Data Handling: Import, Cleaning and Visualisation

Lecture 7:

Data Preparation

Dr. Aurélien Sallin

Welcome back!

Updates

- **Exam for exchange students**: 21.12.2023 at 16:15 in room 01-013.
- Materials on text analysis and image analysis (lecture 6) is for self study
- The **mock exam** is online 💣
  - Central exam from last year
  - Solutions won't be discussed in the lecture: use the forum on Canvas
  - No forum supervision/email during the learning phase guaranteed (from me or the TAs)
- Walk-ins for the **digital exam**: see announcement in Canvas

Part II: Data preparation, analysis, and visualization

Date	Topic
16.11.2023	Data preparation and manipulation
23.11.2023	Basic statistics and data analysis with R
23.11.2023	Exercises/Workshop 4: Data gathering, data import
30.11.2023	Guest Lecture: Matteo Courthoud (Senior Economist and Data Scientist @Zalando)

Part II: Data preparation, analysis, and visualization

Date	Topic
07.12.2023	Visualisation, dynamic documents
07.12.2023	Exercises/Workshop 5: Data preparation and applied data analysis with R
14.12.2023	Guest Lecture: Florian Chatagny (Head of Data Science @Federal Finance Administration in Bern)
21.12.2023	Exercises/Workshop 6: Visualization, dynamic documents
21.12.2023	Summary, Wrap-Up, Q&A, Feedback
21.12.2023	Exam for Exchange Students

Summary and warm up

Summary: Data

Rectangular data

- Import data from text files, csv, tsv, etc.
- Tibbles, data frames in R

Non-rectangular data

- Hierarchical data (xml, html, json)
- Unstructured text data
- Images/Pictures data

A Template/Blueprint

Tell your future self what this script is all about

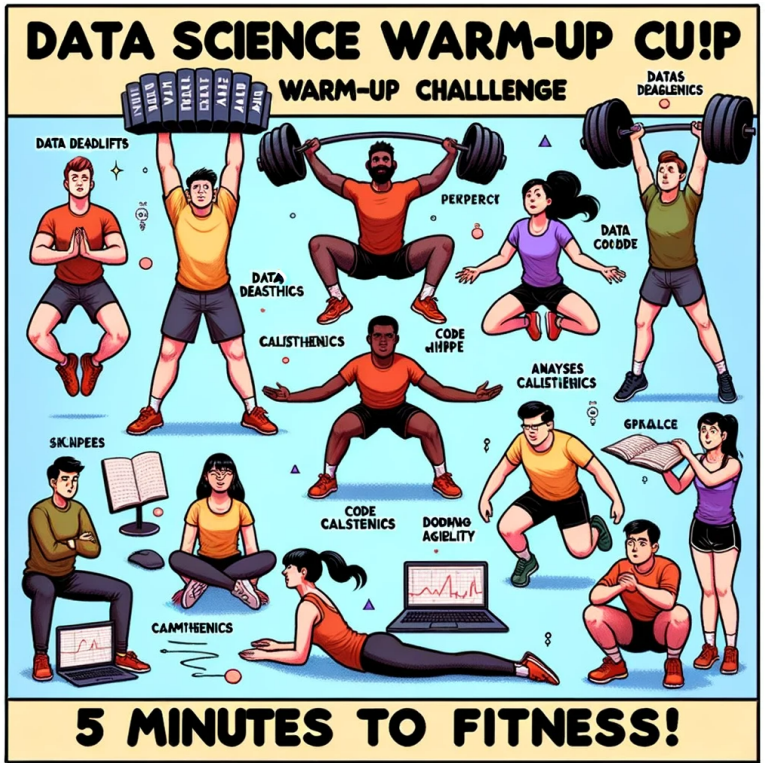
```
########################################################################
# Project XY: Data Gathering and Import
#
# This script is the first part of the data pipeline of project XY.
# It imports data from ...
# Input: links to data sources (data comes in ... format)
# Output: cleaned data as CSV
#
# A. Sallin, St. Gallen, 2023
########################################################################


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

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


# IMPORT RAW DATA FROM CSVs -------------
```

Warm up


DATA SCIENCE WARM-UP CU!P
WARM-UP CHALLENGE
DATA DEADLIFTS
DATAS DEAGLENICS
PERPERCT
DATA COLDDE
DATA DEASTHICS
CALISTHENICS
CODE
ANAYSES CALISTIENICS
SKNPEES
CODE CALSTENICS
AGIELITY
CAMITHENICS
5 MINUTES TO FITNESS!

JSON files: open-ended question

Be the JSON file


```
{
  "students": [
    {
      "id": 19091,
      "firstName": "Peter",
      "lastName": "Mueller",
      "grades": {
          "micro": 5,
          "macro": 4.5,
          "data handling": 5.5
          }
    },
    {
      "id": 19092,
      "firstName": "Anna",
      "lastName": "Schmid",
      "grades": {
          "micro": 5.25,
          "macro": 4,
          "data handling": 5.75
          }
    },
    {
      "id": 19093,
      "firstName": "Noah",
      "lastName": "Trevor",
      "grades": {
          "micro": 4,
          "macro": 4.5,
          "data handling": 5
          }
    }
```

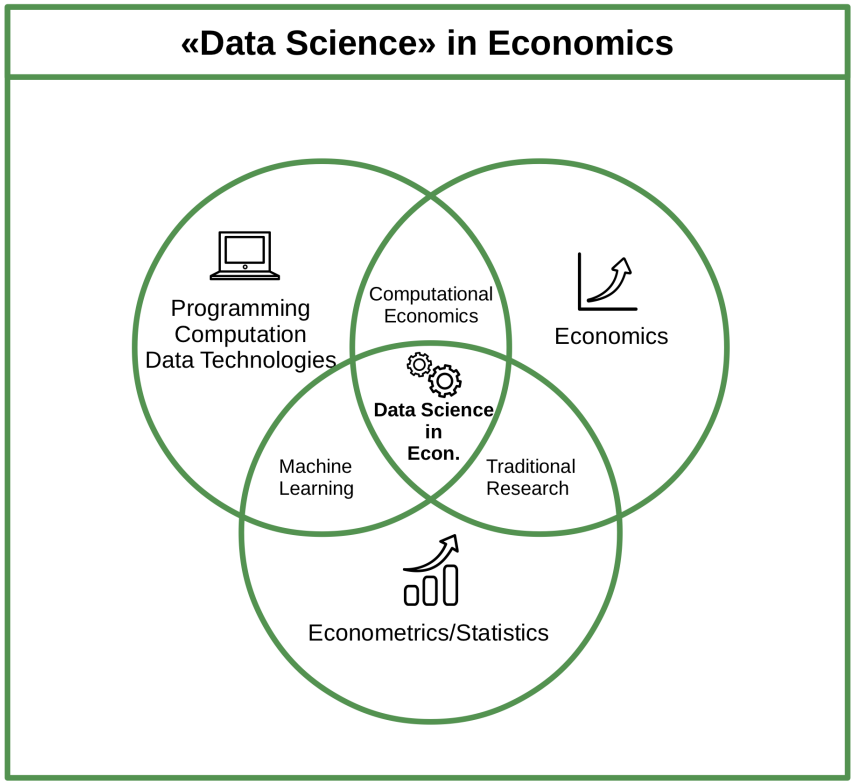
XML

```
<students>
  <student>
    <id>19091</id>
    <firstName>Peter</firstName>
    <lastName>Mueller</lastName>
    <grades>
      <micro>5</micro>
      <macro>4.5</macro>
      <dataHandling>5.5</dataHandling>
    </grades>
  </student>
  <student>
    <id>19092</id>
    <firstName>Anna</firstName>
    <lastName>Schmid</lastName>
    <grades>
      <micro>5.25</micro>
      <macro>4</macro>
      <dataHandling>5.75</dataHandling>
    </grades>
  </student>
  <student>
    <id>19093</id>
    <firstName>Noah</firstName>
    <lastName>Trevor</lastName>
    <grades>
      <micro>4</micro>
      <macro>4.5</macro>
      <dataHandling>5</dataHandling>
    </grades>
  </student>
</students>
```

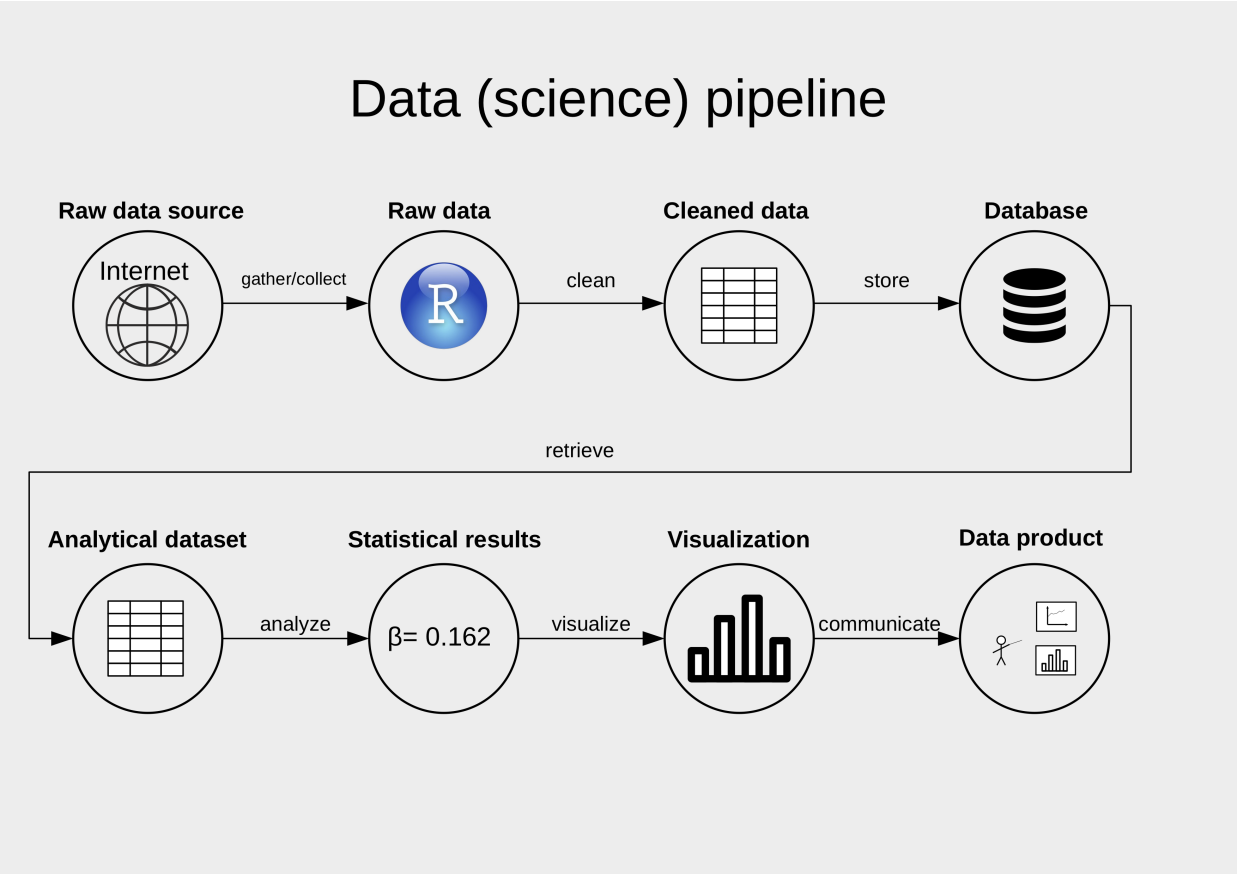
- ‘students’ is the root-node, ‘grades’ are its children

Part II: Data gathering and preparation

Part II: Data gathering and preparation



Part II: Data gathering and preparation



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

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’ (long) format.

“Garbage in garbage out” principle

Move to Nuvolos

Data preparation: three concepts

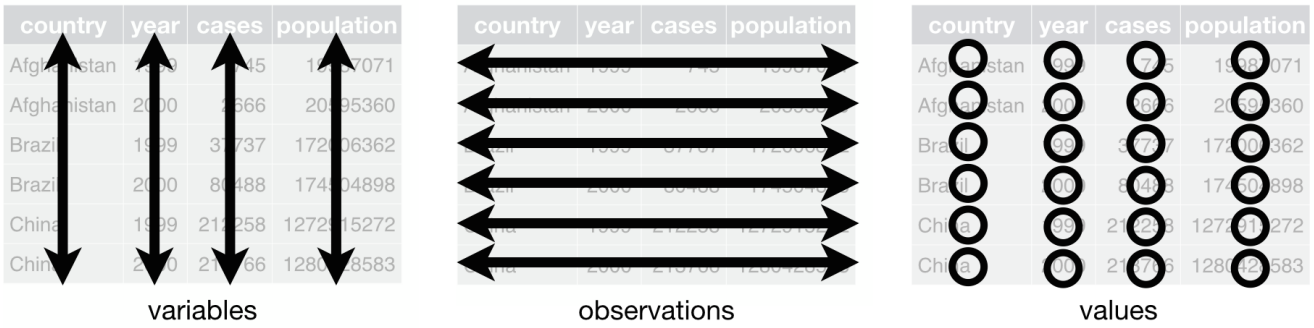
- Tidy data
- Reshaping
- Stacking

Tidy data: some vocabulary

Following Wickham (2014), a tidy dataset is tidy when…

1. Each variable is a column; each column is a variable.
2. Each observation is a row; each row is an observation.
3. Each value is a cell; each cell is a single value.

Tidy data



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

Three examples of non-tidy data (1)

Messy:

```
##       measure Jan.1 Jan.2 Jan.3
## 1 Temperature    20    22    21
## 2    Humidity    80    78    82
```

Tidy:

...

Three examples of non-tidy data (1)

Messy 💩

```
##       measure Jan.1 Jan.2 Jan.3
## 1 Temperature    20    22    21
## 2    Humidity    80    78    82
```

Tidy 😎

```
## # A tibble: 3 × 3
##   Date  Temperature Humidity
##   <chr>       <dbl>    <dbl>
## 1 Jan.1          20       80
## 2 Jan.2          22       78
## 3 Jan.3          21       82
```

Three examples of non-tidy data (2)

Messy:

```
##   year temperature_location
## 1 2019           22C_London
## 2 2019            18C_Paris
## 3 2019             25C_Rome
```

Tidy:

homework..

Three examples of non-tidy data (3)

Messy:

```
##    Student Econ DataHandling Management
## 1 Johannes 5.00          4.0        5.5
## 2   Hannah 5.25          4.5        6.0
## 3     Igor 4.00          5.0        6.0
```

Tidy:

homework..

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

Reshaping: implementation in `R`

- From wide to long: `melt()`, `gather()`,

👉 We'll use `tidyverse::pivot_longer()`.

- From long to wide: `cast()`, `spread()`,

👉 We'll use `tidyverse::pivot_wider()`.

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

Stack/row-bind: implementation in `R`

- Use **`rbind()`** in base **`R`**
  - Requires that the data frames have the same column names and same column classes.

- Use **`bind_rows()`** from **`dplyr()`**
  - More flexible
  - Binds data frames with different column names and classes
  - Automatically fills missing columns with **`NA`**

For these reasons (+ performance, handling or row names, and handling of factors), `dplyr::bind_rows()` is preferred in most applications.

Move to Nuvolos

Summary

Reshaping: summary

"Long" format

country	year	metric
x	1960	10
x	1970	13
x	2010	15
y	1960	20
y	1970	23
y	2010	25
z	1960	30
z	1970	33
z	2010	35

"Wide" format

country	yr1960	yr1970	yr2010
x	10	13	15
y	20	23	25
z	30	33	35

Long and wide data. Source: Hugo Tavares

Reshaping: summary

country	year	metric
x	1960	10
x	1970	13
x	2010	15
y	1960	20
y	1970	23
y	2010	25
z	1960	30
z	1970	33
z	2010	35

```
pivot_wider(names_from = “year”,
            names_prefix = “yr”,
            values_from = “metric”)
```

country	yr1960	yr1970	yr2010
x	10	13	15
y	20	23	25
z	30	33	35

```
pivot_longer(cols = yr1960:yr2010,
             names_to = “year”,
             names_prefix = “yr”
             values_to = “metric”)
```

Long and wide data with code. Source: Hugo Tavares

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/.