DECISION PATTERNS

Resampling

Practical Machine Learning (with R)

UC Berkeley

MODEL PERFORMANCE

Model Performance (thus far)

- Determine performance metric:
  - **RMSE (regression)**
  - **Accuracy (classification)**
- Fit Model
- Calculate statistic ("metric") on Data

"*training*" or "*apparent*" performance will:

- over-fit to training data
- predict very well, unbelievably well
- Not generalize to *new data*.

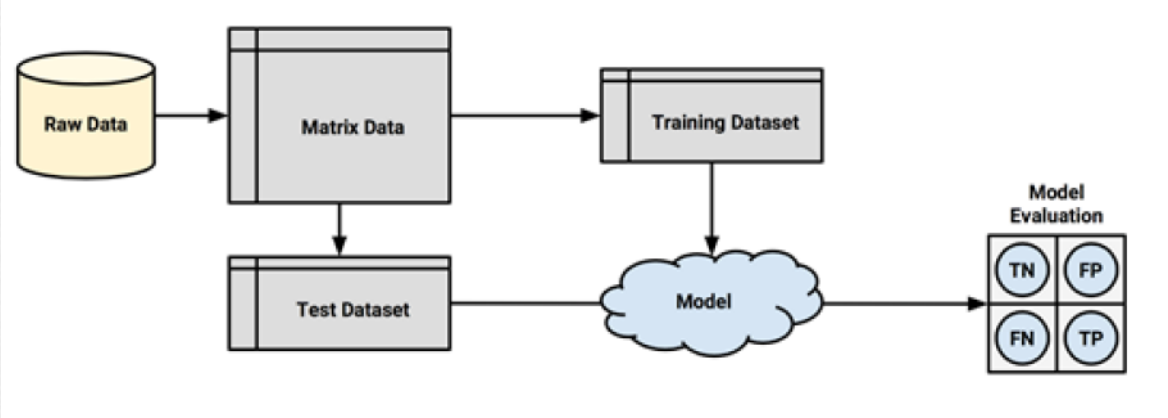
CARDINAL RULE

DO NOT ESTIMATE PERFORMANCE ON YOUR TRAINING DATA

→ Need technique for unbiased estimate for calculating performance

1: Hold Out Method

- Partition data into train and test sets



- What are the partition ratios?
  - Large N: doesn't matter
  - Small N: Need to provide sufficient

IS THERE A BETTER WAY?

Measurements and Statistics

Measurement

Quantification of a phenomena

Deterministic
≠
Stochastic

Statistic

measurement of a *stochastic* phenomena

Examples:

- `mean(x)` <- **x** is generated by a stochastic process
- `sd(x)`

EXERCISE: CALCULATE `sd( mean(x) )`

STATISTICS

- "True" value unknown → uncertainty
- Uncertainty can be measured
  - Variance
  - Standard deviation
  - Confidence Interval
  - …
- Repeated measurements decrease the uncertainty

RESAMPLING

Kuhn benefits of resampling:

- Selection of optimal tuning parameter(s) "With so many choices how do we

- Unbiased estimate of model performance

Resampling Strategies

- Repeated Holdouts
- K-Fold Cross Validation
- Bootstrap

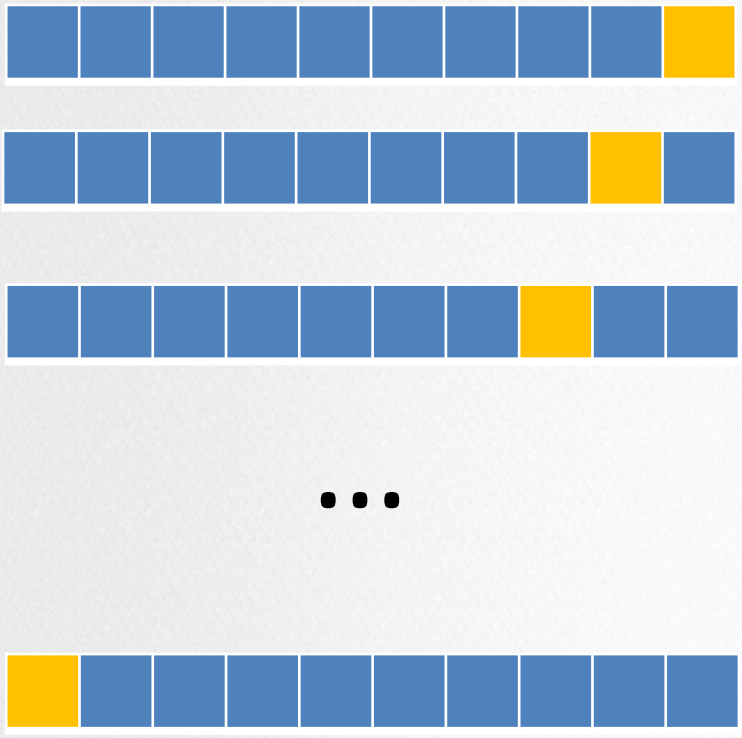
Repeated Holdout

AKA Monte Carlo Splitting

- Split data 75%-25%
  - Fit Model
  - Calculate Performance Metric
  - Repeat with Different Split(K-times)
- Calculate Metric

$$Metric = AVG_i(metric)$$

10-Fold Cross Validation



LOOCV : K→n

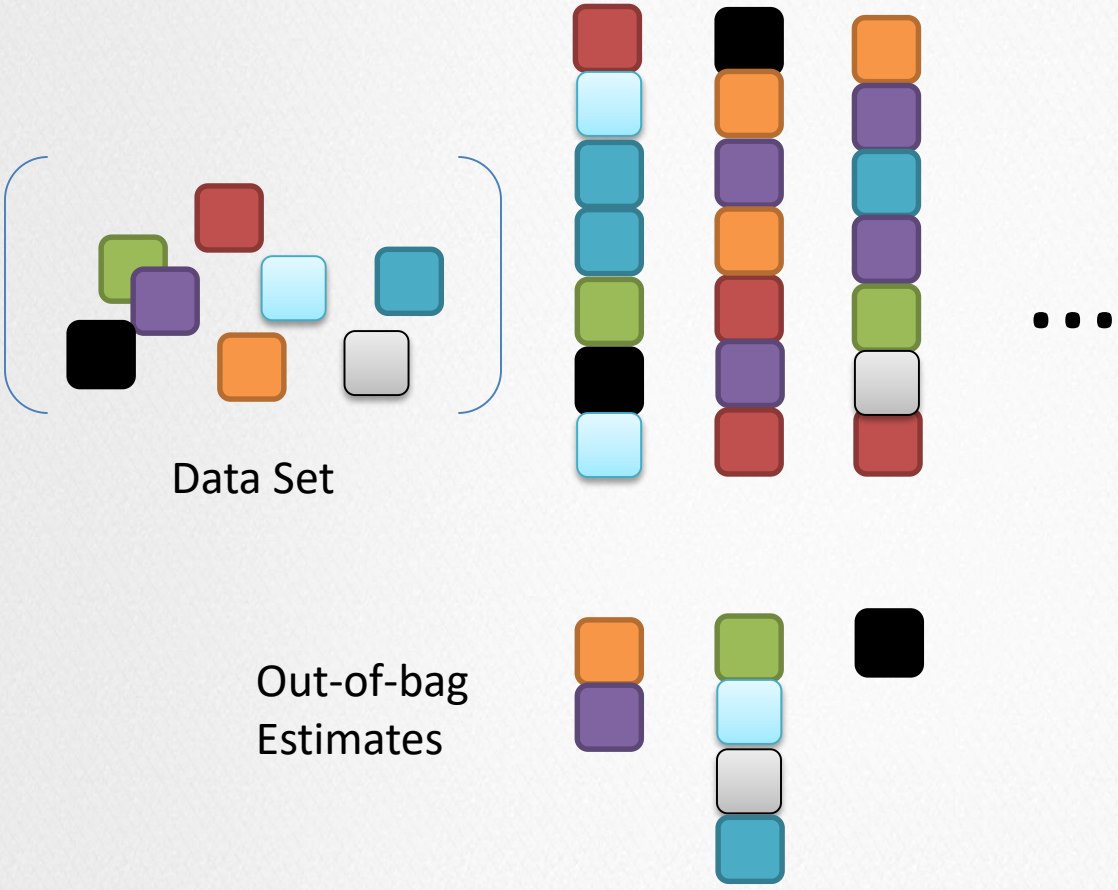
- Split the data set into 10 equal sized samples.
- Leave one sample out (fold)
  - Fit the model
  - calculate the metric on the fold
  - Repeat choosing another sample until done
- Calculate Metric

$$Metric = AVG_i(metric)$$

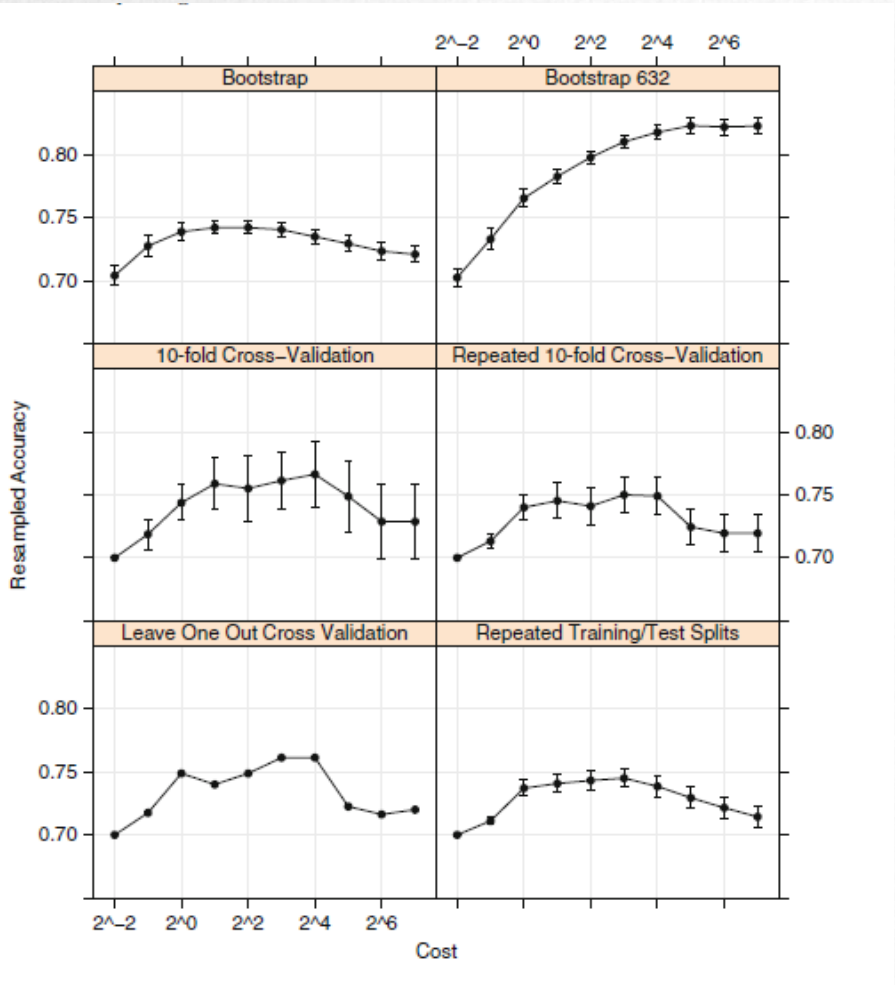
- 5 or 10-fold common

Bootstrap

- "Sampling with Replacement"



Which Is Best?



- There isn't one.

K-fold cross validation

- Higher Variance
- Lower Bias

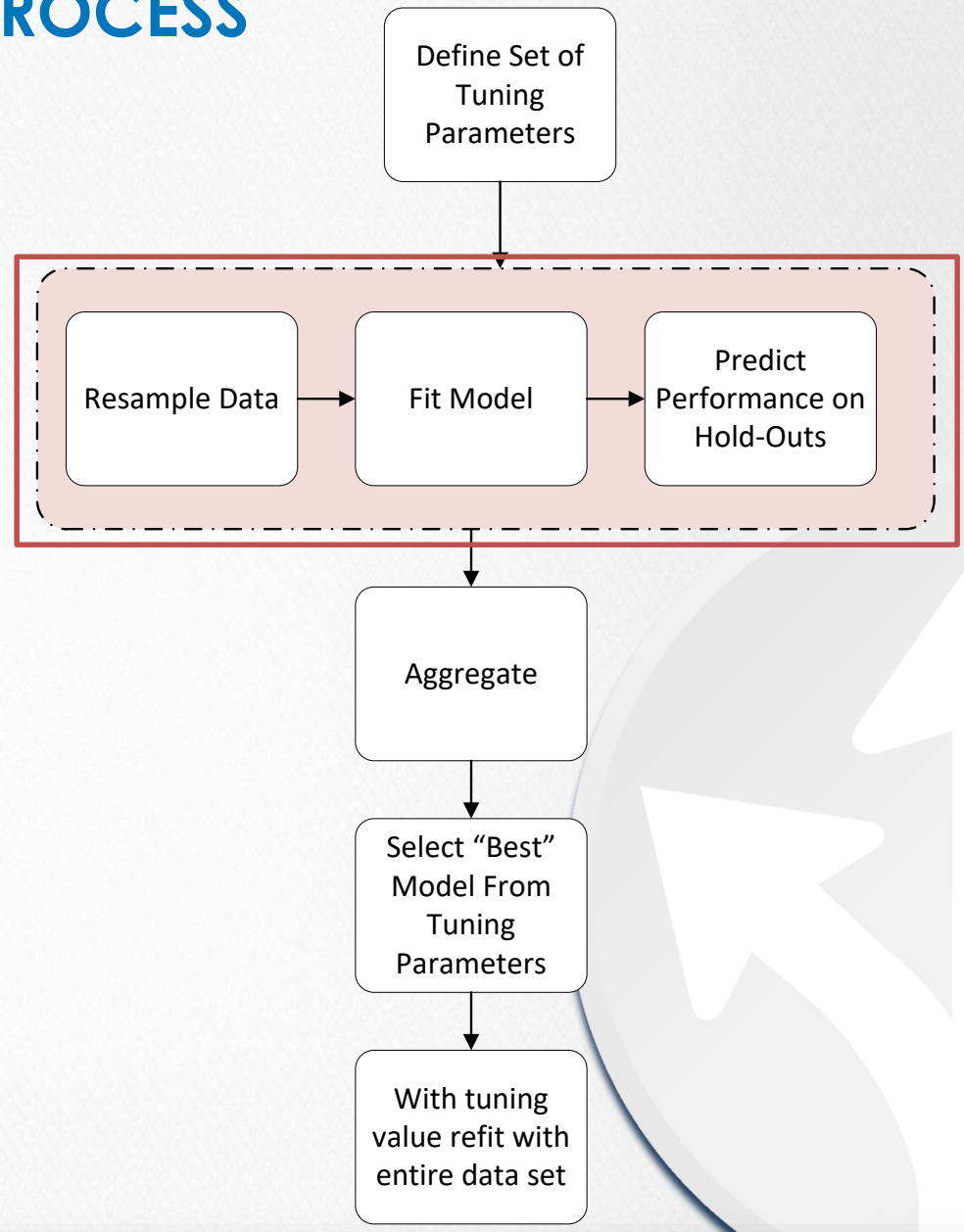
Bootstrap

- Lower Variance
- Higher Bias

Better to employ resampling than worry about not resampling

Kuhn's Resampling Process

Today's Focus



RESAMPLING

- Best Solution (n-permitting)
  - split data into training and test data
  - and do what Kuhn says.

Why(?)

- Easy to interpret defend
- Requires data not be consumed by model
- Computationally easy
- Is generally not (by itself) the most accurate → no confidence

MODEL PERFORMANCE IS *NOT* *TRAINING* PERFORMANCE