Missing data

by Sina Rüeger

Notes & Material

↓

https://github.com/sinarueeger/teaching/tree/master/missing_data

Aim of this video

Discover the different facets of missing data by learning about

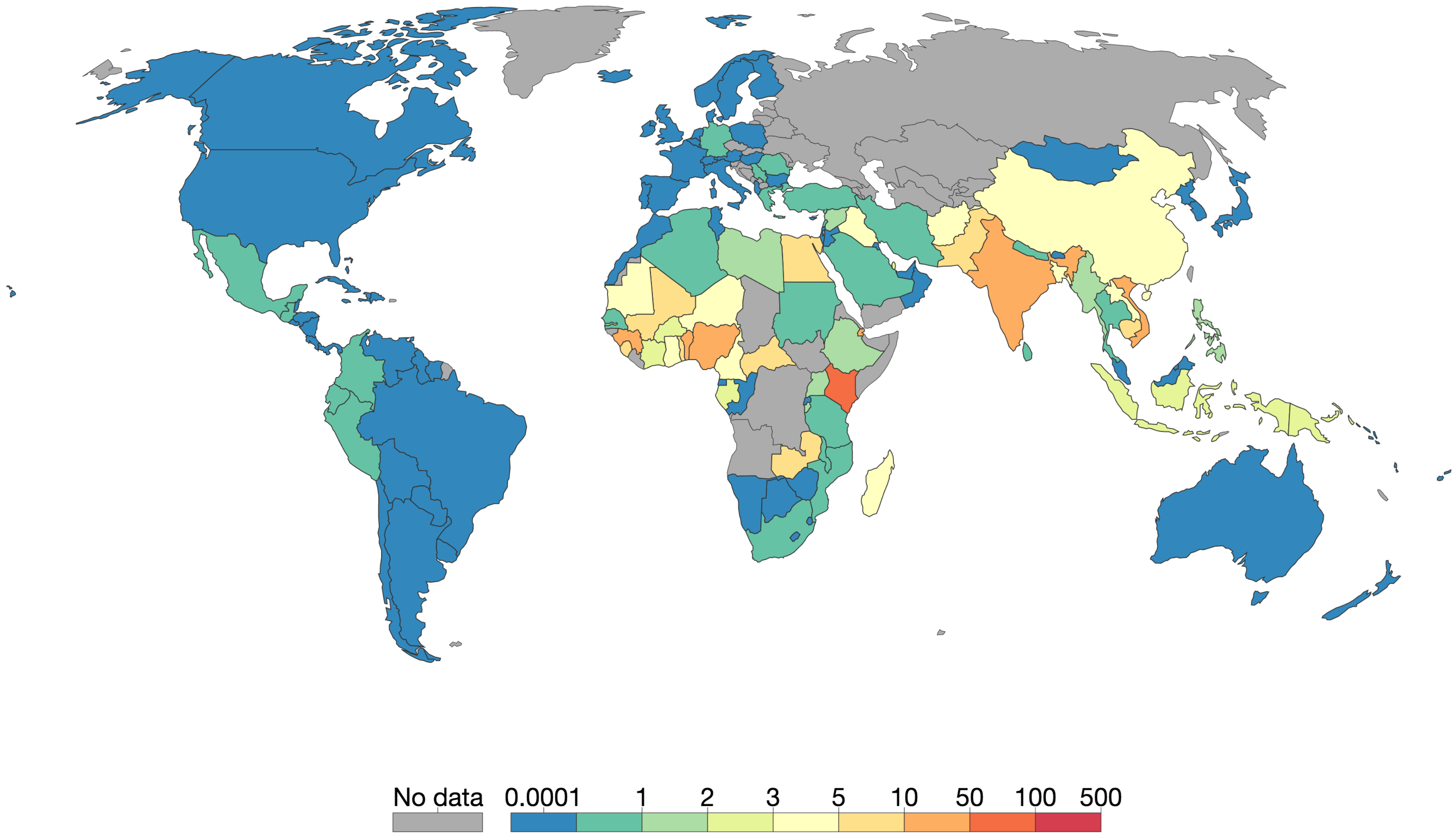
- why missing data occurs
- why missing data can be a problem
- how missing data can be turned in something meaningful

Example "Polio"

Reported paralytic polio cases (per 1 million people), 1990

This includes the wild and vaccine (VAPP) type poliovirus (occurring indigenously and imported)

Our World in Data



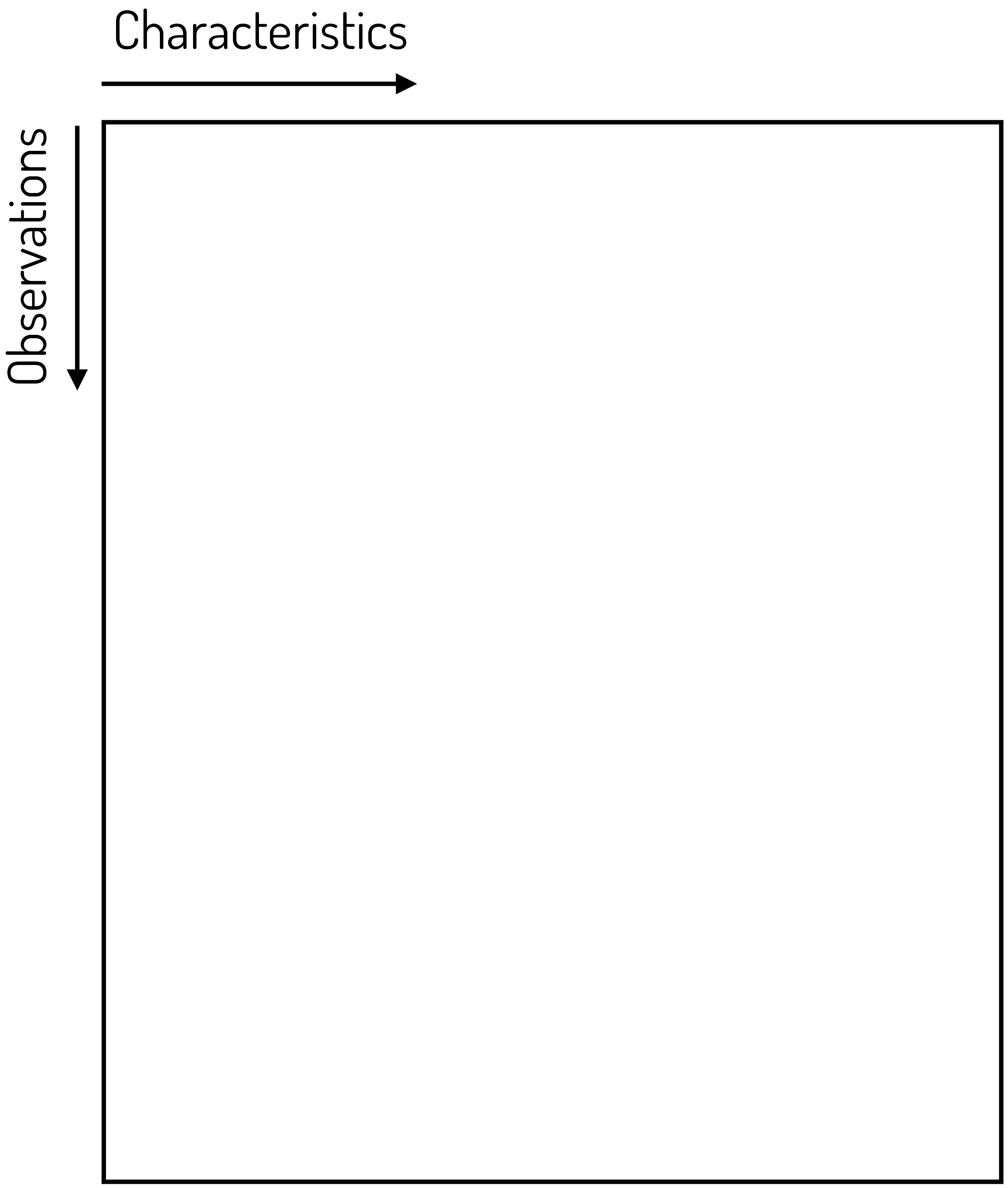
Source: Incidence Report of Selected Vaccine Preventable Diseases (VPDs) - WHO (2017)

OurWorldInData.org/polio/ • CC BY-SA

https://ourworldindata.org/grapher/reported-paralytic-polio-cases-per-1-million-people

Recap: data

- One observation per row
- One characteristics measured per column



Characteristics

Observations

Recap: data

Characteristics

Observations

Observation	Planet name	Surface [km2]	Discovered [dd-month-yyyy]
1	Pluto	1.779×10^{7}	18-Feb-1930
2	Uranus	8.116×10^{9}	13-Mar-1781
3	Saturn	4.270×10^{10}	no data

Information from:
https://en.wikipedia.org/wiki/Pluto
https://en.wikipedia.org/wiki/Uranus
https://en.wikipedia.org/wiki/Neptune

Reasons for missing data

Data not reported

Technical issue > badly measured

No measurement possible

Poorly designed surveys

Hair color

□ Black

□ White

When is missing data a problem

When missingness is not occuring randomly, but somehow related to another variable.

Hair color

□ Black

□ Gray

→

1. People with hair color other than black and gray will have missing values.
2. In this case, missing values can have the following origin:
   - Hair color correlates with ancestry.
   - Gray hair is associated with age.
   - Some people have no hair.

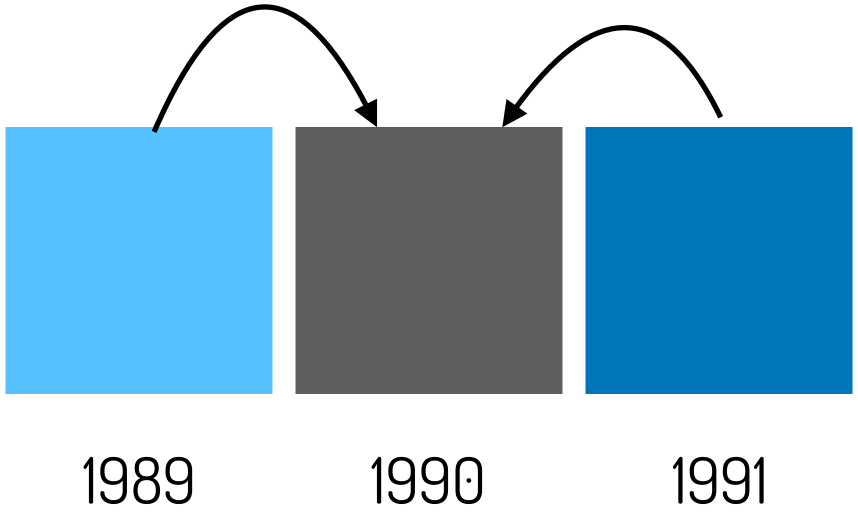
https://en.wikipedia.org/wiki/Human_hair_color

Why is missing data a problem

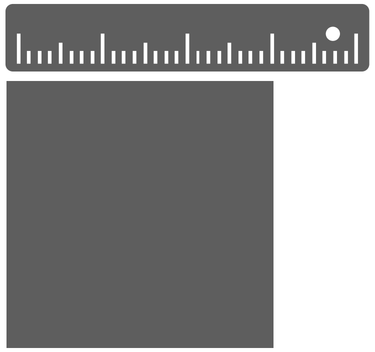
1. Bias
2. Potentially lower sample size

Solutions

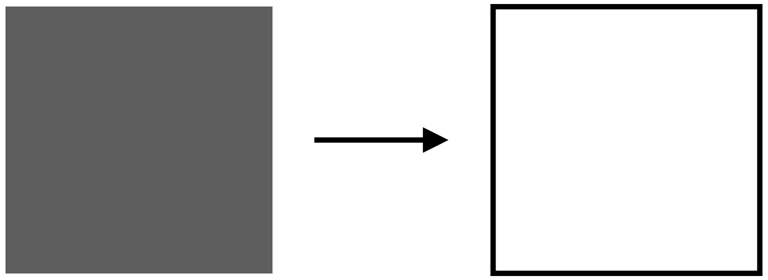
Imputation



Re-measuring



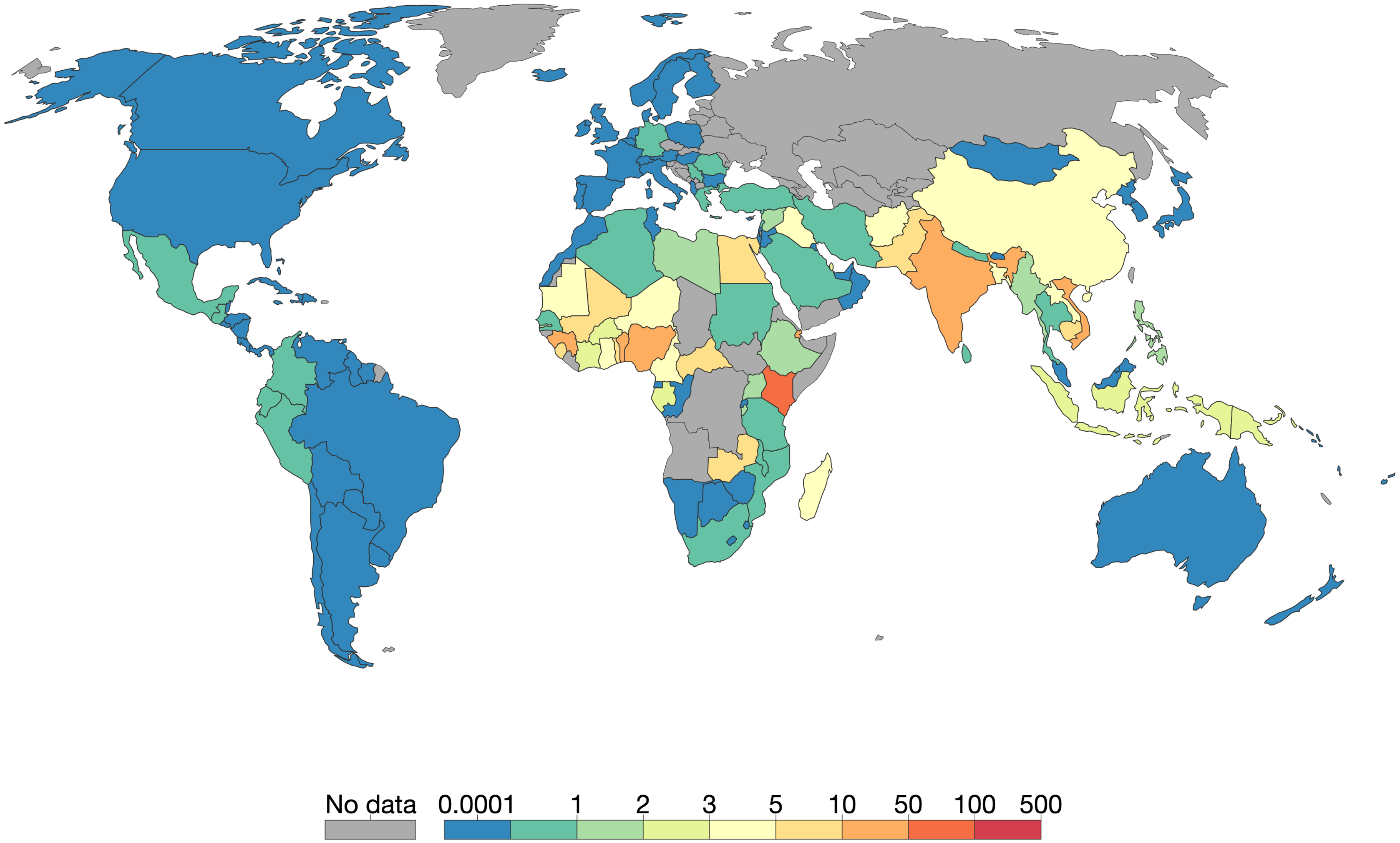
Removal



Example "Polio"

Reported paralytic polio cases (per 1 million people), 1990

This includes the wild and vaccine (VAPP) type poliovirus (occurring indigenously and imported)



Source: Incidence Report of Selected Vaccine Preventable Diseases (VPDs) - WHO (2017)

OurWorldInData.org/polio/ • CC BY-SA

https://ourworldindata.org/polio#data-quality-and-measurement

Summary

✔ why missing data occurs

✔ why missing data can be a problem

✔ how missing data can be turned in something meaningful

Links

Explore human civilisation through the lens of data

https://ourworldindata.org/