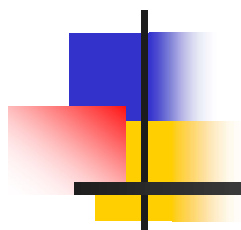




Topic # 01: Data Analytics

Introduction to Topic, Course, & **R**



Instructor: Prof. Arnab Bisi, Ph.D.

Johns Hopkins Carey Business School



Intro to Topic, Course, & **R**

Session 1:

Agenda

- **Part 1:** Topic/Course Introduction
- Break
- **Part 2:** Intro to **R**
- **Part 3:** Working with code

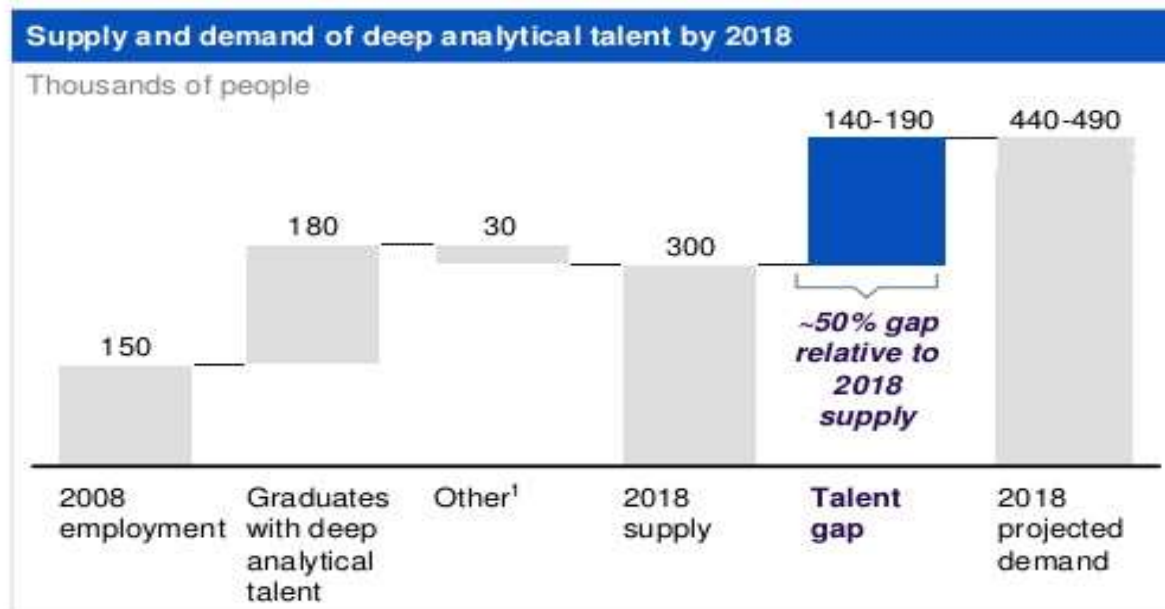


Introduction: Data Analytics

- This is a course about **Data Intelligence**
- We're here to make sure you have the necessary tools for real-world **business analytics**.
- A mix of principles and practice
 - Solid understanding of essential **statistical principles**
 - Concrete analysis ability and best practice guidelines
- What to trust and how to use it

Introduction: Data Analytics

- 2 In the next few years, demand for deep analytical talent in the United States could exceed supply by ~50%



¹ Other supply drivers include attrition (-), immigration (+), and reemploying previously unemployed deep analytical talent (+)

SOURCE: US Bureau of Labor Statistics; US Census; Dun & Bradstreet; interviews; McKinsey Global Institute analysis
McKinsey & Company
Copyright © 2012. All rights reserved | 13



Big Data

- **Big Data: volume, velocity, and variety**
- Demand from Industry: Gartner predicts that by 2018, big data demand will exceed 5 million jobs
- Infrastructure: the future of data management for analytics is the logical **data warehouse**
- Invest in Data Analytics
 - Hardware
 - Software
 - Skills

How to Deal with Big Data



How to Deal with Big Data



How to Deal with Big Data

You know **Big Data Intelligence** is important because ...



- It has a Dilbert cartoon!



What is Data?

“Data is a set of values of **qualitative** or **quantitative** variables; restated, pieces of data are individual pieces of information.”

<http://en.wikipedia.org/wiki/Data>



Data Examples

- What is Data?
 - Something recorded
- Examples
 - Info from sensors
 - Smart phone
 - Photos
 - Texts

■ Everything is data!!!

Grocery Shopping: 1915

- Family owned store
 - Inventory by hand
 - Order each week for following week
 - Informal customer relations
 - 2-3 wholesalers
- Information available
 - Inventory on hand
 - Receipts from last week includes quantity sold and price charged



Grocery Shopping: 2015

- Store as part of chain
 - Inventory updated after each sale
 - Order automated based on sales
 - Customer registered with IT system
 - 20,000 SKU's
- Information available
 - SKU sold, quantity, price, promotions, availability of competing items (this brand and competing brands), prices of all competing items, other contents of basket, shopper demographics (marital status, number of children, annual income, zip code), all prior purchases, form of payment, weather, # of other customers in store, margins on each item sold, use of coupons, form of checkout, waiting time, bag preference, time of day, etc, etc, etc.



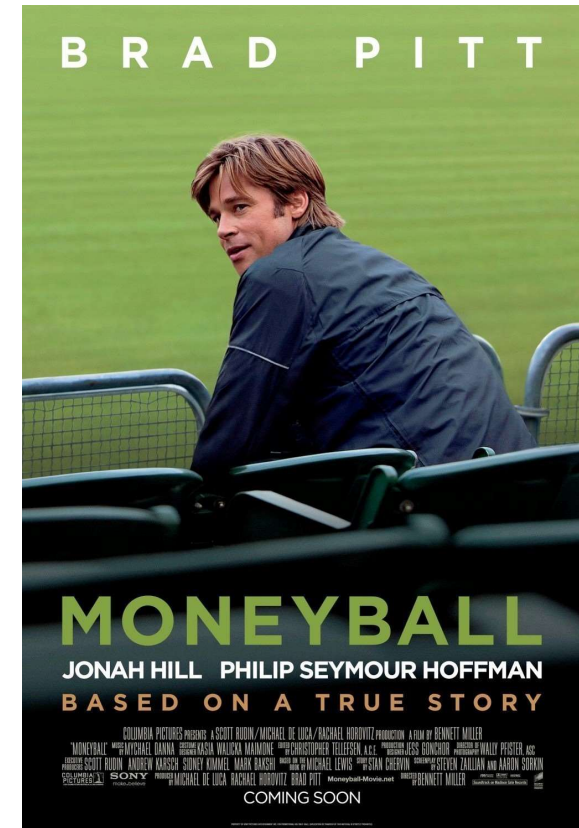


Grocery Shopping

- Lots of data collected and warehoused
 - Web data, e-commerce
 - Purchases at retail stores
 - Bank/credit card transactions
- Computers cheap and powerful
- Competitive pressure is strong
 - Customized services
 - Anticipate customer needs and work to create new ones

Data Analytics: “Money Ball”

- Fast growing area of Analytics
- Sports naturally generate real time data
 - Revenue
 - Costs
 - Performance of individuals, combinations, managers, coaches
 - Measurements in games, weeks, months, seasons, etc.





Is this new?

“In God we trust, all others must bring data.”

“It is not necessary to change.
Survival is not mandatory”

— William Edwards Deming (1900 - 1993)



Example: Advertising Data Set

- Sales (thousand units), advertising budget on TV (\$), Radio (\$), Newspaper (\$)
- 200 records

	TV	Radio	Newspaper	Sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
...				

- Question: can you predict sales given advertising expenditures of (100, 50, 25)?
- `Predict(lm(Sales~TV+Radio+Newspaper, data=ad),
data.frame(TV=100, Radio=50, Newspaper=25),
interval="confidence")`
- Prediction 16.92



Data Analytics

- Data Analytics (DA) is the science of examining raw data with the purpose of drawing conclusions about that information
- Data Analytics focused on *inference*; i.e. the process of deriving a conclusion based solely on what is already known by the researcher
- Inference: the process of arriving at some conclusion that, though it is not logically derivable from the assumed premises, possesses some degree of probability relative to the premises



DA – Discovering Patterns



Prediction is very difficult, especially if it's about
the future.

(Niels Bohr)

izquotes.com



DA – Tasks

- Supervised learning: prediction methods
 - There is one or multiple outcome measures
 - We have some values of the measures on hand
 - Use some variables to predict unknown or future values of other variables
 - Example: Demand forecasting
- Unsupervised learning: description methods
 - There is no outcome measure or none is available
 - Look for interpretable patterns that describe the data
 - Example: market segmentation



DA – Tasks

- Regression [supervised]
- Classification [supervised]
- Clustering [unsupervised]
- Principal Component Analysis [unsupervised]
- Large data sets allow us to do some things that are harder to do with smaller data sets
 - Split sample in useful ways



DA – Steps

- Problem Identification
- Design approach to gather data
- Data Collection
- Data Scrubbing
- Analysis
- Interpretation
- Reporting Results



Data Collection

- Data Collection requires rigorous and systematic design and execution
 - Planning, development, process management
- Data types
 - Surveys, observations, measurements from experiments, real world “natural experiments”
- Much data collection can only be carried out in the messy uncontrolled environments of the real world
 - The search for cause and effect will require tradeoffs between real-world contexts and a controlled environment



Data Cleaning (Scrubbing)

- Data that is incorrectly entered is useless
 - Are characters entered accurately
 - Are numeric values within a reasonable range
- Data that is incomplete is useless (sometimes)
 - Check for missing values
 - Check for duplicates
 - Check for invalid entries
- Technical issues
 - Is software working together correctly
 - Are files combined properly



Data Analysis

- Model definition and development
 - Search for relationships
 - Propose multiple models
- Are you testing a hypothesis?
 - How can your hypothesis be proven wrong
 - Can we reject the opposite of what you want to test
- Use known results to fit a model
- Use other data to test ability to make predictions
- Select model to present

Why the DA Course is Hard

- **R** is a programming language
 - Code writing is not natural
 - Conditioned by apps and menus
 - Lessons build on prior work
- Statistics is Key
 - Review Regression models
 - Hypothesis testing
 - Central limit theorem



Why the DA Course is Not Hard

- Examples across all platforms
 - Text book
 - Slides
 - **R** code provided
- Cut and Paste
 - Homework steps will be very similar to examples
 - Same code can be used after changing variables names



Ask Questions

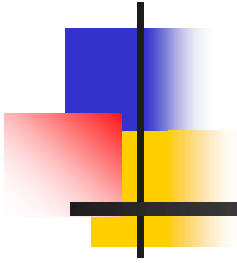
- Don't conclude that staring is learning: try the following
 - Consider text, slides, readings and code
 - Help in **R**-studio
 - Stackoverflow.com
 - Lynda.com
 - YouTube
 - Roger Peng, MarinStatsLectures
 - Discussion board
 - E-mail or call instructor





How to ask a question

- Be specific
 - Bad: What's regression, I'm lost?
 - Better: If the coefficient in my output is -2, does this mean that the objective goes up or down when x changes?
- Show what you did
 - Bad: pages of code
 - Better: Here are my first 10 lines. My error occurs after line 9
- Explain what you expected to get
 - Bad: the answer ain't here
 - Better: the result does not show the confidence intervals
- Explain what you actually got
 - Bad: the screen says "error"
 - Better: object x not found
 - <https://www.youtube.com/watch?v=ZFaWxxzouCY&index=2&list=PL7Tw2kQ2edvpNEGrU0cGKwmdDRKc5A6C4>



Course Information

Topics

Session Organization

Grades



PART 1-B



Course Resources: Part 1

- Text

- Available as PDF via JHU Online Library

- Software

- Download instructions in Syllabus
 - **R** and **R**-Studio

- Video

- Lynda.com
- <http://carey.jhu.edu/about/lynda>
- Up and Running with **R**



Course Resources: Part 2

- Syllabus

- Available under Student Resources/Syllabus

- Slides

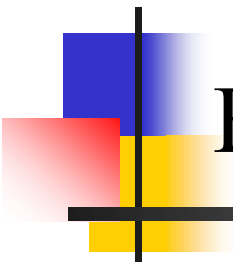
- All slides under Course Documents

- Data Sets

- Data for many examples and in-class exercises

- Readings

- Some extras in place now
 - More may be added at any time



Evaluation and Grading

- Weekly Exercises (12%)
 - Submission of weekly exercises (6)
- Discussions (10%)
- Assignments (32%)
 - 3 homework assignments (Weeks 3, 4, 6)
 - HW 1 is individual assignment
 - HW 2 and 3 are group assignments (4-6 members)
 - Submit via Blackboard: MSWord and **R** files
- Final Project (22%)
 - Group project (4-6 members)
 - Analysis and presentation (Session 7)
- Final Exam (24%)
 - Samples available on Blackboard

Exercise

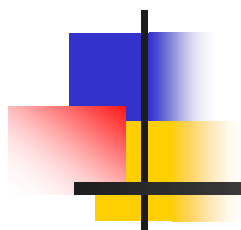
- Install **R** and **R**-studio
- Download data files from Blackboard
 - Create directory for data files
 - Set this as working directory
- 10 mins





Topic # 01: Data Analytics

Intro to **R** with Examples



Instructor: Prof. Arnab Bisi, Ph.D.
Johns Hopkins Carey Business School



R- Background

- S is a statistical high-level and interpreted programming language developed at the **Bell laboratories** around 1975 by **John Chambers**. The commercial implementation of S is called S-PLUS and appeared in 1988.
- **R** is an open-source implementation of S and was created in the early nineties by **Ross Ihaka** and **Robert Gentleman** at the University of Auckland, New Zealand. These days, **R** is maintained by the **R** core team.
- **R** has become very popular particularly in academia but also in industry. Much of **R**'s success story is due to all the packages written for R by the **R**-community.

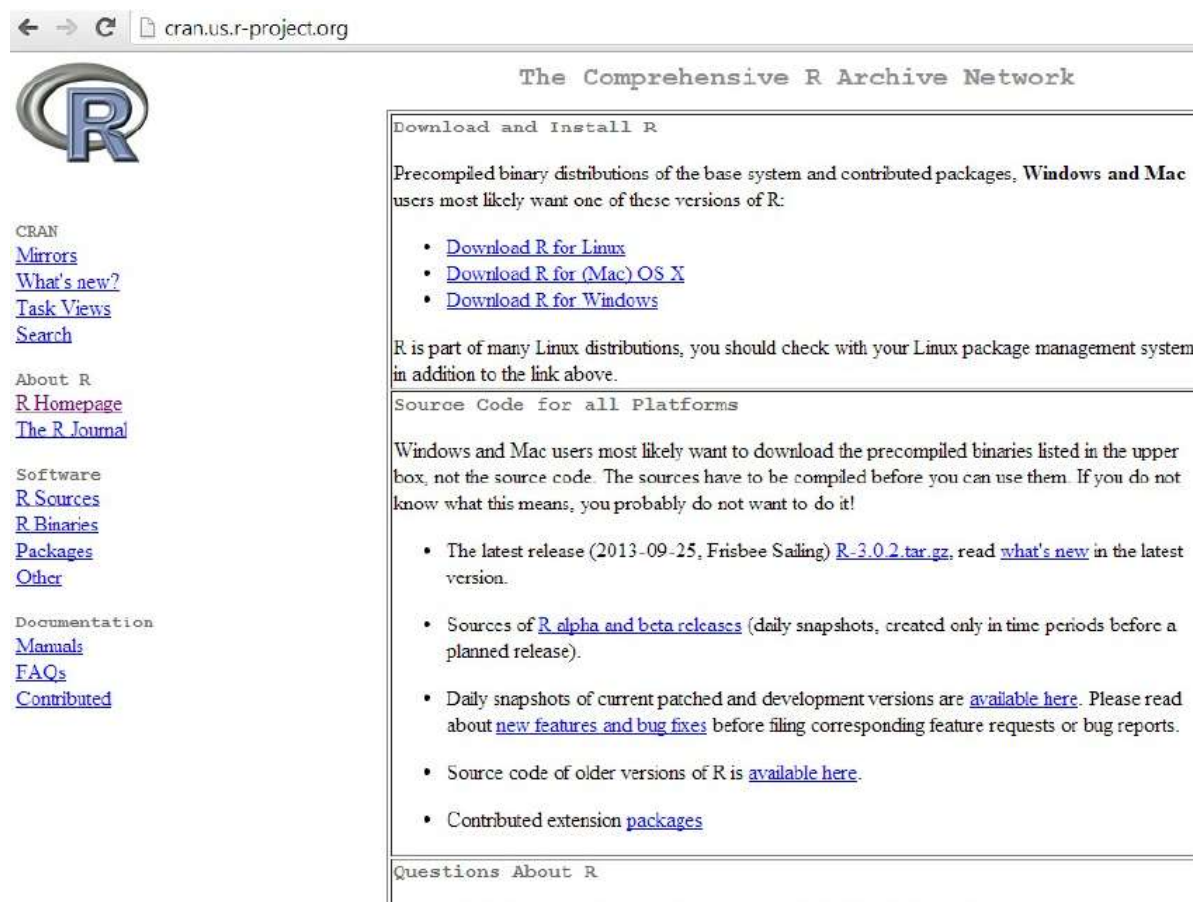


What is **R**?

- A software package
- A programming language
- A toolkit for developing statistical and analytical packages and apps
- An extensive library of statistical and mathematical software and algorithms
- A scripting language

Installing R

- <http://carey.jhu.edu/about/lynda>
- Up and Running with R
- Installing R on your computer & Using RStudio



The screenshot shows the CRAN website with the following content:

CRAN
[Mirrors](#)
[What's new?](#)
[Task Views](#)
[Search](#)

About R
[R Homepage](#)
[The R Journal](#)

Software
[R Sources](#)
[R Binaries](#)
[Packages](#)
[Other](#)

Documentation
[Manuals](#)
[FAQs](#)
[Contributed](#)

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- [Download R for Linux](#)
- [Download R for \(Mac\) OS X](#)
- [Download R for Windows](#)

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2013-09-25, Frisbee Sailing) [R-3.0.2.tar.gz](#), read [what's new](#) in the latest version.
- Sources of [R alpha and beta releases](#) (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are [available here](#). Please read about [new features and bug fixes](#) before filing corresponding feature requests or bug reports.
- Source code of older versions of R is [available here](#).
- Contributed extension [packages](#)

Questions About R



Why Use **R**?

- **R** is free
- **R** is cross-platform and runs on Windows, Mac, and Linux (as well as more obscure systems)
- **R** provides a vast number of useful statistical tools, many of which have been tested
- **R** produces publication-quality graphics in a variety of formats.
- **R** plays well with FORTRAN, C, and scripts in many languages
- **R** scales, making it useful for small and large projects.



Why Use **R**Studio?

- **R**Studio is free
- **R**Studio is cross-platform and looks the same on all machines
- **R**Studio provides a vast number of useful tools, for the novice that handle pathways, input, and output and is consistent across platforms.
- **R**Studio makes it easy to experiment to find the right syntax and then save the correct commands for later.
- **R** plays well with FORTRAN, C, and scripts in many languages
- **R**Studio helps handle the interfaces with Office.



Do We Still Need Excel?

- Excel is almost free
- Excel is cross-platform and looks the same on all machines
- Excel provides a vast number of useful statistical tools, many of which have been tested and work well on “SMALL” data sets.
- Being able to “See” your data offers many advantages
- Does a very good job of sorting, using relative addresses and plotting small data sets



Excel is Extremely useful for Data Scrubbing



R Help Functions

- If you know the name of the function or object on which you want help:
 - `> help(read.csv)`
 - `> help('read.csv')`
 - `> ?read.csv`
- If you do not know the name of the function or object on which you want help:
 - `> help.search('input')`
 - `> RSiteSearch('input')`
 - `> ??input`
- Do Not forget our friend: **Google**
- <http://rseek.org/>
- <http://stackoverflow.com/tour>



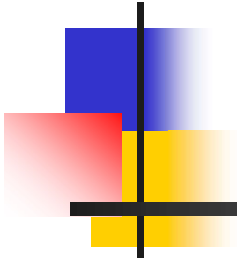
Vector Operations

- Operations on a single vector are typically done element by element
- If the operation involves two vectors:
 - Same length: **R** applies the operation to each pair of elements
 - Different length but one is a multiple of the other: **R** reuses the shorter vector as many times as needed
 - Different length and one is not a multiple of the other: **R** reuses the shorter vector as needed and delivers a warning.



Factors

- A factor is a special type of vector, normally used to hold a categorical variable in many statistical functions.
 - When **R** reads a column of data and it contains a non-numeric element, **R** will treat it as a factor
 - Factors in **R** often appear to be character vectors when printed, but you will notice that they do not have double quotes around them.



Questions, Comments?

Let's move to the Code.