Intro to Supervised Machine Learning in R

Thomas W. Jones
for
District Data Labs

Introductions

Goals & Scope

- What this class is:
  - Designed for you to “do”
  - Focused on code
- What this class is not:
  - Focused on math
  - Theoretically rigorous

To That End…

Emphasis is on *out-of-sample* predictive power

Agenda

- A Motivating Example
- Overview of Supervised Machine Learning
- Your Project: R, R Studio, R Studio Projects
- Classification
- Numeric Prediction
- Q&A or Excursions

Agenda

- **A Motivating Example**
- Overview of Supervised Machine Learning
- Your Project: R, R Studio, R Studio Projects
- Classification
- Numeric Prediction
- Q&A or Excursions

Prediction of Injury in Automobile Accidents

IMPACT RESEARCH

www.impactresearchinc.com

AUTO ALLIANCE

www.nhtsa.gov

Sample Data Set

- National Automotive Sampling System / Crashworthiness Data System (NASS/CDS)
- 10,000 observations sampled from tow-away crashes 2000 - 2014
- Check out data dictionary

Agenda

- A Motivating Example
- **Overview of Supervised Machine Learning**
- Your Project: R, R Studio, R Studio Projects
- Classification
- Numeric Prediction
- Q&A or Excursions

“Supervised” =

“I already have some measured outcome”

Goals of Supervised Machine Learning

- Prediction
- Inference
- Generalizability

Two Types of Prediction

- Classification = I have buckets
    - e.g. - Will the Caps win the playoffs?
- Numeric Prediction = I have numbers
    - e.g. - How much money will I win or lose when the playoffs are over?

Two Types of Models

- Parametric
  - Easy to interpret
  - (Usually) not great predictors
- Non-Parametric
  - Really hard to interpret
  - (Usually) better predictors

Models Covered Here

Supervised Methods in This Course

	Task	Type
Linear Regression	Numeric Prediction	Parametric
Logistic Regression	Classification	Parametric
Decision Trees	Classification	Non-parametric
Regression Trees	Numeric Prediction	Non-parametric
Random Forrests	Classification or Numeric Prediction	Non-parametric
Support Vector Machines	Classification or Numeric Prediction	Non-parametric

Workflow

1. Summarize the data

2. Partition training and test sets

3. Fit or train a model

4. Evaluate the model

5. Deploy the model

Workflow

1. Summarize the data

2. Partition training and test sets

3. Fit or train a model

4. Evaluate the model

5. Deploy the model

Agenda

- A Motivating Example
- Overview of Supervised Machine Learning
- **Your Project: R, R Studio, R Studio Projects**
- Classification
- Numeric Prediction
- Q&A or Excursions

Have you/are you…?

- Running the latest version of R from CRAN?
  - 3.2.5
    - http://www.cran.r-project.org
- Running the latest (stable) version of RStudio?
  - https://www.rstudio.com/products/rstudio/download/

Tour of R

Agenda

- A Motivating Example
- Overview of Supervised Machine Learning
- Your Project: R, R Studio, R Studio Projects
- **Classification**
- Numeric Prediction
- Q&A or Excursions

Classification

- Definition: task of assigning objects to one or more *pre-defined* categories
- Examples:
  - Detect spam from legitimate email
  - Categorize cells as malignant or benign
  - Identify handwritten letters

For Classification You Need…

- Pre-defined outcome/class variable
- Outcome is categorical
- Outcome is tied to predictor variables

Classifier Evaluation 1: Confusion Matrix

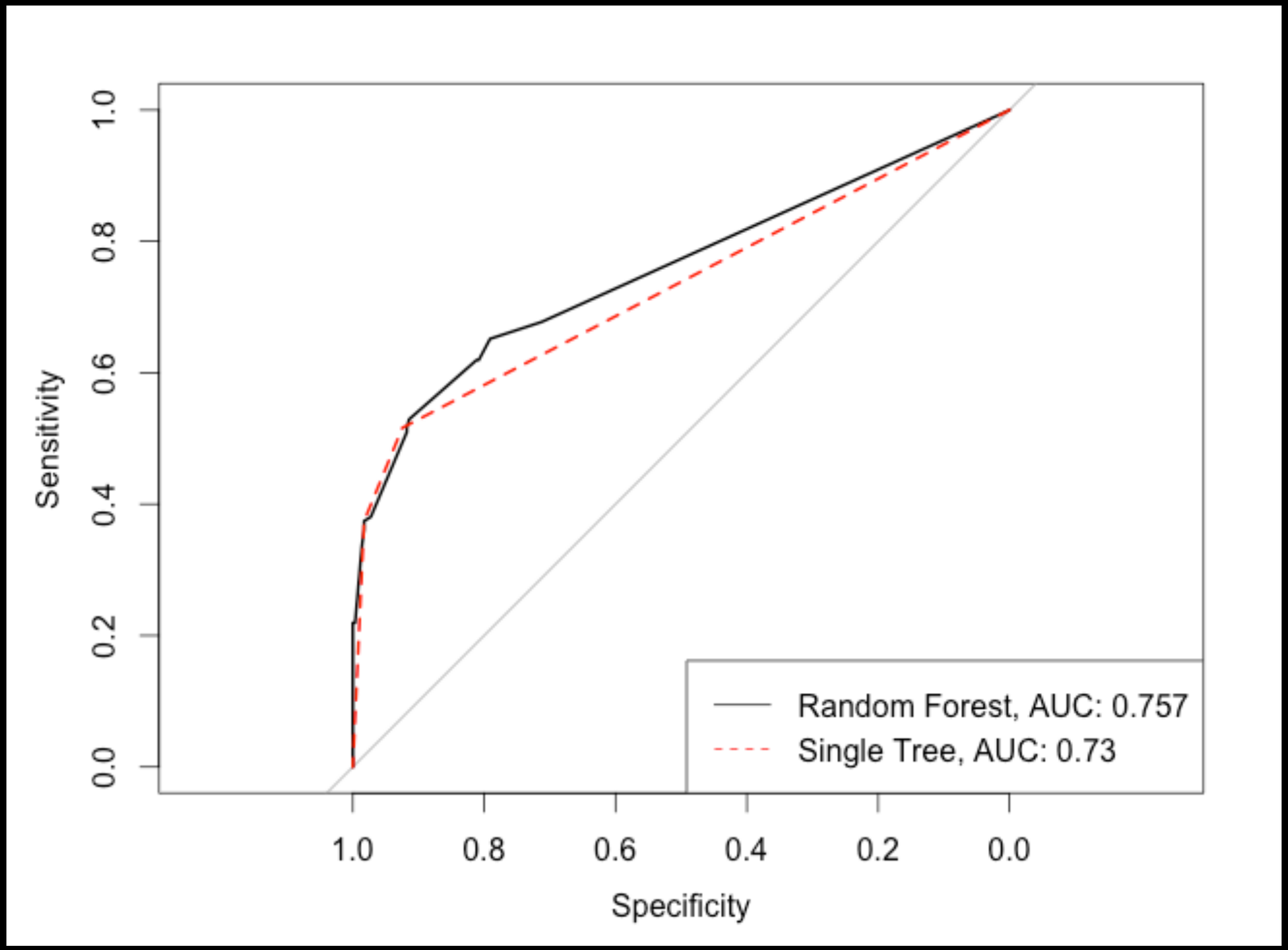
		prediction outcome		
		p	***n***	**total**
actual value	***p'***	True Positive	False Negative	P'
	n'	False Positive	True Negative	N'
	total	P	N	

- TP = True Positives
- TN = True Negatives
- FP = False Positives
- FN = False Negatives

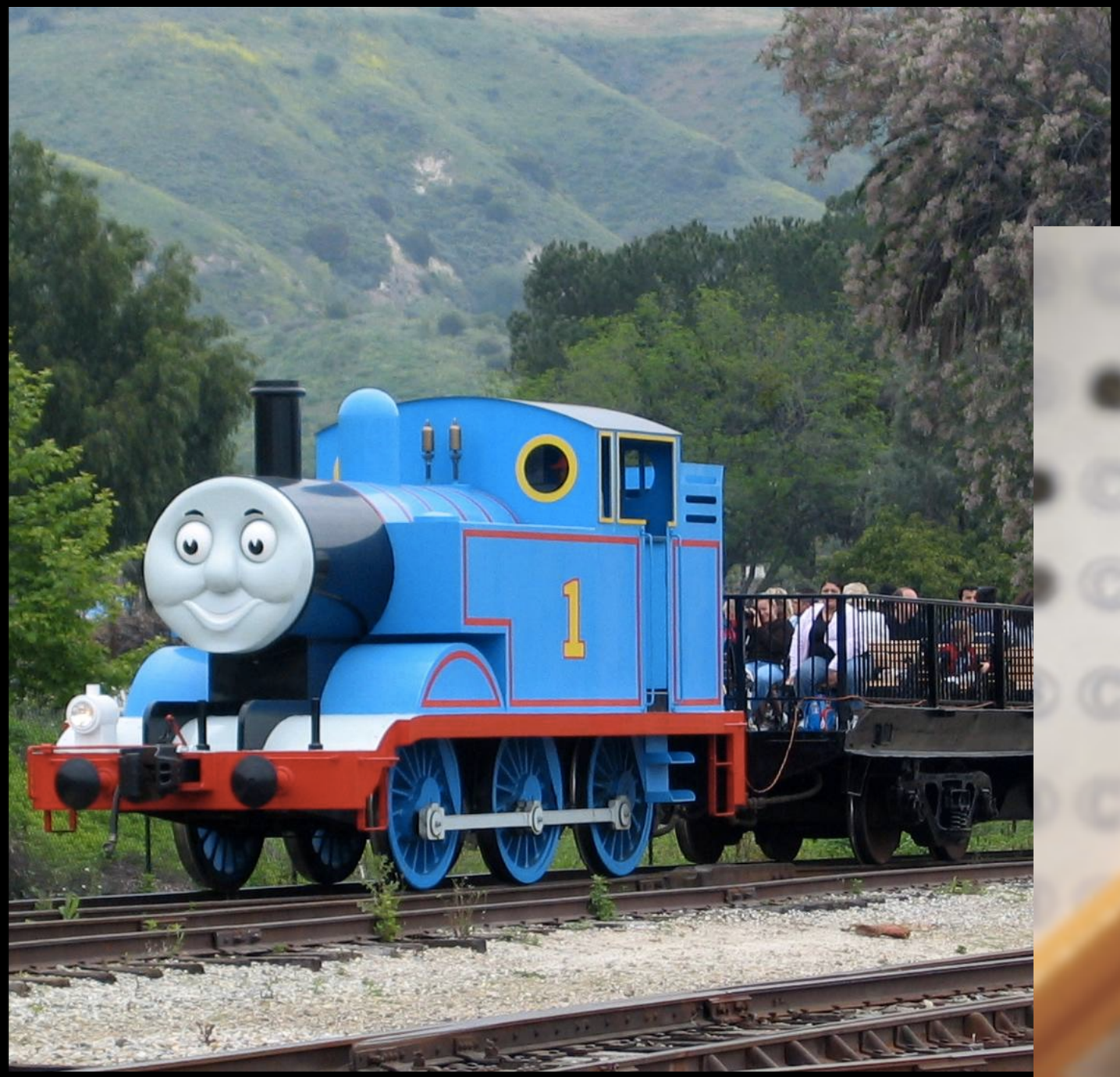
Classifier Evaluation II: Performance Metrics

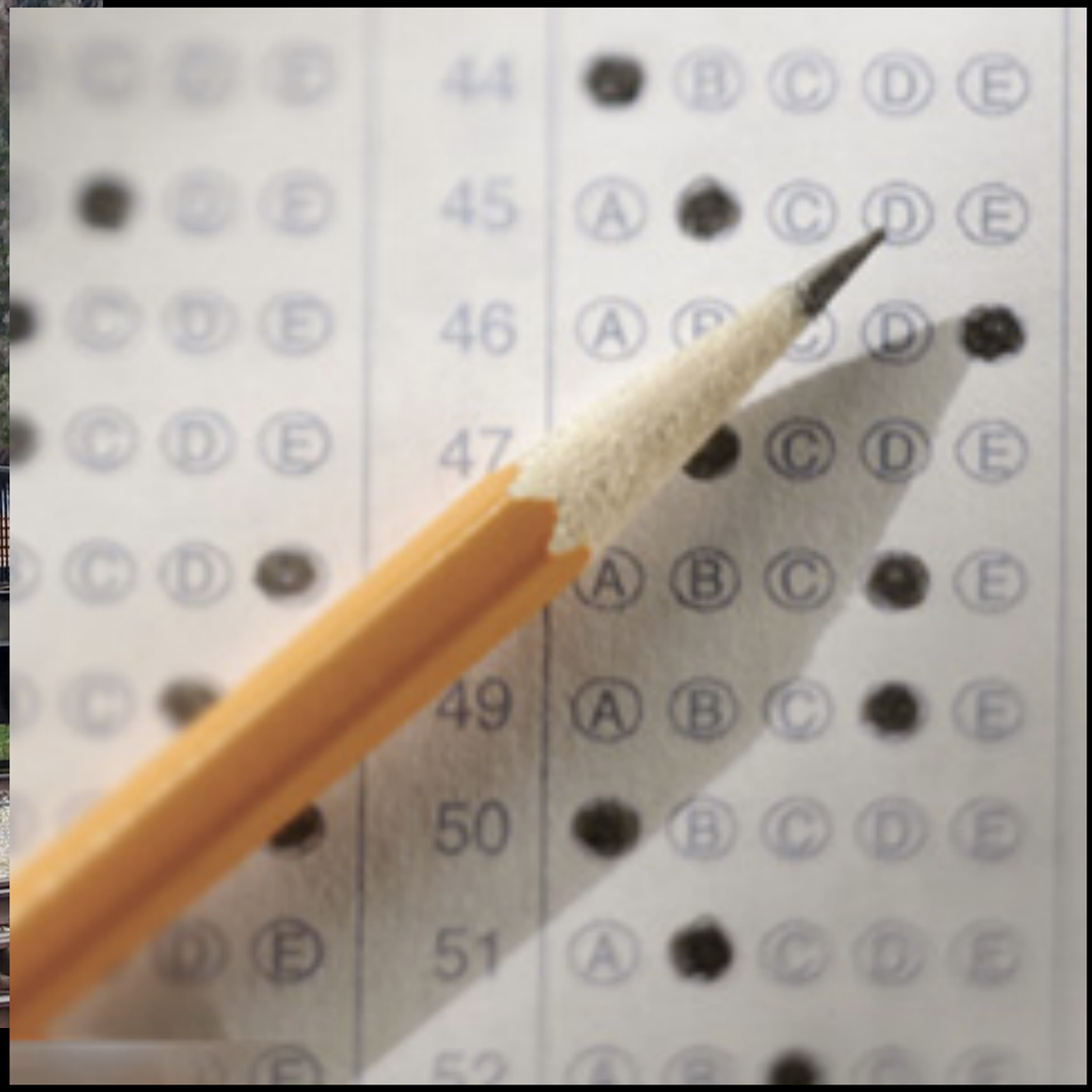
- Precision = TP / (TP + FP)
- Recall/Sensitivity = TP / (TP +FN) = TP / P
- Specificity = TN / (TN + FP) = TN / N
- False Discovery Rate = FP / (FP + TP)

Classifier Evaluation III: The ROC Curve



Training and Testing

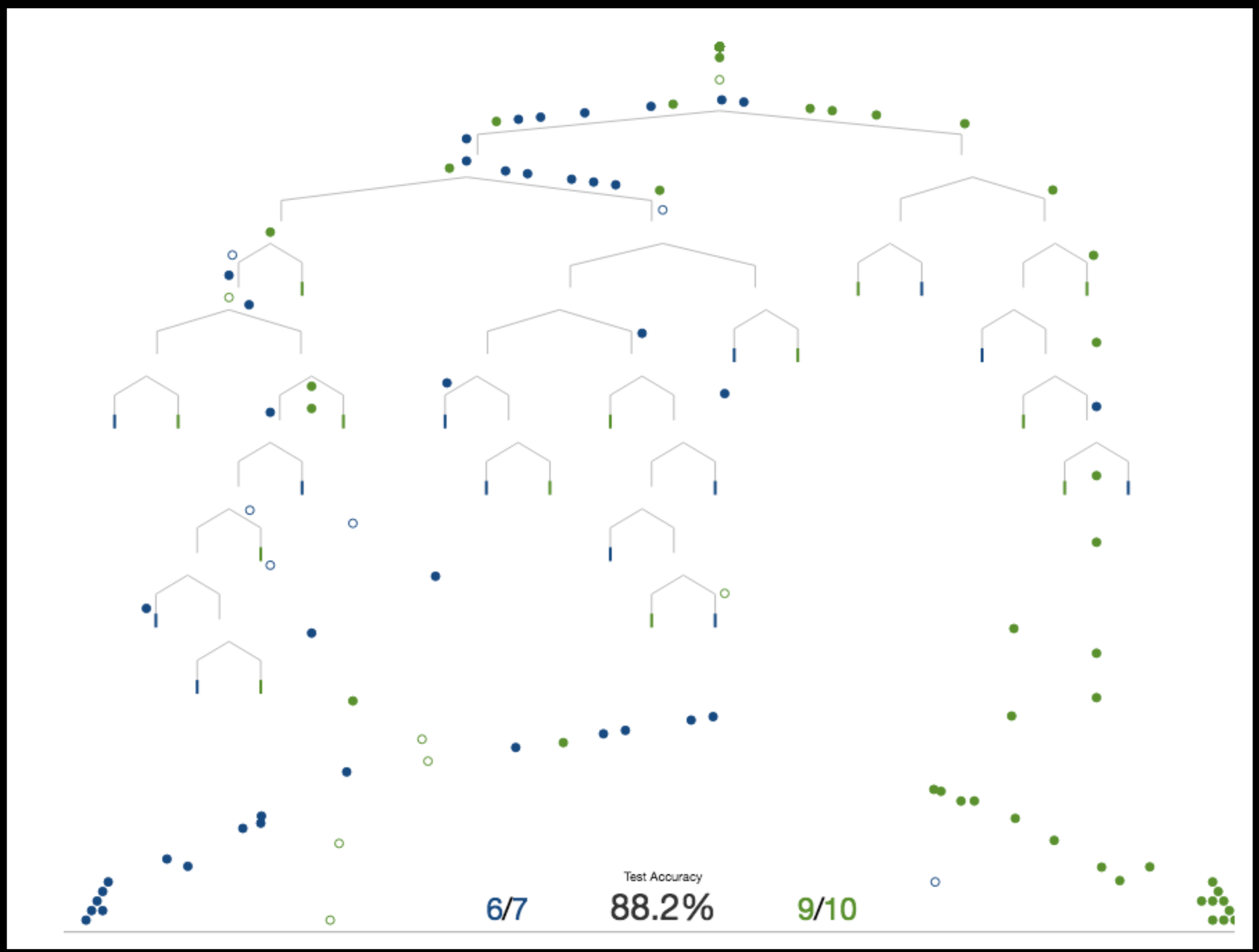




Classifier Evaluation in R

Decision Trees

- Used for:
  - Classification & Numeric Prediction
- Advantages:
  - Common
  - Transparent
- Disadvantages:
  - Prone to over-fitting



http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

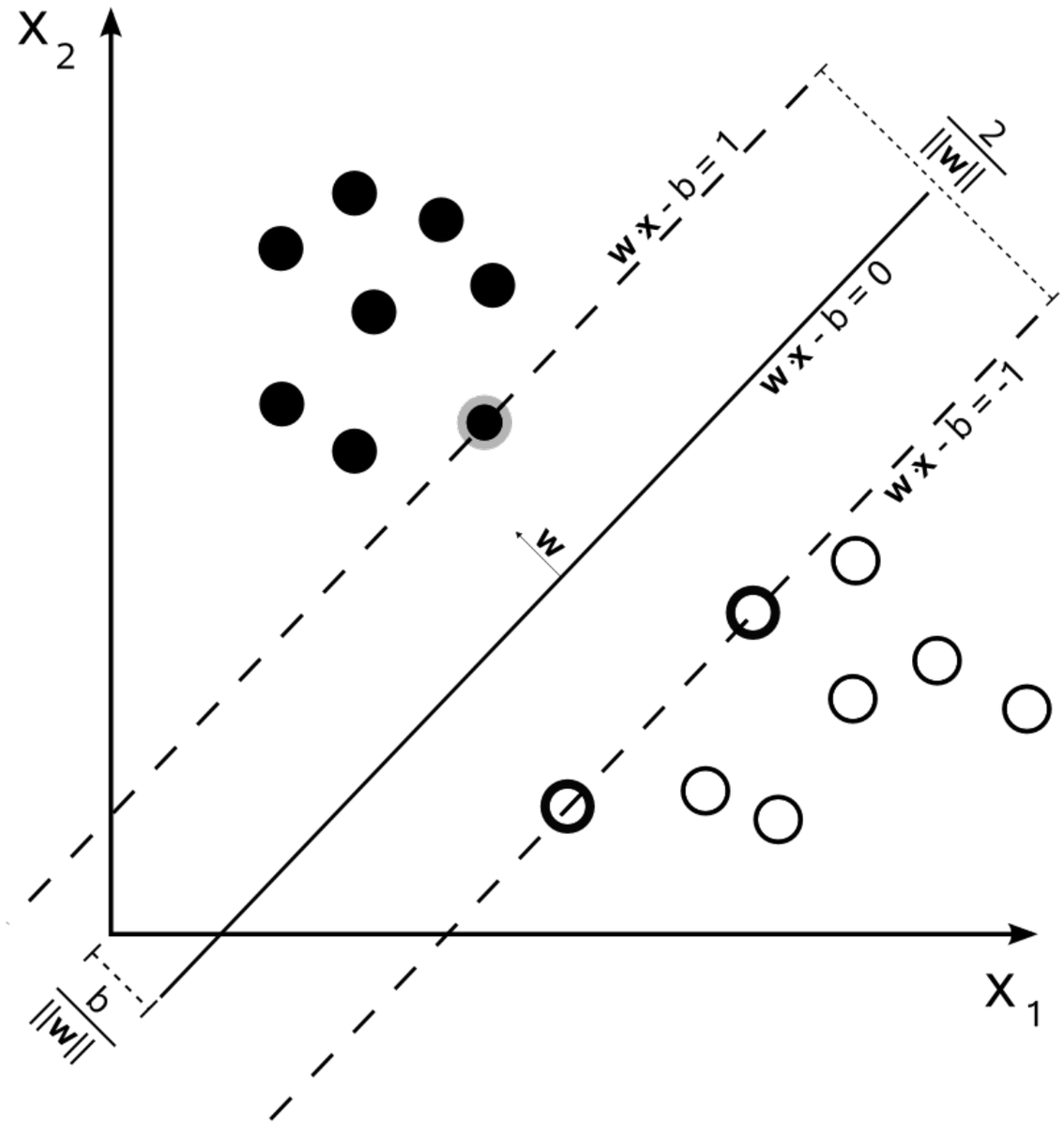
Random Forests

- Used for:
  - Classification & Numeric Prediction
- Advantages:
  - Common
  - Accurate
- Disadvantages:
  - Black box
  - Slowish

Decision Trees & Random Forests in R

SUPPORT VECTOR MACHINES

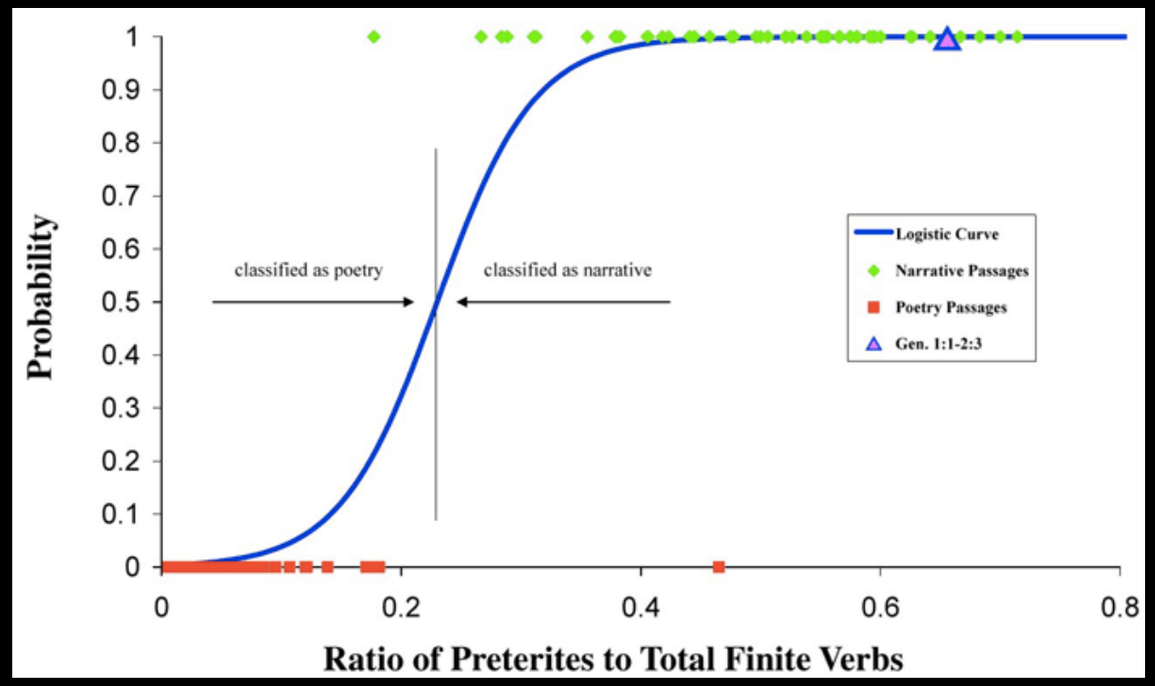
- Used for:
  - Classification & Regression
- Advantages:
  - Accurate
  - Faster than RF
- Disadvantages:
  - Memory hog
  - Black box
  - Class imbalances hurt

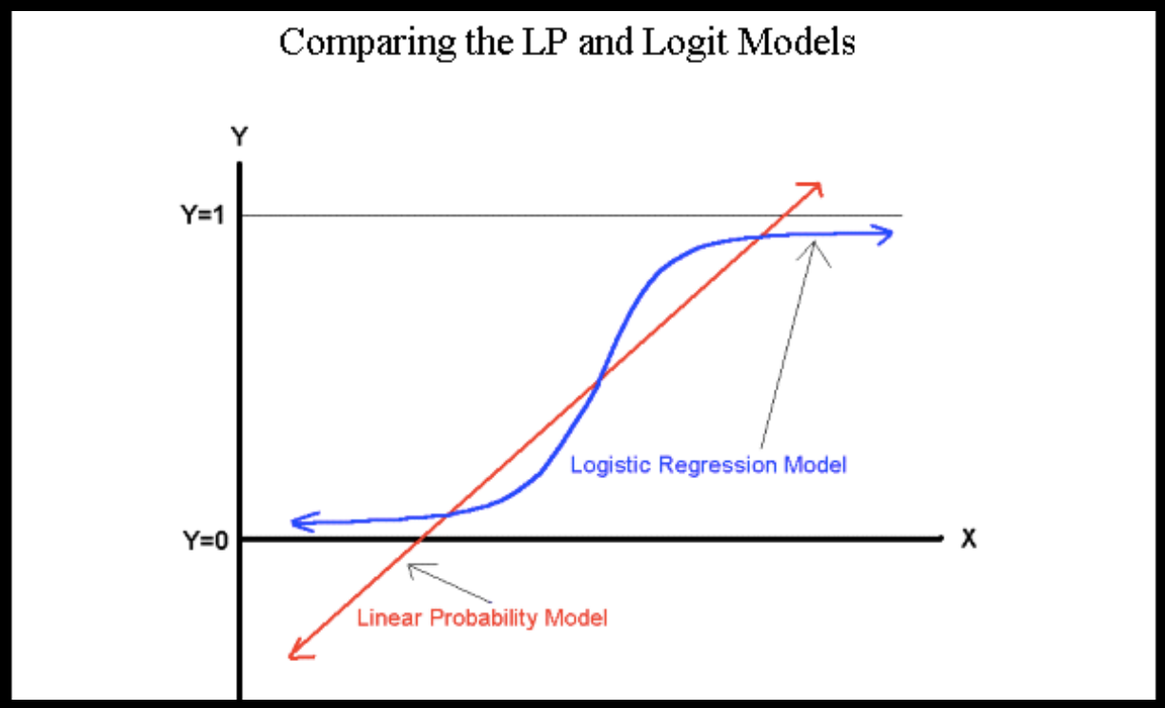


Support Vector Machines in R

Logistic Regression

- Used for:
  - Binary classification
- Advantages:
  - Well understood
  - Transparent
- Disadvantages:
  - Strong assumptions





Logistic Function

$$y = 1 / (1 + e^{-(\alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \varepsilon)})$$

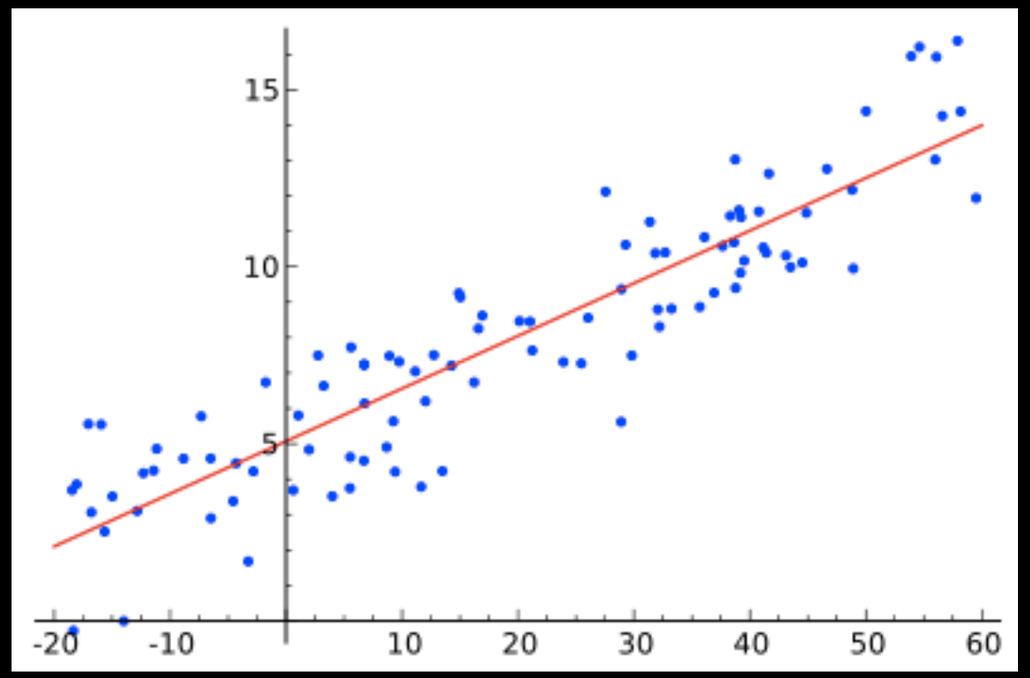
Logistic Regression in R

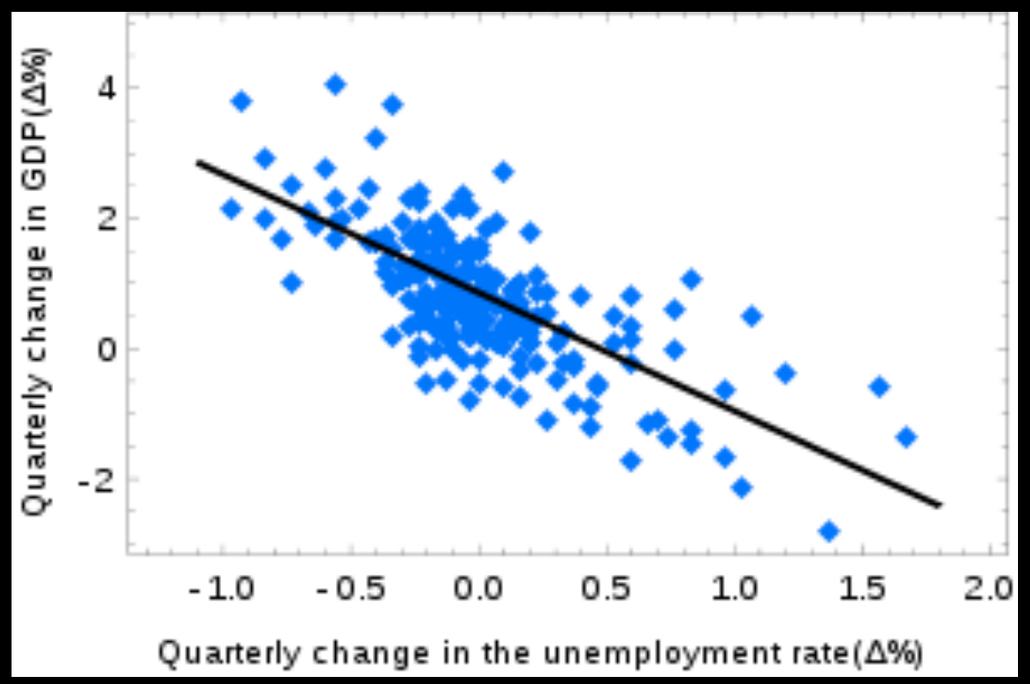
Agenda

- A Motivating Example
- Overview of Supervised Machine Learning
- Your Project: R, R Studio, R Studio Projects
- Classification
- **Numeric Prediction**
- Q&A or Excursions

Linear Regression

- Used for:
  - Linear & semi-linear models
- Advantages:
  - Transparent
  - Fast
  - WELL understood
- Disadvantages:
  - Restrictive assumptions





Remember…

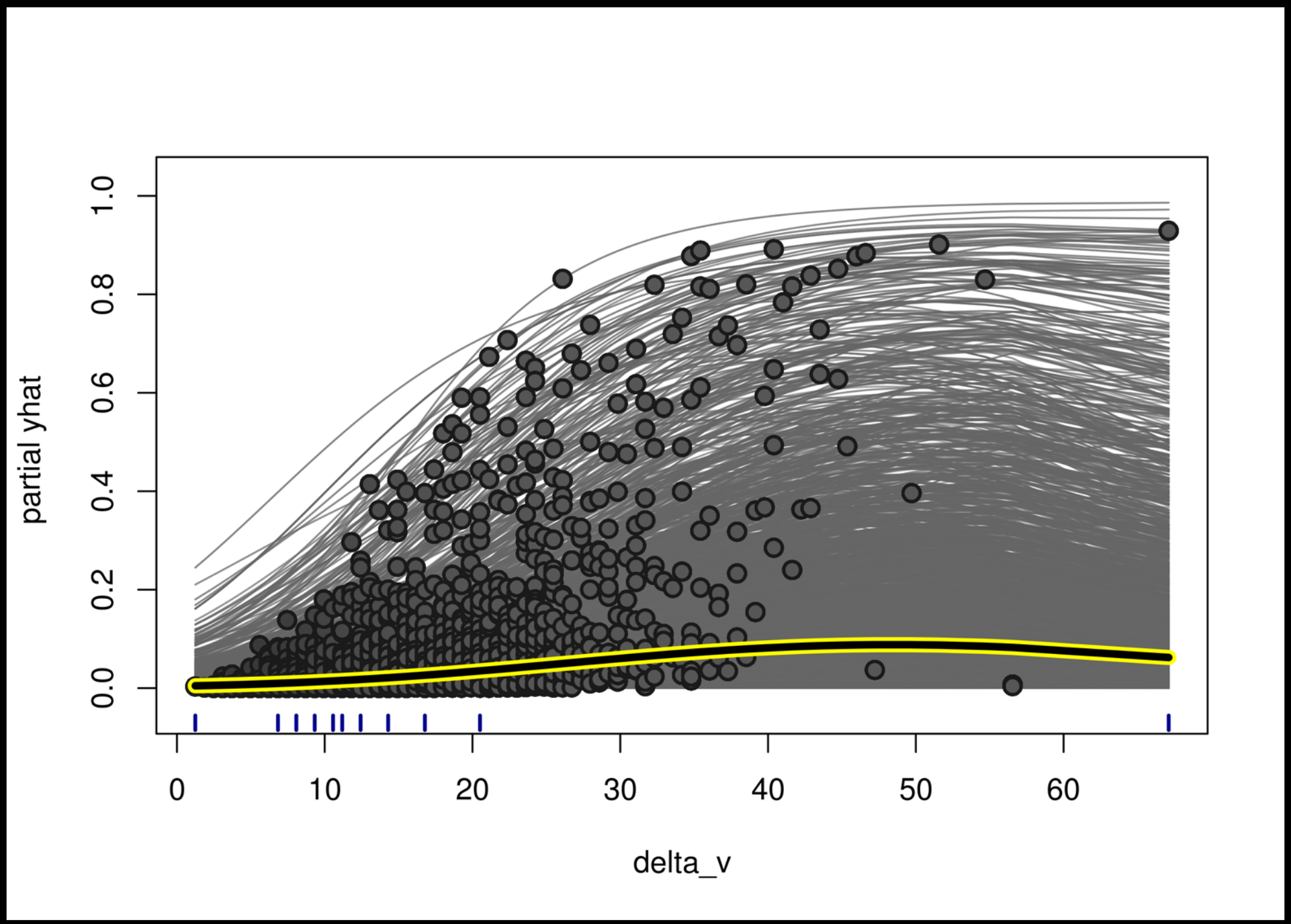
$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \varepsilon$$

Agenda

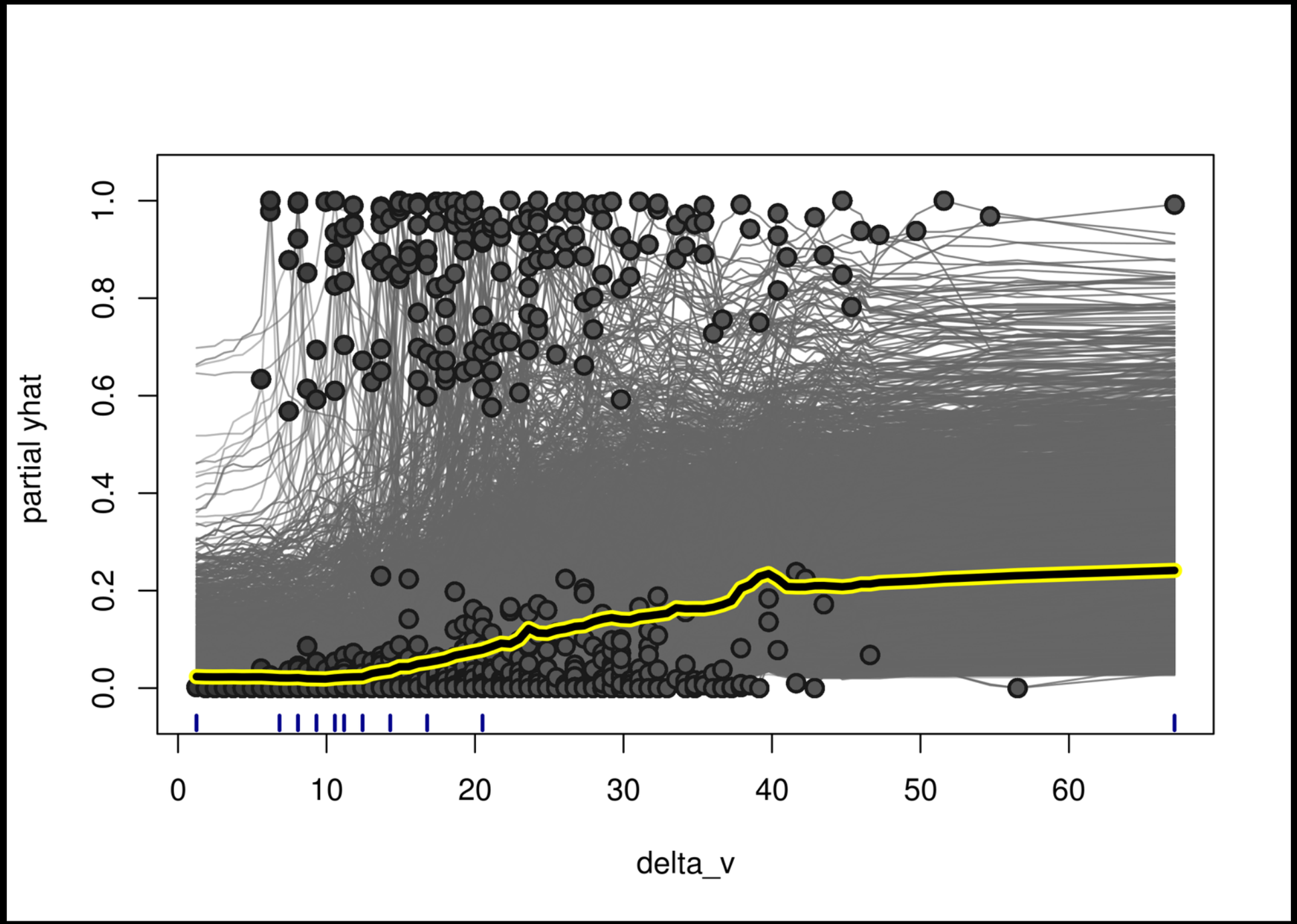
- A Motivating Example
- Overview of Supervised Machine Learning
- Your Project: R, R Studio, R Studio Projects
- Classification
- Numeric Prediction
- **Q&A or Excursions**

Questions?

Parametric ICE Curve



Non-Parametric ICE Curve



Contact

- jones.thos.w@gmail.com
- tjones@impactresearchinc.com
- http://www.impactresearchinc.com
- twitter: @thos_jones
- http://www.biasedestimates.com
- https://GitHub.com/TommyJones