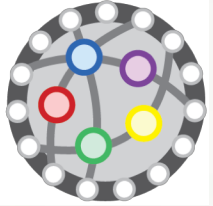


ELAC 2019 INTRODUCTION TO DATA SCIENCE

Day 2

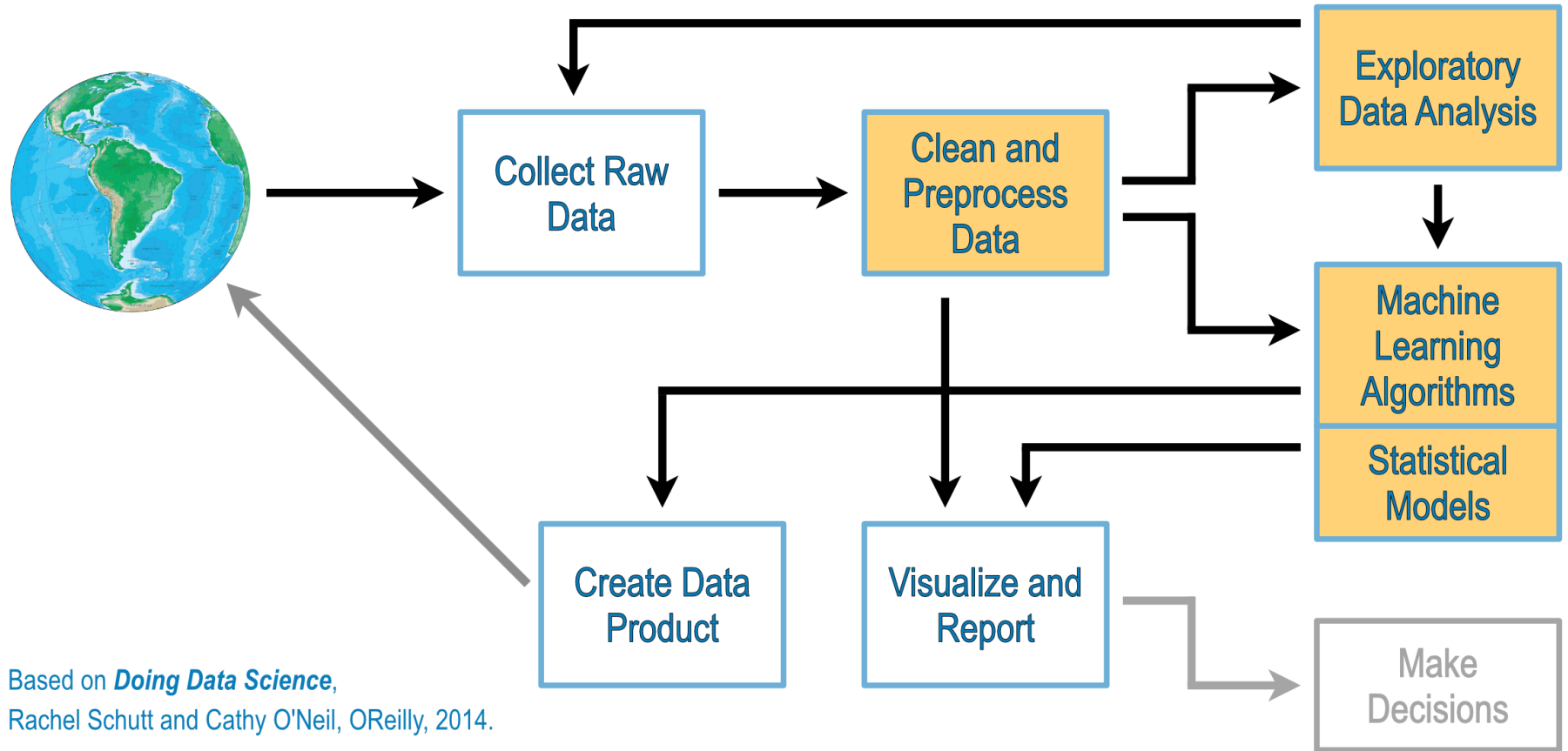
Rafael Santos - rafael.santos@inpe.br
www.lac.inpe.br/~rafael.santos/talks.html

Introduction to Data Science

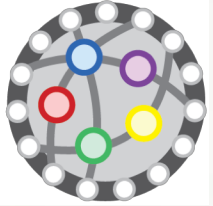


About this Lecture

Where are we?



Introduction to Data Science

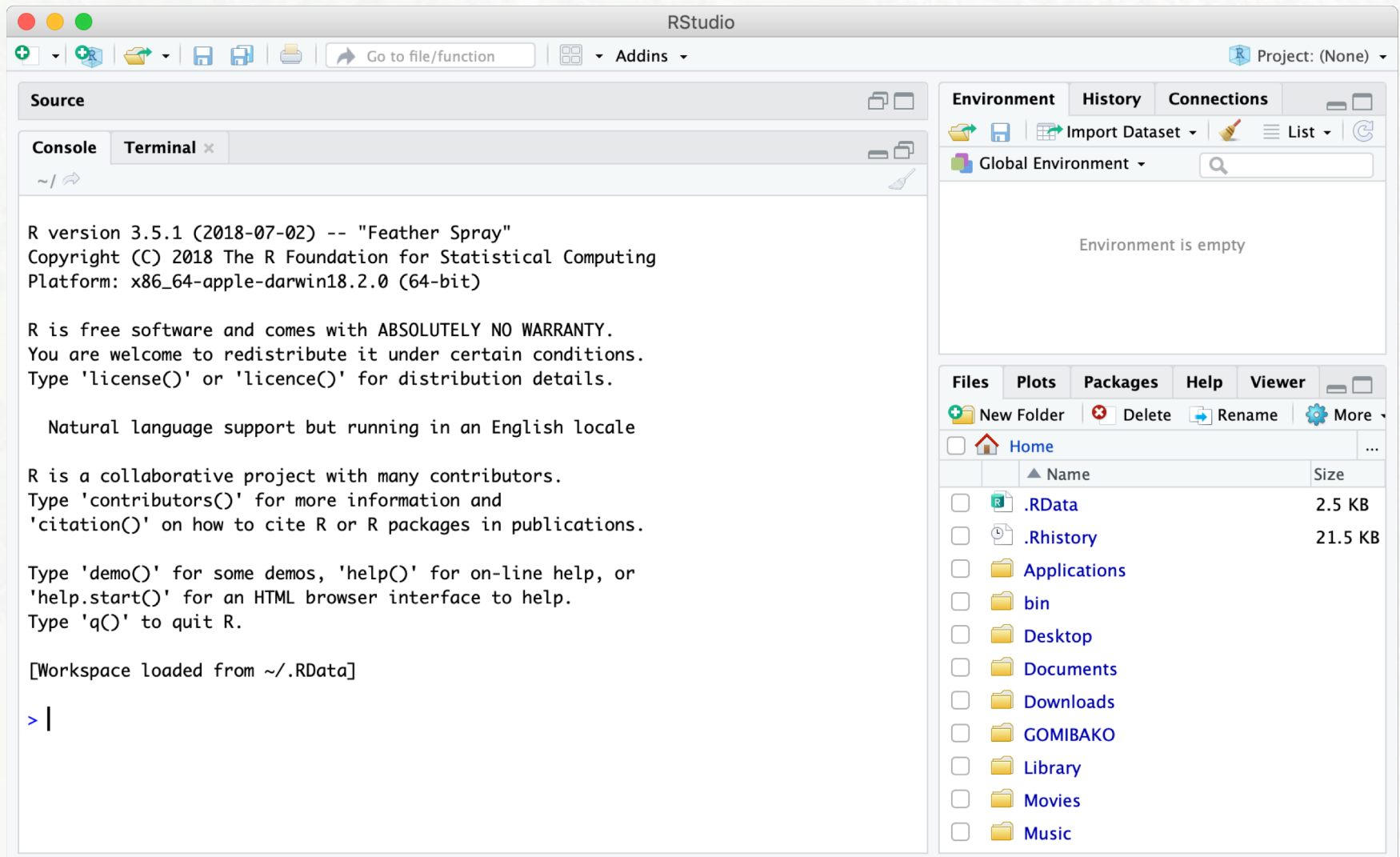


R/RStudio

We'll start with R/RStudio

- Free language, free IDE.
- All major OSs.
- Focus on scripts.
 - ▣ Fancy stuff later!
- Easy to edit/run code, import packages, etc.

Run R/RStudio



The screenshot displays the RStudio application window. The top menu bar includes options like '+', 'Go to file/function', 'Addins', and 'Project: (None)'. The main interface is divided into several panes:

- Source:** Currently empty.
- Console:** Shows the R version 3.5.1 (2018-07-02) -- "Feather Spray" and copyright information. It also displays the R license text and instructions for using help functions.
- Environment:** Shows 'Global Environment' and indicates that the environment is empty.
- Files:** A file explorer showing the home directory with a list of files and folders, including .RData (2.5 KB) and .Rhistory (21.5 KB).

```
R version 3.5.1 (2018-07-02) -- "Feather Spray"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin18.2.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

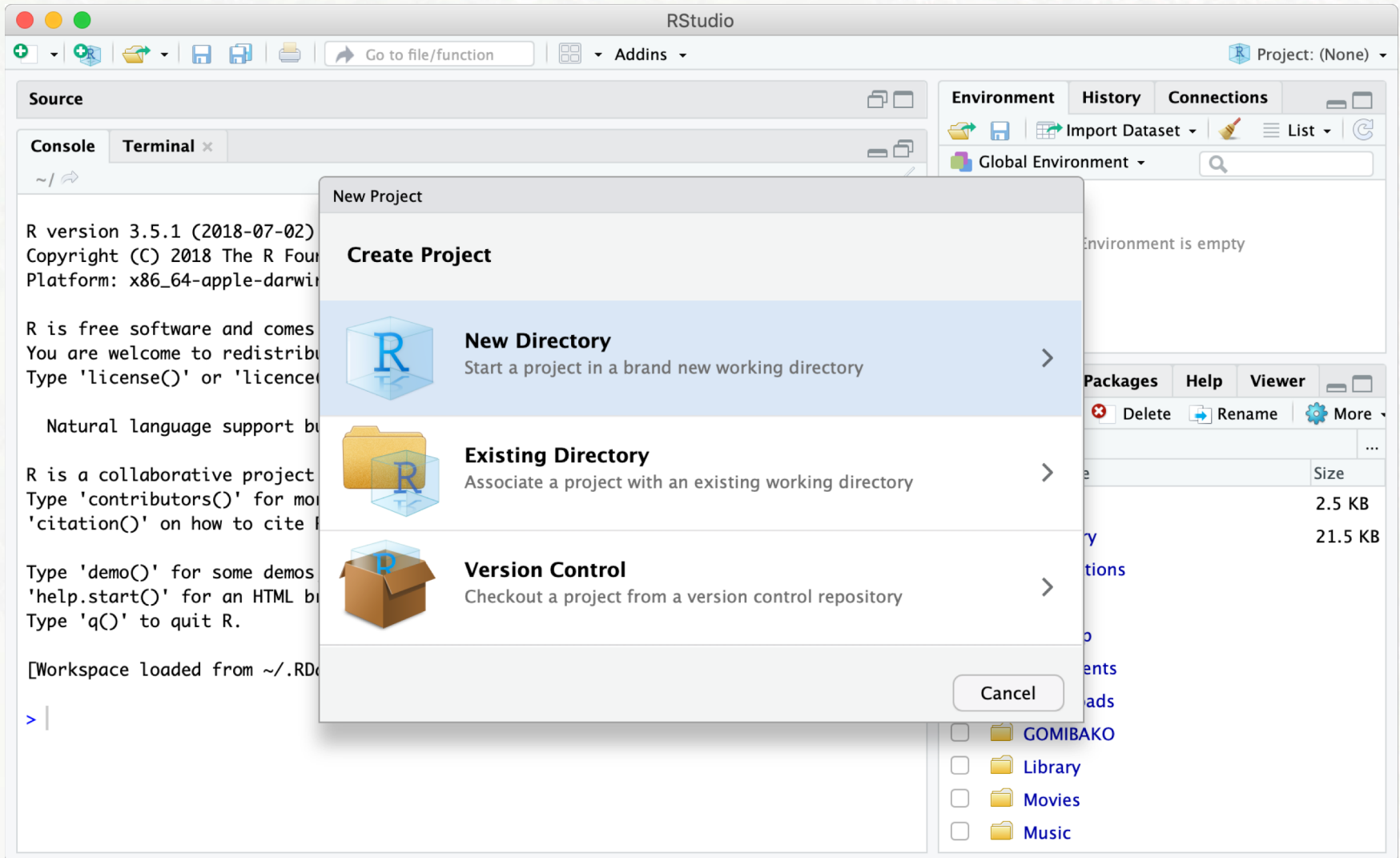
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]
> |
```

Name	Size
.RData	2.5 KB
.Rhistory	21.5 KB
Applications	
bin	
Desktop	
Documents	
Downloads	
GOMIBAKO	
Library	
Movies	
Music	

Create a Project (File/New Project)



Create a Project (File/New Project)

The screenshot shows the RStudio interface with the 'New Project' dialog box open. The dialog box has a 'Back' button and a 'Project Type' section. The 'Project Type' section lists several options, each with a right-pointing arrow:

- New Project
- R Package
- Shiny Web Application
- R Package using Rcpp
- R Package using RcppArmadillo
- R Package using RcppEigen
- R Package using devtools

The 'New Project' option is currently selected. The background shows the RStudio console with the following text:

```
R version 3.5.1 (2018-07-02)
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin18.0.0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]
```

The RStudio interface also shows the 'Environment' pane with 'Global Environment' selected, and the 'Packages' pane with a list of installed packages including 'GOMIBAKO', 'Library', 'Movies', and 'Music'.

Create a Project (File/New Project)

RStudio

Project: (None)

Source

Environment History Connections

Console Terminal x

~/

R version 3.5.1 (2018-07-02)
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin18.0.0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~ /.RData]

> |

Environment is empty

Global Environment

Packages Help Viewer

Delete Rename More

	Size
	2.5 KB
	21.5 KB

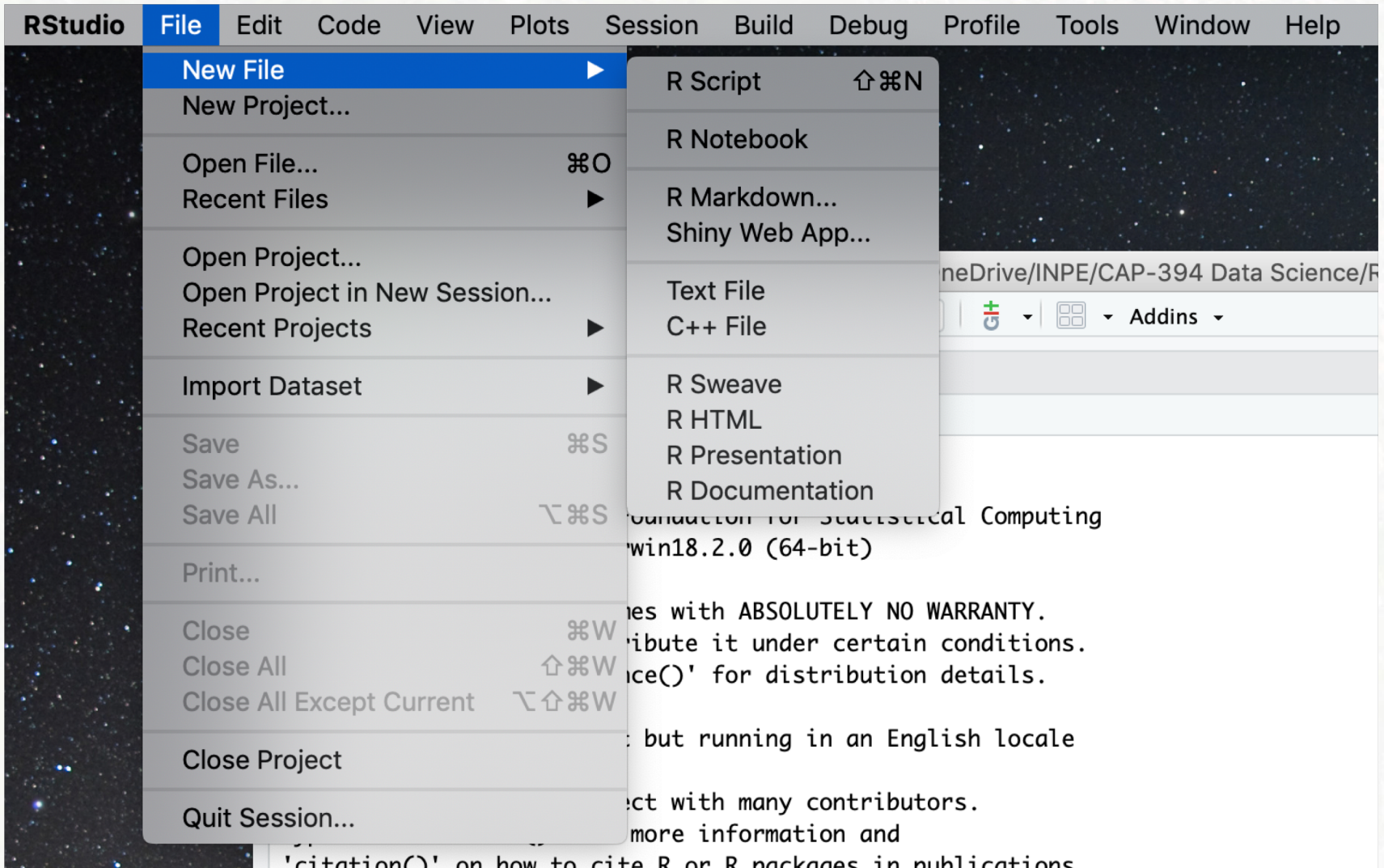
GOMIBAKO

Library

Movies

Music

Create a R Script (File/New File)



Editing/Running a R Script

The screenshot displays the RStudio environment with the following components:

- Source Editor:** Contains a script with four lines of code:

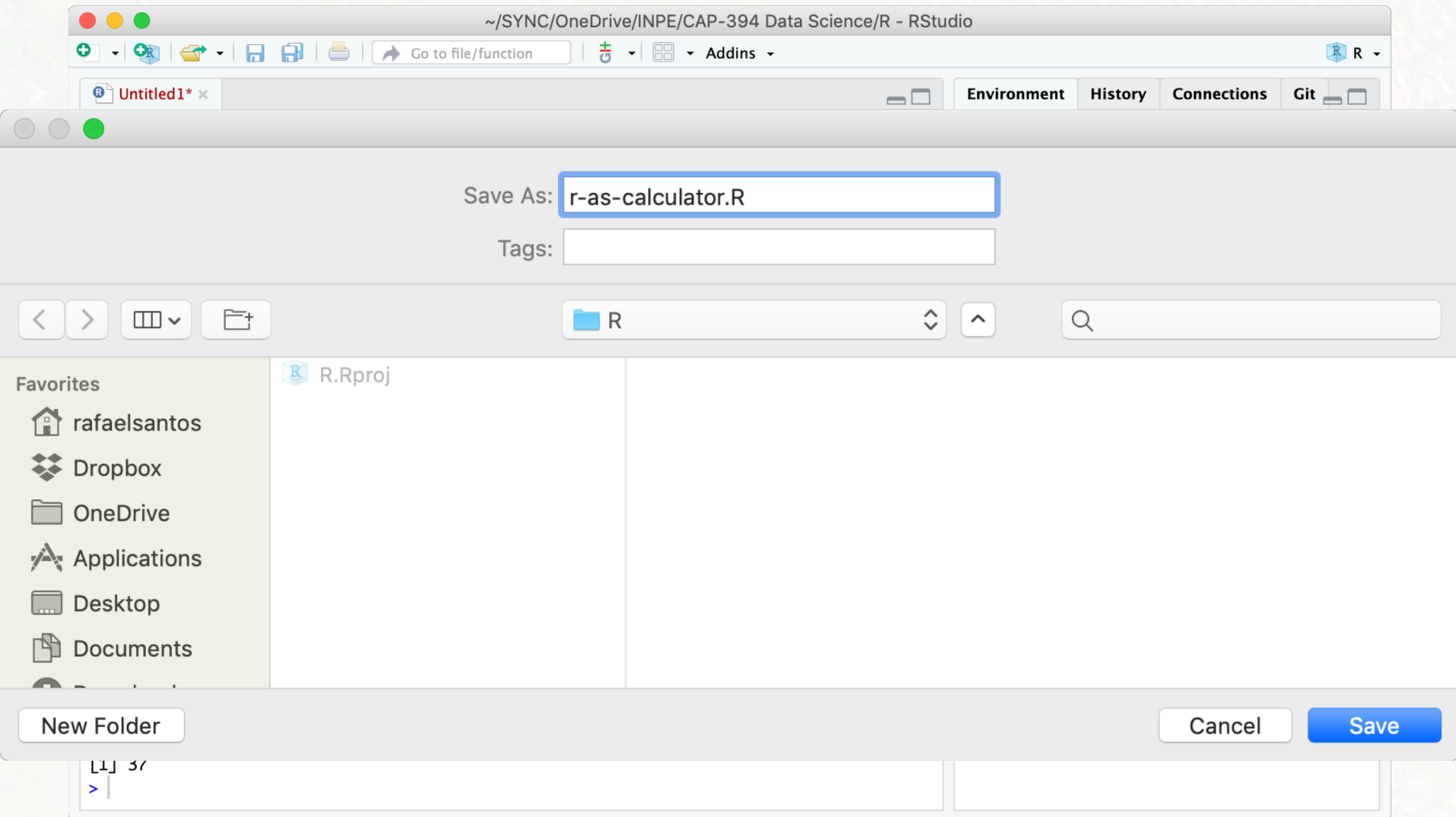
```
1 a <- 20
2 b <- 17
3 a+b
4
```
- Environment Pane:** Shows the current environment with variables:

Variable	Value
a	20
b	17
- Files Pane:** Shows the project structure:

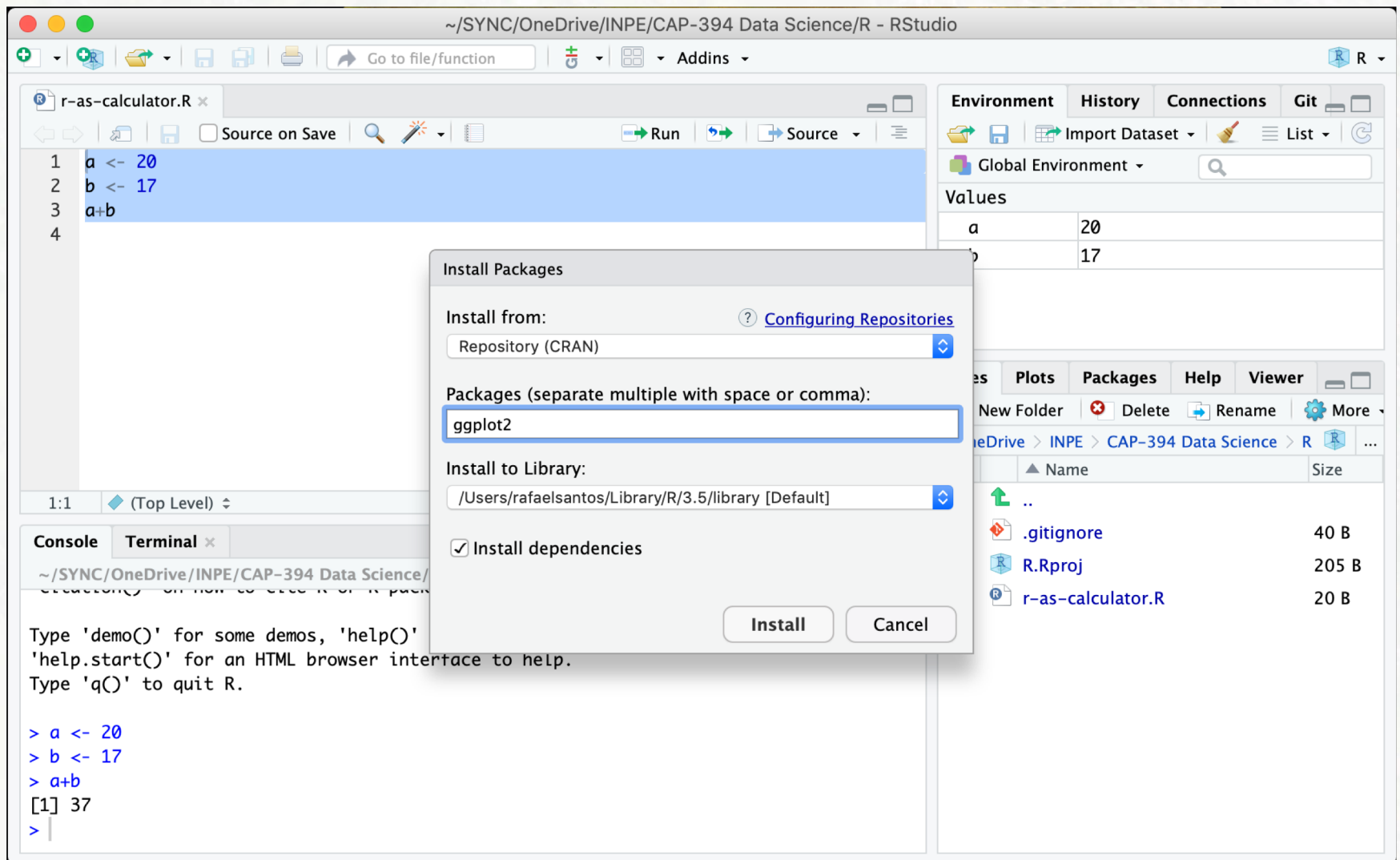
Name	Size
..	
.gitignore	40 B
R.Rproj	205 B
- Console:** Shows the execution output:

```
> a <- 20
> b <- 17
> a+b
[1] 37
>
```

Saving a R Script



Installing Packages in RStudio



The screenshot shows the RStudio interface with the following components:

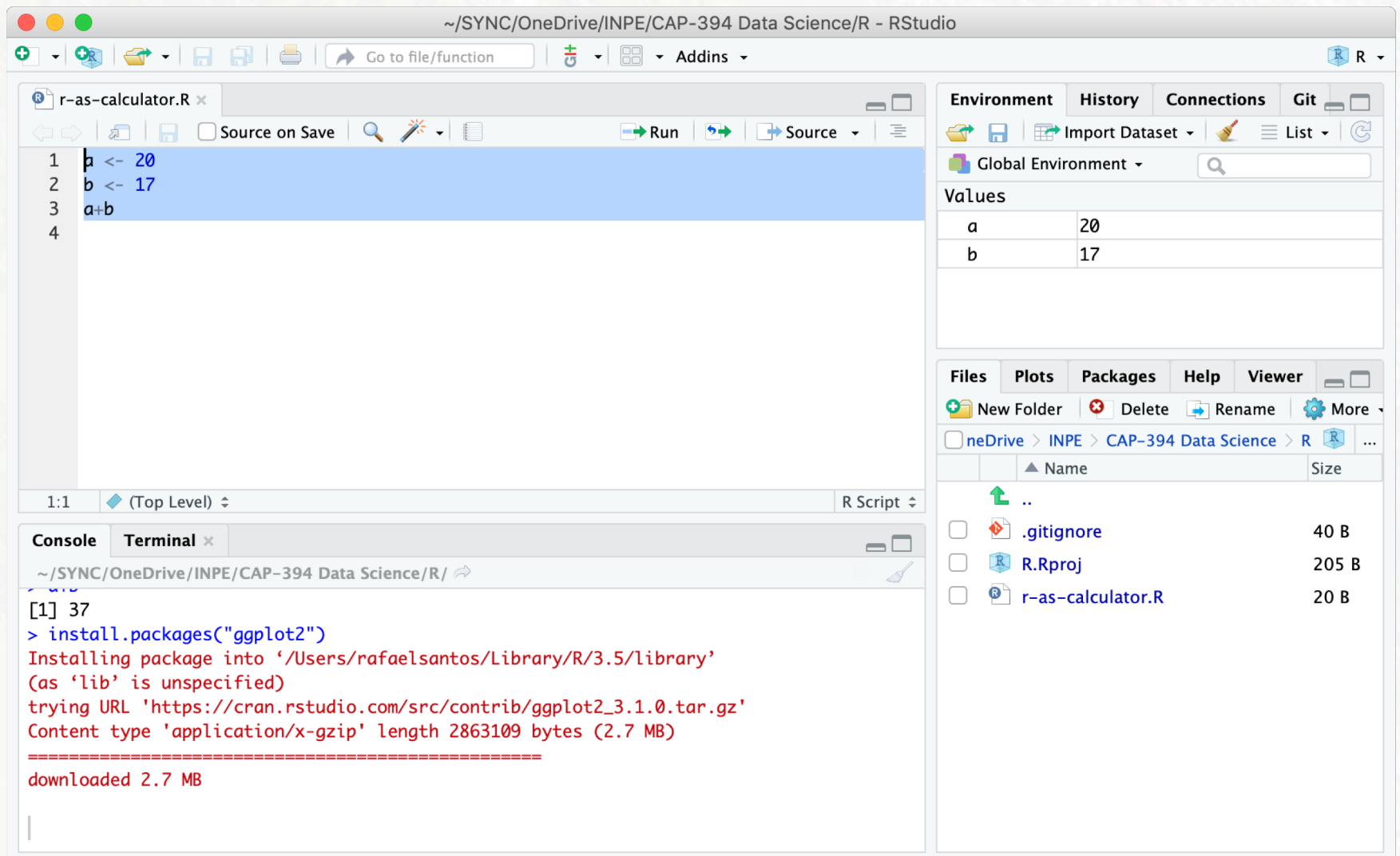
- Script Editor:** A file named `r-as-calculator.R` containing the following R code:

```
1 a <- 20
2 b <- 17
3 a+b
4
```
- Console:** Shows the execution of the script:

```
> a <- 20
> b <- 17
> a+b
[1] 37
> |
```
- Environment Pane:** Displays the current environment with the following values:

Variable	Value
a	20
b	17
- Files Pane:** Shows the project structure:
 - ..
 - .gitignore (40 B)
 - R.Rproj (205 B)
 - r-as-calculator.R (20 B)
- Install Packages Dialog:** A modal dialog box with the following fields and options:
 - Install from:** Repository (CRAN)
 - Packages (separate multiple with space or comma):** ggplot2
 - Install to Library:** /Users/rafaelsantos/Library/R/3.5/library [Default]
 - Install dependencies
 - Buttons: Install, Cancel

Installing Packages in RStudio



The screenshot shows the RStudio interface with a script editor, console, and file browser. The script editor contains the following code:

```
1 a <- 20
2 b <- 17
3 a+b
4
```

The console shows the output of the script and the installation of the `ggplot2` package:

```
[1] 37
> install.packages("ggplot2")
Installing package into '/Users/rafaelsantos/Library/R/3.5/library'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/src/contrib/ggplot2_3.1.0.tar.gz'
Content type 'application/x-gzip' length 2863109 bytes (2.7 MB)
=====
downloaded 2.7 MB
```

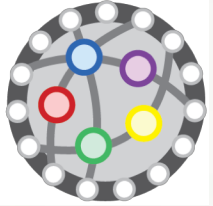
The file browser on the right shows the current directory structure:

Name	Size
..	
.gitignore	40 B
R.Rproj	205 B
r-as-calculator.R	20 B

Installing Packages in RStudio

- Usually everything works OK but...
 - ▣ Installation may not be as smooth as shown.
 - ▣ If R version changes you may need to reinstall.

Introduction to Data Science



Introduction to R

“R Programming”



Introduction to R

<https://github.com/rafaeldcsantos/IntroDataScience>

□ 01-Variables

- ▣ Pros and cons of starting scripts with `rm(list=ls())`

□ 02-PrimitiveTypes

- ▣ Primitives are actually Vectors!
- ▣ `[1] 10` -> Index of first element of vector.

□ 03-Vectors

□ 04-OperationsWithVectors

□ 05-Factors

- ▣ Factors are categorical variables, very important for classification!

Introduction to R

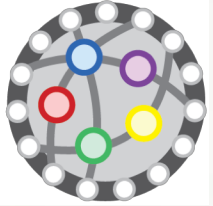
□ **06-DataFrames**

- DataFrames are the essential data structure for Data Science.

□ **07-Control**

- Remember: there are other ways to do conditionals and loops

Introduction to Data Science



Raw Data and Tidy Data





Raw Data

- Data in databases, spreadsheets...
- Images, videos, audio...
- Time series...
- Logs, text, JSON files, XML files...

Based on Coursera's "Getting and Cleaning Data" course.

Tidy Data

- One table with all the data (or linked tables).
 - ▣ Each variable in its column.
 - ▣ Each observation in its row.
 - ▣ Variable names in the first row, with good, clear names.
- “Tidiness” is not an absolute feature!
 - ▣ Depends on what we have and what we want to do.

  incidentLocation	  location
4000 MT PLEASANT AV	(39.2910056,-76.5636502)
1000 W BALTIMORE ST	(39.2889640,-76.6339440)
1000 STAMFORD RD	(39.2963480,-76.7101510)

Based on Coursera’s “Getting and Cleaning Data” course.

Raw Data to Tidy Data

- Create a Code Book that describes how we can get from Raw Data to Tidy Data.
- A simple (formatted) text file with:
 - ▣ Sources for the raw data.
 - ▣ More detail on the variables.
 - ▣ What was {selected, enhanced, preprocessed} and how.
 - ▣ Instruction on how the data was processed.
- **Code Books essential to reproducibility!**

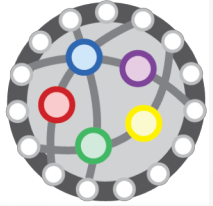
Based on Coursera's "Getting and Cleaning Data" course.

Introduction to R (Part II)

<https://github.com/rafaeldcsantos/IntroDataScience>

- 08-DateTime
- 09-DownloadingFiles
- 10-Strings
- 11-TidyData

Introduction to Data Science



Help!

Help!

□ help(“keyword”)

The screenshot displays the RStudio interface. The top toolbar includes icons for file operations and a search bar labeled 'Go to file/function'. The main editor window shows a script named 'r-as-a-calculator.R' with the following code:

```
1 # Variable assignment with <- or =
2 a <- 12
3 b <- 26
4 # Simple arithmetic expressions
5 a+b
6 max(a,b)
7 # More operators
8 x = 7
9 y = 3
10 x ^ y
11 x %% y # remainder
12 x %/% y # integer division
13 (x %% y) + y * ( x %/% y ) # x
14
```

The console at the bottom left shows the command `> help("sin")` and a prompt `>`. The right-hand pane is split into two sections. The top section, titled 'Environment', shows the 'Global Environment' with a search bar and a table of values:

Values	
a	12
b	26
x	7
y	3

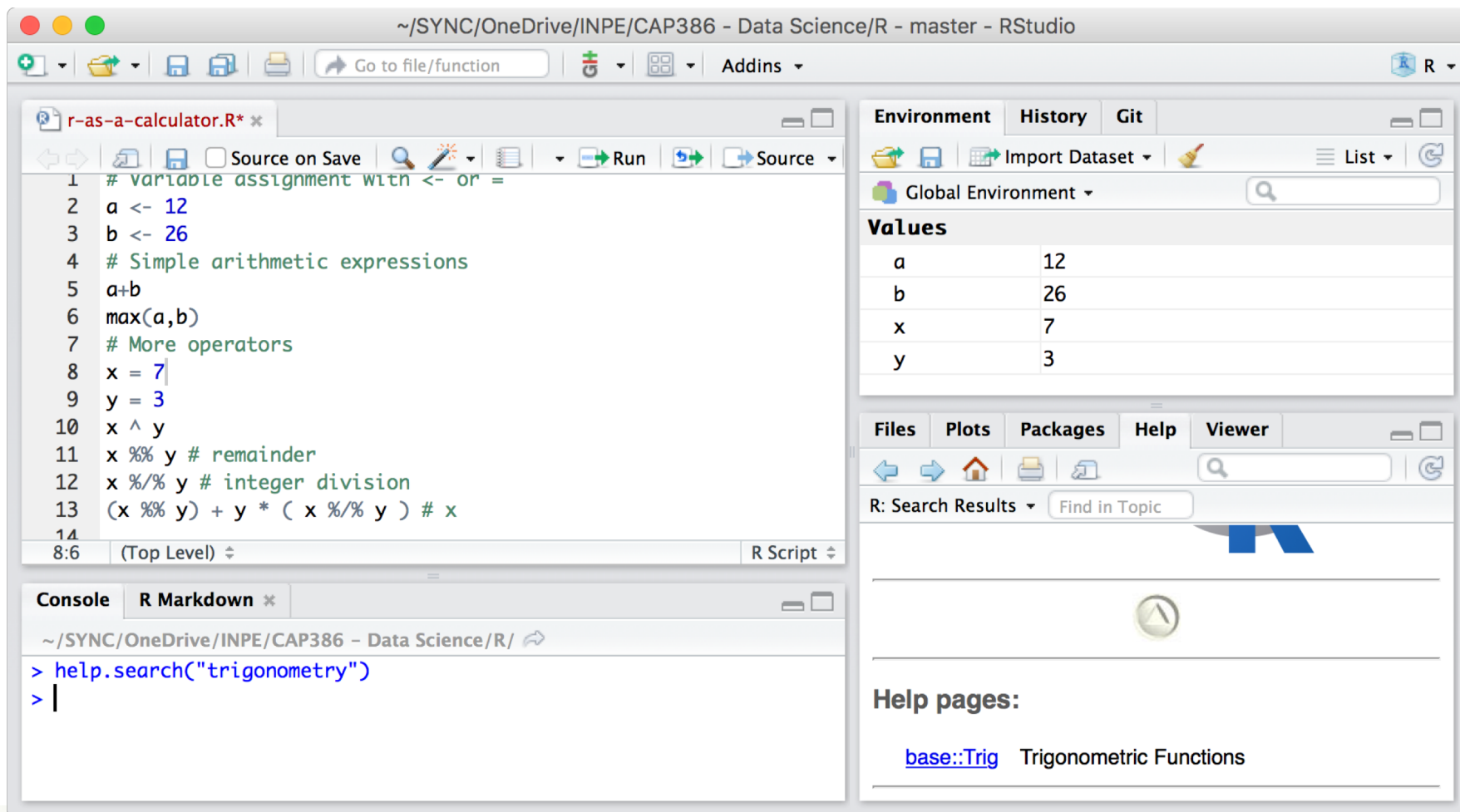
The bottom section, titled 'Help', shows the 'R: Trigonometric Functions' help page. The page title is 'Trigonometric Functions' and it includes a 'Description' section:

Description

These functions give the obvious trigonometric functions. They respectively compute the cosine, sine, tangent, arc-cosine, arc-sine, arc-tangent, and the two-argument arc-

Help!

- `help.search("keyword")` (see also `apropos("keyword")`)



The screenshot displays the RStudio interface. The main editor window shows an R script named `r-as-a-calculator.R` with the following code:

```
1 # Variable assignment with <- or =
2 a <- 12
3 b <- 26
4 # Simple arithmetic expressions
5 a+b
6 max(a,b)
7 # More operators
8 x = 7
9 y = 3
10 x ^ y
11 x %% y # remainder
12 x %/% y # integer division
13 (x %% y) + y * ( x %/% y ) # x
14
```

The console at the bottom left shows the command `> help.search("trigonometry")` being entered.

The right-hand pane is split into two sections. The top section, titled "Environment", shows the "Global Environment" with the following values:

Variable	Value
a	12
b	26
x	7
y	3

The bottom section, titled "Help", shows the "R: Search Results" pane with a search bar and a list of help pages. The first result is:

[base::Trig](#) Trigonometric Functions

Help!

- Much better options, specially for usage and tips and/or more complex scripts:
 - rseek.org
 - stackoverflow.com
- Can I do *X* with R?
 - <https://cran.r-project.org/>
 - <https://cran.r-project.org/web/views/>

Cutting corners to meet arbitrary management deadlines



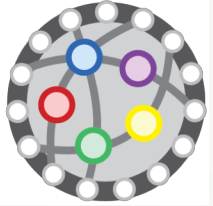
Essential

Copying and Pasting
from Stack Overflow

O'REILLY®

*The Practical Developer
@ThePracticalDev*

Introduction to Data Science



References

References

Proven Recipes for Data Analysis, Statistics, and Graphics



R Cookbook

O'REILLY®

Paul Teetor

2nd Edition



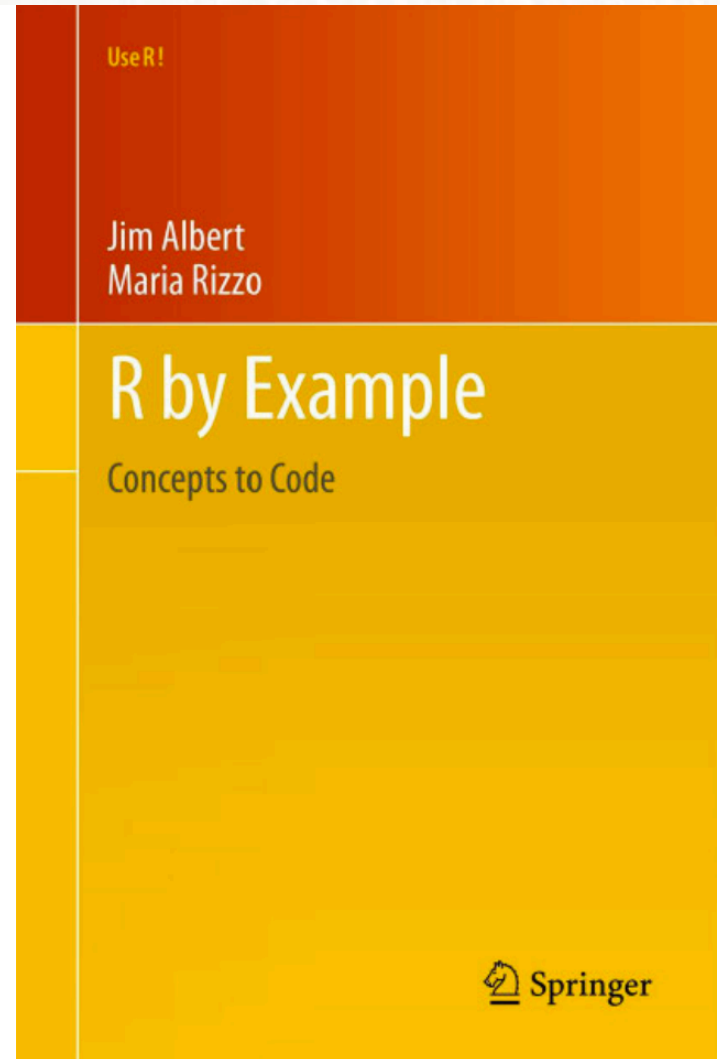
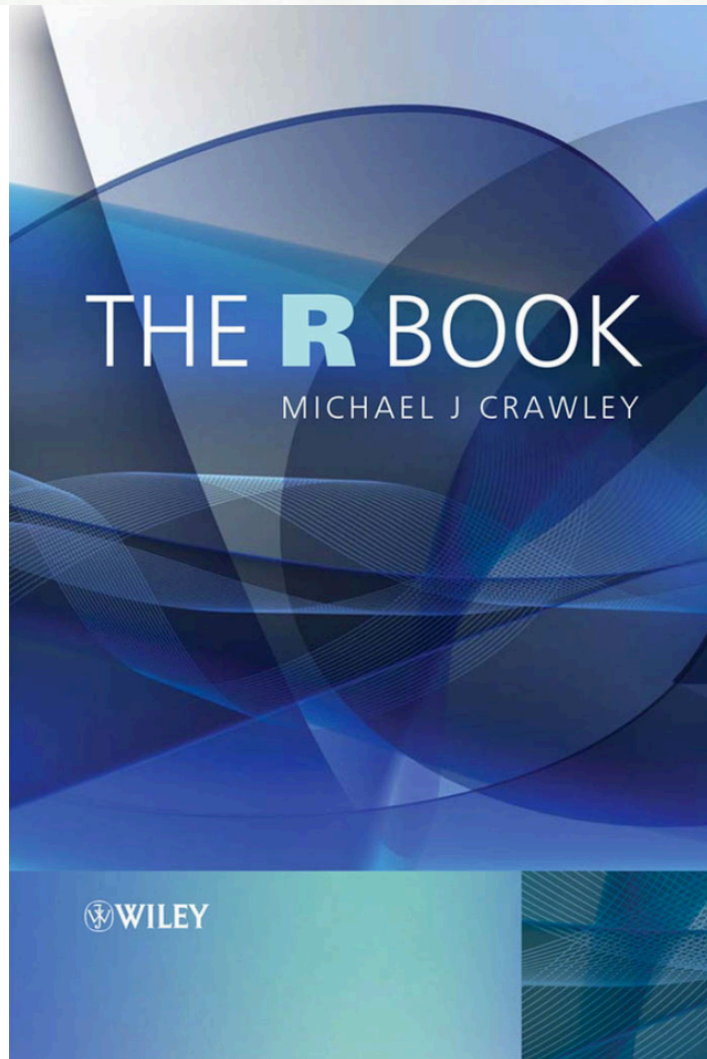
R IN A NUTSHELL

A Desktop Quick Reference

O'REILLY®

Joseph Adler

References

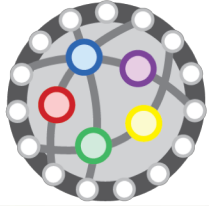


Shameless Advertising

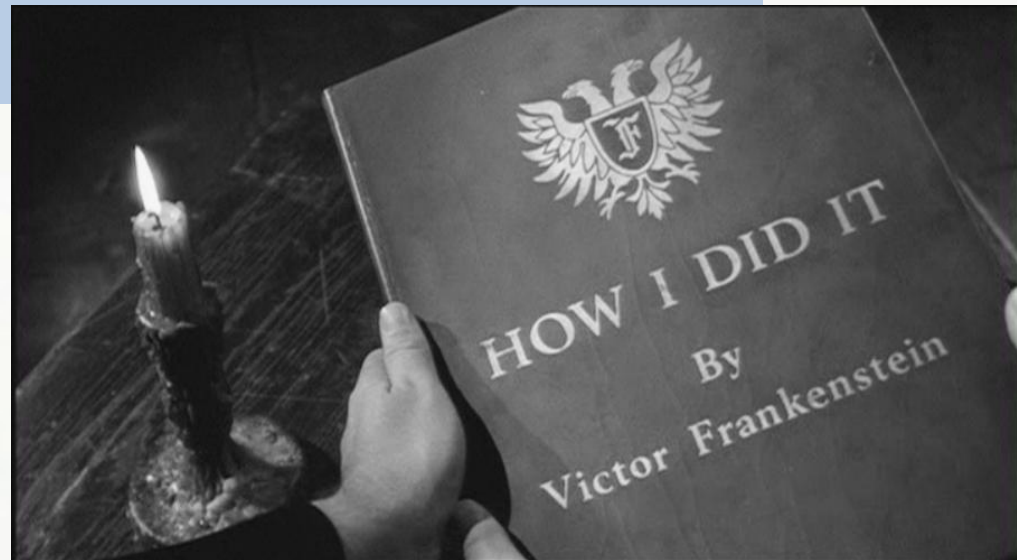
- Applied Computing Graduate Program at INPE:
 - http://www.inpe.br/pos_graduacao/cursos/cap/
- CAP's Annual Workshop (September 2019):
 - <http://www.inpe.br/worcap/>
- Grants!

rafael.santos@inpe.br

Introduction to Data Science



Publishing R code and results on GitHub



Did you do this first?

RStudio

Project: (None)

Environment History Connections

Import Dataset List

Global Environment

New Project

Back Create New Project

Directory name: R

Create project as subdirectory of: ~/SYNC/OneDrive/INPE/CAP-394 Data Science Browse...

Create a git repository

Open in new session

Create Project Cancel

R version 3.5.1 (2018-07-02)
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin18.0.0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help.start()' for an HTML browser interface to help,
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]

> |

Size

2.5 KB

21.5 KB

tions

o

ents

ads

GOMIBAKO

Library

Movies

Music

Then do this (on github.com):

Create a new repository

A repository contains all project files, including the revision history.

Owner

 rafaeldcsantos ▾


Repository name *

IntroDataScience ✓

Great repository names are short and memorable. Need inspiration? How about [expert-waffle?](#)

Description (optional)

 **Public**
Anyone can see this repository. You choose who can commit.

 **Private**
You choose who can see and commit to this repository.

Initialize this repository with a README
This will let you immediately clone the repository to your computer. Skip this step if you're importing an existing repository.

Add .gitignore: **None** ▾

Add a license: **None** ▾



Create repository



NO!

Publishing your R code on GitHub

1. `cd your-directory-with-R-code`
2. `git add .`
3. `git commit -m "First commit"`
4. `git remote add origin`
`https://github.com/rafaeldcsantos/IntroDataScience.git`
5. `git push -u origin master`
6. Check
`https://github.com/rafaeldcsantos/IntroDataScience`

Publishing your R code on GitHub

The screenshot shows the GitHub interface for a repository named 'IntroDataScience' by user 'rafaeldcsantos'. The repository is currently empty, with no description, website, or topics provided. It has 1 commit, 1 branch, 0 releases, and 1 contributor. The repository is on the 'master' branch. There are two files listed: '.gitignore' and 'R.Rproj', both committed for the first time 12 minutes ago. A green button 'Add a README' is visible at the bottom of the repository view.

rafaeldcsantos / IntroDataScience

Unwatch 1 Star 0 Fork 0

Code Issues 0 Pull requests 0 Projects 0 Wiki Insights Settings

No description, website, or topics provided. [Edit](#)

[Manage topics](#)

1 commit 1 branch 0 releases 1 contributor

Branch: master New pull request Create new file Upload files Find file Clone or download

rafaeldcsantos First Commit Latest commit b804d17 12 minutes ago

.gitignore	First Commit	12 minutes ago
R.Rproj	First Commit	12 minutes ago

Help people interested in this repository understand your project by adding a README. [Add a README](#)

(Re)Publishing your R code on GitHub

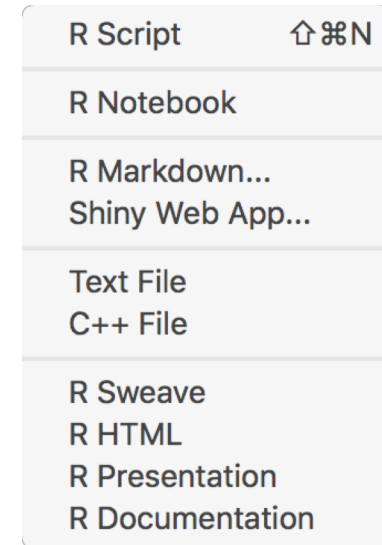
1. `git add .`
2. `git commit -m "Nth commit"`
3. `git push -u origin master`

4. Check

<https://github.com/rafaeldcsantos/IntroDataScience>

Playing nicely with GitHub

- R Scripts: source code only.
- R Notebook: creates a .Rmd and a .html file.
 - ▣ .Rmd: shown as is (no plots)
 - ▣ .html file: shown as source code.
- R HTML: formatted HTML code.
- Solution: create a R Markdown with a GitHub Document template.



Playing nicely with GitHub

- R Markdown with a GitHub Document template.

The screenshot shows the RStudio interface with the 'New R Markdown' dialog box open. The dialog box has a left sidebar with options: Document, Presentation, Shiny, and From Template. The 'From Template' option is selected, and a list of templates is shown on the right. The 'GitHub Document (Markdown)' template is highlighted. The background shows the R console with the R version 3.5.1 (2018-07-02) and the file explorer showing the current project directory.

~ /SYNC/OneDrive/INPE/CAP-394 Data Science/R - RStudio

Console Terminal x

~/SYNC/OneDrive/INPE/CAP-394 Data Science/R/

R version 3.5.1 (2018-07-02) --
Copyright (C) 2018 The R Founda
Platform: x86_64-apple-darwin18

R is free software and comes wi
You are welcome to redistribute
Type 'license()' or 'licence()'

Natural language support but

R is a collaborative project wi
Type 'contributors()' for more
'citation()' on how to cite R o

Type 'demo()' for some demos, '
'help.start()' for an HTML brow
Type 'q()' to quit R.

> |

New R Markdown

Document
Presentation
Shiny
From Template

Template: ? Using R Markdown Templates

Package Vignette (HTML) {rmarkdown}
GitHub Document (Markdown) {rmarkdown}

OK Cancel

Environment History Connections Git

Import Dataset List

Environment is empty

packages Help Viewer

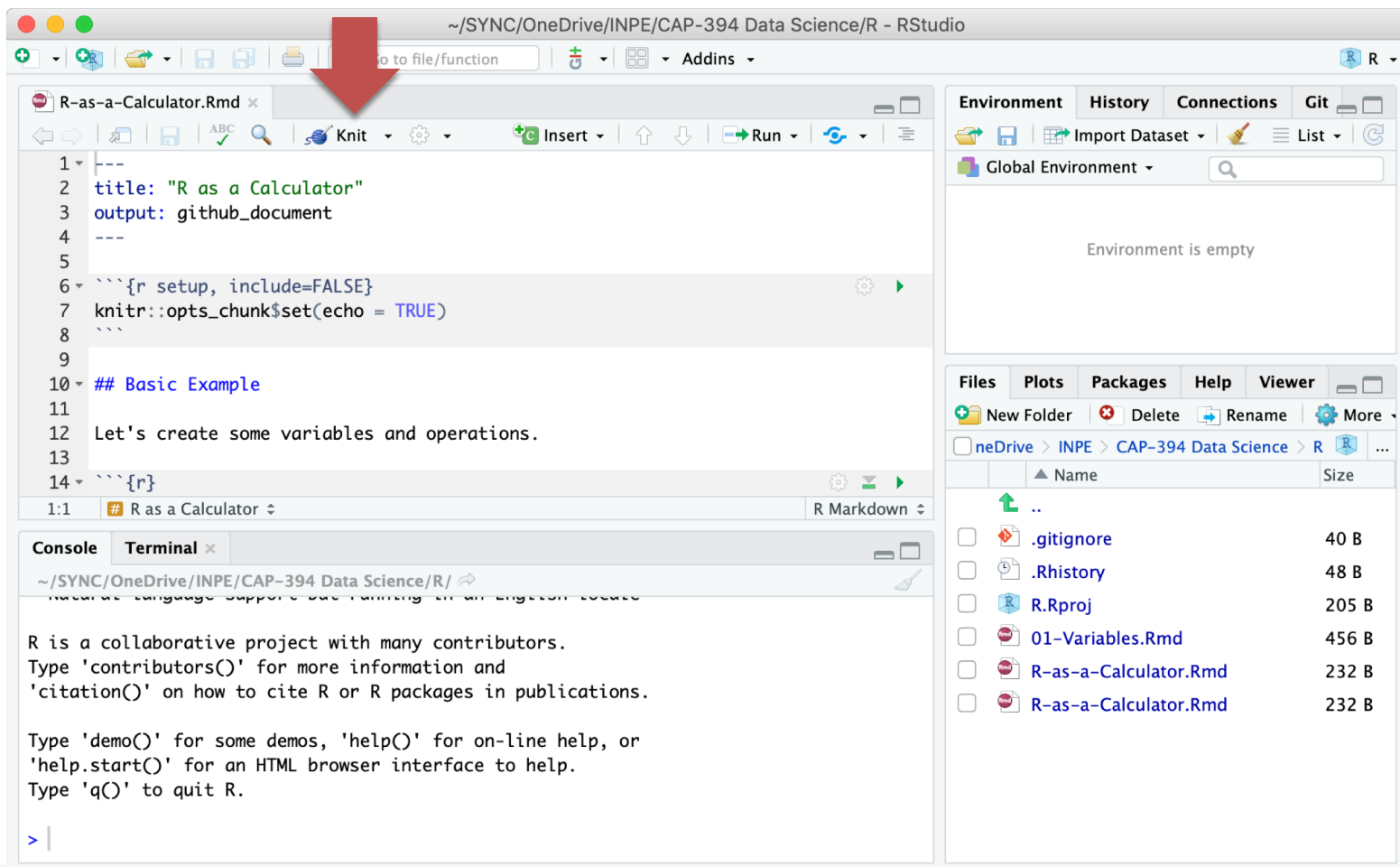
Delete Rename More

CAP-394 Data Science > R ...

	Size
	40 B
	48 B
	205 B
les.Rmd	456 B

Playing nicely with GitHub

- R Markdown with a GitHub Document template.
- Knit it!



The screenshot shows the RStudio interface with the following components:

- Editor:** Displays an R Markdown document titled "R-as-a-Calculator.Rmd". The content includes a title, output format, chunk options, and a code chunk for a basic example.
- Knit Button:** A red arrow points to the "Knit" button in the toolbar.
- Environment:** Shows "Global Environment" with the message "Environment is empty".
- Files:** A file browser showing the project structure, including files like ".gitignore", ".Rhistory", "R.Rproj", "01-Variables.Rmd", "R-as-a-Calculator.Rmd", and "R-as-a-Calculator.Rmd".
- Console:** Shows the R prompt and introductory text about R, including instructions on how to use help and quit.

```
1 ---
2 title: "R as a Calculator"
3 output: github_document
4 ---
5
6 ```{r setup, include=FALSE}
7 knitr::opts_chunk$set(echo = TRUE)
8 ```
9
10 ## Basic Example
11
12 Let's create some variables and operations.
13
14 ```{r}
```

1:1 # R as a Calculator

~ /SYNC/OneDrive/INPE/CAP-394 Data Science/R/

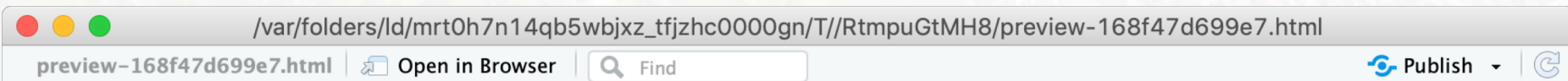
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |

Name	Size
..	
.gitignore	40 B
.Rhistory	48 B
R.Rproj	205 B
01-Variables.Rmd	456 B
R-as-a-Calculator.Rmd	232 B
R-as-a-Calculator.Rmd	232 B

Playing nicely with GitHub



R as a Calculator

Basic Example

Let's create some variables and operations.

```
a <- 12  
b <- 24  
a+b
```

```
## [1] 36
```

Awesome!

Playing nicely with GitHub

1. `git add .`
2. `git commit -m "Nth commit"`
3. `git push -u origin master`

Playing nicely with GitHub

rafaeldcsantos / IntroDataScience

Unwatch 1 Star 0 Fork 0

Code Issues 0 Pull requests 0 Projects 0 Wiki Insights Settings

No description, website, or topics provided. [Edit](#)

[Manage topics](#)

2 commits 1 branch 0 releases 1 contributor

Branch: master New pull request Create new file Upload files Find file Clone or download


rafaeldcsantos Added Examples Latest commit 0644881 a minute ago

.gitignore	First Commit	43 minutes ago
01-Variables.Rmd	Added Examples	a minute ago
R-as-a-Calculator.Rmd	Added Examples	a minute ago
R-as-a-Calculator.md	Added Examples	a minute ago
R.Rproj	First Commit	43 minutes ago




Help people interested in this repository understand your project by adding a README. [Add a README](#)

Playing nicely with GitHub

Branch: master ▾ [IntroDataScience](#) / R-as-a-Calculator.md Find file Copy path

 **rafaeldcsantos** Added Examples 0644881 2 minutes ago

[1 contributor](#)

18 lines (12 sloc) | 165 Bytes Raw Blame History   

R as a Calculator

Basic Example

Let's create some variables and operations.

```
a <- 12
b <- 24
a+b
```

```
## [1] 36
```

Awesome!