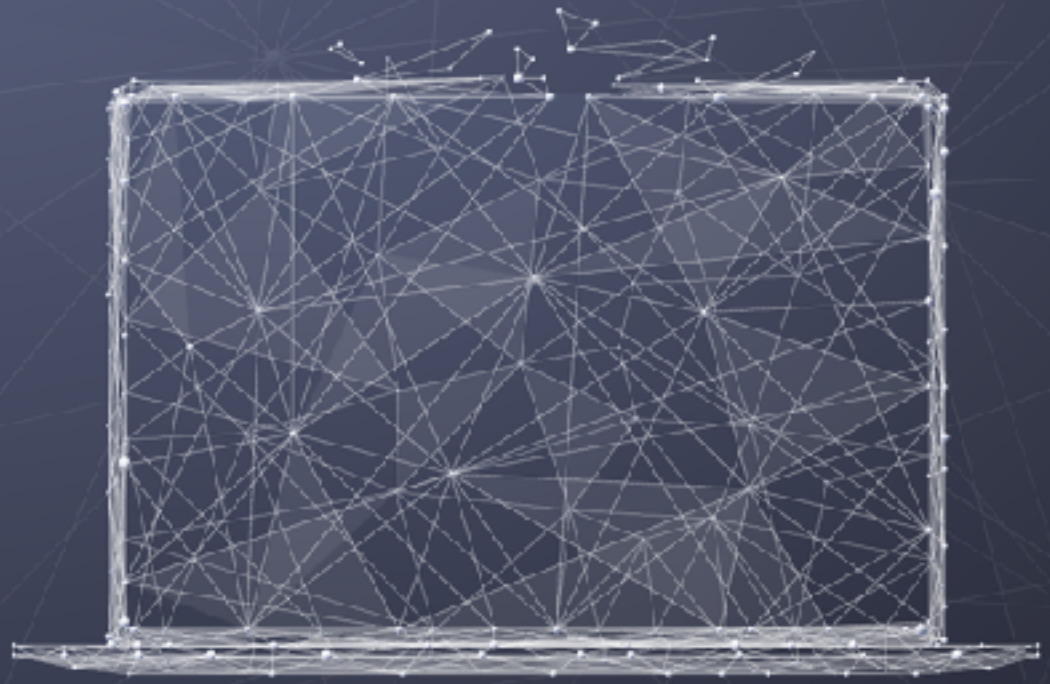
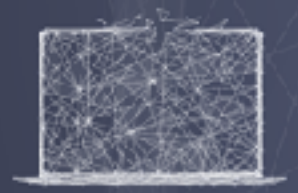
Data Science

Foundations of Decision Making

Evaluating predictive models

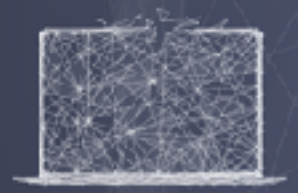


PURDUE UNIVERSITY | College of Science



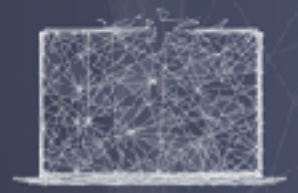
Empirical evaluation

- Given observed accuracy of a model on limited data, how well does this estimate generalize for additional examples?
- Given that one model outperforms another on some sample of data, how likely is it that this model is more accurate in general?
- When data are limited, what is the best way to use the data to both learn and evaluate a model?



Evaluating classifiers

- Goal: Estimate true future error rate
- When data are limited, what is the best way to use the data to both learn and evaluate a model?
- Approach 1
  - Reclassify training data to estimate error rate



Evaluating classifiers

- Goal: Estimate true future error rate
- When data are limited, what is the best way to use the data to both learn and evaluate a model?
- Approach 1
  - Reclassify training data to estimate error rate

Approach 1

Y	X1	X2

Data Set

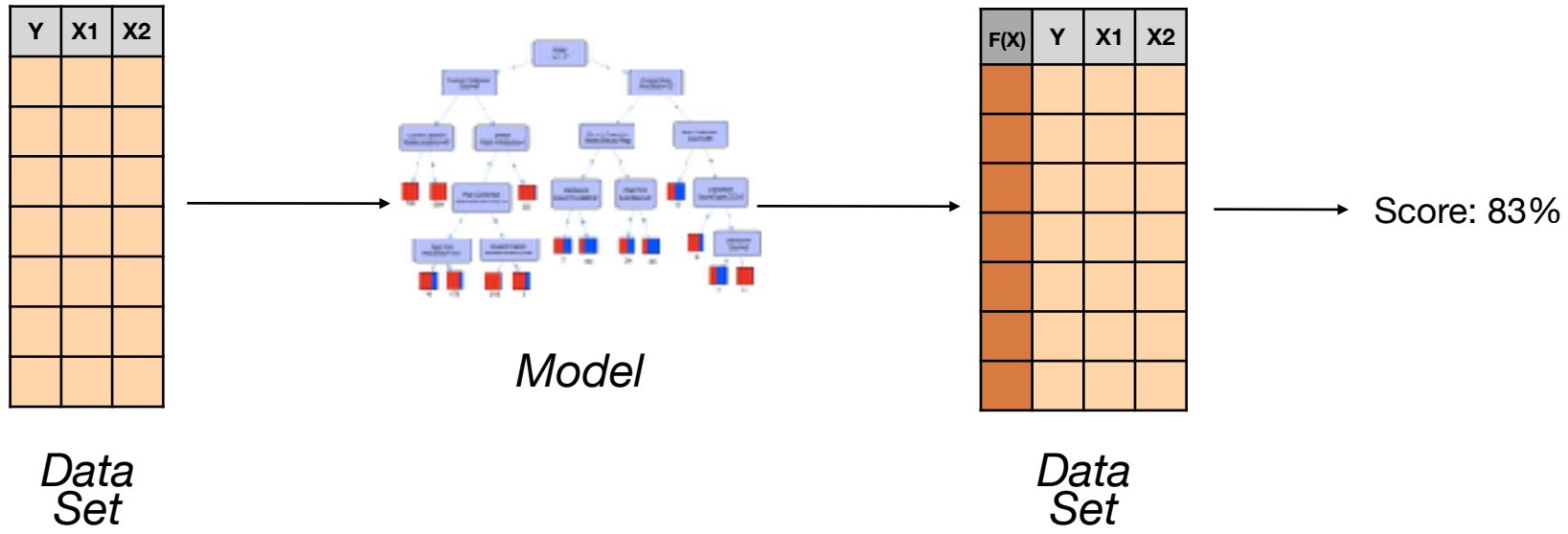
Approach 1

Y	X1	X2

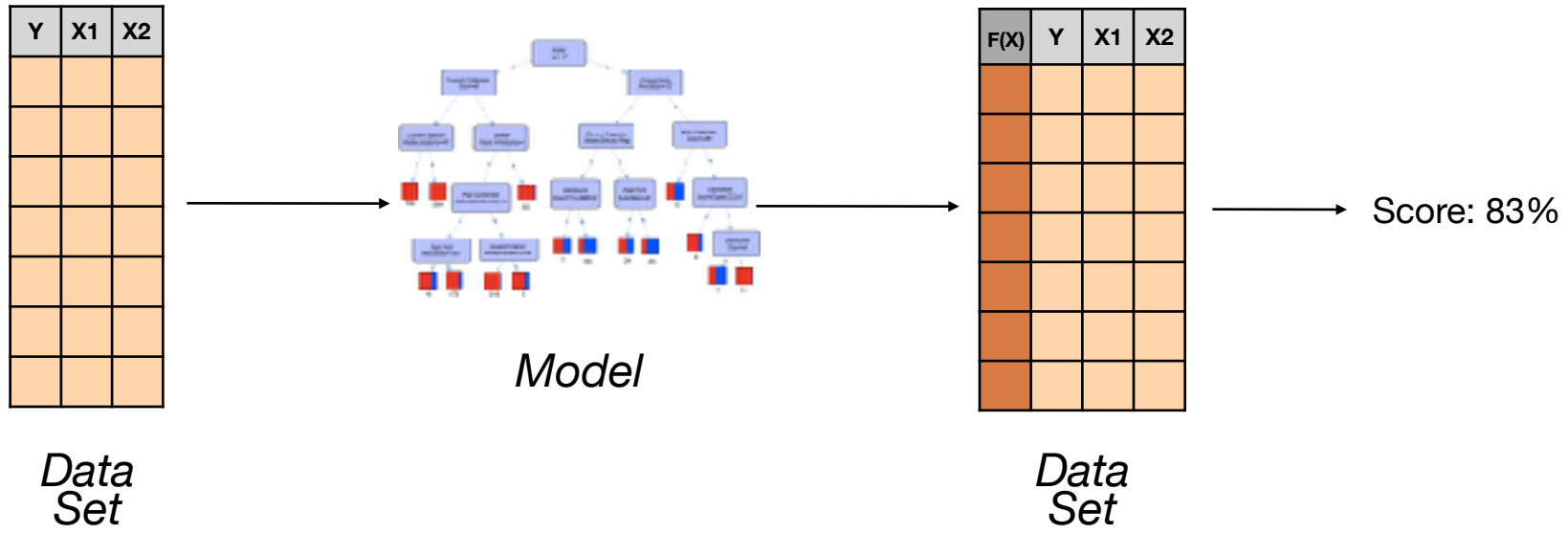
Data Set

Model

Approach 1



Approach 1



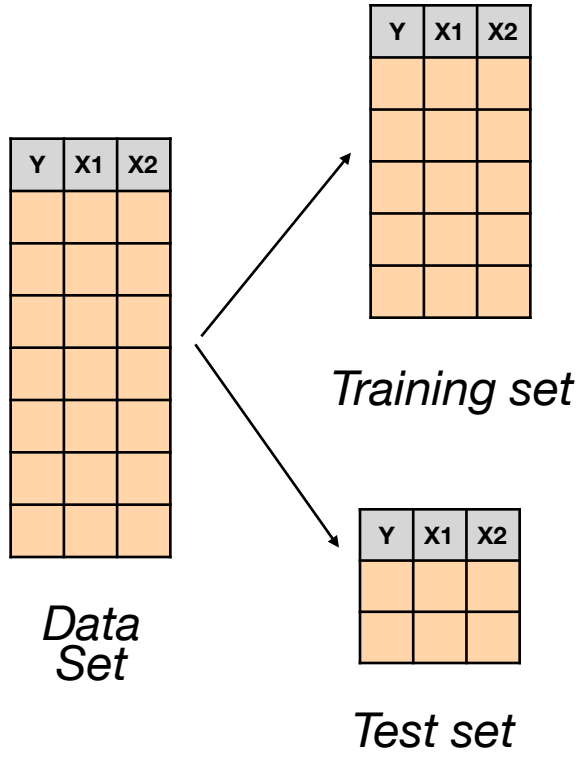
Typically produces a biased estimate of future error on new data because model is ***overfit*** to the training data

Approach 2

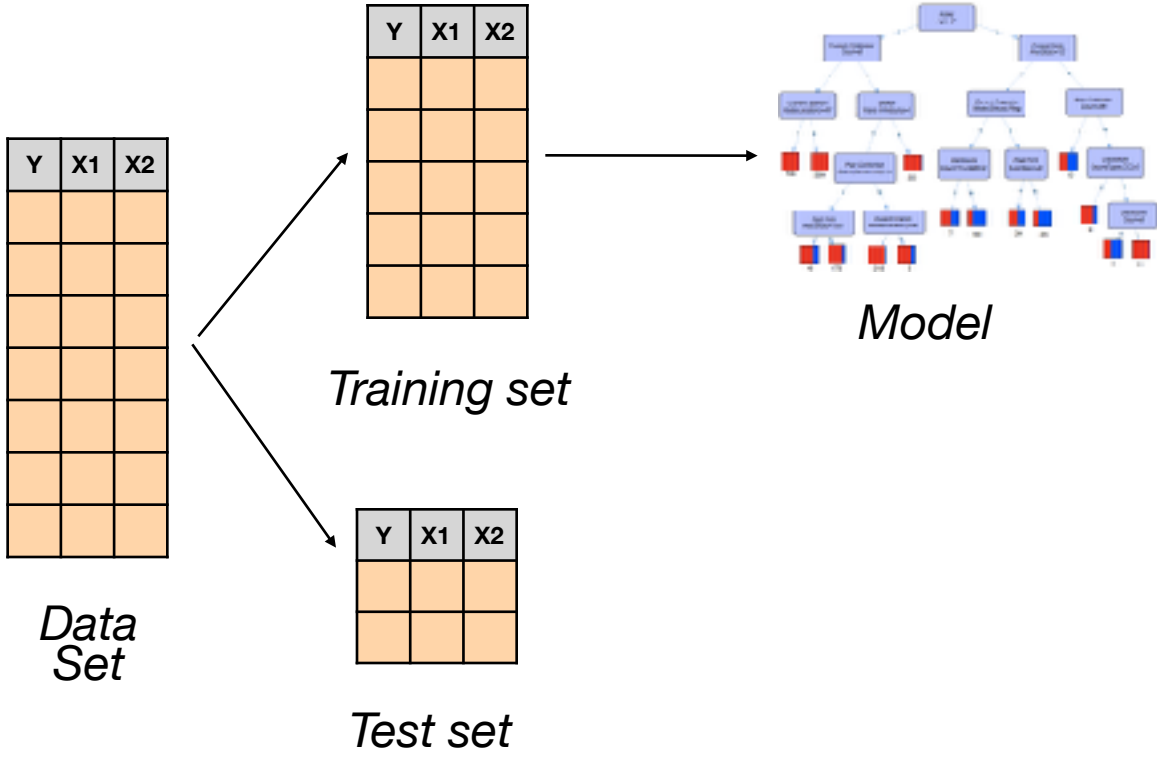
Y	X1	X2

Data Set

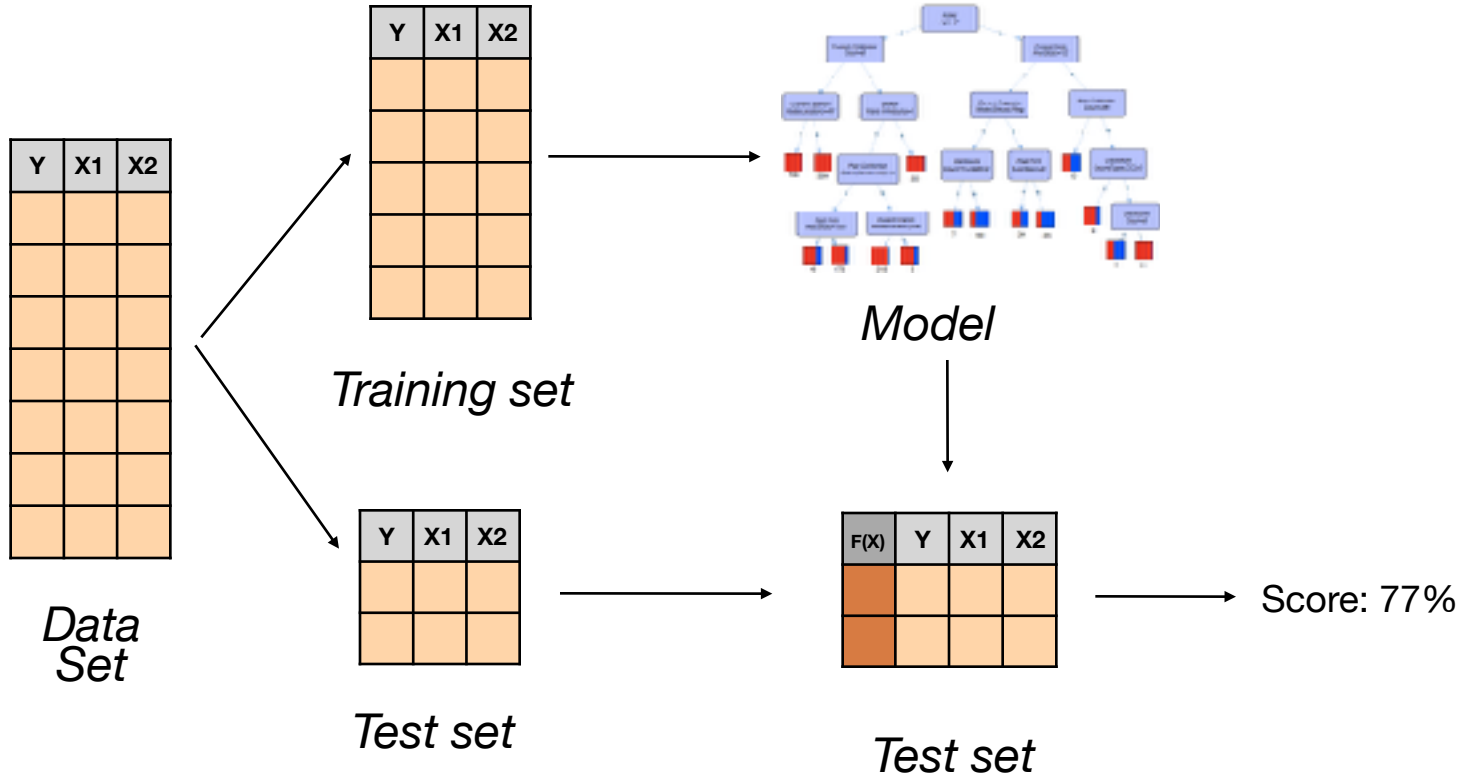
Approach 2



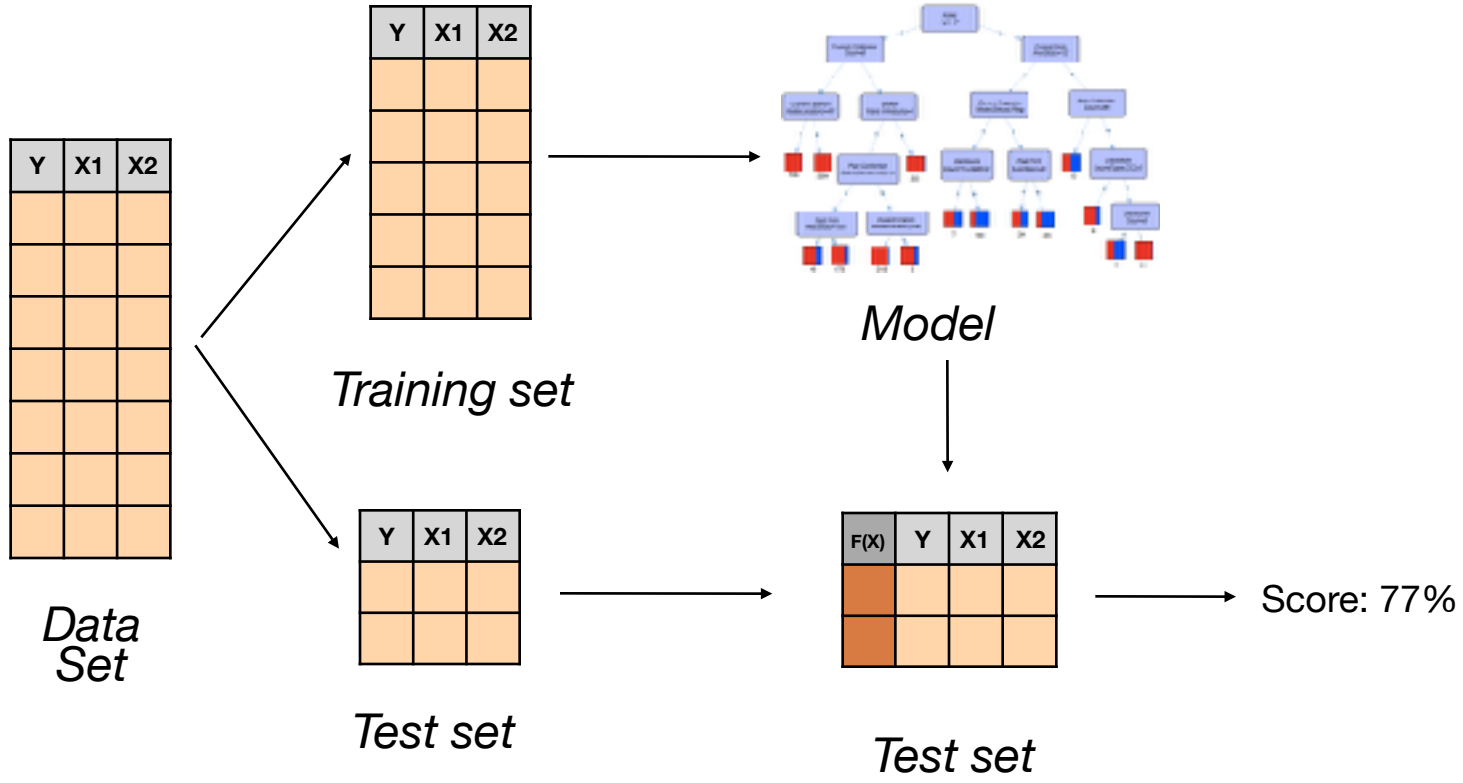
Approach 2



Approach 2

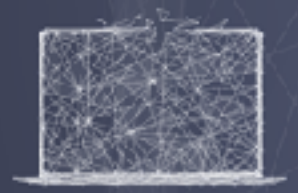

Y X1 X2
Data Set
Y X1 X2
Training set
Model
Y X1 X2
Test set
F(X) Y X1 X2
Test set
Score: 77%

Approach 2



Partition data into training and test sets: quality of error estimate will vary due to size and makeup of test set

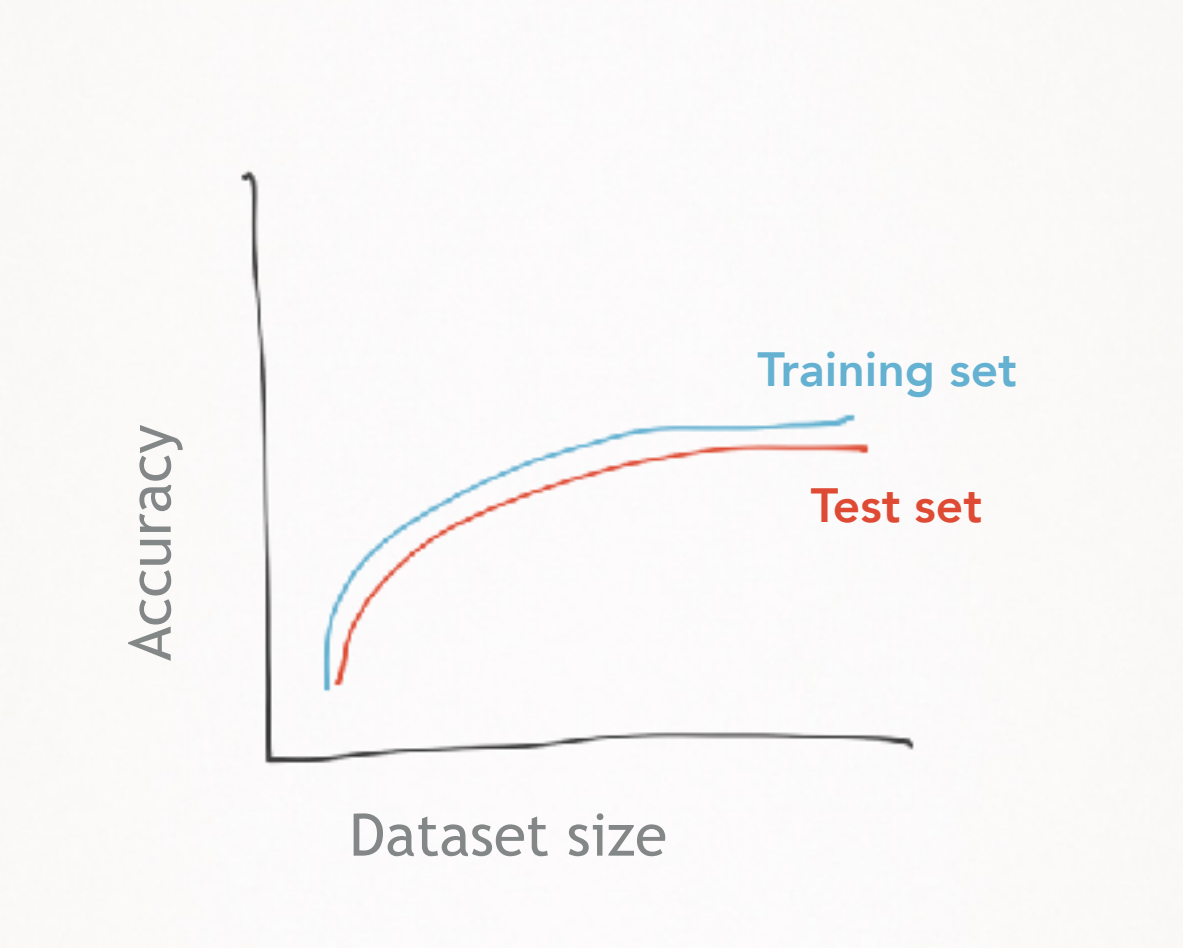
How to know if it's the data or the model that's limiting performance ?



Learning curves

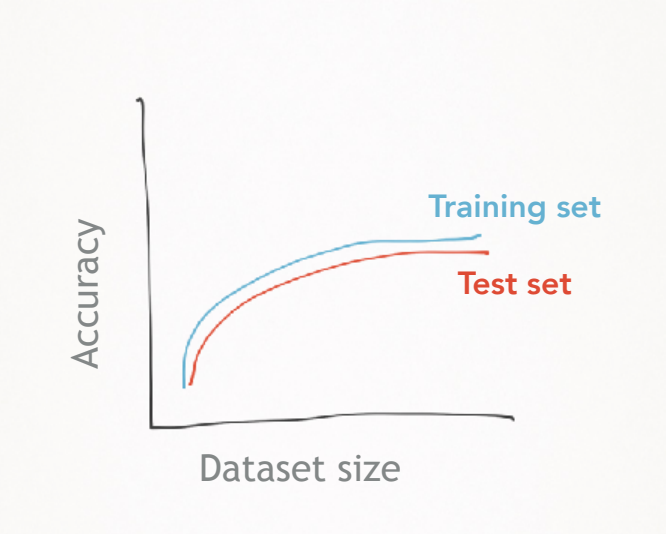
- Goal: See how performance improves with additional training data
- From dataset set S, where |S|=n
  - For i=[10, 20, ... ,100]
    - Randomly sample i% of S to construct sample S’
    - Learn model on S’
    - Evaluate model
  - Plot training set size vs. accuracy

Learning curves illuminate likely causes of error



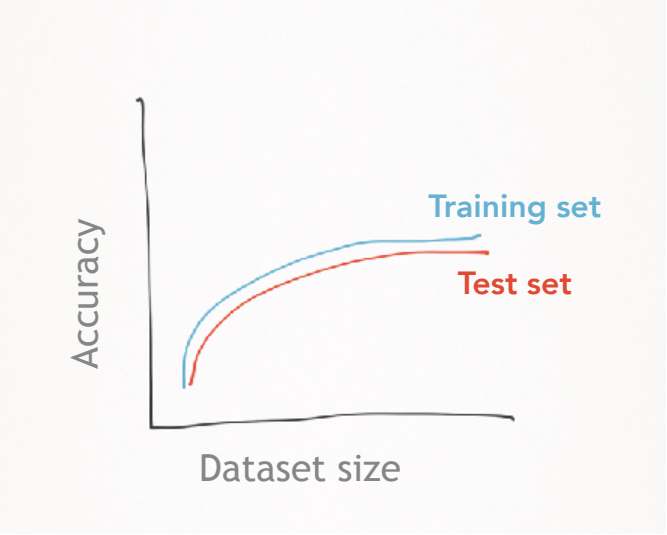
Know when to get more data vs. use different model

Underfitting



Know when to get more data vs. use different model

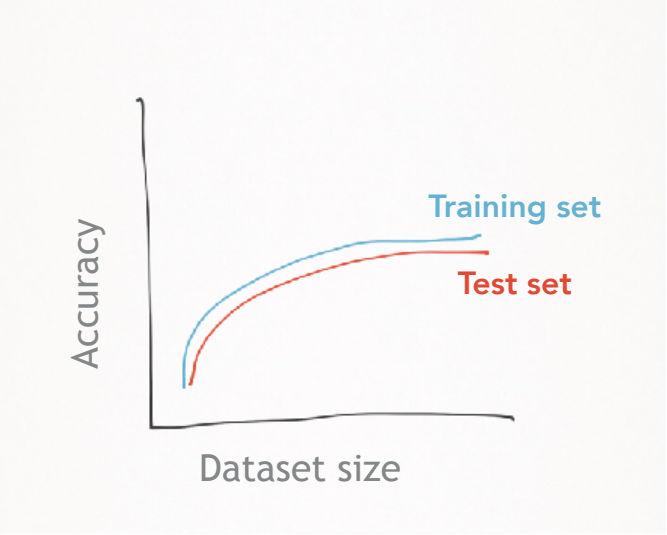
Underfitting



If learning curves flatten early, then additional data is not being exploited the model

Know when to get more data vs. use different model

Underfitting



If learning curves flatten early, then additional data is not being exploited the model

→ increase complexity of model

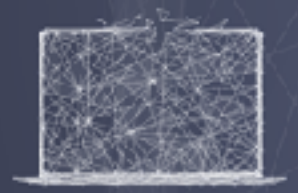
Know when to get more data vs. use different model

Overfitting

Know when to get more data vs. use different model

Overfitting

If accuracy on test data starts to degrade, then model is paying too much attention to idiosyncrasies in training data



Know when to get more data vs. use different model

Overfitting

If accuracy on test data starts to degrade, then model is paying too much attention to idiosyncrasies in training data

→ get more data and/or regularize during learning

Know when to get more data vs. use different model

Just right

Know when to get more data vs. use different model

Just right

If learning curve on training data reaches a high plateau and test performance is similar

Know when to get more data vs. use different model

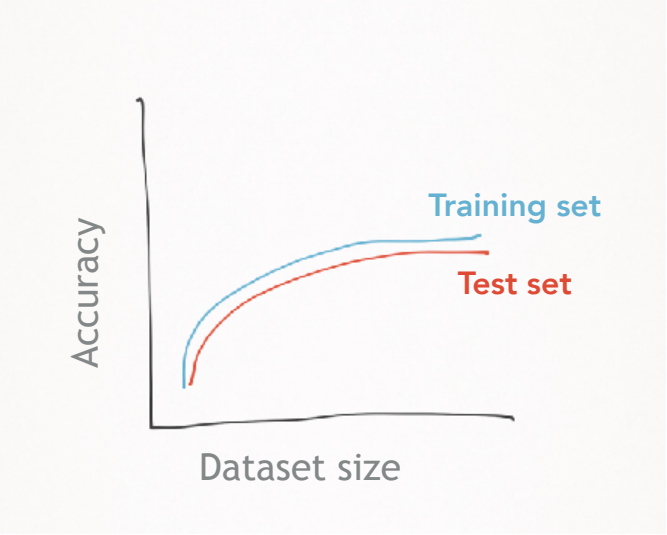
Just right

If learning curve on training data reaches a high plateau and test performance is similar

→ stop to celebrate, this almost never happens!

Know when to get more data vs. use different model

Underfitting



Increase complexity of model

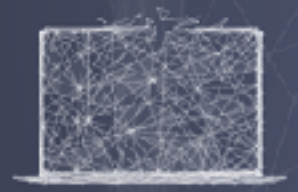
Overfitting

Get more data and/or regularize during learning

Just right

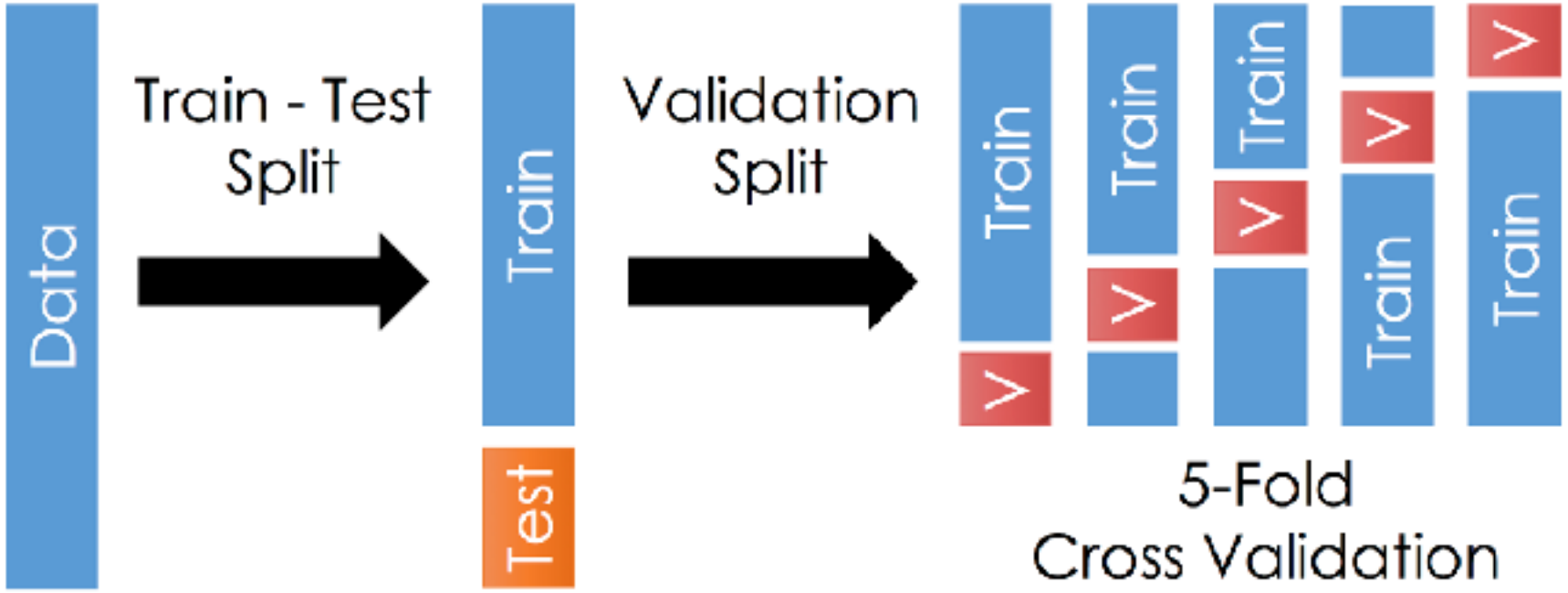
Ideal performance that is almost never observed

PURDUE UNIVERSITY | College of Science

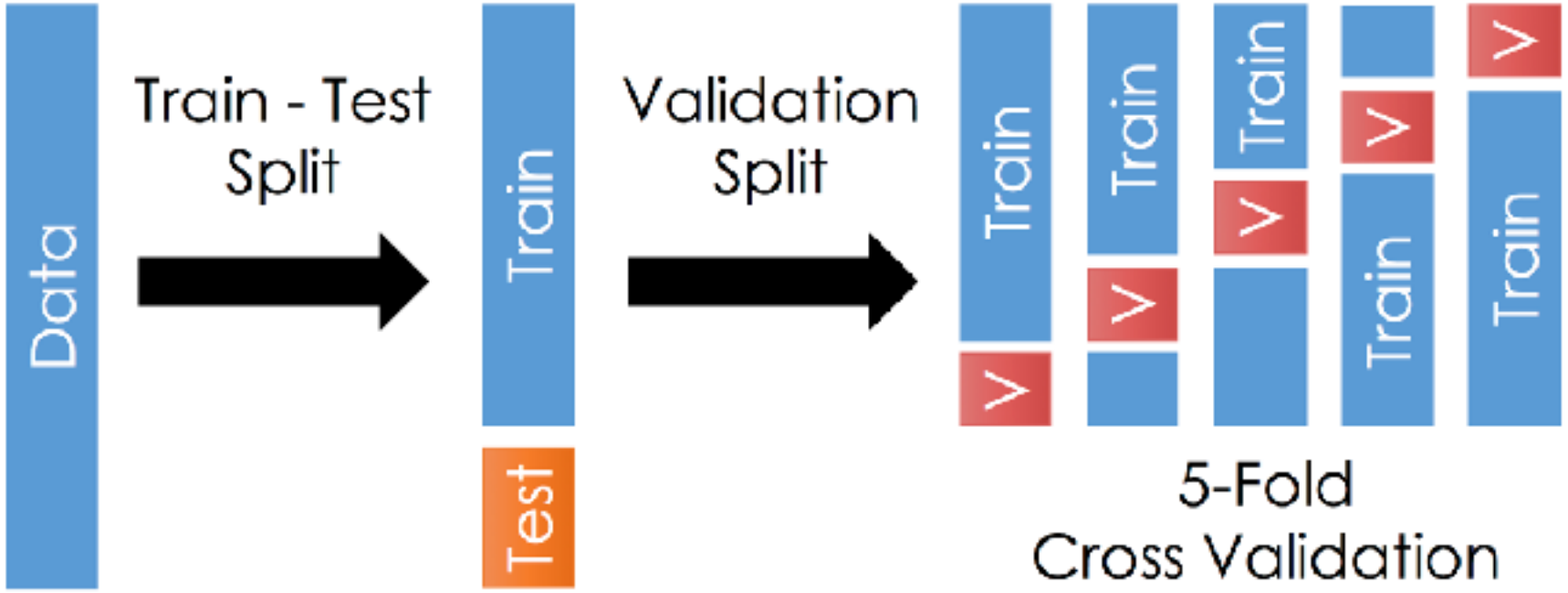


Key to assessing significance: held out test data

Key to assessing significance: held out test data



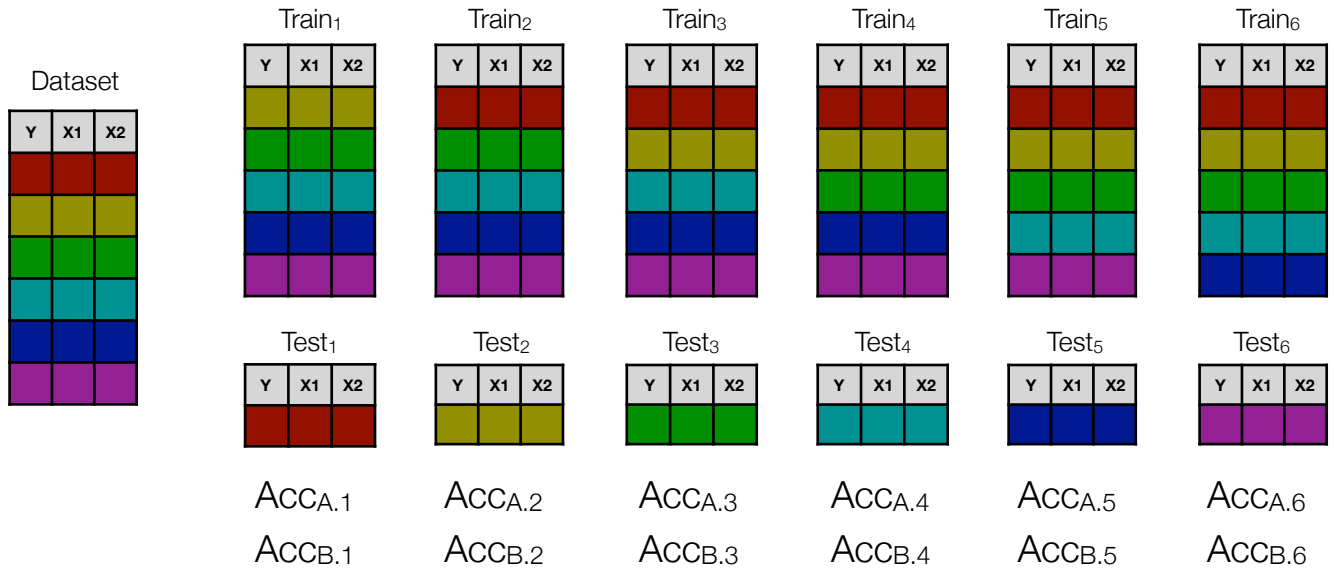
Key to assessing significance: held out test data



Be careful you don't overfit by testing too much on held out data

Evaluating classification algorithms A and B

- Use k-fold cross-validation to get k estimates of error for MA and MB



- Set of errors estimated over the test set folds provides empirical estimate of sampling distribution
- Mean is estimate of expected error

PURDUE UNIVERSITY | College of Science

Assessing significance

- Use paired t-test to assess whether the two distributions of errors are statistically different from each other

$Acc_{A.1}$	$Acc_{B.1}$
$Acc_{A.2}$	$Acc_{B.2}$
$Acc_{A.3}$	$Acc_{B.3}$
$Acc_{A.4}$	$Acc_{B.4}$
$Acc_{A.5}$	$Acc_{B.5}$

- Takes into account both the difference in means and the variability of the scores