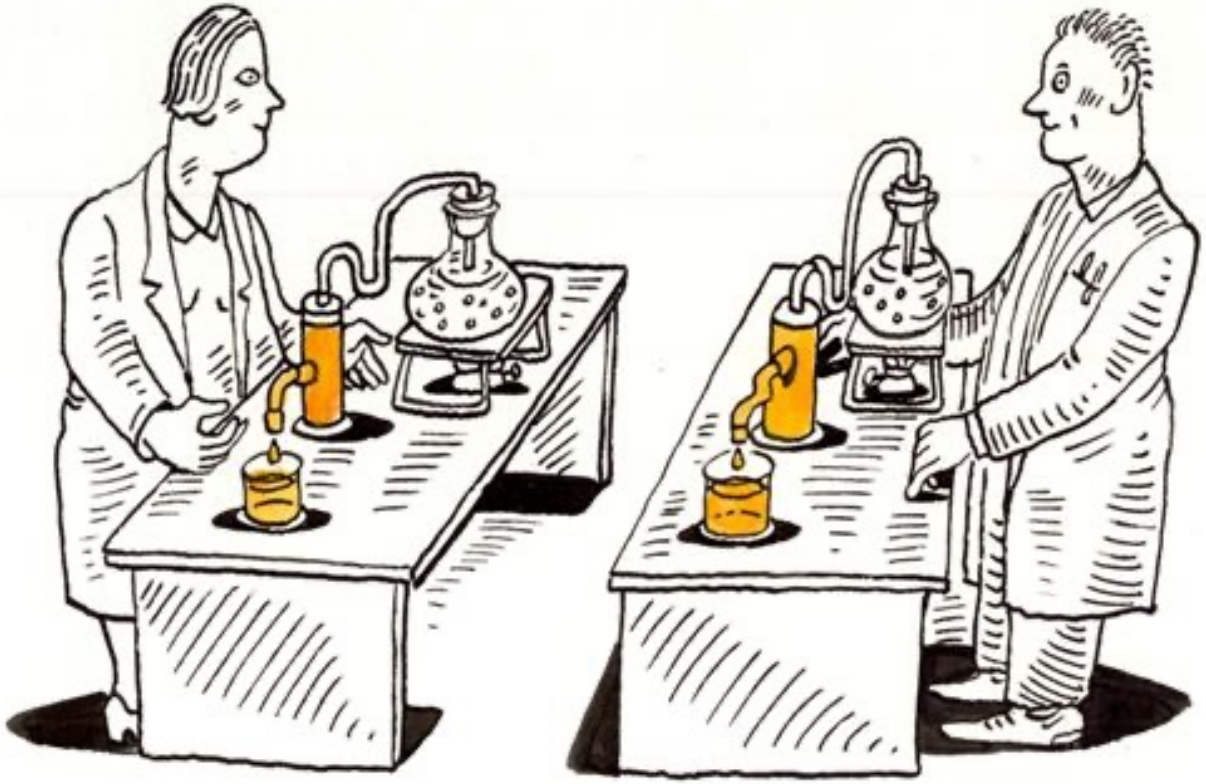
Introduction to Reproducible Science: *Day 1*

Nelson Roque, PhD

Scan for Slides & Code available on GitHub



Background

- Over the last decade I've engaged in professional development and research opportunities to become proficient across topics including data wrangling and modeling of text, image, video, and eye-tracking data, as well as more recently, sensor data.

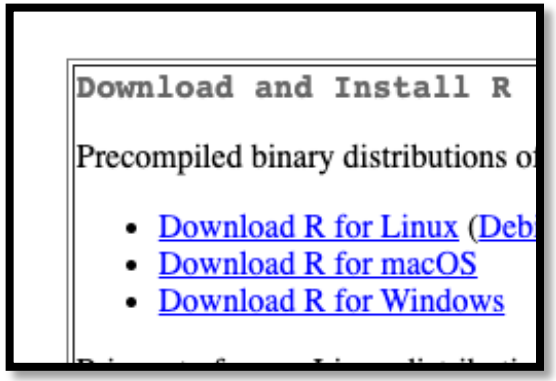
- I look forward to training the next generation of scientists on code-based methods to enhance reproducibility and usability of their research results.

About this Workshop

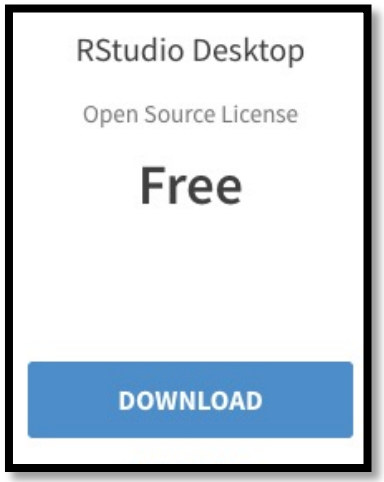
- 3 days: July 6, 8, 11, 11am to 2pm
- In-person, room 301Q, but also recorded
- All slides, code, and recordings will be made available at the website listed below, as well as Github
  - Code & Slides
    - https://github.com/nelsonroque/contextlab_reproduciblescience_workshops
  - Textbook
    - https://nelsonroque.github.io/contextlab_introdatascience_webcourse/

Software Required for Workshop

- R - https://cran.r-project.org/



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



Workshop Learning Objectives

1. Describe various principles, tools and techniques supportive of open and reproducible science.
    1. List and describe the FAIR Principles (https://www.go-fair.org/fair-principles)
2. Develop a code-only pipeline to allow reproducibility of data prep and analyses.
    1. Learn about data wrangling principles (e.g., tidy data)
    2. Import, pre-process & visualize data using R
3. Develop a long-term learning plan for practicing reproducible science tools and techniques.
    1. Share recommended readings, activities for continued learning

Agenda: Day 1

- What is Reproducible Science?
- Reproducible & FAIR Data Workflows
- Tools Supporting Reproducible Science
  - Overview of available tools
  - **Skill 1: Using Endnote for Reference Management**
  - **Skill 2: Using Git (and Github) for code management and collaboration**
- Orientation to R, RStudio, RMarkdown
  - **Skill 3: R syntax primer**
- Data Science: Latest trends
- Long-term Learning Recommendations

Agenda: Day 2

- Data wrangling and visualization of Big Data
  - **Skill 1: Data wrangling the Google Mobility dataset**

- Reproducible survey research
  - Qualtrics survey design tips
  - **Skill 2: Data wrangling Qualtrics data**

- Working with JSON data
  - **Skill 3: cleaning and visualizing keystroke JSON data**

Agenda: Day 3

- Text mining
  - **Skill 1: word and bigram frequency analysis**
  - **Skill 2: generating wordclouds**
  - **Skill 3: sentiment analysis**

- Interacting with APIs and JSON data
  - **Skill 4: querying API for results and data aggregation**

- **Closing Discussion & Q/A**

What is Reproducible Science?

Icebreaker: Introductions

1. Name, lab affiliation
2. What does reproducible science mean to you?
3. Which tools have you previously used?
4. Which tools are you most excited to learn?

Background

- A reproducibility crisis (Ioannidis, 2005; Open Science Collaboration, 2015) has emerged as a threat to the scientific enterprise.
  - Ioannidis, John P A. 2005. “Why Most Published Research Findings Are False.” PLoS Medicine 2 (8): e124. doi:10.1371/journal.pmed.0020124.
  - Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Percie du Sert, N., ... & Ioannidis, J. (2017). A manifesto for reproducible science. *Nature human behaviour*, *1*(1), 1-9.
  - Open Science Collaboration. 2015. “Estimating the Reproducibility of Psychological Science.” Science 349 (6251): aac4716–aac4716. doi:10.1126/science.aac4716.

What is reproducible science?

- Making entire scientific process transparent (as allowable; often required by law, grants).
- Sharing experiment, raw data, and analysis code
- Detailed methods sections
- Sharing stimulus sets
- Being able to walk through a data analysis start (load raw data) to finish (manuscript analyses) in code
- Following FAIR Principles
- Pre-registering hypotheses and analysis plans (e.g., on OSF)

Benefits of reproducible research

- increased likelihood that the research will be correct
- reproducibility makes it easier to check the research
- it is easier to reproduce the research independently
- easier to extend the research
- reusable code and instruction resulting in increased efficiencies

https://www.displayr.com/what-is-reproducible-research

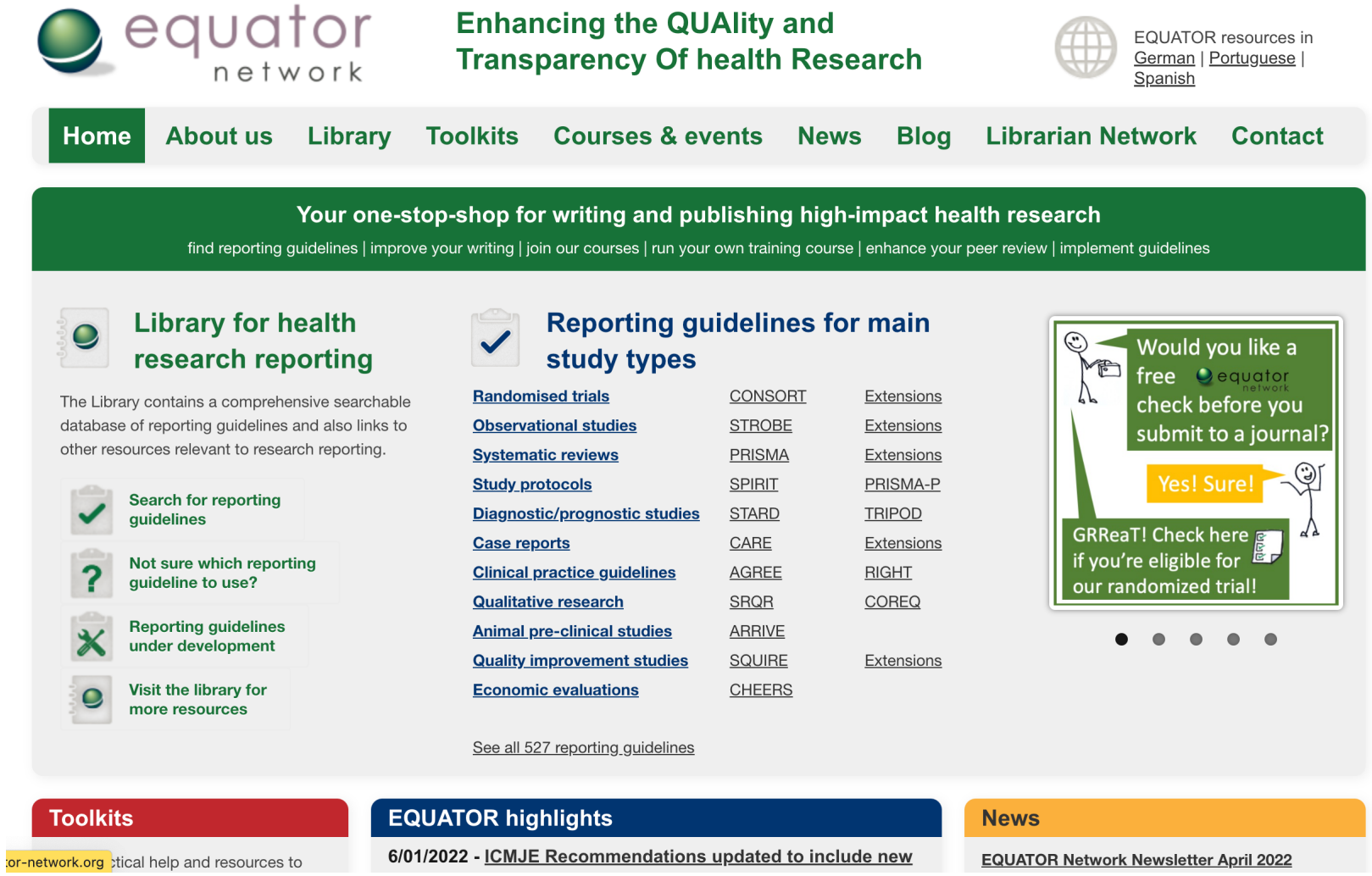
Reproducible versus replicable

- **Reproducible:** when the exact results can be reproduced if given access to the original data, software, or code

- **Replicable:** research results can be reproduced by independent researchers using different methods.

Improving Usability of Reproducible Research

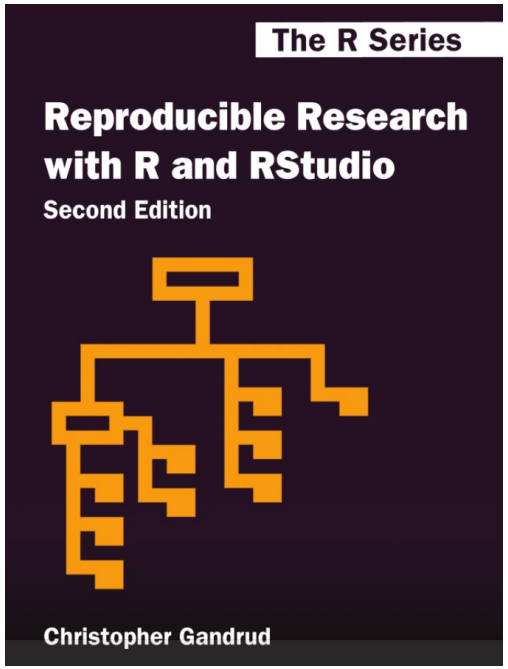
- Hundreds of reporting guidelines exist to enhance usability of research



Reproducible & FAIR Data Workflows

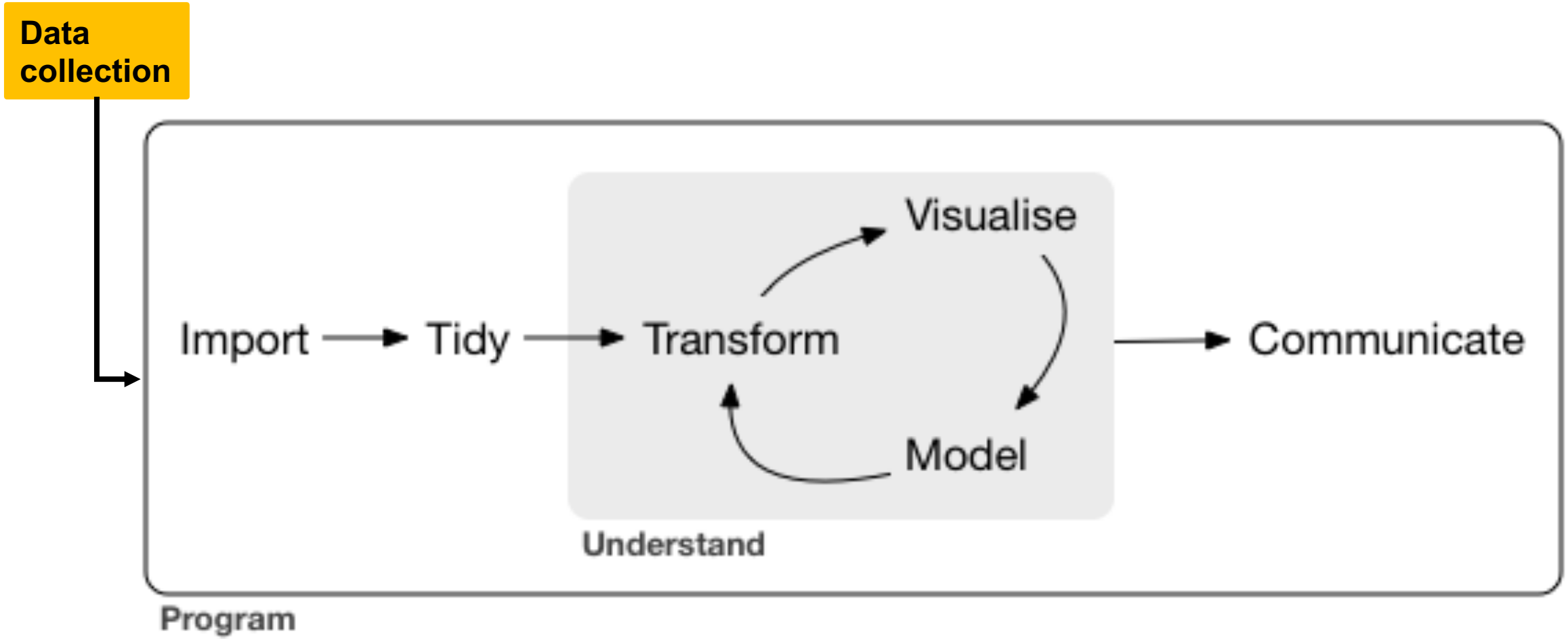
Reproducible Workflows

- Documented, code-based approach to pre-processing and analytics
  - R (RMarkdown), Python (Jupyter notebooks)
- File versioning with clear changelogs
  - Git and Github



https://monashdatafluency.github.io/r-rep-res/

Typical Data Science Workflow



Source: R for Data Science

What is tidy data?

1. Every column is a variable.
2. Every row is an observation.
3. Every cell is a single value.

participant	condition	avg_response_time	perc_accuracy
9991	control	506	90
9992	active	516	96
9993	control	526	99

Source: R for Data Science

Uncleaned Data Problems

- In a perfect data-generating system, data are already tidy

- In practice, we must first determine
  - data structure issues, including:
    - Column headers are values, not variable names.
    - Multiple variables are stored in one column.
    - Variables are stored in both rows and columns.
    - Multiple types of observational units are stored in the same table.
    - A single observational unit is stored in multiple tables.
  - Data quality issues
    - Values out of range
    - Improperly/inconsistently coded response options
    - Inconsistent records counts

FAIR Principles

F

Findable: "The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services”

A

Accessible: "Once the user finds the required data, she/he/they need to know how they can be accessed, possibly including authentication and authorisation.”

I

Interoperable: "The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.”

R

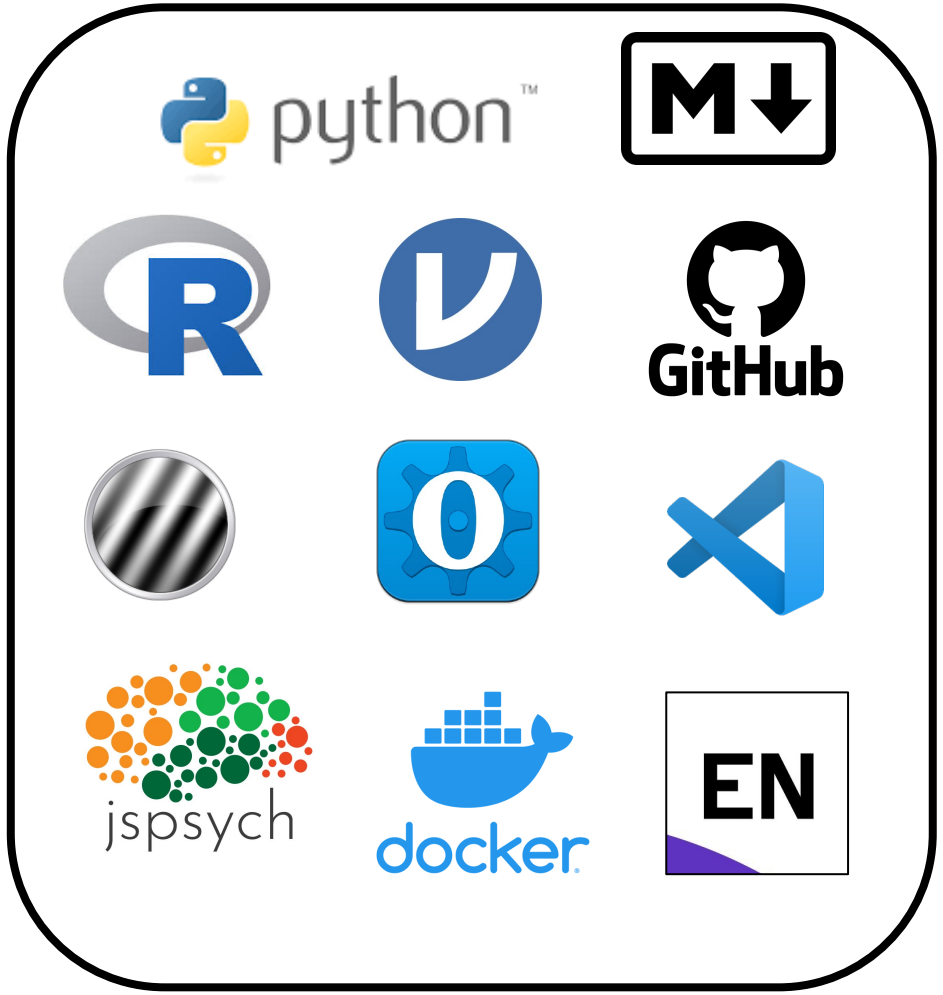
Reusable: "The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings."

https://www.go-fair.org/fair-principles

Tools Supporting Reproducible Science

Tools Supporting Reproducible Science

- R
- Latex, Markdown
- Python, Anaconda
- Github, Github Pages: https://pages.github.com
- Docker, VMs
- Infrastructure as Code (IaC; e.g., AWS CDK)
- IDEs (e.g., Visual Studio Code)
- Open Science Framework
- Documentation software: Docusaurus, https://docusaurus.io
- Code-based experiment creation software: Opensesame, Psychopy, jsPsych



Markdown

Lightweight markup used to add formatting elements to plaintext documents.

Created by John Gruber in 2004, Markdown is now one of the world's most popular markup languages.

Using Markdown is different than using a WYSIWYG editor.

- In Word, you click buttons in GUI to apply formatting
- In Markdown, you add syntax to text to apply formatting

https://www.markdownguide.org/getting-started/

Element	Markdown Syntax

Heading	`# H1`
	`## H2`
	`### H3`

Bold	`**bold text**`
Italic	`*italicized text*`
Blockquote	`> blockquote`

Ordered List	`1. First item`
	`2. Second item`
	`3. Third item`

Unordered List	`- First item`
	`- Second item`
	`- Third item`

Code	`` `code` ``
Horizontal Rule	`---`
Link	`[title](https://www.example.com)`
Image	`![alt text](image.jpg)`

What is Git?

- A version control system to manage code; edits, branching, and collaboration.

- Git providers
  - Open source Git: https://git-scm.com/
  - Github - https://desktop.github.com
    - Get Github Cheatsheets: https://training.github.com/downloads/github-git-cheat-sheet/
  - Gitlab

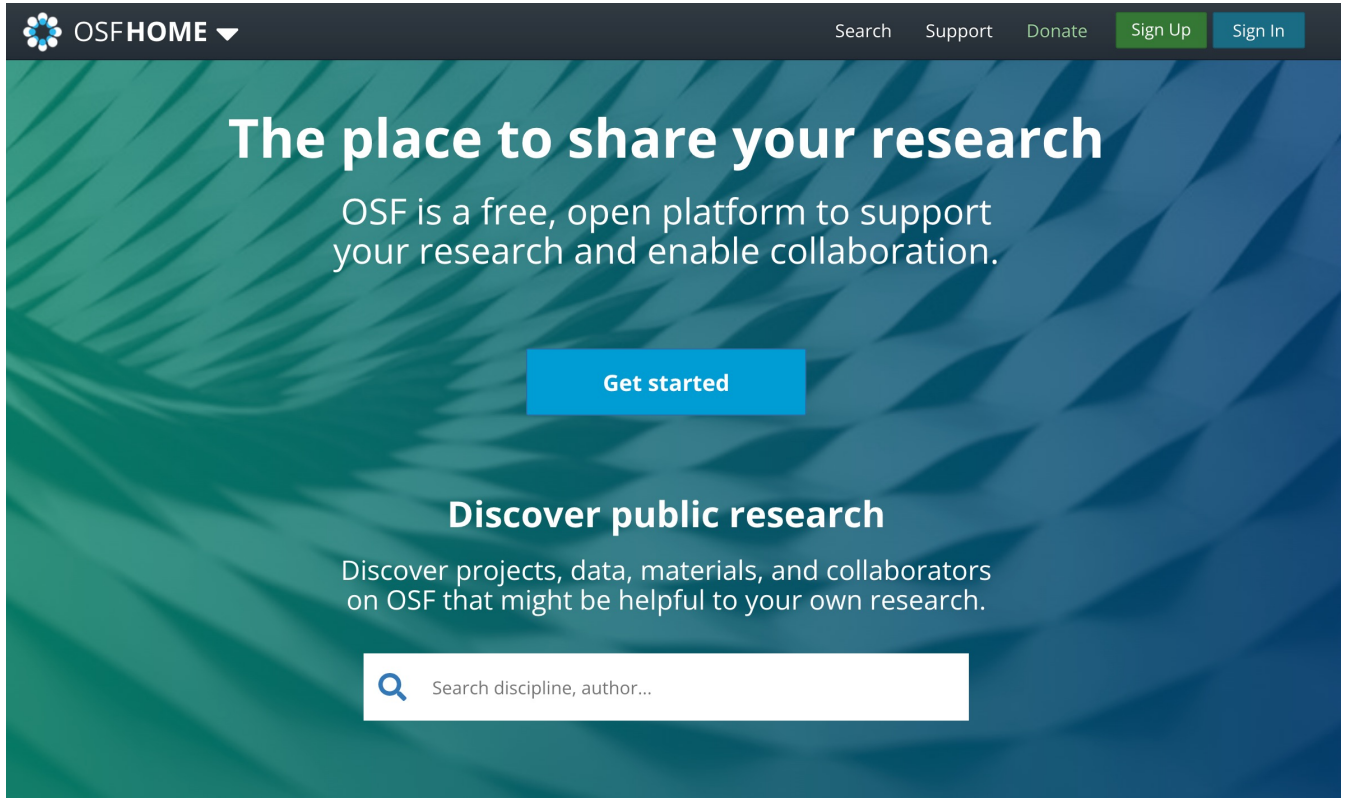
Git These Concepts

- **Repository**: a logical unit of storage for maintaining code/other assets. Can contain many folders, files.
- **.gitignore**: a file (with dot prefix) to specify what files should NOT be pushed to Github
- **.gitkeep**: a file (with dot prefix) to specify that a folder should be kept if blank as part of the project structure. This is helpful when creating templates.
- **commit**: create a record/snapshot of all files (or specific files) at a moment in time. You may tag collaborators, add a title and notes.
- **push**: 'storing' the result of prior commits
- **pull**: 'extracting' the result of prior commits
- **diff**: difference (line-specific) between two commits. Helpful when trying to remember when and where code edits where made.

OSF

- Preregister study design, hypotheses, analysis plans and scripts

- Post-hoc curate documentation for your study across various components (IRB, surveys, stimuli, datasets, codebooks)



PsyArXiv

- Publish pre-prints of your paper
- Supports immediate sharing of research
- https://psyarxiv.com

Let’s jump into Github

Download here:

https://desktop.github.com

Endnote

- **What is Endnote?**
  - a citation manager (like Zotero, Mendeley)
  - Guide from the UCF library with installation links: https://guides.ucf.edu/citations-endnote
- **Why Endnote?**
  - no need for manual tracking (deleting, updating) of references
  - change citation format at click of a button
  - capture PDFs for any citations at click of a button
- **Downloads:**
  - [Mac](http://ezproxy.library.ucf.edu/loggedin/EndNote20SiteInstaller.zip)
  - [Windows](http://ezproxy.library.ucf.edu/loggedin/EndNote20.zip)

Let’s jump into Endnote

Download here:

https://guides.ucf.edu/citations-endnote

Orientation to R, RStudio, RMarkdown

What is R? RStudio?

- **What is R?**
  - Statistical programming language (released ~1993). Similar to `S` (released ~1976)
  - Download R: https://cran.r-project.org

- **What is RStudio?**
  - Integrated development environment (IDE) for R; view data objects, plots, and file directory all in one place.
  - Download Rstudio: https://www.rstudio.com/products/rstudio/download

What are R packages?

- **Libraries of code to accomplish specific functions**
  - e.g., data viz, machine learning, data wrangling, compute effect sizes

- **Think of CRAN as an equivalent to iOS App Store, or Google Play Store**
  - CRAN is where you download published R packages (directly from R, RStudio)

- **A developer can also make packages available directly in source code, or via tools like Github, for example:**

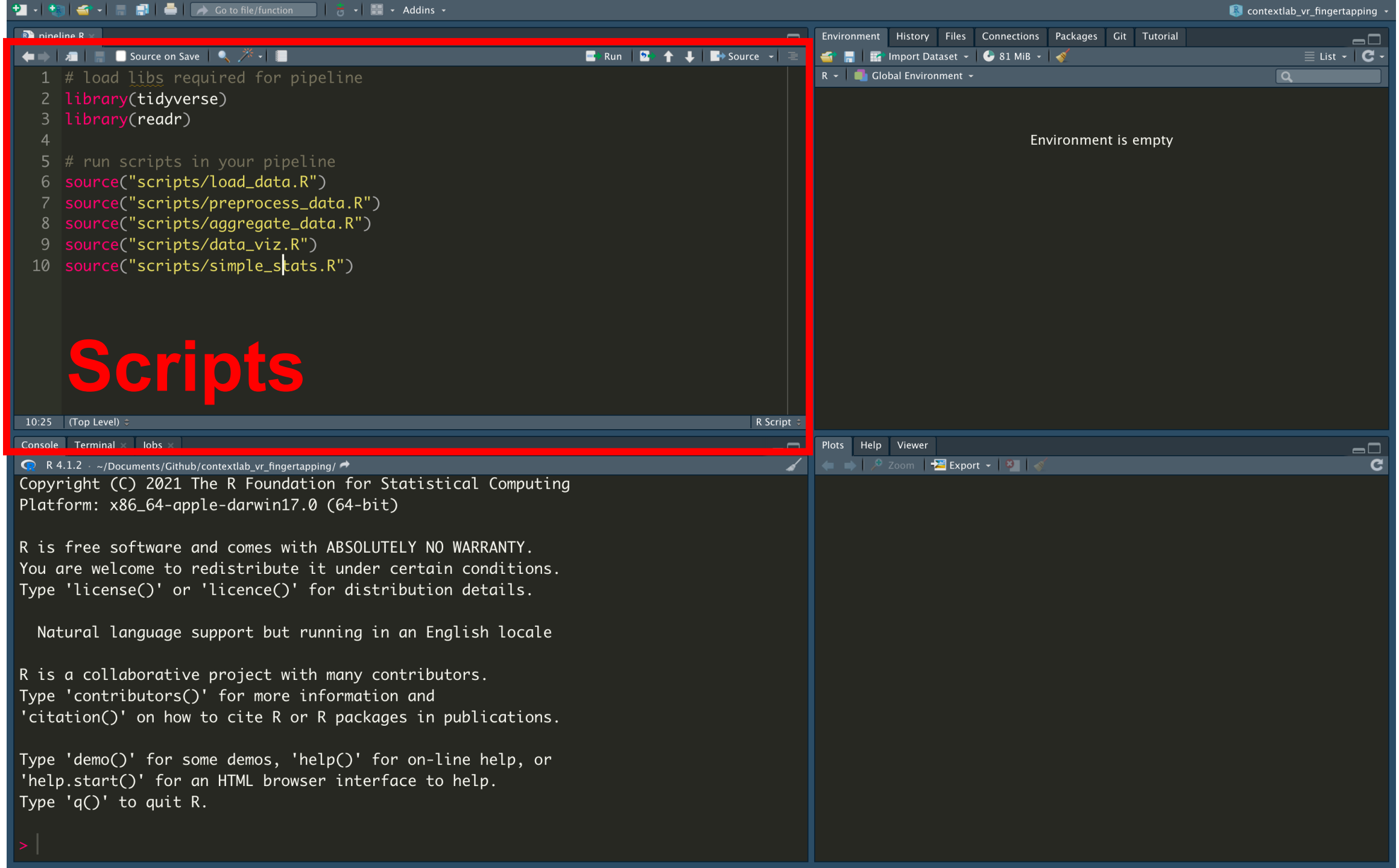
```
install.packages("devtools")
devtools::install_github("nelsonroque/ruf")
```

What is RMarkdown?

- Reproducible notebook (akin to Jupyter Notebooks in Python) where you can write a narrative using Markdown syntax and embed code/plots throughout.
  - Based on Markdown syntax: https://www.markdownguide.org/basic-syntax/

- With R Markdown, you can generate HTML, PDF, MS Word documents, PowerPoint presentations, and books!

- Learn More: http://rmarkdown.rstudio.com


Go to file/function
Addins
contextlab_vr_fingertapping
pipeline.R
Source on Save
Run
Source
1 # load libs required for pipeline
2 library(tidyverse)
3 library(readr)
4
5 # run scripts in your pipeline
6 source("scripts/load_data.R")
7 source("scripts/preprocess_data.R")
8 source("scripts/aggregate_data.R")
9 source("scripts/data_viz.R")
10 source("scripts/simple_stats.R")
Scripts
10:25
(Top Level)
R Script
Environment
History
Files
Connections
Packages
Git
Tutorial
Import Dataset
81 MiB
List
R
Global Environment
Environment is empty
Console
Terminal
Jobs
R 4.1.2 · ~/Documents/Github/contextlab_vr_fingertapping/
Copyright (C) 2021 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin17.0 (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.
>
Plots
Help
Viewer
Zoom
Export

pipeline.R

```r
# load libs required for pipeline
library(tidyverse)
library(readr)

# run scripts in your pipeline
source("scripts/load_data.R")
source("scripts/preprocess_data.R")
source("scripts/aggregate_data.R")
source("scripts/data_viz.R")
source("scripts/simple_stats.R")
```

Environment | History | Files | Connections | Packages | Git | Tutorial

New Folder | New Blank File | Delete | Rename | More

Home > Documents > Github > contextlab_vr_fingertapping > scripts

Name	Size	Modified
..		
aggregate_data.R	1.7 KB	Jun 1, 2022, 7:23 PM
data_viz.R	562 B	Jun 1, 2022, 7:29 PM
load_data.R	1.2 KB	Jun 1, 2022, 7:09 PM
preprocess_data.R	696 B	Jun 1, 2022, 7:34 PM
simple_stats.R	124 B	Jun 1, 2022, 7:18 PM

Files

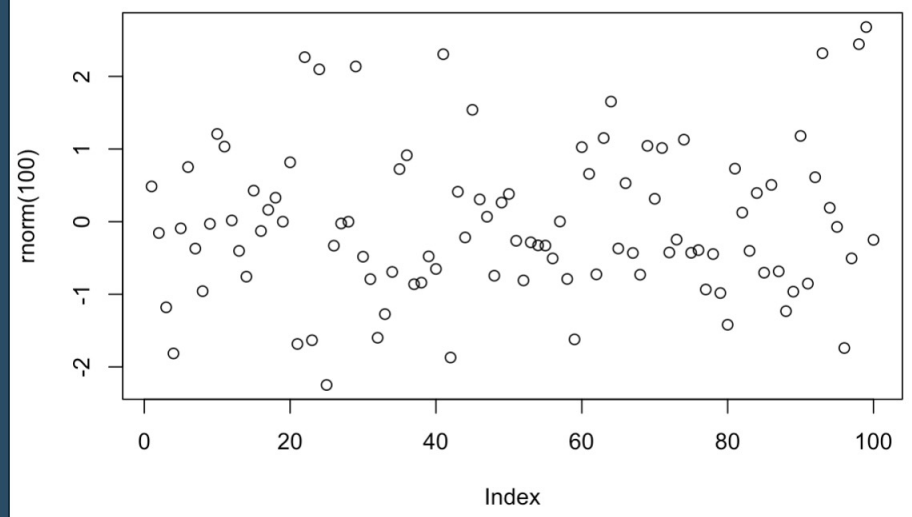
Console | Terminal | Jobs

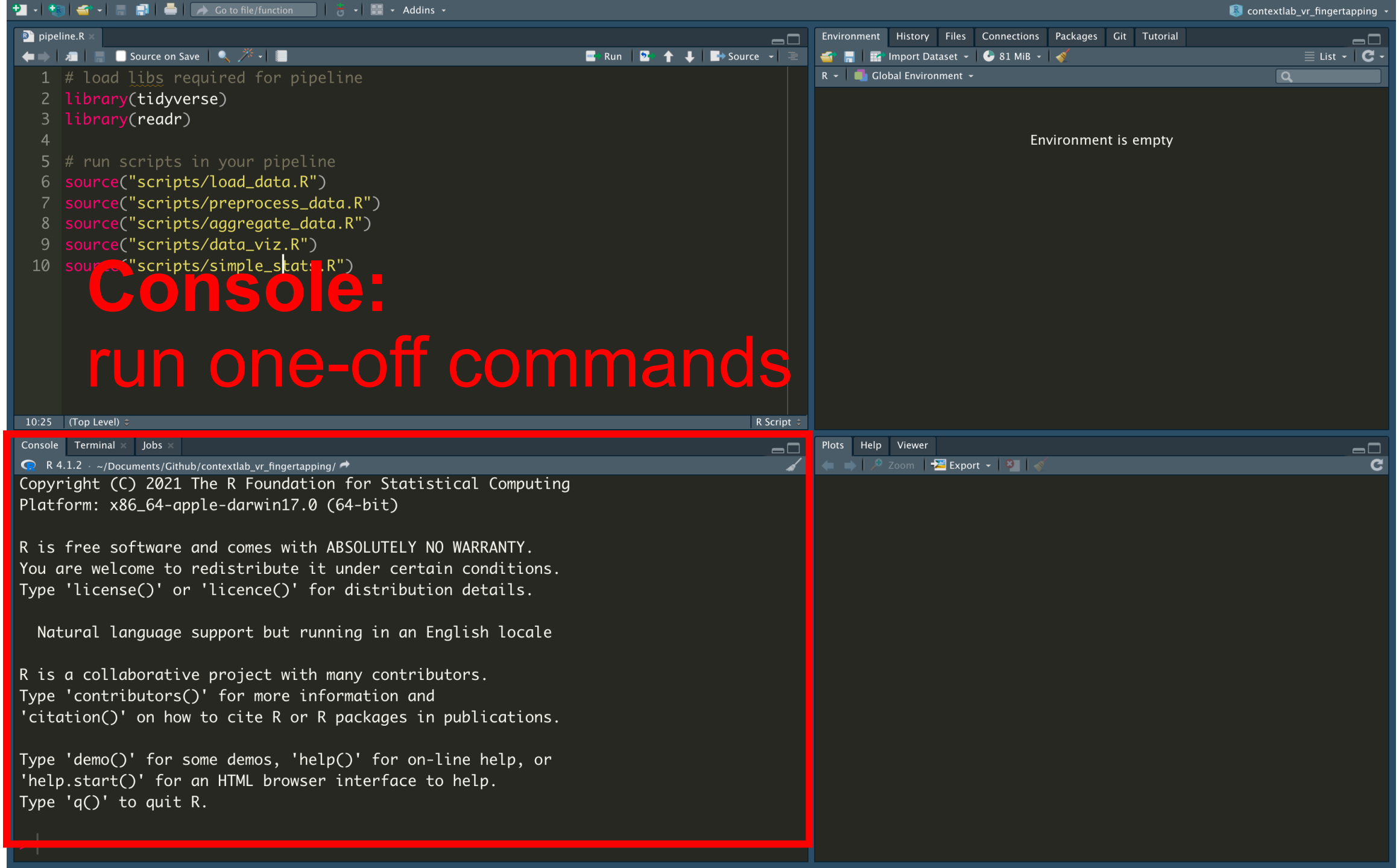
R 4.1.2 · ~/Documents/Github/contextlab_vr_fingertapping/

```
Platform: x86_64-apple-darwin17.0 (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.

> plot(x=rnorm(100))
```

Plots | Help | Viewer

Zoom | Export | Publish




pipeline.R
Source on Save
Run
Source
1 # load libs required for pipeline
2 library(tidyverse)
3 library(readr)
4
5 # run scripts in your pipeline
6 source("scripts/load_data.R")
7 source("scripts/preprocess_data.R")
8 source("scripts/aggregate_data.R")
9 source("scripts/data_viz.R")
10 source("scripts/simple_stats.R")
Console:
run one-off commands
10:25
(Top Level)
R Script
Environment
History
Files
Connections
Packages
Git
Tutorial
Import Dataset
81 MiB
List
R
Global Environment
Environment is empty
Console
Terminal
Jobs
R 4.1.2 · ~/Documents/Github/contextlab_vr_fingertapping/
Copyright (C) 2021 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin17.0 (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.
Plots
Help
Viewer
Zoom
Export

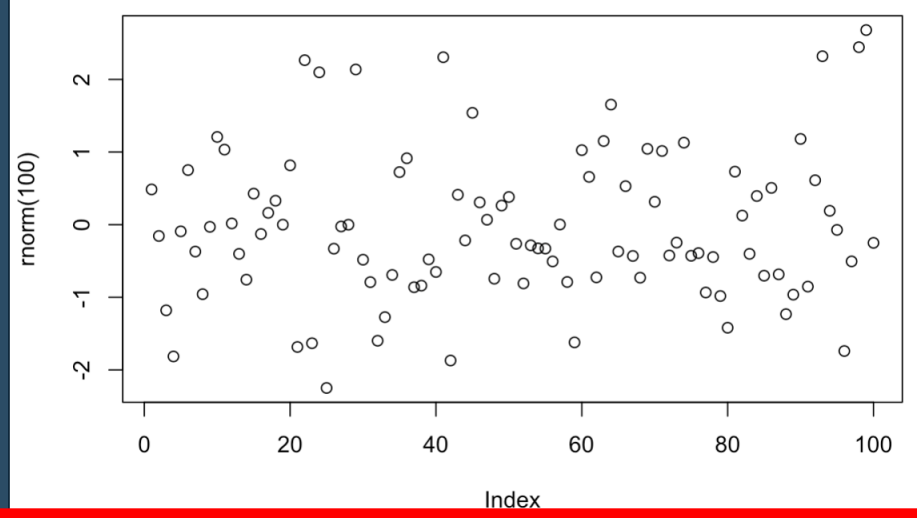
pipeline.R

```r
# load libs required for pipeline
library(tidyverse)
library(readr)

# run scripts in your pipeline
source("scripts/load_data.R")
source("scripts/preprocess_data.R")
source("scripts/aggregate_data.R")
source("scripts/data_viz.R")
source("scripts/simple_stats.R")
```

Environment is empty

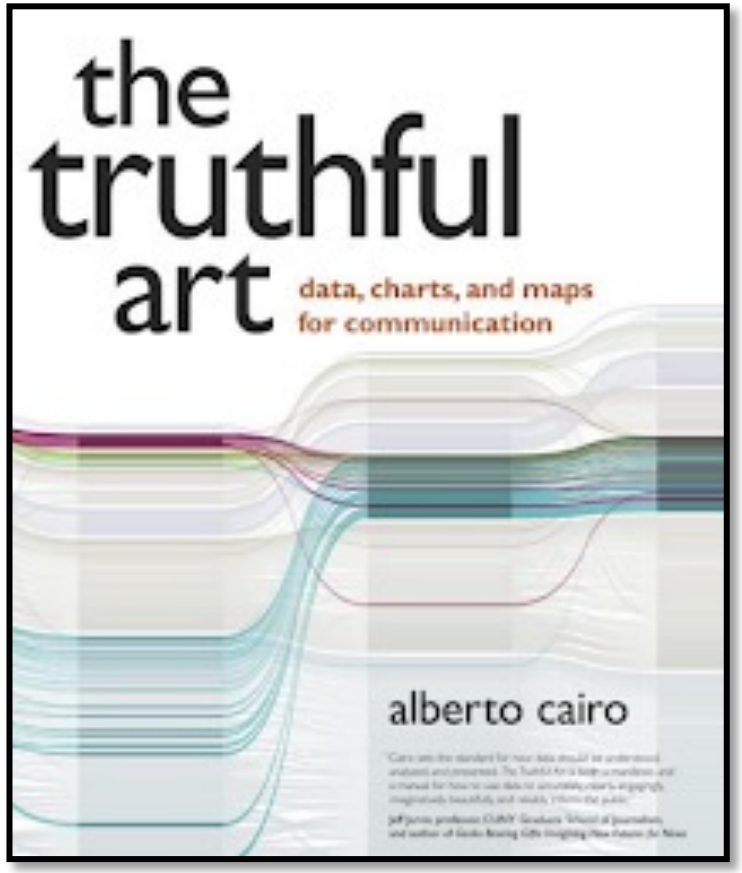
Plots

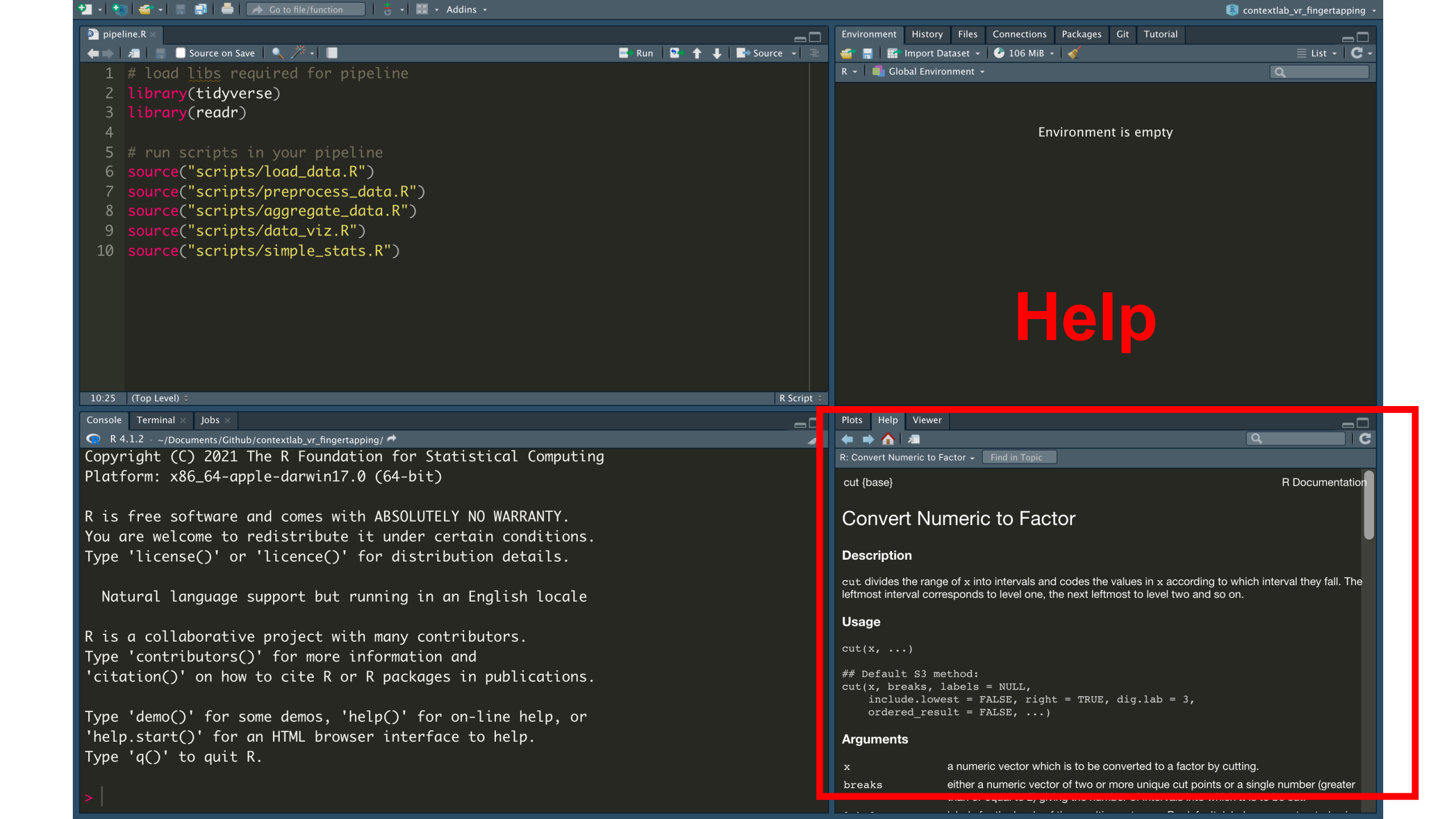
```
Platform: x86_64-apple-darwin17.0 (64-bit)

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.

> plot(x=rnorm(100))
> ``
```



PSA: Five Qualities of Great Visualizations

1. **Truthful:** transparency in decision making and reporting
2. **Functional:** understandable, uncluttered
3. **Beautiful:** plots can be art too
    1. https://playingthearchive.com
4. **Insightful:** reveal trends or relationships
5. **Enlightening:** integration of 1-4 executed succesfully




Go to file/function
Addins
contextlab_vr_fingertapping
pipeline.R
Source on Save
Run
Source
1 # load libs required for pipeline
2 library(tidyverse)
3 library(readr)
4
5 # run scripts in your pipeline
6 source("scripts/load_data.R")
7 source("scripts/preprocess_data.R")
8 source("scripts/aggregate_data.R")
9 source("scripts/data_viz.R")
10 source("scripts/simple_stats.R")
10:25
(Top Level)
R Script
Environment
History
Files
Connections
Packages
Git
Tutorial
Import Dataset
106 MiB
List
R
Global Environment
Environment is empty
Help
Console
Terminal
Jobs
R 4.1.2 · ~/Documents/Github/contextlab_vr_fingertapping/
Copyright (C) 2021 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin17.0 (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.
>
Plots
Help
Viewer
R: Convert Numeric to Factor
Find in Topic
cut {base}
R Documentation
Convert Numeric to Factor
Description
cut divides the range of x into intervals and codes the values in x according to which interval they fall. The leftmost interval corresponds to level one, the next leftmost to level two and so on.
Usage
cut(x, ...)
Default S3 method:
cut(x, breaks, labels = NULL,
include.lowest = FALSE, right = TRUE, dig.lab = 3,
ordered_result = FALSE, ...)
Arguments
x
a numeric vector which is to be converted to a factor by cutting.
breaks
either a numeric vector of two or more unique cut points or a single number (greater

```
# load libs required for pipeline
library(tidyverse)
library(readr)

# run scripts in your pipeline
source("scripts/load_data.R")
source("scripts/preprocess_data.R")
source("scripts/aggregate_data.R")
source("scripts/data_viz.R")
source("scripts/simple_stats.R")

# df$group =  rep(1:ceiling(nrow(df)/6), each = 6)[1:nrow(df)]
```

Environment | History | Files | Connections | Packages | Git | Tutorial

Import Dataset | 337 MiB | List

R | Global Environment

Data

df	100 obs. of 3 variables
fit	Formal class lmerMod

Viewer

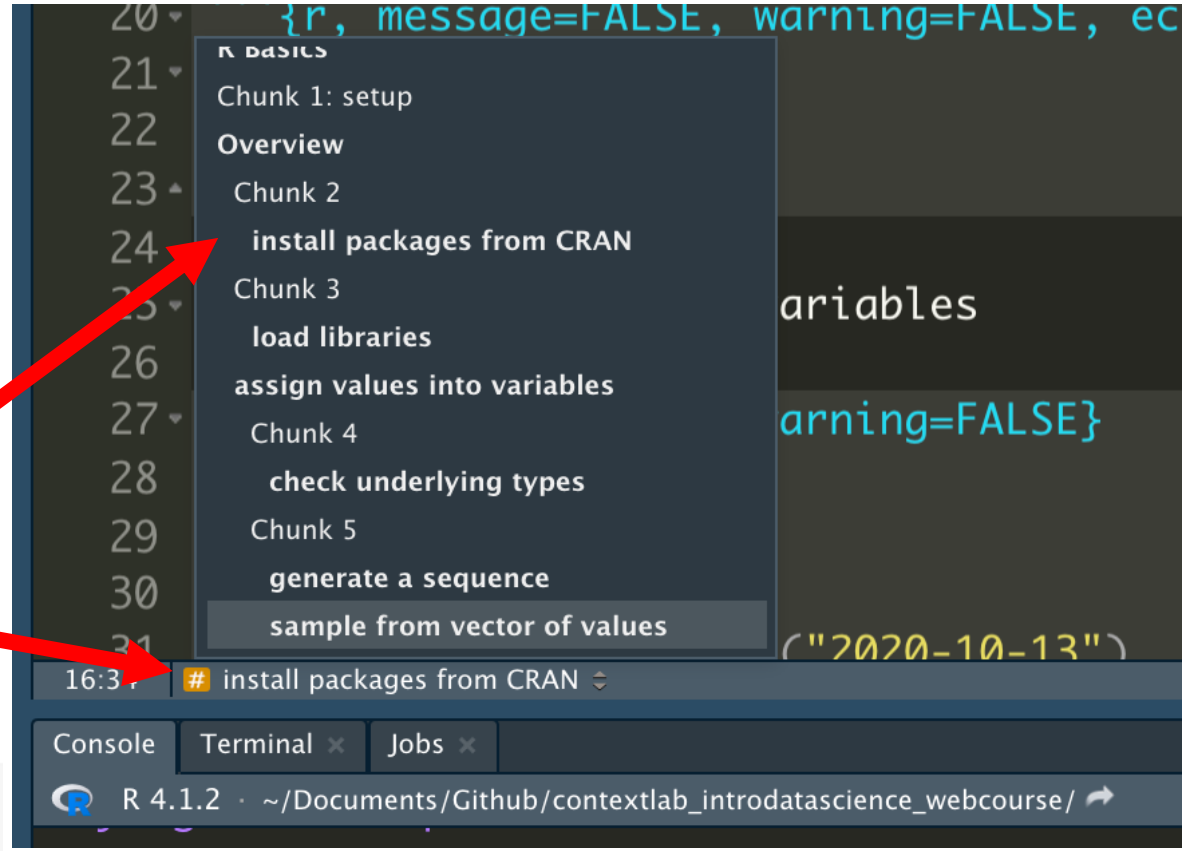
```
boundary (singular) fit: see help('isSingular')
Linear mixed model fit by REML ['lmerMod']
Formula: y ~ x + (1 | group)
   Data: df
REML criterion at convergence: 291.8917
Random effects:
 Groups   Name        Std.Dev.
 group    (Intercept) 0.000
 Residual             1.024
Number of obs: 100, groups:  group, 17
Fixed Effects:
(Intercept)            x
   -0.06577     -0.12750
optimizer (nloptwrap) convergence code: 0 (OK) ; 0 optimizer warnings; 1 lme4 warnings
> fit = lme4::lmer(y ~ x + (1|group), data = df)
boundary (singular) fit: see help('isSingular')
> sjPlot::tab_model(fit)
>
```

	y		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-0.07	-0.27 – 0.14	0.524
x	-0.13	-0.34 – 0.09	0.240
Random Effects			
σ^2	1.05		
$\tau_{00\ group}$	0.00		
N_{group}	17		
Observations	100		
Marginal R^2 / Conditional R^2	0.014 / NA		

Shortcuts

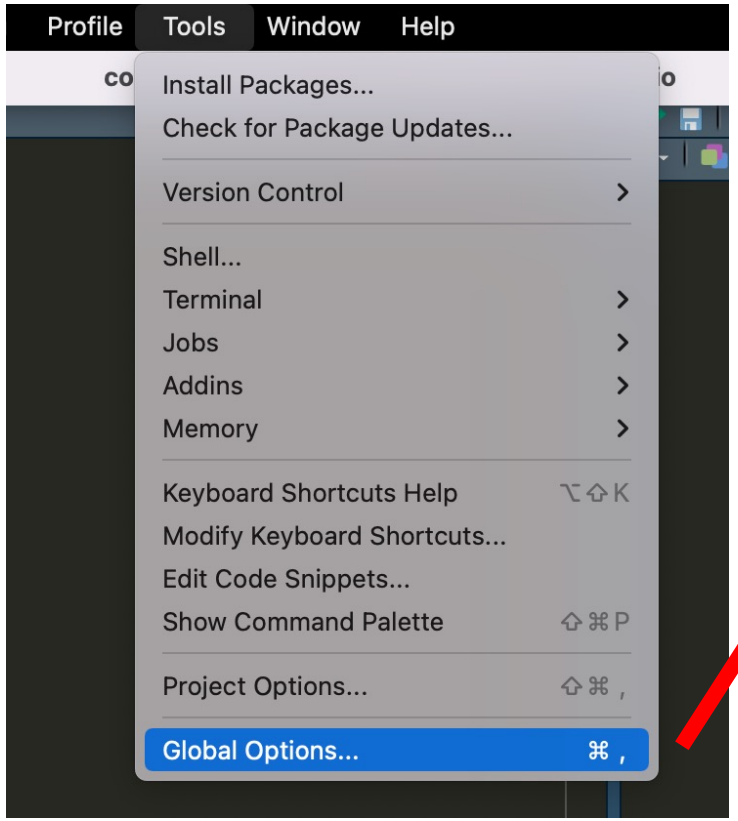
- Command+Enter (Mac) or Ctrl+Enter (Windows) to run the line where your mouse cursor is (or a highlighted selection of code)

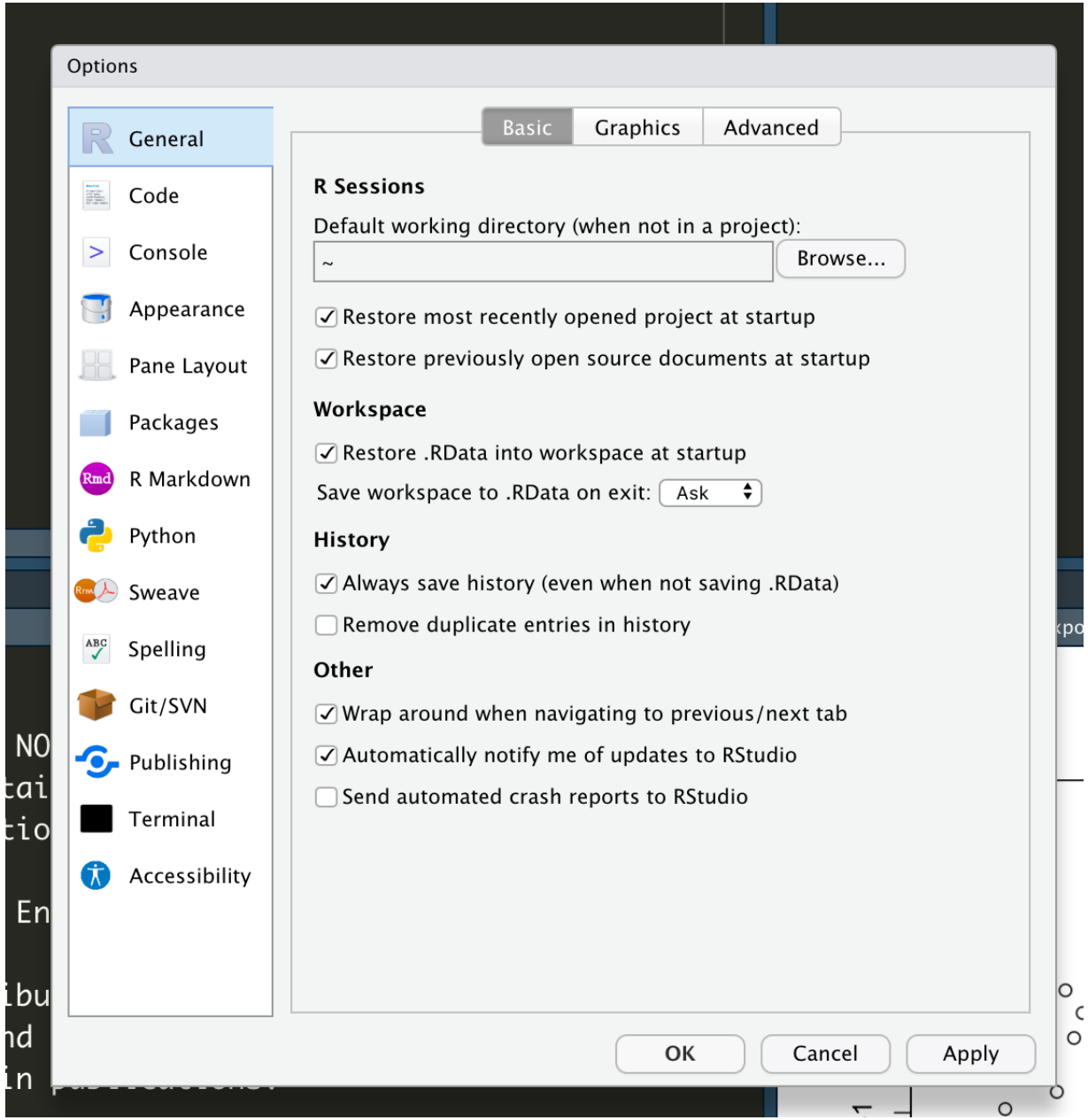
- Add "-----" inside of a comment to create a shortcut




```
15 ```{r}
16 # install packages from CRAN ----
17 install.packages(c("xlsx", "tidyverse", "devtools"))
18 ```
```

Customizing RStudio





Intro to R Syntax

Basic data types & structures

- Character: text data
- Numeric: all real numbers with or without decimal values
- Integer: real values without decimal points
- Logical: TRUE, FALSE
- Raw: raw bytes of data
- List: named or numbered array of entries
- Data frame and tibble: rectangular representation of data
- Vector
- Matrix

https://www.programiz.com/r/data-types

Pseudocode Exercise

You allowed participants on Qualtrics to enter state of residence as free text entry, resulting in e.g., Alabama, as: ‘AL’, ‘Al’, ‘Alabama’, ‘ALABAMA’.

Ideally you want them all to be two upper case letters.

Goal: in words only (no code), describe your approach to solving the data problem above.

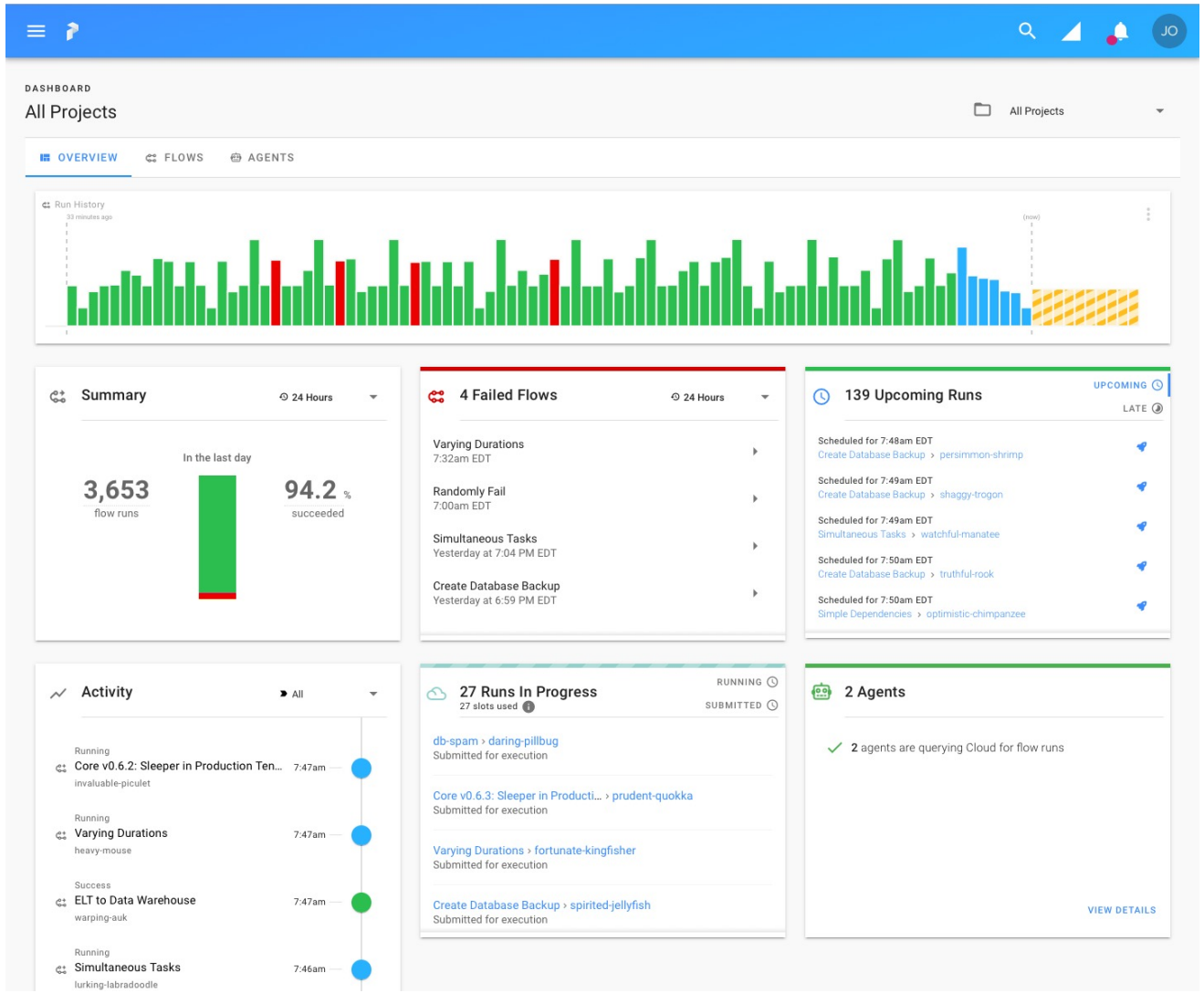
Let’s jump into R

Scan the QR code to view the code we will work with

Data Science: Latest Trends

ETL Pipelines

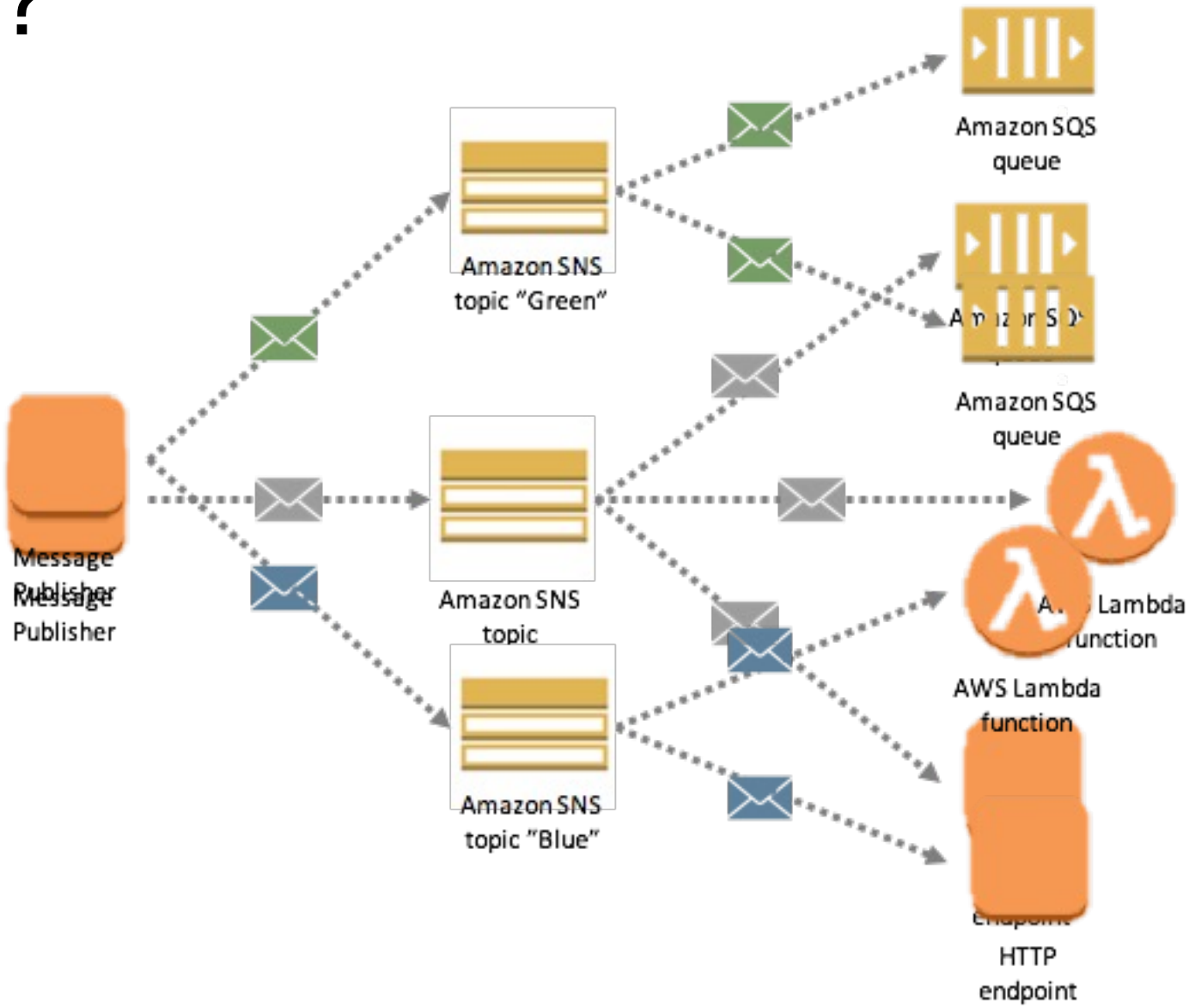
- Extract
  - Programmatically download data from origin (database, file server)
- Transform
  - Manipulate, score, reformat data as needed for computational needs
- Load
  - Re-save transform dataset(s) in new location on origin server (or other server)



What is PubSub?

- Messaging pattern for loosely coupled communication between highly cohesive components

- **Why PubSub?**
  - Scalable, fault-tolerant, and highly available

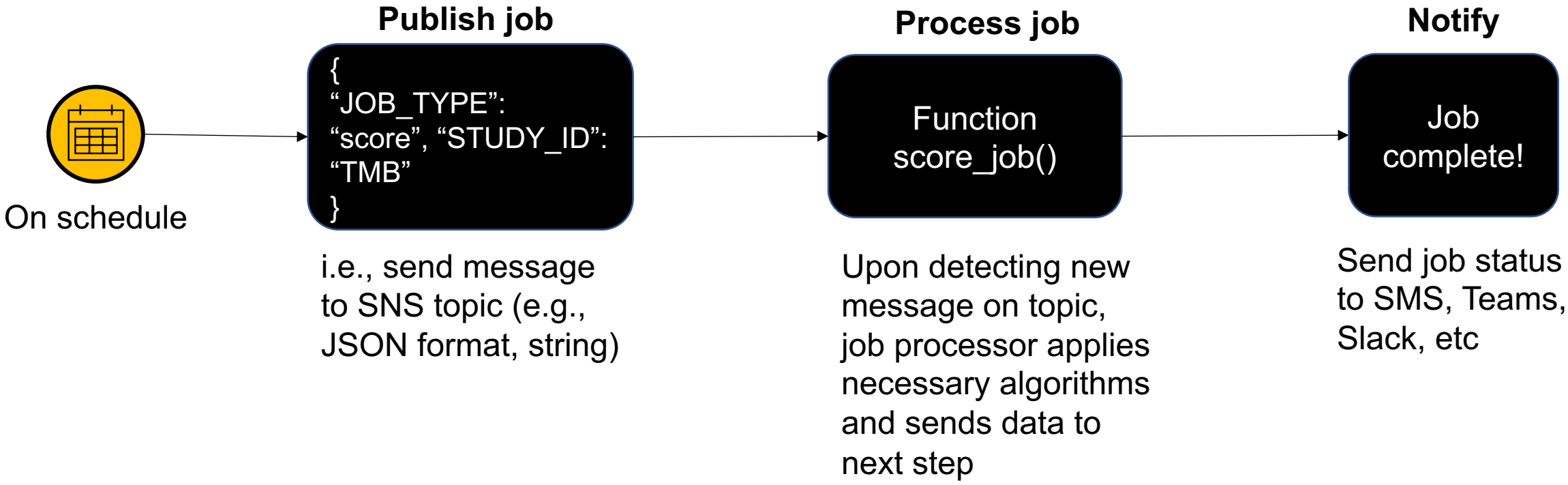


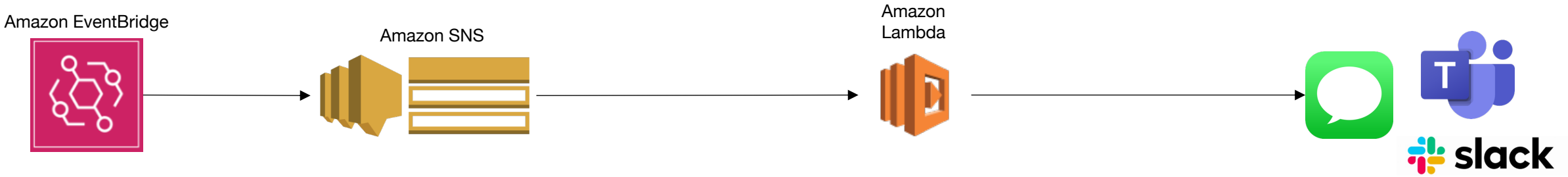
How I’ve Used PubSub Messaging

- Send a message to a topic for each device in a study, and query all data for last X days from device API
  - e.g., Fitbit, Air quality devices

- Send a message to a topic for each cognitive task, query respective data, score it, and publish to API/store in new database table

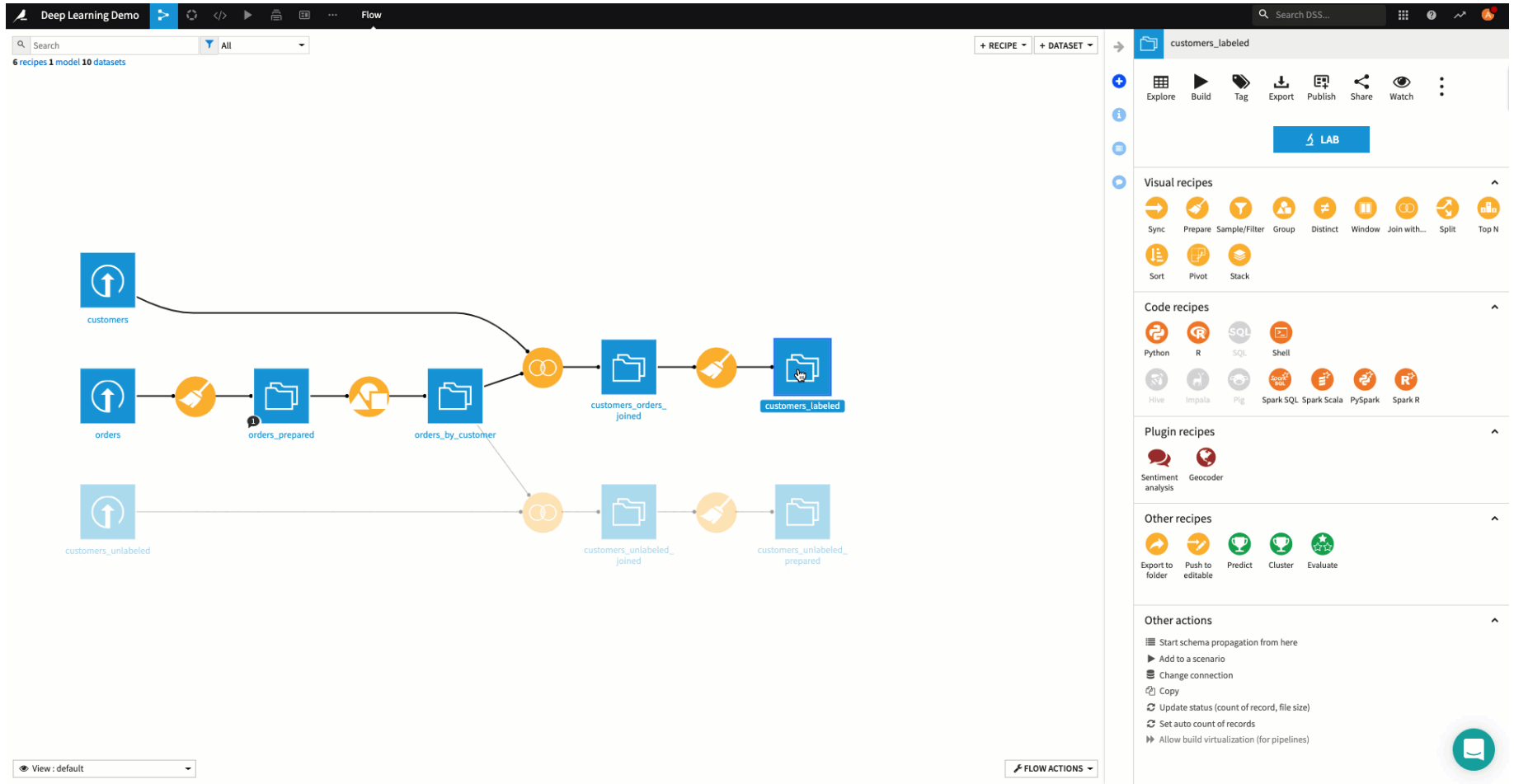
Automating data jobs with PubSub





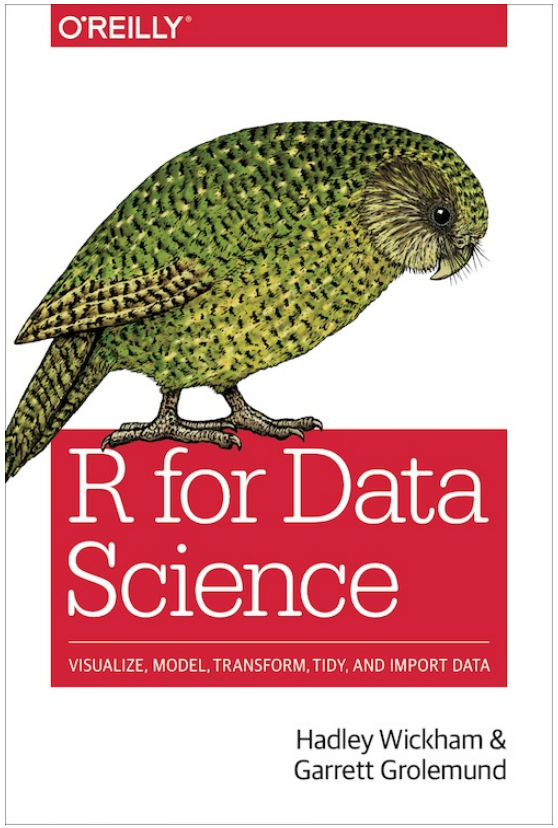
The Future of Data Workflows

Seamless integration platforms, for example:

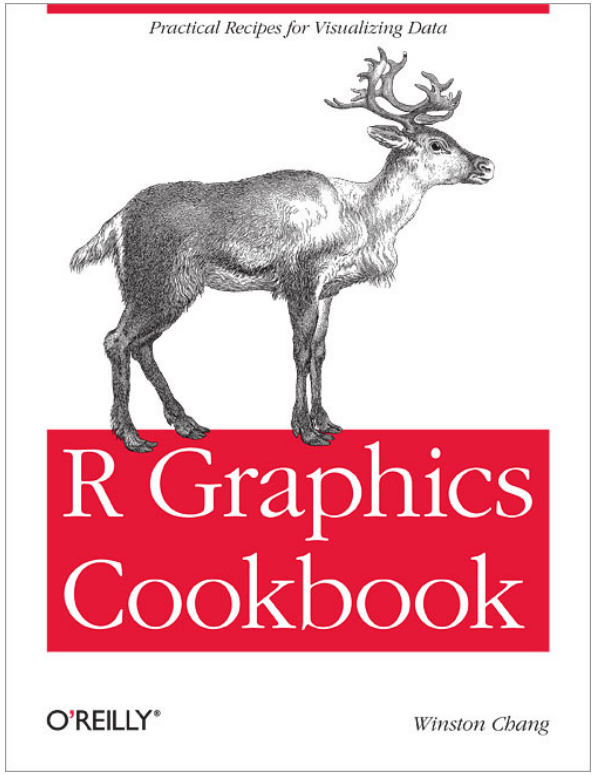


Long-term learning

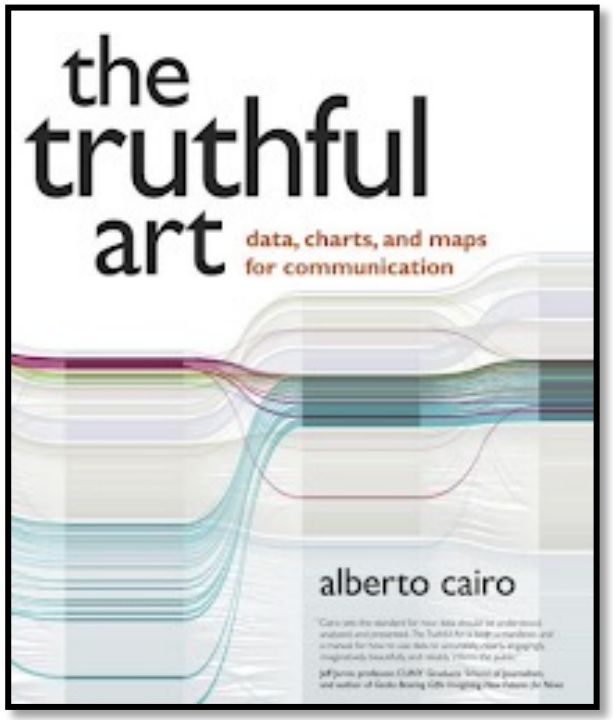
Reading Recommendations



https://r4ds.had.co.nz

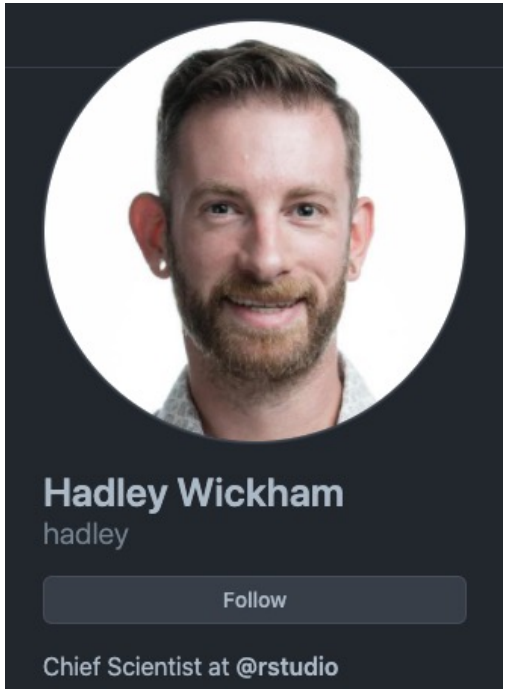


https://r-graphics.org



Tutorial Recommendations

- Anything by Hadley Wickham



- https://www.r-bloggers.com/
- https://www.r-statistics.com/
- https://blog.revolutionanalytics.com/
- https://r-charts.com/
- http://www.cookbook-r.com/Graphs/
- https://plotly.com/r/

Download Open Data

- Kaggle
  - https://www.kaggle.com/datasets
- Data.gov
  - https://www.data.gov/

Download Cheatsheets

- Data Import
  - https://github.com/rstudio/cheatsheets/raw/master/data-import.pdf)
- Data Wrangling Cheatsheet
  - https://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf)
- Data Transformation with dplyr
  - https://github.com/rstudio/cheatsheets/raw/master/data-visualization.pdf)
- String Manipulation
  - https://github.com/rstudio/cheatsheets/raw/master/strings.pdf)
- Work with dates/times
  - https://github.com/rstudio/cheatsheets/raw/master/lubridate.pdf)
- R Markdown
  - https://github.com/rstudio/cheatsheets/raw/master/rmarkdown.pdf)
- More cheatsheets
  - https://www.rstudio.com/resources/cheatsheets/)

thank you

nelson.roque@ucf.edu

nelsonroque.com

https://github.com/nelsonroque