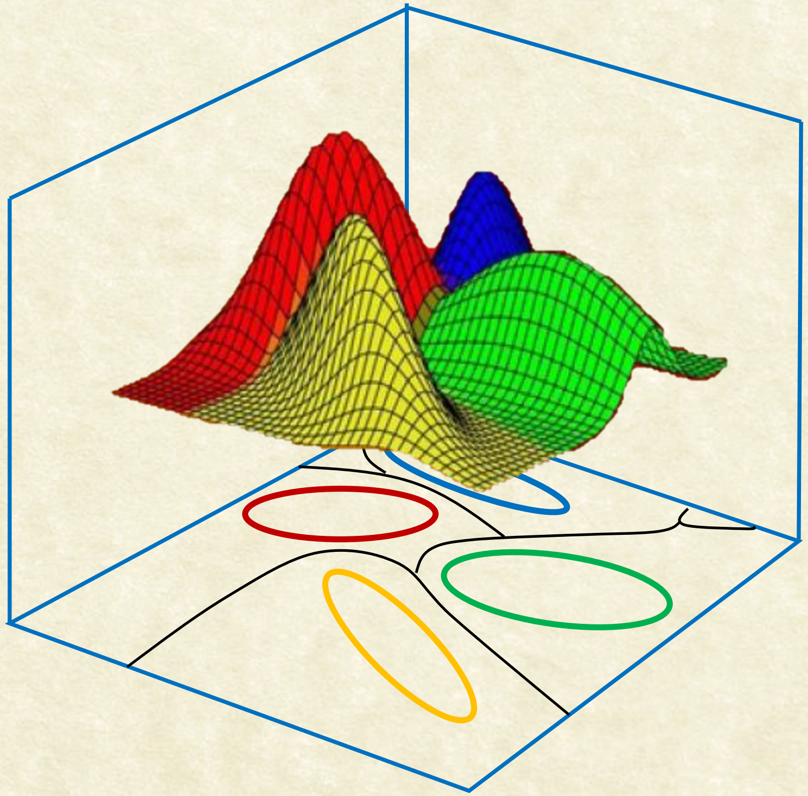
CS7.403: Statistical Methods in AI

Monsoon 2022:

Experimentation and Performance Evaluation



Anoop M. Namboodiri

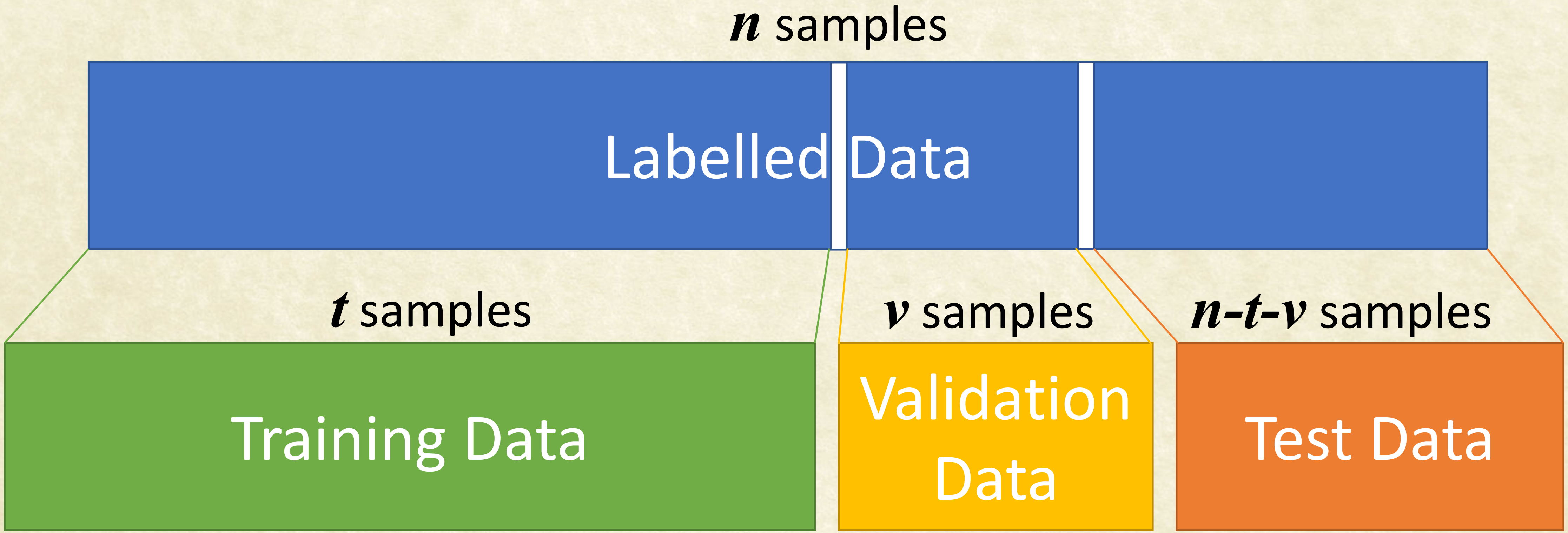
Biometrics and Secure ID Lab, CVIT,

IIIT Hyderabad

Overfitting and Generalization

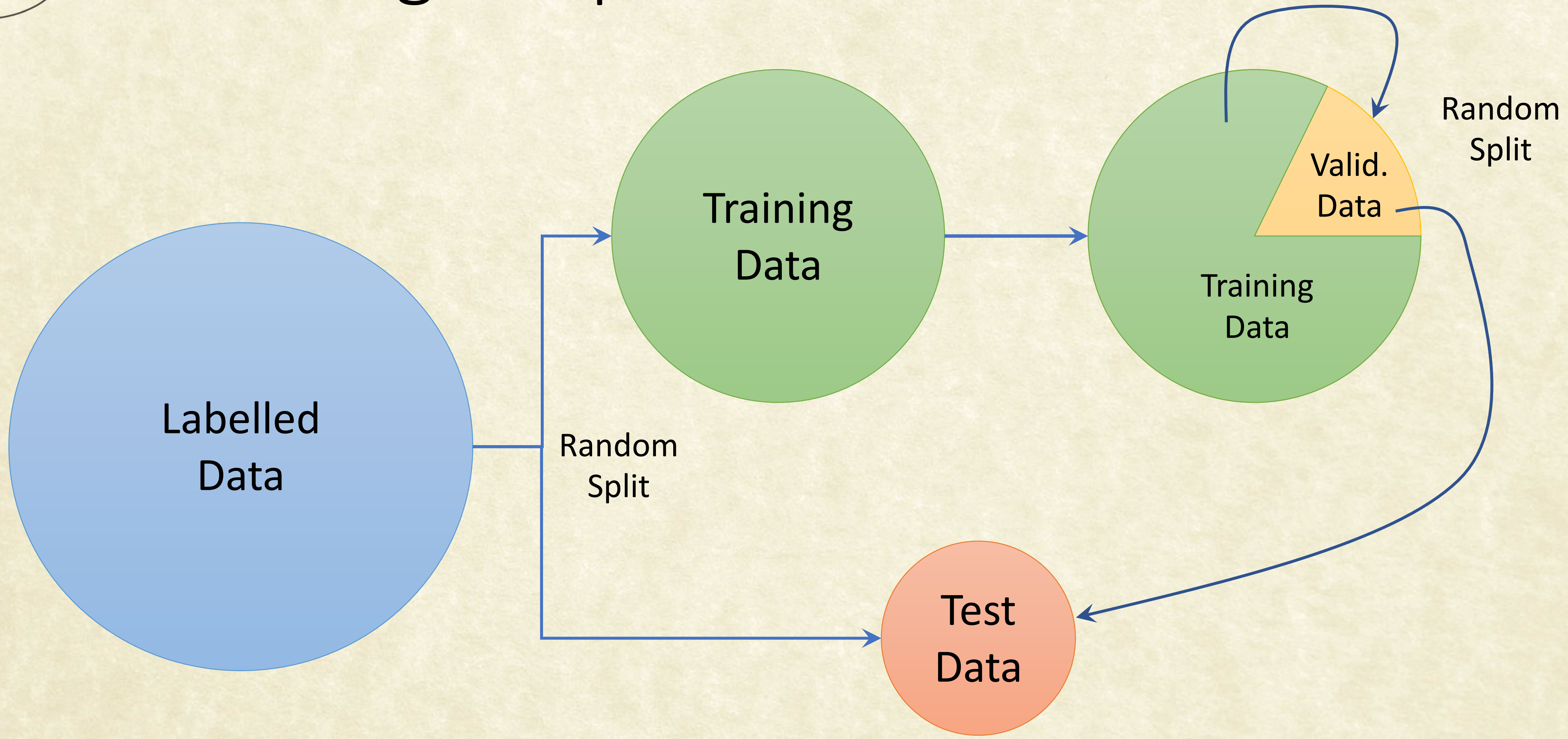
- ***Overfitting***: The process of a model becoming too specific to the training set that its performance on an independent test set becomes poor. Also known as poor ***generalization***.
- Solutions we have seen:
  1. Use of simple classifiers: Linear Models
  2. Large-Margin Classification, SVMs
  3. Ensemble Classifiers
  4. Dimensionality Reduction
- Experimental Check: Cross Validation

Training-Validation-Testing

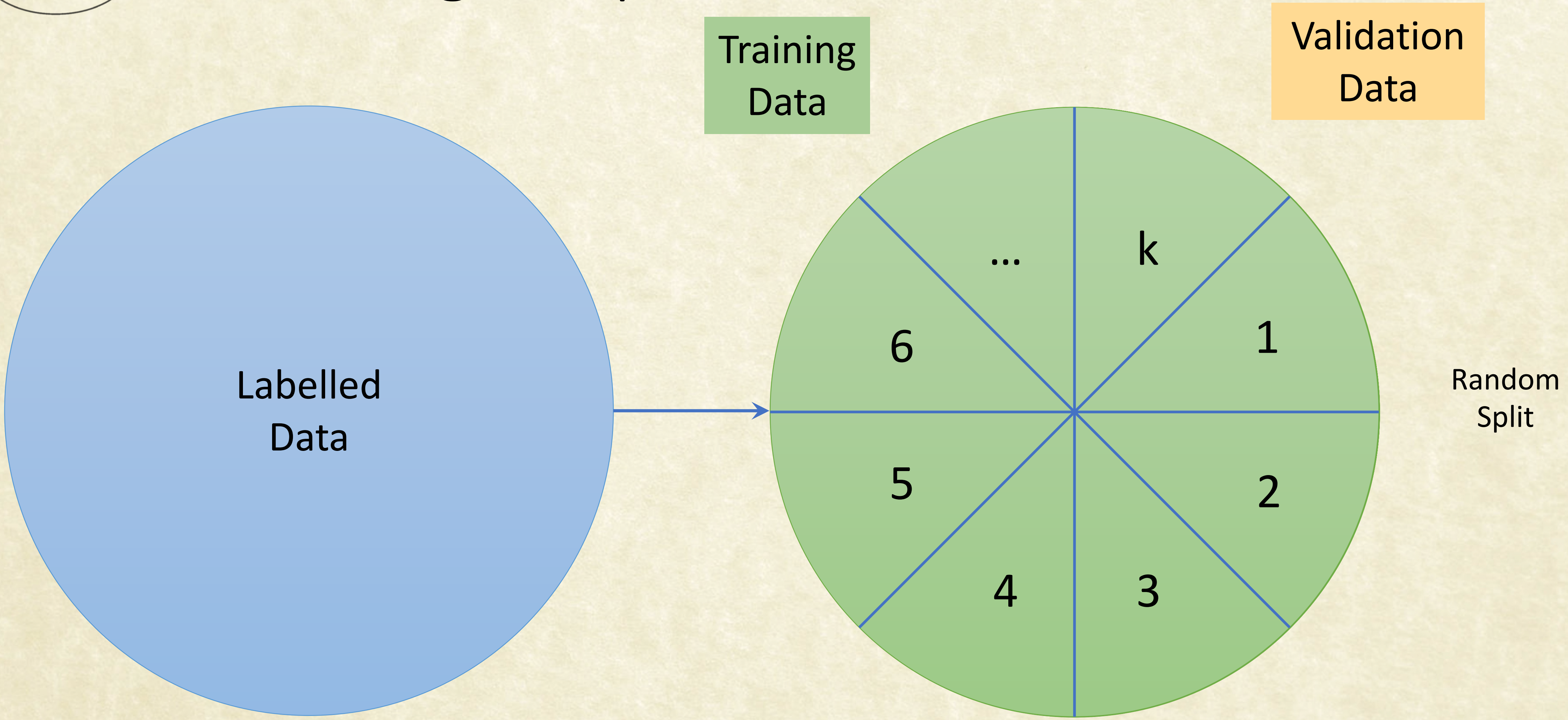


- What if our selection of validation/test set is biased
  - What happens to our model during training?
  - What happens to our estimate of error rate?

Removing the potential Bias

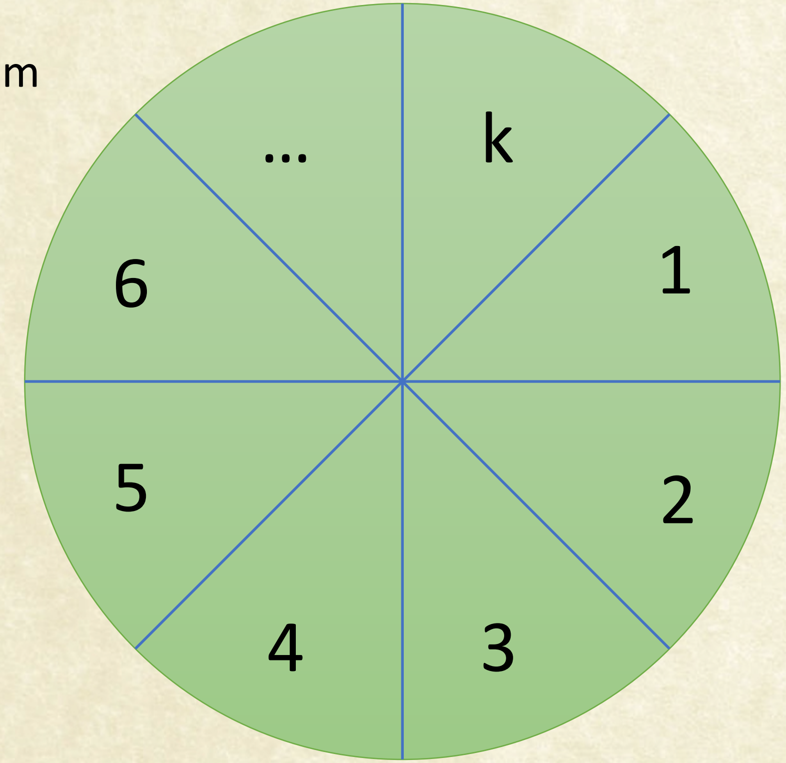


Removing the potential Bias



k-fold Cross Validation

Random Split



Train on N - {1}	Test on {1}	Error: e_1
Train on N - {2}	Test on {2}	Error: e_2
Train on N - {3}	Test on {3}	Error: e_3
Train on N - {4}	Test on {4}	Error: e_4
Train on N - {5}	Test on {5}	Error: e_5
Train on N - {6}	Test on {6}	Error: e_6
	⋮	
Train on N - {k}	Test on {k}	Error: e_k

Extreme case: $|\{i\}| = 1$
Leave-one-out Cross Validation

$$\mu_e = \frac{1}{k}\sum_{i=1}^{k} e_i$$

$$\sigma_e{}^2 = \frac{1}{k}\sum_{i=1}^{k} (e_i - \mu_e)^2$$

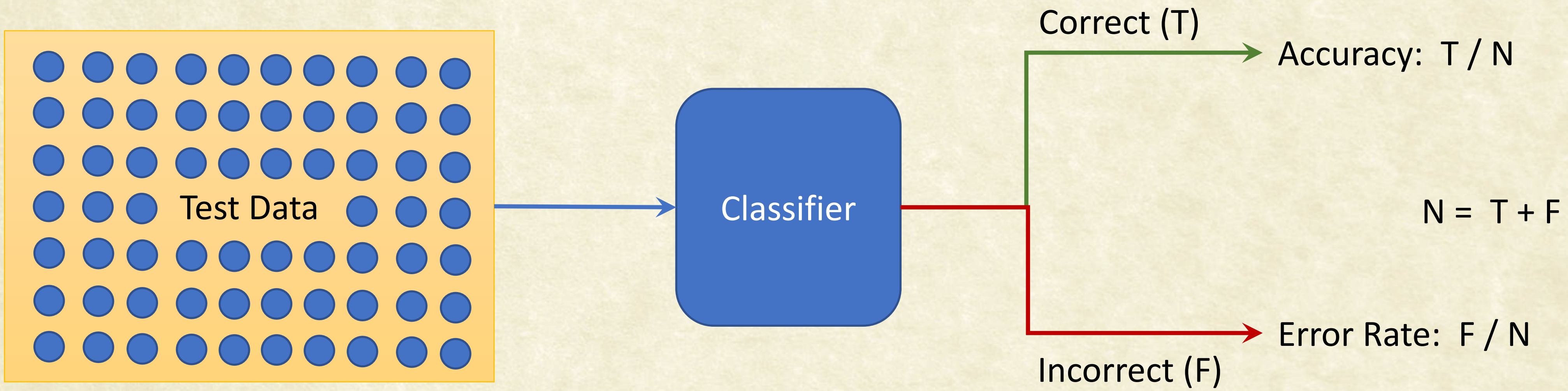
Cross Validation: Notes

- Significantly reduces the bias of error estimate
  - Note: Each error estimate may not be reliable, but the mean is.
- Provides a confidence interval for error estimate
- Can use the whole data for both training and validation
- Does not produce a single trained model
  - May train the model using the whole labelled data
- Takes more time to train

Questions?

Evaluating Classification

Recap: Accuracy



Binary Classification

- Class Labels: {0 , 1} or {- , +}
- Test Samples: 150; {66 , 84}
- Accuracy = (61 + 77) / 150 = 0.92
- Misclassification = (7 + 5) / 150 = 0.08
- True Positive Rate (TP) = 77 / 84 = 0.92
- False Positive Rate (FP) = 5 / 66 = 0.08

N = 150	**Predicted: -**	**Predicted: +**	
Actual: -	TN = 61	FP = 5	66
Actual: +	FN = 7	TP = 77	84
	68	82	

Related Terms

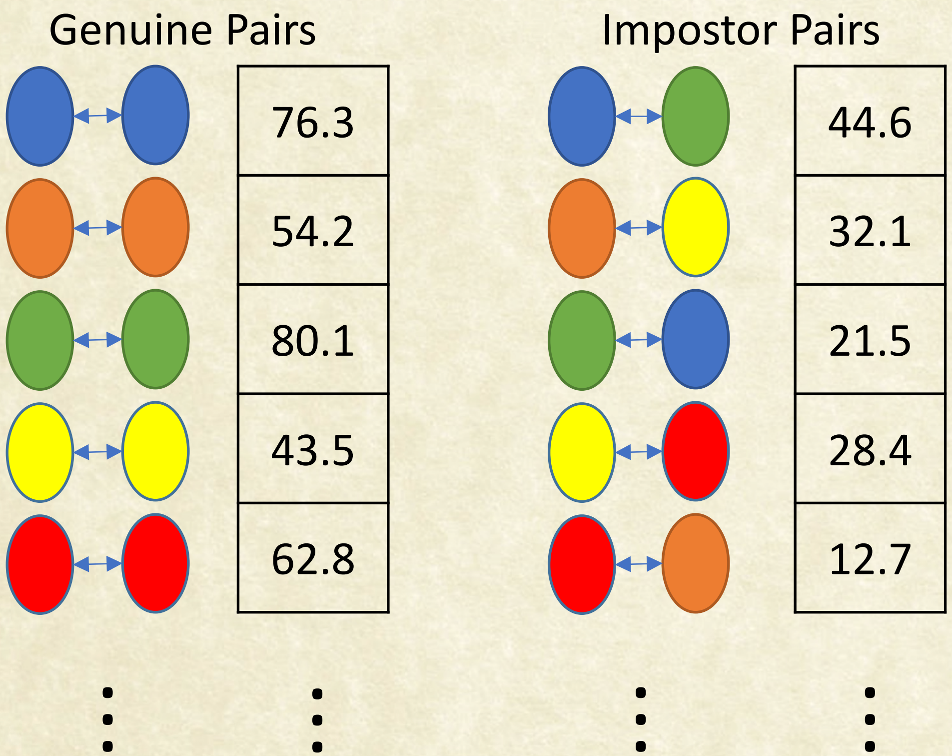
- Hit = 77/84 (TPR)
- Miss = 7/84 (FNR)
- False Alarm = 5/66 (FPR)
- Genuine Reject = 61/66 (TNR)

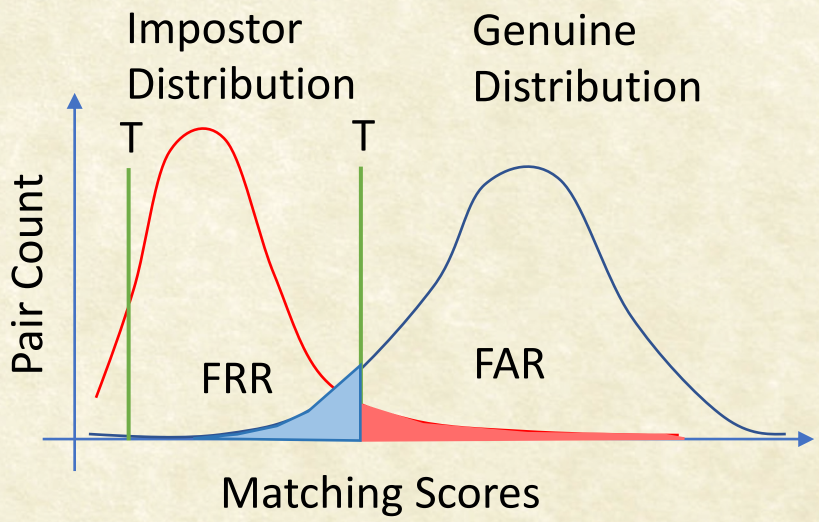
- Precision: 77 / 82
- Recall : 77 / 84 (TPR)

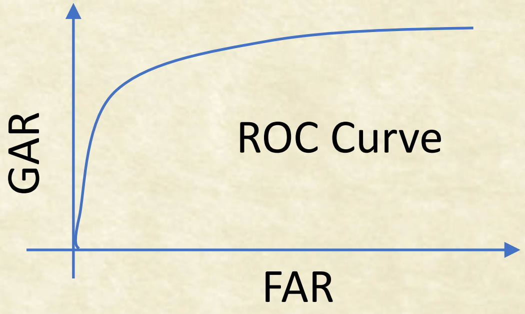
N = 150	Predicted: -	Predicted: +	
Actual: -	TN = 61	FP = 5	66
Actual: +	FN = 7	TP = 77	84
	68	82	

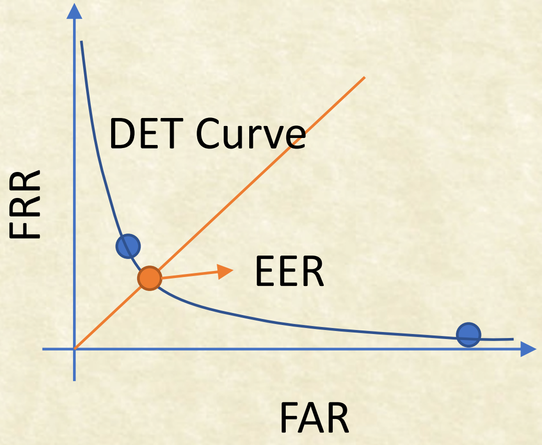
Evaluating Verification

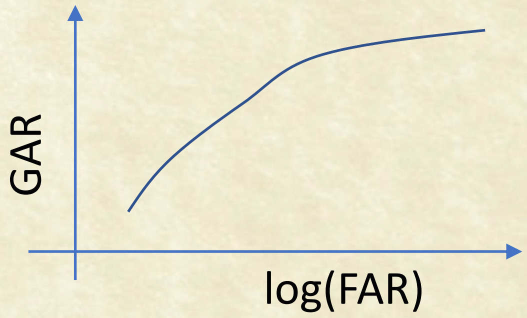
- Test Data:











FMR100 (best FNMR/FMR<=1%)
FMR1000 (best FNMR/FMR<=0.1%)

Which rates are Important?

- Screening for a terminal disease
  - Do not want to miss anyone: Low Miss Rate, High Recall
- Classification between apples and oranges
  - Both types of errors are equally imp.: High Accuracy
- Automatic bombing on detecting a target from a drone
  - Should not hurt civilians: Zero False Alarms
- Giving access to a secure installation
  - Should not give access to unauthorized personnel: Low False Positives

Extension to Multi-class Classifier

Confusion Matrix

Pred → Act ↓	1	2	3	4	5	6
1	98	2	0	0	0	0
2	1	97	1	0	1	0
3	0	0	100	0	0	0
4	0	0	1	99	0	0
5	0	0	0	0	100	0
6	0	1	0	1	2	96

Questions?

Precision vs Recall

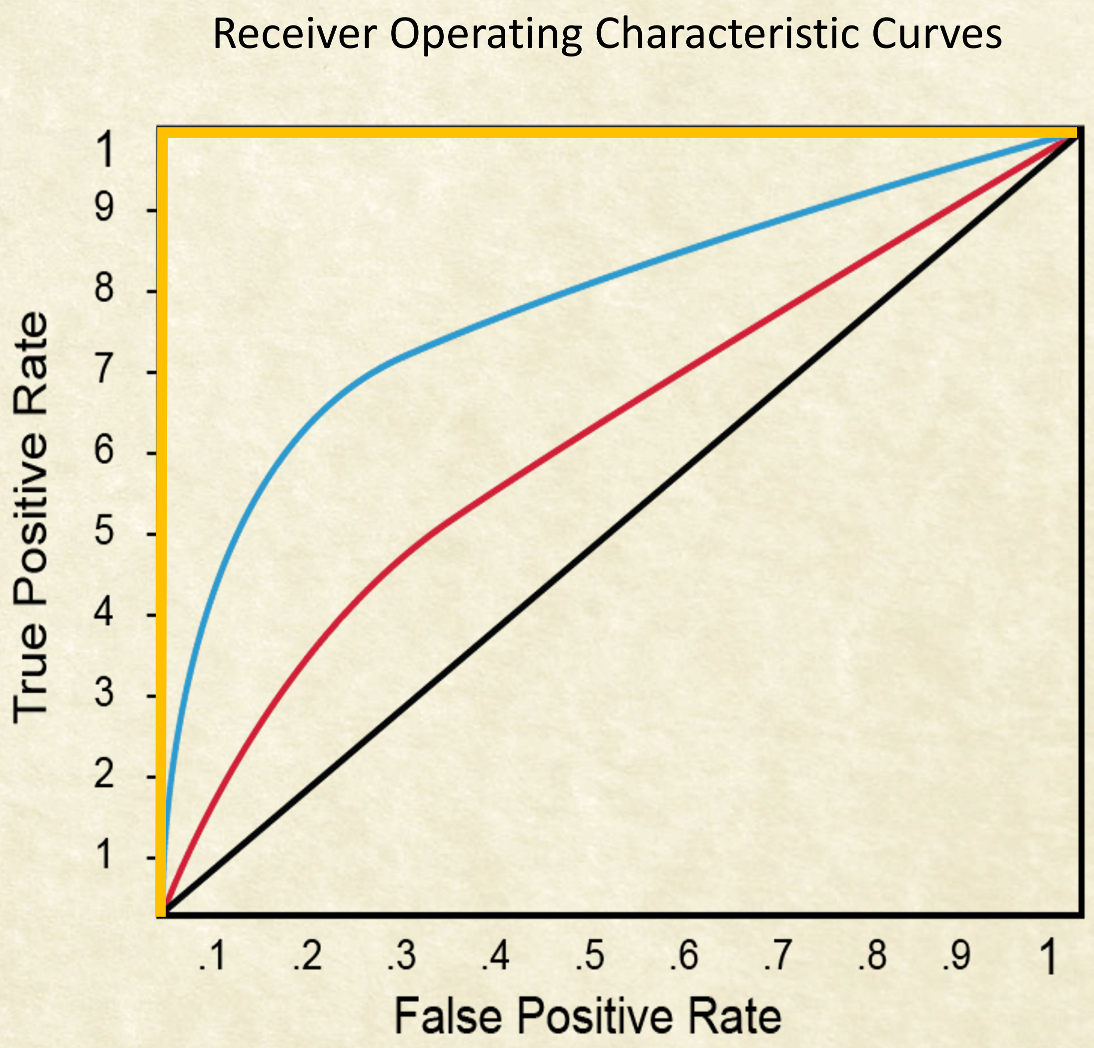
The Tradeoff

Precision Recall Tradeoff

- In most cases one can tradeoff between the two types of errors.
- Consider a search for a document from a database:
- Strategy
  - Retrieve all document that are closer than a threshold in distance
- The Tradeoff
  - Low distance threshold
    - Fewer false positives; High Precision
    - More misses (false negatives), Low Recall
  - High distance threshold
    - Higher Recall, Lower Precision

ROC Curve

- As threshold varies, we move along a curve
- Different representations / distance metrics / algorithms produce different curves
- Blue > Red > Black
- Ideal
- Other variants exist
  - **D**etection **E**rror **T**radeoff Curves
  - FPR vs. FNR
- Semi-log plots



F-1 Measure: A Single Metric

- One classifier has high Precision but lower Recall; Another behaves exactly the opposite
- F-1 Measure (Information Retrieval)

$$F_1 = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}}$$

- Punishes extreme values more
- Definition of Recall and Precision have same numerator, different denominators. A sensible way to combine them is harmonic mean.

Notes on Performance Metrics

- Use the right metric based on the type of problem
- Use a chart that best demonstrates/compares the results
- Use cross-validation whenever possible
  - Report the standard deviation of accuracies
  - Use it to decide if a difference is significant
- Use Single Metrics when appropriate
  - F-Measure
  - Area Under Curve

Questions?