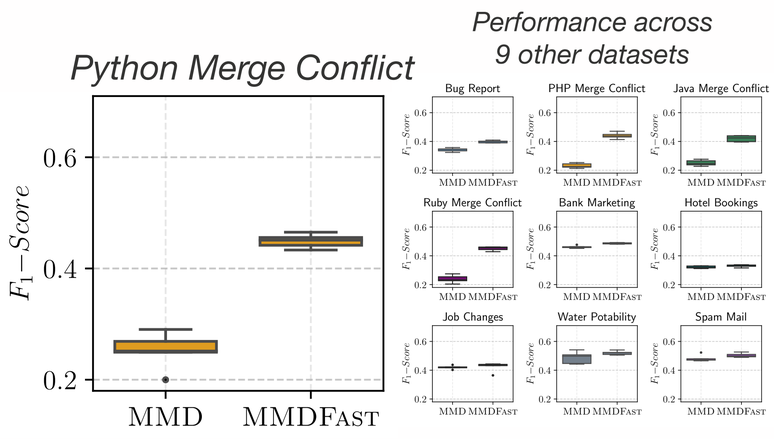
MMDFast *focuses only on the features that are really important.*



MMDFast *produces explanations orders of magnitude faster!*

② More Accurate

Combined with limited-depth decision trees, F_1*-Scores improve significantly!*



MMDFast *produces more accurate explanations!*

③ Actionable Rules

Merge Conflict Prediction:

```
IF: ParallelFileChanges AND num(Commits) > 1
ELIF: num(LinesAdded) > 26
```

Actionable Insight: *Expand training data with commits that make many changes simultaneously.*

Job Change Prediction:

```
IF: CDI < 0.63 AND TrainingHours < 99.5
ELIF: Experience < 5.5
```

Actionable Insight: *The model is less reliable for low-experience or low-CDI applicants. Manually review these cases.*

Eberlein et al.: "Efficient Understanding of Machine Learning Model Mispredictions", ASE 2025.
Cito et. al: „Explaining Mispredictions of Machine Learning Models using Rule Induction", ESEC/FSE 2021.

martin.eberlein@hu-berlin.de www.martin-eberlein.com

Full evaluation study on ***eleven data sets*** *(se and non-se tasks) in the paper:*