DATA SOCIETY™

"If you can't explain it simply, you don't understand it well enough."

- Albert Einstein

Instructor: Dr. Harlan Harris

- Director of Data Science at the Education Advisory Board
- Co-Founder and Co-Organizer of the Data Science DC Meetup
- Co-Founder of Data Community DC, Inc.
- BS in Computer Science from the University of Wisconsin-Madison
- Ph.D. in Computer Science, focusing on Machine Learning and Cognitive Science, from the University of Illinois at Urbana-Champaign

Course syllabus

1. What is Data Science?

2. Programming in R

3. Visualization in R

Setting expectations

Data science takes dedication! You will need to:

1. Take this course ☺
2. Practice
3. Review class material on your own
4. Practice
5. Complete exercises outside of class
6. Practice
7. Share and read latest news

Outline

- What is data science?
- A data scientist’s approach
- Introduction to R
  - Calculations in R
  - Reading data into R
  - Manipulating data in R
- Visualization in R
  - Basic plotting
  - Advanced plotting
  - Building a crime map

What’s going on with data?

- “Every 2 days we create as much information as we did from the dawn of civilization up until 2003” – Eric Schmidt
- Mark Zuckerberg noted that 1 billion pieces of content are shared via Facebook’s Open Graph daily – Facebook earnings call, July 2012
- “1.5 million more data-savvy managers are needed to take full advantage of big data in the United States” – McKinsey & Co.
- A survey reported that more than 37.5% of large organizations said that analyzing big data is their biggest challenge – RainStor, August 2012
- According to Gartner, Big Data will drive $232 billion in spending over the next two years

How is data being used?

Retail	Finance	Marketing	Real estate	Cool
Target:	Kabbage:	Netflix:	Zillow:	Andrew Ng:
The store knows you're pregnant based on what you buy	Makes lending decisions based on Amazon product reviews, etc.	What movie should you watch?	Calculates Zestimate (value of your home)	Machine learning techniques recognize cat faces online using pictures and videos
	amazon reviews	NETFLIX		

Real world applications

- Marketing:
  - How do you classify shoppers who are likely to spend a lot?
  - How do you recommend consumer products based on prior shopping patterns?
  - How do you gauge brand and product perception in real time?
- Healthcare:
  - Do these symptoms suggest a limited possibility of ailments (diagnostics)?
  - Does the patient have a disease based on age, sex, body mass index, results of various blood tests, etc?
  - Detect patterns in spread of disease
- Finance:
  - How do you classify safe vs. unsafe borrowers, what do they have in common?
  - Is this person likely to default on their mortgage?
  - What real estate properties are most similar?
- Politics:
  - Will someone vote Democratic or Republican?
  - Will someone make a political donation?
  - How is a political candidate perceived by a certain demographic in real time?

What is “Big Data”?

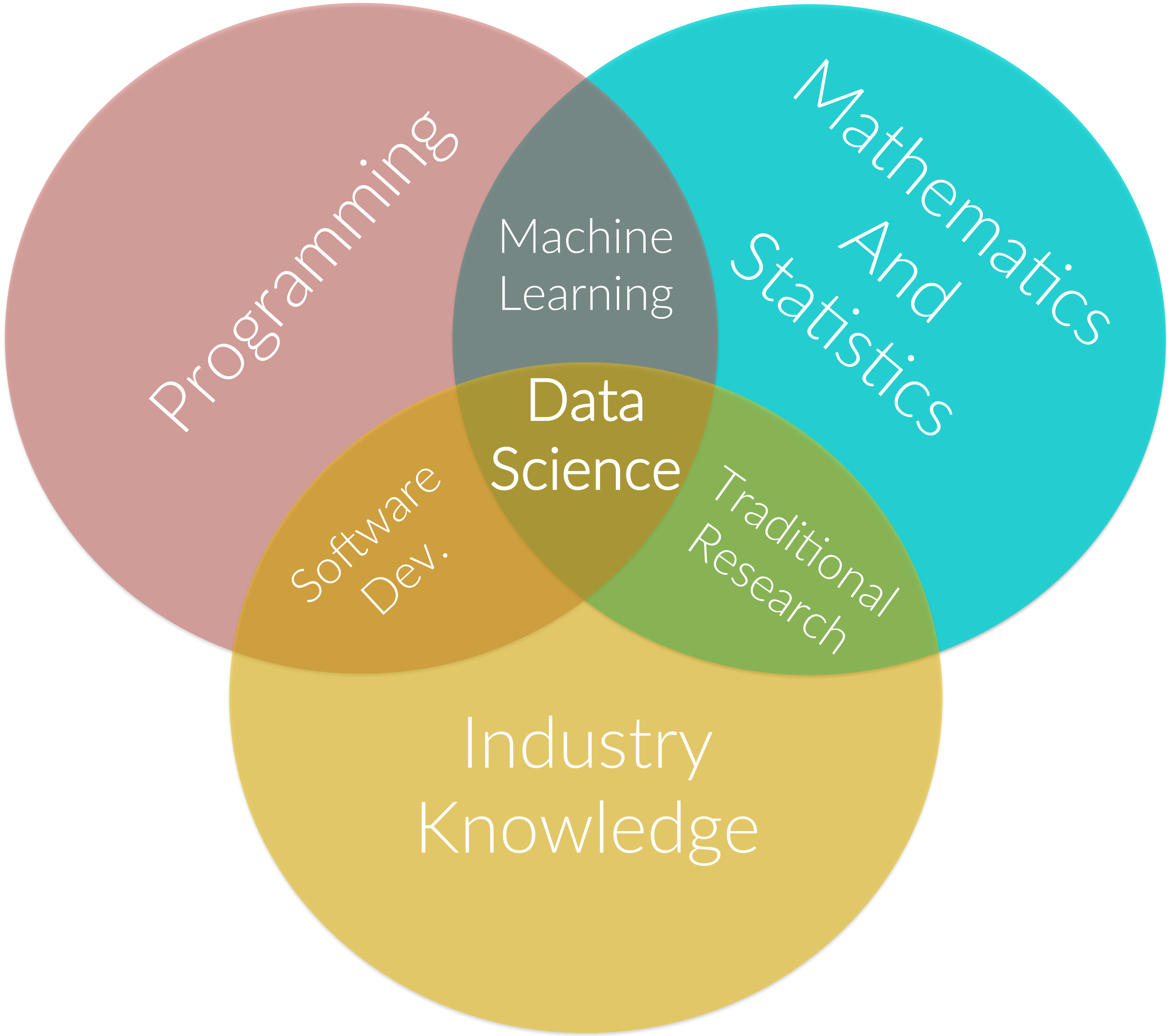
- Big Data is large volumes of information
  - Moving
  - Storing
  - Manipulating
  - Accessing

- It is not:
  - Analysis or insights

That’s why you’re in this class!

What is data science?

- Data science applies the scientific method to analyzing data
- It lies at the intersection of several disciplines
- It draws on domain specific knowledge that makes the analysis of Big Data possible



Who is a data scientist?

- An analyst who can:
  1. Pose the right question
  2. Wrangle the data (gather, clean, and sample data to get a suitable data set)
  3. Manage the data for easy access by the organization
  4. Explore the data to generate a hypothesis
  5. Make predictions using statistical methods such as regression and classification
  6. Communicate the results using visualizations, presentations, and products

Levels of expertise

Data analyst

- Wrangles the data
- Manages the data
- Creates basic analyses and visualizations

Data modeler

- Models to answer specific questions
- Understands the data, its source and structure

Data scientist

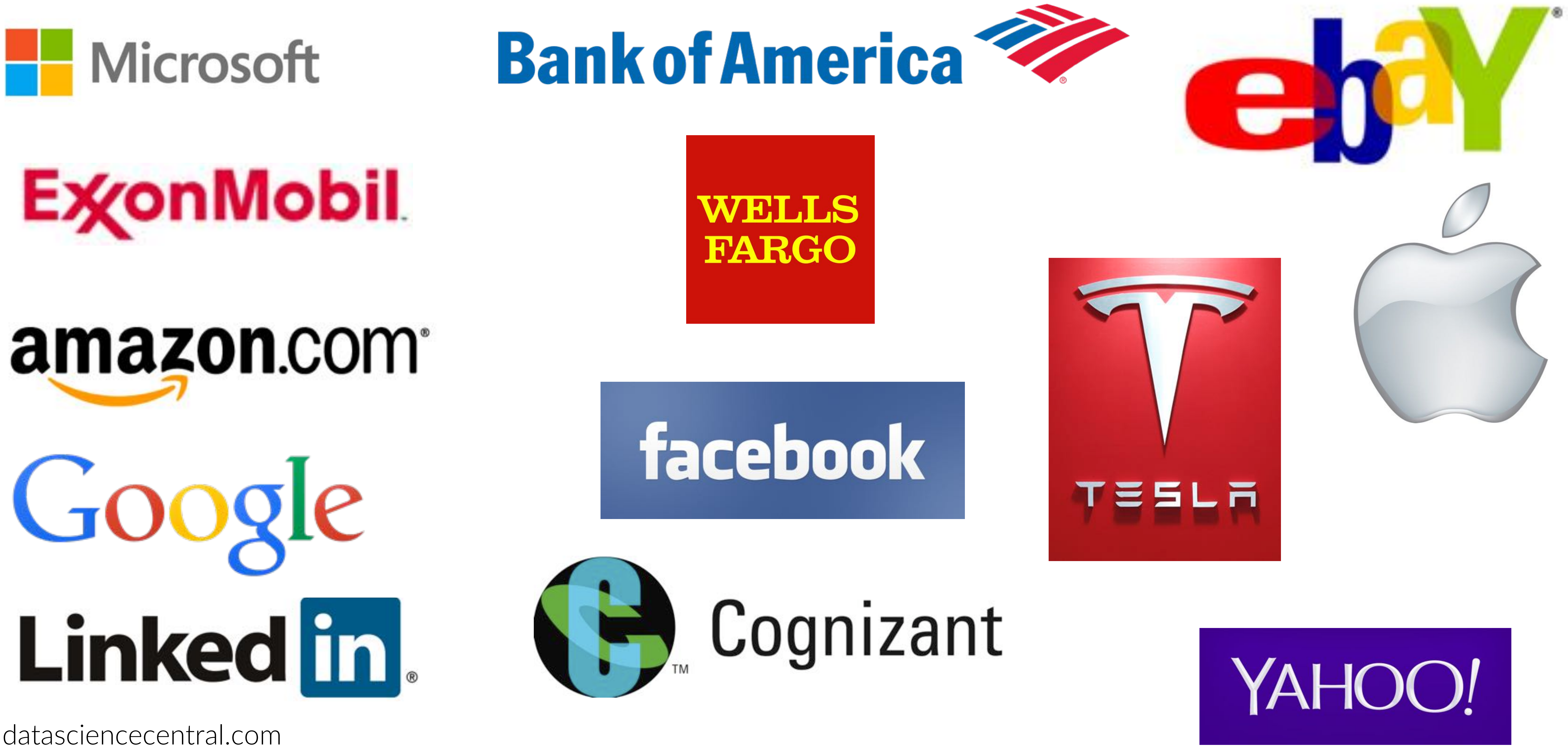
- Asks the right questions
- Looks for patterns in data
- Interprets results critically

Data science job market

Somewhat important ✓ Very important ✔

	A non-data-driven company	The business is just starting to collect data	Data is the product of the company	Company uses data to make decisions
Basic tools	✔	✔	✔	✔
Software engineering		✔	✓	✓
Statistics	✓	✓	✔	✔
Machine learning			✔	✔
Data processing		✔	✓	✔
Data visualization and communication	✔	✓	✓	✔
Thinking like a data scientist	✓	✓	✔	✔

Who hires data scientists?



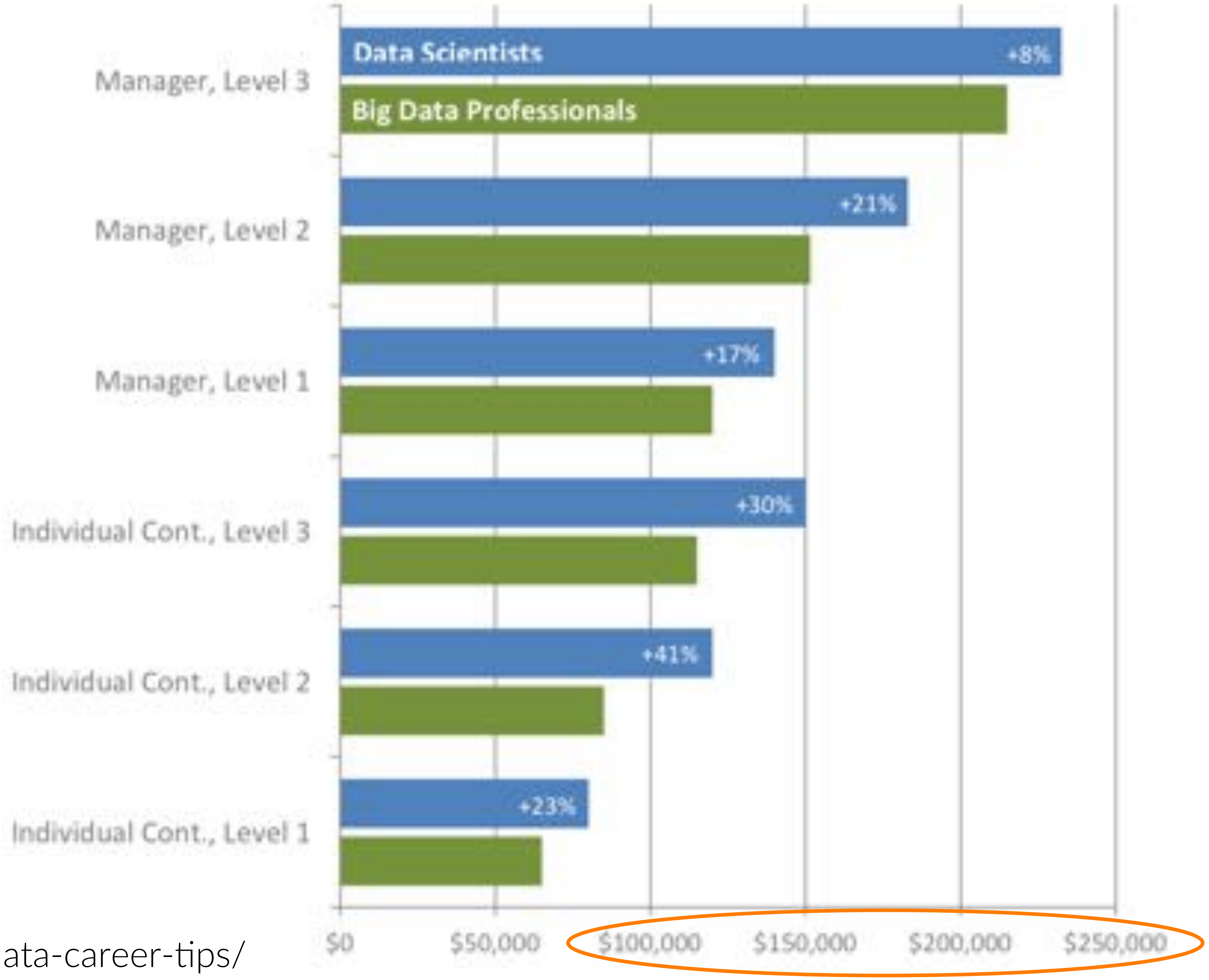
Source: datasciencecentral.com

How much do data scientists make?

- According to a Burtch Works 2014 data science job market survey:

> *"Data scientists earn a median salary that can be up to 40% higher than predictive analytics professionals at the same job level"*

- The graphic on the right provides detail on median salaries by experience level

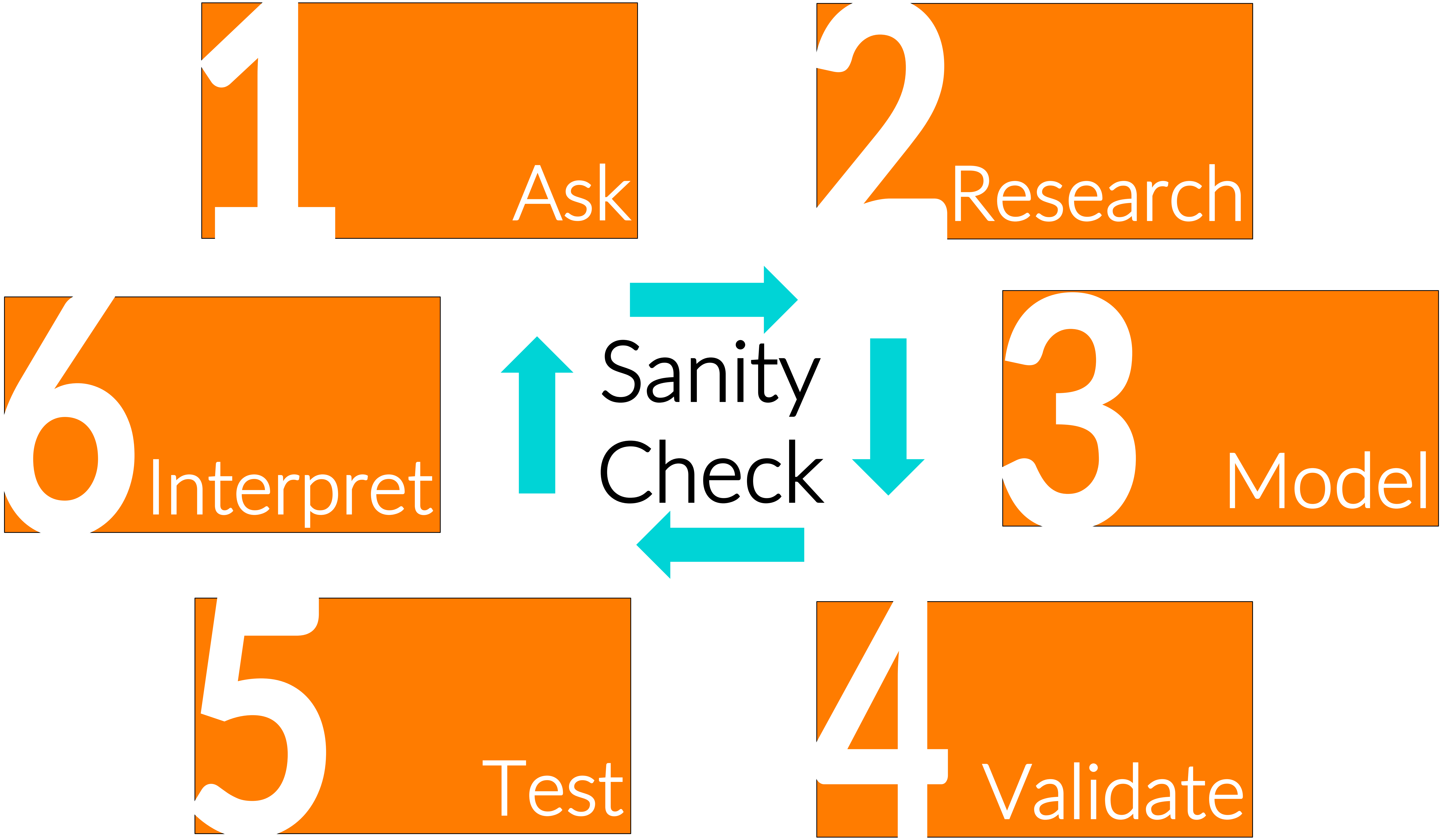


Source: http://www.burtchworks.com/big-data-analyst-salary/big-data-career-tips/

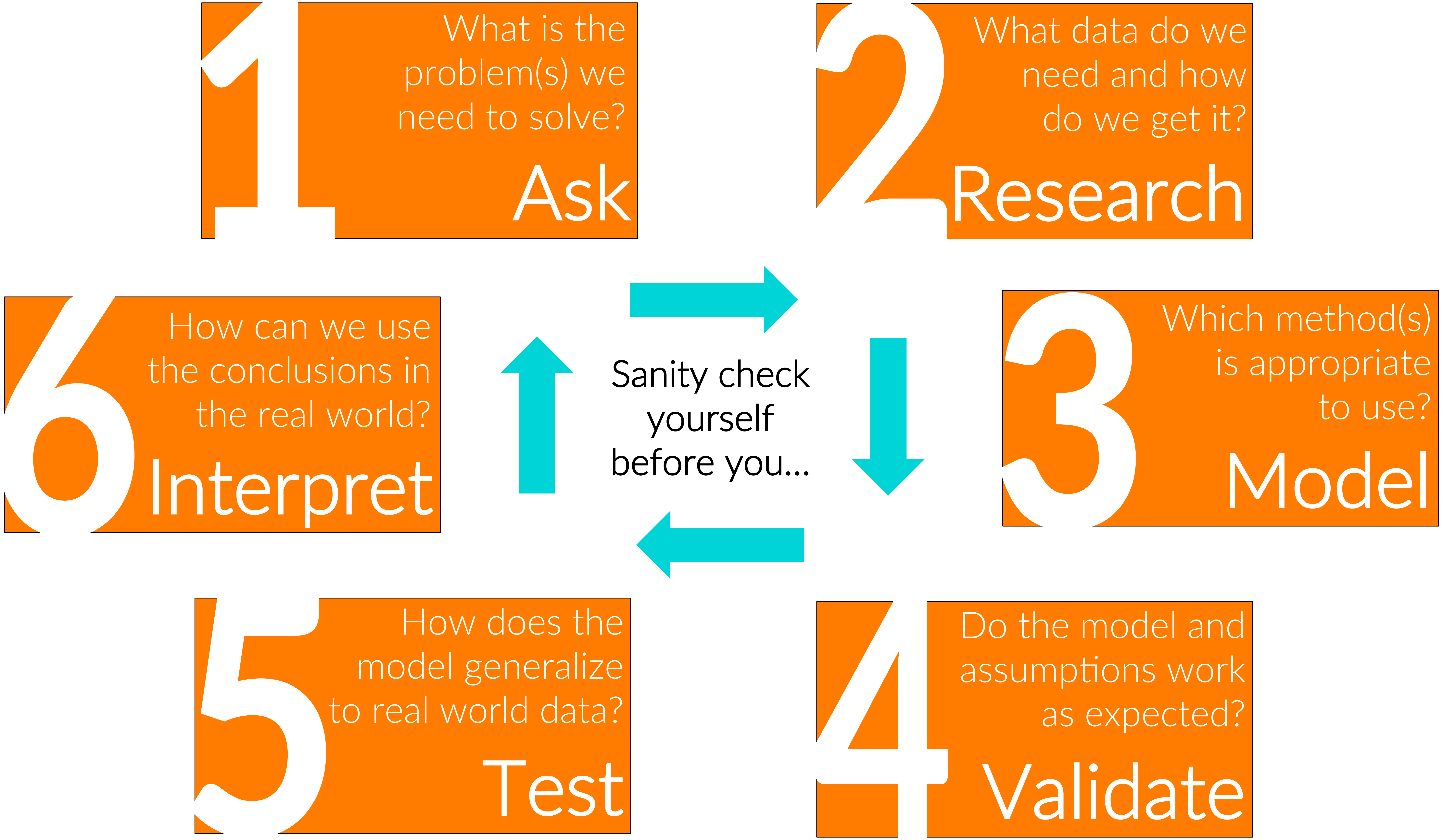
Outline

- What is data science?
- A data scientist’s approach
- Introduction to R
  - Calculations in R
  - Reading data into R
  - Manipulating data in R
- Visualization in R
  - Basic plotting
  - Advanced plotting
  - Building a crime map

Data science control cycle



Data science control cycle



1 What is the problem(s) we need to solve? Ask

2 What data do we need and how do we get it? Research

3 Which method(s) is appropriate to use? Model

4 Do the model and assumptions work as expected? Validate

5 How does the model generalize to real world data? Test

6 How can we use the conclusions in the real world? Interpret

Sanity check yourself before you…

For every job there is a tool

Data aggregation

- Hadoop
- Spark
- SQL
- …

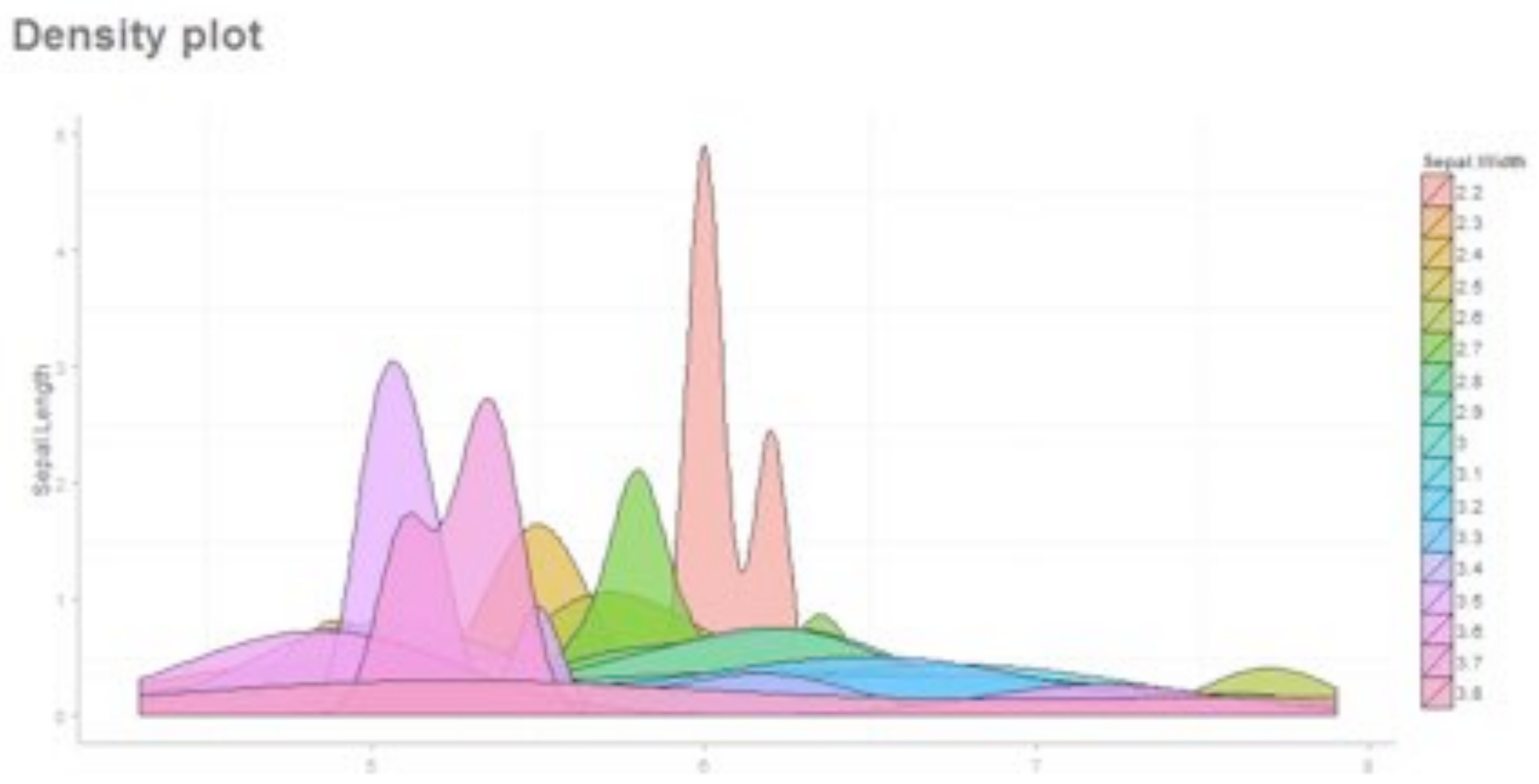
Data analysis

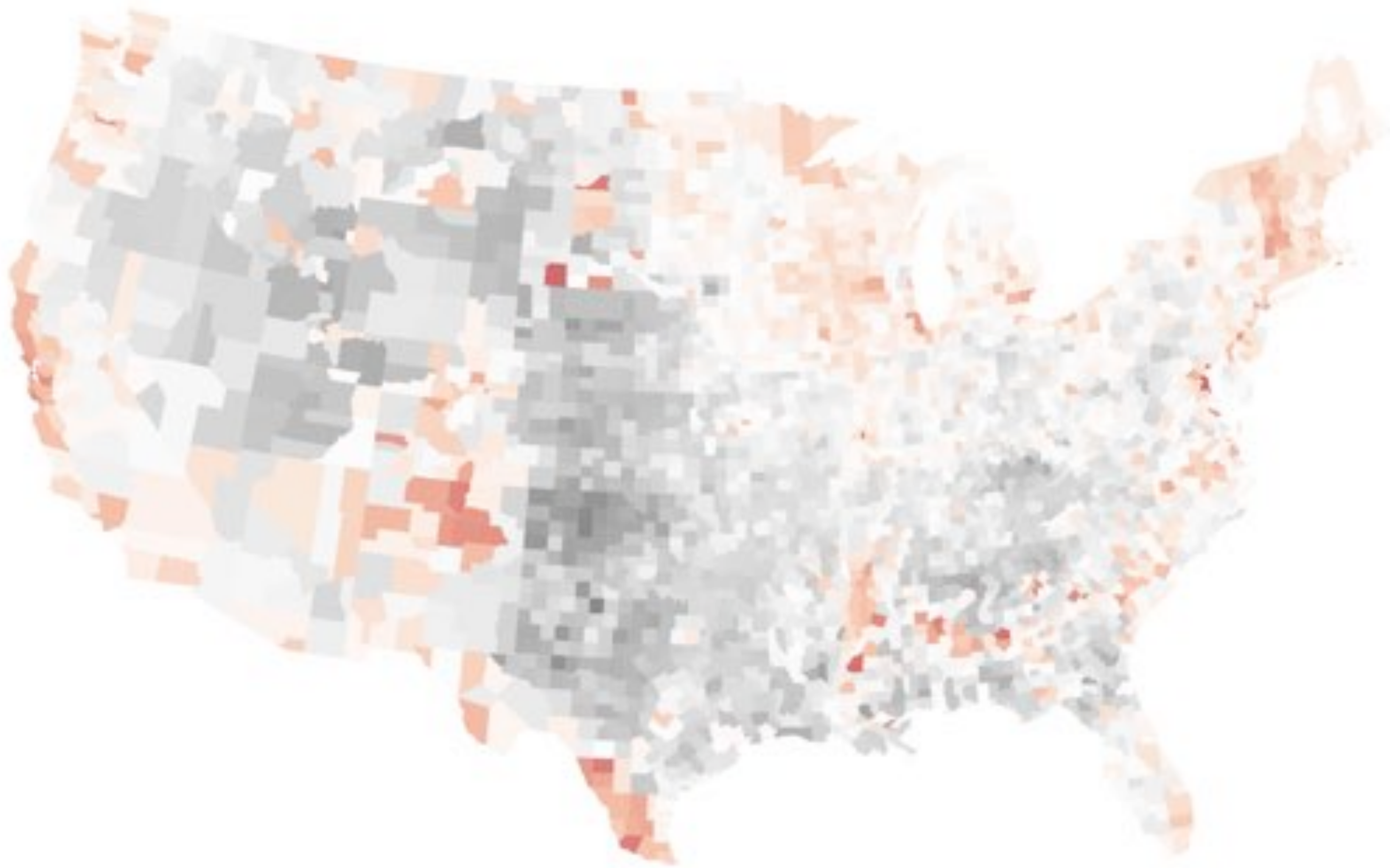
- R
- SAS
- Dell Statistica
- SPSS
- Matlab
- Python
- Google Prediction API
- …

For every job there is a tool

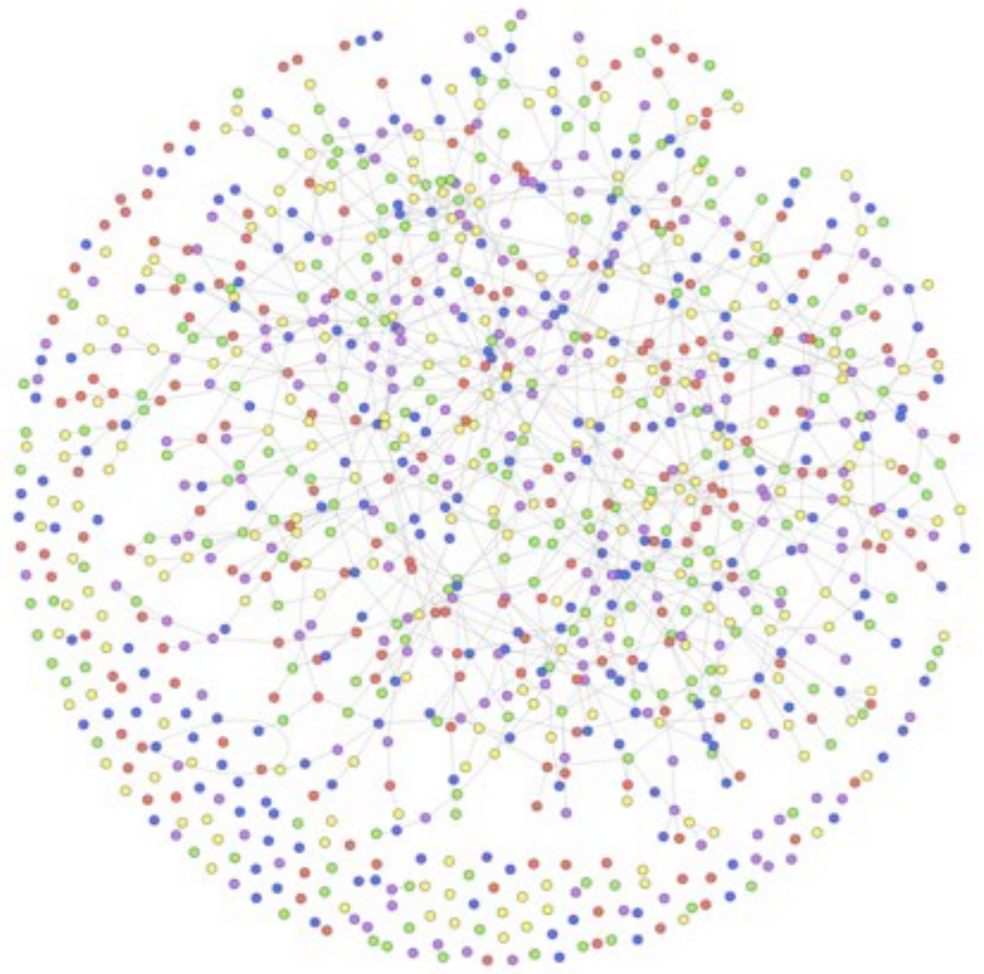
Visualization

- R
- Tableau
- iVEDiX





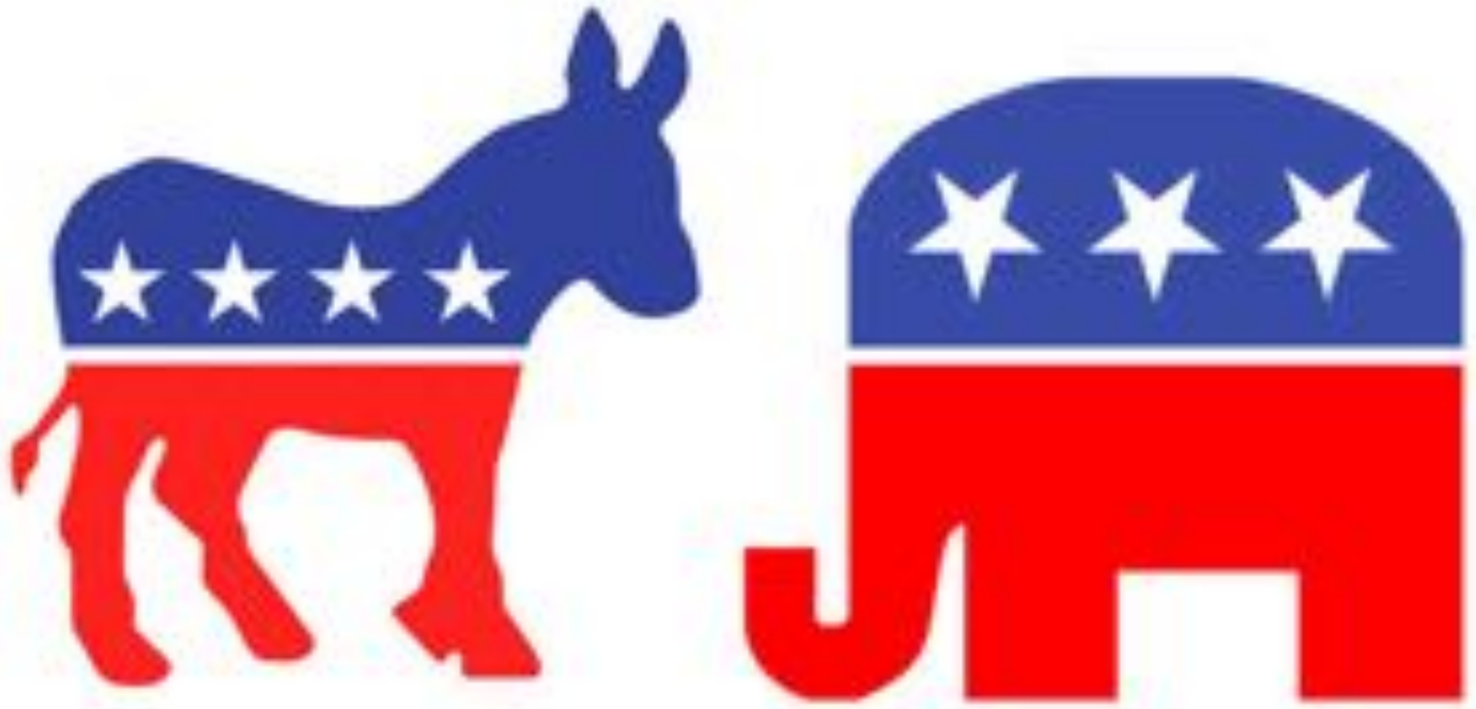
All of these visualizations were created in R



Supervised machine learning

Pattern discovery when inputs (x) and outputs (y) are known

Input x:
Voter



Output y:
Political affiliation

Examples: Classification and regression are supervised machine learning

Unsupervised machine learning

The data inputs *(x)* have no target outputs *(y)*

Input x:
Voter

→ ?

Output y:
Not given
(to be discovered)

We want to impose structure on the inputs *(x)* to say something meaningful about the data

Machine vs. human

	Machine	Human
Understanding context		✓
Thinking through the problem		✓
Asking the right questions		✓
Selecting the right tools		✓
Performing calculations quickly	✓	
Performing repetitive tasks	✓	
Following pre-defined rules	✓	
Interpreting results		✓

Outline

- What is data science?
- A data scientist’s approach
- Introduction to R
  - Calculations in R
  - Reading data into R
  - Manipulating data in R
- Visualization in R
  - Basic plotting
  - Advanced plotting
  - Building a crime map

Why use R

1. De facto standard among professional statisticians
2. Comparable and often superior in power to commercial products (SAS, SPSS)
3. Available for the Windows, Mac, and Linux operating systems
4. R is a general-purpose programming language, so you can use it to automate analyses
5. Create dynamic graphics and visualization
6. Large community of users, many are prominent scientists: www.r-bloggers.com
7. Pre-made packages to run data analyses contributed by user base (over 6,500 packages)

Source: http://cran.r-project.org/web/packages/

Uses of R

1. Can be used to analyze and visualize data

2. Can be used to write software

3. Can be used to create data products and applications

In this course, we will focus on how to analyze and visualize data

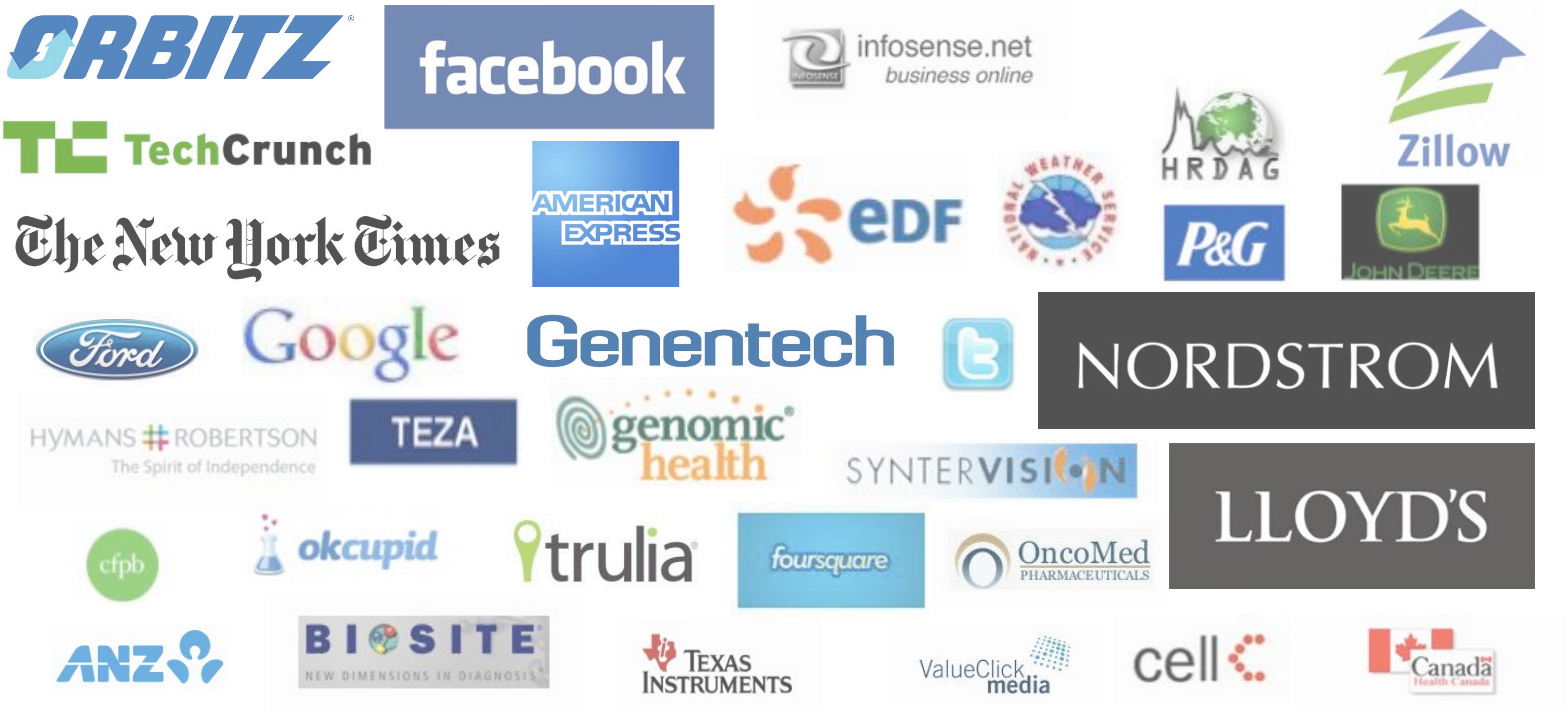
THE TWO STATES OF EVERY PROGRAMMER
I AM A GOD.
I HAVE NO IDEA WHAT I'M DOING.

Data formats R can read

- Can work with many types of data

Microsoft® SQL Server®
Excel
unstructured
information
data
{JSON}
XML
<HTML>
sas®

Companies that use R



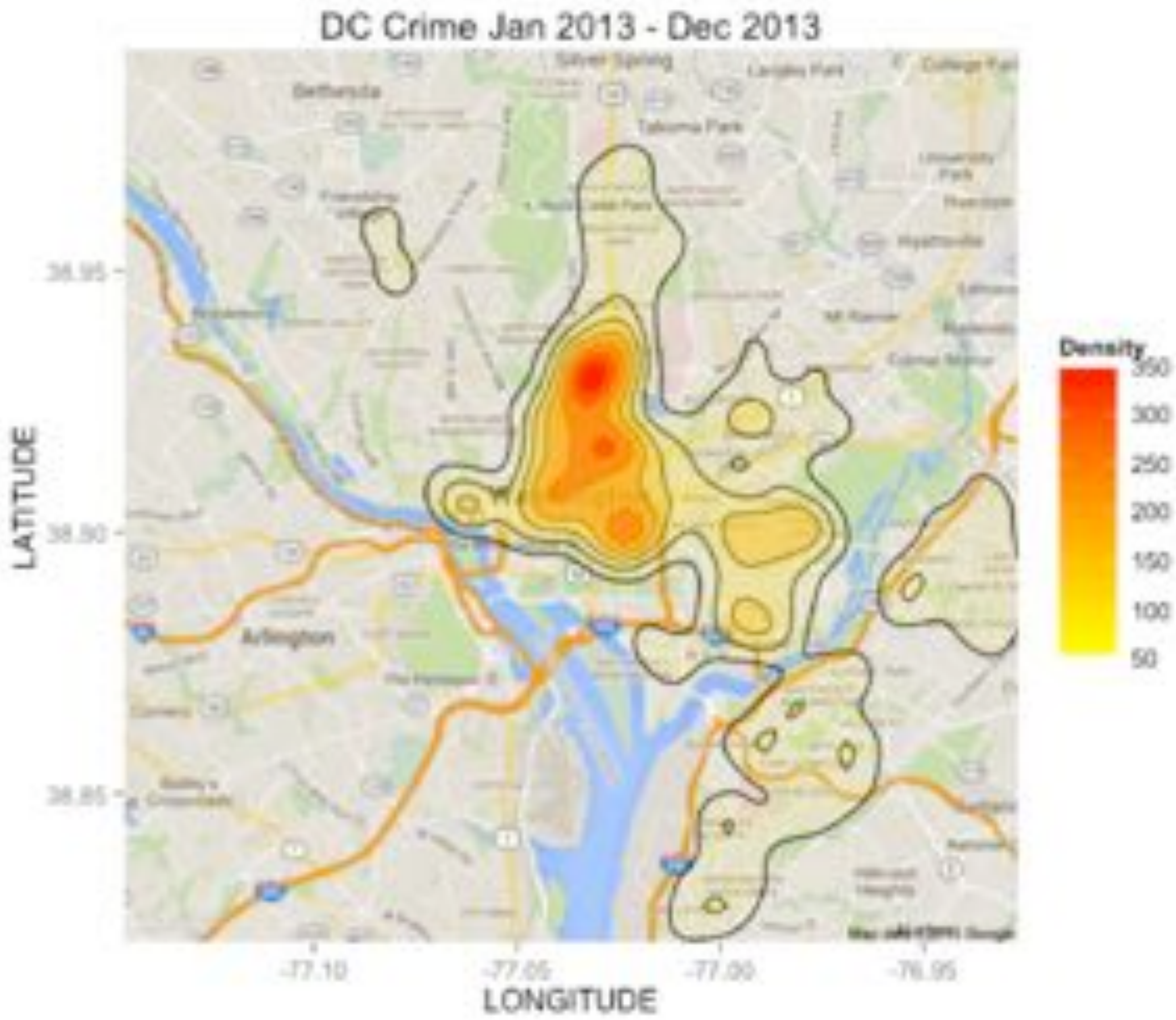
R vs. Excel

	R	Excel
Data capacity	R can read files as big as several gigabytes and trillions of data points; only limitation is your RAM	Excel can't read more than 1,048,576 rows and 16,384 columns (2011 version), files over ~300 megabytes can be very slow to work with
Customization	Can create custom visualizations through code, very flexible	Drop down menus limit ability to manipulate charts and graphs
Analyzing data	Powerful, pre-built packages that speed up work flow	Less flexible built-in analytic abilities that can be augmented by macros
Modeling	Data analysis and statistical models	Complex financial and accounting models
Seeing data	Built-in spreadsheet viewer	Easy to use spreadsheet interface
Usability	Direct commands similar to Excel "if-statements"	Keyboard shortcuts and slower point-and-click functionality

Visualizations in R

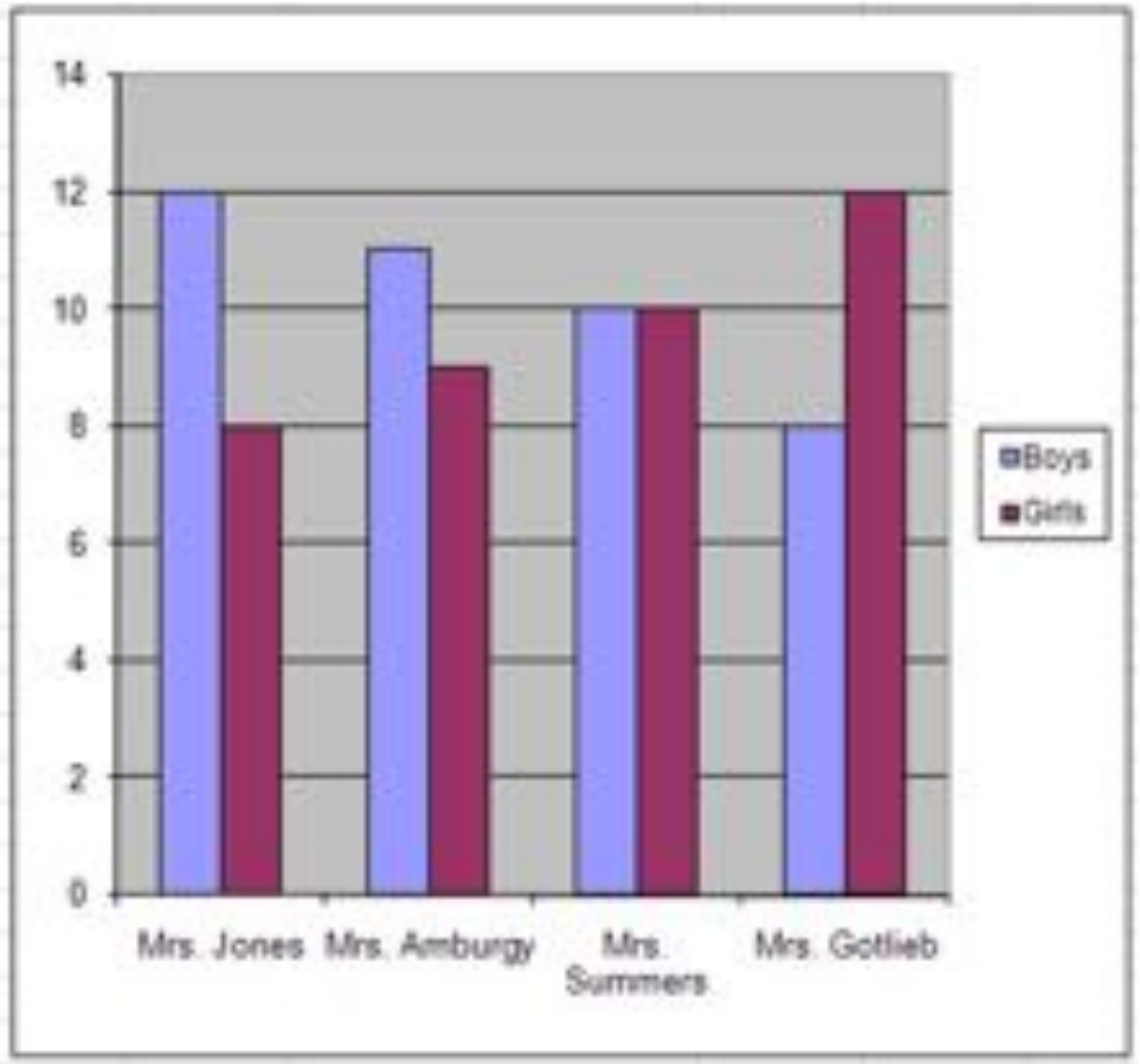
R

Simple customizable code: flexible



Excel

Drag and drop: rigid

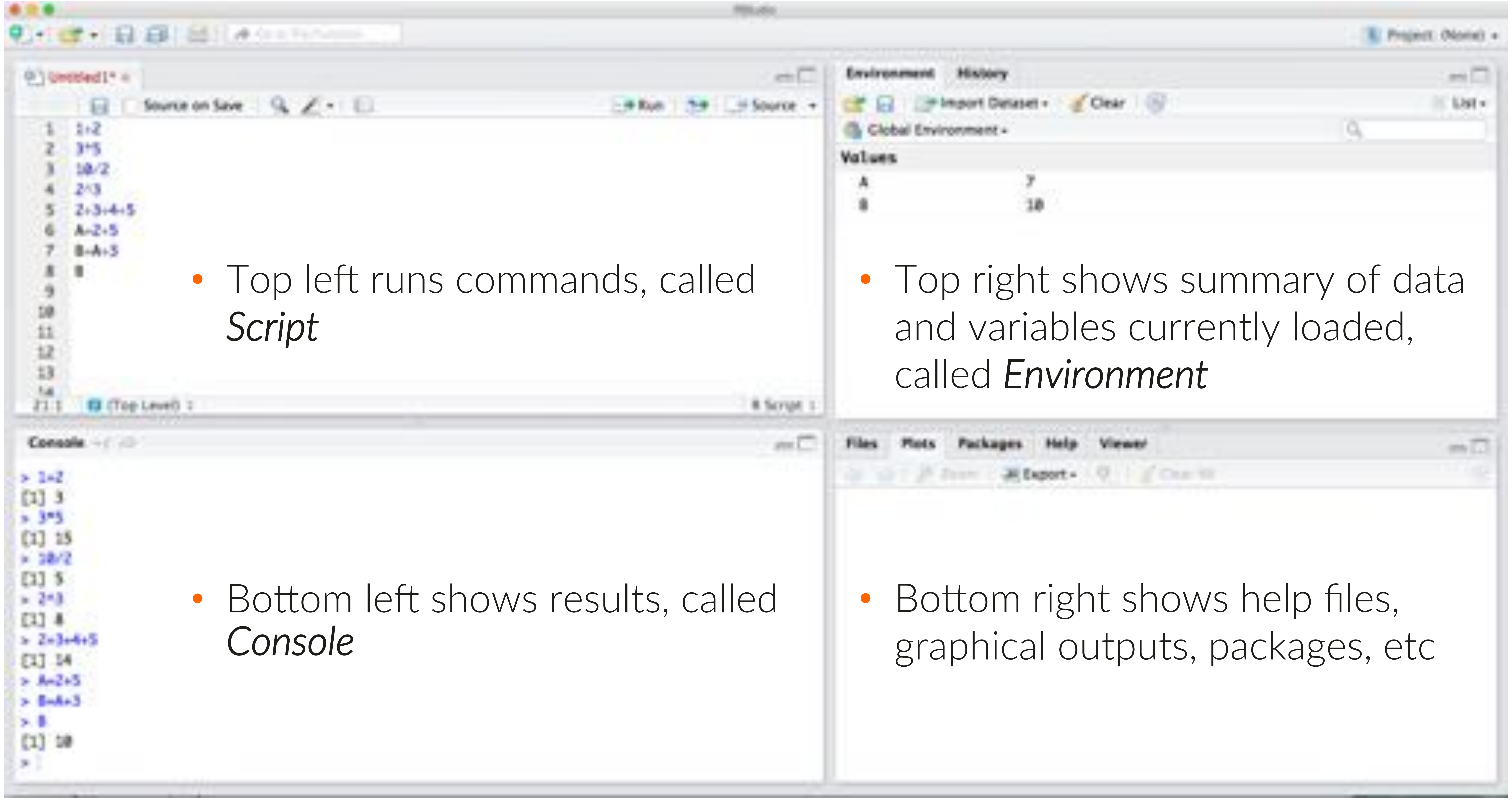


R vs. Python

- R has more convenient statistical packages to analyze data than Python
  - More than any other software tool, over 6,500 as of April 2015
- R is easier to learn for non-programmers than Python, less code is required to perform tasks
- Python is used by many data scientists to build data products (they also tend to be computer scientists)
- Python can be easier to integrate into web applications

Source: http://cran.r-project.org/web/packages/

RStudio overview



```
1+2
3*5
10/2
2^3
2+3+4+5
A=2+5
B=A+3
B
```

```
Values
  A                7
  B                10
```

```
> 1+2
[1] 3
> 3*5
[1] 15
> 10/2
[1] 5
> 2^3
[1] 8
> 2+3+4+5
[1] 14
> A=2+5
> B=A+3
> B
[1] 10
>
```

- Top left runs commands, called *Script*
- Top right shows summary of data and variables currently loaded, called *Environment*
- Bottom left shows results, called *Console*
- Bottom right shows help files, graphical outputs, packages, etc

Working with R: Comments

- Hashmarks are used to add comments and annotate your code

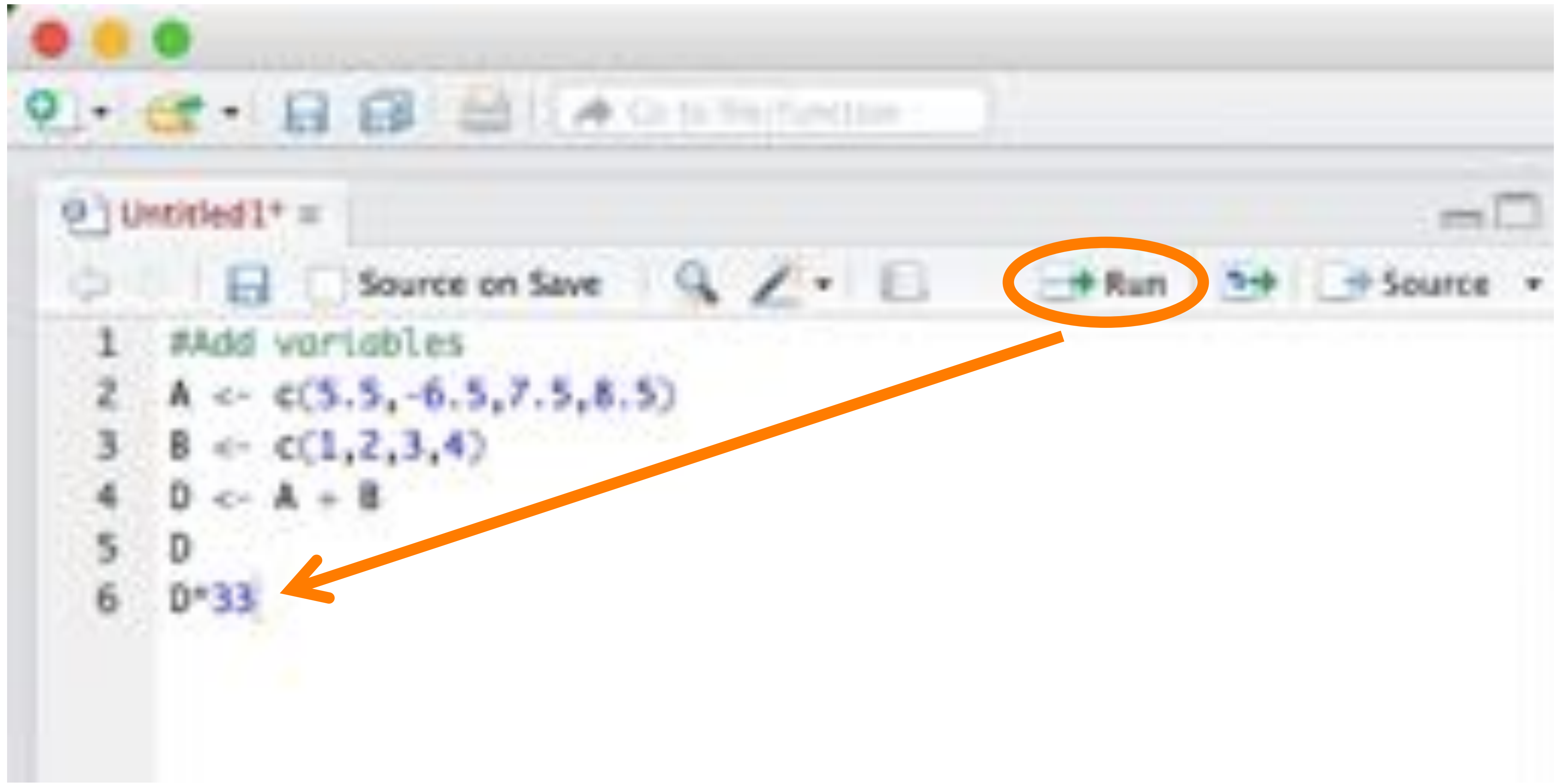
Script

```
# Comments need to start with a hashmark, but don’t need to end with one

# Hashmarks show up in green and are included to explain your code
```

- It’s good practice to annotate your code
    - You can go back later and understand what you were doing

Executing commands in R

- Code is executed when you press "Run" in the top right hand corner of the script window
- R runs the line of code where your cursor is located
- You can also highlight multiple lines to run at once



```
#Add variables
A <- c(5.5,-6.5,7.5,8.5)
B <- c(1,2,3,4)
D <- A + B
D
D*33
```

Note: R is case sensitive

Outline

- What is data science?
- A data scientist's approach
- Introduction to R
  - Calculations in R
  - Reading data into R
  - Manipulating data in R
- Visualization in R
  - Basic plotting
  - Advanced plotting
  - Building a crime map

Working with R: variables

- A series of numbers (think columns in Excel) can be defined using the arrow (**<-**) or equals (**=**) sign

Script

```
# Define variables with arrow
A <- c(5.5, -6.5, 7.5, 8.5)
B <- c(1, 2, 3, 4)
```

or

Script

```
# Define variables with equals sign
A = c(5.5, -6.5, 7.5, 8.5)
B = c(1, 2, 3, 4)
```

- The command `c( )` stands for "concatenate" (join) a series of numbers

Basic operations in R

Adding

- Just use + sign

Script

```
# Add variables
 A = c(5.5, -6.5, 7.5, 8.5)
 B = c(1, 2, 3, 4)
 D = A + B
 D
```

Console

```
> D
 [1] 6.5  -4.5  10.5  12.5
```

Multiplying

- Just use * sign

Script

```
# Multiply variables
 E = D*33
 E
```

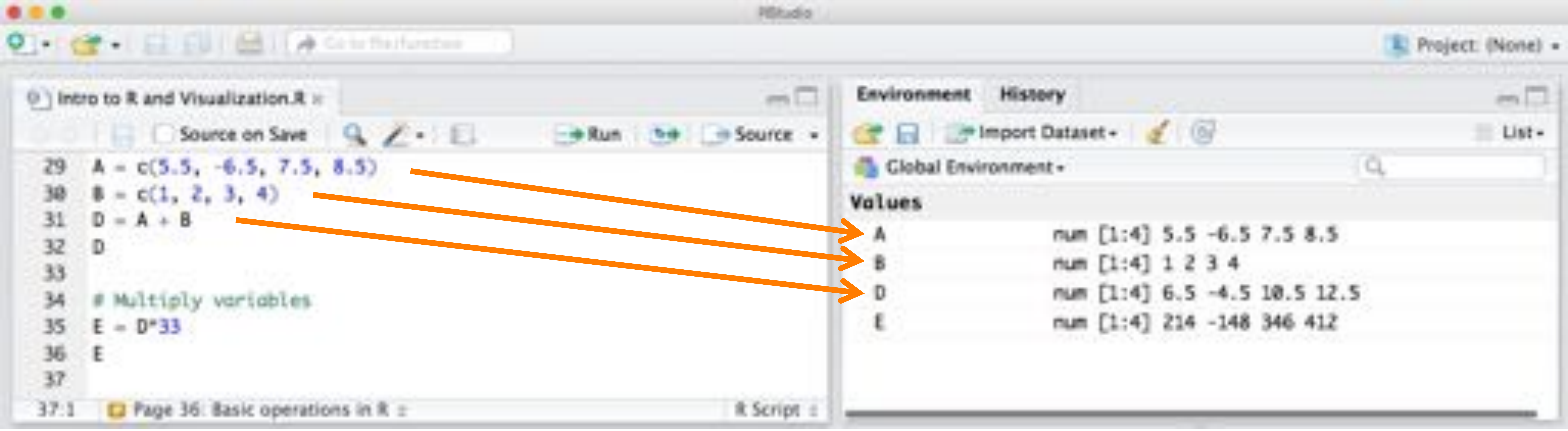
Console

```
> E
 [1] 214.5  -148.5  346.5  412.5
```

Enter formulas in top left window (script)
Output is shown in bottom left window (console)

Working with R: variables

- When a variable is named (instantiated), R stores it in its “environment” and can use it for subsequent operations

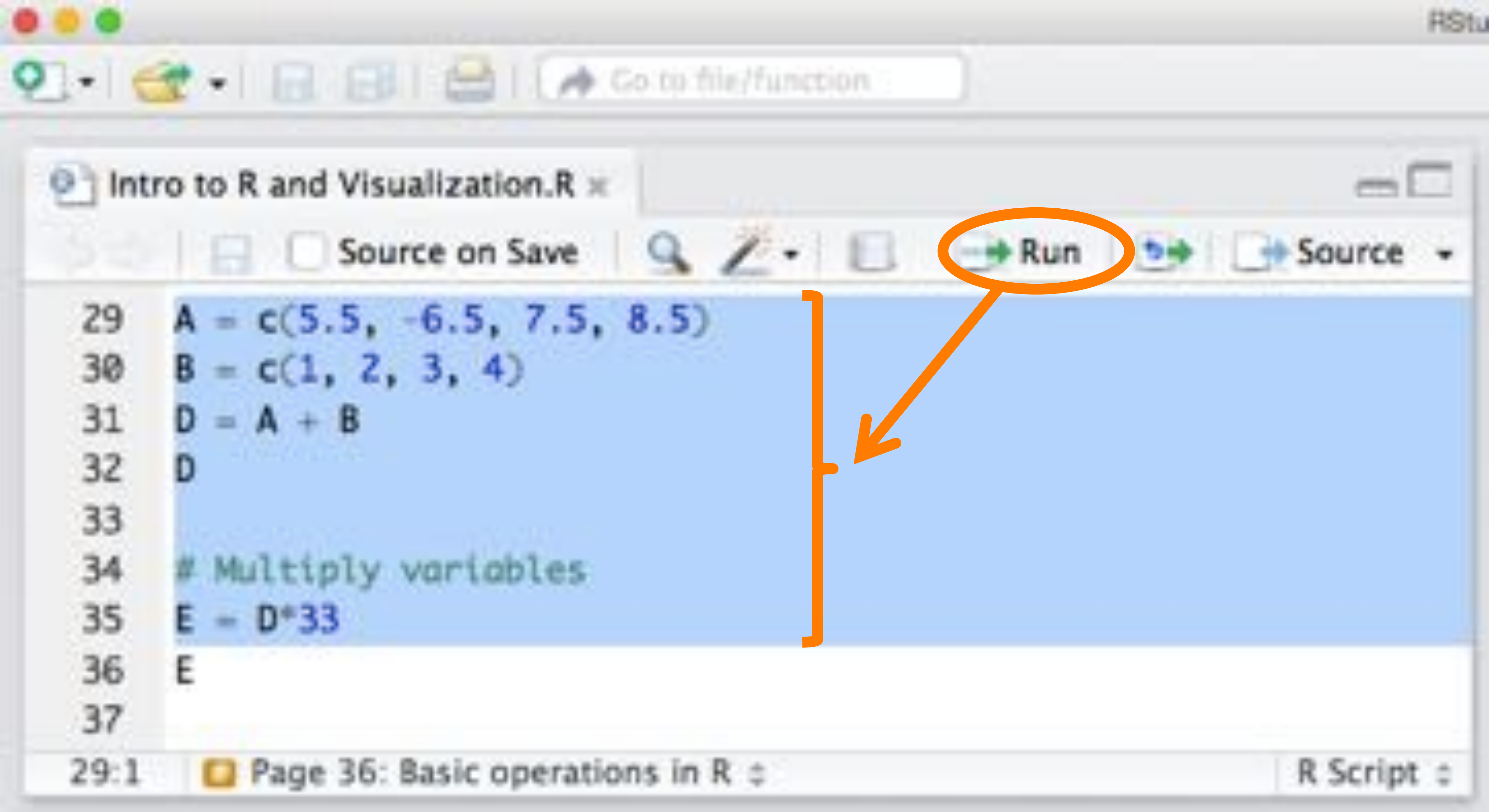


```
A = c(5.5, -6.5, 7.5, 8.5)
B = c(1, 2, 3, 4)
D = A + B
D

# Multiply variables
E = D*33
E
```

```
Values
A                 num [1:4] 5.5 -6.5 7.5 8.5
B                 num [1:4] 1 2 3 4
D                 num [1:4] 6.5 -4.5 10.5 12.5
E                 num [1:4] 214 -148 346 412
```

R can run several lines of code

- You can highlight several lines of code and press “Run” to execute all of them
- Highlighting can be done either with the mouse or by holding “Shift” and using the arrow keys
- You can execute a command by pressing “Ctrl” + “Enter” for PCs or “Command” + “Enter” for Macs

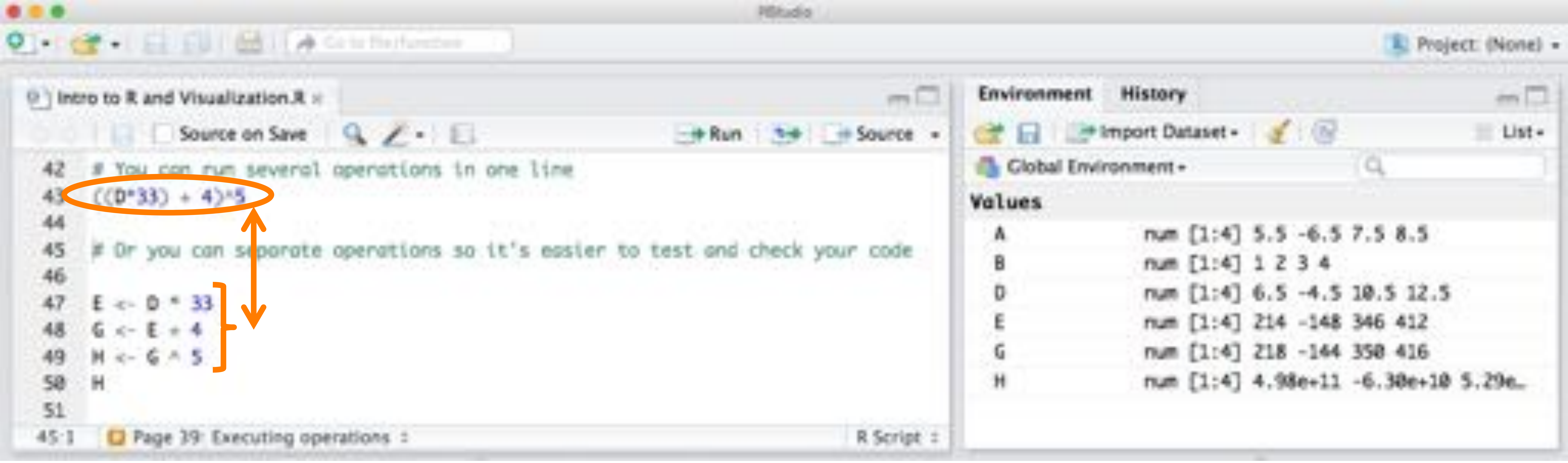

RStu
Go to file/function
Intro to R and Visualization.R
Source on Save
Run
Source
29:1
Page 36: Basic operations in R
R Script

```
A = c(5.5, -6.5, 7.5, 8.5)
B = c(1, 2, 3, 4)
D = A + B
D

# Multiply variables
E = D*33
E
```

Troubleshooting: if you have trouble with this, try restarting R, restarting your computer, or reinstalling R

Executing operations

- You can run several operations in 1 line of code
- Or you can separate steps and instantiate new variables to check your code more easily



```
# You can run several operations in one line
((D*33) + 4)^5

# Or you can separate operations so it's easier to test and check your code

E <- D * 33
G <- E + 4
H <- G ^ 5
H
```

```
Values
  A          num [1:4] 5.5 -6.5 7.5 8.5
  B          num [1:4] 1 2 3 4
  D          num [1:4] 6.5 -4.5 10.5 12.5
  E          num [1:4] 214 -148 346 412
  G          num [1:4] 218 -144 350 416
  H          num [1:4] 4.98e+11 -6.30e+10 5.29e…
```

Outline

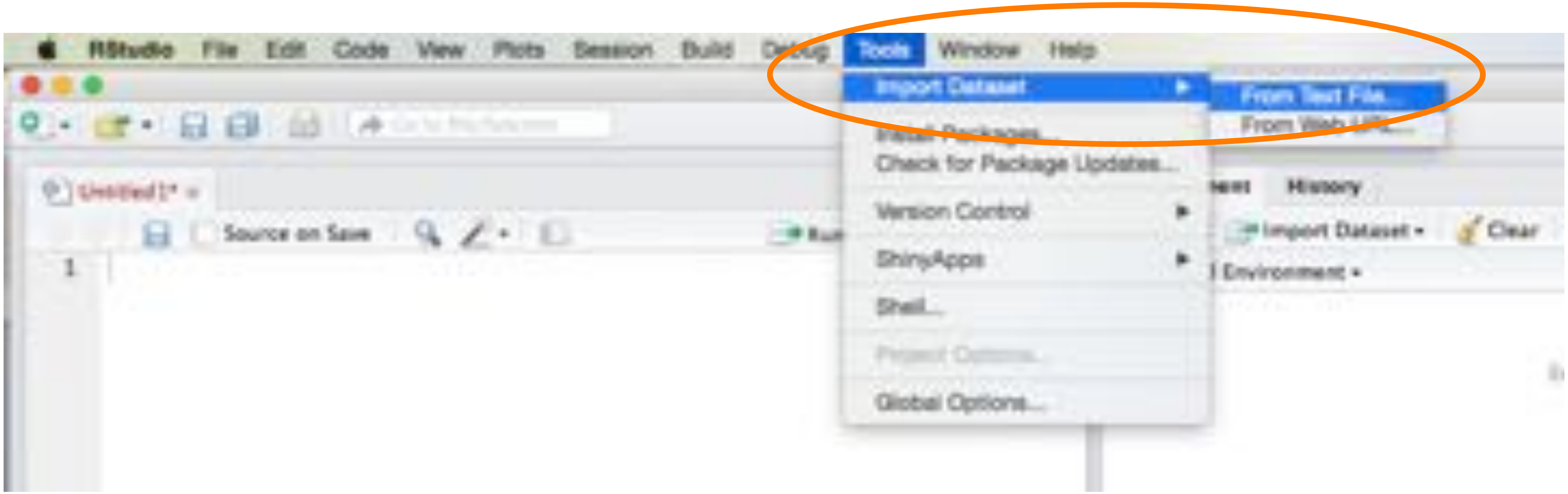
- What is data science?
- A data scientist’s approach
- Introduction to R
  - Calculations in R
  - Reading data into R
  - Manipulating data in R
- Visualization in R
  - Basic plotting
  - Advanced plotting
  - Building a crime map

A note about data

- Data can be found on a variety of sites on the internet
- Processing data stored in different formats is covered in a separate course
- For the purposes of this course, we will provide all the data sets already cleaned

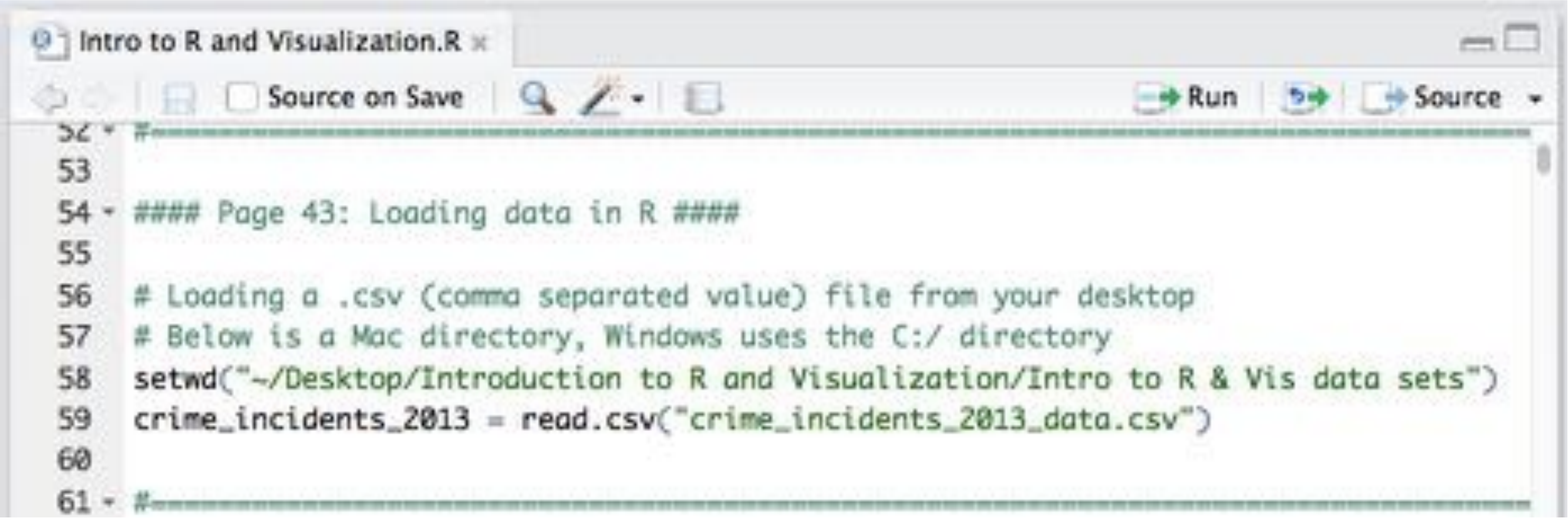
Loading data in R

- Loading data from your computer
  - Point and click



Loading data in R

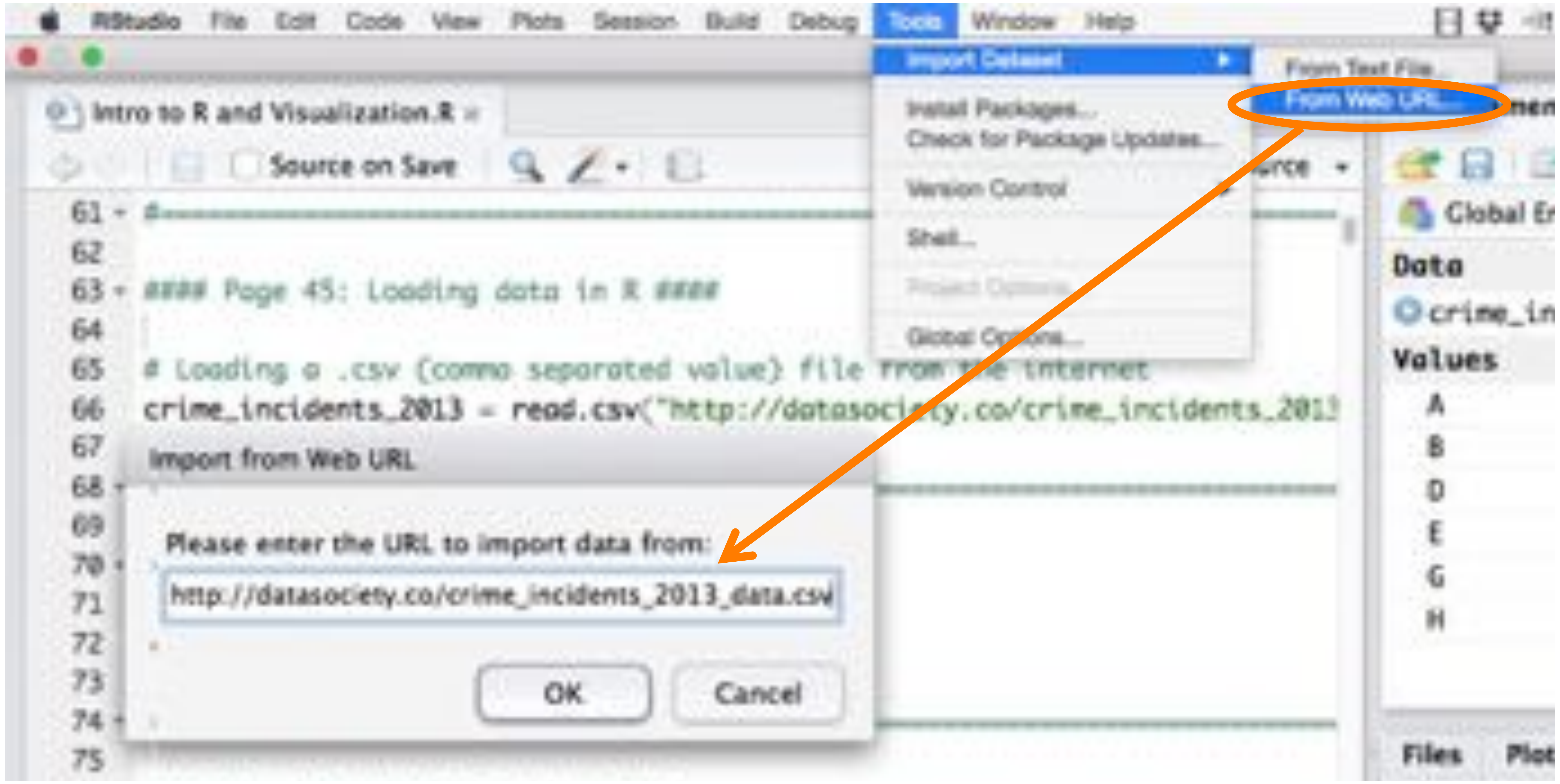
- Loading data from your computer
  - Enter code into script window
  - `crime_incidents_2013` is instantiated as the label of the data set



```
#---------------------------------------------------------------------------

#### Page 43: Loading data in R ####

# Loading a .csv (comma separated value) file from your desktop
# Below is a Mac directory, Windows uses the C:/ directory
setwd("~/Desktop/Introduction to R and Visualization/Intro to R & Vis data sets")
crime_incidents_2013 = read.csv("crime_incidents_2013_data.csv")

#---------------------------------------------------------------------------
```

Loading data in R

- Loading data from the internet
  - Point and click
  - `crime_incidents_2013` is instantiated as the label of the data set



```
#----------------------------------------------------------------

#### Page 45: Loading data in R ####

# Loading a .csv (comma separated value) file from the internet
crime_incidents_2013 = read.csv("http://datasociety.co/crime_incidents_201
```

Import from Web URL

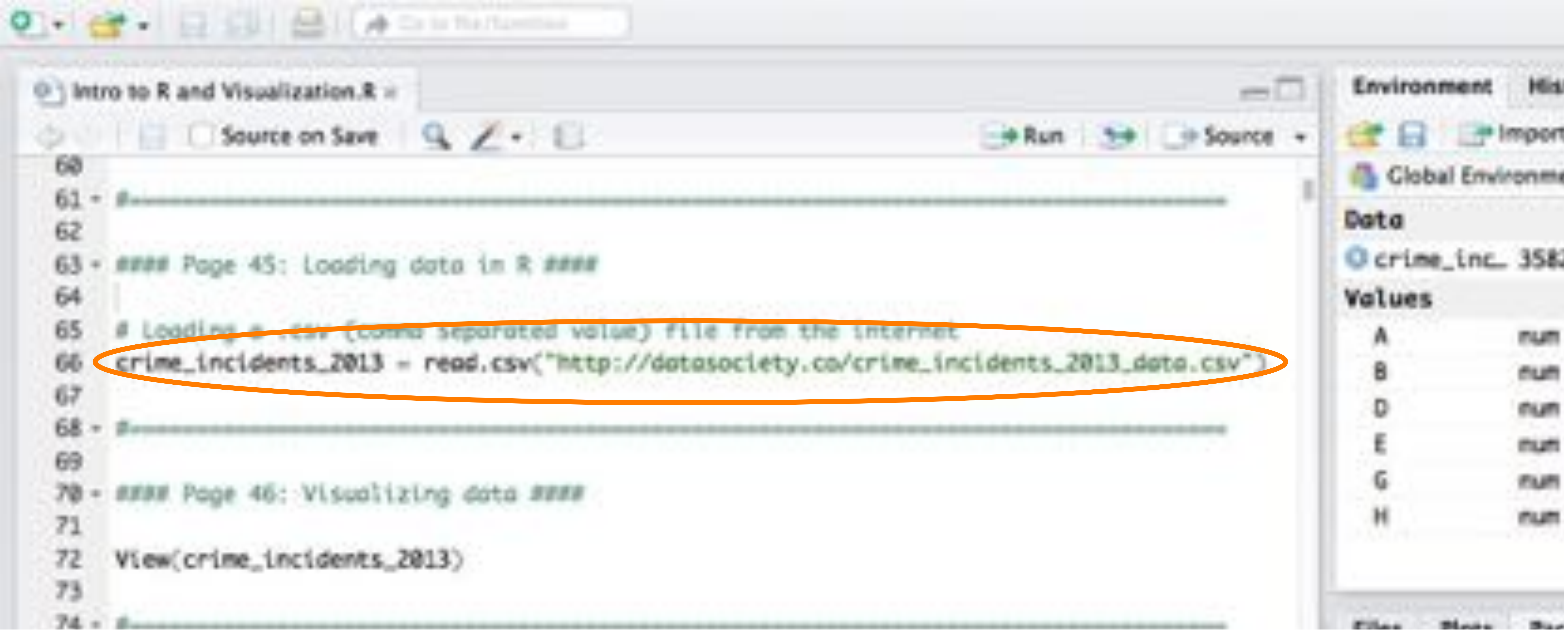
Please enter the URL to import data from:

http://datasociety.co/crime_incidents_2013_data.csv

OK Cancel

Loading data in R

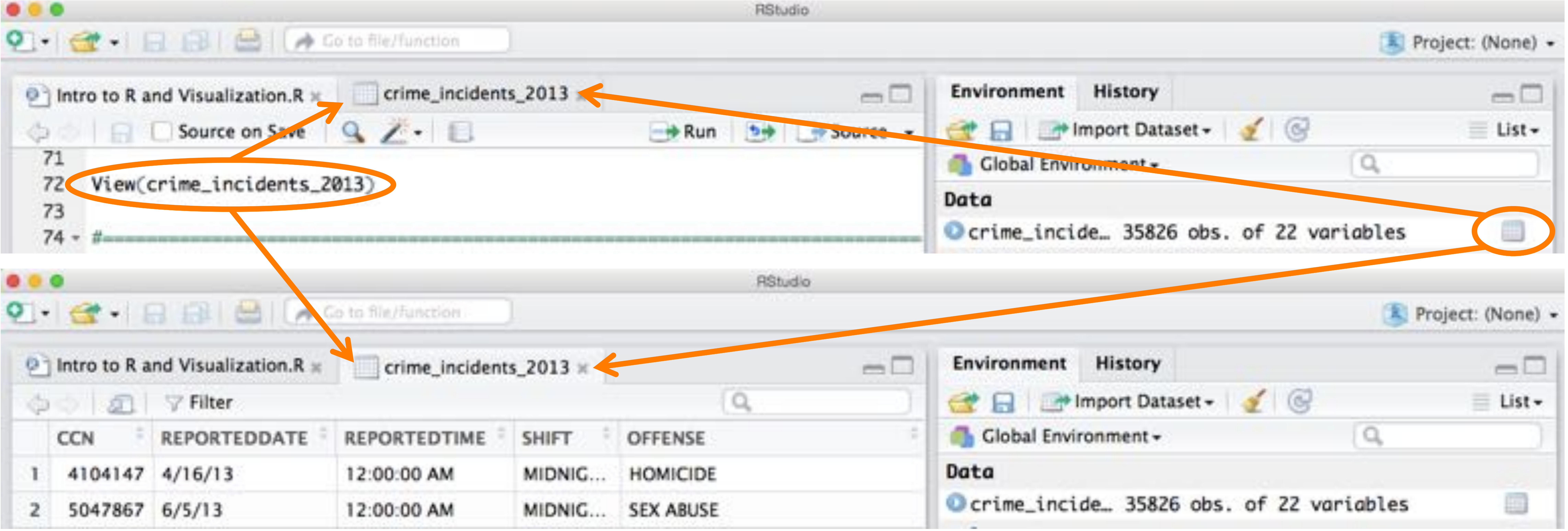
- Loading data from the internet
  - Enter code into script window
  - `crime_incidents_2013` is instantiated as the label of the data set



```r
#------------------------------------------------------------------------------

#### Page 45: Loading data in R ####

# Loading a .csv (comma separated value) file from the internet
crime_incidents_2013 <- read.csv("http://datasociety.co/crime_incidents_2013_data.csv")

#------------------------------------------------------------------------------

#### Page 46: Visualizing data ####

View(crime_incidents_2013)

#------------------------------------------------------------------------------
```

Visualizing data

- Once data is loaded, you can see it as a spreadsheet by either:
  - Pressing the "spreadsheet" button in the top right window
  - Using the `View()` function in the script window



```
View(crime_incidents_2013)
#
```

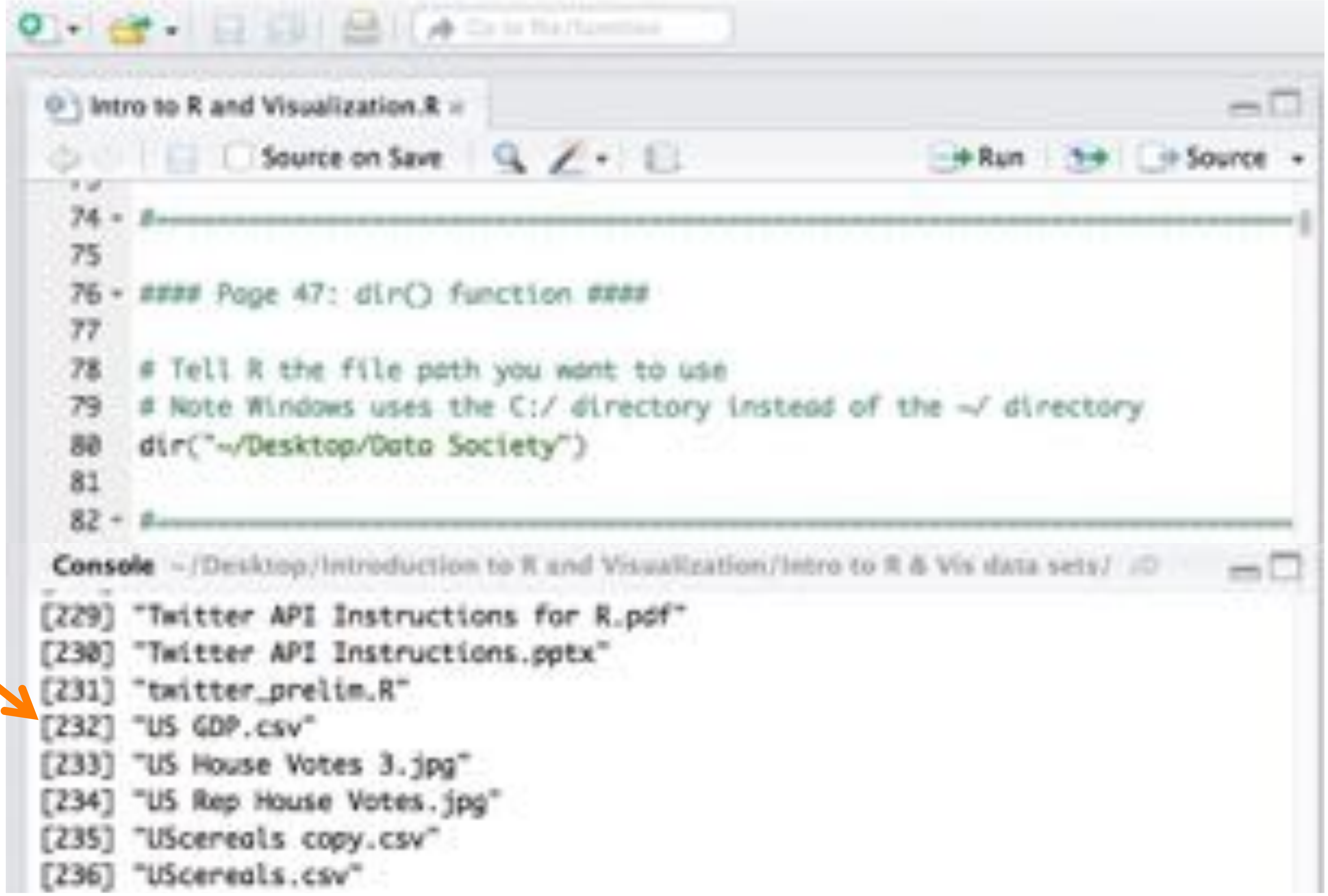
Data

crime_incide… 35826 obs. of 22 variables

	CCN	REPORTEDDATE	REPORTEDTIME	SHIFT	OFFENSE
1	4104147	4/16/13	12:00:00 AM	MIDNIG…	HOMICIDE
2	5047867	6/5/13	12:00:00 AM	MIDNIG…	SEX ABUSE

`dir()` function

- Lists all the files in a particular directory


Go to file/function
Intro to R and Visualization.R
Source on Save
Run
Source
Console ~/Desktop/Introduction to R and Visualization/Intro to R & Vis data sets/

```
#--------------------------------------------------------------------------

#### Page 47: dir() function ####

# Tell R the file path you want to use
# Note Windows uses the C:/ directory instead of the ~/ directory
dir("~/Desktop/Data Society")

#--------------------------------------------------------------------------
```

```
[229] "Twitter API Instructions for R.pdf"
[230] "Twitter API Instructions.pptx"
[231] "twitter_prelim.R"
[232] "US GDP.csv"
[233] "US House Votes 3.jpg"
[234] "US Rep House Votes.jpg"
[235] "UScereals copy.csv"
[236] "UScereals.csv"
```

Number of the file in the list

R can read many types of files

Script

```
read.csv("filename.csv")           # read Excel files converted to csv format
read.table("filename")             # reads a table from a text file
read.spss("filename.spss")         # reads SPSS files
read.dta("filename.dta")           # reads Stata files
read.ssd("filename.ssd")           # reads SAS files
read.octave("filename.octave")     # read Octave files
read.mtp("filename.mtp")           # read Minitab files
read.systat("filename.systat")     # read Systat files
read.JPEG("filename.jpg")          # read JPEG image files
```

Note: this requires us to install package called 'jpeg', we will cover packages later