Introduction to Data Science

Domitilla Brandoni, CINECA

06/2023

Univerza v Ljubljani

TU WIEN TECHNISCHE UNIVERSITÄT WIEN

CINECA

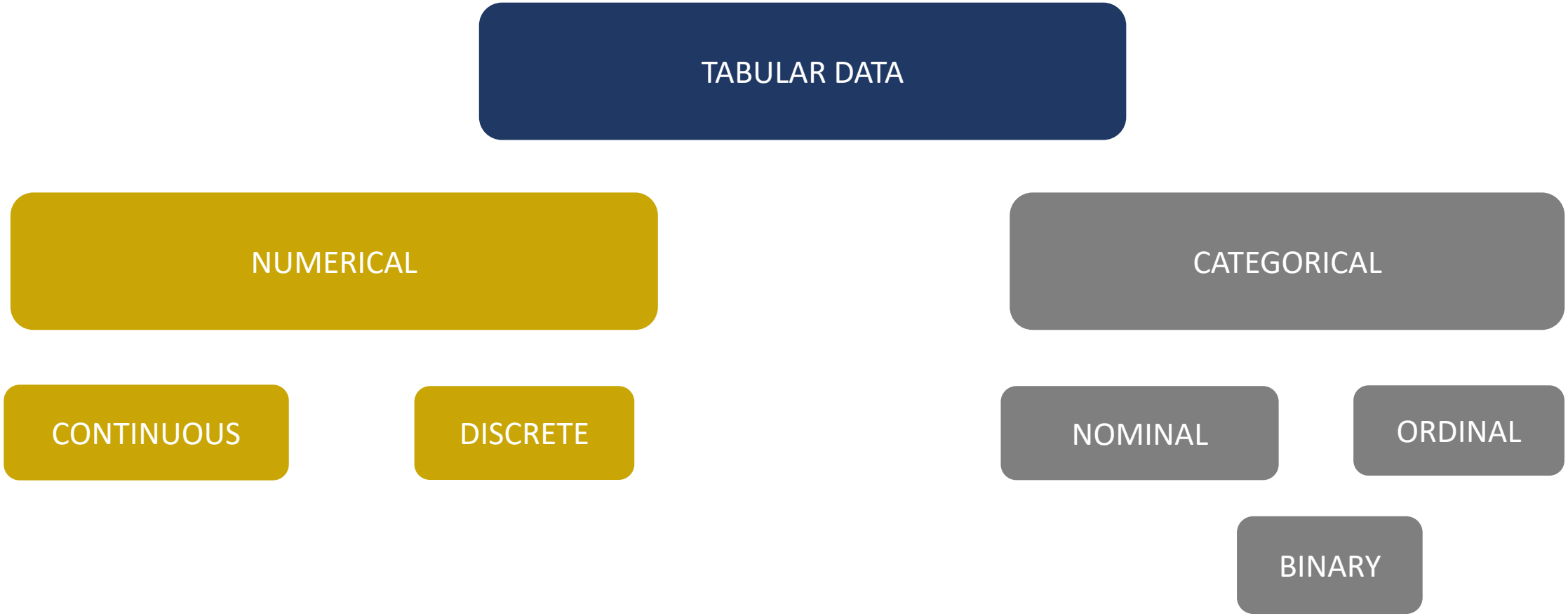
VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

Co-funded by the Erasmus+ Programme of the European Union

This project has been funded with support from the European Commission.
This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

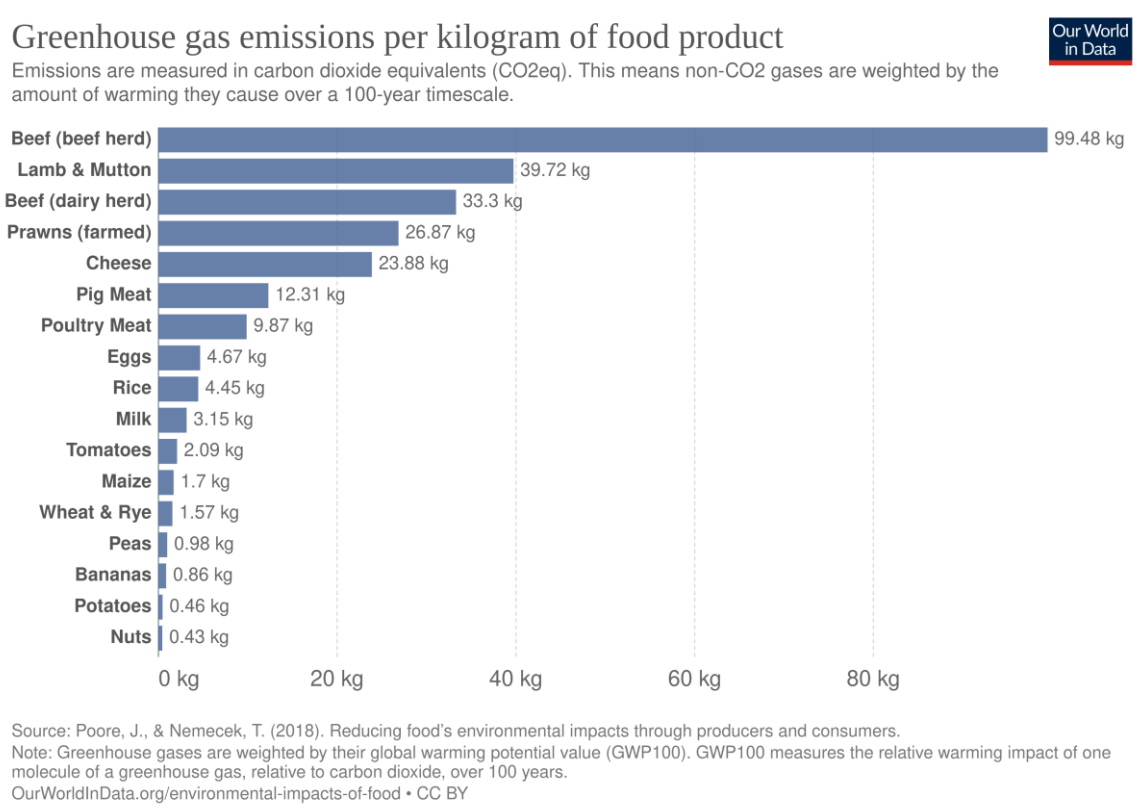
Data formats

Domitilla Brandoni, CINECA

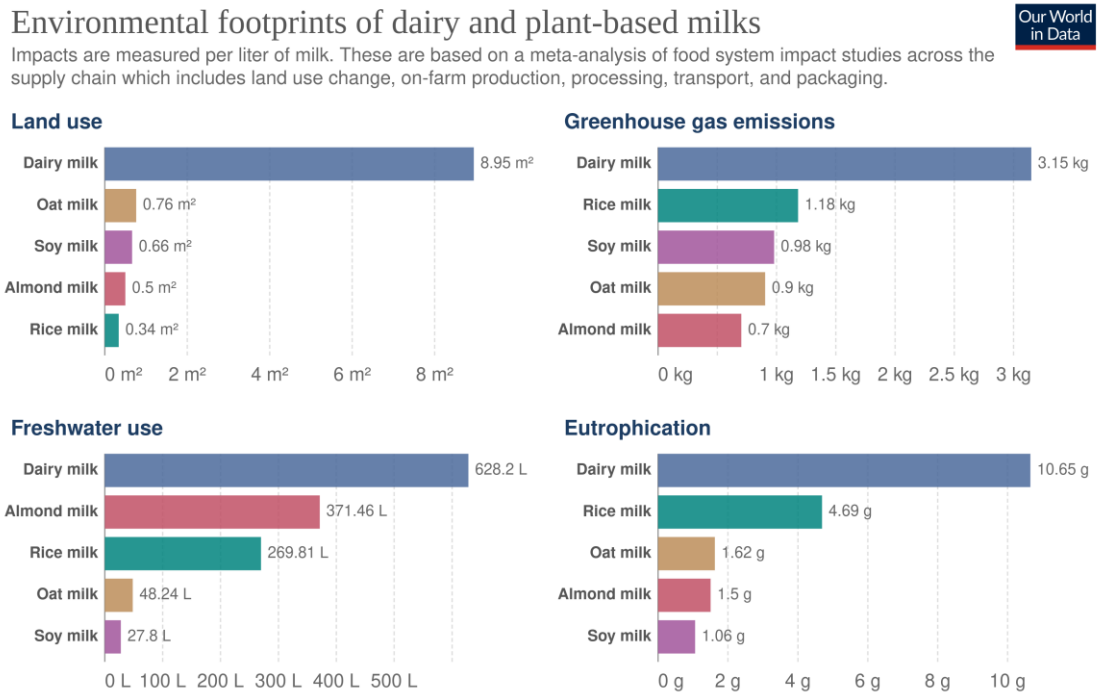


Numerical continuous data

Domitilla Brandoni, CINECA



Source: Poore, J., & Nemecek, T. (2018). Reducing food's environmental impacts through producers and consumers.
Note: Greenhouse gases are weighted by their global warming potential value (GWP100). GWP100 measures the relative warming impact of one molecule of a greenhouse gas, relative to carbon dioxide, over 100 years.
OurWorldInData.org/environmental-impacts-of-food • CC BY

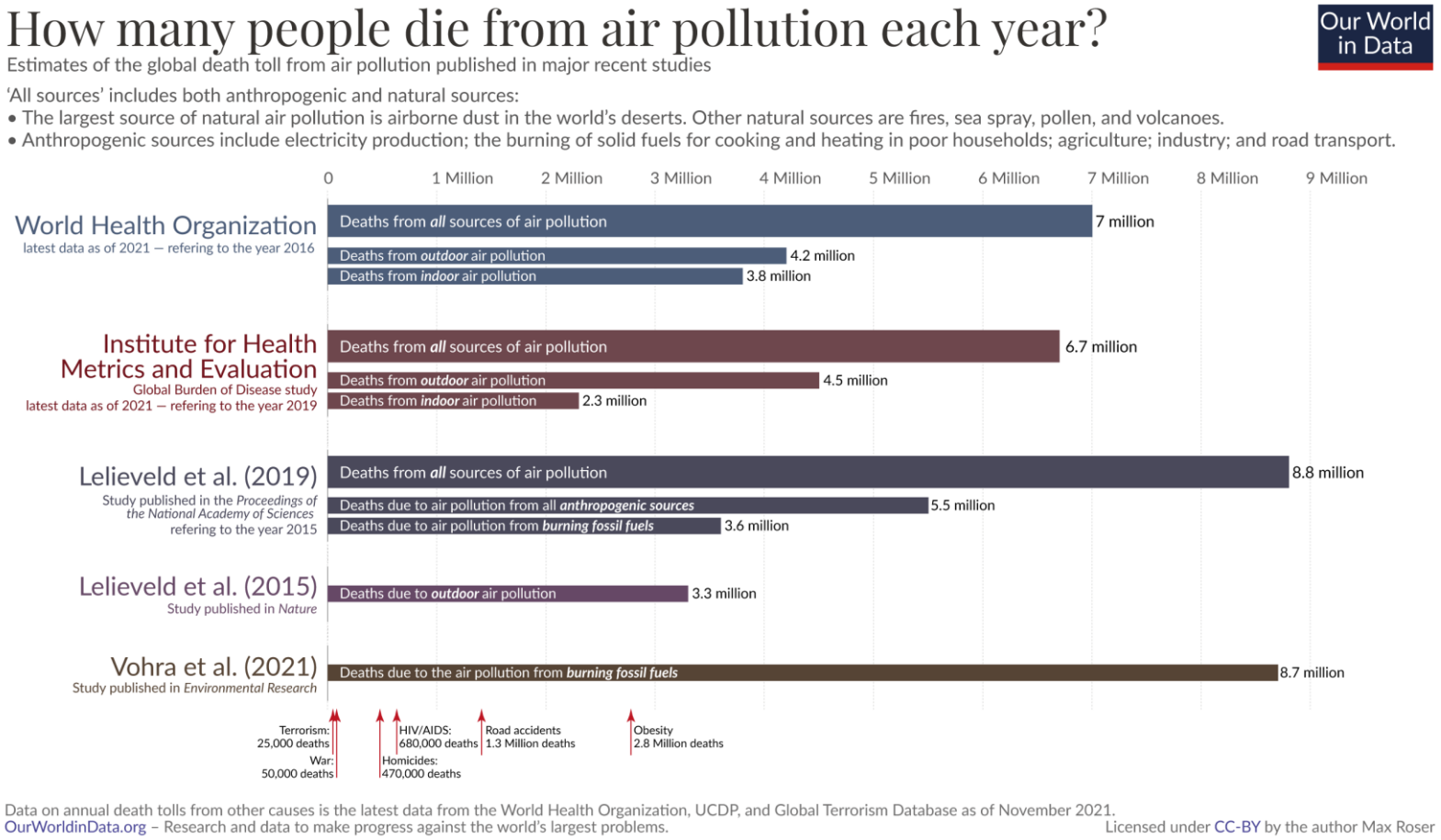


Source: Poore, J., & Nemecek, T. (2018). Reducing food's environmental impacts through producers and consumers. Science.
OurWorldInData.org/environmental-impacts-of-food • CC BY

https://ourworldindata.org/environmental-impacts-of-food

Numerical discrete data

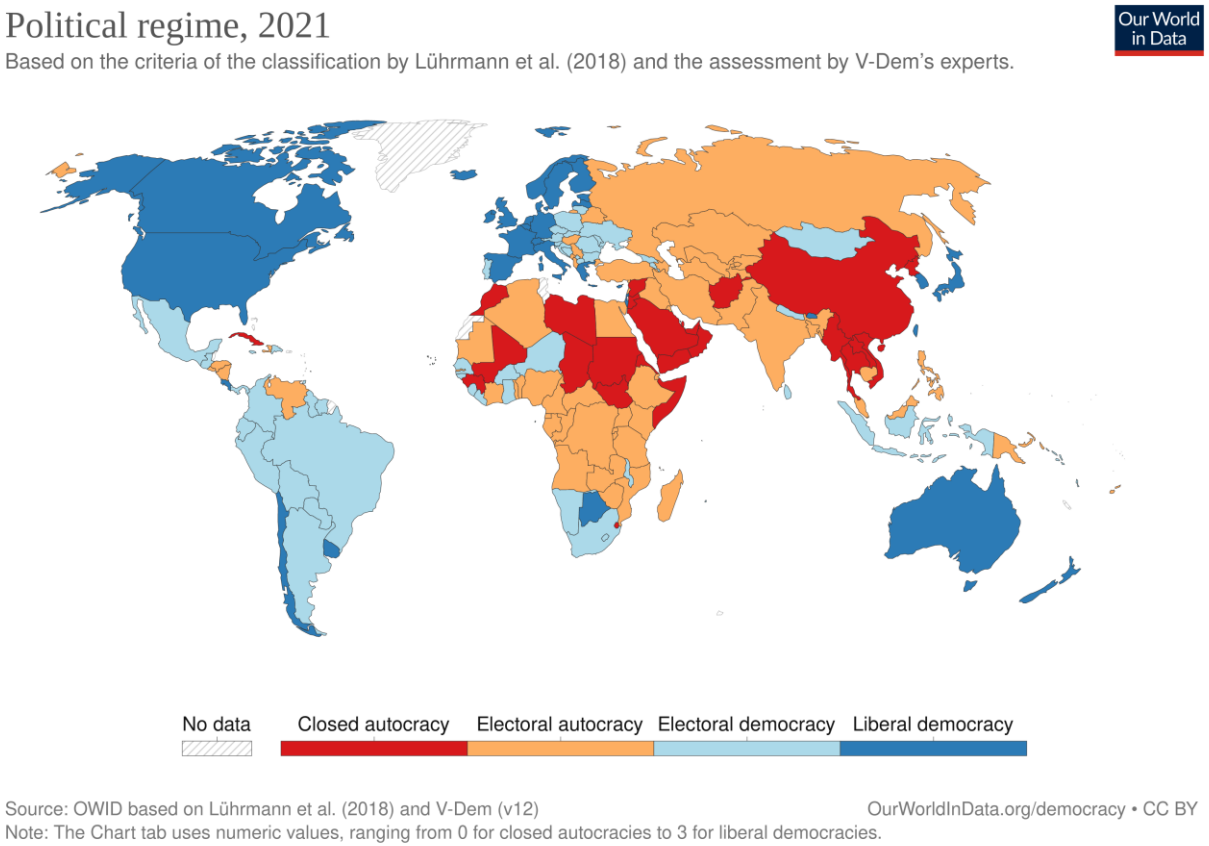
Domitilla Brandoni, CINECA



https://ourworldindata.org/data-review-air-pollution-deaths

Nominal data

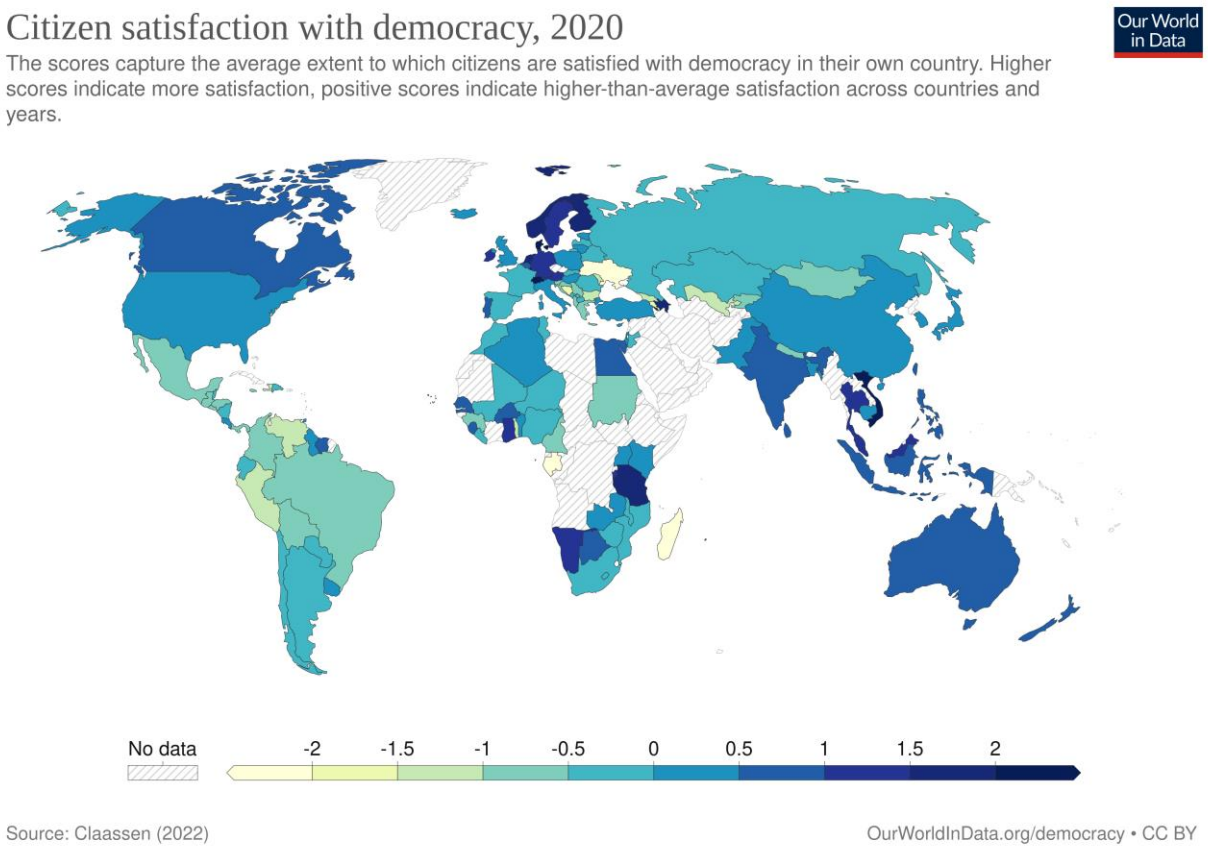
Domitilla Brandoni, CINECA



https://ourworldindata.org/democracy

Ordinal data

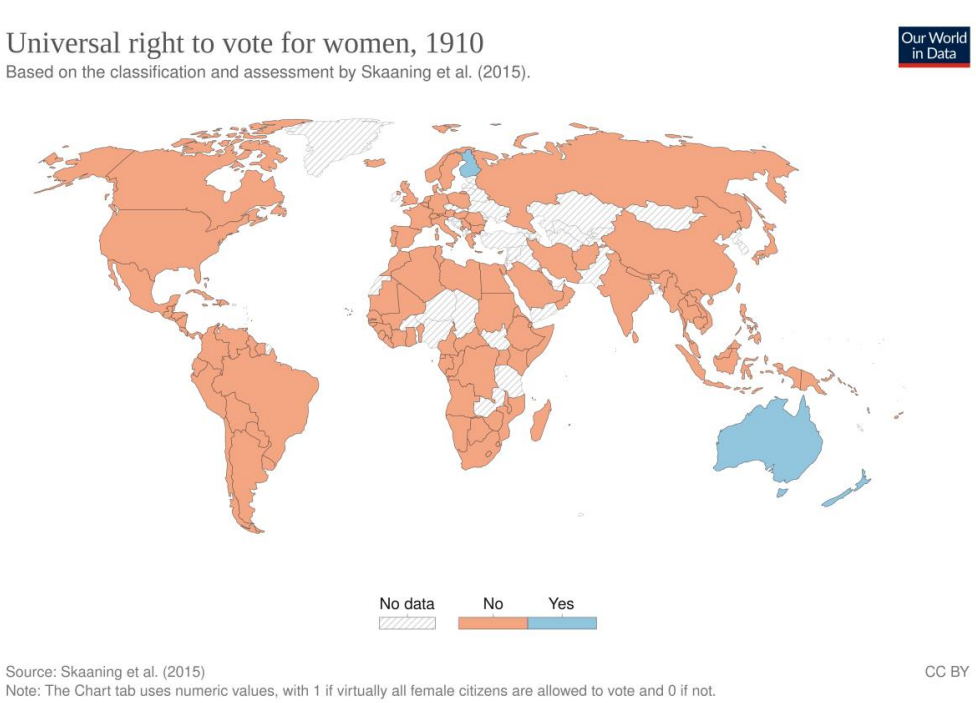
Domitilla Brandoni, CINECA

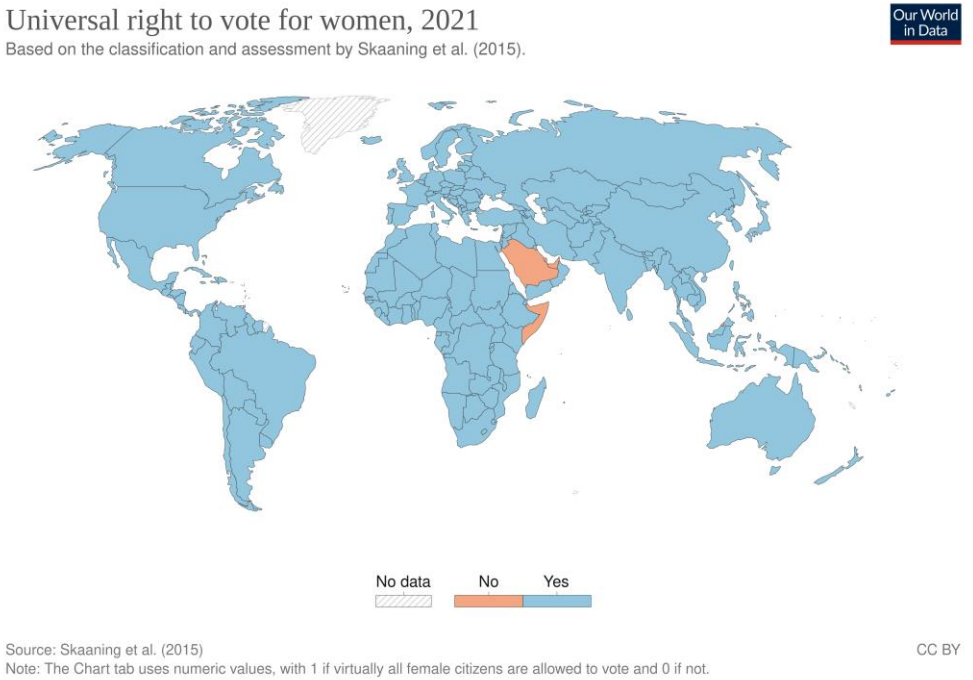


https://ourworldindata.org/influenza-deaths

Binary data

Domitilla Brandoni, CINECA

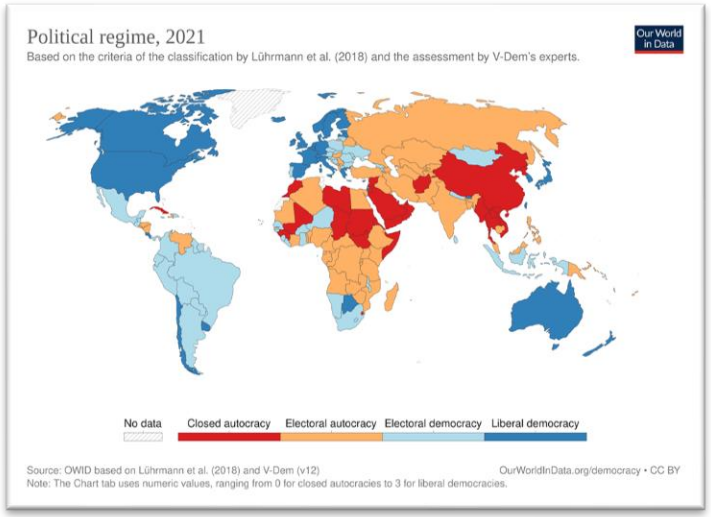


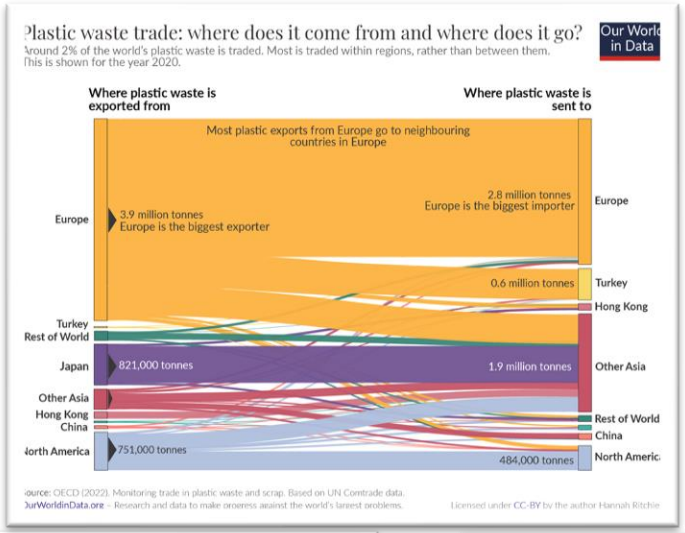


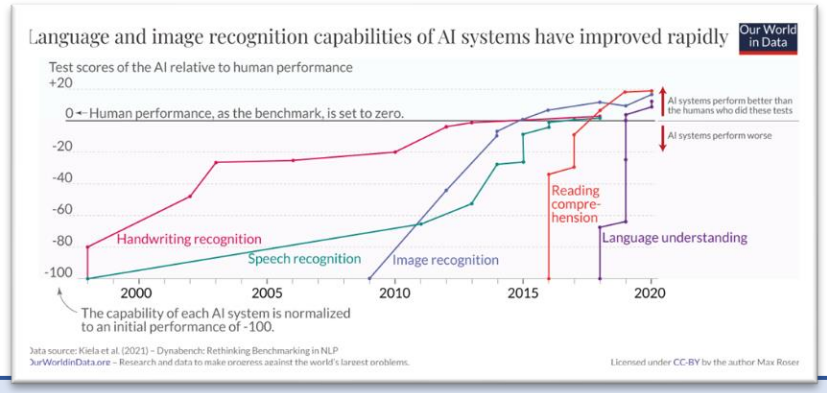
https://ourworldindata.org/democracy

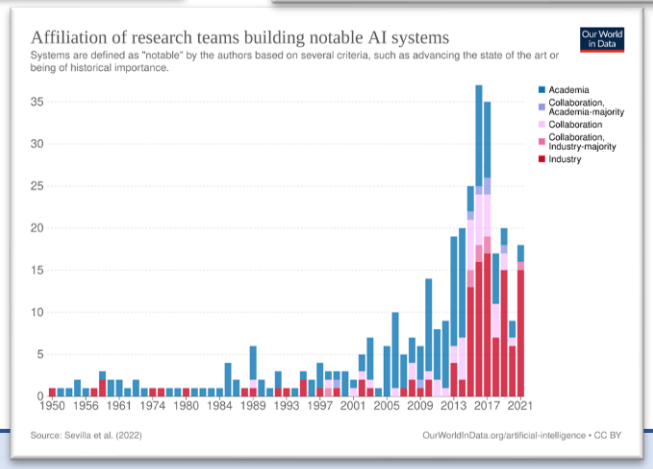
Each format has its own visualization

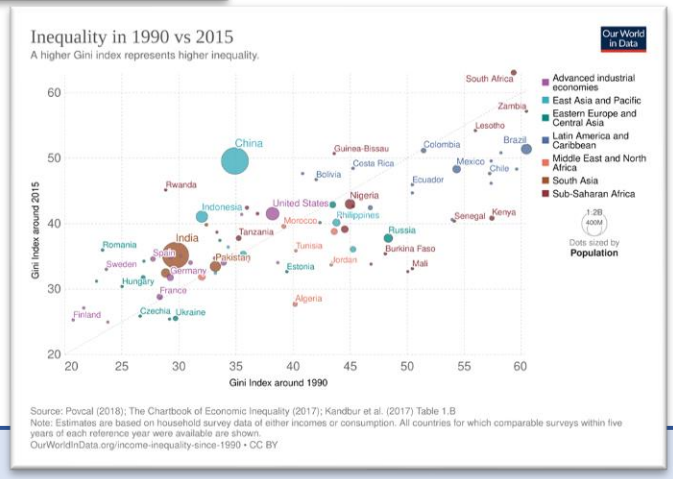
Domitilla Brandoni, CINECA







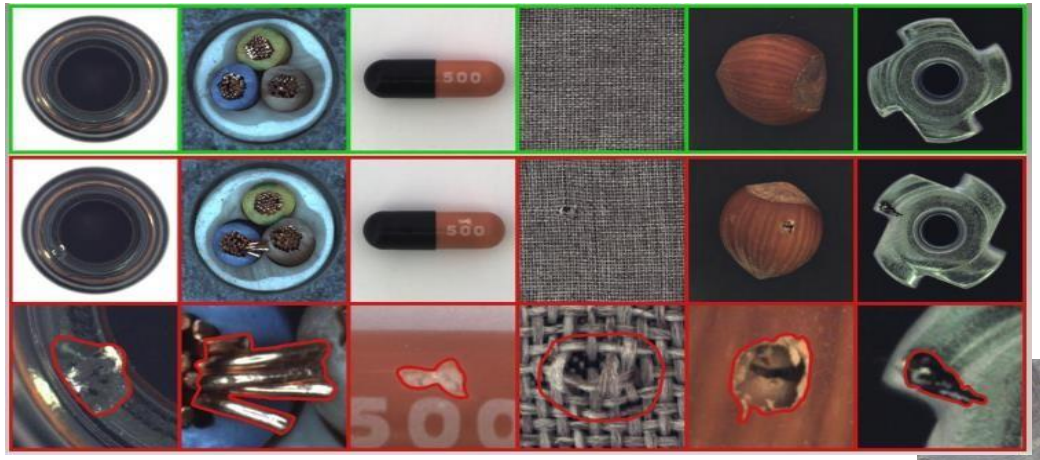




https://ourworldindata.org/

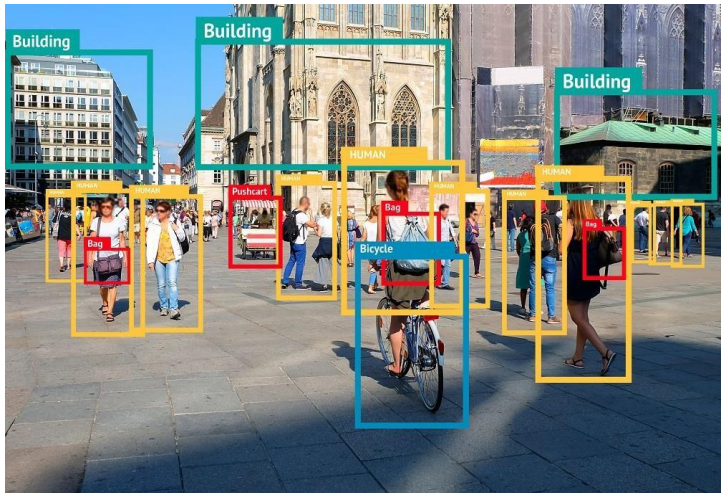
Non tabular data: images

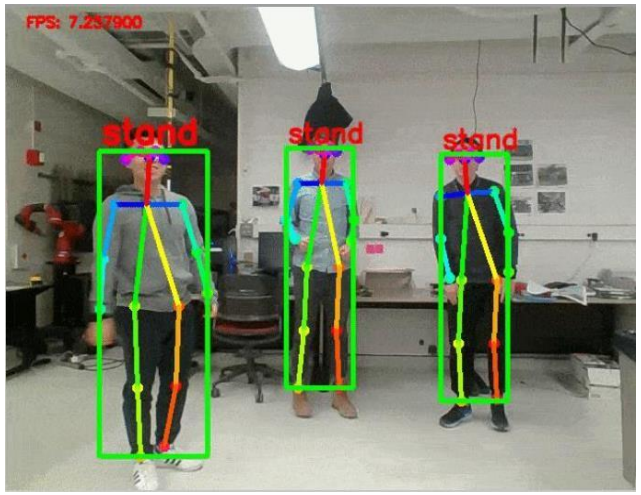
Domitilla Brandoni, CINECA

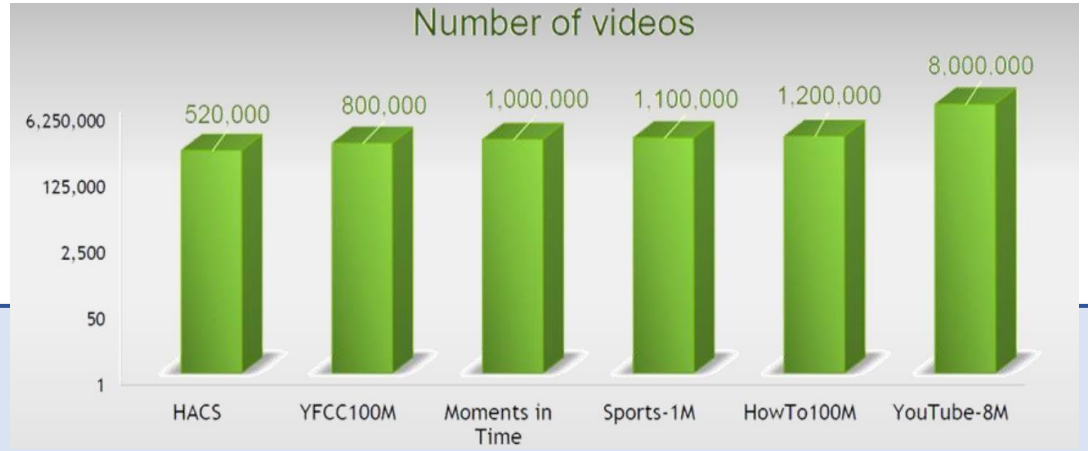


Non tabular data: videos

Domitilla Brandoni, CINECA

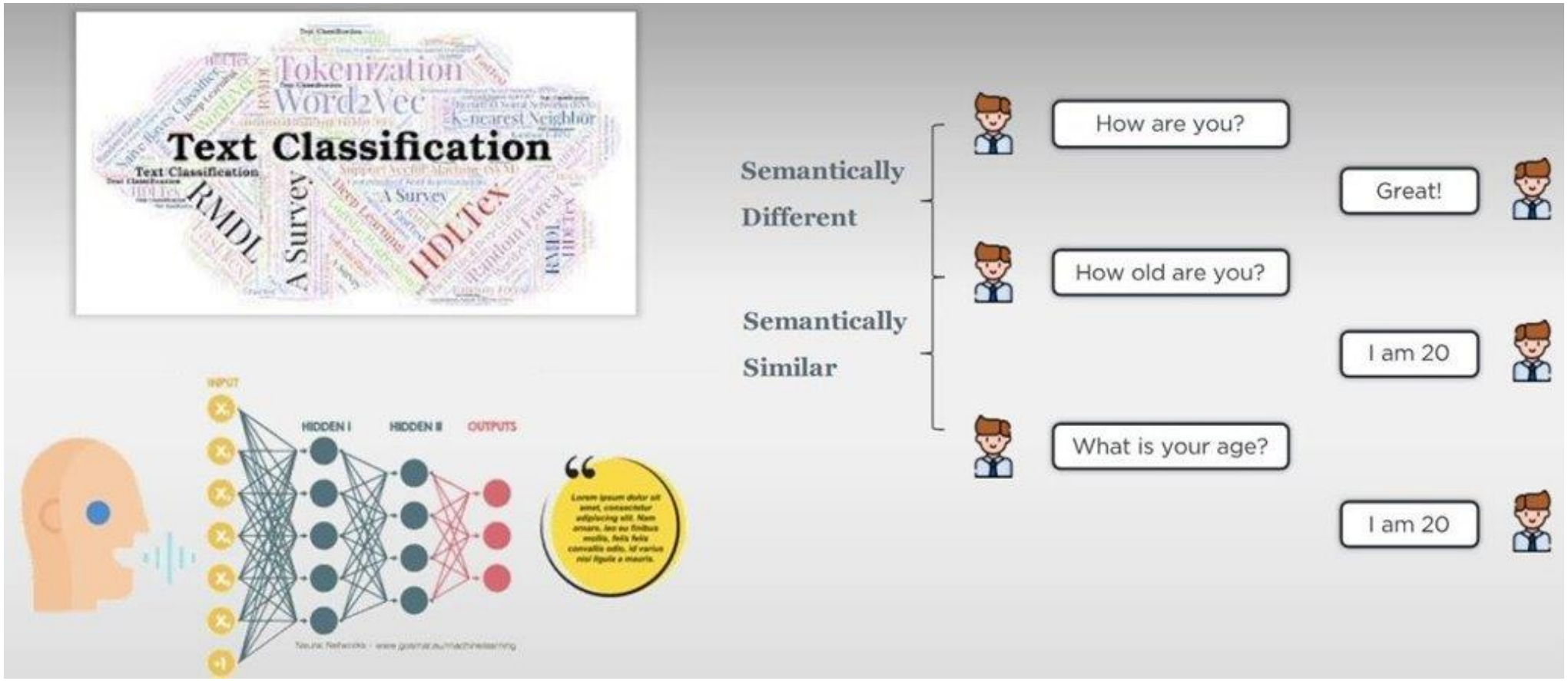






Non tabular data: text & audio

Domitilla Brandoni, CINECA



Where to find GOOD data

Domitilla Brandoni, CINECA

https://ourworldindata.org/

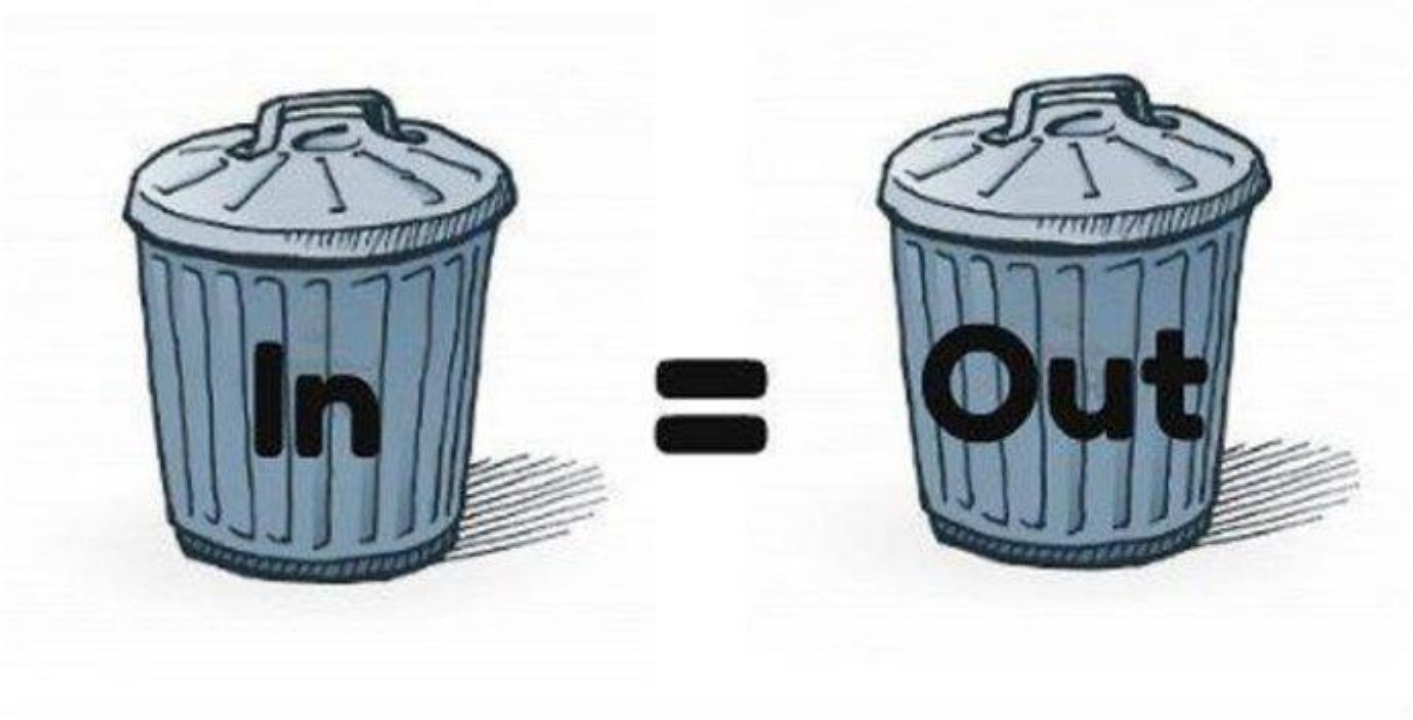
https://www.kaggle.com/

https://paperswithcode.com/datasets

first_steps_with_data.ipynb

Garbage in, garbage out

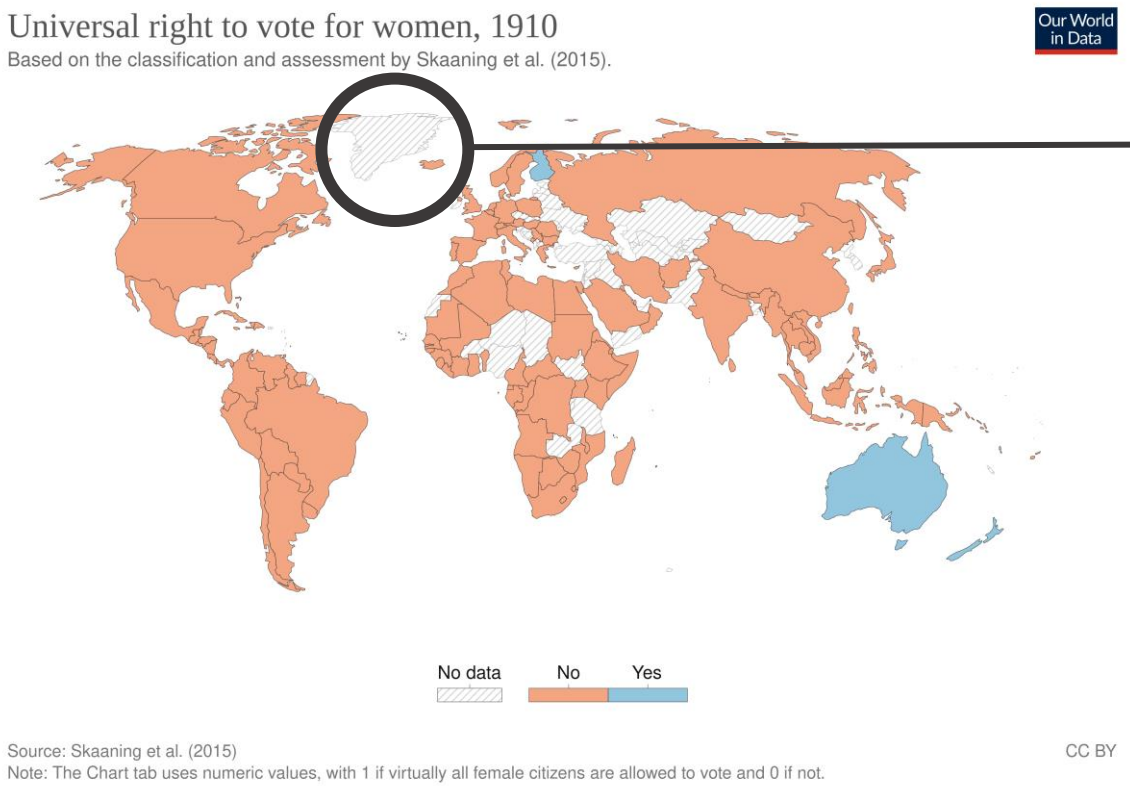
Domitilla Brandoni, CINECA



https://candysdirt.com/2016/05/27/property-taxes-garbage-garbage/

Missing data

Domitilla Brandoni, CINECA



Missing data

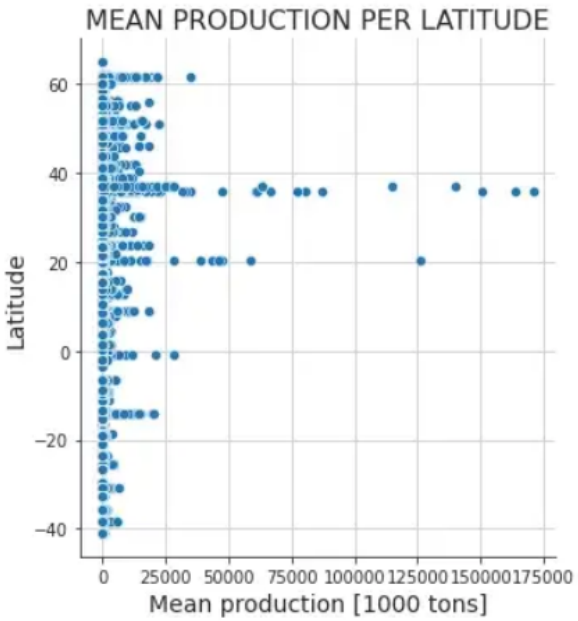
https://ourworldindata.org/democracy

Outliers

Domitilla Brandoni, CINECA

Outliers: data points **significantly** different from the others.

GRAPHICAL APPROACH



Z SCORE

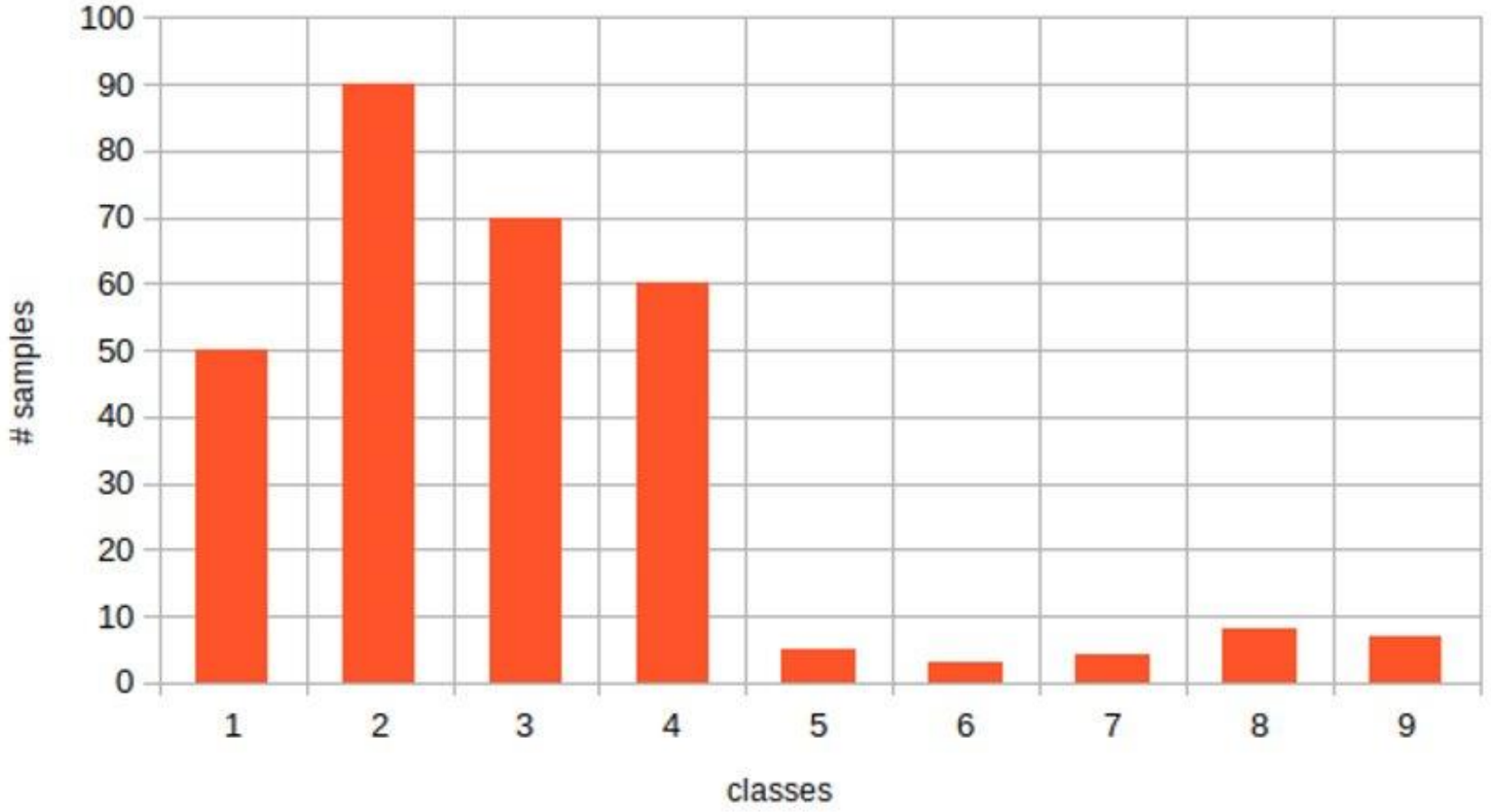
$Z=(x-m)/s$

- m=mean
- s=standard deviation
- x=data point

https://towardsdatascience.com/how-to-detect-outliers-in-a-data-science-project-17f39653fb17

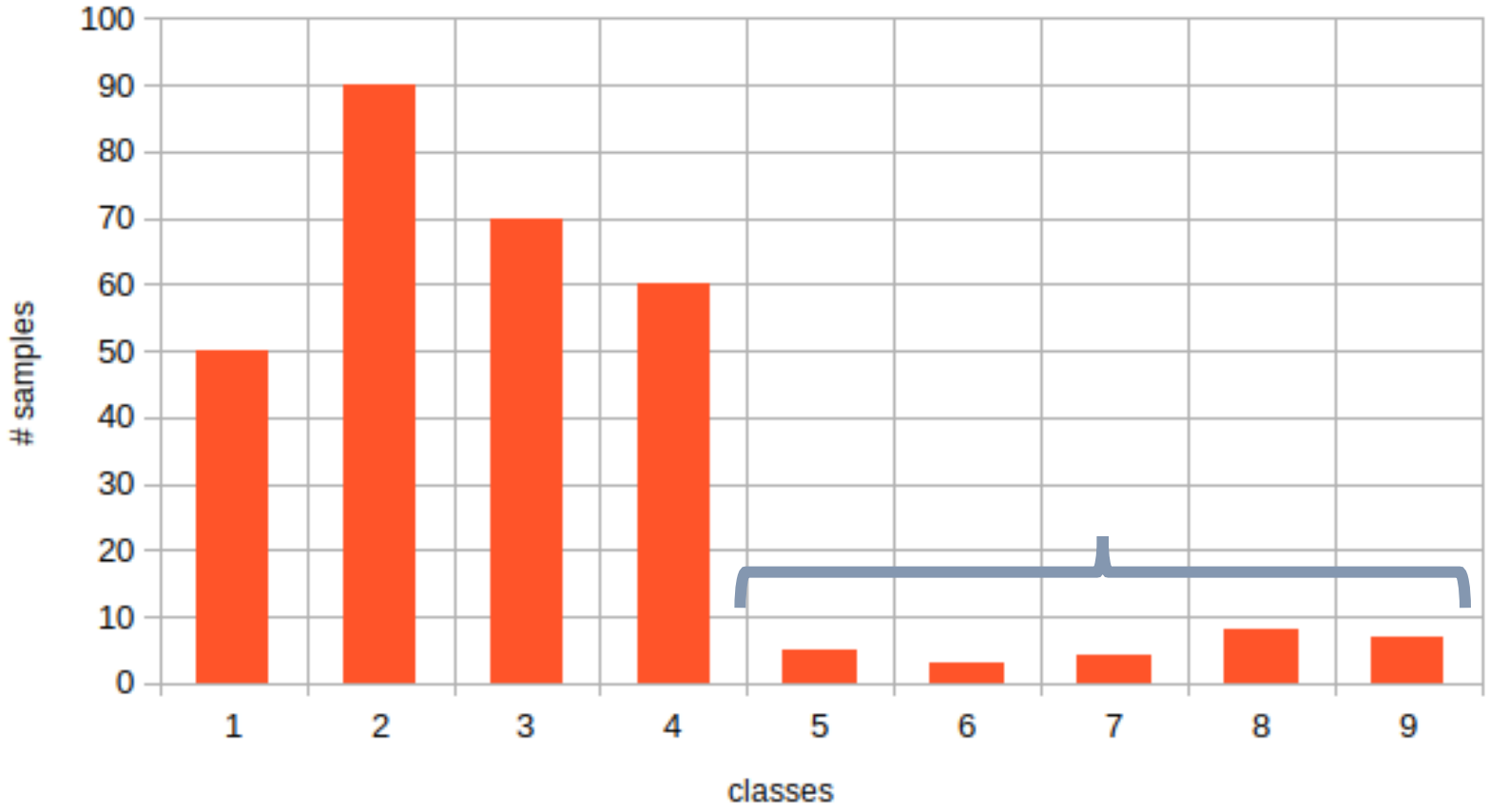
Unbalanced data

Domitilla Brandoni, CINECA



Unbalanced data

Domitilla Brandoni, CINECA



Rare events

Domitilla Brandoni, CINECA

(A) **Cow: 0.99**, Pasture: 0.99, Grass: 0.99, No Person: 0.98, Mammal: 0.98

(B) No Person: 0.99, Water: 0.98, Beach: 0.97, Outdoors: 0.97, Seashore: 0.97

(C) No Person: 0.97, **Mammal: 0.96**, Water: 0.94, Beach: 0.94, Two: 0.94

Fig. 1. Recognition algorithms generalize poorly to new environments. Cows in 'common' contexts (e.g. Alpine pastures) are detected and classified correctly (A), while cows in uncommon contexts (beach, waves and boat) are not detected (B) or classified poorly (C). Top five labels and confidence produced by ClarifAI.com shown.

Wrong labels

Domitilla Brandoni, CINECA

	snake	cat	dog	spider
snake	50	0	0	0
cat	0	40	20	0
dog	0	10	30	0
spider	0	0	0	50

Wrong labels

Domitilla Brandoni, CINECA

	snake	mammal	spider
snake	50	0	0
mammal	0	50	0
spider	0	0	50

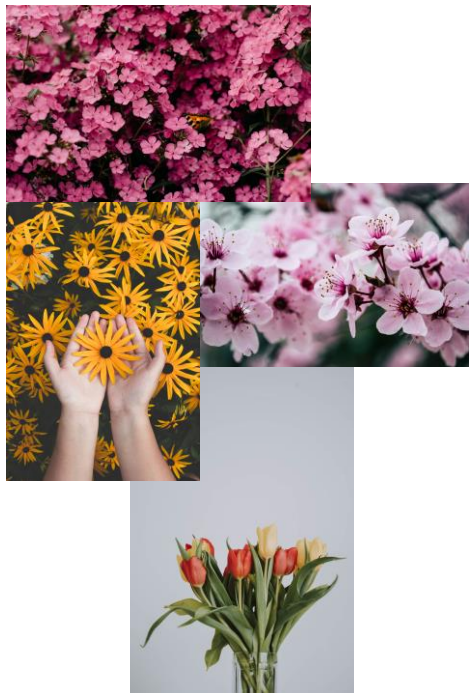
Numerical continuois data

Domitilla Brandoni, CINECA

Wikipedia

Any problems?

Domitilla Brandoni, CINECA



MIT News
ON CAMPUS AND AROUND THE WORLD

SUBSCRIBE BROWSE SEARCH NEWS

Study finds gender and skin-type bias in commercial artificial-intelligence systems

Examination of facial-analysis software shows error rate of 0.8 percent for light-skinned men, 34.7 percent for dark-skinned women.

Watch Video

Larry Hardesty | MIT News Office
February 11, 2018

PRESS INQUIRIES

Joy Buolamwini, a researcher in the MIT Media Lab's Civic Media group

Photo: Bryce Vickmark

https://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212

Exercise

Domitilla Brandoni, CINECA

Same procedure of the hands-on session but with another dataset

- Download the dataset
- Look at the variables
- Plot some values
- Explore the data

Thank you for your attention!

http://sctrain.eu/

Co-funded by the Erasmus+ Programme of the European Union

This project has been funded with support from the European Commission.
This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.