Data Science and Machine Learning in Python

Stephan Weyers

Fachhochschule Dortmund
University of Applied Sciences and Arts

Topics covered in the online lectures

Fachhochschule Dortmund
University of Applied Sciences and Arts

Part 1: Data Science

	Date	Topics covered
1	Apr 13th	Course introduction Data Science motivation How to use Jupyter Notebook Python types and lists Loops, if/else, functions
2	Apr 20th	Python tuples, lists, dictionaries Functions Numpy basics, operations Image processing
3	Apr 27th	Pandas Series, DataFrame Pandas basic operations Import/export files
4	May 4th	Principles of data visualization Data cleaning and preparation Join, combine and reshape data
5	May 11th	**Volkswohl Bund dataset** **Johanna Dahlbeck** **14:30-15:30 German time** Data visualization in Python How to write Data Science reports Data aggregation and grouping

Part 2: Machine Learning

	Date	Topics covered
6	Jun 1st	Introduction to supervised learning Classification and regression scikit-learn k-Nearest Neighbors Linear Models
7	Jun 8th	Decision trees Random forests and gradient boosting Support vector machines Neural networks
8	Jun 15th	Introduction to unsupervised learning Preprocessing and scaling Dimensionality reduction Principal component analysis
9	Jun 22nd	k-means clustering Hierarchical clustering DBSCAN
10	Jun 29th	Representing data Engineering features
11	Jul 6th	Model evaluation and improvement Text data analysis

Agenda for online lecture 4

Fachhochschule Dortmund
University of Applied Sciences and Arts

Session	Topic	Mode	Materials used	Time (min)
14:30-16:00	Tutorial Marius Meiners	Q&A		10
	Data preparation / cleaning	Lecture / Q&A	Lecture slides	20
	Inventory step 1 & 2	Lecture / Q&A	Inventory Excel	10
	Inventory step 1	Lecture / Q&A	Lecture 04 notebook	10
	Inventory step 2	Team work in break-out rooms	Lecture 04 notebook	25
	Inventory step 2	Presentation of results	Lecture 04 notebook	10
16:10-17:40	Importance of Visualization	Lecture / Q&A	Lecture slides	5
	Visualization: Trustworthy	Discussion in main room	Lecture slides	15
	Visualization: Accessibility	Discussion in main room	Lecture slides	15
	Visualization: Elegance	Discussion in main room	Lecture slides	15
	Inventory step 3 & 4	Lecture / Q&A	Inventory Excel	10
	Inventory step 3	Lecture / Q&A	Lecture 04 notebook	10
	Inventory step 4	Team work in break-out rooms	Lecture 04 notebook	15
17:50-19:20	Inventory step 5 & 6 & 7	Lecture / Q&A	Inventory Excel	15
	Inventory step 5	Lecture / Q&A	Lecture 04 notebook	10
	Inventory step 4 & 6 & 7	Team work in break-out rooms	Lecture 04 notebook	40
	Inventory step 4 & 6 & 7	Presentation of results	Lecture 04 notebook	10
	Questions W02	Q&A		20
	Team grouping, deadlines	Lecture / Q&A		5

Data Preparation – Selection

Fachhochschule Dortmund
University of Applied Sciences and Arts

		Description	Examples
Selection of cases	**Validity**	Restriction to valid cases	▪ Only active customers ▪ No “test-buyers”
	Relevance	Restriction to relevant cases related to target	▪ Only contract customers, no pre-paid customers
	Sample	Restriction to sample in case of large dataset	▪ Random sampling vs. stratified sampling
	Partition	Split up into homogeneous segments	▪ Certain customer segments, e.g. high-value
Selection of variables	**Quality**	Exclude variable of low quality	▪ Zero variance ▪ Many missing values
	Time	Use variables time-related to target variable	▪ Use only sales data of last 12 months
	Correlation	Exclude variables with high correlation among each other	▪ Weight and height ▪ Orders and returns
	Relevance	Exclude variables with no relevance for target	▪ Postal code ▪ Insurance ID

Source: Prof. Dr. Michael Bücker – Data Mining

Data Cleaning

Fachhochschule Dortmund
University of Applied Sciences and Arts

ID	Name	Income	Age	Gender	Target
1	Müller	50.570	45	M	1
2	Meyer	19.032	21	M	0
3	Schneider	43.452		F	0
4	Jäger	75.976	37	Male	1
5	Muller	50.570	45	M	J
6	Langer	47.414	140	Femnale	1
7	Schäfer	228.011	48	F	0
8	Simon	36.976	27	M	1

Question for discussion

- Which errors can you observe in the dataset?

Source: Prof. Dr. Michael Bücker – Data Mining

Data Preparation – Cleaning

Fachhochschule Dortmund
University of Applied Sciences and Arts

Options for missing values

- Delete (e.g. remove all customers without email-address)
- Replace (e.g. by mean values by segment)
- Accept (use missing as new category)

Options for outliers

- Delete (e.g. remove all customers sales above 1 Mio. EUR)
- Replace (e.g. truncate all sales above 10.000 EUR to 10.000 EUR)
- Accept (maybe outliers are especially interesting for analysis)
- Binning (use categories like 0-20, 20-50, 50-100, 100-500, >500)

Options for values with errors

- Delete (e.g. remove all customers with sales < 0 EUR)
- Replace (e.g. replace “US” by “United States”)
- Accept (if reasonable, e.g. negative sales = returns)

Source: Prof. Dr. Michael Bücker – Data Mining

Example 1 – raw data

X	Y
0.9	0.5
2.7	1.1
6.7	28.6
10.9	32.8
6.0	15.7
6.3	19
7 .0	32.6

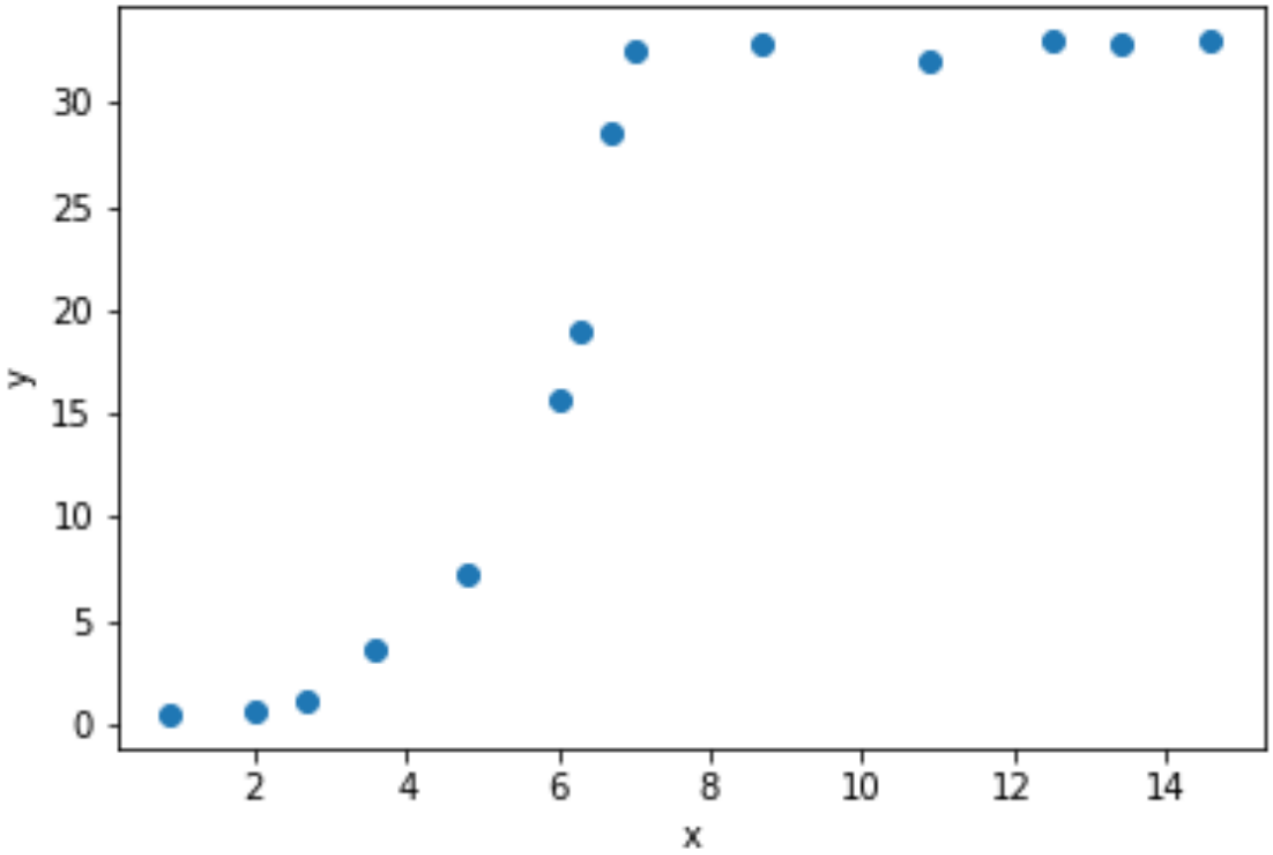
X	Y
8.7	32.3
4.8	7.3
12.5	33.1
13.4	32.9
2.0	0.75
3.6	3.6
14.6	33

	X	Y
Median	6.5	23.8
Mean	7.2	19.5
STD.DEV	4.2	13.6

Correlation = 0.88

Source: UCSanDiegoX DSE200x - Python for Data Science

Example 1 – Visualized



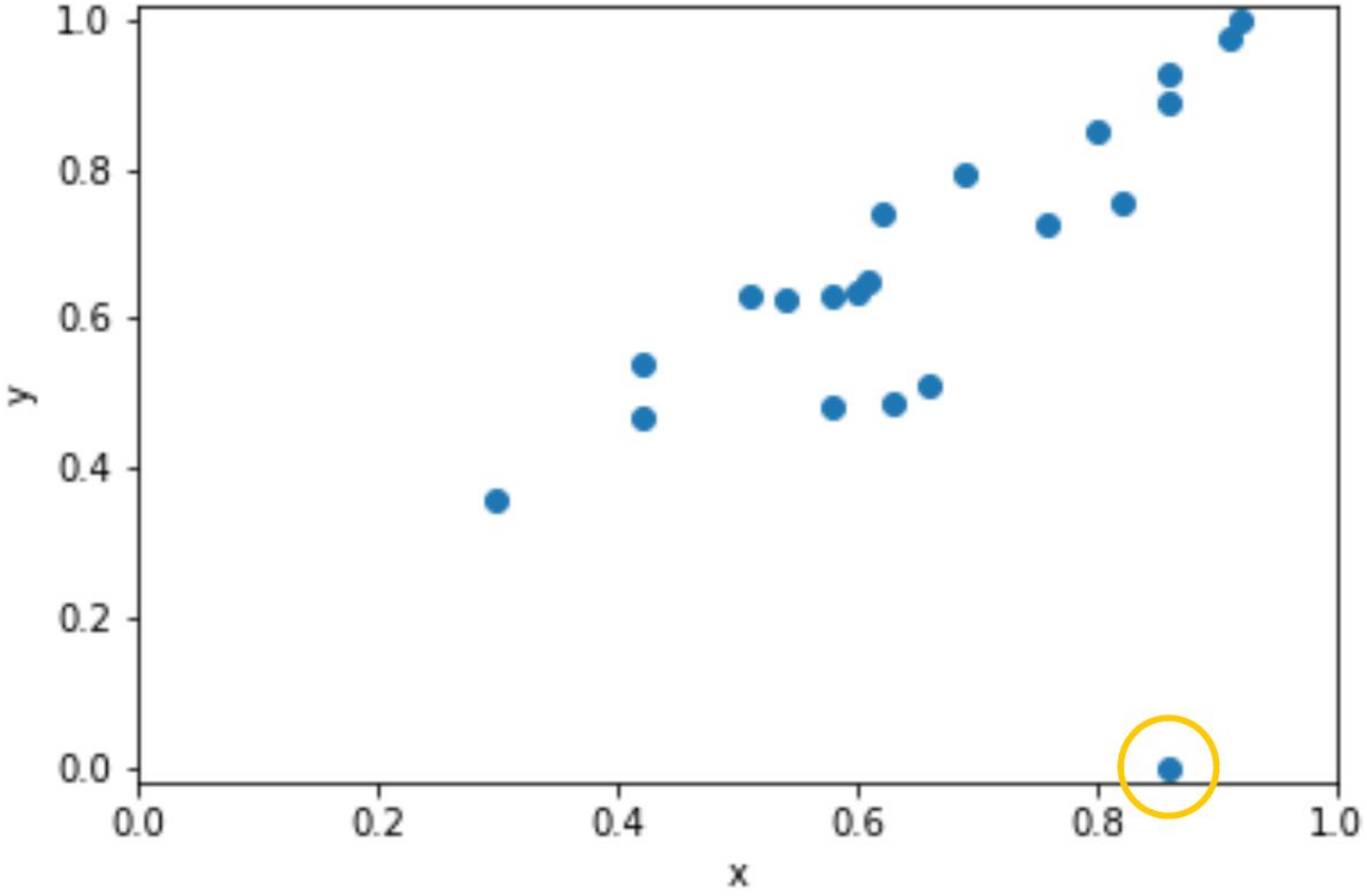
Source: UCSanDiegoX DSE200x - Python for Data Science

Example 2 – raw data

X	Y	X	Y
0.42	0.46750	0.66	0.51250
0.54	0.62500	0.63	0.48750
0.42	0.53750	0.92	1.00000
0.86	0.92750	0.86	0.88750
0.60	0.63750	0.91	0.97500
0.51	0.63125	0.82	0.75625
0.30	0.35625	0.86	0.00000
0.61	0.65000	0.80	0.85000
0.58	0.63125	0.69	0.79375
0.76	0.72500	0.62	0.74000
0.58	0.48125		

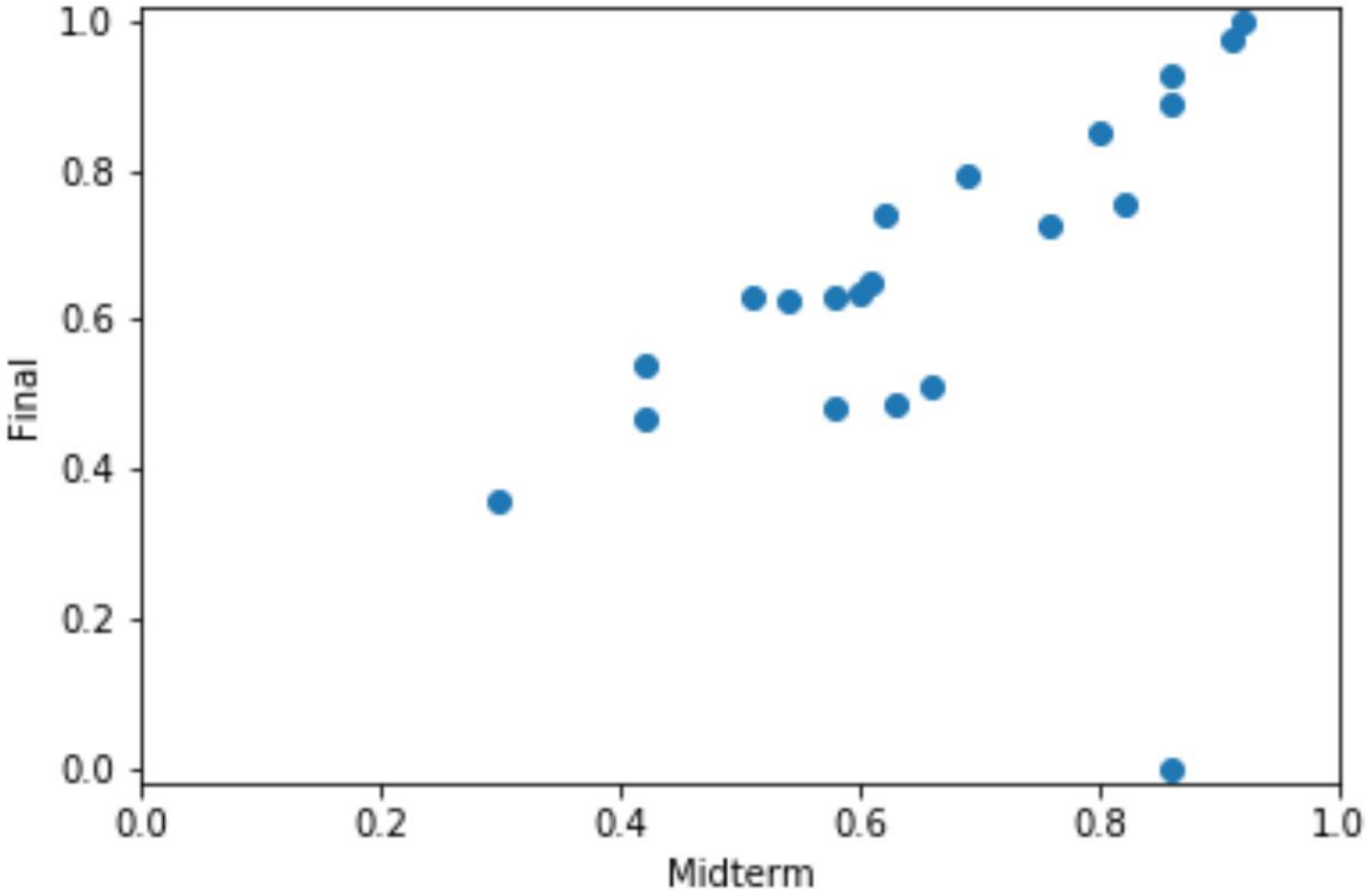
Source: UCSanDiegoX DSE200x - Python for Data Science

Example 2 – Visualized



Source: UCSanDiegoX DSE200x - Python for Data Science

Example 2 – Visualized



Source: UCSanDiegoX DSE200x - Python for Data Science

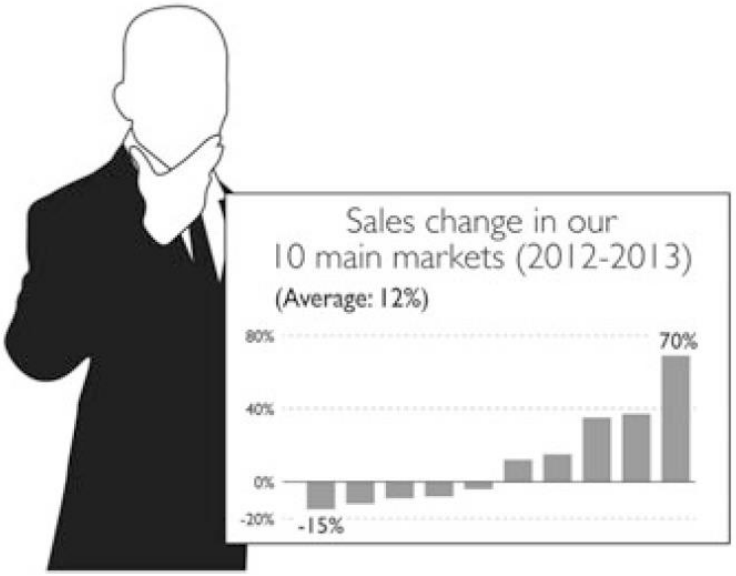
Principles of Good Design

Good data visualization

- Trustworthy
- Accessible
- Elegant

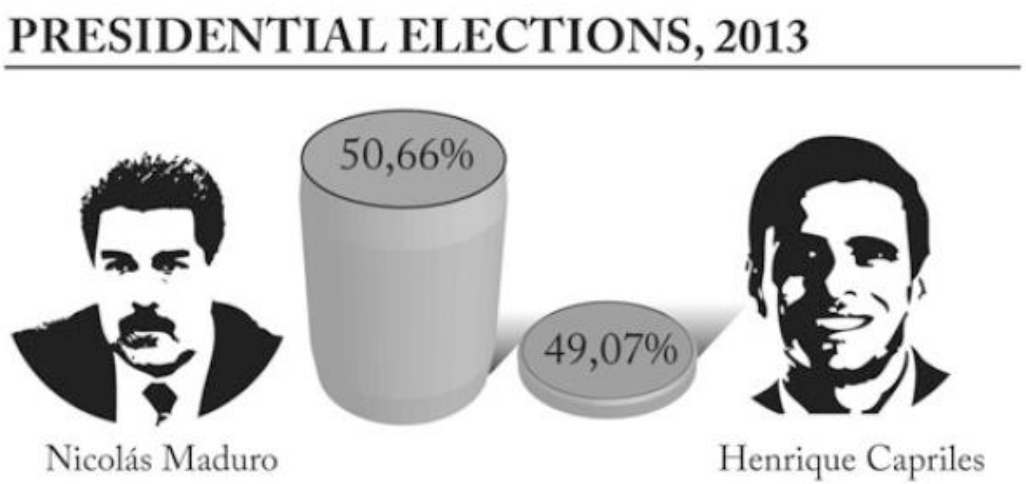
Source: Kirk, A. Data Visualisation: A handbook for Data Driven Design. SAGE publications, 2016.

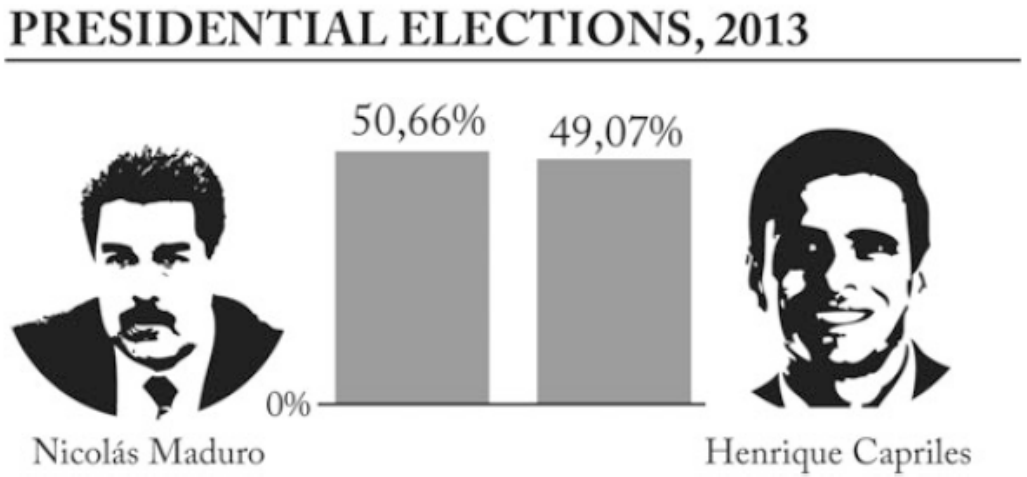
Trustworthy:



Source: Cairo, A. (2015). Graphics lies, misleading visuals. In New Challenges for Data Design (pp. 103-116). Springer London.

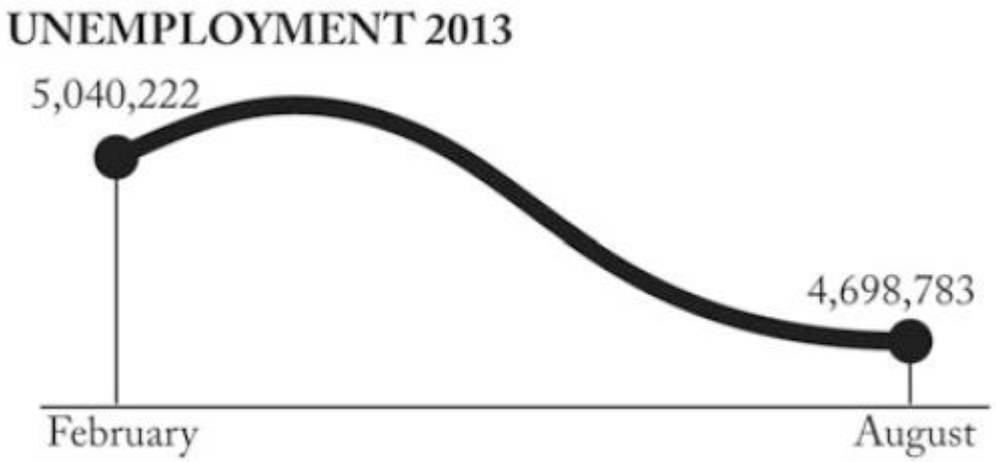
Trustworthy:

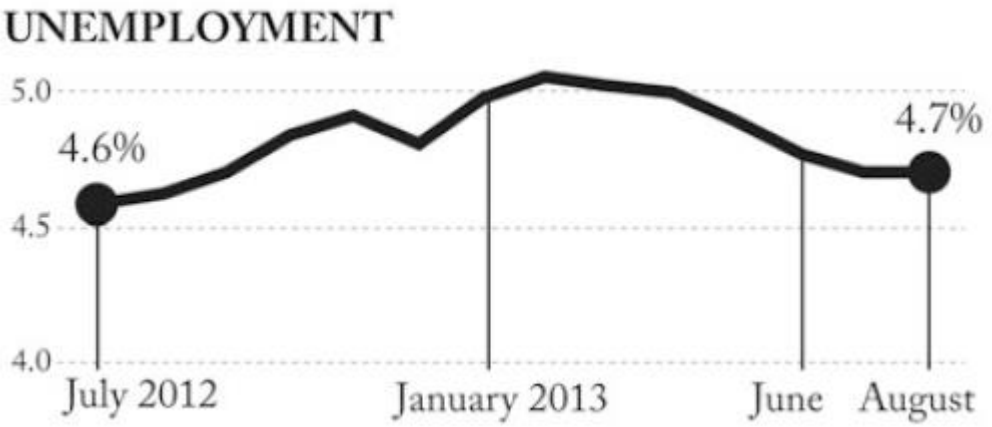




Source: Cairo, A. (2015). Graphics lies, misleading visuals. In New Challenges for Data Design (pp. 103-116). Springer London.

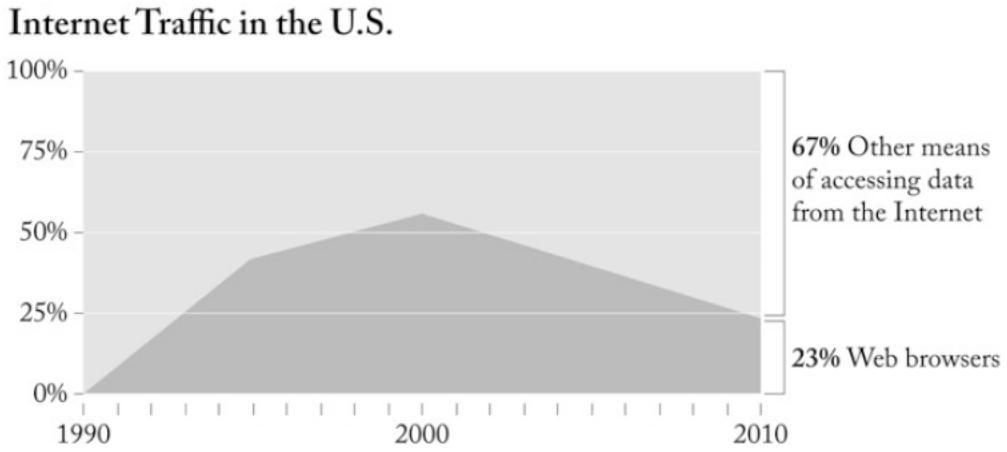
Trustworthy:

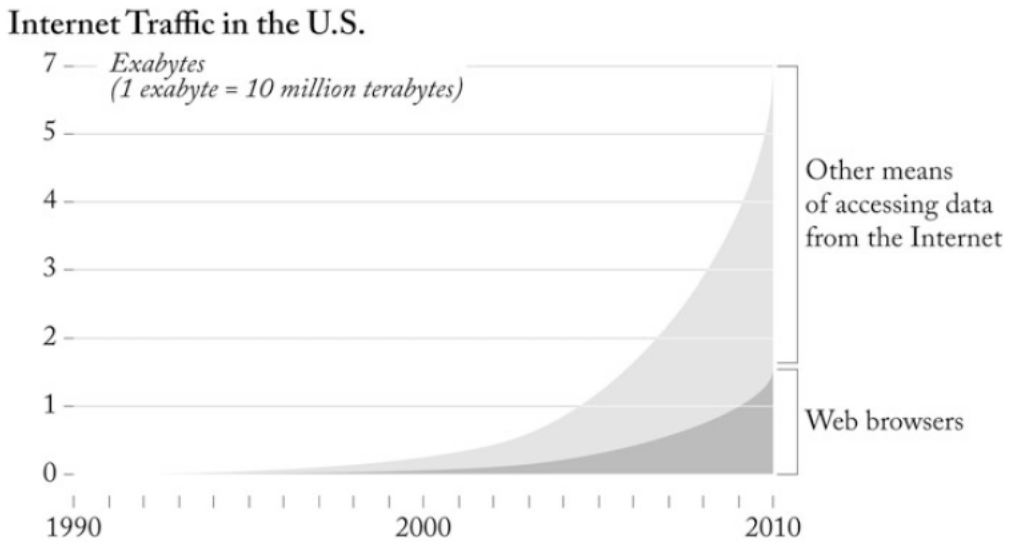




Source: Cairo, A. (2015). Graphics lies, misleading visuals. In New Challenges for Data Design (pp. 103-116). Springer London.

Trustworthy:



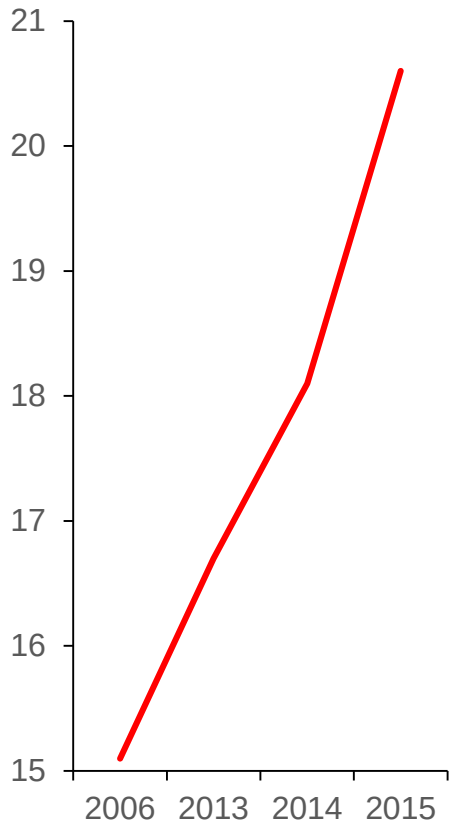


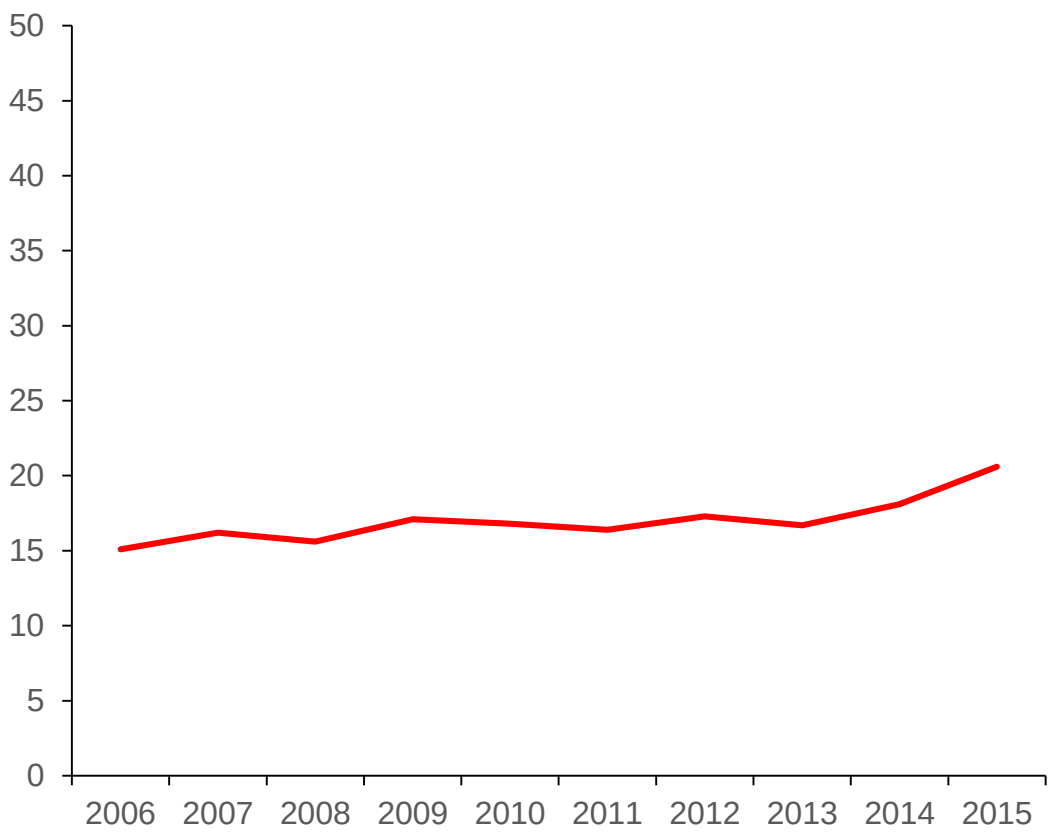
Source: Cairo, A. (2015). Graphics lies, misleading visuals. In New Challenges for Data Design (pp. 103-116). Springer London.

Trustworthy:

Fachhochschule Dortmund
University of Applied Sciences and Arts

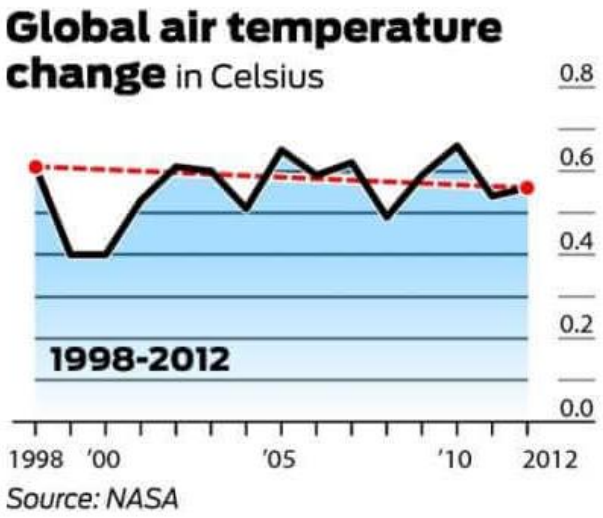
Proportion of women on supervisory and management boards



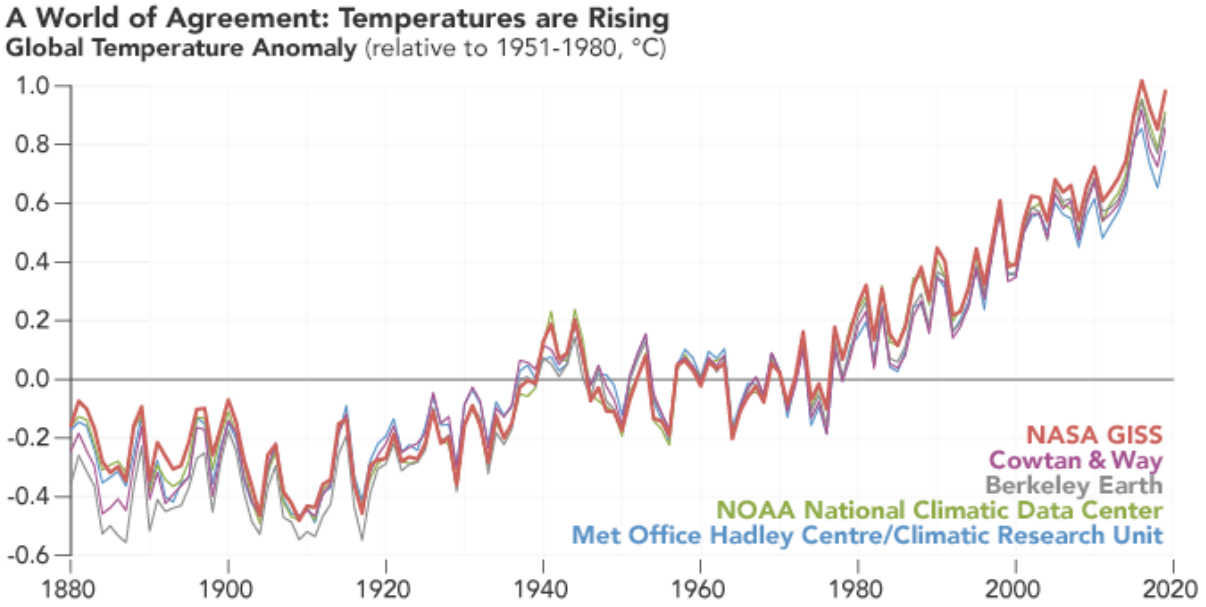


Source: https://www.wiwo.de/politik/deutschland/statistik-das-auge-interpretiert-mit/13580898.html

Trustworthy:

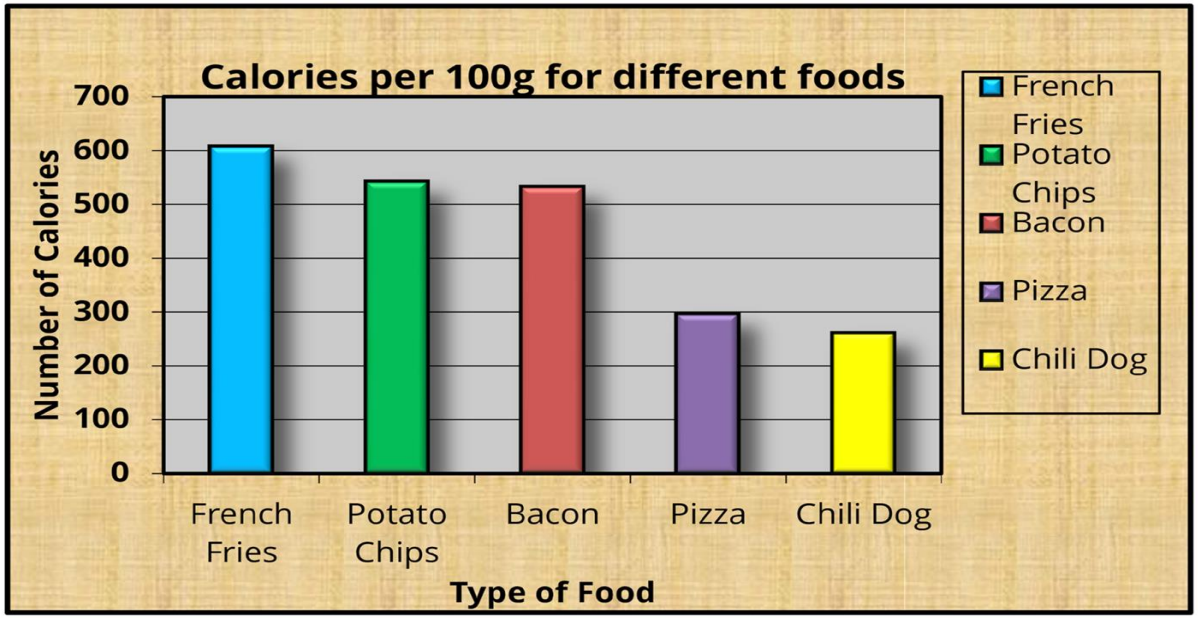


Source:
https://www.datapine.com/blog/misleading-statistics-and-data/



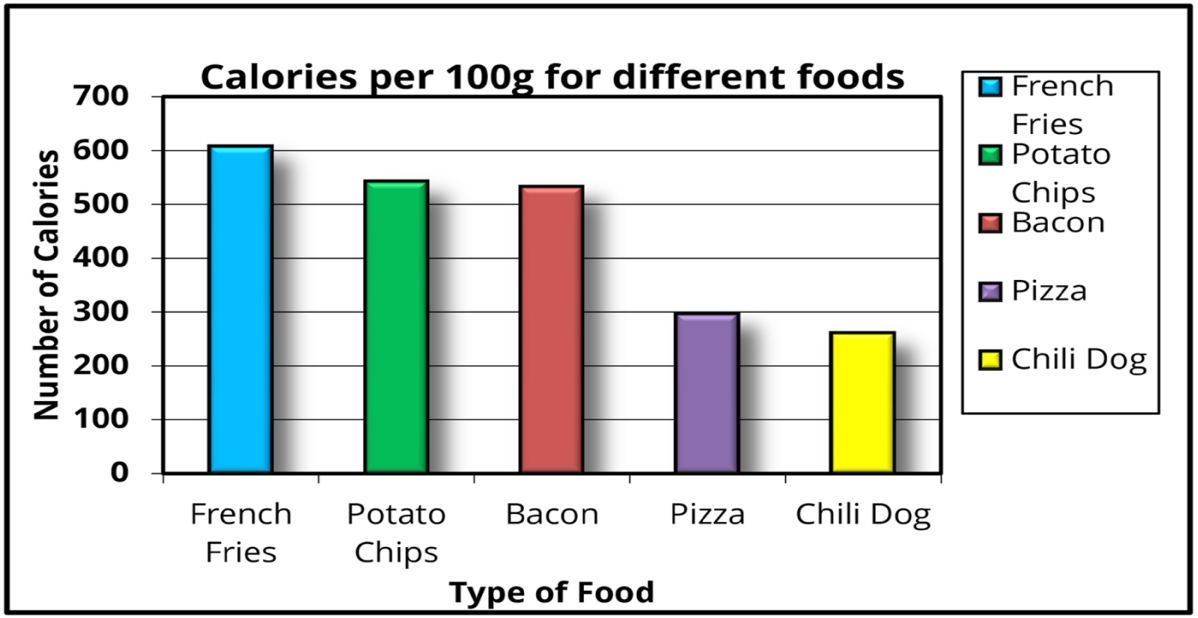
Source: https://earthobservatory.nasa.gov/world-of-change/global-temperatures

Accessibility: Data-To-Ink-Ratio



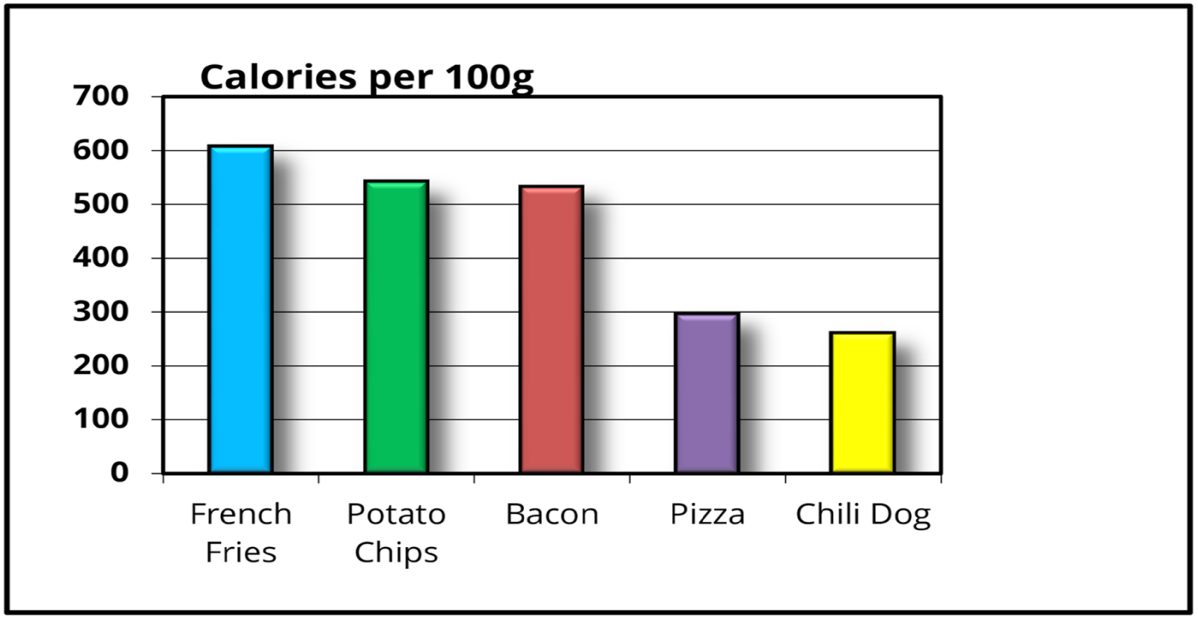
Source: https://www.darkhorseanalytics.com/blog/data-looks-better-naked

Accessibility: Data-To-Ink-Ratio



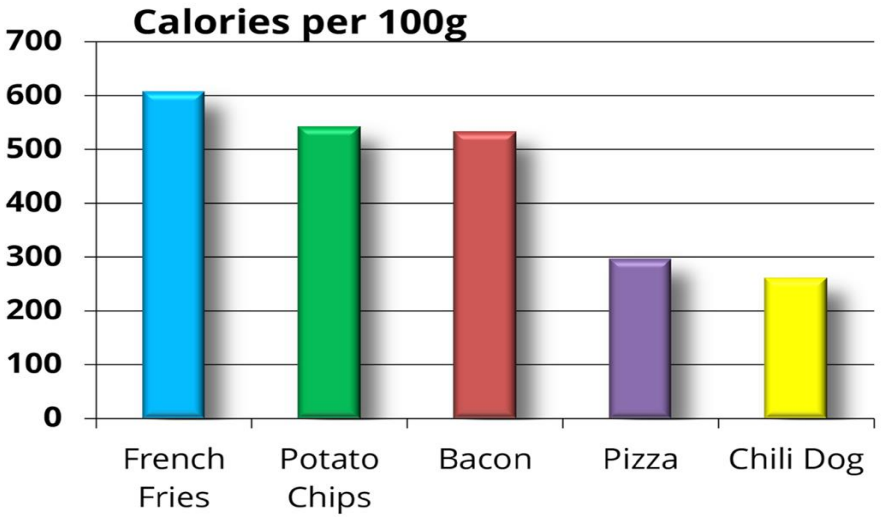
Source: https://www.darkhorseanalytics.com/blog/data-looks-better-naked

Accessibility: Data-To-Ink-Ratio



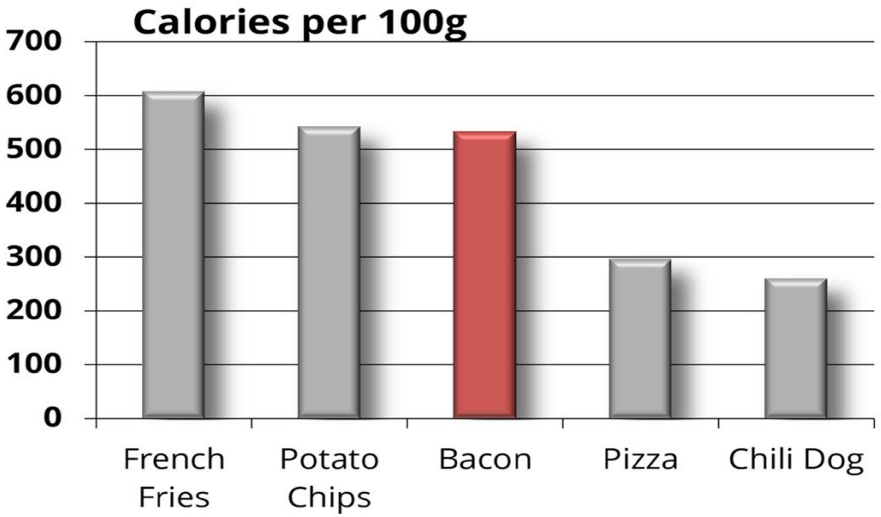
Source: https://www.darkhorseanalytics.com/blog/data-looks-better-naked

Accessibility: Data-To-Ink-Ratio



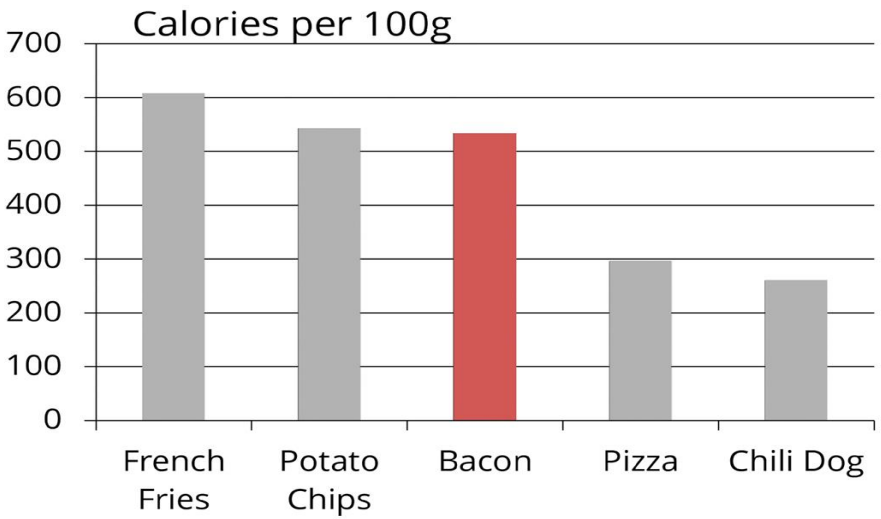
Source: https://www.darkhorseanalytics.com/blog/data-looks-better-naked

Accessibility: Data-To-Ink-Ratio



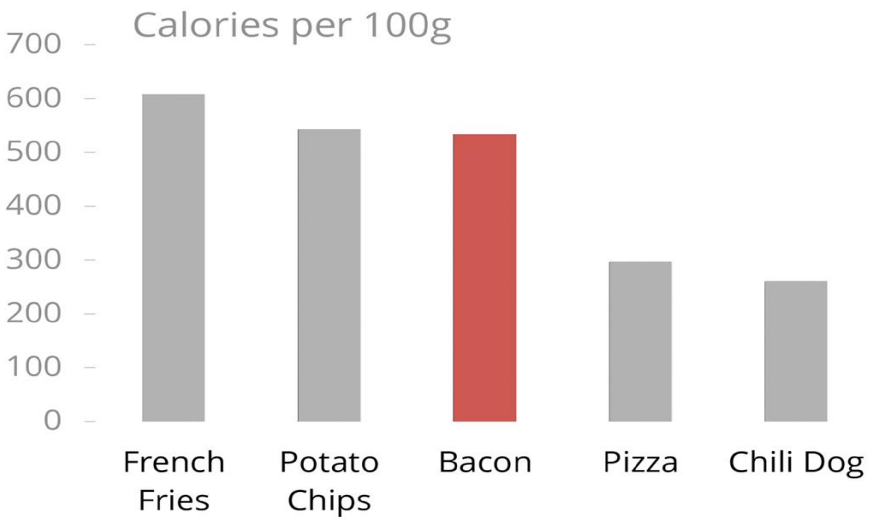
Source: https://www.darkhorseanalytics.com/blog/data-looks-better-naked

Accessibility: Data-To-Ink-Ratio



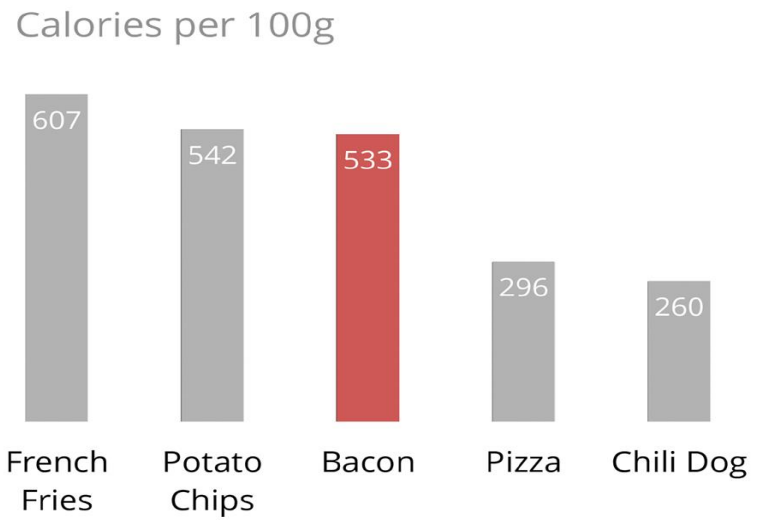
Source: https://www.darkhorseanalytics.com/blog/data-looks-better-naked

Accessibility: Data-To-Ink-Ratio



Source: https://www.darkhorseanalytics.com/blog/data-looks-better-naked

Accessibility: Data-To-Ink-Ratio

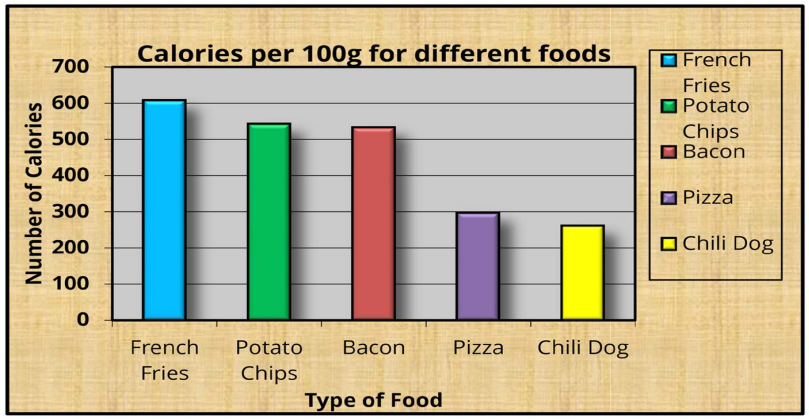


Source: https://www.darkhorseanalytics.com/blog/data-looks-better-naked

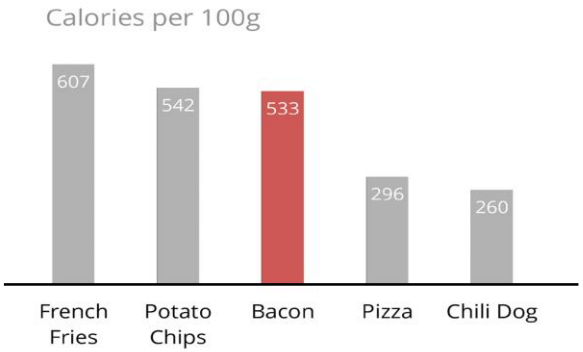
Accessibility: Data-To-Ink-Ratio

Fachhochschule Dortmund
University of Applied Sciences and Arts

Before



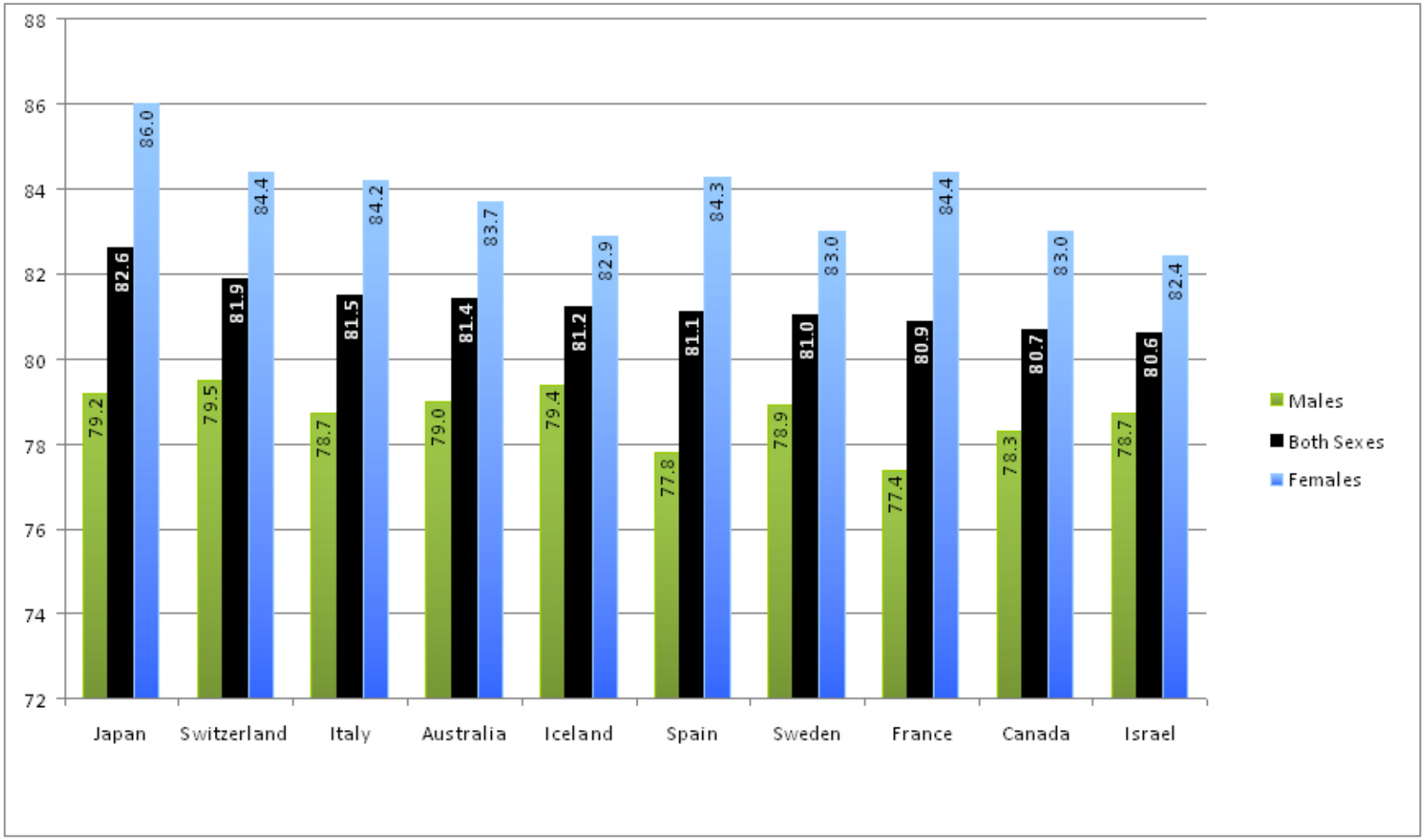
After



Source: https://www.darkhorseanalytics.com/blog/data-looks-better-naked

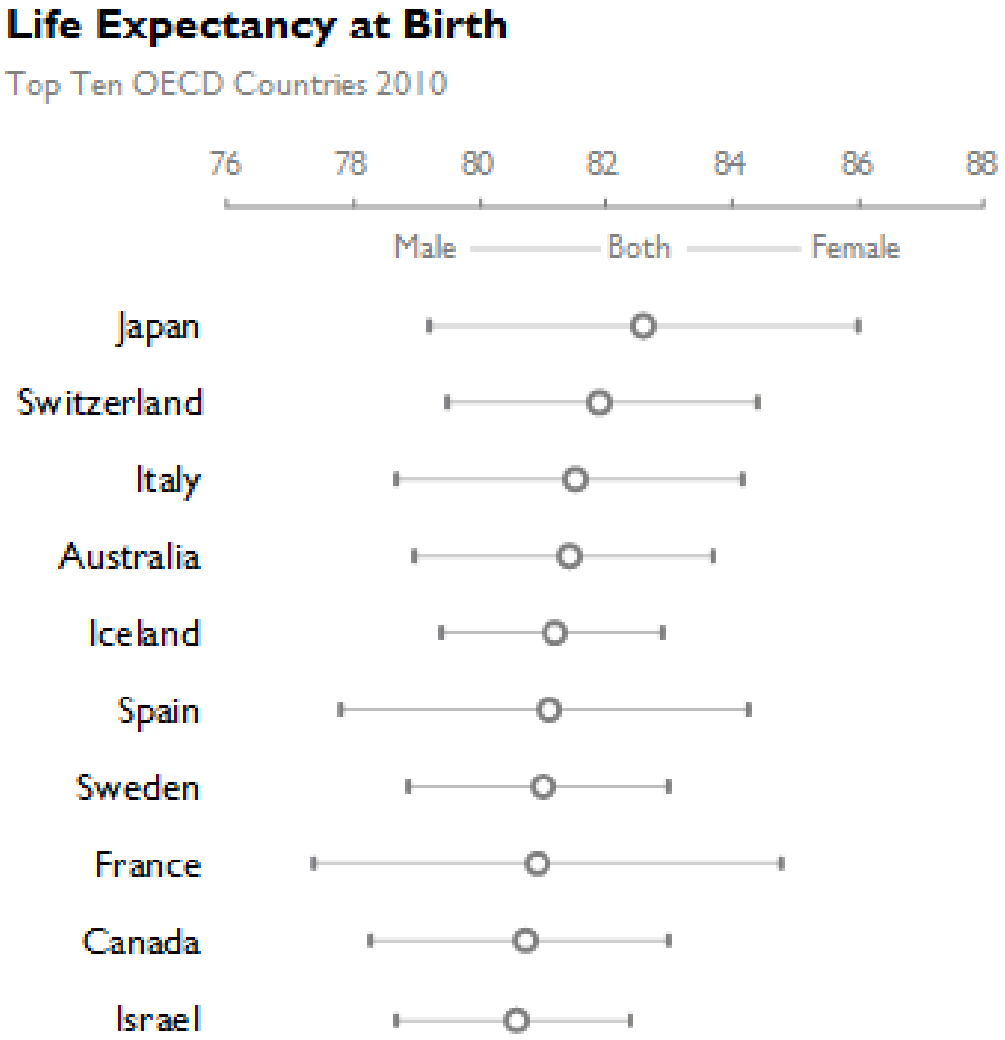
Accessibility: Too Many Bars – Original Chart

Fachhochschule Dortmund
University of Applied Sciences and Arts



Source: https://www.darkhorseanalytics.com/blog/too-many-bars

Accessibility: Too Many Bars – Alternative Chart



Source: https://www.darkhorseanalytics.com/blog/too-many-bars

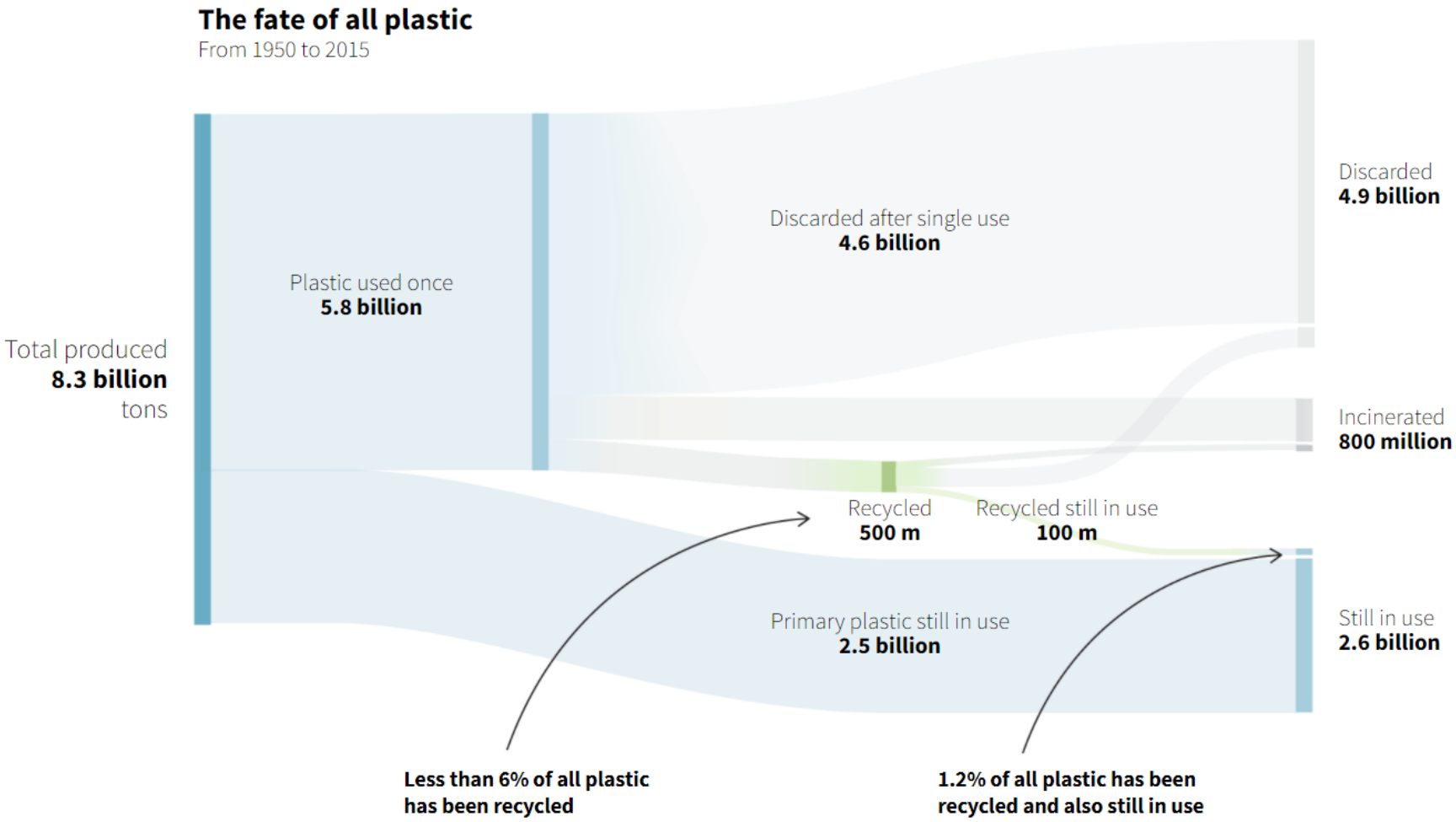
Elegance: Cholera Outbreak London 1854

Fachhochschule Dortmund
University of Applied Sciences and Arts

Source: https://www.visualcapitalist.com/wp-content/uploads/2018/04/data-viz-cholera.html

Elegance: The Fate of Plastic

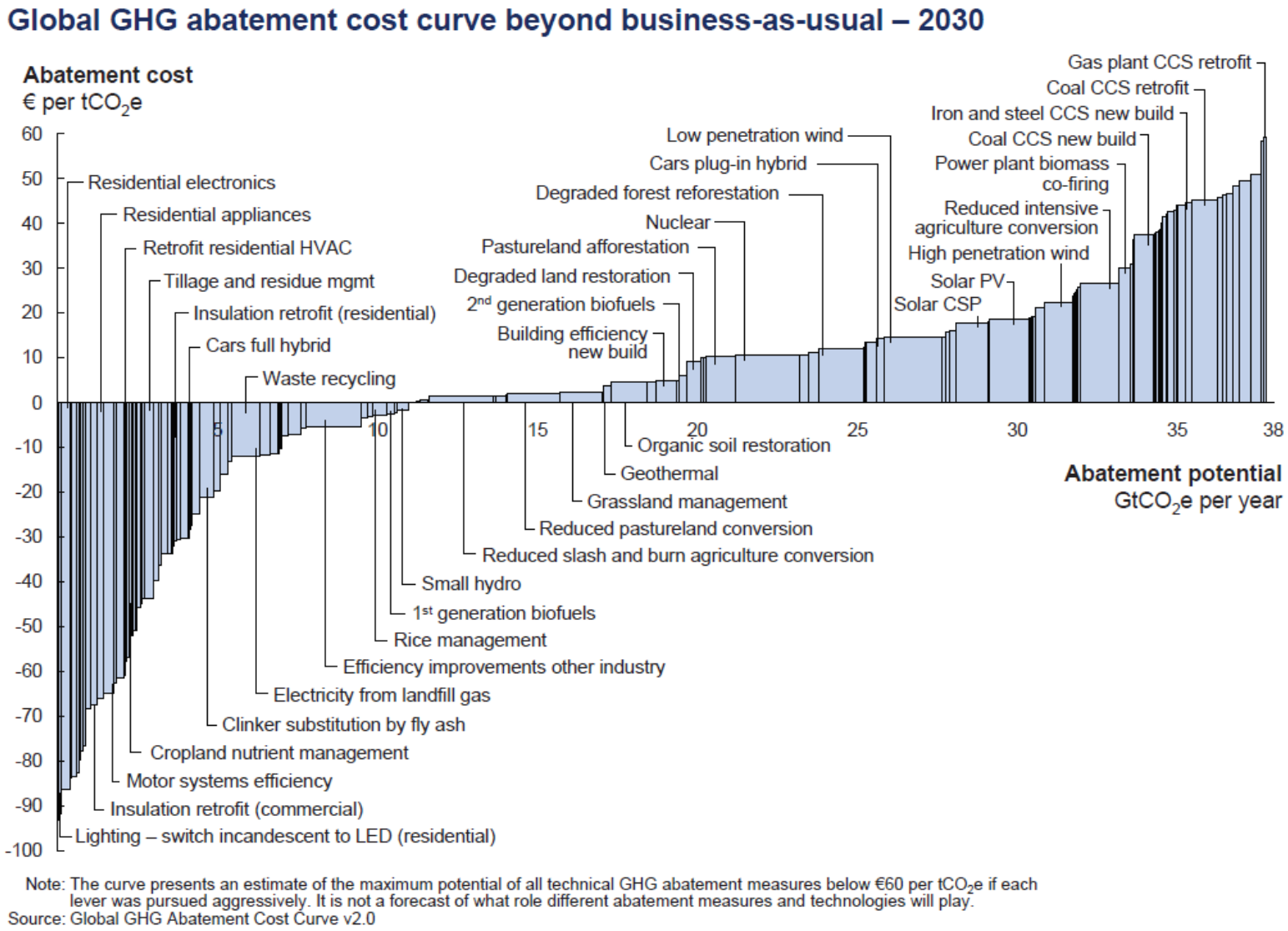
Fachhochschule Dortmund
University of Applied Sciences and Arts



Source: https://graphics.reuters.com/ENVIRONMENT-PLASTIC/0100B275155/index.html

Elegance: GHG Abatement Cost Curves

Fachhochschule Dortmund
University of Applied Sciences and Arts



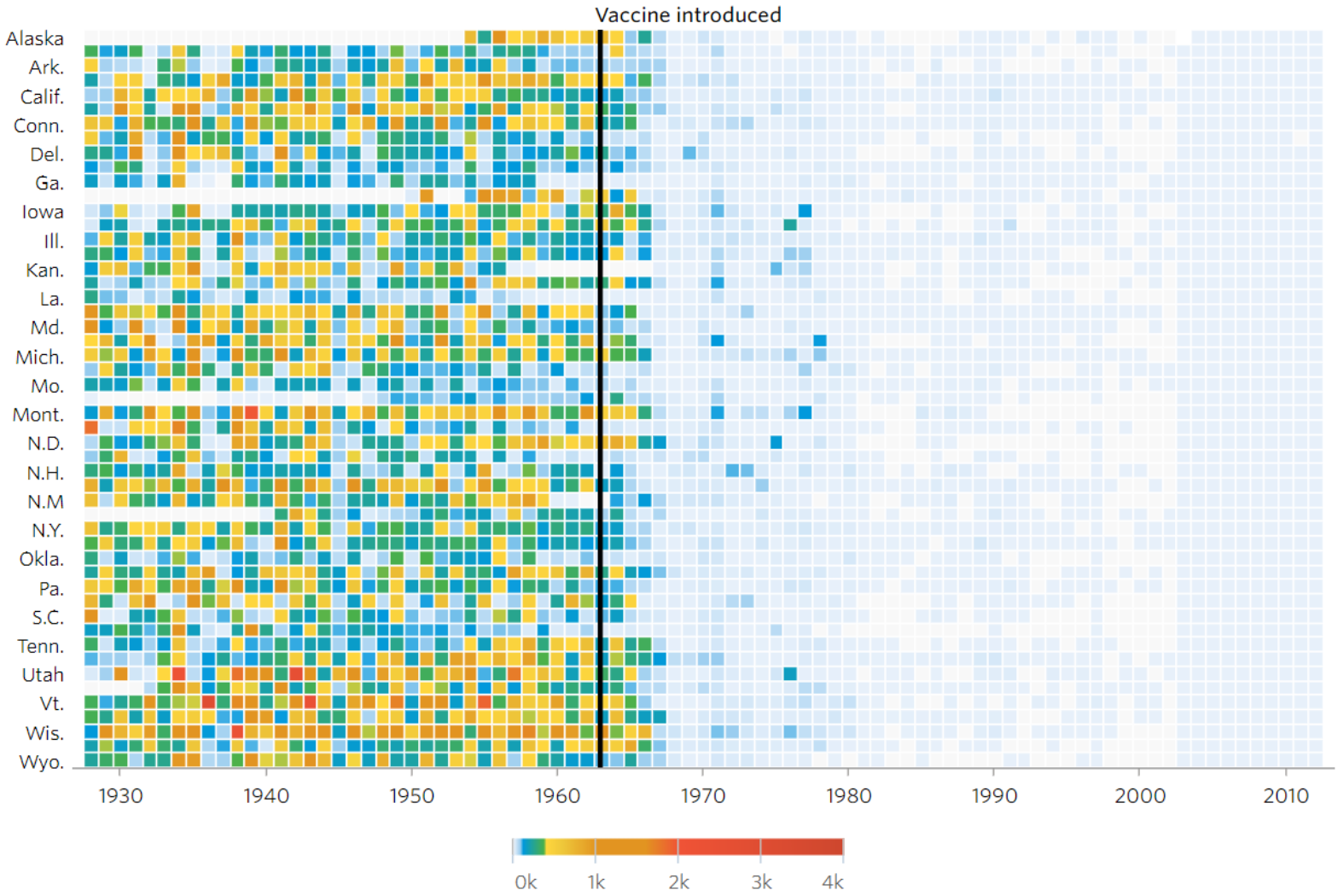
Note: The curve presents an estimate of the maximum potential of all technical GHG abatement measures below €60 per tCO_2e if each lever was pursued aggressively. It is not a forecast of what role different abatement measures and technologies will play.
Source: Global GHG Abatement Cost Curve v2.0

Source: McKinsey & Company, Pathways to a low-carbon economy - version 2 of the global greenhouse gas abatement cost curve, 2009

Trustworthy? Accessible? Elegant?

Fachhochschule Dortmund
University of Applied Sciences and Arts

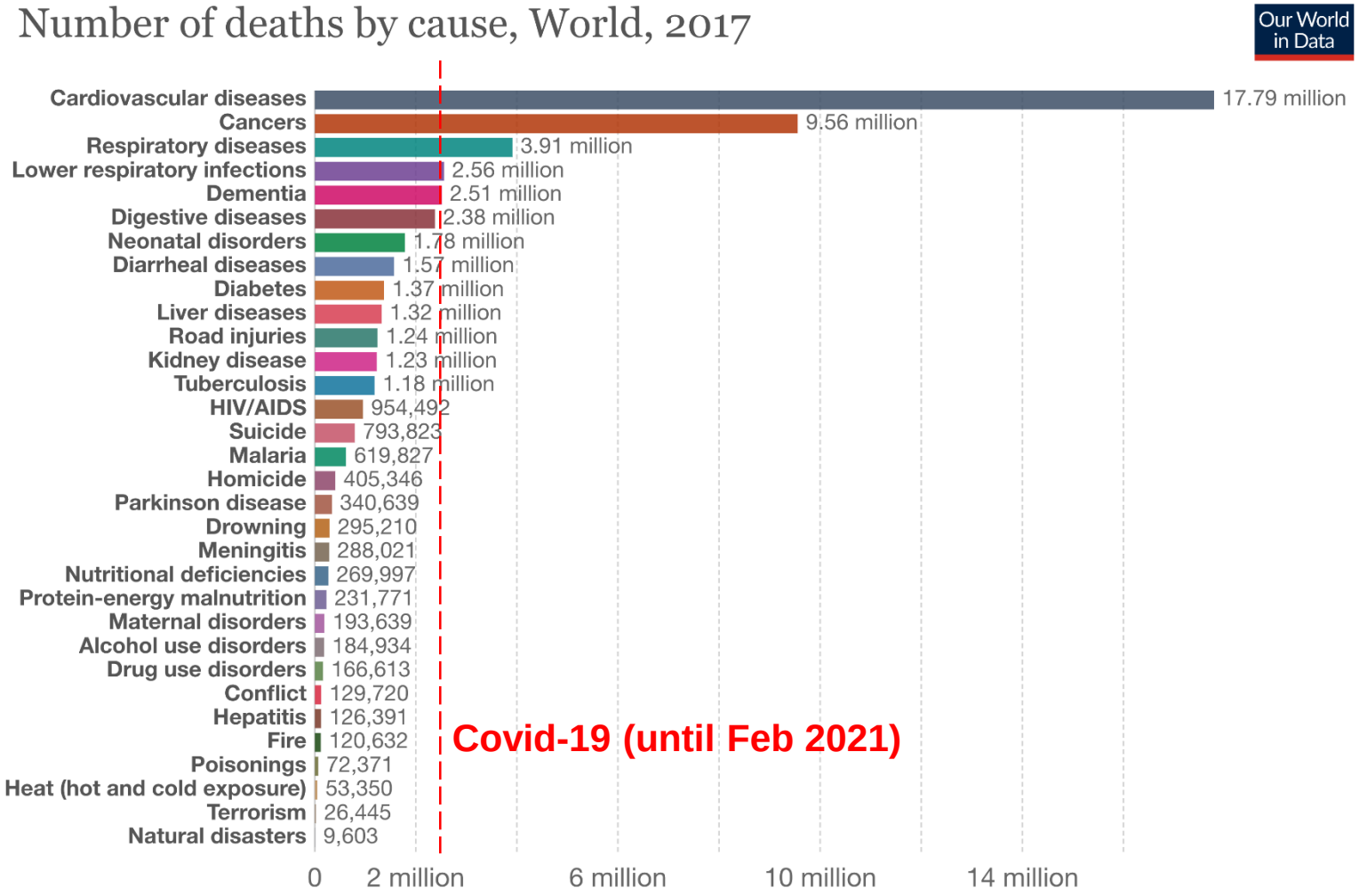
Number of measles cases per 100,000 people



Source: http://graphics.wsj.com/infectious-diseases-and-vaccines/?mc_cid=711ddeb86e

Trustworthy? Accessible? Elegant?

Fachhochschule Dortmund
University of Applied Sciences and Arts



Deadlines for Submission and Distribution of Grading

Fachhochschule Dortmund
University of Applied Sciences and Arts

Student task	Deliverables	Deadline	Work	Share of grade
W01 Assignment	Code and results	Apr 26th	Team A	5.0%
W02 Case Study	Code / presentation slides	May 15th	Team B	18.0%
W02 Case Study	Peer review*	May 24th	Individual	2.0%
W03 Assignment	Code and results	May 22nd	Team B	5.0%
W04 Assignment	Code and results	Jun 7th	Team C	10.0%
W05 Assignment	Code and results	Jun 14th	Team D	7.0%
W06 Assignment	Code and results	Jun 28th	Team D	13.0%
W07 Case Study	Code / presentation slides	Jul 12th	Team D	22.0%
W07 Case Study	Peer review*	Jul 26th	Individual	3.0%
DataCamp 1	Finish course	May 9th	Individual	2.5%
DataCamp 2	Finish course	May 30th	Individual	2.5%
DataCamp 3	Finish course	Jun 20th	Individual	2.5%
DataCamp 4	Finish course	Jul 11th	Individual	2.5%

* Peer review is mandatory. Quality of peer review itself is graded. Not providing peer review at all would result in high point deduction