

# About the Course

Predictive Modeling & Statistical Learning

Gaston Sanchez

CC BY-SA 4.0

Official Title

# Stat 154: Modern Statistical Prediction and Machine Learning

I prefer something like ...

An Introduction to  
Predictive Modeling  
and Statistical Learning

What is this course about?

# Machine Learning (ML)?

**Machine learning** is the subfield of computer science that, according to Arthur Samuel in 1959, gives "computers the ability to learn without being explicitly programmed."

Machine learning - Wikipedia

[https://en.wikipedia.org/wiki/Machine\\_learning](https://en.wikipedia.org/wiki/Machine_learning)

# Not this type of ML

**Machine learning** is the subfield of computer science that, according to Arthur Samuel in 1959, gives "computers the ability to learn without being explicitly programmed."

**Machine learning - Wikipedia**

[https://en.wikipedia.org/wiki/Machine\\_learning](https://en.wikipedia.org/wiki/Machine_learning)

## Simply put

- ▶ focus on Predictive Models ( $\approx 85\%$ )
- ▶ with a pinch of Descriptive Methods ( $\approx 15\%$ )
- ▶ from a Statistical Learning standpoint

# How I think of Statistical Learning

*Data analysis and model-building techniques from cross-pollination between Statistics, Applied Math, and Computer Science, with contributions and applications from all scientific corners (Life sciences + Social sciences + other)*



# Statistical Learning

2 main branches:

Supervised    -vs-    Unsupervised

---

## Statistics

## Machine Learning

---

Predictive  
methods

Supervised  
learning

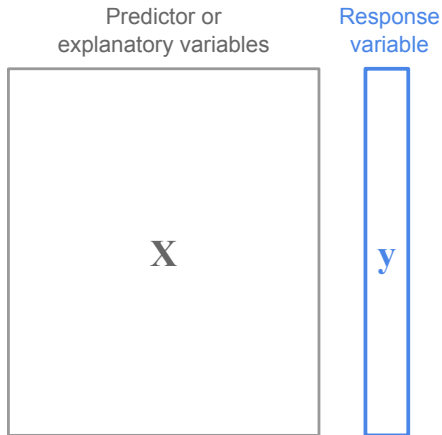
---

Descriptive  
methods

Unsupervised  
learning

---

# Predictive Methods (Supervised)



# Supervised Methods

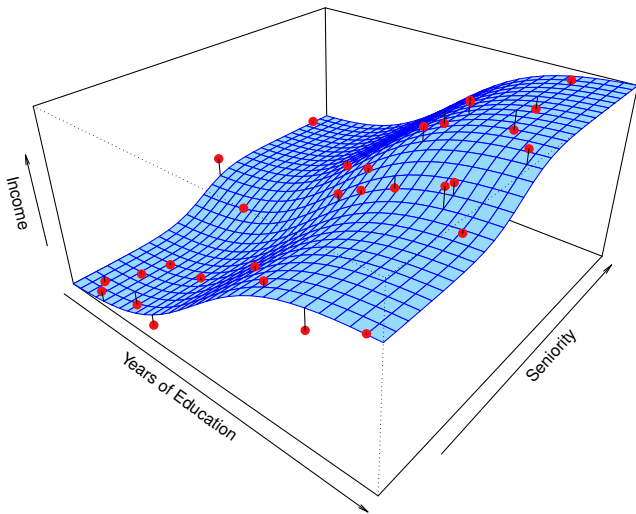
## Two flavors

- ▶ **Regression:** quantitative target variable
- ▶ **Classification:** qualitative target variable

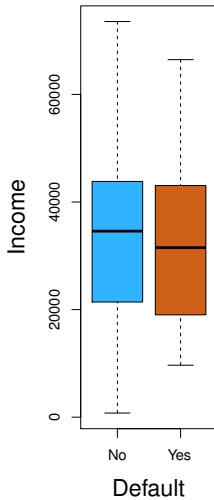
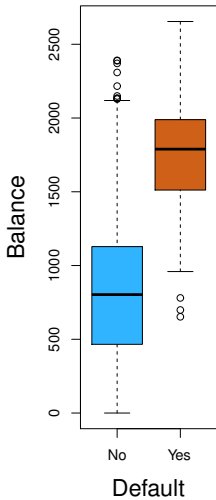
## Supervised or Predictive

Build models and procedures for regression and classification tasks, and assess the predictive accuracy of those models and procedures when applied to new data.

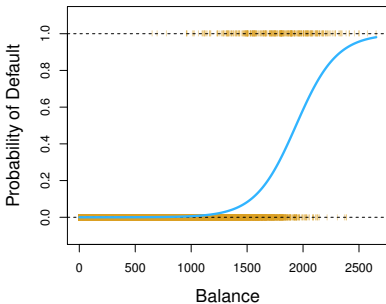
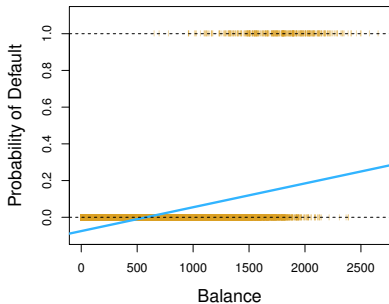
# Regression Example



# Classification



# Classification



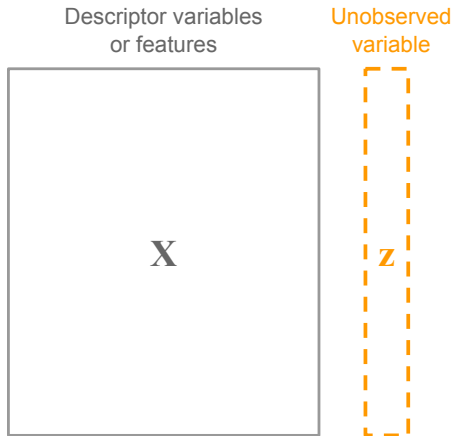
# Supervised Learning

Problems in which the learning algorithm receives a set of continuous or categorical input variables and a correct output variable (which is observed or provided by an explicit “teacher”) and tries to find a function of the input variables to approximate the known output variable: a continuous output variable yields a regression problem, whereas a categorical output variable yields a classification problem.

*Izenman, 2008*



# Descriptive Methods (Unsupervised)



# Unsupervised Methods

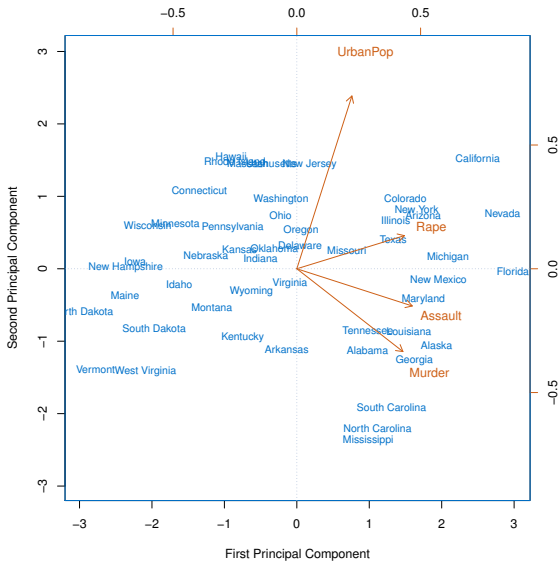
## Descriptive Methods

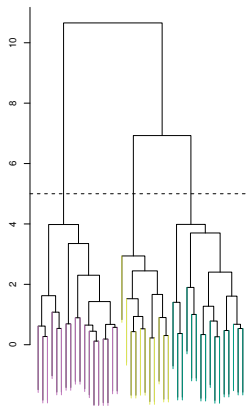
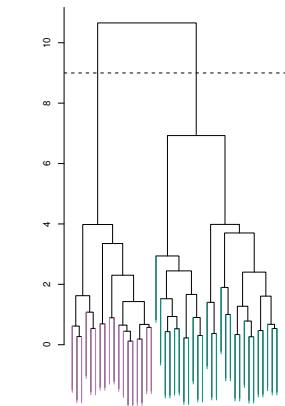
- ▶ **Ordering/Scaling:** finding systematic patterns of variation
- ▶ **Clustering:** finding groups in data

## Unsupervised or Descriptive

Search data sets and discover unexpected structures or relationships, patterns, trends, clusters, and outliers in the data.

# Ordination/Scaling (PCA)



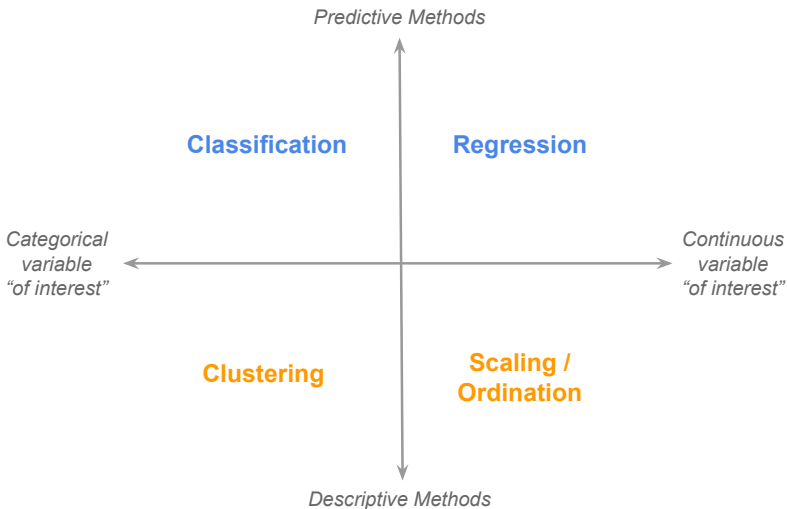


# Unsupervised Learning

Problems in which there is no information available (i.e. no explicit “teacher”) to define an appropriate output variable.

*Izenman, 2008*

# Predictive and Descriptive Methods



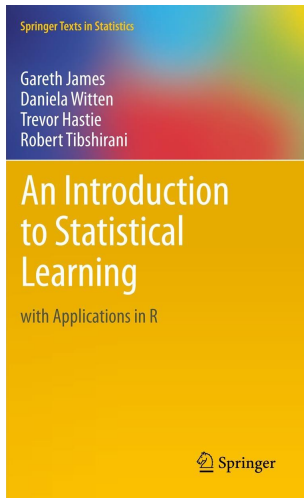
# A word of caution

Sometimes there might not be a clear distinction between supervised and unsupervised learning. Often, a given method mixes both types of approaches.

# Course Topics

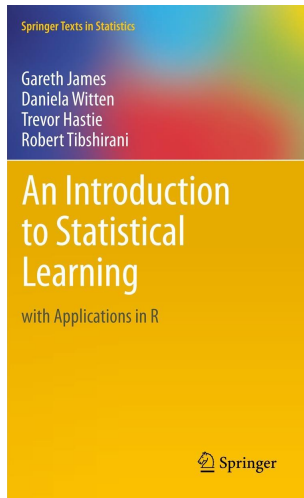


# Primary Textbooks

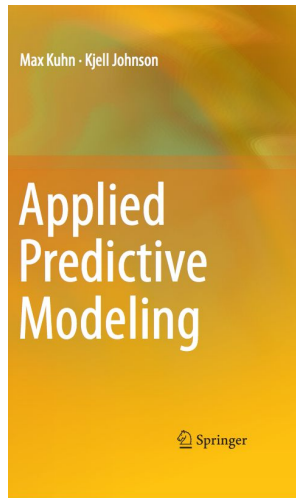


ISL

# Primary Textbooks



ISL



APM

# Course in a nutshell

Not necessarily in this order

- ▶ Matrix algebra housekeeping
- ▶ Data Preprocessing
- ▶ Principal Components Analysis
- ▶ Regression
  - Linear (and related) Models
  - Regression Trees and extensions
- ▶ Classification
  - Linear (and related) Models
  - Classification Trees and extensions
- ▶ Process of predictive model building
- ▶ Clustering

# Remarks

- ▶ Cover the theory and concepts of each method
- ▶ How-and-Why does a method work?
- ▶ When does a method (may) fail?
- ▶ Learn how to interpret the outputs
- ▶ **Coding**: write code/scripts/functions for most methods
- ▶ Get to know packaged functionalities

# Github repo

- ▶ username: **ucb-stat154**
- ▶ repository: **stat154-fall-2017**

<https://github.com/ucb-stat154/stat154-fall-2017>

I'll be uploading/updating the repo's content as we move on with the course

# Prereqs

# Prereqs

- ▶ Math 53: multivariate calculus
- ▶ Math 54: linear algebra
- ▶ Stat 134: probability
- ▶ Stat 133: computing with data

# Two Assumptions

I'm assuming 2 things about you:

Matrix Algebra    &    R basics



# Matrix Algebra

You should have been exposed to concepts such as:

- ▶ Vector Spaces
- ▶ Inner Products
- ▶ Matrix Multiplication
- ▶ Linear Dependency
- ▶ Rank
- ▶ Trace, Determinant
- ▶ Inverse
- ▶ *etc*

# R Basics

You should have been exposed to:

- ▶ R vector's, list's, data.frame's
- ▶ Subscripting and indexing (i.e. bracket notation)
- ▶ Writing functions: `function() {...}`
- ▶ Conditionals: `if {...} else {...}`
- ▶ Loops: `for`, `while`, `repeat`
- ▶ Graphics: `base`, `ggplot2`, etc
- ▶ RStudio familiarity

# Matrix Algebra

You should have been exposed to concepts such as:

- ▶ Vector Spaces
- ▶ Inner Products
- ▶ Matrix Multiplication
- ▶ Linear Dependency
- ▶ Rank
- ▶ Trace, Determinant
- ▶ Inverse
- ▶ *etc*

# Expectations

At the end of the course

- ▶ Understand theory and concepts
- ▶ Being able to interpret results
- ▶ Being able to implement algorithms in R (scripting, programming)
- ▶ Implement Full pipeline (with prepacked tools)
- ▶ Move on to more specialized techniques

# Please read the course policies



<https://github.com/ucb-stat154/stat154-fall-2017/blob/master/syllabus/policies.md>