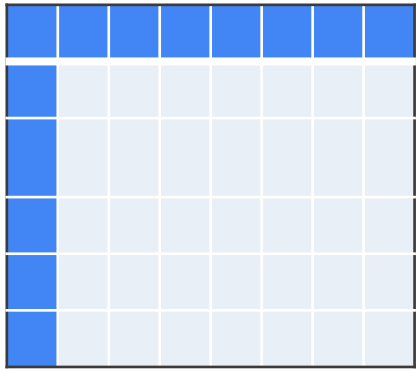
2482612

What’s in a Dataset?

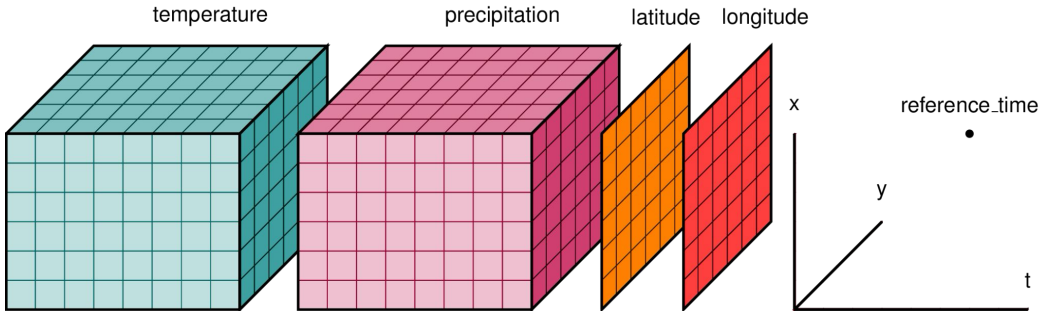
Rectangular and non-rectangular data

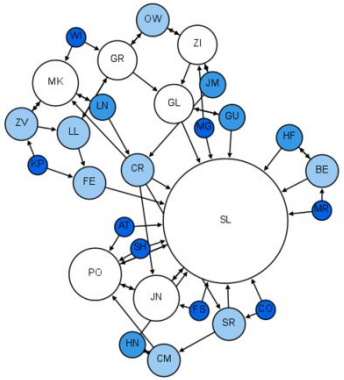
Data come in many different shapes.

Rectangular data



Non-rectangular data





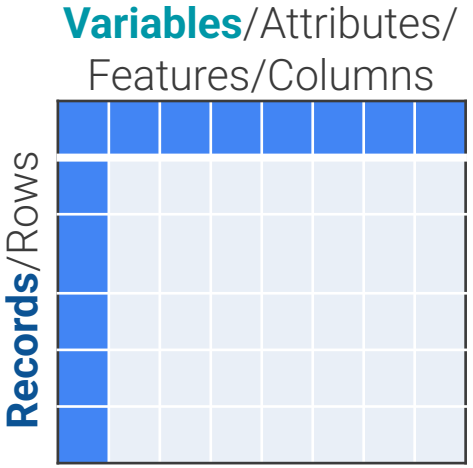
Current U.S Senators

2482612

Rectangular data

We prefer rectangular data for analysis (why?)

- Regular structures are easy to manipulate and analyze
- A big part of data cleaning is about transforming data to be more rectangular
- Two kinds of rectangular data: **tables** and **matrices**

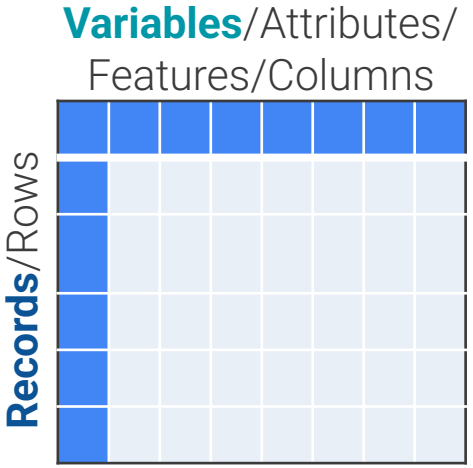


2482612

Rectangular data

Tables (a.k.a. dataframes in R/Python and relations in SQL)

- Named columns with different types
- Manipulated using data transformation grammars (e.g., dplyr, pandas)



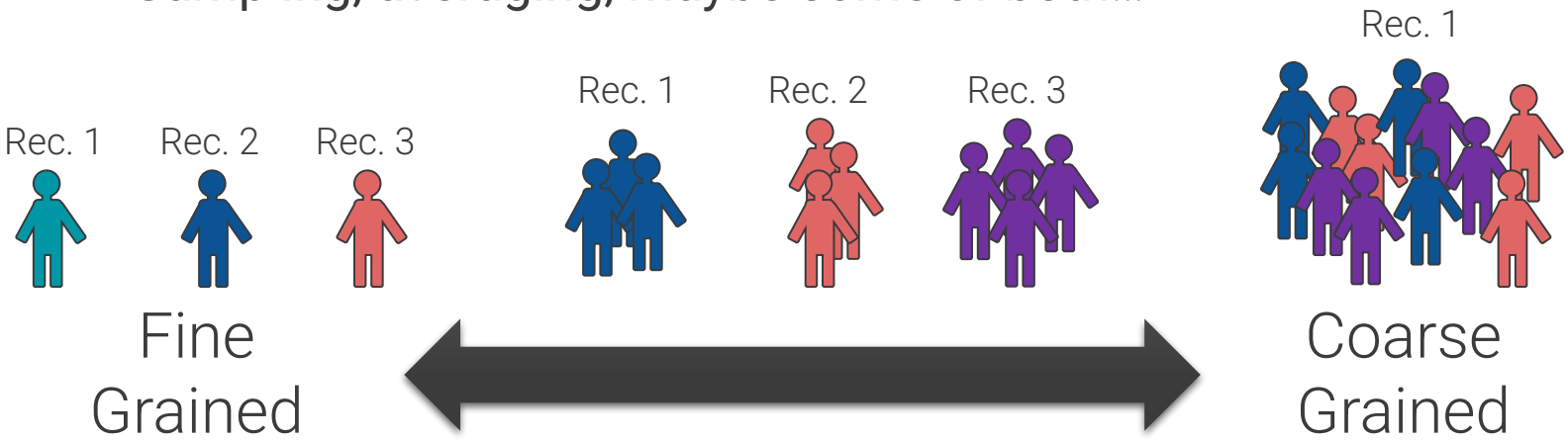
Matrices

- Numeric data of the same type
- Manipulated using linear algebra

2482612

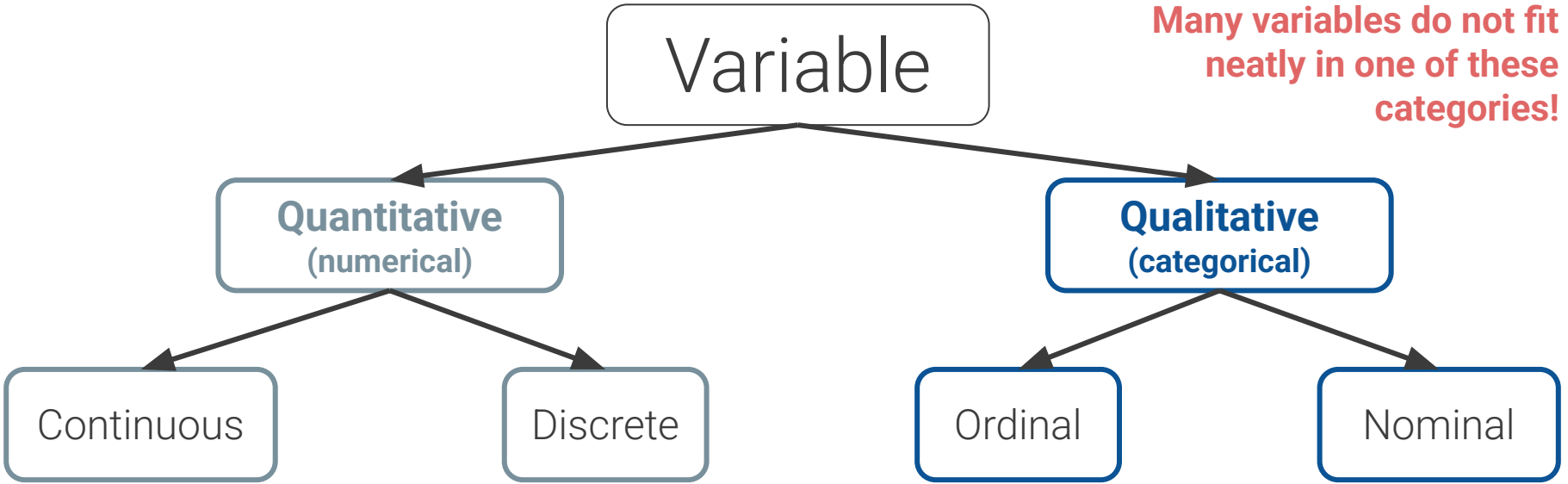
Granularity: how fine/coarse is each record?

- What does each **record** represent?
  - Examples: a purchase, a person, a group of users
- Do all records capture granularity at the same level?
- If the data are **coarse**, how were the records aggregated?
  - Sampling, averaging, maybe some of both…



2482612

Variables have types



Many variables do not fit neatly in one of these categories!

Could be measured to arbitrary precision.

Examples:
- Price
- Temperature

Finite possible values

Examples:
- Number of siblings
- Number of classes

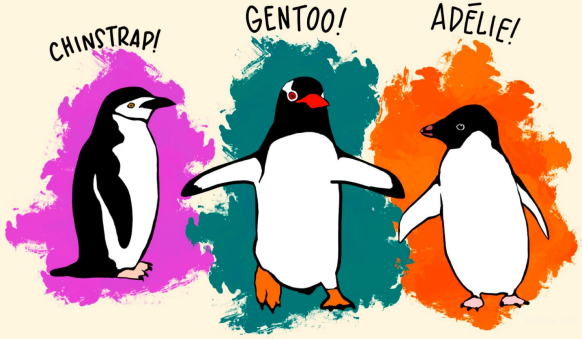
Categories with some semantic ordering.

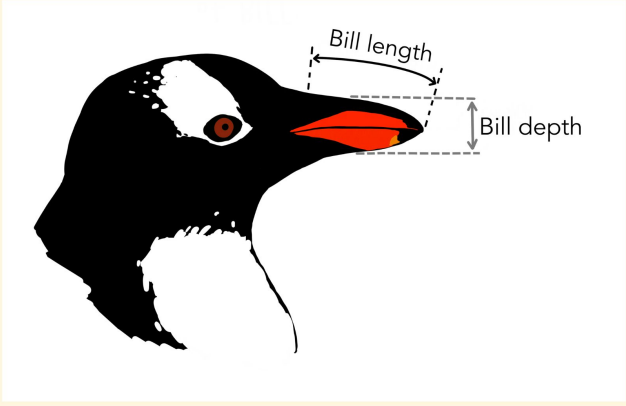
Examples:
- Preferences
- Level of education

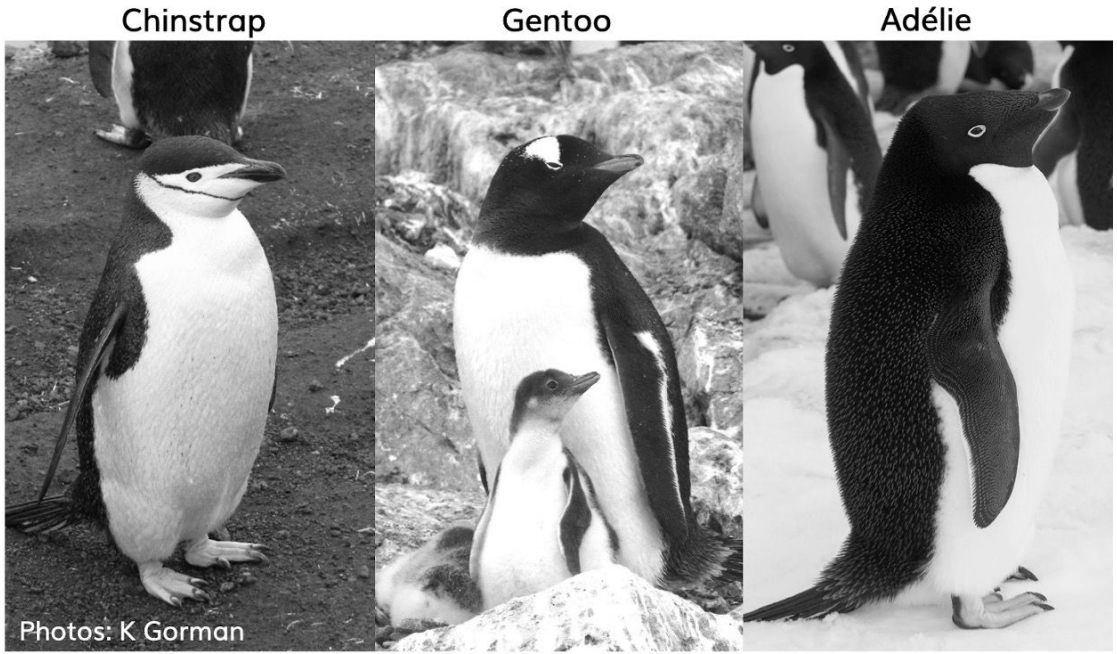
Categories with no specific ordering.

Examples:
- Political Affiliation
- UM C Number

Palmer Penguins



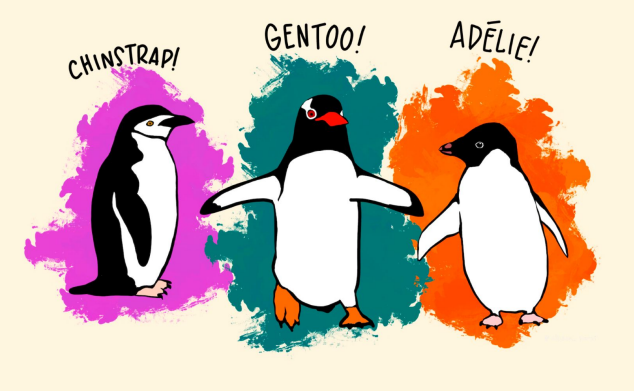


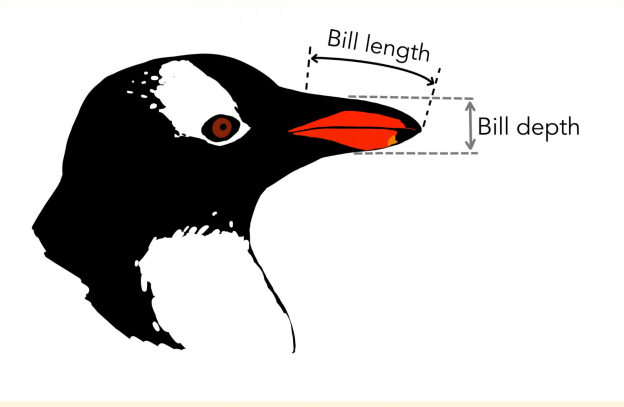


2482612

2482612

Palmer Penguins





Size measurements for adult foraging penguins near Palmer Station, Antarctica

- **Species**: Penguin species
- **Island**: Island in Palmer Archipelago
- **Bill length** (mm)
- **Bill depth** (mm)
- **Flipper length** (mm)
- **Body mass** (g)
- **Sex**
- **Year**: study year

Read more about the data here.

2482612

`penguins` is rectangular data

Records represent individual penguins of same granularity.

Variables

A tibble: 344 × 8

species <fctr>	island <fctr>	bill_length_mm <dbl>	bill_depth_mm <dbl>	flipper_length_mm <int>	body_mass_g <int>	sex <fctr>	year <int>
Adelie	Torgersen	39.1	18.7	181	3750	male	2007
Adelie	Torgersen	39.5	17.4	186	3800	female	2007
Adelie	Torgersen	40.3	18.0	195	3250	female	2007
Adelie	Torgersen	*NA*	*NA*	*NA*	*NA*	*NA*	2007
Adelie	Torgersen	36.7	19.3	193	3450	female	2007
Adelie	Torgersen	39.3	20.6	190	3650	male	2007
Adelie	Torgersen	38.9	17.8	181	3625	female	2007
Adelie	Torgersen	39.2	19.6	195	4675	male	2007
Adelie	Torgersen	34.1	18.1	193	3475	*NA*	2007
Adelie	Torgersen	42.0	20.2	190	4250	*NA*	2007

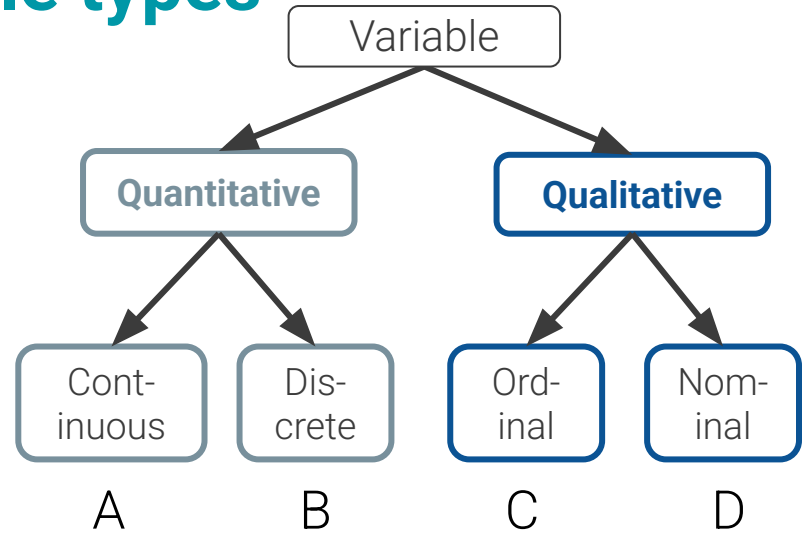
1–10 of 344 rows

Previous 1 2 3 4 5 6 ... 35 Next

2251603

Review: `penguins` and variable types

Q	Variable	Data Type
1	species	**D. Qualitative Nominal**
2	island	**D. Qualitative Nominal**
3	bill_length_mm	**A. Quantitative Cont.**
4	body_mass_g	**B. Quantitative Discrete**
5	sex	**D. Qualitative Nominal**
6	year	**C. Qualitative Ordinal**



These examples show how “shaggy” these categories are!
We will revisit variable types in our data visualization unit.