+

Lecture 1 – Introduction

IEOR 142 – Introduction to Machine Learning and Data Analytics

Fall 2018 – Paul Grigas

Today's Agenda

- Motivation and Objectives
- Google's Search Engine
- Course Information
- Introduction to R
- Introduction to Statistical Learning

Importance of Analytics

- Tremendous growth in the amount and variety of data, and in computational power
- Organizations are becoming much more data-driven
  - Number of Chief Data Officers in companies doubled from 2012 to 2014
- Study of 330 public North American companies:
  - Companies in top third of industry in data-driven decision making were **5% more productive and 6% more profitable** than competitors

What is Data Analytics?

- Using **data** to build **models** that lead to better **decisions**
- … and ultimately add **value**
- Analytics is both "big data" and "small data"
- Hopefully you have seen some examples before
  - We will take a deep dive into **methodology, applications, and practical skills**

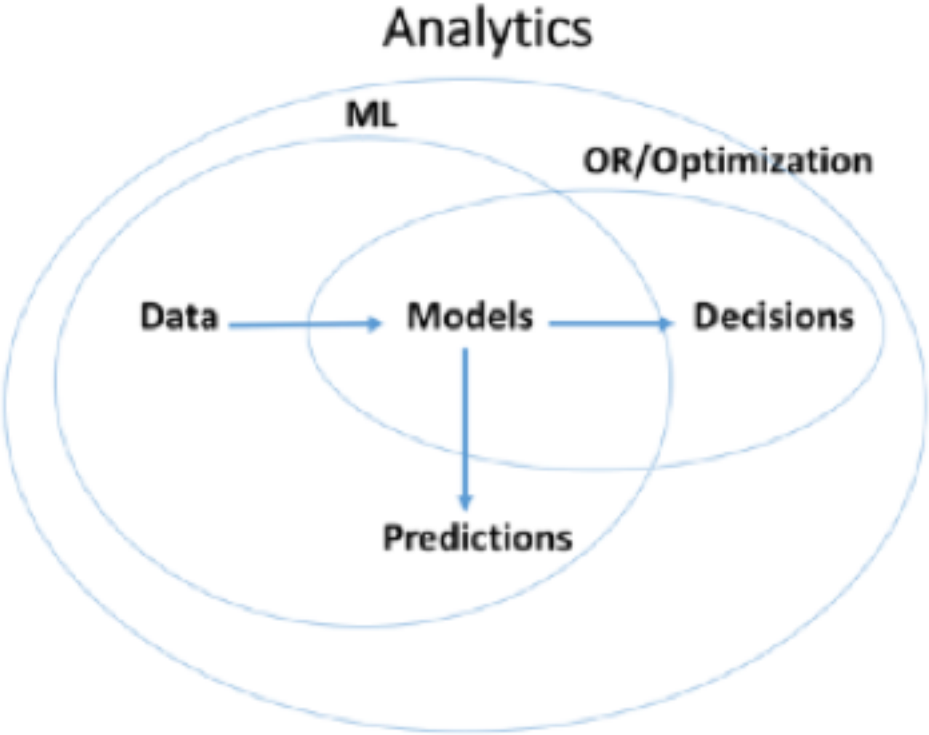
What is Data Analytics?

- Descriptive – finding patterns in data
  - Summary statistics
  - Visualization
  - Clustering, etc.
- Predictive – predicting outcomes
  - Linear Regression
  - Logistic regression, linear discriminant analysis
  - CART, Random Forests, Boosting, etc.
- Prescriptive – making decisions
  - Data-driven optimization

What is Statistical Machine Learning?

- A synonym for data analytics?
- A vast set of tools for **understanding data**
- Data analytics places more emphasis on **decisions**
  - Not a 100% fair statement
- Distinction seems to be more historical than anything else
  - Lines are increasingly blurred in recent years

Analytics, ML, and Optimization – One Perspective



(Image from Dimitris Bertsimas, Editor-in-Chief of the *INFORMS Journal on Optimization.*)

Key Themes

- Analytics provides a key competitive edge to both individuals and companies
- **Our perspective:**
  - Analytics is not a series of black boxes, but there is no need to reinvent the bicycle
  - (Rather, we will teach you how to ride the bicycle)
  - Teach analytics techniques through real-world examples and real data
  - Problem → data → tool(s) → solutions → decisions

+ Course Objectives

- Teach you how to identify opportunities for creating value through the use of data analytics
- Teach you how to interpret, understand, and use the results of analytical models
- Convey enough about the inner workings of analytics methods so that you can make an **informed** choice about which method to use
- Give you practical experience **applying** analytics methods and **communicating** their results
- Inspire you to leverage analytics in your future career

Analytics and Google's Search Engine

Google's Search Engine

- "I think Google should be like a Swiss Army knife: clean, simple, the tool you want to take everywhere." – Marissa Mayer

Brief History of Google

- 1996 – Sergei Brin and Larry Page, then graduate students at Stanford, started Google
- 2001 – Eric E. Schmidt, then CEO of Novell, became CEO of Google
  - Many new products launched since then
- 2004 – Google IPO: 19,605,052 shares at $85 per share
- Alphabet's market cap yesterday: $841.15B
  - Behind only Amazon and Apple

How Did Search Engines Work in the Early 1990s?

- Simplest model:
  - User searches for "analytics"
  - Rank webpages according to the number of times they use the term "analytics"
  - What can go wrong?

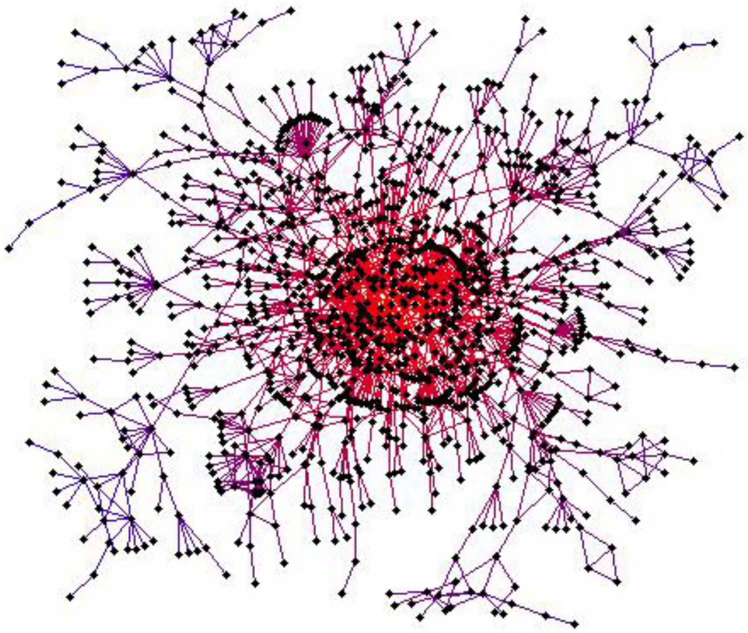
- Somewhat improved model (?):
  - Let *T = 1* if "analytics" appears in the title of a given webpage, *T = 0* otherwise
  - Rank webpages according to score = *50*T + # of times "analytics" appears in the page*
  - Does this help?

How Does Google's Search Engine Work?

- First determine/estimate which pages are visited most frequently
  - create a ranked list of all ~30 trillion pages – the "PageRank"
- Response to search query is comprised of those pages with the queried word combinations, ranked by popularity of page
- **Challenge:** how to reliably estimate the popularity of each and every web page?

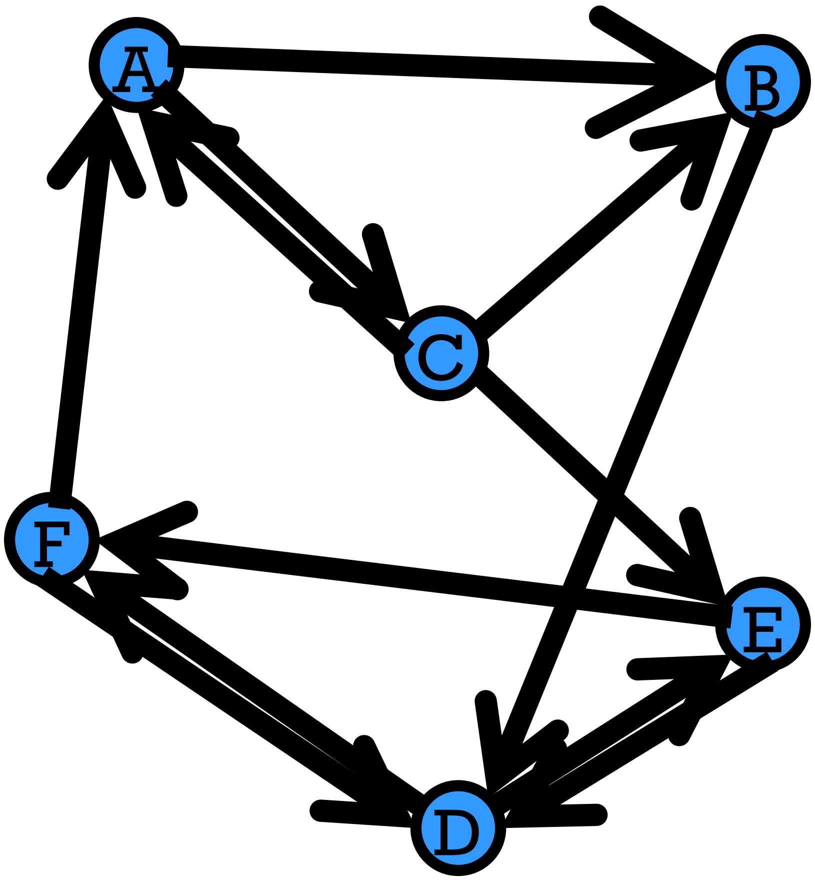
World Wide Web is a Network

- We need a new way of thinking
- View the World Wide Web as a **network**
  - **Data:** webpages and who they link to



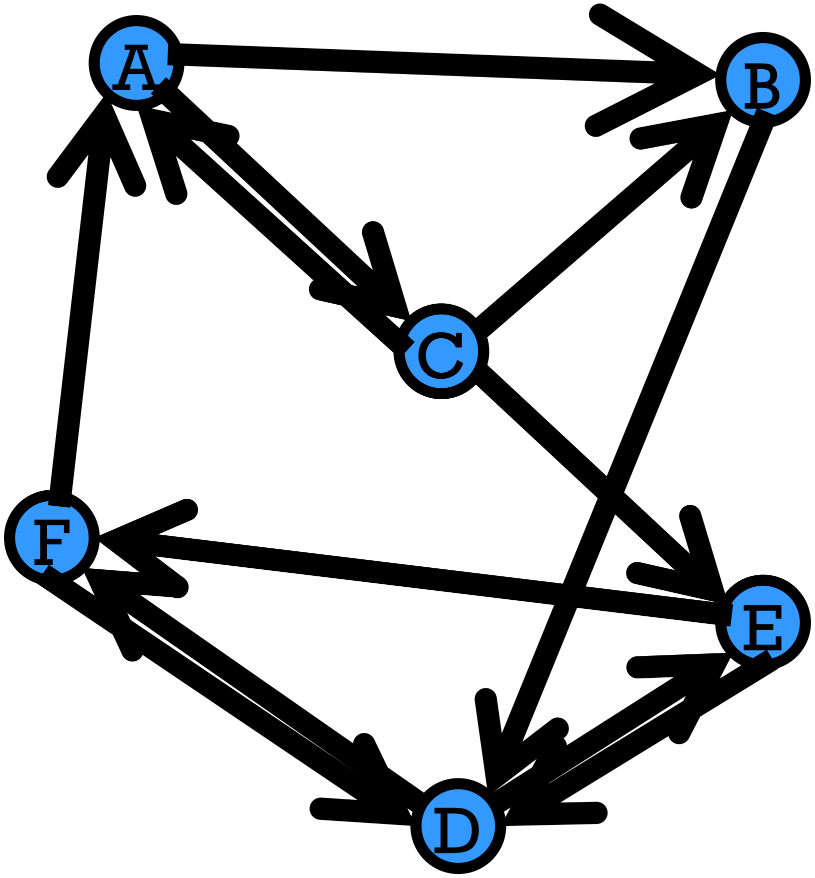
World Wide Web is a Network

- Vertices are pages
- Arcs are hyperlinks



Web User Model

- Model: users go from page to page by randomly clicking on a hyperlink on the current page
- If I start at page C, I go to A, B, or E each with probability 1/3



Web User Model

- Think of a hyperlink as a recommendation: a hyperlink from my homepage to yours is my endorsement of your webpage
- A page is more important if it receives many recommendations
- But status of recommender also matters

Web User Model

- An endorsement from Warren Buffett probably does more to strengthen a job application than 20 endorsements from 20 unknown teachers and colleagues
- However, if the interviewer knows that Warren Buffett is very generous with praise and has written 20,000 very strong recommendations, then the endorsement drops in importance

Slightly More Sophisticated Web User Model

- Model: users go from page to page by randomly clicking on a hyperlink on the current page
- Slightly more sophisticated model:
  - First flip a weighted coin with *P(heads) = 0.85*
  - If heads, now randomly click on a hyperlink from the current page
  - If tails, randomly go to any other page on the web

Assessing the Popularity of a Webpage

- Imagine a user who keeps surfing the web according to the previous model, *ad infinitum*
- The popularity of a given webpage is the proportion of time she spends on that page
  - (The probability of that webpage in the stationary distribution of the Markov Chain)
- Method is "easy" to compute
  - Classical approaches from Markov Chain theory
  - ~30 trillion webpages creates a challenge from a computational engineering point of view

The PageRank Algorithm

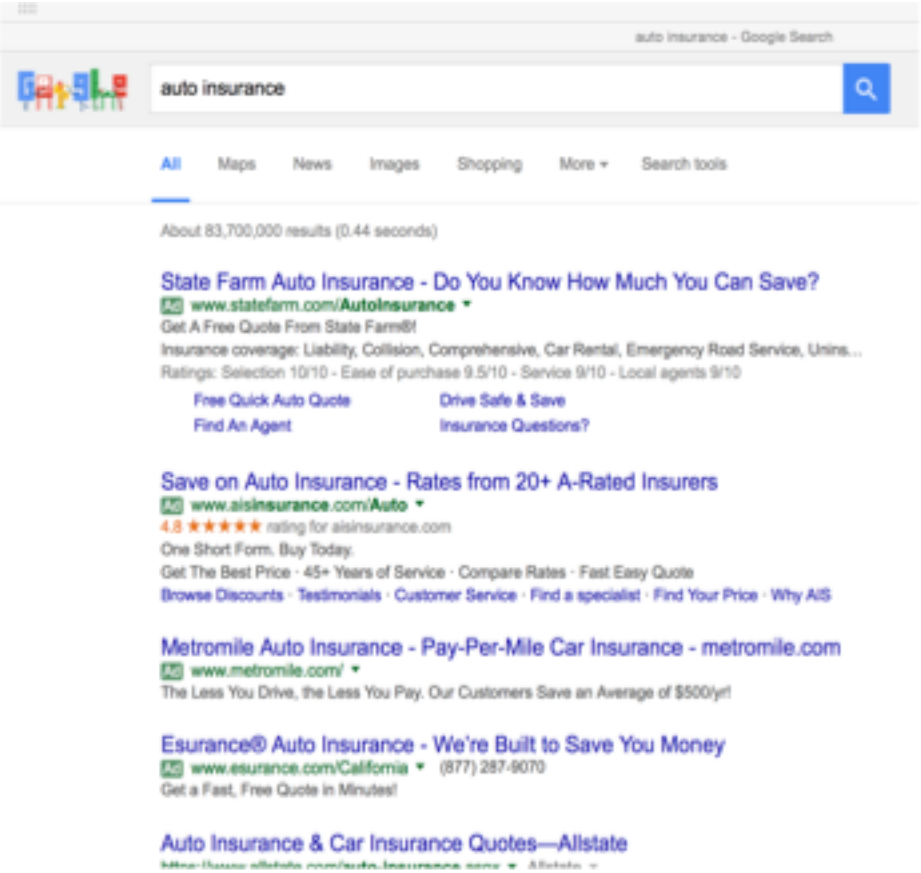
- Brin and Page proposed the idea of ranking pages using this method
- As always, in practice there are other issues to deal with:
  - Web users can spend lots of time or little time on a page
  - Not all users click on all hyperlinks with equal likelihood
  - Other issues
- However, this method forms the heart of Google's Search Engine

The Edge of Simplicity

- Simple but powerful use of "basic" Markov Chain theory
- Brilliantly implemented and led to the rise of Google
- Would there be a search business without the mathematics?
- Without a good search engine, how valuable would the web be?

What about Ads?

- Sponsored search advertising is Google's main source of revenue
- Ads are ordered by *bid * predicted click-through rate * quality score*
- How does Google predict the click-through rate (CTR)?



About 83,700,000 results (0.44 seconds)

State Farm Auto Insurance - Do You Know How Much You Can Save?
Ad www.statefarm.com/AutoInsurance ▾
Get A Free Quote From State Farm®!
Insurance coverage: Liability, Collision, Comprehensive, Car Rental, Emergency Road Service, Unins...
Ratings: Selection 10/10 - Ease of purchase 9.5/10 - Service 9/10 - Local agents 9/10
Free Quick Auto Quote
Drive Safe & Save
Find An Agent
Insurance Questions?

Save on Auto Insurance - Rates from 20+ A-Rated Insurers
Ad www.aisinsurance.com/Auto ▾
4.8 ★★★★★ rating for aisinsurance.com
One Short Form. Buy Today.
Get The Best Price · 45+ Years of Service · Compare Rates · Fast Easy Quote
Browse Discounts · Testimonials · Customer Service · Find a specialist · Find Your Price · Why AIS

Metromile Auto Insurance - Pay-Per-Mile Car Insurance - metromile.com
Ad www.metromile.com/ ▾
The Less You Drive, the Less You Pay. Our Customers Save an Average of $500/yr!

Esurance® Auto Insurance - We're Built to Save You Money
Ad www.esurance.com/California ▾ (877) 287-9070
Get a Fast, Free Quote in Minutes!

Auto Insurance & Car Insurance Quotes—Allstate
https://www.allstate.com/auto-insurance.aspx ▾ Allstate ▾

+ Analytics for AdWords Auctions

- Many different sources of data for this prediction problem:
  - What is the text of the ad?
  - What is the demographic information of the user?
  - What is the historical CTR of this ad?
  - If the ad is new, what is the historical CTR of other related ads?
- We will take a deep dive into internet advertising later in the course
  - Considering both the CTR prediction problem and the decision problems faced by Google and advertisers alike
- We will also examine methods for incorporating text data and temporal data into predictions

Bringing it All Together

- In this course, you will learn about a wide range of data analytics techniques
- Many times, several techniques need to be integrated together to effectively address a business problem
- We will illustrate many applications of business analytics so you can see how the tools might be effectively applied

Course Information

Course Information

- **Instructor:** Paul Grigas
  - I am an assistant professor in the IEOR Department, joined in Fall 2016
- **GSIs:** Meng Qi and Nathan Vermeersch
  - PhD/MS students in the IEOR Department
- **Reader:** Andrew Ding
  - MS student in the IEOR Department
- Office hours:
  - Start next week
  - TBD and will be posted on bCourses

Course Information

- **Website(s):** bCourses and Piazza
- **Prerequisites:**
  - Basic fluency in statistics, probability, and ideally optimization
  - Implicit in the above is also a working knowledge of linear algebra and calculus
  - Some programming experience
- **Recommended Readings:**
  - Chapters from *An Introduction to Statistical Learning with Applications in R* by James, Witten, Hastie, and Tibshirani (available online)
  - Excerpts from *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data* by Hadley Wickham and Garrett Grolemund (available online)
- **Supplementary Readings:**
  - Chapters from *The Analytics Edge* by Bertsimas, O'Hair, and Pulleyblank
  - Chapters from *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* by Trevor Hastie, Robert Tibshirani and Jerome Friedman

Course Information

- **Grading:**
  - Homework Assignments: 40%
  - Final Project: 35%
  - Midterm Exam: 25%
- **Lecture:**
  - We will post lecture notes ahead of class
  - Participation is encouraged
- **Discussion:**
  - Optional, but attendance is encouraged
  - The main goal of discussion sections is to demo R code
  - We will post notes and associated R code snippets

Homework Assignments

- Discussion/some collaboration is allowed, but final product must be completed *individually*
- ~5 homework assignments
- Assignments will be due on Gradescope by 6 am of the next day
- Assignments will be a mix of conceptual exercises and data analysis exercises in R
- Late assignments will not be accepted but we will drop the lowest score

+ Midterm

- In class on October 18
- Purpose of the midterm is to gauge your understanding of the material so far
- Conceptual questions similar to the homework and also some conceptual R based questions
  - No need to memorize R commands
- More details as the date approaches

Final Project

- Final project carries a lot of weight and is a crucial part of the course
- Teams of 3-4
- You propose a topic
- We are available for advice and will provide sample resources to help you find data if necessary

Some Example Projects

- Predict code blue events in an ICU, improve response times
- Predict the winning team in the video game DOTA 2, and then optimize hero selection strategy
- Predict which players will participate in the NBA all-star game, based on mid-season performance
- Predict energy demand and then optimize solar panel placement

Final Project Timeline

- **October 26:** submit project proposal
  - We will provide e-mail feedback
- **November 29:** 5-minute project presentations
  - During lecture (plus maybe some extra time before lecture or on a different day)
- **December 14:** four page (+ appendices) final report due

Course Content

- **Methods:** linear regression, logistic regression, linear discriminant analysis, ROC analysis, cross validation, bootstrapping, CART, random forests, boosting, text analytics, time series analysis, clustering, regularization and feature engineering, PCA, neural networks, data manipulation and visualization in R, integration with optimization, others
- **Applications:** predicting wine quality and prices, loan defaults, customer retention, parole violators, click-through rates, sentiment of tweets, sales volume, housing prices. Analyzing social networks, customer characteristics, recommendation systems, internet advertising spending, and more

IEOR 142 vs. CS 189 vs. DS 100 vs. ??

- IEOR 142 emphasizes methodology, some mathematical details, real-world datasets, applications, and data science skills
- CS 189 covers a wide range of machine learning methods
  - As compared to IEOR 142, CS 189 has relatively more weight on mathematical and implementation details and less weight on applications and data science tools
- DS 100 covers key principles and techniques of data science
  - As compared to IEOR 142, DS 100 has relatively less emphasis on ML and more emphasis on DS skills
- IEOR 142 applications are sometimes more of a "business analytics" flavor

+

Introduction to R

Why do we use R?

- Some course goals:
  - Understand the complexity of data and **how to deal with data**
  - Understand, use, and **think critically** about the results of analytics models
  - **Create your own** analytics models
- We can reach these goals effectively by using R

What is R?

- R is a software **environment** for data analysis, statistical computing, and graphics
  - Natural to use, complete data analyses in just a few lines
  - Can create almost any analytics model imaginable
- R is a sophisticated programming language
  - More R packages are consistently being developed
    - Data wrangling and transformation techniques, integration with databases and web scraping, improved ways to communicate results, etc.
  - We will explore some advanced features later in the course
- R is an **ecosystem/community** for Data Science
  - Check out #rstats on Twitter

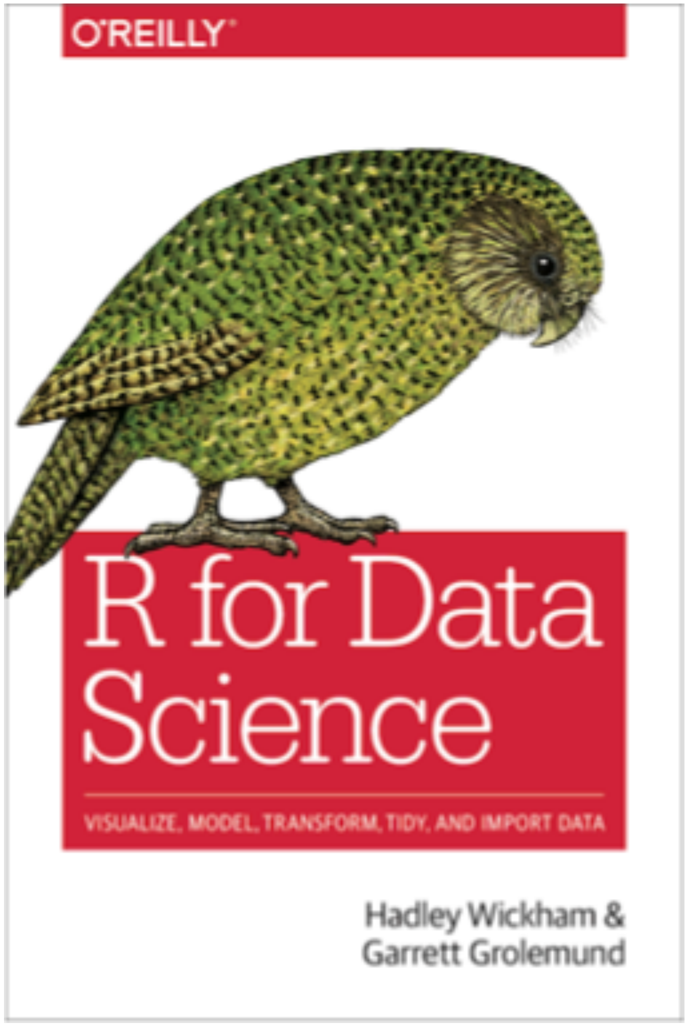
Why use R?

- There are many choices for data analysis software/programming languages
  - Python, SAS, Stata, SPSS, Excel, MATLAB, Minitab, Julia, …
- R is free (open source)
- Easy to re-run previous work and make adjustments
- Excellent graphics and visualizations
- Many companies and organizations that use R
  - Check out: https://www.kaggle.com/surveys/2017

+ R vs. Python?

- "Generally, I think R and python are much more similar than they are different. I'm not really interested in the debates about which one you should learn. Obviously, I think learning R is the right choice, but you can be effective with either. My main advice is to focus on one and get good at it. That's a much more effective way of learning than dabbling in both. (Of course, once you get good in one, you can learn the other, but do it in serial, not parallel)" – Hadley Wickham, Chief Scientist at RStudio

R for Data Science



- Excellent reference on practical tools for how to do data science with R
- We will cover some aspects of this book, but not nearly everything
- Fun and enjoyable to read and follow along with R

What is RStudio?

- RStudio is an integrated development environment (IDE) for R
- RStudio puts everything in one place and will enable us to be more productive and efficient

+ Guidelines About R in this Course

- This is **not** a programming course
  - Focus of this course is about learning data analytics tools and applications
- We will distribute scripts that serve as coding templates for many tasks that you will need to perform
- The project (and likely some assignments) will require you to solve programming problems in R
  - Use Stack Overflow, Google, etc.!
  - All *experienced* programmers search for help online every single day
  - Last resort: ask the teaching team for help (Piazza and office hours are best)

Introduction to Statistical Learning

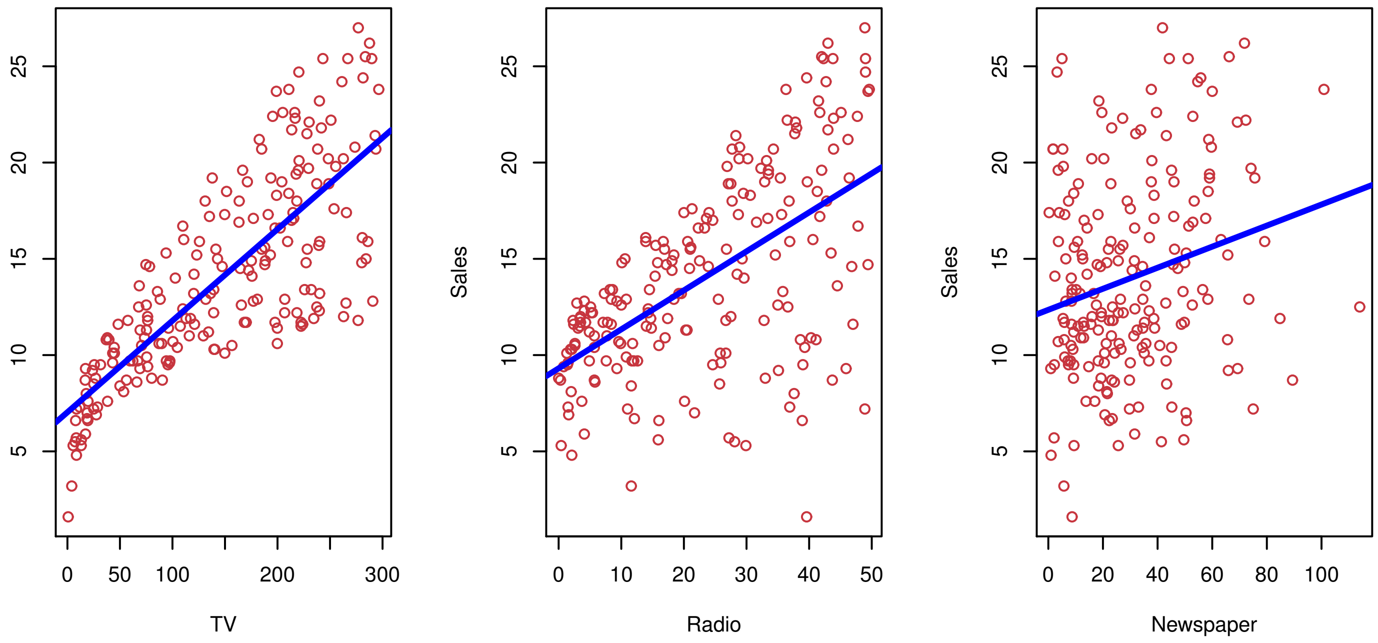
What is Statistical Learning?

- Data analytics is a process
  - From data to decisions
- Statistical learning is a set of **tools** for modeling and understanding complex datasets
  - Like optimization is a set of tools for modeling and solving complex decision problems
- Statistical Learning is inherently a more mathematical subject
  - Data analytics uses mathematics
  - Effective use of data analytics requires fluency in the language of statistical learning

Advertising Budget Planning Example

- Imagine that you are consulting for a company that wants to improve sales of a particular product
- 200 different markets
- For each market, we collected data on:
  - Sales of the product in dollars
  - Advertising budgets for three different types of media: TV, radio, and newspaper
- The company would like to know how advertising budget decisions impact sales

Advertising Budget Planning Example



General Statistical Learning Model

- Input variables: $X = (X_1, X_2, \ldots, X_p)$
  - Also often called features, predictors, or independent variables

- Output variable: Y
  - Also often called response or dependent variable

- Collected data in the form of n pairs:
  - $(x_i, y_i) \quad i = 1, \ldots, n$
  - $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$

A General Statistical Learning Model

- We presume that there is some relationship between X and Y:

$$Y = f(X) + \epsilon$$

- ϵ is a random error term that is **independent** of X and has mean 0
- f is a fixed but **unknown** function that represents the **systematic** information that X provides about Y
- (Supervised) statistical learning is a set of tools for estimating f

+ Why estimate f ?

- Two reasons to estimate f
  - Prediction
  - Inference
- If we have a good estimate for f, call it $\hat{f}$, then we can use $\hat{f}$ to make a prediction for a new value of X

$$\hat{Y} = \hat{f}(X)$$

- Think of the advertising example

Statistical Learning for Prediction

- True model: $Y = f(X) + \epsilon$
- Our prediction: $\hat{Y} = \hat{f}(X)$
- What does the accuracy of our prediction depend on?
  - Reducible error: $\hat{f}$ is not a perfect estimate of f
  - Irreducible error: $\mathrm{Var}(\epsilon) > 0$
- The aim of statistical learning techniques is to reduce the reducible error!

Statistical Learning for Inference

- Inference: how does Y change when X changes?
  - Which predictor variables are associated with the response?
  - What is the relationship between the response and each associated predictor? Positive or negative?
  - Is a linear equation adequate to describe the relationship between X and Y?
- These are all essentially questions about the **behavior** (e.g., slope) of f
- In this course, we will mostly be concerned with prediction, but inference is important!

How do we estimate f?

- As always, we start with data:
  - $(x_i, y_i) \quad i = 1, \ldots, n$
  - Often called the **training data**
- A statistical learning method is a procedure, applied to the training data, for estimating f
  - We'll cover a lot of these methods in this course
- Broadly speaking, two classes of methods:
  - Parametric methods
  - Non-parametric methods

Parametric Methods

- Start by assuming a particular functional form for f
  - For example, assume that f is linear:

$$f(X) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

  - f is parameterized by $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$

- Now apply a method that uses the training data to estimate β
  - We sometimes call this fitting the model
  - Classic example: ordinary least squares, i.e., linear regression
  - We will consider more sophisticated approaches as well

Parametric Methods

- Advantages of Parametric Methods:
  - Simplifies the problem of estimating f to the problem of estimating β
  - Potentially relatively less data needed to produce a reliable estimate of β

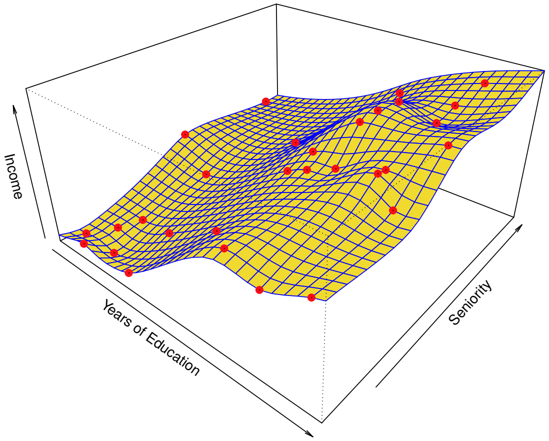
- Major Disadvantage of Parametric Methods:
  - The true functional form of f is usually more complicated than the model we chose
  - This may be remedied by selecting a flexible model class, but this comes at the danger of *overfitting*

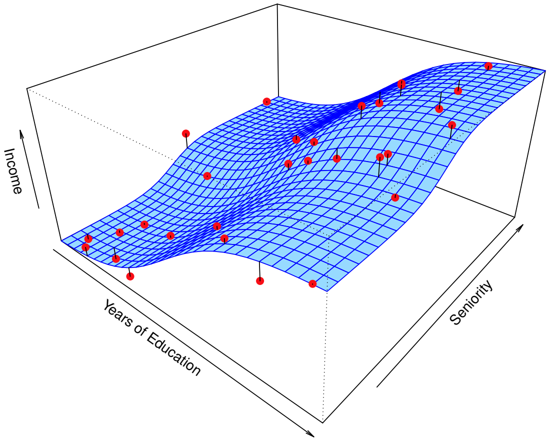
+ Non-parametric Methods

- Of course, non-parametric methods do not make parametric assumptions about f
- No explicit functional form is assumed
  - Allows for greater **flexibility**
  - Runs a greater risk of overfitting if you are not careful
  - Generally requires more data to produce an accurate estimate

Tradeoff Between Flexibility and Interpretability

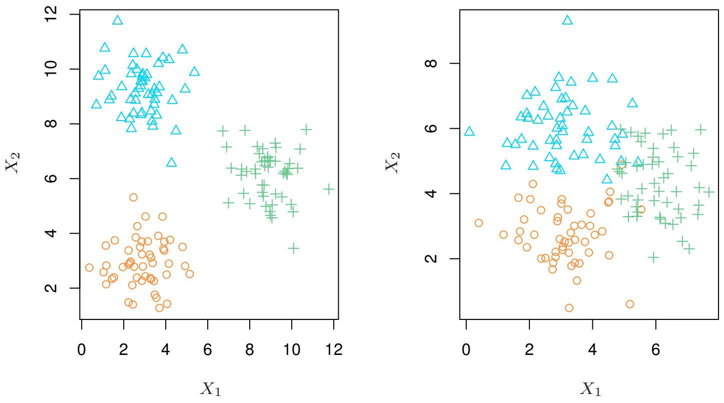
- Why not just always use flexible, non-parametric methods?
  - One reason is that parametric models are more interpretable and thus better for inference
  - Even if you don't care about inference, non-parametric methods may overfit the training data





Supervised vs. Unsupervised Learning

- So far, we have been looking at supervised learning – estimating f
- Unsupervised learning is about finding **structure** or **patterns** in data
  - There is no longer a response variable, we just observe data $X_1, X_2, \ldots, X_n$
  - Examples in this course:
    - Clustering
    - PCA



Regression vs. Classification

- Two important classes of supervised learning problems
- Regression involves predicting a continuous response variable
  - The value of a household
  - Next week: the quality of wine
- Classification involves predicting a binary yes/no outcome
  - Did the user click on the ad or not?
  - Is that email spam?
- Overarching themes but different methods for each

+ Looking Ahead

- In this course, we will explore a plethora of statistical learning methods in the contexts of:
  - Regression
  - Classification
  - Unsupervised learning

+ Next Week

- We will take a deep dive into linear regression
  - Simple linear regression, multiple linear regression, other issues
- We will see how linear regression can be used to "beat the experts" at predicting the quality of wine
  - We will learn how to run and analyze linear regression models in R

- Some of the figures in this presentation are taken from “An Introduction to Statistical Learning, with applications in R” (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani