

# Introduction to Data Science

Domitilla Brandoni, CINECA

06/2023

Univerza v Ljubljani



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.

TABULAR DATA

NUMERICAL

CONTINUOUS

DISCRETE

CATEGORICAL

NOMINAL

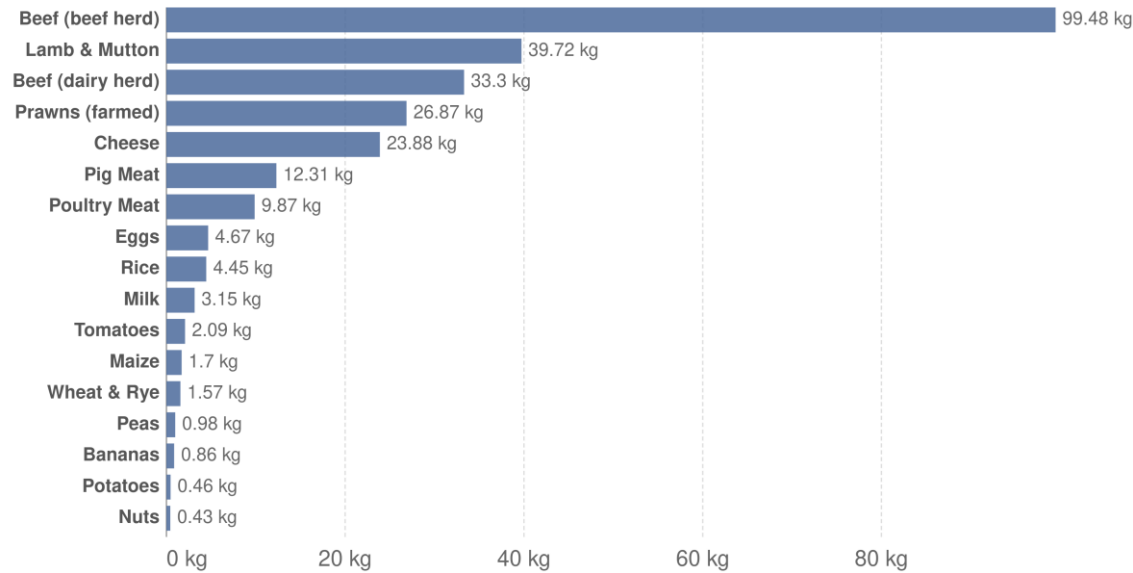
ORDINAL

BINARY

## Greenhouse gas emissions per kilogram of food product

Emissions are measured in carbon dioxide equivalents (CO<sub>2</sub>eq). This means non-CO<sub>2</sub> gases are weighted by the amount of warming they cause over a 100-year timescale.

Our World in Data

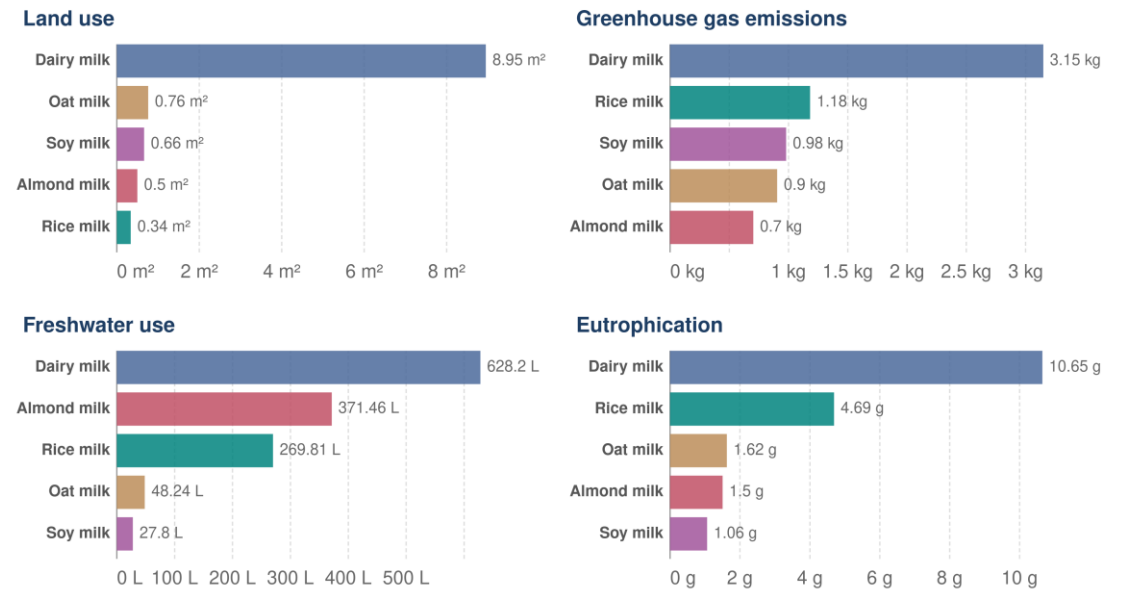


Source: Poore, J., & Nemecek, T. (2018). Reducing food's environmental impacts through producers and consumers. Science.  
 Note: Greenhouse gases are weighted by their global warming potential value (GWP100). GWP100 measures the relative warming impact of one molecule of a greenhouse gas, relative to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

## Environmental footprints of dairy and plant-based milks

Impacts are measured per liter of milk. These are based on a meta-analysis of food system impact studies across the supply chain which includes land use change, on-farm production, processing, transport, and packaging.

Our World in Data



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

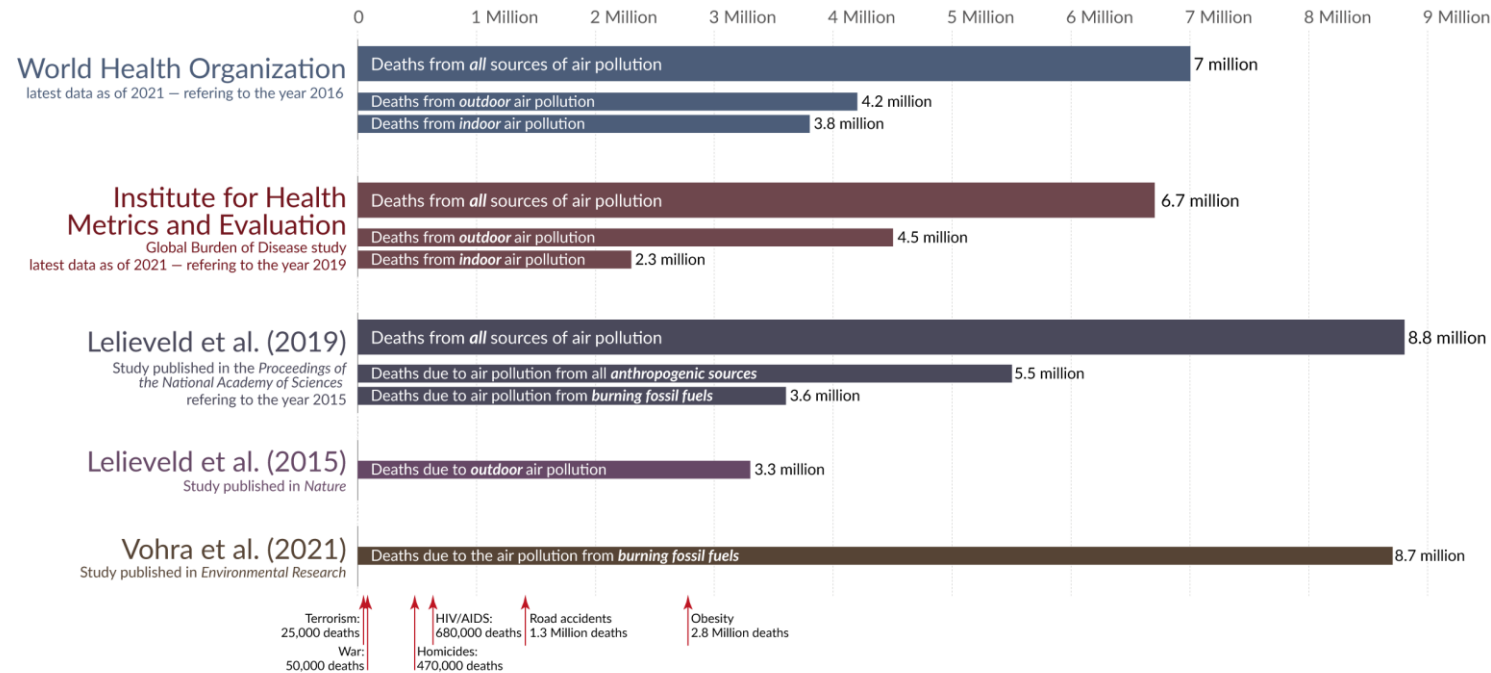
## How many people die from air pollution each year?



Estimates of the global death toll from air pollution published in major recent studies

'All sources' includes both anthropogenic and natural sources:

- The largest source of natural air pollution is airborne dust in the world's deserts. Other natural sources are fires, sea spray, pollen, and volcanoes.
- Anthropogenic sources include electricity production; the burning of solid fuels for cooking and heating in poor households; agriculture; industry; and road transport.



Data on annual death tolls from other causes is the latest data from the World Health Organization, UCDP, and Global Terrorism Database as of November 2021.

OurWorldinData.org – Research and data to make progress against the world's largest problems.

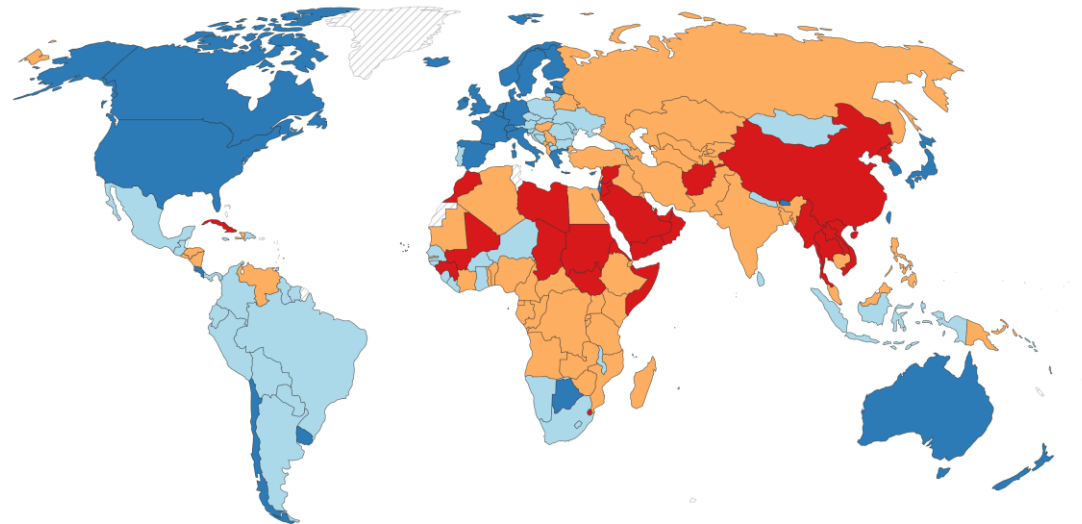
Licensed under CC-BY by the author Max Roser

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

## Political regime, 2021

Based on the criteria of the classification by Lührmann et al. (2018) and the assessment by V-Dem's experts.

Our World  
in Data



No data Closed autocracy Electoral autocracy Electoral democracy Liberal democracy

Source: OWID based on Lührmann et al. (2018) and V-Dem (v12)

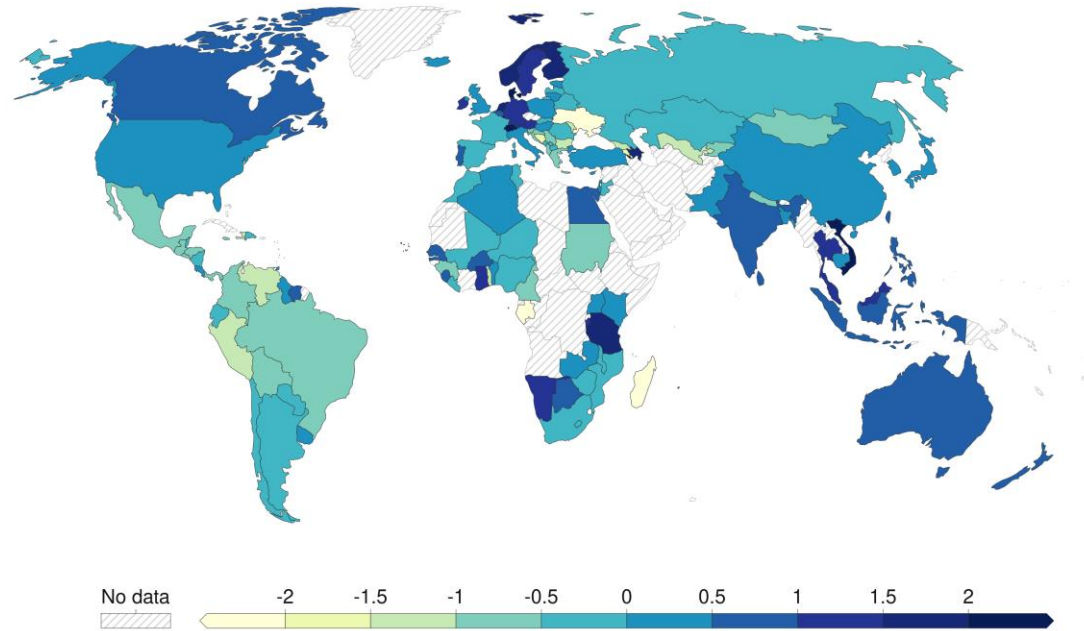
OurWorldInData.org/democracy • CC BY

Note: The Chart tab uses numeric values, ranging from 0 for closed autocracies to 3 for liberal democracies.

<https://ourworldindata.org/democracy>

## Citizen satisfaction with democracy, 2020

The scores capture the average extent to which citizens are satisfied with democracy in their own country. Higher scores indicate more satisfaction, positive scores indicate higher-than-average satisfaction across countries and years.



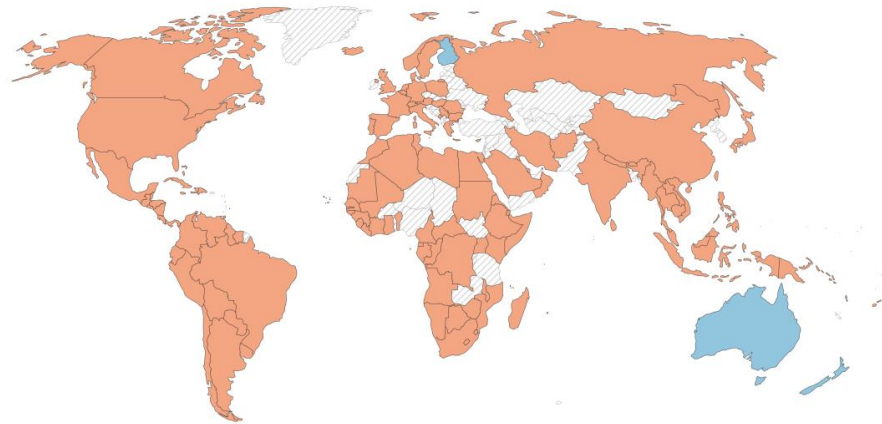
Source: Claassen (2022)

OurWorldInData.org/democracy • CC BY

## Universal right to vote for women, 1910

Based on the classification and assessment by Skaaning et al. (2015).

Our World  
in Data



No data No Yes

Source: Skaaning et al. (2015)

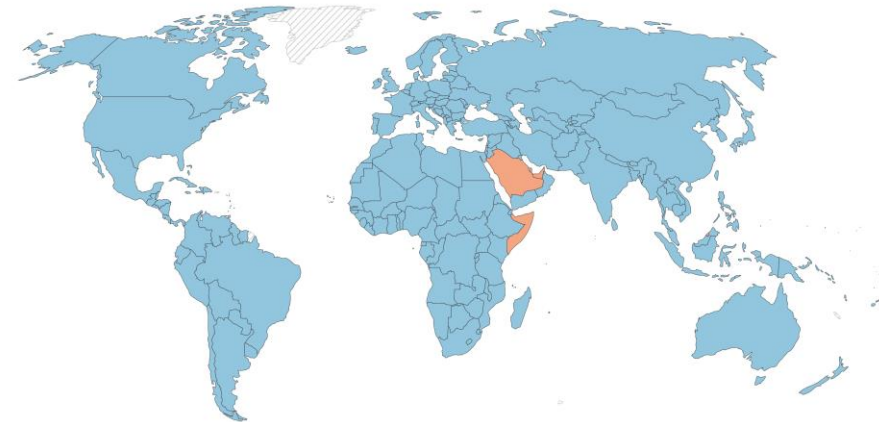
Note: The Chart tab uses numeric values, with 1 if virtually all female citizens are allowed to vote and 0 if not.

CC BY

## Universal right to vote for women, 2021

Based on the classification and assessment by Skaaning et al. (2015).

Our World  
in Data



No data No Yes

Source: Skaaning et al. (2015)

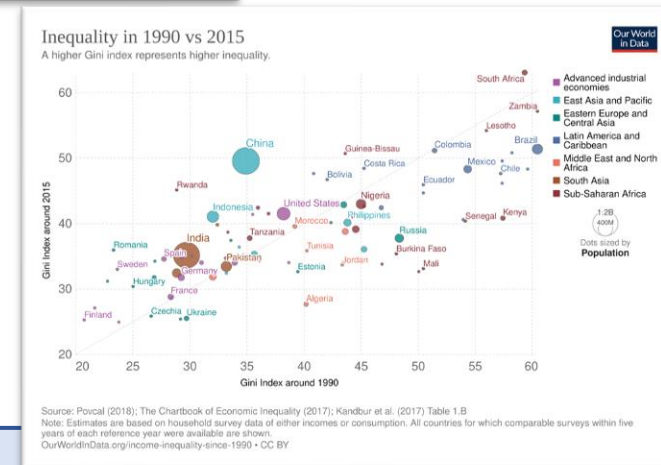
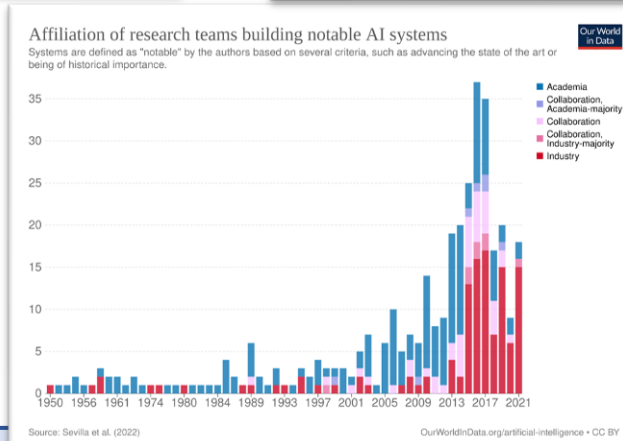
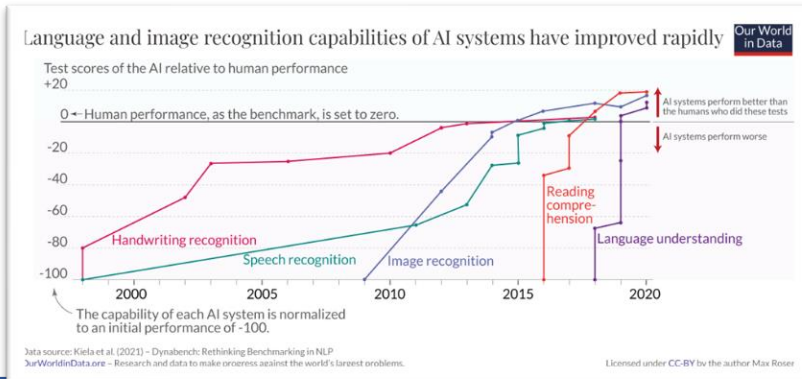
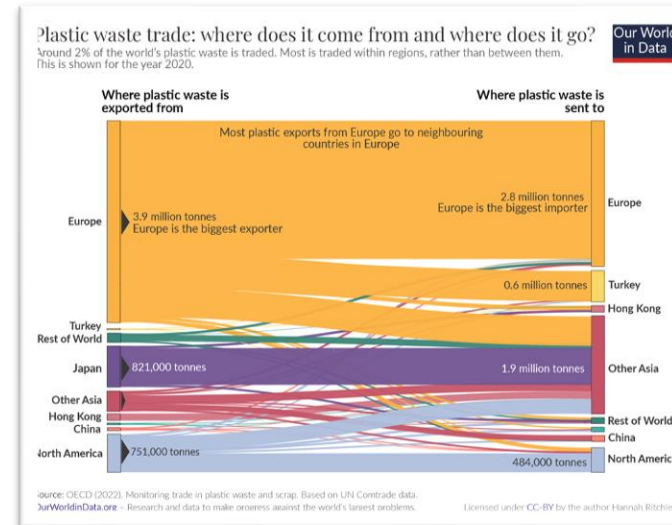
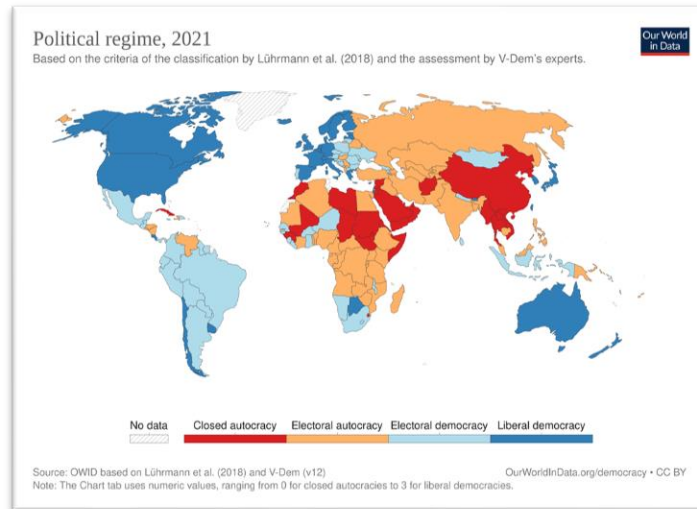
Note: The Chart tab uses numeric values, with 1 if virtually all female citizens are allowed to vote and 0 if not.

CC BY

<https://ourworldindata.org/democracy>

# Each format has its own visualization

Domitilla Brandoni, CINECA

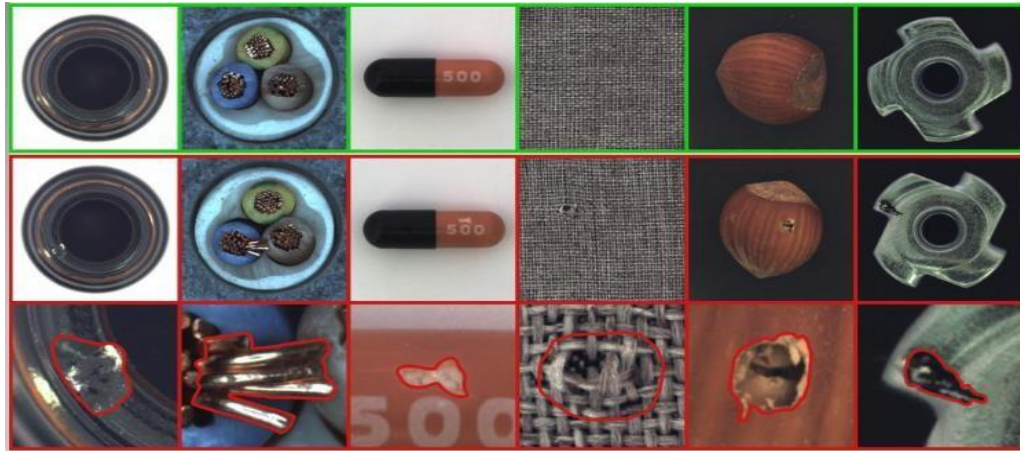


<https://ourworldindata.org/>



# Non tabular data: images

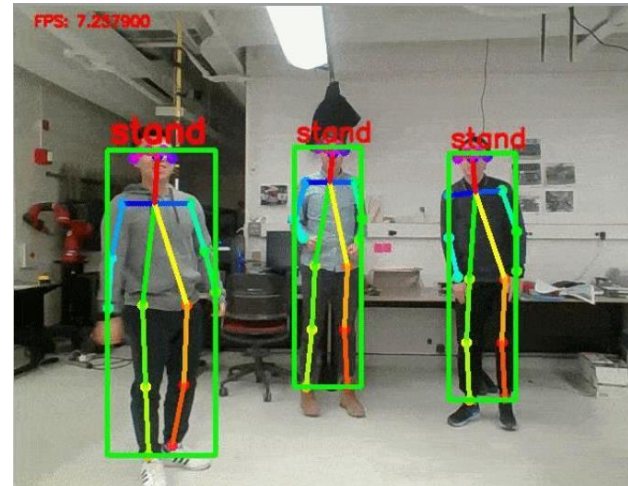
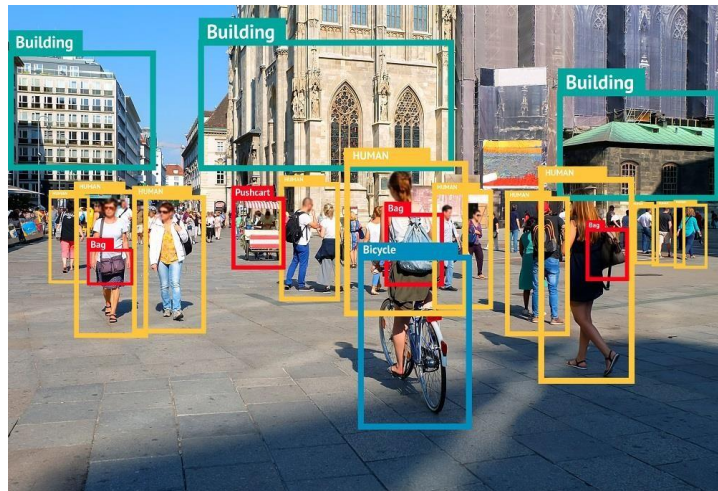
Domitilla Brandoni, CINECA



- Buildings
- Misc.
- Roads
- Tracks
- Trees
- Crops
- Standing Water
- Large Vehicles
- Small Vehicles

# Non tabular data: videos

Domitilla Brandoni, CINECA



The slide is divided into three main sections:

- Text Classification:** A word cloud centered around the text "Text Classification". Other prominent words include "Tokenization", "Word2Vec", "K-nearest Neighbor", "A Survey", "HDLTex", "RMIDL", "Deep Learning", "Support Vector Machine (SVM)", "Random Forest", "HMM", "Bayesian", "Naive Bayes", "Classification", "Learning", "RMIDL", "HDLTex", "Random Forest", "HMM", "Bayesian", "Naive Bayes", "Classification", "Learning", "RMIDL", "HDLTex", "Random Forest", "HMM", "Bayesian", "Naive Bayes", "Classification", "Learning".
- Neural Network Diagram:** A diagram showing a human head profile on the left with sound waves entering an "INPUT" layer of 5 nodes. These connect to "HIDDEN I" (4 nodes), "HIDDEN II" (4 nodes), and "OUTPUTS" (3 nodes). A yellow callout bubble contains the text: "Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nam ornare, leo eu finibus euismod, velit velit convallis odio, sit curabitur nisi ligula a mauris." Below the diagram is the text "Neural Networks - www.gismat.ai/machinelearning".
- Semantic Similarity Exercise:** A diagram with a bracket on the left. The top half is labeled "Semantically Different" and shows a question "How are you?" with a response "Great!". The bottom half is labeled "Semantically Similar" and shows a question "How old are you?" with a response "I am 20", and another question "What is your age?" with a response "I am 20". Each question and response is enclosed in a rounded rectangle and accompanied by a small cartoon character icon.

# Where to find GOOD data

Domitilla Brandoni, CINECA

<https://ourworldindata.org/>

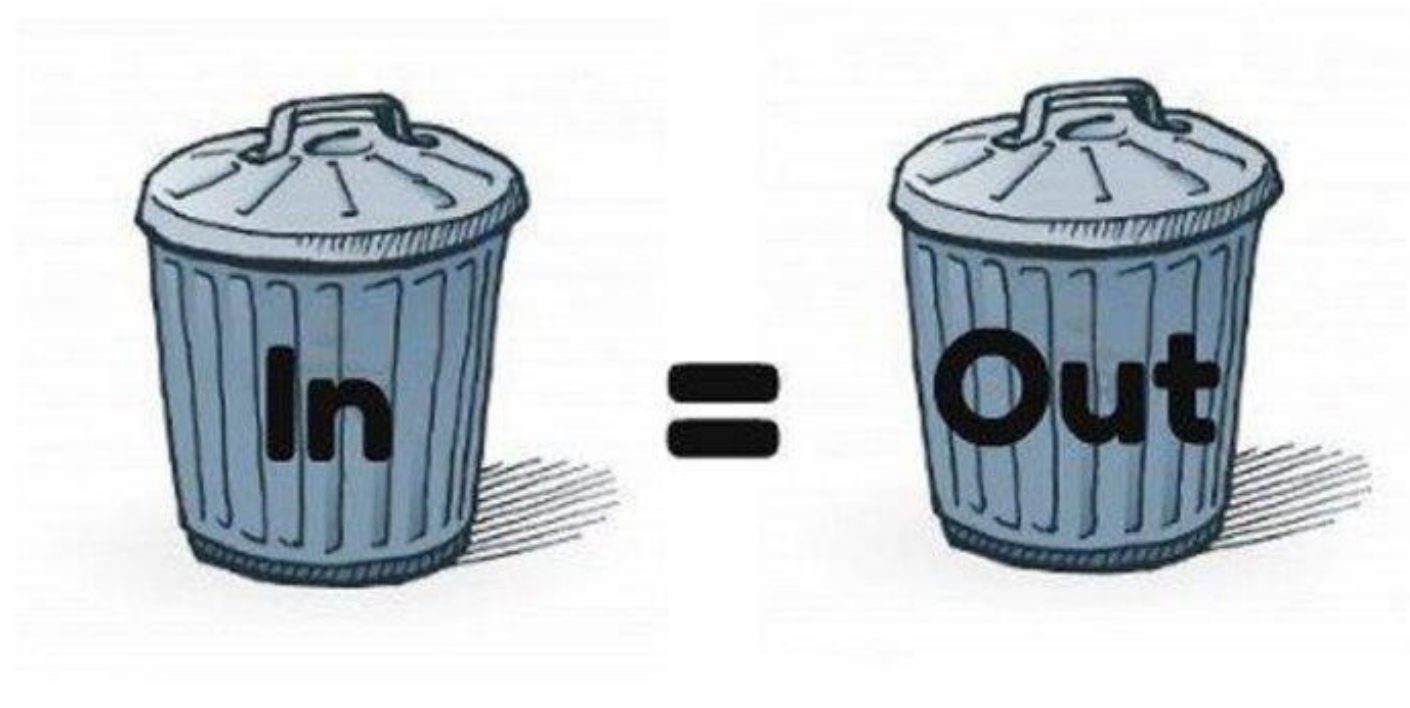
<https://www.kaggle.com/>

<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/>

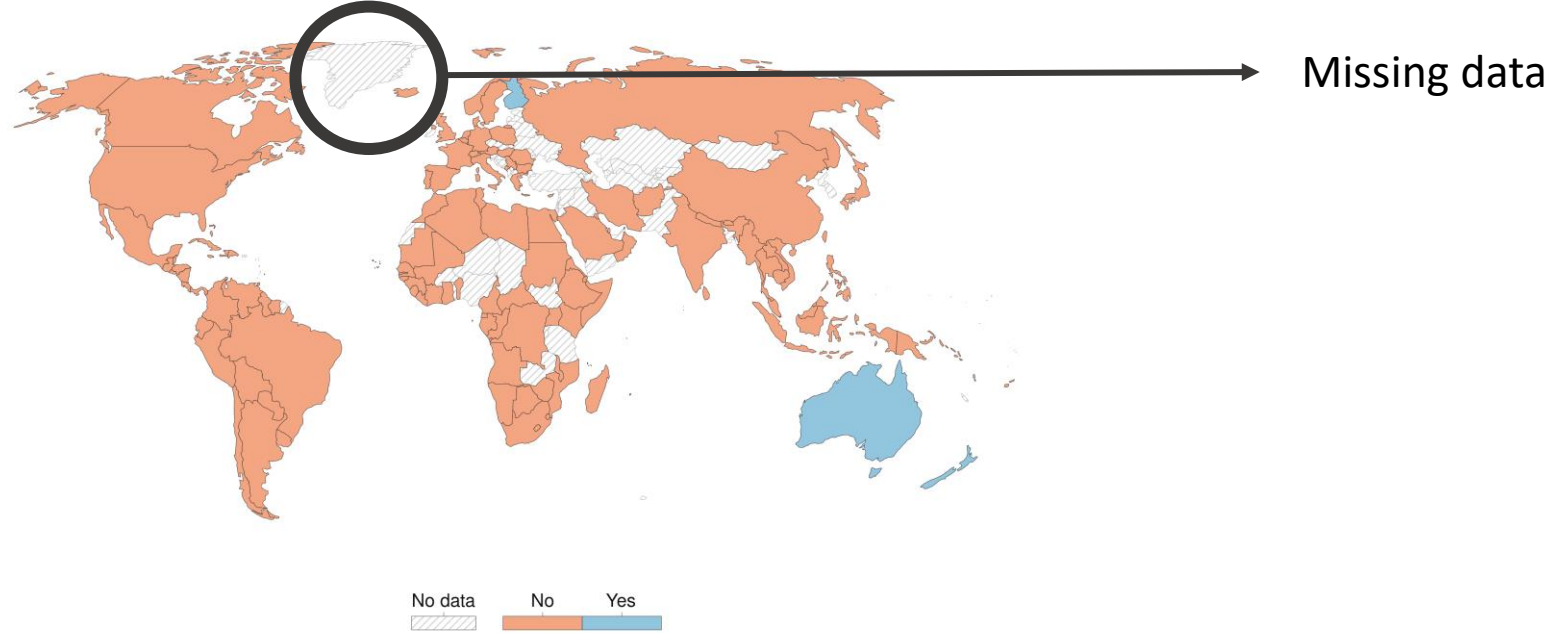
# Missing data

Domitilla Brandoni, CINECA

## Universal right to vote for women, 1910

Based on the classification and assessment by Skaaning et al. (2015).

Our World  
in Data



Source: Skaaning et al. (2015)

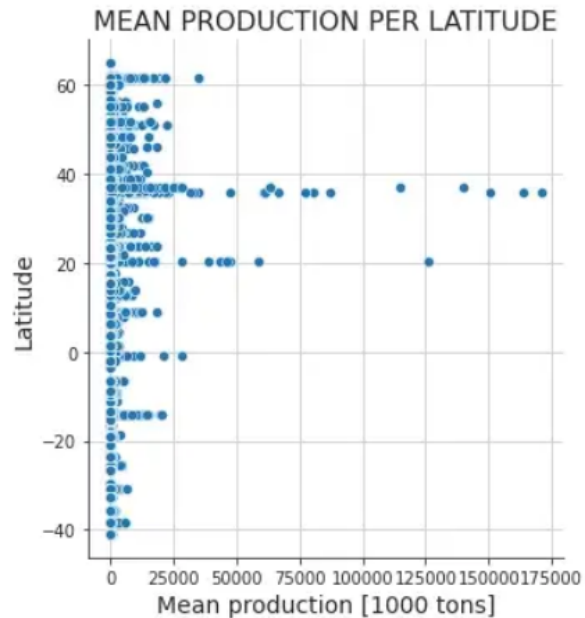
Note: The Chart tab uses numeric values, with 1 if virtually all female citizens are allowed to vote and 0 if not.

CC BY

<https://ourworldindata.org/empowerment>

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

## GRAPHICAL APPROACH



## Z SCORE

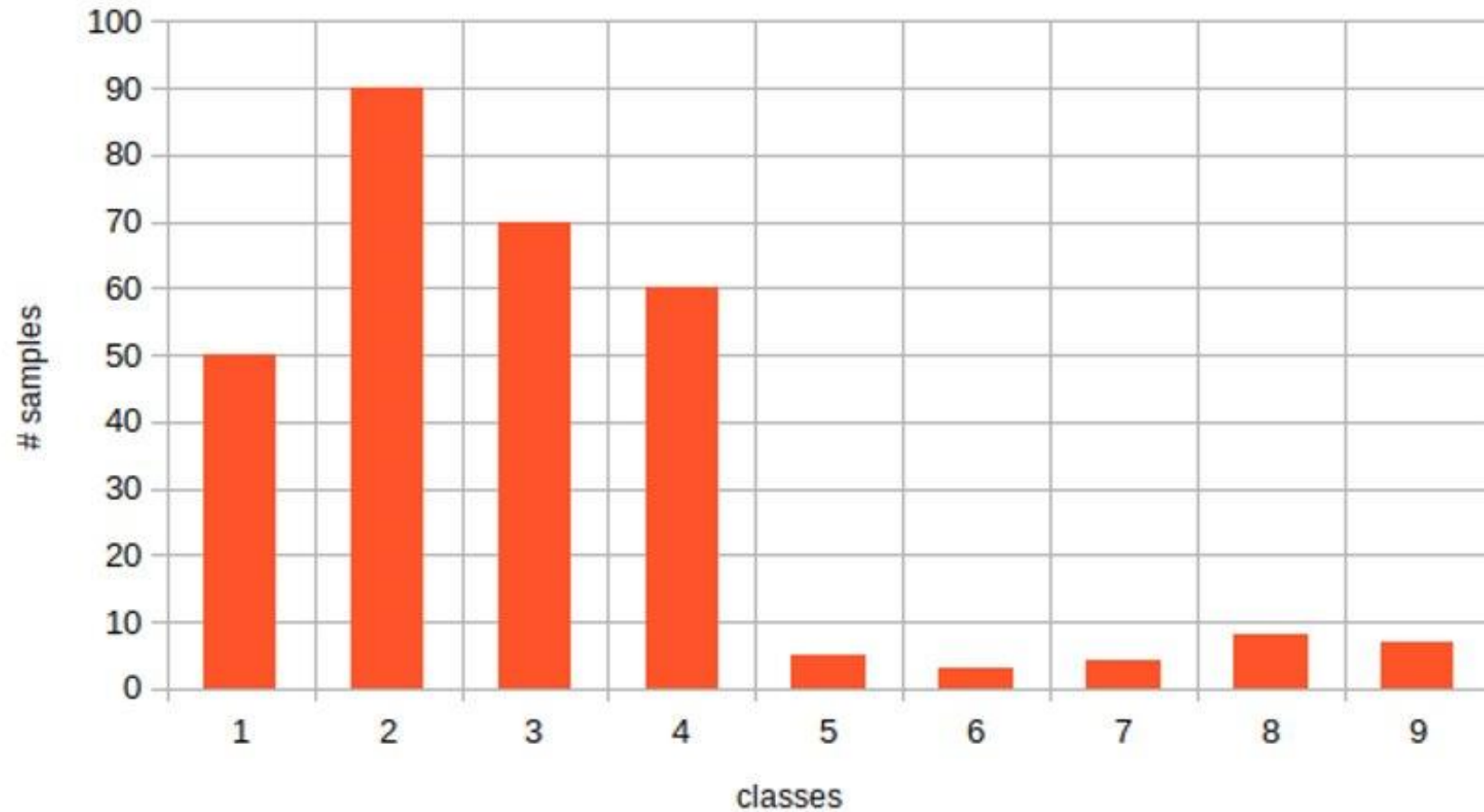
$$Z=(x-m)/s$$

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



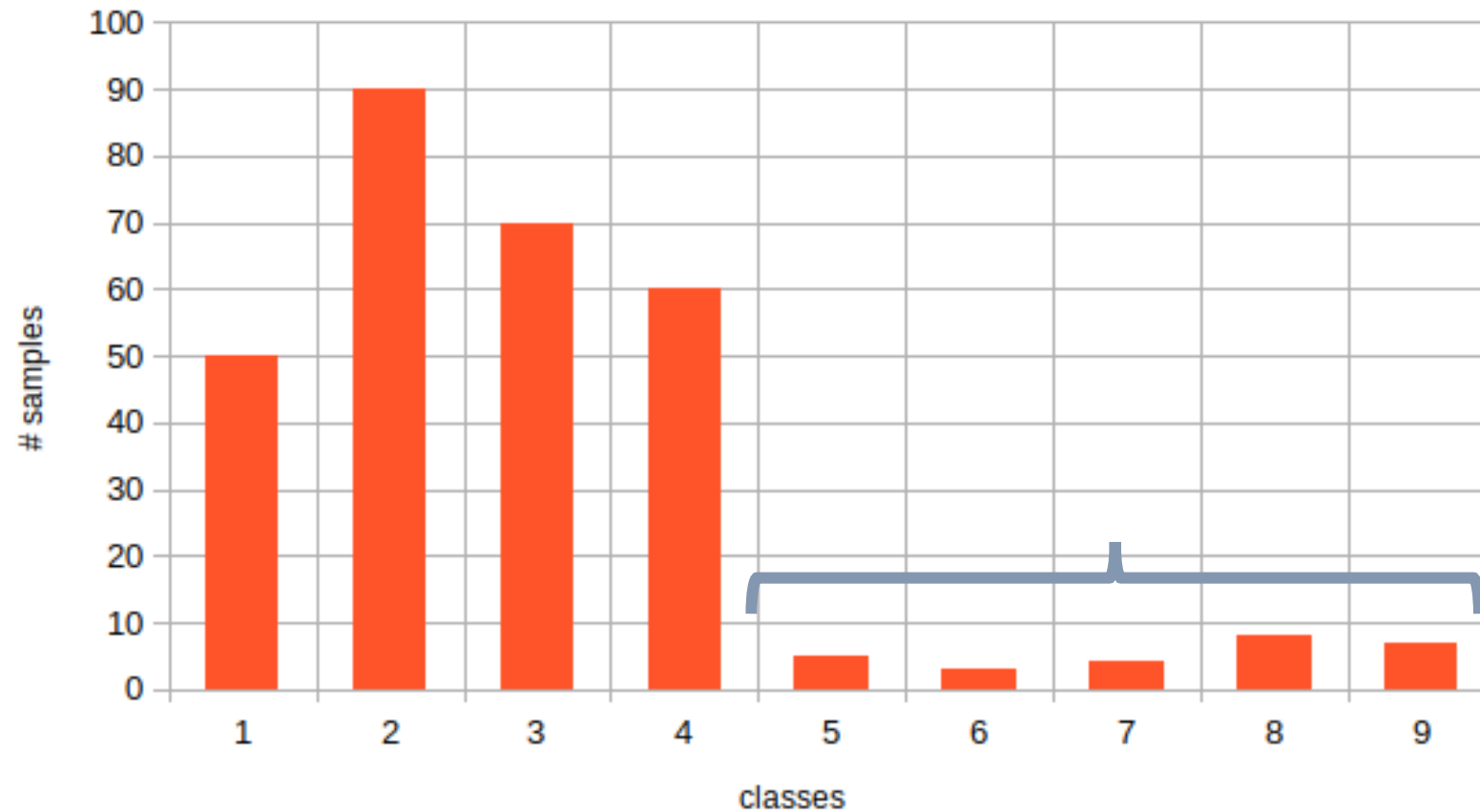
# Unbalanced data

Domitilla Brandoni, CINECA



# Unbalanced data

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

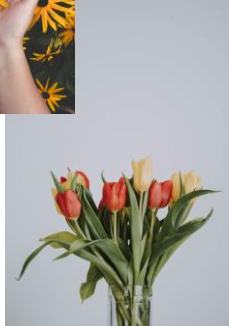
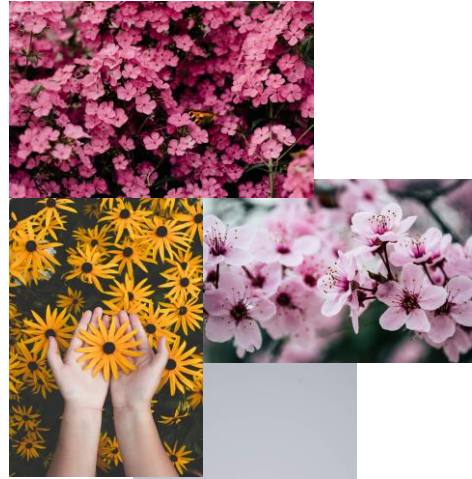
# Numerical continuous data

Domitilla Brandoni, CINECA



# 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



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/>

Univerza v Ljubljani



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.