Riskified R training

March 2019

Materials by Adi Sarid https://adisarid.github.io and http://www.sarid-ins.com

Sarid
Research Services

2 Instructor – Adi Sarid

Professional: Market Research, Data Scientist, Operations Research, Educator

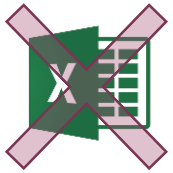
Academia: Mathematics, Statistics and Operations Research (Bsc, MA, Phd-in-process)

Software: R, Python

Sarid
Research Services

Course goals

- Novice:
  - I’m not afraid to use R.
  - When I have a problem with data, I will be comfortable using R to solve it.



- Intermediate:
  - Formal knowledge
  - Strengthen the basics (functions, iterations)
  - Get everyone on the same page of state-of-the-art
    - dplyr, tidyr, ggplot2, purrr, etc.

What will we learn?

- Introduction
  - The data science process
  - Rstudio IDE, Base Syntax
- Visualization (telling stories with charts)
  - ggplot2 – theory and practice
- Intrdoduction to tidyverse
- Solving business problems
  - Modelling, optimization, classification/regression ROC
- Iterations purrr-ing functions
  - Functional programming with and iterations with map)
- Additional topics – as time permits

How will we learn?

- Github repo. To download (clone) it, use:

```
git clone https://github.com/adisarid/Riskified_training Riskified_training
```

- To pull updates use (inside the directory): `git pull`
- (Consider forking your own copy)
- Sticky notes
- Please make sure you have:
  - Latest R (3.5.3) – https://www.r-project.org/
  - Rstudio IDE (https://www.rstudio.com/products/rstudio/download/)
  - git (https://git-scm.com/)
  - Enthusiasm and curiousity! (it’s going to be fun)

Additional sources

- R for Data Science by Hadley Wickham & Garrett Grolemund: https://r4ds.had.co.nz/
- Advanced R by Hadley Wickham: https://adv-r.hadley.nz/
- RStudio cheatsheets (dead tree copies + link)
- Sign-up to R-Bloggers mailing list
- We will use a lot of data sets from kaggle

7 The foundations

(Of a data science project)

Materials by Adi Sarid https://adisarid.github.io and http://www.sarid-ins.com

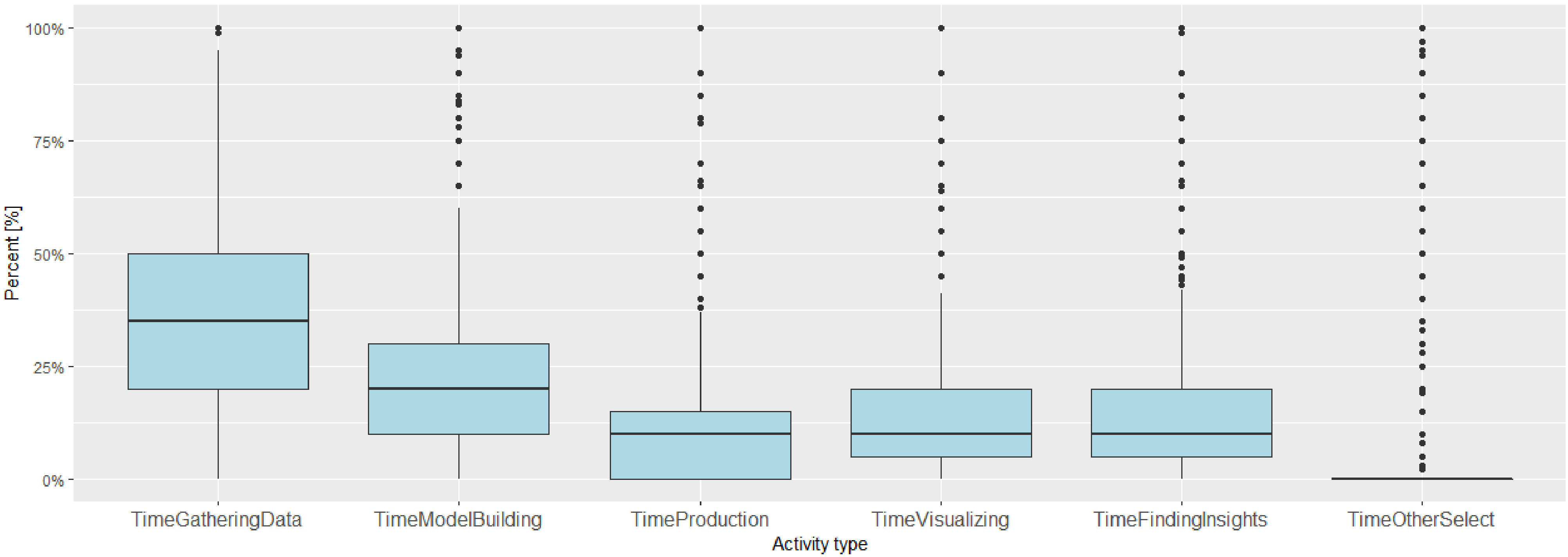
Sarid Research Services

A “quiz”, in pairs:

- Rank the following activities starting from the one which takes up most of your time to the one which takes up the least:
  - Gathering and preparing data
  - Visualizing
  - Finding insights
  - Building models
  - Putting things into production
  - Other activities

Here's what 45,000+ kaggle members thought (and what is a "boxplot")

At work, on average, what percentage of your time is devoted to...?



Data from Kaggle's 2017 members' survey

Materials by Adi Sarid https://adisarid.github.io and http://www.sarid-ins.com

Arrange this into a workflow model:

Communicate

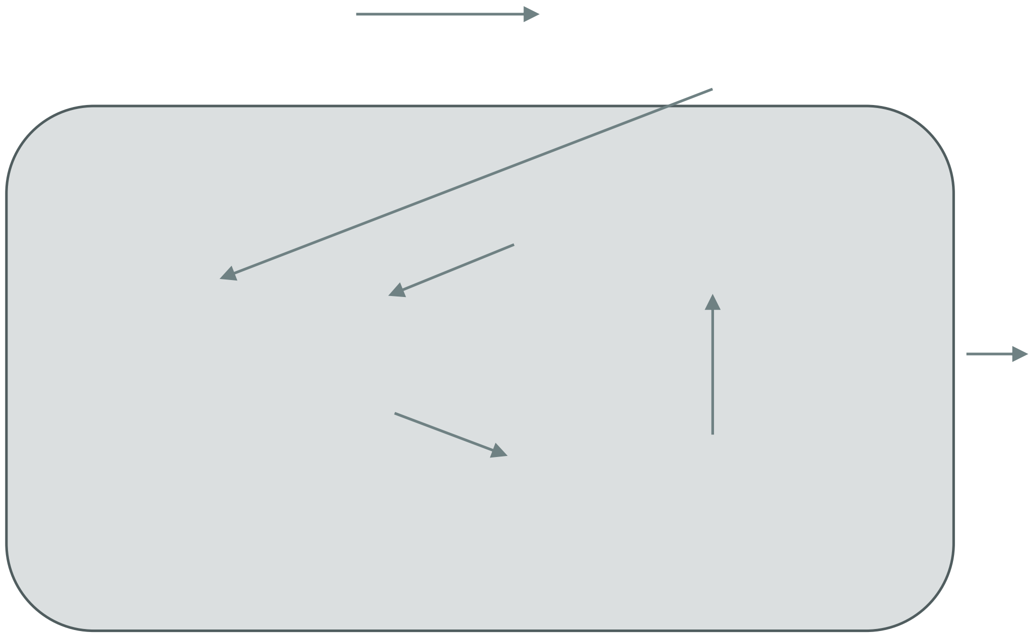
Import

Model

Tidy

Transform

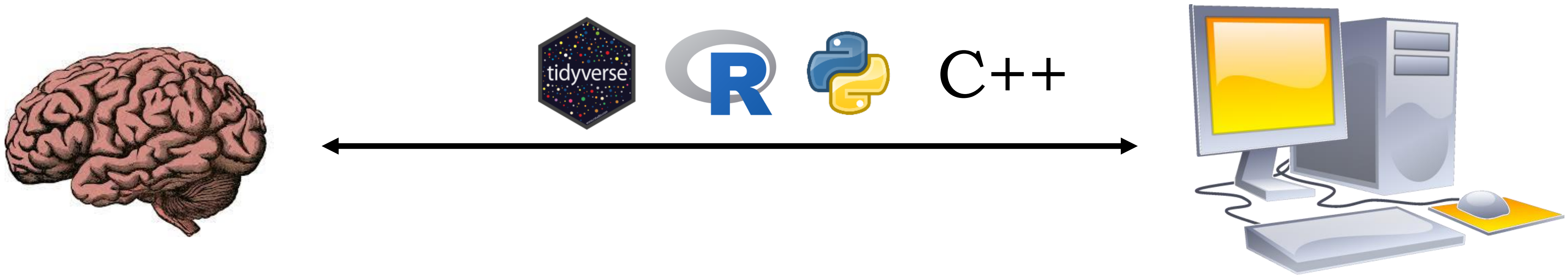
Visualize



* R for Data Science, chapter 1

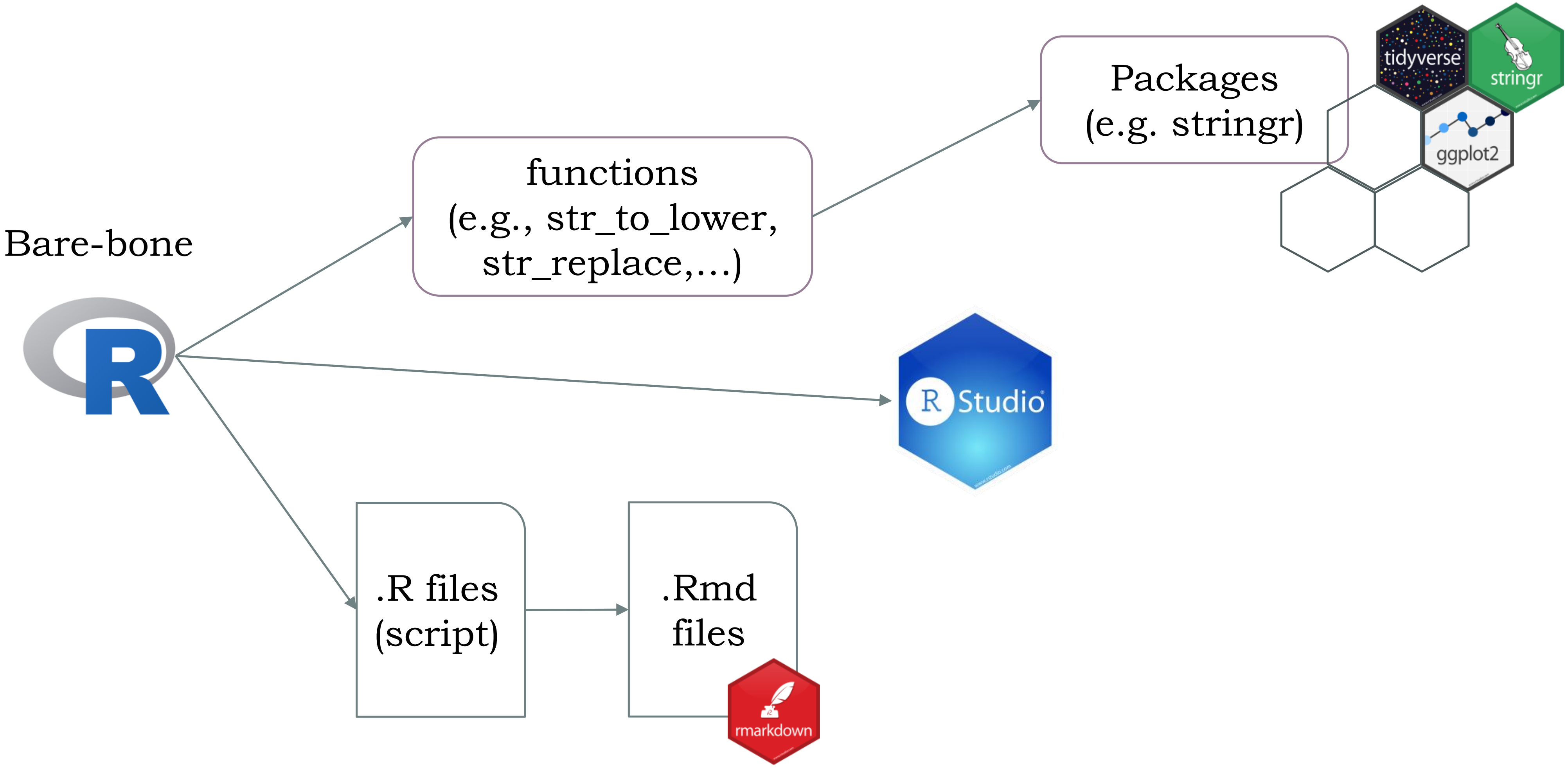
What is R?

- A free software environment for statistical computing and graphics (r-project.org)
- An analogy I adopted (from Garret Grolemund)



C++

Some terms



Bare-bone

functions (e.g., str_to_lower, str_replace,…)

Packages (e.g. stringr)

.R files (script)

.Rmd files

RStudio

Exercise

- Clone the repository (if you know what fork means, do it)
- Open up 00-Introduction.Rmd from the exercises folder.
- Follow the instructions.

- Novice – 30 minutes
- Intermediate – 15 minutes