slides originally by
Dr. Richard Burns,
modified by
Dr. Stephanie Schwartz

CROSS VALIDATION AND SAMPLING

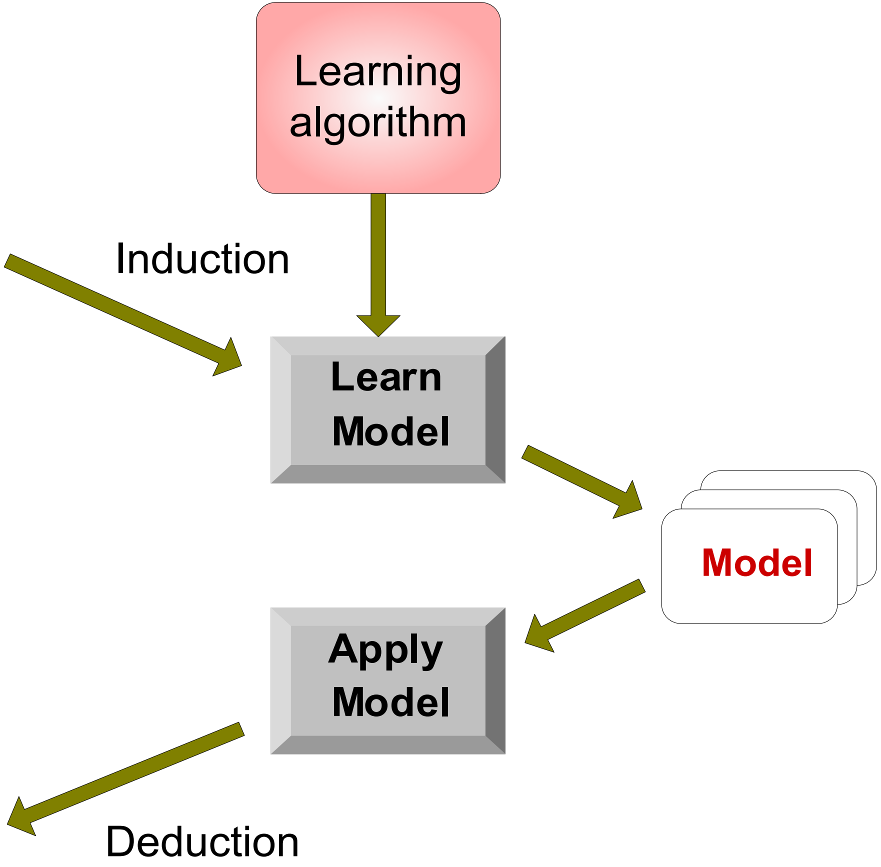
CSCI 452: Data Mining

Training Set vs. Test Set

- From Week 2:
- Overall dataset can be divided into:
  1. Training set – used to build model
  2. Test set – evaluates model

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set



Learning algorithm

Induction

Learn Model

Model

Apply Model

Deduction

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

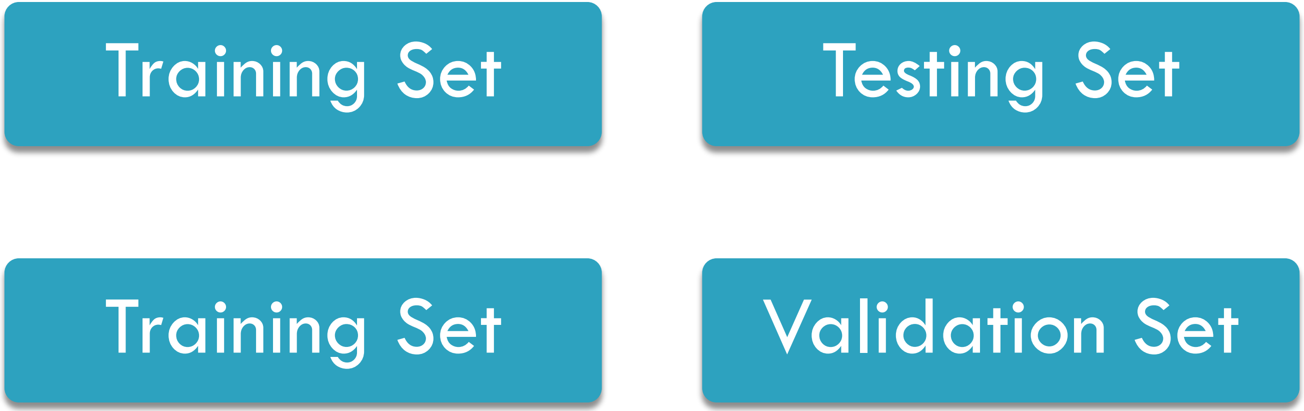
Test Set

Evaluation

- Besides:
  - Training set, testing set
- Sometimes also hear:
  - <u>Validation Set</u>
  - <u>Cross Validation</u>

Validation Set

- A set of data observations used to estimate the test error rate.
  - Like test set, validation set is held out of training data



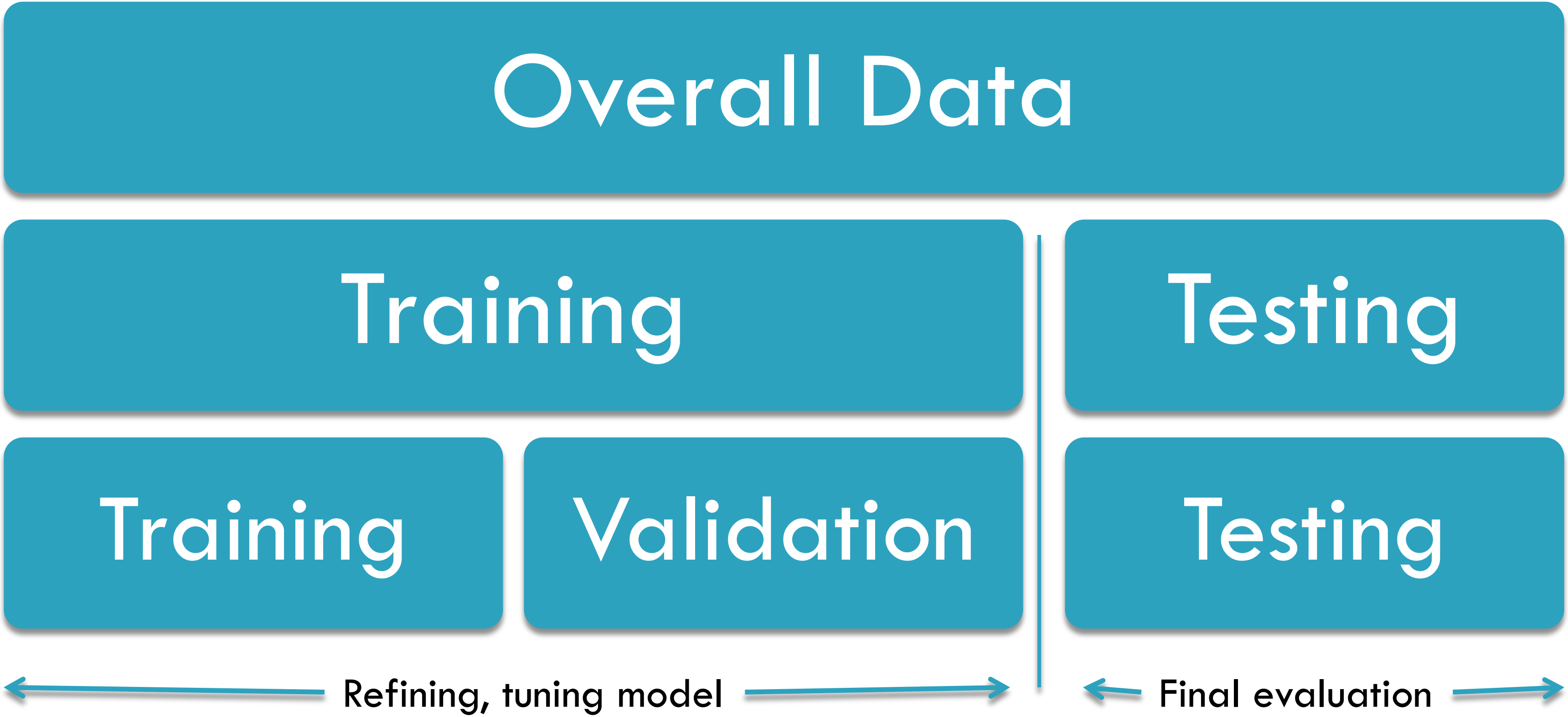
Training Set

Testing Set

Training Set

Validation Set

Training / Validation / Testing Set What's the Difference?



Overall Data

Training

Testing

Training

Validation

Testing

Refining, tuning model

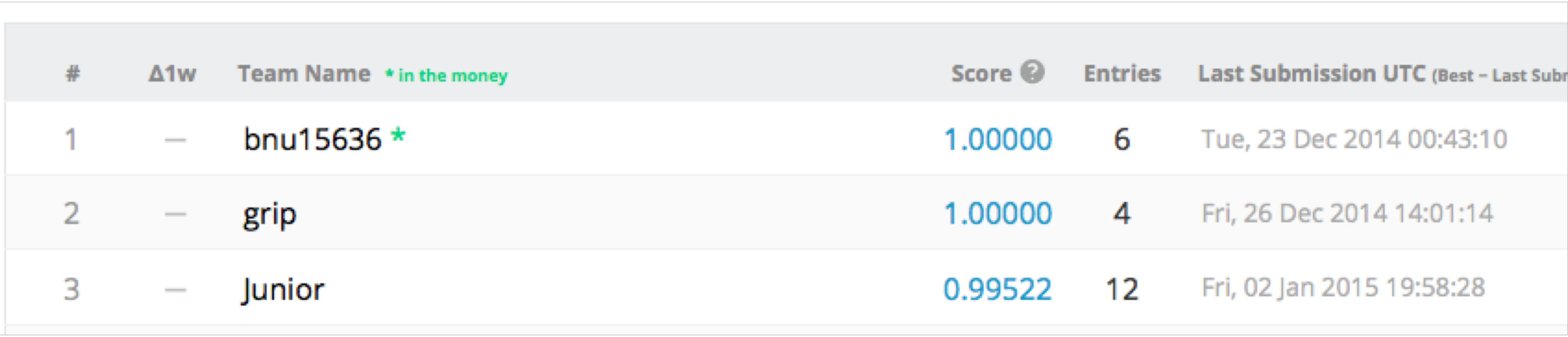
Final evaluation

Beware!

- Once we start using a validation/testing set to evaluate our model, and we want to improve our model:
  - *(making improvements to model)*
  - *error rate is decreasing: .4, .325, .31, .29, .256*
- We might be tuning our model to the validation/testing set
  - If the validation / testing set isn't changing
- This is the motivation for a final, completely held out testing set

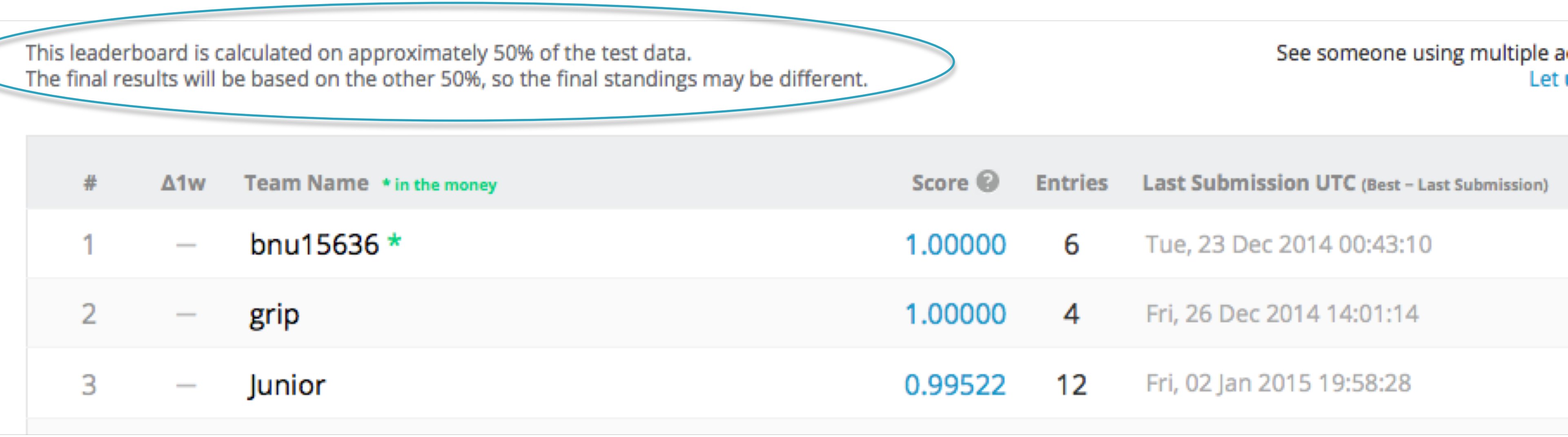
Kaggle - Titanic

- How is it possible for these two models to be perfect?
  - *Answer: It's tuning to the "test set"*
  - *How? Keep trying different predictions…*
  - Would not expect this model to definitely "be the best"



#	Δ1w	Team Name * in the money	Score	Entries	Last Submission UTC (Best – Last Subr
1	—	bnu15636 *	1.00000	6	Tue, 23 Dec 2014 00:43:10
2	—	grip	1.00000	4	Fri, 26 Dec 2014 14:01:14
3	—	Junior	0.99522	12	Fri, 02 Jan 2015 19:58:28

What does Kaggle do?

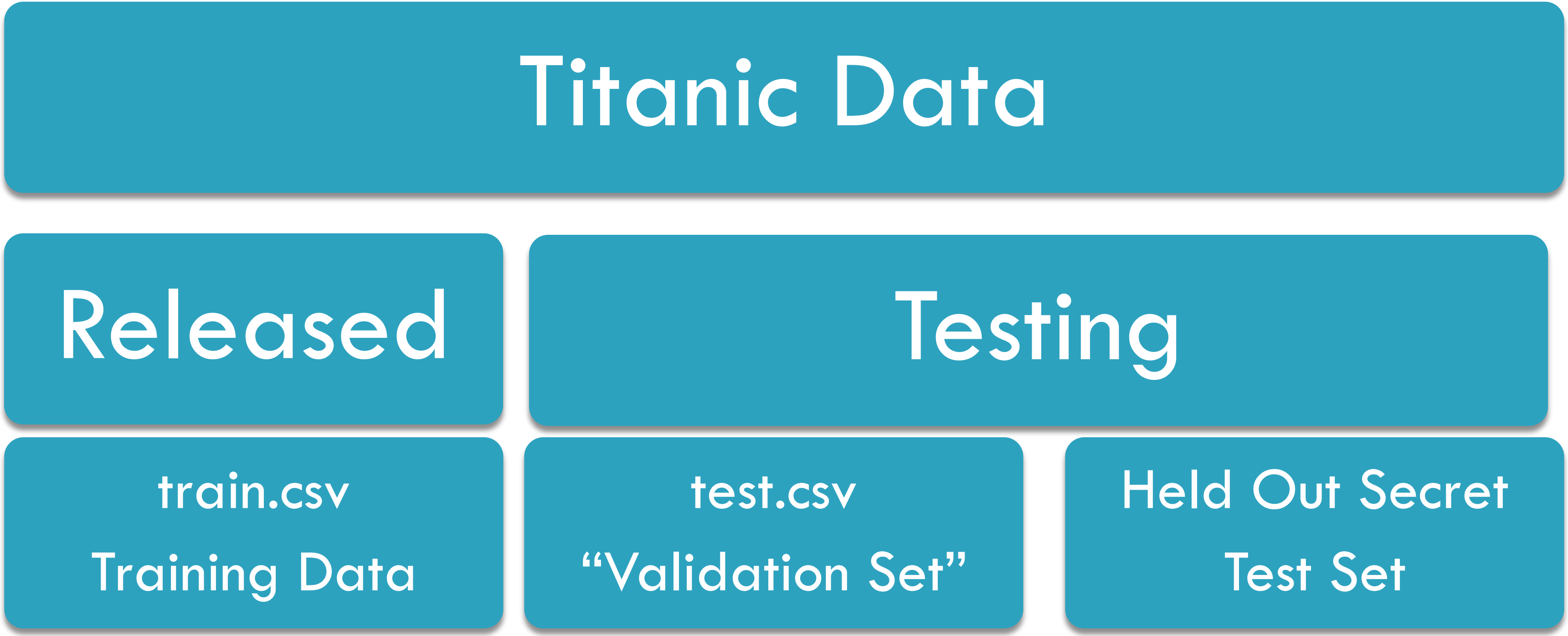


This leaderboard is calculated on approximately 50% of the test data.
The final results will be based on the other 50%, so the final standings may be different.

See someone using multiple a
Let

#	Δ1w	Team Name * in the money	Score ?	Entries	Last Submission UTC (Best – Last Submission)
1	—	bnu15636 *	1.00000	6	Tue, 23 Dec 2014 00:43:10
2	—	grip	1.00000	4	Fri, 26 Dec 2014 14:01:14
3	—	Junior	0.99522	12	Fri, 02 Jan 2015 19:58:28

Kaggle - Titanic



Titanic Data

Released

Testing

train.csv

Training Data

test.csv

“Validation Set”

Held Out Secret

Test Set

We don’t actually know
the correct predictions.

Issues with the "Holdout" Method

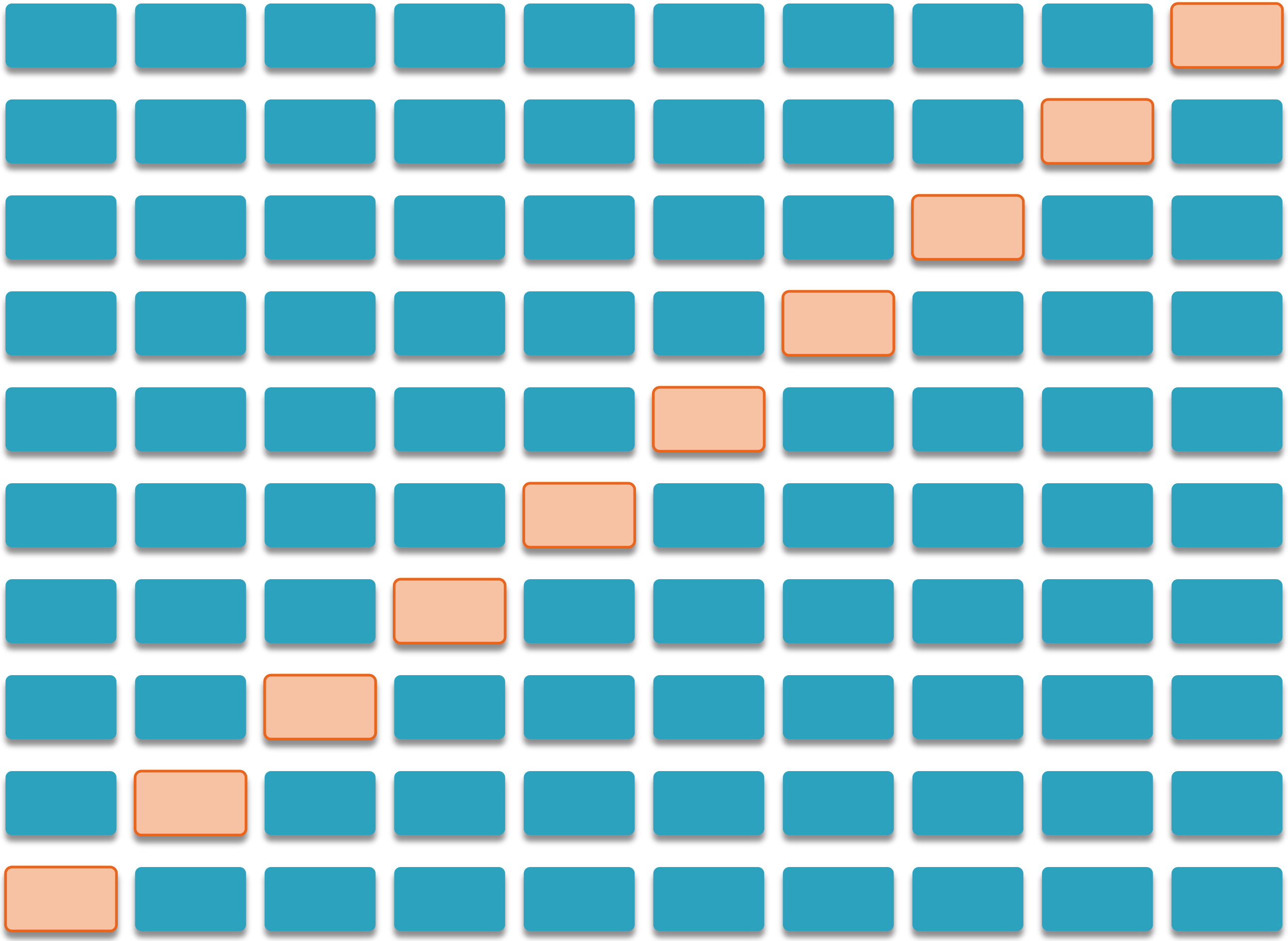
- Separate sets for training and validation/testing

1. Fewer labeled examples are available for training because some records are held out for testing
2. Model may be highly dependent on composition of training and testing sets
   - *Small training sets will have greater variance*
   - *Small testing sets will be less reliable (will have wider confidence intervals)*

Cross Validation

- Widely used alternative to single Training Set + Validation Set
- Multiple evaluations using different portions of data for training and validating/testing
- k-fold Cross Validation
  - *k = number of folds* (integer)
  - *k = 10* is common
- More computationally expensive

10-fold Cross Validation



□ 1/10th of training data

🟦 X 9 = 90% used for training

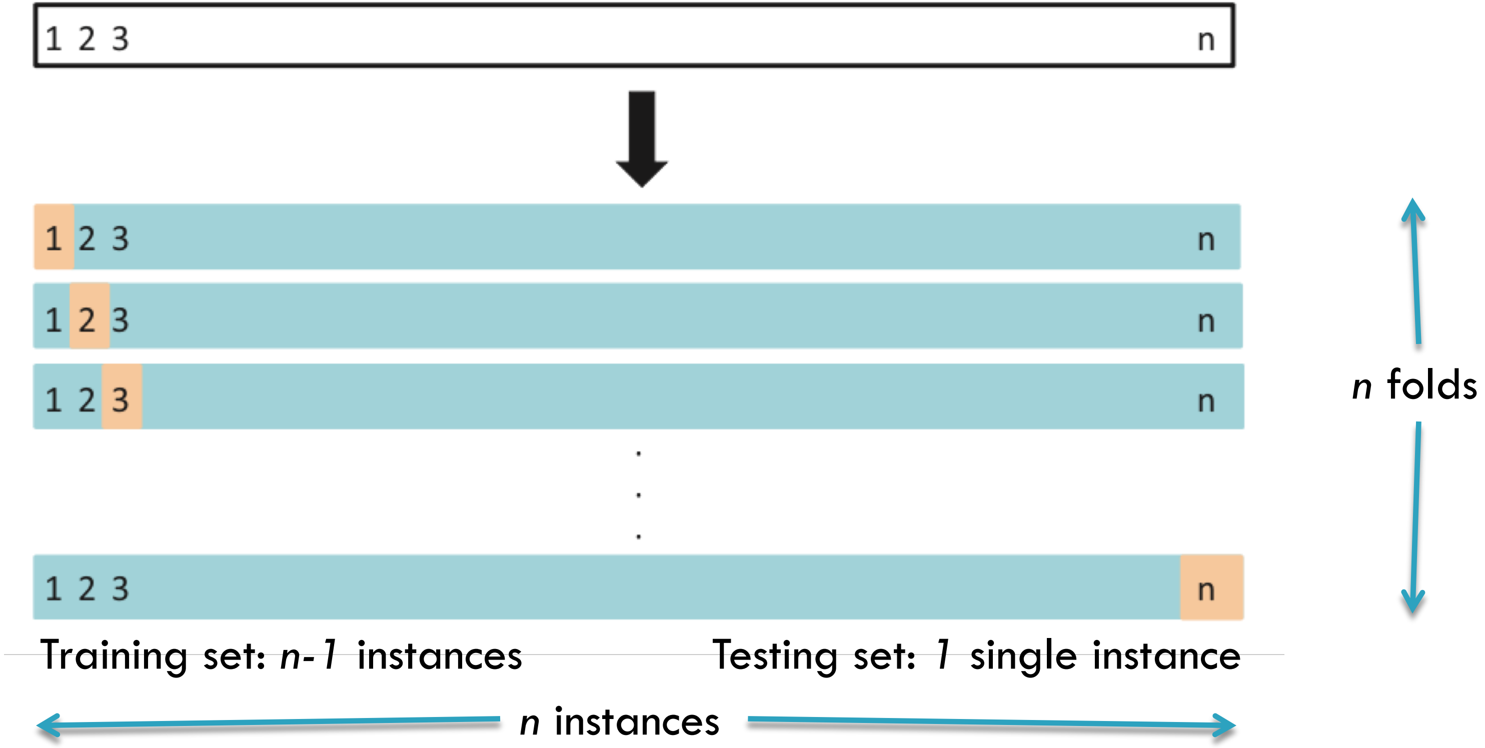
🟧 10% used for testing

- Repeat *k* times
- Average results
- Each instance will be used <u>once</u> in testing

Leave-One-Out Cross-Validation

- <u>k-fold Cross-Validation</u>
  - k = *number of folds* (integer)
  - $k = 10$ is common
- <u>Leave-One-Out Cross-Validation</u>
  - *Extreme:* $k = n$, where there are n observations in training+validation set
  - Significantly more computationally expensive

Leave-One-Out Cross-Validation



Training set: *n-1* instances

Testing set: *1* single instance

n instances

Cross-Validation Error Estimate

$$CV_k = \frac{1}{k}\sum_{i=1}^{k} ErrorRate_i$$

Average the error rate for each fold.

In leave-one-out cross-validation, since each test contains only one record, the variance of the estimated performance will be high. (Usually either 100% or 0%.)

Preprocessing: Sampling

- Common approach for selecting a <u>subset</u> of data objects to be analyzed
  - Select only some instances for the training set, instead of all of them
- *Motivation #1*: reduce dataset size so that more computationally expensive algorithm can be used
- Wait? Won't our learned models get worse since we aren't using all of the training data that we can have?
  - using a sample will work if the <u>sample is representative</u>
  - *Example*: mean of subset ("sample") is approximate to mean of original data ("population")

Preprocessing: Sampling

- Variety of sampling techniques
- Data analysts needs to choose:
  1. Sample size to generate
  2. Sampling technique to use

Sampling Approaches

- <u>Simple Random Sampling:</u> equal probability of selecting any particular item
- <u>Weighted Sampling:</u> probabilities are not uniform

Sampling Approaches

- Sampling without replacement - as each item is sampled, it is removed from the population
- Sampling with replacement - the same object/instance can be picked more than once

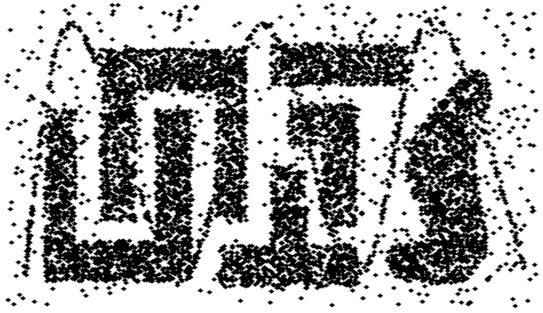
Stratified Sampling

- Simple Random Sampling can fail to represent objects that are less frequent
  - *Example problem:* a Spam-Notspam dataset where 99% of the instances in the dataset are NotSpam
  - Class Imbalance
- Some data mining techniques require proper representation of all object types
- Stratified Sampling:
  - Starts with prespecified groups of objects
  - Equal numbers of objects are drawn from each group

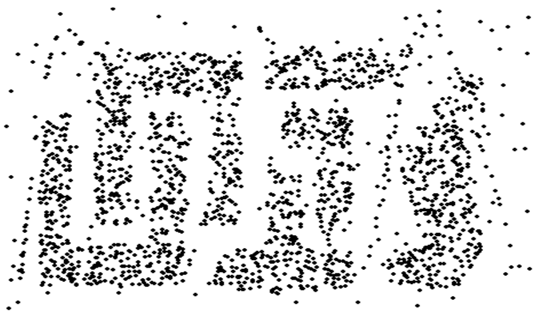
Sampling and the Loss of Information

Once a sampling technique has been selected, it is still necessary to choose the sample size.

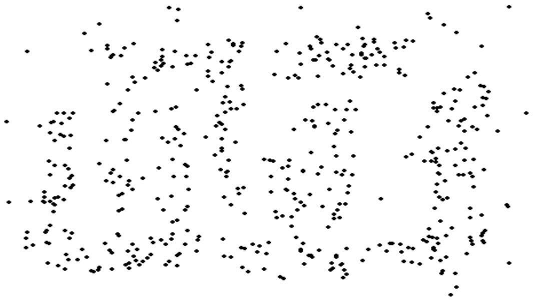
- Large sample sizes increase the probability that they are representative
  - Don’t have same computational benefit that smaller samples have
- Small sample sizes may lose patterns present in the full data



8000 points



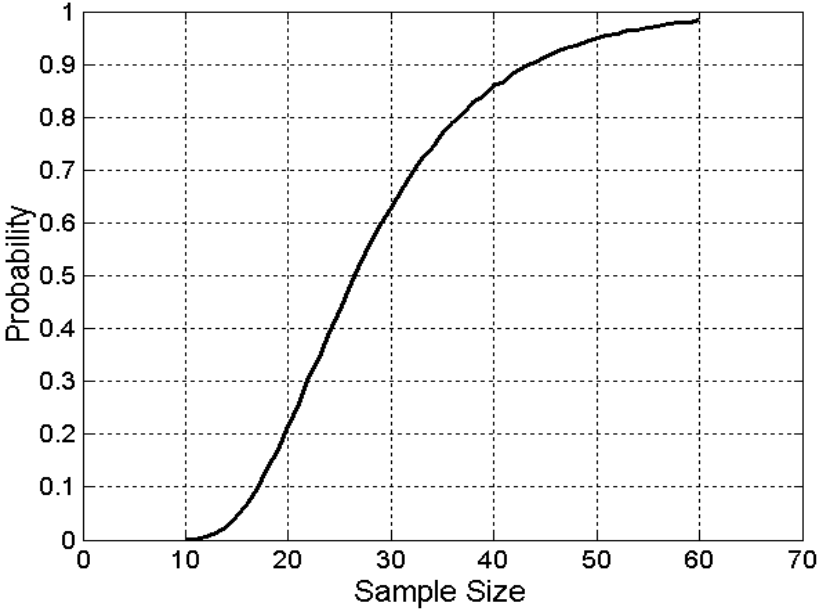
2000 Points



500 Points

Determining the Proper Sample Size

- What sample size is necessary to get at least one object from each of 10 groups? (Assuming clustering is being learned)



References

- *An Introduction to Statistical Learning*, 1st edition, James et al.
- *Introduction to Data Mining*, 1st edition, Tam et al.