**MA477: Data Science**
**Lesson 5 Outline — 16 January 2026**
United States Military Academy, West Point
Instructor: MAJ Patrick Kuiper

# 1 Administrative

- Quiz and Pset next week

# 2 Student Review

# 3 Lesson 5 Resampling, CV, Bootstrapping - Lesson Objectives

By the end of this lesson, cadets will be able to:

- Understand how resampling methods are used in model assessment.
- Know how to use cross-validation to estimate the test error of a statistical model.
- Know the differences between leave-one-out cross-validation and k-fold cross-validation.
- Understand how the bias-variance tradeoff applies when deciding on the number of folds.

# 4 Limited data discussion

1. With limited data, what types of error can we reduce?
2. Given we are limited to the data available what can we do to improve our model?
3. What are the costs of this method?

# K-Fold Cross-Validation (CV)

**Idea.** Cross-validation estimates how well a model will generalize to new data. Instead of training once on a single train/test split, we repeatedly train and validate on different parts of the data.

## Procedure (K-fold CV)

Assume we have $n$ labeled examples and choose an integer $K$ (often 5 or 10).

1. Randomly shuffle the dataset (optional but common).
2. Split the data into $K$ **folds** of (roughly) equal size.
3. For each round $i = 1, 2, \ldots, K$:
   - Use fold $i$ as the **validation set**.
   - Use the other $K - 1$ folds as the **training set**.
   - Train the model on the training set and compute a validation score on fold $i$.
4. Average the $K$ validation scores to get the cross-validation estimate:

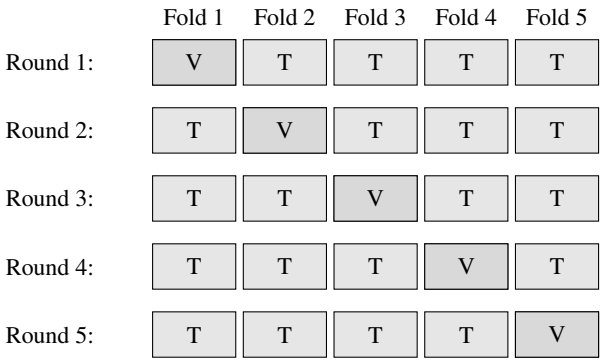
$$\text{CV score} = \frac{1}{K} \sum_{i=1}^{K} s_i,$$

   where $s_i$ is the chosen metric (accuracy, MSE, etc.) on validation fold $i$.

**Why it helps.** Each example is used for validation exactly once and for training $K - 1$ times, so the estimate is typically more stable than a single split.

## A simple visual (5-fold CV)

Legend: **T** = training fold, **V** = validation fold




## Common variations

- **Stratified K-fold:** keeps class proportions similar in each fold (important for classification).
- **Repeated K-fold:** repeats K-fold CV multiple times with different shuffles.
- **Leave-one-out (LOOCV):** $K = n$ (each fold has one example); can be expensive.