Thinking About Data

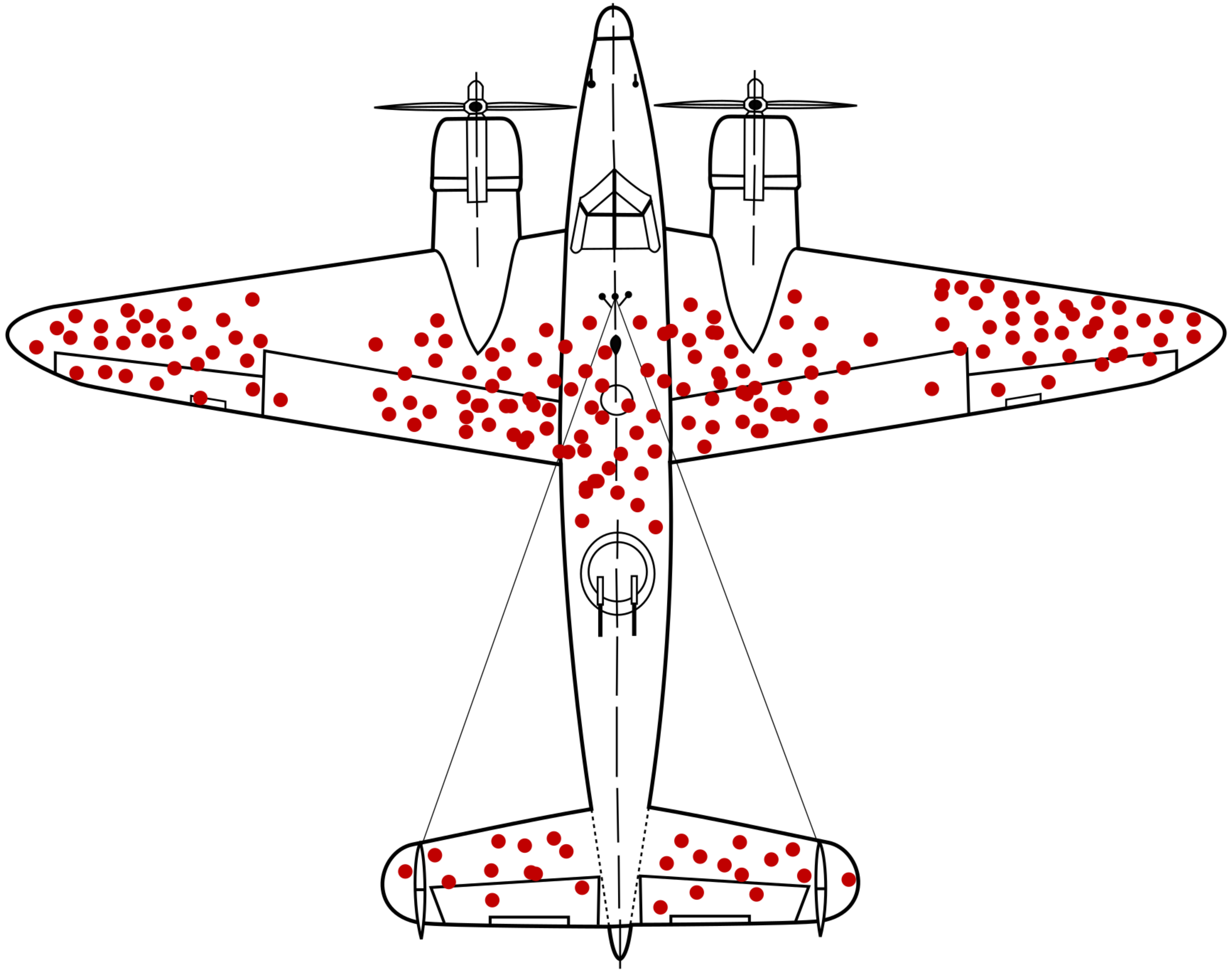
Aaron R. Williams & Alex Engler

There Is Not Such Thing As Raw Data

“I believe that often when people think of datasets, they think of them as being the truth, facts, raw information, something not to be questioned, but I really want everybody to question their data before they go out and use it.”

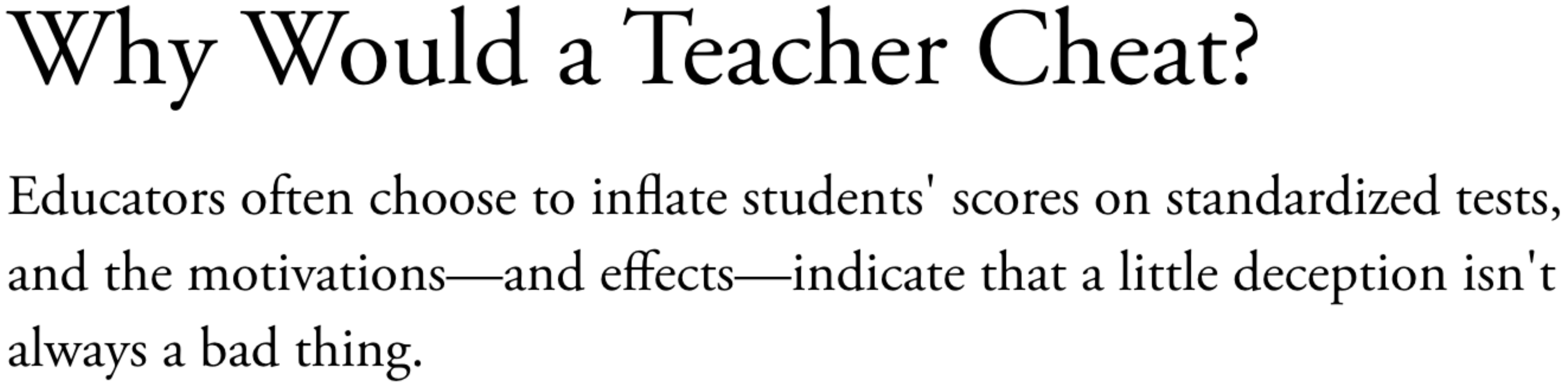
~ Sarah Williams

Data Are the Result of a Process



Source: Cameron Moll

Data Are the Result of a Process



Why Would a Teacher Cheat?

Educators often choose to inflate students' scores on standardized tests, and the motivations—and effects—indicate that a little deception isn't always a bad thing.

Source: The Atlantic

The Process

Experimental Data: Data collected through a process actively controlled by a researcher with interventions

Non-experimental Data: Data observed and collected outside of a controlled experiment

The Process

Type	Typical Process	Example	Strengths	Challenges
Census	Data gathered by measuring characteristics about every unit	Decennial Census	Contains the population	Cost+++
Survey	Data gathered with a questionnaire distributed by a probabilistic design	ACS	Content	Cost and nonresponse
Administrative	Data for a process other than research	Initial Unemployment Insurance Claims	Detail and accuracy	Representativeness of the population of interest, privacy
Extracted	Social media, text extraction, computer vision	Billion Prices	Volume	Representativeness of the population of interest, accuracy
Corporate	Typically administrative data	Credit Bureau Data	Detail and accuracy	Representativeness of the population of interest, access

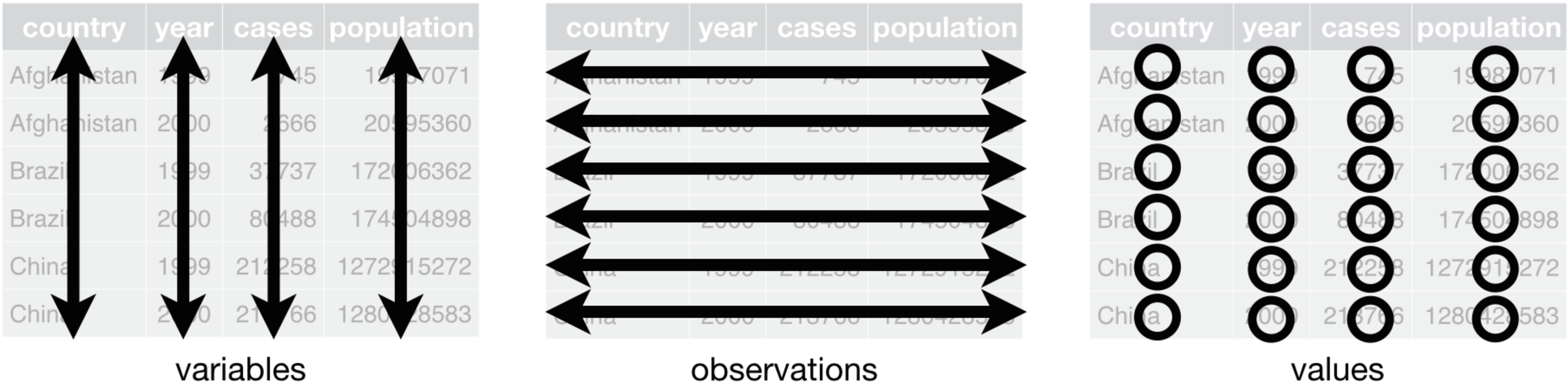
Types of Tabular Data

Type	Rows
Cross-sectional	Observations at one point in time
Pooled cross sections	Different observations at multiple points in time (often with adjustments to variables)
Panel/longitudinal	Observations at two or more points in time
Time series	One or a few observations at many points in time

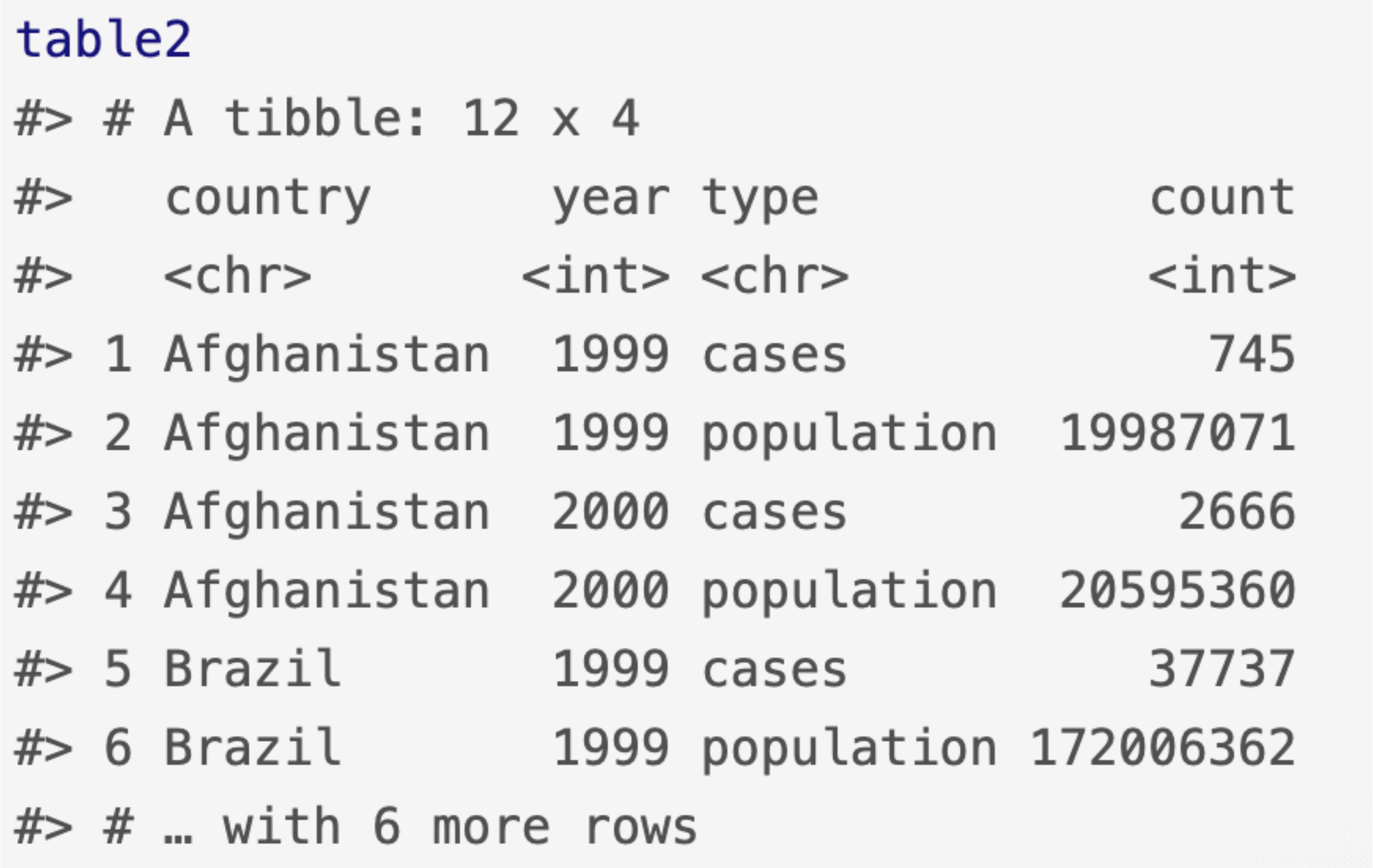
Types of Tabular Data

Type	Rows
Survival	Observations at multiple points in time with varying numbers of rows per observation
Geospatial	Observations with one or more points, lines, or polygons
Hierarchical	Observations at different levels of analysis in one data set (i.e. students and schools)
Spatiotemporal	Geospatial and panel data

Tidy Data


country
year
cases
population
variables
observations
values

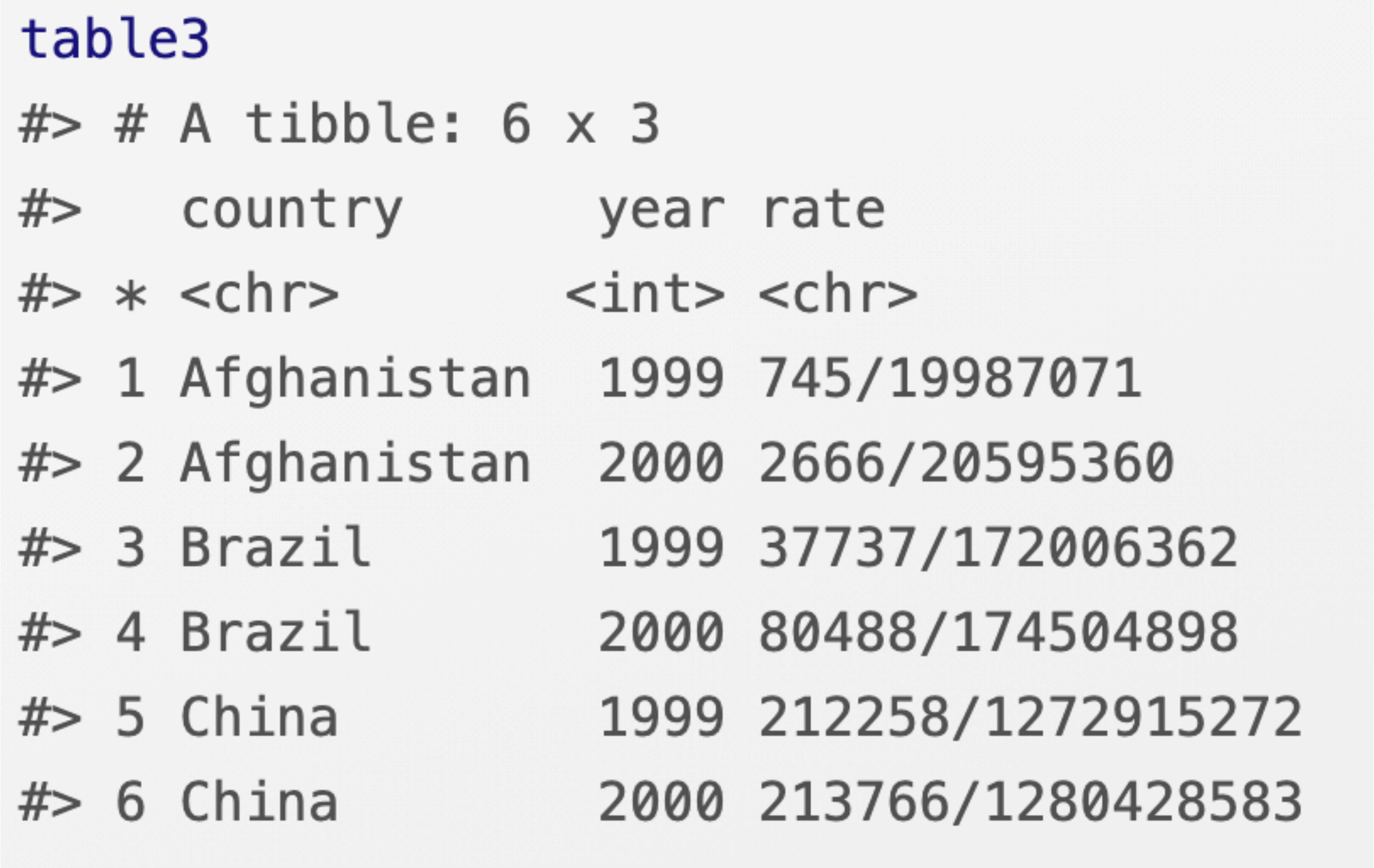
Tidy Data



```
table2
#> # A tibble: 12 x 4
#>   country      year type           count
#>   <chr>       <int> <chr>          <int>
#> 1 Afghanistan  1999 cases            745
#> 2 Afghanistan  1999 population  19987071
#> 3 Afghanistan  2000 cases           2666
#> 4 Afghanistan  2000 population  20595360
#> 5 Brazil       1999 cases          37737
#> 6 Brazil       1999 population 172006362
#> # … with 6 more rows
```

Source: R for Data Science

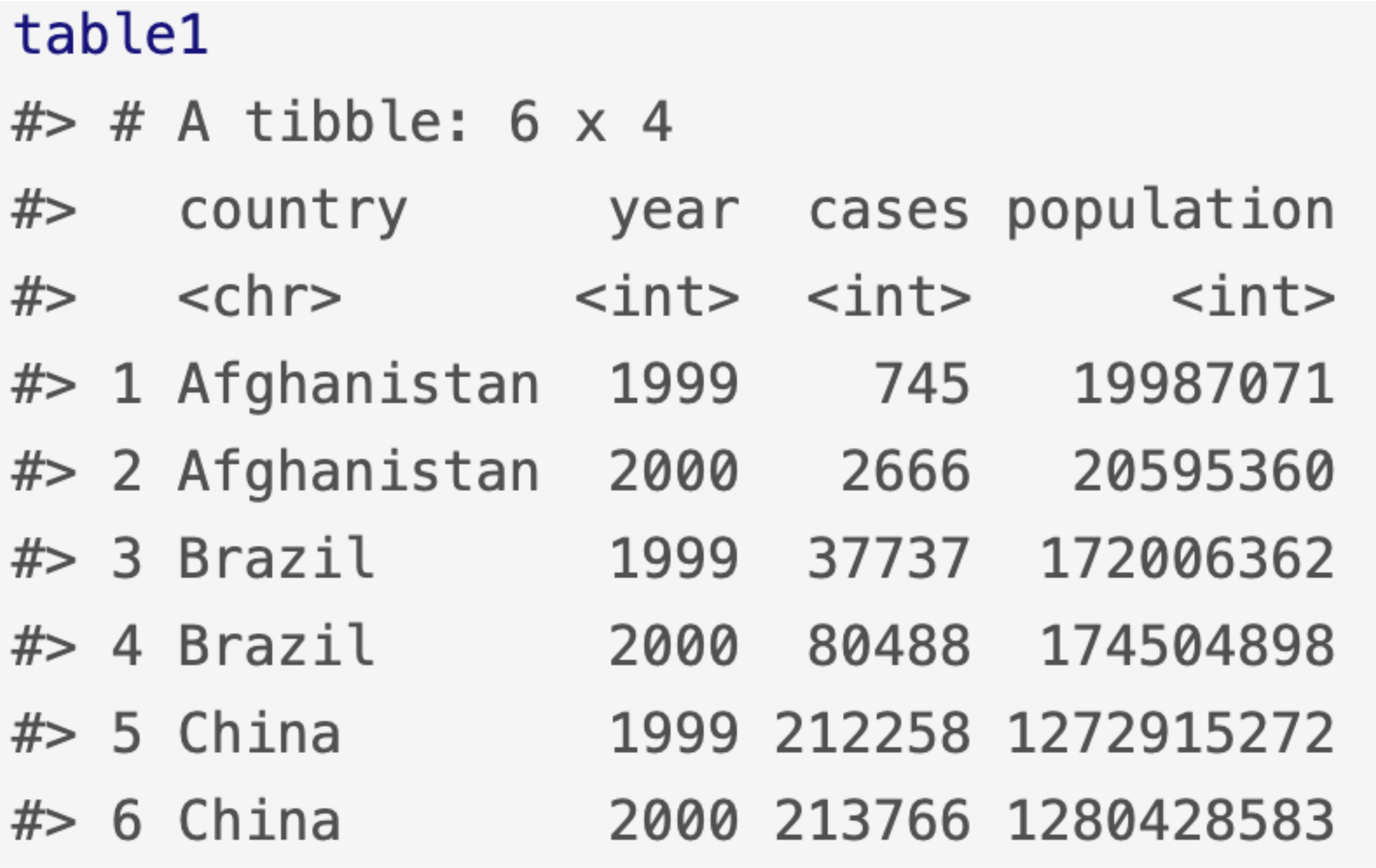
Tidy Data



```
table3
#> # A tibble: 6 x 3
#>   country      year rate
#> * <chr>       <int> <chr>
#> 1 Afghanistan  1999 745/19987071
#> 2 Afghanistan  2000 2666/20595360
#> 3 Brazil       1999 37737/172006362
#> 4 Brazil       2000 80488/174504898
#> 5 China        1999 212258/1272915272
#> 6 China        2000 213766/1280428583
```

Source: R for Data Science

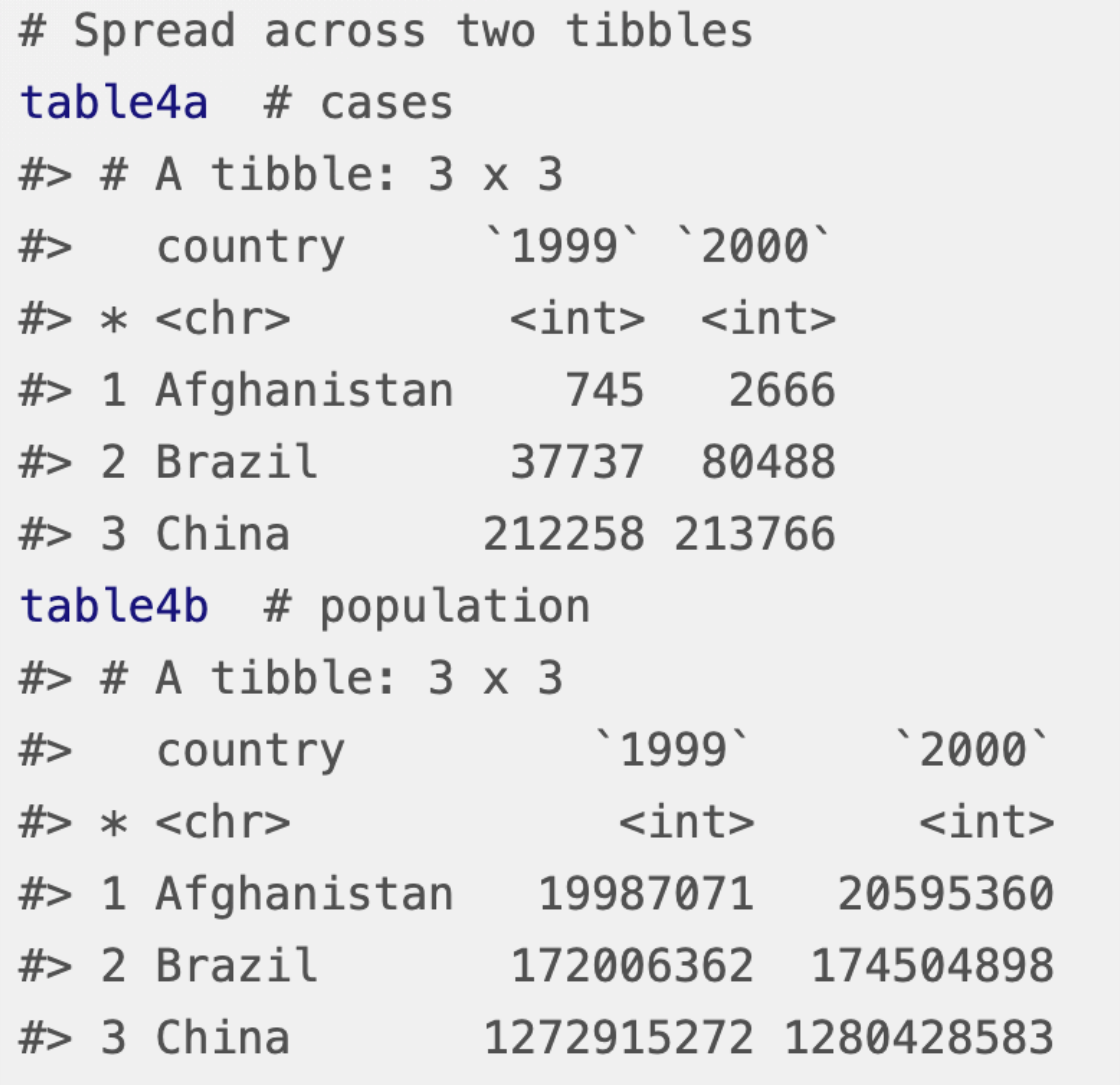
Tidy Data



```
table1
#> # A tibble: 6 x 4
#>   country      year  cases population
#>   <chr>       <int>  <int>      <int>
#> 1 Afghanistan  1999    745   19987071
#> 2 Afghanistan  2000   2666   20595360
#> 3 Brazil       1999  37737  172006362
#> 4 Brazil       2000  80488  174504898
#> 5 China        1999 212258 1272915272
#> 6 China        2000 213766 1280428583
```

Source: R for Data Science

Tidy Data



```
# Spread across two tibbles
table4a  # cases
#> # A tibble: 3 x 3
#>   country     `1999` `2000`
#> * <chr>        <int>  <int>
#> 1 Afghanistan    745   2666
#> 2 Brazil       37737  80488
#> 3 China       212258 213766
table4b  # population
#> # A tibble: 3 x 3
#>   country         `1999`     `2000`
#> * <chr>            <int>      <int>
#> 1 Afghanistan   19987071   20595360
#> 2 Brazil       172006362  174504898
#> 3 China       1272915272 1280428583
```

Source: R for Data Science

Always Read the Documentation

> “Sometimes you will think it is unnecessary. What else could this column possibly mean? Oh, sweet summer child. **Always read the documentation.**”

Alex Engler

Always Create Good Documentation

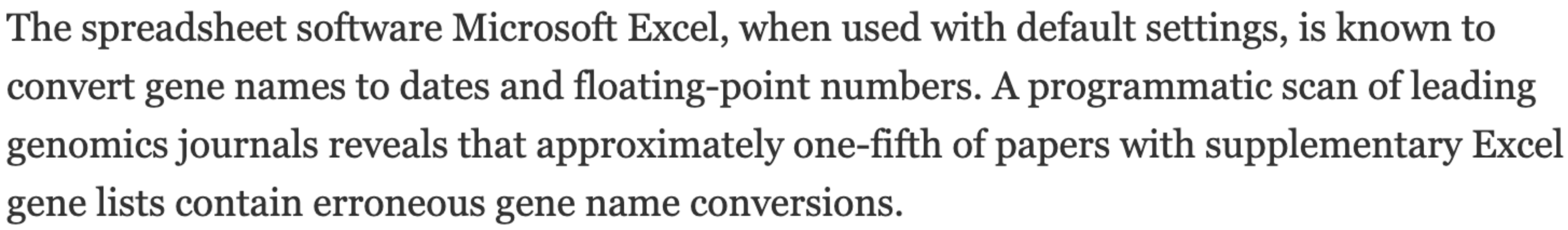
Data Dictionary

- Definition of a row (unit of analysis) (level of data)
- Definition of how to uniquely identify a row (could be multiple columns, but hopefully it’s just one ID)
- Time period of the data
- Definitions of variables and universes of questions (including skip patterns)
- Missing value codes and reasons for missingness (structural? nonresponse?)
- Weights and information about survey designs
- Other information that is likely obscure but crucial to an analysis

Data Storage

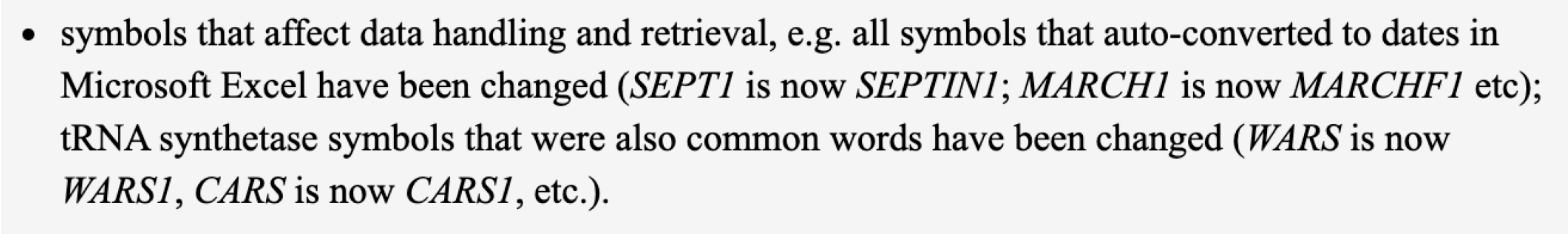
- .csv
- .xlsx
- .pdf
- APIs
- JSON
- Slides

Excel is Bad



The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

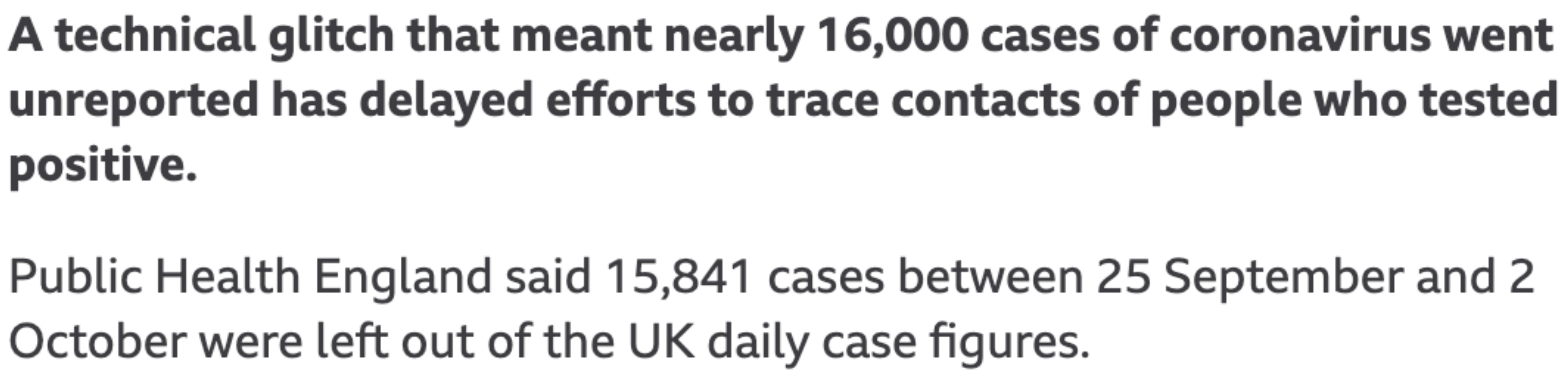
Source: Ziemann, Eren, and El-Osta



- symbols that affect data handling and retrieval, e.g. all symbols that auto-converted to dates in Microsoft Excel have been changed (*SEPT1* is now *SEPTIN1*; *MARCH1* is now *MARCHF1* etc); tRNA synthetase symbols that were also common words have been changed (*WARS* is now *WARS1*, *CARS* is now *CARS1*, etc.).

Source: Bruford, Braschi, Denny, Jones, Seal, and Tweedie

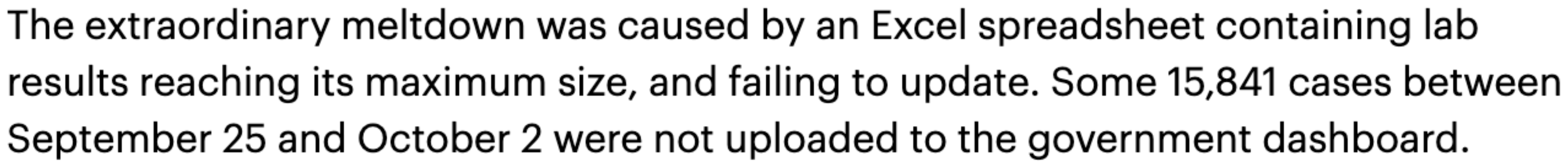
Excel is Terrible



A technical glitch that meant nearly 16,000 cases of coronavirus went unreported has delayed efforts to trace contacts of people who tested positive.

Public Health England said 15,841 cases between 25 September and 2 October were left out of the UK daily case figures.

Source: BBC



The extraordinary meltdown was caused by an Excel spreadsheet containing lab results reaching its maximum size, and failing to update. Some 15,841 cases between September 25 and October 2 were not uploaded to the government dashboard.

Source: DailyMail

Excel Was Good for Science Exactly Once

Does Contact Tracing Work? Quasi-Experimental Evidence from an Excel Error in England

“This paper exploits quasi-random variation in COVID-19 contact tracing. Between September 25 and October 2, 2020, a total of 15,841 COVID-19 cases in England (around 15 to 20% of all cases) were not immediately referred to the contact tracing system due to a data processing error.”

“Conservative estimates suggest that the failure of timely contact tracing due to the data glitch is associated with more than 125,000 additional infections and over 1,500 additional COVID-19-related deaths. Our findings provide strong quasi-experimental evidence for the effectiveness of contact tracing.”

Source: Fetzer and Graber (2021) in the Proceedings of the National Academy of Sciences

Parting Advice

- Read the data dictionary
- Watch out for missing value encodings!
- Validate against published statistics and popular publications
- Defensive programming
- Read the data dictionary