Erasmus
School of
Economics

Machine Learning

FEM31002

Introduction

Part 3

Ilker Birbil

birbil@ese.eur.nl

ERASMUS UNIVERSITEIT ROTTERDAM

FAQ

- Which predictor or classifier should we choose?
- What is the generalization performance?
- What do we mean by unseen (*future*) data?
- How to select the parameters of the algorithms?
- How about trying and comparing multiple algorithms?
- What is the setup for fair performance evaluation and comparison?

In this lecture notes, I have borrowed many ideas from the work of Sebastian Raschka below. We have also used his code repository to prepare our Python scripts.

Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning, S. Raschka, arXiv:1811.1280v2, 3 Dec 2018.

Steps to Delivery

- **Generalization Performance** (algorithm and its hyperparameters are fixed)

  Example 1: *K*-Nearest Neighbors (fixed *K*)

  Example 2: Curve fitting with *m*th order polynomial (fixed *m*)

- **Model Selection** (algorithm is fixed, its 'best' parameters are sought)

  Hyperparameters: *K*=?, *m*=?, number of nodes=?, tree depth=?, …

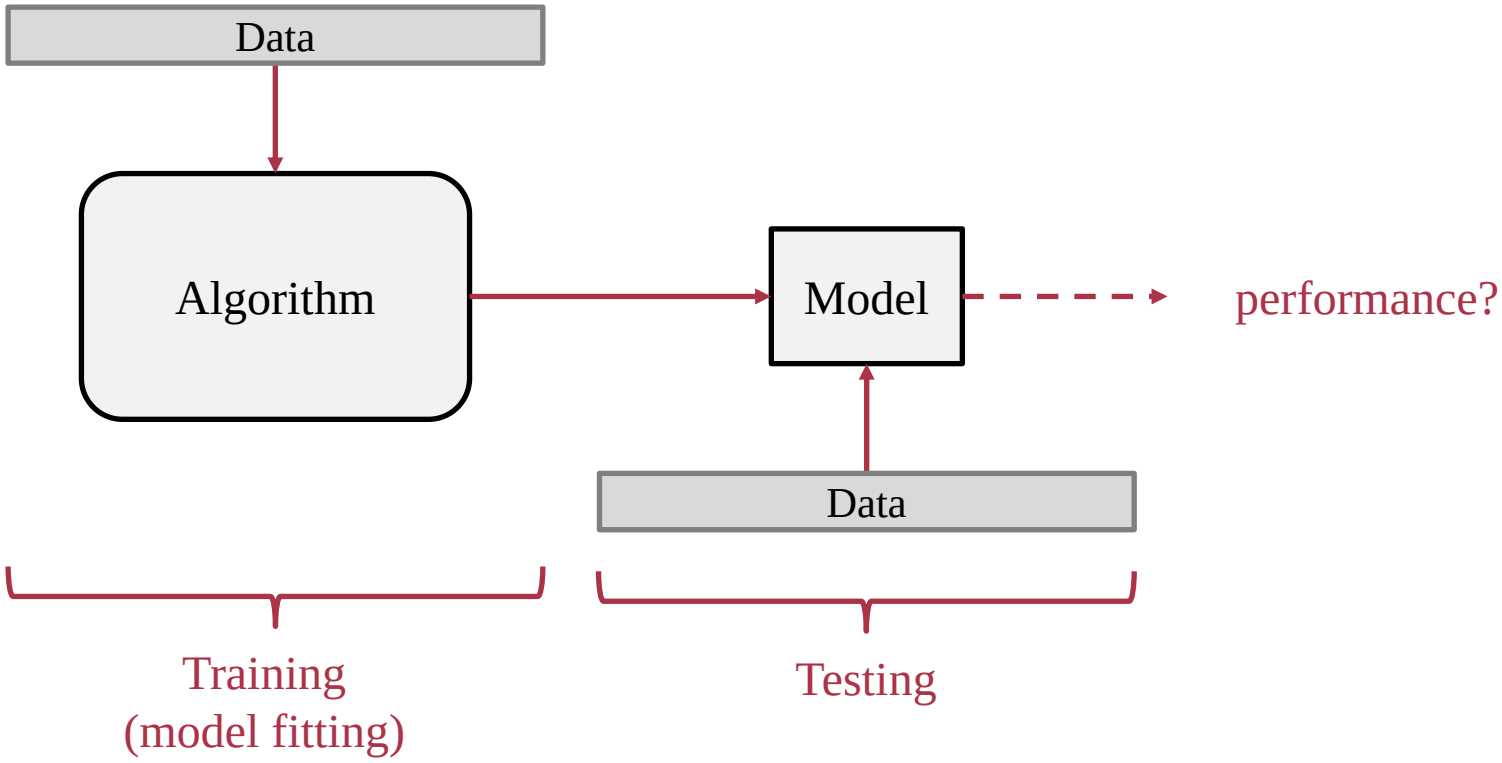
  Different hyperparameters (≠ model parameters) lead to different models

- **Algorithm Selection**

  K-NN?, curve fitting?, neural networks?, decision trees?, …

Generalization Performance

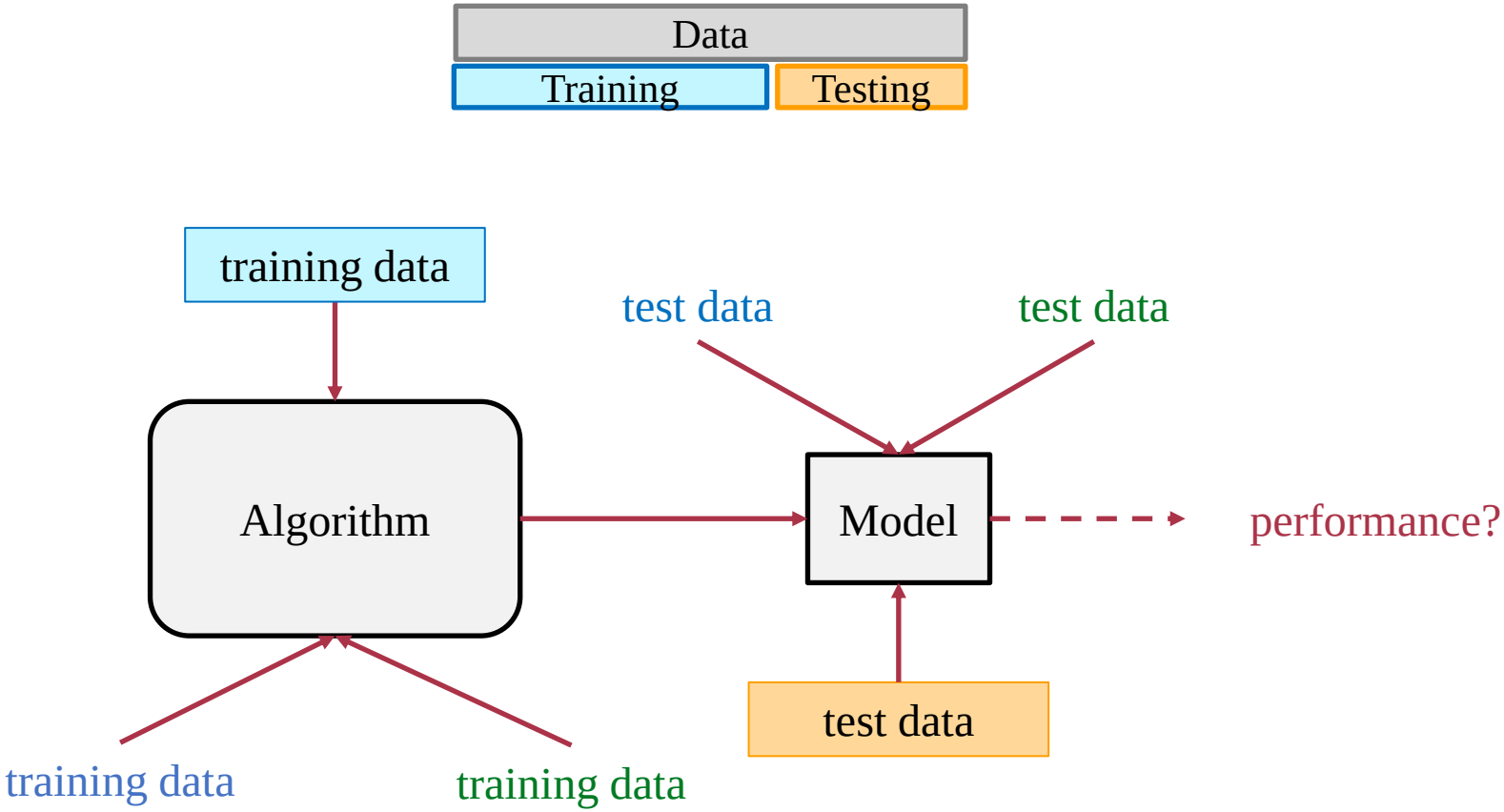
Resubstitution



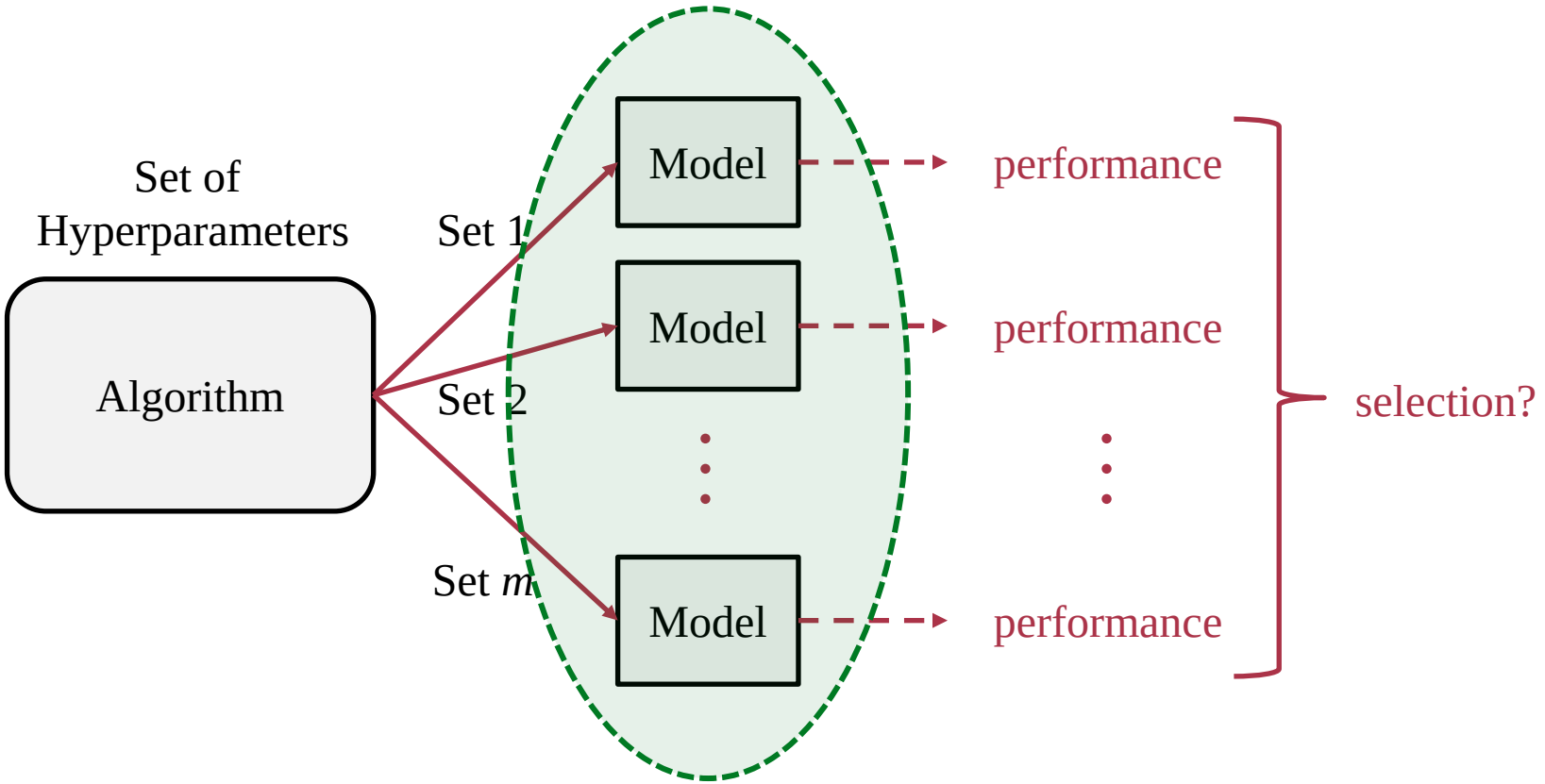
optimistic bias

Generalization Performance

Holdout

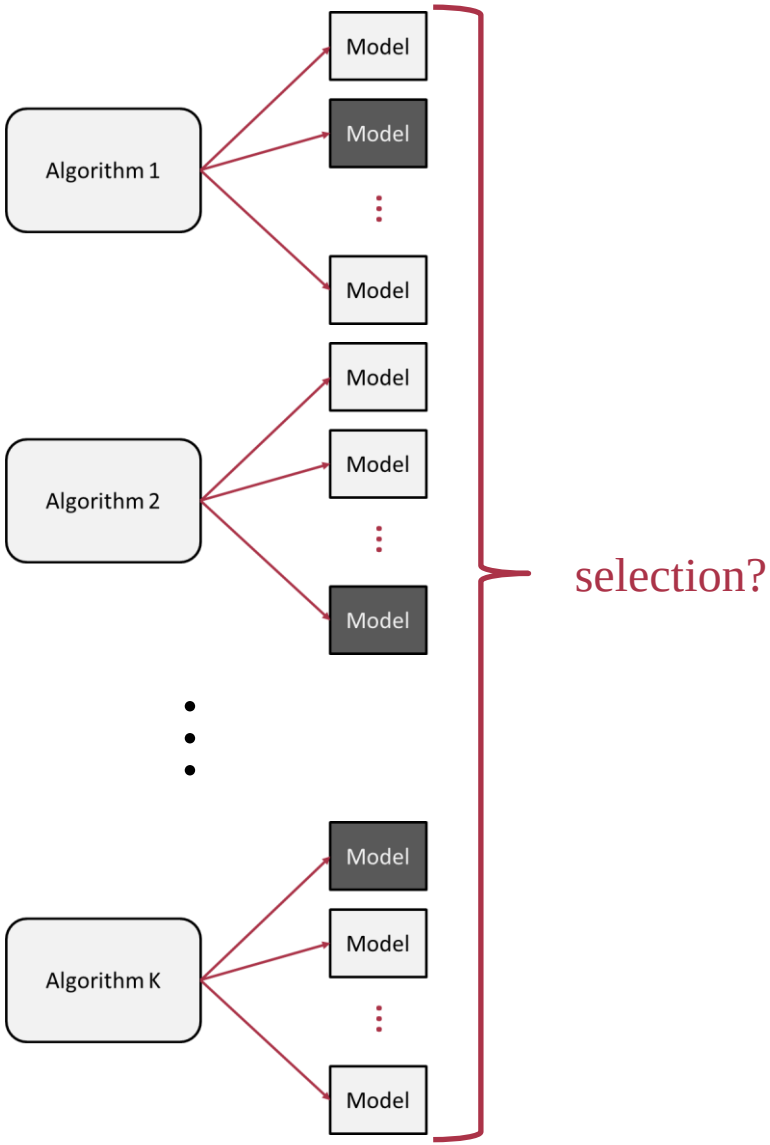


Model Selection

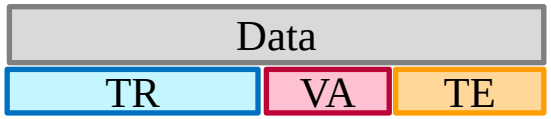


$$\hat{Y} = \hat{f}(X)$$

Algorithm Selection



- Train (TR) - Validation (VA) - Test (TE)

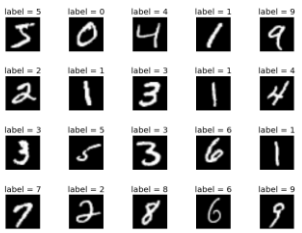


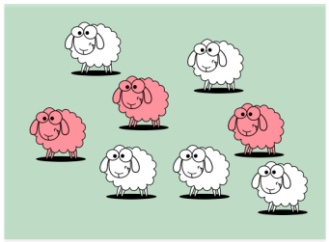
- Statistical tests
- Resampling
  - Cross-validation
  - Bootstrapping
- Bias – Variance trade-off

Settings and Rules

$$\{(x_i, y_i) : 1, \ldots, n\}$$

- Classification problem with i.i.d. data





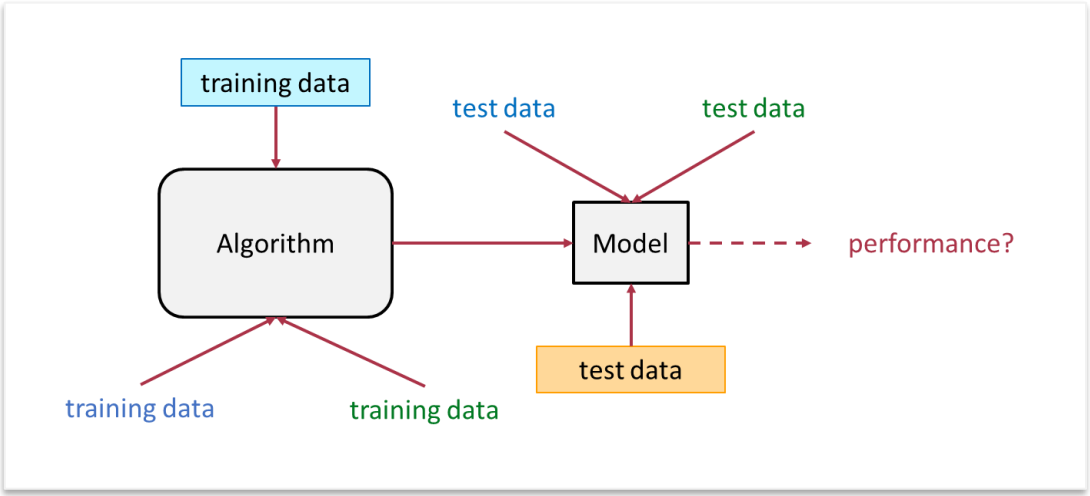
- **Accuracy:** Fraction of correctly classified points

$$\delta = 1 - \frac{1}{n} \sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

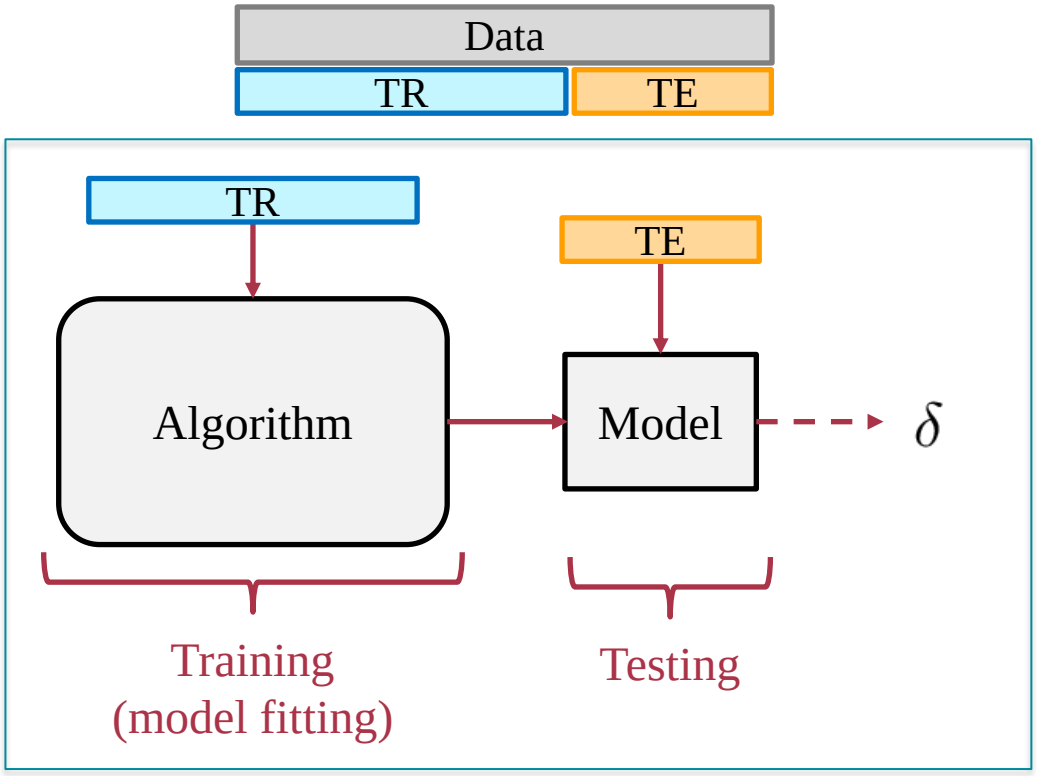
- To avoid bias **use the test set only once**

Generalization Performance

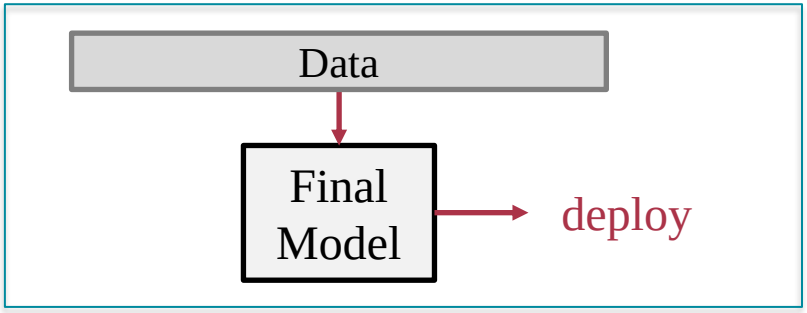
(algorithm and its parameters are fixed)



Holdout



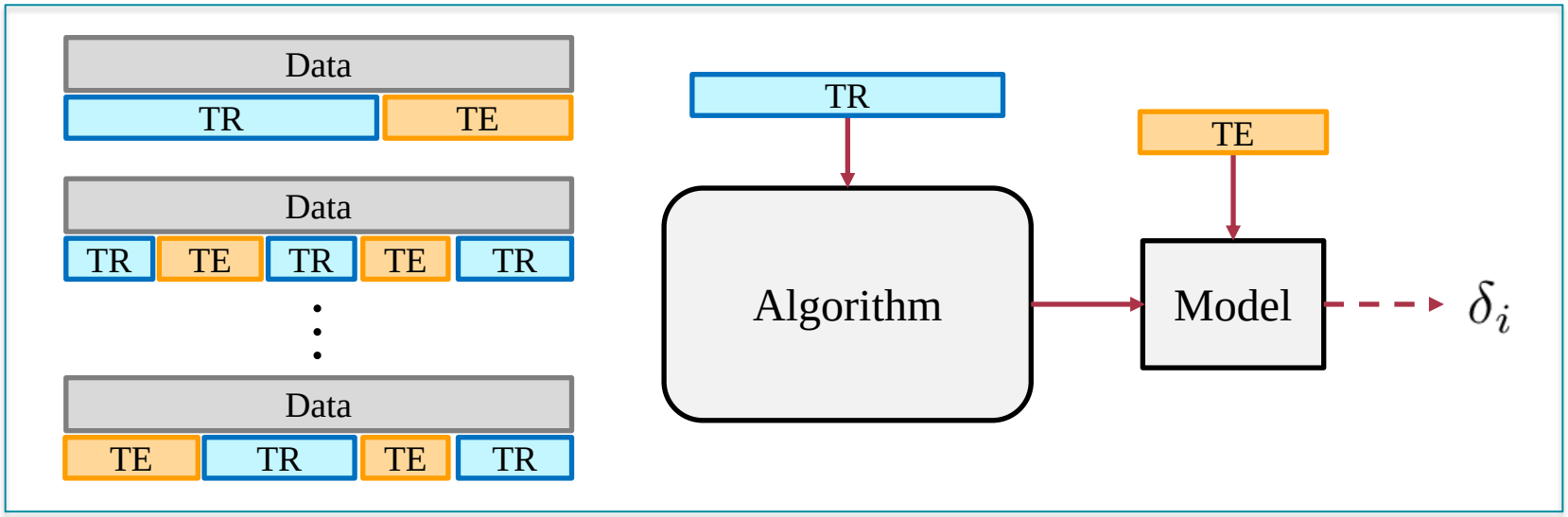
Final Training



How about uncertainty due to random data split?

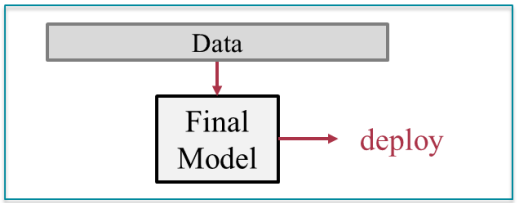
Repeated Holdout

$$i = 1, \ldots, k$$



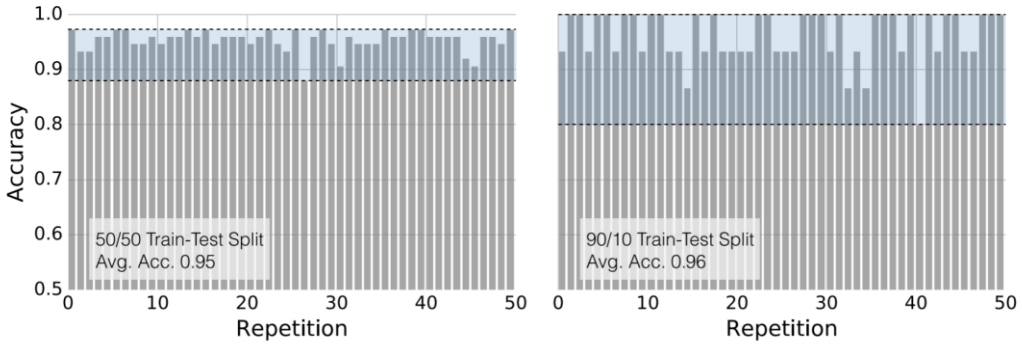
$$\delta = \frac{1}{k} \sum_{i=1}^{k} \delta_i$$

Final Training



Notes on (Repeated) Holdout

- Single split: point estimate (high variance)
- Repeated splits: mean estimate (low variance)
- Large Test Set – Pessimistic Bias **vs.** Small Test Set – High Variance



*

- Stratification: Preserving the class ratios in training and test sets
- Other resampling methods: cross-validation, bootstrap

* *Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning*, S. Raschka, arXiv:1811.1280v2, pg.15, 3 Dec 2018. (link)