



Data handling

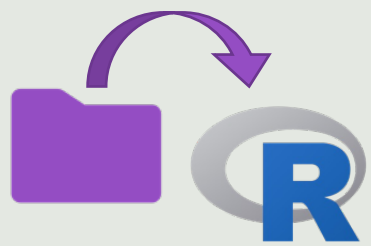
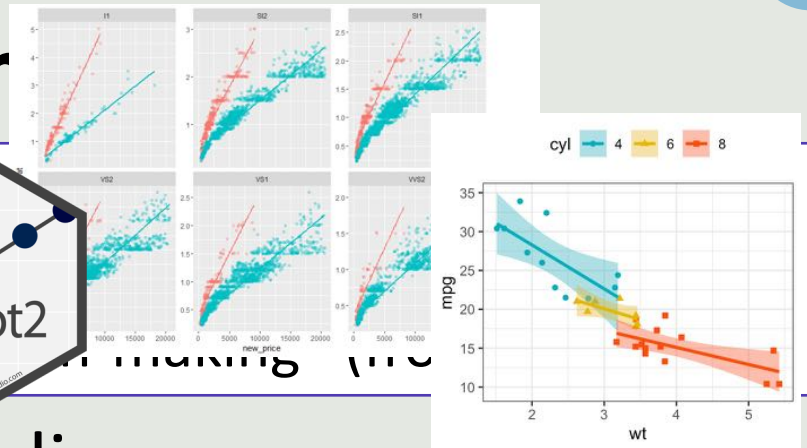
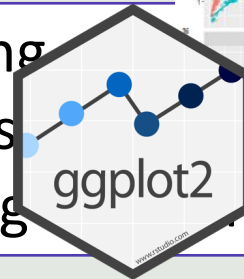
Kanan Saikai

NJC stat seminar series Part 1

July 16, 2021

What is data analysis

“Data analysis is a process of inspecting and modeling data with the goal of discovering patterns, informing conclusions, and supporting decision making (the



Import

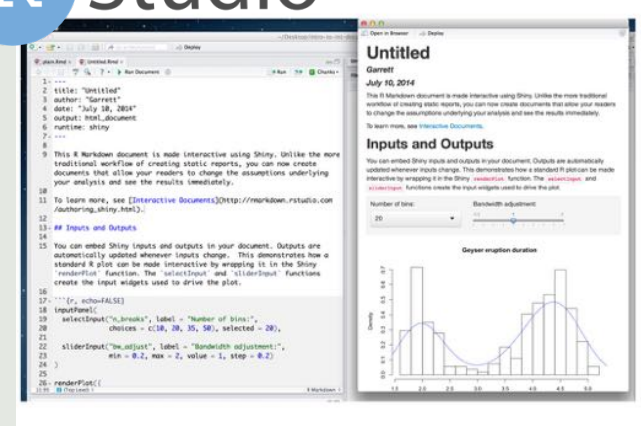
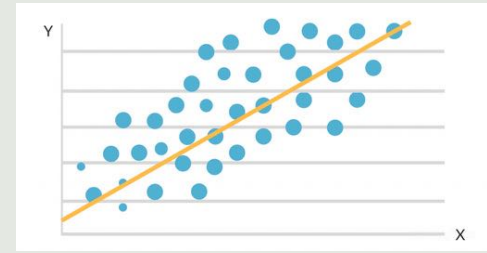
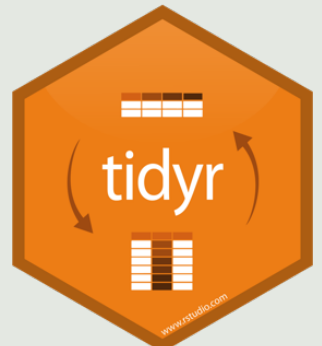
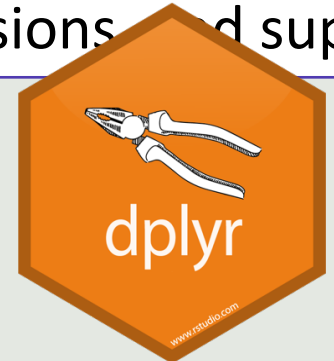
Tidy

Transform

Visualize

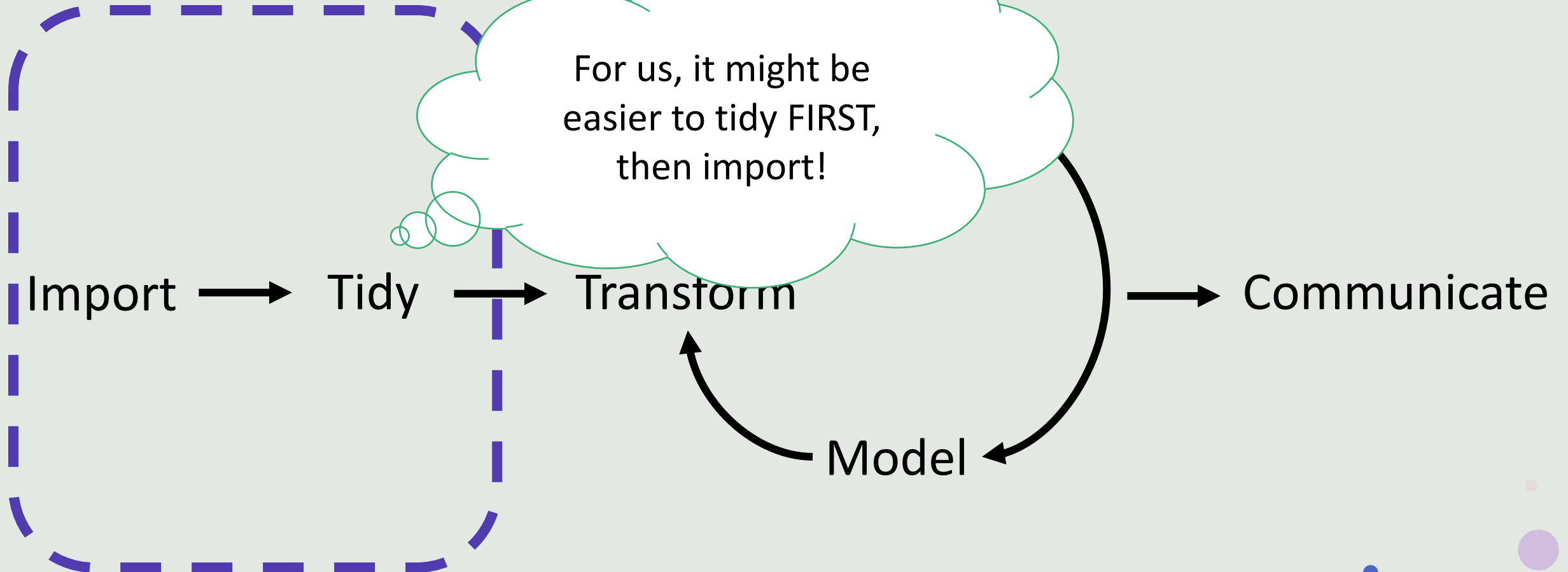
Communicate

Model



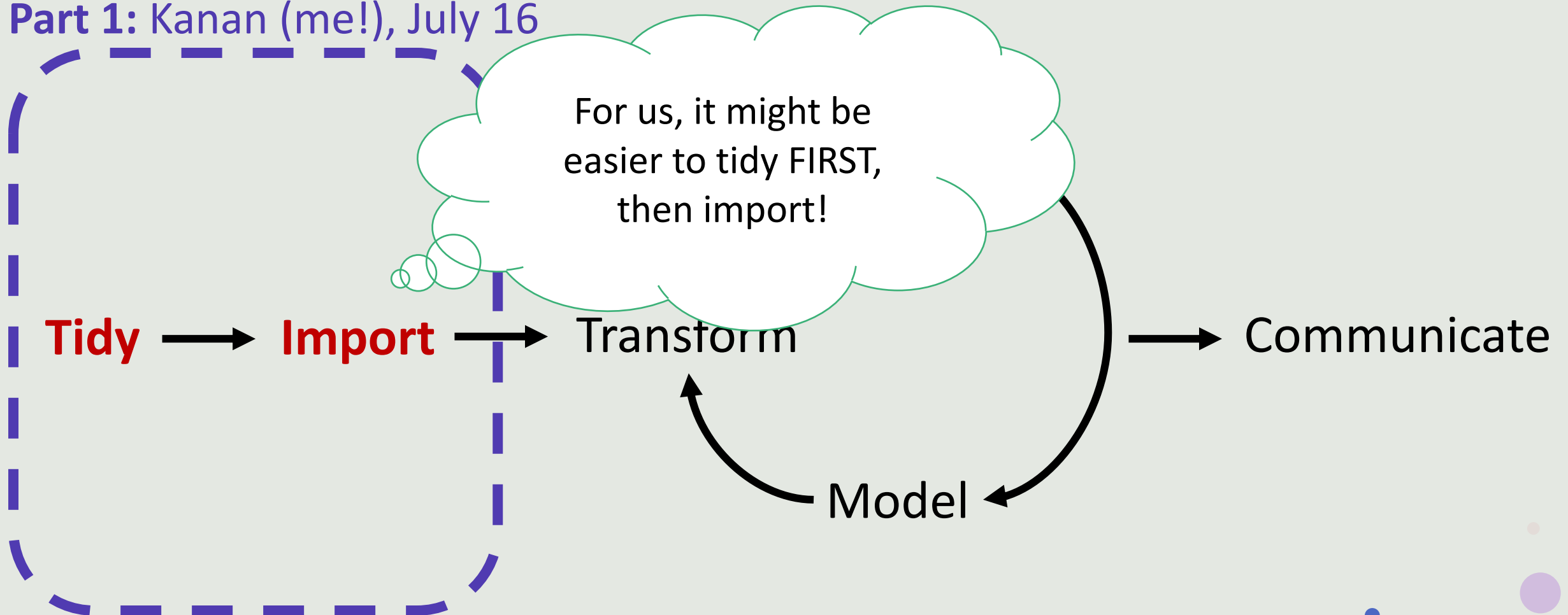
What is data analysis?

Part 1: Kanan (me!), July 16



What is data analysis?

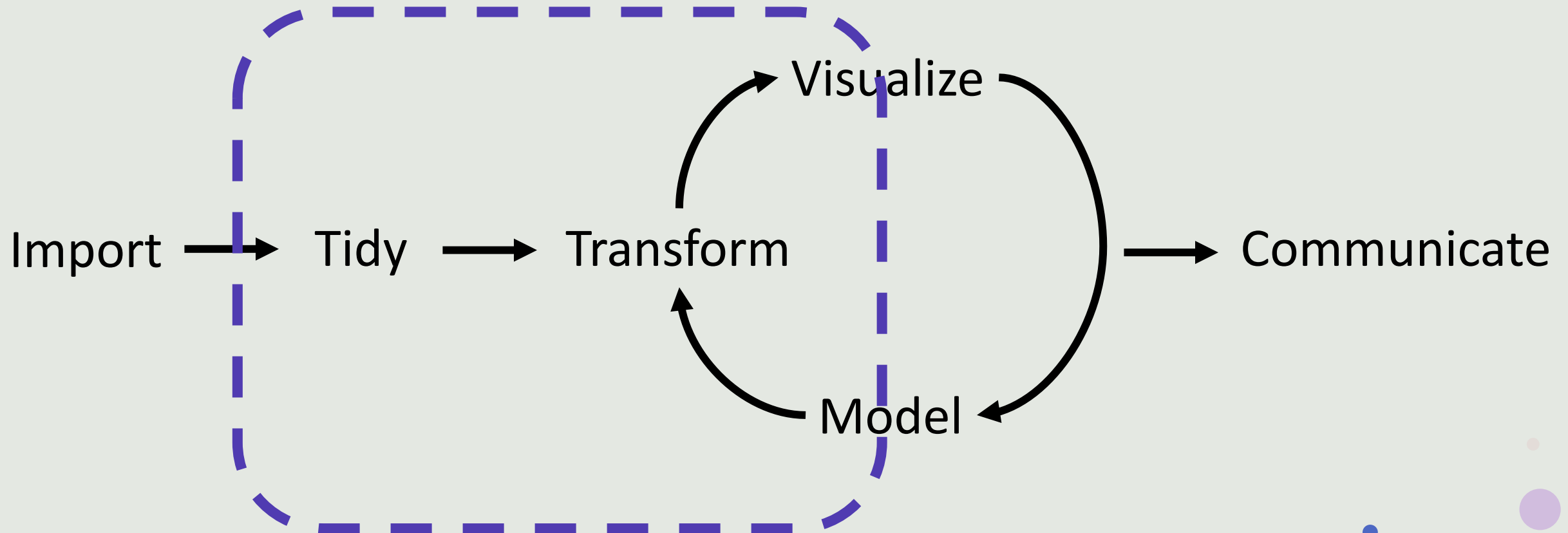
Part 1: Kanan (me!), July 16



What is data analysis?

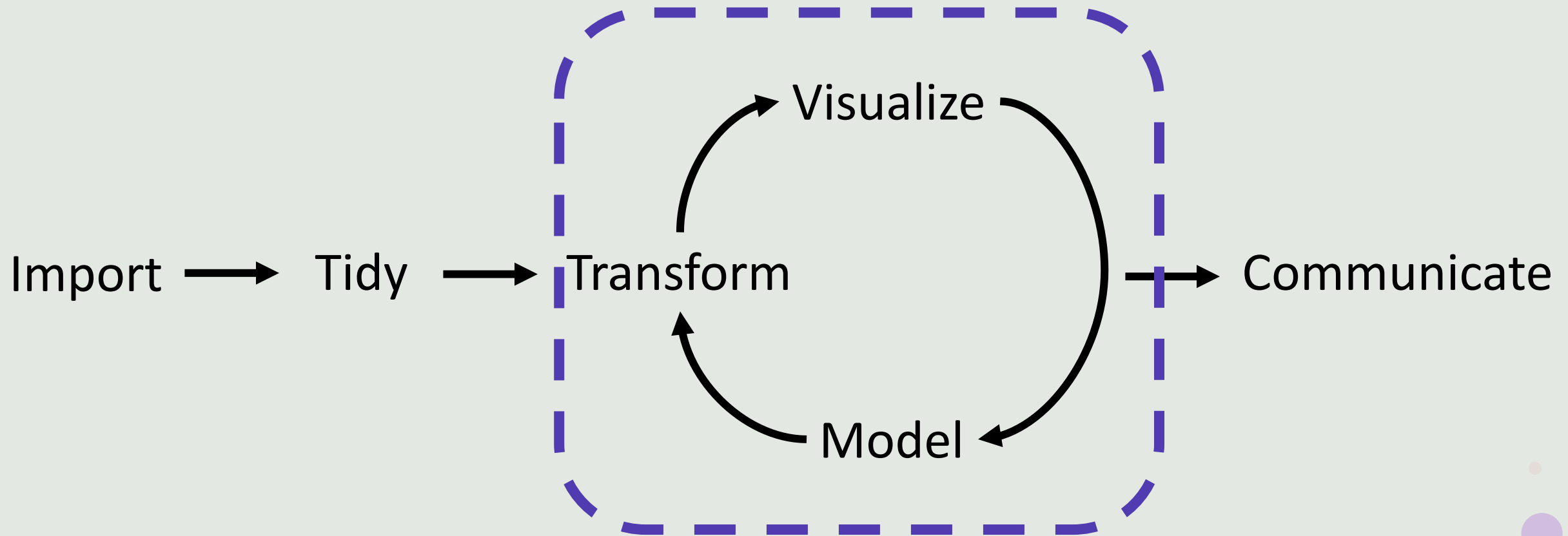
Part 2: Dr. Masatoshi Katabuchi, July 23

Plant Ecologist @ Xishuangbanna Tropical Botanical Garden

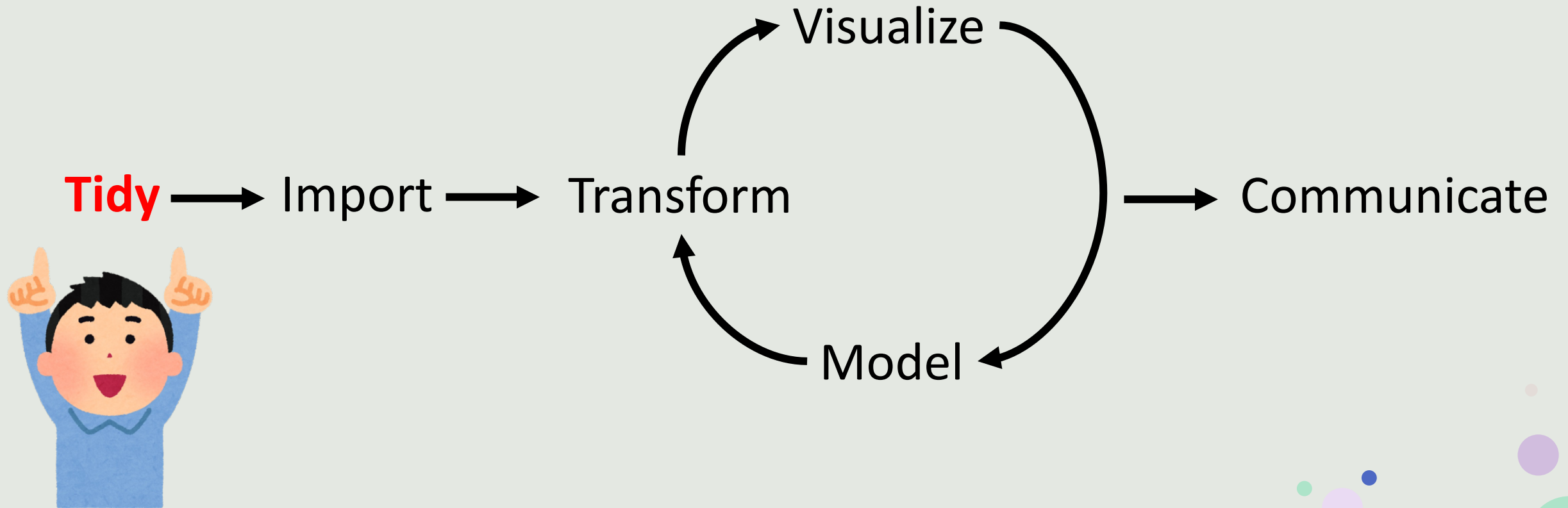


What is data analysis?

Part 3: Dr. Hyunseung Kang, July 30
Statistician @ University of Wisconsin-Madison



Let's tidy!



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

| country | year | cases | population |
|---------|------|-------|------------|
| Algeria | 1999 | 15 | 1809071 |
| Algeria | 2000 | 166 | 2035360 |
| Brazil | 1999 | 31737 | 17206362 |
| Brazil | 2000 | 81688 | 17404898 |
| China | 1999 | 21258 | 127215272 |
| China | 2000 | 21706 | 12902583 |

variables

| country | year | cases | population |
|---------|------|-------|------------|
| Algeria | 1999 | 15 | 1809071 |
| Algeria | 2000 | 166 | 2035360 |
| Brazil | 1999 | 31737 | 17206362 |
| Brazil | 2000 | 81688 | 17404898 |
| China | 1999 | 21258 | 127215272 |
| China | 2000 | 21706 | 12902583 |

observations

| country | year | cases | population |
|---------|------|-------|------------|
| Algeria | 1999 | 15 | 1809071 |
| Algeria | 2000 | 166 | 2035360 |
| Brazil | 1999 | 31737 | 17206362 |
| Brazil | 2000 | 81688 | 17404898 |
| China | 1999 | 21258 | 127215272 |
| China | 2000 | 21706 | 12902583 |

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 |

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



Introduction of R & R Studio



A programming language + software that interprets it

A screenshot of the R Console window. It shows the R version 4.1.0 (2021-05-18) and the platform x86_64-apple-darwin17.0 (64-bit). The console displays the R startup message, including the license and the fact that R is free software. It also shows warnings about missing namespaces (ggplot2, scales, plotly, complot) and the restoration of the workspace and history. The console is titled 'R Console' and has a search bar at the top.

```
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.

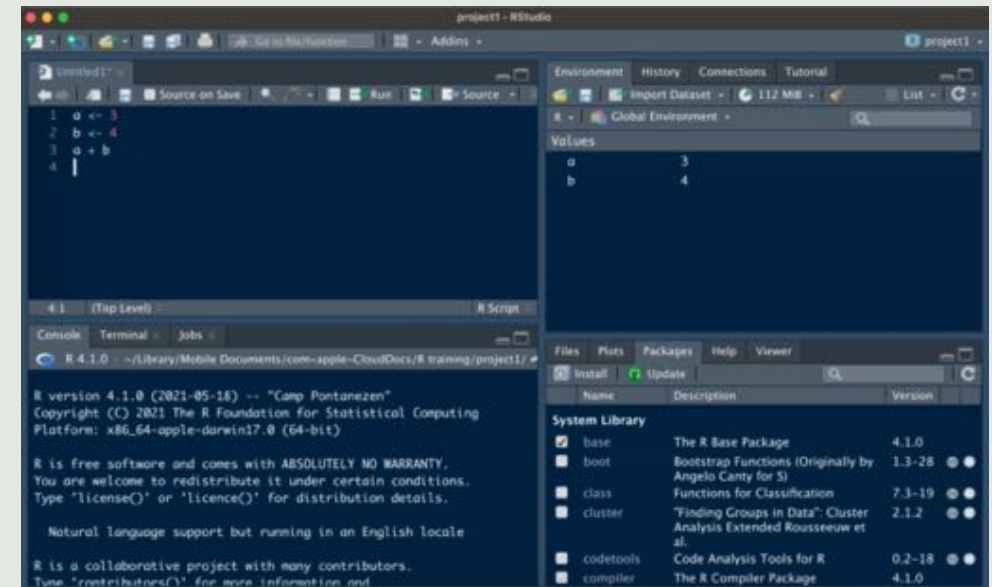
[R.app GUI 1.76 (7976) x86_64-apple-darwin17.0]

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 'complot' 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



<https://www.r->

R Studio®

The screenshot shows the R Studio interface with the following components and annotations:

- Script Editor:** Contains R code:

```
1 a <- 3
2 b <- 4
3 a + b
4 |
```

 A white box labeled "Script" is placed over this area.
- Environment:** Shows the "Global Environment" with variables `a` (value 3) and `b` (value 4). A white box labeled "Environment" is placed over this area.
- Console:** Displays the R version (4.1.0) and startup messages. A white box labeled "Console" is placed over this area.
- Files:** Shows a list of installed packages in the "System Library". A white box labeled "Files" is placed over this area.
- Action Annotations:** A large red arrow points from the "Script" area down to the "Console". Above the arrow, the word "Write" is in red, and below it, "Ctrl + Enter" is in red. To the right of the arrow, the word "Run" is in red.

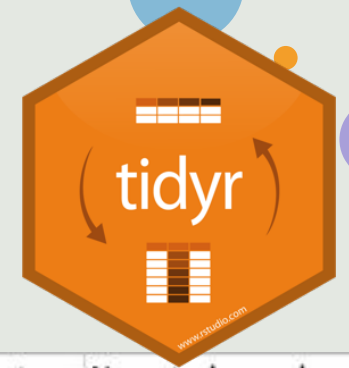
| Name | Description | Version |
|--|---------------------------|---------|
| <input checked="" type="checkbox"/> base | The R Base Package | 4.1.0 |
| <input type="checkbox"/> boot | Boots | 1.3-28 |
| <input type="checkbox"/> class | Functional | 7.3-19 |
| <input type="checkbox"/> cluster | "Find" Cluster | 2.1.2 |
| <input type="checkbox"/> codetools | Code Analysis Tools for R | 0.2-18 |
| <input type="checkbox"/> compiler | The R Compiler Package | 4.1.0 |

Tidyverse



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

Key

Value

| 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 |

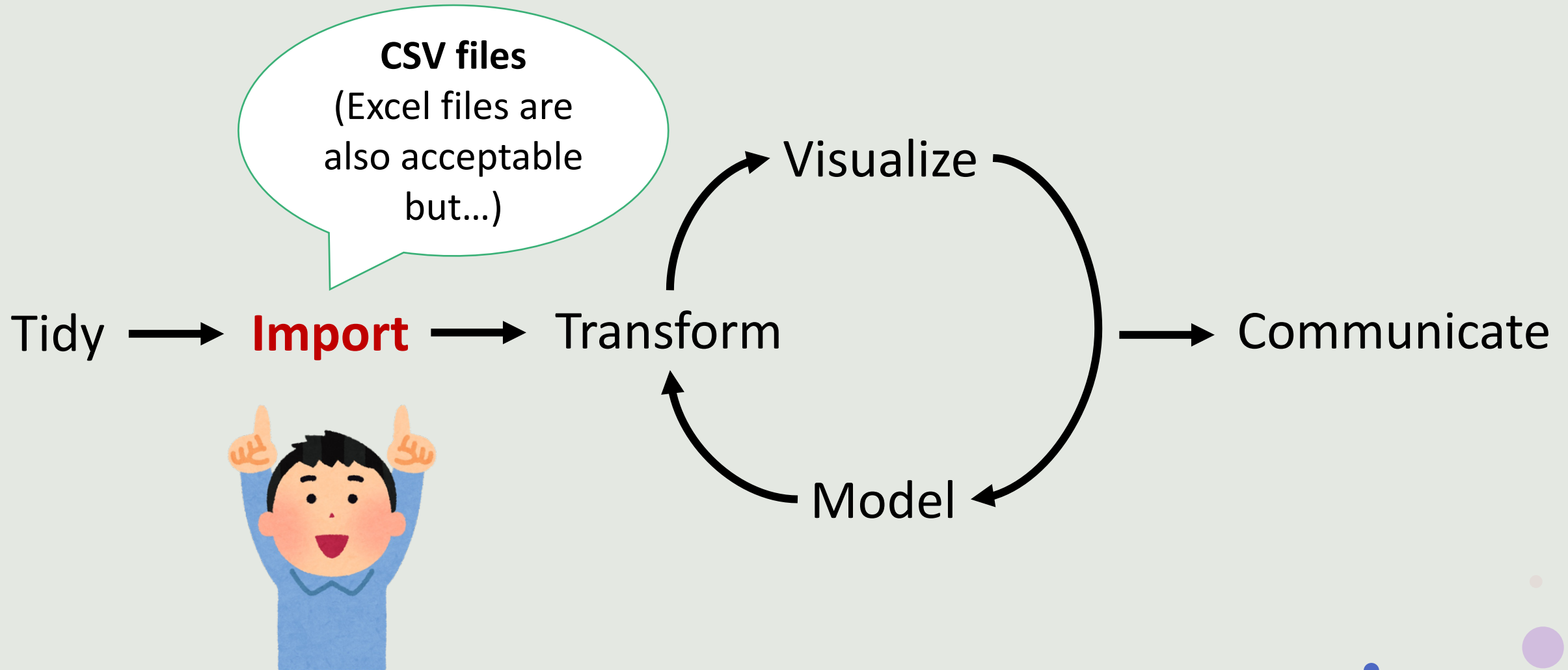
Demonstration



Tidying data using {tidyr}

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

Let's import your data to R



But before you import... set up a working directory

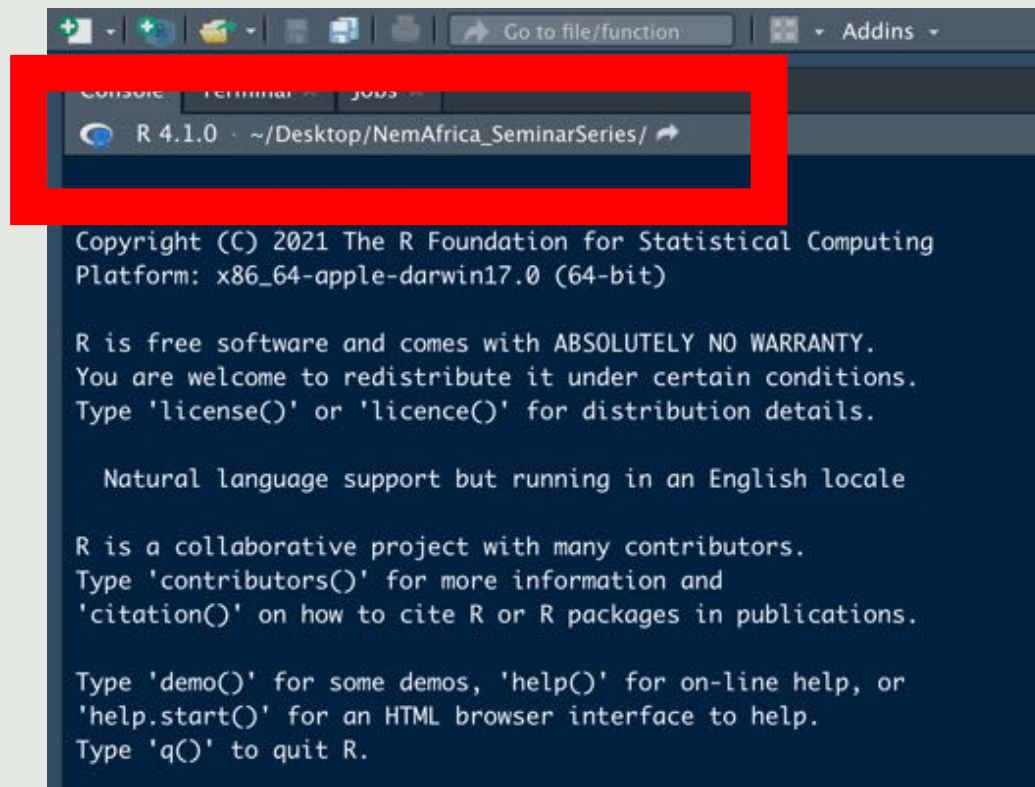
- Open Rstudio
- **File > New project > New directory > Empty project**
- Enter a name for this new folder
- Choose a convenient location
- Click “Create project”

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

Set working directory: **setwd()**

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



Importing data

CSV file is probably the best

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

Default package for importing csv file

```
18 library(readr)
```

A function to read csv file

```
19 read_csv("exercise_1.csv")
```

- Require {readr} package

Importing excel file is still possible but not common..

```
21 library(readxl)
```

A function to read excel file

```
22 read_excel("exercise_1.xlsx")
```

- 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} Import only |
| Excel | writexl | - | write_excel() | Export only |

Tidyverse

Let's install Tidyverse packages for next week session.

```
10 install.packages("Tidyverse")|
```



R coding practices using {swirl}

```
10 install.packages("swirl")
11 library(swirl)
12 swirl()
```



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.

- 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.

See you next week!

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_Gz9Jo660jADxQ78deas5LeRXejqn2z1yHlv2gLQ/edit#slide=id.gc7dee91765_1_10

Data Carpentry R basics by Tobin Magle

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

