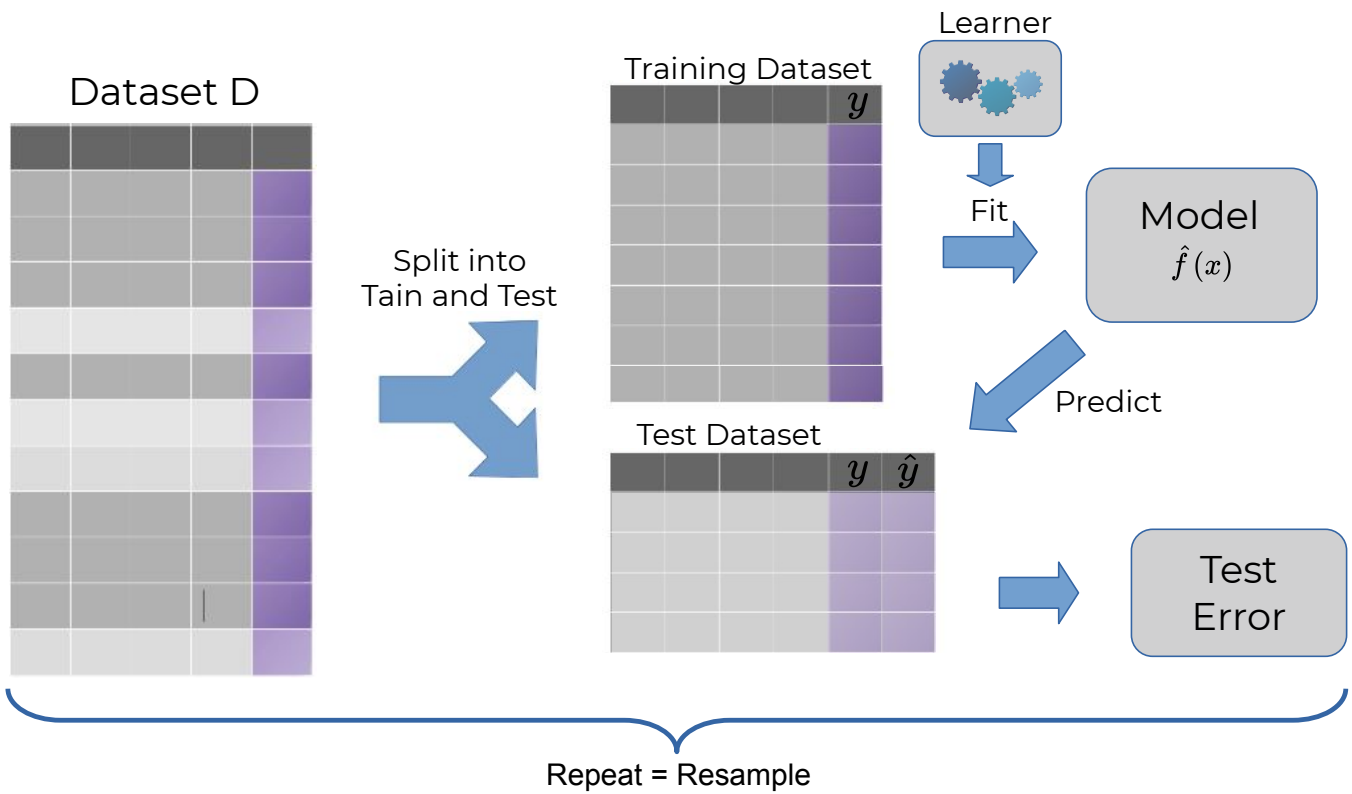
# Introduction to Machine Learning

# Evaluation: Resampling

compstat-lmu.github.io/lecture_i2ml

# RESAMPLING

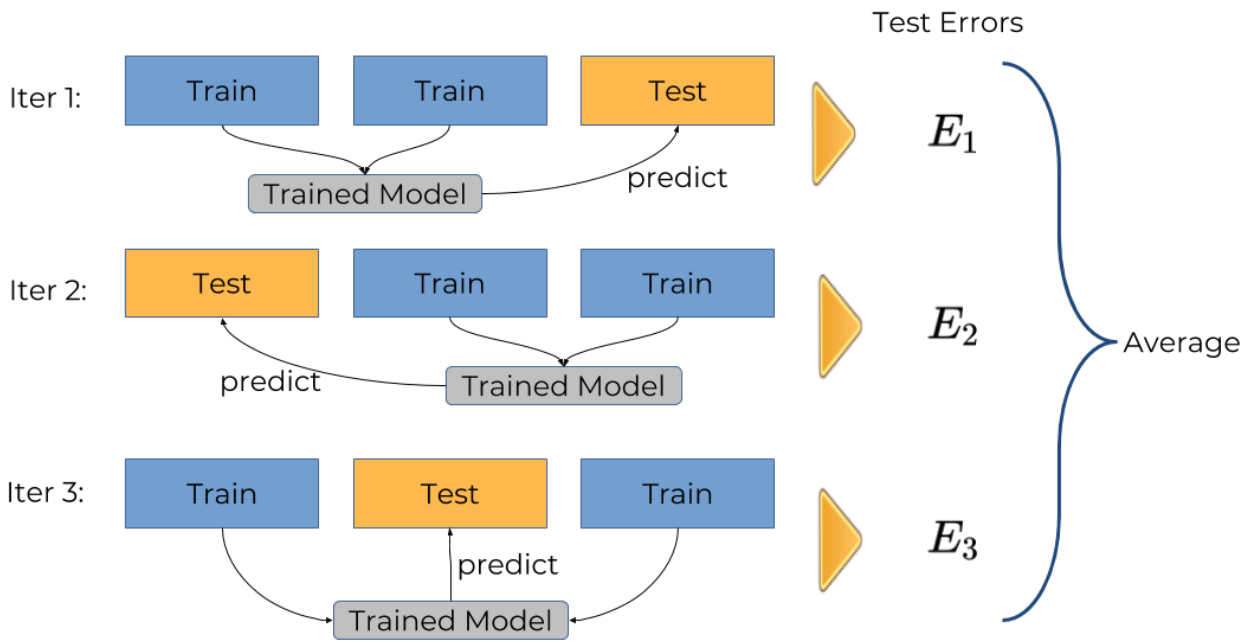
- Aim: Assess the performance of learning algorithm.
- Make training sets large (to keep the pessimistic bias small), and reduce variance introduced by smaller test sets through many repetitions / averaging of results.



# CROSS-VALIDATION

- Split the data into $k$ roughly equally-sized partitions.
- Use each part once as test set and join the $k-1$ others for training
- Obtain $k$ test errors and average.

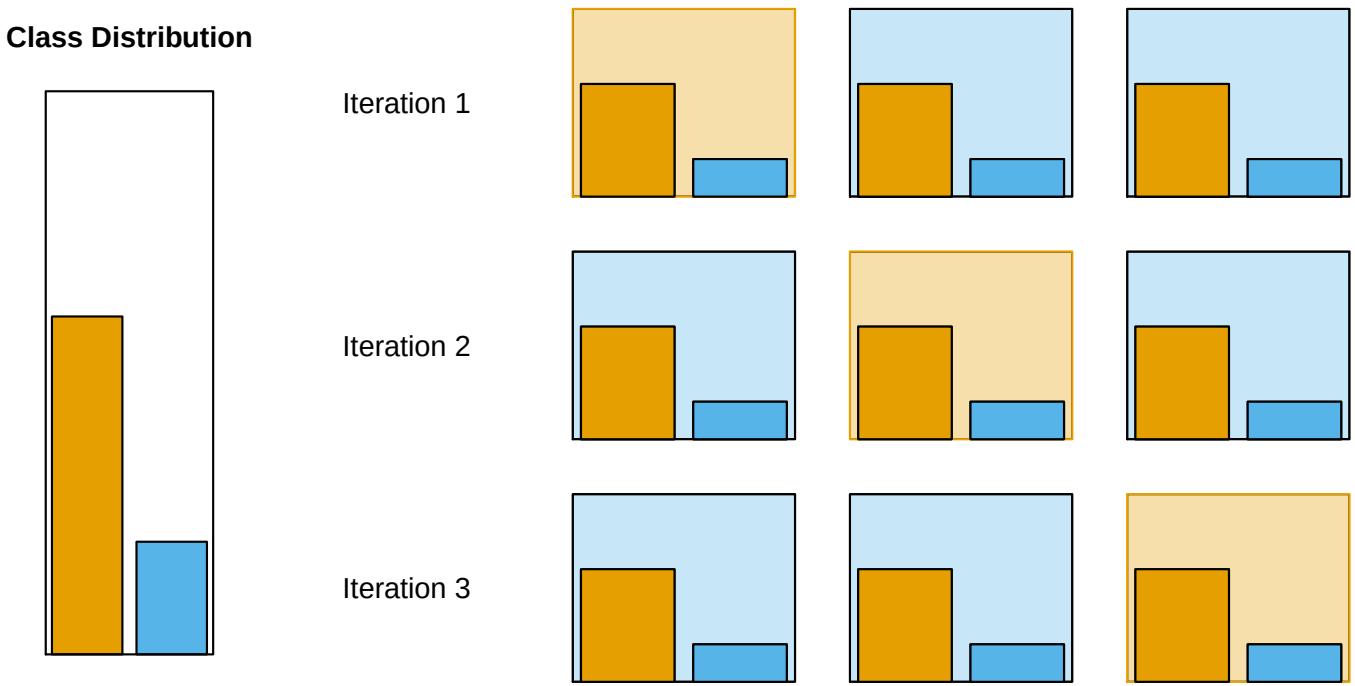
Example: 3-fold cross-validation:



# CROSS-VALIDATION - STRATIFICATION

Stratification tries to keep the distribution of the target class (or any specific categorical feature of interest) in each fold.

Example of stratified 3-fold Cross-Validation:

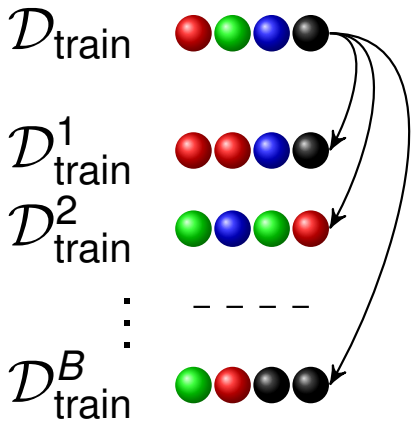


# CROSS-VALIDATION

- 5 or 10 folds are common
- $k = n$ is known as leave-one-out (LOO) cross-validation
- Estimates of the generalization error tend to be pessimistically biased
  size of the training sets is $n - (n/k) < n$)
  bias increases as $k$ gets smaller.
- The $k$ performance estimates are dependent, because of the structured overlap of the training sets.
  $\Rightarrow$ variance of the estimator increases for very large $k$ (close to LOO), when training sets nearly completely overlap.
- Repeated $k$-fold CV (multiple random partitions) can improve error estimation for small sample size.

# BOOTSTRAP

The basic idea is to randomly draw $B$ training sets of size $n$ with replacement from the original training set $\mathcal{D}_{\text{train}}$:



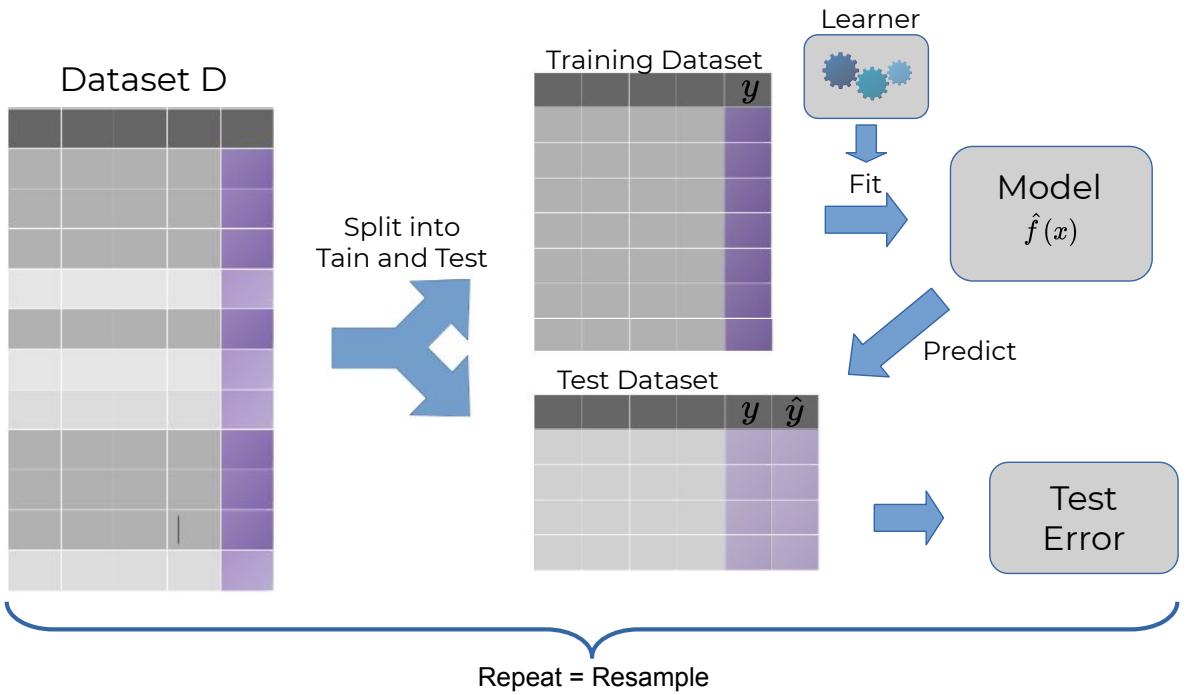

We define the test set in terms of out-of-bag observations $\mathcal{D}^b_{\text{test}} = \mathcal{D}_{\text{train}} \setminus \mathcal{D}^b_{\text{train}}$.

# BOOTSTRAP

- Typically, $B$ is between 30 and 200.
- The variance of the bootstrap estimator tends to be smaller than the variance of $k$-fold CV.
- The more iterations, the smaller the variance of the estimator.
- Tends to be pessimistically biased (because training sets contain only about 63.2% unique the observations).
- Bootstrapping framework allows for inference (e.g. detect significant performance differences between learners).
- Extensions exist for very small data sets, that also use the training error for estimation: B632 and B632+.

# SUBSAMPLING

- Repeated hold-out with averaging, a.k.a. monte-carlo CV
- Similar to bootstrap, but draws without replacement
- Typical choices for splitting: $4/5$ or $9/10$ for training




- The smaller the subsampling rate, the larger the pessimistic bias.
- The more subsampling repetitions, the smaller the variance.

# RESAMPLING DISCUSSION

In ML we fit, at the end, a model on all our given data.

**Problem:** We need to know how well this model performs in the future, but no data is left to reliably do this.
$\Rightarrow$ Approximate using holdout / CV / bootstrap / resampling estimate

**But:** pessimistic bias because we don't use all data points

Final model is (usually) computed on all data points.

# RESAMPLING DISCUSSION

- 5CV or 10CV have become standard
- Do not use Hold-Out, CV with few iterations, or subsampling with a low subsampling rate for small samples, since this can cause the estimator to be extremely biased, with large variance.
- If $n < 500$, use repeated CV
- A $\mathcal{D}$ with $|\mathcal{D}| = 100.000$ can have small sample properties if one class has few observations
- Research indicates that subsampling has better properties than bootstrapping. The repeated observations can can cause problems in training.