Biostatistics
Week XIV

Ege Ülgen, MD, PhD

5 January 2023

ACIBADEM
MEHMET ALİ AYDINLAR
ÜNİVERSİTESİ

Two issues

- **Model selection**
  - How do we select the **optimal** model for a given problem
- **Validation**
  - Once we have chosen a model, how do we estimate its **true error rate**?
  - (the error rate when tested on the entire population)

Naïve Approach

- Use the entire training data to select the optimal model, *then* estimate the error rate
- Two fundamental problems:
  - The final model will likely **overfit** the training data (i.e., it will not be able to generalize to new data)
  - The error rate estimate will be overly optimistic (lower than the true error rate)

The Holdout Method

Training Set
Test Set

- Split dataset in two groups
  - Training set: used to obtain the model
  - Test set: used to estimate the error rate of the trained model
- Since it is a single train-and-test experiment, the holdout estimate of error rate will be misleading if we happen to get an "**unfortunate**" split

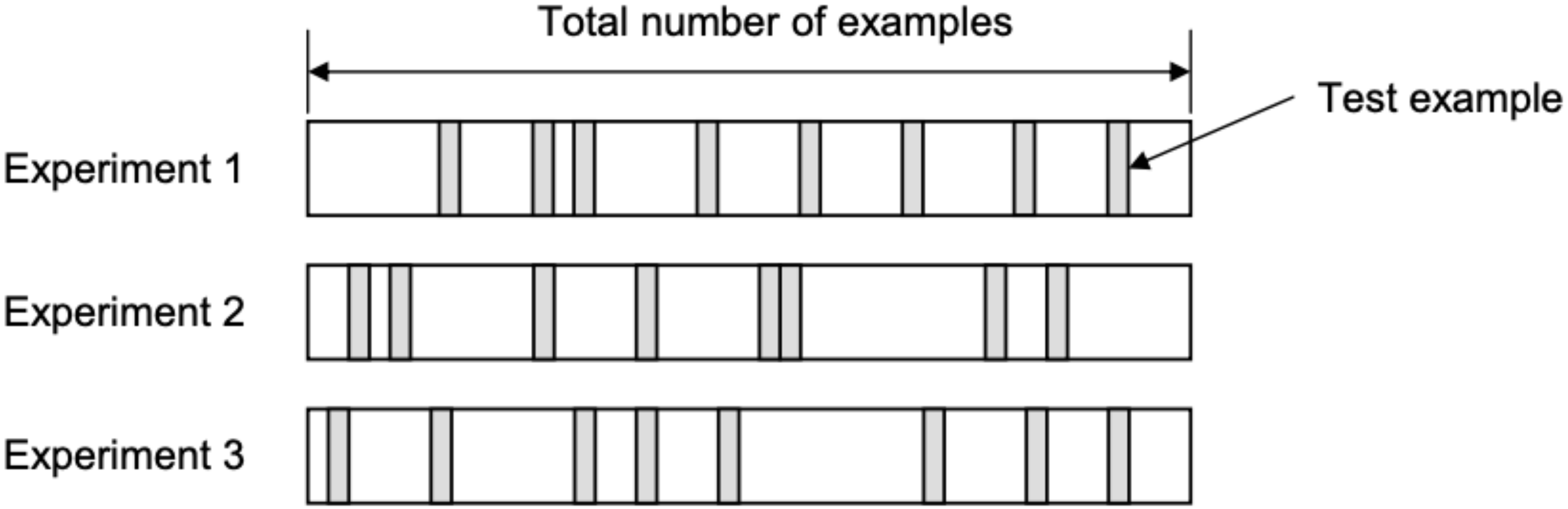
Resampling Methods

- The limitations of the holdout can be overcome with a family of **resampling methods** at the expense of higher computational cost:
  - Cross Validation
    - Random Subsampling
    - K-Fold Cross-Validation
    - Leave-one-out Cross-Validation
  - Bootstrap

https://www.ic.unicamp.br/~rocha/teaching/2011s2/mc906/aulas/pr_l13.pdf

Random Subsampling

- Random subsampling performs K data splits of size m (without replacement)

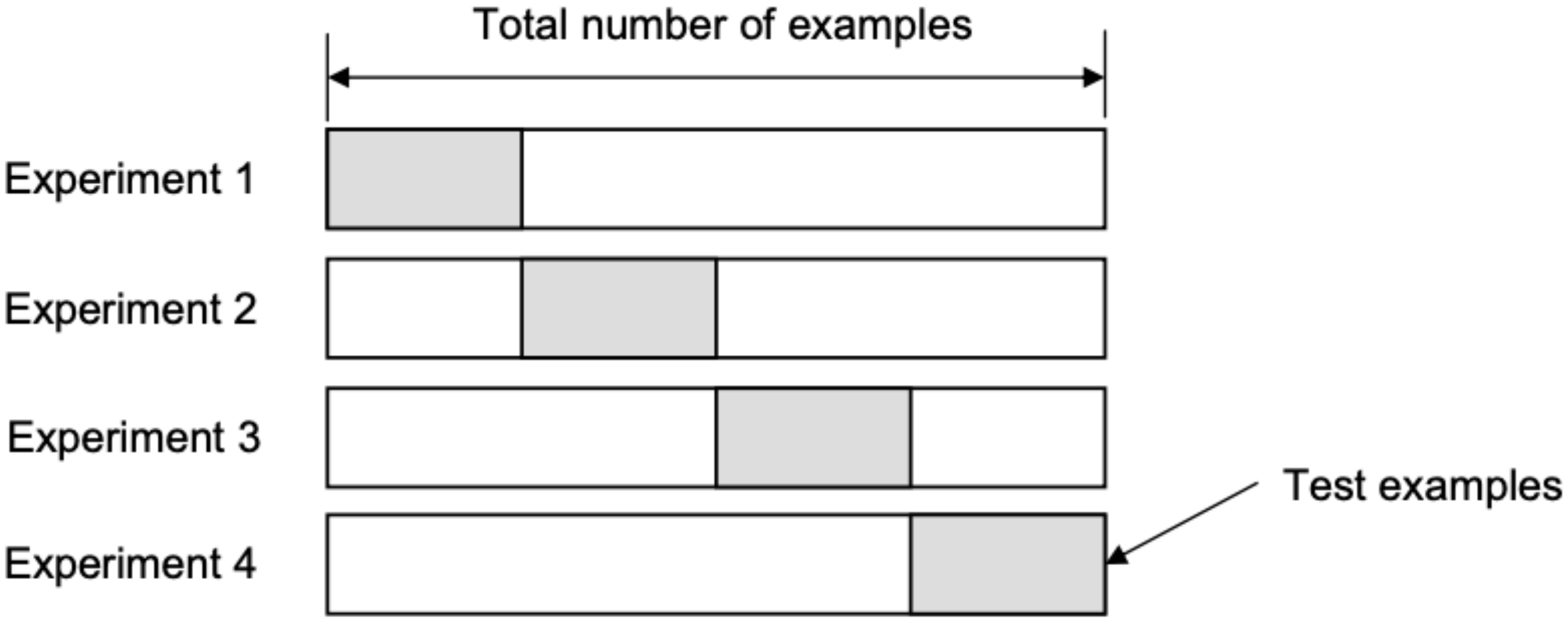


$$E = \frac{1}{K}\sum_{i=1}^{K} E_i$$

https://www.ic.unicamp.br/~rocha/teaching/2011s2/mc906/aulas/pr_l13.pdf

K-Fold Cross-Validation

- Each instance is eventually used for both training and testing

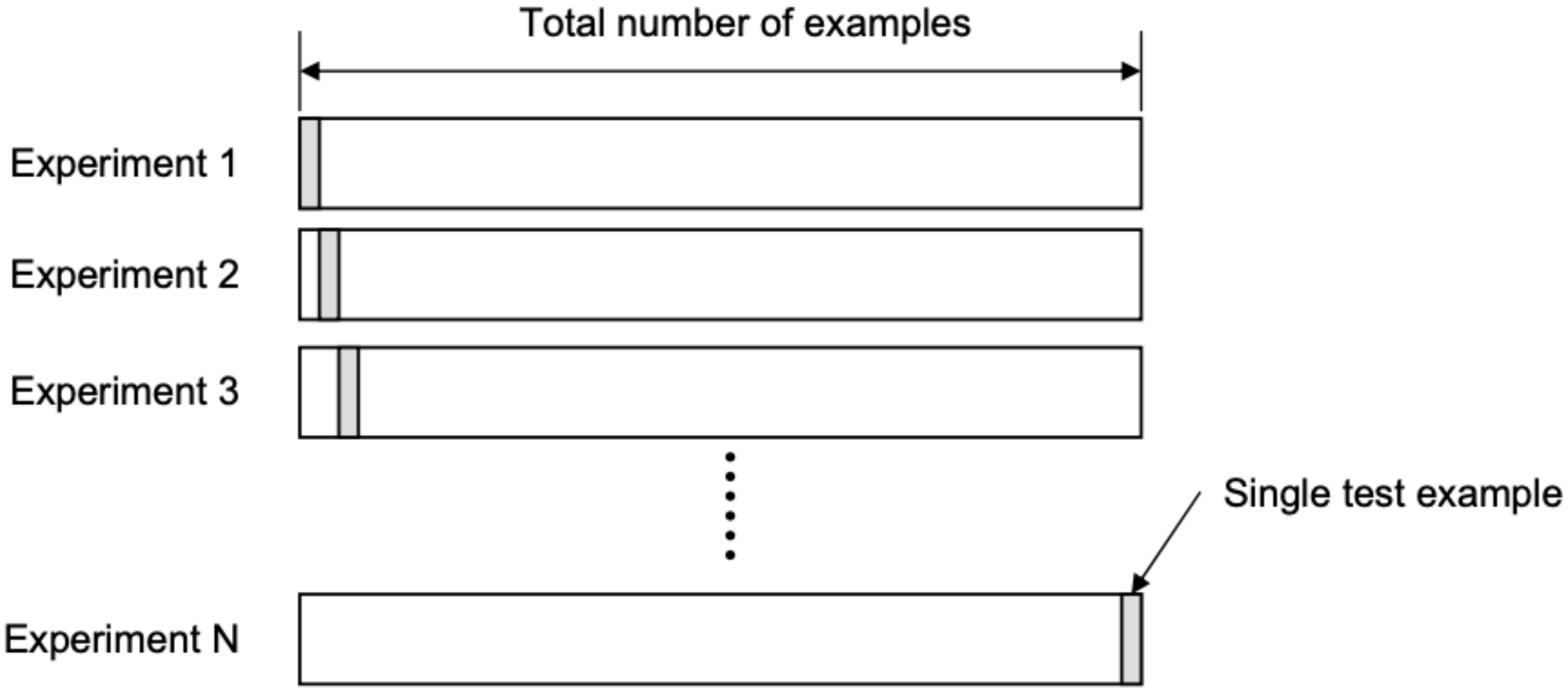


$$E = \frac{1}{K}\sum_{i=1}^{K} E_i$$

https://www.ic.unicamp.br/~rocha/teaching/2011s2/mc906/aulas/pr_l13.pdf

Leave-one-out Cross-Validation

- Leave-one-out is the degenerate case of K-Fold Cross Validation, where K is chosen as the total number of examples



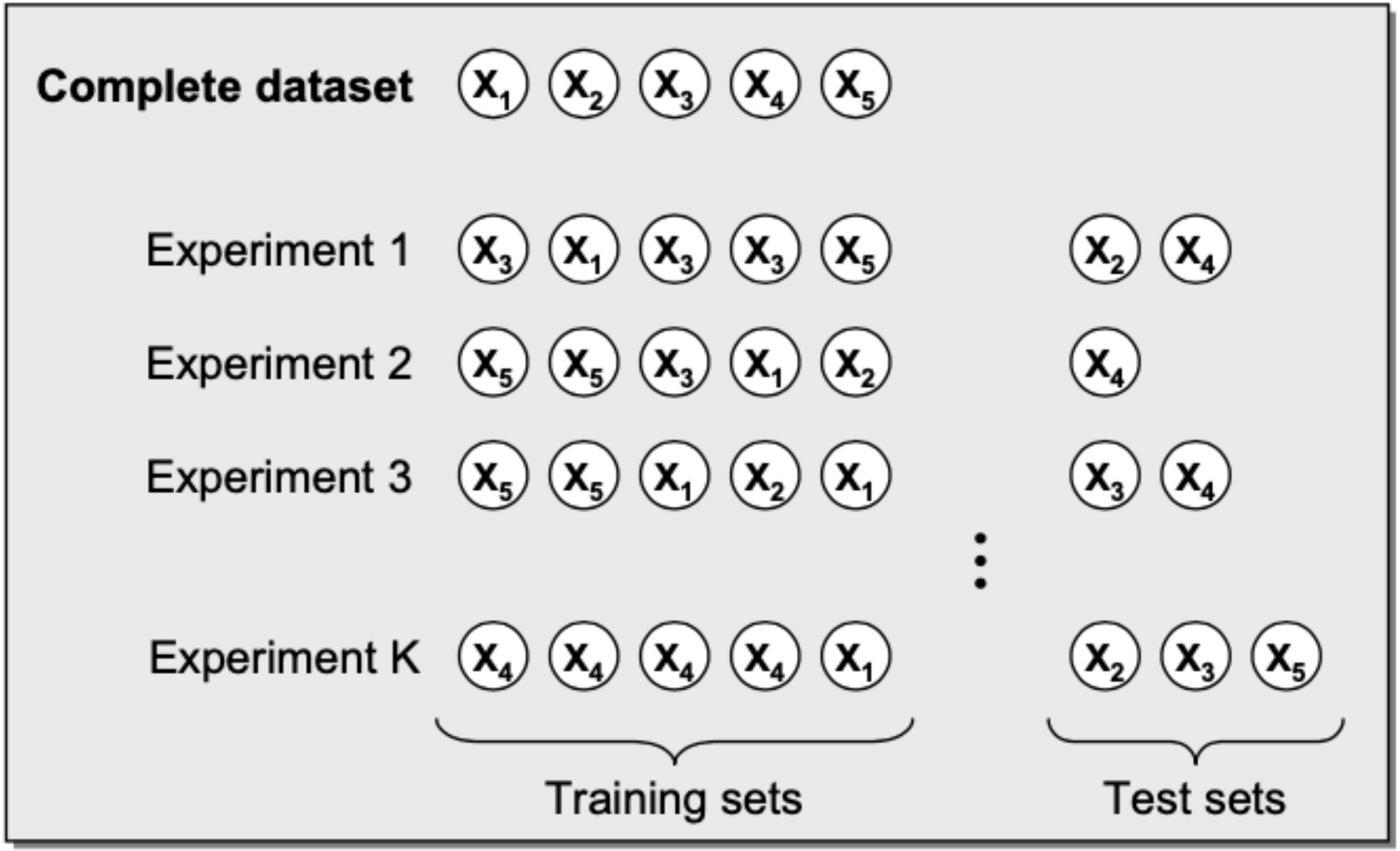
$$E = \frac{1}{K}\sum_{i=1}^{K} E_i$$

Tips on Cross Validation

- With larger K
  - The bias of the true error rate estimator will be small (the estimator will be very accurate)
  - The variance of the true error rate estimator will be large
  - The computational time will be very large as well
- A common choice for K-fold CV is K = 10
- For large datasets, even 3-Fold Cross Validation will be quite accurate
- For very sparse datasets, we may have to use leave-one-out in order to train on as many examples as possible

https://www.ic.unicamp.br/~rocha/teaching/2011s2/mc906/aulas/pr_l13.pdf

The Bootstrap



Complete dataset X_1 X_2 X_3 X_4 X_5

Experiment 1 X_3 X_1 X_3 X_3 X_5 X_2 X_4

Experiment 2 X_5 X_5 X_3 X_1 X_2 X_4

Experiment 3 X_5 X_5 X_1 X_2 X_1 X_3 X_4

⋮

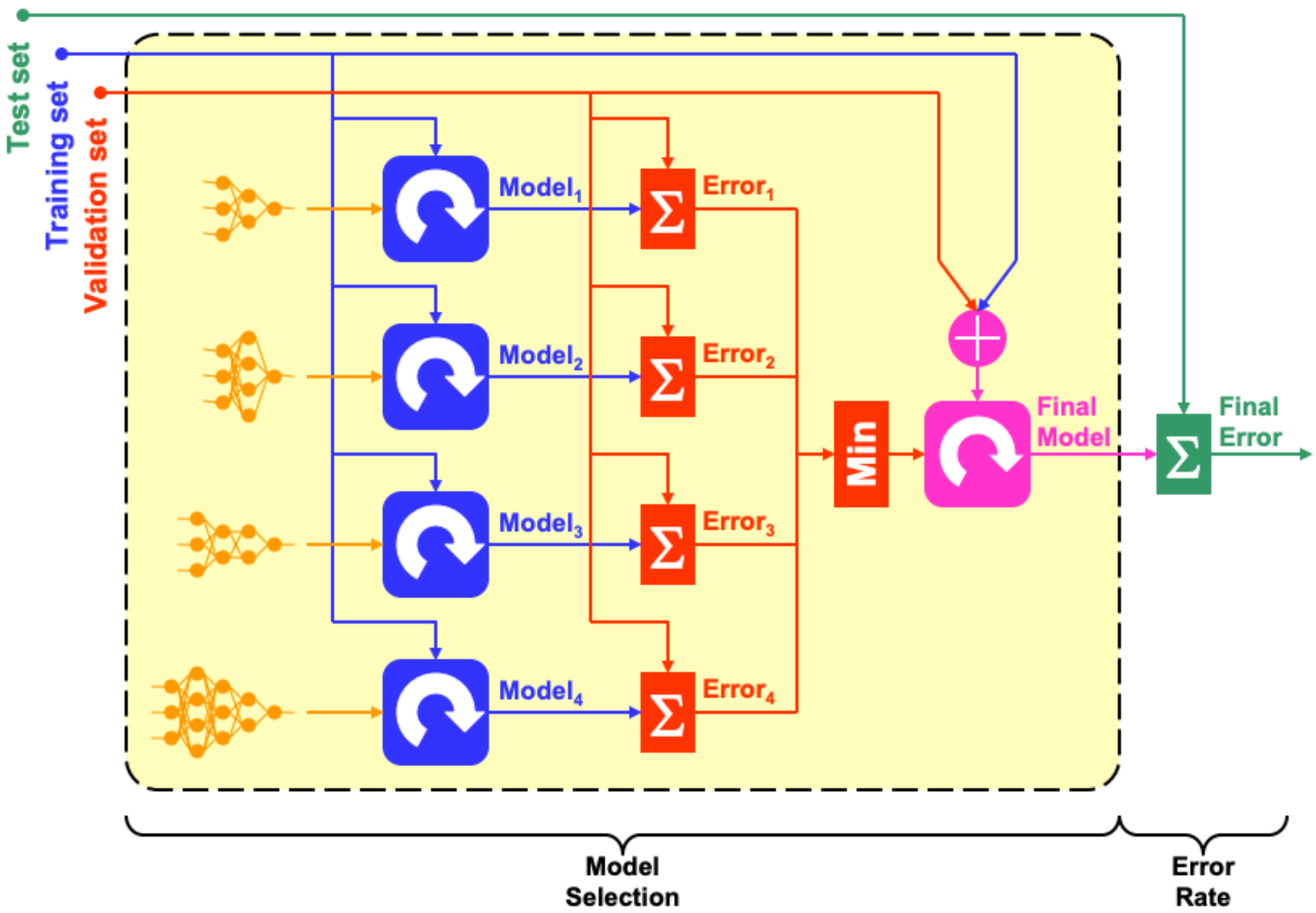
Experiment K X_4 X_4 X_4 X_4 X_1 X_2 X_3 X_5

Training sets

Test sets

https://www.ic.unicamp.br/~rocha/teaching/2011s2/mc906/aulas/pr_l13.pdf

Three-way data splits



Brief Summary

- Resampling methods:
  - Cross Validation
    - Random Subsampling
    - K-Fold Cross-Validation
    - Leave-one-out Cross-Validation
  - Bootstrap