Improving Model Selection by Employing the Test Data

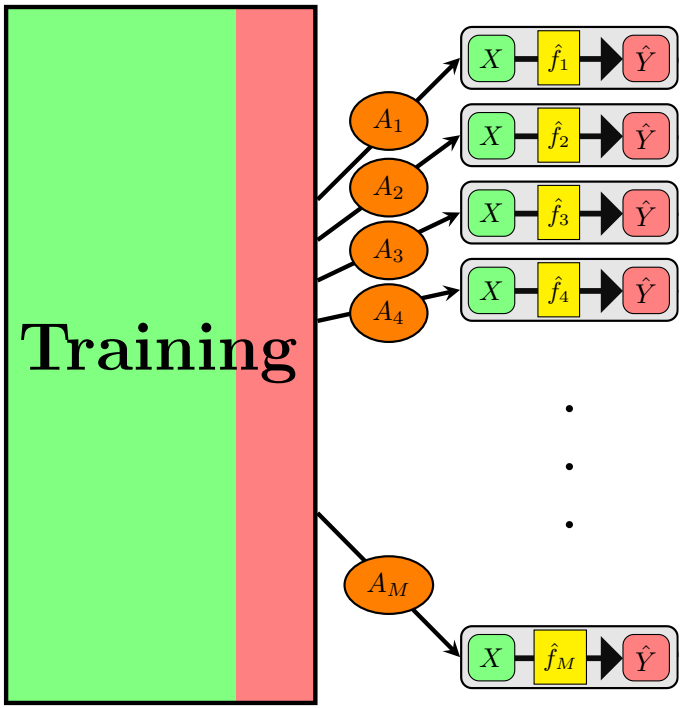
Max Westphal, Werner Brannath

University of Bremen, Germany
Institute for Statistics
Research Training Group π^3

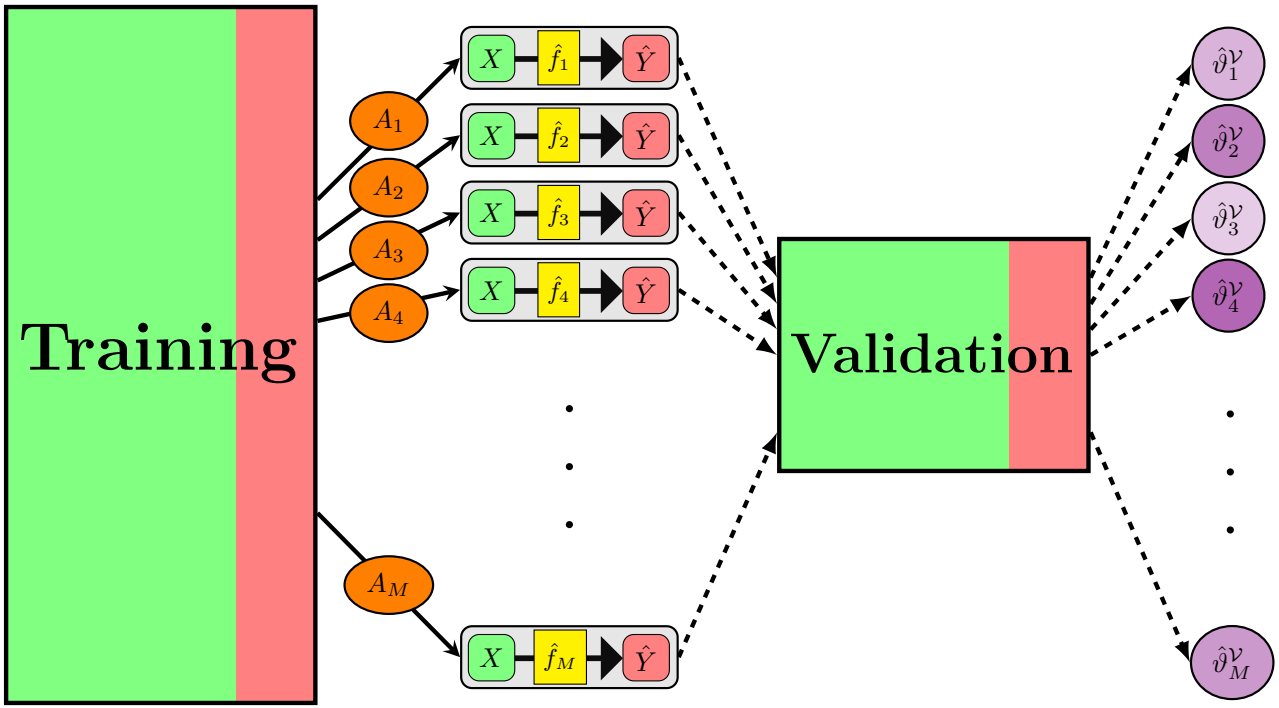
mwestphal@uni-bremen.de
https://github.com/maxwestphal/

ICML 2019, Long Beach
June 11, 2018

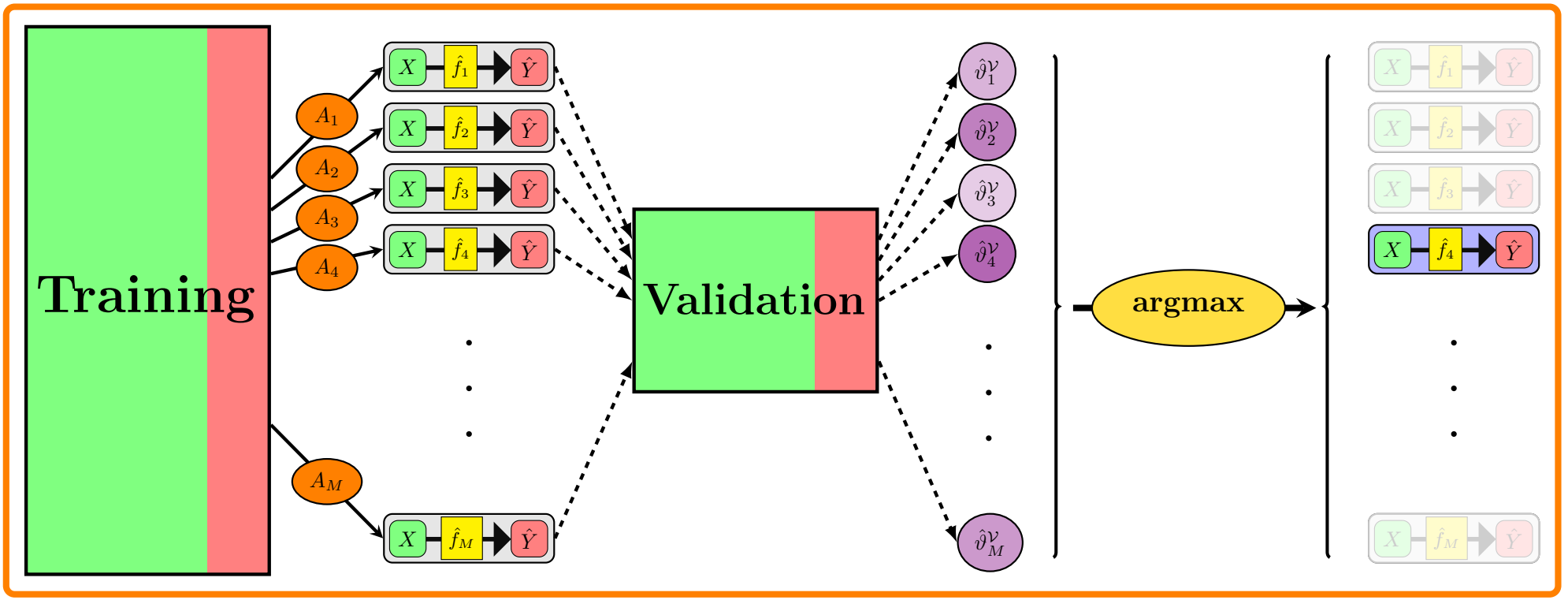
Train-Validation-Test Split



Train-Validation-Test Split



Train-Validation-Test Split

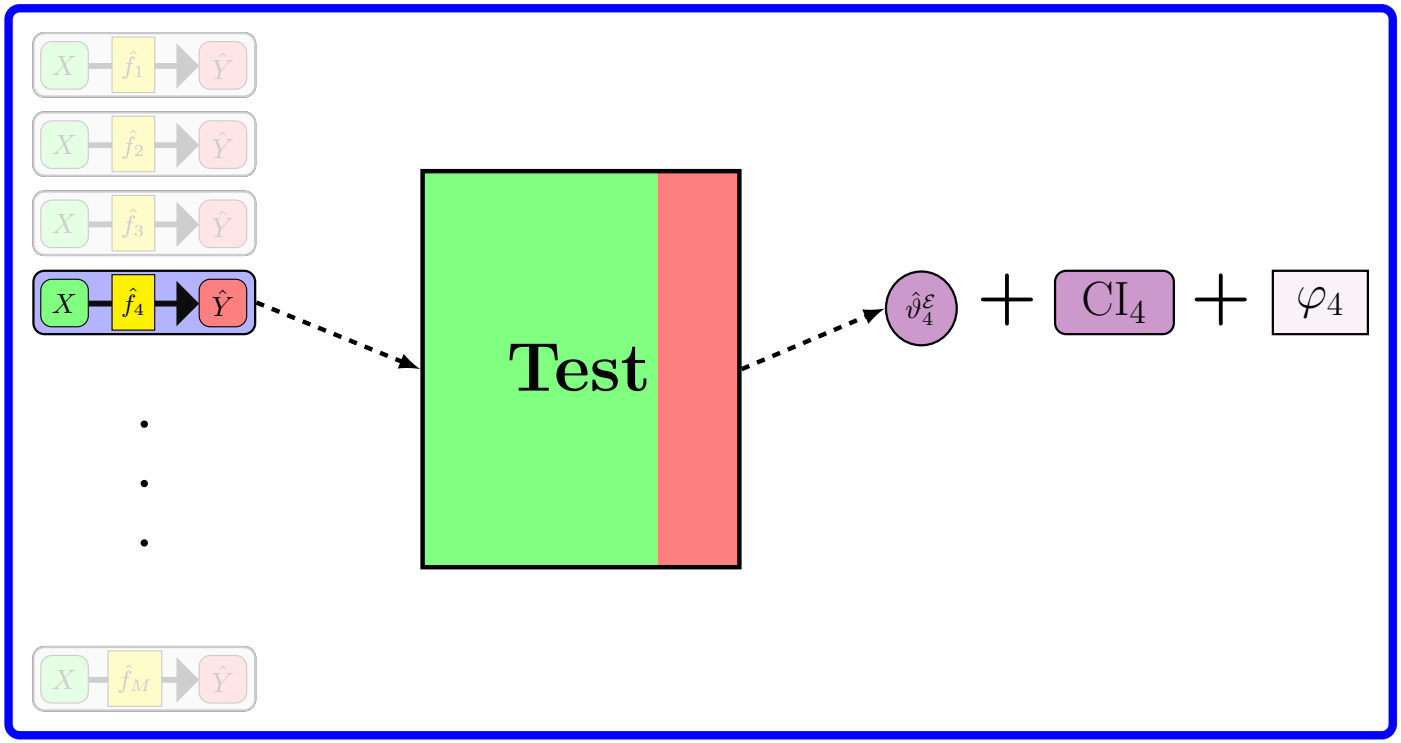


Learning

Train-Validation-Test Split

- Particularly in regulated environments, we need a reliable performance assessment before implementing a prediction model in practice.
- Example: disease diagnosis / prognosis based on clinical data
- Usually recommended strategy:

Evaluate a single final model on independent test data.


X
$\hat{f}_1$
$\hat{Y}$
X
$\hat{f}_2$
$\hat{Y}$
X
$\hat{f}_3$
$\hat{Y}$
X
$\hat{f}_4$
$\hat{Y}$
X
$\hat{f}_M$
$\hat{Y}$
Test
$\hat{\vartheta}_4^{\varepsilon}$
+
CI_4
+
φ_4

Evaluation

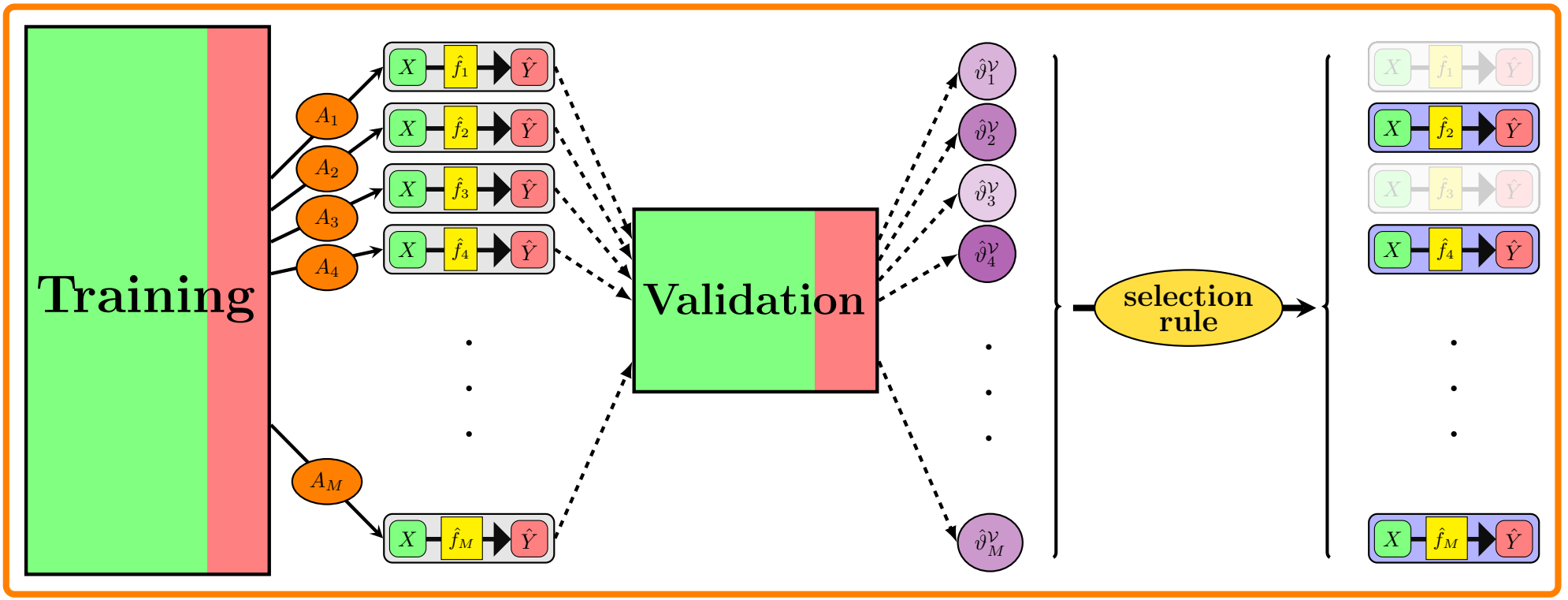
Train-Validation-Test Split

- Particularly in regulated environments, we need a reliable performance assessment before implementing a prediction model in practice.
- Example: disease diagnosis / prognosis based on clinical data
- Usually recommended strategy:

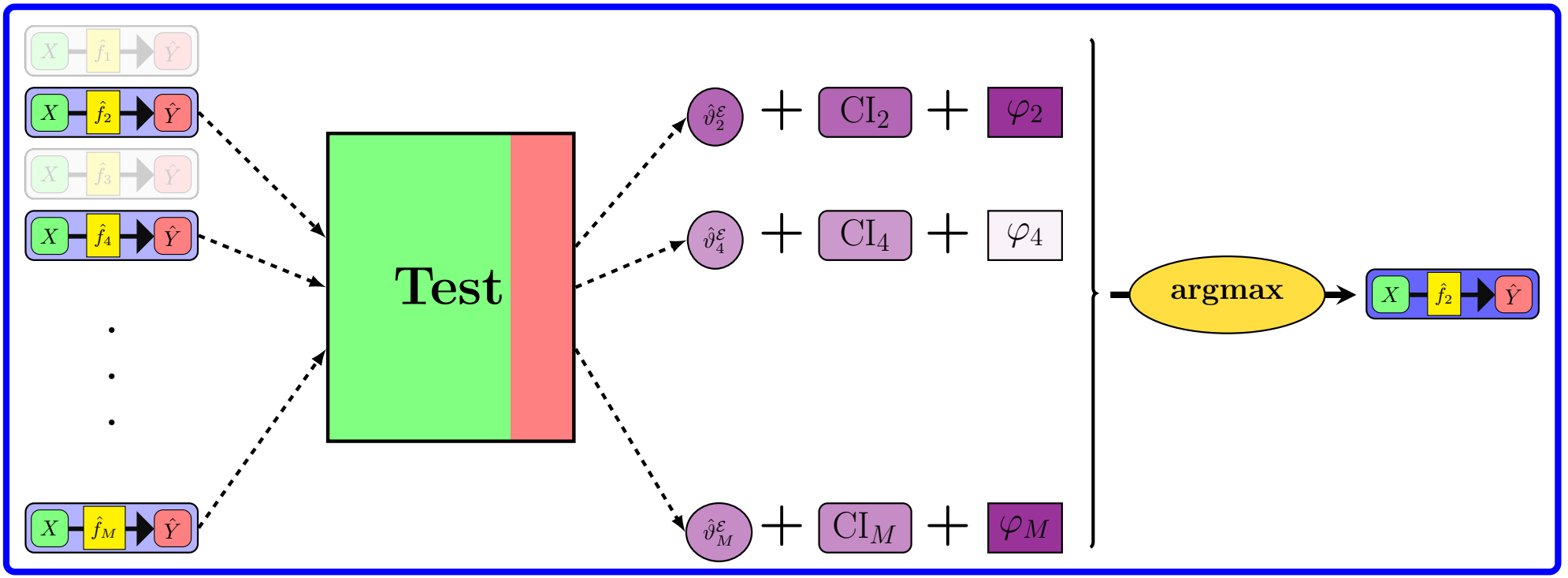
Evaluate a single final model on independent test data!

- Easy-to-use strategy, allowing for a reliable performance assessment and simple inference.
- However, we have no way to fix a bad model selection after having observed the test data.

Simultaneous Model Evaluation

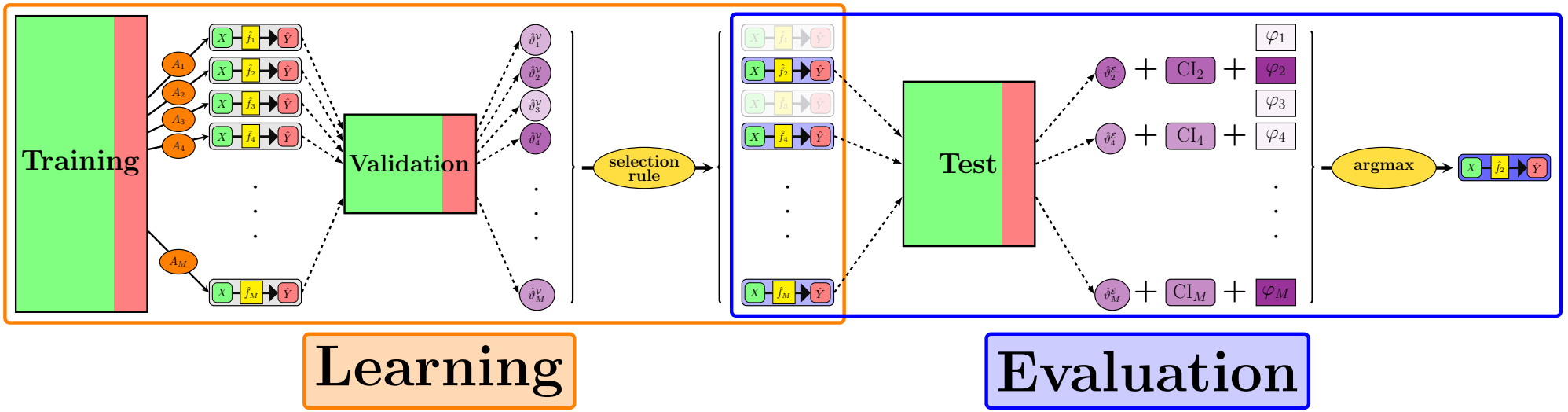


Learning


X
$\hat{f}_1$
$\hat{Y}$
X
$\hat{f}_2$
$\hat{Y}$
X
$\hat{f}_3$
$\hat{Y}$
X
$\hat{f}_4$
$\hat{Y}$
X
$\hat{f}_M$
$\hat{Y}$
Test
$\hat{\vartheta}_2^{\varepsilon}$
CI_2
$\varphi 2$
$\hat{\vartheta}_4^{\varepsilon}$
CI_4
$\varphi 4$
$\hat{\vartheta}_M^{\varepsilon}$
CI_M
φM
argmax
X
$\hat{f}_2$
$\hat{Y}$

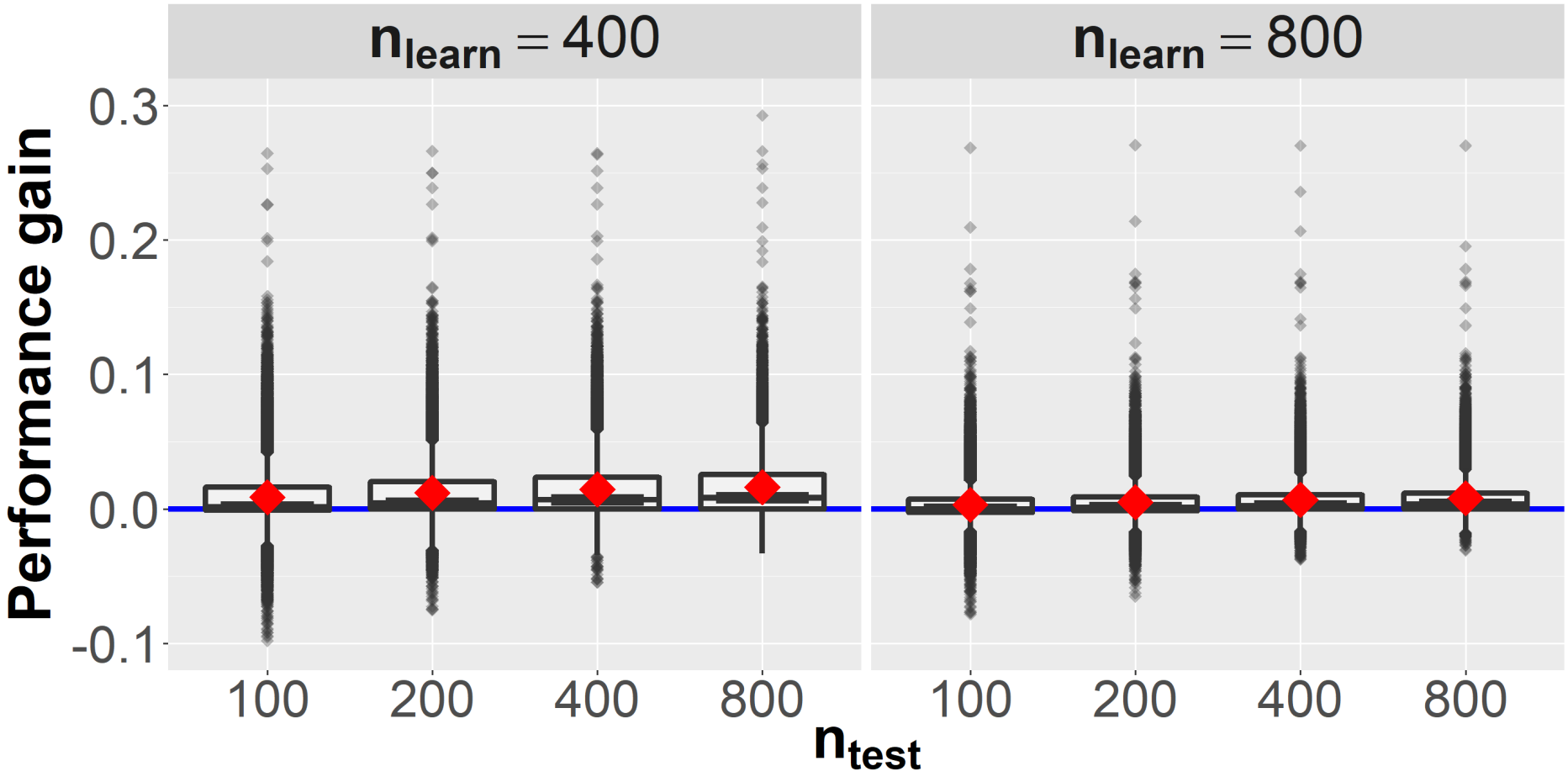
Evaluation

Simulation study

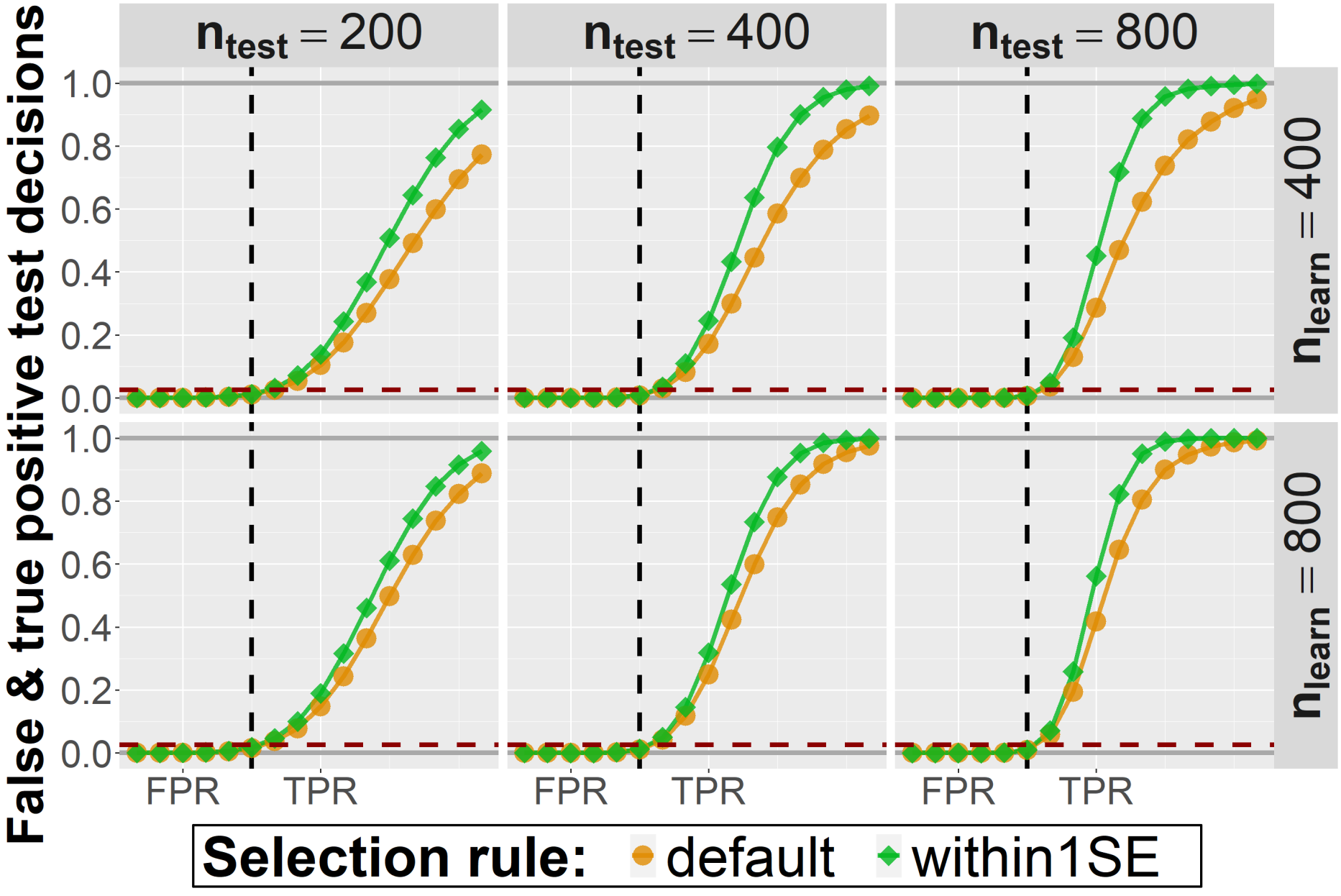


- **Idea:** simulate data and train, select and evaluate binary classifiers in different scenarios
  - 24 artificial classification tasks
  - 72,000 replications of complete ML pipeline
  - 28,800,000 distinct models (EN, CART, SVM, XGB)
- **Goal:** comparison of different evaluation strategies
  - **default**: best validation model only
  - **within 1 SE**: all models within 1 SE of best validation model

Simulation Results



Simulation Results



When in doubt, delay the final model choice to the test data.

- Improvements in model performance and probability to correctly identify a good model in all investigated scenarios.
- Adjustment for multiple comparisons via approximate parametric procedure taking into account model similarity (maxT-approach).

Questions & feedback welcomed!

- mwestphal@uni-bremen.de
- https://github.com/maxwestphal/

POSTER #123 (Pacific Ballroom)