IRT SAINT EXUPÉRY

Data Exploration

Antonin POCHE

Mouhcine MENDIL

“Torture the data, and it will confess to anything”.
Ronald Coase

What is Machine Learning ?

Definition of Machine Learning:

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

TOM MITCHELL

What is Machine Learning ?

Exemples of Machine Learning (T. Mitchell)

Use case	Chess	Handwriting recognition	Financial trading
Task T	Playing chess	Recognizing and classiying handwritten words within images	Predict stock prices
Performance measure P	Percent of games won against opponents	Percent of words correctly classified	Average absolute error between predicted and real prices
Training experience E	Playing practice games against itself	A database of handwritten words with given classification labels	Historical stock prices and market states

- The machine learns how to perform tasks without being explicitly programmed.
- Instead, it learns from the **<u>data</u>** to build rules and knowledge.

What is Machine Learning ?

Exemples of Machine Learning (T. Mitchell)

Use case	Chess	Handw re	Financial trading
Task T	Playing chess	R ying ds within ages	Predict stock prices
Performance measure P	Percent of game o	ercent of words correctly classified	Average absolute error between predicted and real prices
Training experience E	F es against elf	A database of handwritten words with given classification labels	Historical stock prices and market states

WE NEED DATA !!

- The machine learns how to perform tasks without being explicitly programmed.
- Instead, it learns from the **data** to build rules and knowledge.

What is « Data »?

Definitions of data:

[…] result of an observation made on a population or a sample.

Dodge (2007)

- Data are a collection of […] values that convey information, describing quantity, quality, fact, statistics, other basic units of meaning.

Wikipedia

What is « Data »?

More concretly

- Data can be a number or a symbol that inform on an individual, an object or an observation.

Example

- 5000 is a number (not interesting by itself)
- « My salary is 5000€ » is data (information on M. Dupont)

Variable

- A variable is a mathematical object related to a given concept (e.g. salary)
- It can take different values coming from different observations/individuals/objects
  - Example 1: X_i salary of an person i in France (1500€, 4000€, 500000€, …)
  - Example 2: $X_{k,t}$ salary evolution of a person k with time t (3500€, 4000€, 4500€, 5000€, …)

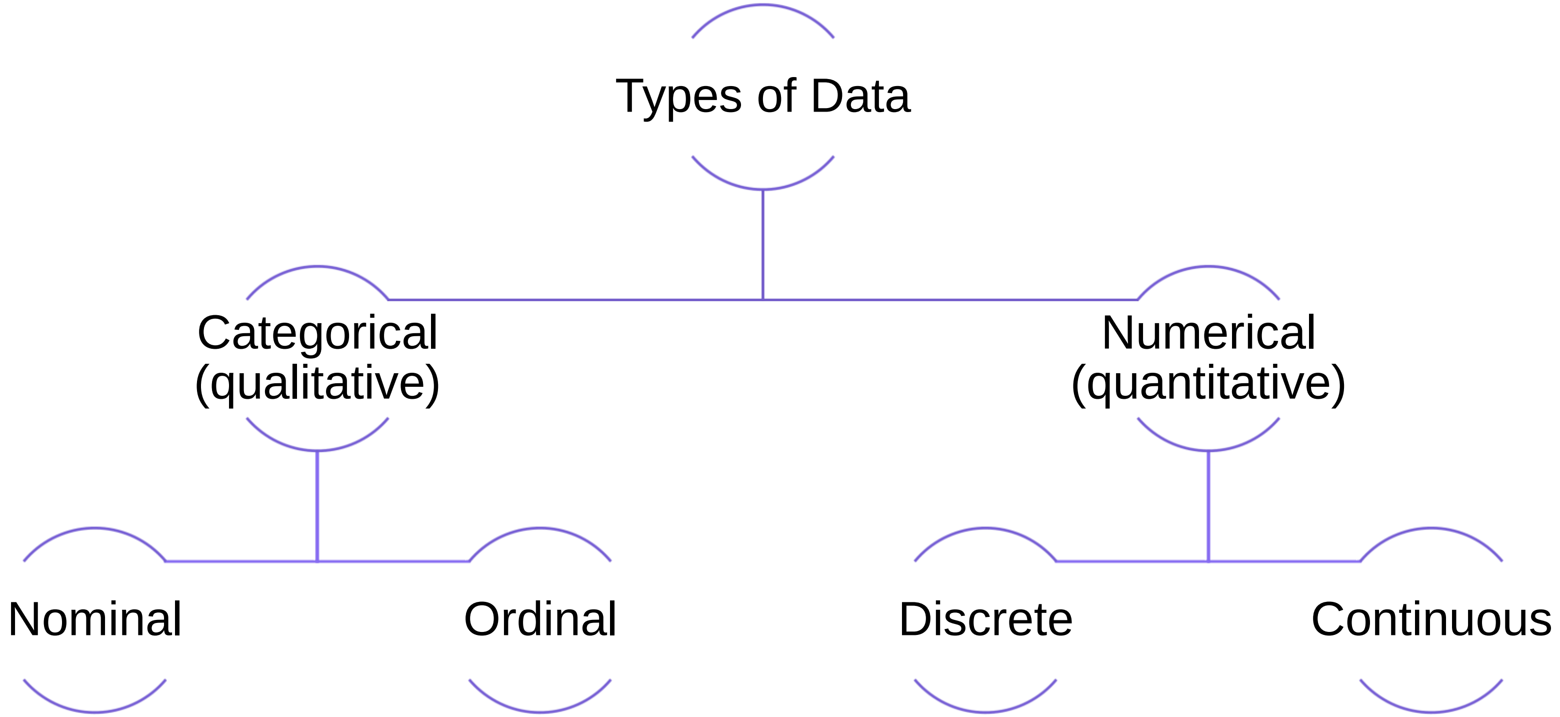
FRENCH INSTITUTES OF TECHNOLOGY

Origin of « Data »

- Every day, we create roughly 2.5 quintillion bytes of data
- Data come from anything observable through **surveys**, **sensors**, **logs**, etc…

- **Private** data: e.g., emails and pictures
- **Public** data: open data, open APIs, web (scraping)
- Data **operations**: modeling, storage, access and processing
  - Different formats: audio, video, time series, tabular, text …
  - Data centers, data lakes, data hubs …

Types of Data



Types of Data

Categorical (qualitative)

Numerical (quantitative)

Nominal

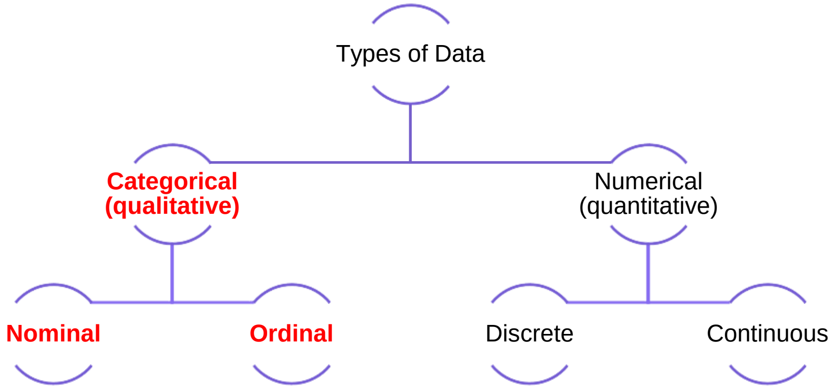
Ordinal

Discrete

Continuous

Categorical Data

- Describe qualitative characteristics (e.g., gender)
- Do not provide any quantitative value
- Do not support arithmetical operations



Nominal Data

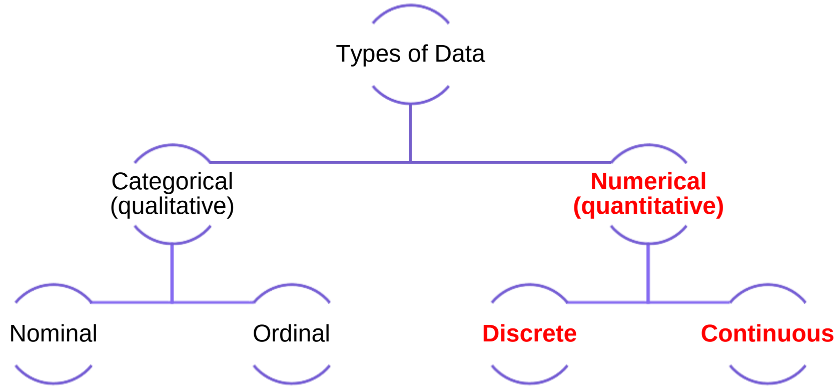
- No intrinsic ordering: can not be compared
- Examples: gender, marital status, color

Ordinal Data

- Have a logical sequential order
- Examples: clothes size (S, M, L, XL…), satisfaction (low, medium, high), opinion (strongly disagree, disagree, agree, strongly agree)

Numerical Data

- Expressed in numerical values (price, height …)
- Support arithmetical operations and statistical analysis



Discrete Data

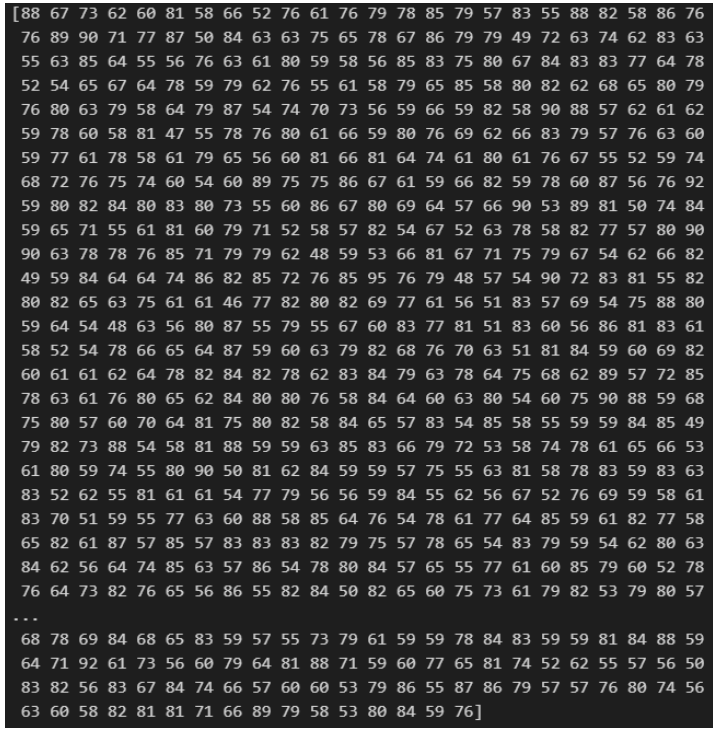
- Can have only finite (or countable) values (integer numbering)
- Examples: number of children, number of rooms in a house

Continuous Data

- Can take infinite number of values (real values)
- Examples: weight, temperature, unemployment rate

Statistics

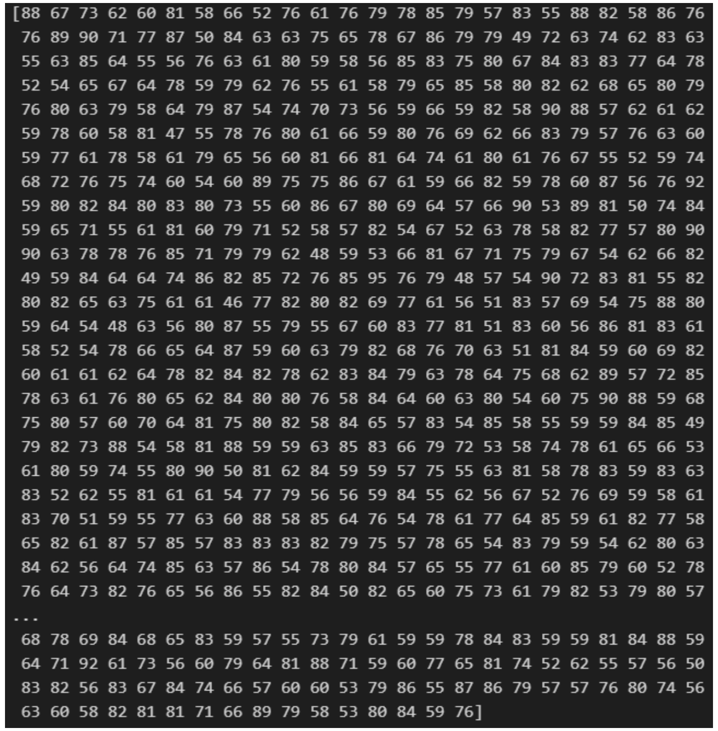
- Consider the weights of a group of 1000 persons.
- Hard to describe the data by just looking at the numbers ☹ !



```
[88 67 73 62 60 81 58 66 52 76 61 76 79 78 85 79 57 83 55 88 82 58 86 76
 76 89 90 71 77 87 50 84 63 63 75 65 78 67 86 79 79 49 72 63 74 62 83 63
 55 63 85 64 55 56 76 63 61 80 59 58 56 85 83 75 80 67 84 83 83 77 64 78
 52 54 65 67 64 78 59 79 62 76 55 61 58 79 65 85 58 80 82 62 68 65 80 79
 76 80 63 79 58 64 79 87 54 74 70 73 56 59 66 59 82 58 90 88 57 62 61 62
 59 78 60 58 81 47 55 78 76 80 61 66 59 80 76 69 62 66 83 79 57 76 63 60
 59 77 61 78 58 61 79 65 56 60 81 66 81 64 74 61 80 61 76 67 55 52 59 74
 68 72 76 75 74 60 54 60 89 75 75 86 67 61 59 66 82 59 78 60 87 56 76 92
 59 80 82 84 80 83 80 73 55 60 86 67 80 69 64 57 66 90 53 89 81 50 74 84
 59 65 71 55 61 81 60 79 71 52 58 57 82 54 67 52 63 78 58 82 77 57 80 90
 90 63 78 78 76 85 71 79 79 62 48 59 53 66 81 67 71 75 79 67 54 62 66 82
 49 59 84 64 64 74 86 82 85 72 76 85 95 76 79 48 57 54 90 72 83 81 55 82
 80 82 65 63 75 61 61 46 77 82 80 82 69 77 61 56 51 83 57 69 54 75 88 80
 59 64 54 48 63 56 80 87 55 79 55 67 60 83 77 81 51 83 60 56 86 81 83 61
 58 52 54 78 66 65 64 87 59 60 63 79 82 68 76 70 63 51 81 84 59 60 69 82
 60 61 61 62 64 78 82 84 82 78 62 83 84 79 63 78 64 75 68 62 89 57 72 85
 78 63 61 76 80 65 62 84 80 80 76 58 84 64 60 63 80 54 60 75 90 88 59 68
 75 80 57 60 70 64 81 75 80 82 58 84 65 57 83 54 85 58 55 59 59 84 85 49
 79 82 73 88 54 58 81 88 59 59 63 85 83 66 79 72 53 58 74 78 61 65 66 53
 61 80 59 74 55 80 90 50 81 62 84 59 59 57 75 55 63 81 58 78 83 59 83 63
 83 52 62 55 81 61 61 54 77 79 56 56 59 84 55 62 56 67 52 76 69 59 58 61
 83 70 51 59 55 77 63 60 88 58 85 64 76 54 78 61 77 64 85 59 61 82 77 58
 65 82 61 87 57 85 57 83 83 83 82 79 75 57 78 65 54 83 79 59 54 62 80 63
 84 62 56 64 74 85 63 57 86 54 78 80 84 57 65 55 77 61 60 85 79 60 52 78
 76 64 73 82 76 65 56 86 55 82 84 50 82 65 60 75 73 61 79 82 53 79 80 57
...
 68 78 69 84 68 65 83 59 57 55 73 79 61 59 59 78 84 83 59 59 81 84 88 59
 64 71 92 61 73 56 60 79 64 81 88 71 59 60 77 65 81 74 52 62 55 57 56 50
 83 82 56 83 67 84 74 66 57 60 60 53 79 86 55 87 86 79 57 57 76 80 74 56
 63 60 58 82 81 81 71 66 89 79 58 53 80 84 59 76]
```

Statistics

- Consider the weights of a group of 1000 persons.
- Hard to describe the data by just looking at the numbers ☹ !



```
[88 67 73 62 60 81 58 66 52 76 61 76 79 78 85 79 57 83 55 88 82 58 86 76
 76 89 90 71 77 87 50 84 63 63 75 65 78 67 86 79 79 49 72 63 74 62 83 63
 55 63 85 64 55 56 76 63 61 80 59 58 56 85 83 75 80 67 84 83 83 77 64 78
 52 54 65 67 64 78 59 79 62 76 55 61 58 79 65 85 58 80 82 62 68 65 80 79
 76 80 63 79 58 64 79 87 54 74 70 73 56 59 66 59 82 58 90 88 57 62 61 62
 59 78 60 58 81 47 55 78 76 80 61 66 59 80 76 69 62 66 83 79 57 76 63 60
 59 77 61 78 58 61 79 65 56 60 81 66 81 64 74 61 80 61 76 67 55 52 59 74
 68 72 76 75 74 60 54 60 89 75 75 86 67 61 59 66 82 59 78 60 87 56 76 92
 59 80 82 84 80 83 80 73 55 60 86 67 80 69 64 57 66 90 53 89 81 50 74 84
 59 65 71 55 61 81 60 79 71 52 58 57 82 54 67 52 63 78 58 82 77 57 80 90
 90 63 78 78 76 85 71 79 79 62 48 59 53 66 81 67 71 75 79 67 54 62 66 82
 49 59 84 64 64 74 86 82 85 72 76 85 95 76 79 48 57 54 90 72 83 81 55 82
 80 82 65 63 75 61 61 46 77 82 80 82 69 77 61 56 51 83 57 69 54 75 88 80
 59 64 54 48 63 56 80 87 55 79 55 67 60 83 77 81 51 83 60 56 86 81 83 61
 58 52 54 78 66 65 64 87 59 60 63 79 82 68 76 70 63 51 81 84 59 60 69 82
 60 61 61 62 64 78 82 84 82 78 62 83 84 79 63 78 64 75 68 62 89 57 72 85
 78 63 61 76 80 65 62 84 80 80 76 58 84 64 60 63 80 54 60 75 90 88 59 68
 75 80 57 60 70 64 81 75 80 82 58 84 65 57 83 54 85 58 55 59 59 84 85 49
 79 82 73 88 54 58 81 88 59 59 63 85 83 66 79 72 53 58 74 78 61 65 66 53
 61 80 59 74 55 80 90 50 81 62 84 59 59 57 75 55 63 81 58 78 83 59 83 63
 83 52 62 55 81 61 61 54 77 79 56 56 59 84 55 62 56 67 52 76 69 59 58 61
 83 70 51 59 55 77 63 60 88 58 85 64 76 54 78 61 77 64 85 59 61 82 77 58
 65 82 61 87 57 85 57 83 83 83 82 79 75 57 78 65 54 83 79 59 54 62 80 63
 84 62 56 64 74 85 63 57 86 54 78 80 84 57 65 55 77 61 60 85 79 60 52 78
 76 64 73 82 76 65 56 86 55 82 84 50 82 65 60 75 73 61 79 82 53 79 80 57
...
 68 78 69 84 68 65 83 59 57 55 73 79 61 59 59 78 84 83 59 59 81 84 88 59
 64 71 92 61 73 56 60 79 64 81 88 71 59 60 77 65 81 74 52 62 55 57 56 50
 83 82 56 83 67 84 74 66 57 60 60 53 79 86 55 87 86 79 57 57 76 80 74 56
 63 60 58 82 81 81 71 66 89 79 58 53 80 84 59 76]
```

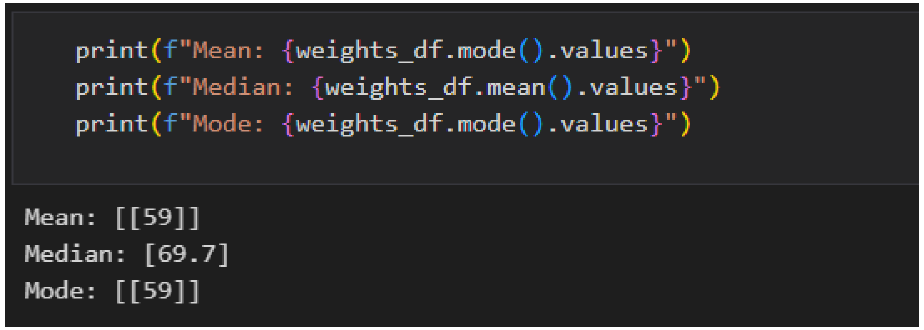
Descriptive statistics

- Summarizes or describes the characteristics of a data set.

- Consists of three basic categories of measures:
  - **Central tendency**: describes the center of the data set (mean, median, mode)
  - **Variability (or spread)**: describes the dispersion of the data set (variance, standard deviation)
  - **Frequency distribution**: describes the occurrence of data within the data set (count)

Statistics

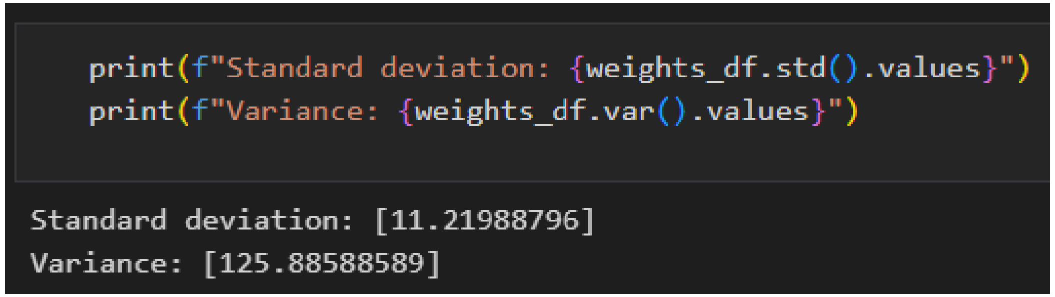
Central tendencies



```
print(f"Mean: {weights_df.mode().values}")
print(f"Median: {weights_df.mean().values}")
print(f"Mode: {weights_df.mode().values}")
```

```
Mean: [[59]]
Median: [69.7]
Mode: [[59]]
```

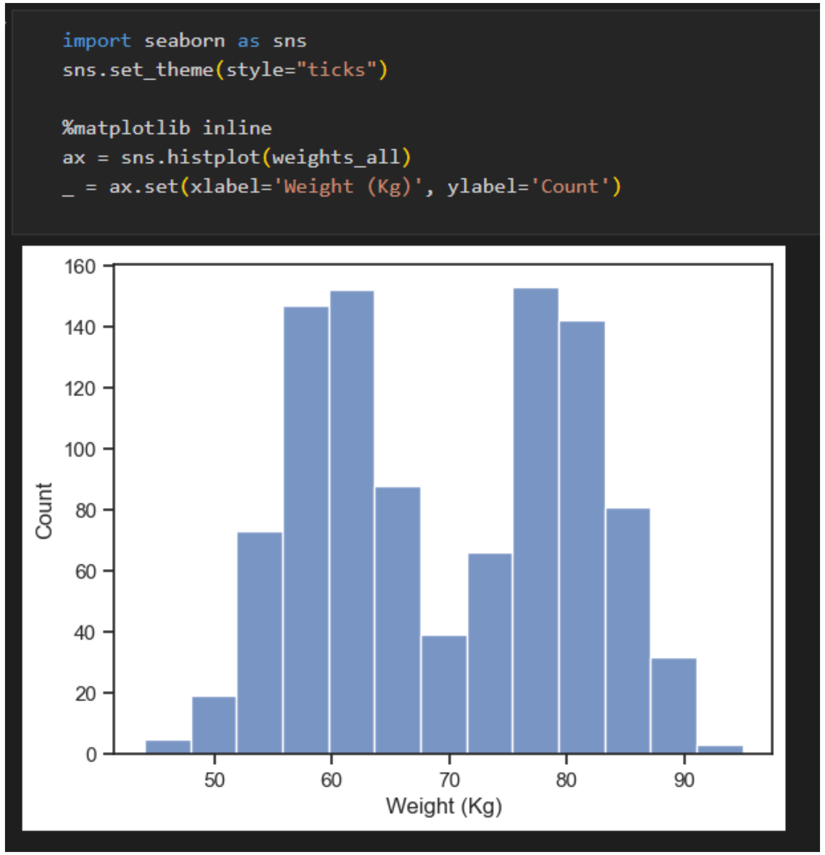
Variability



```
print(f"Standard deviation: {weights_df.std().values}")
print(f"Variance: {weights_df.var().values}")
```

```
Standard deviation: [11.21988796]
Variance: [125.88588589]
```

Frequency Distribution



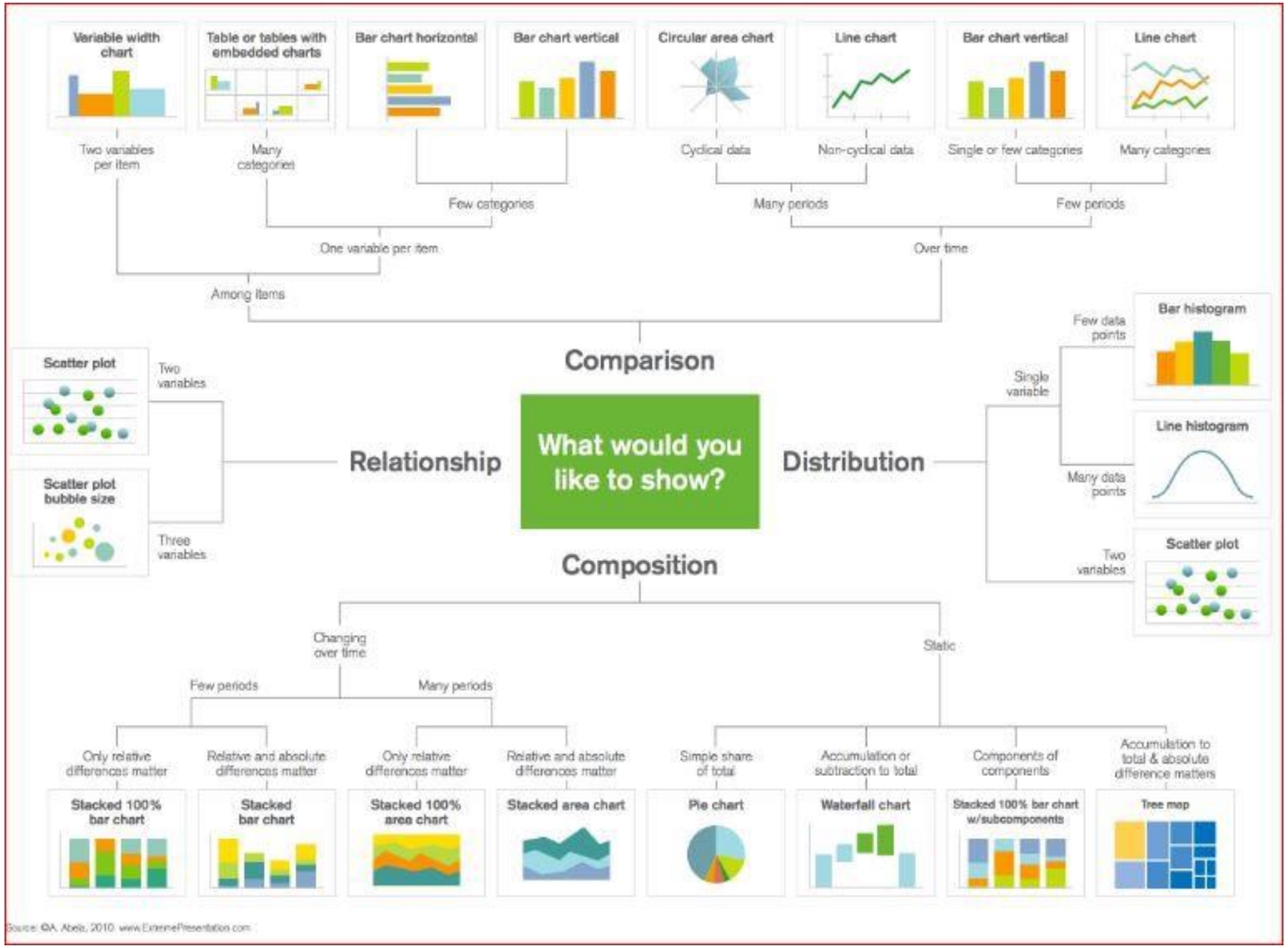
```
import seaborn as sns
sns.set_theme(style="ticks")

%matplotlib inline
ax = sns.histplot(weights_all)
_ = ax.set(xlabel='Weight (Kg)', ylabel='Count')
```

Data Visualization

- Translates information into a visual context, such as a map or graph.

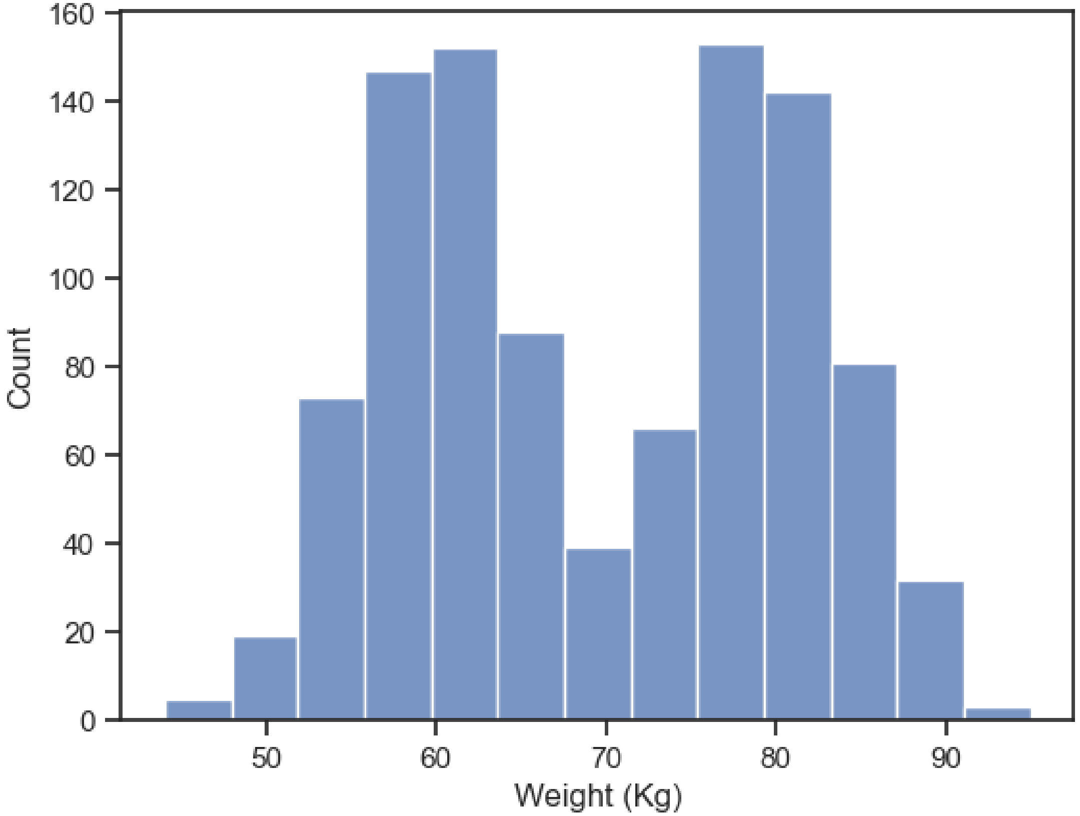
- Makes data easier for the human brain to understand and pull insights from.



[Source: www.kdnuggets.com]

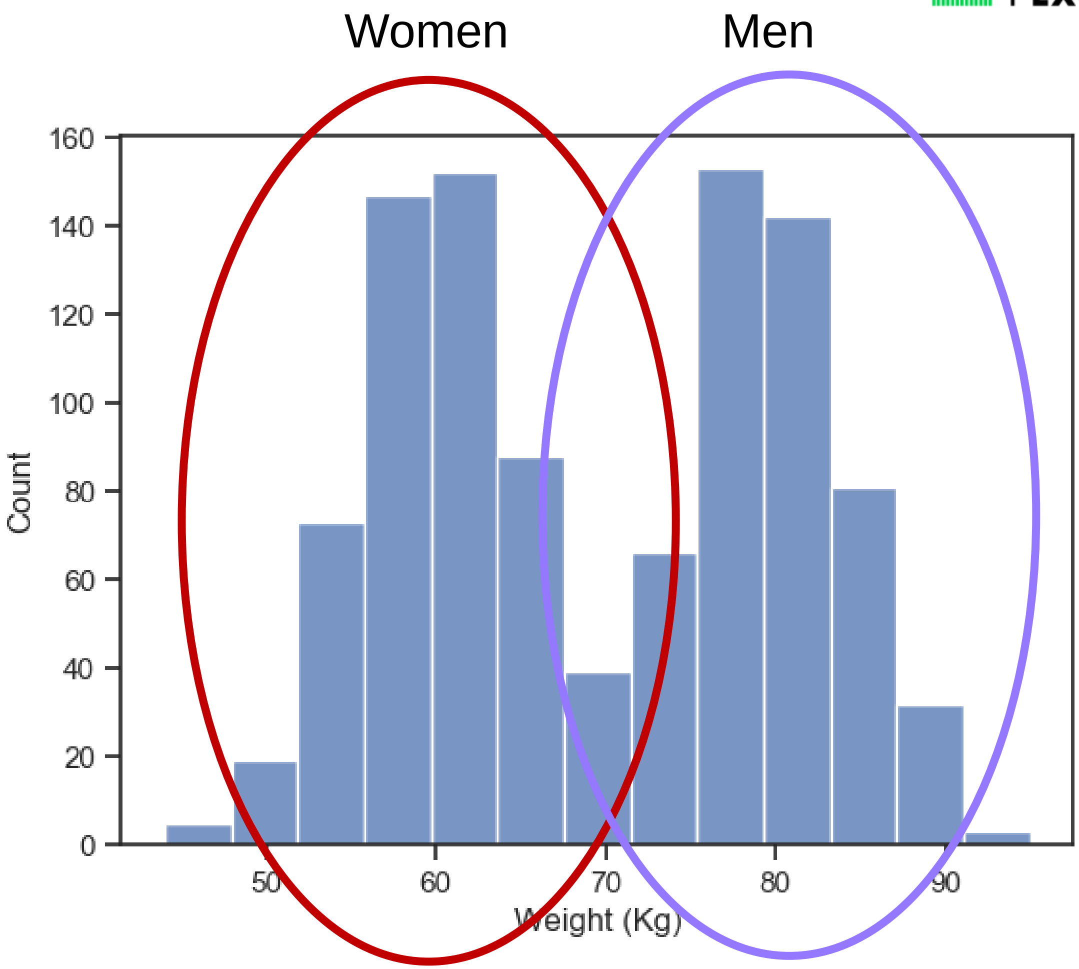
Patterns

- Bimodal distribution (two peaks).
  - Why ?



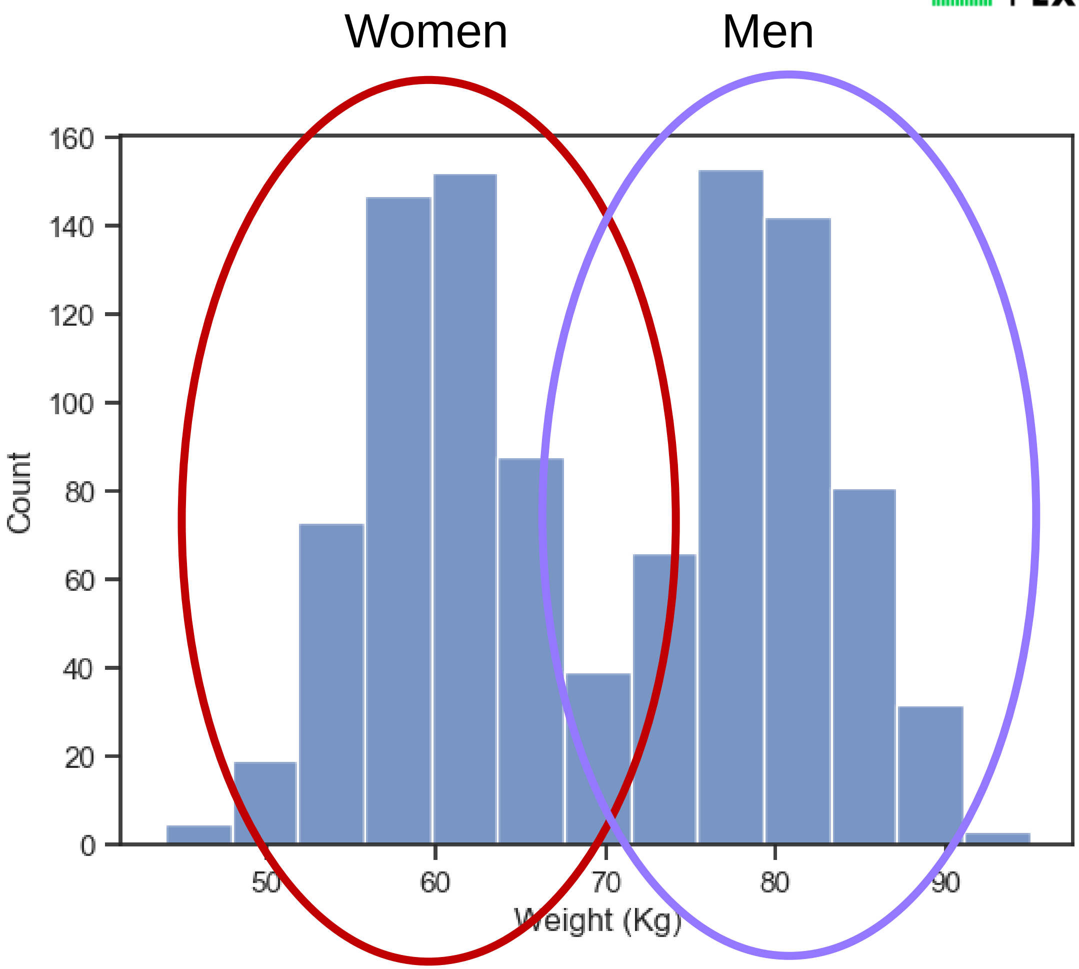
Patterns

- Bimodal distribution (two peaks)
  - Why ?
- It’s a Gaussian mixture:
  - Subpopulations of men and women



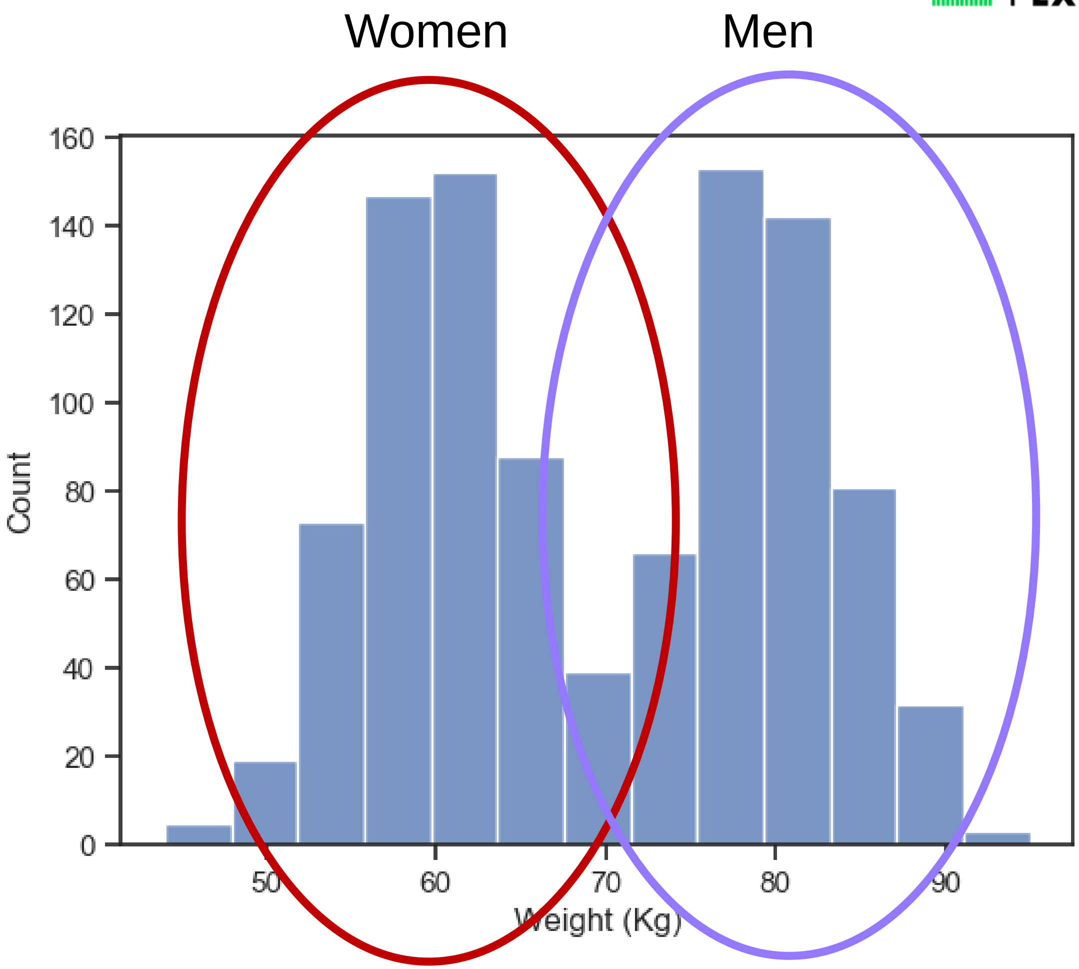
Patterns

- Bimodal distribution (two peaks)
  - Why ?
- It’s a Gaussian mixture:
  - Subpopulations of men and women
- Gender and weight are linked
- Consider a new person whose weight is unknown
  - What’s a straighforward predictor of his/her weight ?



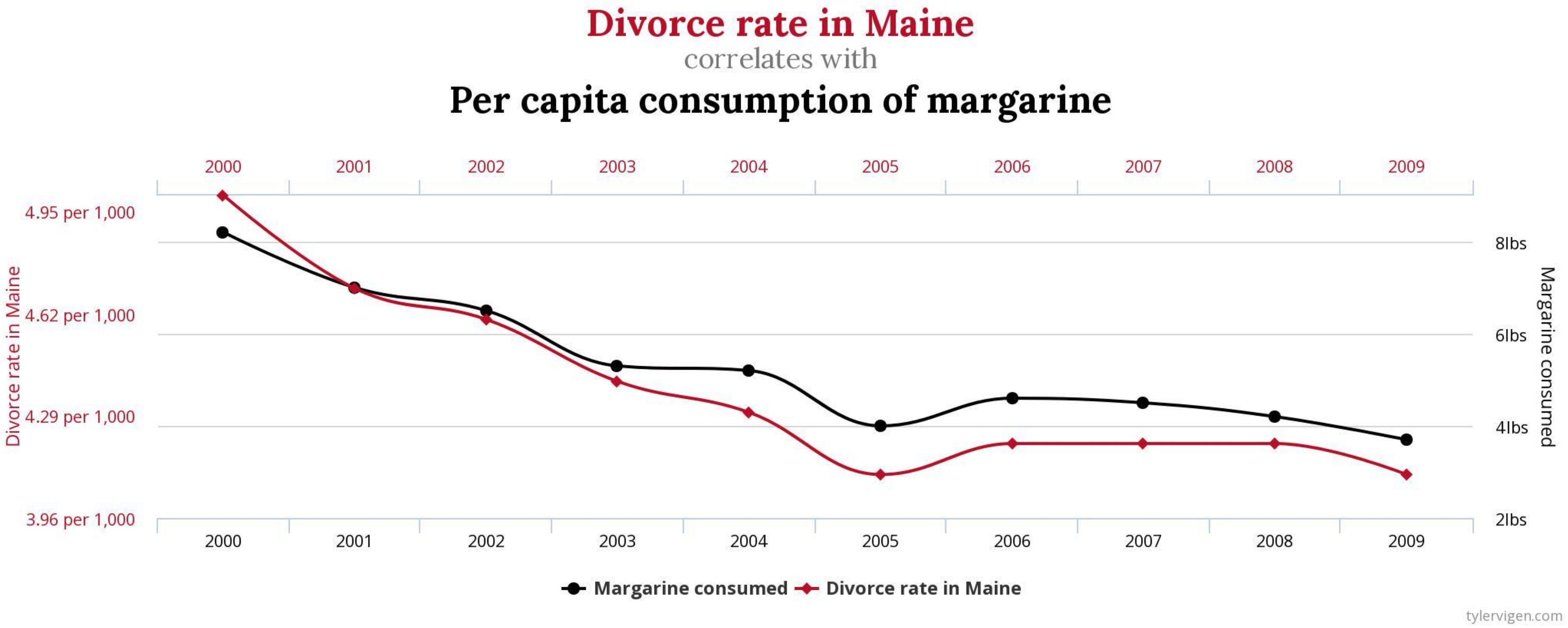
Patterns

- Bimodal distribution (two peaks)
  - Why ?
- It’s a Gaussian mixture:
  - Subpopulations of men and women
- Gender and weight are linked
- Consider a new person whose weight is unknown
  - What’s a straighforward predictor of his/her weight ?



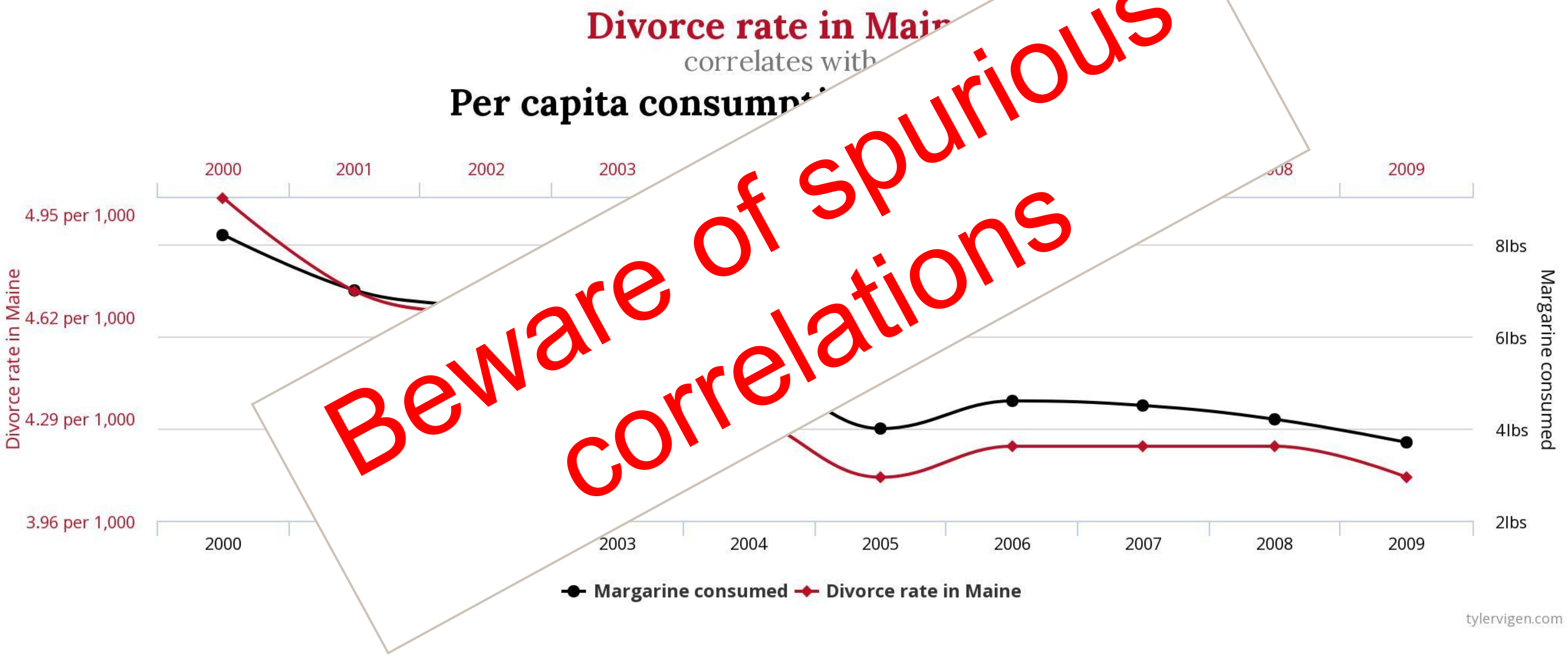
Inferential statistics allows you to make predictions (“inferences”) from that data by leveraging the underlying patterns in a sample and generalizing them to a larger population.

Patterns, but …



[Source: www.tylervigen.com]

Patterns, but …

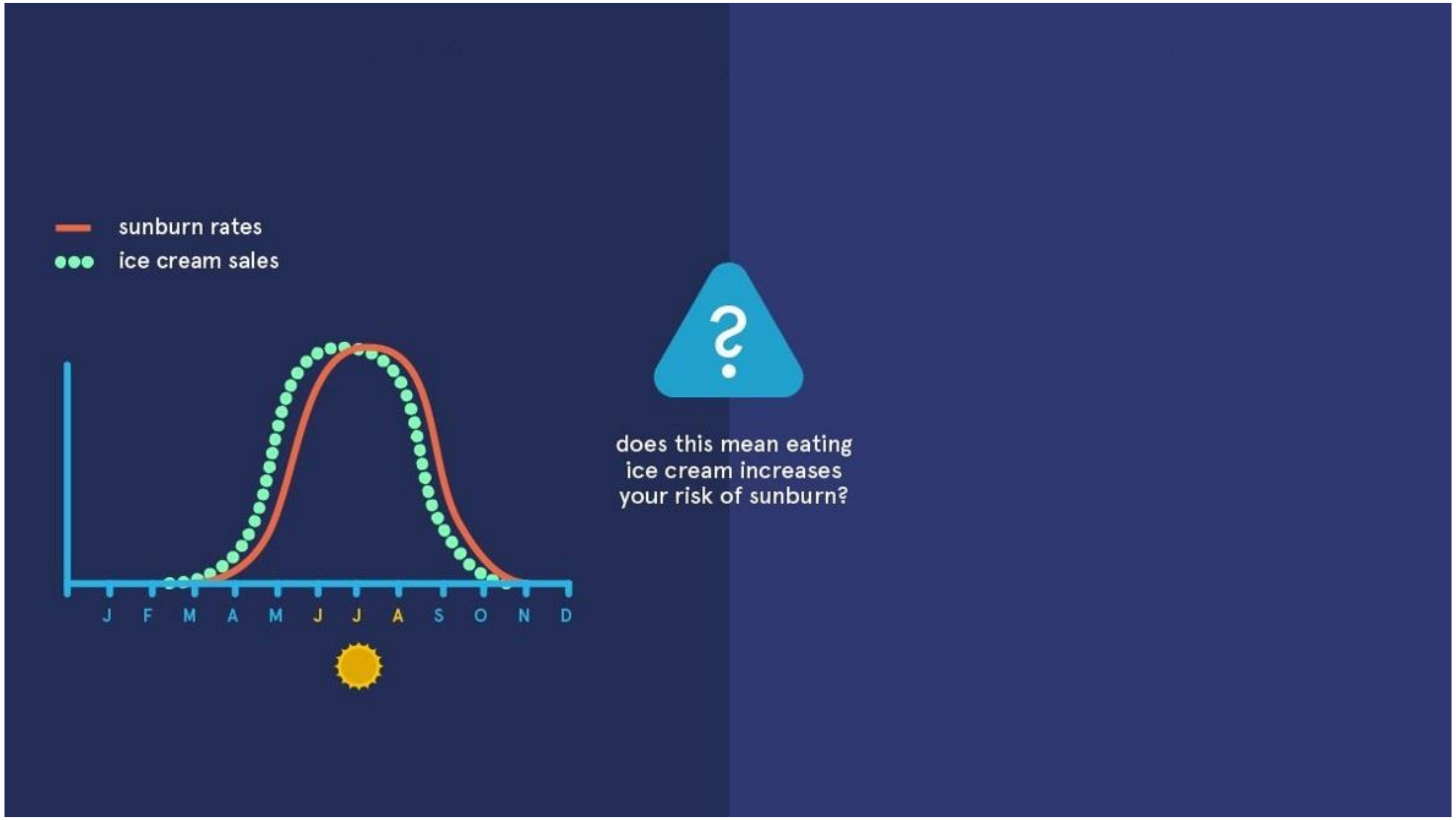


Beware of spurious correlations

[Source: www.tylervigen.com]

Patterns, but …

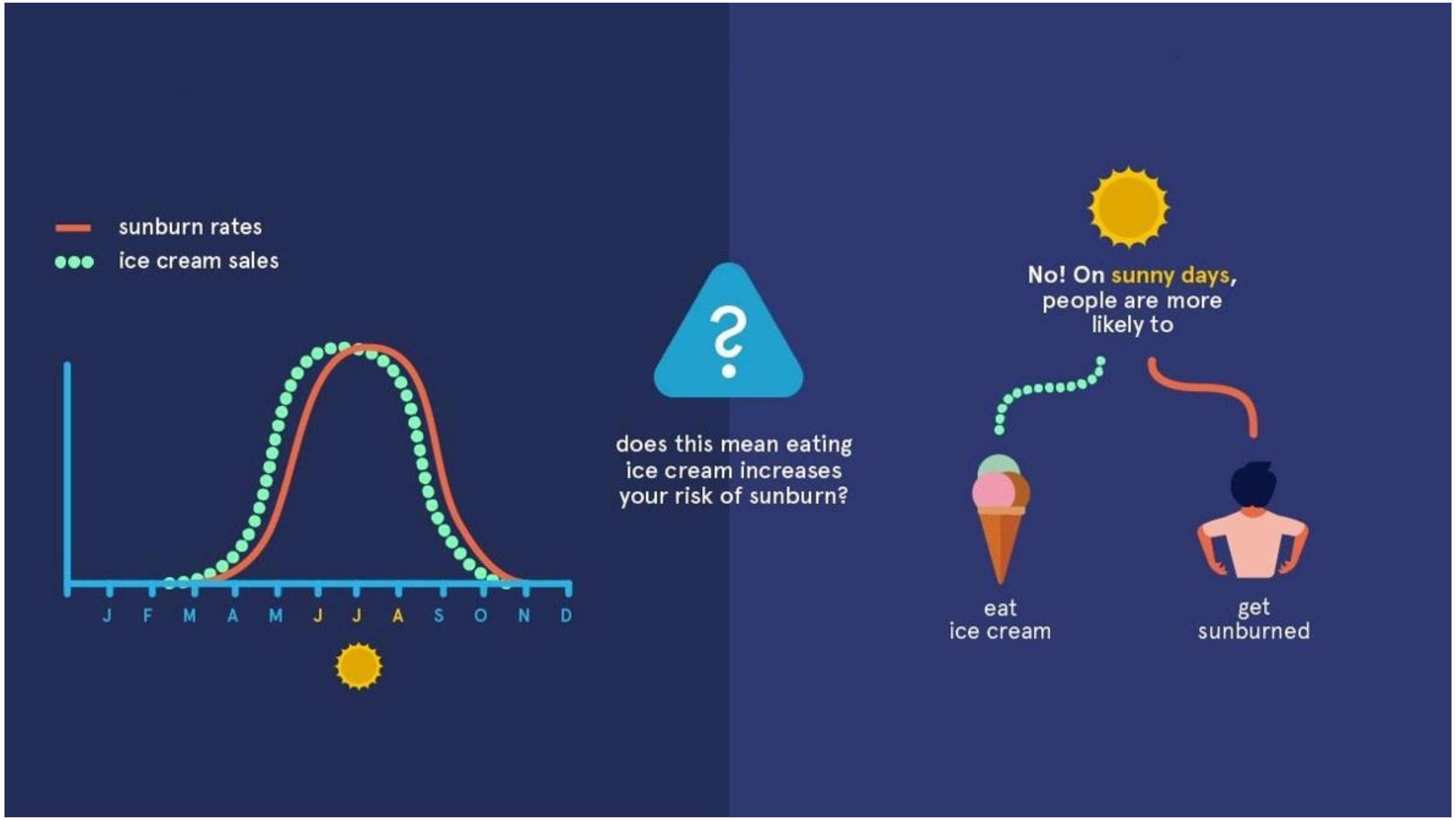
- You collect data on **sunburns** and **ice cream consumption**.
- You find that higher ice cream consumption is associated with a higher probability of sunburn. Does that mean ice cream consumption causes sunburn?



[Source: European Food Information Council]

Patterns, but ...

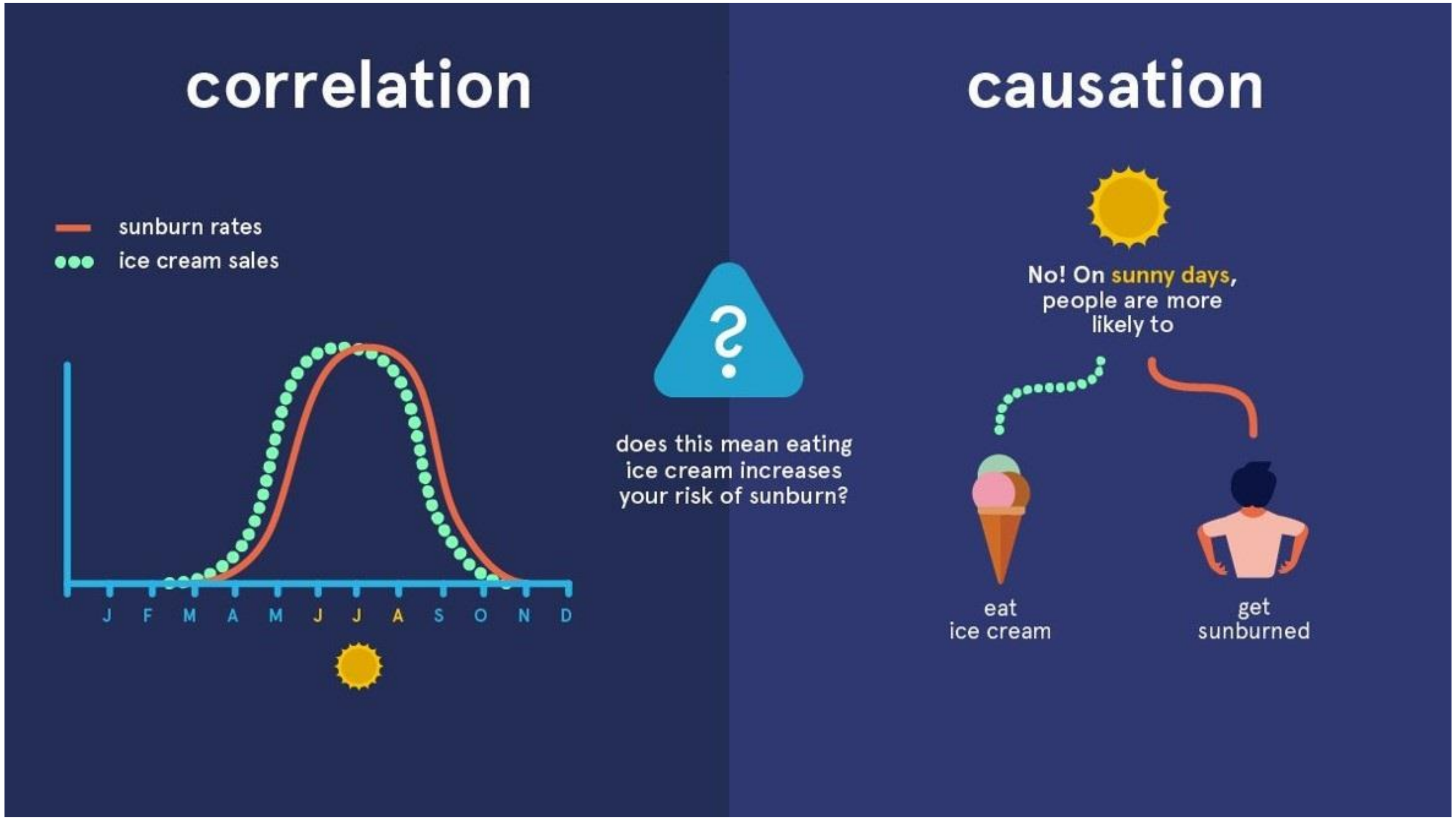
- You collect data on **sunburns** and **ice cream consumption**.
- You find that higher ice cream consumption is associated with a higher probability of sunburn. Does that mean ice cream consumption causes sunburn?


sunburn rates
ice cream sales
J F M A M J J A S O N D
does this mean eating ice cream increases your risk of sunburn?
No! On sunny days, people are more likely to
eat ice cream
get sunburned

[Source: European Food Information Council]

Patterns, but …

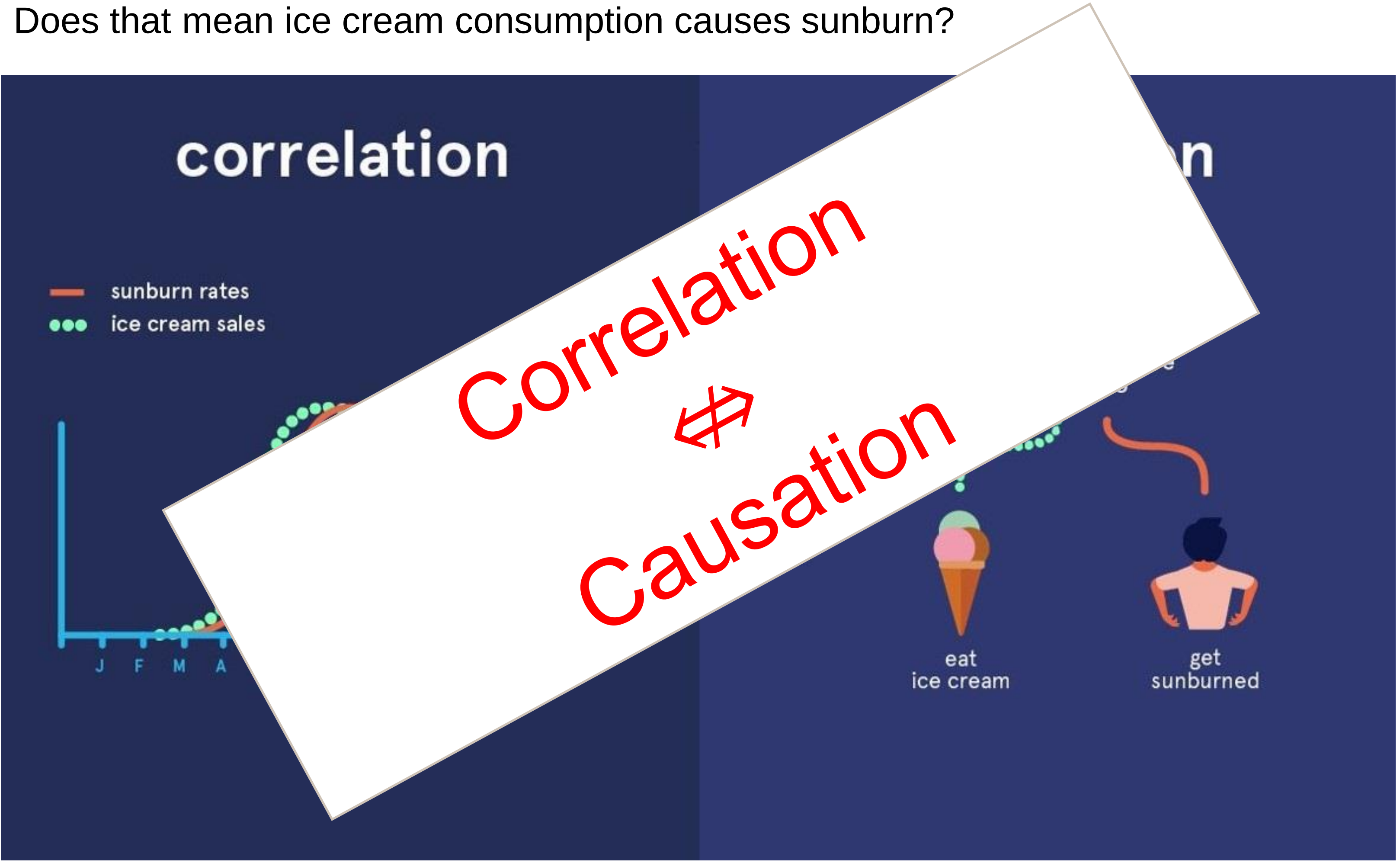
- You collect data on **sunburns** and **ice cream consumption**.
- You find that higher ice cream consumption is associated with a higher probability of sunburn. Does that mean ice cream consumption causes sunburn?


correlation
sunburn rates
ice cream sales
J F M A M J J A S O N D
does this mean eating ice cream increases your risk of sunburn?
causation
No! On sunny days, people are more likely to
eat ice cream
get sunburned

[Source: European Food Information Council]

Patterns, but …

- You collect data on **sunburns** and **ice cream consumption**.
- You find that higher ice cream consumption is associated with a higher probability of sunburn. Does that mean ice cream consumption causes sunburn?



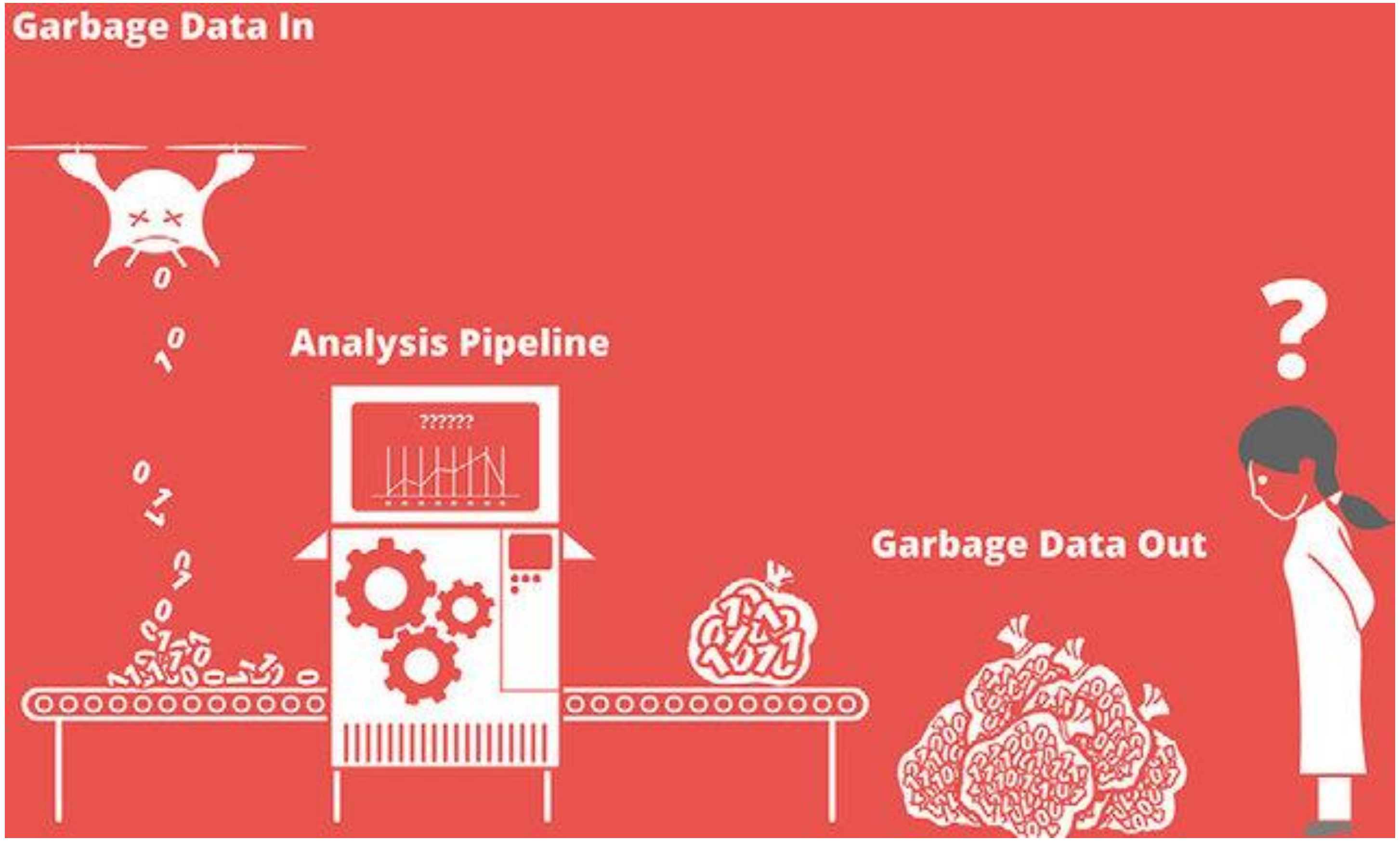
Correlation ⇎ Causation

[Source: European Food Information Council]

Data Quality

Examples:

- Incomplete data
- Inconsistent data
- Incorrect data



[Source: The Plant Phenome Journal]

Data Preprocessing

Preprocessing prepares the data for machine learning algorithms

Data cleaning

Missing values

Noisy samples

Outliers

…

Data transformation

Categorical data encoding

Feature scaling

Attribute selection

…

Data reduction

Dimension reduction

🎖 Time to practice ! 🎖