WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

Data Science and Machine Learning 2187 & 2087: Data Wrangling

Max Thomasberger,

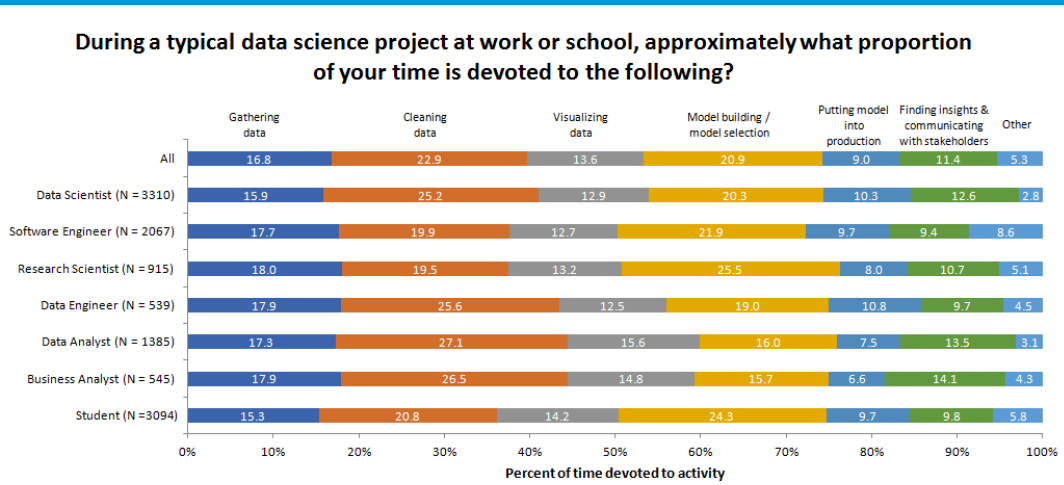
October 13, 2020

Inconvenient Truth No. 22:

You will spend a considerable amount of time wrangling with your data

Figure 1: Listen to Yoda, he's been doing this for ages

How do data scientist spend their time?



Note: Data are from the 2018 Kaggle ML and Data Science Survey. You can learn more about the study here: http://www.kaggle.com/kaggle/kaggle-survey-2018.
A total of 23859 respondents completed the survey; the percentages in the graph are based on a total of 15937 respondents who provided an answer to this question. Only selected job titles are presented.

Gathering Data for social science research I

- Public Databases by offical actors:
  - Eurostat, Fred, Worldbank, Open Data, etc.
  - Files in various formats (txt, excel, csv, stata, spss, raster files, etc.).
  - Data access via an API, mostly json format.
- Public Databases by private actors and NGOs:
  - Facebook movement data, World Pop, etc.
  - Files in various formats (txt, excel, csv, stata, spss, raster files, etc.).
  - Data access via an API, mostly json format.
- Closed Data/Microdata by governmental agencies, statistical bureaus, etc.:
  - Complicated “vetting” procedure.
  - Scientific use files in various formats (txt, excel, csv, stata, spss, etc.).
  - Unstructured data
  - Data “hidden” in databases
  - access sometimes only allowed “on location”.

Gathering Data for social science research II

- Data used/provided by academic publications in various formats (excel, csv, stata, spss, etc.)
  - Harvard Dataverse, Nature Scientific Data, Academic Torrents, etc.
- Private Data providers
  - Cooperation with companies, Platforms like Kaggle, etc.
  - Company data is often "hidden" inside databases.
  - Files in various formats (txt, excel, csv, stata, spss, raster files, etc.).
  - Data access via an API, mostly json format.
- Gathering your own data
  - Experiments
  - Surveys
  - Webscraping, API access, Google Analytics, etc.

Figure 2: A nice image about the pain we are about to face

The data wrangling process starts

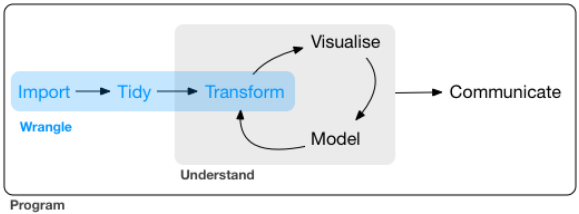


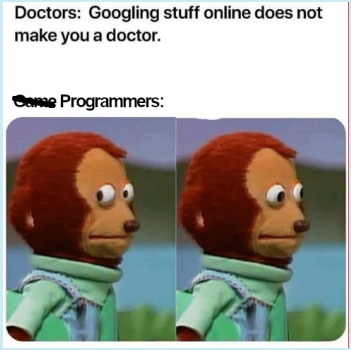
Figure 3: Phases of a typical data science project (R for data science)

- How to read the data in?
- Data is spread across multiple files and sources.
- Variable definitions aren't clean/consistent.
- The data isn't displayed correctly.
- You have to create variables, you have to transform variables, you have to aggregate variables.
- You have to bring the data into the format the packages actually needs.

Sounds pretty awful but in reality:

We all use Google

1. Learning this can be fun!
2. It is easy once you've learned the basics
3. Stack Exchange and Google are your friends (almost every problem already ocurred)



The main ressource for this part of the course is

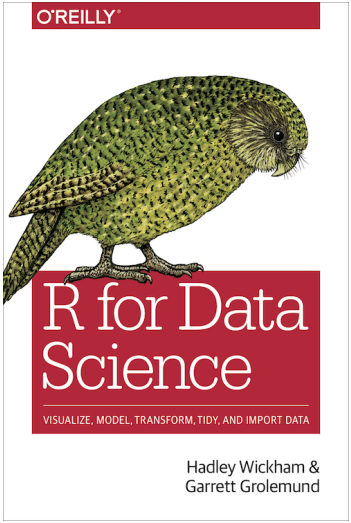


Figure 4: Get it for free at: https://r4ds.had.co.nz/

In this course we are focusing on the tidyverse for doing this stuff

Figure 5: A great collection of R-packages maintained by the guys/gals from R-studio

Tidyverse or not?

"The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures."

www.tidyverse.org

- Succint and readable synthax which is important for reproducible research and collaborations
- Great for small data sets handled inside RAM (a few hundred Megabytes to 1-2 Gb)
- Best plotting package out there (imo)
- Maintained by R-Studio and Hadley Wickham
- Big community and a lot of documentation/information
- Connectors to SQL, data.table, spark and hadoop

For big(ish) data other tools are better suited:

- data.table package
- Databases like PostgreSQL or MySQL
- Clustering solutions like Spark, Hadoop, etc.