CS 66: Machine Learning

Prof. Sara Mathieson

Spring 2019

Outline for March 4

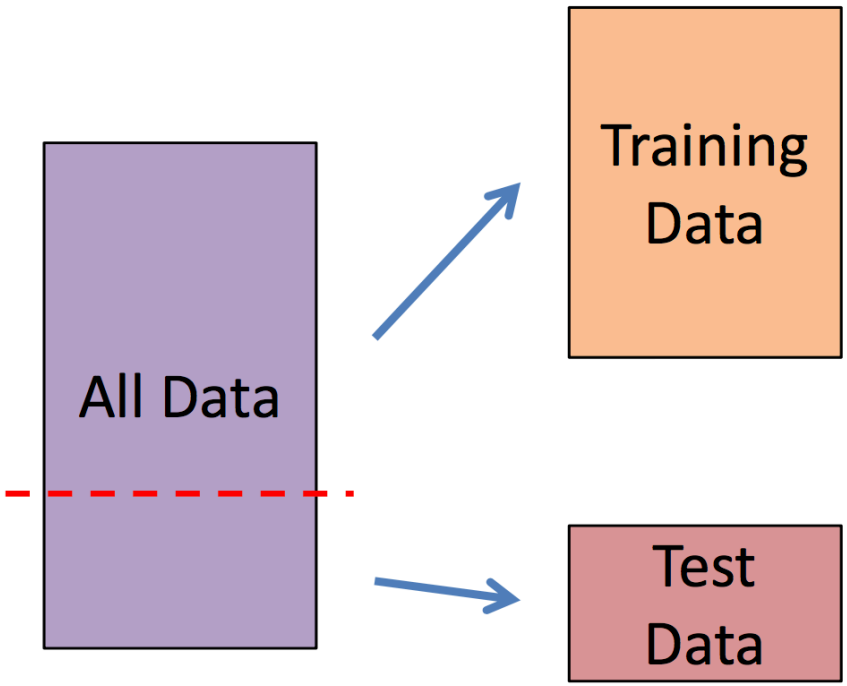
- Cross-validation
- Ensemble methods
  - Bagging
  - Boosting
  - Random Forests

- **Lab 4 due Friday**
- Office hours **TODAY** 12:30-2pm

Outline for March 4

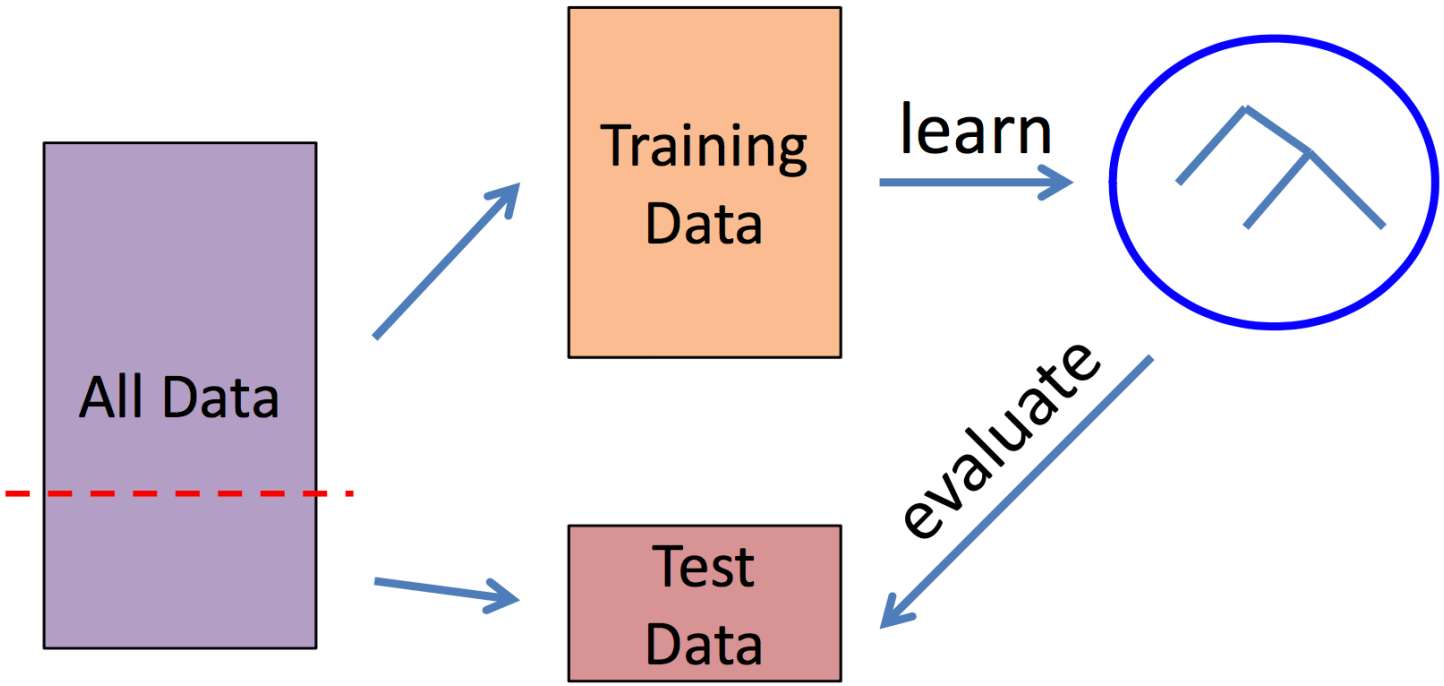
- Cross-validation
- Ensemble methods
  - Bagging
  - Boosting
  - Random Forests

Evaluation in Practice



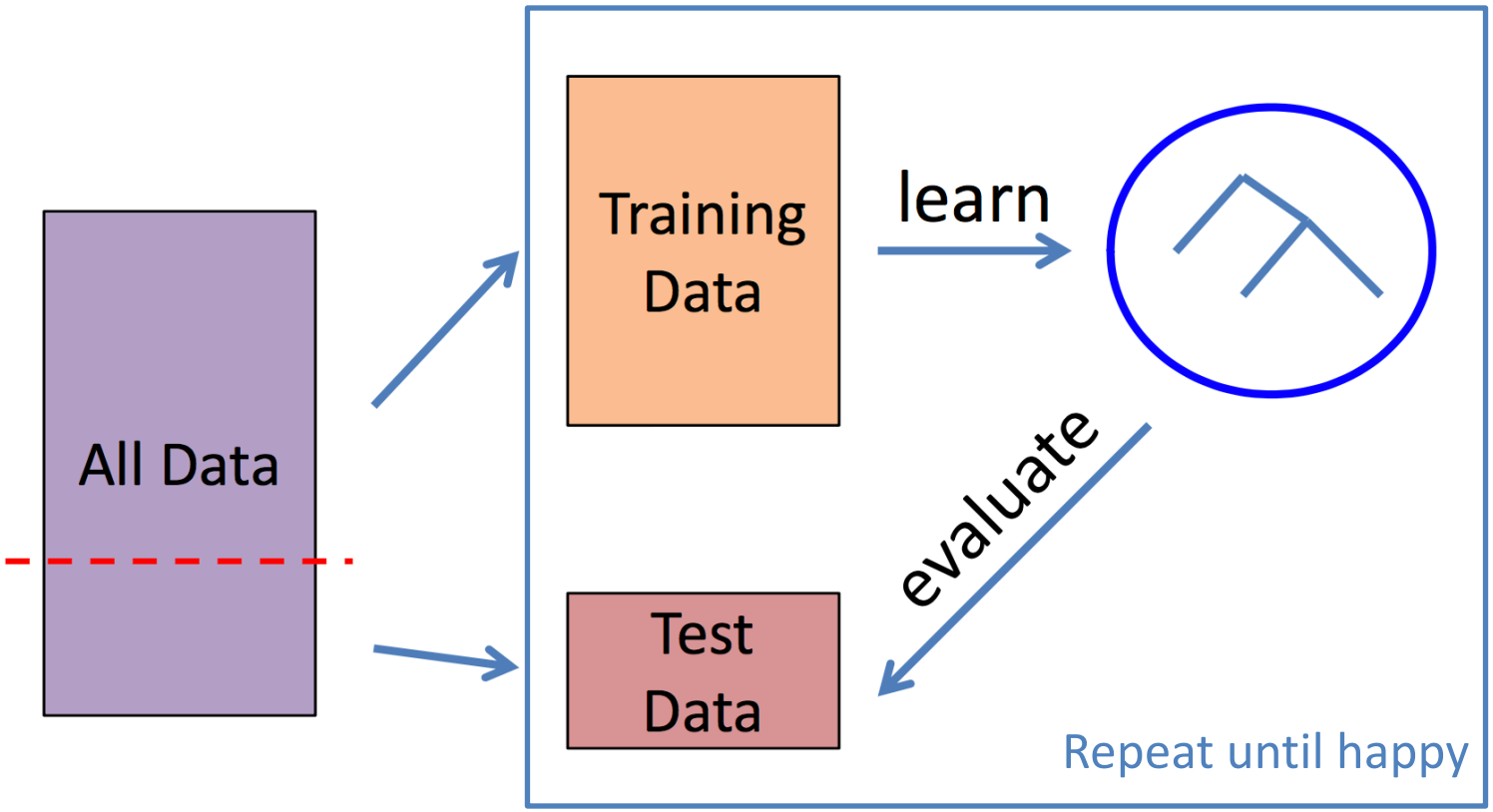
Modified from Jessica Wu

Evaluation in Practice



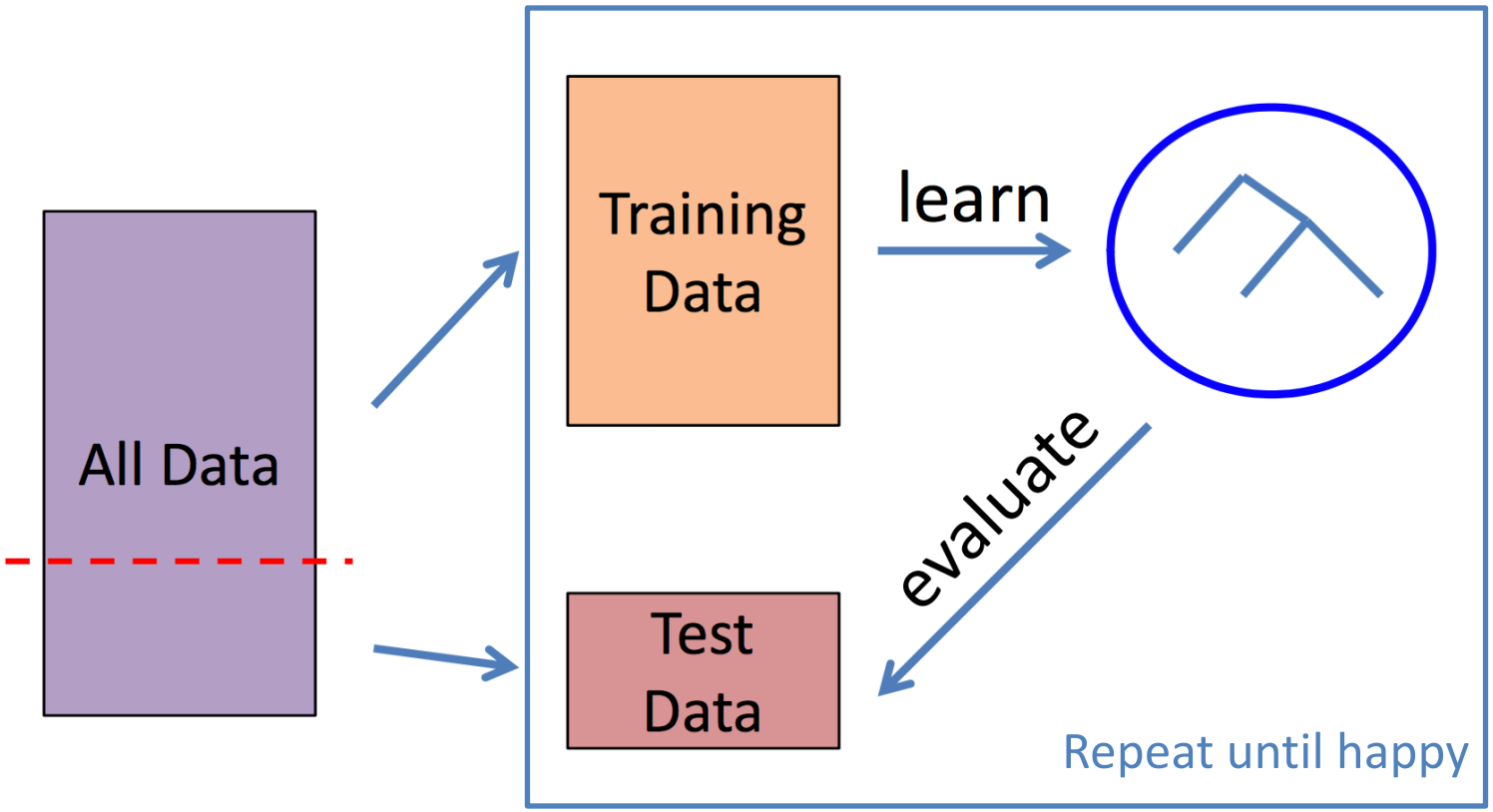
Modified from Jessica Wu

Evaluation in Practice



Modified from Jessica Wu

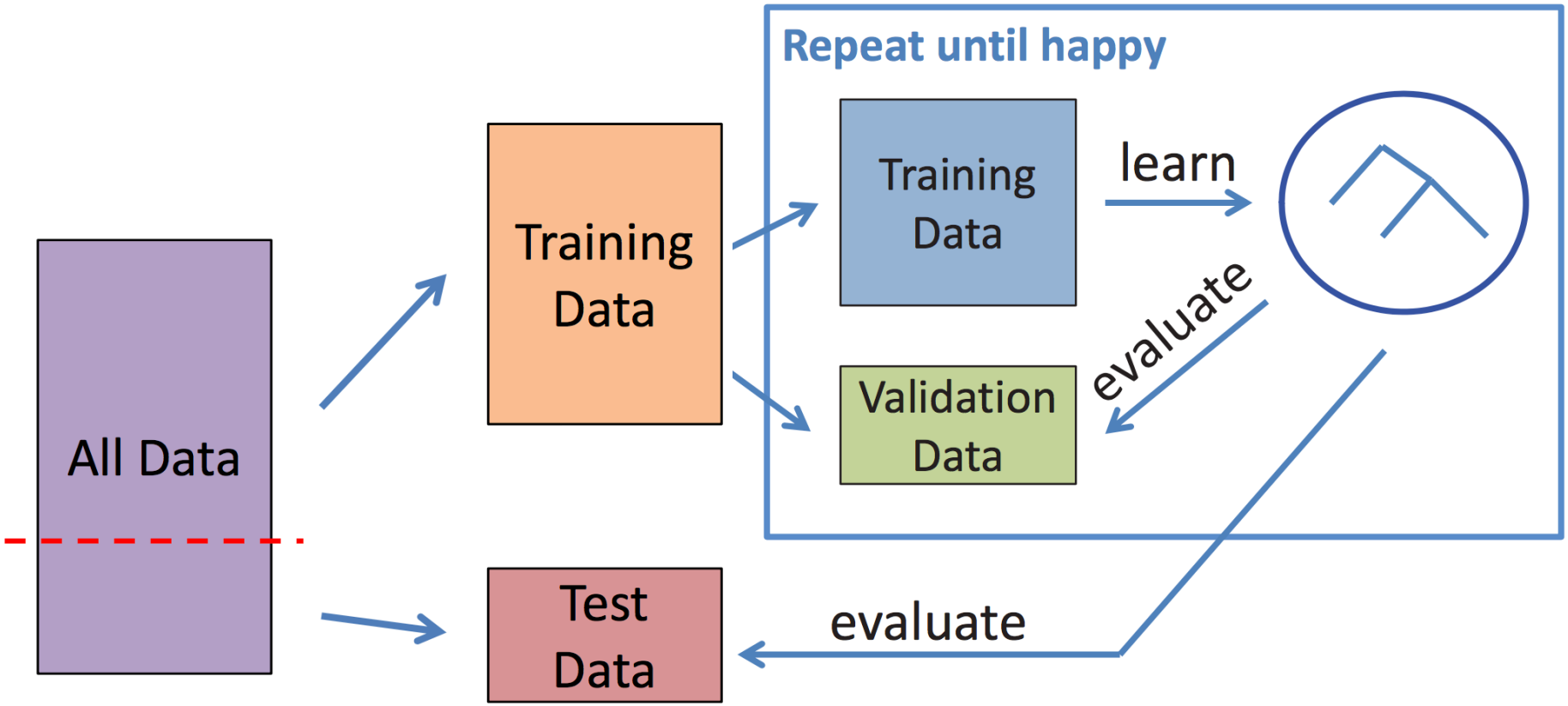
Evaluation in Practice



NO! Using test data as part of the model selection process

Modified from Jessica Wu

Better: use a *validation* dataset



Modified from Jessica Wu

k-fold Cross Validation

- Why just choose one particular "split" of data?
  - in principle, we should do this multiple times since performance may be different for each split

Slide: Jessica Wu

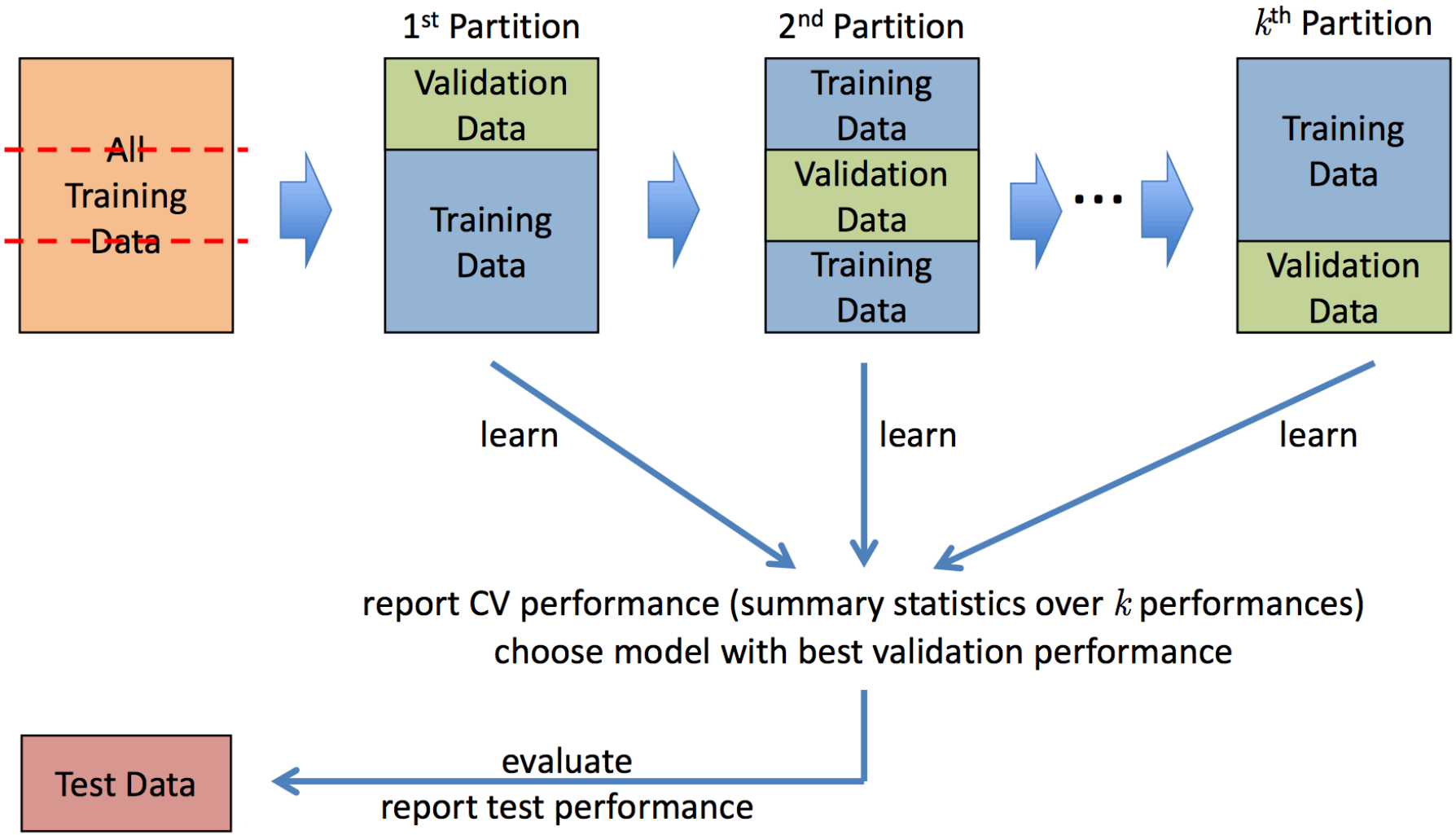
k-fold Cross Validation

- Why just choose one particular "split" of data?
  - in principle, we should do this multiple times since performance may be different for each split

- k-Fold Cross-Validation (e.g., $k = 10$)
  - randomly partition full data set of n instances into **k disjoint subsets** (each roughly of size n/k)
  - choose each fold in turn as validation set; train model on the other $k-1$ folds and evaluate
  - compute statistics over k test performances, or choose best of k models

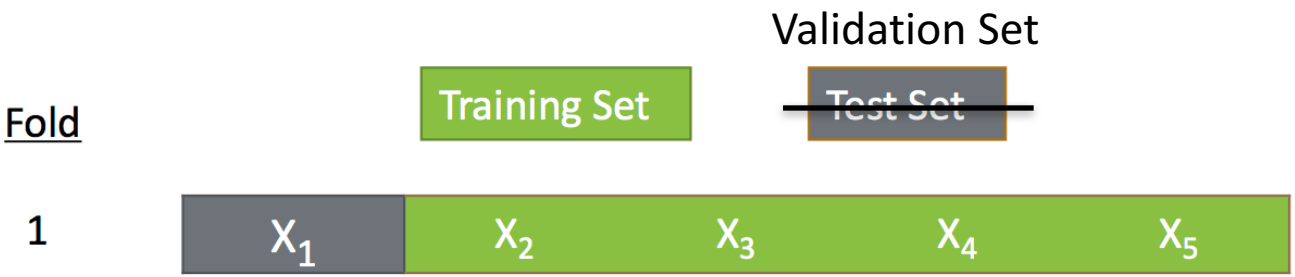
Slide: Jessica Wu

k-fold Cross Validation



Slide: Jessica Wu

k-fold Cross Validation



Evaluation

Accuracy(P_1) = 11/20

Slide: Ameet Soni

k-fold Cross Validation

Slide: Ameet Soni

k-fold Cross Validation

Evaluation

Accuracy(P_1) = 11/20

Accuracy(P_2) = 17/20

Accuracy(P_3) = 16/20

Accuracy(P_4) = 13/20

Accuracy(P_5) = 16/20

Generalization: average accuracy across all folds = 73/100 = 73%

Slide: Ameet Soni

Discussion

1) What are the costs of *k*-fold cross validation?

2) Pros and cons of no longer having one model?

3) How to choose *k*?

Discussion

1) What are the costs of *k*-fold cross validation?

    - Computational, especially if training takes a long time

2) Pros and cons of no longer having one model?

    - Con: might be hard to interpret
    - Pro: might be able to average results

3) How to choose *k*?

    - Large *k* can be good for small datasets (i.e. where *n* is small)
    - Tradeoff between computation and reducing variance
    - Many choose *k*=10 in practice :)

Cross Validation: other considerations

- Can use cross-validation to choose hyper parameters

- Leave-one-out cross validation (LOOCV)
  - Special case of $k=n$
  - Train using $n-1$ examples, evaluate on remaining
  - Repeat n times

- Can do multiple trials of CV

The Short Way
(that Many People Actually Use)

- Split into only training data + validation data
- Train on training data, evaluate on validation data
- Report cross-validation performance
  - possibly also training performance

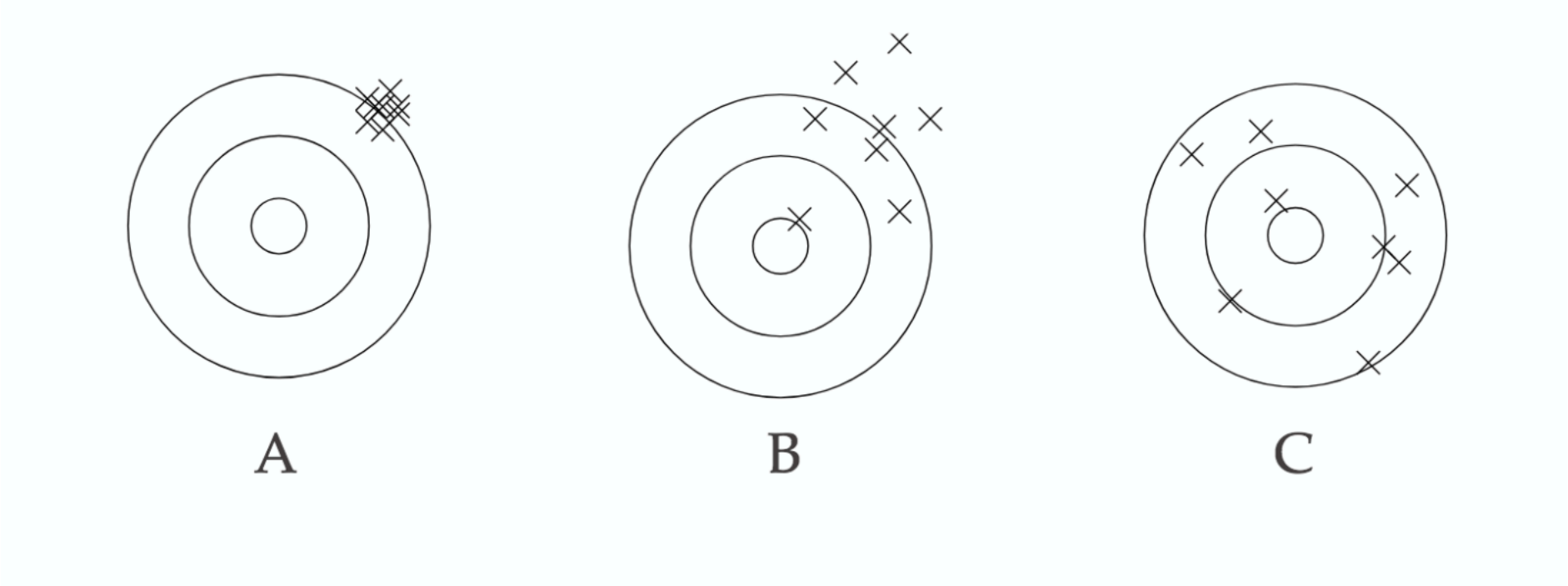
- Why is this used?
  - might not be enough data to create held-out test set
  - you cannot trust that authors did not peek at test data anyway =P

Slide: Jessica Wu

Outline for March 4

- Cross-validation
- Ensemble methods
  - Bagging
  - Boosting
  - Random Forests

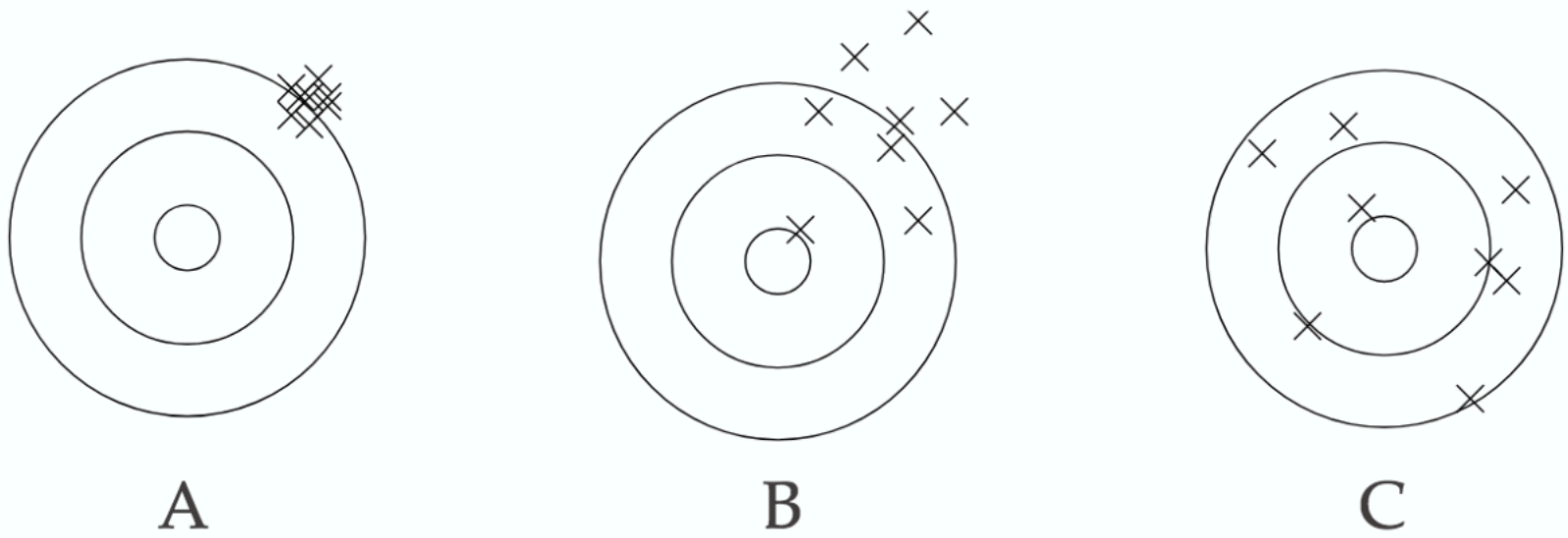
Quiz: recap bias and variance



Label each picture with variance (high or low) and bias (high or low)

Example from Ameet Soni

Quiz: recap bias and variance

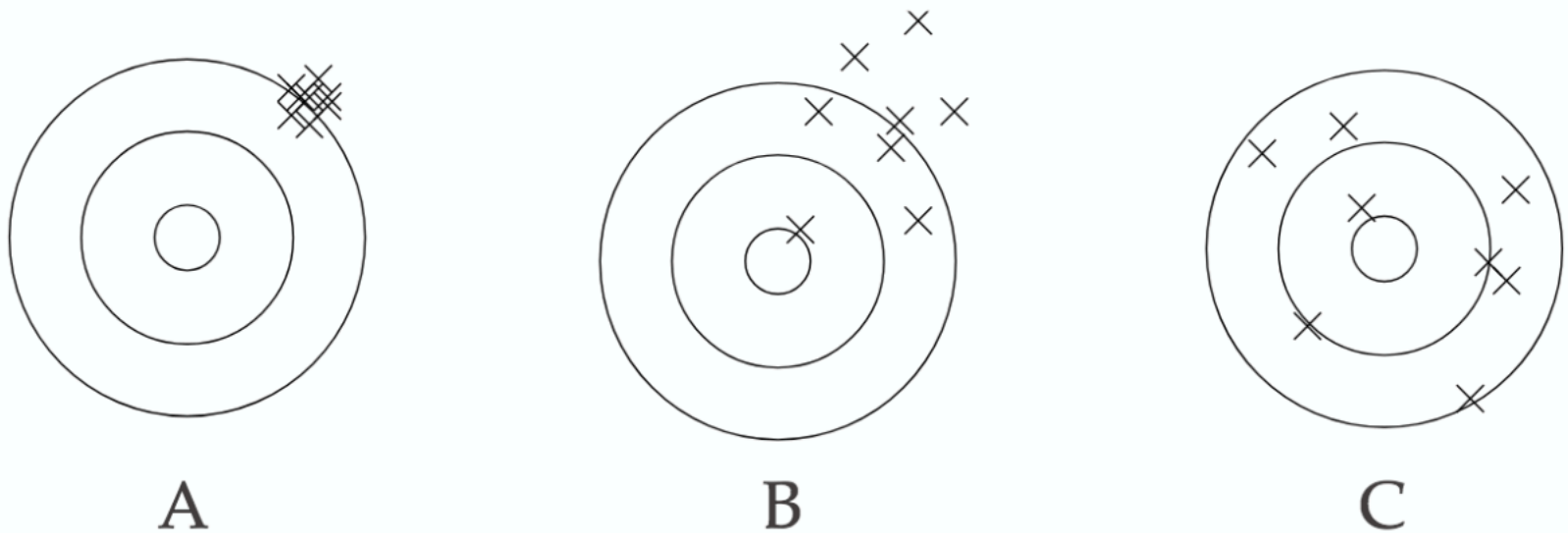


Variance: low
Bias: high

Label each picture with variance (high or low) and bias (high or low)

Example from Ameet Soni

Quiz: recap bias and variance

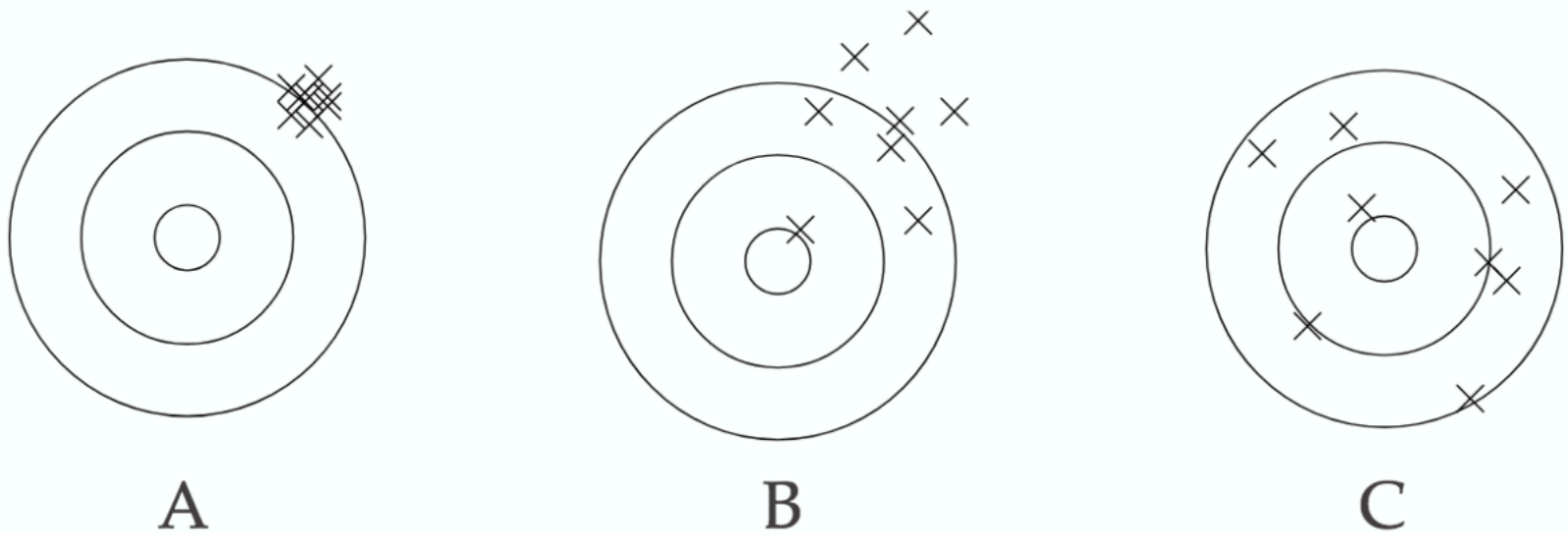


Variance:	low	Variance:	high
Bias:	high	Bias:	high

Label each picture with variance (high or low) and bias (high or low)

Example from Ameet Soni

Quiz: recap bias and variance

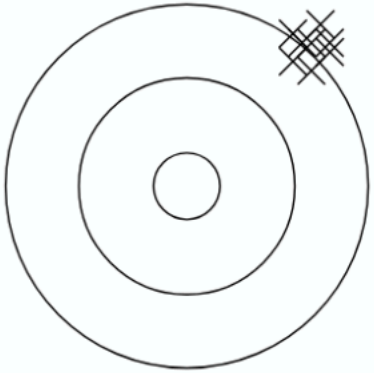


Variance:	low	Variance:	high	Variance:	high
Bias:	high	Bias:	high	Bias:	low

Label each picture with variance (high or low) and bias (high or low)

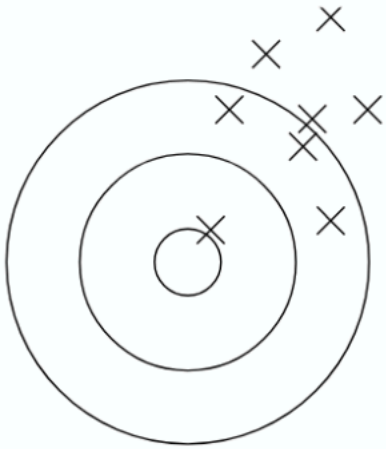
Example from Ameet Soni

Quiz: recap bias and variance



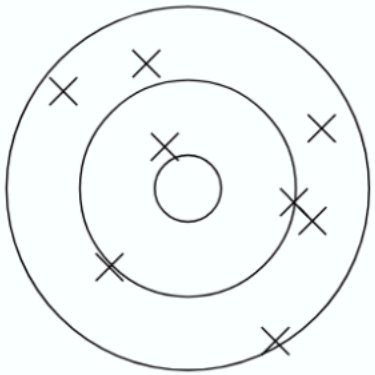
A

Variance: low
Bias: high



B

Variance: high
Bias: high



C

Variance: high
Bias: low

This is the type of classifier we want to average!

Label each picture with variance (high or low) and bias (high or low)

Example from Ameet Soni

Ensemble Idea

- Average the results from several models with high variance and low bias
  - Important that models be diverse (don't want them to be wrong in the same ways)

- If n observations each have variance s^2, then the mean of the observations has variance s^2/n (reduce variance by averaging!)

Learning Theory

Let *H* be the hypothesis space

Three sources of limitations for traditional classifiers:

- Statistical - *H* is too large relative to size of data
  - Many hypotheses can fit the data by chance
- Computational - *H* is too large to completely search for "best" model
- Representational - *H* is not expressive enough

Modified from Ameet Soni

Learning Theory

- Statistical: Average of unstable models (high variance) has more stability
- Computational: searching from multiple starting points is better approximation than one starting point
- Representational: sum of many models can represent more hypotheses than an individual model

Modified from Ameet Soni

Learning Theory

- Statistical: Average of unstable models (high variance) has more stability
- Computational: searching from multiple starting points is better approximation than one starting point
- Representational: sum of many models can represent more hypotheses than an individual model

Ensembles can address all 3!

Modified from Ameet Soni

Learning Theory

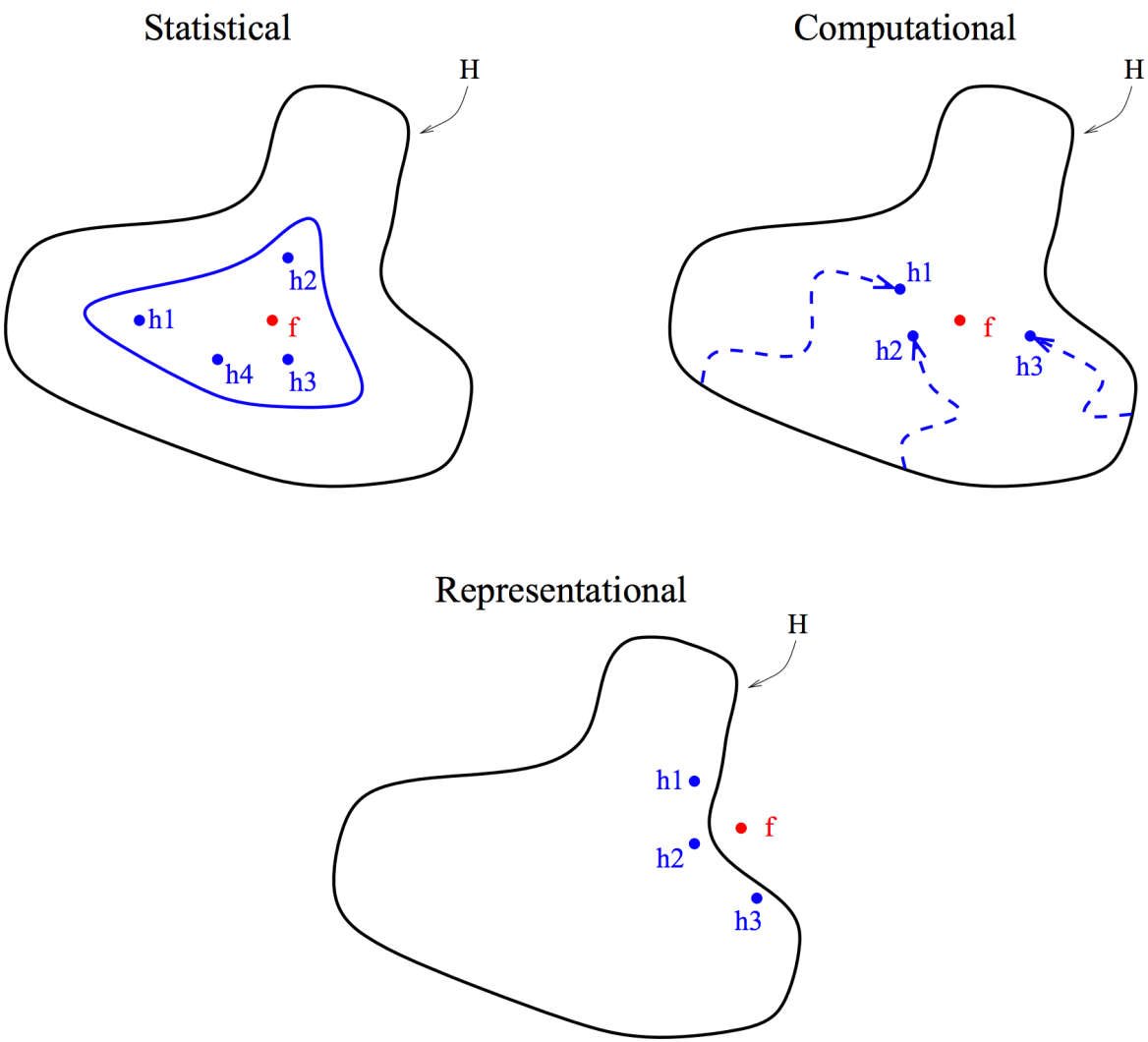
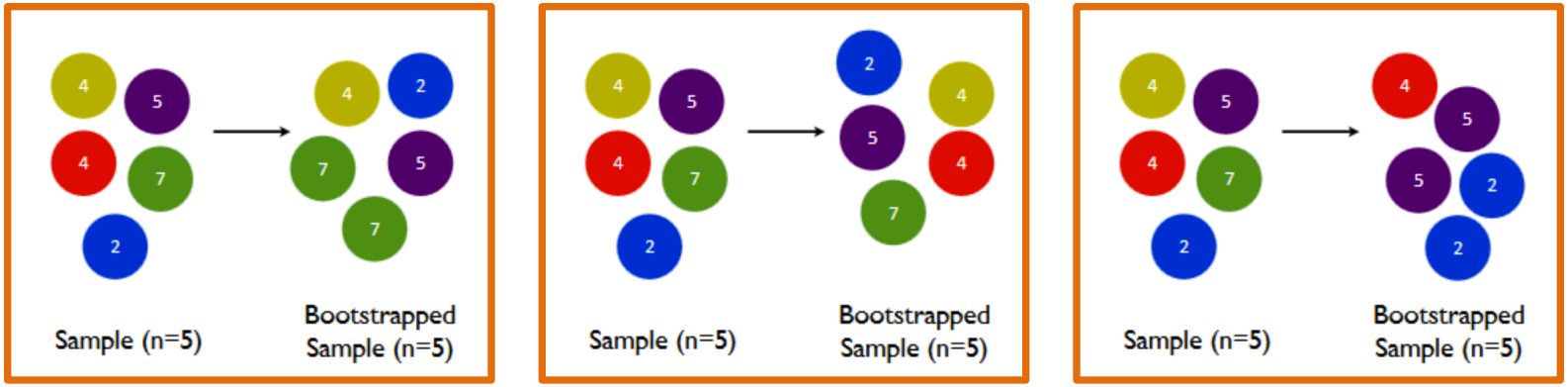


Figure from Tom Dietterich

Bagging Algorithm

- Bagging = Bootstrap Aggregation [Brieman, 1996]
- *Bootstrap* (randomly sample <u>with replacement</u>) original data to create many different training sets
- Run base learning algorithm on each new data set independently



Desmond Ong, Stanford

Modified from Ameet Soni