# Data handling

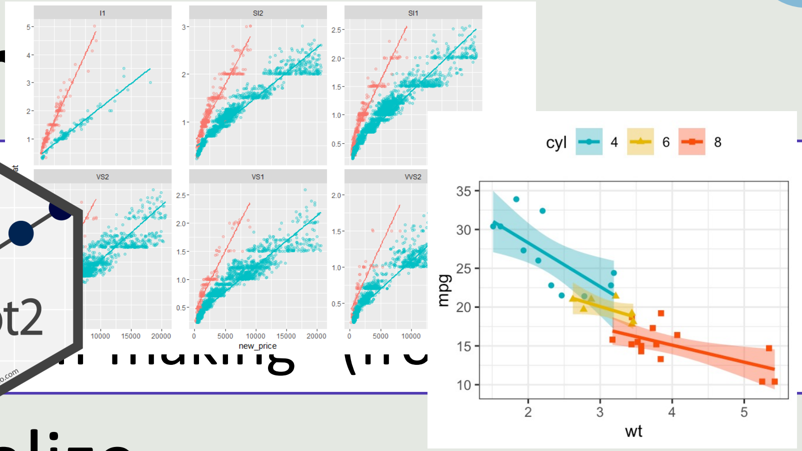
Kanan Saikai

NJC stat seminar series Part 1

July 16, 2021

# What is data ar

“Data analysis is a process of inspecting
and modeling data with the goal of dis
informing conclusions d supporting n making" (Wi

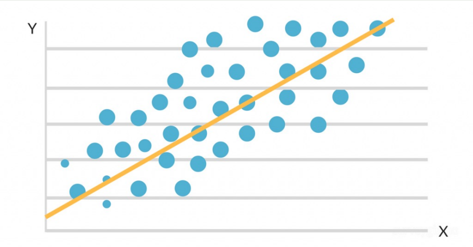


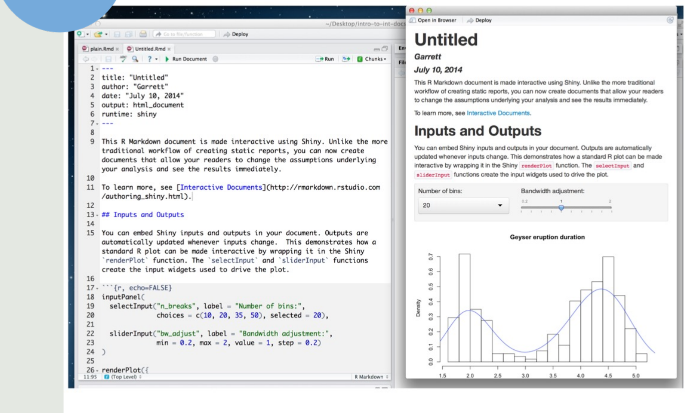
Import → Tidy → Transform

Visualize

Model

→ Communicate





~~~
---
title: "Untitled"
author: "Garrett"
date: "July 10, 2014"
output: html_document
runtime: shiny
---

This R Markdown document is made interactive using Shiny. Unlike the more
traditional workflow of creating static reports, you can now create
documents that allow your readers to change the assumptions underlying
your analysis and see the results immediately.

To learn more, see [Interactive Documents](http://rmarkdown.rstudio.com
/authoring_shiny.html).

## Inputs and Outputs

You can embed Shiny inputs and outputs in your document. Outputs are
automatically updated whenever inputs change.  This demonstrates how a
standard R plot can be made interactive by wrapping it in the Shiny
`renderPlot` function. The `selectInput` and `sliderInput` functions
create the input widgets used to drive the plot.

```{r, echo=FALSE}
inputPanel(
  selectInput("n_breaks", label = "Number of bins:",
              choices = c(10, 20, 35, 50), selected = 20),

  sliderInput("bw_adjust", label = "Bandwidth adjustment:",
              min = 0.2, max = 2, value = 1, step = 0.2)
)

renderPlot({
~~~

**Untitled**

*Garrett*

*July 10, 2014*

This R Markdown document is made interactive using Shiny. Unlike the more traditional workflow of creating static reports, you can now create documents that allow your readers to change the assumptions underlying your analysis and see the results immediately.

To learn more, see Interactive Documents.

**Inputs and Outputs**

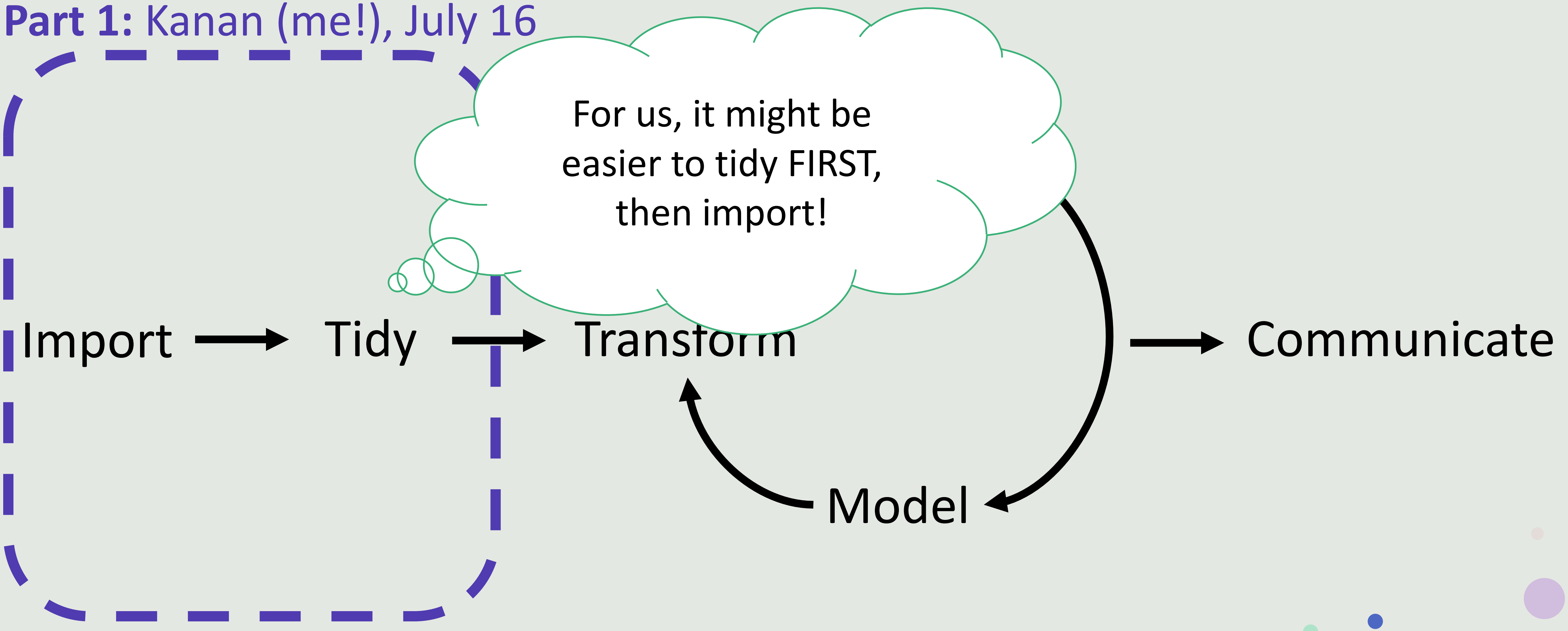
You can embed Shiny inputs and outputs in your document. Outputs are automatically updated whenever inputs change. This demonstrates how a standard R plot can be made interactive by wrapping it in the Shiny `renderPlot` function. The `selectInput` and `sliderInput` functions create the input widgets used to drive the plot.

Number of bins: 20

Bandwidth adjustment: 1

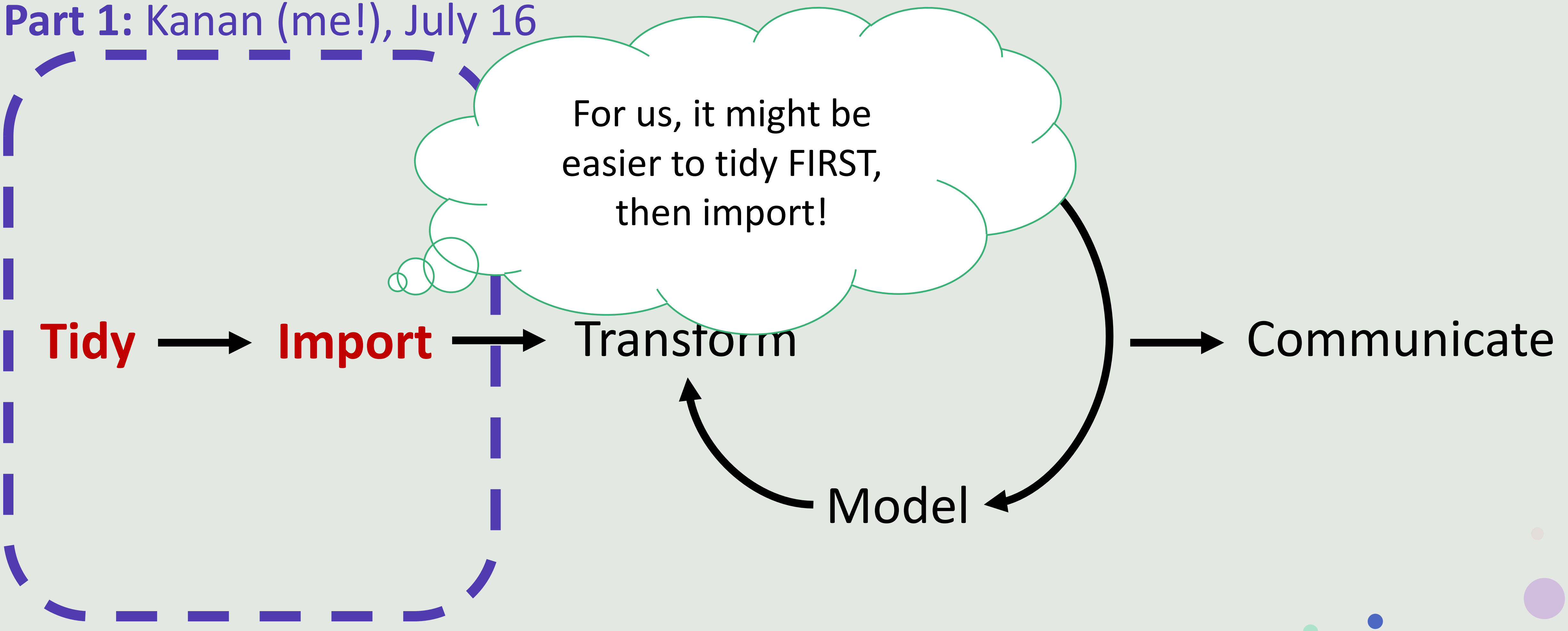
Hadley Wickham and Garrett Grolemund (2016) R for Data Science, O'Reilly Media, Inc.

# What is data analysis?


Part 1: Kanan (me!), July 16
For us, it might be easier to tidy FIRST, then import!
Import
Tidy
Transform
Model
Communicate

Hadley Wickham and Garrett Grolemund (2016) R for Data Science, O'Reilly Media, Inc.

# What is data analysis?



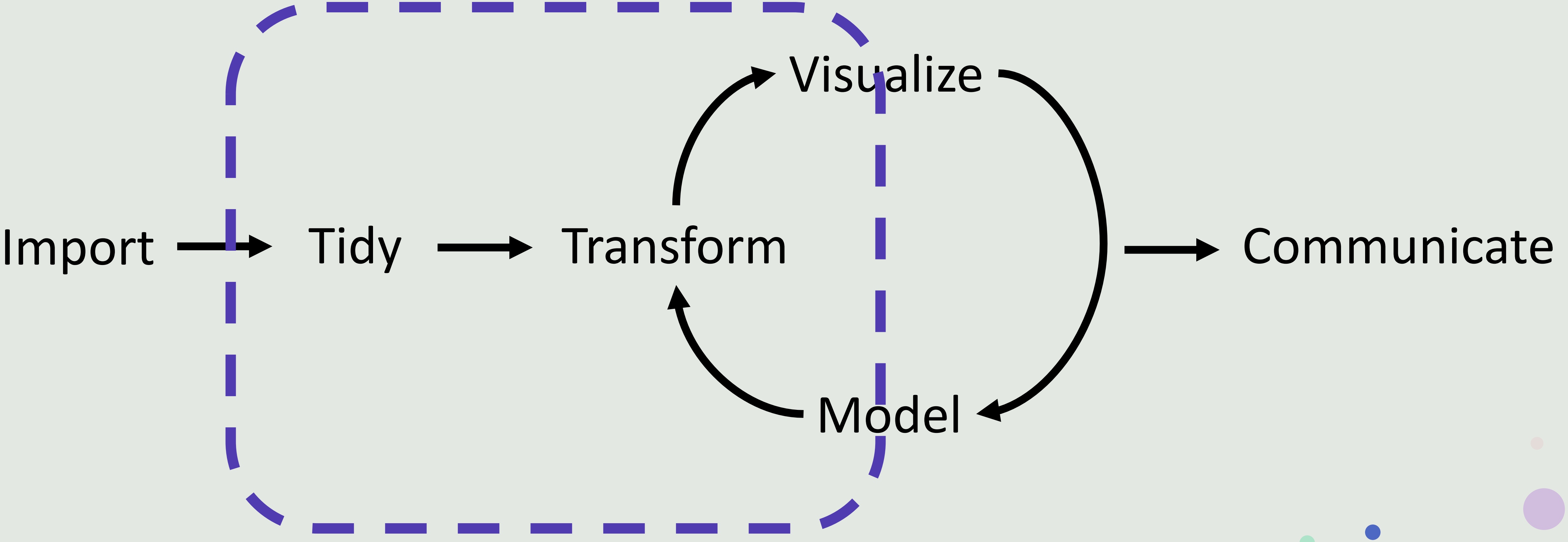
For us, it might be easier to tidy FIRST, then import!

Hadley Wickham and Garrett Grolemund (2016) R for Data Science, O'Reilly Media, Inc.

# What is data analysis?

**Part 2:** Dr. Masatoshi Katabuchi, July 23

Plant Ecologist @ Xishuangbanna Tropical Botanical Garden

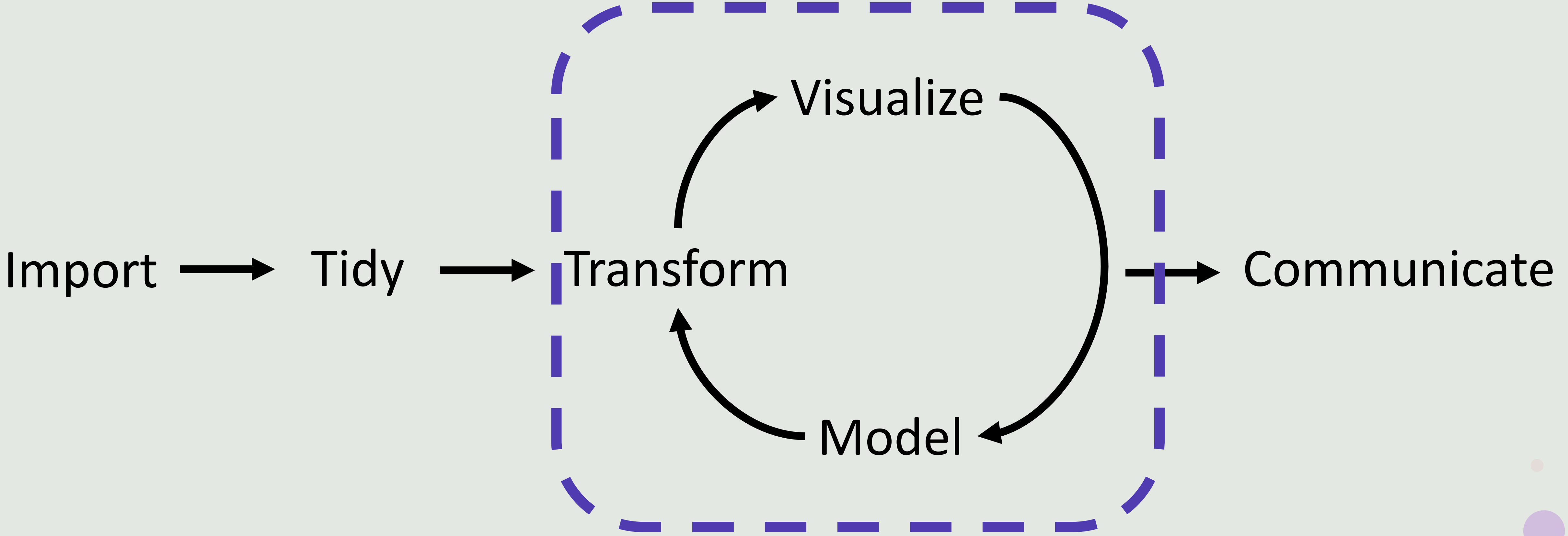


Hadley Wickham and Garrett Grolemund (2016) R for Data Science, O'Reilly Media, Inc.

# What is data analysis?

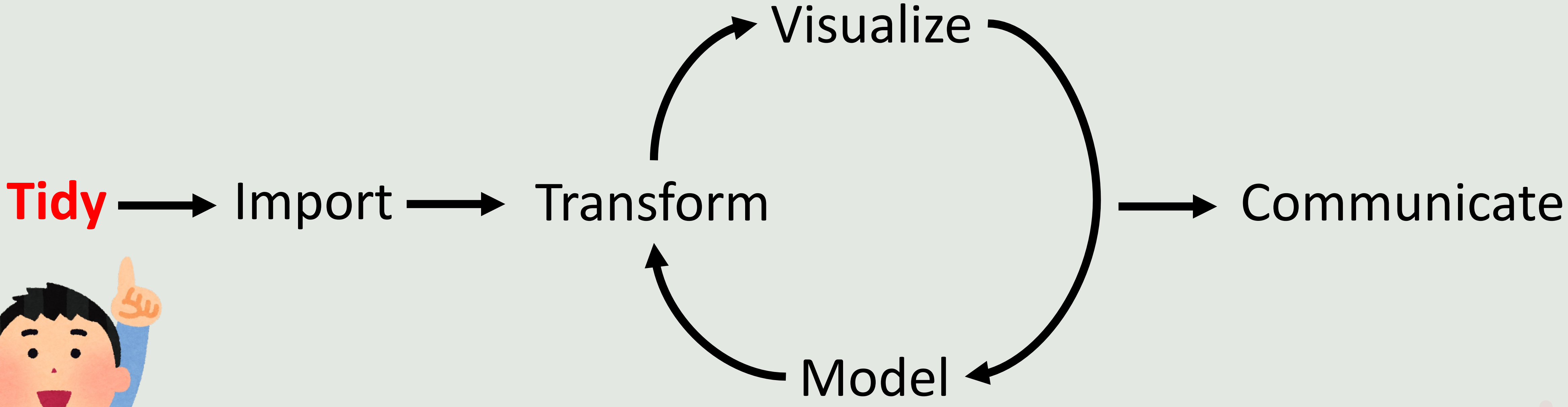
**Part 3:** Dr. Hyunseung Kang, July 30

Statistician @ University of Wisconsin-Madison



Hadley Wickham and Garrett Grolemund (2016) R for Data Science, O'Reilly Media, Inc.

# Let’s tidy!

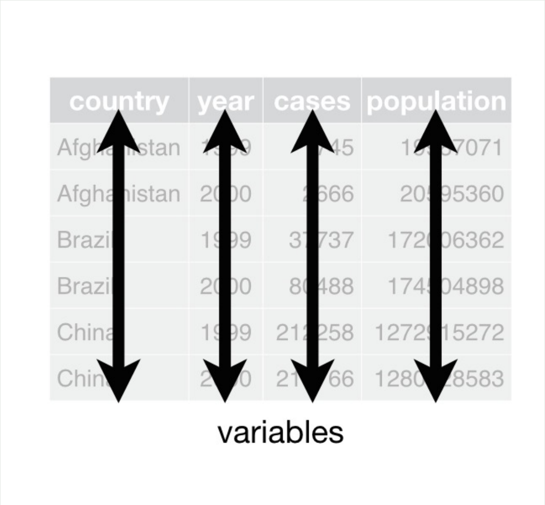


Hadley Wickham and Garrett Grolemund (2016) R for Data Science, O'Reilly Media, Inc.

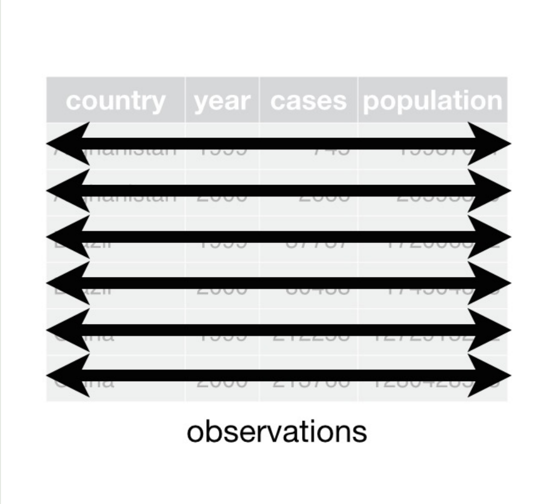
# Let’s tidy your data!

But what is tidy data?

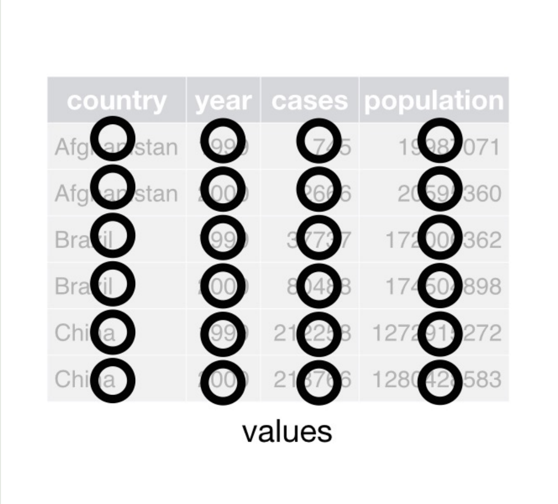
- Each variable must have its own column
- Each observation must have its own row
- Each value must have its own cell



variables



observations



values

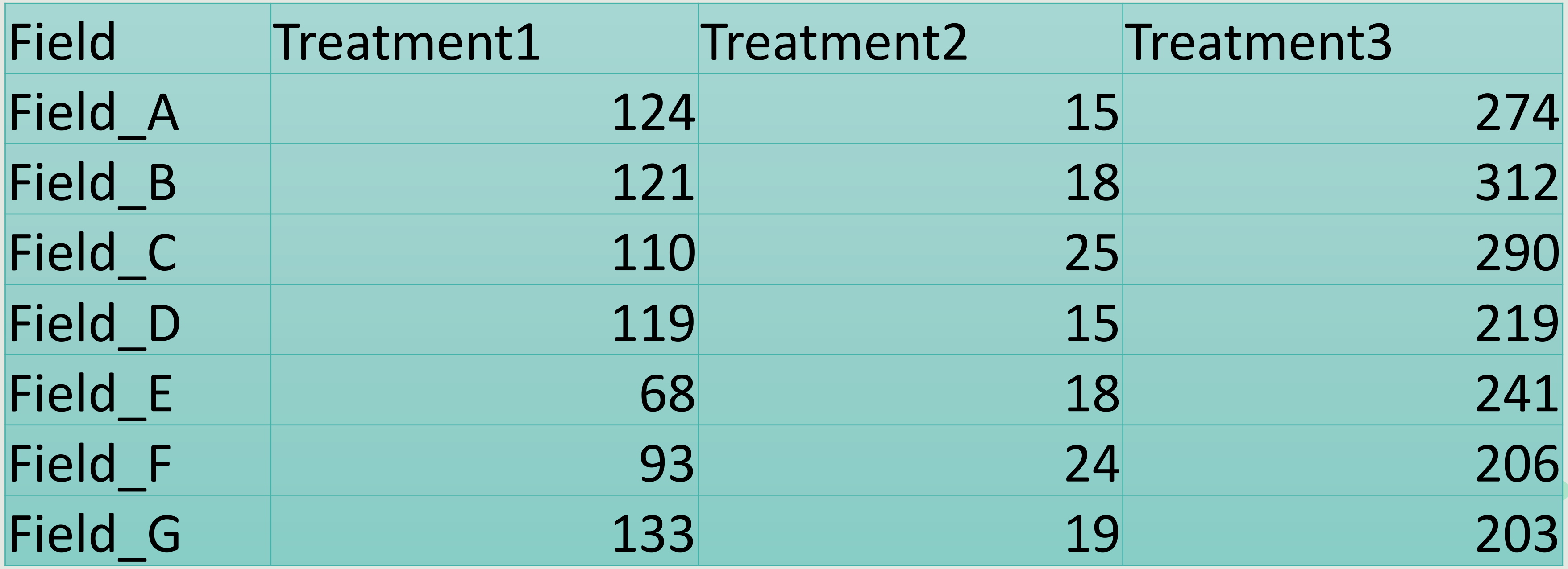
TokyoR#91 material from Masatoshi Katabuchi

Wickham, Hadley. 2014. “Tidy Data.” Journal of Statistical Software, Articles 59 (10): 1–23

# Exercise 1

Download the file “exercise_1.csv” :
https://www.dropbox.com/s/68jloxnvdcblfx2/exercise_1.csv?dl=0

1. Explain why this data is untidy.
2. Rearrange the data frame to make it tidy.



| Field | Treatment1 | Treatment2 | Treatment3 |
|---|---:|---:|---:|
| Field_A | 124 | 15 | 274 |
| Field_B | 121 | 18 | 312 |
| Field_C | 110 | 25 | 290 |
| Field_D | 119 | 15 | 219 |
| Field_E | 68 | 18 | 241 |
| Field_F | 93 | 24 | 206 |
| Field_G | 133 | 19 | 203 |

# Let’s tidy your data!

## Non-tidy data

| Field | Treatment_1 | Treatment_2 |
|---|---:|---:|
| Field_A | 124 | 15 |
| Field_B | 121 | 18 |
| Field_C | 110 | 25 |

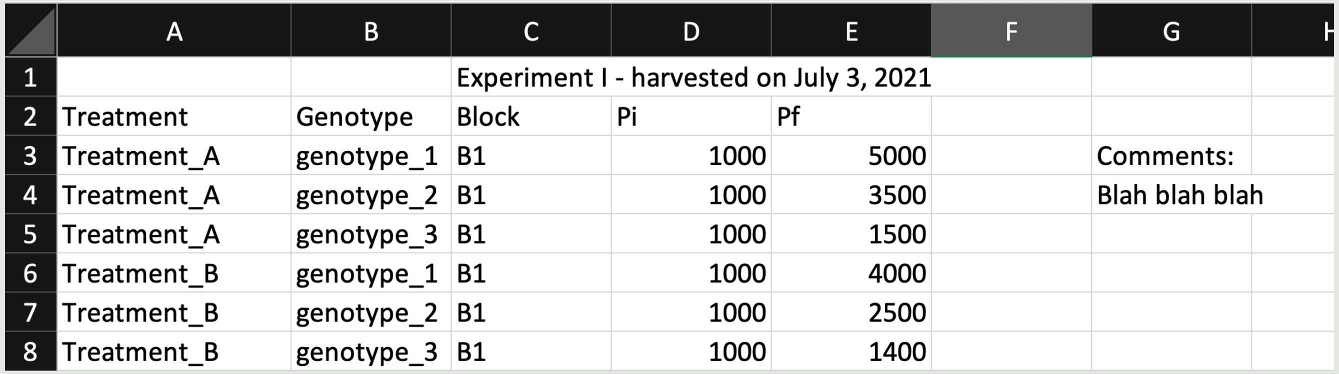
## Tidy data

| Field | Treatment | Nematode number |
|---|---|---|
| Field_A | Treatment_1 | 124 |
| Field_B | Treatment_1 | 121 |
| Field_C | Treatment_1 | 110 |
| Field_A | Treatment_2 | 15 |
| Field_B | Treatment_2 | 18 |
| Field_C | Treatment_2 | 25 |

TokyoR#91 material from Masatoshi Katabuchi
Wickham, Hadley. 2014. “Tidy Data.” Journal of Statistical Software, Articles 59 (10): 1–23

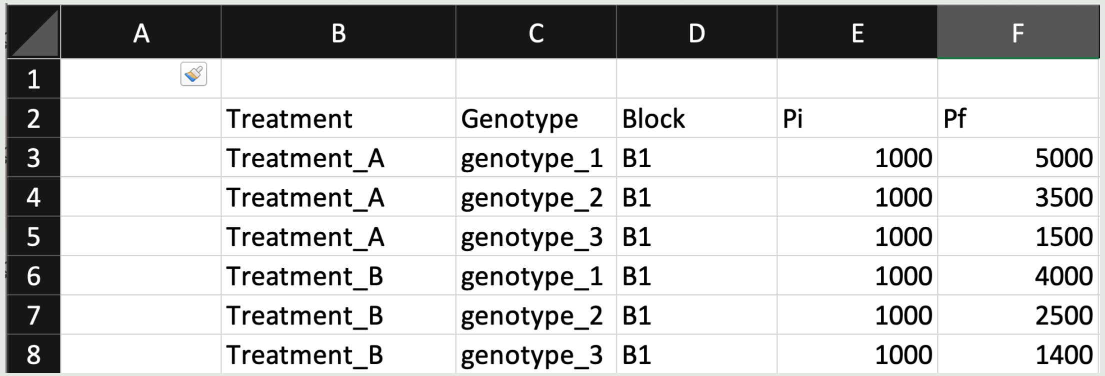
# Other common mistakes

## With comments / titles



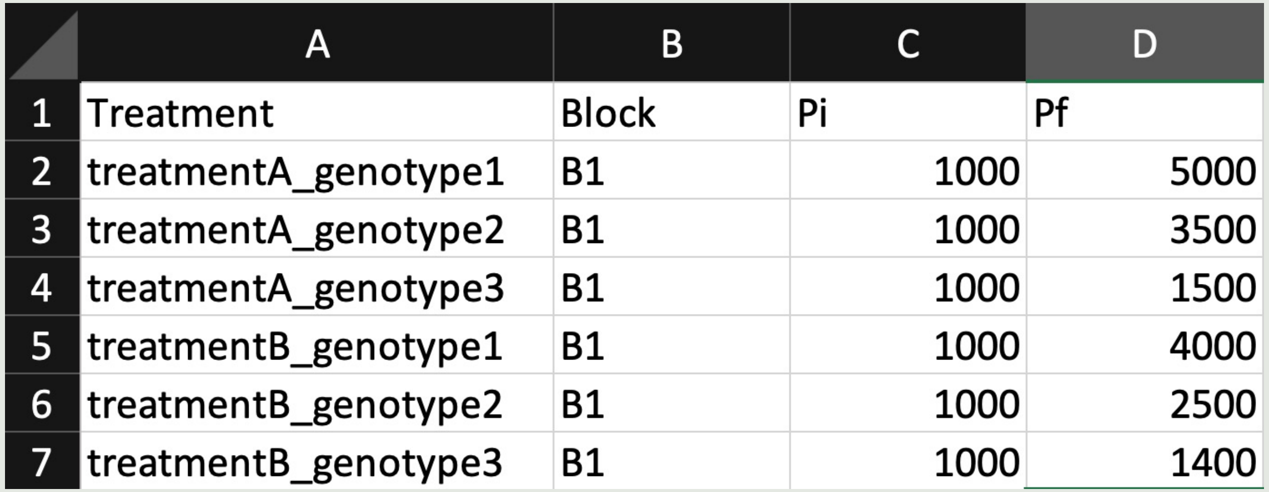
|   | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 |  |  | Experiment I - harvested on July 3, 2021 |  |  |  |  |  |
| 2 | Treatment | Genotype | Block | Pi | Pf |  |  |  |
| 3 | Treatment_A | genotype_1 | B1 | 1000 | 5000 |  | Comments: |  |
| 4 | Treatment_A | genotype_2 | B1 | 1000 | 3500 |  | Blah blah blah |  |
| 5 | Treatment_A | genotype_3 | B1 | 1000 | 1500 |  |  |  |
| 6 | Treatment_B | genotype_1 | B1 | 1000 | 4000 |  |  |  |
| 7 | Treatment_B | genotype_2 | B1 | 1000 | 2500 |  |  |  |
| 8 | Treatment_B | genotype_3 | B1 | 1000 | 1400 |  |  |  |

## No data entry in the first row / first column



|   | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 |  |  |  |  |  |  |
| 2 |  | Treatment | Genotype | Block | Pi | Pf |
| 3 |  | Treatment_A | genotype_1 | B1 | 1000 | 5000 |
| 4 |  | Treatment_A | genotype_2 | B1 | 1000 | 3500 |
| 5 |  | Treatment_A | genotype_3 | B1 | 1000 | 1500 |
| 6 |  | Treatment_B | genotype_1 | B1 | 1000 | 4000 |
| 7 |  | Treatment_B | genotype_2 | B1 | 1000 | 2500 |
| 8 |  | Treatment_B | genotype_3 | B1 | 1000 | 1400 |

## Variables are combined for one column



|   | A | B | C | D |
|---|---|---|---|---|
| 1 | Treatment | Block | Pi | Pf |
| 2 | treatmentA_genotype1 | B1 | 1000 | 5000 |
| 3 | treatmentA_genotype2 | B1 | 1000 | 3500 |
| 4 | treatmentA_genotype3 | B1 | 1000 | 1500 |
| 5 | treatmentB_genotype1 | B1 | 1000 | 4000 |
| 6 | treatmentB_genotype2 | B1 | 1000 | 2500 |
| 7 | treatmentB_genotype3 | B1 | 1000 | 1400 |

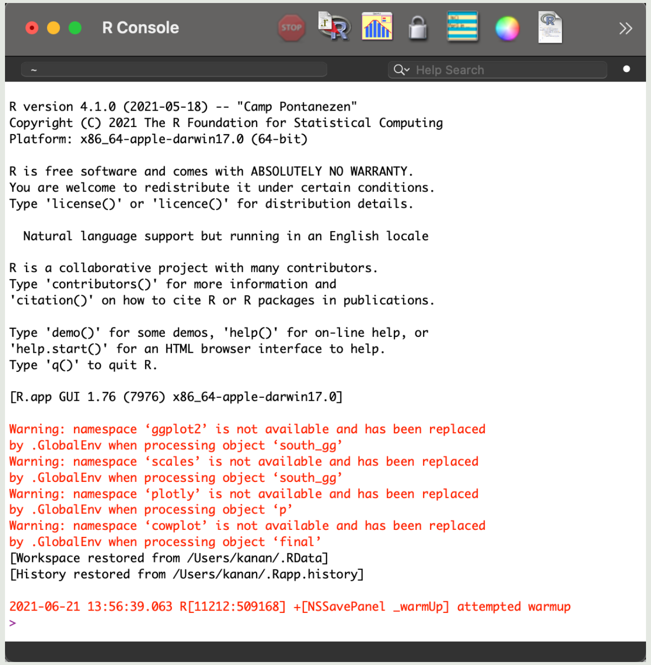
# Why learn R?

- Free, open source, cross platform
- 10,000+ “packages”
- Works on many data types
- Produced high-quality graphics
- Reproducibility and repeatability

Materials from Data Carpentry – R basics by Tobin Magle

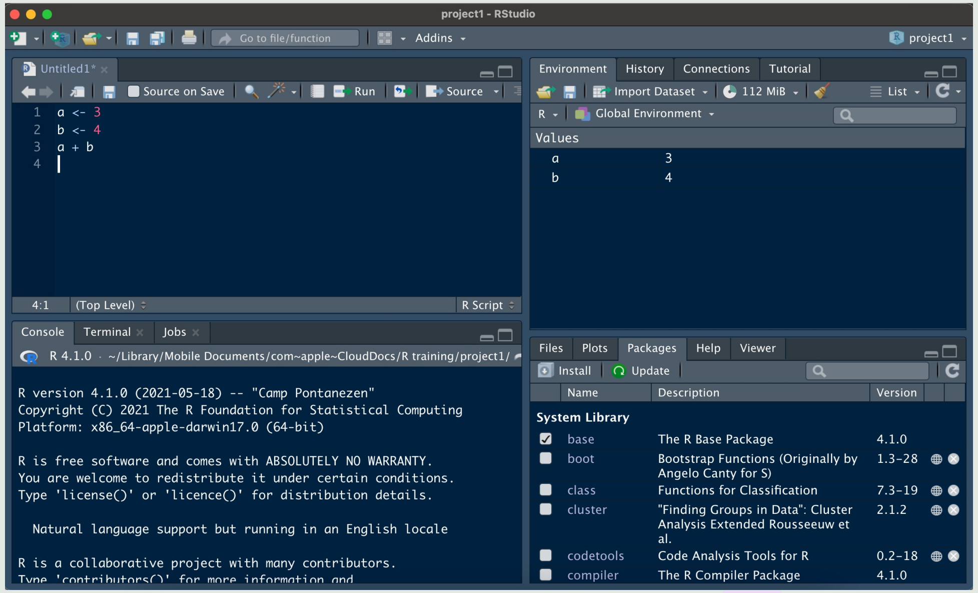
# Introduction of R & R Studio

A programming language + software that interprets it



```
R version 4.1.0 (2021-05-18) -- "Camp Pontanezen"
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.

[R.app GUI 1.76 (7976) x86_64-apple-darwin17.0]

Warning: namespace ‘ggplot2’ is not available and has been replaced
by .GlobalEnv when processing object ‘south_gg’
Warning: namespace ‘scales’ is not available and has been replaced
by .GlobalEnv when processing object ‘south_gg’
Warning: namespace ‘plotly’ is not available and has been replaced
by .GlobalEnv when processing object ‘p’
Warning: namespace ‘cowplot’ is not available and has been replaced
by .GlobalEnv when processing object ‘final’
[Workspace restored from /Users/kanan/.RData]
[History restored from /Users/kanan/.Rapp.history]

2021-06-21 13:56:39.063 R[11212:509168] +[NSSavePanel _warmUp] attempted warmup
>
```

A popular software to write R scripts and interact with the R software



```
a <- 3
b <- 4
a + b
```

Values

| | |
|---|---|
| a | 3 |
| b | 4 |

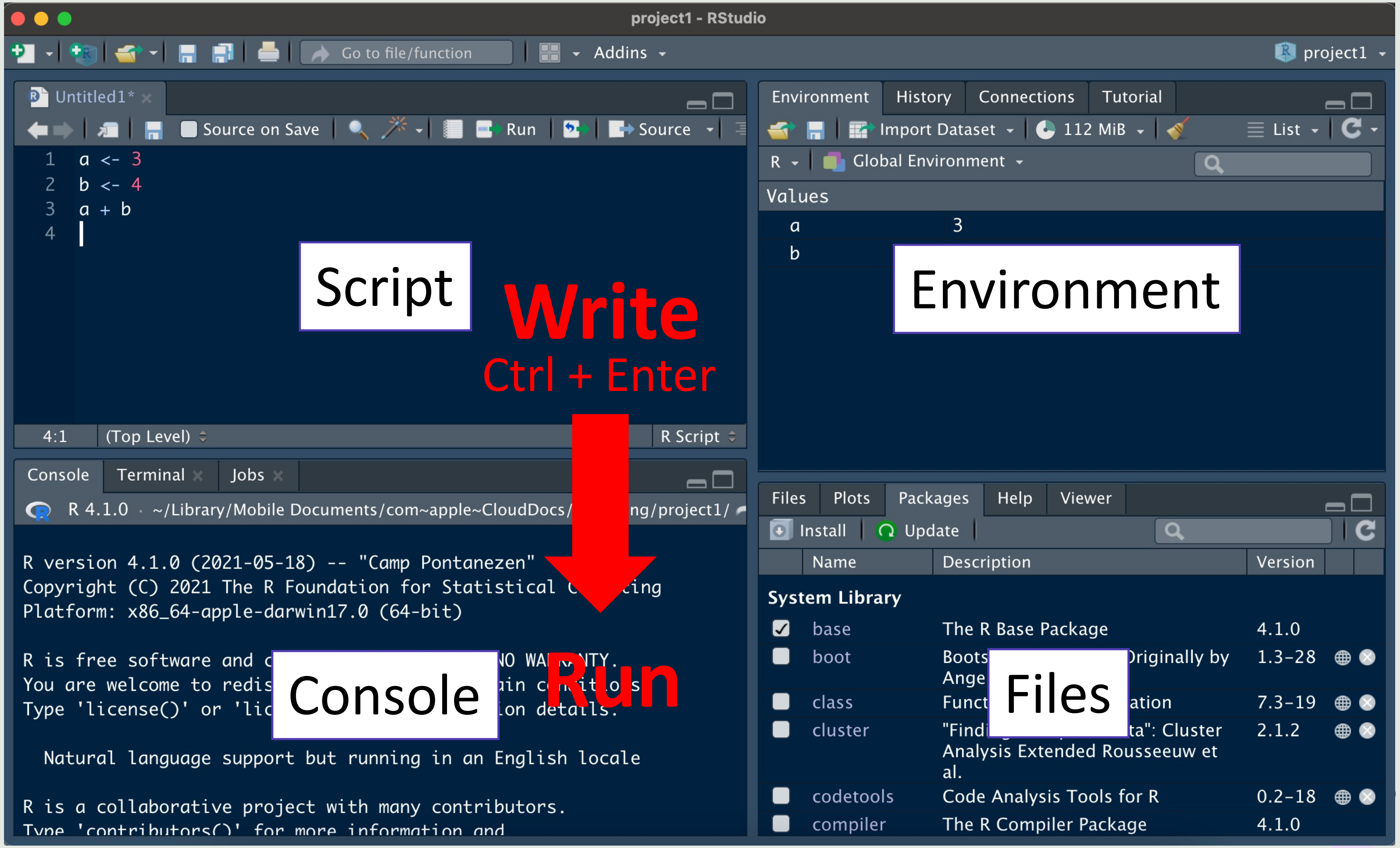
R 4.1.0 · ~/Library/Mobile Documents/com~apple~CloudDocs/R training/project1/

```
R version 4.1.0 (2021-05-18) -- "Camp Pontanezen"
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
```

| | Name | Description | Version |
|---|---|---|---|
| | **System Library** | | |
| ☑ | base | The R Base Package | 4.1.0 |
| □ | boot | Bootstrap Functions (Originally by Angelo Canty for S) | 1.3–28 |
| □ | class | Functions for Classification | 7.3–19 |
| □ | cluster | "Finding Groups in Data": Cluster Analysis Extended Rousseeuw et al. | 2.1.2 |
| □ | codetools | Code Analysis Tools for R | 0.2–18 |
| □ | compiler | The R Compiler Package | 4.1.0 |

https://www.r-

# RStudio®



```
a <- 3
b <- 4
a + b
```

```
R version 4.1.0 (2021-05-18) -- "Camp Pontanezen"
Copyright (C) 2021 The R Foundation for Statistical C    ing
Platform: x86_64-apple-darwin17.0 (64-bit)

R is free software and c                  NO WA   NTY.
You are welcome to redis                  ain c  it o s
Type 'license()' or 'lic                  ion details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
```

Values

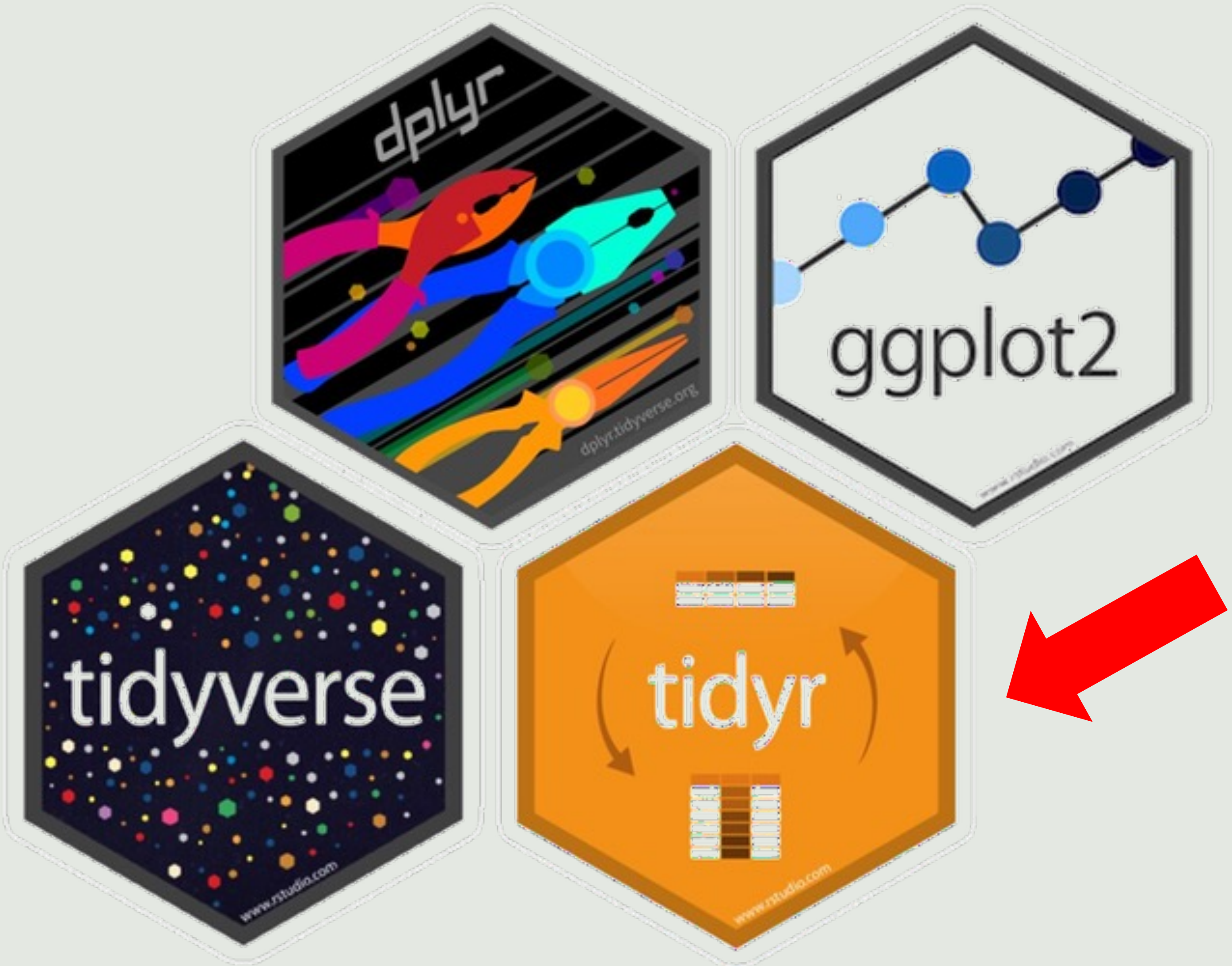
| | |
|---|---|
| a | 3 |
| b | |

System Library

| | Name | Description | Version |
|---|---|---|---|
| ☑ | base | The R Base Package | 4.1.0 |
| □ | boot | Boots Originally by Ange | 1.3-28 |
| □ | class | Funct ation | 7.3-19 |
| □ | cluster | "Findi ta": Cluster Analysis Extended Rousseeuw et al. | 2.1.2 |
| □ | codetools | Code Analysis Tools for R | 0.2-18 |
| □ | compiler | The R Compiler Package | 4.1.0 |

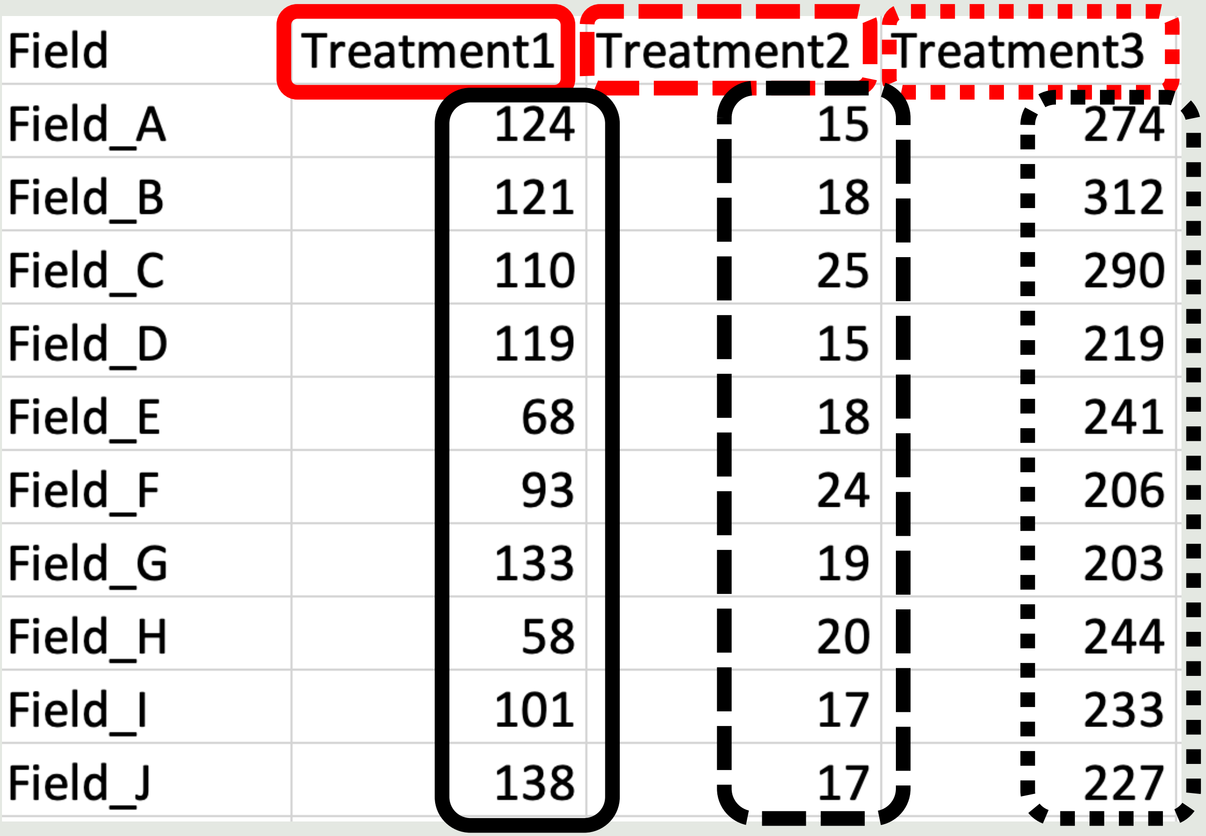
https://www.r-

# Tidyverse


dplyr
dplyr.tidyverse.org
ggplot2
www.rstudio.com
tidyverse
www.rstudio.com
tidyr
www.rstudio.com

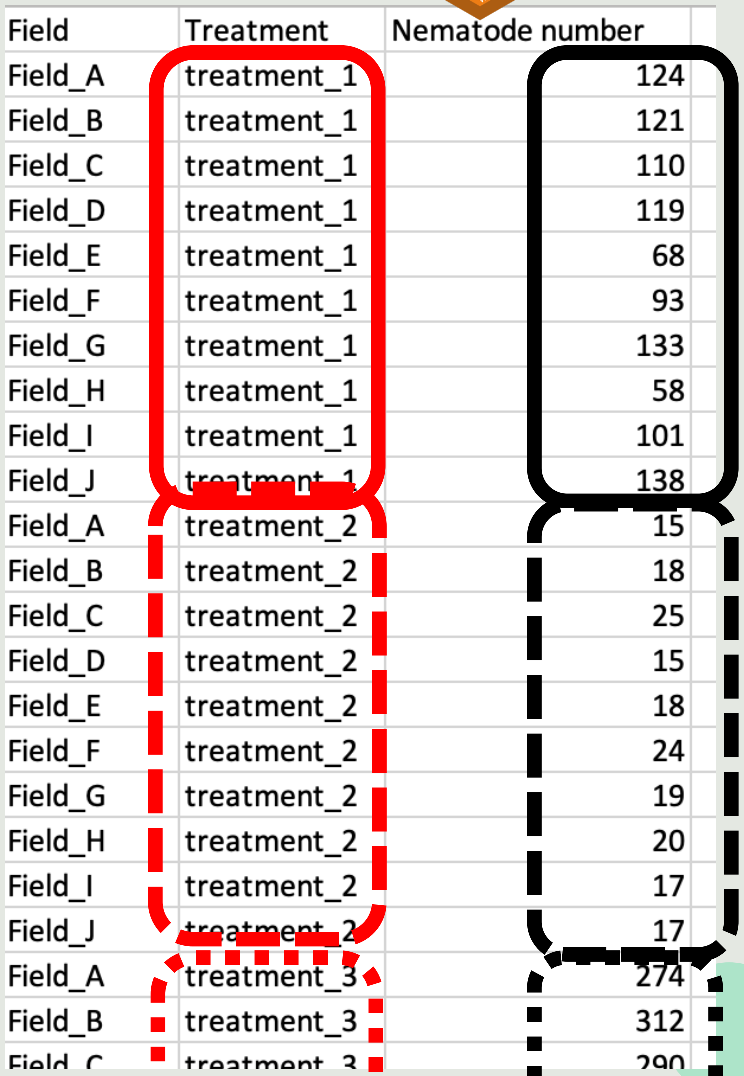
# Tidying data using {tidyr}

Let's tidy the data from the exercise 1 using {tidyr}!



| Field | Treatment1 | Treatment2 | Treatment3 |
|---|---|---|---|
| Field_A | 124 | 15 | 274 |
| Field_B | 121 | 18 | 312 |
| Field_C | 110 | 25 | 290 |
| Field_D | 119 | 15 | 219 |
| Field_E | 68 | 18 | 241 |
| Field_F | 93 | 24 | 206 |
| Field_G | 133 | 19 | 203 |
| Field_H | 58 | 20 | 244 |
| Field_I | 101 | 17 | 233 |
| Field_J | 138 | 17 | 227 |

Gathering
Spreading



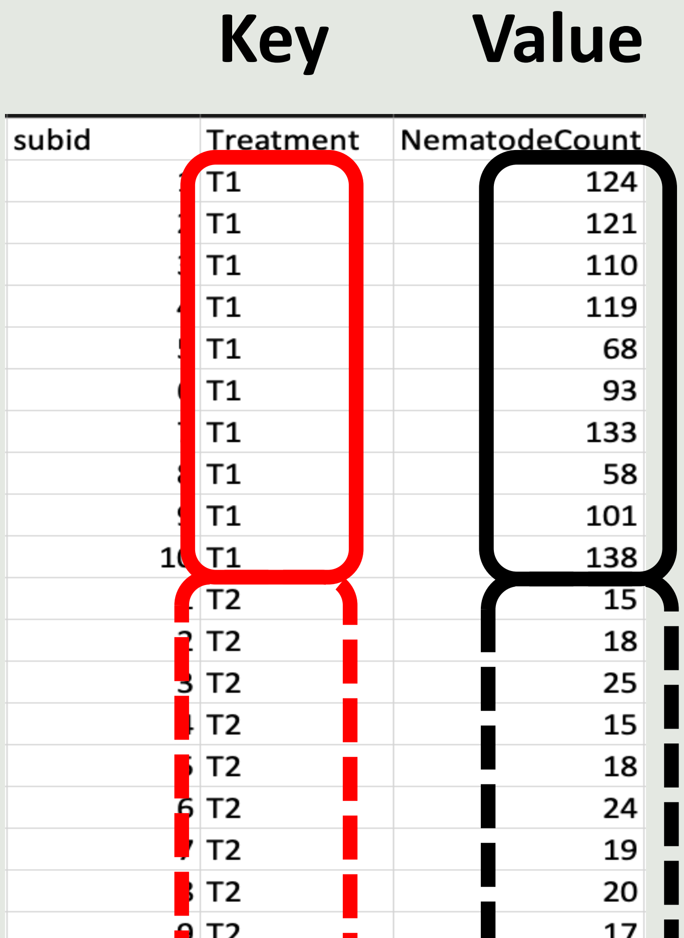
| Field | Treatment | Nematode number |
|---|---|---|
| Field_A | treatment_1 | 124 |
| Field_B | treatment_1 | 121 |
| Field_C | treatment_1 | 110 |
| Field_D | treatment_1 | 119 |
| Field_E | treatment_1 | 68 |
| Field_F | treatment_1 | 93 |
| Field_G | treatment_1 | 133 |
| Field_H | treatment_1 | 58 |
| Field_I | treatment_1 | 101 |
| Field_J | treatment_1 | 138 |
| Field_A | treatment_2 | 15 |
| Field_B | treatment_2 | 18 |
| Field_C | treatment_2 | 25 |
| Field_D | treatment_2 | 15 |
| Field_E | treatment_2 | 18 |
| Field_F | treatment_2 | 24 |
| Field_G | treatment_2 | 19 |
| Field_H | treatment_2 | 20 |
| Field_I | treatment_2 | 17 |
| Field_J | treatment_2 | 17 |
| Field_A | treatment_3 | 274 |
| Field_B | treatment_3 | 312 |
| Field_C | treatment_3 | 290 |

# gather()

- Use when column names are not names of variables, but values of a variable.

- Input:

  **data,**
  **key** column (created from col names),
  **values** column (fill the key variable),
  A range of columns to gather

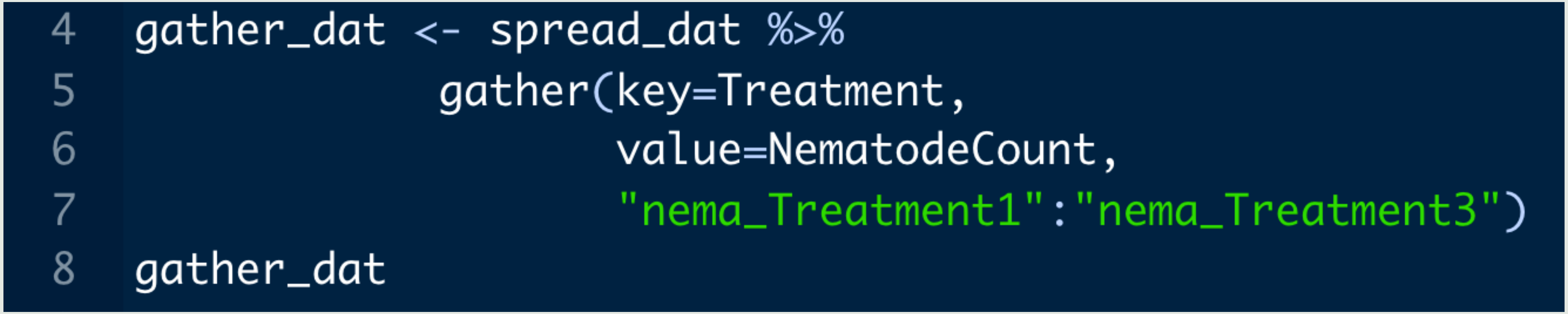


| subid | Treatment | NematodeCount |
|---|---|---|
| 1 | T1 | 124 |
| 2 | T1 | 121 |
| 3 | T1 | 110 |
| 4 | T1 | 119 |
| 5 | T1 | 68 |
| 6 | T1 | 93 |
| 7 | T1 | 133 |
| 8 | T1 | 58 |
| 9 | T1 | 101 |
| 10 | T1 | 138 |
| 1 | T2 | 15 |
| 2 | T2 | 18 |
| 3 | T2 | 25 |
| 4 | T2 | 15 |
| 5 | T2 | 18 |
| 6 | T2 | 24 |
| 7 | T2 | 19 |
| 8 | T2 | 20 |
| 9 | T2 | 17 |

Materials from Data Carpentry – R basics by Tobin Magle

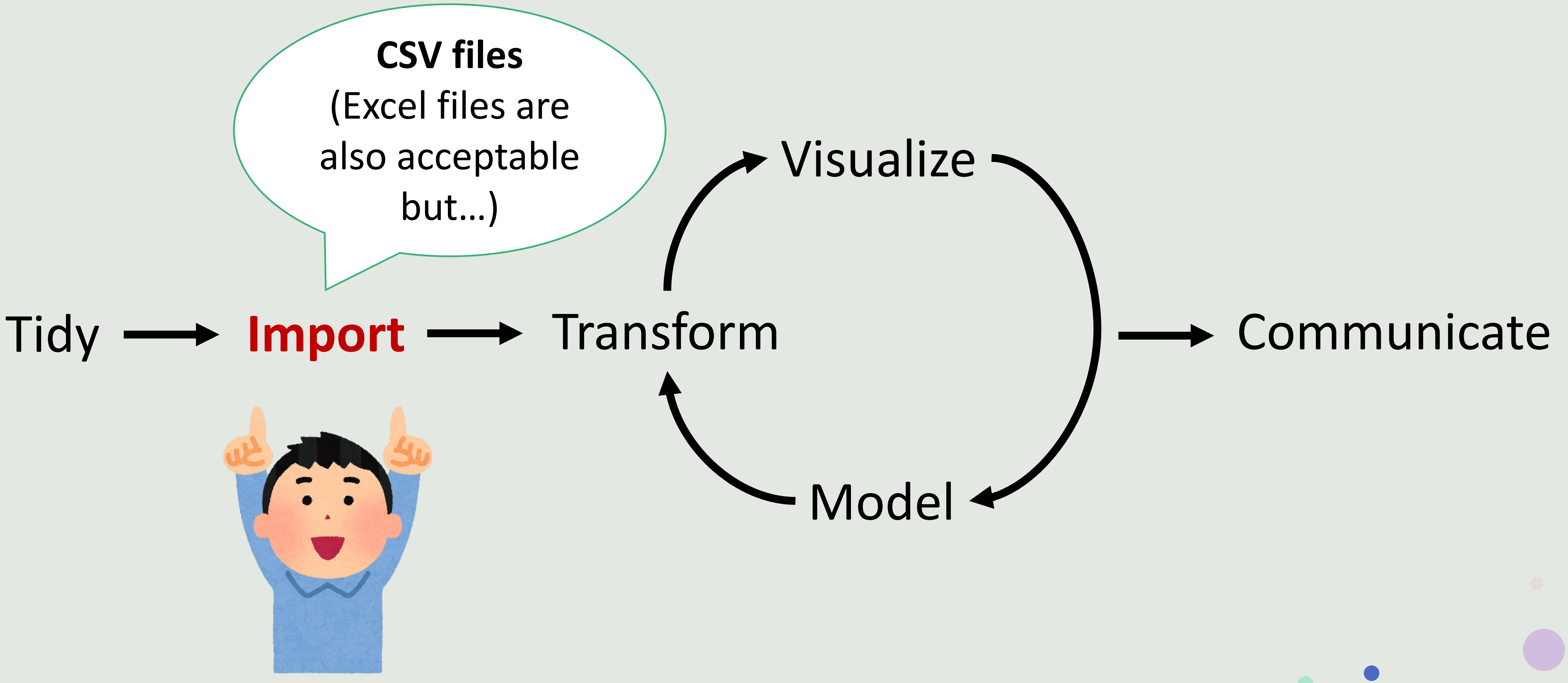
# Demonstration

# Tidying data using {tidyr}



```
gather_dat <- spread_dat %>%
            gather(key=Treatment,
                   value=NematodeCount,
                   "nema_Treatment1":"nema_Treatment3")
gather_dat
```

# Let’s import your data to R



**CSV files**
(Excel files are
also acceptable
but…)

Tidy → **Import** → Transform → Visualize → Model → Transform → Communicate

# But before you import… set up a working directory

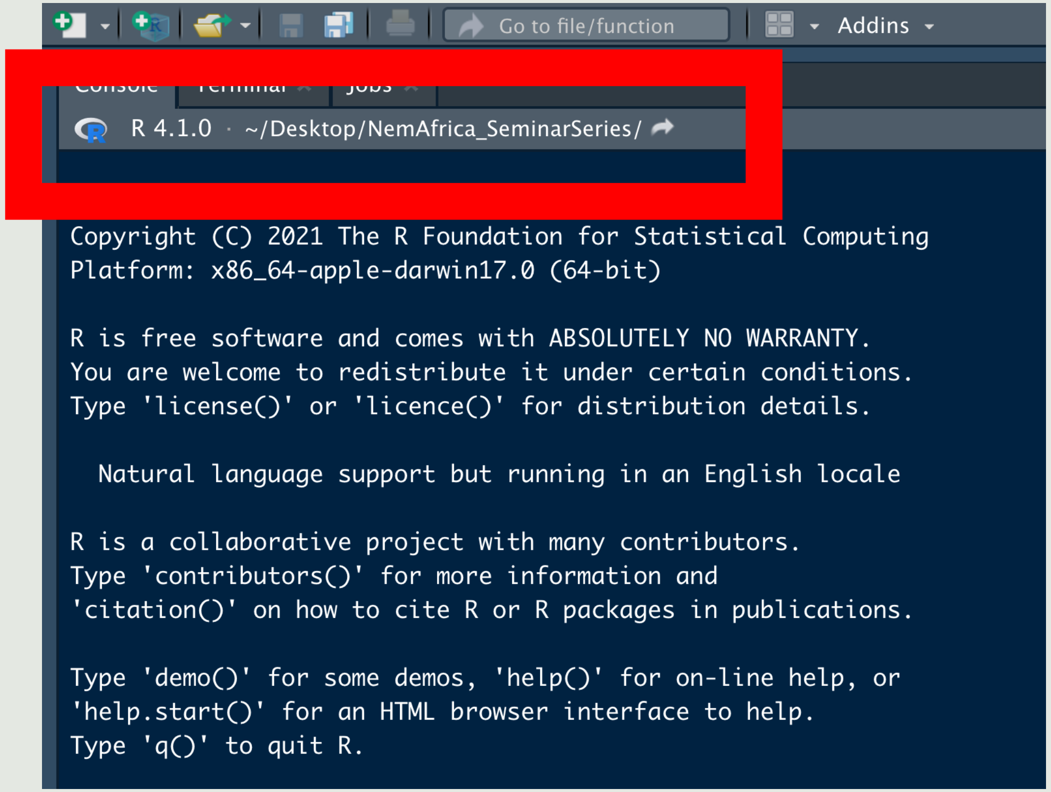
- Open Rstudio
- **File > New project > New directory** > Empty project
- Enter a name for this new folder
- Choose a convenient location
- Click “Create project”

Check which the working directory is: **getwd()**
Set working directory: **setwd()**

https://www.r-

# But before you import… create a new R script

- **File > New File > R script**
- Save it in your project directory
- Look on the top left of the R Studio window to see where it’s saved



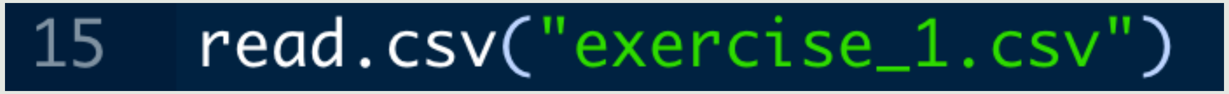
R 4.1.0 · ~/Desktop/NemAfrica_SeminarSeries/

```
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.
```

https://www.r-

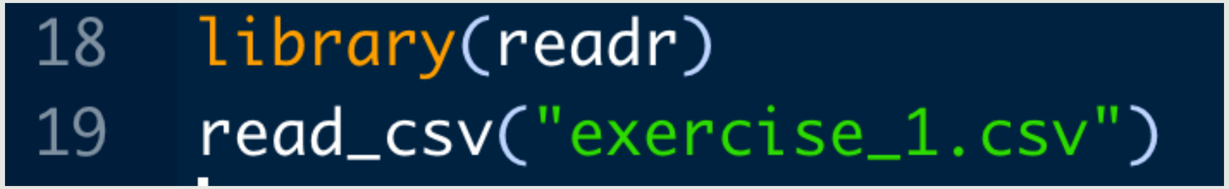
# Importing data

## CSV file is probably the best



```r
read.csv("exercise_1.csv")
```

Default package for importing csv file

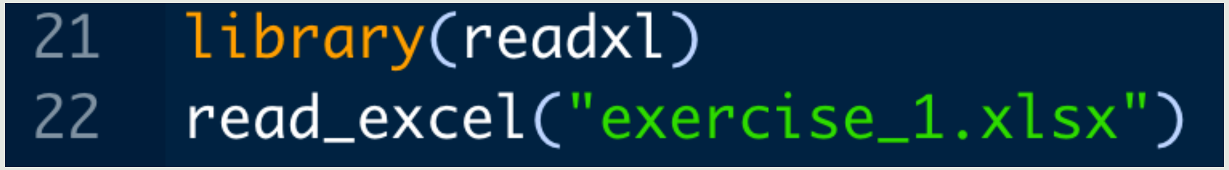


```r
library(readr)
read_csv("exercise_1.csv")
```

A function to read csv file

- Require {readr} package

## Importing excel file is still possible but not common..



```r
library(readxl)
read_excel("exercise_1.xlsx")
```

A function to read excel file

- Require {readxl} package

# Exercise 2

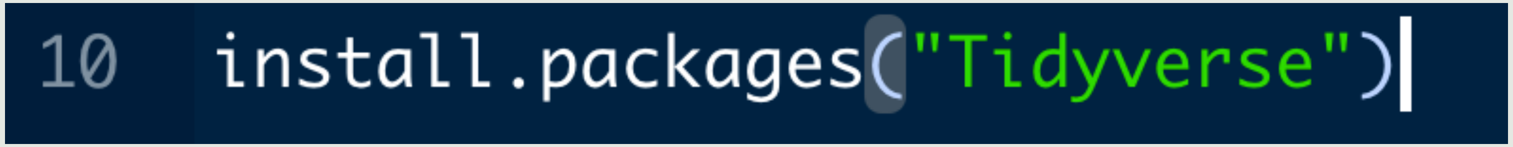
1. Create a working directory and a new R script.
2. Import the file you tidied in the exercise 1 to R.

# Data export

| File type | Package | Import function | Export function | |
|---|---|---|---|---|
| CSV | Default | read.csv() | write.csv() | |
| CSV | readr | read_csv() | write_csv() | Part of {tidyverse} |
| Excel | readxl | read_excel() | - | Part of {tidyverse}<br>Import only |
| Excel | writexl | - | write_excel() | Export only |

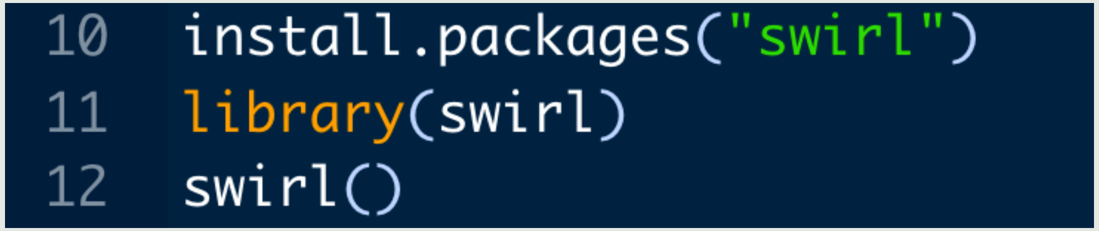
# Tidyverse

Let’s install Tidyverse packages for next week session.

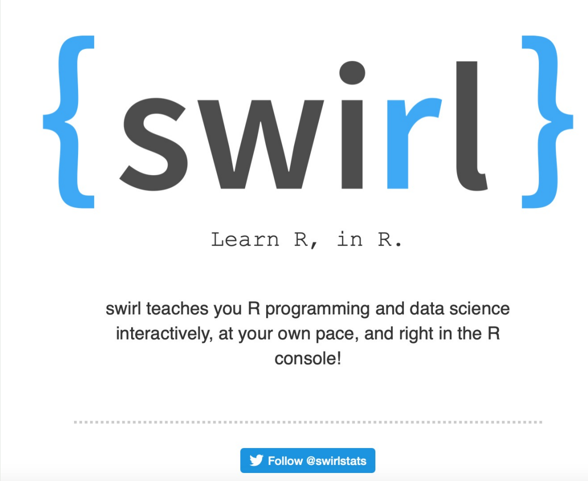


```r
install.packages("Tidyverse")
```

# R coding practices using {swirl}

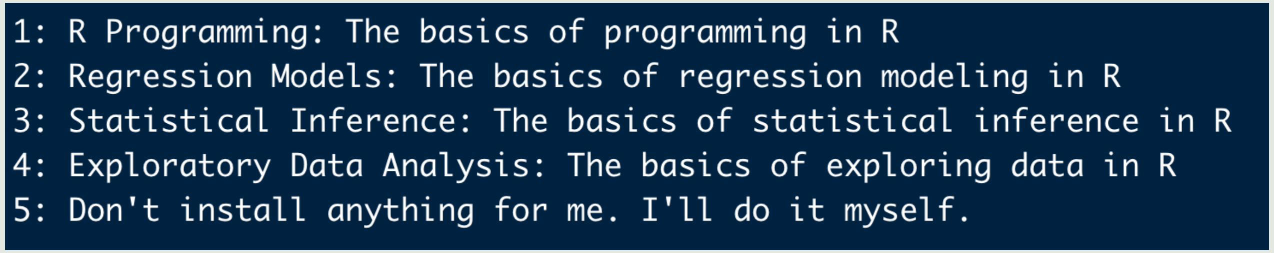


```r
install.packages("swirl")
library(swirl)
swirl()
```



Learn R, in R.

swirl teaches you R programming and data science interactively, at your own pace, and right in the R console!



```
1: R Programming: The basics of programming in R
2: Regression Models: The basics of regression modeling in R
3: Statistical Inference: The basics of statistical inference in R
4: Exploratory Data Analysis: The basics of exploring data in R
5: Don't install anything for me. I'll do it myself.
```

**To familiarize R, {swirl} is a great place to start!**

Please install {swirl} and learn the basic R coding as well as data analysis.

Knowing basic codes becomes handy for the following sessions.

**See you next week!**

https://swirlstats.com

# Thanks to

**Functional Programming** by Sara Altman, Bill Behrman and Hadley Wickham

https://github.com/dcl-docs/prog

**Introduction to Data Handling @TokyoR91** by Masatoshi Katabuchi

https://mattocci27.github.io/assets/TokyoR91/data_handling.html#1

**BeginnerR Special データの読み書き@TokyoR91** by Osamu Machida

https://docs.google.com/presentation/d/1XQk_Gz9Jo660jADxQ78deas5LeRXejqn2z1yHIv2gLQ/
edit#slide=id.gc7dee91765_1_10

**Data Carpentry R basics** by Tobin Magle

https://datacarpentry.org/R-ecology-lesson/