Erasmus
School of
Economics

Machine Learning

FEM31002

Introduction

Part 4

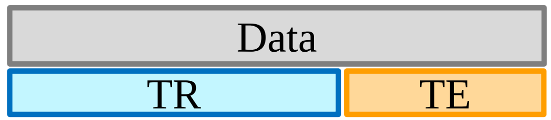
Ilker Birbil

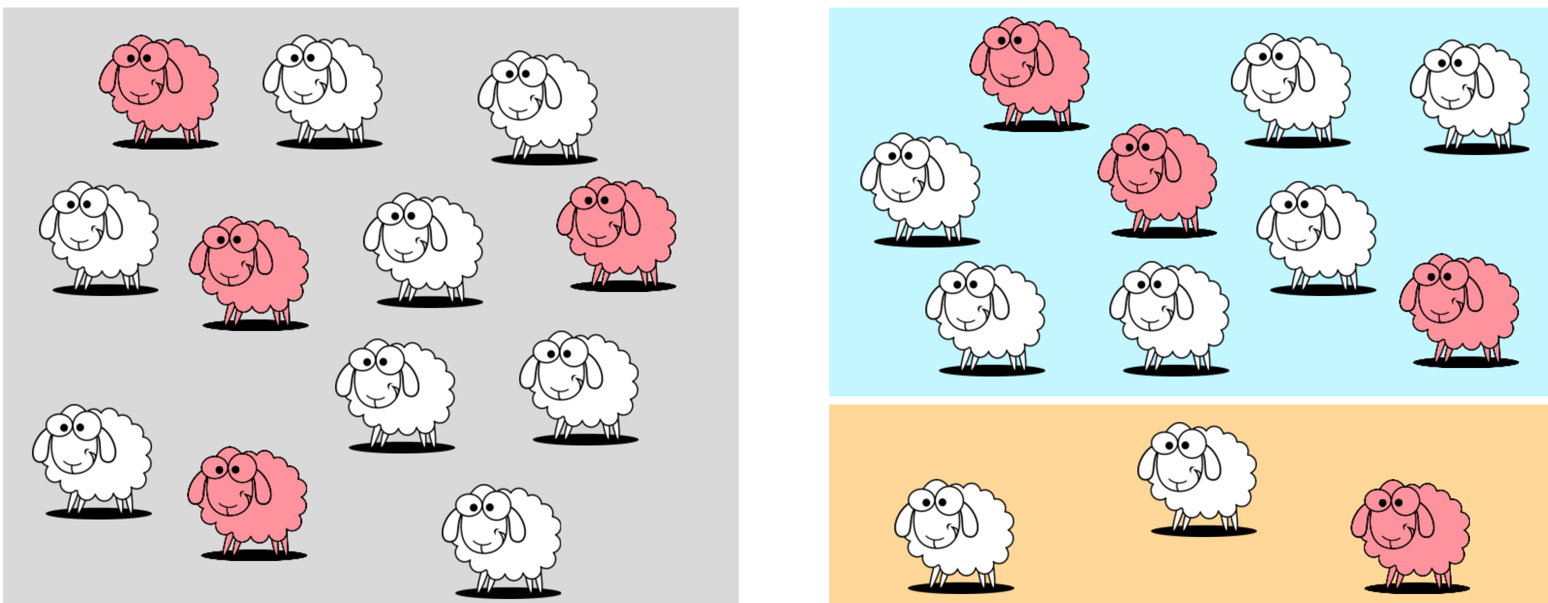
birbil@ese.eur.nl

ERASMUS UNIVERSITEIT ROTTERDAM

Resampling

Stratification

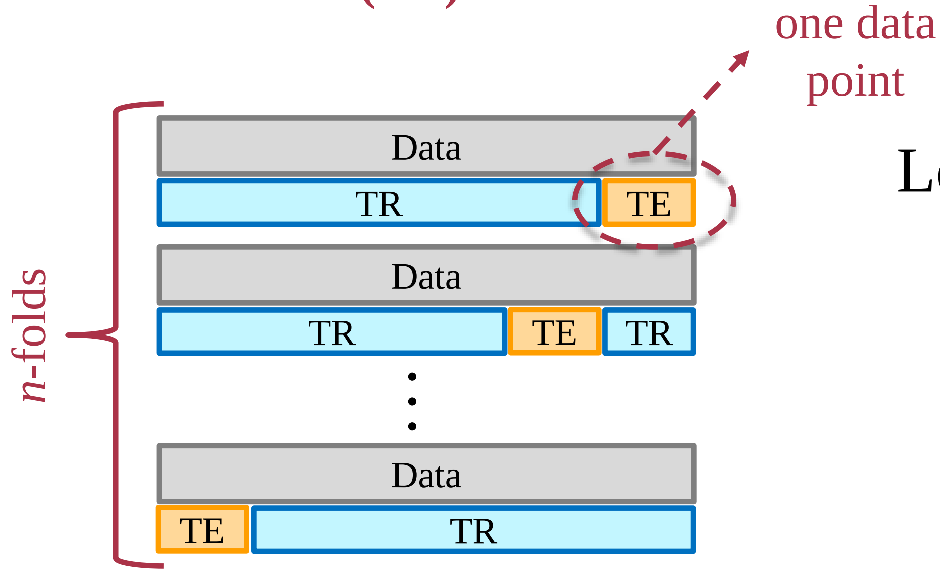




Resampling

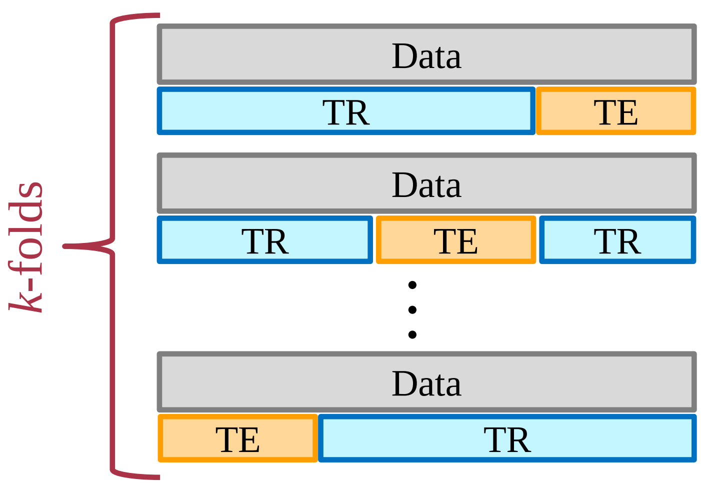
Cross-Validation (CV)

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



Leave-One-Out Cross-Validation (LOOCV)

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



k-Fold Cross-Validation

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

Data is often shuffled before applying CV

Resampling

Bias-Variance Trade-off

Test Error

k ↑	Variance ↑	Bias ↓

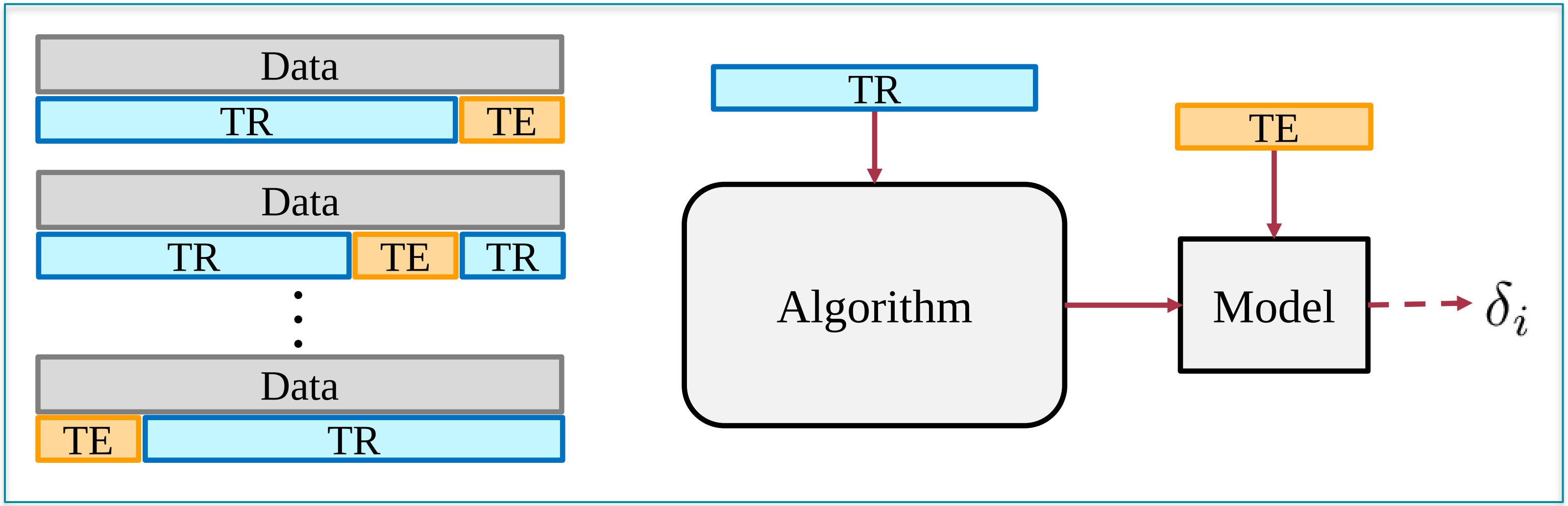
$\left(\frac{k-1}{k}\right)n$: training set size

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

Recall: $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)$

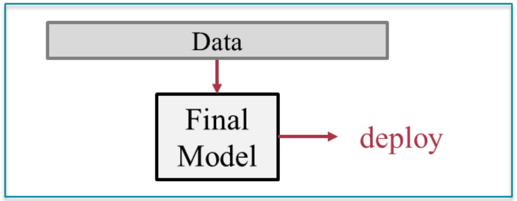
k-fold Cross-Validation

$i = 1, \ldots, k$

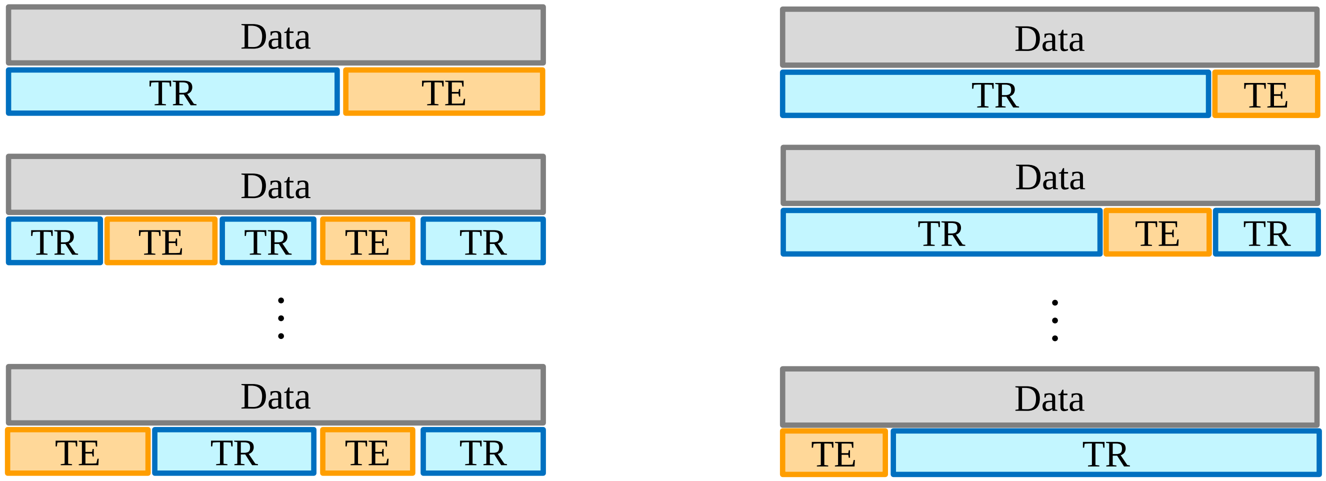


$$\delta = \frac{1}{k} \sum_{i=1}^{k} \delta_i$$

Final Training



Repeated Holdout vs. *k*-fold Cross-Validation



Resampling

Bootstrap

1 2 3 n

3 9 5 3 3 9

8 9 5 12 1 2

⋮

1 9 5 6 1 37

- with replacement
- training set sample size is n
- the rest (out-of-bag) can be used for testing
- training set can be used for testing (resubstitution)

- allows collection of statistics (*e.g.*, variance of regression parameters)
- plays nicely with small data sets

Resampling

Statistics

optimistic bias

$$\underbrace{\delta^r = \frac{1}{b}\sum_{j=1}^{b} \delta_j^r}_{\text{resubstitution accuracy}}$$

$$\underbrace{\delta^h = \frac{1}{b}\sum_{j=1}^{b} \delta_j^h}_{\text{holdout accuracy}}$$

pessimistic bias

$$\underbrace{\delta^\bullet \pm t\sqrt{\frac{1}{b-1}\sum_{j=1}^{b}(\delta_j^\bullet - \delta^\bullet)^2}}_{\text{confidence interval (under normality assumption)}}$$

Example: $b = 100, t_{95} = 1.984$

Resampling

.632 Estimate

$$\mathbb{P}(\text{a sample is chosen}) = 1 - \left(1 - \frac{1}{n}\right)^n \underset{n>>0}{\approx} 0.632$$

$$\delta = \frac{1}{b}\sum_{j=1}^{b}(0.632\ \delta_j^h + 0.368\ \delta_j^r)$$

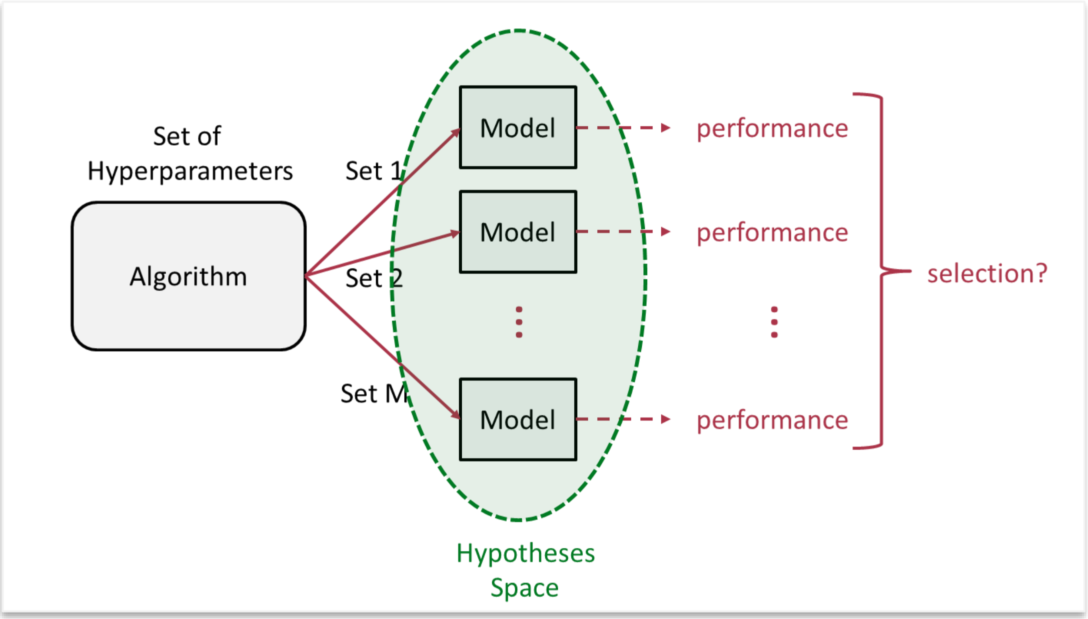
$$= 0.632\ \delta^h + 0.368\ \delta^r$$

slightly optimistic bias

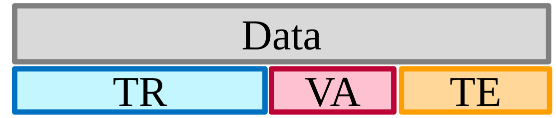
The case where the convex combination weights are not fixed but evaluated with the dataset is called the .632+ Bootstrap Method (Efron and Tibshirani, 1997).

Model Selection

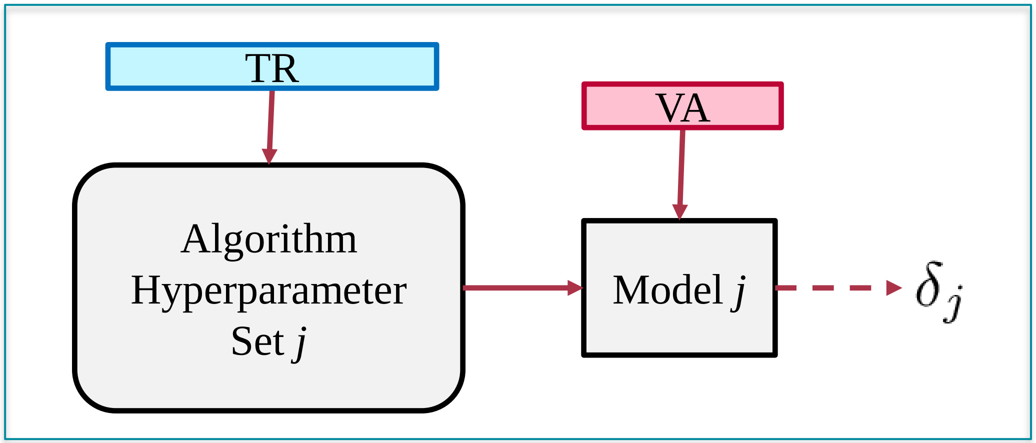
(algorithm is fixed, its ‘best’ parameters are sought)



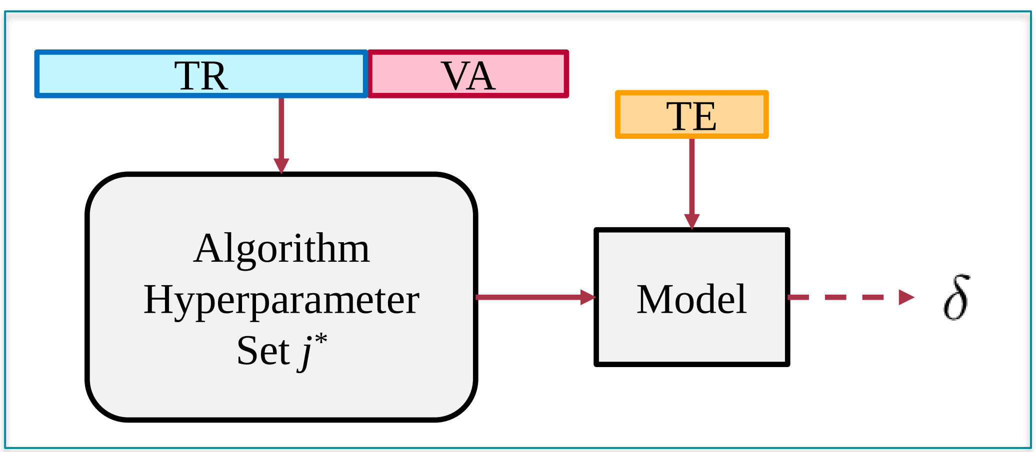
Three-way Holdout



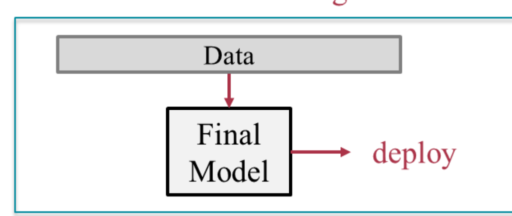
$$j = 1, \ldots, m$$



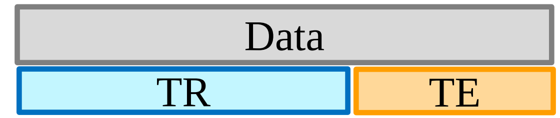
$$j^* = \arg\max\{\delta_j : j = 1, \ldots, m\}$$



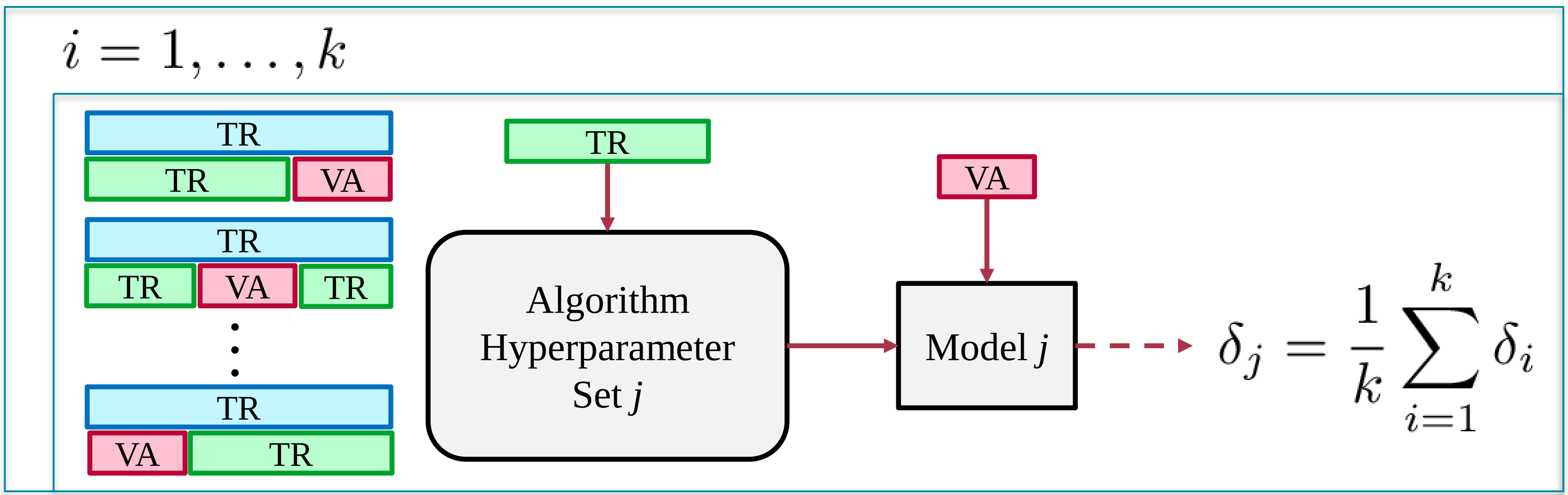
Final Training



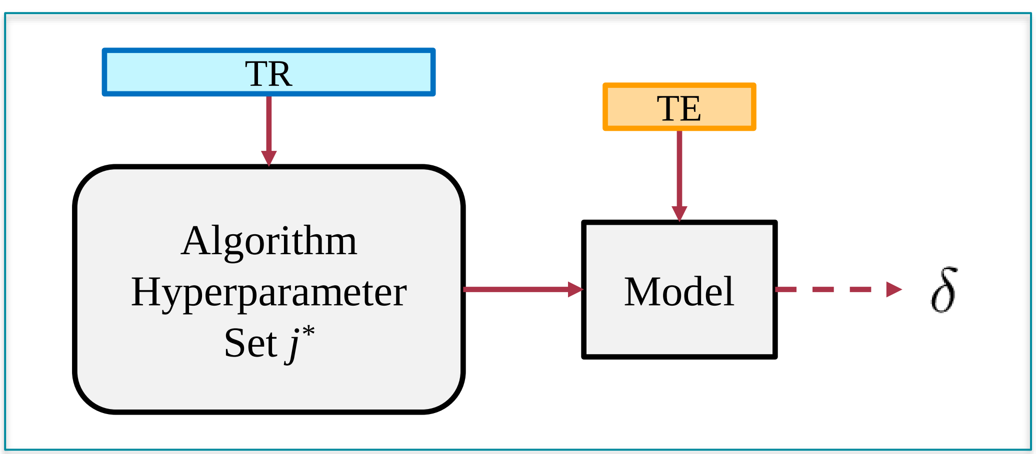
k-fold Cross-Validation

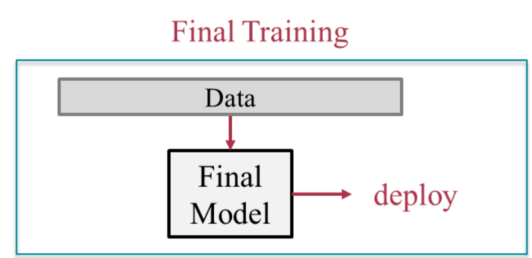


$$j = 1, \ldots, m$$



$$j^* = \arg\max\{\delta_j : j = 1, \ldots, m\}$$



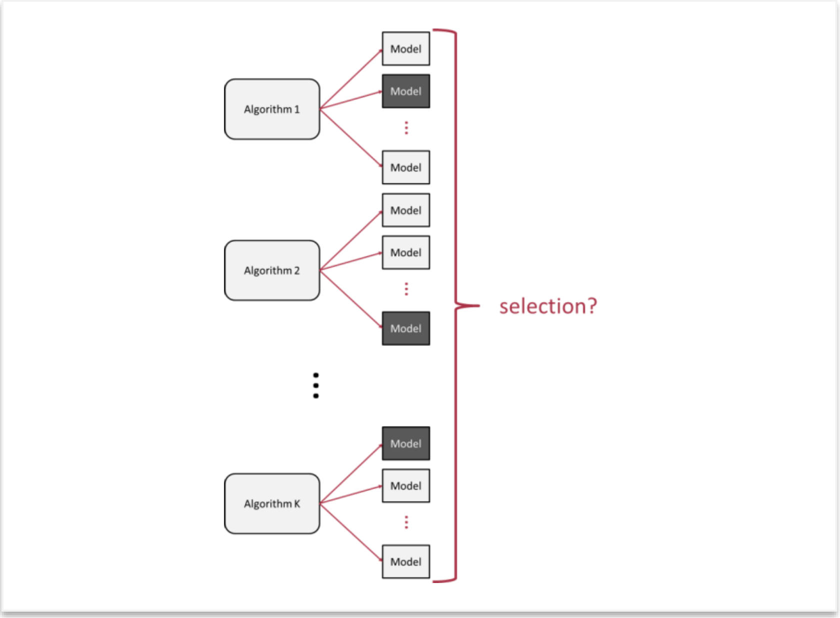


Model Selection Notes

- Overall generalization performance depends on the test set
- *k*-fold CV takes a long time with large datasets (or slow algorithms)
- Three-way holdout is faster when dataset is large
- Holdout method is occasionally called as 2-fold CV (not exactly true)
- There is no universal *k* value in *k*-fold CV (usually 5 or 10, though)

- Roughly: LOOCV (small dataset), *k*-fold CV or three-way holdout (large dataset)

Algorithm Selection



Model Comparison

- Comparing two models
  - Difference tests based on *z*-scores
  - McNemar Test
- Comparing multiple models
  - Cochran *Q* Test
  - *F*-test
  - Paired *t*-tests, combined *F*-tests
  - Nested cross-validation
- And more…

Comparing Two Models

McNemar Test

		Model 2 ✓	Model 2 ✗
Model 1	✓	a	b
	✗	c	d

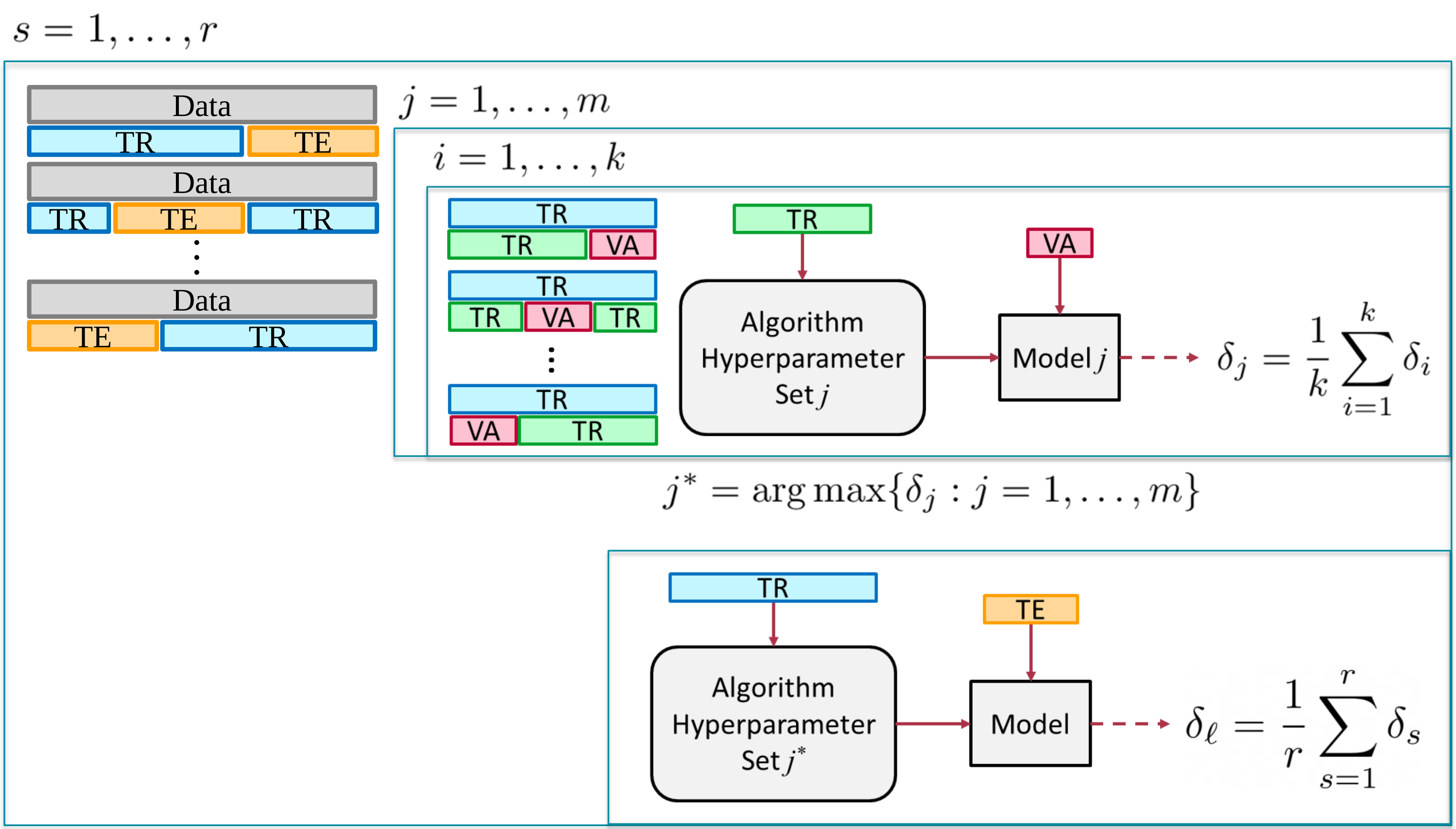
$$a + b + c + d = n$$

$$H_0 : \mathbb{P}(b) = \mathbb{P}(c)$$

$$\text{Test statistic: } \chi^2 = \frac{(|b - c| - 1)^2}{b + c}$$

1. Pick a significance level (e.g., 0.05)
2. Evaluate *p*-value
3. Accept or reject the null hypothesis

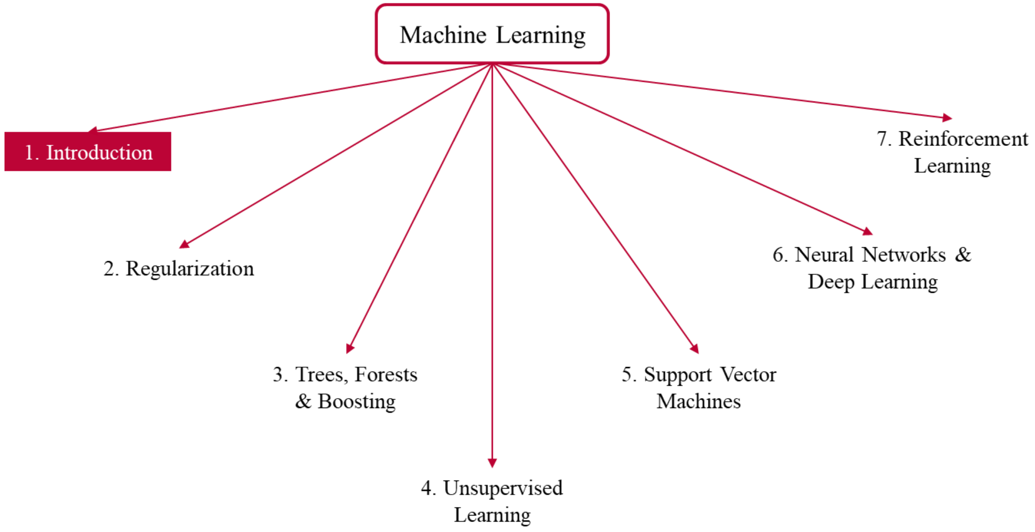
Nested *k*-fold Cross-Validation



δ_ℓ : Performance of algorithm $\ell = 1, \ldots, K$

Then select the best performing algorithm (model) among *K*

Outline



- Overview of the course
- Supervised Learning vs. Unsupervised Learning
- Train-test errors and overfitting
- Bias vs. Variance
- Bayes Classifier vs. K-Nearest Neighbor (KNN)
- Cross validation and bootstrap
- Model evaluation and algorithm comparison