CSE P 590

Building Data Analysis Pipelines

Fall 2024

Data wrangling

Today

- Wide vs. long data
- Tidy data
- Data encoding
- Data wrangling: live demo and in-class exercise

Wide vs. long data

Example study: completing coding tasks

Study design

- Two participants
  - S1
  - S2
- Three observations
  - T1: morning
  - T2: noon
  - T3: afternoon

Example study: wide format

Study design

- Two participants
  - S1
  - S2
- Three observations
  - T1: morning
  - T2: noon
  - T3: afternoon

Wide format

ID	T1	T2	T3
S1	0.2	0.4	0.6
S2	0.1	0.3	0.5

Example study: long format

Study design

- Two participants
  - S1
  - S2
- Three observations
  - T1: morning
  - T2: noon
  - T3: afternoon

Wide format

ID	T1	T2	T3
S1	0.2	0.4	0.6
S2	0.1	0.3	0.5

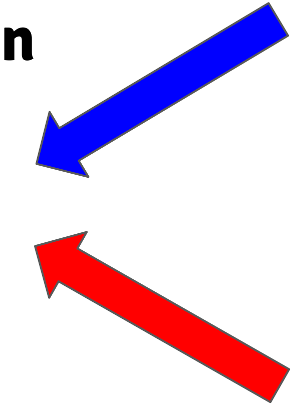
Long format

ID	Time	Value
S1	T1	0.2
S1	T2	0.4
S1	T3	0.6
S2	T1	0.1
S2	T2	0.3
S2	T3	0.5

Example study: data aggregation

Computing the median

ID	Median
S1	0.4
S2	0.3



Wide format

ID	T1	T2	T3
S1	0.2	0.4	0.6
S2	0.1	0.3	0.5

Long format

ID	Time	Value
S1	T1	0.2
S1	T2	0.4
S1	T3	0.6
S2	T1	0.1
S2	T2	0.3
S2	T3	0.5

Wide vs. long data format: why do we care?

Questions

1. Does the study design dictate the data layout?
2. What are the pros and cons for each data layout?
3. Why do we care about the data layout?

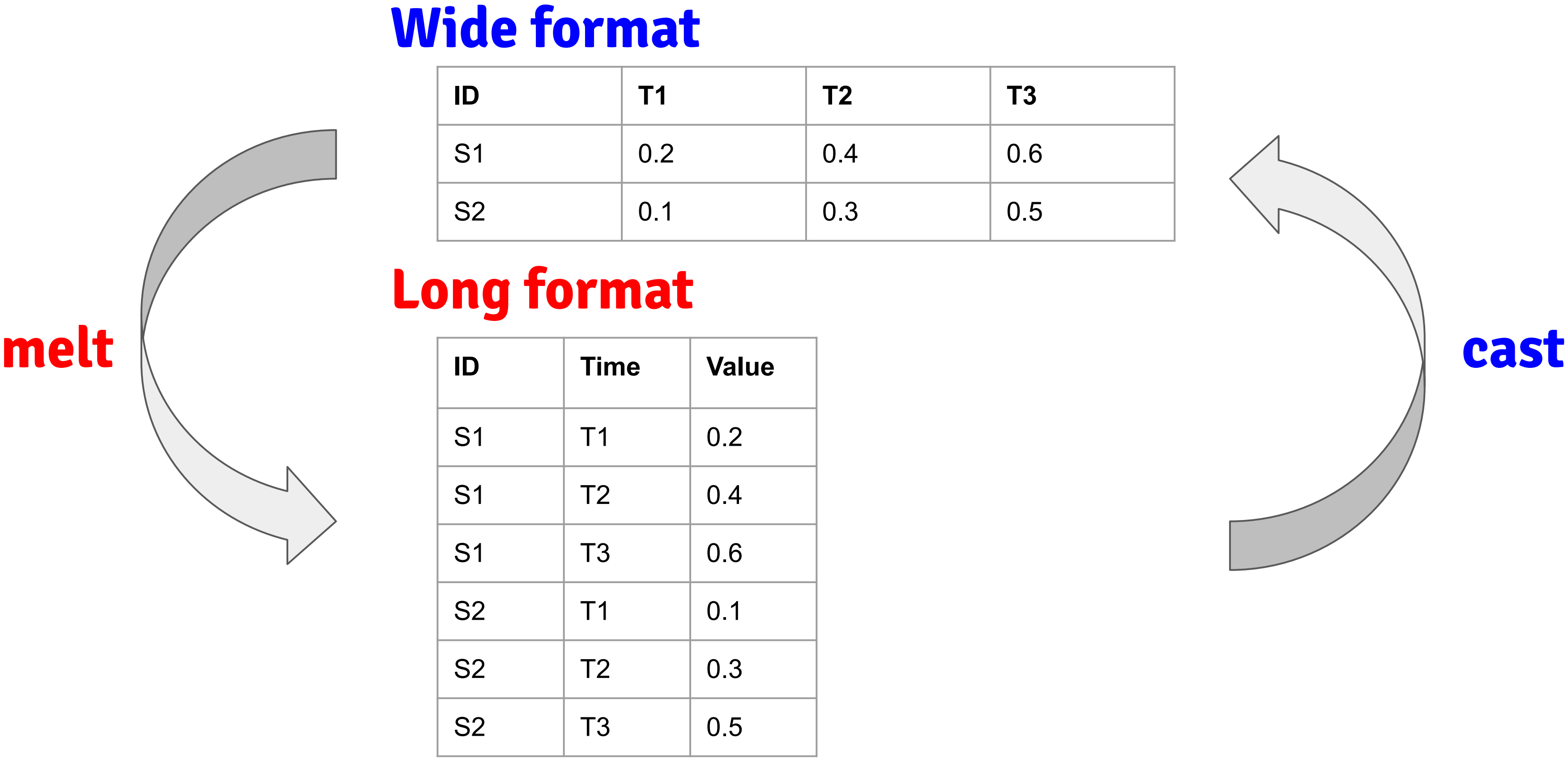
Wide format

ID	T1	T2	T3
S1	0.2	0.4	0.6
S2	0.1	0.3	0.5

Long format

ID	Time	Value
S1	T1	0.2
S1	T2	0.4
S1	T3	0.6
S2	T1	0.1
S2	T2	0.3
S2	T3	0.5

Cast: convert long to wide format



Wide format

ID	T1	T2	T3
S1	0.2	0.4	0.6
S2	0.1	0.3	0.5

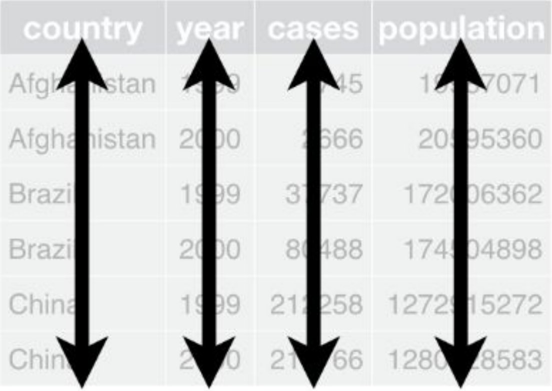
Long format

ID	Time	Value
S1	T1	0.2
S1	T2	0.4
S1	T3	0.6
S2	T1	0.1
S2	T2	0.3
S2	T3	0.5

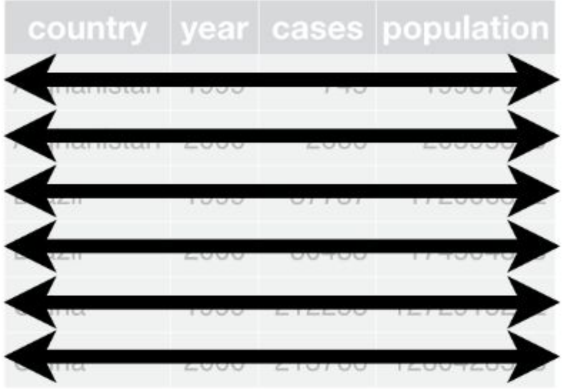
Tidy data

Tidy data: three rules

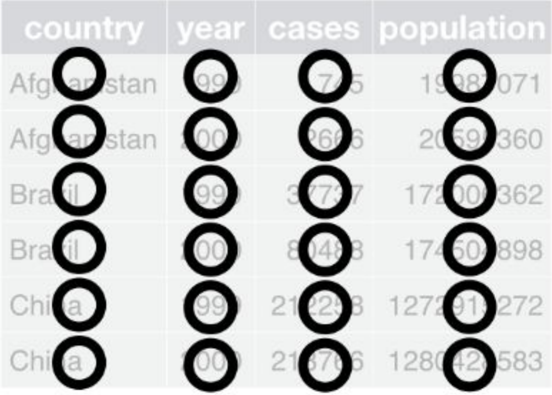
1. Each **variable** has its own **column**.
2. Each **observation** has its own **row**.
3. Each **value** has its own **cell**.



variables



observations

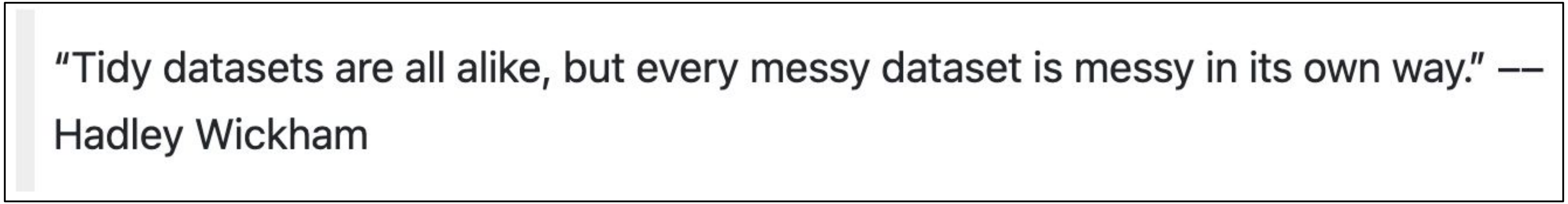


values

Tidy data: advantages

Advantages of tidy data

- Consistent data structure → easier to learn related tools (uniformity).
- Variables in columns → easier to take advantage of vectorized code.
- Tidyverse packages are designed to work with tidy data.



> “Tidy datasets are all alike, but every messy dataset is messy in its own way.” –– Hadley Wickham

Data encoding

Data encoding: recall the types of variables

- **Categorical** (nominal)
  - Unordered set of values
  - Example: [HCI, PLSE, Robotics, UbiComp]
- **Dichotomous** (dichotomized or “natural” dichotomy)
  - Categorical with exactly two possible values
  - Example: [Day, Night]
- **Ordinal**
  - Ordered set of values (no assumption about equidistant values)
  - Example: [low, medium, high]
- **Continuous/Interval**
  - Ordered values (equidistant values)
  - Example: [0..100]

Data encoding: best practices

General advice

- Be explicit about data types (in data sources and code)
- Use factors with fixed (known) factor levels
    - Avoid encoding factors as integers or strings
- Check for incomplete or corrupted data
    - NAs are everywhere
- Let domain knowledge guide decisions about encoding
    - Binning of continuous data (e.g., response time)
    - Categorical vs. ordinal vs. continuous data

Data wrangling: live demo

Data wrangling: in-class exercise