

# Data Science in the

Amelia McNamara

[amelia.mn](http://amelia.mn)

@AmeliaMN



*Data Science in the tidyverse* is licensed under a [Creative Commons Attribution 4.0 International License](#). Based on work at <https://github.com/cwickham/data-science-in-tidyverse> and <https://github.com/rstudio-education/master-the-tidyverse>

**HELLO**  
my name is

**Amelia**



@AmeliaMN

**HELLO**  
my name is

**Hadley**



@hadleywickham

**HELLO**  
my name is

Jesse



@kierisi

**HELLO**  
my name is

Irene



@i\_stevens

**HELLO**  
my name is

**Ben**



@baumerben

# Your Turn

Introduce yourself to your neighbors:

- Who are you?
- What you do with data?
- How would you describe your experience with R?



No sticky note: "I'm happily working on it"



**Blue** sticky note: "I'm all done and ready to move on"



**Orange** sticky note: "I'm stuck, can someone help me?"

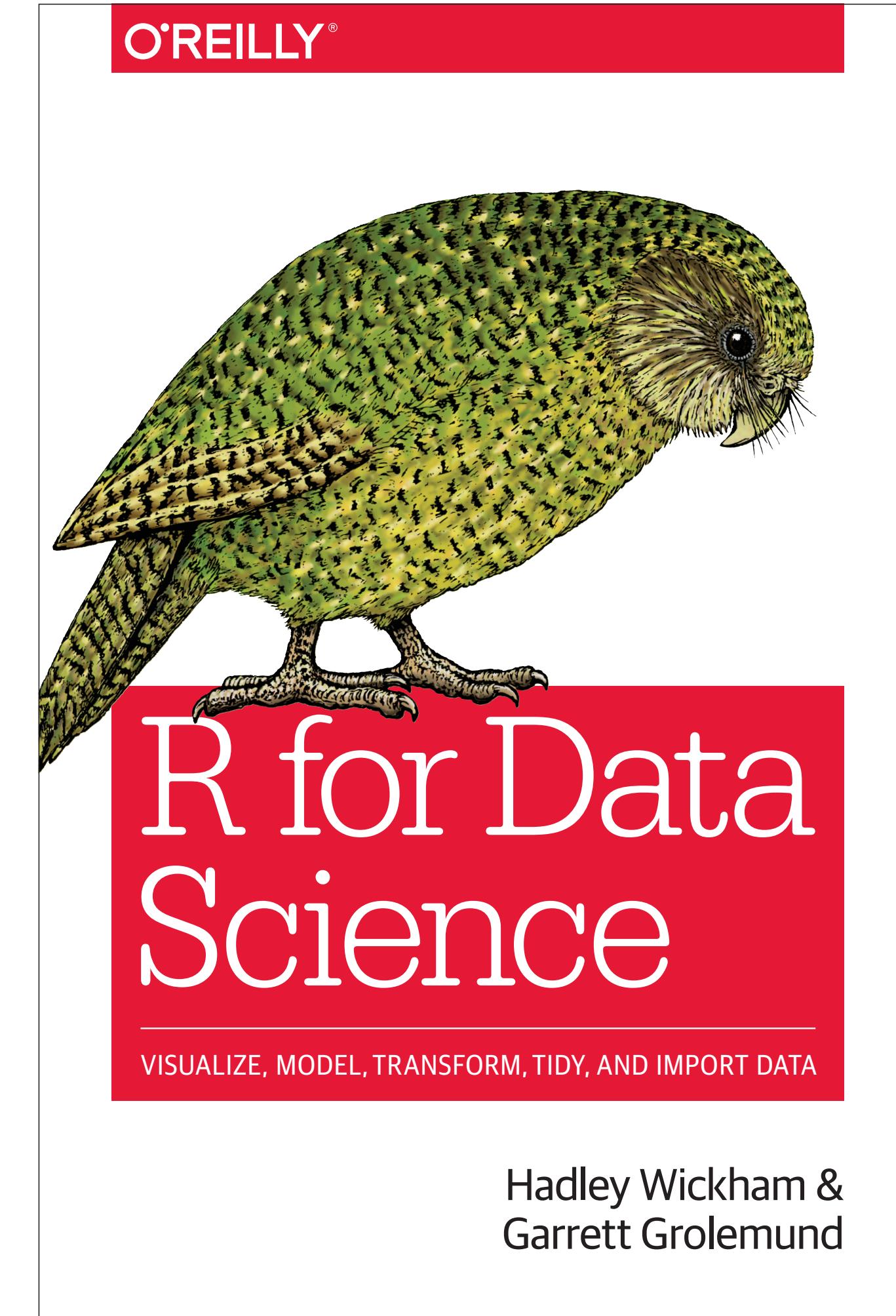
Alternatively, flag one of us down



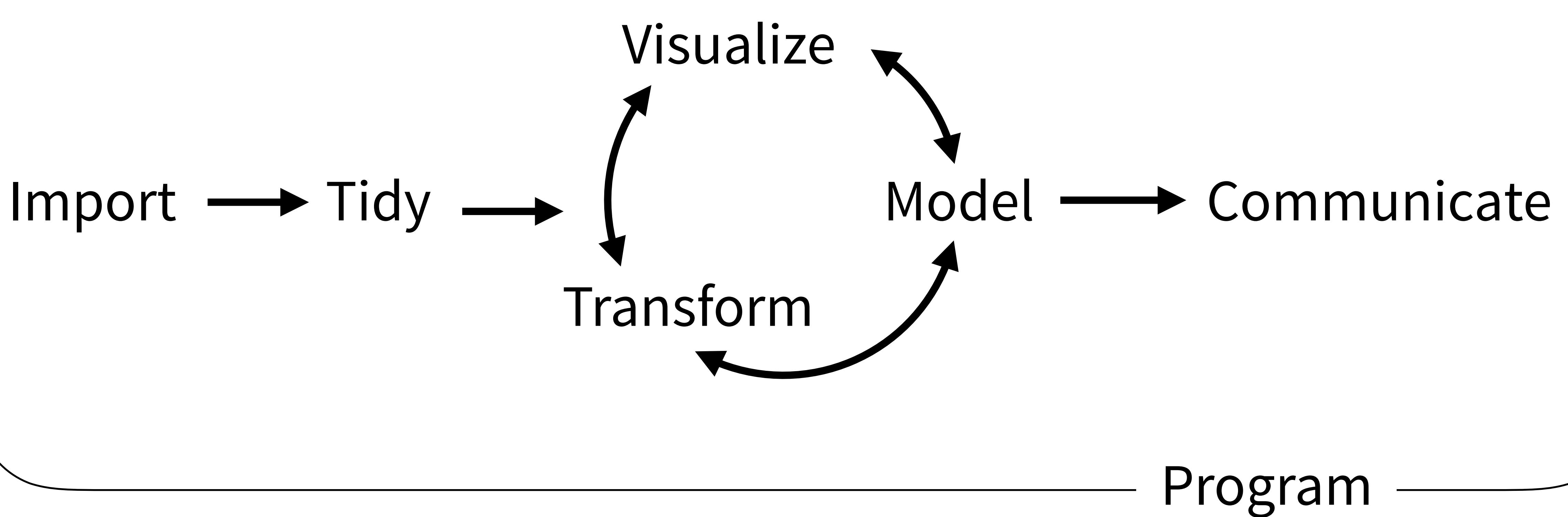
This class is heavily based on  
R for Data Science

<http://r4ds.had.co.nz/>

Links to the relevant  
sections of the book

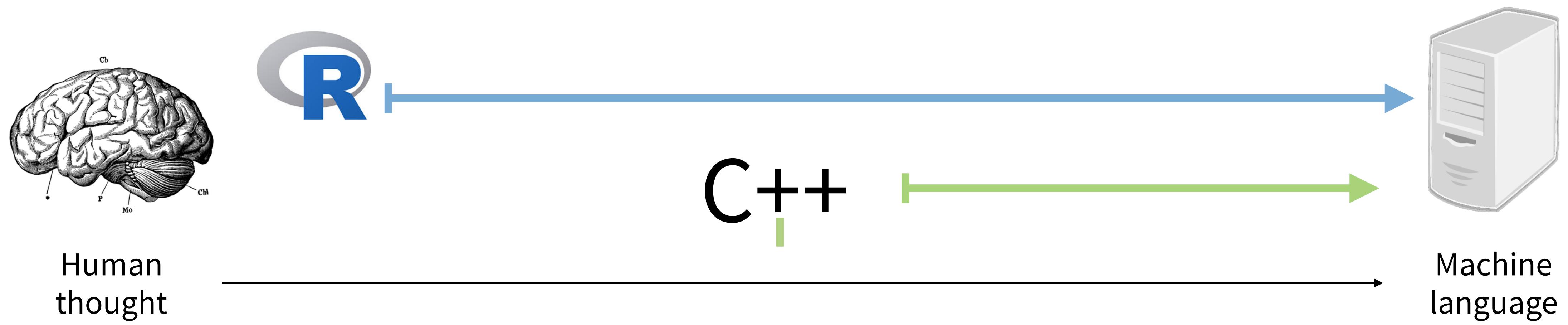


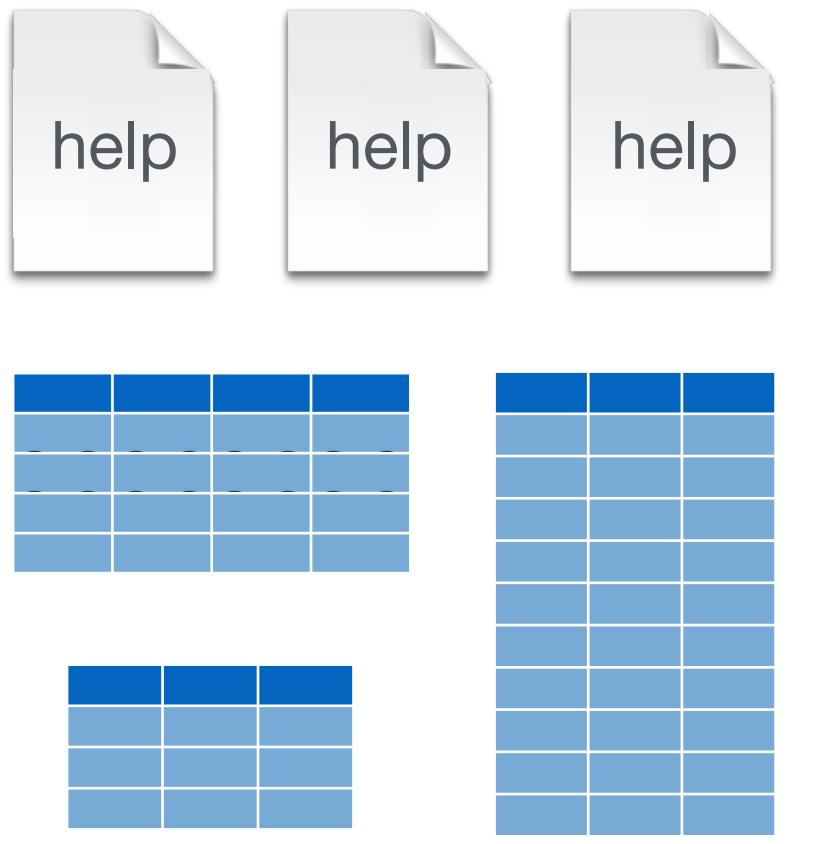
# (Applied) Data Science



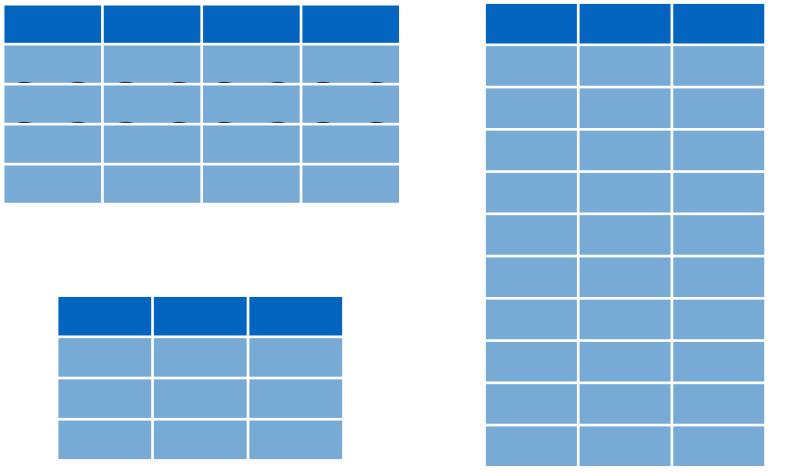
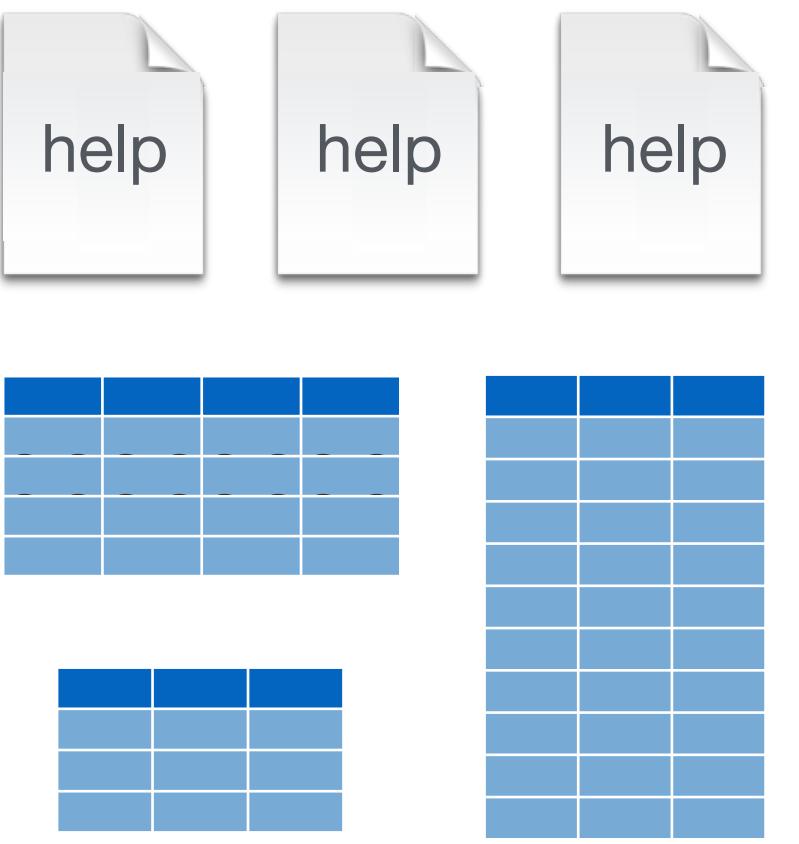


# R - A computer language for scientists

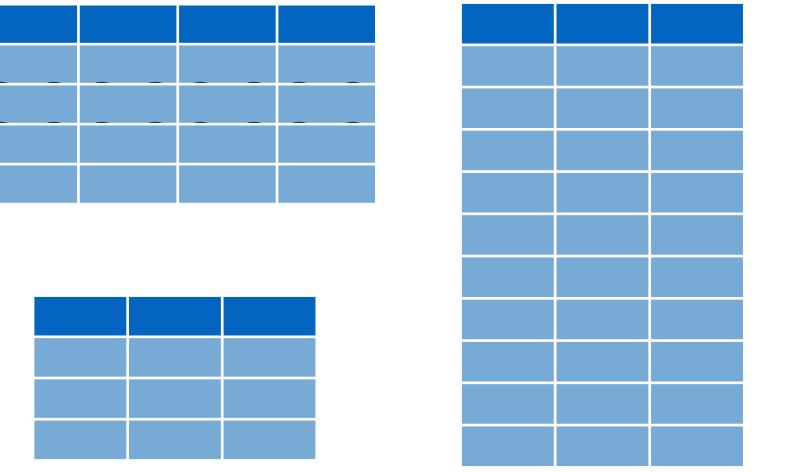
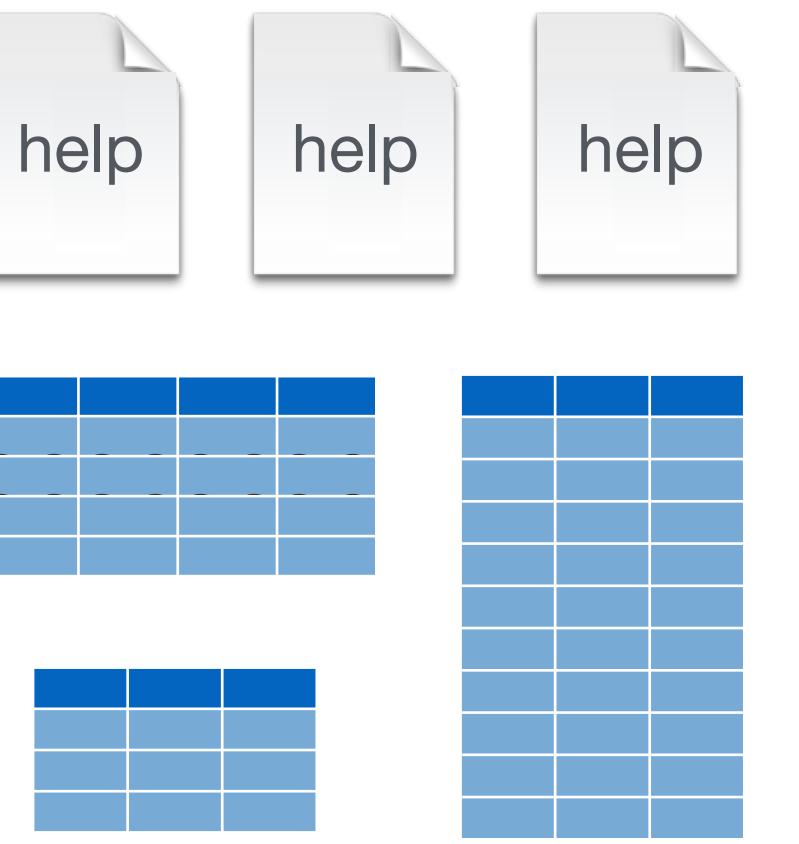




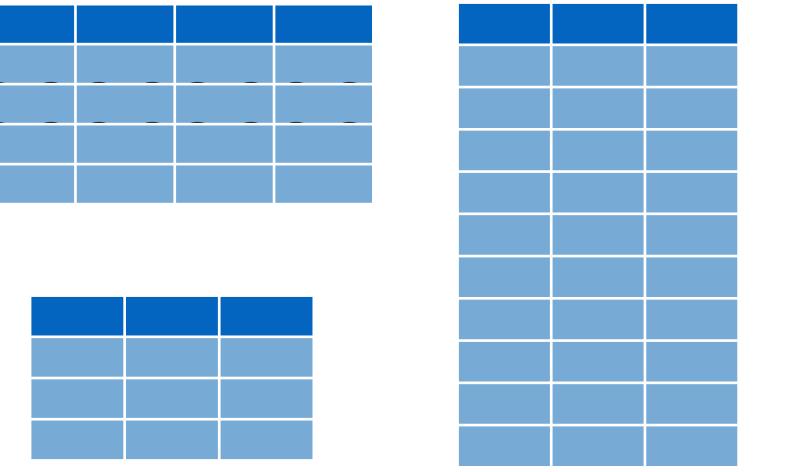
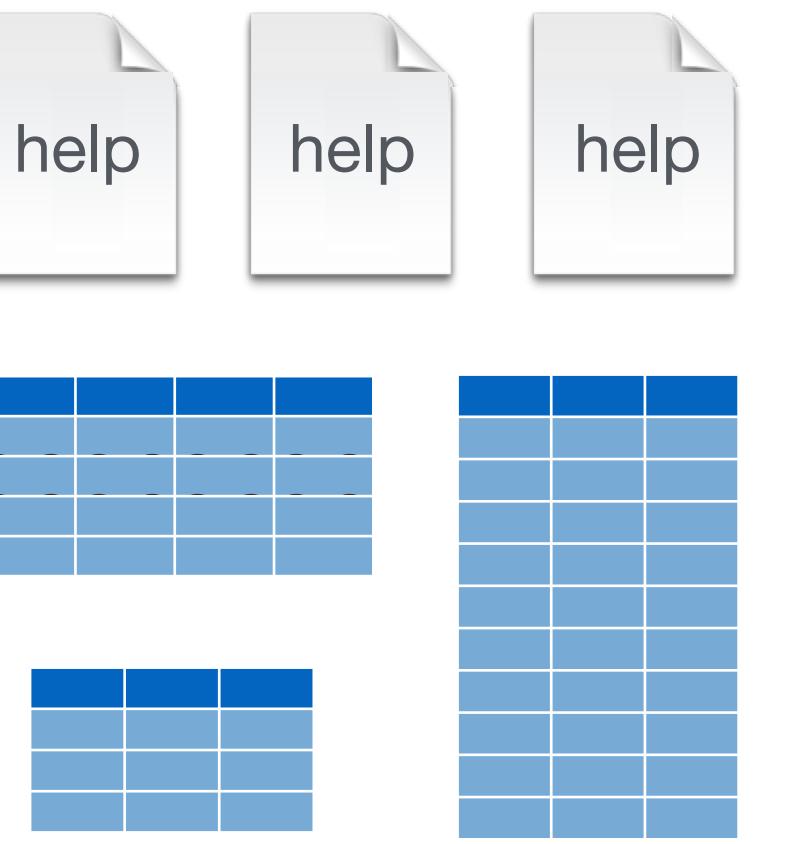
`function1()`  
`function2()`  
`function3()`  
`function4()`



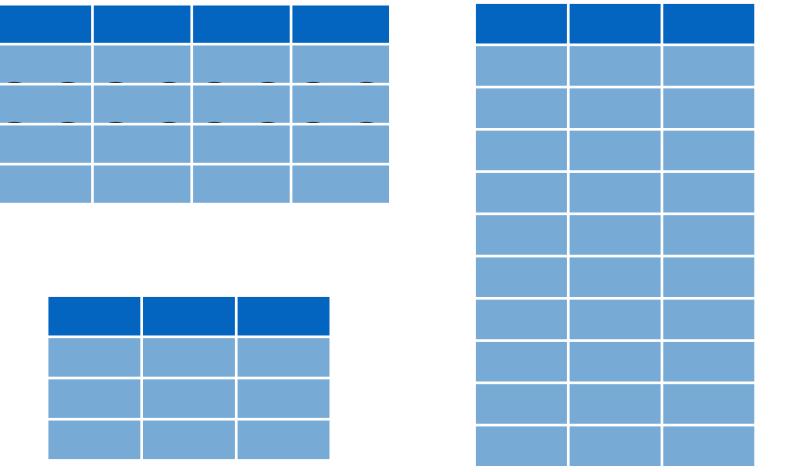
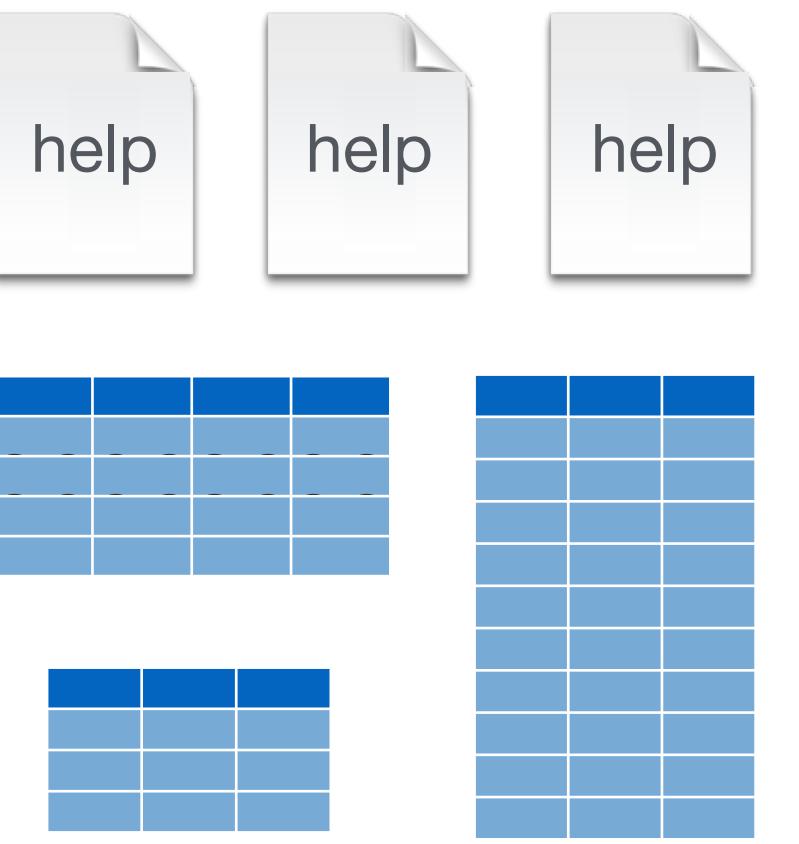
function1()  
function2()  
function3()  
function4()



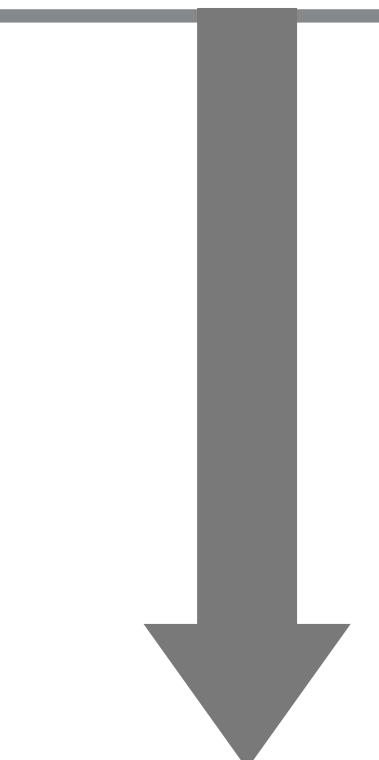
function5()  
function6()  
function7()  
function8()



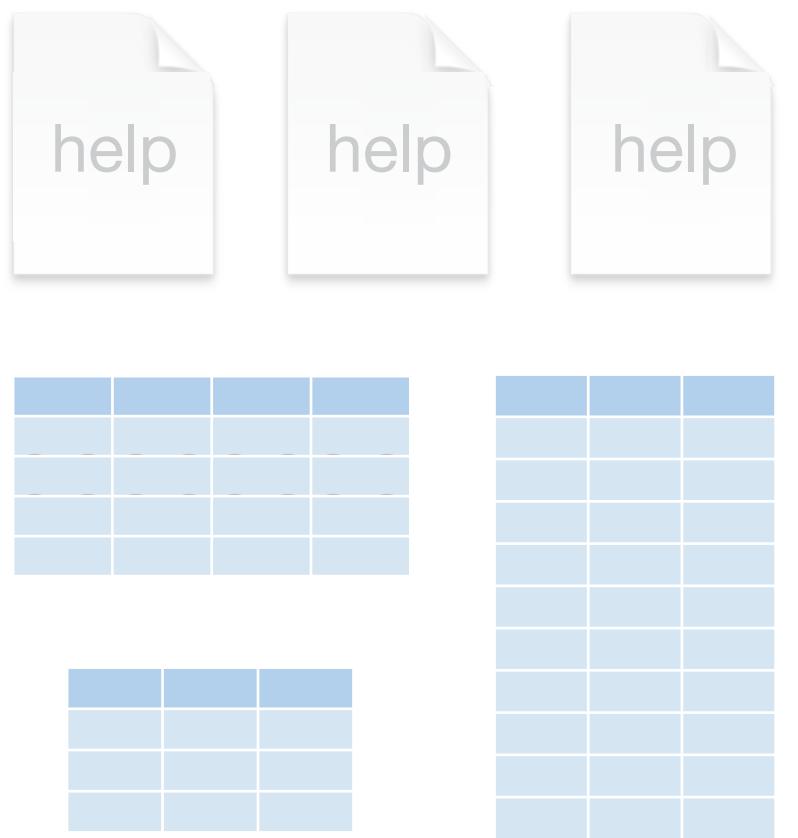
function9()  
functionA()  
functionB()  
functionC()



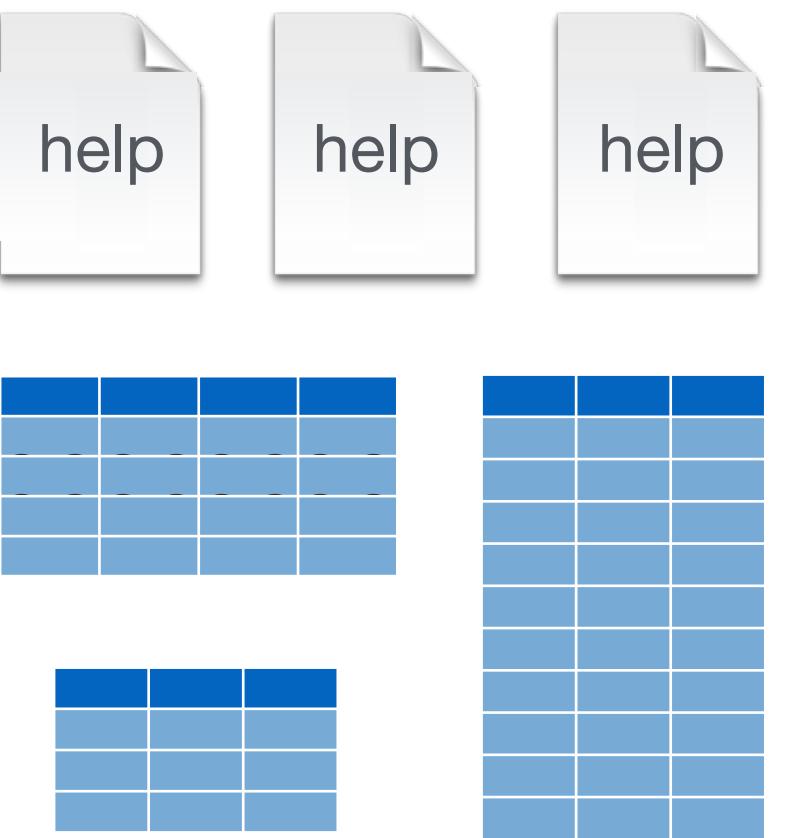
functionD()  
functionE()  
functionF()  
functionG()



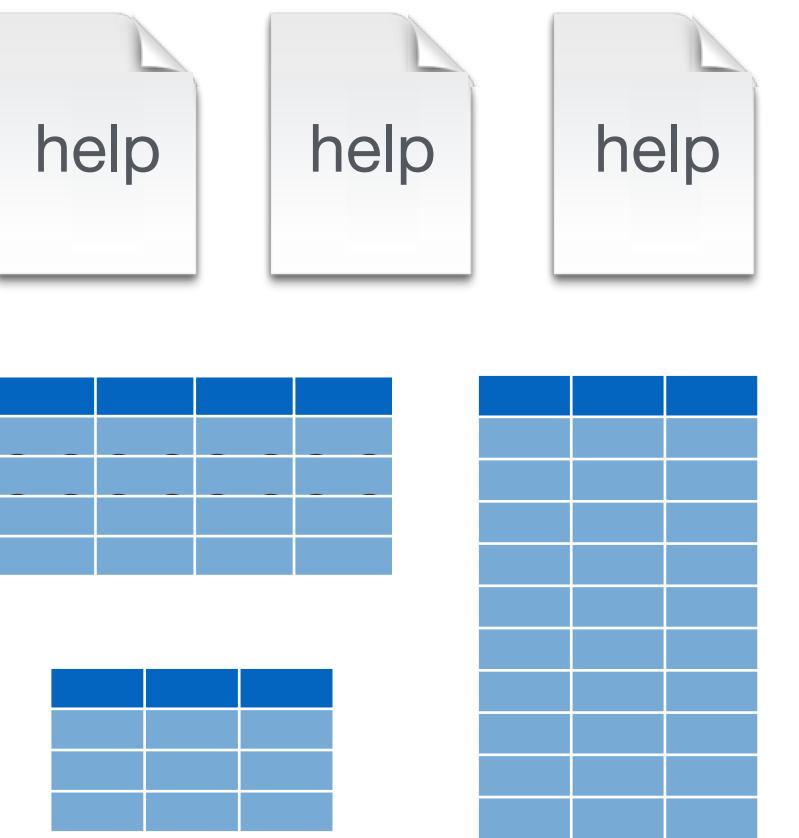
Base R



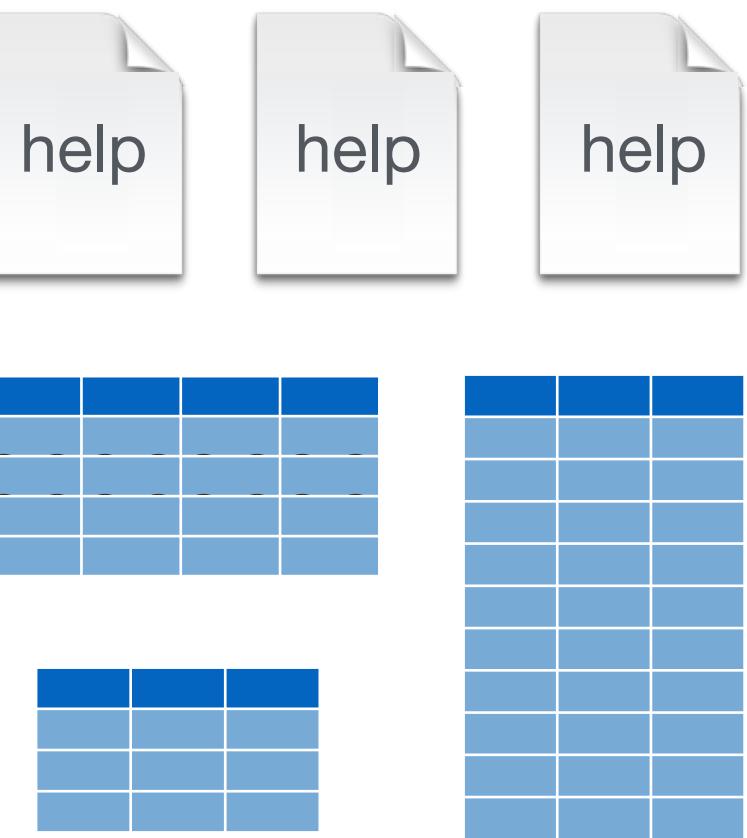
function1()  
function2()  
function3()  
function4()



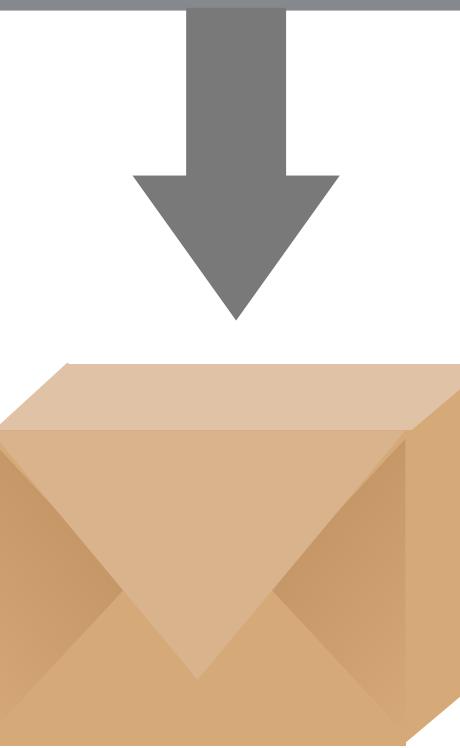
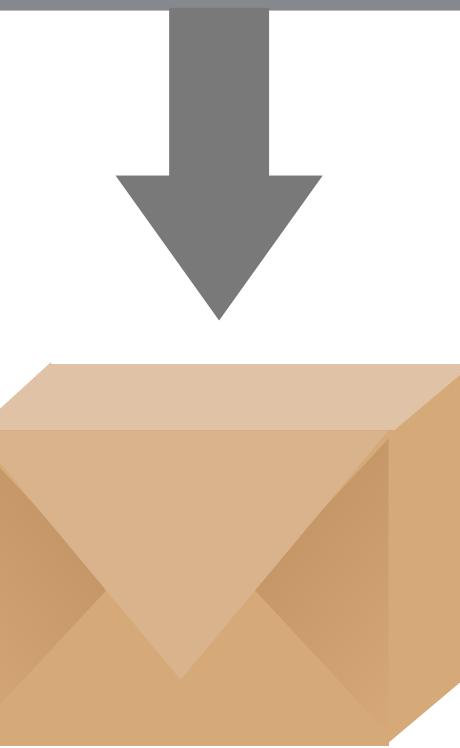
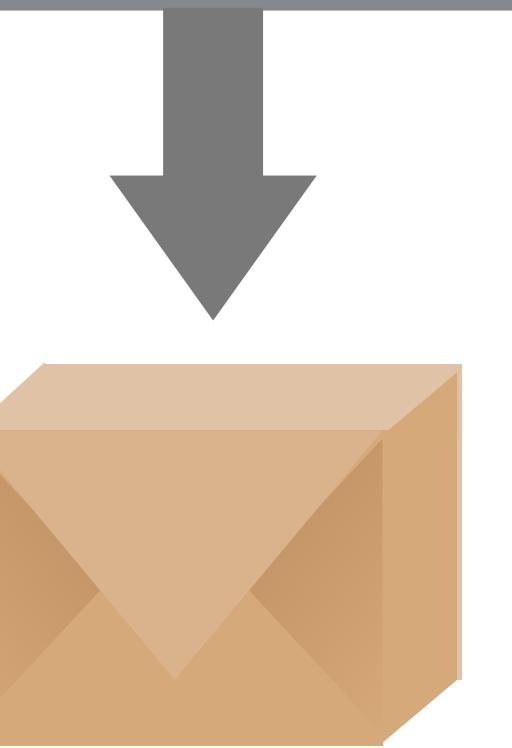
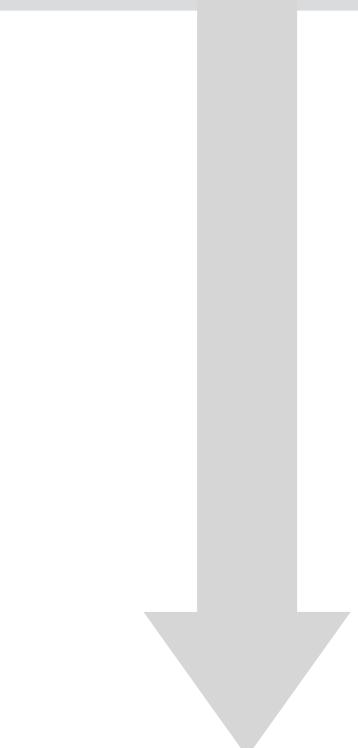
function5()  
function6()  
function7()  
function8()



function9()  
functionA()  
functionB()  
functionC()



functionD()  
functionE()  
functionF()  
functionG()



Base R

R Packages



The Comprehensive R Archive x ! Garrett

Secure | <https://cran.r-project.org>

[!\[\]\(1bcfefa5ce11f5709f61f09cca20172c\_img.jpg\)](#)

[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](#)

## Available CRAN Packages By Name

[A](#) [B](#) [C](#) [D](#) [E](#) [F](#) [G](#) [H](#) [I](#) [J](#) [K](#) [L](#) [M](#) [N](#) [O](#) [P](#) [Q](#) [R](#) [S](#) [T](#) [U](#) [V](#) [W](#) [X](#) [Y](#) [Z](#)

<a href="#">A3</a>	Accurate, Adaptable, and Accessible Error Metrics for Predictive Models
<a href="#">abyyR</a>	Access to Abbyy Optical Character Recognition (OCR) API
<a href="#">abc</a>	Tools for Approximate Bayesian Computation (ABC)
<a href="#">ABCAnalysis</a>	Computed ABC Analysis
<a href="#">abc.data</a>	Data Only: Tools for Approximate Bayesian Computation (ABC)
<a href="#">abcdeFBA</a>	ABCDE_FBA: A-Biologist-Can-Do-Everything of Flux Balance Analysis with this package
<a href="#">ABCOptim</a>	Implementation of Artificial Bee Colony (ABC) Optimization
<a href="#">ABCp2</a>	Approximate Bayesian Computational Model for Estimating P2
<a href="#">ABC.RAP</a>	Array Based CpG Region Analysis Pipeline
<a href="#">abcrf</a>	Approximate Bayesian Computation via Random Forests
<a href="#">abctools</a>	Tools for ABC Analyses
<a href="#">abd</a>	The Analysis of Biological Data
<a href="#">abf2</a>	Load Gap-Free Axon ABF2 Files
<a href="#">ABHgenotypeR</a>	Easy Visualization of ABH Genotypes
<a href="#">abind</a>	Combine Multidimensional Arrays
<a href="#">abjutils</a>	Useful Tools for Jurimetric Analysis Used by the Brazilian Jurimetrics Association
<a href="#">abn</a>	Modelling Multivariate Data with Additive Bayesian Networks
<a href="#">abodOutlier</a>	Angle-Based Outlier Detection

# Using packages

**1**

```
install.packages("foo")
```

Downloads files to "computer"  
**1 x per "computer"**

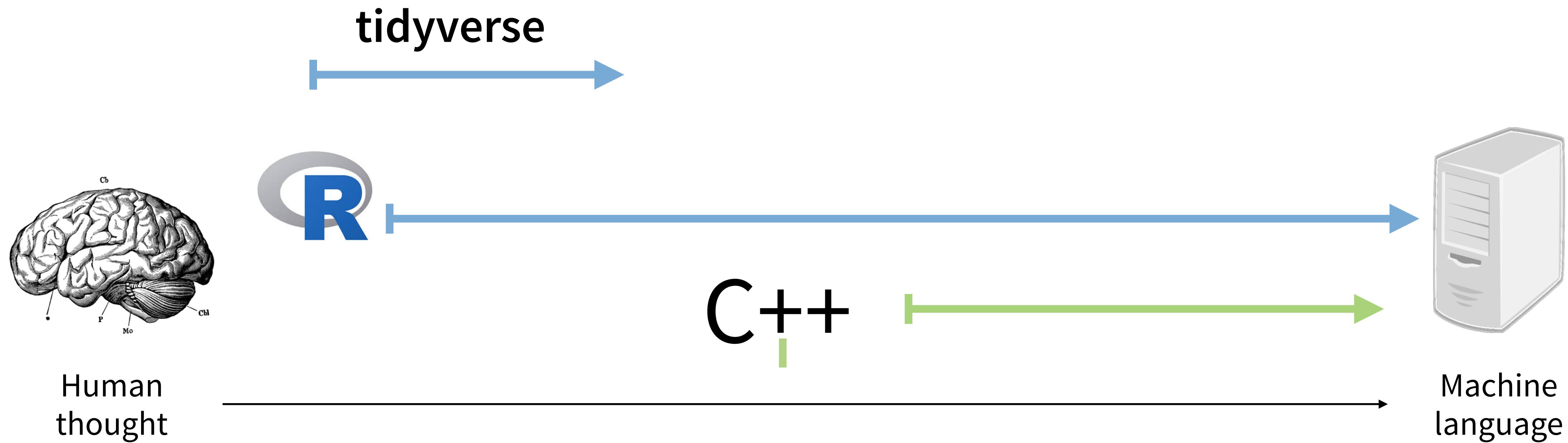
**2**

```
library("foo")
```

Loads package  
**1 x per R Session**

I've done this  
for you for this  
workshop

# The tidyverse - A set of R packages to unify some data science tasks



# tidyverse.org

The screenshot shows a Mac OS X browser window displaying the [tidyverse.org](https://www.tidyverse.org) website. The page has a dark blue header with the word "Tidyverse" in white. Below the header is a navigation bar with links for "Packages", "Articles", "Learn", "Help", and "Contribute". On the left side, there is a graphic composed of several hexagonal icons representing different R packages: `dplyr` (orange, with a pliers icon), `ggplot2` (grey, with a line plot icon), `readr` (blue, with a document icon), `tidyverse` (dark blue, with a grid icon), `tidyr` (orange, with a circular arrow icon), and `purrr` (white with a cat icon). To the right of the graphic, the text reads: "R packages for data science. The tidyverse is an opinionated **collection of R packages** designed for data science. All packages share an underlying philosophy and common APIs." Below this text is a section titled "Install the complete tidyverse with:" containing the command `install.packages("tidyverse")`.

Tidyverse

Packages Articles Learn Help Contribute

R packages for data science

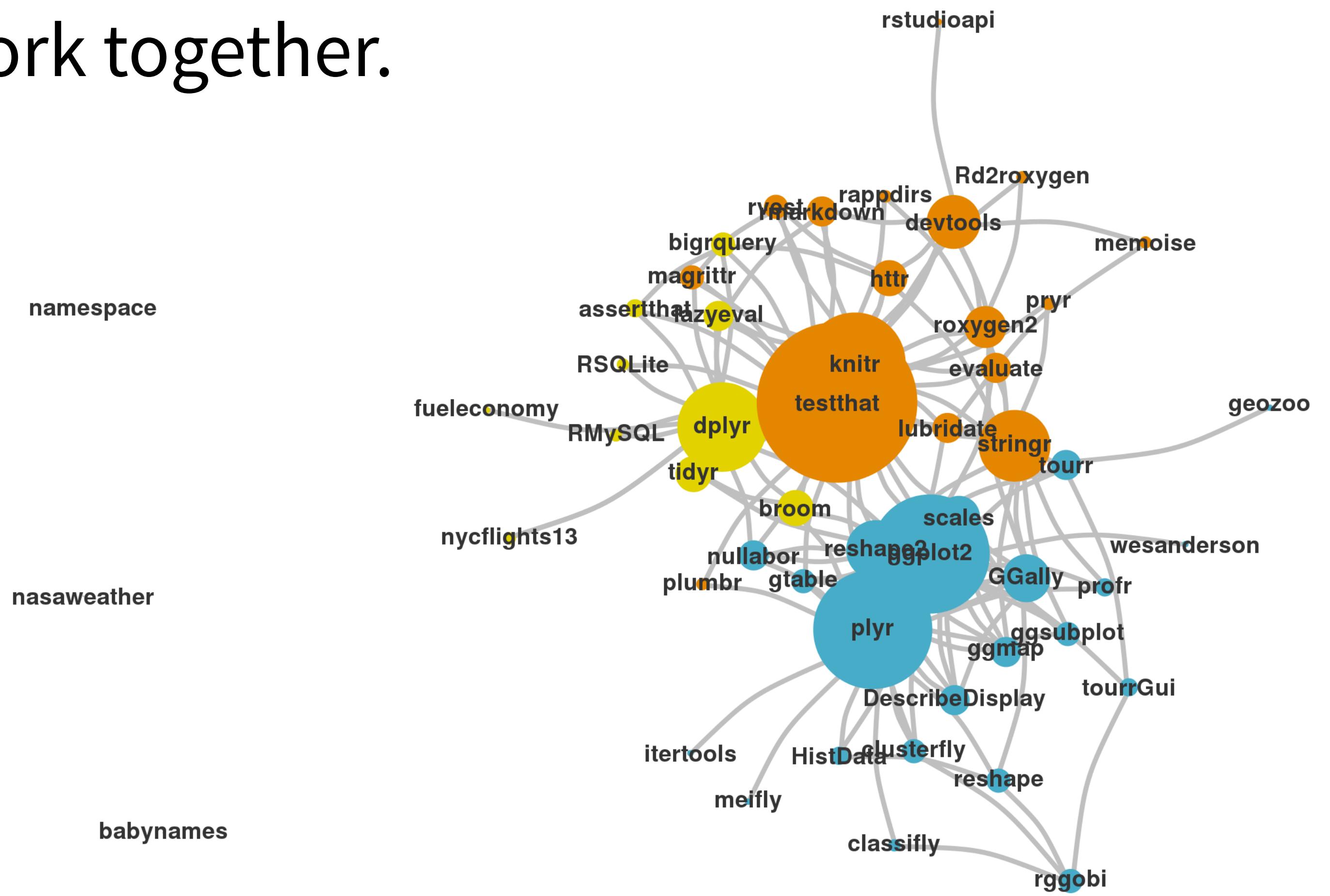
The tidyverse is an opinionated **collection of R packages** designed for data science. All packages share an underlying philosophy and common APIs.

Install the complete tidyverse with:

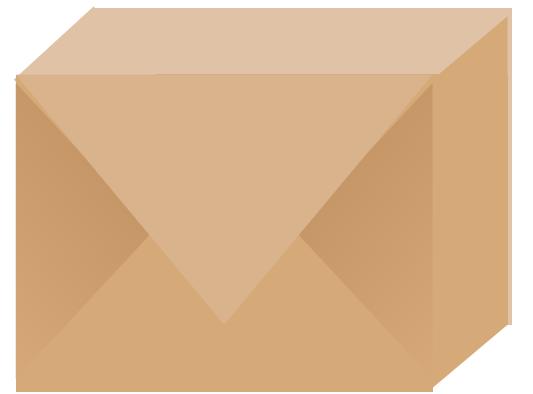
```
install.packages("tidyverse")
```

# The Tidyverse

A collection of modern R packages that share common philosophies, embed best practices, and are designed to work together.



# tidyverse



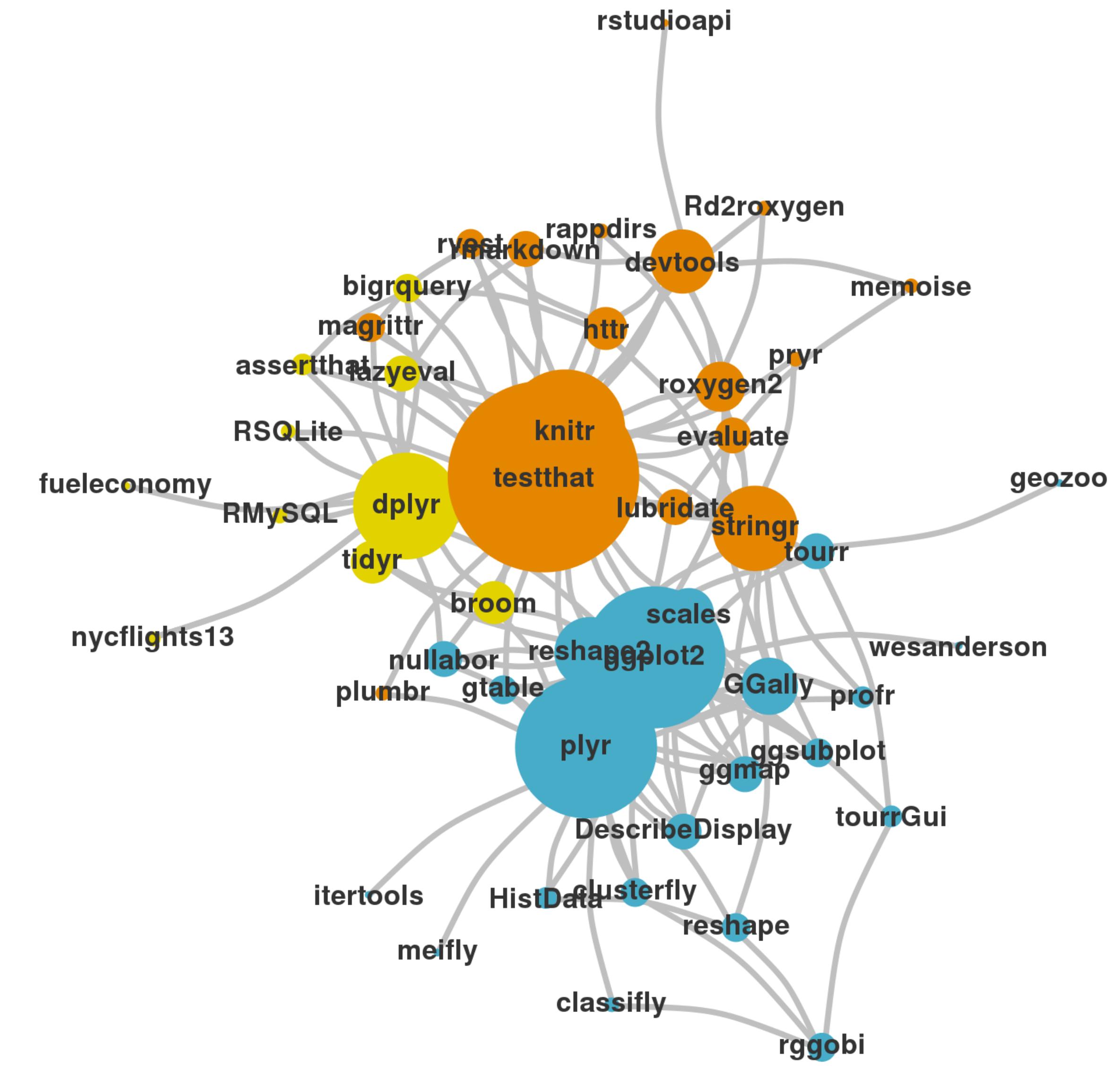
An R package that serves as a short cut for installing and loading the components of the tidyverse.

```
library("tidyverse")
```

```
install.packages("tidyverse")
```

does the equivalent of

```
install.packages("ggplot2")
install.packages("dplyr")
install.packages("tidyverse")
install.packages("readr")
install.packages("purrr")
install.packages("tibble")
install.packages("hms")
install.packages("stringr")
install.packages("lubridate")
install.packages("forcats")
install.packages("DBI")
install.packages("haven")
install.packages("httr")
install.packages("jsonlite")
install.packages("readxl")
install.packages("rvest")
install.packages("xml2")
install.packages("modelr")
install.packages("broom")
```



```
install.packages("tidyverse")
```

does the equivalent of

```
install.packages("ggplot2")
install.packages("dplyr")
install.packages("tidyr")
install.packages("readr")
install.packages("purrr")
install.packages("tibble")
install.packages("hms")
install.packages("stringr")
install.packages("lubridate")
install.packages("forcats")
install.packages("DBI")
install.packages("haven")
install.packages("httr")
install.packages("jsonlite")
install.packages("readxl")
install.packages("rvest")
install.packages("xml2")
install.packages("modelr")
install.packages("broom")
```

```
library("tidyverse")
```

does the equivalent of

```
library("ggplot2")
library("dplyr")
library("tidyr")
library("readr")
library("purrr")
library("tibble")
```

# Day 1

Introduction and  
Visualize Data

8:30 - 10:30

Morning Break

10:30 - 11:00

Visualize and Transform

11:00 - 12:30

Lunch

12:30 - 2:00

Transform

2:00 - 3:30

Afternoon Break

3:30 - 4:00

Tidy Data/  
Case Study

4:00 - 5:30

# Day 2

Data types	8:30 - 10:30
Morning Break	10:30 - 11:00
Iteration	11:00 - 12:30
Lunch	12:30 - 2:00
Modeling	2:00 - 3:30
Afternoon Break	3:30 - 4:00
Organize and wrap-up	4:00 - 5:30



# RStudio: a software program

1. like Microsoft Word, Excel, etc.
2. built to help you write R code, run R code, and analyze data with R
3. text editor, version control, keyboard shortcuts, debugging tools, and much more

# R Notebooks

(Let's start!)



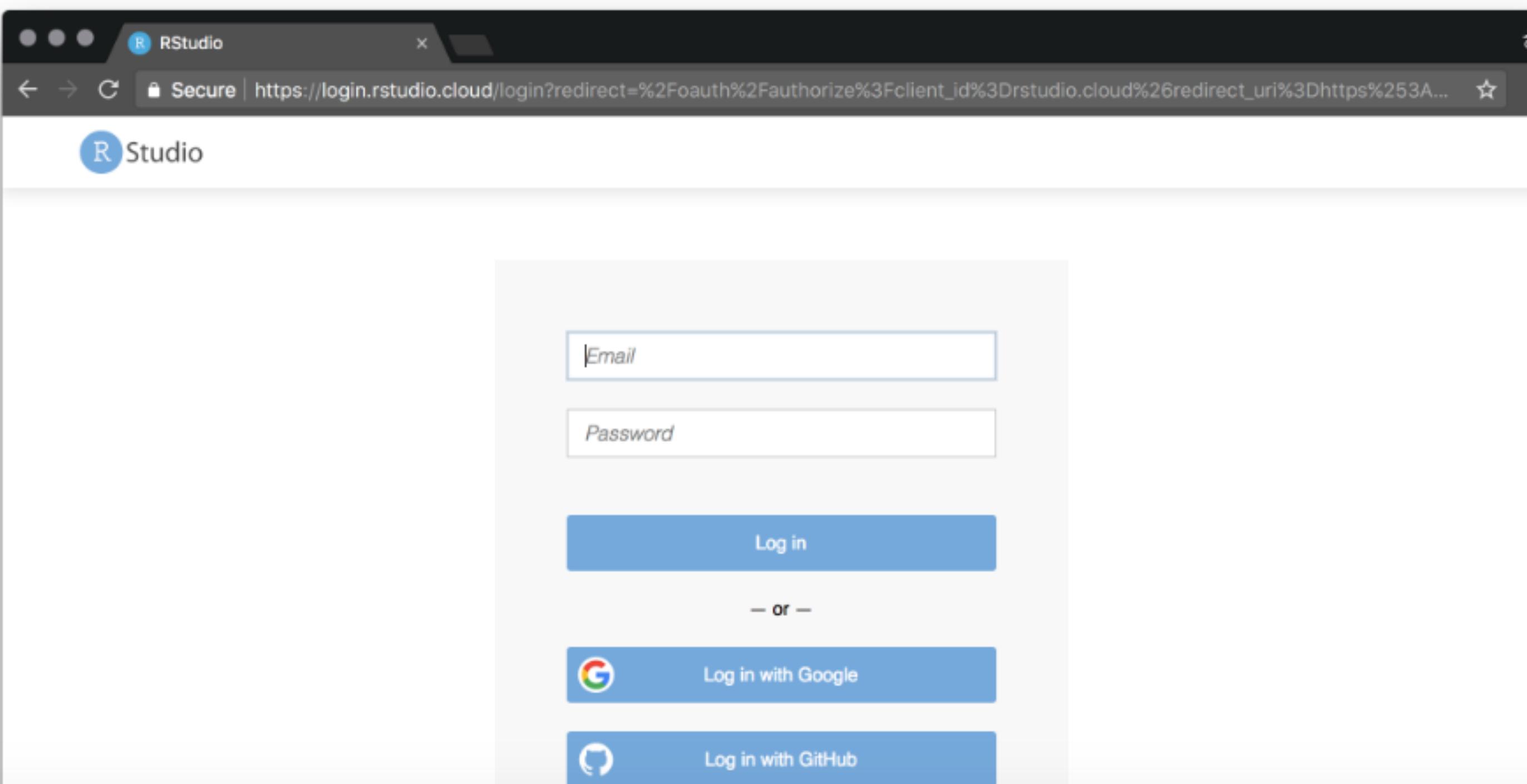
# Data Science in the Tidyverse

Go to <http://bit.ly/rstudio2019setup>  
and follow instructions

## To get started:

To get set up follow these steps:

1. Visit the project at <https://rstudio.cloud/project/163983>
2. Log in using google, github, shinyapps.io or "Sign Up".



# Your Turn

Instructions with screenshots at [bit.ly/rstudio2019setup](https://bit.ly/rstudio2019setup)

First, if you haven't already

- visit <https://rstudio.cloud/project/163983>
- Log In / Sign Up
- "Save a copy" of the project
- Open data-science-in-the-tidyverse.Rproj

When you have your copy of the project, let us know by putting up the  
**Blue** post-it.

Then, open 00-Getting-started.Rmd and look around



# R Notebooks

An authoring format for Data Science.

The screenshot shows the RStudio interface with an R Notebook open. The notebook file is titled "R-Notebook.Rmd". The code editor pane contains the following R Markdown code:

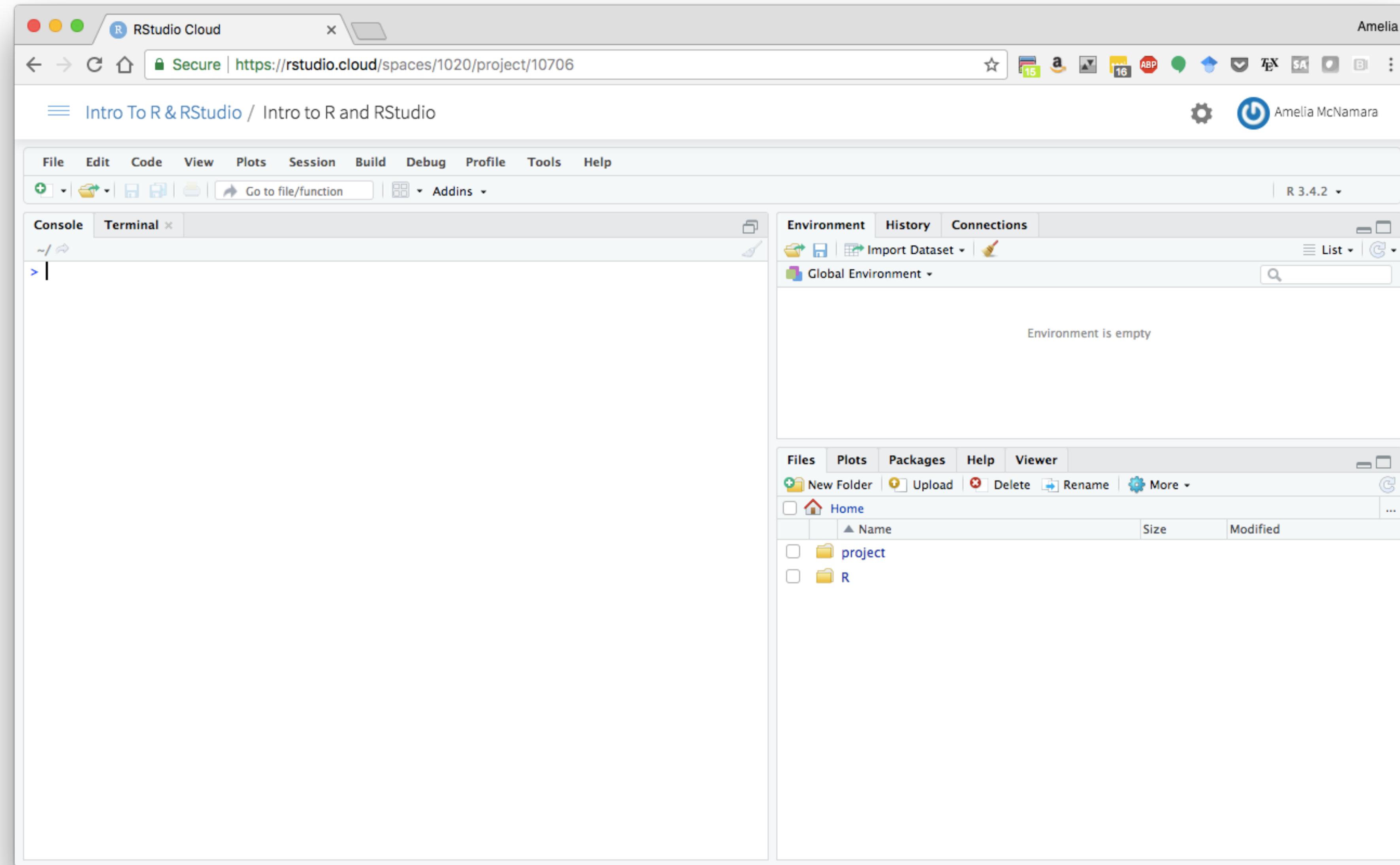
```
1 ---  
2 title: "R Notebook"  
3 output: html_notebook  
4 ---  
5  
6 Text written in **markdown**  
7  
8 ```{r}  
9 # code written in R  
10 (x <- rnorm(7))  
11 ````  
12  
13 Text written in _markdown_  
14  
15 ```{r}  
16 # code written in R  
17 hist(x)  
18 ````  
19  
20 (Top Level) ◊
```

The code editor has syntax highlighting for R and Markdown. A tooltip "Click to run all code chunks above" points to the green play button icon at the top of the code editor. Another tooltip "Click to run code in chunk" points to the green play button icon within the code chunk. The console pane below shows the output of the R code:

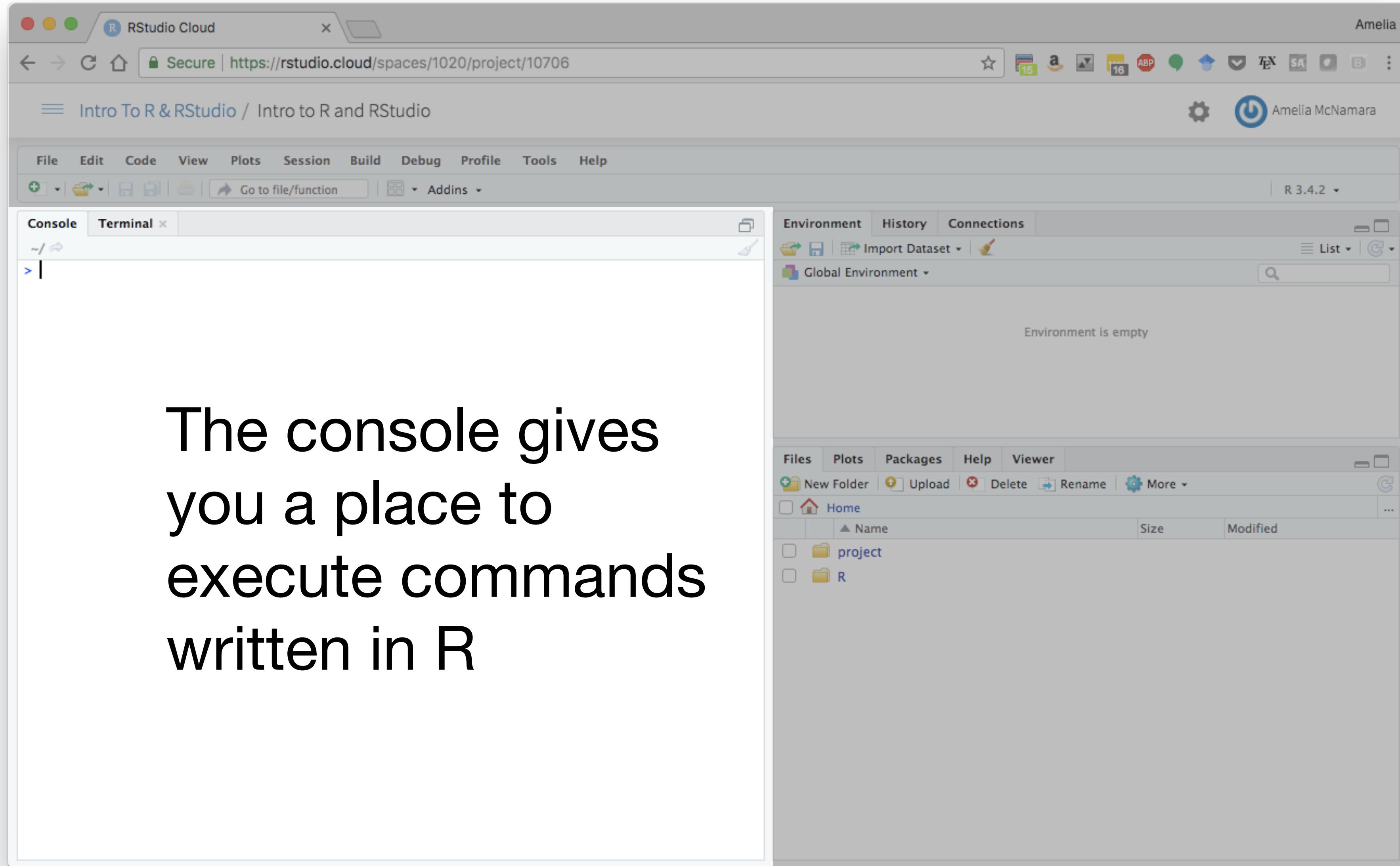
```
[1] -1.2 1.0 -0.5 0.9 -0.6 -1.1 -1.5
```

A tooltip "Code result" points to the console output. The status bar at the bottom right indicates "R Markdown".

# RStudio



# RStudio

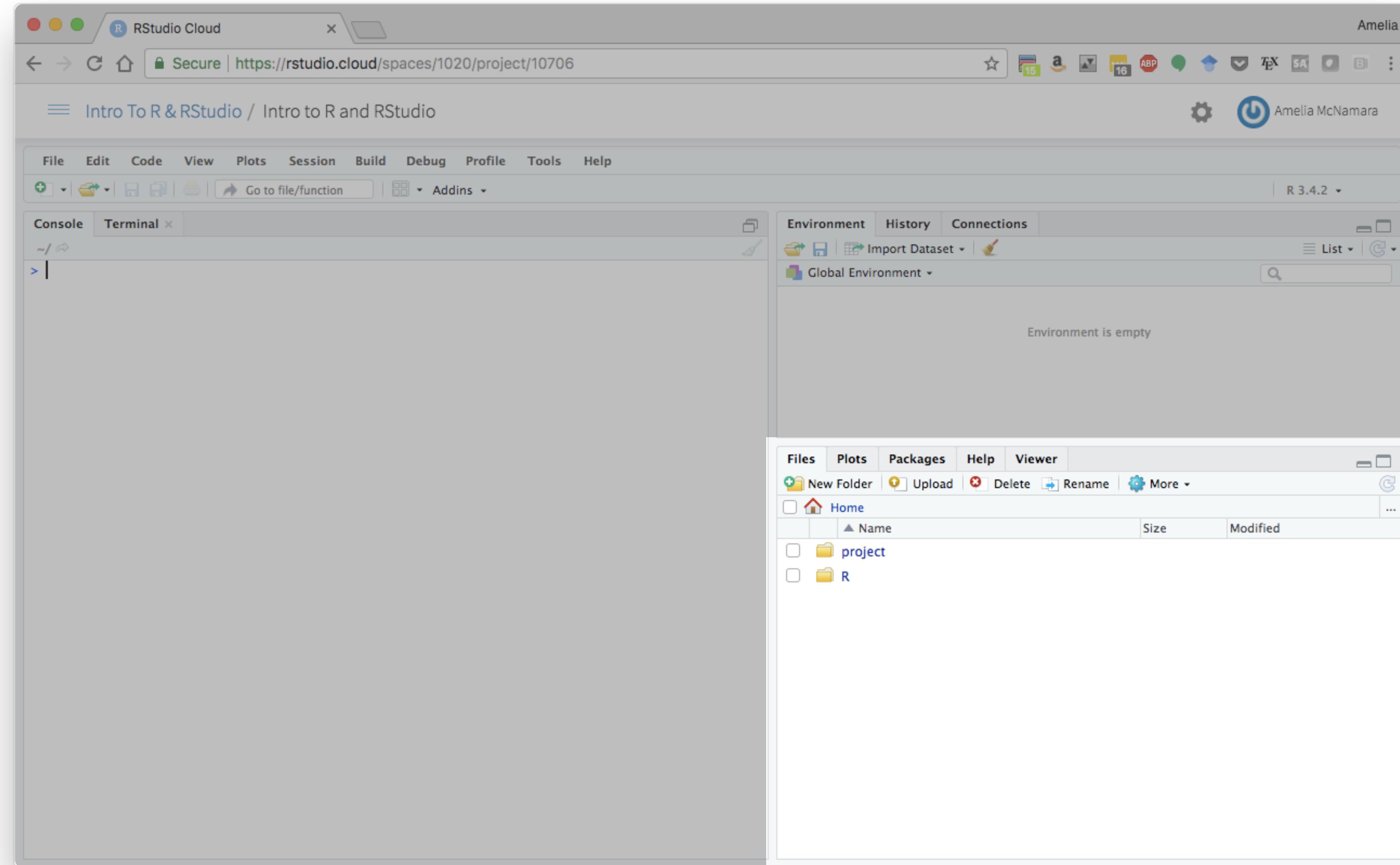


The screenshot shows the RStudio Cloud interface. The top navigation bar includes tabs for File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Help. The main menu bar has options like Go to file/function and Addins. The top right corner shows the user's name, Amelia McNamara. The interface is divided into several panes:

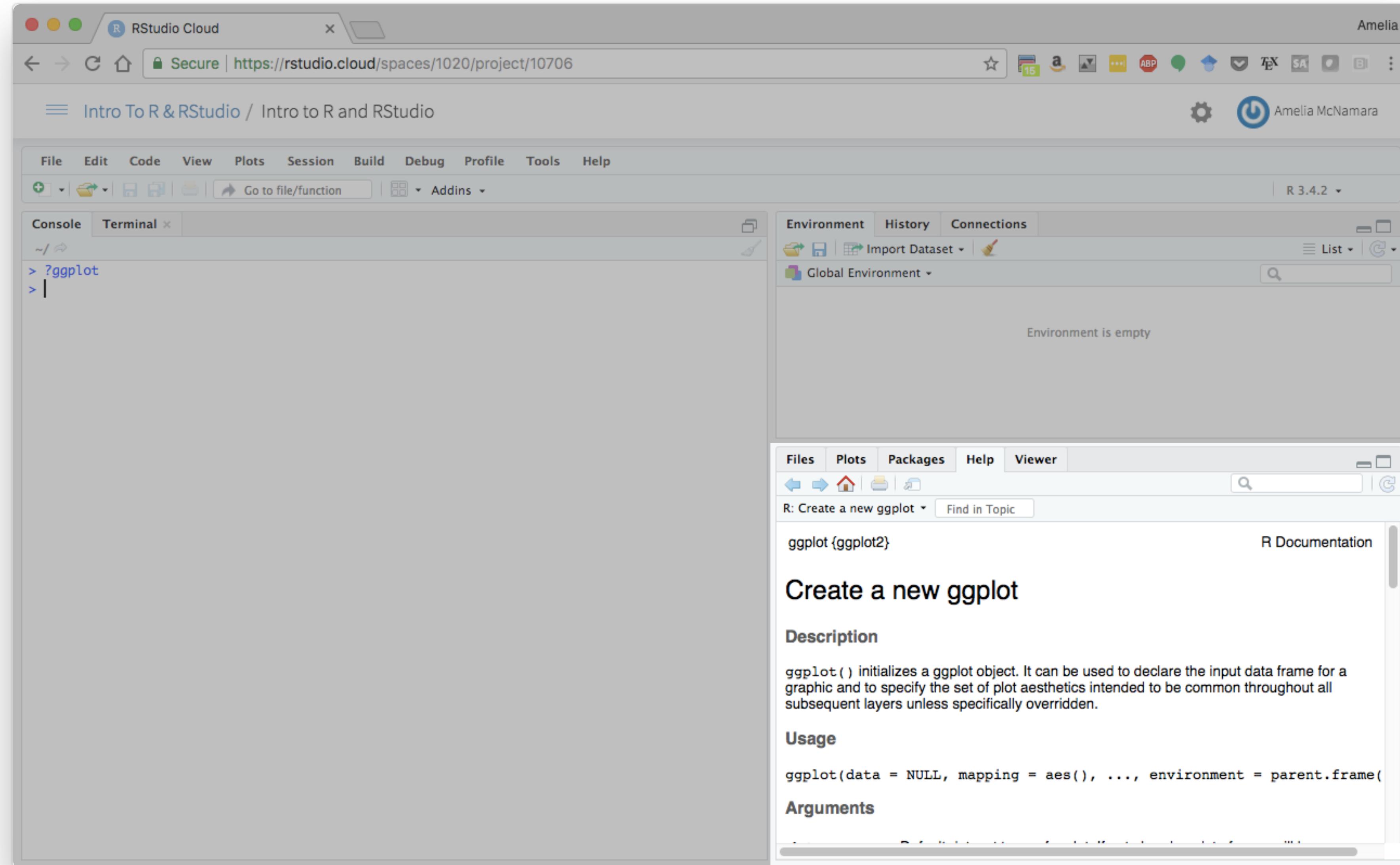
- Console:** Shows a single line starting with '> |'.
- Environment:** Displays the message "Environment is empty".
- Files:** Shows a directory structure with Home, project, and R folders. Actions like New Folder, Upload, Delete, Rename, and More are available.
- Plots:** Currently empty.
- Packages:** Currently empty.
- Help:** Currently empty.
- Viewer:** Currently empty.

**Text overlay:** The central area contains the text: "The console gives you a place to execute commands written in R".

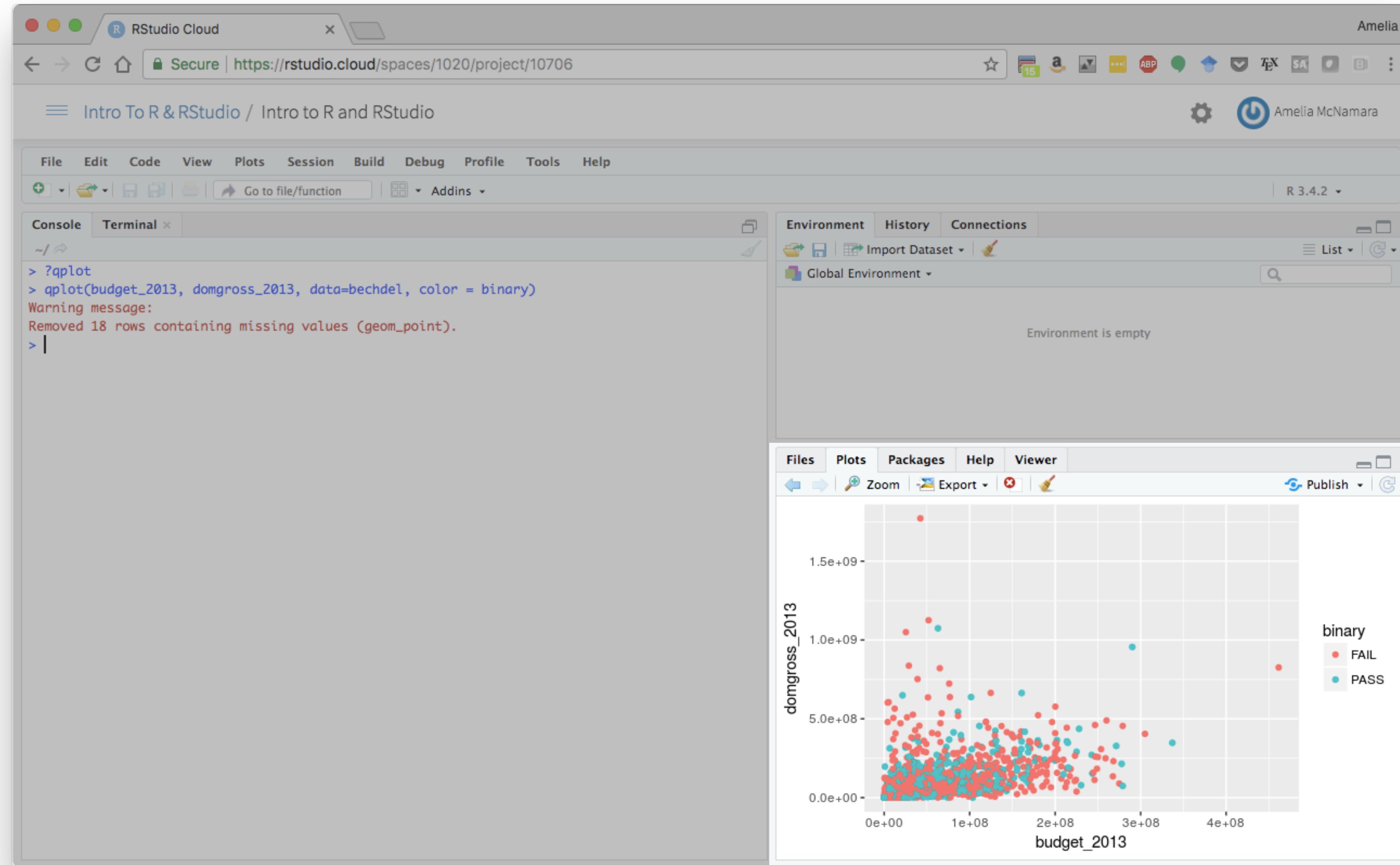
# RStudio



# RStudio



# RStudio



# RStudio

RStudio Cloud | Secure | https://rstudio.cloud/spaces/1020/project/10706

Intro To R & RStudio / Intro to R and RStudio

Console Terminal

```
> ?qplot
> qplot(budget_2013, domgross_2013, data=bechdel, color = binary)
Warning message:
Removed 18 rows containing missing values (geom_point).
> |
```

Environment History Connections

```
dechdel %>% skim(domgross_2013)
library(skimr)
data(bechdel)
bechdel %>% skim(domgross_2013)
bechdel %>% skim(clean_test)
qplot(budget_2013, domgross_2013, data=bechdel, color = binary)
lm(domgross_2013~budget_2013, data=bechdel)
?qplot
qplot(budget_2013, domgross_2013, data=bechdel, color = binary)
```

Files Plots Packages Help Viewer

Zoom Export Publish

domgross\_2013

budget\_2013

binary

- FAIL
- PASS

# RStudio

RStudio Cloud | Secure | https://rstudio.cloud/spaces/1020/project/10706

Intro To R & RStudio / Intro to R and RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Untitled1 x Go to file/function Addins R 3.4.2

```
1 ---  
2 title: "Untitled"  
3 output: html_document  
4 ---  
5  
6 ```{r setup, include=FALSE}  
7 knitr::opts_chunk$set(echo = TRUE)  
8  
9  
10 ## R Markdown  
11  
12 This is an R Markdown document. Markdown is a simple formatting syntax for authoring  
HTML, PDF, and MS Word documents. For more details on using R Markdown see  
http://rmarkdown.rstudio.com.  
13  
14 When you click the **Knit** button a document will be generated that includes both  
2:1 # Untitled
```

Environment History Connections

Global Environment

Environment is empty

Files Plots Packages Help Viewer

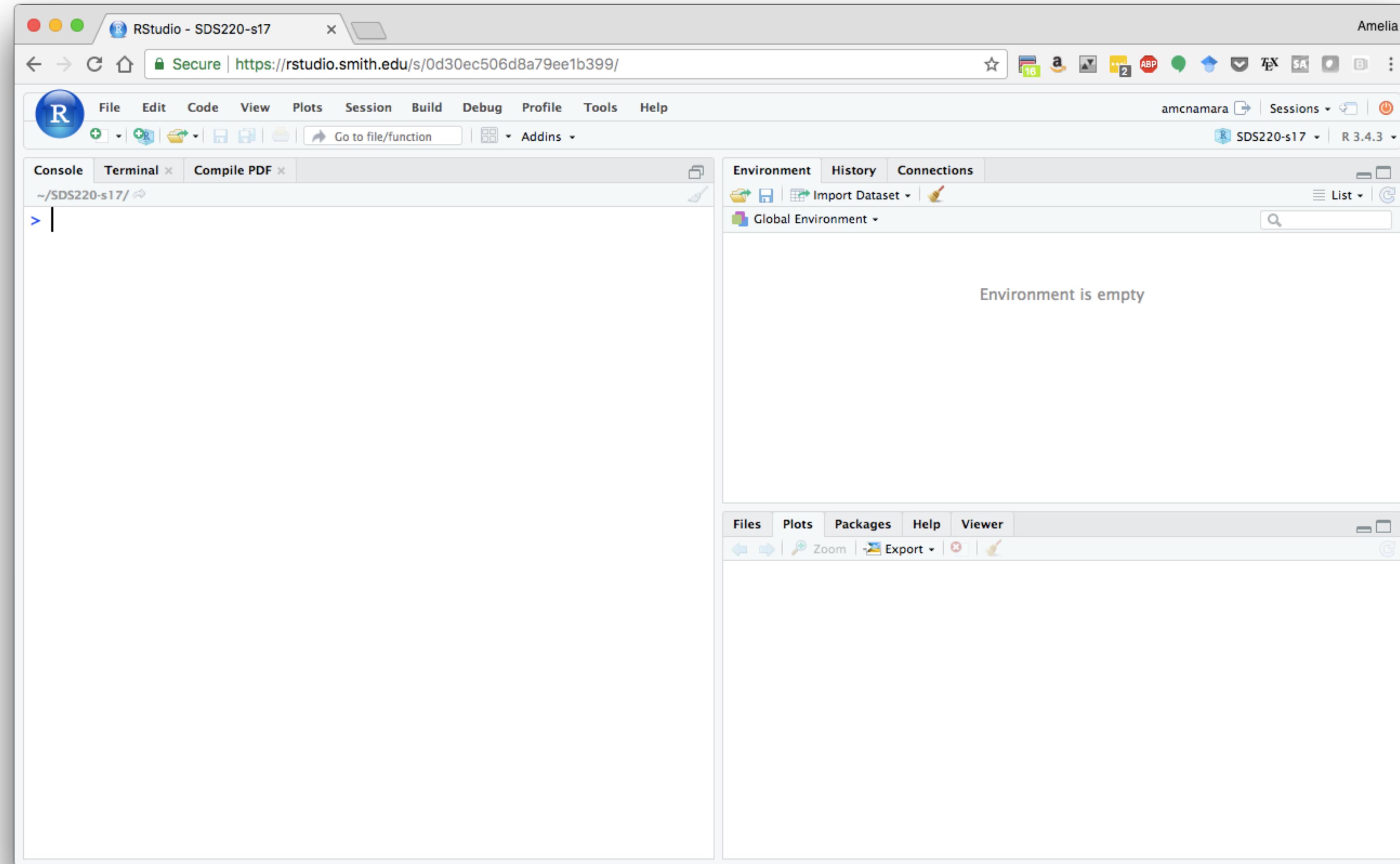
binary

FAIL PASS

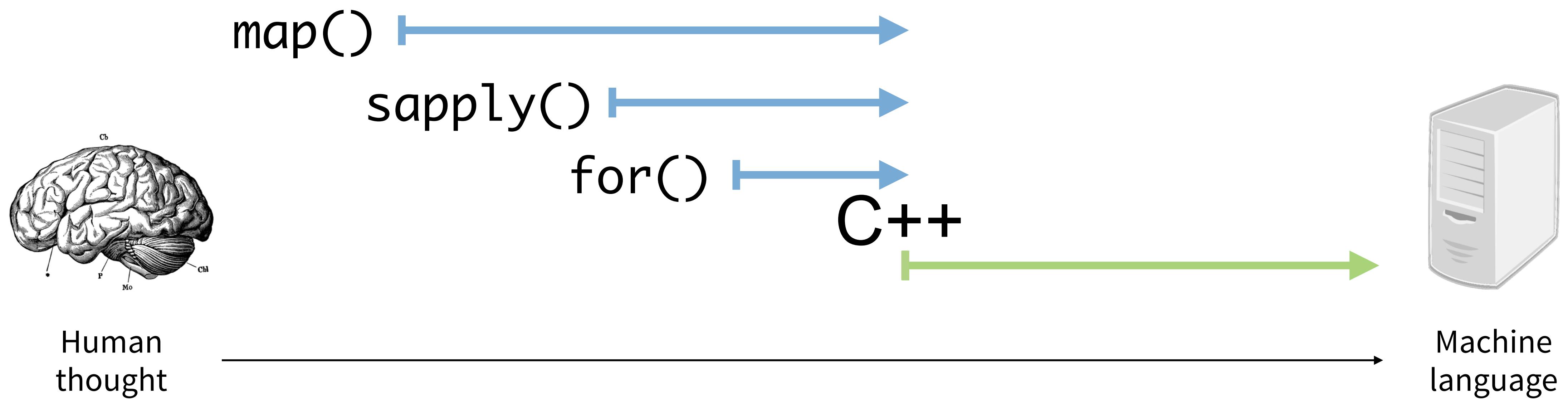
Console Terminal

```
> ?qplot  
> qplot(budget_2013, domgross_2013, data=bechdel, color = binary)  
Warning message:  
Removed 18 rows containing missing values (geom_point).  
>
```

# RStudio



# R - A computer language for scientists



# R Syntax Comparison :: CHEAT SHEET

## Dollar sign syntax

```
goal(data$x, data$y)
```

### SUMMARY STATISTICS:

one continuous variable:  
`mean(mtcars$mpg)`

one categorical variable:  
`table(mtcars$cyl)`

two categorical variables:  
`table(mtcars$cyl, mtcars$am)`

one continuous, one categorical:  
`mean(mtcars$mpg [mtcars$cyl==4])`  
`mean(mtcars$mpg [mtcars$cyl==6])`  
`mean(mtcars$mpg [mtcars$cyl==8])`

### PLOTTING:

one continuous variable:  
`hist(mtcars$disp)`

one categorical variable:  
`barplot(table(mtcars$cyl))`

two continuous variables:  
`plot(mtcars$disp, mtcars$mpg)`

two categorical variables:  
`mosaicplot(table(mtcars$am, mtcars$cyl))`

one continuous, one categorical:  
`histogram(mtcars$disp [mtcars$cyl==4])`  
`histogram(mtcars$disp [mtcars$cyl==6])`  
`histogram(mtcars$disp [mtcars$cyl==8])`

boxplot(mtcars\$disp [mtcars\$cyl==4])  
boxplot(mtcars\$disp [mtcars\$cyl==6])  
boxplot(mtcars\$disp [mtcars\$cyl==8])

### WRANGLING:

subsetting:  
`mtcars [mtcars$mpg > 30, ]`

making a new variable:  
`mtcars$efficient [mtcars$mpg > 30] <- TRUE`  
`mtcars$efficient [mtcars$mpg < 30] <- FALSE`

## Formula syntax

```
goal(y~x|z, data=data, group=w)
```

### SUMMARY STATISTICS:

one continuous variable:  
`mosaic::mean(~mpg, data=mtcars)`

one categorical variable:  
`mosaic::tally(~cyl, data=mtcars)`

two categorical variables:  
`mosaic::tally(cyl~am, data=mtcars)`

one continuous, one categorical:  
`mosaic::mean(mpg~cyl, data=mtcars)`

tilde

### PLOTTING:

one continuous variable:  
`lattice::histogram(~disp, data=mtcars)`

one categorical variable:  
`lattice::bwplot(~disp, data=mtcars)`

one categorical variable:  
`mosaic::bargraph(~cyl, data=mtcars)`

two continuous variables:  
`lattice::xyplot(mpg~disp, data=mtcars)`

two categorical variables:  
`mosaic::bargraph(~am, data=mtcars, group=cyl)`

one continuous, one categorical:  
`lattice::histogram(~disp|cyl, data=mtcars)`

lattice::bwplot(cyl~disp, data=mtcars)

The variety of R syntaxes give  
you many ways to “say” the  
same thing

read across the cheatsheet to see how different  
syntaxes approach the same problem

## Tidyverse syntax

```
data %>% goal(x)
```

### SUMMARY STATISTICS:

one continuous variable:  
`mtcars %>% dplyr::summarize(mean(mpg))`

one categorical variable:  
`mtcars %>% dplyr::group_by(cyl) %>%  
dplyr::summarize(n())`

the pipe

two categorical variables:  
`mtcars %>% dplyr::group_by(cyl, am) %>%  
dplyr::summarize(n())`

one continuous, one categorical:  
`mtcars %>% dplyr::group_by(cyl) %>%  
dplyr::summarize(mean(mpg))`

**PLOTTING:**  
one continuous variable:  
`ggplot2::qplot(x=mpg, data=mtcars, geom = "histogram")`

`ggplot2::qplot(y=disp, x=1, data=mtcars, geom="boxplot")`

one categorical variable:  
`ggplot2::qplot(x=cyl, data=mtcars, geom="bar")`

two continuous variables:  
`ggplot2::qplot(x=disp, y=mpg, data=mtcars, geom="point")`

two categorical variables:  
`ggplot2::qplot(x=factor(cyl), data=mtcars, geom="bar") +  
facet_grid(.~am)`

one continuous, one categorical:  
`ggplot2::qplot(x=disp, data=mtcars, geom = "histogram") +  
facet_grid(.~cyl)`

`ggplot2::qplot(y=disp, x=factor(cyl), data=mtcars,  
geom="boxplot")`

**WRANGLING:**  
subsetting:  
`mtcars %>% dplyr::filter(mpg > 30)`

making a new variable:  
`mtcars <- mtcars %>%  
dplyr::mutate(efficient = if_else(mpg > 30, TRUE, FALSE))`



# Data Science in the

Amelia McNamara

[amelia.mn](http://amelia.mn)

@AmeliaMN



*Data Science in the tidyverse* is licensed under a [Creative Commons Attribution 4.0 International License](#). Based on work at <https://github.com/cwickham/data-science-in-tidyverse> and <https://github.com/rstudio-education/master-the-tidyverse>