Lecture 1: Introduction

UNIVERSITY OF
SAN FRANCISCO

James D. Wilson

BSDS 100 - Intro to Data Science with R

Outline

- Course Overview
- What is Data Science?
  - Where is Data Science?
  - A brief history
- The R Programming Language and RStudio
  - Comparison with other programming languages
  - Installation
  - Handy Shortcuts

Part I: Course Overview

A Little About Me

- Ph.D. Statistics and Operations Research (UNC Chapel Hill, ’15)
  - Research focused on statistical analysis of networks
  - Explore, model, and analyze network data (e.g., social networks)
- M.S. Mathematical Sciences (Clemson University, ’10)
- B.S. Mathematics and Chemistry (Campbell University ’08)

A Little About Me

Classes I teach:

- BSDS 100 - Intro to Data Science with R
- MATH 106 - Business Statistics
- MATH 370 - Probability with Applications
- MATH 373 - Statistical Learning
- MSAN 601 - Linear Regression Analysis
- MSAN 630 - Advanced Computational Statistics
- MSAN 700 - Social Network Analysis

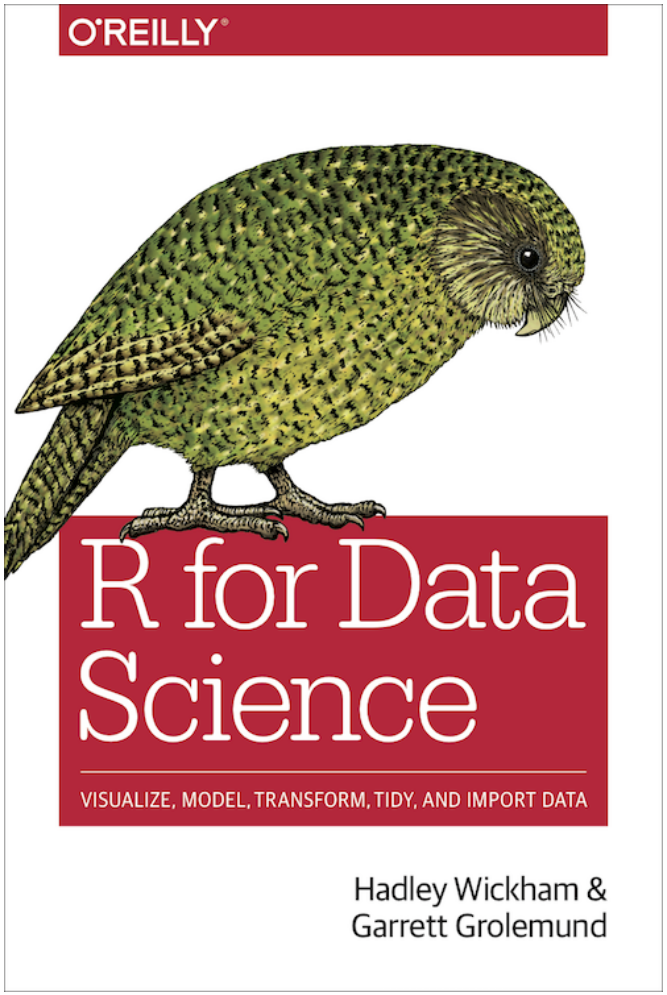
A Little About Me

- Born and raised in NC (near Raleigh)
- Live in Rockridge, Berkeley.
- A huge college basketball fan! (Go Heels!)
- Have loved college football since 2008 (Go Tigers!)
- Enjoy tasting beers (bourbon-barrel stouts are my favorite).

Course Description and Syllabus

All lecture notes, the syllabus, assignments, and course description are available at this site: `https://github.com/jdwilson4/Intro-Data-Science-2017`

Main Text



Available online here: `http://r4ds.had.co.nz/index.html`

Other Resources

The Best Place for Answers to R Questions?

Part II: What is Data Science?

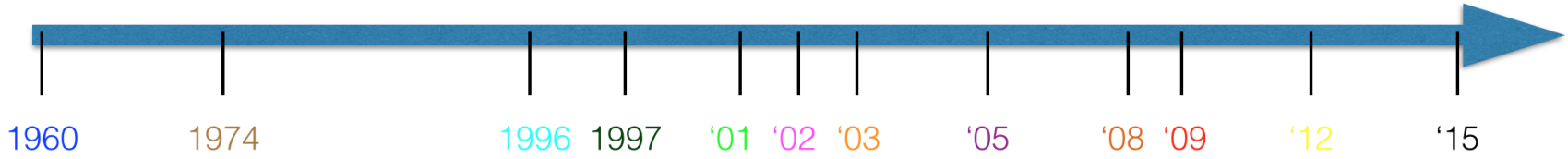
What is Data Science?

- **Wikipedia**: “the extraction of knowledge from data."
- A precise definition is a bit unclear and has faced much controversy... (we’ll see more on this in a moment)
- Practitioners tend to agree on the *components* of data science:
  - gathering and cleaning data
  - database management
  - exploratory analysis
  - predictive modeling
  - data summary and visualization

Where is Data Science?

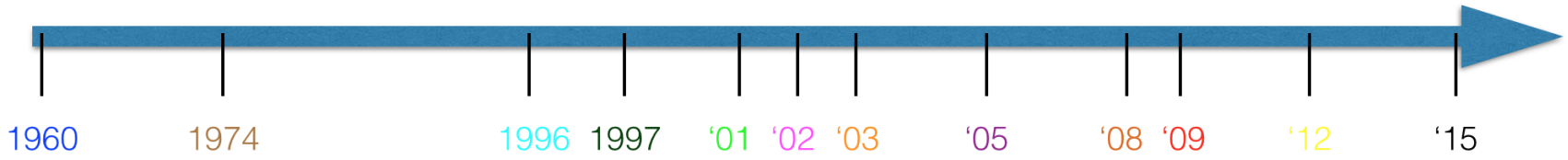
- Twitter feed, December 2014

The Evolution of Data Science



- 1960: Peter Naur (CS Ph.D.) published *Datalogy: the science of data and its place in education.*
- 1974: Peter Naur published *Concise Survey of Computer Methods*.
  - defines data science as “the science of dealing with data, once they have been established."
  - continues to say that “... the relation of the data to what they represent is delegated to other fields and sciences."

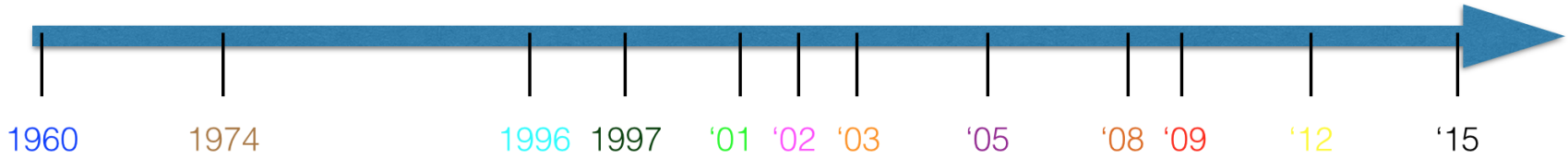
The Evolution of Data Science



- 1996: International Federation of Classification Societies meet in Tokyo and for the first time include “data science" in the conference title: “Data science, classification, and related methods."

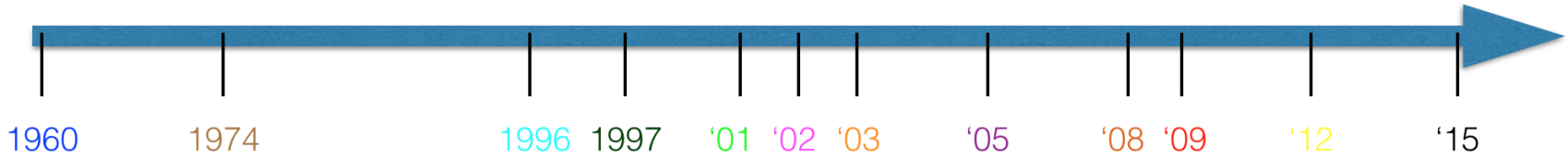
- 1997: C.F. Jeff Wu gave the inaugural lecture “Statistics = Data Science?" for appointment to the H. C. Carver Professorship at the University of Michigan.

The Evolution of Data Science



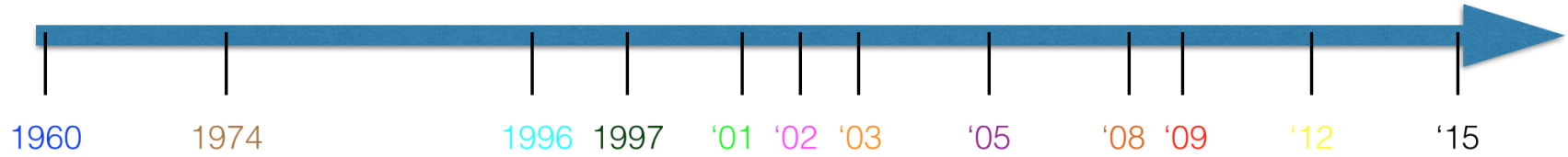
- 2001: William Cleveland (Bell Labs) published *Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics*.
  - Sets forth 6 areas for a university department involving statistics.
- 2002: *Data Science Journal* is launched
  - Focus on data systems, publications on internet, and applications
- 2003: *Journal of Data Science* is launched
  - Focus on application of statistical and quantitative methods

The Evolution of Data Science



- 2005: National Science board redefines data scientists:
  - “The information and computer scientists, data and software programmers, disciplinary experts, ... who are crucial to successful management of a digital data collection whose primary activity is to conduct creative inquiry and analysis"
- 2008: DJ Patil (LinkedIn) and Jeff Hammerbacher (Facebook) coined the term "data scientist" to define their jobs

The Evolution of Data Science



- January, 2009: Hal Varian (chief economist at Google) writes that "... the sexy job in the next 10 years will be statisticians."
- October, 2012: Harvard Business Review publishes "Data Scientist: The Sexiest Job of the 21st Century."
- February 5th, 2015: DJ Patil appointed as the first Chief Data Scientist in the White House.

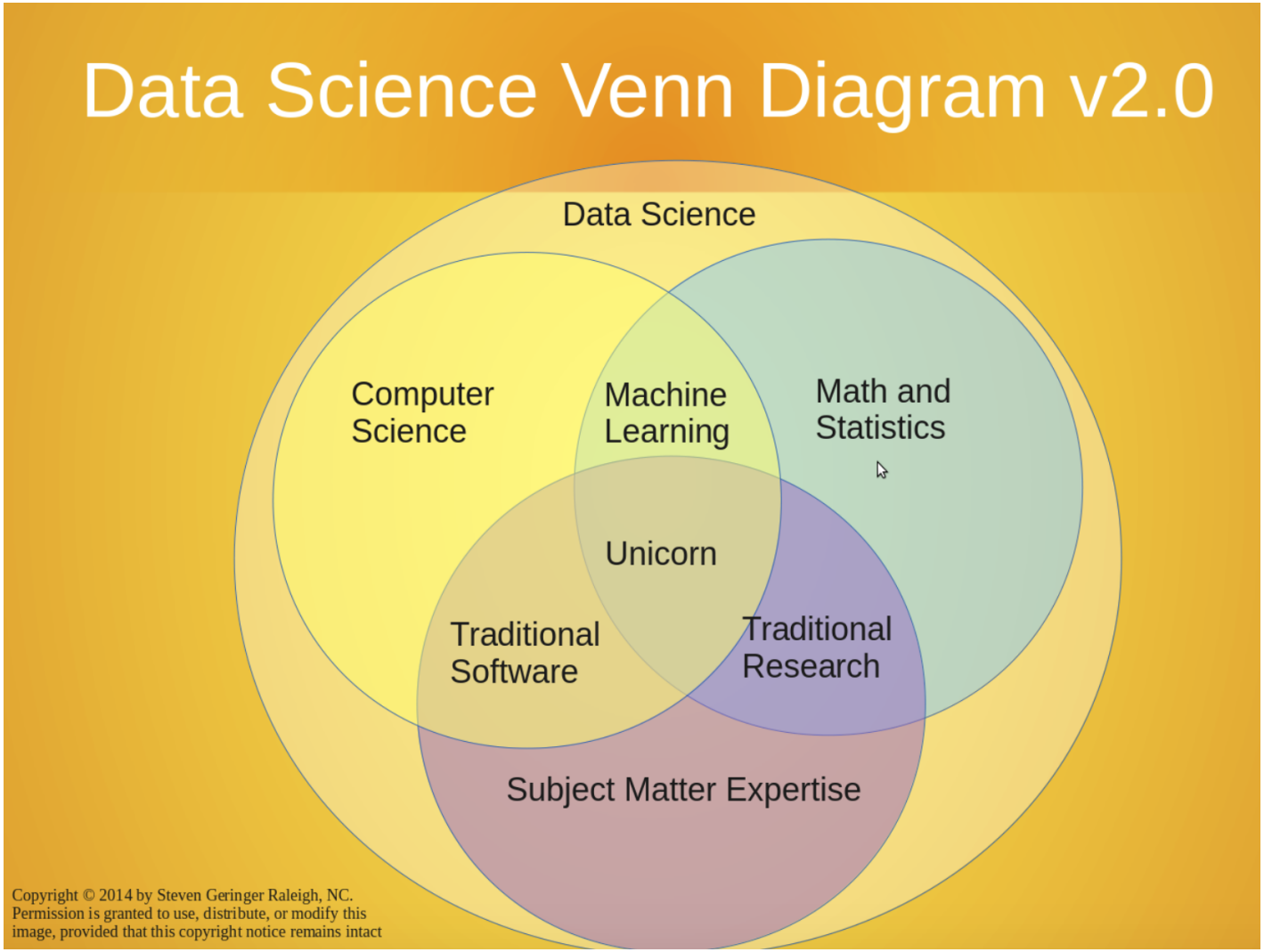
Applications

Marketing analytics, sports analytics, biotechnology, social experiments, e-commerce, government analysis, ...

Why Data Science?

- Size, complexity, and amount of data
  - Predicted $\approx$ 40 trillion gigabytes of data in 2020; up from 130 billion in 2005!
  - Big data requires innovative techniques for analysis
- *McKinsey*: "The U.S. faces a shortage of 140K - 190K people with analytical expertise and 1.5 million managers and analysts with the skills to understand and make decisions based on the analysis of big data." (May, 2011)
- *Harvard Business Review*: "Data Scientist: The Sexiest Job of the 21st Century." (October, 2012)

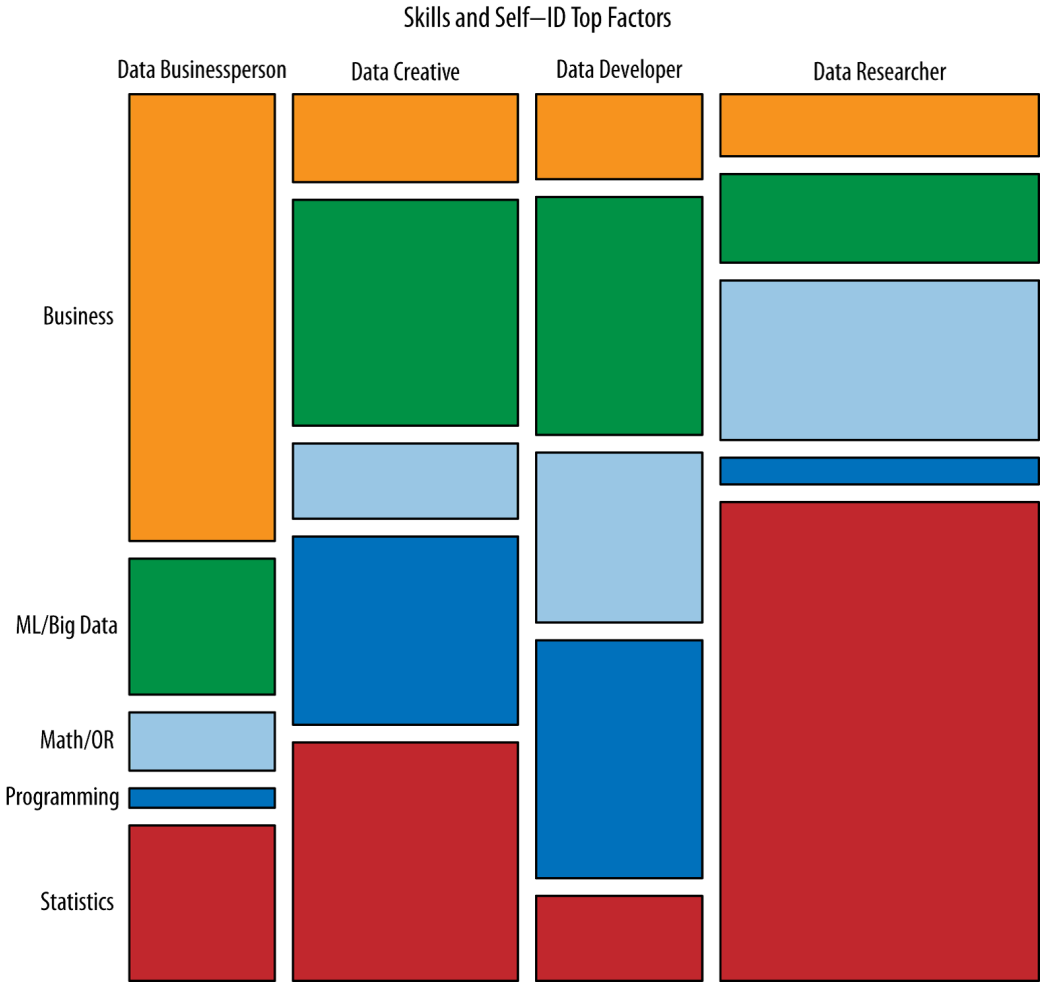
Data Scientists: The unicorn industries want?



Data Scientists: The unicorn industries want?

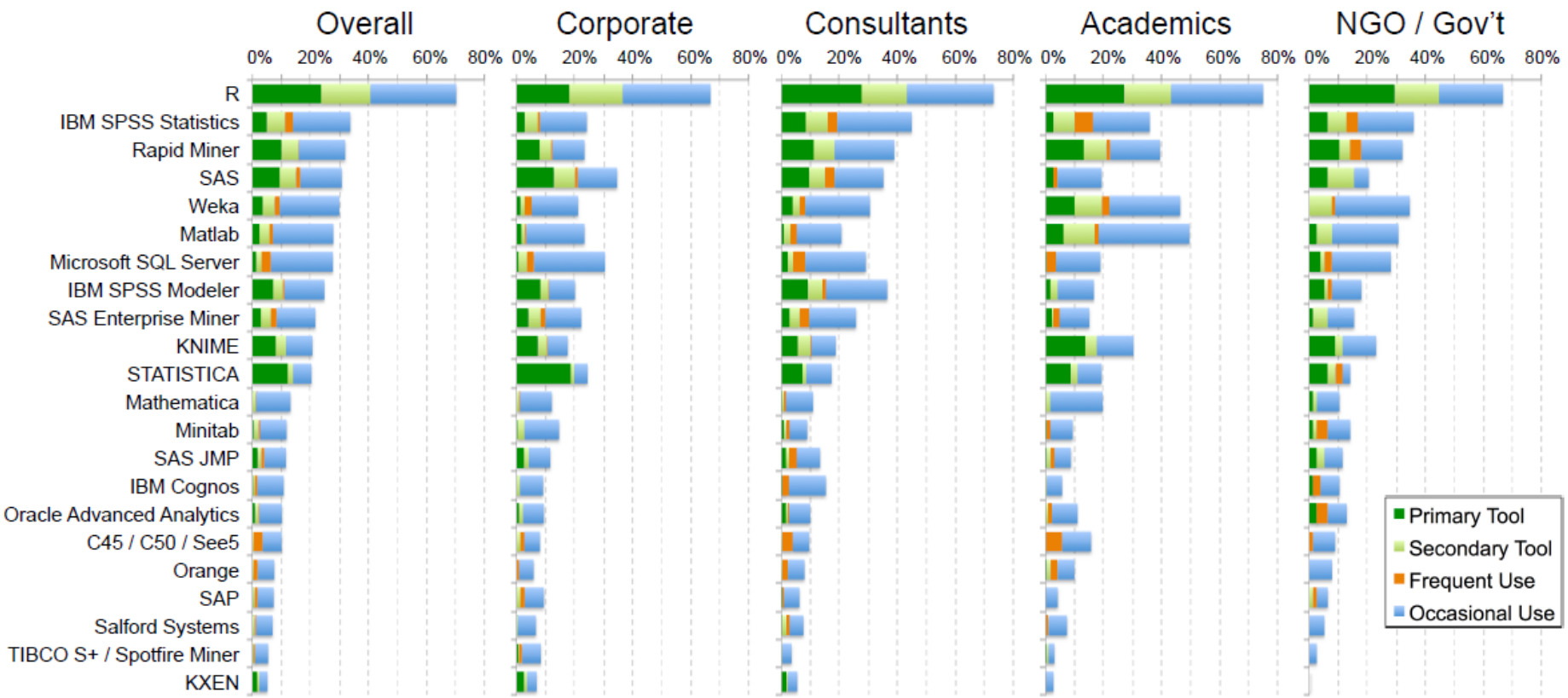
- The field is inherently interdisciplinary
  - mathematical statistics
  - computer science
  - domain expertise
- The magical Unicorn: having all three skills
  - In 2014, these jobs go unfilled for 6 months or longer on average
- Has lead to the development of data science *teams*
  - hope is to merge skills of analysts

The Analysts of Data Science



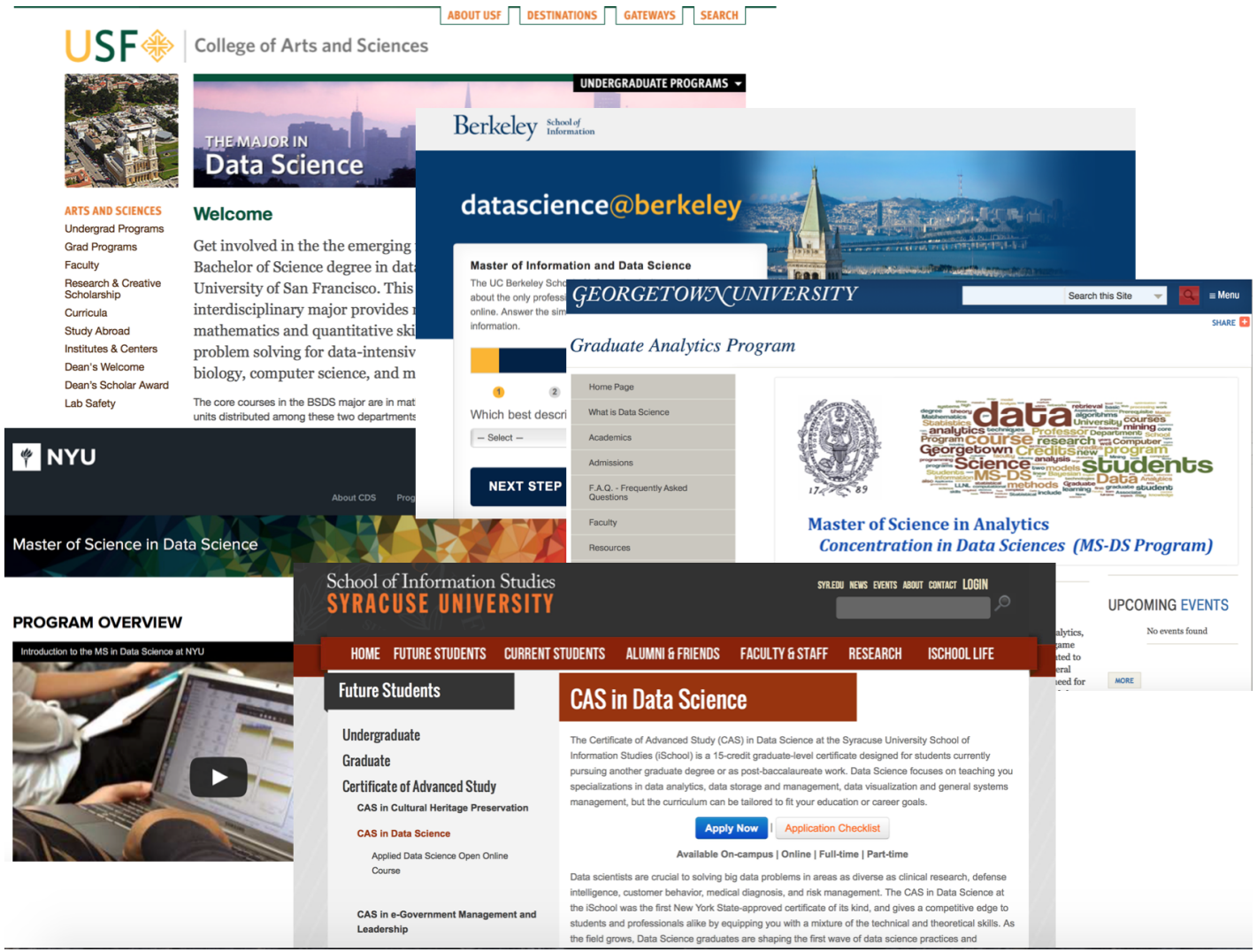
"Analyzing the Analyzers (2013) by Harry, Murphy, and Vaisman."

Software: A Data Scientist's first weapon

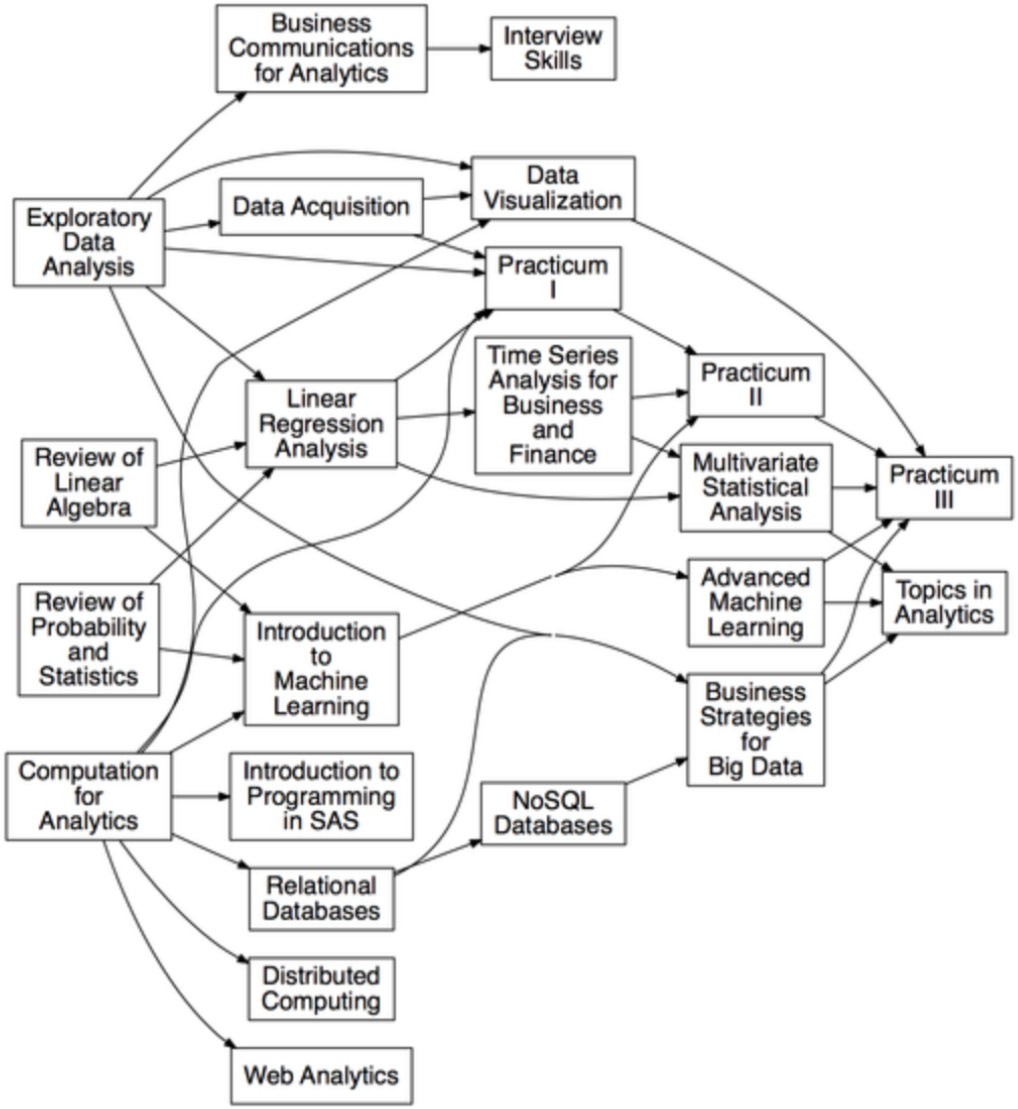


-www.datasciencecentral.com

Data Science in Academia



Academic Programs



A Data Scientist's Toolkit

Harvard's data science toolkit:

1. **Wrangle the data**: gather, clean, and sample data
2. **Manage the data**: access big data quickly and reliably
3. **Explore the data**: to make a hypothesis
4. **Make predictions**: statistical methods
5. **Communicate the results**: visualization, presentations, summaries

Get Involved! Great Resources

- Flowingdata.com
  - Contemporary visualization and data manipulation techniques
- Kaggle.com
  - Kaggle competitions: win money for solving problems!
- Coursera.org
  - Free online courses in data science and machine learning
  - 972 courses. Great resource for coding, data analysis, etc.
  - Recent notable course: "The Data Scientist's Toolbox."

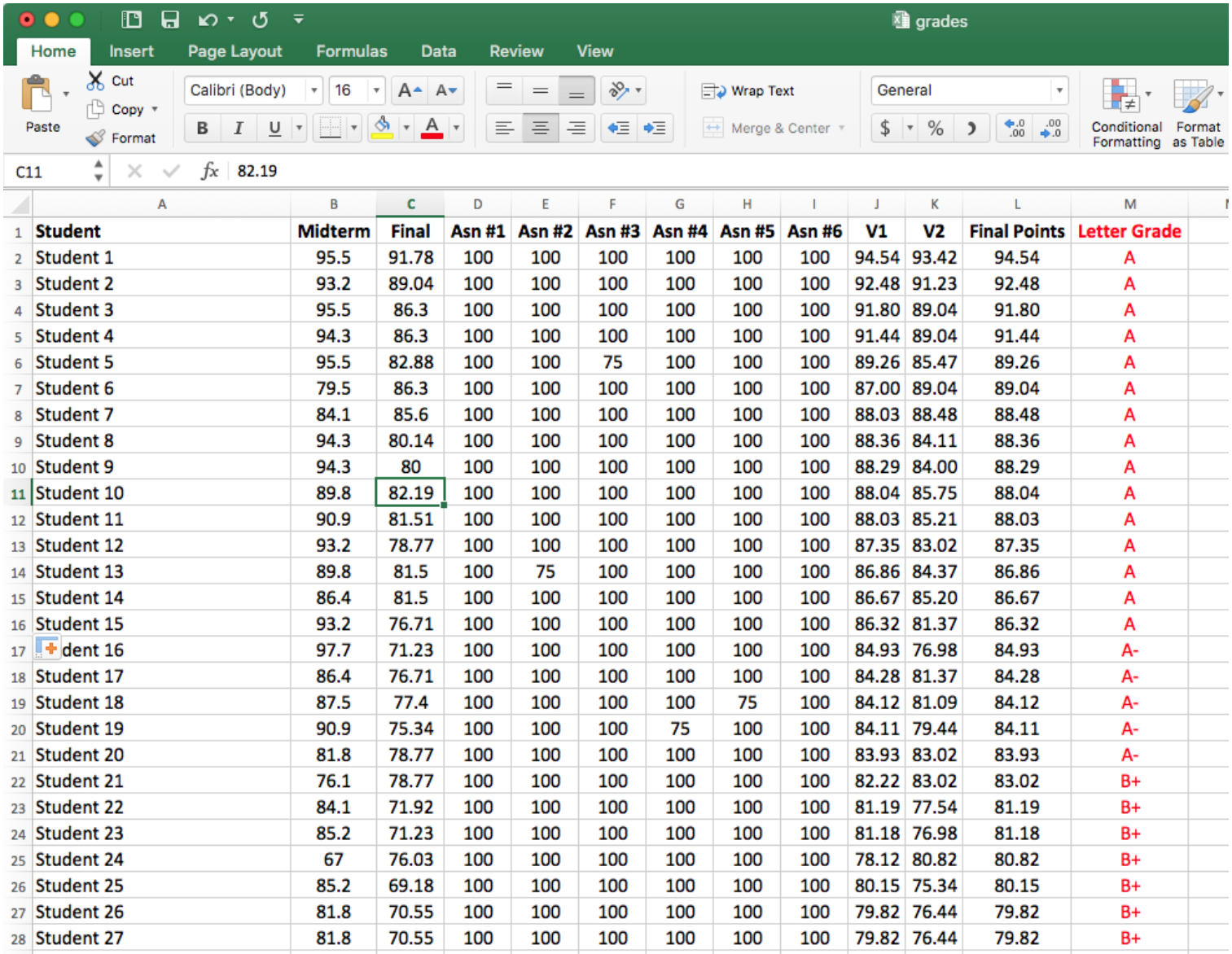
Part III: R and RStudio

Why Use R?

- Open source (free)
- Runs on just about any platform
- Great visualization capabilities (`ggplot2`)
- Read/write from/to various data sources
- Scripting language (interpreted)
- Massive library of data manipulation and statistical packages

But... what about Excel?

Excel is Great for Certain Things...



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	**Student**	**Midterm**	**Final**	**Asn #1**	**Asn #2**	**Asn #3**	**Asn #4**	**Asn #5**	**Asn #6**	**V1**	**V2**	**Final Points**	**Letter Grade**
2	Student 1	95.5	91.78	100	100	100	100	100	100	94.54	93.42	94.54	A
3	Student 2	93.2	89.04	100	100	100	100	100	100	92.48	91.23	92.48	A
4	Student 3	95.5	86.3	100	100	100	100	100	100	91.80	89.04	91.80	A
5	Student 4	94.3	86.3	100	100	100	100	100	100	91.44	89.04	91.44	A
6	Student 5	95.5	82.88	100	100	75	100	100	100	89.26	85.47	89.26	A
7	Student 6	79.5	86.3	100	100	100	100	100	100	87.00	89.04	89.04	A
8	Student 7	84.1	85.6	100	100	100	100	100	100	88.03	88.48	88.48	A
9	Student 8	94.3	80.14	100	100	100	100	100	100	88.36	84.11	88.36	A
10	Student 9	94.3	80	100	100	100	100	100	100	88.29	84.00	88.29	A
11	Student 10	89.8	82.19	100	100	100	100	100	100	88.04	85.75	88.04	A
12	Student 11	90.9	81.51	100	100	100	100	100	100	88.03	85.21	88.03	A
13	Student 12	93.2	78.77	100	100	100	100	100	100	87.35	83.02	87.35	A
14	Student 13	89.8	81.5	100	75	100	100	100	100	86.86	84.37	86.86	A
15	Student 14	86.4	81.5	100	100	100	100	100	100	86.67	85.20	86.67	A
16	Student 15	93.2	76.71	100	100	100	100	100	100	86.32	81.37	86.32	A
17	dent 16	97.7	71.23	100	100	100	100	100	100	84.93	76.98	84.93	A-
18	Student 17	86.4	76.71	100	100	100	100	100	100	84.28	81.37	84.28	A-
19	Student 18	87.5	77.4	100	100	100	100	75	100	84.12	81.09	84.12	A-
20	Student 19	90.9	75.34	100	100	100	75	100	100	84.11	79.44	84.11	A-
21	Student 20	81.8	78.77	100	100	100	100	100	100	83.93	83.02	83.93	A-
22	Student 21	76.1	78.77	100	100	100	100	100	100	82.22	83.02	83.02	B+
23	Student 22	84.1	71.92	100	100	100	100	100	100	81.19	77.54	81.19	B+
24	Student 23	85.2	71.23	100	100	100	100	100	100	81.18	76.98	81.18	B+
25	Student 24	67	76.03	100	100	100	100	100	100	78.12	80.82	80.82	B+
26	Student 25	85.2	69.18	100	100	100	100	100	100	80.15	75.34	80.15	B+
27	Student 26	81.8	70.55	100	100	100	100	100	100	79.82	76.44	79.82	B+
28	Student 27	81.8	70.55	100	100	100	100	100	100	79.82	76.44	79.82	B+

...but Not Everything

Sample Data

- Six columns of data with ~ 1.05 million rows
- Column 5: `startDate`
- Column 6: `endDate`
- **Objective**: test to see if `endDate` < `startDate`

RESULTS

- **Excel**: good luck...
- `R`: 33 min (poor coding technique)
- `R`: 58.5 sec (improved coding technique)

...but Not Everything

Sample Data

- Six columns of data with ~ 1.05 million rows
- Column 5: `startDate`
- Column 6: `endDate`
- **Objective**: test to see if `endDate` < `startDate`

RESULTS

- **Excel**: good luck...
- `R`: 33 min (poor coding technique)
- `R`: 58.5 sec (improved coding technique)

How about Python?

Vectorization in R

- Vectorized code saves time asking `type` questions
- There is an optimized engine—a basic linear algebra system (BLAS)—that is highly efficient at solving linear algebra problems
- A lot or R functions are written in C (or variants)
- MATLAB, Mathematica and the NumPy package for Python are also vectorized

http://www.noamross.net/blog/2014/4/16/vectorization-in-r–why.html

Installing R

- RStudio is a nice, user-friendly integrated development environment (IDE), but can be quirky at times — still highly recommended and what I will use in class
- You can even run R from a terminal window if you wish
- Download and install at this website:

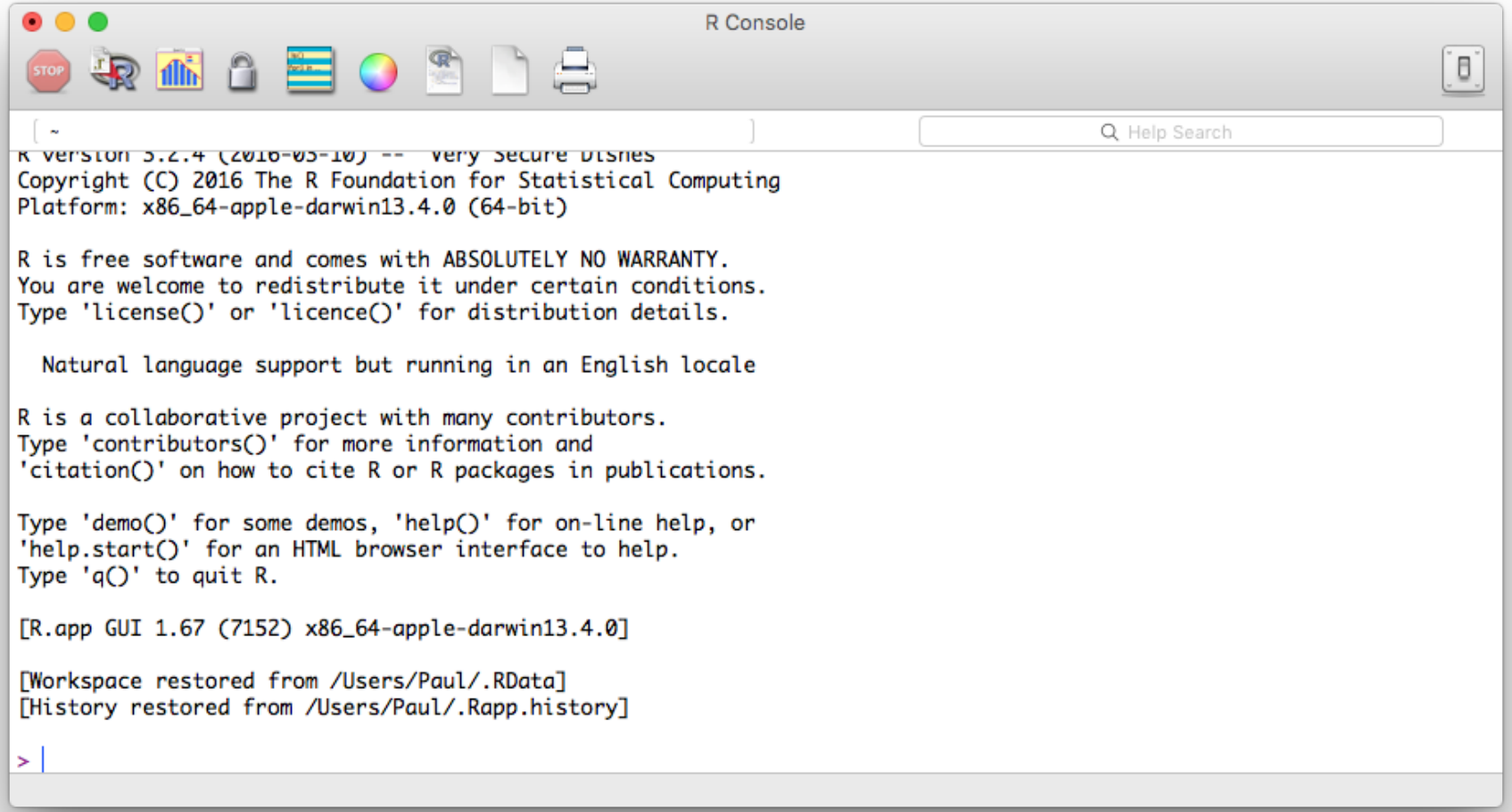
  `https://www.r-project.org`
- **Important**: You will have to re-install R from time-to-time to maintain the newest version so that code remains compatible! New versions generally come out every 4 - 6 months.

Installing RStudio

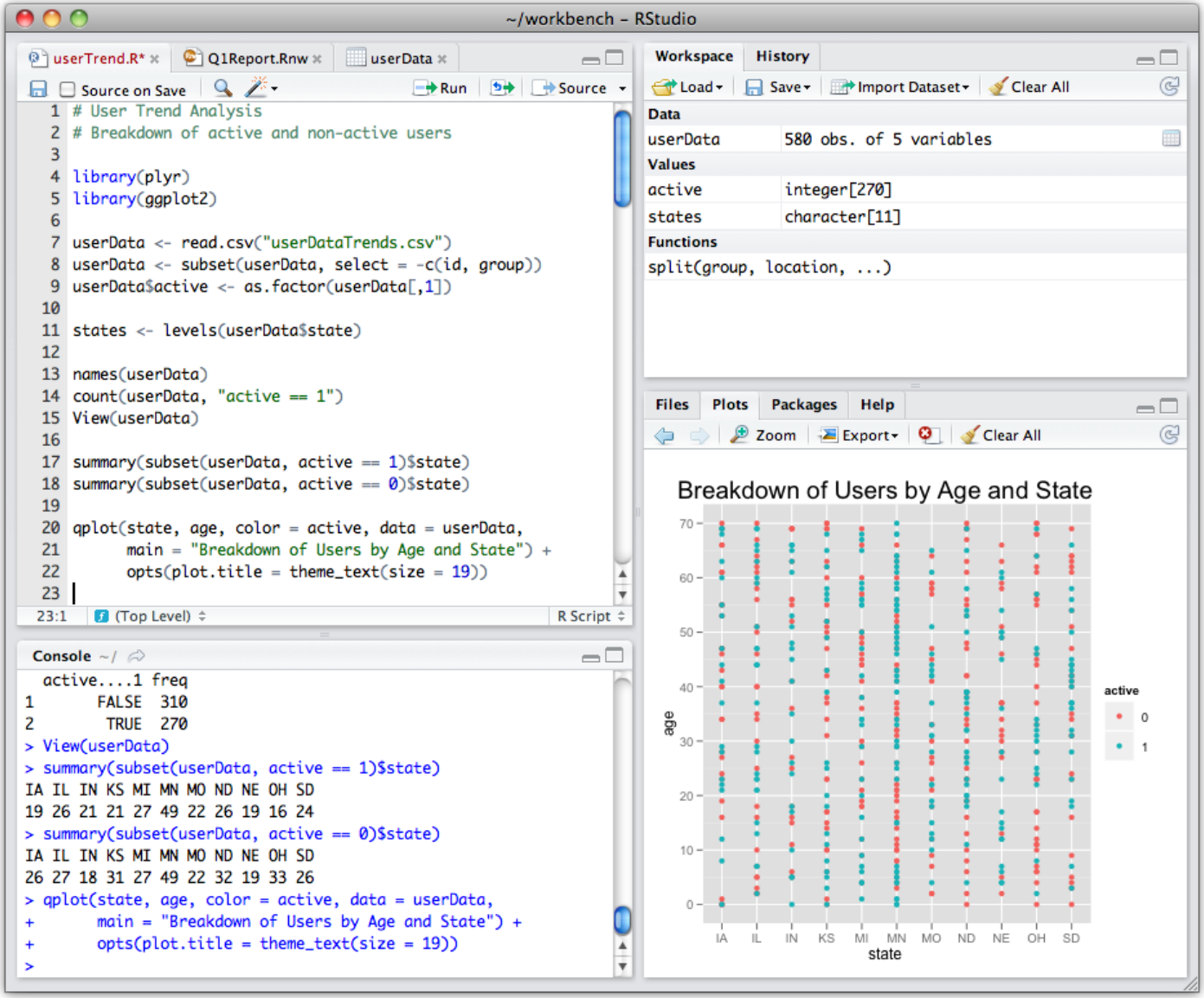
- RStudio has a very nice graphical user interface (GUI) that is easier to use than base R
- We will be using this throughout the course
- Make sure that you have R installed first. Then, download and install at this website:

  https://www.rstudio.com/products/RStudio/

The R Graphical User Interface (GUI)



The RStudio GUI



RStudio has Four Panels

- Console (bottom left)
- Scripting/Viewing (top left)
- Files/Packages/Help/Viewer (top right)
- Environment/History/Plots (bottom right)

Really Advanced Calculators

At their core, `R` and `R`Studio are just calculators! Try the following

$3 + 2 = ?$ $\log(10) = ?$ $\sqrt{32} = ?$

```
> 3 + 2
> log(10)
> sqrt(32)
```

Notes on R

- R is case-sensitive
- I require you to use the assignment operator `'<-'` instead of the equality operator `'='` for all submitted code, even though both work, e.g.,

Syntax	*Comments*
`x <- 5`	standard syntax, required
`x = 5`	poor syntax, not permitted
`5 -> x`	awkward syntax, not permitted (but it works)

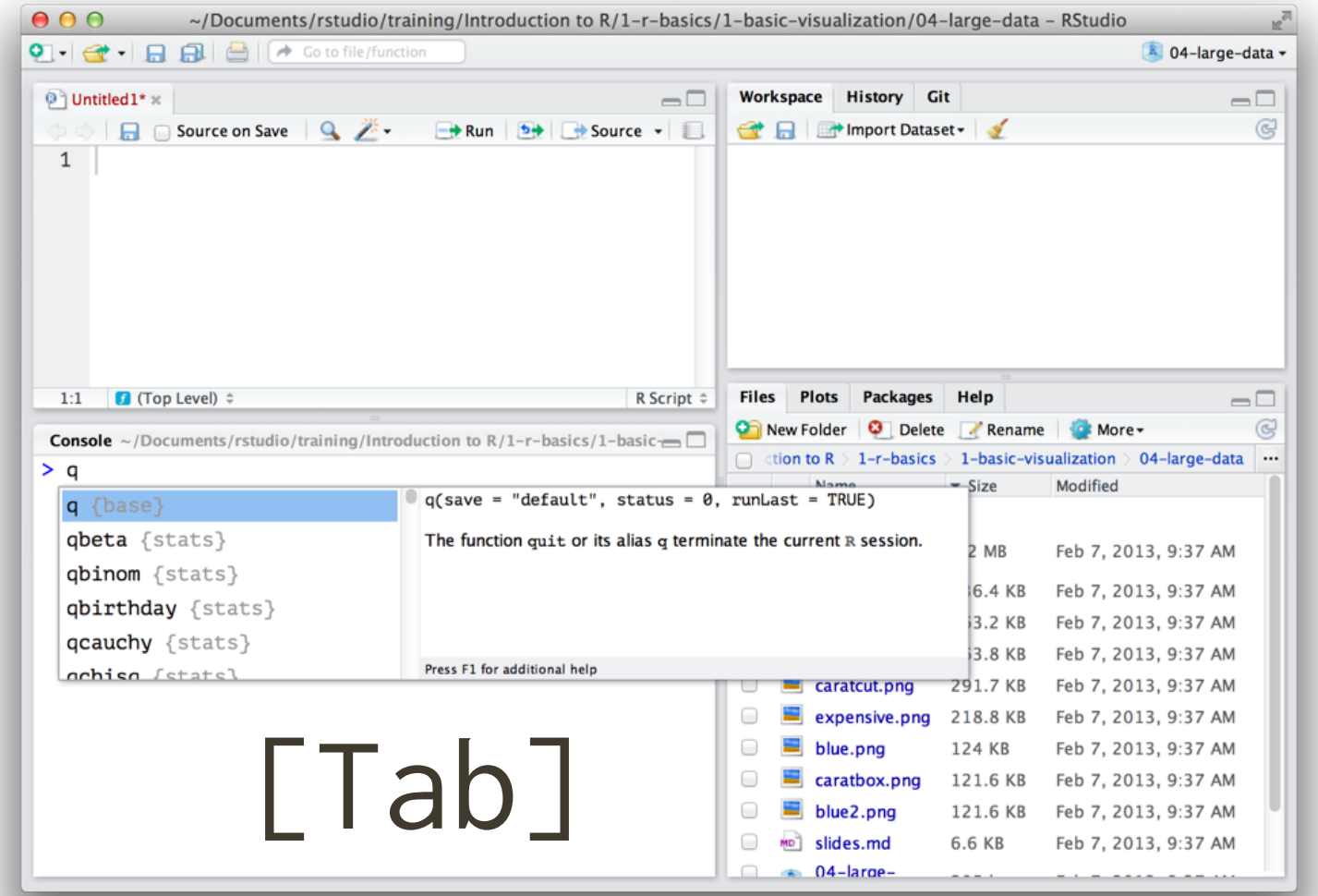
Basic R Help Functions

Function	Action
`?foo`	Help on the function `foo`
`??foo`	Search the help system for instances of the function `foo`
`data()`	List all available example datasets contained in currently loaded packages
`getwd()`	List the current working directory
`ls()`	List the objects in the current directory

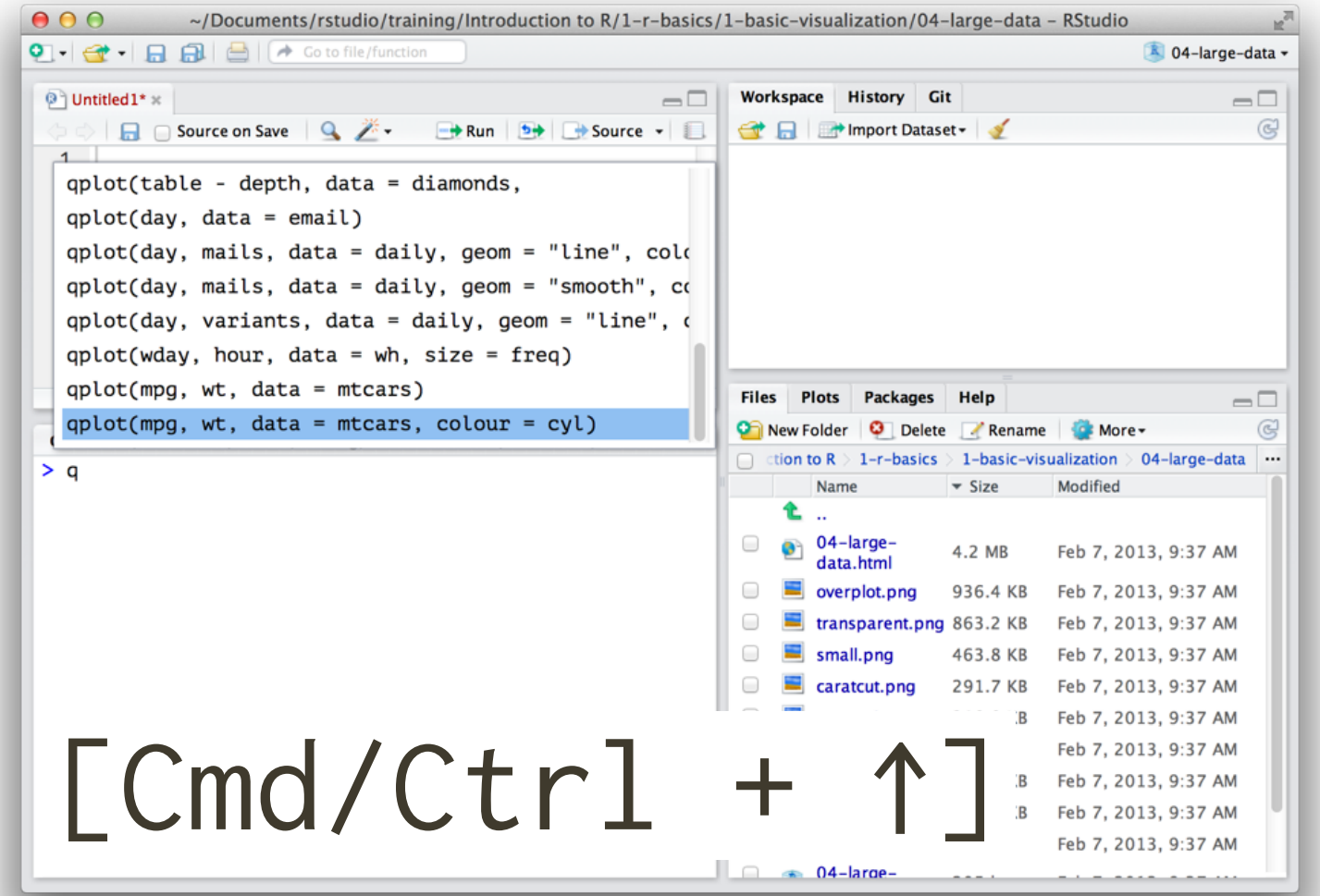
Basic R Workspace Functions

Function	Action
`getwd()`	List the current working directory.
`setwd("mydirectory")`	Change the current working directory to *mydirectory*.
`ls()`	List the objects in the current workspace.
`rm(objectlist)`	Remove (delete) one or more objects.
`help(options)`	Learn about available options.
`options()`	View or set current options.
`history(#)`	Display your last # commands (default = 25).
`savehistory("myfile")`	Save the commands history to *myfile* (default = `.Rhistory`).
`loadhistory("myfile")`	Reload a command's history (default = `.Rhistory`).
`save.image("myfile")`	Save the workspace to myfile (default = `.RData`).
`save(objectlist, file="myfile")`	Save specific objects to a file.
`load("myfile")`	Load a workspace into the current session (default = `.RData`).
`q()`	Quit R. You'll be prompted to save the workspace.

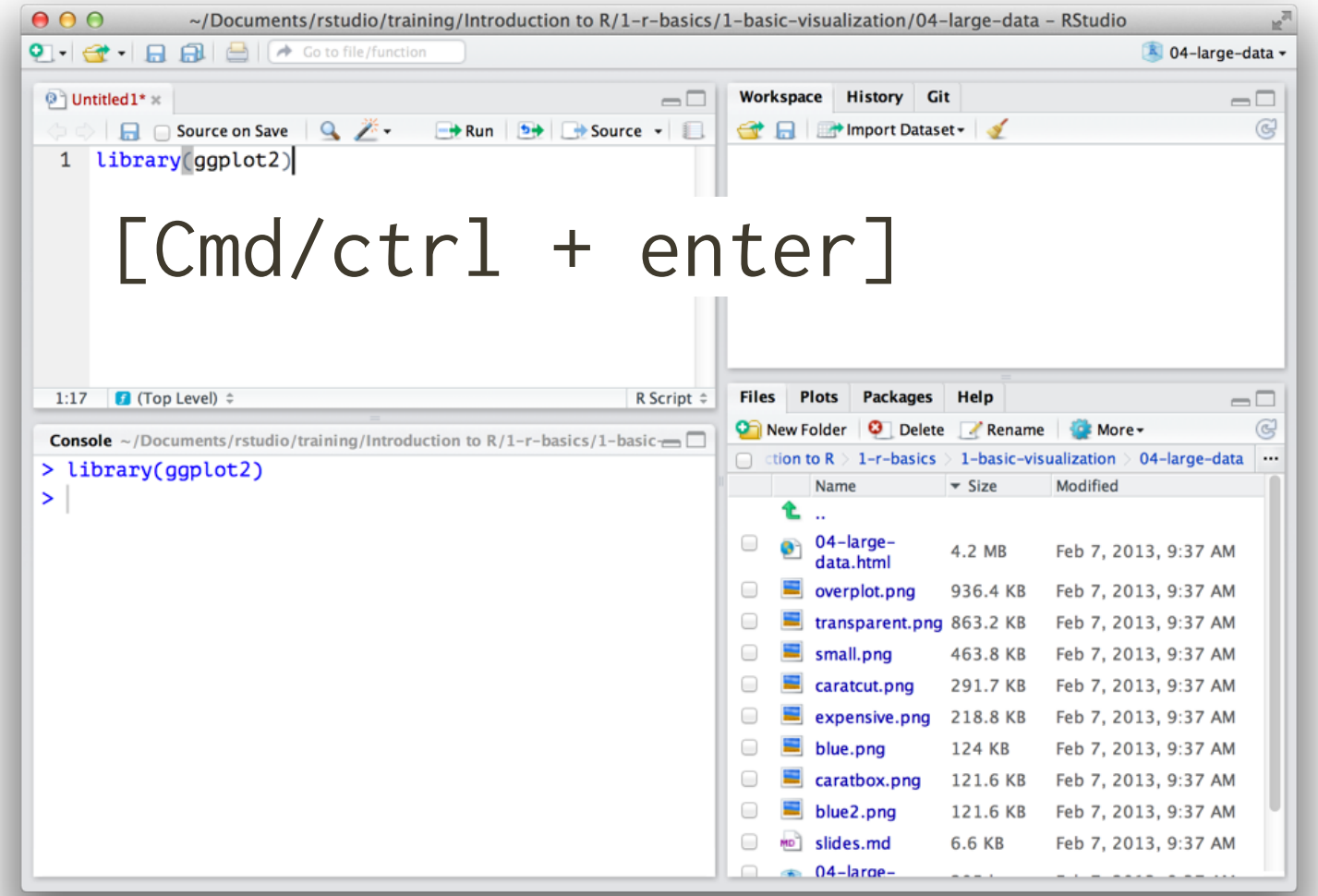
Useful R Keyboard Shortcuts: Autocomplete



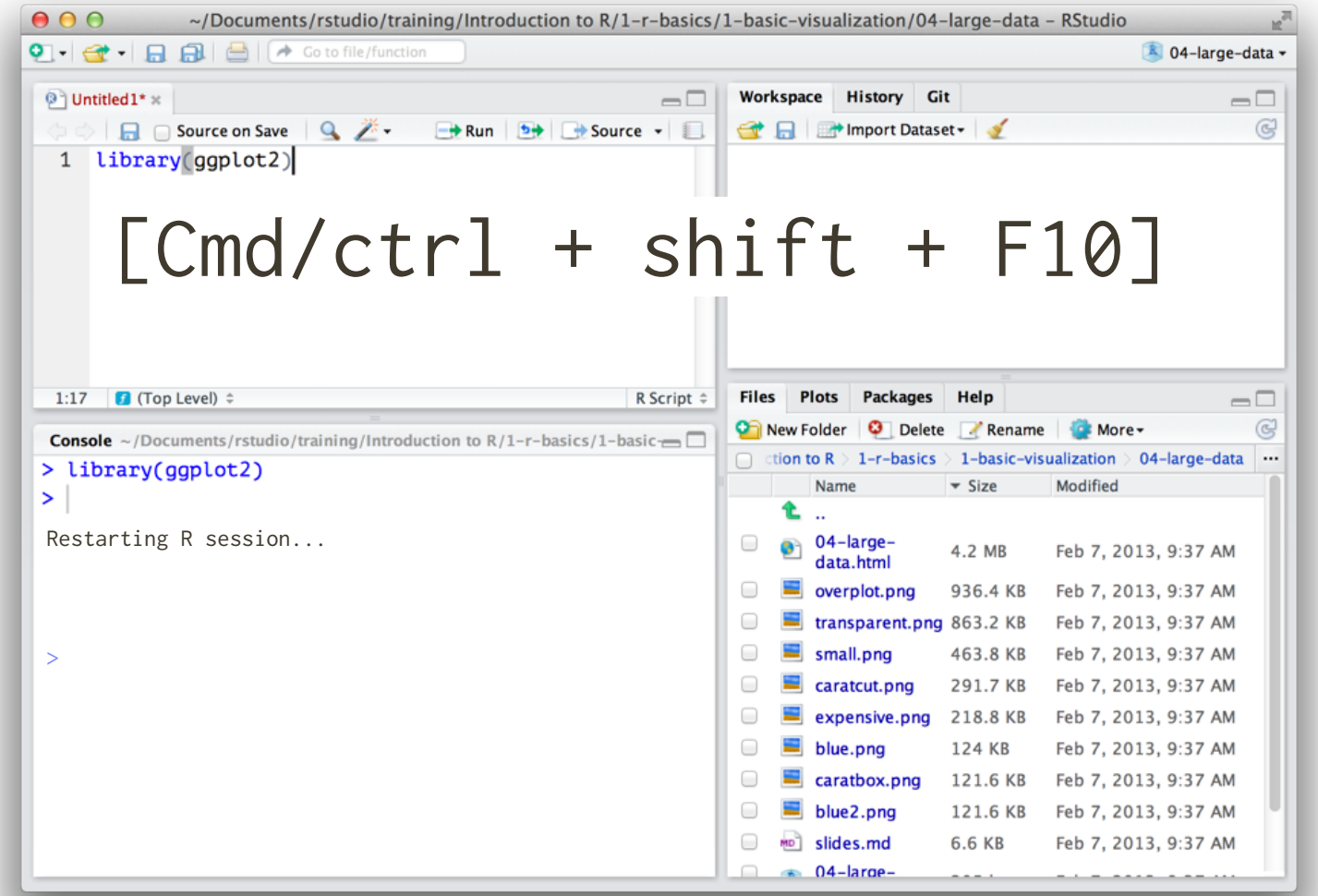
Useful R Keyboard Shortcuts: History



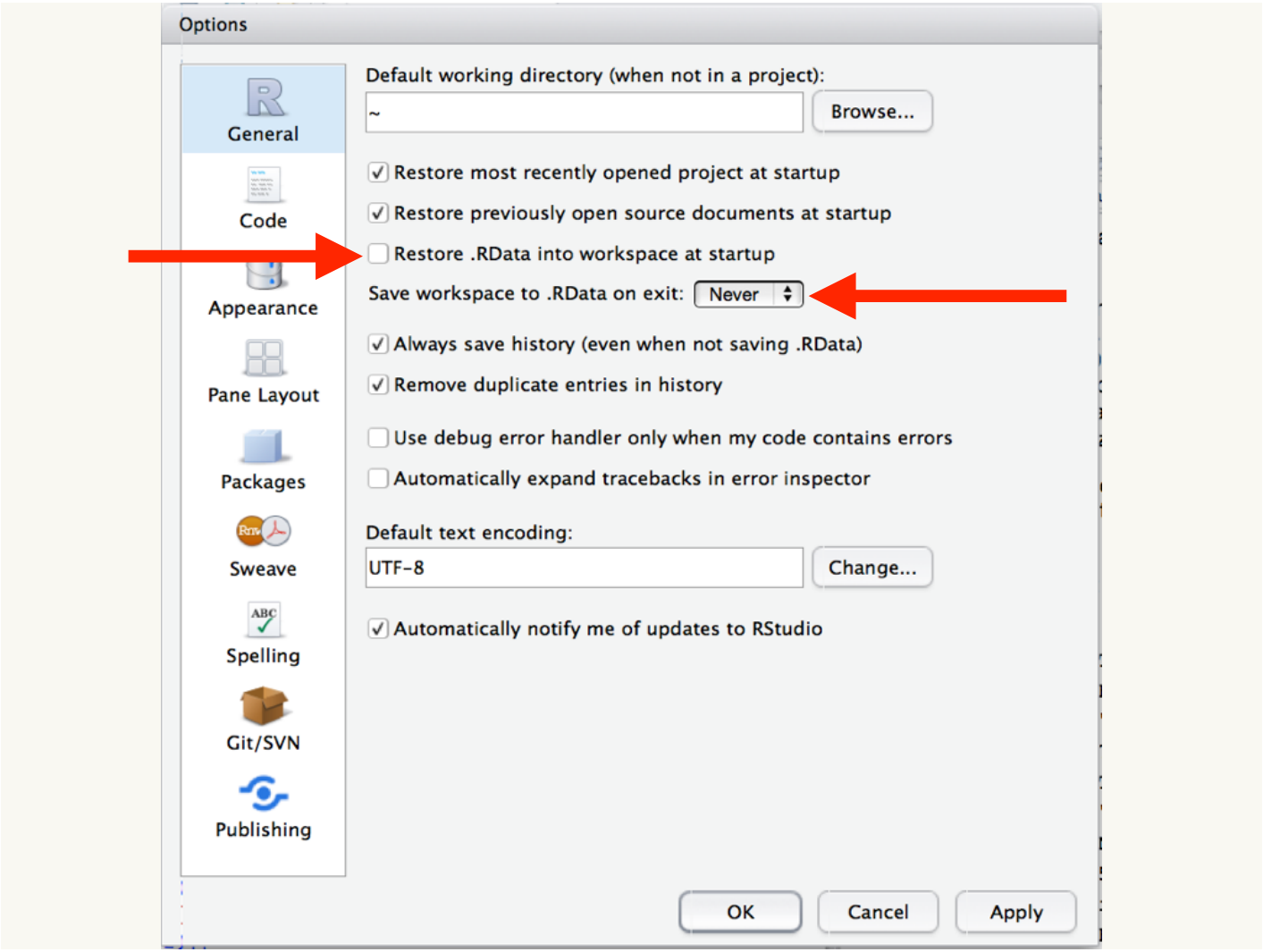
Useful R Keyboard Shortcuts: Execute Code



Useful R Keyboard Shortcuts: Restarting an R Sess



IMPORTANT R Setting



A Brief Digression

- Whenever writing code, you want to be sure to clear your environment to ensure the fidelity of your results
- In each and every `R` script file I write, I always include the following two lines of code

```
rm(list=ls())
cat("\014")
```

1. `rm(list=ls())` removes all objects in the current environment
2. On a Mac, `cat("\014")` clear the console windows (same as `ctrl + l`)

Data Sets in R

- R comes built in with multiple data sets you can play with
- Many (most?) packages also have data sets
- `data()` will bring up a list of all data sets available across all loaded packages
- `help(<nameOfDataSet>)` will provide you a detailed description of the data set in question

How Big is *Big Data* in R?

- R holds data in memory, effectively limiting data to the amount of RAM a computer has access to
- It is not uncommon to work with a data set containing 100,000,000 elements (e.g., 100,000 observations of 1,000 variables or 1,000,000 observations of 100 variables) without difficulty
- Approximations depend on what type of data is contained in each variable, e.g., a data set with 2.2 million records and twenty variables, which takes approximately one minute to load into memory

How Big is *Big Data* in R?

- Also depends on what techniques and/or functions will be applied to the data
- The more complex and memory intensive the task, the smaller the data will be required to be
- Basic plotting requires far less computational exertion than a complex statistical learning model
- **Common Definition**: *Big Data* refers to any data set that cannot be loaded into working memory on your personal computer

Assignment 1

- Read Chapter 1 of Doing Data Science by Cathy O'Neil and Rachel Schutt entitled "What is Data Science?" here. Be ready to answer questions about this in the next class.