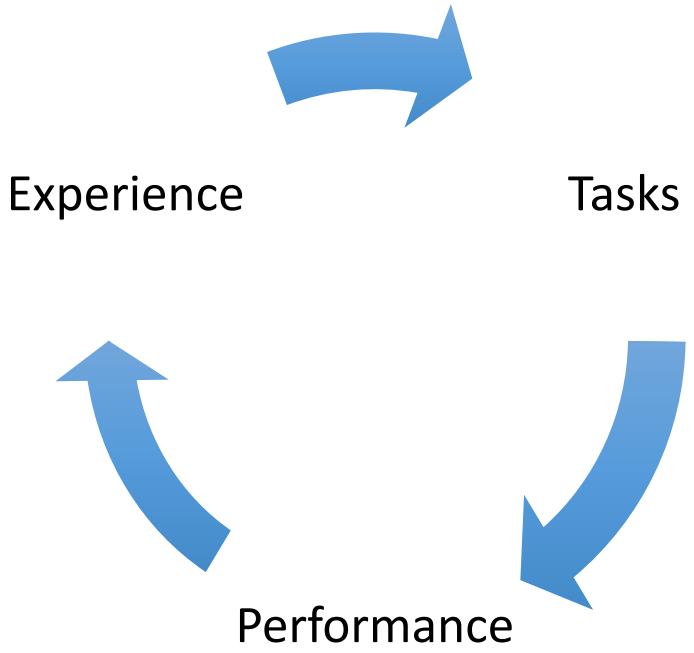
DIAG 2016
Machine Learning:
Classification Techniques in R
AJG Project

adip@umich.edu

What is Machine Learning?

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."



- The humans provide : definition of Tasks (T) and measure of Performance (P)
- The machine leverages Experience (E) to get better at T as measured by P
- ML is also known as 'Statistical Learning'

What is Classification

- In machine learning terms, categorizing data points is a **classification** task.
  - Binary Classification involves two classes e.g. 0 or 1
  - Multi-class Classification involves multiple classes
- A classifier algorithm (procedure) implements a ‘loss function’ (performance) that is optimized using the training data (experience)
- This is known as **Supervised Learning** as the data is labelled
  - Unsupervised learning is used in clustering (e.g. Market Segmentation)

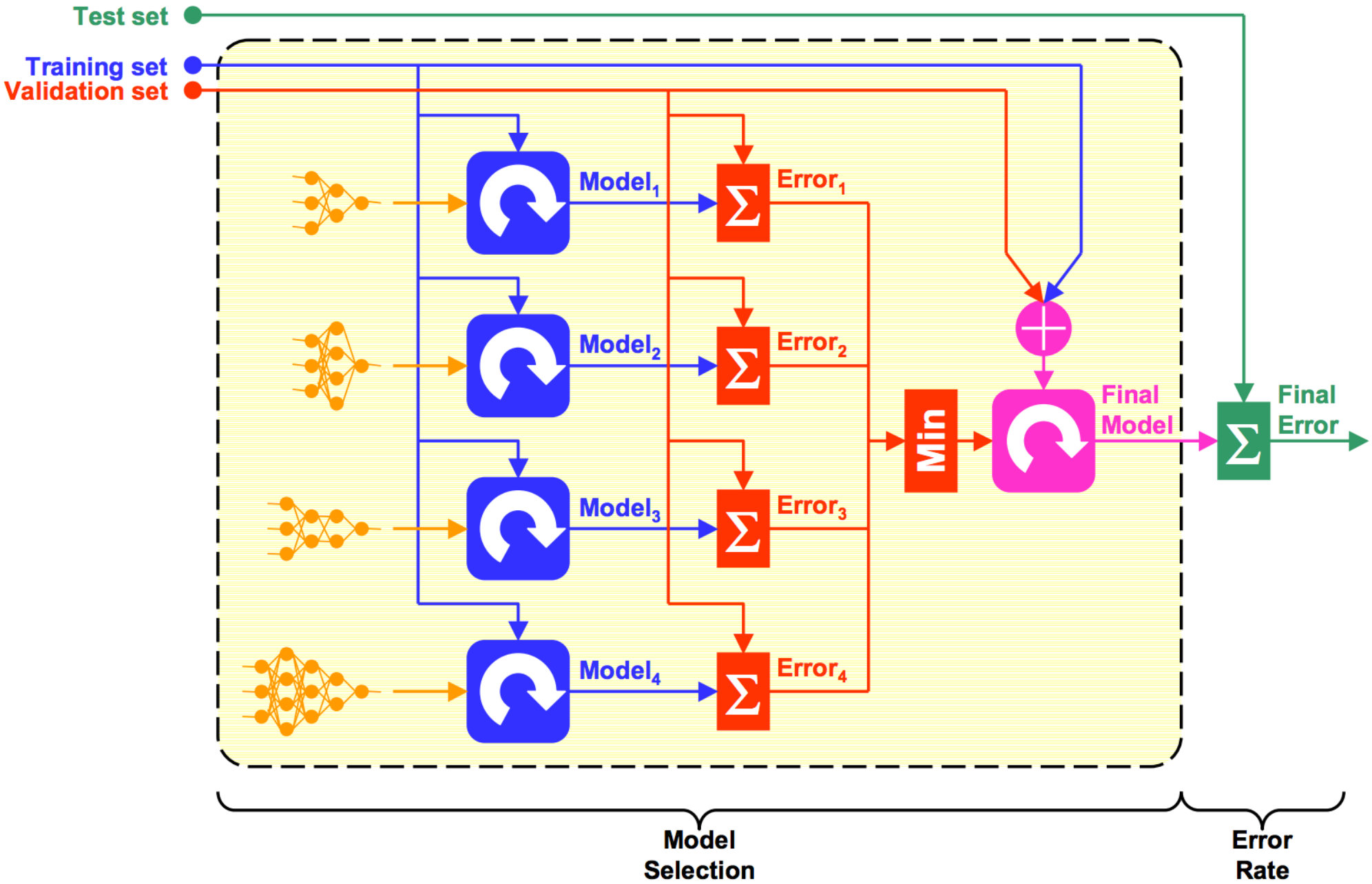
Different Kinds of Classification Methods

- **Logistic Regression**
- **Classification and Regression Trees (CART)**
  - Also known as **'Decision Trees'** in some domains
- SVM
- **Random Forest**
- Etc.

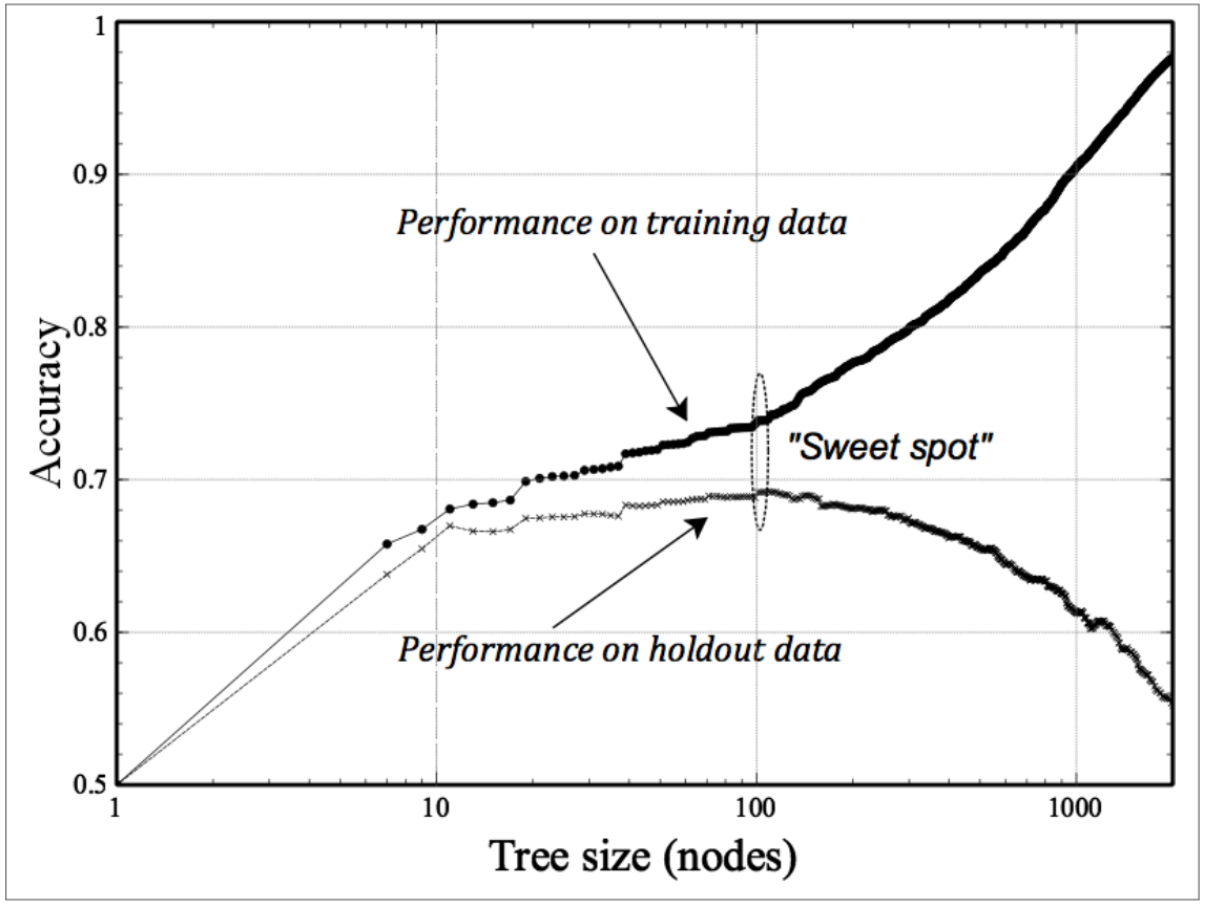
Overfitting is the WORST THING EVER

- Overfitting is when your model ‘memorizes’ the data set and uses the data to predict itself! That is USELESS
- To avoid this, data sets are split into three:
  - **Training Set** is a set of **labelled** data used to train a Machine Learning Model
  - **Validation Set** is a set of **labelled** data used to measure the Model Performance
  - **Test Set** is a set of **un-labelled** data used to make predictions via the trained and validated model

Three-way data splits



Tune to max(perf) && min(overfitting)



How to measure model performance

- Confusion Matrix!

n=165	**Predicted: NO**	**Predicted: YES**
Actual: NO	50	10
Actual: YES	5	100

- **true positives (TP):** These are cases in which we predicted yes (they have the disease), and they do have the disease.
- **true negatives (TN):** We predicted no, and they don't have the disease.
- **false positives (FP):** We predicted yes, but they don't actually have the disease. (Also known as a "Type I error.")
- **false negatives (FN):** We predicted no, but they actually do have the disease. (Also known as a "Type II error.")

- http://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/

Model Performance

- **Accuracy:** Overall, how often is the classifier correct?
  - (TP+TN)/total = (100+50)/165 = 0.91
- **Misclassification Rate:** Overall, how often is it wrong?
  - (FP+FN)/total = (10+5)/165 = 0.09
  - equivalent to 1 minus Accuracy
  - also known as "Error Rate"
- **True Positive Rate:** When it's actually yes, how often does it predict yes?
  - TP/actual yes = 100/105 = 0.95
  - also known as "Sensitivity" or "Recall"
- **False Positive Rate:** When it's actually no, how often does it predict yes?
  - FP/actual no = 10/60 = 0.17
- **Specificity:** When it's actually no, how often does it predict no?
  - TN/actual no = 50/60 = 0.83
  - equivalent to 1 minus False Positive Rate

Interactive Tutorial

- Lets do the Kaggle ‘Titanic Challenge’
  - You can do it yourself later on www.kaggle.com
- Predict which Passengers survived the Titanic disaster

Questions?

adip@umich.edu