Data Handling: Import, Cleaning and Visualisation

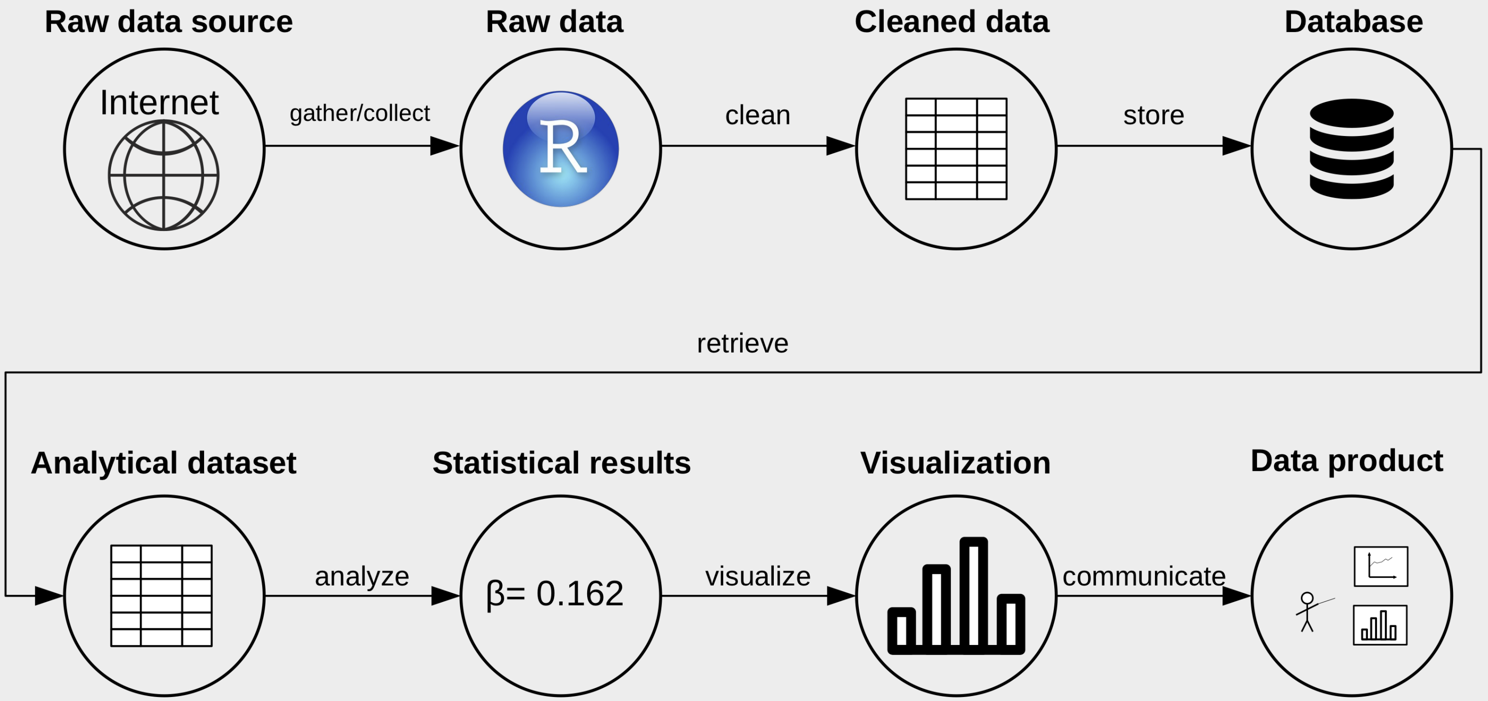
Lecture 9:

Data Preparation

Prof. Dr. Ulrich Matter

Status

Data (science) pipeline



Raw data source

Internet

gather/collect

Raw data

clean

Cleaned data

store

Database

retrieve

Analytical dataset

analyze

Statistical results

β= 0.162

visualize

Visualization

communicate

Data product

Recap: Data Import

Sources/formats in economics

- CSV (typical for rectangular/table-like data)
- Variants of CSV (tab-delimited, fix length etc.)
- XML and JSON (useful for complex/high-dimensional data sets)
- HTML (a markup language to define the structure and layout of webpages)
- Unstructured text

Parsing CSVs

- Recall the introduction to data structures and data types in R
- How does R represent data in RAM?
    - **Structure**: `data.frame`/`tibble`, etc.
    - **Types**: `character`, `numeric`, etc.
- Parsers in `read_csv()` guess the data **types**.

A Template/Blueprint

```r
#######################################################################
# Data Handling Course: Example Script for Data Gathering and Import
#
# Imports data from ...
# Input: links to data sources (data comes in ... format)
# Output: cleaned data as CSV
#
# U. Matter, St.Gallen, 2019
#######################################################################


# SET UP --------------
# load packages
library(tidyverse)

# set fix variables
INPUT_PATH <- "/rawdata"
OUTPUT_FILE <- "/final_data/datafile.csv"
```

Goals for today

Goals for today: cognitive goals

- Recognize where the problems are in a given dataset, and what is in the way of a proper analysis of the data.
- Organize your work: what needs to be addressed first?

Goals for today: skills

- Use simple string-operations to clean text variables.
- Reshape datasets from wide to long (and vice versa).
- Apply row-binding/stacking of datasets

Data Preparation

DEVELOPING EMPLOYEES

Prioritize Which Data Skills Your Company Needs with This 2×2 Matrix

by Chris Littlewood

OCTOBER 18, 2018 **UPDATED** OCTOBER 23, 2018

The dataset is imported, now what?

- In practice: still a long way to go.
- Parsable, but messy data: Inconsistencies, data types, missing observations, wide format.

The dataset is imported, now what?

- In practice: still a long way to go.
- Parsable, but messy data: Inconsistencies, data types, missing observations, wide format.
- **Goal** of data preparation: Dataset is ready for analysis.
- **Key conditions**:
  1. Data values are consistent/clean within each variable.
  2. Variables are of proper data types.
  3. Dataset is in ‘tidy’ (in long format, more on this after the break)!

Move to Nuvolos

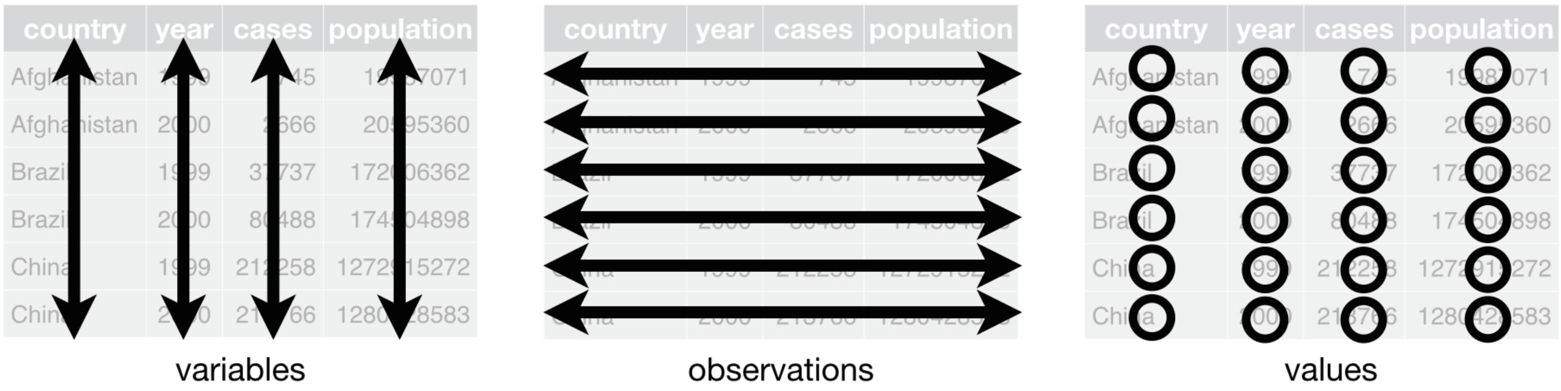
nuvolos

Tidy data: some vocabulary

Following Wickham (2014):

- Dataset: Collection of values (numbers and strings).
- Every value belongs to a variable and an observation.
- Variable: Contains all values that measure the same underlying attribute across units.
- Observation: Contains all values measured on the same unit (e.g., a person).

Tidy data



Tidy data. Source: Wickham and Grolemund (2017), licensed under the Creative Commons Attribution-Share Alike 3.0 United States license.

Reshaping: the concept

Name	sales Jan	sales Feb
Andy	50	54
Claire	60	59

Name	month	sales
Andy	Jan	50
Andy	Feb	54
Claire	Jan	60
Claire	Feb	59

Stack/row-bind: the concept

ID	X	Y
1	a	50
2	b	10

ID	Z
3	M
4	O

ID	X	Z
5	c	P

ID	X	Y	Z
1	a	50	NA
2	b	10	NA
3	NA	NA	M
4	NA	NA	O
5	c	NA	P

Move to Nuvolos

nuvolos

Q&A

References

Wickham, Hadley. 2014. “Tidy Data.” **Journal of Statistical Software** 59 (10): 1–23. https://doi.org/10.18637/jss.v059.i10.

Wickham, Hadley, and Garrett Grolemund. 2017. Sebastopol, CA: O’Reilly. http://r4ds.had.co.nz/.