Data Scientist's Toolbox

Course #1 from Data Science: Foundations using R

Shiyin Tan, 2021.08.31

Week 1: Data Science Fundamentals
- 1.1 Why Automated Videos?
- 1.2 What is Data Science?
- 1.3 What is Data?
- 1.4 Getting Help
- 1.5 The Data Science Process
  - ! 一些有趣的例子 !

Week 2: R and RStudio
- 2.1 Installing R
- 2.2 Installing RStudio
- 2.3 RStudio Tour
- 2.4 R Packages
  - 2.4.1 Where to find packages? Repository.
  - 2.4.2 How to find packages?
  - 2.4.3 How to install packages?
  - 2.4.4 Other operations
- 2.5 Projects in R
  - 2.5.1 Create a new project
  - 2.5.2 Open and Close a project
  - 2.5.3 Setup folders in projects

Week 3: Version Control and GitHub
- 3.1 Version Control
- 3.2 GitHub and Git
  - 3.2.1 GitHub
  - 3.2.2 Git
- 3.3 Linking GitHub and RStudio
  - Step 1: Create & View SSH RSA Key in RStudio
  - Step 2: Add SSH key to GitHub
  - Step 3: Create a new repository in GitHub
  - Step 4: Create your personal access token in GitHub
  - Step 5: Link GitHub repository to RStudio
- 3.4 Projects Under Version Control

Week 4: R Markdown, Scientific Thinking, and Big Data

Week 1: Data Science Fundamentals

1.1 Why Automated Videos?

- Elements for online open course

    - Tutorials
    - Slides
    - Assessments (markup language)
    - Videos
- R packages for videos: ari + didactr
    - Ari: script + slides, narrates using Amazon Polly (voice synthesis)
    - Didactr: automates steps

1.2 What is Data Science?

- Definition: using data to answer (novel) questions
- **Qualities of DS: 3V**
    - **Volume**: More data is becoming increasingly available
    - **Velocity**: Data is being generated at an astonishing rate
    - **Variety**: The data we can analyze comes in many forms
- **DS = Intersection of 3 fields**
    - **Substantive Expertise** (identify questions and data)
    - **Hacking Skills** (programming)
    - **Math & Statistics Knowledge**
- Data Scientist
    - Daryl Morey: general manager of a US basketball team, the Houston Rockets
    - Hilary Mason: FastForward labs, mining the web and understanding the way that humans interact with each other through social media
    - Nate Silver: **FiveThirtyEight**, uses statistical analysis - hard numbers - to tell compelling stories about elections, politics, sports, science, economics and lifestyle
- Example: 2009, Google analyzed commonly searched terms that had strong correlation with the CDC flu outbreaks.

1.3 What is Data?

- Definition of Data:
    - Cambridge Dictionary: Information, especially facts or numbers, collected to be examined and considered and used to help decision-making.
    - Wiki: **A set of values of qualitative定性的 or quantitative定量的 variables**
- Examples of Data Sources
    - sequencing data, population census, electronic medical records, images/videos
    - lack of tidy data sets
- Ask The Right Questions First

1.4 Getting Help

- Steps for getting help in this course
    - manuals / help files / FAQs
    - google
    - course forum, search the archives first

- Coding problems
    1. red error message: error messages, help(), forum
    2. unwanted outputs: debug, ask peers, **rubber duck debugging**
- Ask the right question
    - Forums: stackoverflow, cross validated
    - Provied detailed info:
      steps to reproduce problems, expected output, actual output, error message, version of products (R, packages, OS, etc.)
    - the more specific the question, the faster the answer
    - be ourteous
- Other resources
    - How To Ask Questions The Smart Way
    - Roger Peng's video on "Getting Help"

1.5 The Data Science Process

- DS process
    - Form Question
    - Get Data
    - Analyze Data: exploring, modeling
    - Draw Conclusion
    - Show Results

! 一些有趣的例子 !

- **Hilary: the most poisoned baby name in US history**
  By 数据科学家**Hilary Parker**
- Predicting Spatial Risk of Opioid Overdoses in Providence, RI
- **Text analysis of Trump's tweets confirms he writes only the (angrier) Android half**
- Where to Live in the US
- Sexual Health Clinics in Toronto

Week 2: R and RStudio

2.1 Installing R

- CRAN = Comprehensive R Archive Network
- Why use R?
    - Popularity
    - Free
    - Extensive functionality
    - Community: Stackoverflow, cross validated

2.2 Installing RStudio

- RStudio is a graphical user interface for R

2.3 RStudio Tour

略

2.4 R Packages

- R Packages
  - Package = a collection of functions, data, and code conveniently provided in a nice complete format
  - now 14,300+ packages available
  - Packages $\in$ Library

2.4.1 Where to find packages? Repository.

- Repository = a central location where many developed packages are located and available for download
- three big repositories:
  1. **CRAN (Comprehensive R Archive Network):** R's main repository (>12,100 packages available!)
  2. **BioConductor:** A repository mainly for bioinformatic-focused packages
  3. **GitHub:** A very popular, open source repository (not R specific!)

2.4.2 How to find packages?

- CRAN Task Views
- RDocumentation
- Google: task + R package

2.4.3 How to install packages?

```
'''
CRAN Repository
'''
install.packages("ggplot2") # both single and double quotes are OK
install.packages('ggplot2')
install.packages(c("ggplot2", "devtools", "lme4")) # install multiple packages
# or Tools menu -> Install Packages...

'''
Bioconductor
'''
# from coursera
source("http://bioconductor.org/biocLite.R")
boicLite("GenomicFeatures")
```

```
# from https://bioconductor.org/install/
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install(version = "3.13")
BiocManager::install(c("GenomicFeatures", "AnnotationDbi"))

'''
GitHub
'''
# take note of both the package name and the author of the package on GitHub
install.packages("devtools")
library(devtools)
install_github("author/package")
```

2.4.4 Other operations

```
'''Load packages'''
library(ggplot2) # Do not put the package name in quotes!
# or "Packages" tag in RStudio


'''Unload packages'''
detacj("package:ggplot2", unload=TRUE)
# or "Packages" tag in RStudio


'''check installed packages'''
installed.packages()
library()
# or "Packages" tag in RStudio


'''update packages'''
old.packages() # check outdated packages
update.packages() # update all outdated packages
install.packages("ggplot2") # update specific package


'''uninstall packages'''
remove.package("ggplot2")
# or "Packages" tag in RStudio


'''check R version'''
# first open R/Rstudio, pay attention to the console
version
sessionInfo() # great to put on forum when posting questions
```

```
'''learn about functions in packages'''
help()
help(package = "ggplot2")
# or "Packages" tag in RStudio
browseVignettes()
browseVignettes("ggplot2")
```

2.5 Projects in R

- Project in R = Creates a folder with saved environment
- Benefits of R projects
    - Easy organization
    - Easy sharing
    - Easy to start back up on a project

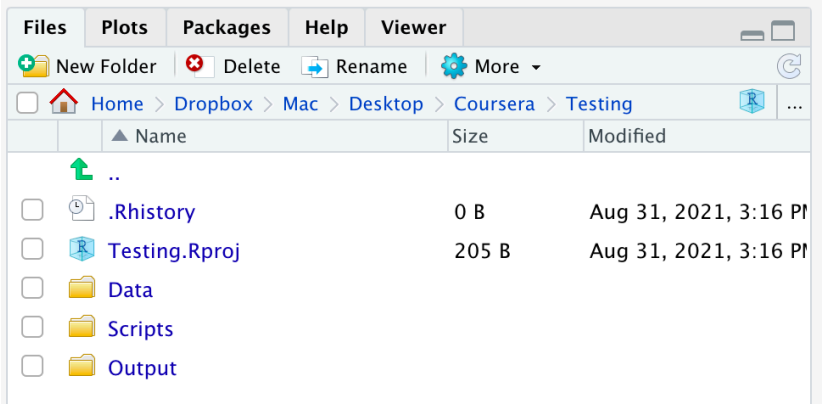
2.5.1 Create a new project

- **File** > **New Project...** > **New Directory** > **New Project** > enter Directory name and choose a location > **Create Project**

2.5.2 Open and Close a project

- Open * 3
    1. Click the **.Rproj** file
    2. **File** > **Open Project...**
    3. The drop-down list in up-right corner > **Open Project...**
- Close * 3
    1. **Exit** RStudio
    2. **File** > **Close Project**
    3. The drop-down list in up-right corner > **Close Project**
- Switch projects / Multiple projects open at once
    - The drop-down list in up-right corner > **Open Project in New Session...**

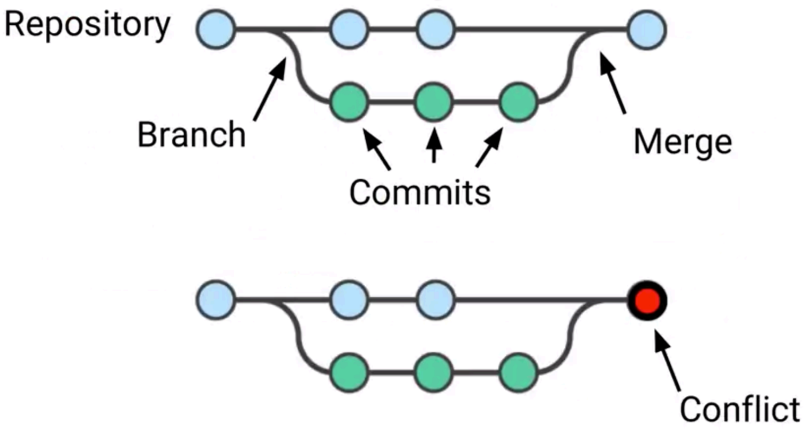
2.5.3 Setup folders in projects



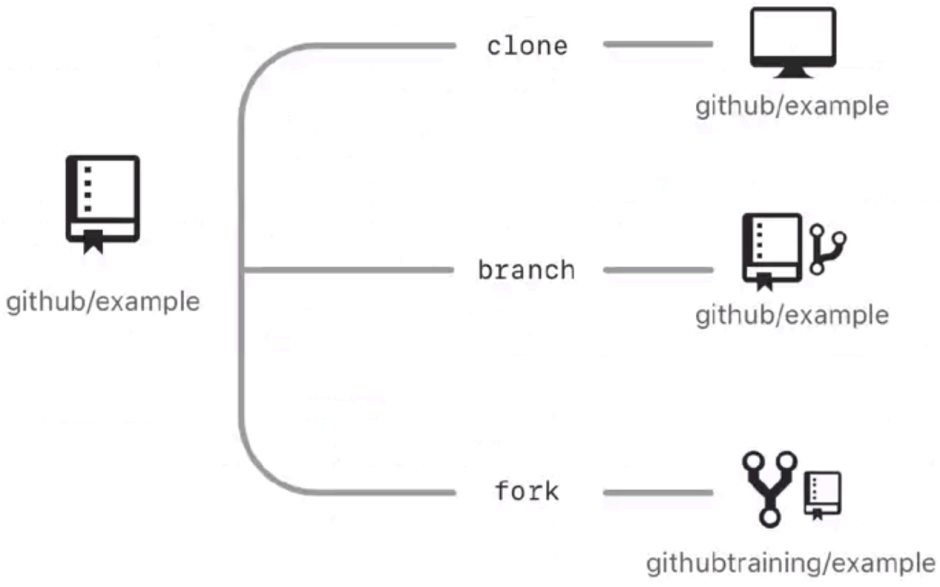
Week 3: Version Control and GitHub

3.1 Version Control

- GitHub = Online Git, Git = local version control software
    - Repository(Repo): project folder (private / public)
    - Commit: snapshot of files, changes of files and reasons for change
    - Push: update repository online
    - Pull: download repository and update local version
    - Staging: the act of preparing a file for commit



    - Branch: two simultaneous copies of a same file
    - Merge: incorporate independent edits into a single file
    - Conflict: cannot merge the edits automatically, have to merge them manually
    - Clone: copy an existing repository
    - Fork: make a personal copy of a repository from others



- Best Practice
    - Purposeful, single issue commits
    - Informative commit messages
    - Pull and puch often

- Comics
  - PHD Comics
  - xkcd

3.2 GitHub and Git

3.2.1 GitHub

- Account
  - Email: tanshiyin11@yeah.net
  - Password: 4******_
- More to learn:
  - To learn more about the power of Pull Requests, we recommend reading the GitHub flow Guide. You might also visit GitHub Explore and get involved in an Open Source project.
  - Tip: Check out our other Guides, YouTube Channel and On-Demand Training for more on how to get started with GitHub.

3.2.2 Git

- 官网给出的`brew install git`运行出错，查到了下载安装包的网址

```
$ git config --global user.name "Tan Shiyin"
$ git config --global user.email tanshiyin11@yeah.net
$ git config --list
core.excludesfile=~/.gitignore
core.legacyheaders=false
core.quotepath=false
...
user.name=Tan Shiyin
user.email=tanshiyin11@yeah.net
```

3.3 Linking GitHub and RStudio

Step 1: Create & View SSH RSA Key in RStudio

RStudio:

- **Tools** > **Global Options** > **Git/SVN** (> Browse... and verify Git executable) > **Create RSA Key...**
- **Tools** > **Global Options** > **Git/SVN** > **View public key** > copy your key
  - My Public Key

```
ssh-rsa
AAAAB3NzaC1yc2EAAAADAQABAAABAQCrMSFY7aVdzmYE/2V3IHYB1bIxd1wpKkbAabFn4Skxek/gumU
v9wjW0V3Ww1nlg2PSuAnqWzzLAR4dwrDUugqiMROea5HbViB1M0ZF0k2uUxXITwE5TRmAIvy/cVv7zu
LtXwZe8/AJnYcZdfCIrvSYVXq3sz7bG7FX2Nh+m3GnAzfjgm9zC61wHqRp1vk/a6lStMgDuu0rBUf8j
DWWdibKBZ6oL9tZsTErQFH3P7sWFp/zCv9uysL6yLFXCEb4IZTsk6N34WQv3cfqhSPYtt3HsMgiOdOW
WmmN9jKr/+lj+HqOxYkyIPzIo7Eh0EuuzJEIQUZDMC3ksT6/B1E/roK1
tanshiyin@tanshiyindeMacBook-Pro.local
```

Step 2: Add SSH key to GitHub

GitHub:

- **Settings** > **SSH and GPG keys** > **New SSH key** > paste your key and give a title > **Add SSH key**

Step 3: Create a new repository in GitHub

Step 4: Create your personal access token in GitHub

GitHub:

- **Settings** > **Developer Settings** > **Personal Access Token** > **Generate New Token** > fill up the form > **Generate token** > copy your token (Make sure to copy your personal access token now. You won't be able to see it again!)
    - My token

```
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
```

Step 5: Link GitHub repository to RStudio

GitHub:

- Copy the **URL** for your new repository (e.g. https://github.com/Mariana-Tan/testing-RStudio)

RStudio:

- Create a new R project: **File** > **New Project** > **Version Control** > **Git** > paste **Repository URL**, name your project and choose its directory > **Create Project**
- Create a new R script: **File** > **New File** > **R Script** > coding > save it
- Push R script to GitHub: **Git** in the environment quadrant > click the checkbox under **Stage** for your R script > **Commit** > write commit message > **Commit** > **Push** > enter your **GitHub username** and GitHub password (paste your **personal aceess token** generated in GitHub)

3.4 Projects Under Version Control

Week 4: R Markdown, Scientific Thinking, and Big Data