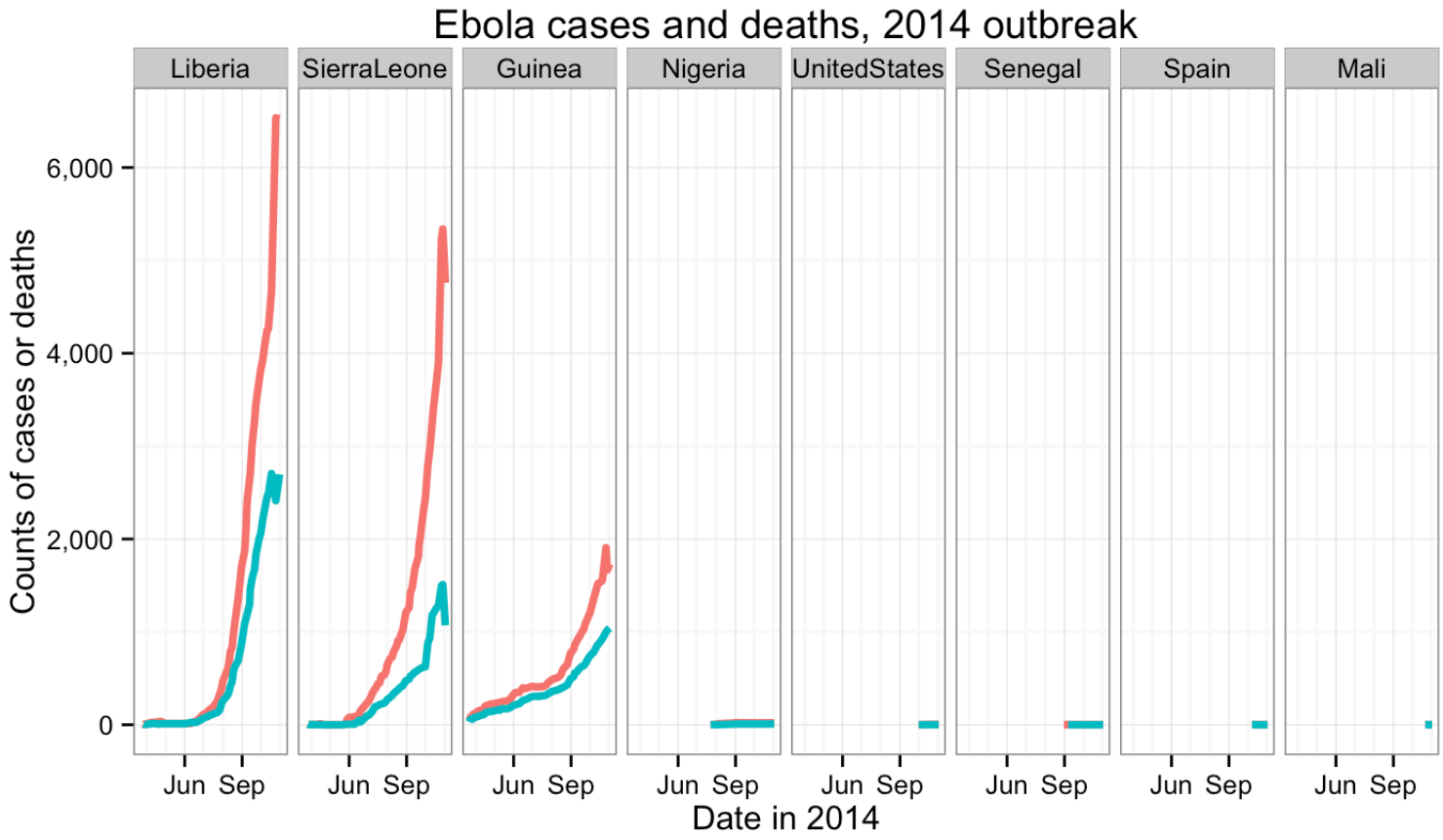
Meeting 1: R and Ebola

Brooke Anderson

November 7, 2014

2014 Ebola outbreak



First, catch your rabbit

GitHub

Caitlin River's repo of data for the 2014 Ebola outbreak:

Caitlin River's Ebola repo

First, catch your rabbit

Basic approach:

- Download data to your computer
- Make sure R is working in the directory with your data
- Read data into R

Fancier approach:

- Ask R to read data straight from GitHub

Full details of basic approach for beginners

Basic approach

Step 1: Download the data to your computer.

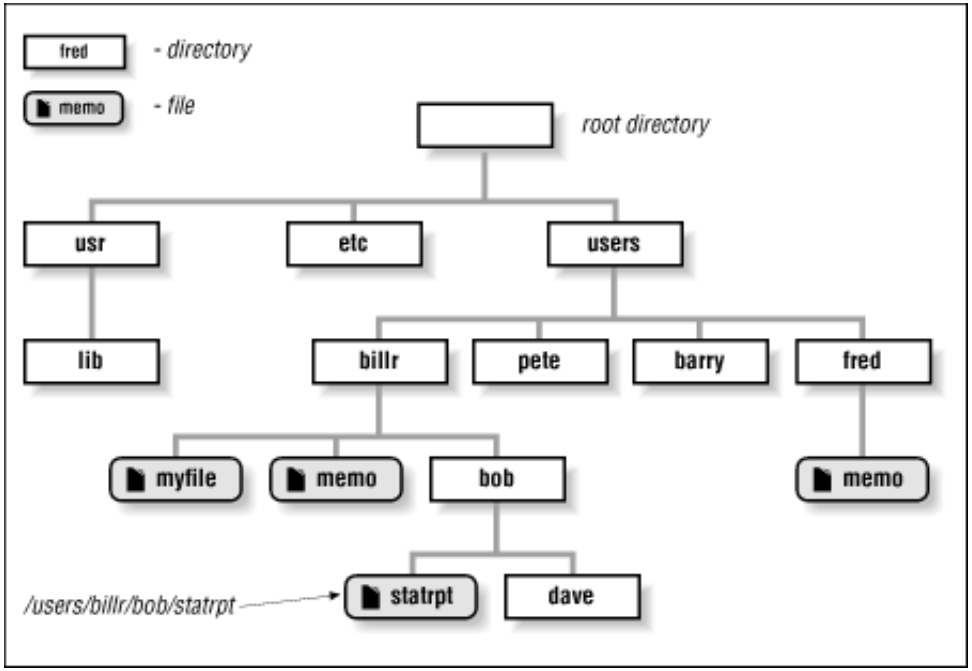
To download a datafile from GitHub, check out the "Raw" button on the page…

Note: Make sure you save with the extension ".csv", and don't let your computer add on ".txt"

Basic approach

Step 2: Anytime you work in R, R will run from within a directory somewhere on your computer.

Let's review directories:



Basic approach

The default is usually your home directory (for example, mine is "/Users/brookeanderson"). You can easily change your working directory…

Once you open R, what is your working directory?

```
getwd()
```

```
## [1] "/Users/brookeanderson/FallRMeetings"
```

```
setwd("/Users/brookeanderson")
getwd()
```

```
## [1] "/Users/brookeanderson"
```

Basic approach

Check your working directory now:

```
getwd()
```

```
## [1] "/Users/brookeanderson/FallRMeetings"
```

If you're in the right directory, you should see our data file if you list the files in the directory:

```
list.files()
```

```
## [1] "country_timeseries.csv" "EbolaCenters.csv"
## [3] "ebolavirus_cds.nex"     "Figures"
## [5] "Meeting1Code.r"         "Meeting1Notes_files"
## [7] "Meeting1Notes.html"     "Meeting1Notes.Rmd"
## [9] "README.md"
```

Basic approach

Step 3: Now we can read the data into R. It's a very basic command:

```
ebola <- read.table("country_timeseries.csv", sep = ",",
                    header = TRUE)
ebola[1:3, 1:5]
```

```
##         Date Day Cases_Guinea Cases_Liberia Cases_SierraLeone
## 1  11/2/2014 225         1731            NA              4759
## 2 10/31/2014 222           NA          6525                NA
## 3 10/29/2014 220         1667            NA              5338
```

But why does this work?

Basic approach

R can read in data from *a lot* of different formats. The only catch: you need to tell R how to do it.

Most basically, we'll look at flat files:

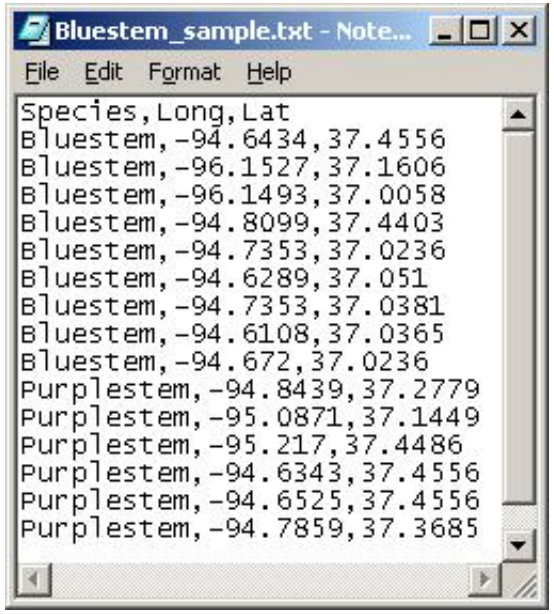
Fixed width files

Delimited files

- ".csv": Comma-separated values
- ".tab", ".tsv": Tab-separated values
- Other possible delimiters: colon, semicolon, pipe ("|")

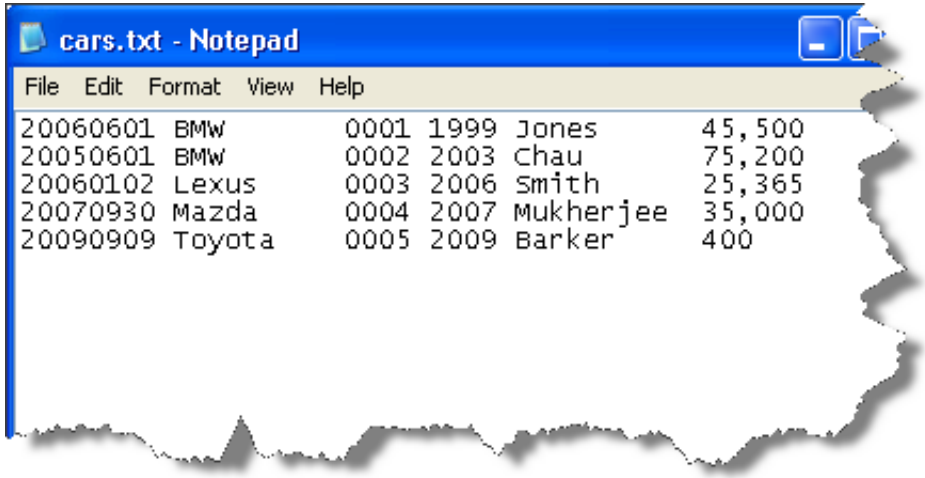
See if you can identify what types of files the following files are…

What type of file?




```
Species,Long,Lat
Bluestem,-94.6434,37.4556
Bluestem,-96.1527,37.1606
Bluestem,-96.1493,37.0058
Bluestem,-94.8099,37.4403
Bluestem,-94.7353,37.0236
Bluestem,-94.6289,37.051
Bluestem,-94.7353,37.0381
Bluestem,-94.6108,37.0365
Bluestem,-94.672,37.0236
Purplestem,-94.8439,37.2779
Purplestem,-95.0871,37.1449
Purplestem,-95.217,37.4486
Purplestem,-94.6343,37.4556
Purplestem,-94.6525,37.4556
Purplestem,-94.7859,37.3685
```

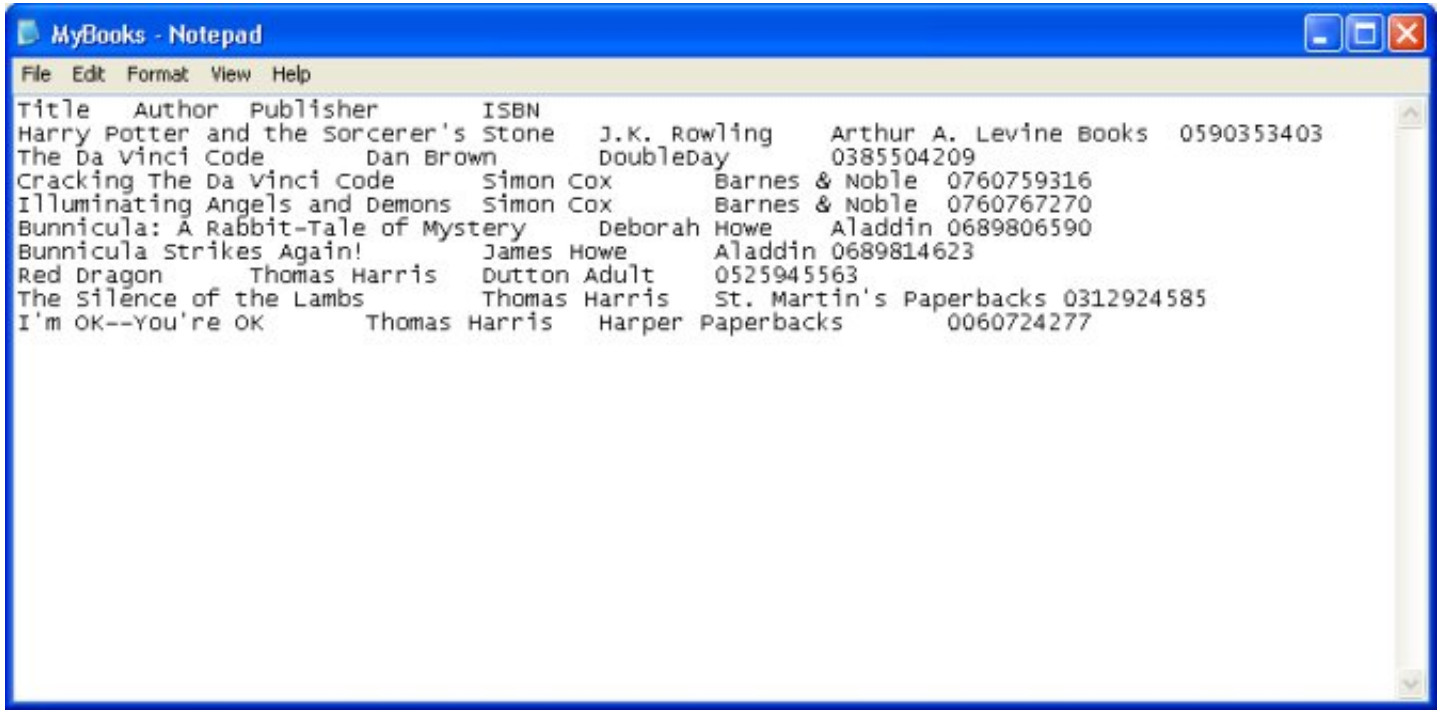
What type of file?



What type of file?

```
H|20110606|pizza.txt|
D|10|Chicken Pesto|20|23|30|5.5|7.4|9.9||
D|10|Meatball|10|53|60|6.5|8.4|10.9|
D|10|Fire Cracker|3|13|60|5.8|7.9|11.9|
D|10|Spinach|1|2|5|5.5|7.0|8.8|
D|10|BBQ Chicken|35|102|95|6.5|7.9|10.9|
D|10|Vegetarian|5|13|28|4.5|7.9|9.5|
D|10|Mexican|11|33|36|5.5|7.4|9.9|
D|10|The Monaco|22|53|7|5.5|7.5|8.9|
D|10|Chilli Prawn|5|5|6|5.5|7.4|9.9|
D|10|Chefs Special|8|18|40|5.8|7.8|9.8|
D|10|Marinara|3|17|41|5.5|7.4|9.0|
D|10|Supreme|50|52|58|5.5|7.4|9.2|
D|10|Margherita|9|19|87|5.0|7.0|8.0|
D|10|Napoli|60|85|66|5.2|7.2|9.2|
D|10|Caprice|31|32|38|5.5|7.4|9.3|
D|10|Ham and Pineapple|18|39|28|5.8|7.0|9.0|
T|16|
```

What type of file?



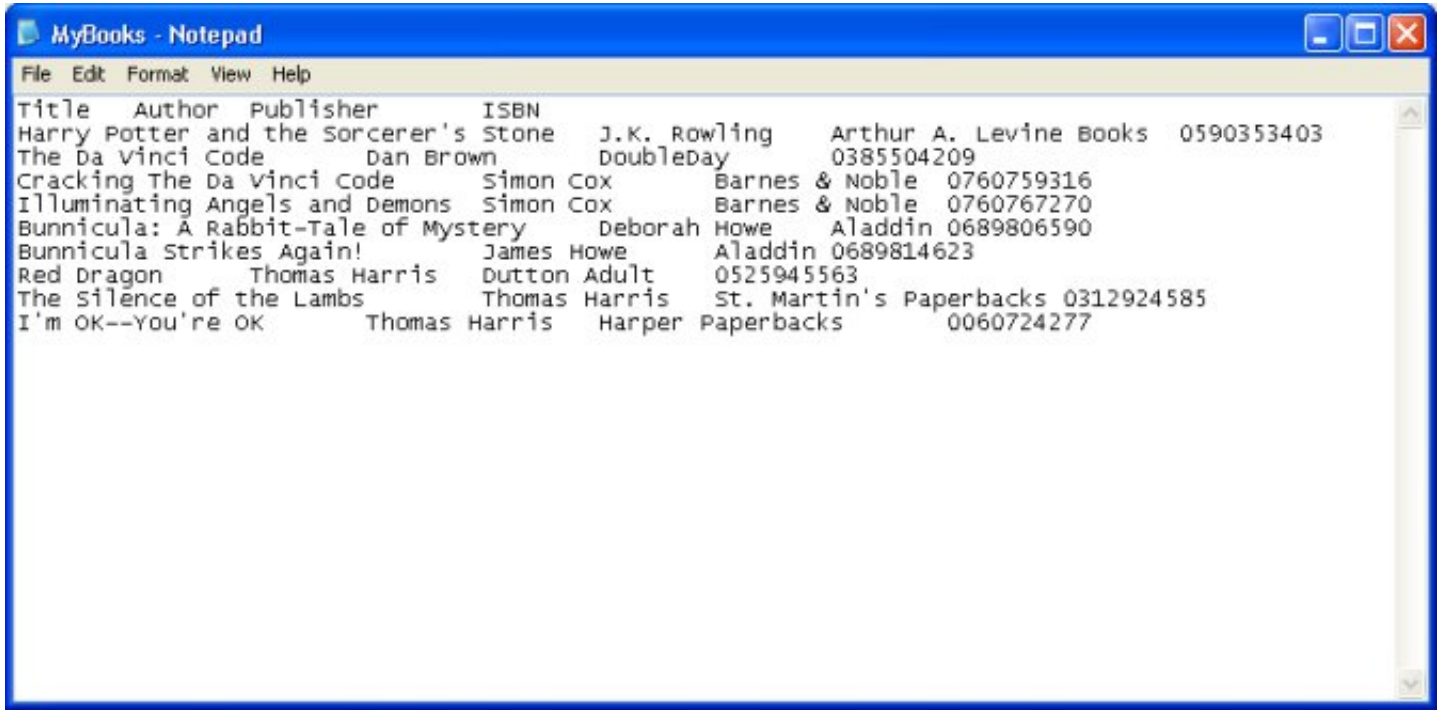

```
Title   Author  Publisher       ISBN
Harry Potter and the Sorcerer's Stone   J.K. Rowling    Arthur A. Levine Books  0590353403
The Da Vinci Code       Dan Brown       DoubleDay       0385504209
Cracking The Da Vinci Code      Simon Cox       Barnes & Noble  0760759316
Illuminating Angels and Demons  Simon Cox       Barnes & Noble  0760767270
Bunnicula: A Rabbit-Tale of Mystery     Deborah Howe    Aladdin 0689806590
Bunnicula Strikes Again!        James Howe      Aladdin 0689814623
Red Dragon      Thomas Harris   Dutton Adult    0525945563
The Silence of the Lambs        Thomas Harris   St. Martin's Paperbacks 0312924585
I'm OK--You're OK       Thomas Harris   Harper Paperbacks       0060724277
```

What type of file?

File Edit Format View Help

```
Title,Subtitle,Larger Work,Contributor #1,Contributor #2,Contributor #3,Contributor
#4,Genre,Publisher,Published Location,Date
Published,Instrumentation,Key,Location,Indiana Connection,Sheet Music
Consortium,Notes,Complete
"""A"" You're Adorable",The alphabet song,,Buddy Kaye,Sidney Lippman,Fred Wise,,Popular
standard,Laurel Music Corporation,"New York, NY",1948,Voice and
piano/guitar or ukulele,C Major,,None,Yes,Perry Como pictured on cover,
"Aba Daba Honey Moon, The",,"""Two Weeks with Love"" Motion Picture",Arthur Fields,Walter
Donovan,,,"Popular Standard, Movie Selection",Leo Feist Inc.,"New
York, NY",1942,Voice and Piano,C Minor,,None,Yes,,
Abi Bezunt,,"""Mamele"" Motion Picture",Abraham Ellstein,Molly Picon,,,"Popular Standard,
Movie Selection",Metro Music Co.,"New York, NY",1939,Voice and
Piano,E Minor,,None,No,Molly Picon pictured on cover,
Abdul the Bulbul Ameer,,,Bob Kaai,Jim Smock,,,Popular Standard,Calumet Music Co.
,"Chicago, IL",1935,"Voice, Piano, Hawaiian Guitar, Ukulele",G
Major,,None,Yes,Ben Pollack pictured on cover,
About A Quarter to Nine,,"""Go Into Your Dance"" Motion Picture",Harry Warren,Al
Dubin,,,"Popular Standard, Movie Selection",M. Witmark & Sons,"New York,
NY",1935,"Voice, Piano, Guitar, Ukelele",E Minor,,None,No,Al Jolson and Ruby Keeler
pictured on cover,
Absent,,,John. W. Metcalf,Catherine Young Glen,,,Popular Standard,Arthur P.
Schmidt,"Boston, MA",1899,Voice and Piano,G Major,,None,Yes,,
The Academy Two-Step,,,Barclay Walker,,,,Popular Standard,Carlin & Lennox,"Indianapolis,
IN",,Piano,F Major,,Composer,No,,
Ac-cent-tchu-ate the Positive,Mister In Between,"""Here Come the Waves"" Motion
Picture",Harold Arlen,Johnny Mercer,,,"Popular Standard, Movie
Selection",Edwin H. Morris & Co.,"New York, NY",1944,"Voice, Piano, Guitar",F
Major,,None,Yes,Bing Crosby and Betty Hutton pictured on cover,
Across the Alley From the Alamo,,,Joe Greene,,,,Popular Standard,Leslie Music
```

What type of file?




```
Title   Author  Publisher       ISBN
Harry Potter and the Sorcerer's Stone   J.K. Rowling    Arthur A. Levine Books  0590353403
The Da Vinci Code       Dan Brown       DoubleDay       0385504209
Cracking The Da Vinci Code      Simon Cox       Barnes & Noble  0760759316
Illuminating Angels and Demons  Simon Cox       Barnes & Noble  0760767270
Bunnicula: A Rabbit-Tale of Mystery     Deborah Howe    Aladdin 0689806590
Bunnicula Strikes Again!        James Howe      Aladdin 0689814623
Red Dragon      Thomas Harris   Dutton Adult    0525945563
The Silence of the Lambs        Thomas Harris   St. Martin's Paperbacks 0312924585
I'm OK--You're OK       Thomas Harris   Harper Paperbacks       0060724277
```

What type of file?

```
1000233   Miralda          John
1000234   Faley            Nick
1000235   Baylog           Cathy
1000236   Gallardo         Mike
1000237   Christian        Daniel
1000238   Baufield         Daniel
1000239   Frazier          Robert
1000240   Garrido          Edward
1000241   Williams         Zachary
1000242   Morel            David
          Padilla          Damian
1000244   Rosenberg        Wayne
1000245   Blanchard        Phong S
1000246   Wiggins          David
1000247   Miller           Jeffrey
1000248   Coon             Terry
1000249   Chretien         Walter
1000250   Myers            Timothy

1000233   Miralda          John
1000234   Faley            Nick
1000235   Baylog           Cathy
```

Basic approach

R can read any of these types of files using one of the `read.table` and `read.fwf` functions. Find out more about those functions with:

```
?read.table
?read.fwf
```

Basic approach

Now let's read the data in and assign it (`<-`) to an object named `ebola`:

```
ebola <- read.table("country_timeseries.csv", sep = ",",
                    header = TRUE)
```

Notice that the function is `read.table`, and we've specified a value for the options `sep` and `header`. We'll talk about functions and options more later…

Question for you: What would have happened if we hadn't assigned the data we were reading in to `ebola`?

Basic approach

If this worked, you should have an object in your R session named `ebola`. You can check out the beginning of it:

```
ebola[1:3, 1:5]
```

```
##         Date Day Cases_Guinea Cases_Liberia Cases_SierraLeone
## 1  11/2/2014 225         1731            NA              4759
## 2 10/31/2014 222           NA          6525                NA
## 3 10/29/2014 220         1667            NA              5338
```

Basic approach

There are a number of other functions you can use to check out your data. For example, try:

```
head(ebola)
tail(ebola)
summary(ebola)
str(ebola)
ebola[1,]
```

Details of fancier approach

Fancier approach

But, wait. **Everyone** is using GitHub for sharing data now. Surely there's a simpler way??

This is R. Of course there is!

Fancier approach

If you don't know how to do it, try Googling:

"read data into r github"

Check out the first result:

"data- Read a CSV from github into R - Stack Overflow"

Fancier approach

It turns out that you can read it straight from GitHub using the `RCurl` package.

If you don't have the package yet, you'll need to install it:

```
install.packages("RCurl")
```

Then, to use it in your R session, you'll need to call it:

```
library(RCurl)
```

Now you have all the `RCurl` tools available in your session.

Fancier approach

Now all it takes to read the data in is:

```
github.page <- getURL("https://raw.githubusercontent.com/cmrivers/ebola/master/co
ebola.2 <- read.csv(text = github.page)
```

Note: It runs off the page, but the full "https" for the `getURL` *function is just [the web address of the raw data we want from GitHub](#):*

Fancier approach

```r
ebola.2[1:3, 1:5]
```

```
##         Date Day Cases_Guinea Cases_Liberia Cases_SierraLeone
## 1  11/2/2014 225         1731            NA              4759
## 2 10/31/2014 222           NA          6525                NA
## 3 10/29/2014 220         1667            NA              5338
```

What kind of data can you get into R?

The sky is the limit…

- Tables on webpages (e.g., the table near the end of this page)
- Files from other statistical packages (SAS, Excel, Stata, SPSS)
- Data in a database (e.g., SQL)
- Really crazy data formats used in other disciplines (e.g., netCDF files from climate folks, MRI data stored in Analyze, NIfTI, and DICOM formats)
- Data through APIs (e.g., GoogleMaps, Twitter)
- Incrediably messy data using `scan` and `readLines`

Find out more in Chapter 3 of The R Book.

Challenges for the more advanced

Challenge 1

GitHub user BrcMapsTeam has geojson data on the locations of Ebola medical centers in West Africa as well as a link to a GoogleDocs dataset with the same information. See if you can get this data into R from one of these two sources.

Challenge 2

GitHub user evogytis has a repo called ebolaGuinea2014 with `.nex` files with protein coding sequences from Ebola genomes. He ultimately is using this data for phylogenetic analysis of the Ebola outbreak in Guinea. See if you can read one of these files into R. Also, see if you can figure out what tools exist for doing phylogenetic analysis in R.