

The background of the slide features a complex, abstract network diagram. It consists of numerous nodes of varying sizes and colors (dark blue, light blue, and grey) connected by thin, light grey lines. Some nodes are highlighted with larger, concentric circles. The overall aesthetic is modern and technological, suggesting themes of data, research, and interconnectedness.

# SETTING UP TOOLS AND WORKFLOW FOR TEACHING REPRODUCIBLE RESEARCH, BIG DATA AND DATA MINING IN NURSING AND PUBLIC HEALTH

---

Melinda Higgins, PhD; Emory University, Professor  
Vicki Hertzberg, PhD (co-instructor, Professor)

<https://melindahiggins2000.github.io/N741bigdata/>

COURSE NUMBER, TITLE:

COURSE DESCRIPTION

COURSE OBJECTIVES

TEACHING AND LEARNING

# N741 Big Data Analytics

## COURSE NUMBER, TITLE:

NRSG 741, Big Data Analytics for Healthcare

## COURSE DESCRIPTION

This course will describe the concepts underlying the field of study identified as big data analytics along with its application in healthcare. The theoretical underpinnings of these concepts will be presented along with applications in healthcare, including knowledge discovery, precision medicine/nursing, and the development of targeted interventions to improve health outcomes. Commonly used methods in big data analytics will be reviewed, and the challenges related to gathering, analyzing, visualizing, and interpreting big data will be discussed. Hands-on computer laboratory experience with these techniques relevant to an identified area will be included.

# COURSE CHECKLIST

- Software
- Version Control
- Environment
- Workflow
- Reproducible Research
- Tidyverse vs/& Base R
- To GUI or not to GUI
- Data Mining
- Datasets, Data Repositories
- R Packages
- Student Exemplars
- Resources

# SOFTWARE

R

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



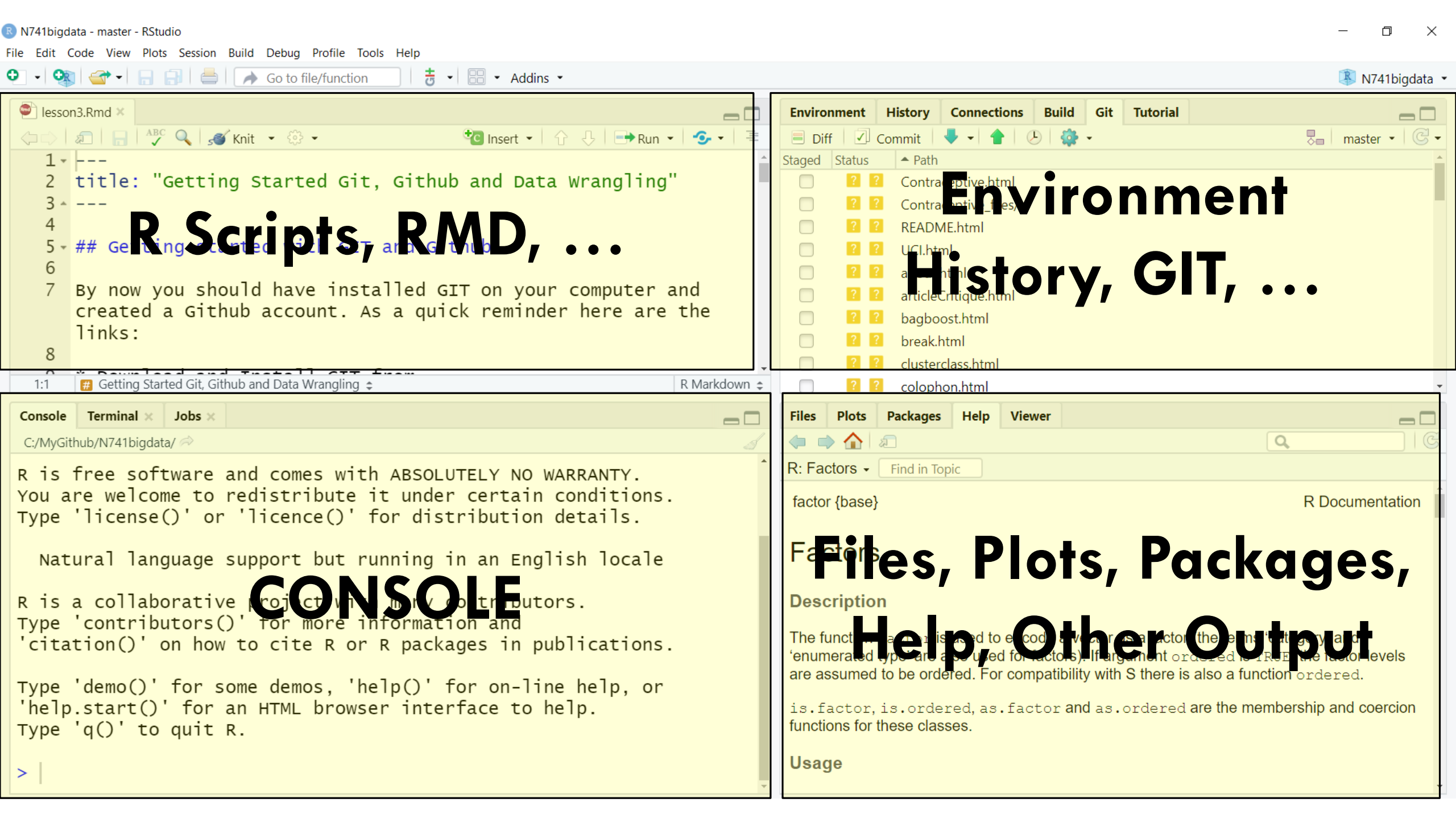
Rstudio

<https://rstudio.com/products/rstudio/download/>

Git

<https://git-scm.com/>





R Scripts, RMD, ...

Environment  
History, GIT, ...

CONSOLE

Files, Plots, Packages,  
Help, Other Output

# VERSION CONTROL



Github, <https://github.com/>

[Gitlab, <https://about.gitlab.com/>]



“Happy Git and GitHub for the User”

by Jenny Bryan, [<https://happygitwithr.com/>]

## History for [N741bigdata](#) / [\\_site.yml](#)

Commits on Jan 15, 2020

**update links to hmwk**



melindahiggins2000 committed on Jan 15 ✓



[97c868a](#)



**add files for 2020**



melindahiggins2000 committed on Jan 15 ✓



[4fffc4c](#)



Commits on Apr 24, 2019

**add files networks lecture**



melindahiggins2000 committed on Apr 24, 2019 ✓



[96ab8d1](#)



Commits on Apr 17, 2019

**add hmwk8 files**



melindahiggins2000 committed on Apr 17, 2019 ✓



[847b8c2](#)



# ENVIRONMENT(S)

PC & Macs (no Linux to date)

Rstudio.cloud, <https://rstudio.cloud/>

**\*\* new pricing updates Aug 3 \*\***

AWS – demos only



Local R/Rstudio server (we haven't done – maybe future)

<https://rstudio.com/products/rstudio/#rstudio-server>

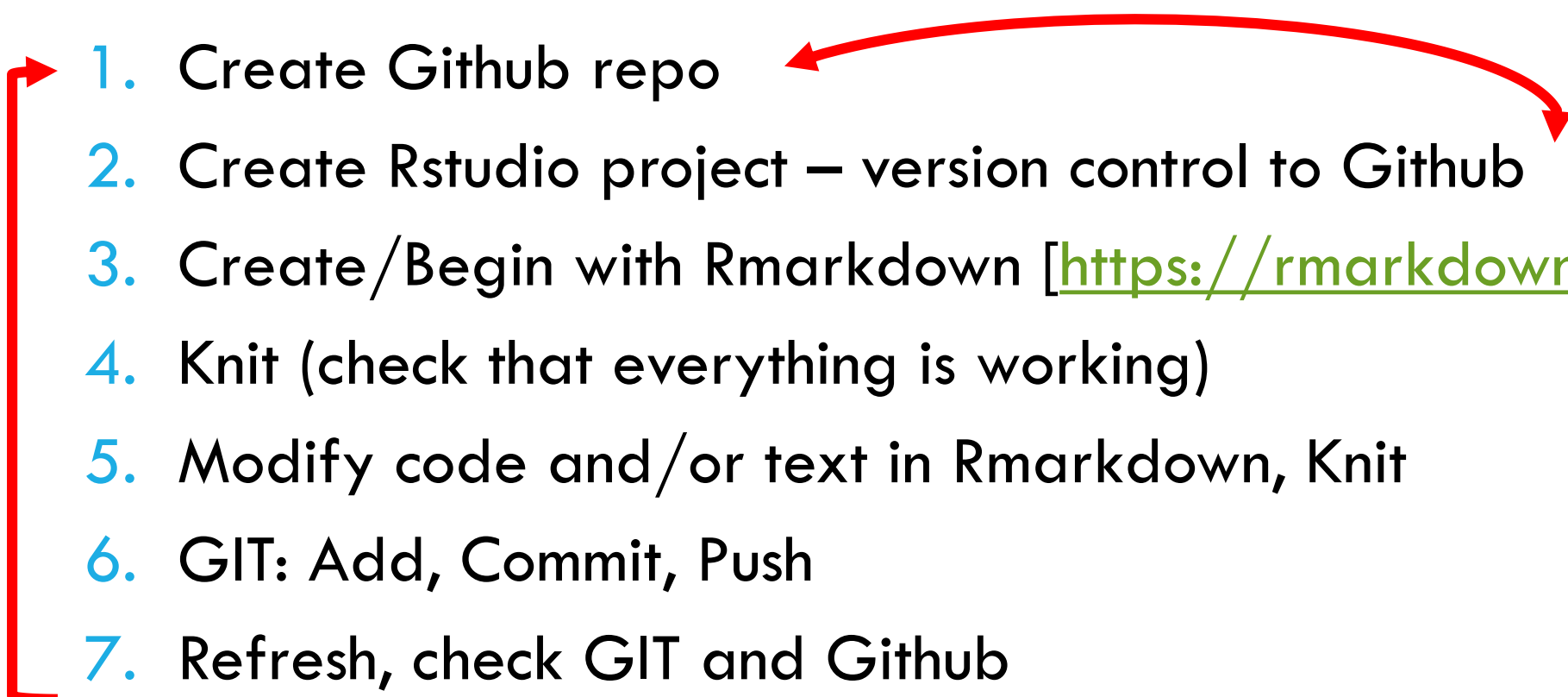




# REPRODUCIBLE RESEARCH

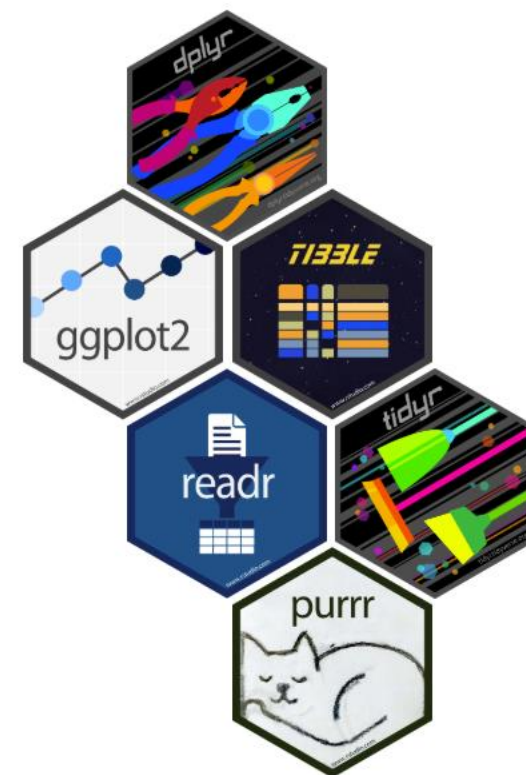
- Emphasized from day 1
- All exercises and homeworks link Github ↔ Rstudio project
- Rmarkdown: data, code, document immediately linked
- Learn to “knit” to multiple formats <https://rmarkdown.rstudio.com/>
  - documents – HTML, PDF, DOC
  - slides – HTML (ioslides, slidy), PDF (Beamer)
  - others – e.g. dashboards

# WORKFLOW

- 
1. Create Github repo
  2. Create Rstudio project – version control to Github
  3. Create/Begin with Rmarkdown [<https://rmarkdown.rstudio.com/>]
  4. Knit (check that everything is working)
  5. Modify code and/or text in Rmarkdown, Knit
  6. GIT: Add, Commit, Push
  7. Refresh, check GIT and Github

# TIDYVERSE VS/ & BASE R

- Tidyverse – packages that work well together
  - **dplyr** - pipe %>% workflow
  - **ggplot2** – build graphs with + layers
- Base R
  - tibble data frames  $\neq$  data.frame
  - data import **haven** vs **foreign** (SAS, SPSS or Stata files)
  - “haven labeled” variables
  - factors (pros and cons – useful to have both)
  - selecting variables (dplyr::select() and dplyr::pull() versus \$ versus [,2] – useful to know all of these)



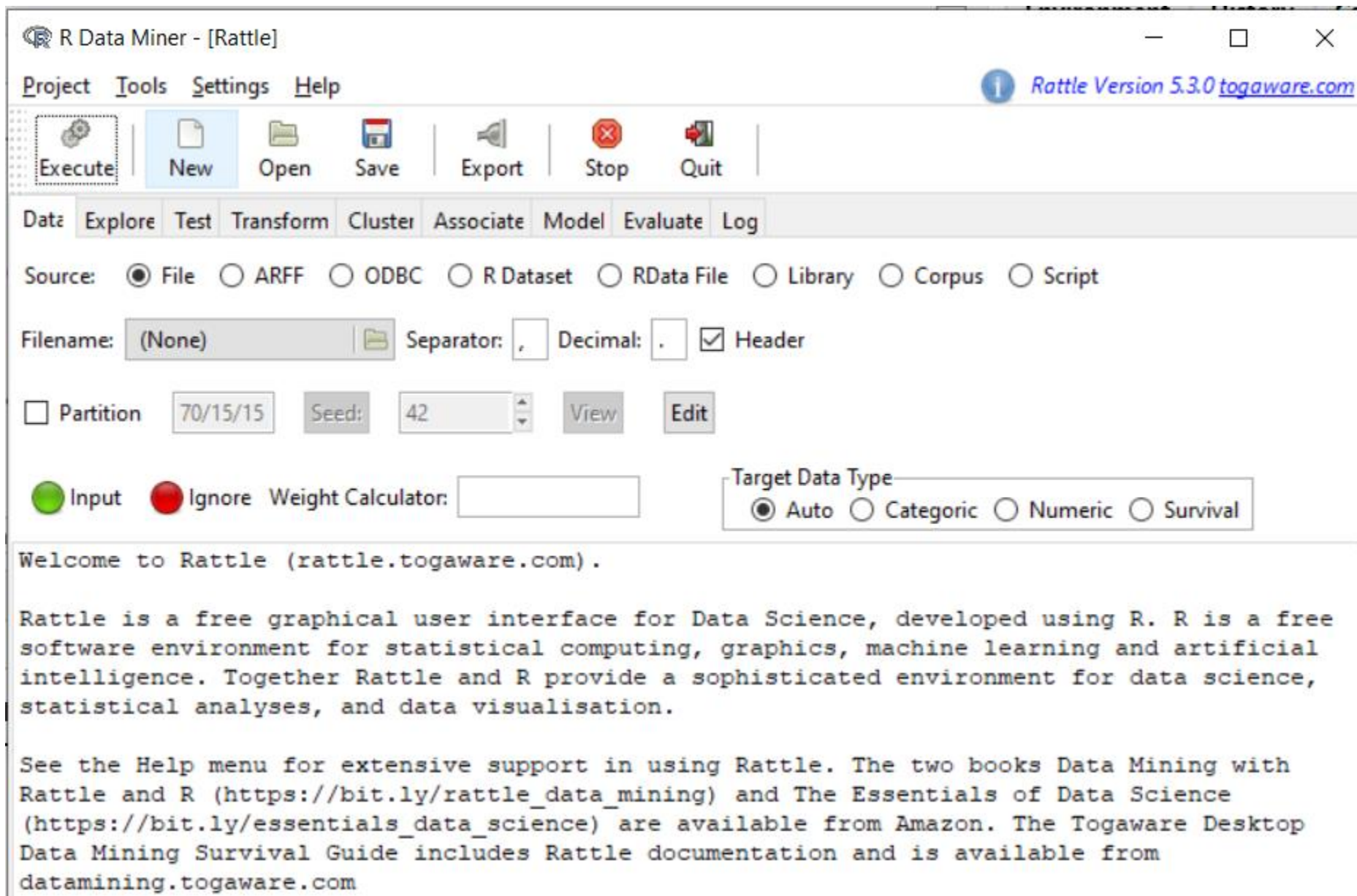
# TO GUI OR NOT TO GUI

- no GUI – all code
  - every step is captured and documented
  - Rmarkdown always begins with clean environment supports reproducible research workflow

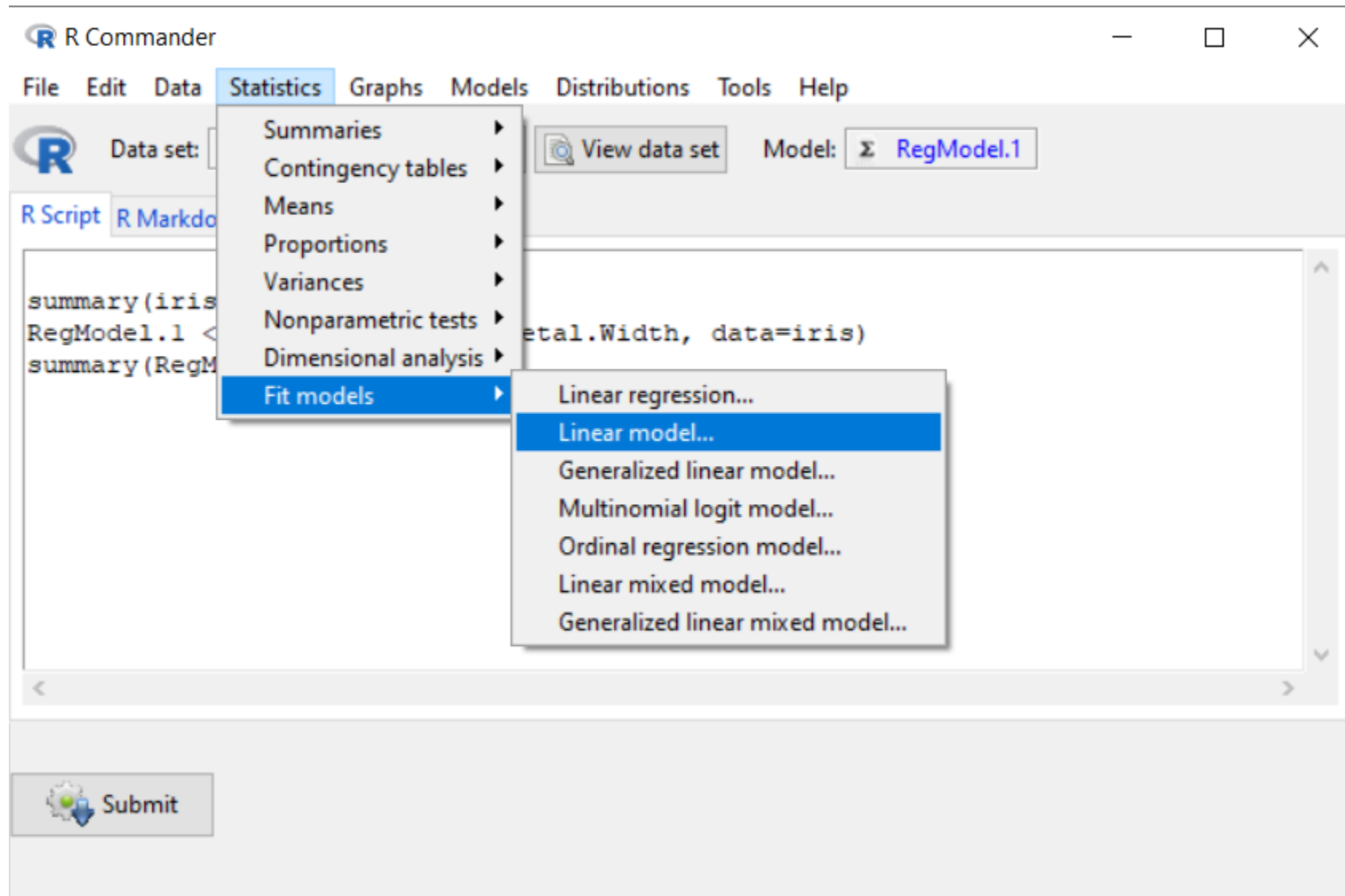
# TO GUI OR NOT TO GUI

- GUIs - packages: **rattle** and **Rcmdr**
  - very helpful for beginners
  - provides insights into data mining
  - **rattle**, <https://rattle.togaware.com/>
    - saves all R code
- **Rcmdr**, <https://www.rcommander.com/>
  - saves all R code
  - also creates a draft Rmarkdown file

<https://rattle.togaware.com/>



<https://www.rcommander.com/>



# STUDENT EXEMPLARS

2017, 2018, 2019, 2020 Spring Semesters

Public Health Datasets – cancer registries, CDC, Medicare, WHO, Pew Research Center

Research datasets – funded grants

Regression (linear and logistic)

Text mining – from Twitter, Web scraping (MLB.com)

Microbiome – American Gut Project (diversity, classification)

Others – PCA, CART, random forests, networks



# SUPPLEMENTAL SLIDES

My contact info:

[Melinda.higgins@emory.edu](mailto:Melinda.higgins@emory.edu)

<https://melindahiggins.netlify.app/>

<http://nursing.emory.edu/faculty-and-research/directory/profile.html?id=980>

# DATASETS, DATA REPOSITORIES

- **Gapminder** package, <https://cran.r-project.org/web/packages/gapminder/>
- UCI (Univ of CA Irvine) Machine Learning Repository, <https://archive.ics.uci.edu/ml/index.php>
- Datasets in R packages:
  - **carData**
  - **rpart**
  - **NHANES ...**

# DATA MINING TOPICS

- exploratory stats and graphs
  - `summary()`
  - `describe()` from Hmisc package
  - histograms, boxplots, scatterplots, facets
- regression – linear: `lm()` and logistic: `glm()`
- unsupervised – PCA, MDS, K-means
- supervised – CART, random forests, SVM, KNN
- **rattle** package – workflow: training/testing
- **Rcmdr** – workflow: build model then evaluate model

# HELPFUL R PACKAGES

- **tidyverse** – mainly **dplyr**, **ggplot2**, **readr**
- **foreign** – importing of SAS, SPSS, Stata
- **Hmisc** – lots of useful functions from Frank Harrell, <https://cran.r-project.org/web/packages/Hmisc/index.html>
- **table1** – great for making Rmarkdown summary tables, <https://cran.r-project.org/web/packages/table1/index.html>
- **knitr**, **Rmarkdown**, **printr**, **kablextra**
- **tinytex** - create PDFs without full LaTeX installation!!

# RESOURCES

- Happy Git and Github for the UseR, <https://happygitwithr.com/>
- Stat 545, <https://stat545.com/> and <https://stat545.stat.ubc.ca/>
- Quick R, <https://www.statmethods.net/>
- R Graphics Cookbook, <https://r-graphics.org/> and <http://www.cookbook-r.com/Graphs/>

# RESOURCES

- Rstudio education, <https://education.rstudio.com/>
- Datacamp for the classroom, <https://www.datacamp.com/groups/education>
- Github education, <https://education.github.com/>
- Gitlab for education, <https://about.gitlab.com/solutions/education/>
- Mine Cetinkaya-Rundel, <https://mine-cetinkaya-rundel.github.io/teach-r-online/> - also see **ghclass** R package for managing students in Github