Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

Elements of Machine Learning & Data Science

Introduction to Data Science

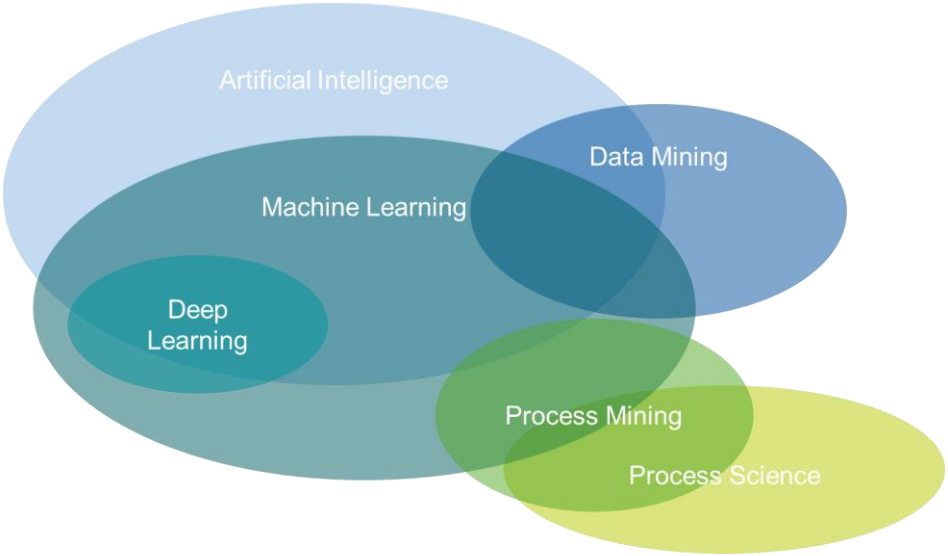
Lecture 6

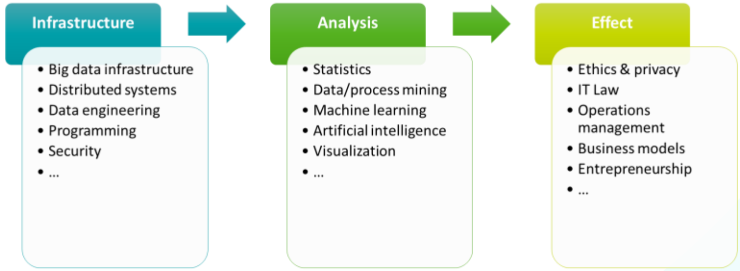
Prof. Wil van der Aalst

Marco Pegoraro, M.Sc.
Leah Tacke genannt Unterberg, M.Sc.

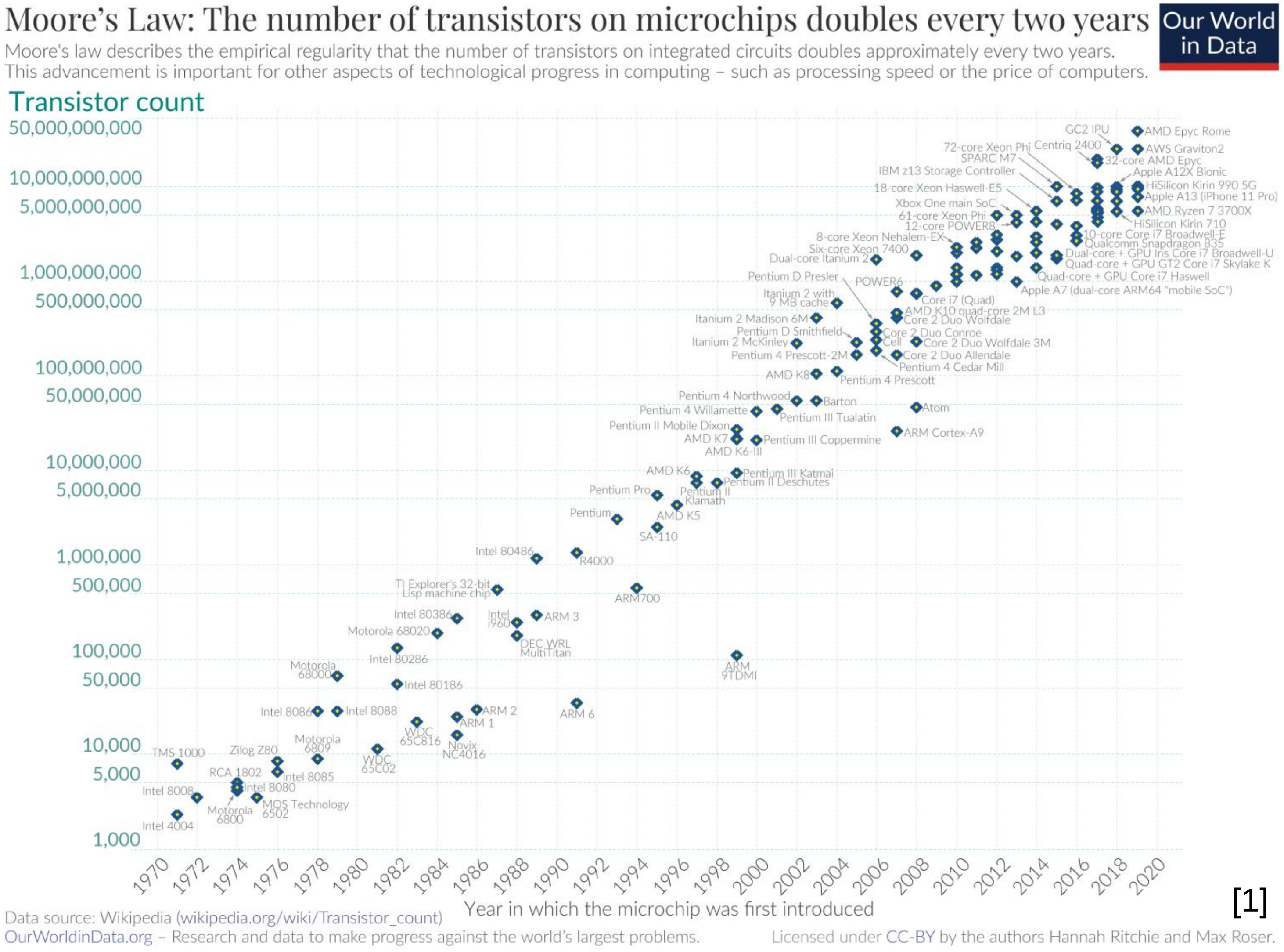
Outline

1. Introduction
2. Tabular Data
3. Data Science Process
4. Challenges
5. Data Types
6. Descriptive Statistics
7. Interpretative Pitfalls
8. Basic Visualizations
9. Feature Transformations
10. “How to lie with statistics”



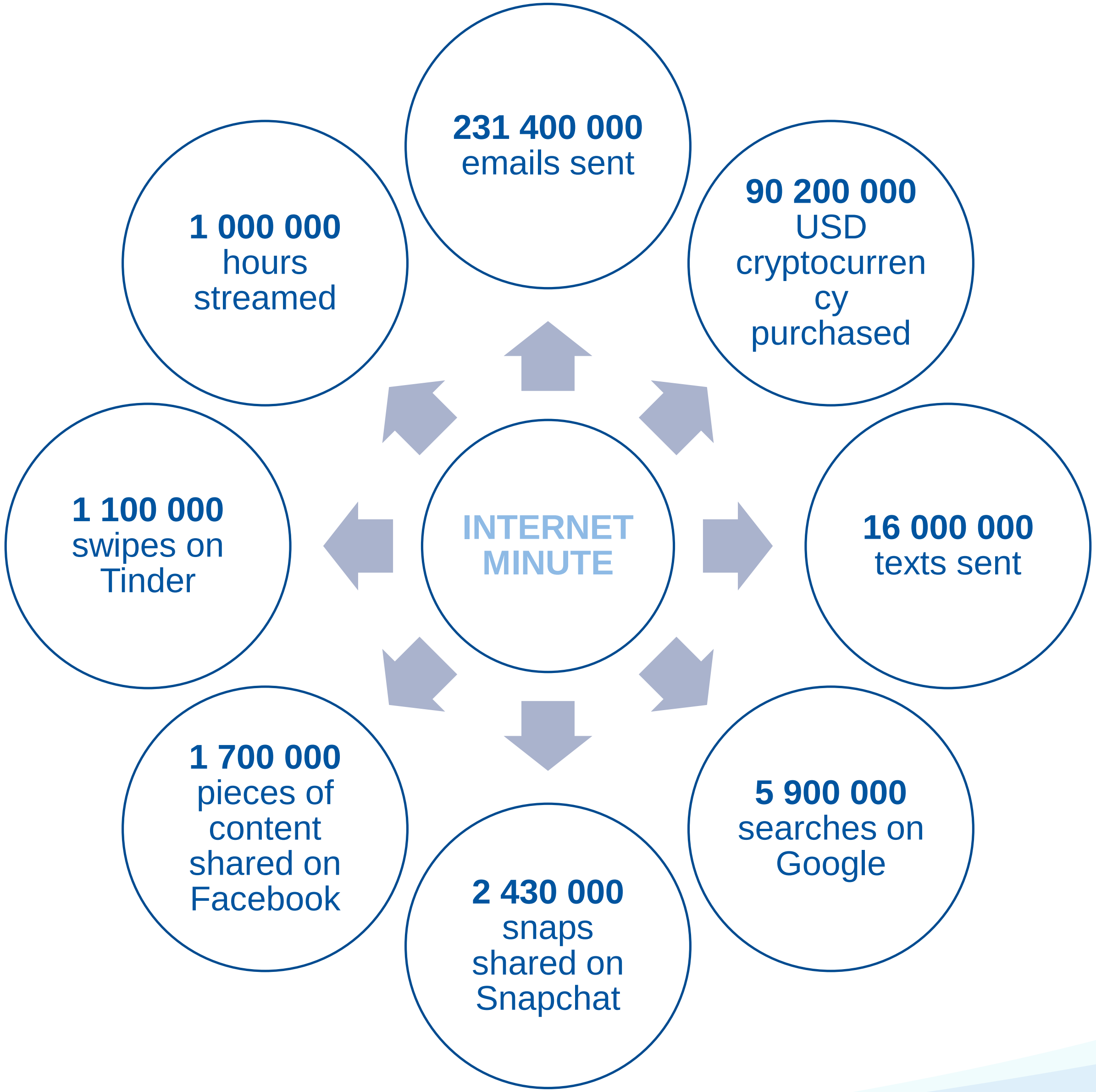


Motivation – Impact and Size of Data



[1]

Motivation – Impact and Size of Data



[2] Statista, as of 27.03.2023

Motivation – Data Scientist

data scientist

The Data Science Pipeline

Infrastructure

- Big data infrastructure
- Distributed systems
- Data engineering
- Programming
- Security
- …

Analysis

- Statistics
- Data/process mining
- Machine learning
- Artificial intelligence
- Visualization
- …

Effect

- Ethics & privacy
- IT Law
- Operations management
- Business models
- Entrepreneurship
- …

The Data Science Pipeline

Infrastructure → **Analysis** → **Effect**

- Big data infrastructure
- Distributed systems
- Data engineering
- Programming
- Security
- …

Challenge:
making things
scalable & instant

The Data Science Pipeline

Infrastructure

Analysis

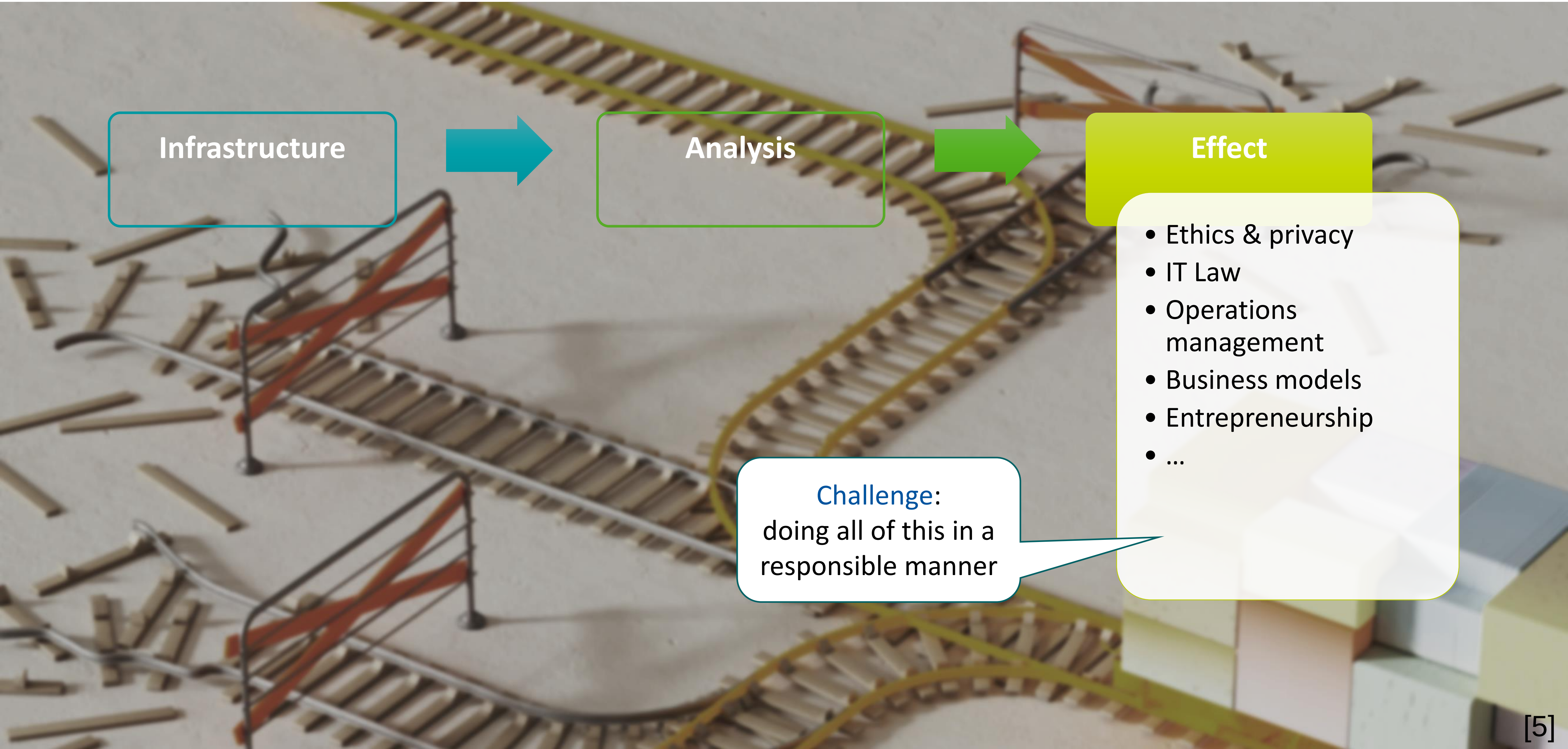
- Statistics
- Data/process mining
- Machine learning
- Artificial intelligence
- Visualization
- …

Effect

Challenge:
providing answers
to known &
unknown unknowns

[4]

The Data Science Pipeline



Infrastructure → Analysis → **Effect**

- Ethics & privacy
- IT Law
- Operations management
- Business models
- Entrepreneurship
- …

Challenge:
doing all of this in a responsible manner

[5]

The Data Science Pipeline

Focus of this course

Infrastructure

- Big data infrastructure
- Distributed systems
- Data engineering
- Programming
- Security
- …

Analysis

- Statistics
- Data/process mining
- Machine learning
- Artificial intelligence
- Visualization
- …

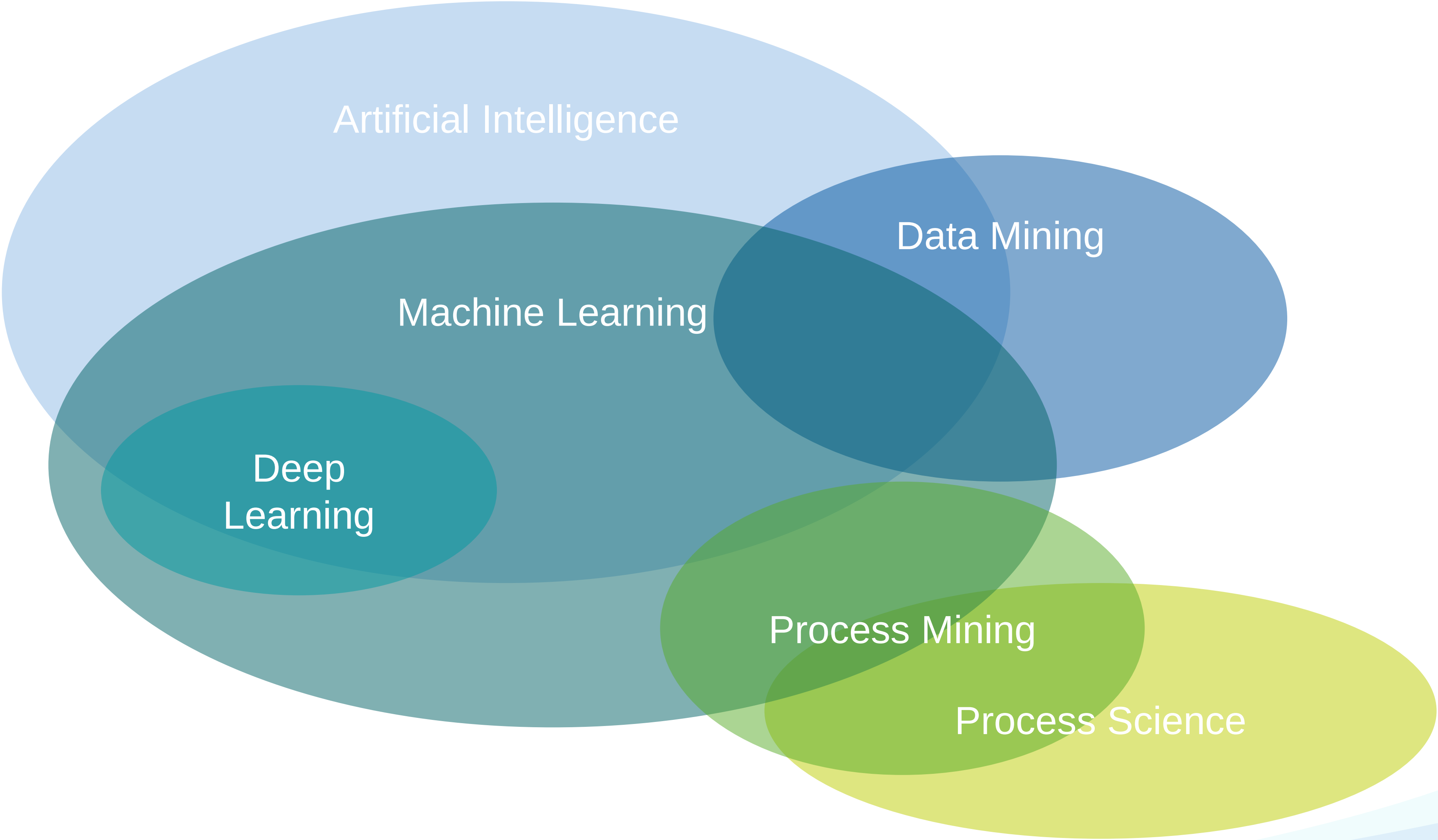
Effect

- Ethics & privacy
- IT Law
- Operations management
- Business models
- Entrepreneurship
- …

Terminology

- Many different names (statistics, data analytics, data mining, machine learning, artificial intelligence, predictive analytics, process mining, etc.) are used to refer to the key disciplines that contribute to data science
- Unfortunately, the areas these names describe are heavily overlapping and context dependent

Terminology



Data Science: A Definition

“Data science is an interdisciplinary field aiming to turn data into real value. Data may be structured or unstructured, big or small, static or streaming. Value may be provided in the form of predictions, automated decisions, models learned from data, or any type of data visualization delivering insights. Data science includes data extraction, data preparation, data exploration, data transformation, storage and retrieval, computing infrastructures, various types of mining and learning, presentation of explanations and predictions, and the exploitation of results taking into account ethical, social, legal, and business aspects.”

Page 10, Wil van der Aalst. Process Mining: Data Science in Action. Springer-Verlag, Berlin, 2016.

What actually is the *Data* in Data Science?

Example

- A restaurant owner wants to analyze the performance of their menu items …
- You have collected the following data:

price	calories	vegetarian	spicy	bestseller
12.99	800	Yes	No	Yes
9.99	600	Yes	Yes	No
14.99	1000	No	Yes	No
11.99	700	No	No	Yes
8.99	500	Yes	No	No

Features

- Features are raw or derived (mean, median, max, min, rank, etc.)
- Time is a special feature:
  - It cannot decrease
  - We often want to predict the future based on the past
  - Vital in temporal data analysis (time series data, event data, sequential data, …)

Example – Unlabeled Data

Unlabeled – no target feature selected

features

instances

price	calories	vegetarian	spicy	bestseller
12.99	800	Yes	No	Yes
9.99	600	Yes	Yes	No
14.99	1000	No	Yes	No
11.99	700	No	No	Yes
8.99	500	Yes	No	No

Example – Labeled Data

Labeled – designated target feature

descriptive features

target feature (class label)

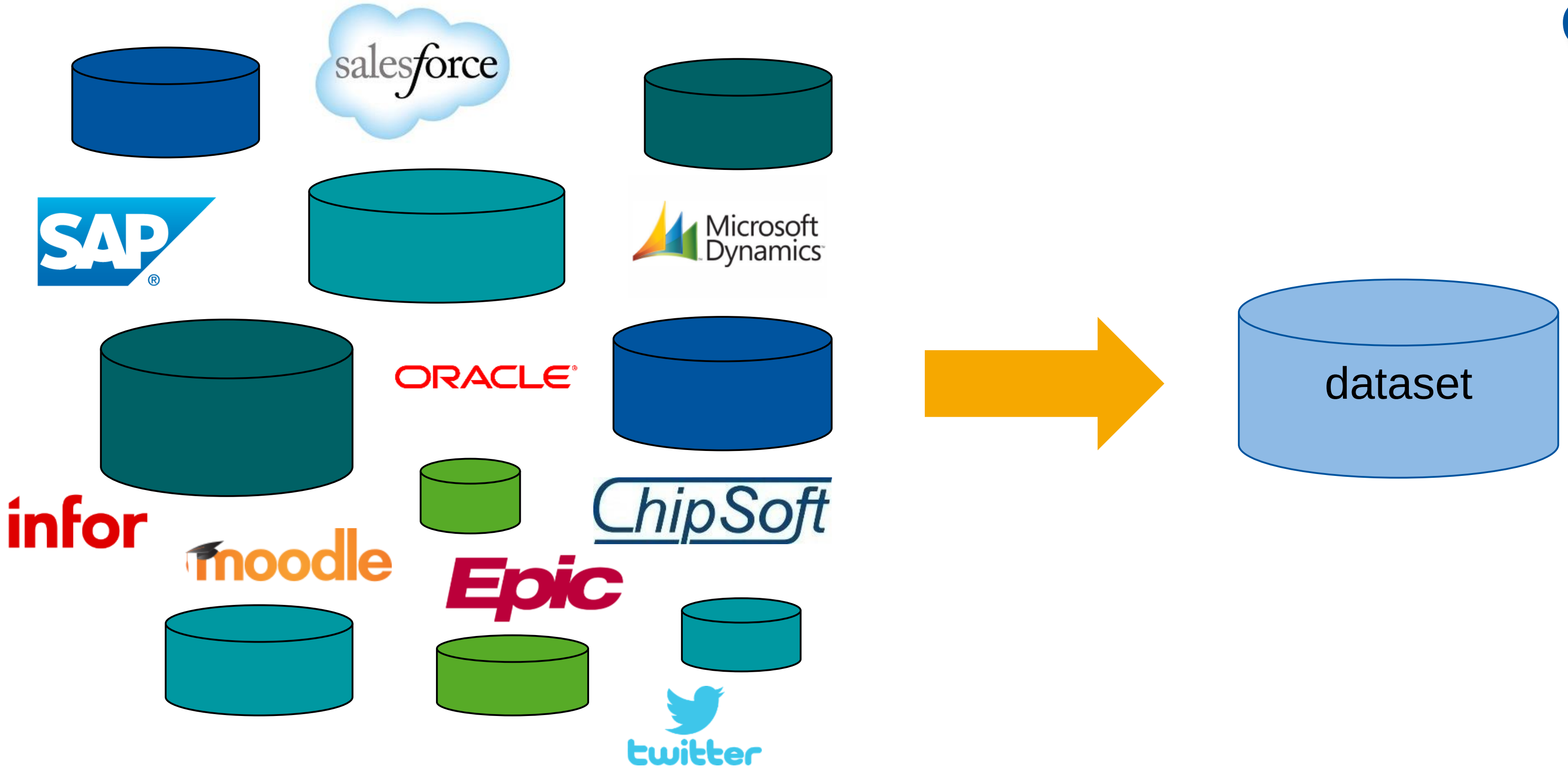
price	calories	vegetarian	spicy	bestseller
12.99	800	Yes	No	**Yes**
9.99	600	Yes	Yes	**No**
14.99	1000	No	Yes	**No**
11.99	700	No	No	**Yes**
8.99	500	Yes	No	**No**

instances

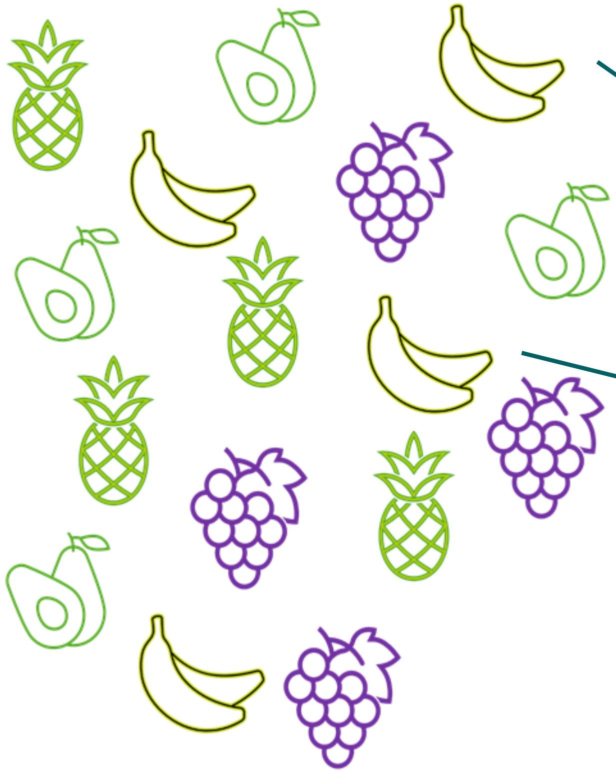
features

Extracting Data

80/20



Feature Extraction



fruit	color	weight [g]	…
Banana	Yellow	128	…
Avocado	Green	236	…
Banana	Yellow	176	…
Grapes	Purple	567	…
…	…	…	…

Example – Instances and Features

- Rows – instances
- Columns – features

features

instances

price	calories	vegetarian	spicy	bestseller
12.99	800	Yes	No	Yes
9.99	600	Yes	Yes	No
14.99	1000	No	Yes	No
11.99	700	No	No	Yes
8.99	500	Yes	No	No

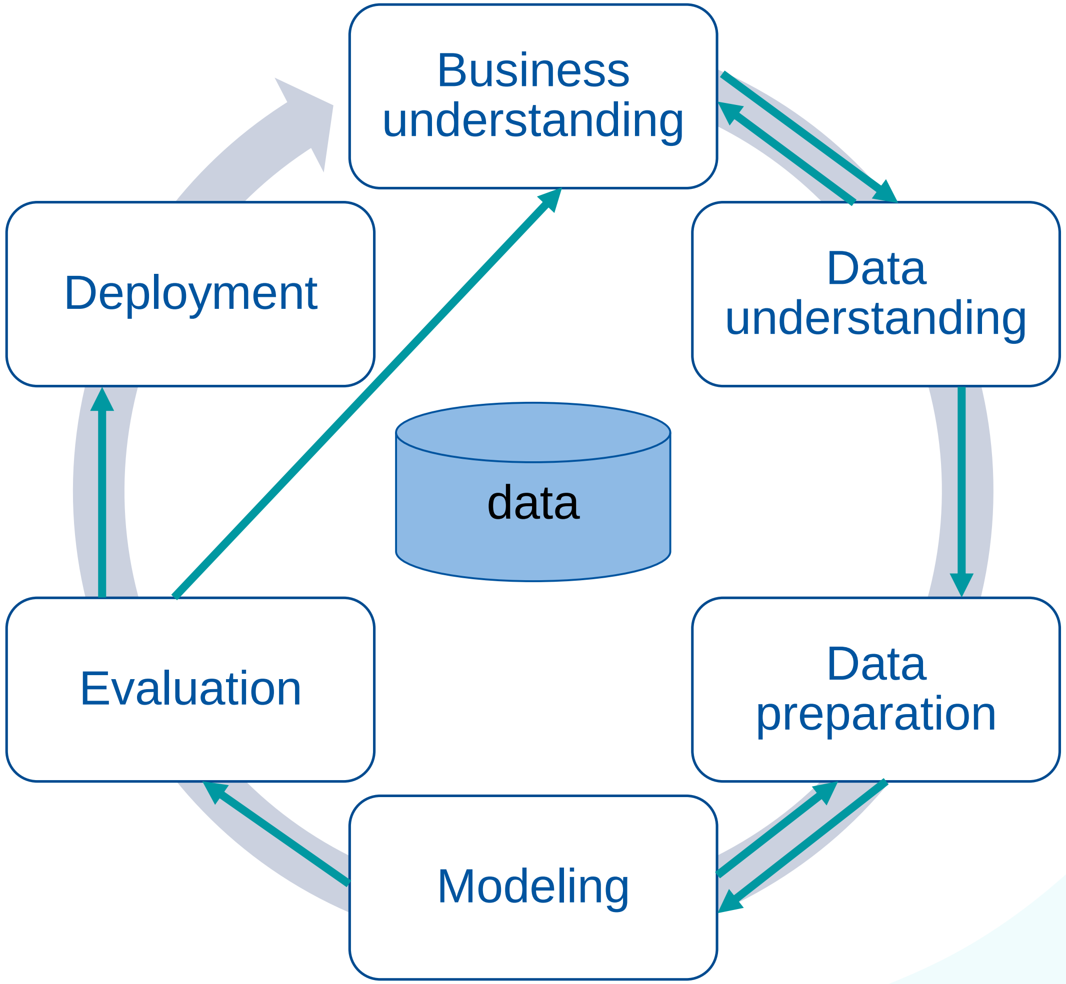
Data Science Is Complex and Requires a Structured Approach

“Data science is an interdisciplinary field aiming to turn data into real value. Data may be structured or unstructured, big or small, static or streaming. Value may be provided in the form of predictions, automated decisions, models learned from data, or any type of data visualization delivering insights. Data science includes data extraction, data preparation, data exploration, data transformation, storage and retrieval, computing infrastructures, various types of mining and learning, presentation of explanations and predictions, and the exploitation of results taking into account ethical, social, legal, and business aspects.”

Wil van der Aalst. Process Mining: Data Science in Action. Springer-Verlag, Berlin, 2016.

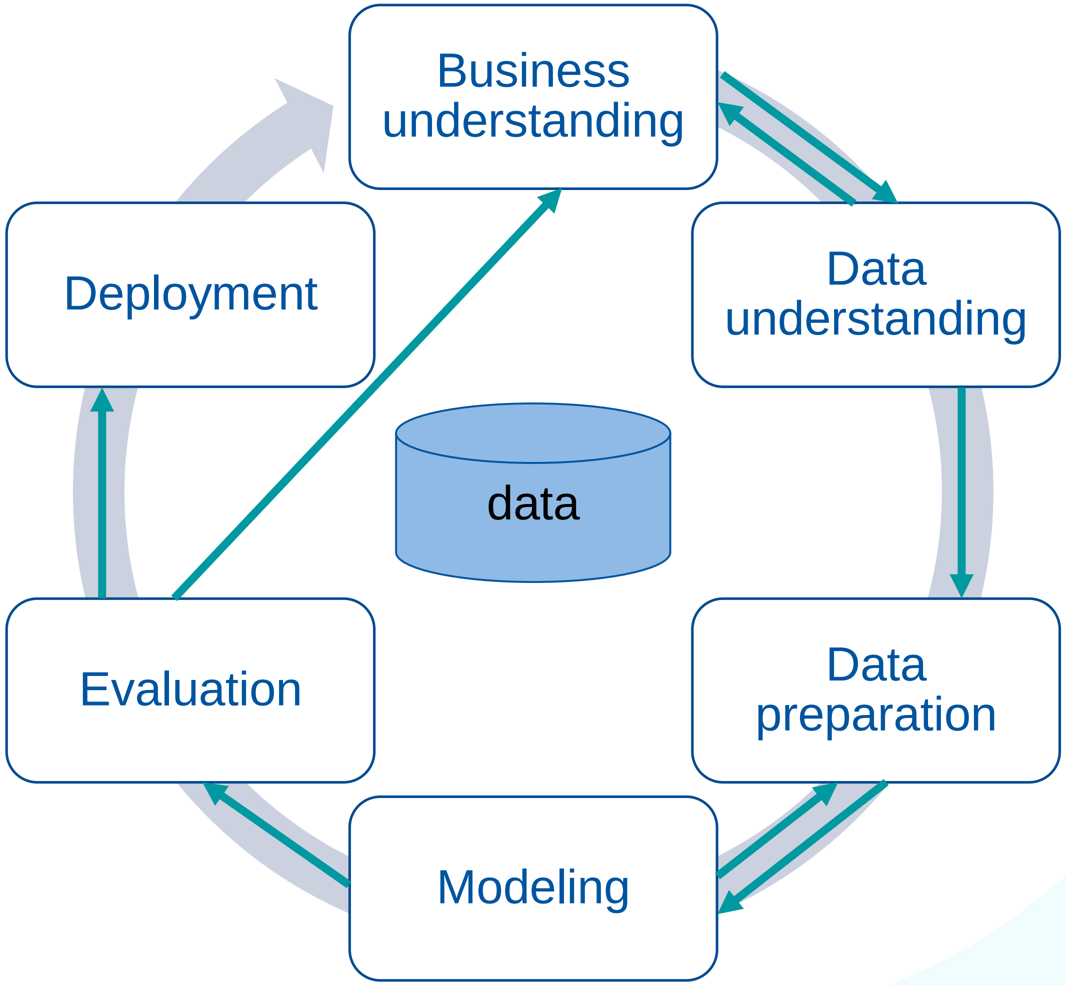
Cross-Industry Standard Process for Data Mining (CRISP-DM)

- Developed in the late 90s
- Its structure is quite obvious
- Details: Pete Chapman (1999) ‘The CRISP-DM User Guide’
- Any similar life-cycle models



Cross-Industry Standard Process for Data Mining (CRISP-DM)

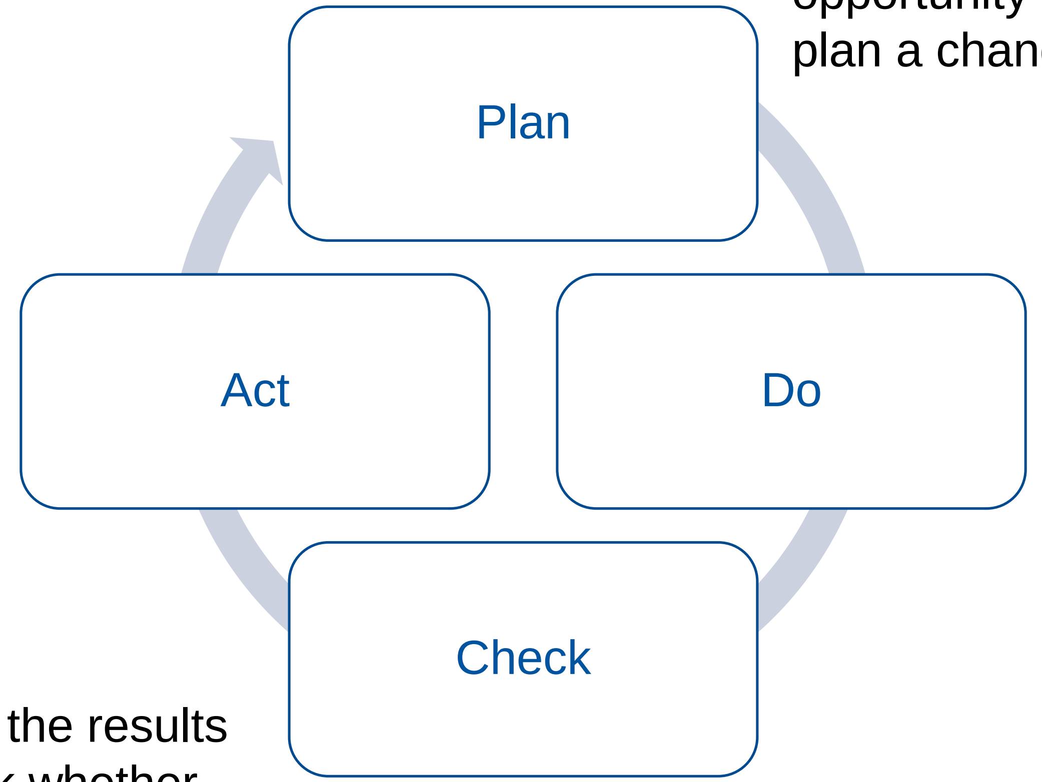
1. Business understanding – What does the organization need?
2. Data understanding – What data do we have?
3. Data preparation – How do we prepare the data for analysis?
4. Modeling – What modeling techniques should we apply?
5. Evaluation – Which model best meets the business objectives?
6. Deployment – How do stakeholders access and use the results?



Plan-Do-Check-Act (PDCA)

Recognize an opportunity and plan a change.

Implement the change



Test the change. Carry out a small-scale study.

Review the results to check whether the change helped

Also known as the Shewhart Cycle or Deming Cycle

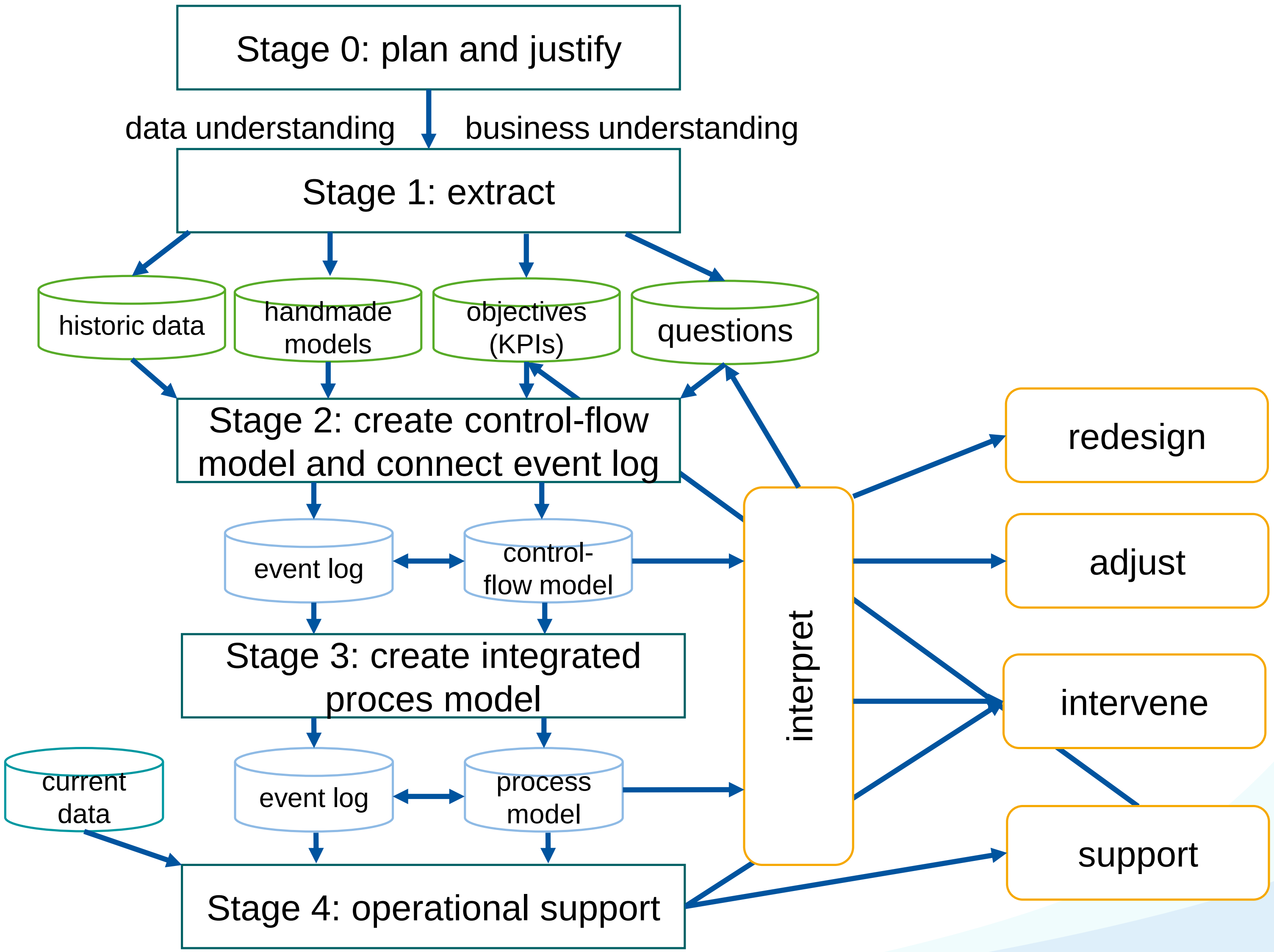
Define-Measure-Analyze-Improve-Control (DMAIC)

Define	Measure	Analyze	Improve	Control
• Launch team • Establish charter • Plan project • Gather VOC/VOB • Plan for change	• Document the process • Collect baseline data • Narrow project focus	• Analyze data • Identify root causes • Indentify and remove waste	• Generate solutions • Evaluate solutions • Optimize solutions • Pilot • Plan and implement	• Control the proces • Validate project benefits

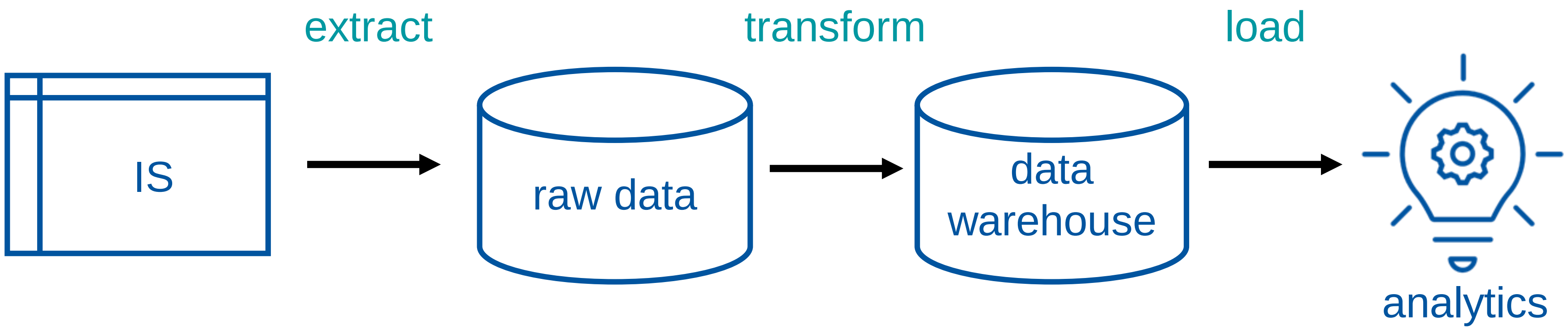
Often used as part of the Six Sigma methodology

L* Lifecycle Model

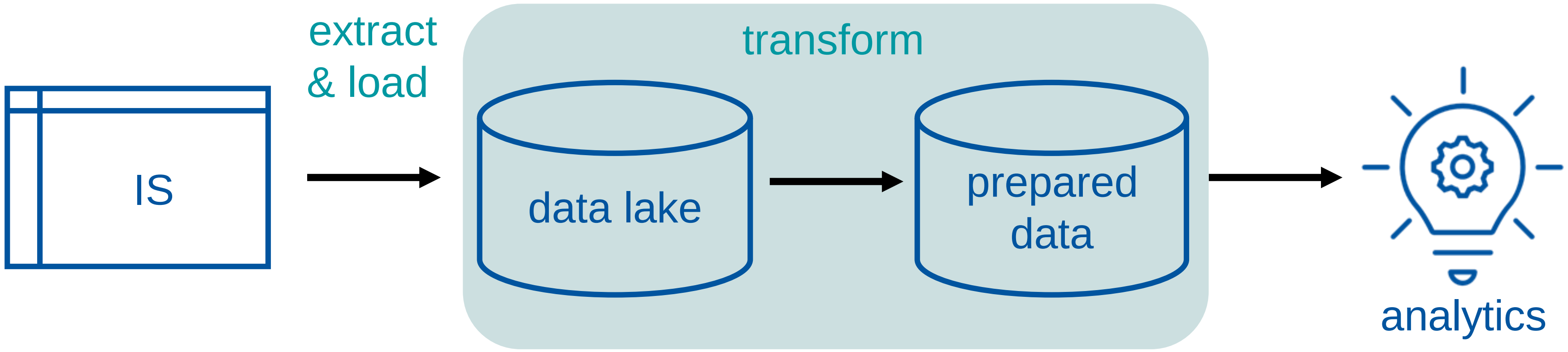
Specific for process mining



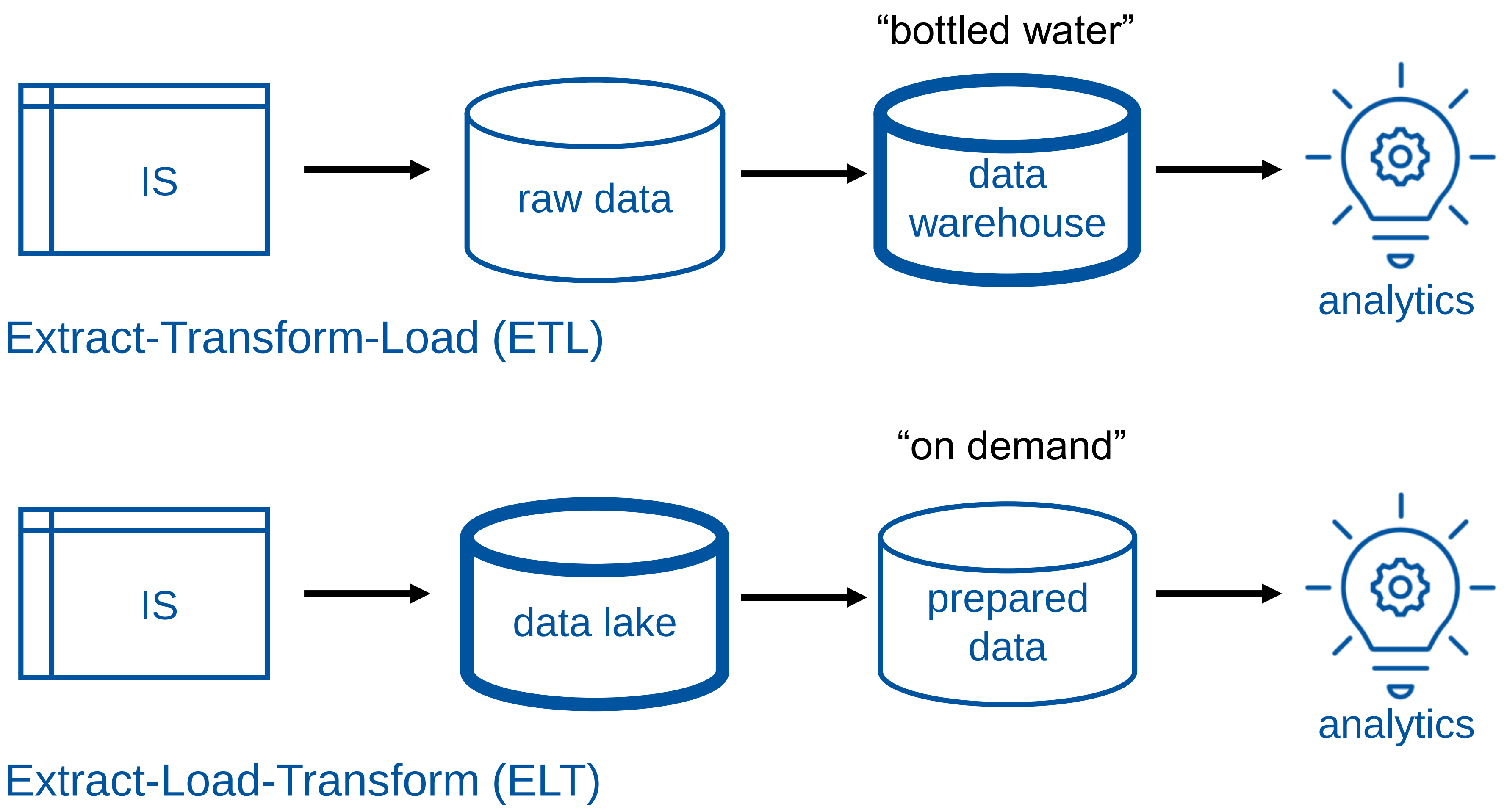
Extract-Transform-Load (ETL)



Extract-Load-Transform (ELT)



Differences



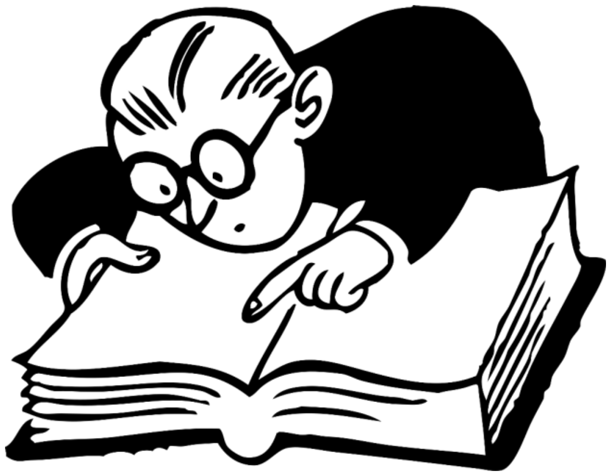
Organizational Issues

- Project or a continuous effort?
- Involve all stakeholders (users, customers, process owners, managers, board level, etc.)
- Positive Return-on-Investment (ROI) requires actionable insights
- Prepare for resistance (privacy concerns, data quality excuses, fear of transparency, etc.)
- Requires change management

Important, but … our focus will be on data science techniques

Finding Data

- There may be hundreds or thousands of tables
- There may exist many different entities that are less or not at all relevant

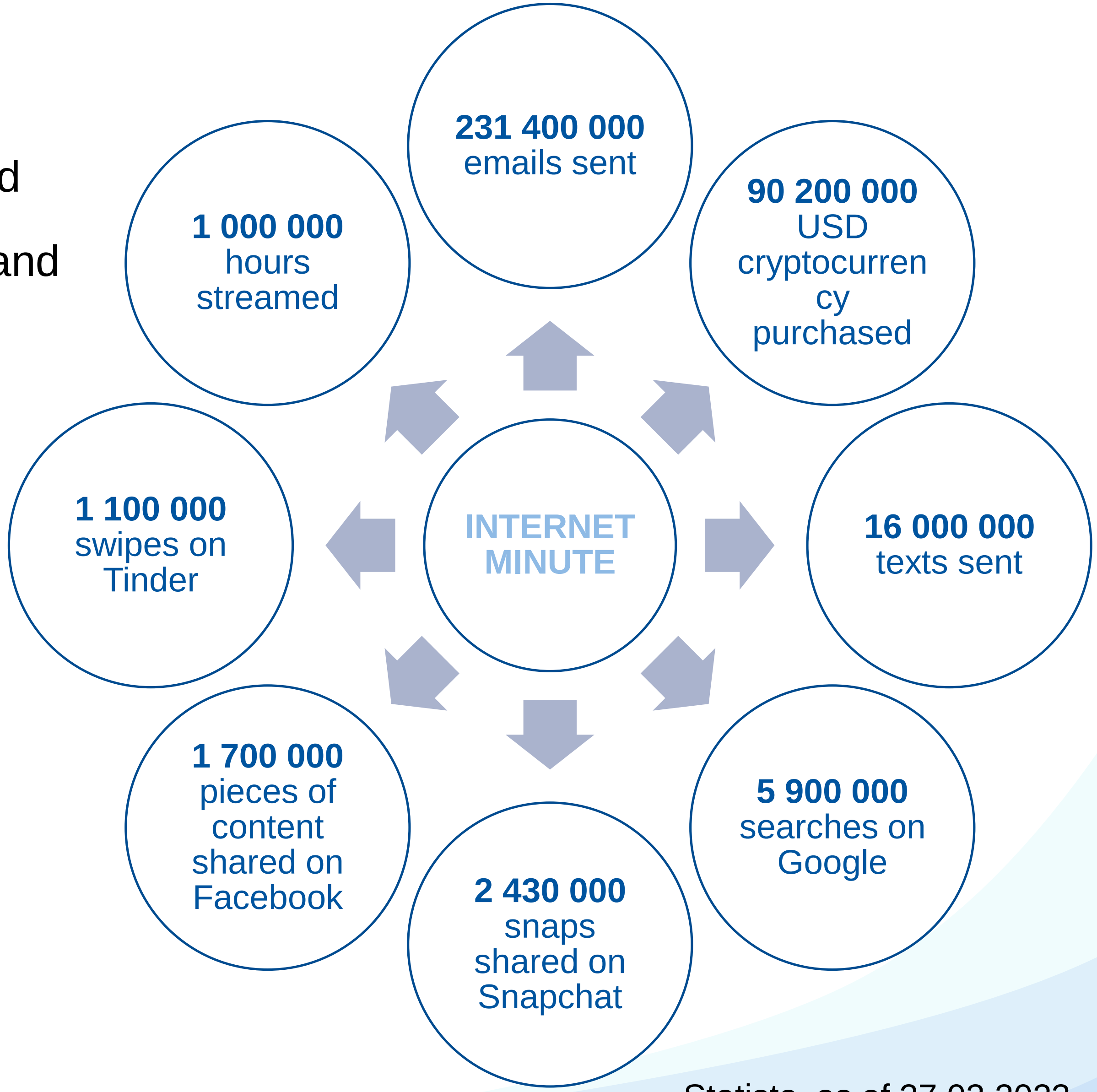


Preparing the Data

- Reorganizing data, filtering data, etc.
- Extracting relevant features
- Normalization (elimination of the effects of varying scales and units in different features, allowing for more accurate comparisons)
- Sampling (making data smaller or removing/changing a sample bias)

Big Data

- Lots of data (e.g. transactions) are recorded
- Need to have the ability to save, compare and analyze the collected data
- Requires distribution and concurrency



Statista, as of 27.03.2023

Streaming Data

- Data is generated continuously and processed in real-time
- Data is not stored in a database for later analysis
- Challenge: processing the data in real-time, need to handle the volume and velocity

Streaming Data

- Data is generated continuously and processed in real-time
- Data is not stored in a database for later analysis
- Challenge: processing the data in real-time, need to handle the **volume** and **velocity**

Source: De Agostini Editorial/Getty Images

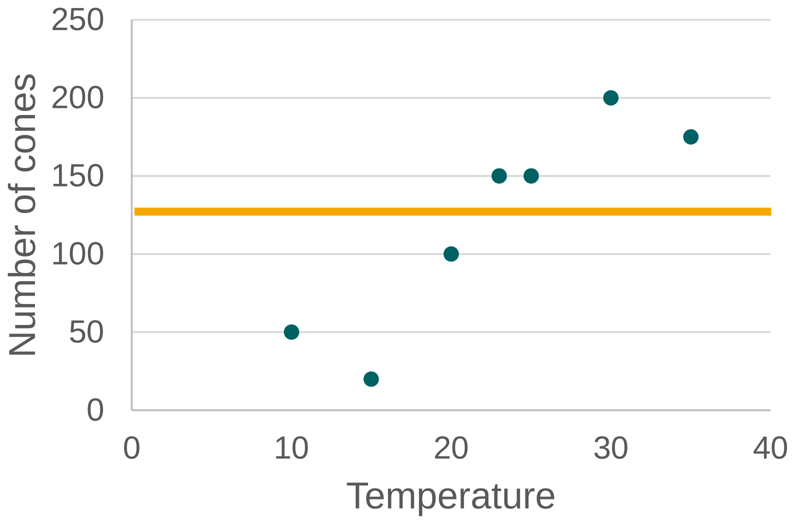
Source: NatGeo

Data Quality

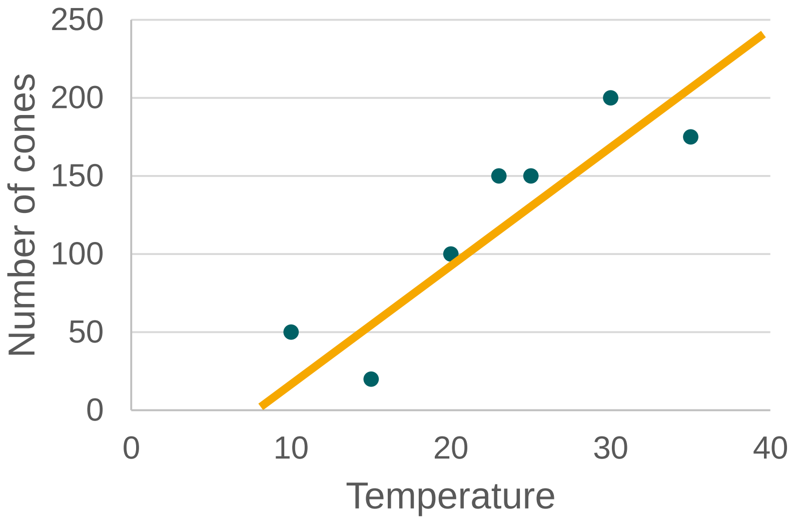
- Data may be:
  - Incomplete
  - Invalid
  - Inconsistent
  - Imprecise
  - Outdated
- Challenge: detecting and handling such issues

Overfitting and Underfitting

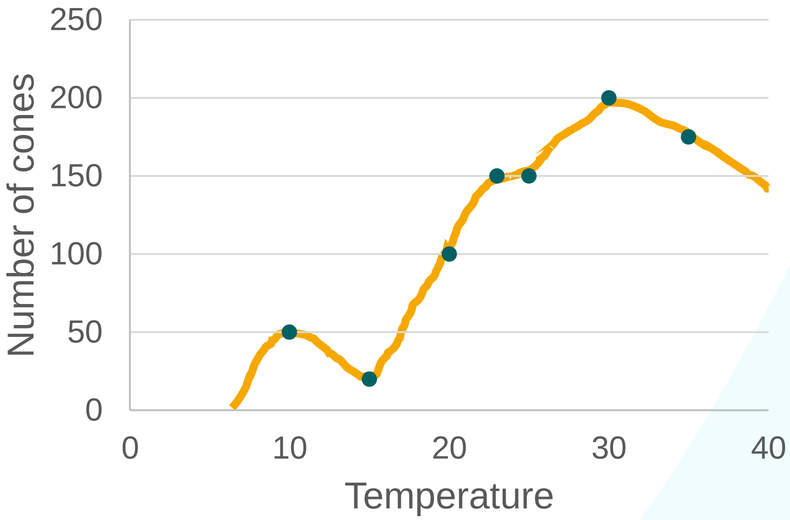
Underfitting



Optimal



Overfitting



Concept Drift

- Properties of the data change over time and thus the performance of a model decreases
- The data that the model is trained on no longer represents the real-world data
- Challenge: when to update the model with new data

Turning Insights into Action

- Predicting the inevitable does not help much
- What can be influenced?
- Is there still time?

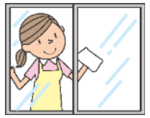
Concerns – Responsible Data Science

- Responsible Data Science advocates the development of techniques, algorithms, tools, laws, ethical/social principles for ensuring fairness, accuracy, confidentiality and transparency covering the whole data science pipeline

Concerns – Responsible Data Science

- Responsible Data Science advocates the development of techniques, algorithms, tools, laws, and ethical/social principles for ensuring fairness, accuracy, confidentiality and transparency covering the whole data science pipeline


- Fairness: How to avoid unfair conclusions even if they are true?
- Accuracy: How to answer questions with a guaranteed level of accuracy?
- Confidentiality: How to answer questions without revealing secrets?
- Transparency: How to clarify answers such that they become indisputable?

Ill-posed Problems

- A problem is well-posed if
  - A solution exists
  - The solution is unique
- Problems in data science are often ill-posed:
  - Many possible models explaining observed phenomena
  - Data set is just a sample and does not represent the whole population
  - Noise in the data set
  - The result needs to generalize to have predictive or explanatory value

Data Types

Tabular Data

Feature values can have various types - knowing these data types is essential for correct data analysis and data processing!

features

Numerical feature

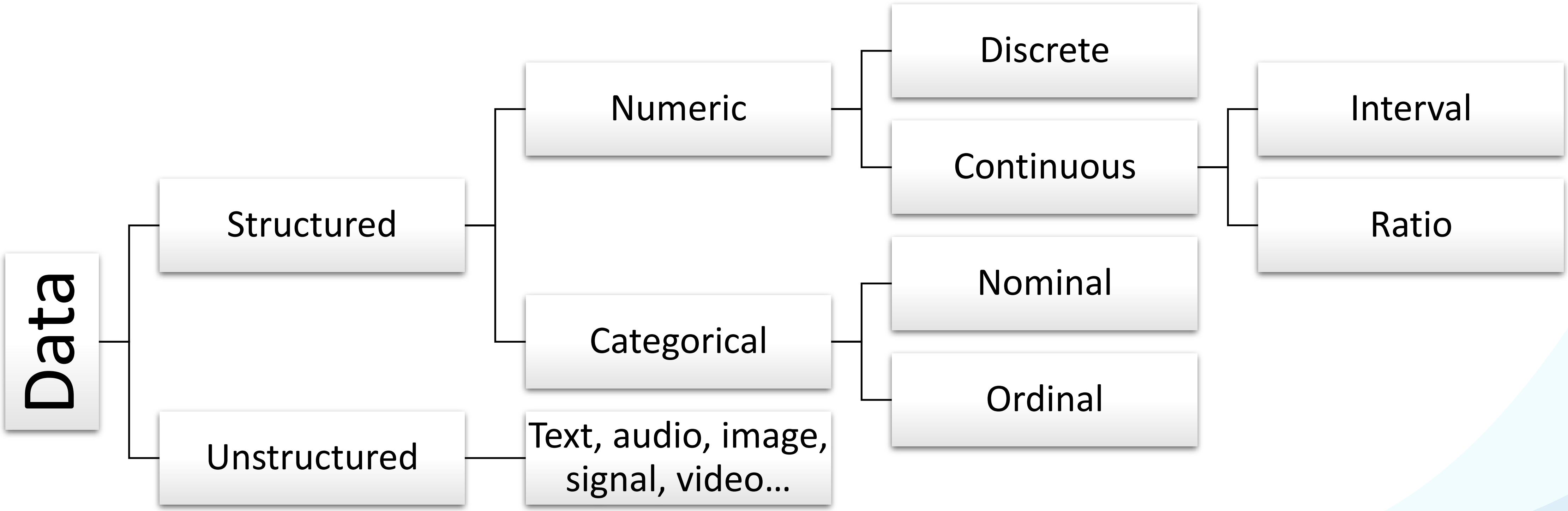
Categorical feature

instances

price	calories	vegetarian	spicy	bestseller
12.99	800	Yes	No	**Yes**
9.99	600	Yes	Yes	**No**
14.99	1000	No	Yes	**No**
11.99	700	No	No	**Yes**
8.99	500	Yes	No	**No**

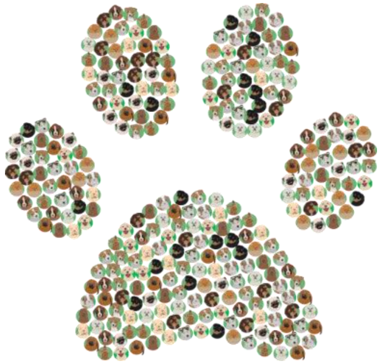
Data Types Overview

Feature values can have various types - knowing these data types is essential for correct data analysis and data processing!

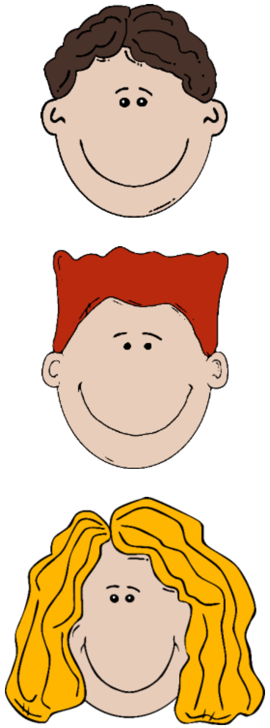


Data Types - Nominal

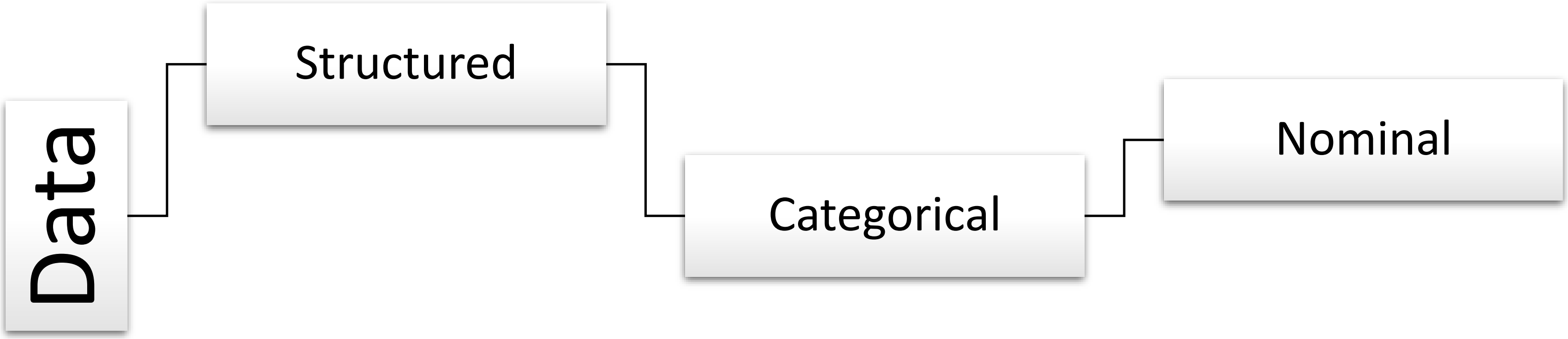
- Represents category, code or state
- Ordering of the values has no meaning
  (e.g., blonde hair is not better than brown hair)



dog breed



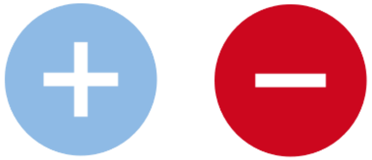
hair color



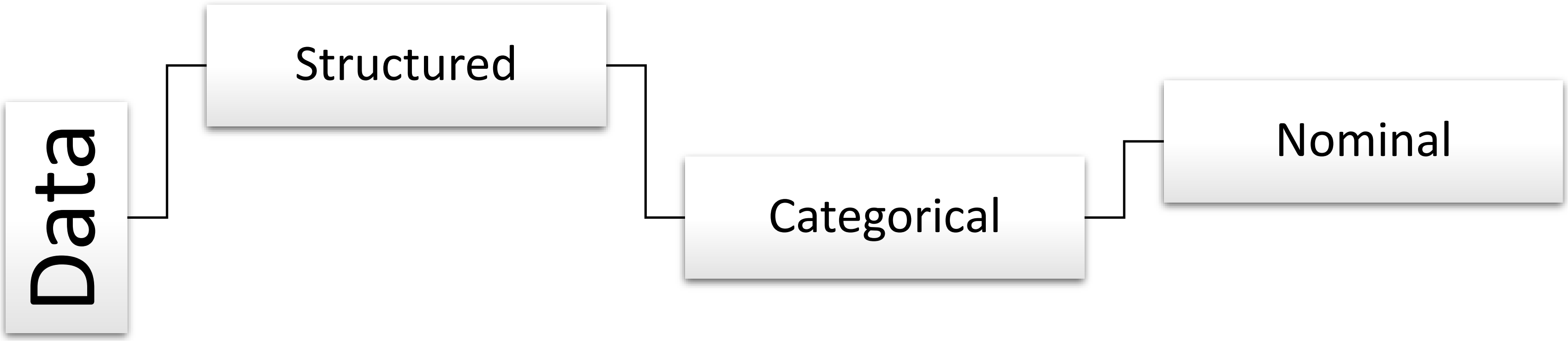
Data Types - Binary

- Special case of nominal: Binary
- Only two categories (often 0 and 1)
- Symmetric: both values are equal (subjectively or frequency based)
- Asymmetric: one value is normal/default, the other exception

light off/on



test status positive/negative

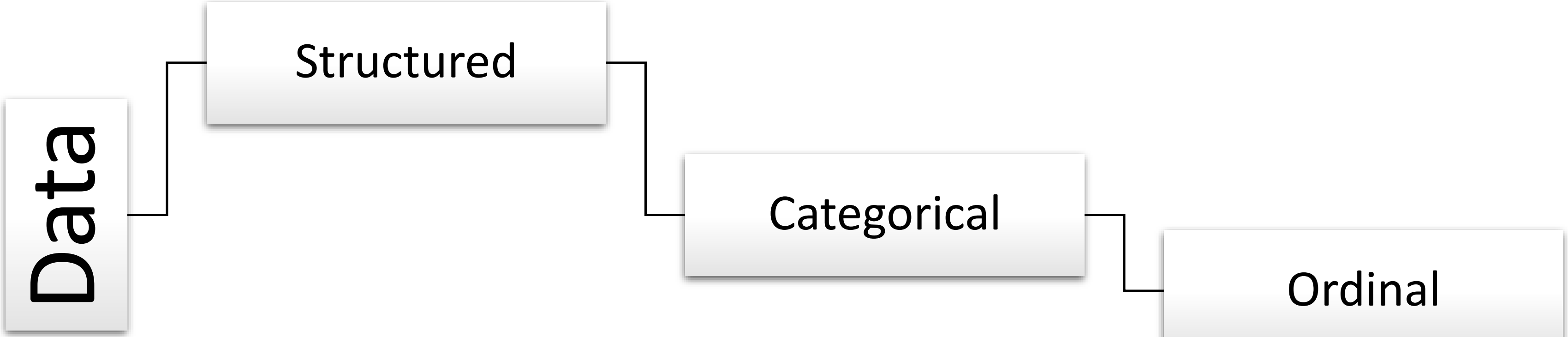


Data Types - Ordinal

- Values have a meaningful order
  - high, medium, low
  - excellent, good, satisfactory, poor
  - lightning fast, quick, slow
  - strongly agree, agree, indifferent, disagree, strongly disagree
- The difference between successive values cannot be quantified

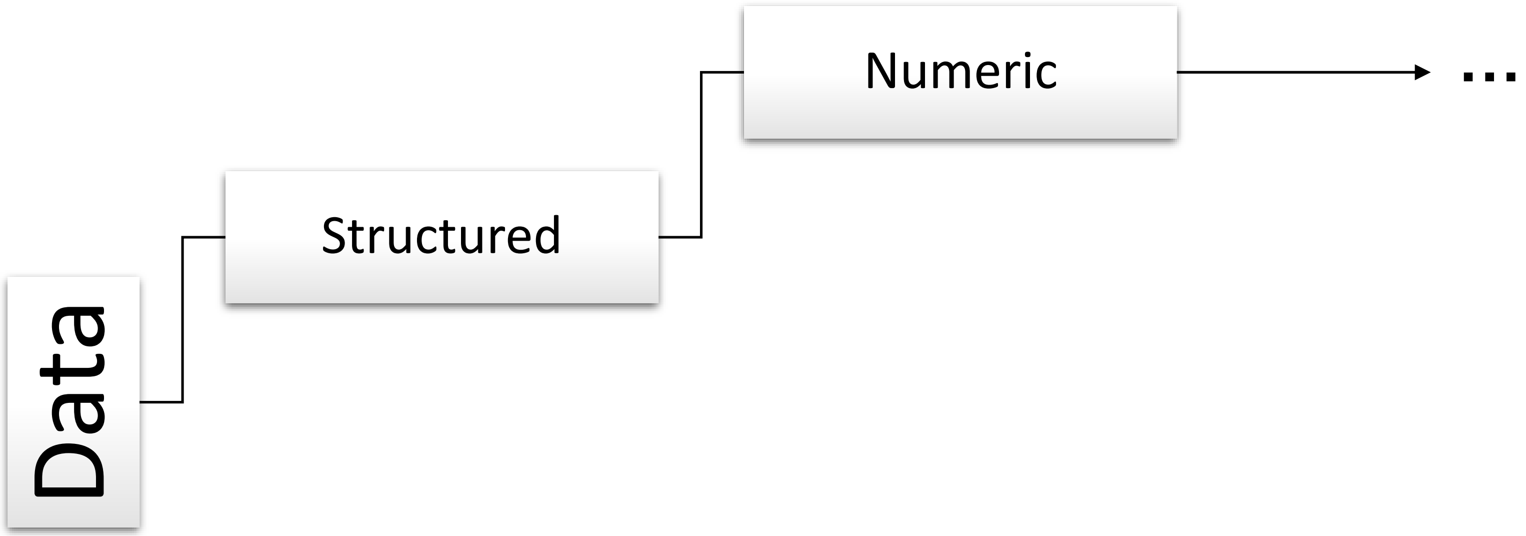
grades

customer satisfaction



Data Types - Numeric

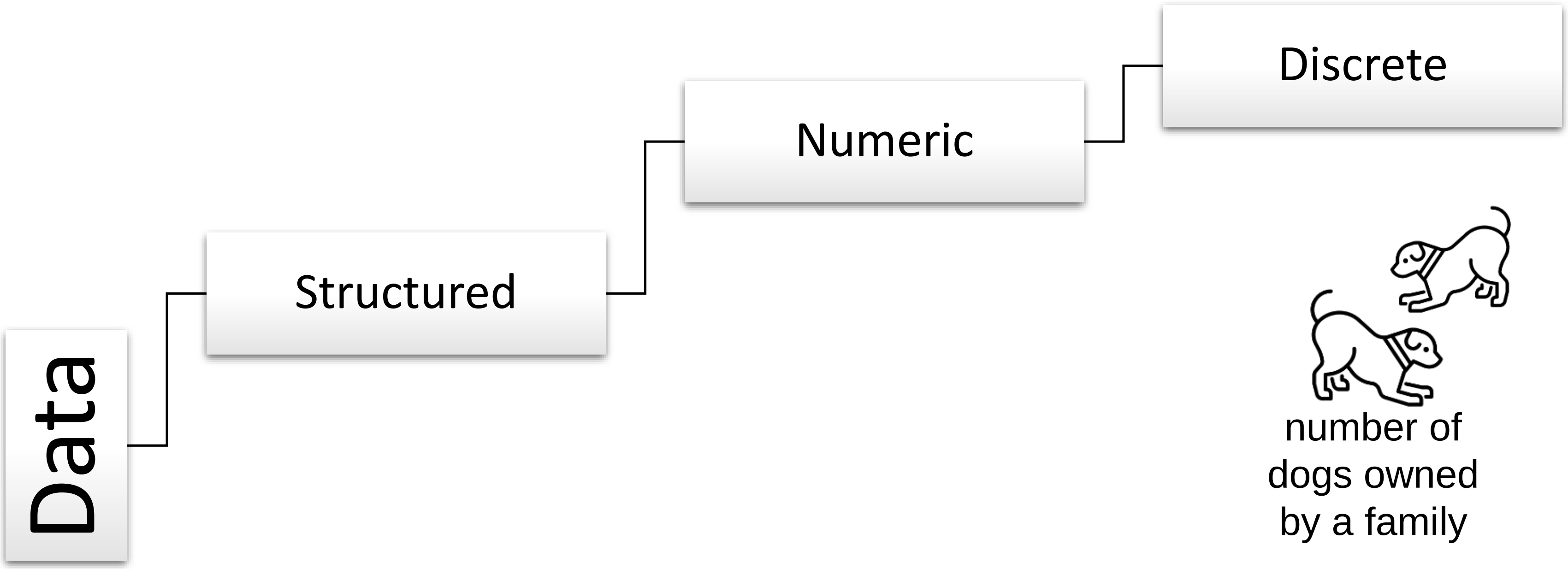
- Measurable quantities
- Differences can be quantified
- Mean, median, mode, variance, etc. can be computed



Data Types – Discrete

- Numeric
- Can be counted

number of rooms in a house



number of trees in a forest

number of dogs owned by a family

Data Types - Interval

- Scale of equal-sized units with quantifiable difference between the units
- A zero may not exist, values may go negative

coordinates

temperature
(Celsius, Fahrenheit)

Data

Structured

Numeric

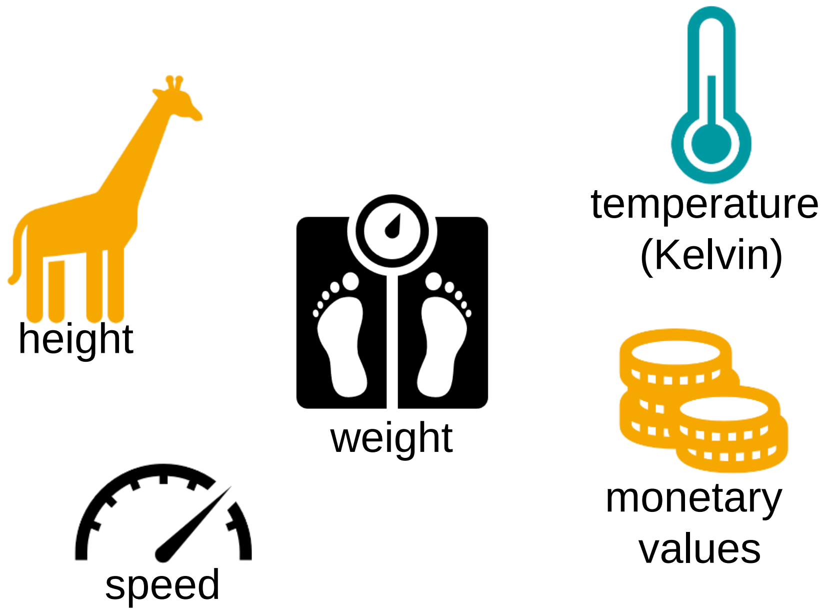
Continuous

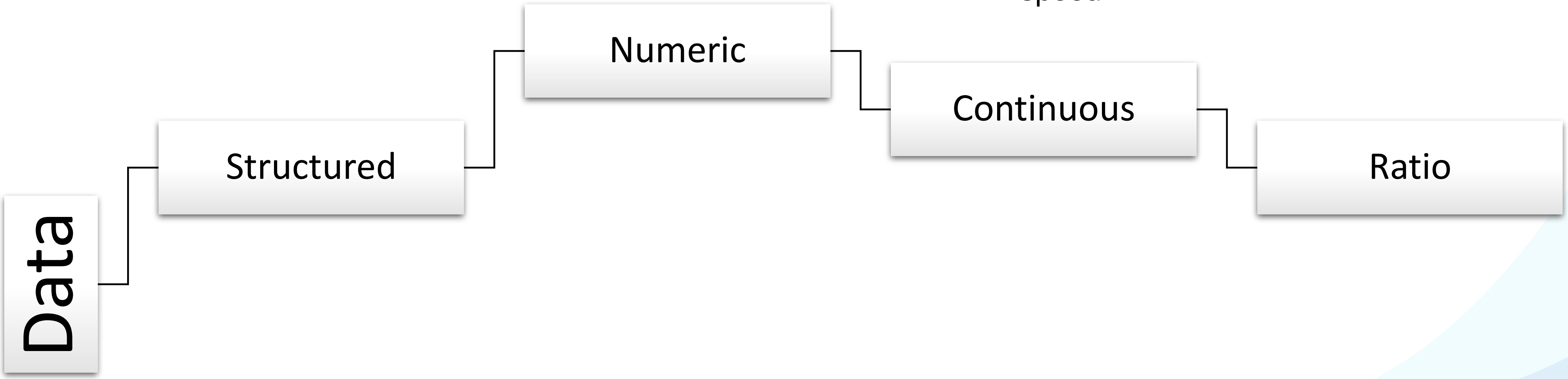
Interval

IQ

Data Types - Ratio

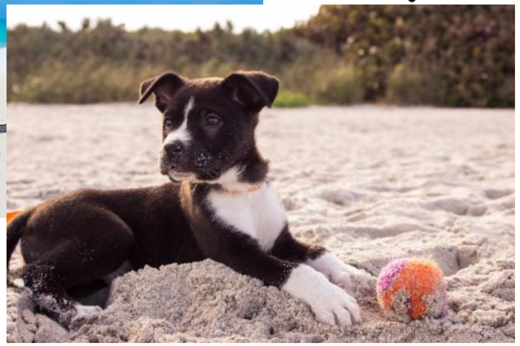
- Multiples/ratios can be identified (e.g., three times as heavy, four times as fast, etc.)
- The scale ends at zero (0 kg, 0 km/h, 0 Kelvin)





Data Types - Unstructured

- Text, audio, video, etc.
- Can be turned into structural data
- Examples: multiset of words or n-grams to describe a text, or pixel information for images



Data

Unstructured

Text, audio, image, signal, video…

Data Types - Unstructured

- Extremely prevalent in Big Data
- Huge opportunity for novel transformation/extraction approaches
  - e.g. NLP
- Misnomer, as data may be structured, just not to an appreciable degree under the current viewpoint

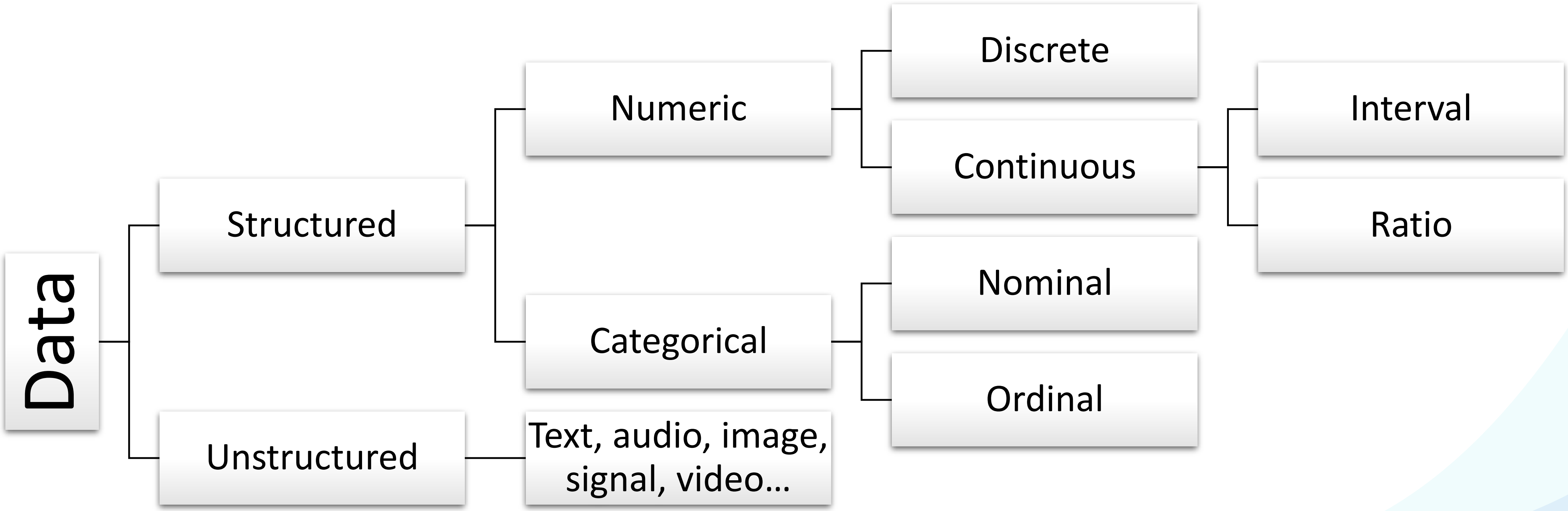
Data

Unstructured

Text, audio, image, signal, video…

Data Types Overview

- Data types are essential for correct data analysis and data processing!



Conclusion

- Data can be unstructured (e.g., text) but turned into e.g., vectors
- Most techniques are based on tabular data (especially the basic ones)
- The data type is vital for the correct data processing and analysis

price	calories	vegetarian	spicy	bestseller
12.99	800	Yes	No	**Yes**
9.99	600	Yes	Yes	**No**
14.99	1000	No	Yes	**No**
11.99	700	No	No	**Yes**
8.99	500	Yes	No	**No**

Descriptive Statistics Repetition

Individual Features - Continuous

x
1.5
2.7
3.1
4.2
5.5
6.9
7.6
8.1
9.3
10.0

Count = 10
Number of instances

Usually denoted by N in this course

Individual Features - Continuous

x
1.5
2.7
3.1
4.2
5.5
6.9
7.6
8.1
9.3
10.0

Count = 10
Number of instances

Cardinality = 10
Number of unique values

Individual Features - Continuous

x
1.5
2.7
3.1
4.2
5.5
6.9
7.6
8.1
9.3
10.0

Count = 10
Number of instances

Cardinality = 10
Number of unique values

Min = 1.5
Minimum value

Individual Features - Continuous

x
1.5
2.7
3.1
4.2
5.5
6.9
7.6
8.1
9.3
10.0

Count = 10
Number of instances

Cardinality = 10
Number of unique values

Min = 1.5
Minimum value

Max = 10.0
Maximum value

Individual Features - Continuous

x
1.5
2.7
3.1
4.2
5.5
6.9
7.6
8.1
9.3
10.0

Count = 10
Number of instances

Cardinality = 10
Number of unique values

Min = 1.5
Minimum value

Max = 10.0
Maximum value

$$\bar{x} = \frac{\sum_{n=1}^{N} x_n}{N}$$

Mean = 5.89
Sum of all values divided by count

Individual Features - Continuous

x
1.5
2.7
3.1
4.2
5.5
6.9
7.6
8.1
9.3
10.0

Count = 10
Number of instances

Cardinality = 10
Number of unique values

Min = 1.5
Minimum value

Max = 10.0
Maximum value

Mean = 5.89
Sum of all values divided by count

Median = 6.2
Middle value / mean of two middle values

Individual Features - Continuous

x
1.5
2.7
3.1
4.2
5.5
6.9
7.6
8.1
9.3
10.0

Count = 10
Number of instances

Cardinality = 10
Number of unique values

Min = 1.5
Minimum value

Max = 10.0
Maximum value

Mean = 5.89
Sum of all values divided by count

Median = 6.2
Middle value / mean of two middle values

Variance ≈ 8.621
Average squared distance of each value from the mean

$$var(x) = \frac{\sum_{n=1}^{N}(x_n - \bar{x})^2}{N-1}$$

Individual Features - Continuous

x
1.5
2.7
3.1
4.2
5.5
6.9
7.6
8.1
9.3
10.0

Count = 10
Number of instances

Cardinality = 10
Number of unique values

Min = 1.5
Minimum value

Max = 10.0
Maximum value

Mean = 5.89
Sum of all values divided by count

Median = 6.2
Middle value / mean of two middle values

Variance ≈ 8.621
Average squared distance of each value from the mean

Standard deviation ≈ 2.936
How spread out the data is
(the square root of the variance)

$$std(x) = \sqrt{var(x)}$$

Individual Features - Continuous

	x
10th	1.5
	2.7
25th	3.1
	4.2
50th	5.5
	6.9
	7.6
75th	8.1
	9.3
100th	10.0

Count = 10
Number of instances

Cardinality = 10
Number of unique values

Min = 1.5
Minimum value

Max = 10.0
Maximum value

Mean = 5.89
Sum of all values divided by count

Median = 6.2
Middle value / mean of two middle values

Variance ≈ 8.621
Average squared distance of each value from the mean

Standard deviation ≈ 2.936
How spread out the data is
(the square root of the variance)

p^{th} percentile

$x_n \text{ with } n = \lceil \frac{p}{100} \cdot N \rceil$

Value at or below (or strictly below) which p percent of the instances are located

10^{th} percentile = 1.5
25^{th} percentile = 3.1 — First quartile (Q_1)
50^{th} percentile = 5.5 — Second quartile (Q_2)
75^{th} percentile = 8.1 — Third quartile (Q_3)
100^{th} percentile = 10

Individual Features - Categorical

x
A
B
A
C
B
B
C
A
C
B

Count = 10

Number of instances

Individual Features - Categorical

x
A
B
A
C
B
B
C
A
C
B

Count = 10
Number of instances

Cardinality = 3
Number of unique values

Individual Features - Categorical

x
A
B
A
C
B
B
C
A
C
B

Count = 10
Number of instances

Cardinality = 3
Number of unique values

Mode = B
Value that appears most frequently

Multiple Features - Covariance

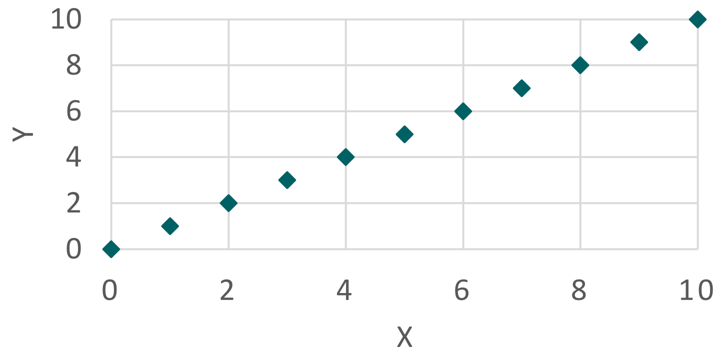
x	y
1.5	4.2
2.7	4.9
3.1	7.1
4.2	9.8
5.5	12.3
6.9	14.7
7.6	16.5
8.1	18.2
9.3	20.9
10.0	22.6

$$Cov(x,y) = \frac{1}{N-1}\sum_{n=1}^{N}((x_n - \bar{x}) \cdot (y_n - \bar{y}))$$

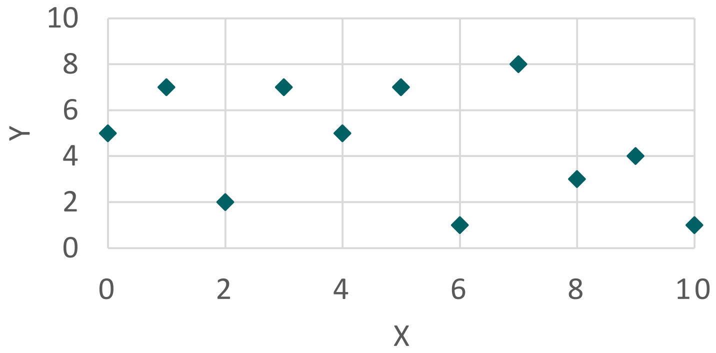
Cov(x,y) ≈ 19.134

$$+ \,\&\, + \Rightarrow +$$
$$+ \,\&\, - \Rightarrow -$$
$$- \,\&\, + \Rightarrow -$$
$$- \,\&\, - \Rightarrow +$$

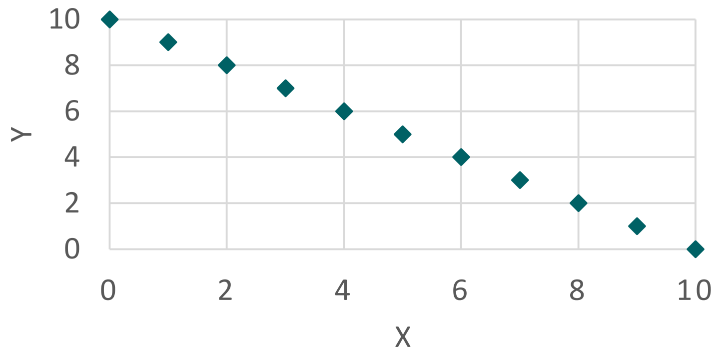
Multiple Features – Correlation



maximal positive correlation



no correlation



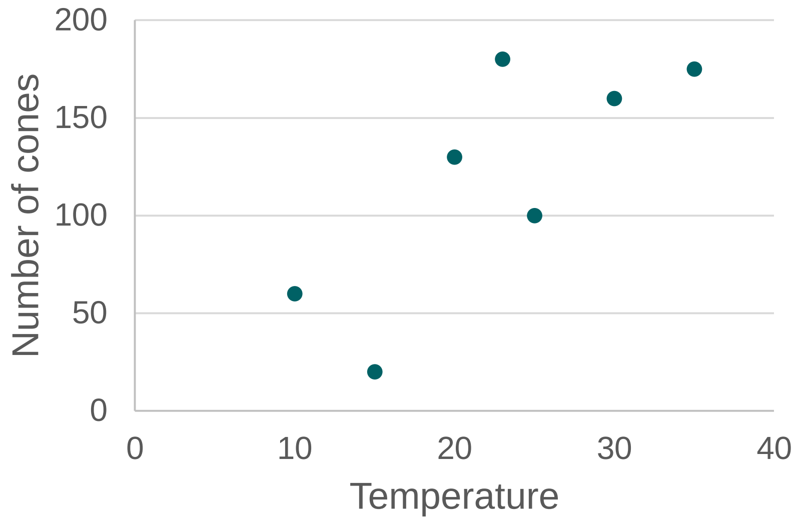
maximal negative correlation

$$Corr(x,y) = \frac{Cov(x,y)}{\sqrt{Var(x)} \cdot \sqrt{Var(y)}}$$

Between -1 and 1
> 0: positive correlation
< 0: negative correlation
≈ 0: independent

Multiple Features – Correlation (Example)

Temperature (◦C)	Number of cones
10	60
15	10
20	185
23	150
25	150
30	200
35	175



$$Corr(x, y) = \frac{Cov(x,y)}{\sqrt{Var(x)}\cdot\sqrt{Var(y)}} = \frac{419.88}{8.54\cdot 63.69} = 0.77$$

Temperature

Number of cones

Strong positive correlation

Multiple Features – Correlation Matrix

$$\text{Features } a, b, ..., z$$

$$\begin{array}{cc} & \begin{array}{cccc} a & b & & z \end{array} \\ \begin{array}{c} a \\ b \\ \\ z \end{array} & \begin{bmatrix} Corr(a,a) & Corr(a,b) & \dots & Corr(a,z) \\ Corr(b,a) & Corr(b,b) & \dots & Corr(b,z) \\ \dots & \dots & \dots & \dots \\ Corr(z,a) & Corr(z,b) & \dots & Corr(z,z) \end{bmatrix} \end{array}$$

Multiple Features – Correlation Matrix

$$\text{Features } a, b, ..., z$$

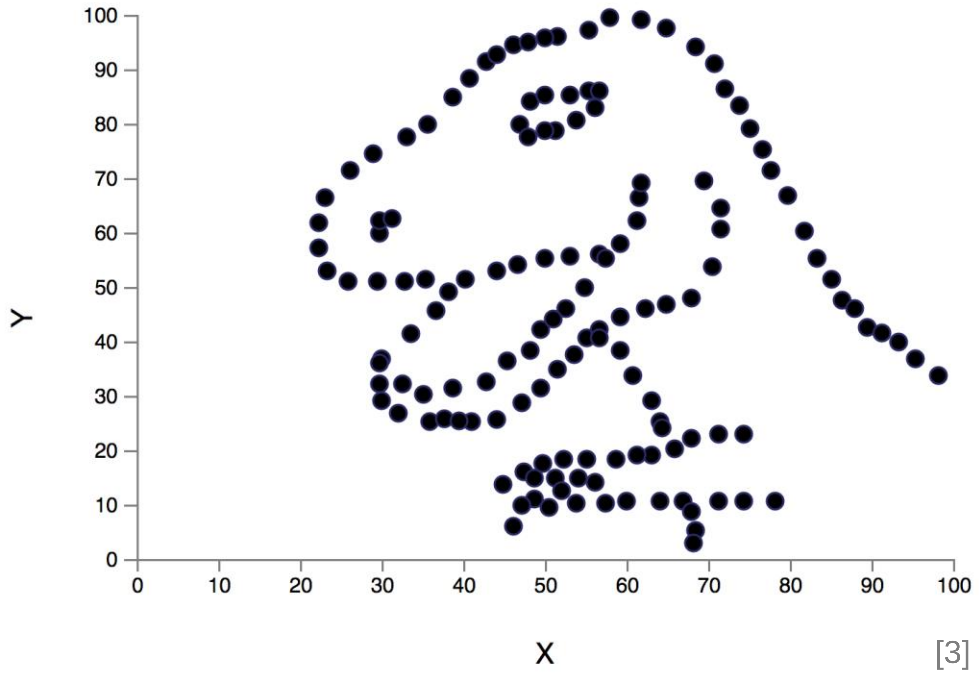
$$\begin{array}{cc} & \begin{array}{cccc} a & b & & z \end{array} \\ \begin{array}{c} a \\ b \\ \\ z \end{array} & \begin{bmatrix} 1.0 & 0.90 & \ldots & 0.35 \\ 0.90 & 1.0 & \ldots & 0.30 \\ \ldots & \ldots & \ldots & \ldots \\ 0.35 & 0.30 & \ldots & 1.0 \end{bmatrix} \end{array}$$

What can we say about this distribution?

x	y
55.3846	97.1795
51.5385	96.0256
46.1538	94.4872
42.8205	91.4103
40.7692	88.3333
38.7179	84.8718
35.641	79.8718
33.0769	77.5641
28.9744	74.4872
26.1538	71.4103
…	…

$$Count = 142$$
$$Mean(x) = 54.2633$$
$$Std(x) = 16.7651$$
$$Mean(y) = 47.8323$$
$$Std(y) = 26.9354$$
$$Corr(x, y) = -0.0645$$

Datasaurus



[3]

$$Count = 142$$
$$Mean(x) = 54.2633$$
$$Std(x) = 16.7651$$
$$Mean(y) = 47.8323$$
$$Std(y) = 26.9354$$
$$Corr(x, y) = -0.0645$$

Anscombe's Quartet

Dataset 1		Dataset 2		Dataset 3		Dataset 4	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

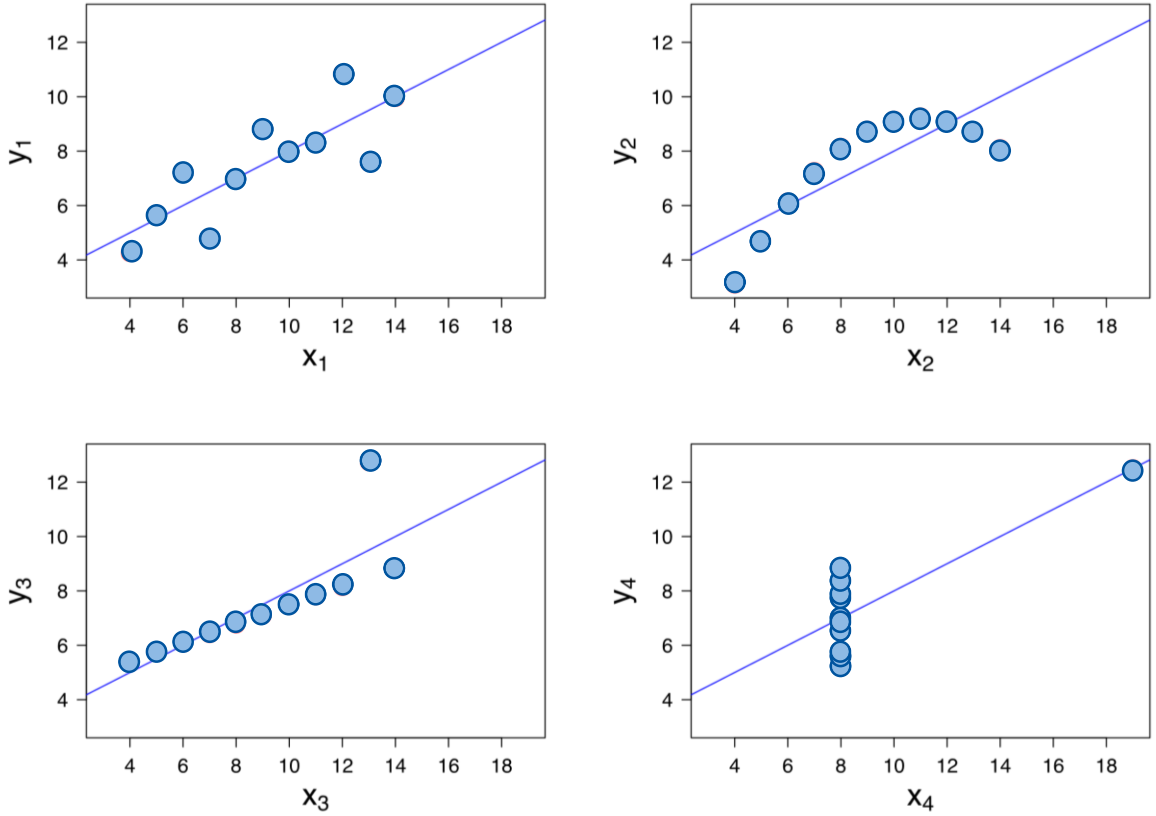
[4]

$$Mean(x) = 9$$
$$Var(x) = 11$$
$$Mean(y) = 7.5$$
$$Var(y) = 4.125$$
$$Corr(x, y) = 0.816$$
$$\text{Linear regression line: } y = \frac{1}{2}x + 3$$

Anscombe’s Quartet

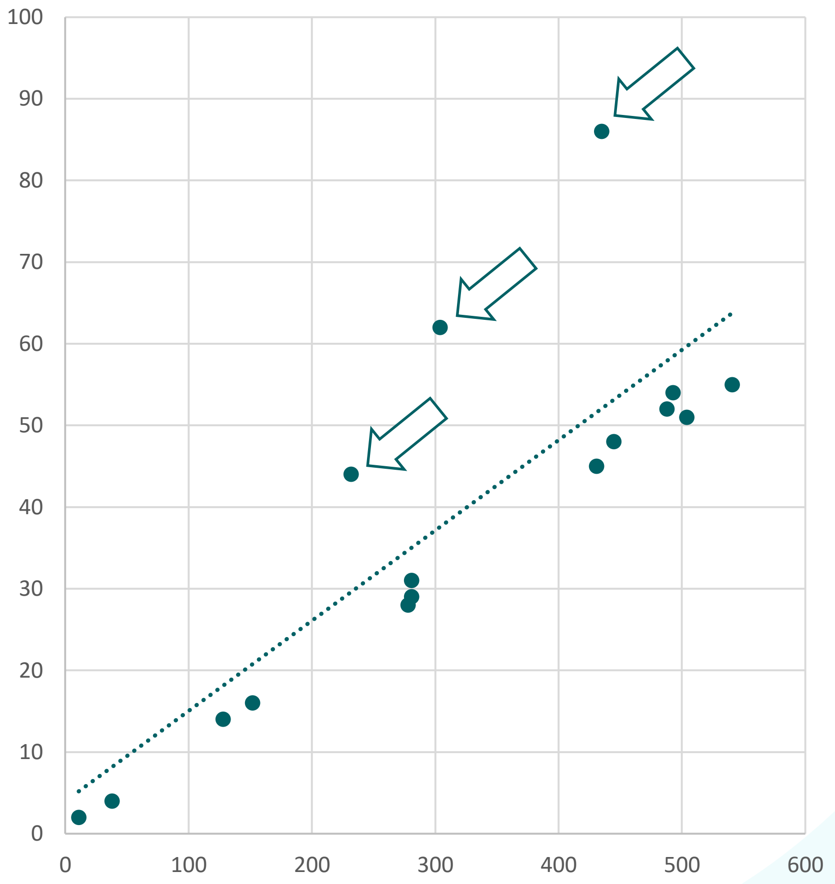
Dataset 1		Dataset 2		Dataset 3		Dataset 4	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

[4]



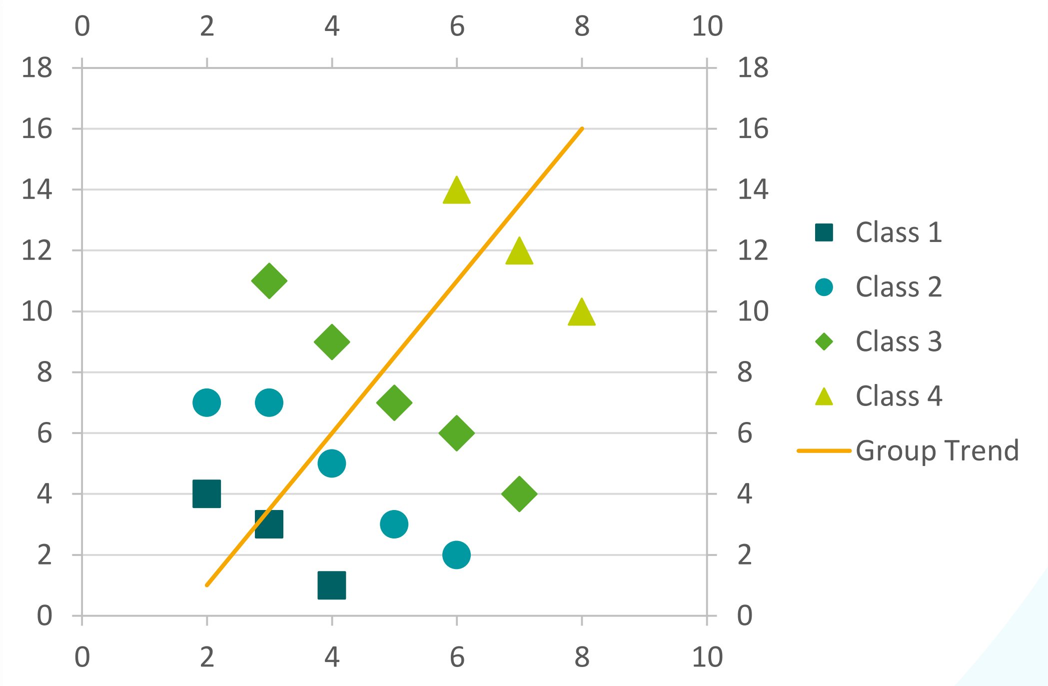
Outliers

- Outlier: an observation that lies an abnormal distance away from other values
- Can have a significant impact on measures such as mean, variance or standard deviation
- It is important to identify and deal with outliers before performing any analysis
  → visualize and explore our data first!



Simpson's Paradox

A trend appears in several different groups of data but disappears or reverses when these groups are combined.



Simpson's Paradox

	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
Total	12,763	41%	8,442	44%	4,321	35%

Aggregated

Department	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
A	933	64%	**825**	62%	108	82%
B	585	63%	**560**	63%	25	68%
C	918	35%	325	37%	**593**	34%
D	792	34%	417	33%	375	35%
E	584	25%	191	28%	**393**	24%
F	714	6%	373	6%	341	7%
Total	**4526**	**39%**	**2691**	**45%**	**1835**	**30%**

Legend:

greater percentage of successful applicants than the other gender

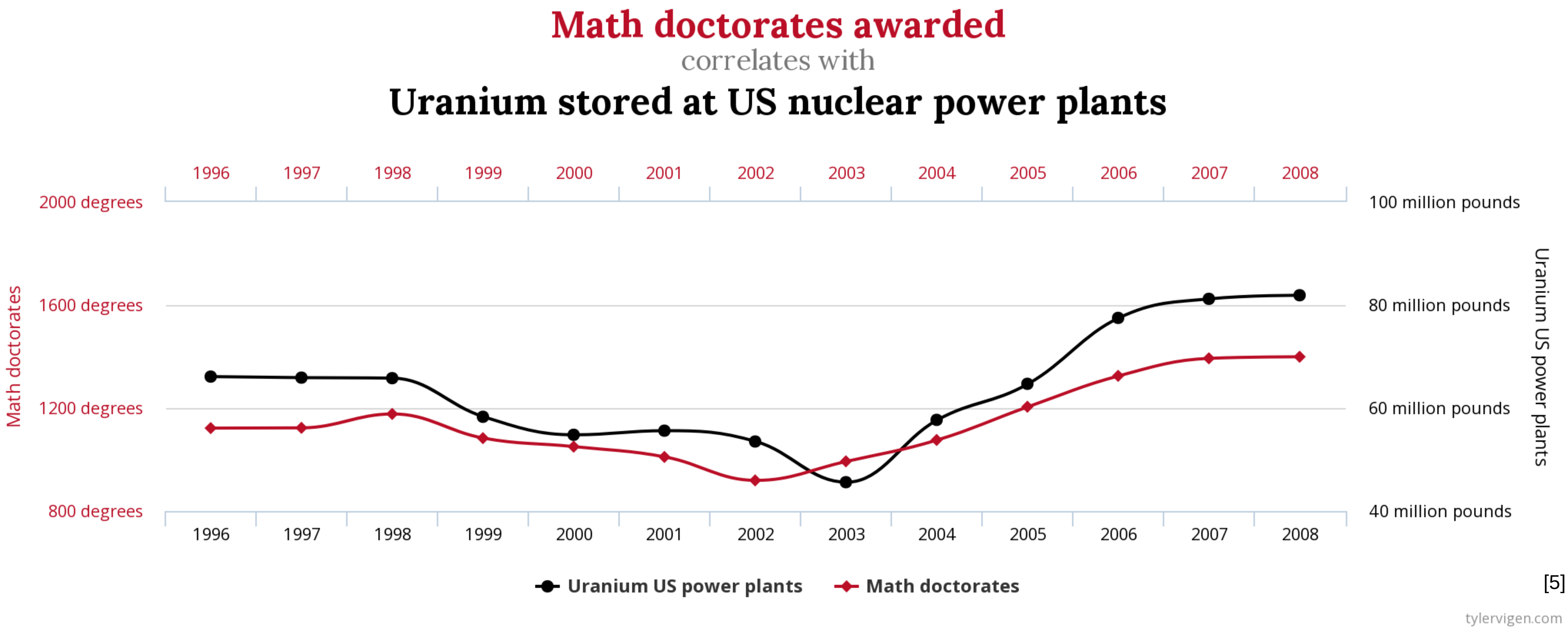
greater number of applicants than the other gender

bold - the two 'most applied for' departments for each gender

By department (six largest)

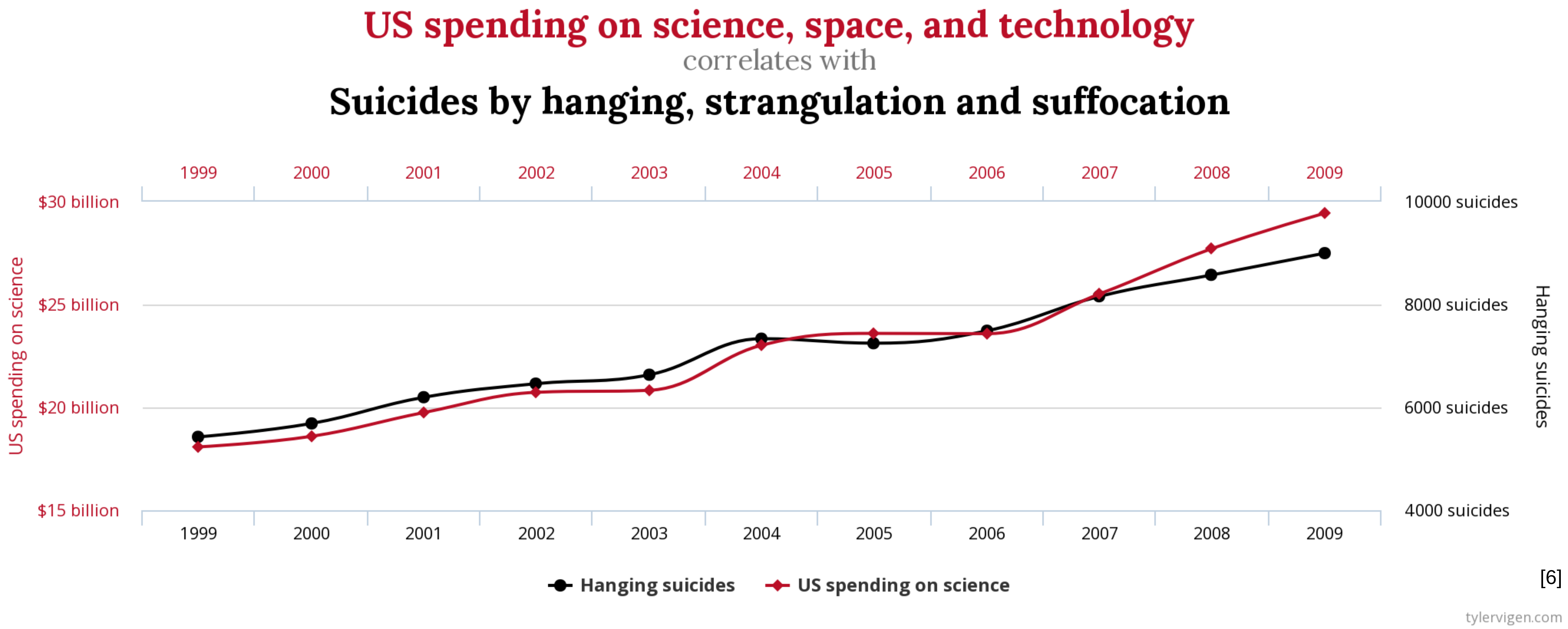
UC Berkeley admission data, 1973

Spurious Correlations



[5]

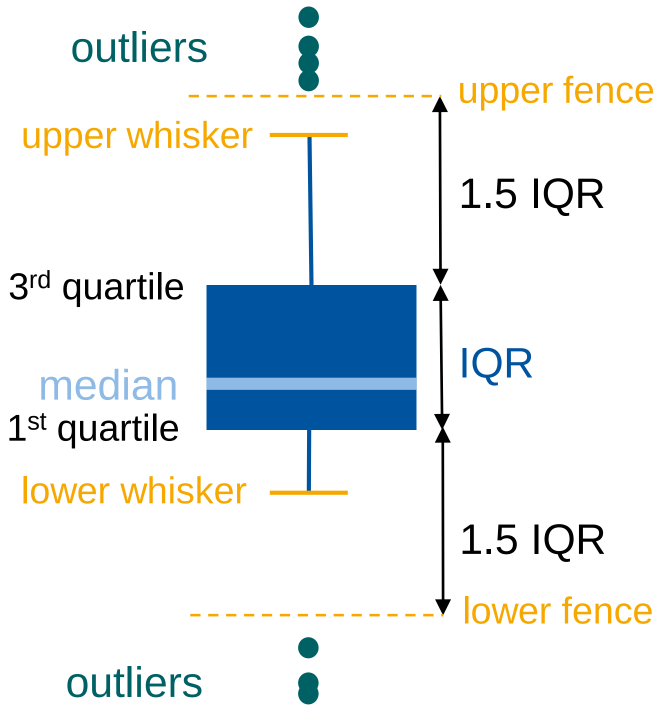
Spurious Correlations



[6]

Box Plot

- Median value (middle), depicted by bar
- IQR – Interquartile Range (covers 50% of middle instances), depicted by box
- Upper fence – 3rd quartile + 1.5 IQR
  Upper whisker – maximal value below upper fence
- Lower fence – 1st quartile - 1.5 IQR
  Lower whisker – minimal value above lower fence
- Outliers – drawn separately



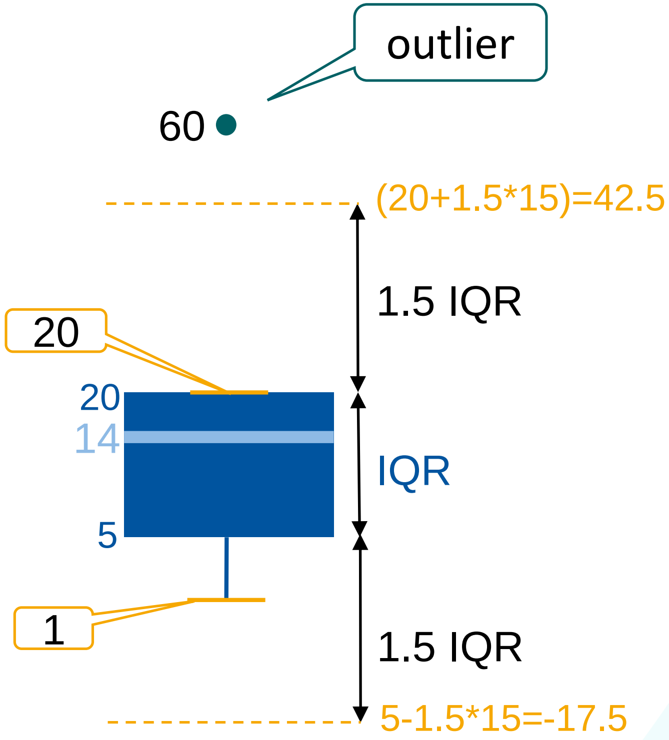
Box Plot - Example

Index	x
1	1
2	3
3	5
4	7
5	10
6	14
7	15
8	17
9	20
10	20
11	60

- Median: 14
- 1st quartile: 5

  x_n with $n = \lceil \frac{25}{100} \cdot 11 \rceil = \lceil 2.75 \rceil = 3$
- 3rd quartile: 20

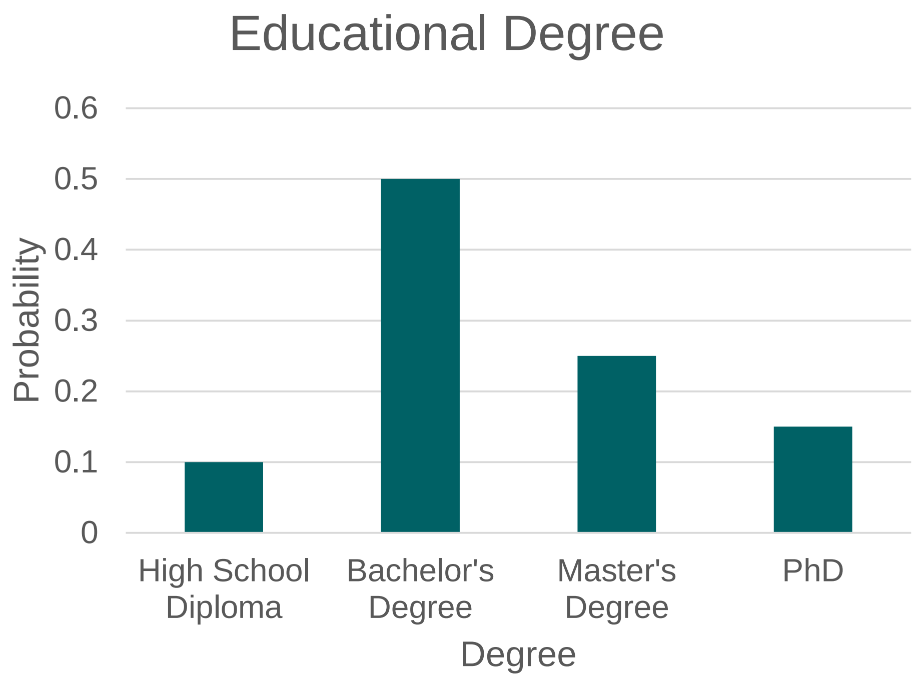
  x_n with $n = \lceil \frac{75}{100} \cdot 11 \rceil = \lceil 8.25 \rceil = 9$
- IQR: 20-5 = 15

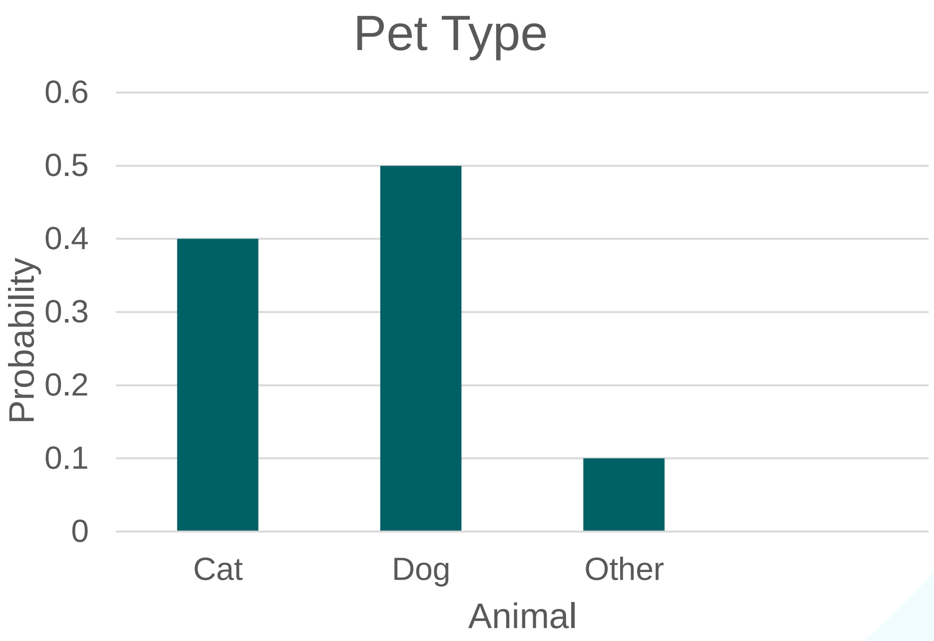


- Upper whisker: maximal value below $20 + 1.5 \cdot 15 = 42.5$ → 20
- Lower whisker: minimal value above $5 - 1.5 \cdot 15 = -17.5$ → 1

Histograms - Visualizations of Distributions

Categorical features



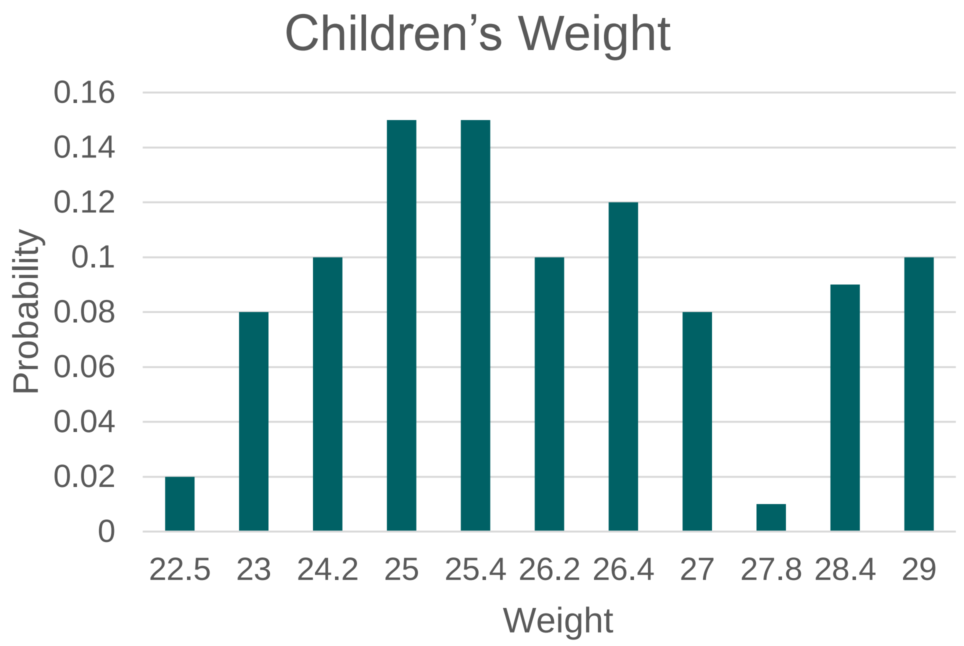


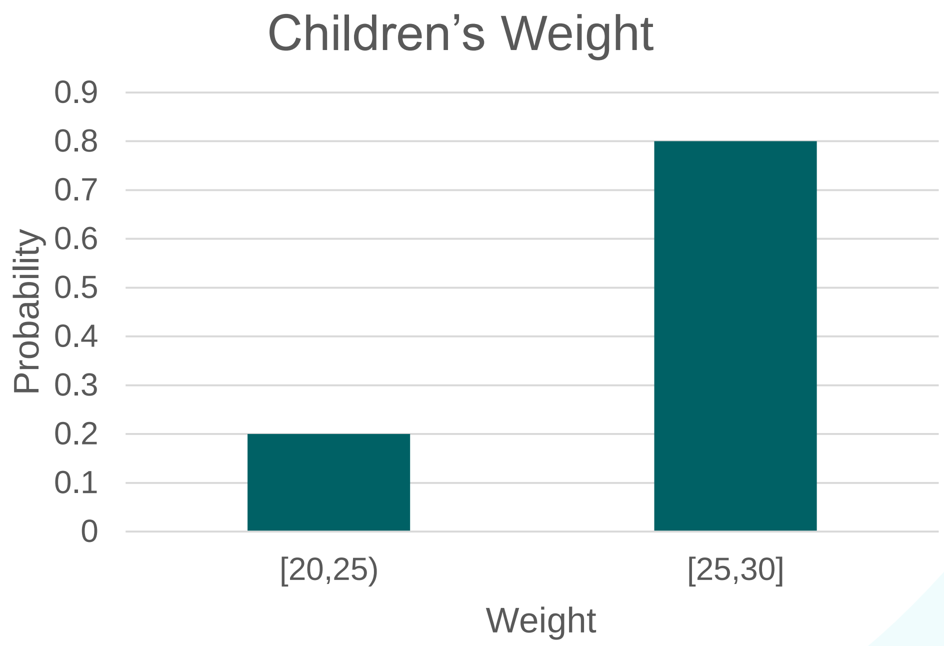
Histograms - Visualizations of Distributions

Continuous features

Histograms - Visualizations of Distributions

Continuous features

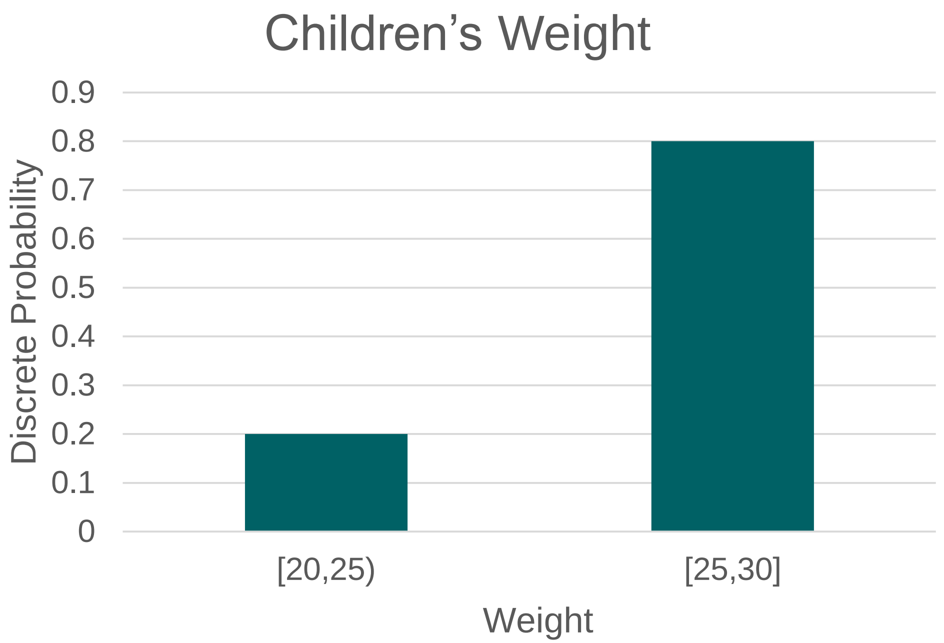




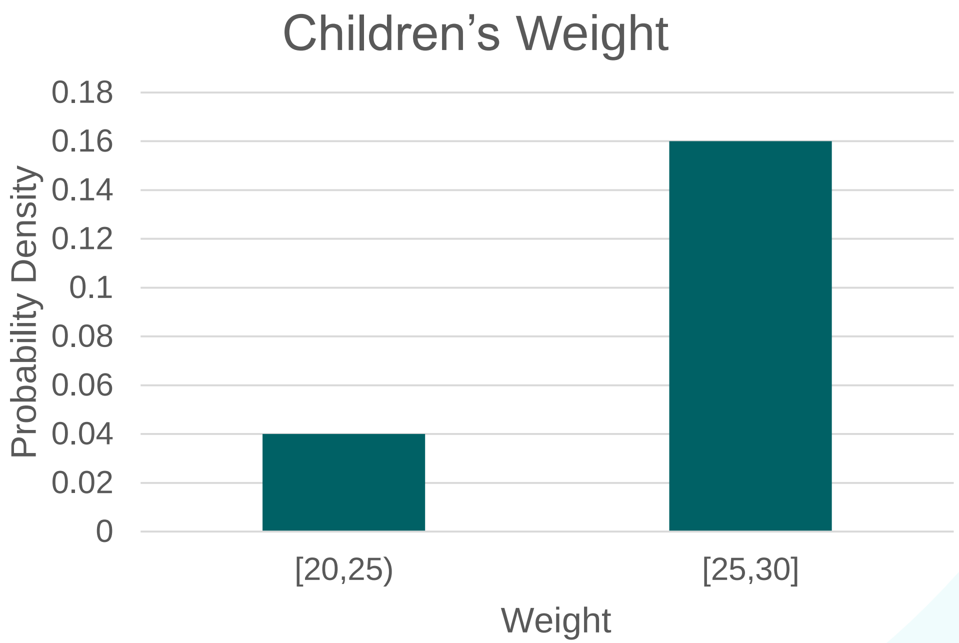
Histograms – Watch out for Normalization!

- Discrete probability distribution over intervals
- Normalized over population
- Sums to **1** [over discrete intervals]

- Continuous probability density over values
- Normalized over population **and** bin width
- Integrates to **1** [over continuous range]

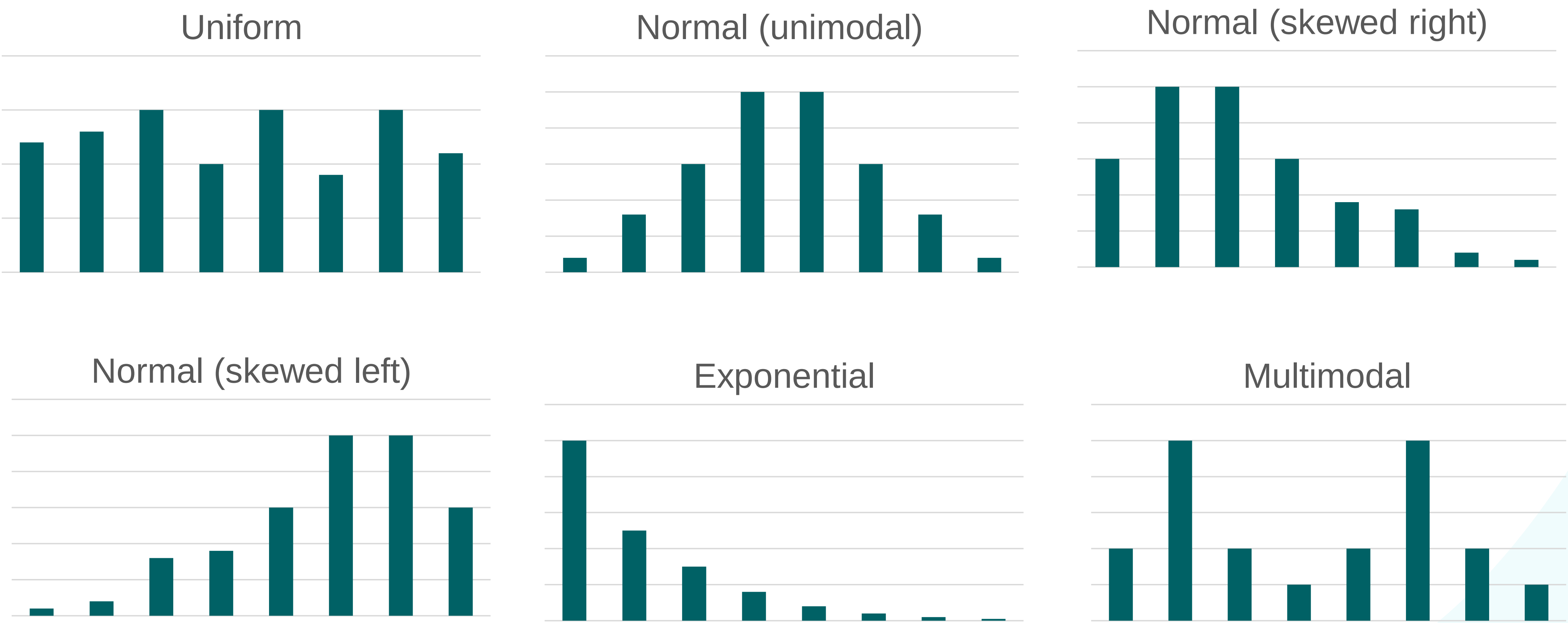


"probability that a child's weight is between 20 to 25 or between 25 to 30"



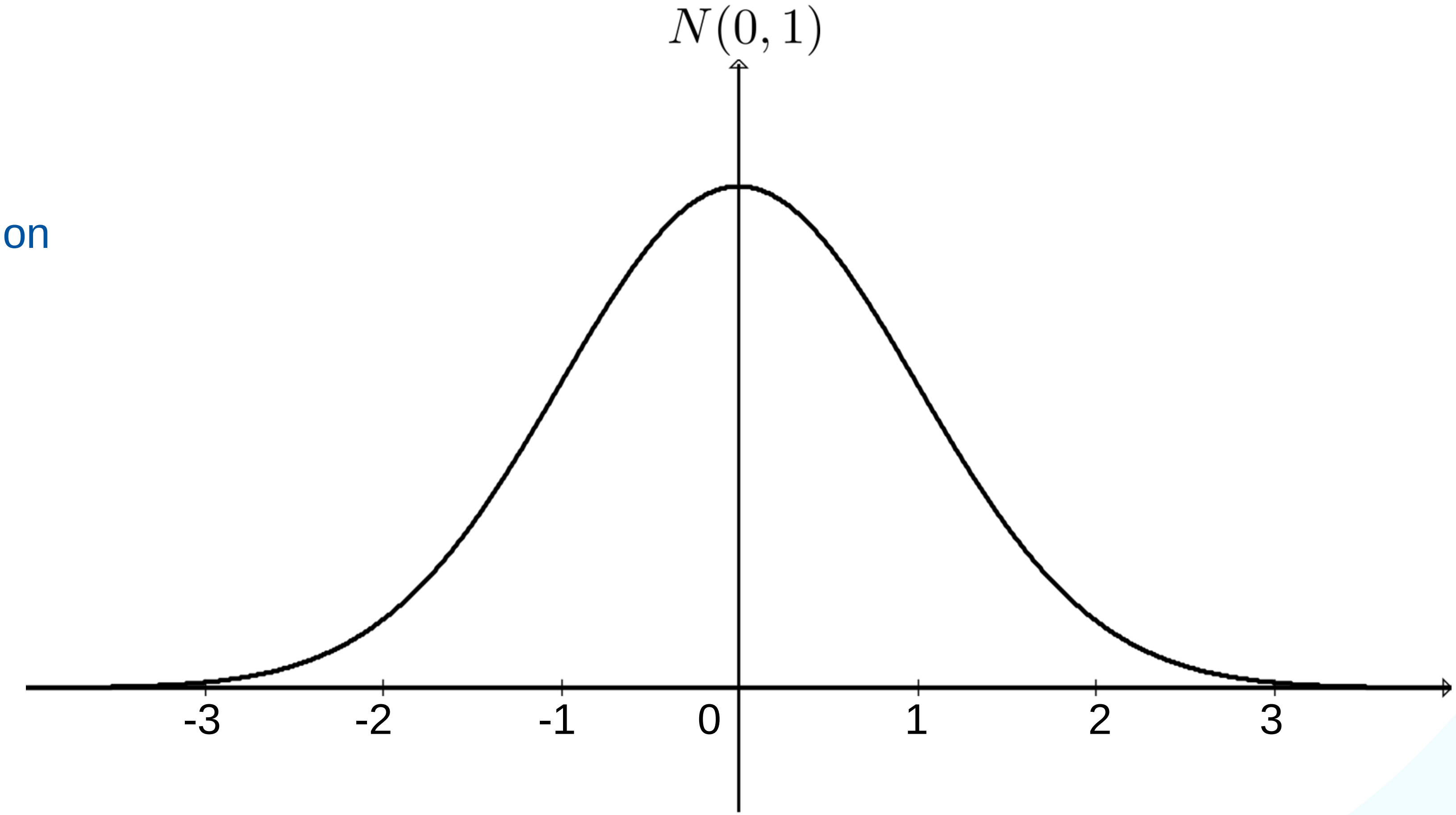
"probability of a child's weight over the reals"

Different Types of Histograms



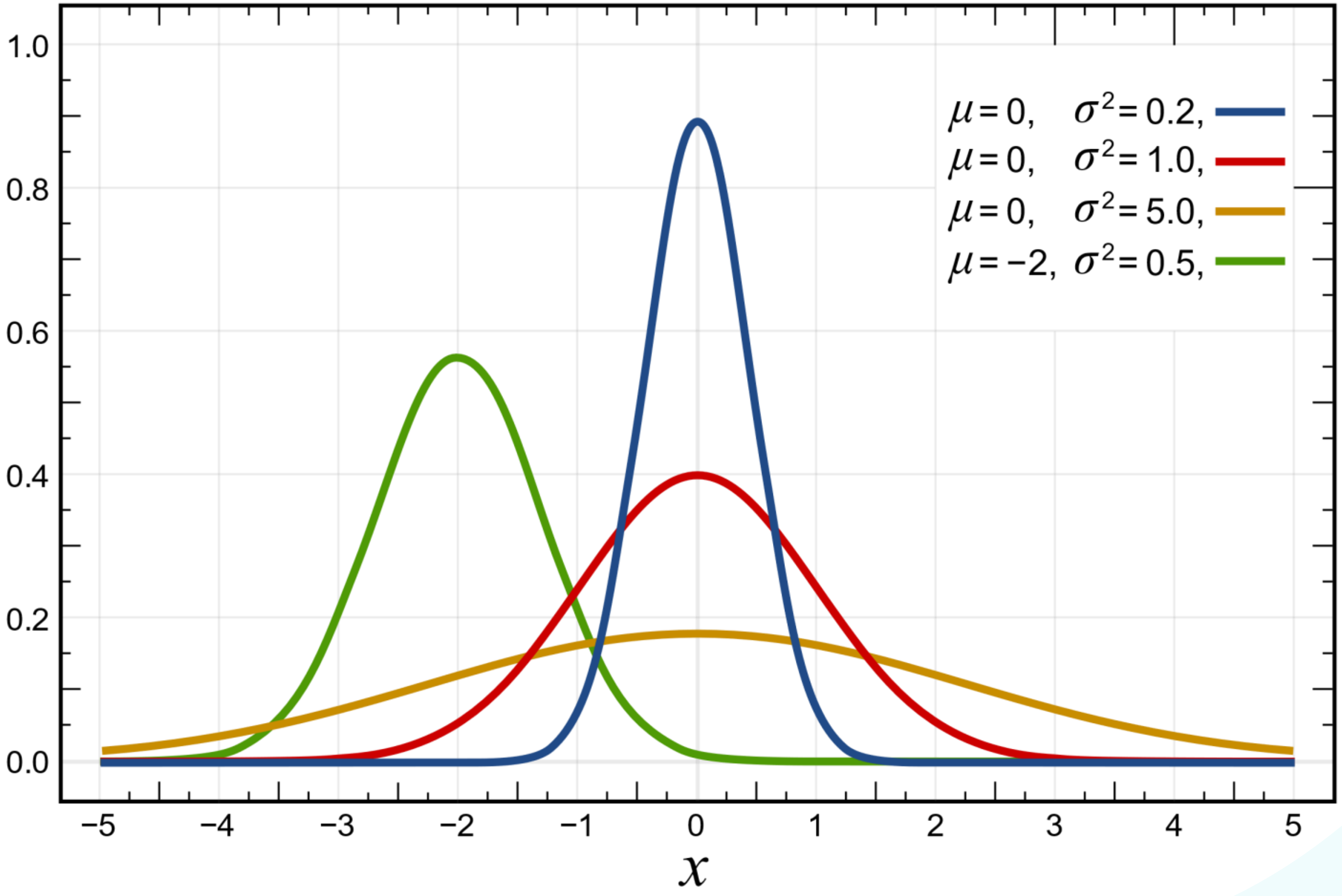
Normal (Gaussian) Distribution

- $N(\mu, \sigma^2)$
- μ - mean
- σ - standard deviation



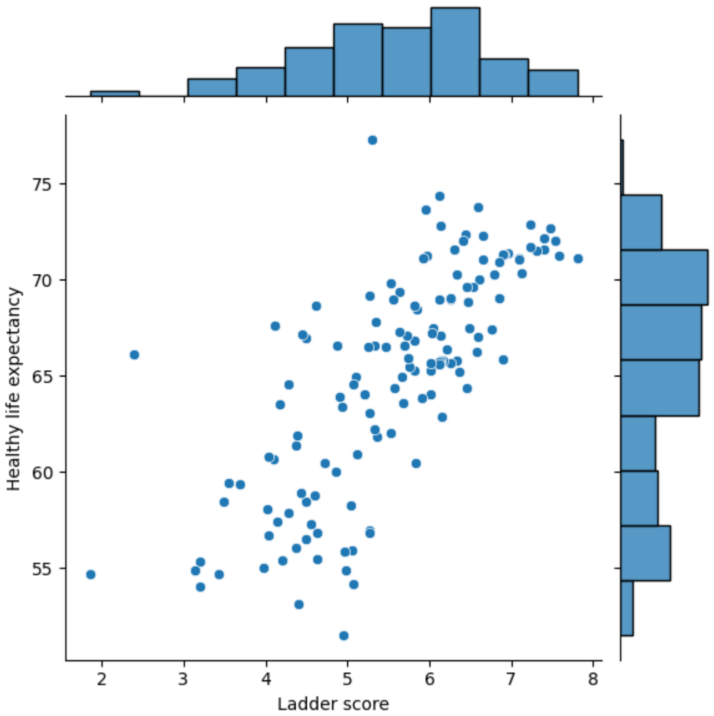
Normal (Gaussian) Distribution

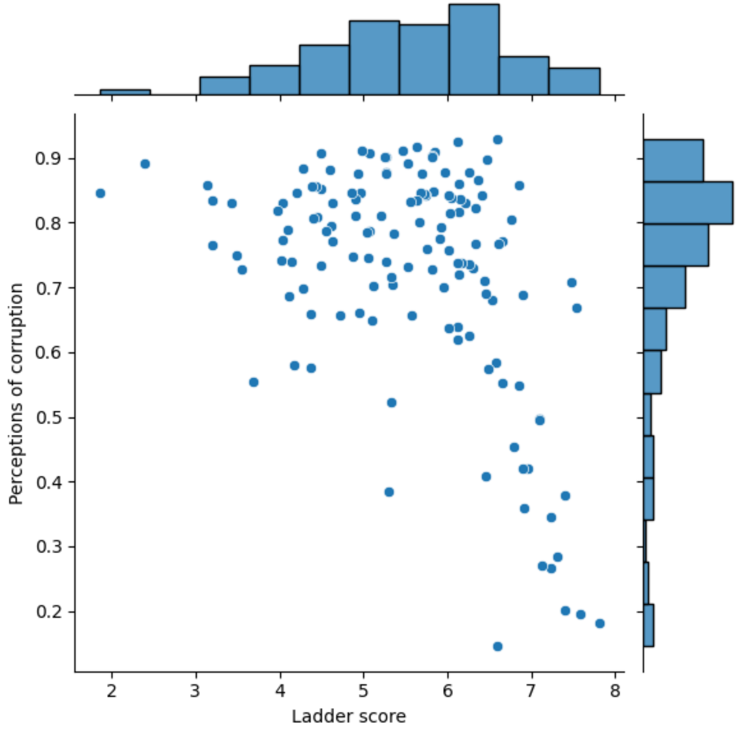
- $N(\mu, \sigma^2)$
- μ - mean
- σ - standard deviation

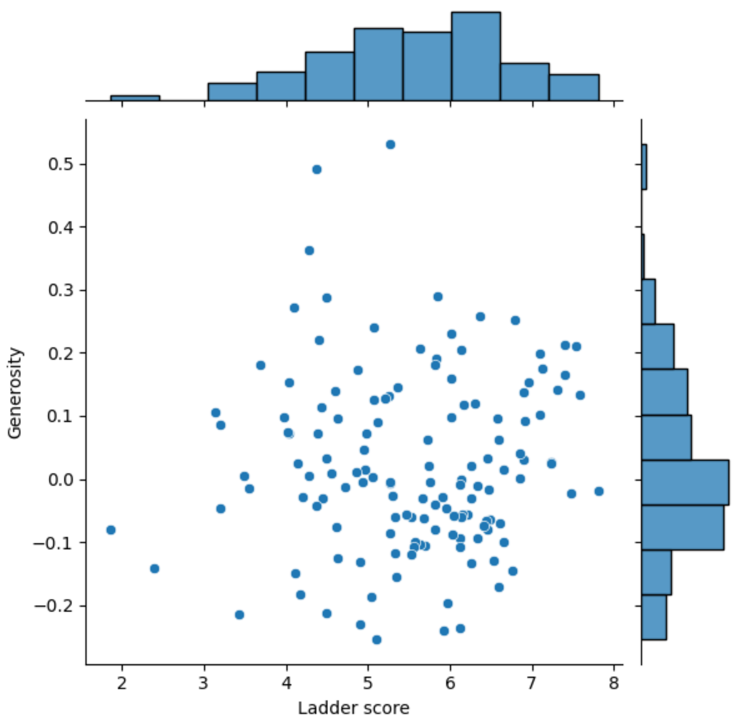


Scatter Plot - Correlation

World Happiness Report 2023 [3]







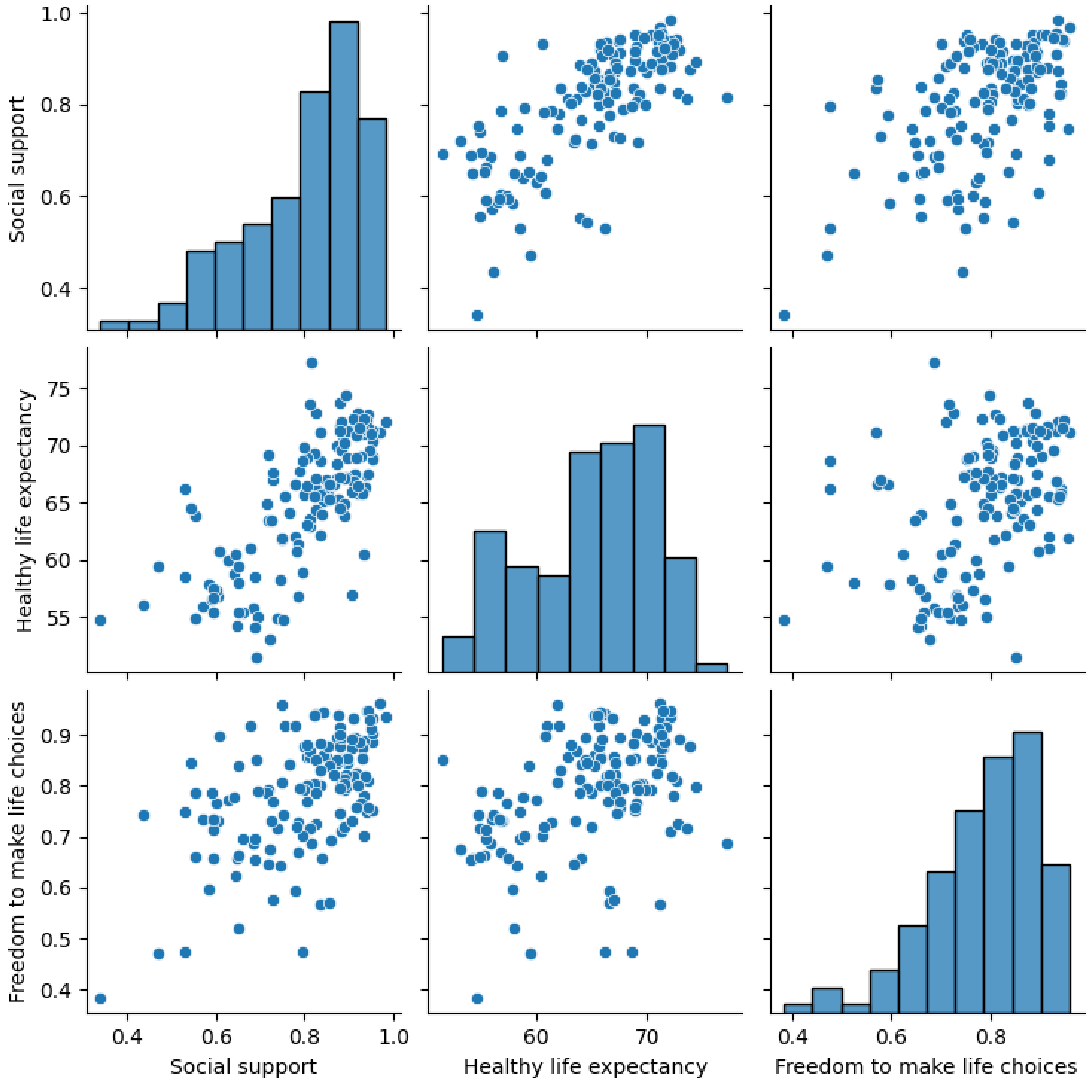
Positive correlation

Negative correlation

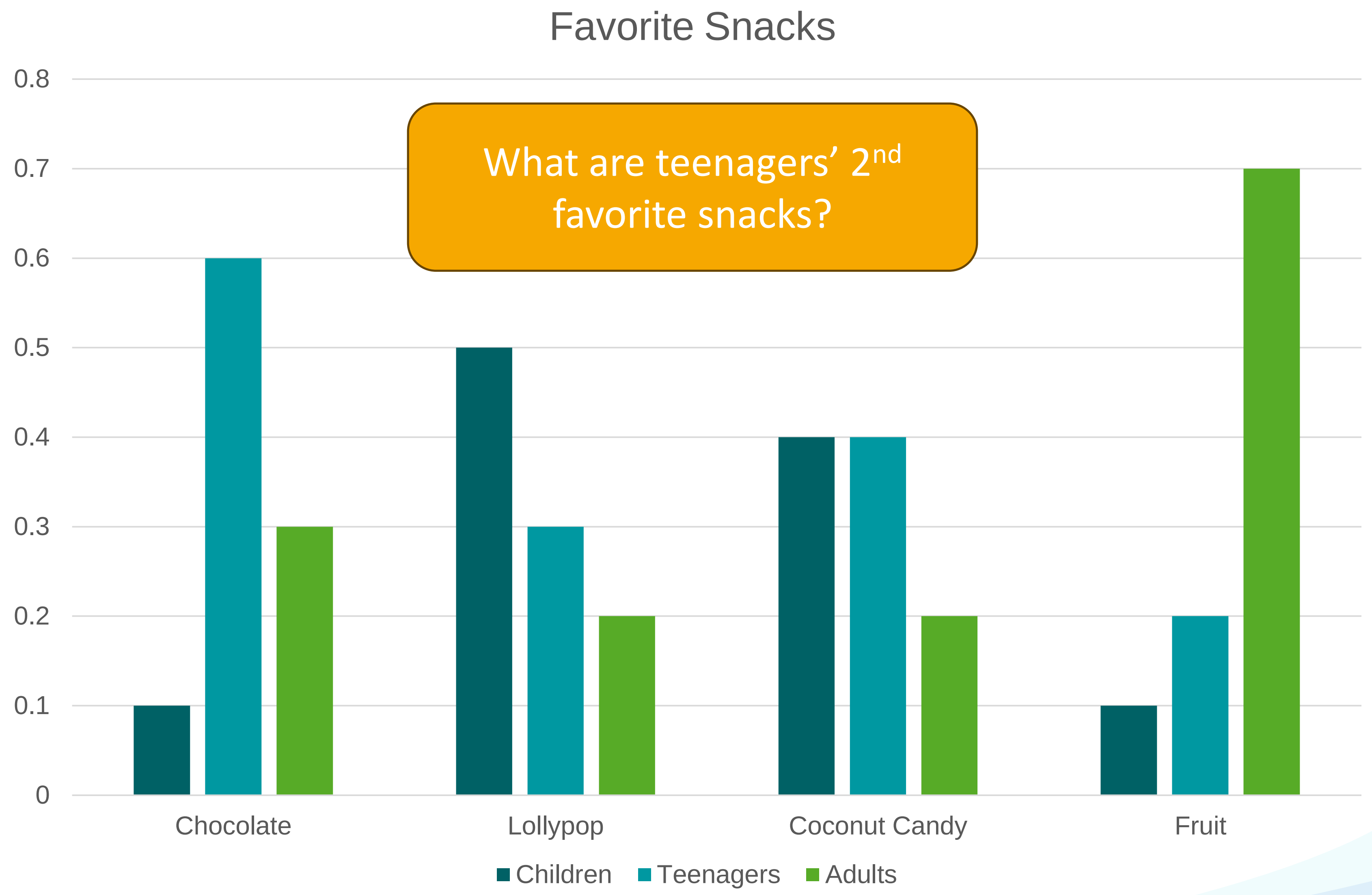
No correlation

Scatter Plot Matrix

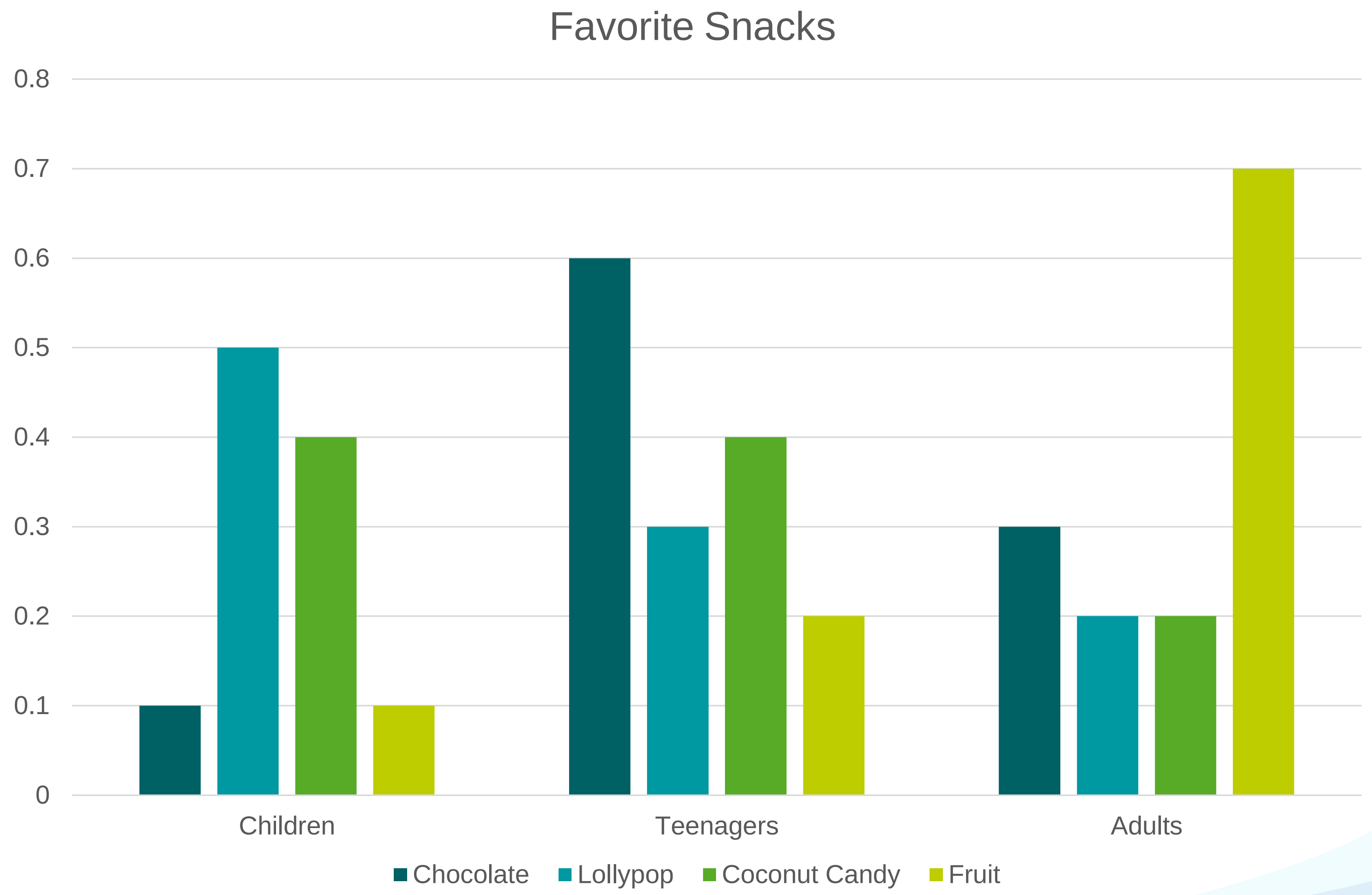
World Happiness Report 2023



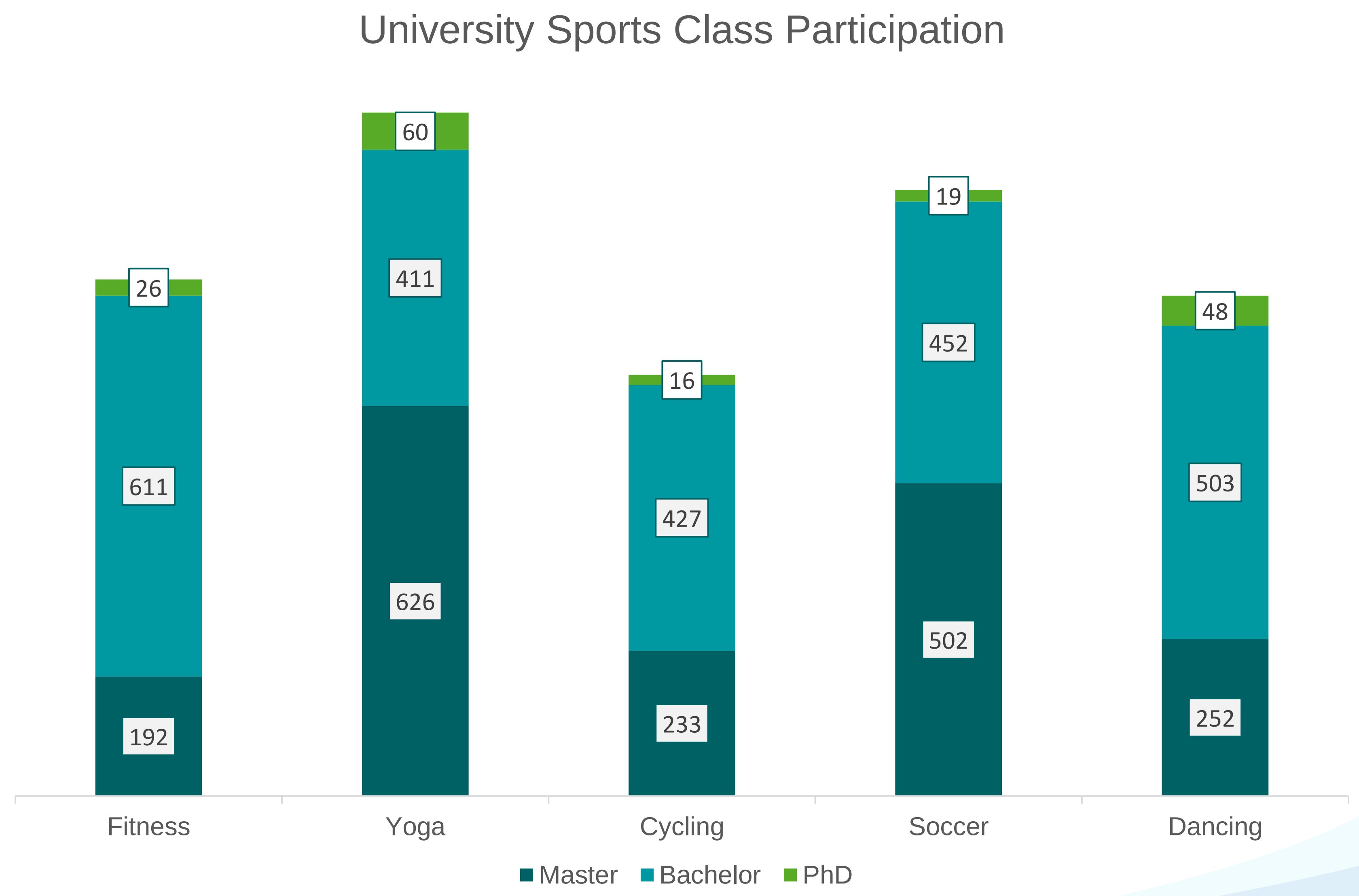
Faceting: Collection of Bar Plots



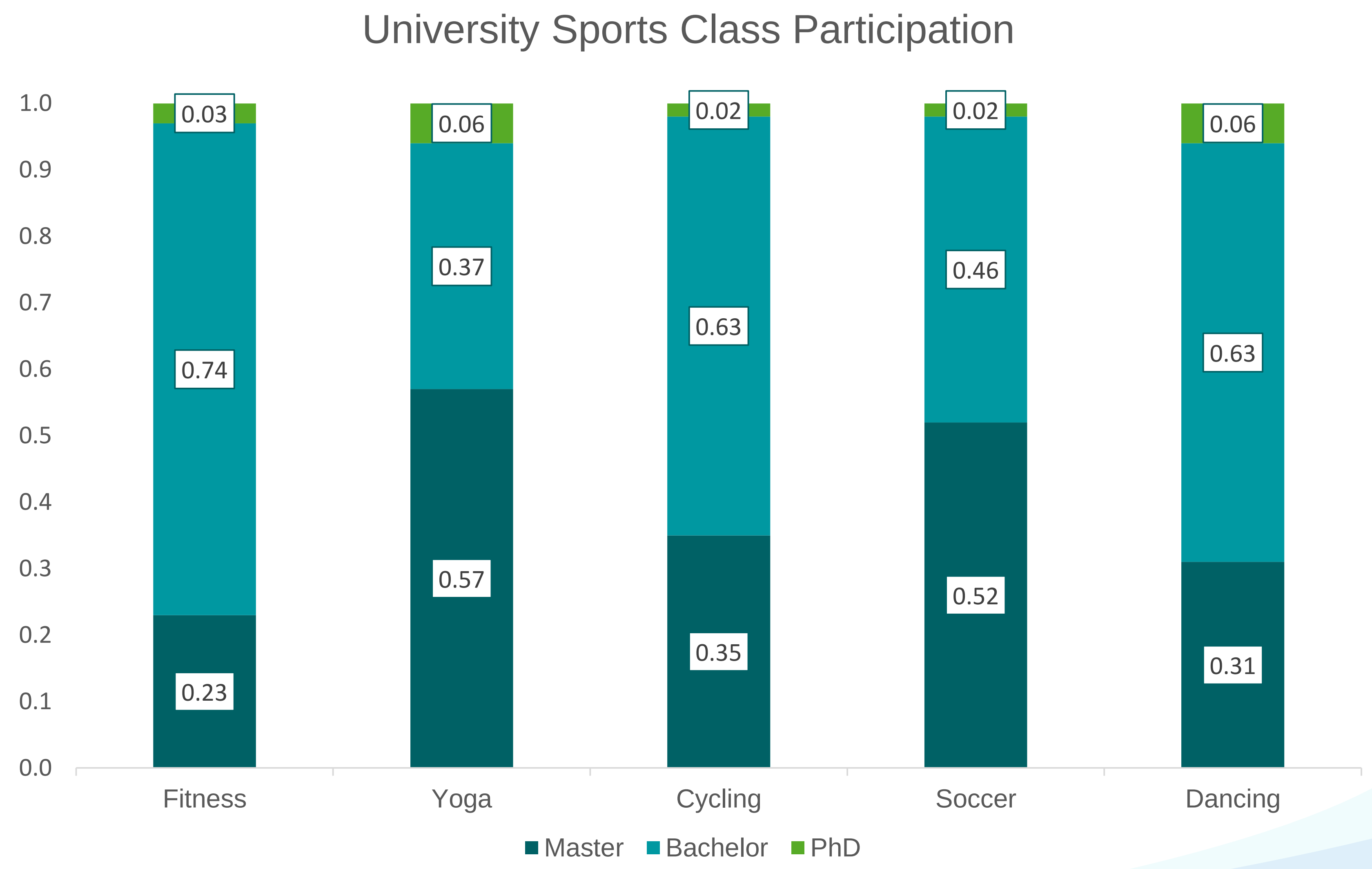
Faceting: Change of Focus



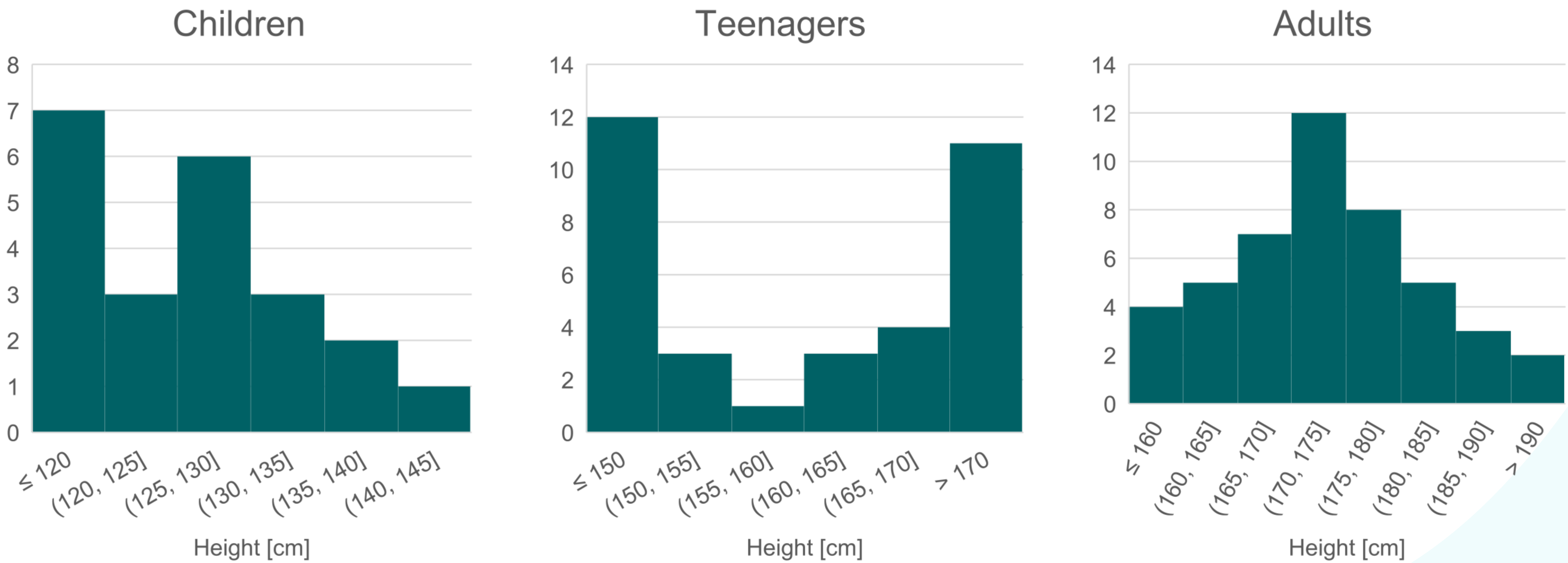
Stacked Bar Plots



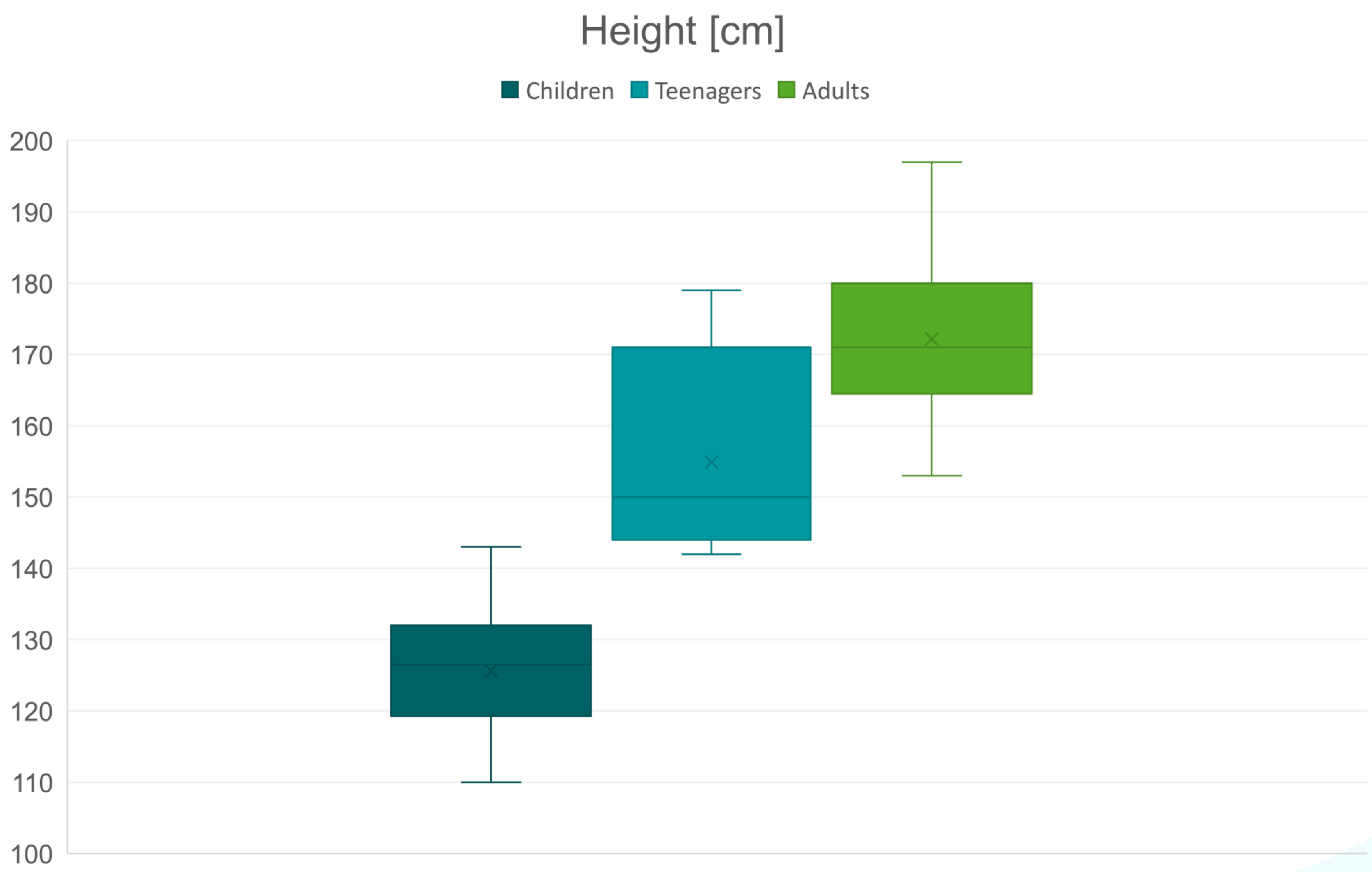
Stacked Bar Plots



Collection of Histograms

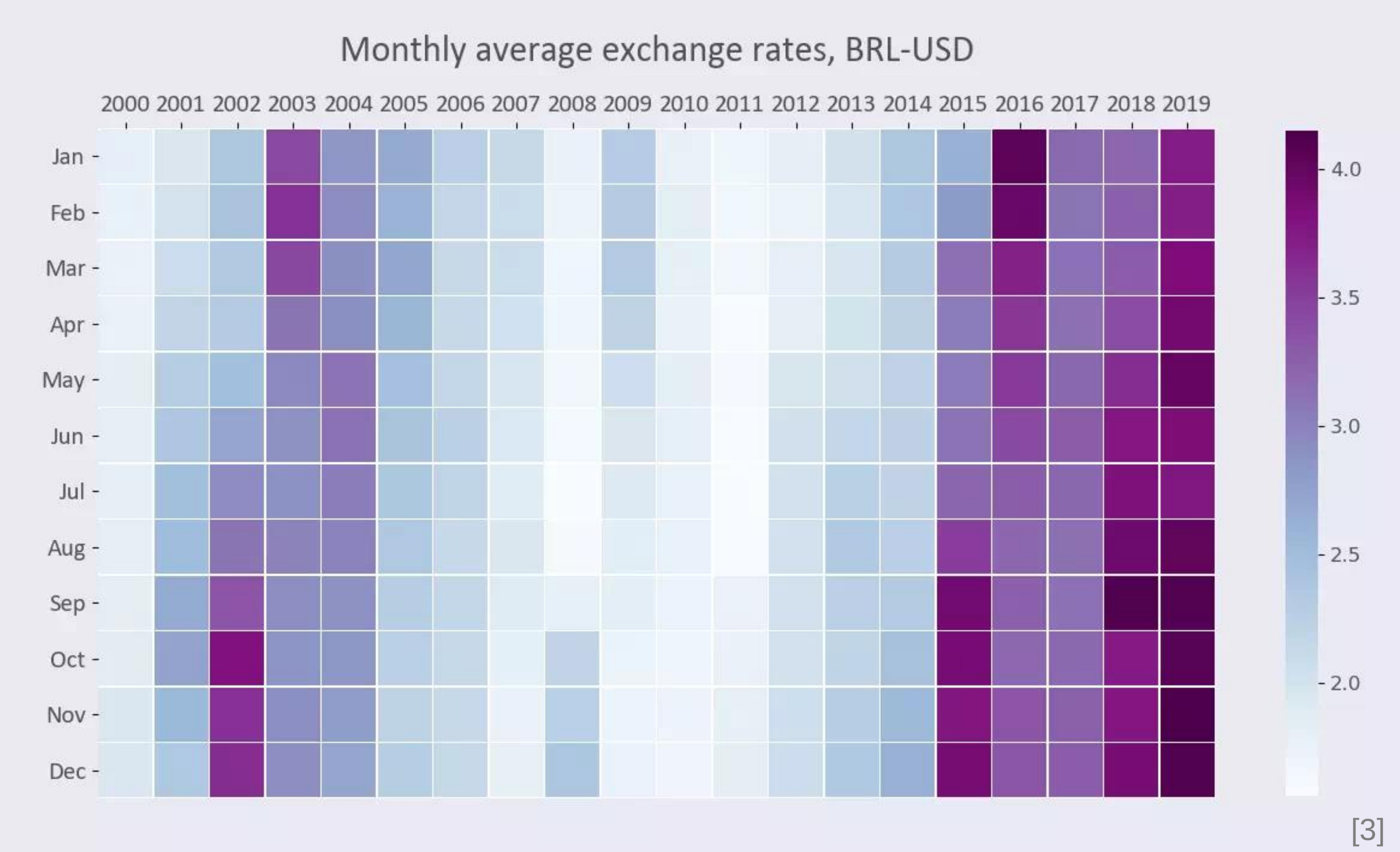


Collection of Box Plots



Advanced Visualizations - Examples

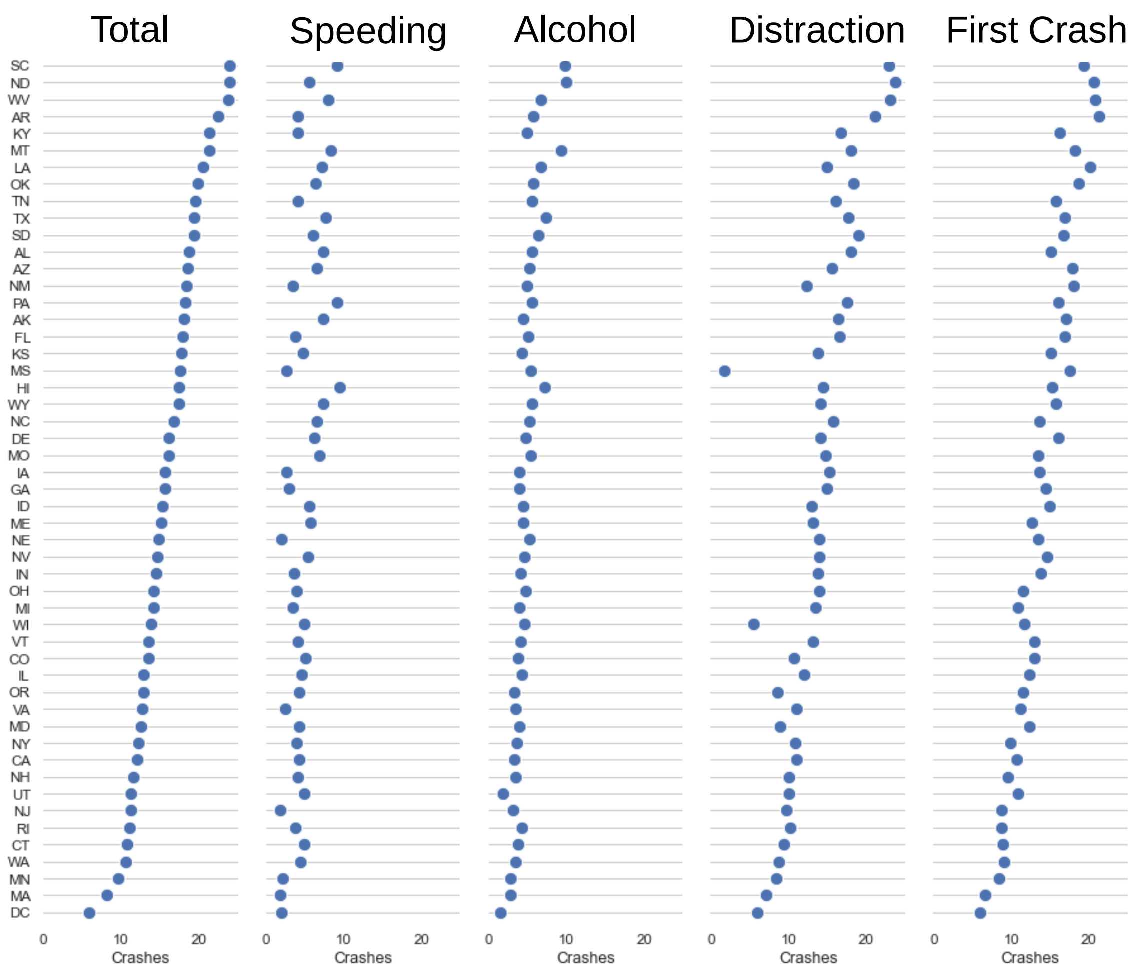
Heatmap



[3]

Advanced Visualizations - Examples

Dot Plot with Several Variables



[4]

Fatal Collisions per Billion Miles

-

Comparison of US States

Value of Good Visualizations

- Understanding and analyzing data more quickly and easily
- Communicating to others more effectively
- Identifying outliers, anomalies and other unexpected patterns in data
- Making clear decisions – identifying key insights

Feature Transformations

Dealing with Categorical Features

f_1	f_2	f_3	**class**
high	true	88	**A**
high	false	76	**B**
medium	false	32	**B**
low	true	89	**C**
high	true	21	**C**
medium	true	45	**A**

Categorical target feature

Categorical descriptive features (f_1,f_2)

One-Hot Encoding

f_1	f_2	f_3	**class**
high	true	88	**A**
high	false	76	**B**
medium	false	32	**B**
low	true	89	**C**
high	true	21	**C**
medium	true	45	**A**

Standard one-hot encoding: introduce a 0/1 feature for every possible value

- high – (1,0,0)
- medium – (0,1,0)
- low – (0,0,1)

One-Hot Encoding: Standard

f_1 - high	f_1 - medium	f_1 - low	f_2	f_3	class
1	0	0	true	88	**A**
1	0	0	false	76	**B**
0	1	0	false	32	**B**
0	0	1	true	89	**C**
1	0	0	true	21	**C**
0	1	0	true	45	**A**

Standard one-hot encoding: introduce a 0/1 feature for every possible value

- high – (1,0,0)
- medium – (0,1,0)
- low – (0,0,1)

One-Hot Encoding: Common Variant

f_1 - dummy$_0$	f_1 - dummy$_1$	f_2	f_3	class
1	0	true	88	**A**
1	0	false	76	**B**
0	1	false	32	**B**
0	0	true	89	**C**
1	0	true	21	**C**
0	1	true	45	**A**

k-1 one-hot encoding:
- high – (1,0)
- medium – (0,1)
- low – (0,0)

\+ preferable where co-linearity of features is problematic

\- introduces asymmetry, e.g., see *low*

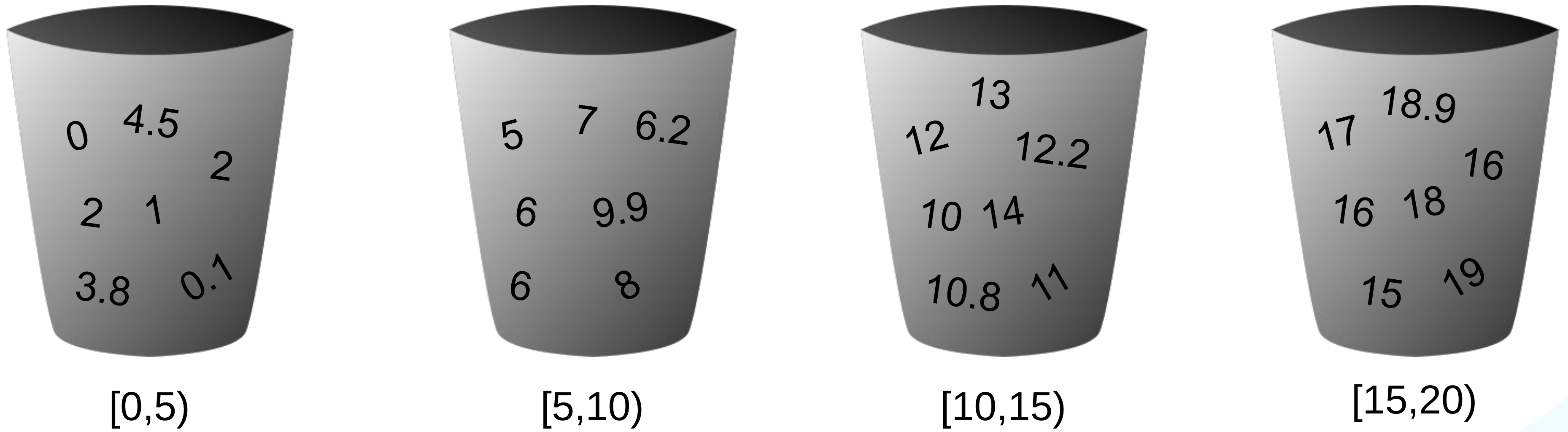
One-Hot Encoding – Special Cases

f_1 - high	f_1 - medium	f_1 - low	f_2	f_3	class
1	0	0	true	88	A
1	0	0	false	76	B
0	1	0	false	32	B
0	0	1	true	89	C
1	0	0	true	21	C
0	1	0	true	45	A

- Binary values (true, false) can be translated to a single numeric value (1, 0) [example of k-1 encoding]
- Note that categorical variables with a clear order (ordinal) may be translated to a single numeric value (e.g., excellent = 1.0, good = 0.7, average = 0.5, poor = 0.3, horrible = 0.0)

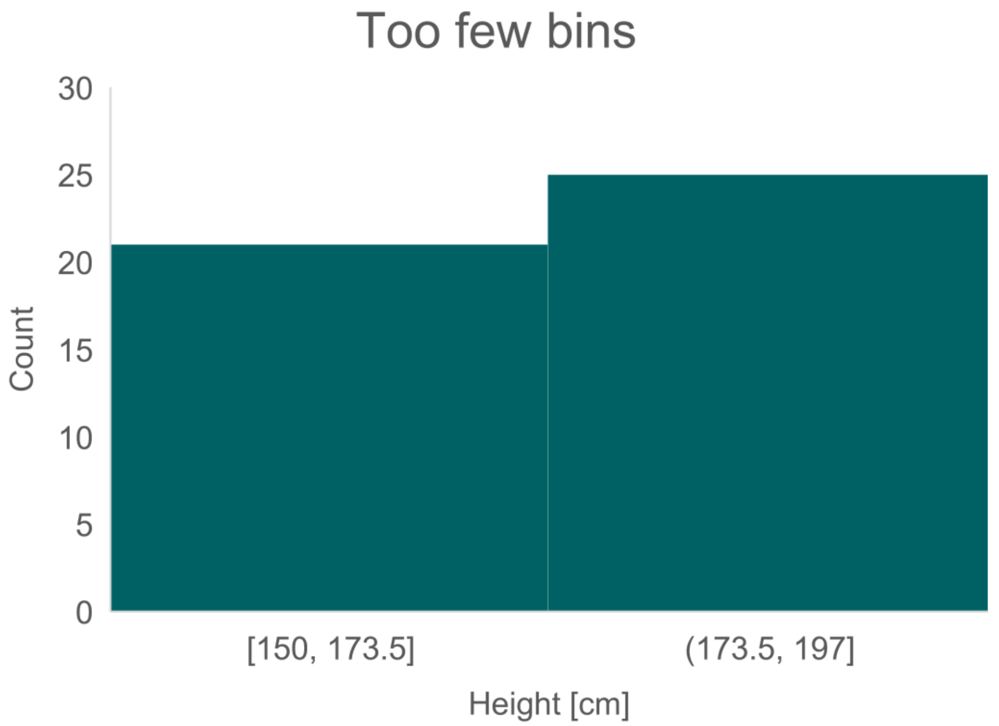
Dealing with Continuous Features - Binning

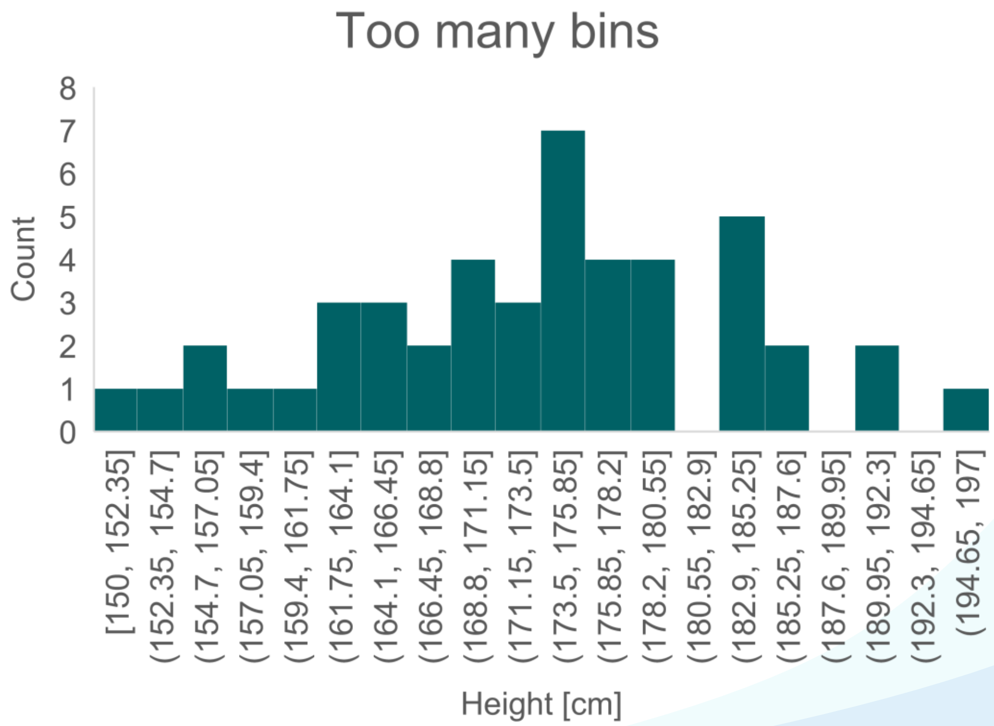
- Binning is used to transform continuous features into categorical
- A bin is a range, e.g., [0,5), [5,10), [10,15), [15,20)
- Choosing the right bins (their number and size) is crucial (e.g., to create meaningful histograms)



Binning – Number of Bins

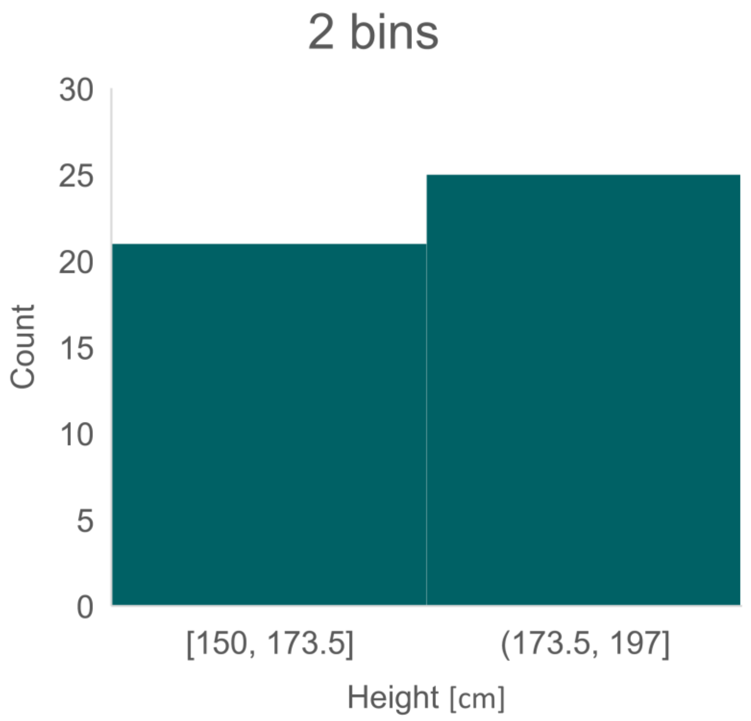
- Too few bins may lead to the loss of information (underfitting)
- Too many bins may lead to sparseness – bins that are empty or have just a few instances (overfitting)

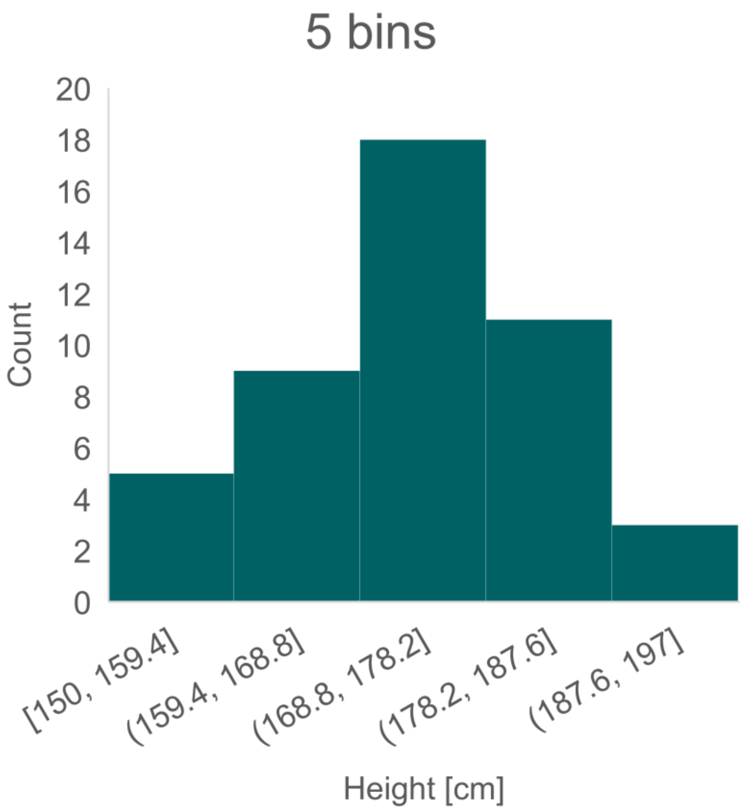


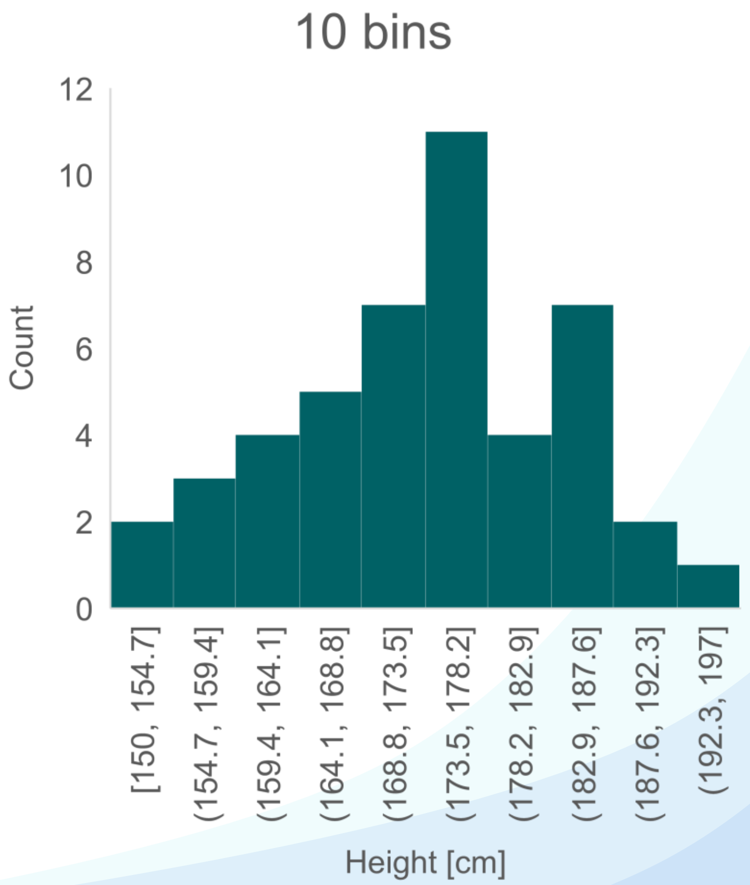


Equal Width Binning

Bins have a fixed width, but the numer of items per bin may vary







Equal Width Binning - Example

Tree Age [years]	Tree Height [m]
9	26
51	96
47	61
77	118
64	91
2	6
48	60
13	31
9	11
29	86
90	107
80	88

Apply equal width binning to the feature Tree Height with a bin width of 29. The lowest bin boundaries should coincide with the smallest value.

Equal Width Binning - Example

Tree Age [years]	Tree Height [m]
9	26
51	96
47	61
77	118
64	91
2	6
48	60
13	31
9	11
29	86
90	107
80	88

Apply equal width binning to the feature Tree Height with a bin width of 29. The lowest bin boundaries should coincide with the smallest value.

1. Sort the data

Equal Width Binning - Example

Tree Age [years]	Tree Height [m]
2	6
9	11
9	26
13	31
48	60
47	61
29	86
80	88
64	91
51	96
90	107
77	118

Apply equal width binning to the feature Tree Height with a bin width of 29. The lowest bin boundaries should coincide with the smallest value.

1. Sort the data

Equal Width Binning - Example

Tree Age [years]	Tree Height [m]
2	6
9	11
9	26
13	31
48	60
47	61
29	86
80	88
64	91
51	96
90	107
77	118

Apply equal width binning to the feature Tree Height with a bin width of 29. The lowest bin boundaries should coincide with the smallest value.

1. Sort the data
2. Distribute elements to bins

Equal Width Binning - Example

Tree Age [years]	Tree Height [m]
2	6
9	11
9	26
13	31
48	60
47	61
29	86
80	88
64	91
51	96
90	107
77	118

Apply equal width binning to the feature Tree Height with a bin width of 29. The lowest bin boundaries should coincide with the smallest value.

1. Sort the data
2. Distribute elements to bins:
   6+29=35 → [6,35)

Equal Width Binning - Example

Tree Age [years]	Tree Height [m]
2	6
9	11
9	26
13	31
48	60
47	61
29	86
80	88
64	91
51	96
90	107
77	118

Apply equal width binning to the feature Tree Height with a bin width of 29. The lowest bin boundaries should coincide with the smallest value.

1. Sort the data
2. Distribute elements to bins:
   6+29=35 → [6,35)
   35+29=64 → [35,64)

Equal Width Binning - Example

Tree Age [years]	Tree Height [m]
2	6
9	11
9	26
13	31
48	60
47	61
29	86
80	88
64	91
51	96
90	107
77	118

Apply equal width binning to the feature Tree Height with a bin width of 29. The lowest bin boundaries should coincide with the smallest value.

1. Sort the data
2. Distribute elements to bins:
   6+29=35 → [6,35)
   35+29=64 → [35,64)
   64+29=93 → [64,93)

Equal Width Binning - Example

Tree Age [years]	Tree Height [m]
2	6
9	11
9	26
13	31
48	60
47	61
29	86
80	88
64	91
51	96
90	107
77	118

Apply equal width binning to the feature Tree Height with a bin width of 29. The lowest bin boundaries should coincide with the smallest value.

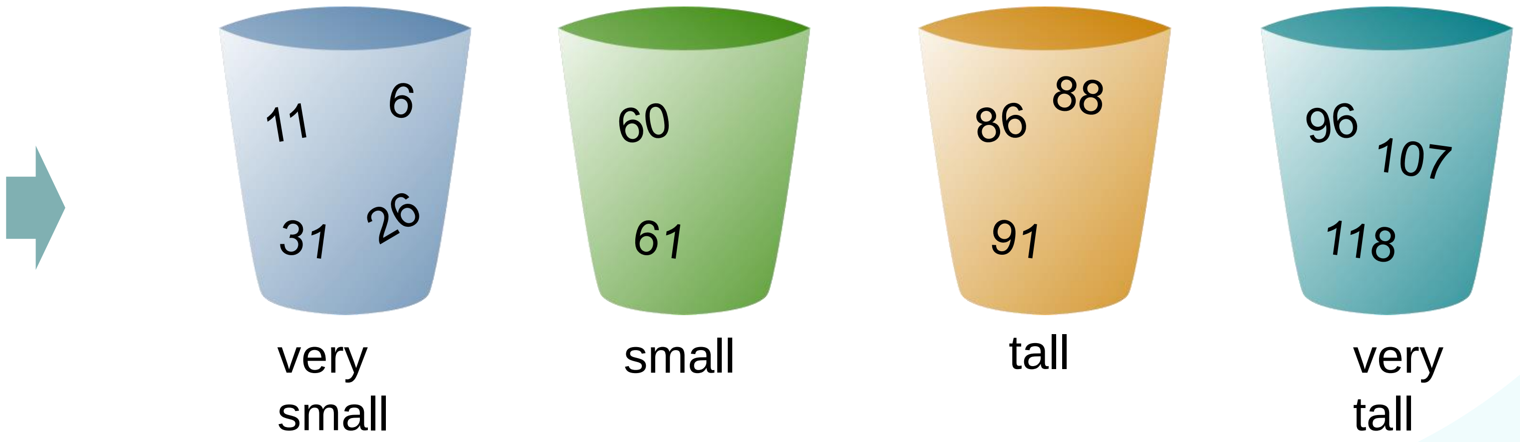
1. Sort the data
2. Distribute elements to bins:
   6+29=35 → [6,35)
   35+29=64 → [35,64)
   64+29=93 → [64,93)
   93+29=122 → [93,122)

Equal Width Binning - Example

Tree Age [years]	Tree Height [m]
2	6
9	11
9	26
13	31
48	60
47	61
29	86
80	88
64	91
51	96
90	107
77	118

Apply equal width binning to the feature Tree Height with a bin width of 29. The lowest bin boundaries should coincide with the smallest value.

1. Sort the data
2. Distribute elements to bins:

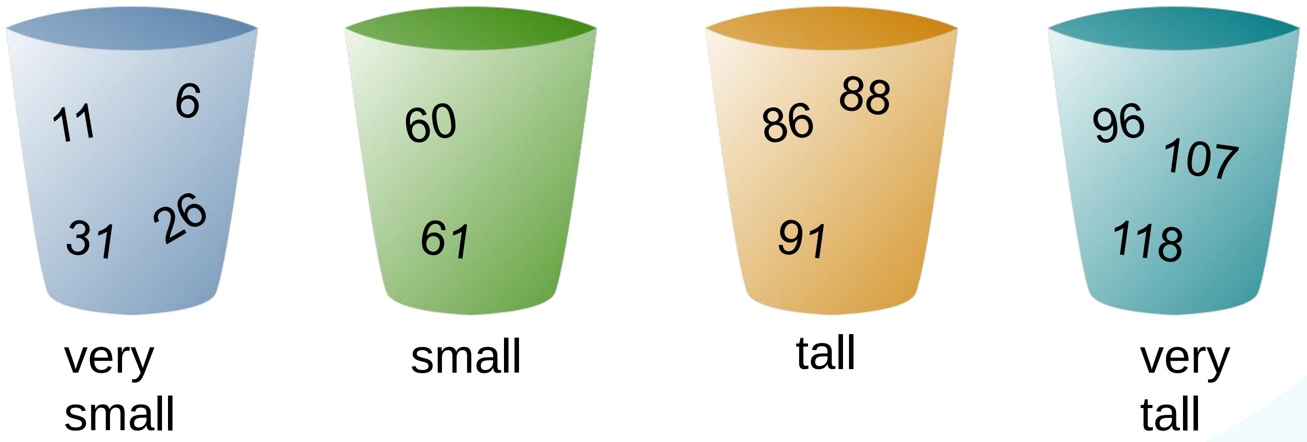


Equal Width Binning - Example

Tree Age [years]	Tree Height [m]
2	very small
9	very small
9	very small
13	very small
48	small
47	small
29	tall
80	tall
64	tall
51	very tall
90	very tall
77	very tall

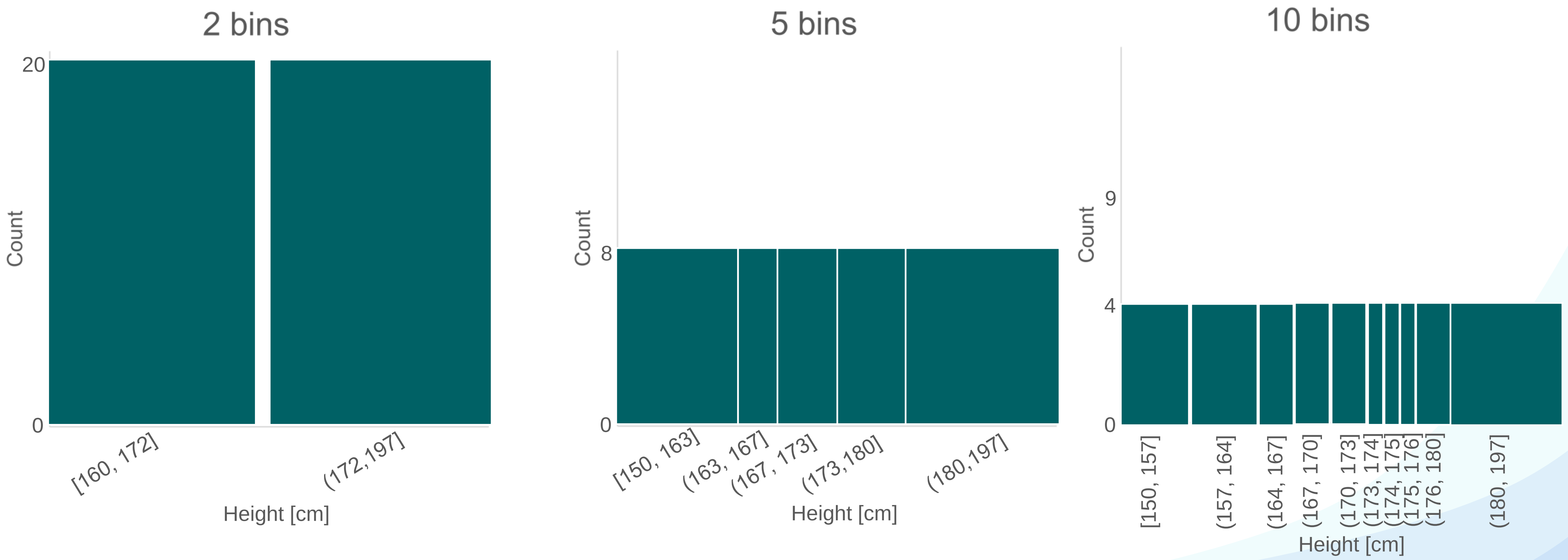
Apply equal width binning to the feature Tree Height with a bin width of 29. The lowest bin boundaries should coincide with the smallest value.

1. Sort the data
2. Distribute elements to bins:



Equal Frequency Binning

Bins vary in width, but the numer of items per bin is fixed



Equal Frequency Binning – Example

Tree Age [years]	Tree Height [m]
9	26
51	96
47	61
77	118
64	91
2	6
48	60
13	31
9	11
29	86
90	107
80	88

Apply equal frequency binning to the feature Tree Age with an element frequency of 4.

Equal Frequency Binning – Example

Tree Age [years]	Tree Height [m]
9	26
51	96
47	61
77	118
64	91
2	6
48	60
13	31
9	11
29	86
90	107
80	88

Apply equal frequency binning to the feature Tree Age with an element frequency of 4.

1. Sort the data

Equal Frequency Binning – Example

Tree Age [years]	Tree Height [m]
2	6
9	26
9	11
13	31
29	86
47	61
48	60
51	96
64	91
77	118
80	88
90	107

Apply equal frequency binning to the feature Tree Age with an element frequency of 4.

1. Sort the data

Equal Frequency Binning – Example

Tree Age [years]	Tree Height [m]
2	6
9	26
9	11
13	31
29	86
47	61
48	60
51	96
64	91
77	118
80	88
90	107

Apply equal frequency binning to the feature Tree Age with an element frequency of 4.

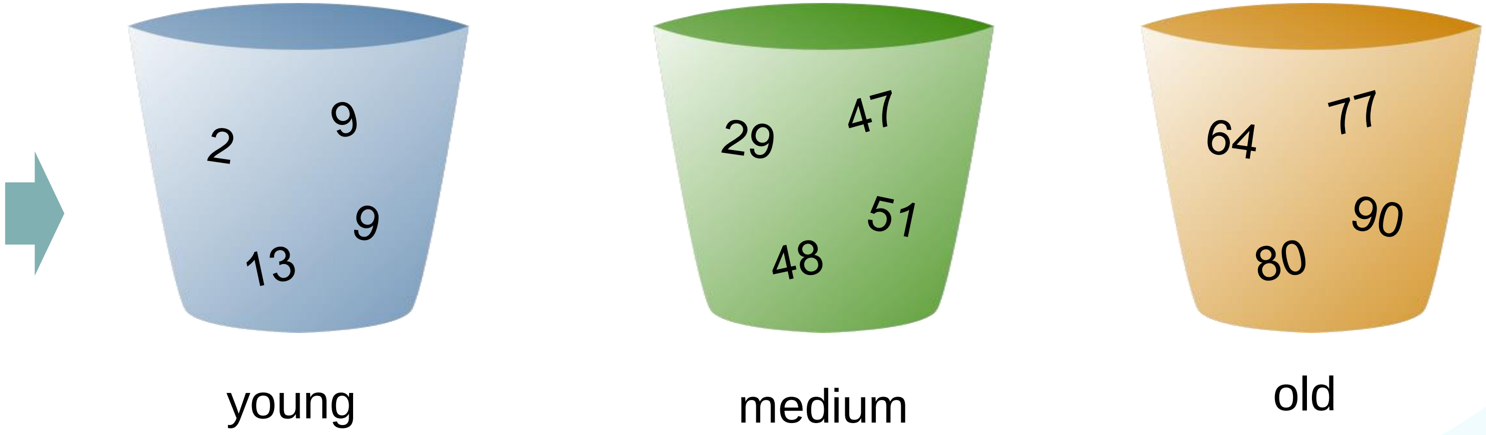
1. Sort the data
2. Distribute elements to bins

Equal Frequency Binning – Example

Tree Age [years]	Tree Height [m]
2	6
9	26
9	11
13	31
29	86
47	61
48	60
51	96
64	91
77	118
80	88
90	107

Apply equal frequency binning to the feature Tree Age with an element frequency of 4.

1. Sort the data
2. Distribute elements to bins

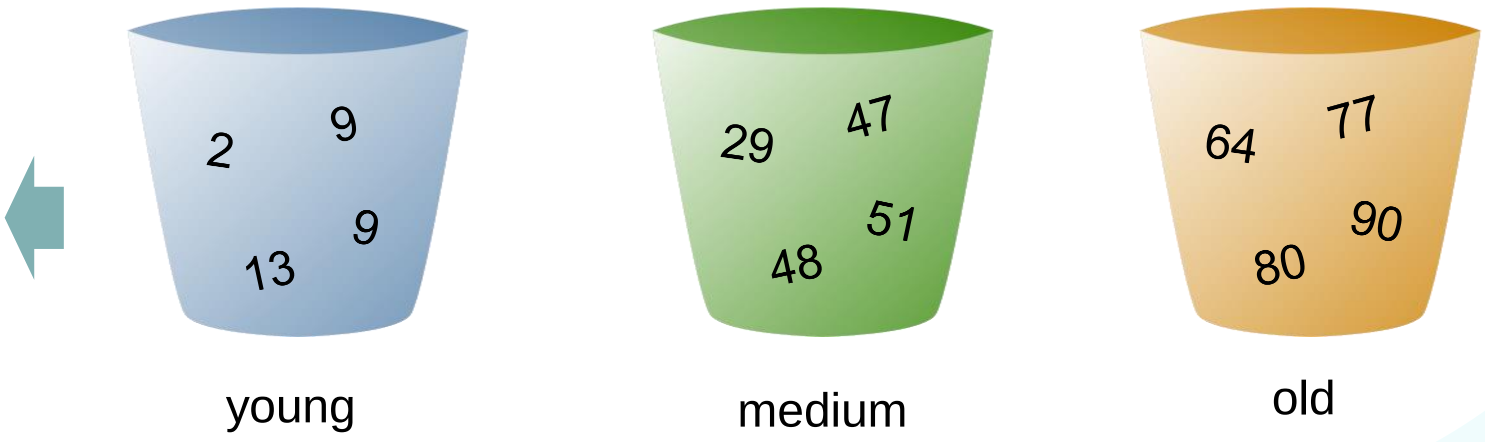


Equal Frequency Binning – Example

Tree Age [years]	Tree Height [m]
young	6
young	26
young	11
young	31
medium	86
medium	61
medium	60
medium	96
old	91
old	118
old	88
old	107

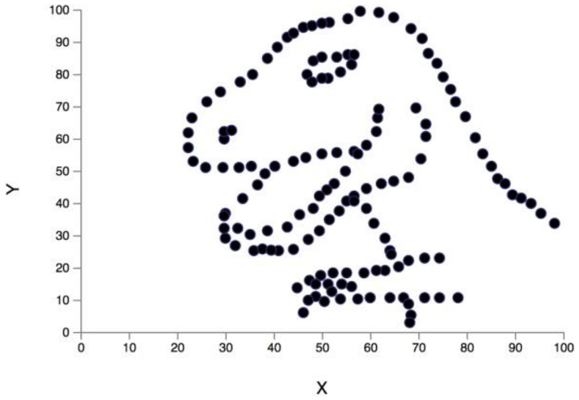
Apply equal frequency binning to the feature Tree Age with an element frequency of 4.

1. Sort the data
2. Distribute elements to bins



Key Points

- Raw data has no value, we need to extract information
- Not just known unknowns, also unknown unknowns
- Visual exploration is a vital first step
  (initial understanding, spotting data quality problems, building trust, etc.)
  - Humans have pretty good visual pattern recognition abilities – use them!



How to Lie with Statistics

... or, how to avoid misleading information and visualizations.

“There are lies, damn lies, and statistics”

- Anonymous

Mark Twain?

Benjamin Disraeli?

How to Lie with Statistics

... or, how to avoid misleading information and visualizations.

- Design choice in presenting data and statistics have a huge impact!

- …Even if what is shown is **technically true**

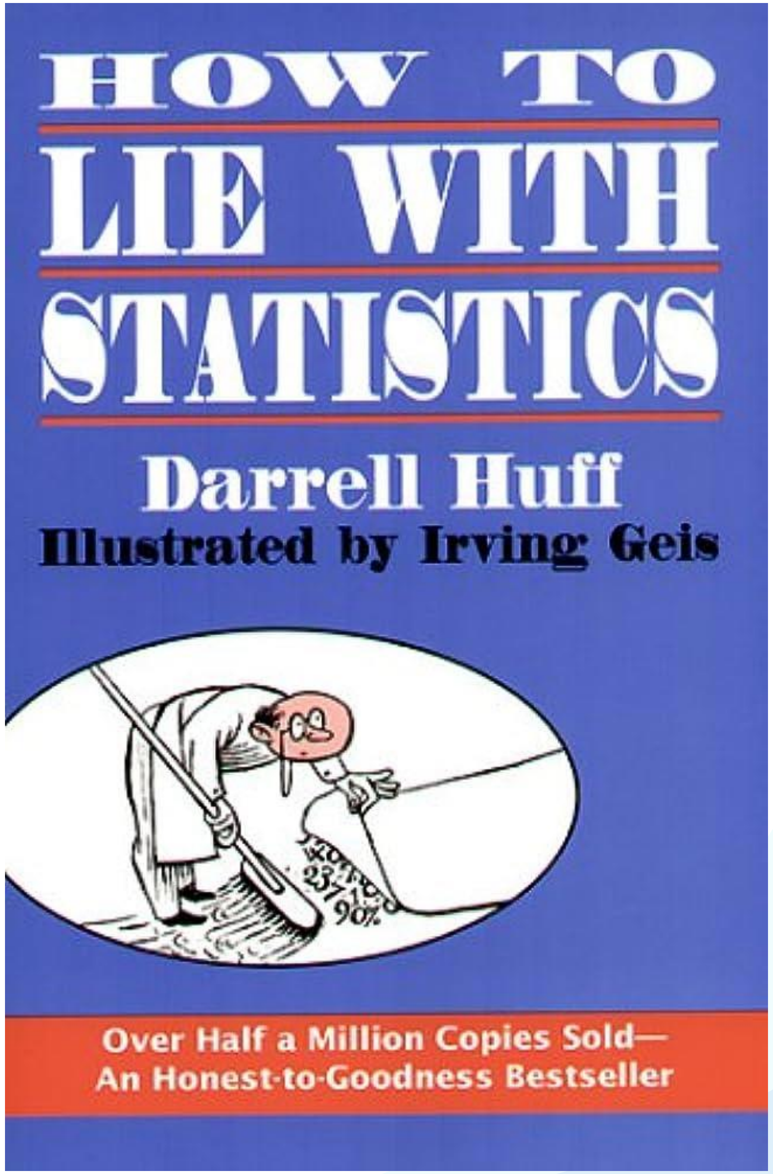
How to Lie with Statistics

... or, how to avoid misleading information and visualizations.

- Design choices in presenting data and statistics have a huge impact!

- …Even if what is shown is **technically true**

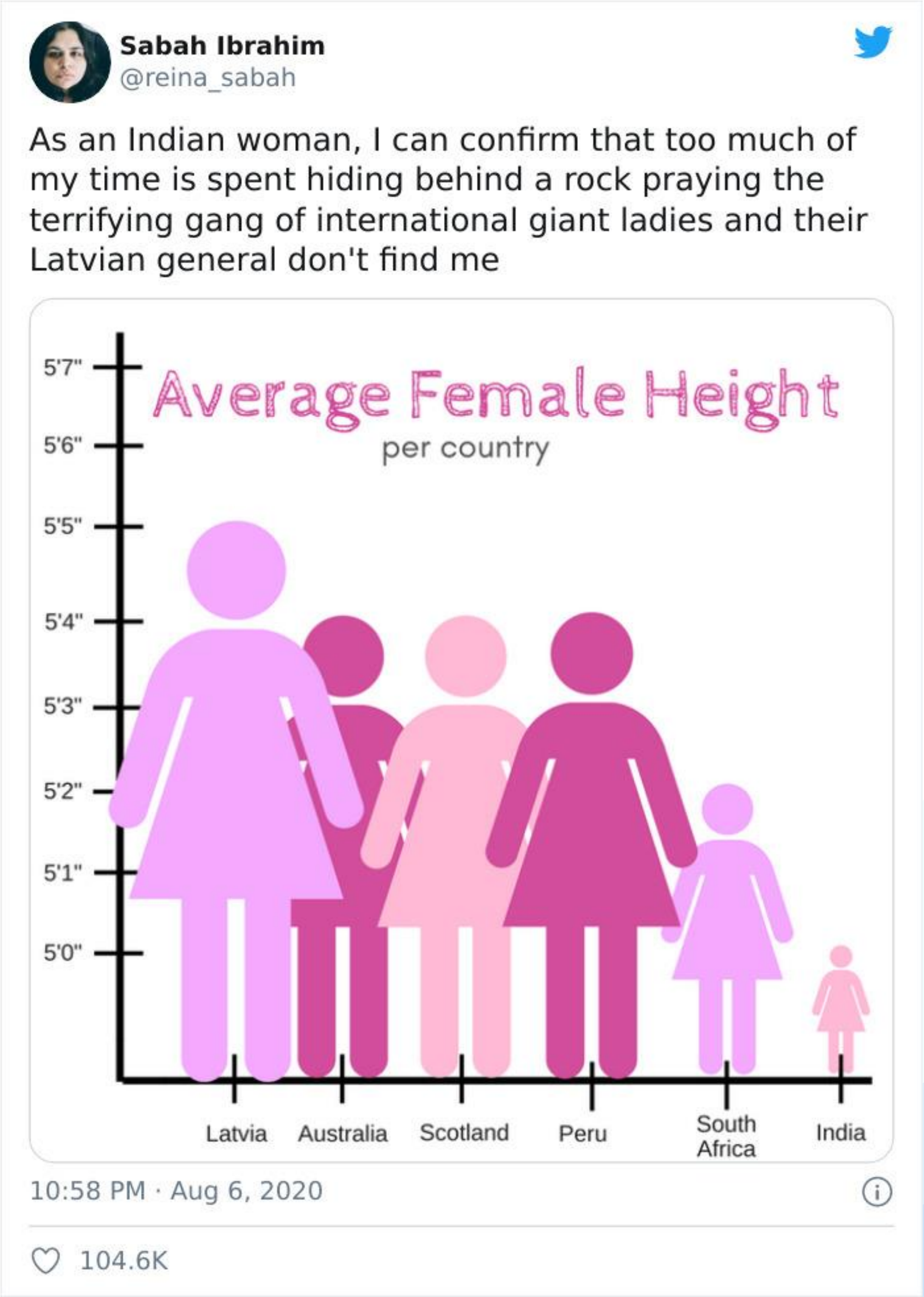
- For an extreme example, search the case of **Sally Clark** (Discretion advised)



How to Lie with Statistics

... or, how to avoid misleading information and visualizations.

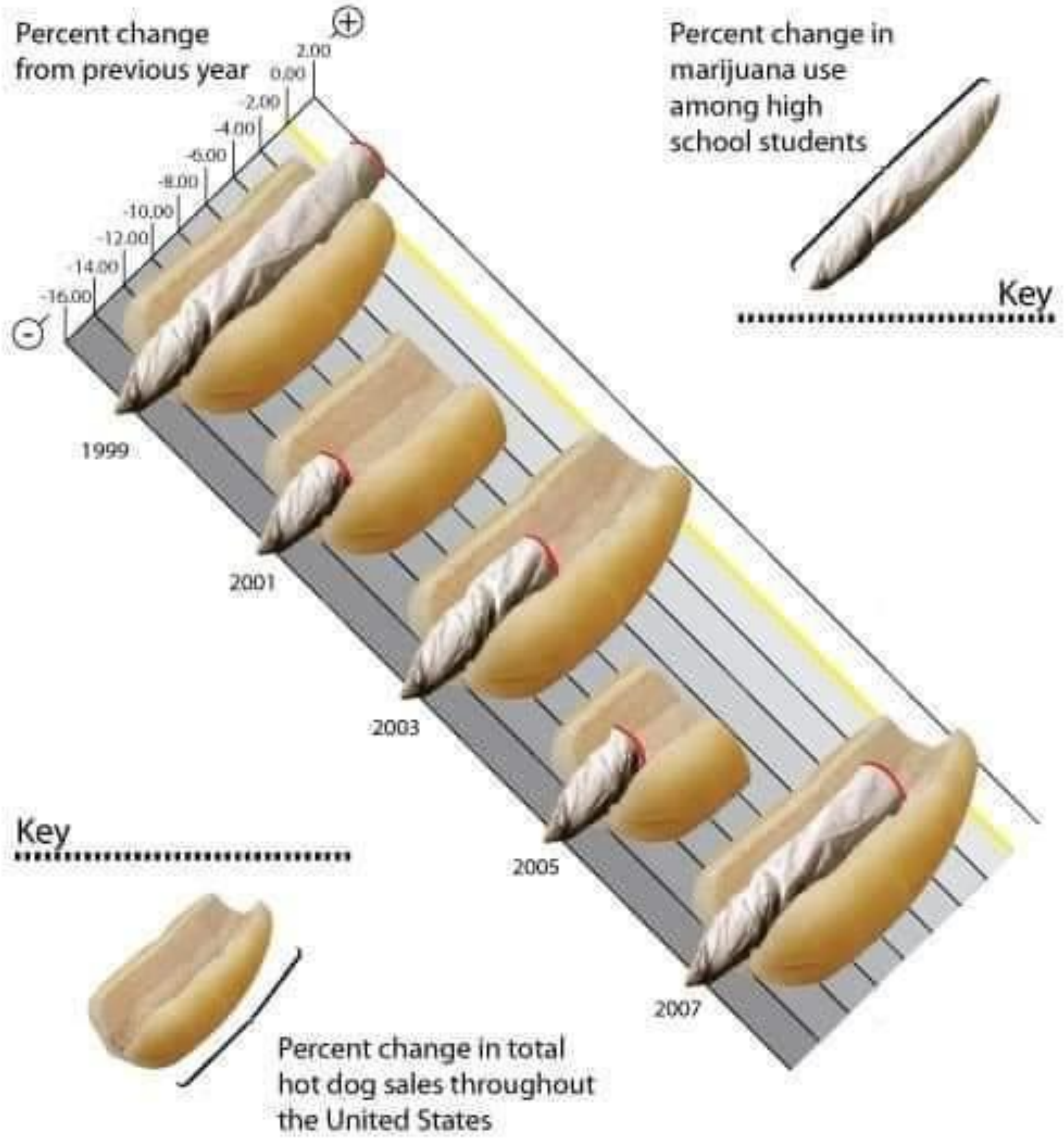
In some cases, rather than lying, the design is just **hilariously bad.**



How to Lie with Statistics

... or, how to avoid misleading information and visualizations.

In some cases, rather than lying, the design is just **hilariously bad.**

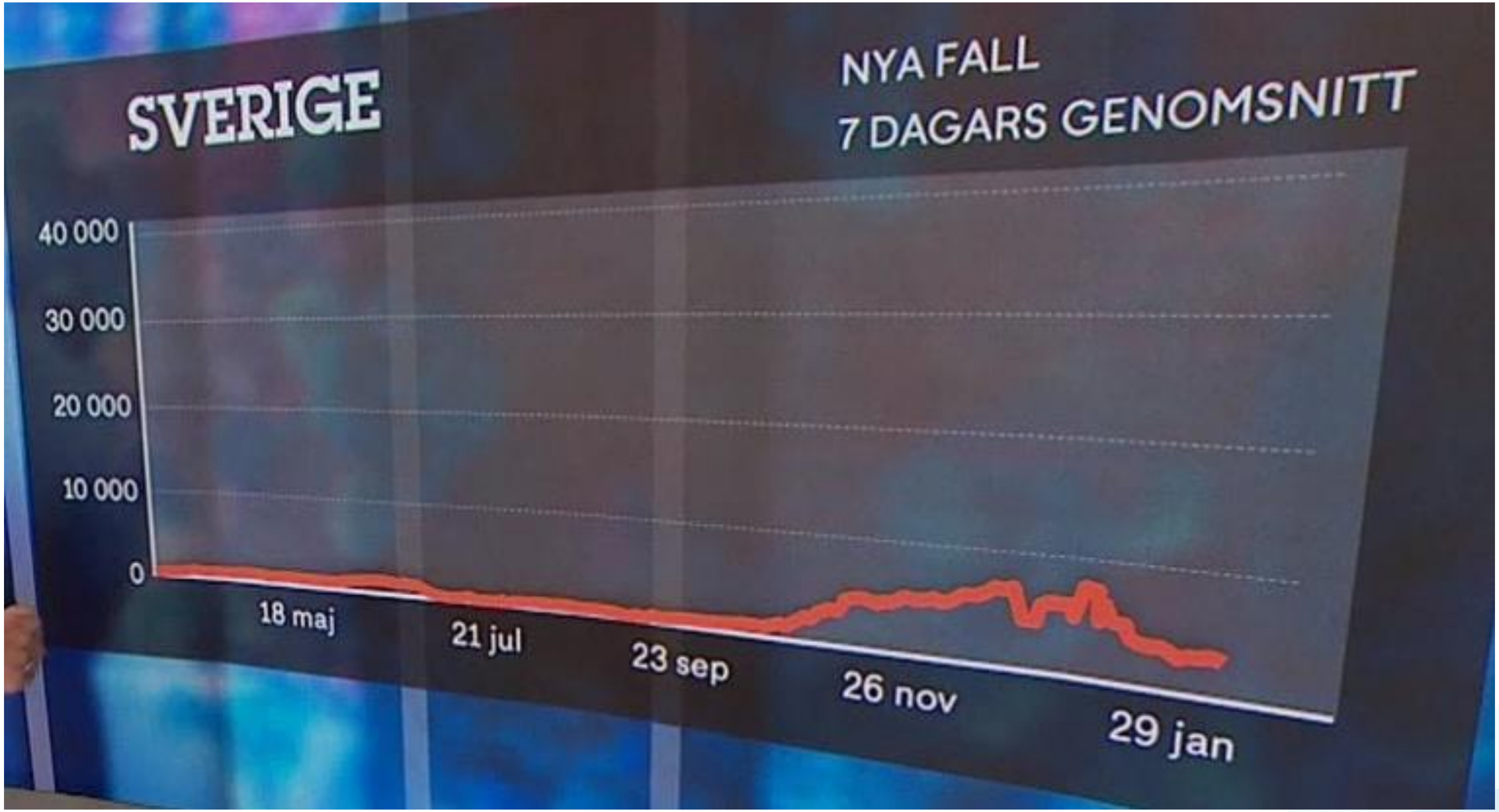


How to Lie with Statistics

... or, how to avoid misleading information and visualizations.

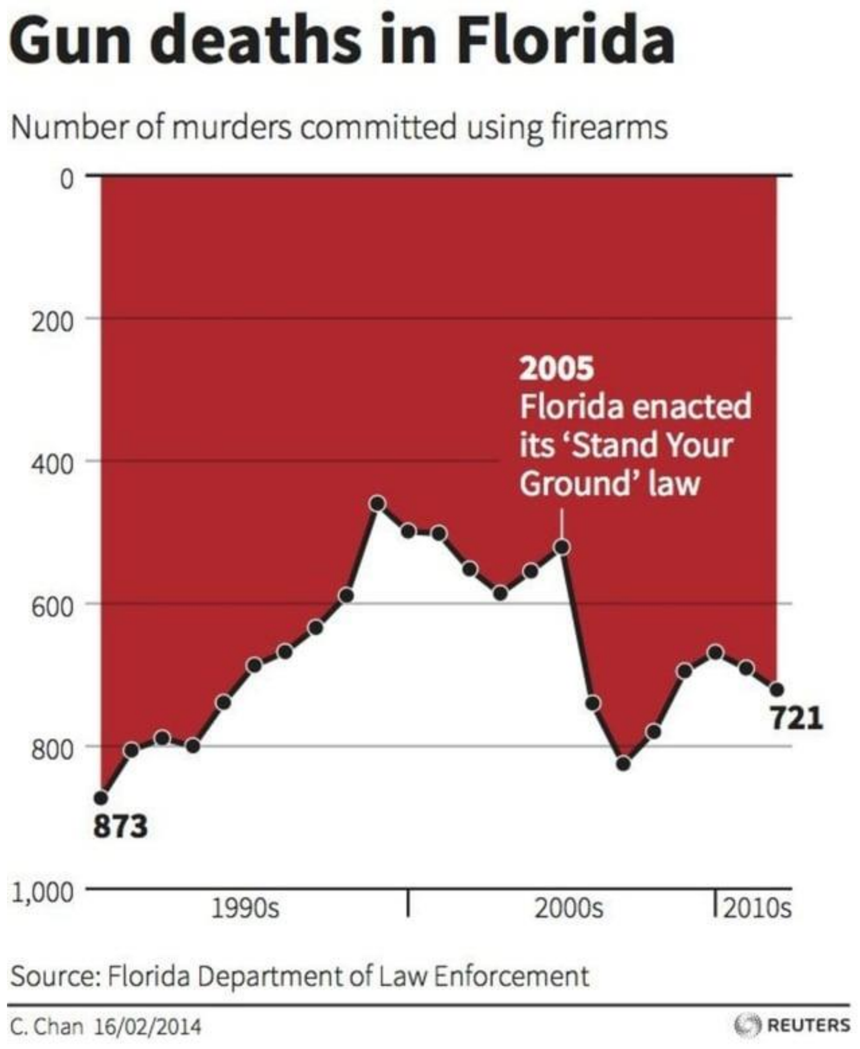
How to Lie with Statistics

... or, how to avoid misleading information and visualizations.



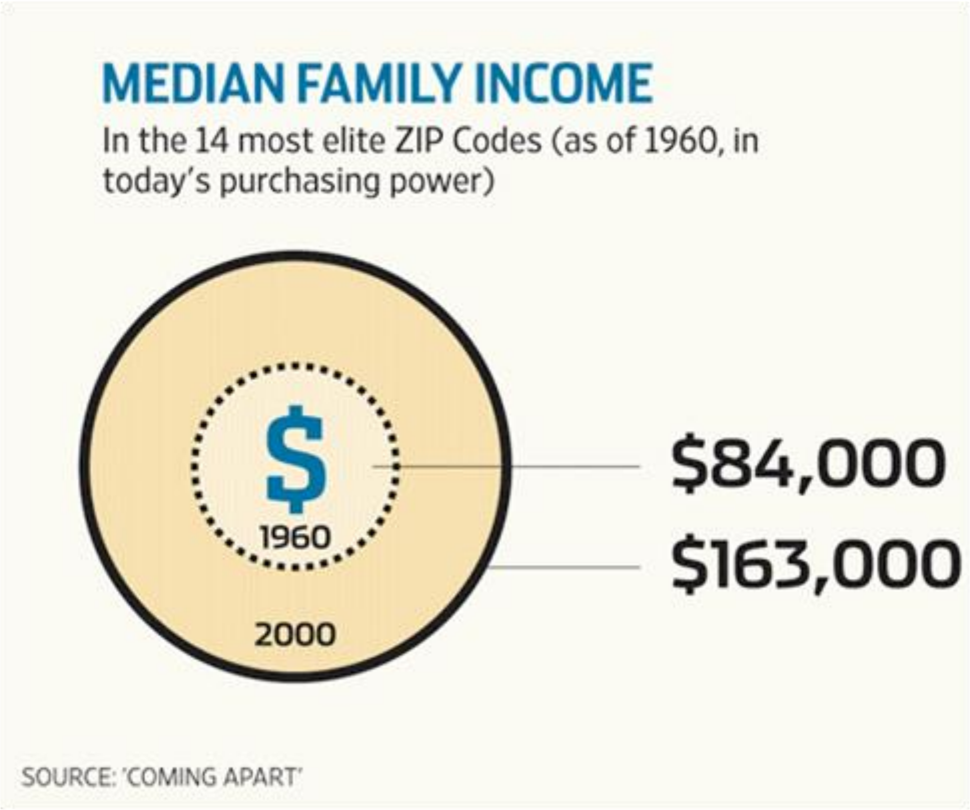
How to Lie with Statistics

... or, how to avoid misleading information and visualizations.



How to Lie with Statistics

... or, how to avoid misleading information and visualizations.

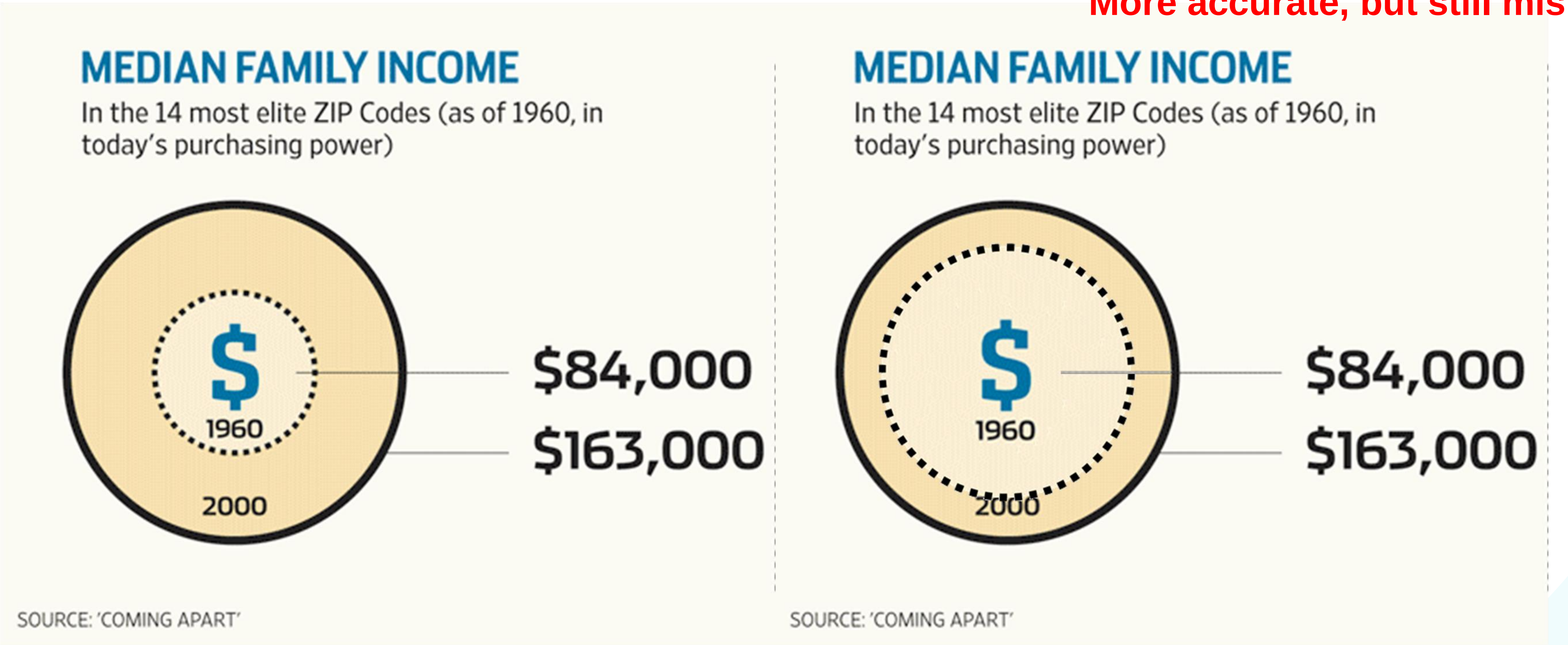


The Wall Street Journal, 2012

How to Lie with Statistics

... or, how to avoid misleading information and visualizations.

More accurate, but still misleading!

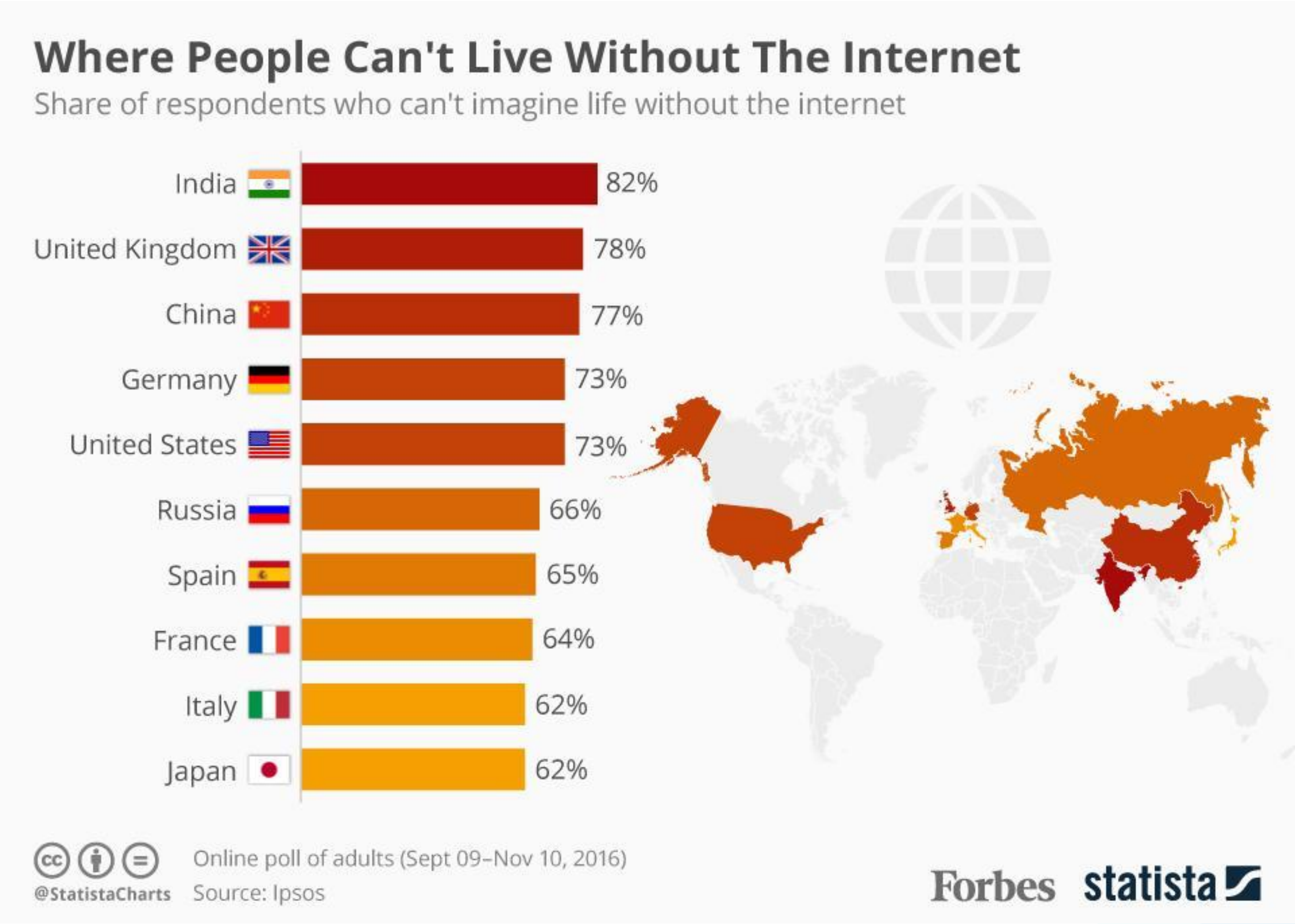


Inaccurate graph as it appeared in the *Wall Street Journal* (1/21/2012)

Accurate graph constructed by EvalBlog.com

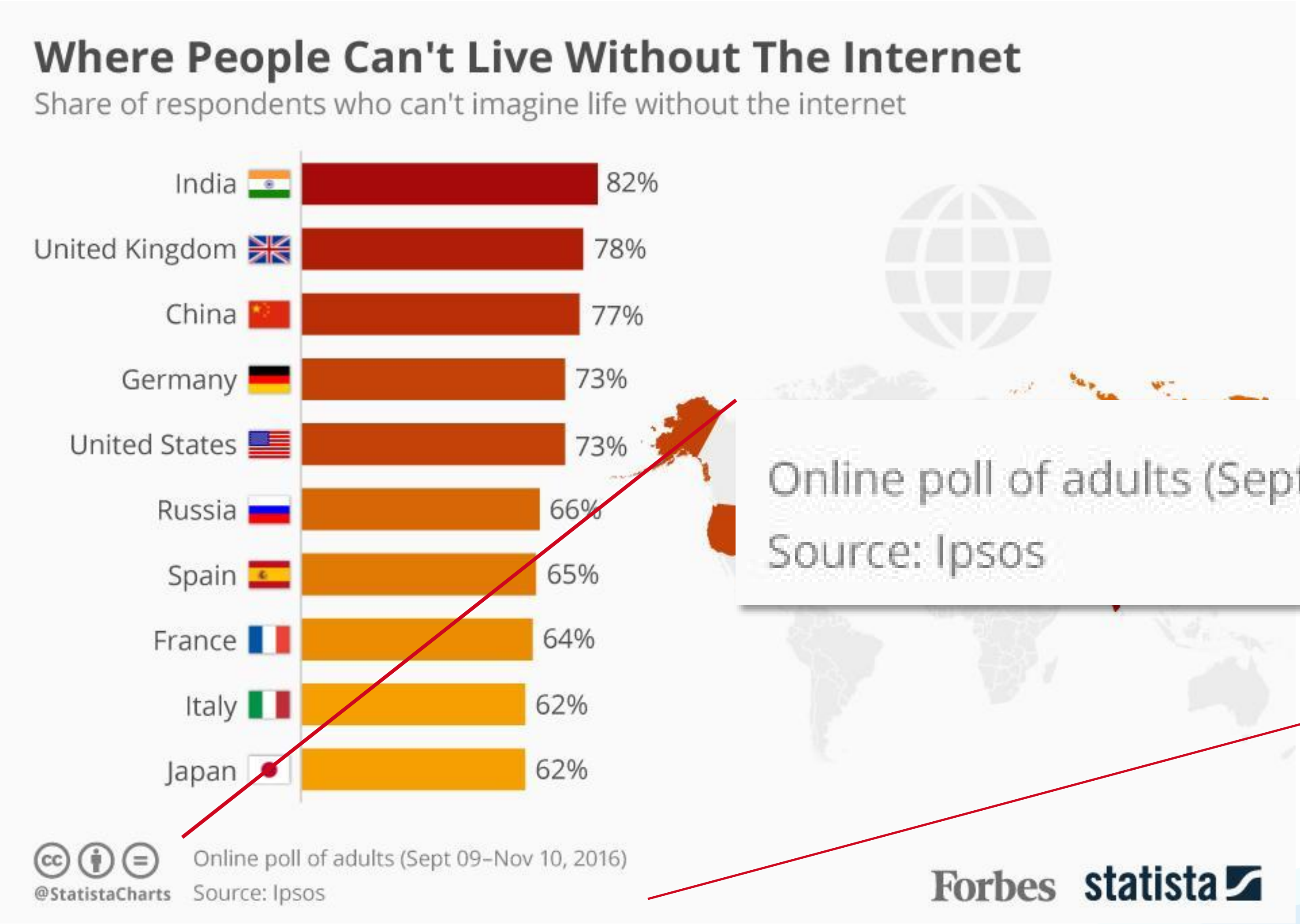
How to Lie with Statistics

... or, how to avoid misleading information and visualizations.



How to Lie with Statistics

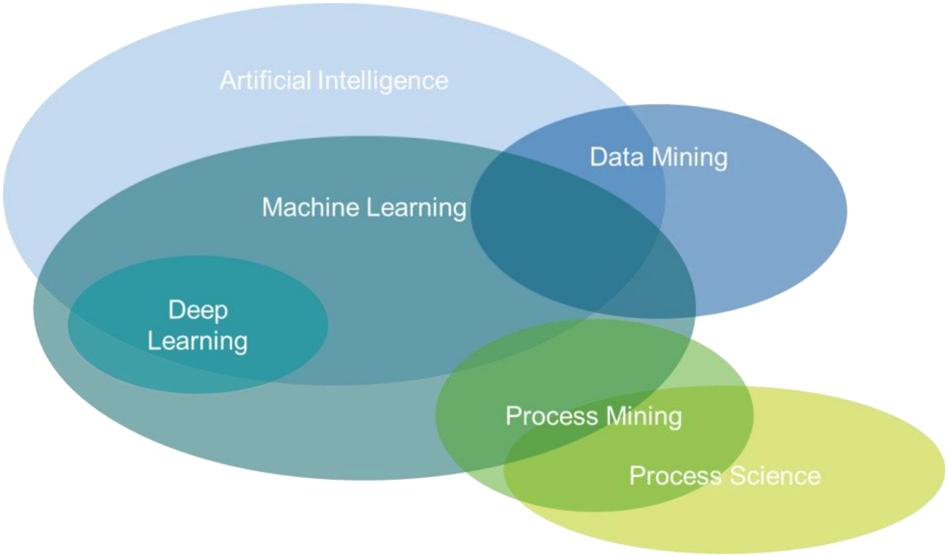
... or, how to avoid misleading information and visualizations.

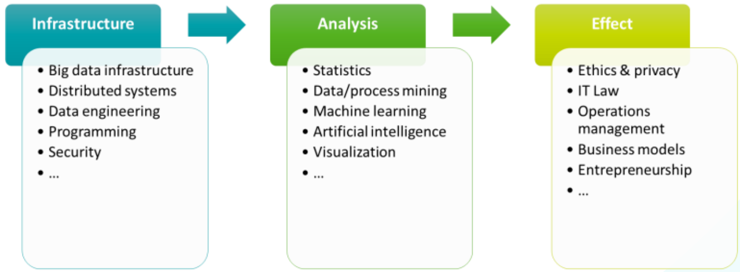


Online poll of adults (Sept 09–Nov 10, 2016)
Source: Ipsos

Wrap up

- Data is vital, but hard to manage!
- Obtaining insights is a looping process, not a one-off application of algorithms
- Various criticalities, such as noise and bias
- Visualization is always fundamental…
- …and comes with its own challenges!





Next up: Decision Trees