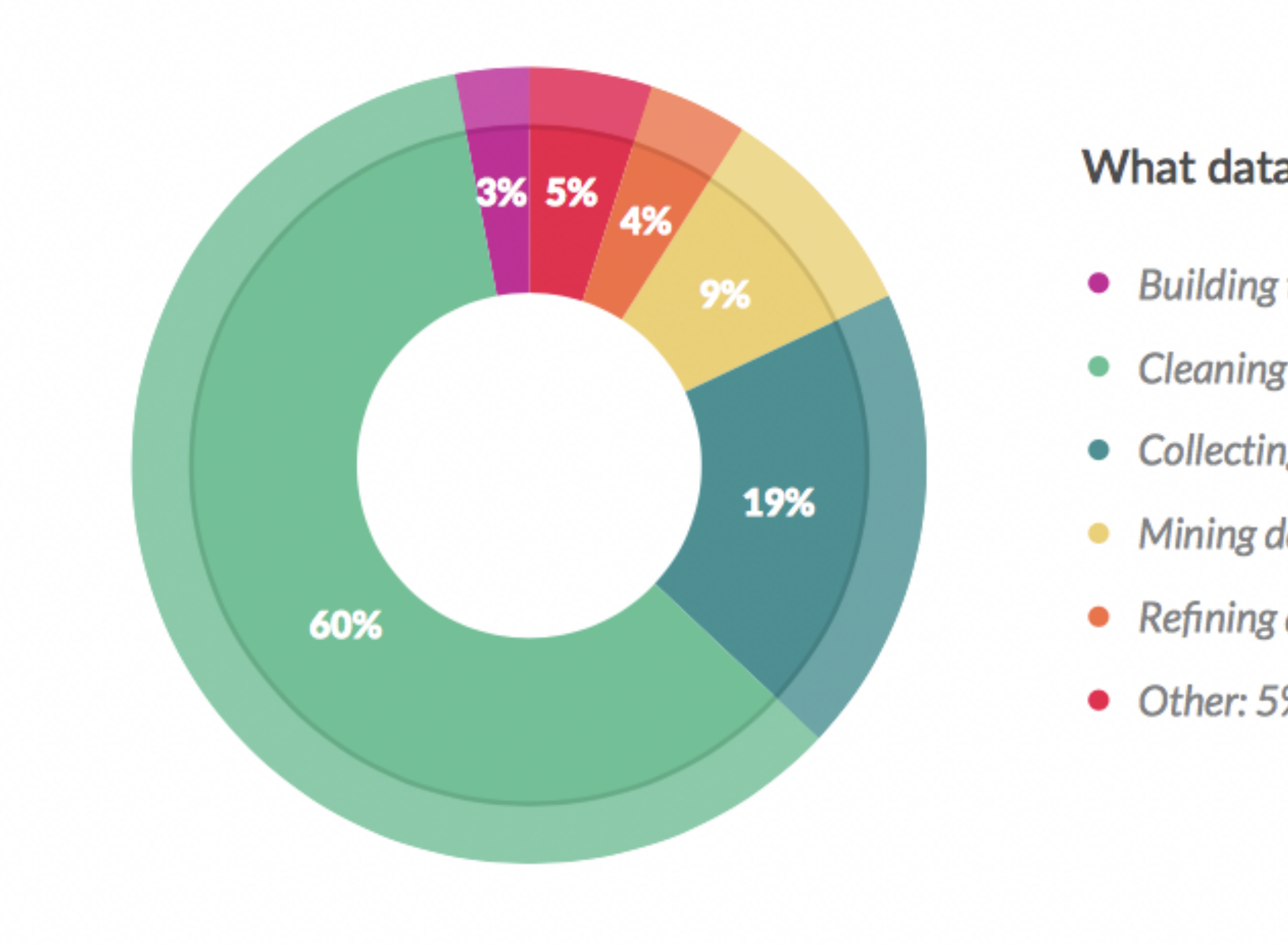
Data Processing for D

Kylie A. Bem

Northeastern Uni
Khoury College of Compu

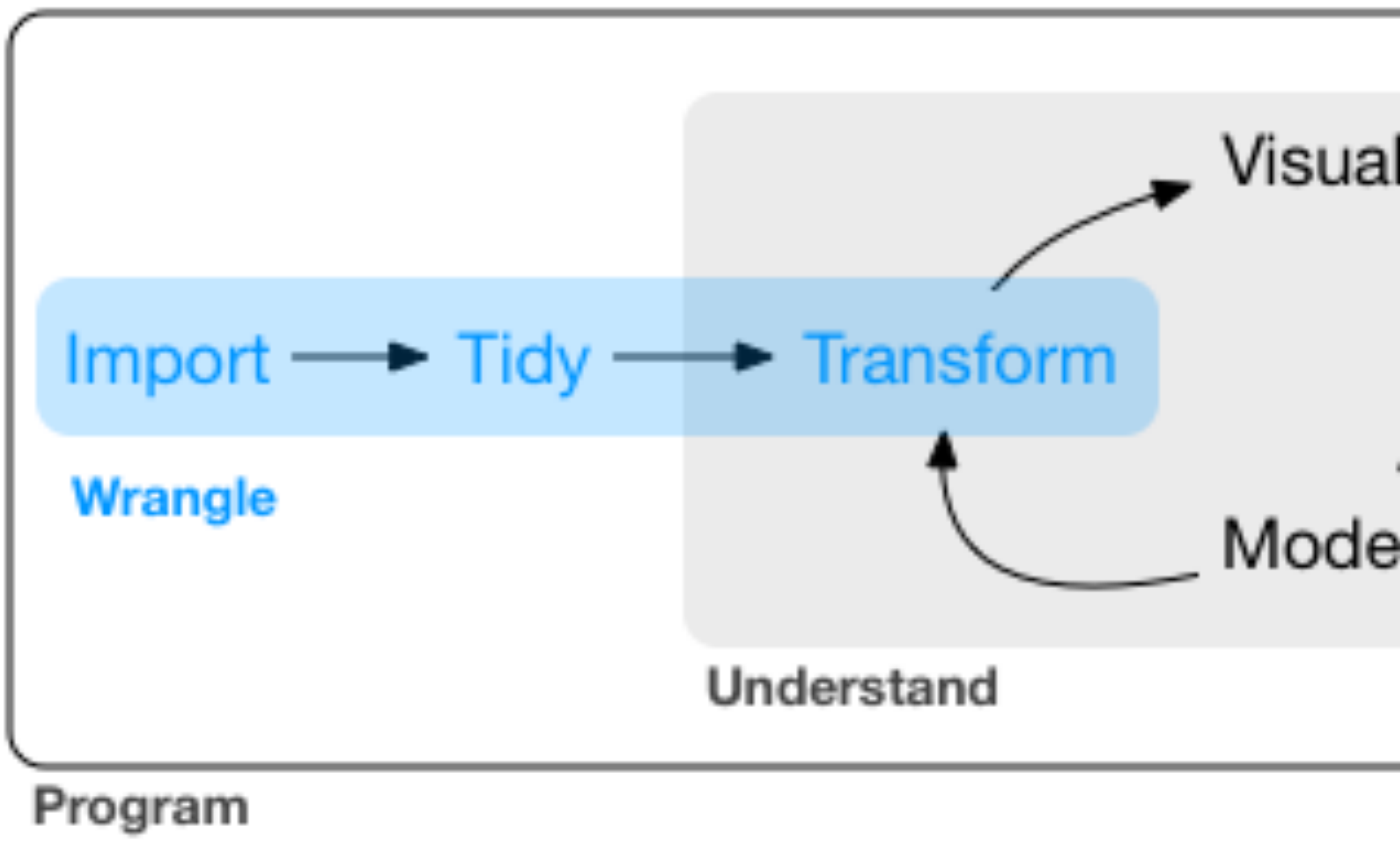
Northeastern U

Wrangling c



https://visit.figure-eight.com/rs/416-ZBE-142/images/Crov

Wrangling c



R for Data Science, Wickham an

Learning gc

- Types of data
- Structuring data for dat
- Data wrangling and dat
- Summarizing data

TYPES OF D

Data comes in ma

- **Structured** data is highl easy to query, transform
- **Semi-structured** data h organization but require
- **Unstructured** data is un requires significant tidyi

Unstructured

- Text, video, images, etc.
- Vast majority of data in abundant on the intern
- Requires significant pro useable for data analysis

Semi-structure

- Structured text, JSON, I
- Follows a structure (e.g requires transformation
- Structured elements of purpose besides data a
- Required amount of pr

Structured c

- Tables in a database or
- High level of organizatio
- May follow a schema (b
- Easy to query, transform

Tabular da

- Most common kind of s
- Follows a "table" forma
  - Rows and columns
  - Values in cells
- Tables in a RDBMS
- Data frames in R, Pytho

STRUCTURIN

Data mode

- A **data model** is a concept
  organize elements of data
- A **data model** is analogous
  in computer programming

Common data

- Relational data
- Key-value pairs
- Graphics and networks
- Arrays and matrices
- Tree structures

Common data

- Relational data
- Key-value pairs
- Graphics and networks
- Arrays and matrices
- Tree structures

Goals of structur

- Make the data easier to
- Ideal structure may diffe the desired computatio
  - Exploratory analysis — "tidy" ta
  - Machine learning — arrays anc
- May need to transform different data models

“Tidy” dat

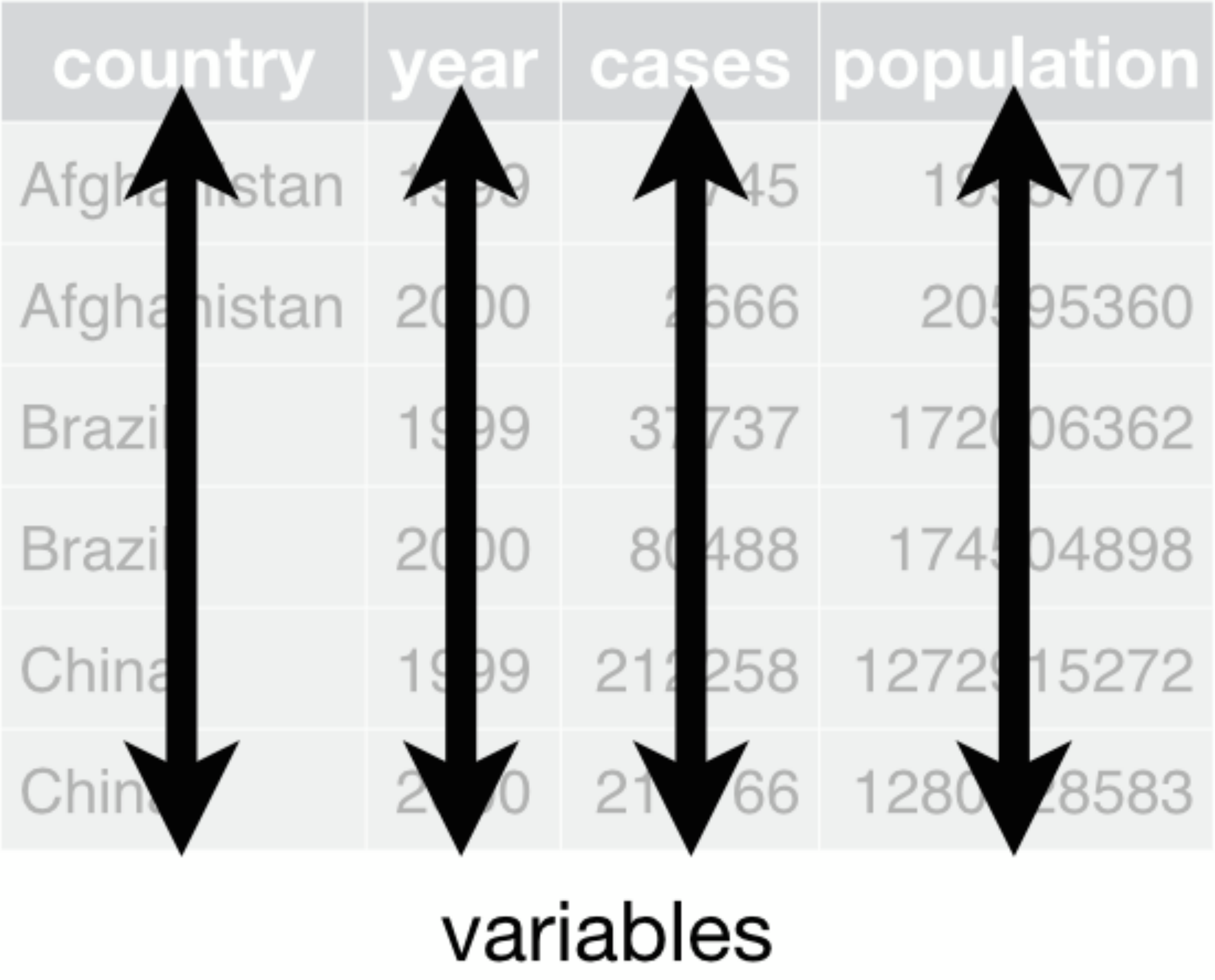
- Each variable forms a co
- Each observation forms
- Each value is a cell
  - Stricter: *Each type of observatio*

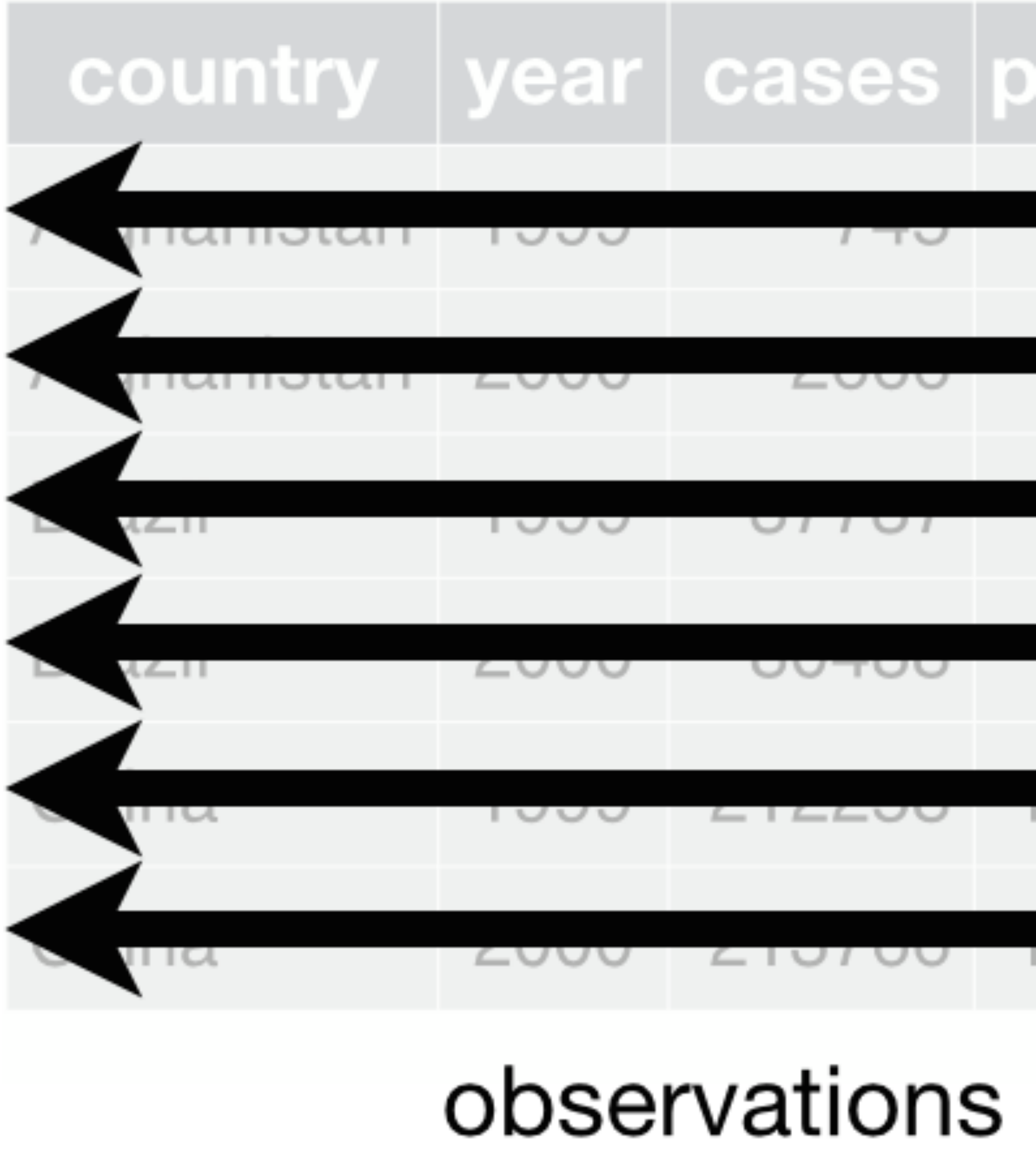
Hadley Wickham. “Tidy Data.” *Journal of S*

Useful definit

- A dataset is a collection o
- An **observational unit** is a on which values are meas
- A **variable** is a quantity, qu that is measured
- An **observation** is a set of made under similar condit

Tidy data





R for Data Science, Wickham an

Why tidy da

- Easy to query, transform
- Consistent format allow a variety of tools (e.g., c
- Relationship to RDBMS
  - Concept of "tidy" data mirrors but framed in language of statis

“Messy” da

- “Messy” data frequently
- Storage and/or comput
  - Messy form may be more com
  - Matrices/arrays preferable for s
- Ease of data entry
  - Data entry by hand
  - Recording instrument
- “Tidy” form not easily e

Common symptoms o

- Column headers are values
- Multiple variables are store
- Variables are stored in both
- A single observation is stor

Is it tidy?

```
## #   tibble: 12 x 4
##    country      year type
##    <chr>       <int> <chr>
##  1  fghanistan  1999 cases
##  2  fghanistan  1999 popul
##  3  fghanistan  2000 cases
##  4  fghanistan  2000 popul
##  5 Brazil       1999 cases
##  6 Brazil       1999 popul
##  7 Brazil       2000 cases
##  8 Brazil       2000 popul
##  9 China        1999 cases
## 10 China        1999 popul
## 11 China        2000 cases
## 12 China        2000 popul
```

Is it tidy? —

```
## #   tibble: 12 x 4
##    country      year type
##    <chr>       <int> <chr>
##   1  fghanistan  1999 cases
##   2  fghanistan  1999 popul
##   3  fghanistan  2000 cases
##   4  fghanistan  2000 popul
##   5 Brazil       1999 cases
##   6 Brazil       1999 popul
##   7 Brazil       2000 cases
##   8 Brazil       2000 popul
##   9 China        1999 cases
## 10 China        1999 popul
## 11 China        2000 cases
## 12 China        2000 popul
```

“cases” and “population” should

Is it tidy?

```
## #   tibble: 6 x 3
##   country        year ra
## * <chr>         <int> <c
## 1  fghanistan   1999 74
## 2  fghanistan   2000 26
## 3 Brazil        1999 37
## 4 Brazil        2000 80
## 5 China         1999 21
## 6 China         2000 21
```

Is it tidy? —

```
## #   tibble: 6 x 3
##   country      year ra
## * <chr>       <int> <c
## 1  fghanistan  1999 74
## 2  fghanistan  2000 26
## 3 Brazil       1999 37
## 4 Brazil       2000 80
## 5 China        1999 21
## 6 China        2000 21
```

“rate” column encodes two variables

Is it tidy?

```
## #   tibble: 3 x 3
##   country      `1999` `
## * <chr>         <int>
## 1  fghanistan    745
## 2 Brazil        37737
## 3 China        212258 2
```

```
## #   tibble: 3 x 3
##   country          `199
## * <chr>             <in
## 1  fghanistan    199870
## 2 Brazil        1720063
## 3 China        12729152
```

Is it tidy? —

```
## #   tibble: 3 x 3
##   country     `1999` `
## * <chr>        <int>
## 1  fghanistan    745
## 2 Brazil       37737
## 3 China       212258 2

## #   tibble: 3 x 3
##   country         `199
## * <chr>            <in
## 1  fghanistan   199870
## 2 Brazil       1720063
## 3 China       12729152
```

observations in multiple tables; co

Is it tidy?

```
## #   tibble: 6 x 4
##   country       year  ca
##   <chr>        <int>  <i
## 1  fghanistan  1999
## 2  fghanistan  2000    2
## 3 Brazil       1999  37
## 4 Brazil       2000  80
## 5 China        1999 212
## 6 China        2000 213
```

Is it tidy? —

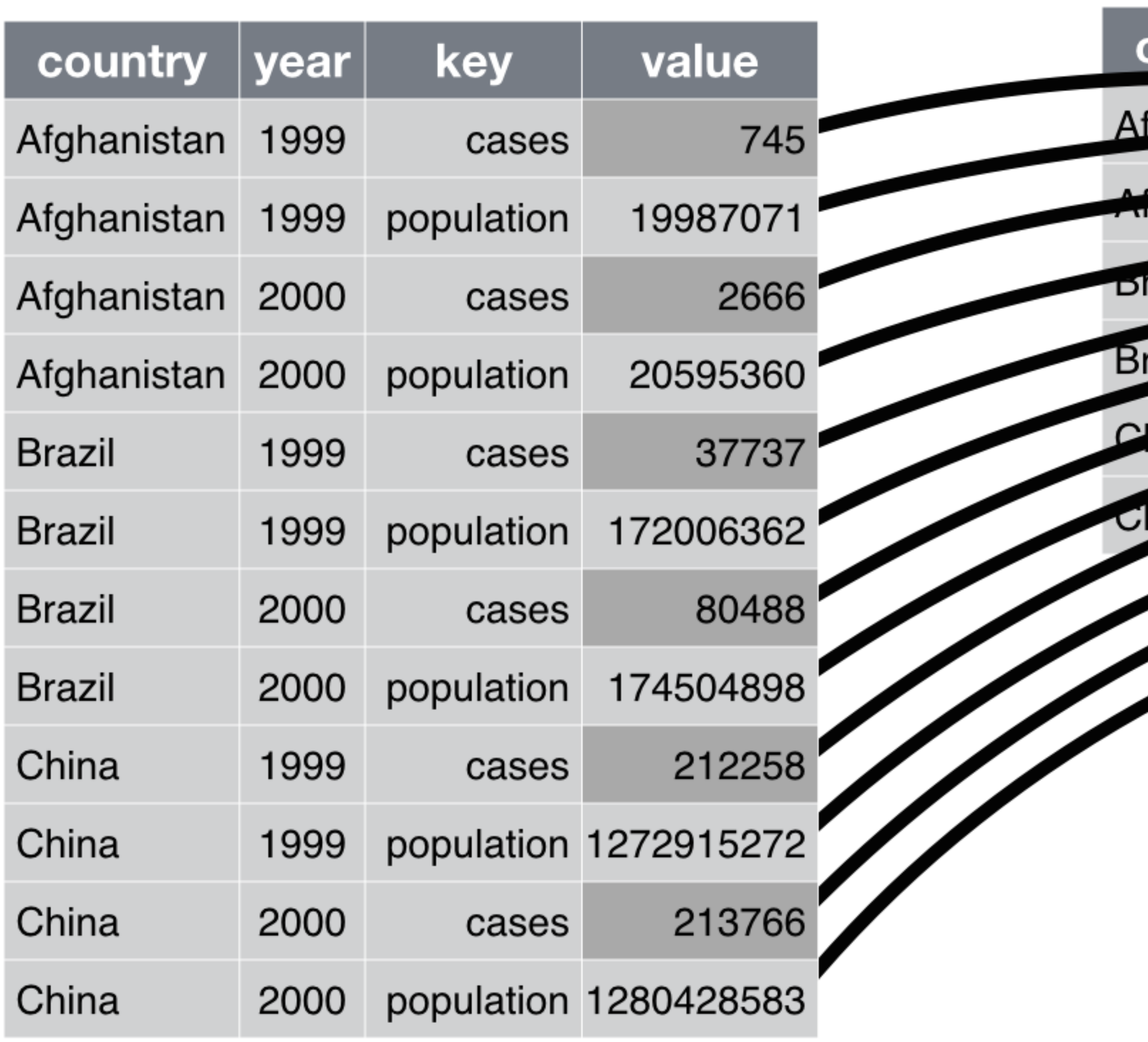
```
## #   tibble: 6 x 4
##   country      year  ca
##   <chr>       <int>  <i
## 1  fghanistan  1999
## 2  fghanistan  2000   2
## 3 Brazil       1999  37
## 4 Brazil       2000  80
## 5 China        1999 212
## 6 China        2000 213
```

Tidying da

- Pre-requisite step to an
- Makes additional data c
- Reshape the dataset int
  - "Wider" — more columns
  - "Longer" — more rows
- Process improperly coc

Going “wid

- Single observations (country-year) scat
- Values of “key” column should be varia



country	year	key	value
Afghanistan	1999	cases	745
Afghanistan	1999	population	19987071
Afghanistan	2000	cases	2666
Afghanistan	2000	population	20595360
Brazil	1999	cases	37737
Brazil	1999	population	172006362
Brazil	2000	cases	80488
Brazil	2000	population	174504898
China	1999	cases	212258
China	1999	population	1272915272
China	2000	cases	213766
China	2000	population	1280428583

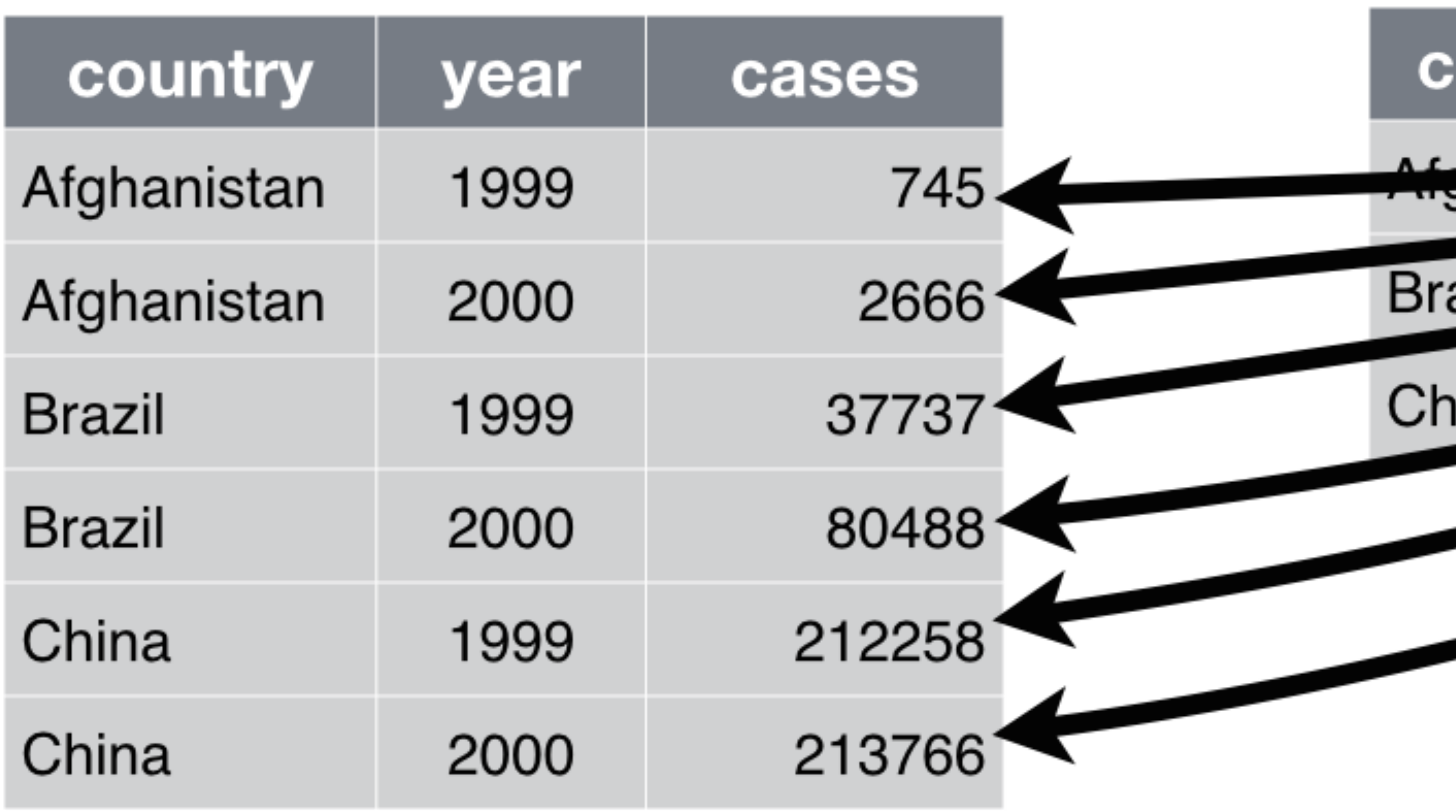
table2

R for Data Science, Wickham an

Going “long

- Single variable (“cases”) spread acros

- Column names are values (1999 and

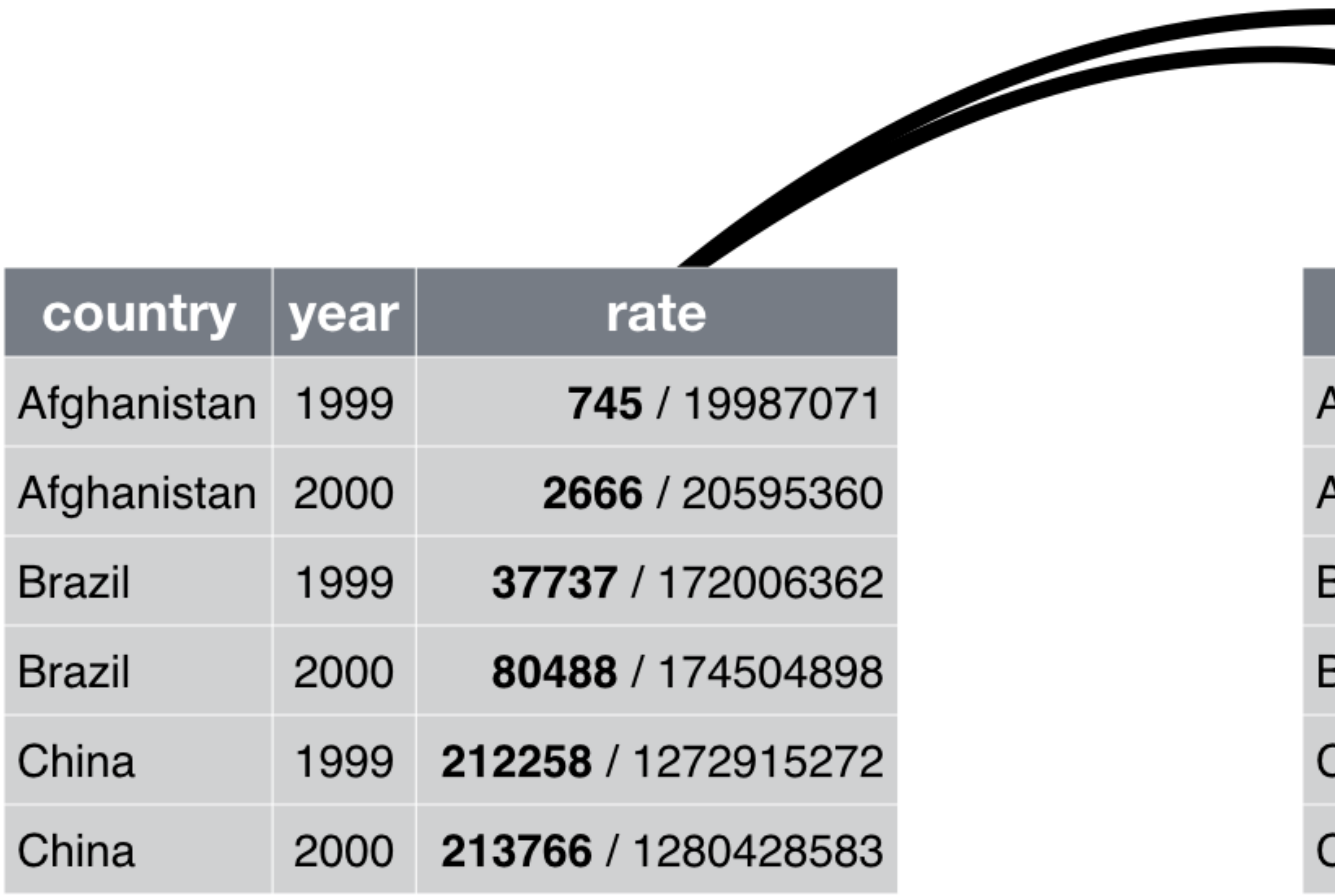


country	year	cases
Afghanistan	1999	745
Afghanistan	2000	2666
Brazil	1999	37737
Brazil	2000	80488
China	1999	212258
China	2000	213766

R for Data Science, Wickham an

Process improperly-co

- Single column (“rate”) encodes two var
- Strings used to represent quantitative (r



country	year	rate
Afghanistan	1999	**745** / 19987071
Afghanistan	2000	**2666** / 20595360
Brazil	1999	**37737** / 172006362
Brazil	2000	**80488** / 174504898
China	1999	**212258** / 1272915272
China	2000	**213766** / 1280428583

table3

R for Data Science, Wickham an

TIDYR

Summary: "tidy

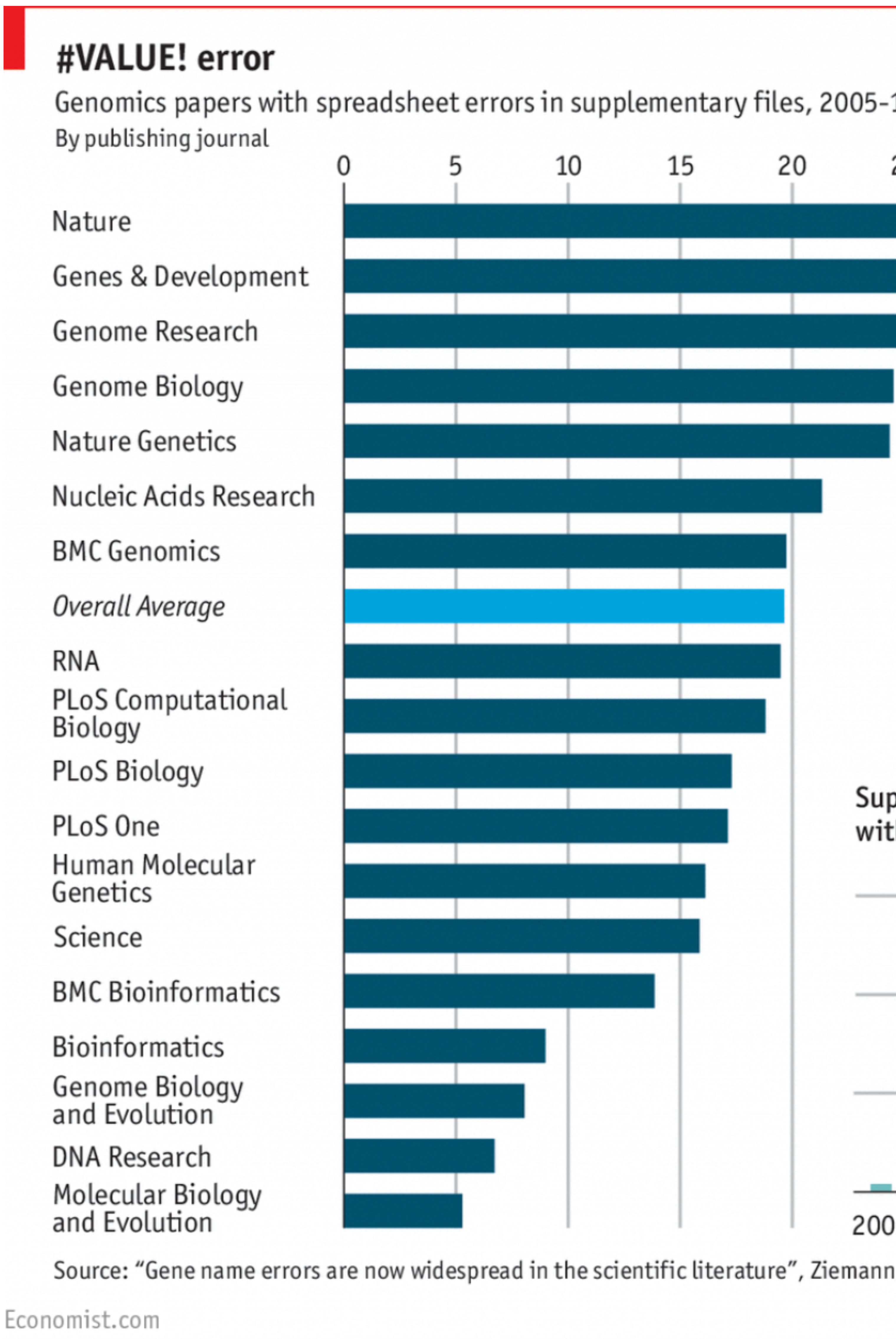
- Easy to query, transform
- Frames database norma in language of statistical

DATA CLEA

Wrangling c

- Post-structuring, additio is often necessary on re
- Consistency
  - Dates and strings often need to
  - Label levels of categorical varia
- Missing data
  - Check for patterns of missing c
  - To impute or not to impute

Never trust a spr



#VALUE! error

Genomics papers with spreadsheet errors in supplementary files, 2005-1

By publishing journal

Source: “Gene name errors are now widespread in the scientific literature”, Ziemann

Economist.com

Strings

- Trim/pad white space
- Normalization and punc
    - Singular vs plural, verb forms, et
- Capitalization/case-foldi
    - Proper vs common nouns
- Special characters and e

Dates and ti

- Consistent input forma
  - MM/DD/YY vs DD-MM-YYYY,
- Convert to appropriate
- Consider time zones
- Be careful of your assur
  - Leap years vs leap seconds, oh
  - ***Use a good library!!!***

Missing da

- Why is the data missing
- What to do about it

Types of missin

- Unit non-response
  - Entire rows of data are missing
  - Usually not directly observed i
  - Very dangerous — sampling bi
- Item non-response
  - Missing values/cells in a colum
  - Can be directly inspected in th

Patterns of missi

- Missing Completely at Rand
  - Missing data are non-systematic and
- Missing at Random (MAR)
  - Missing data are independent of their missingness are related to features of
- Missing Not at Random (MN
  - Missing data are dependent on their

Patterns of missing da

- Missing Completely at Rand
  - Data missing, randomly
- Missing at Random (MAR)
  - Data from earlier years more likely to
- Missing Not at Random (MN
  - Data values near zero more likely to

Methods of imput

- Do nothing
  - Easiest
  - Adequate for some visualizatio
  - Not always possible or approp
- Mean/median/mode im
  - Easy
  - Distorts data — underestimate
  - Appropriate for MCAR and M

Methods of imput

- Zero/constant imputati
  - Easy
  - Introduces bias to the data
  - Can be appropriate for certain
- Algorithmic/model-base
  - Difficult
  - Can be more accurate and less
  - Many methods to choose from

Missing data: final

- Look for patterns of missir
  - Understand why data is missing
- How does the missingness
  - Does it introduce bias?
- Do you need to impute th
  - How does it impact the analysis if y
- Always report what you di

DATA TRANSFOR
SUMMARIZA

Key tasks

- Select columns of interest
- Filter/subset data based o
- Order/rank rows based o
- Transform data and create
- Group and aggregate sum

DPLYR