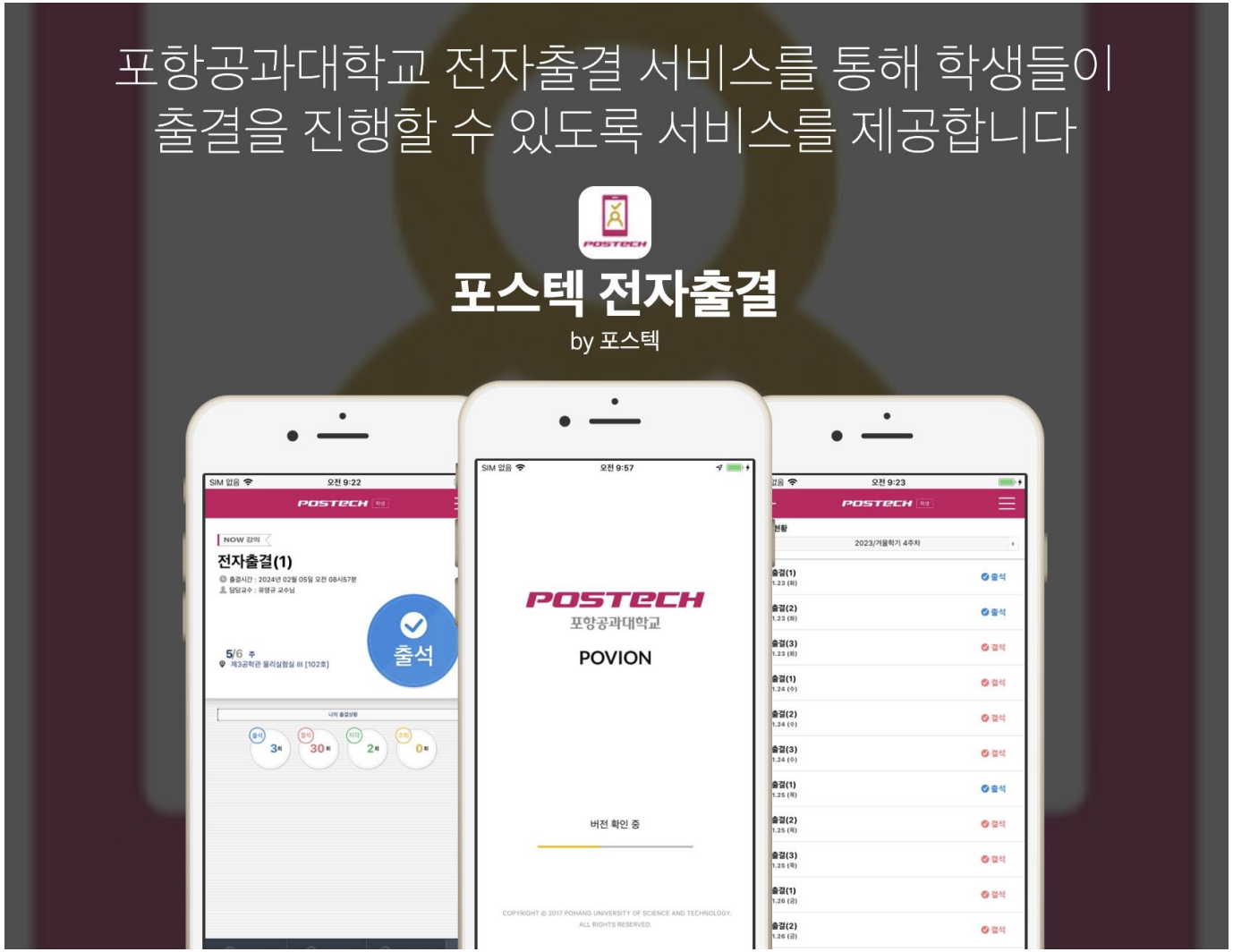
IMEN891K/IDSC501 Data Engineering

3. **Data Exploration**

Fall 2025

Instructor: **Sangwoo Mo**

Recap: Attendance

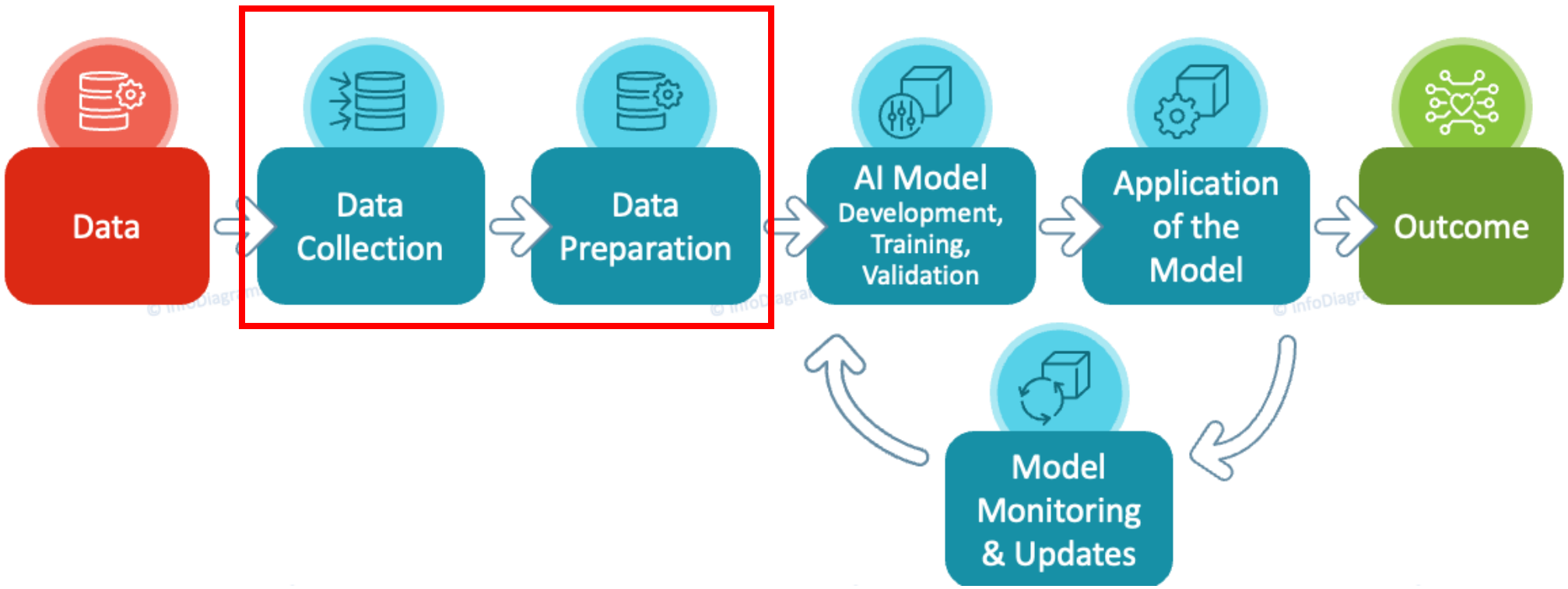


Recap: Team Project

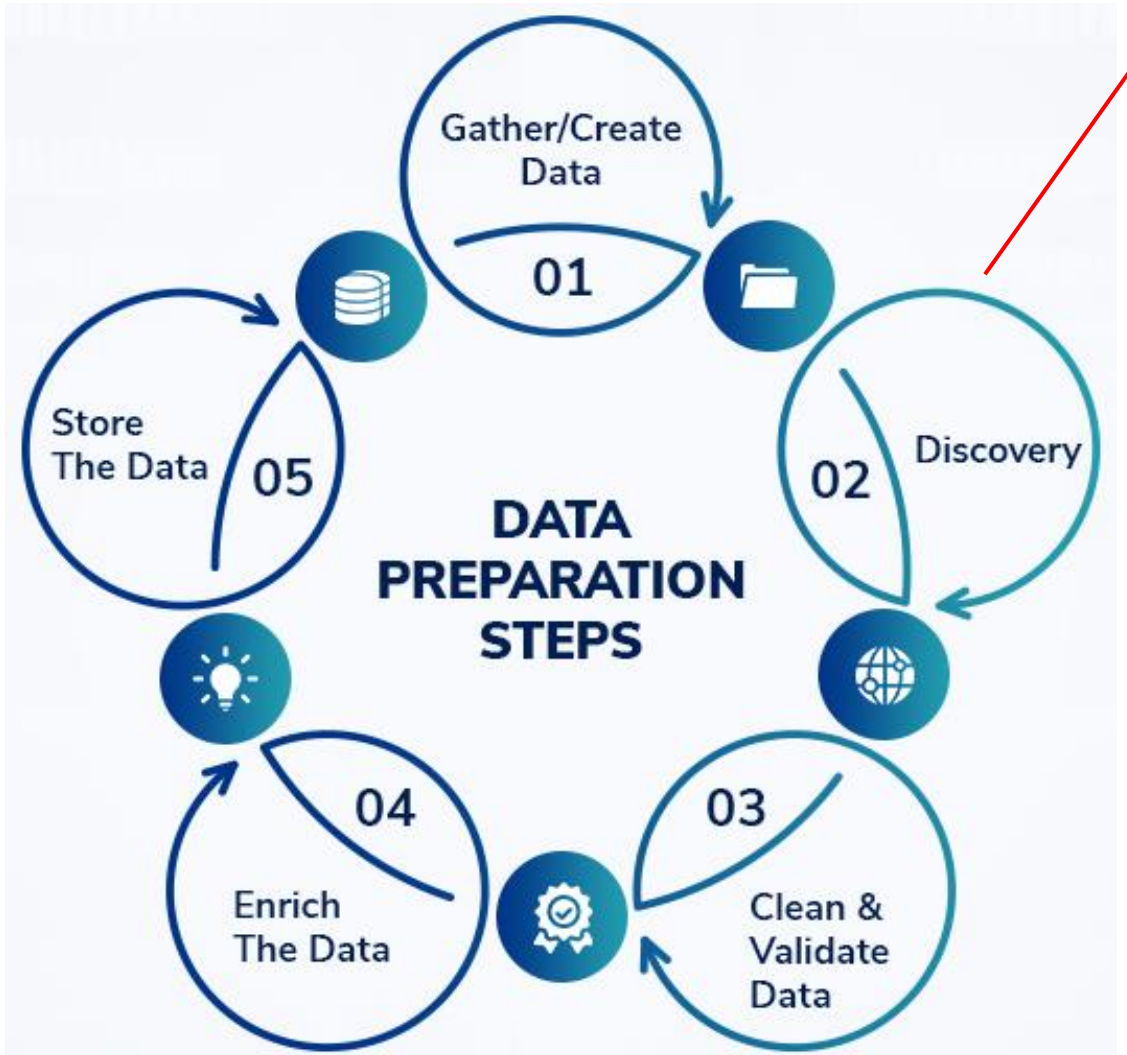
Build your team! **(Deadline: Sep 18)**



Recap: Course Overview

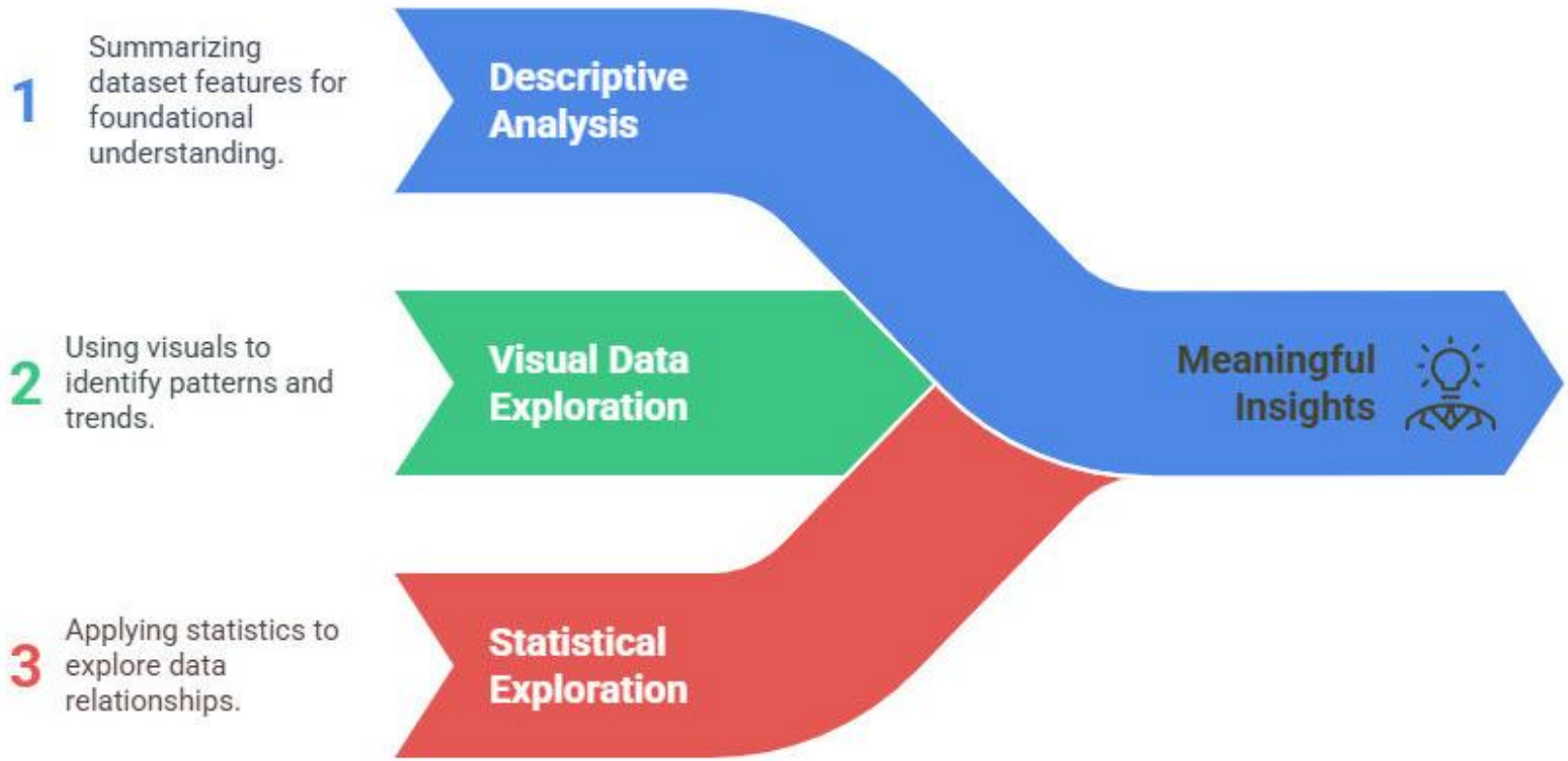


Recap: Course Overview

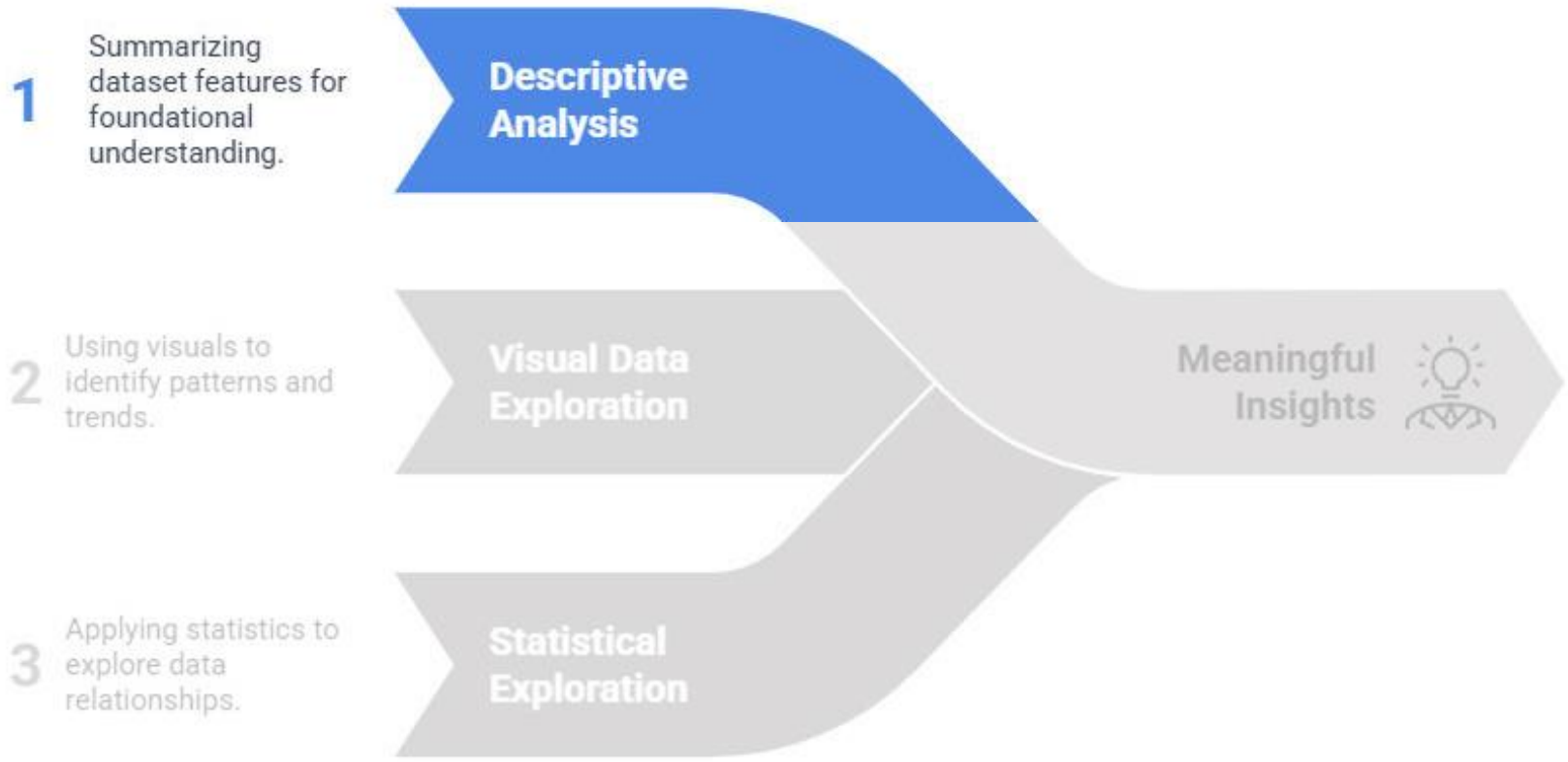


Lecture 3

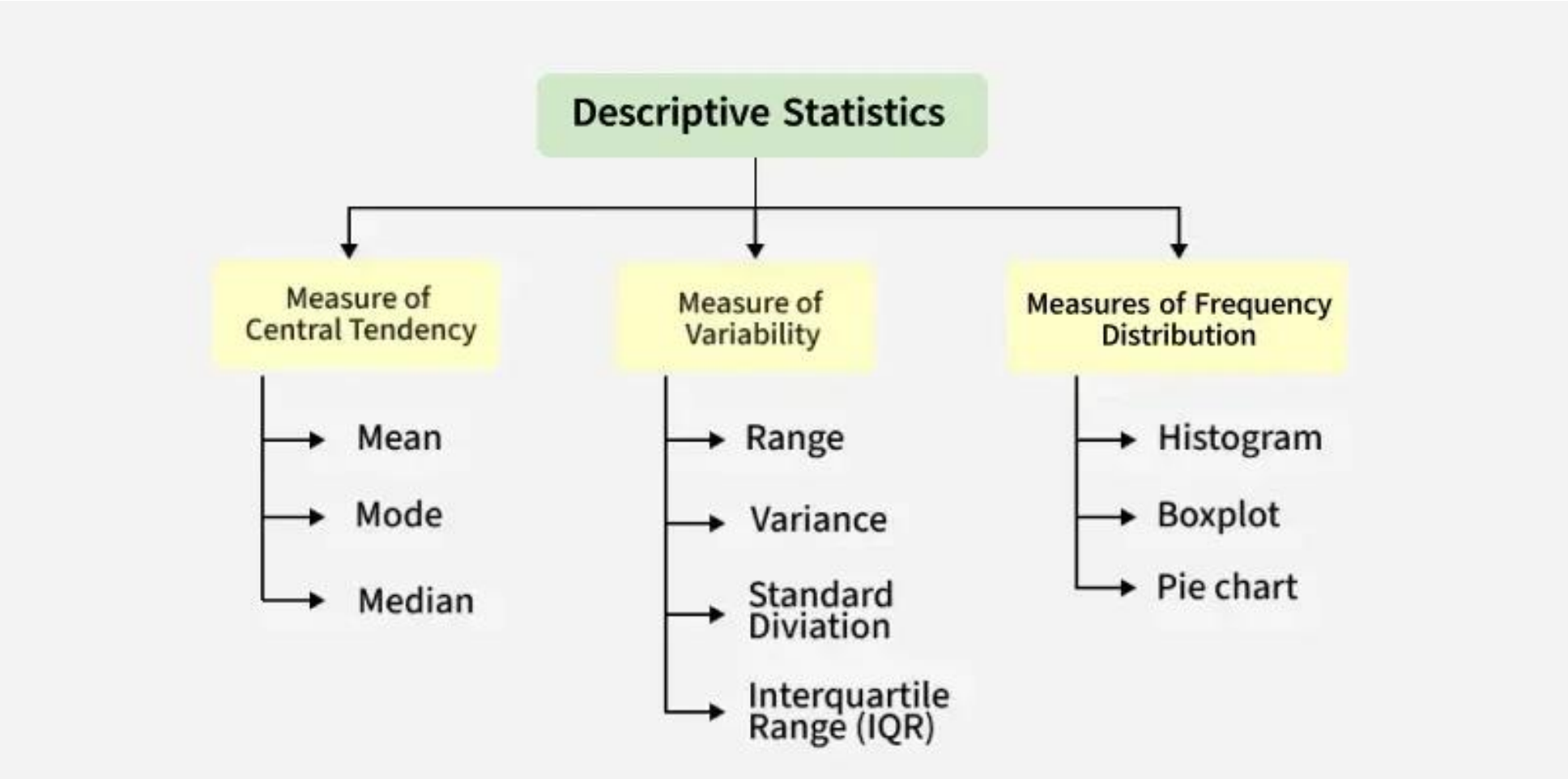
Outline: Data Exploration



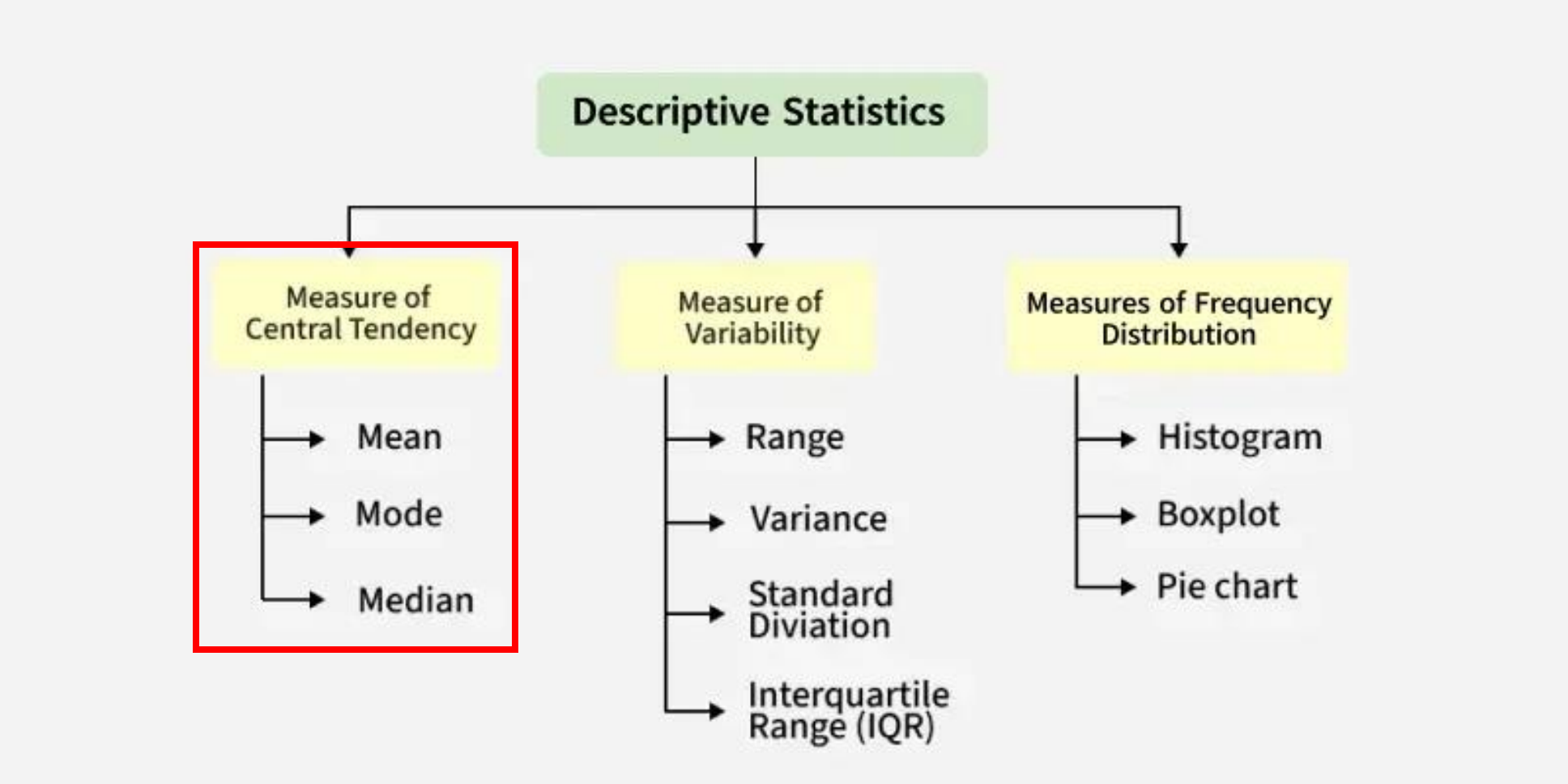
Outline: Data Exploration



Descriptive Statistics



Statistics - Central Tendency



Mean/Median/Mode

Mean

Add all the numbers then divide by the amount of numbers

9, 3, 1, 8, 3, 6

9 + 3 + 1 + 8 + 3 + 6 = 30

30 ÷ 6 = 5

The mean is 5

Median

Order the set of numbers, the median is the middle number

9, 3, 1, 8, 3, 6

1, 3, 3, 6, 8, 9

The median is 4.5

Mode

The most common number

9, 3, 1, 8, 3, 6

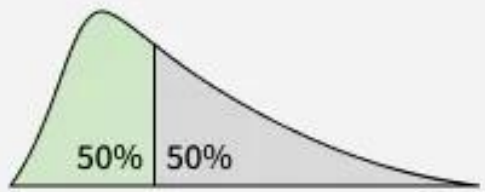
The mode is 3

Mean/Median/Mode

MODE Vs **MEDIAN** Vs **MEAN**



Mode is the most frequent value.

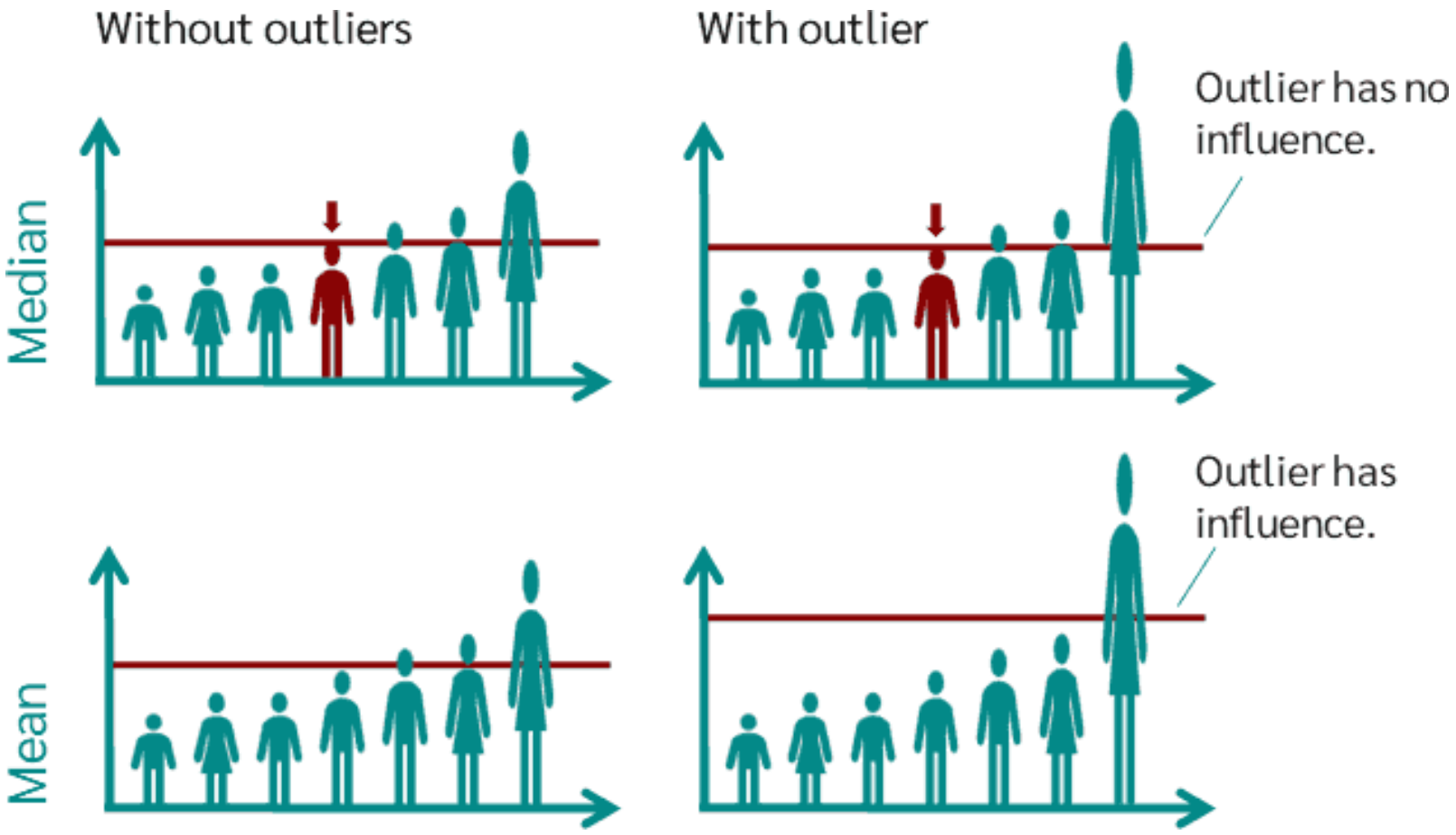


Meadian is the middle value.



Mean is the average or norm.

Robust Statistics - Mean vs. Median

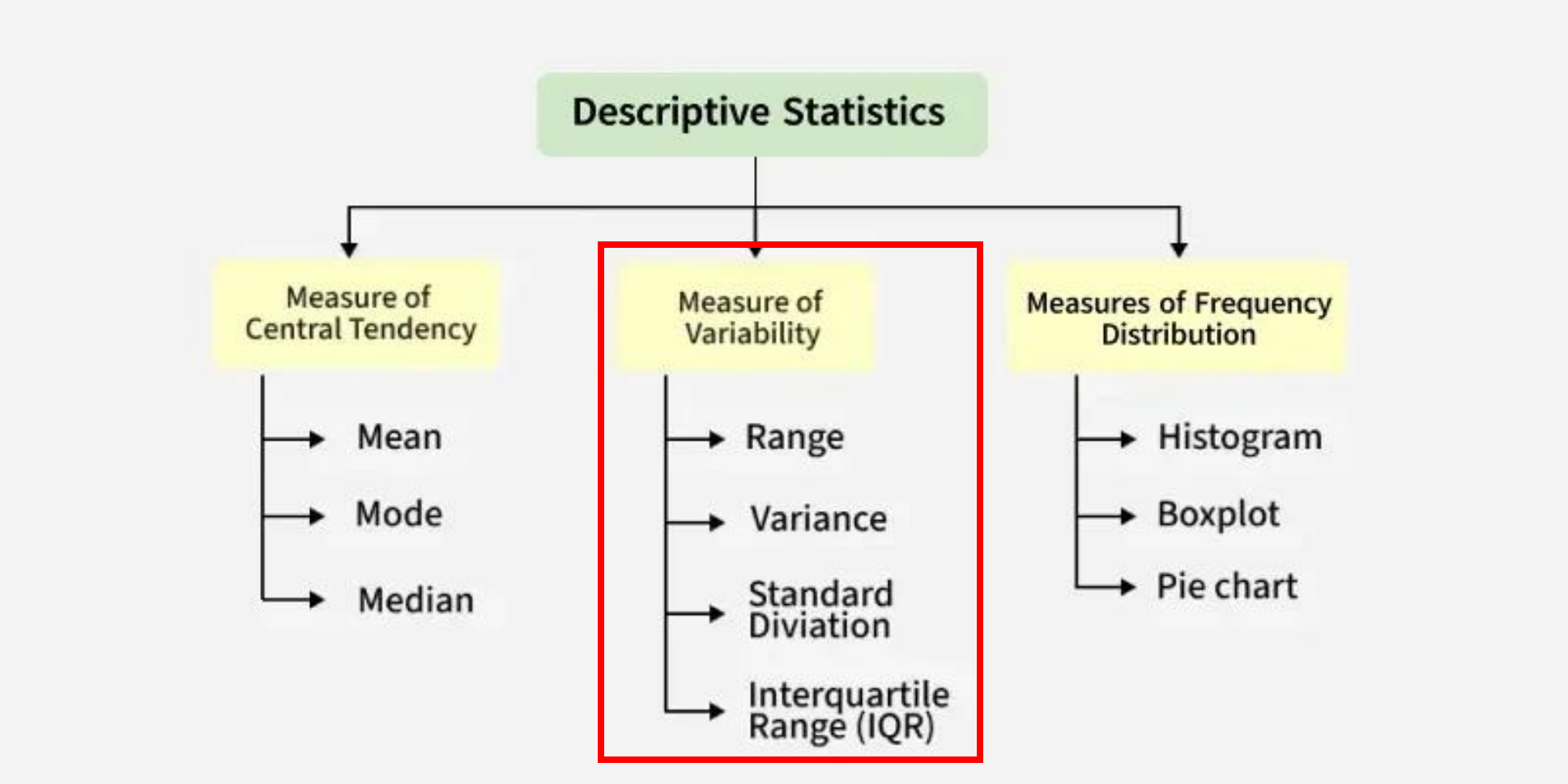


Robust Statistics - Trimmed Mean

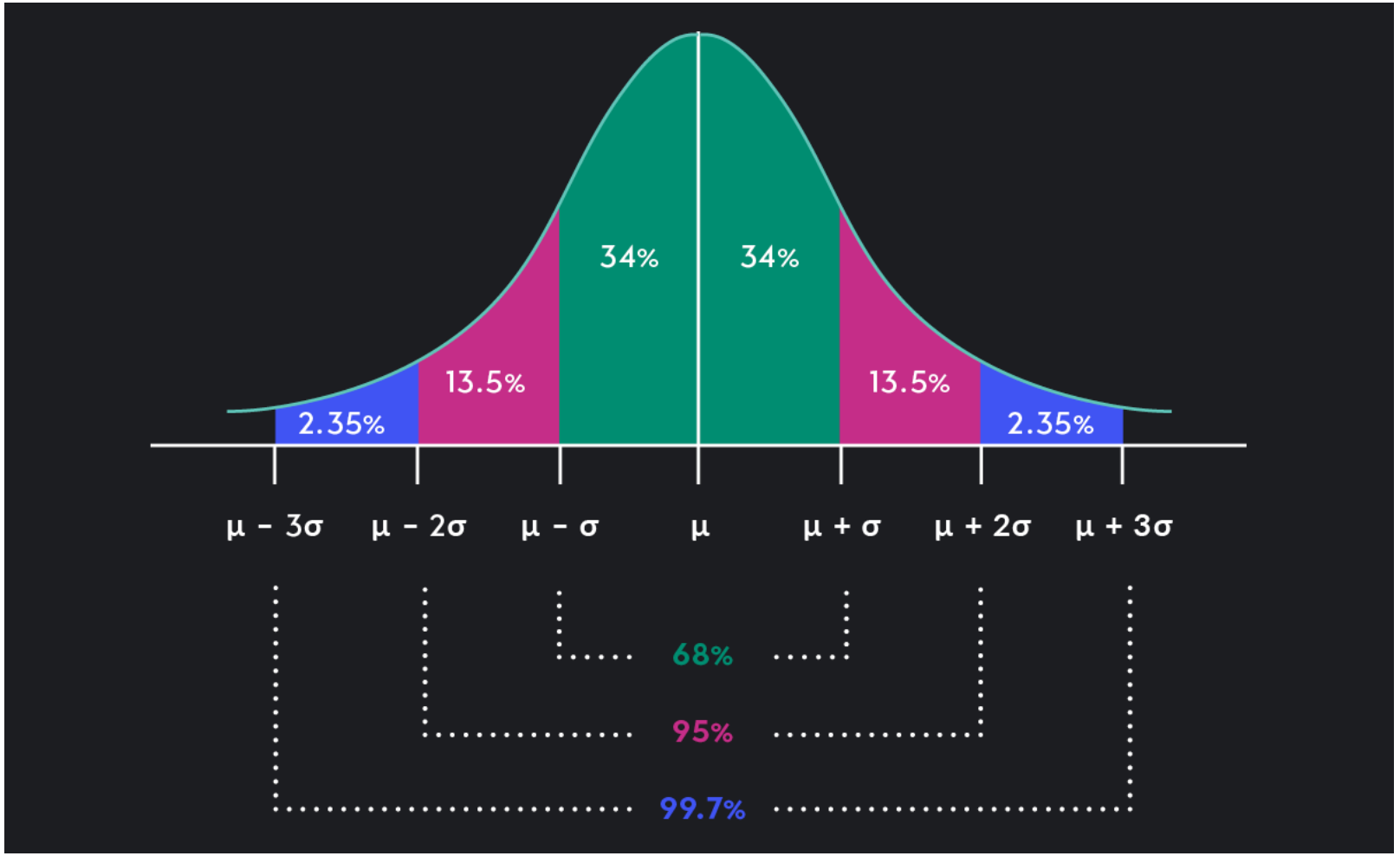
"Trimmed mean is a central tendency measure that cuts down the smallest and highest values before applying the standard averaging formula for greater accuracy."



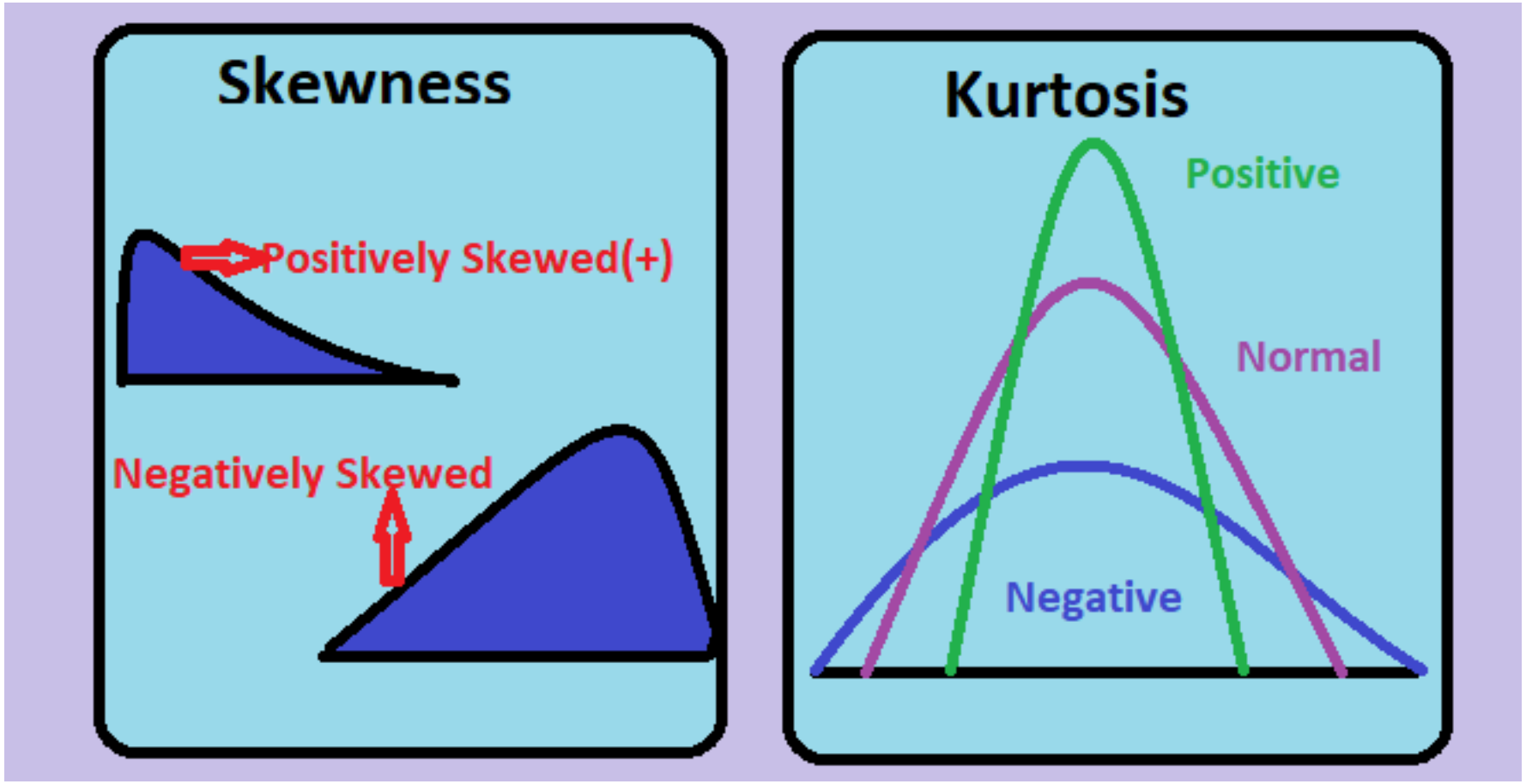
Statistics - Variability



Standard Deviation



Skewness / Kurtosis

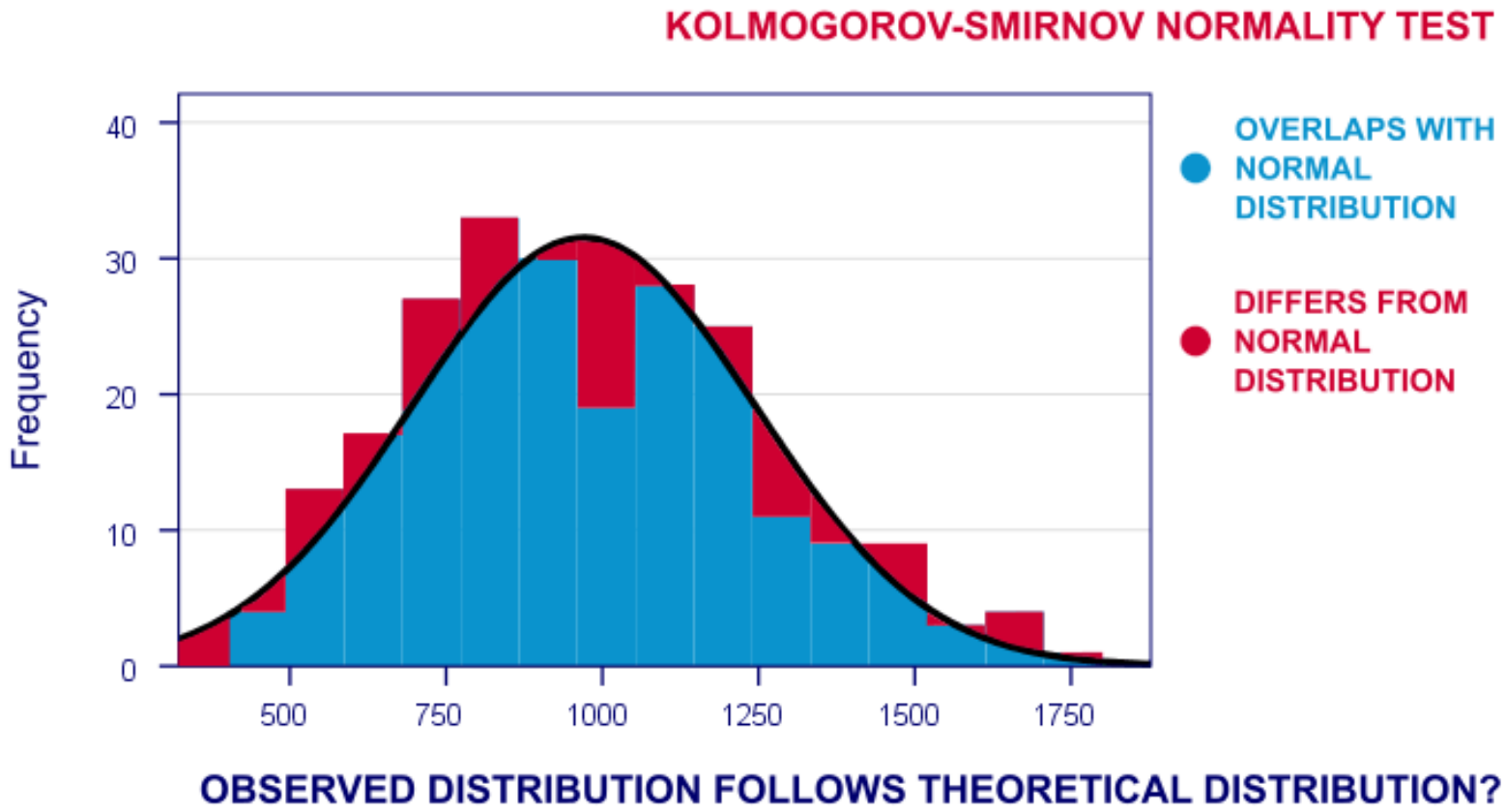


Moments

Moment number	Name	Measure of	Formula
1	Mean	Central tendency	$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N}$
2	Variance (Volatility)	Dispersion	$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N}$
3	Skewness	Symmetry (Positive or Negative)	$Skew = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{(X_i - \bar{X})}{\sigma} \right]^3$
4	Kurtosis	Shape (Tall or flat)	$Kurt = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{(X_i - \bar{X})}{\sigma} \right]^4$

Where X is a random variable having N observations (i = 1,2,…,N).

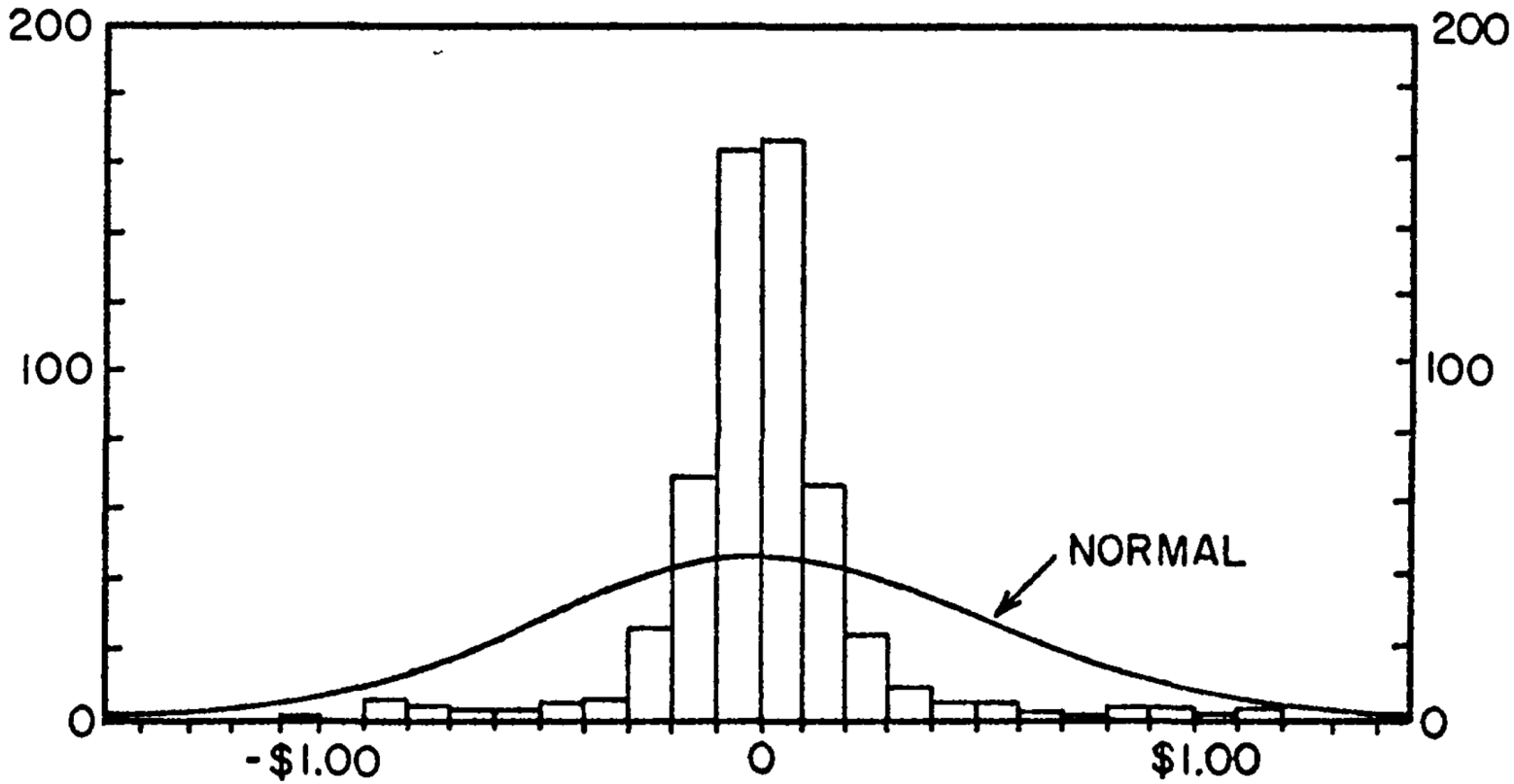
Normality Test



Case Study: Heavy-tailed Distribution

Case Study: Heavy-tailed Distribution

Price changes are more “peaked” than Gaussian



Mandelbrot. The Variation of Certain Speculative Prices. The Journal of Business, 1963.

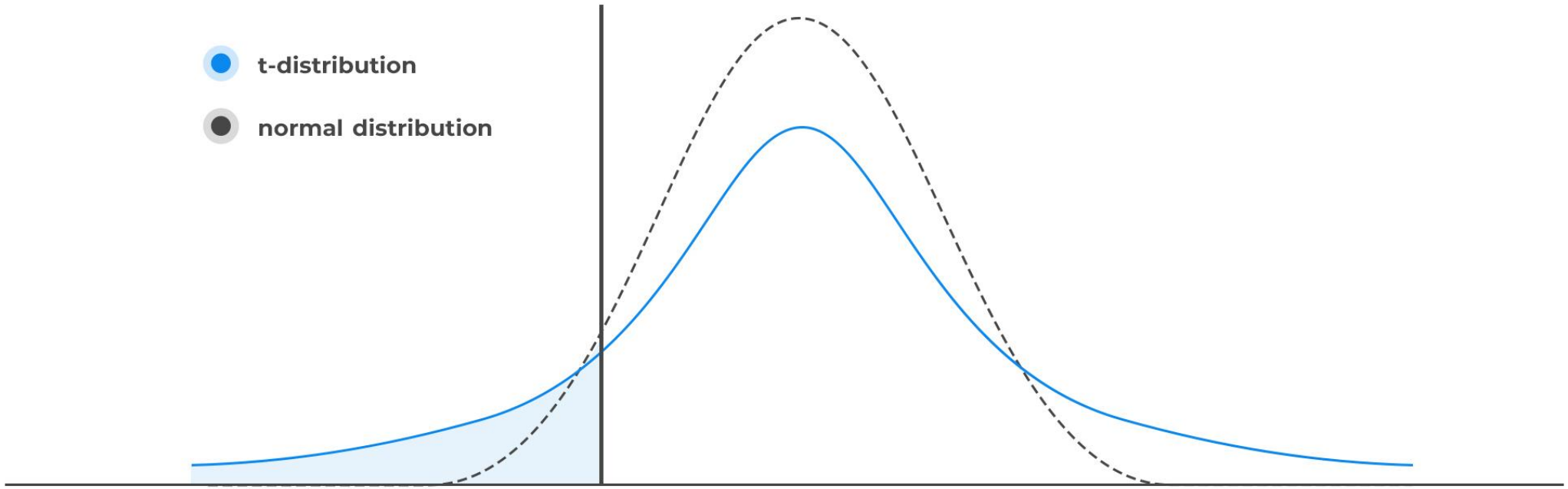
Case Study: Heavy-tailed Distribution

Wrong assumptions triggered a catastrophic crisis

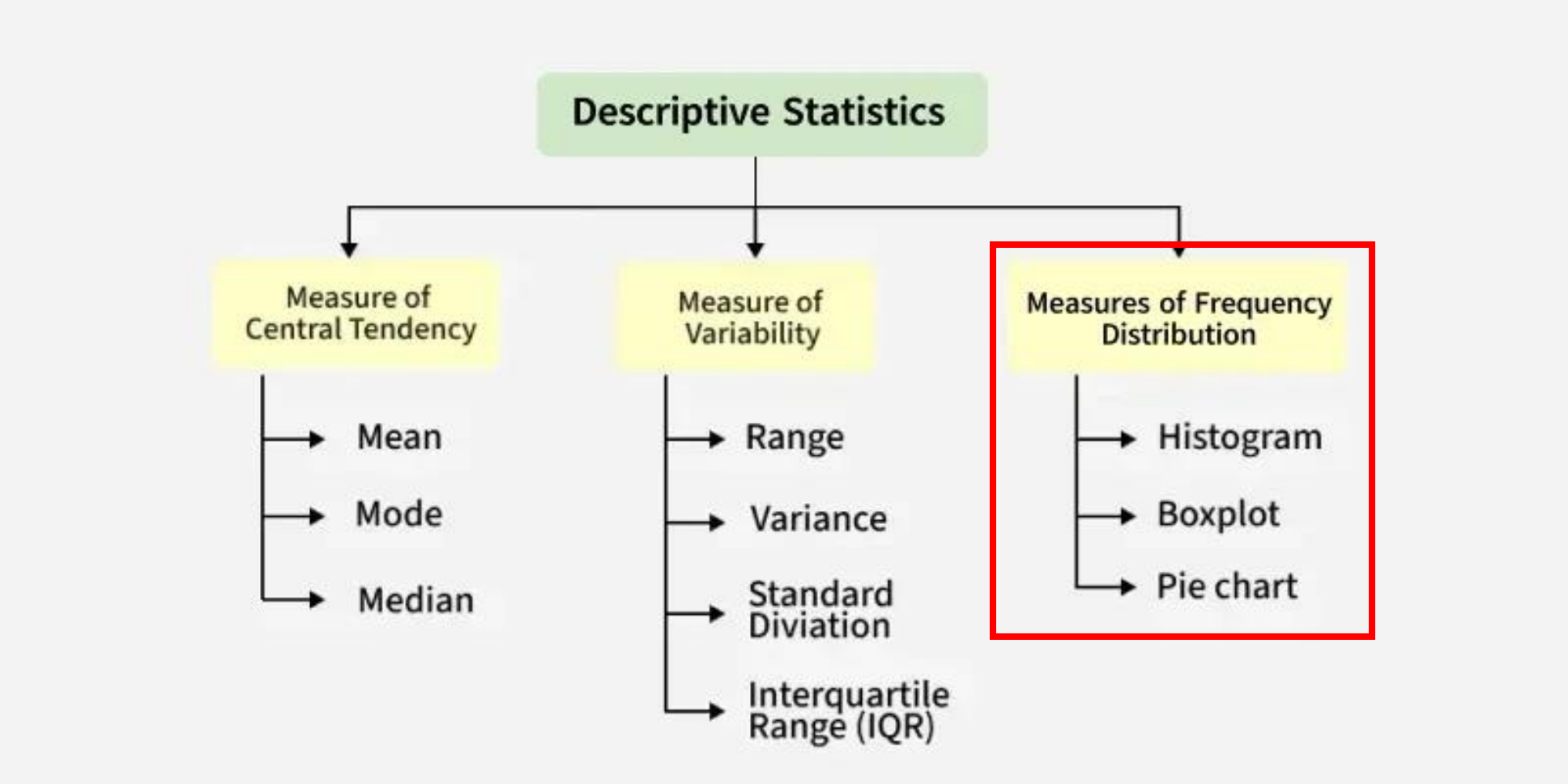


Case Study: Heavy-tailed Distribution

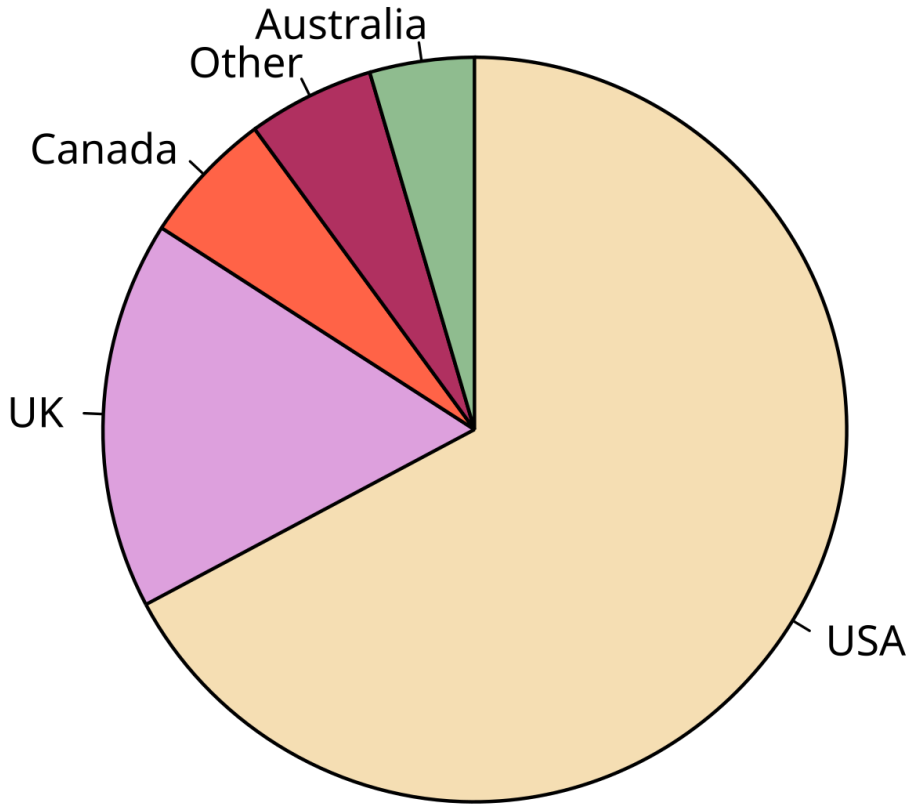
Necessity of advanced algorithms



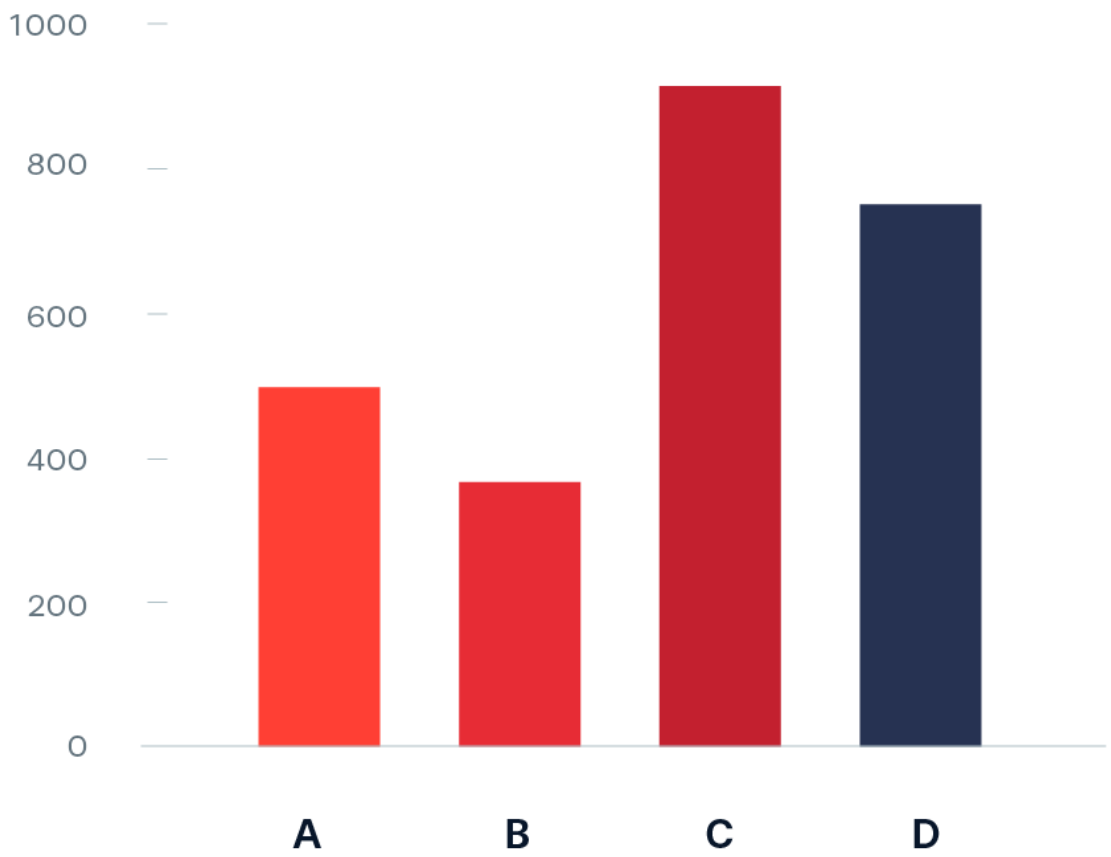
Statistics - Frequency



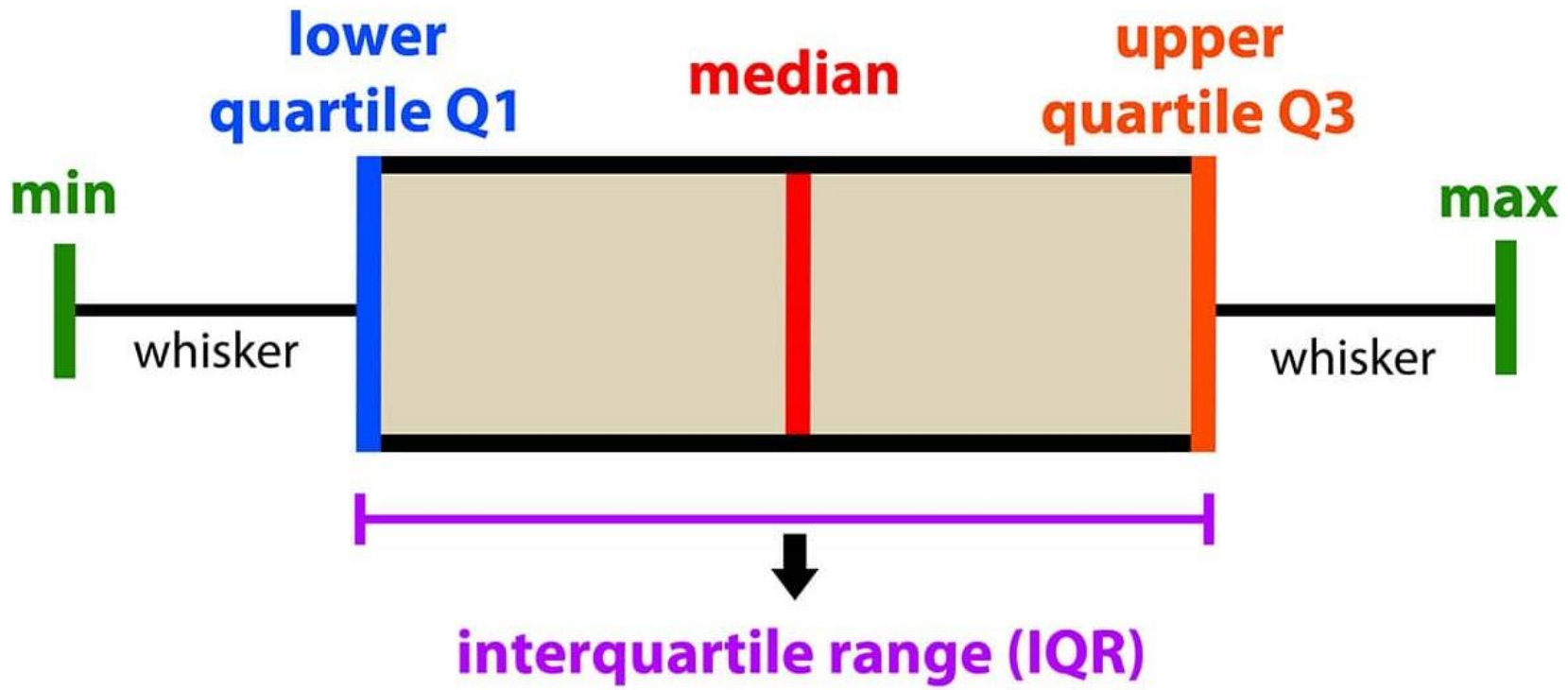
Pie Chart



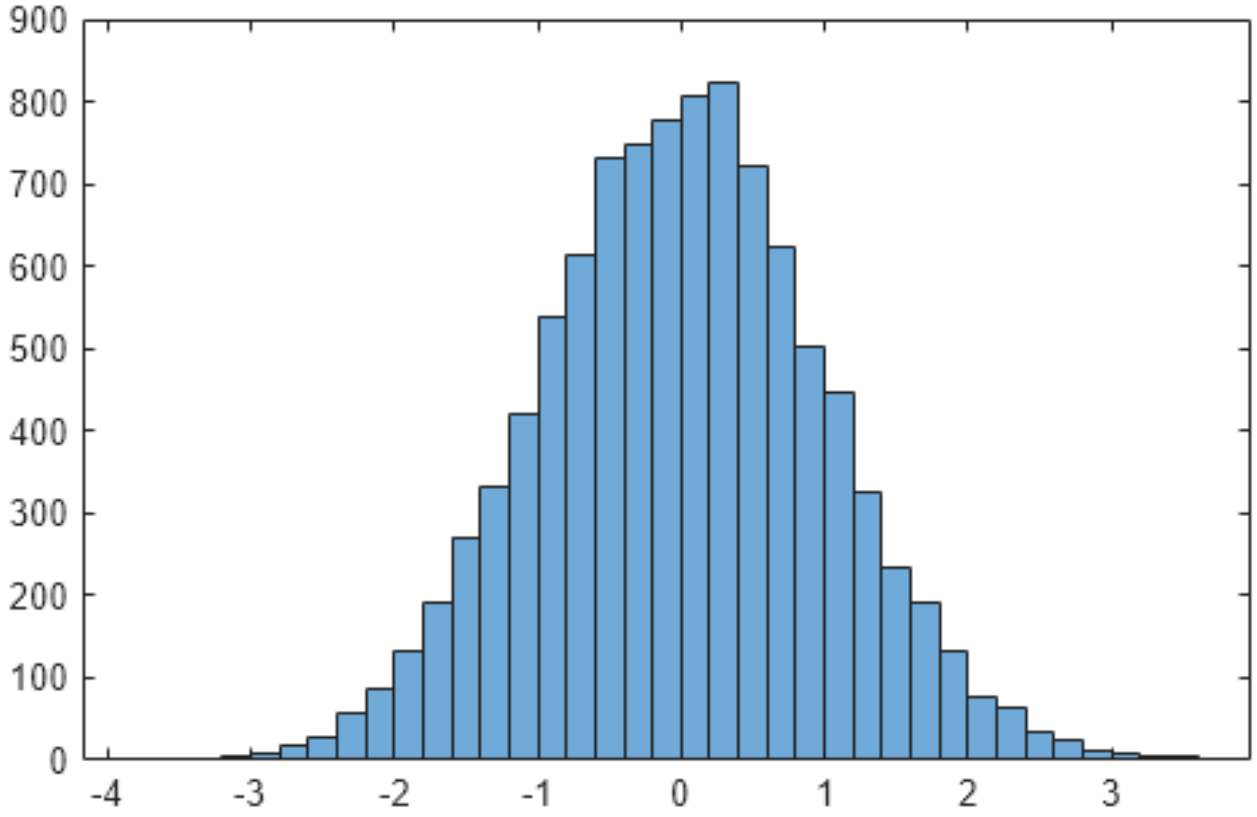
Bar Chart



Box Plot

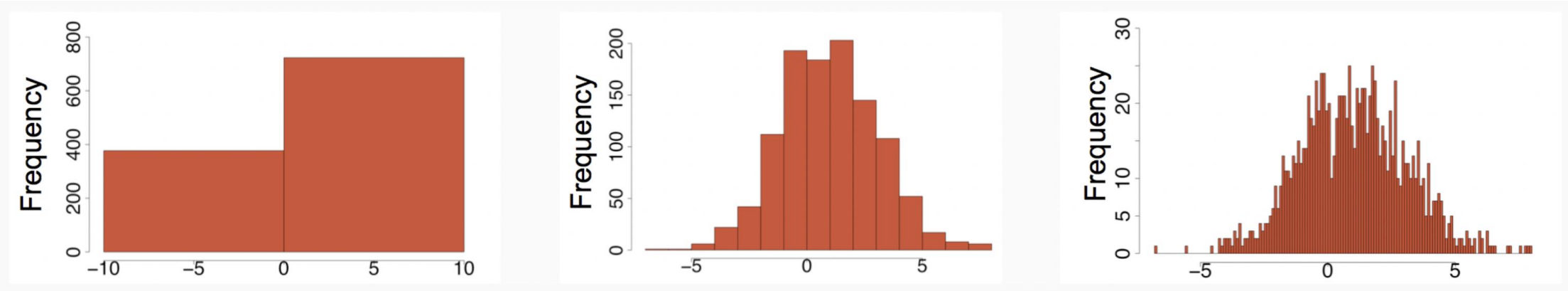

lower
quartile Q1
median
upper
quartile Q3
min
max
whisker
whisker
interquartile range (IQR)

Histogram (Empirical distribution, in a fancier word)



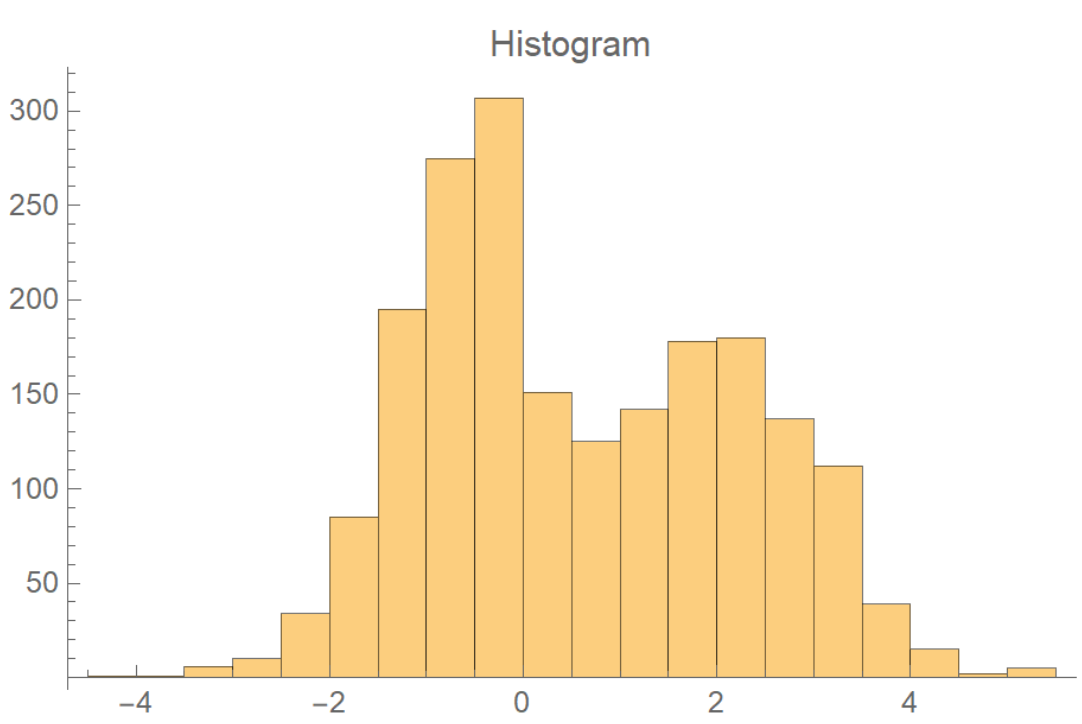
Histogram (Empirical distribution, in a fancier word)

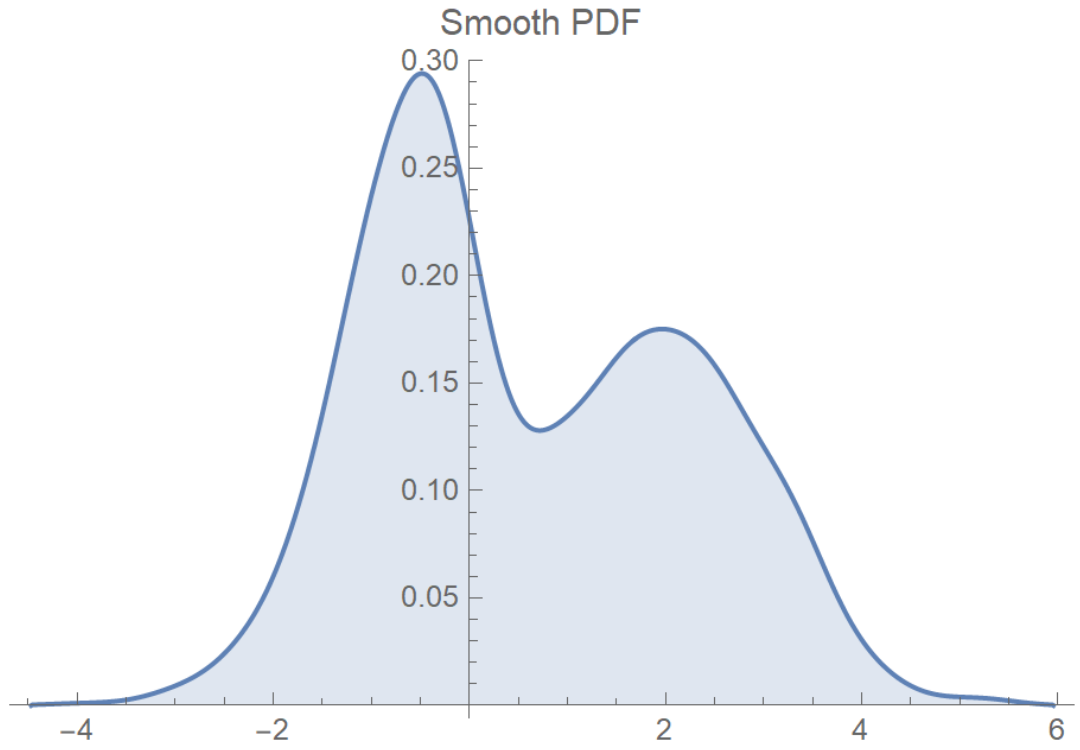
...is sensitive to # of bins



Density Estimation

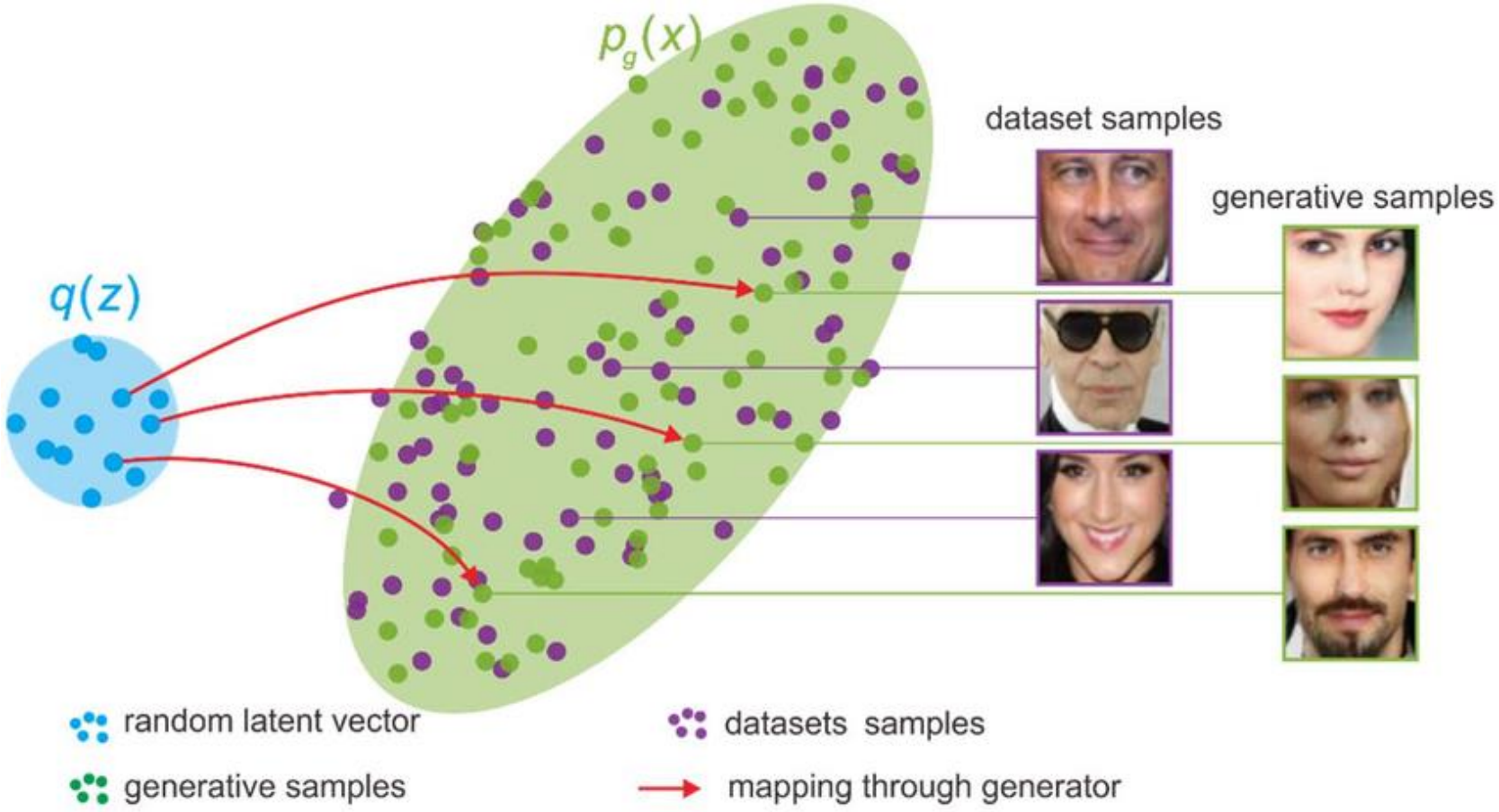
= “Smoothed” empirical distribution



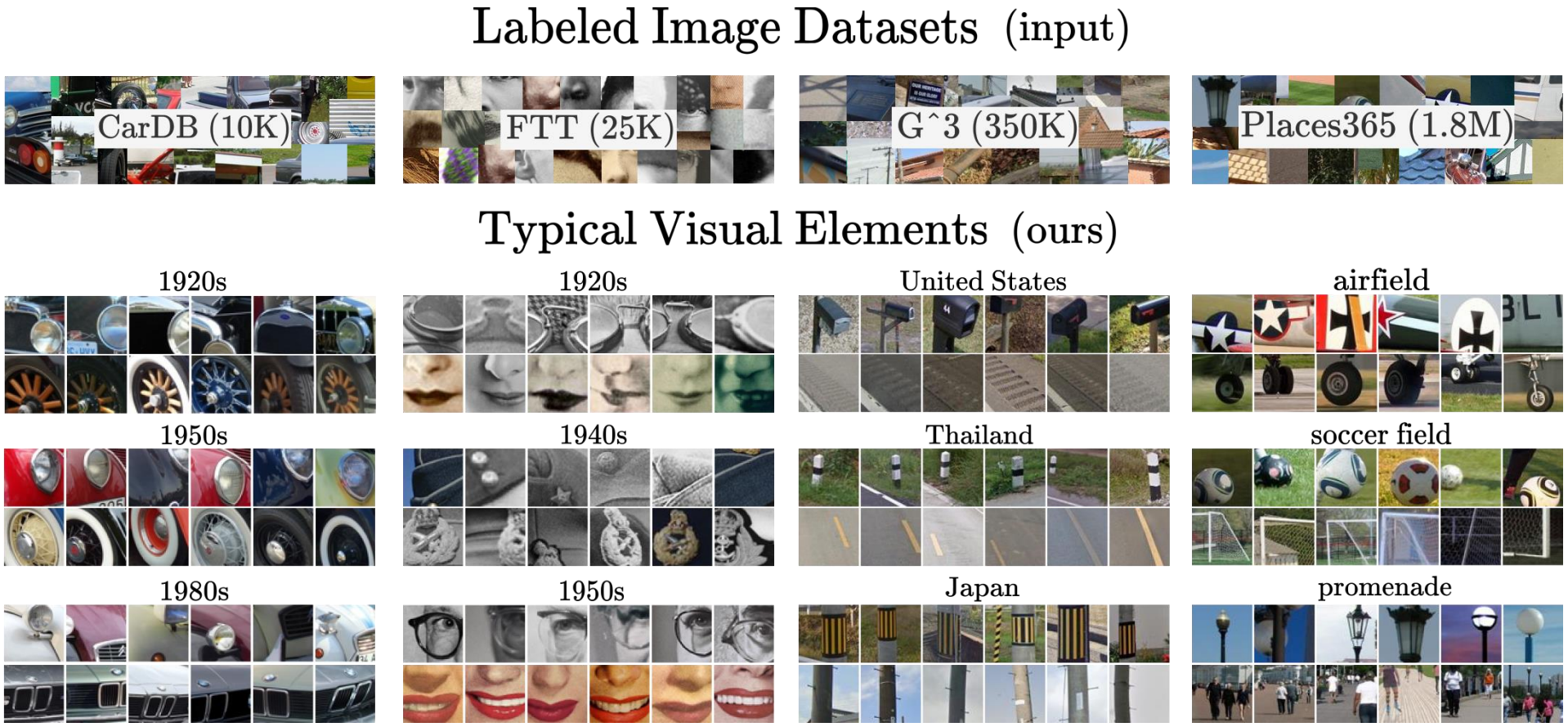


Generative Models

= “Deep” density estimation

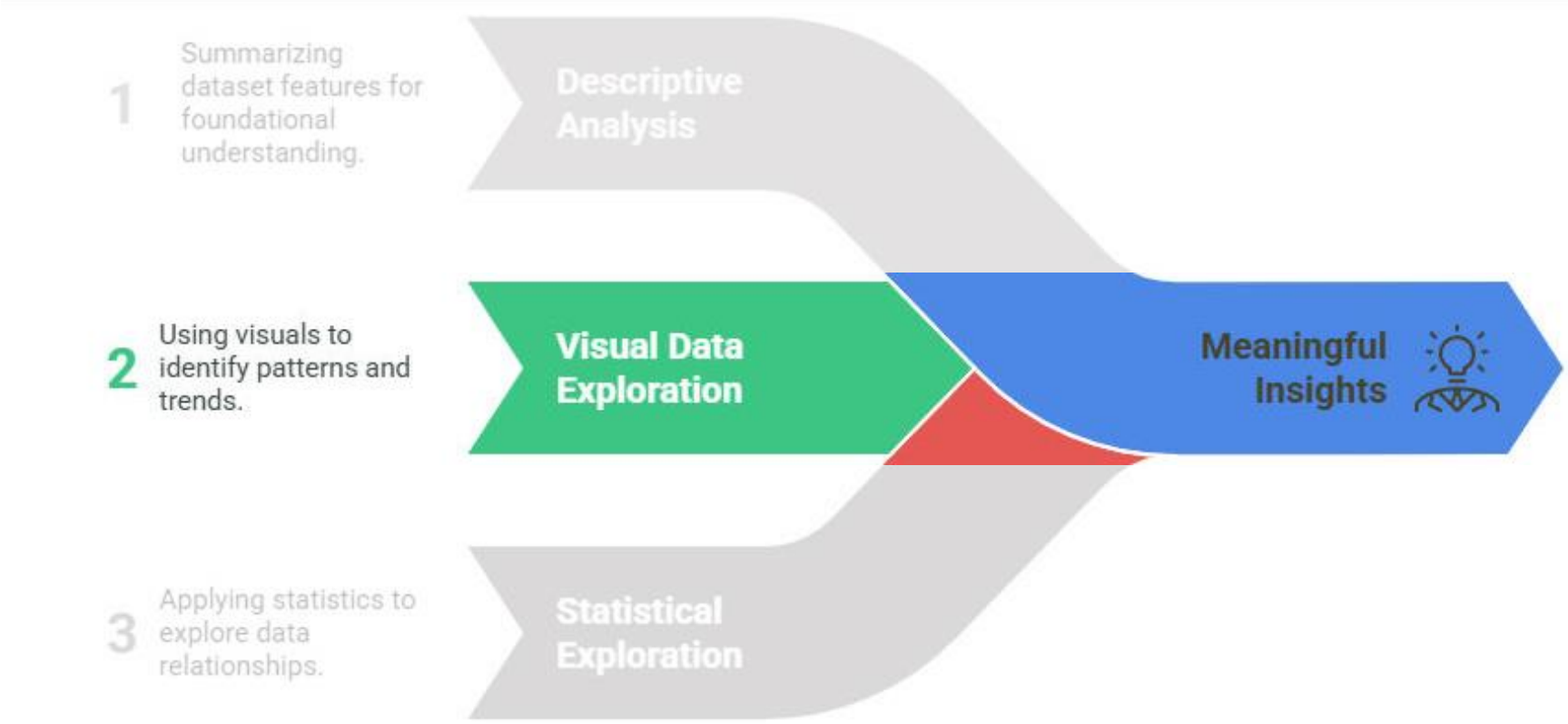


Case Study: GenAI as Data Explorers

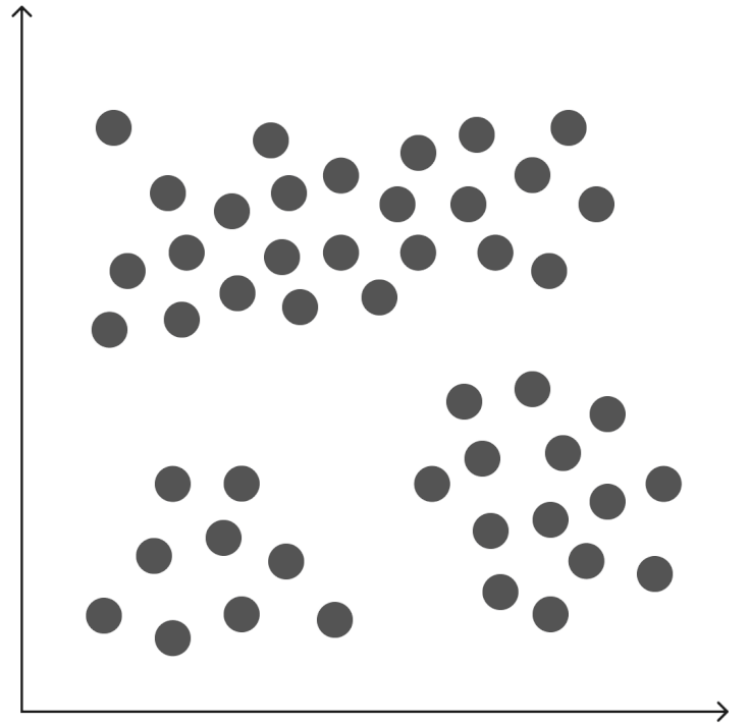


Diffusion Models as Data Mining Tools. ECCV, 2024.

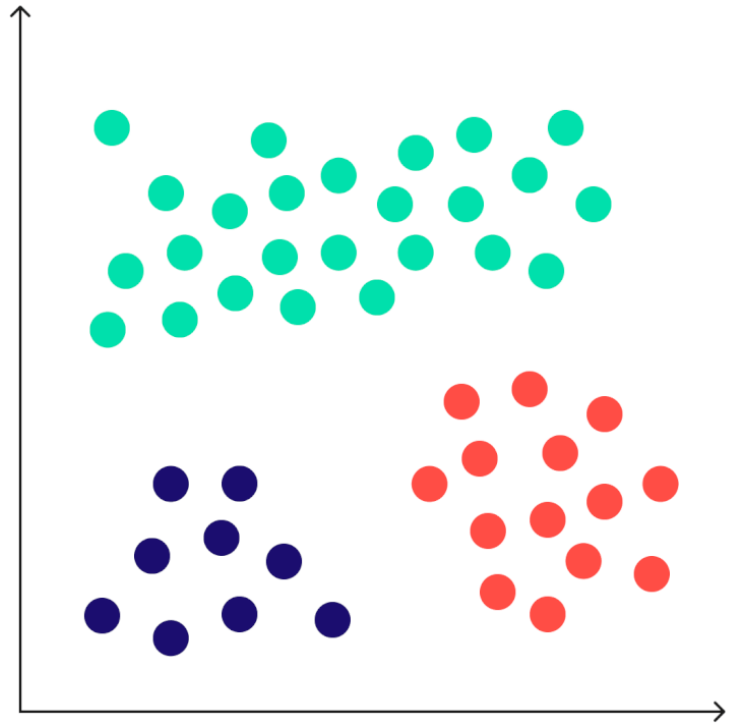
Outline: Data Exploration



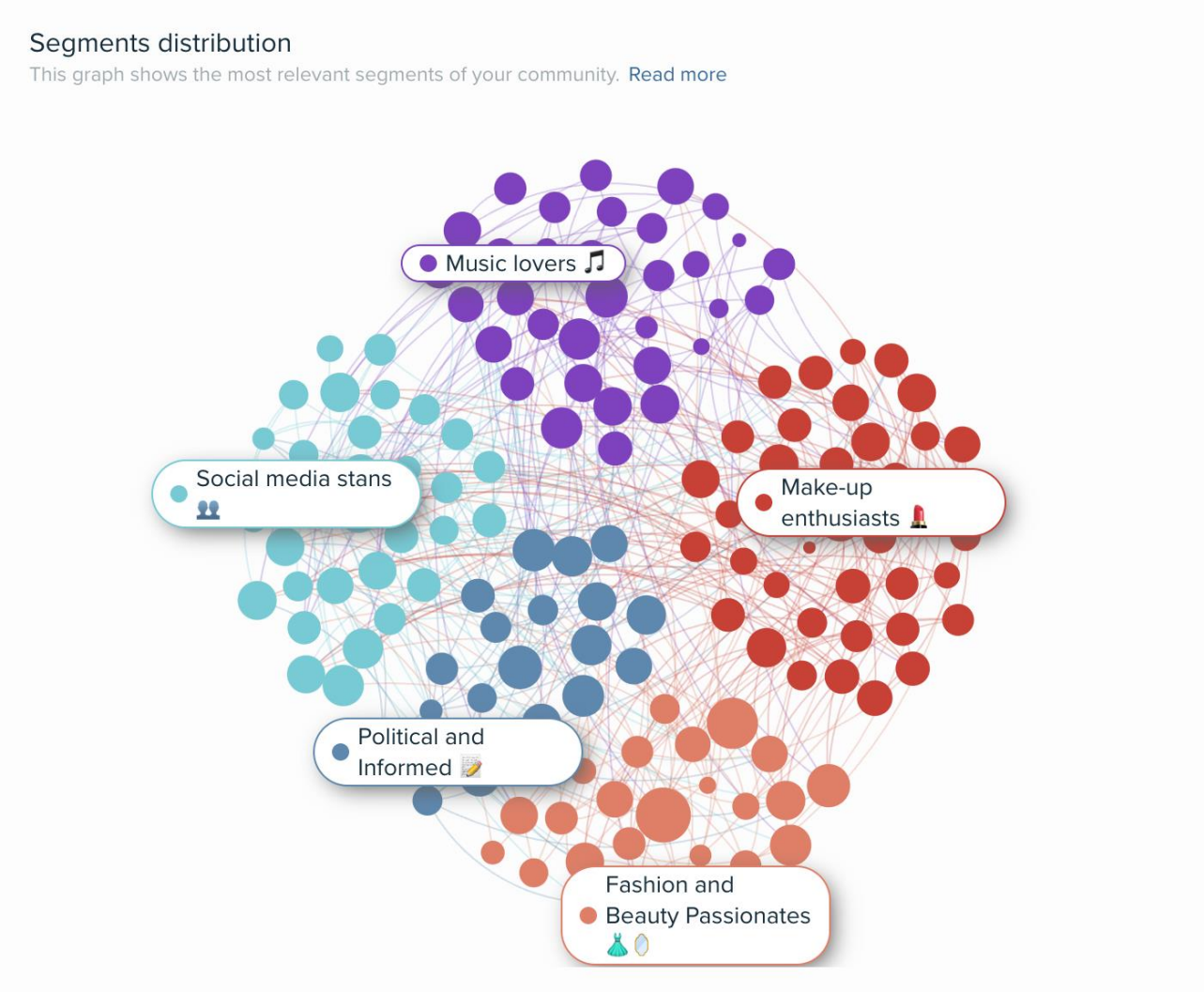
Data Manifold



Data Manifold w/ Clustering



Case Study: Segmentation & Persona



Top bio keywords

love 4% | life 3% | makeup 3% | ❤️ 2% | beauty 2% | lover 2% | ig 2% | mom 2% | artist 2% | like 1%

Distinctive affinities

ColourPop Cosmetics | Morphe | Kylie Jenner | Rihanna | Anastasia Beverly Hills

Top hashtags

#metgala #thelittlemermaid #renaissanceworldtour #suga #eurovision #tstheerastour #agustd #bts

Top characteristics

Age	18-24	32.10%
Gender	Female	83.34%
Country	United States of America	54.73%
Interests	Education	73.65%

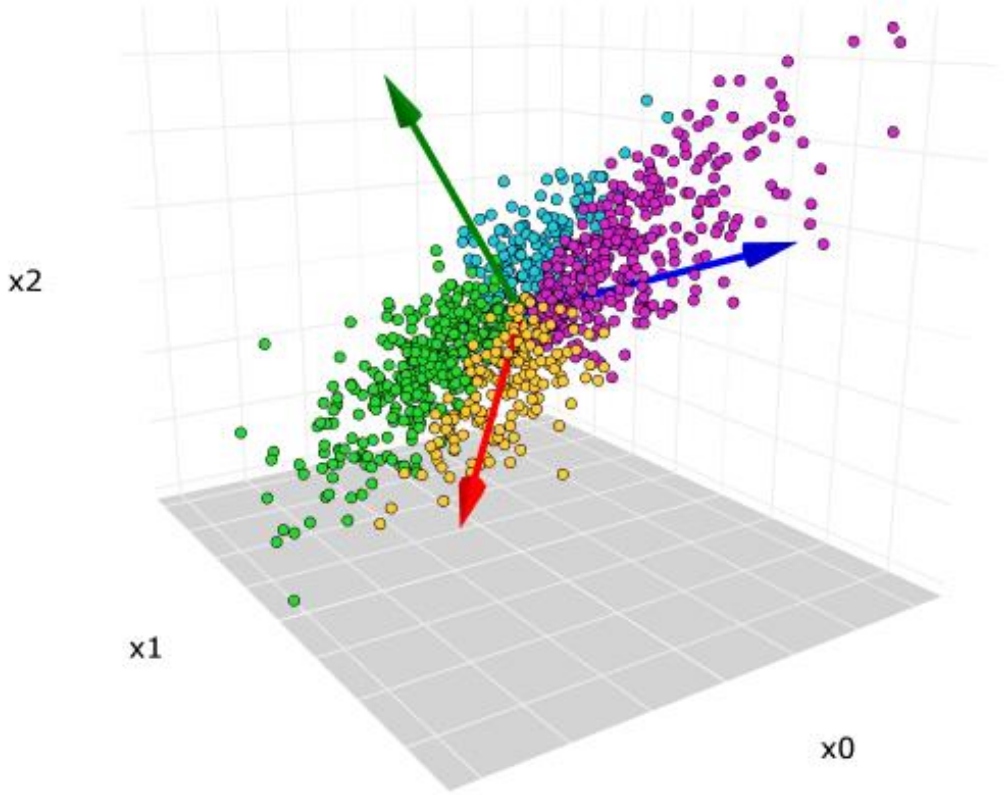
View more details →

https://resources.audiense.com/en/blog/how-to-create-effective-marketing-segmentation-and-persona-strategies

Dimensionality Reduction

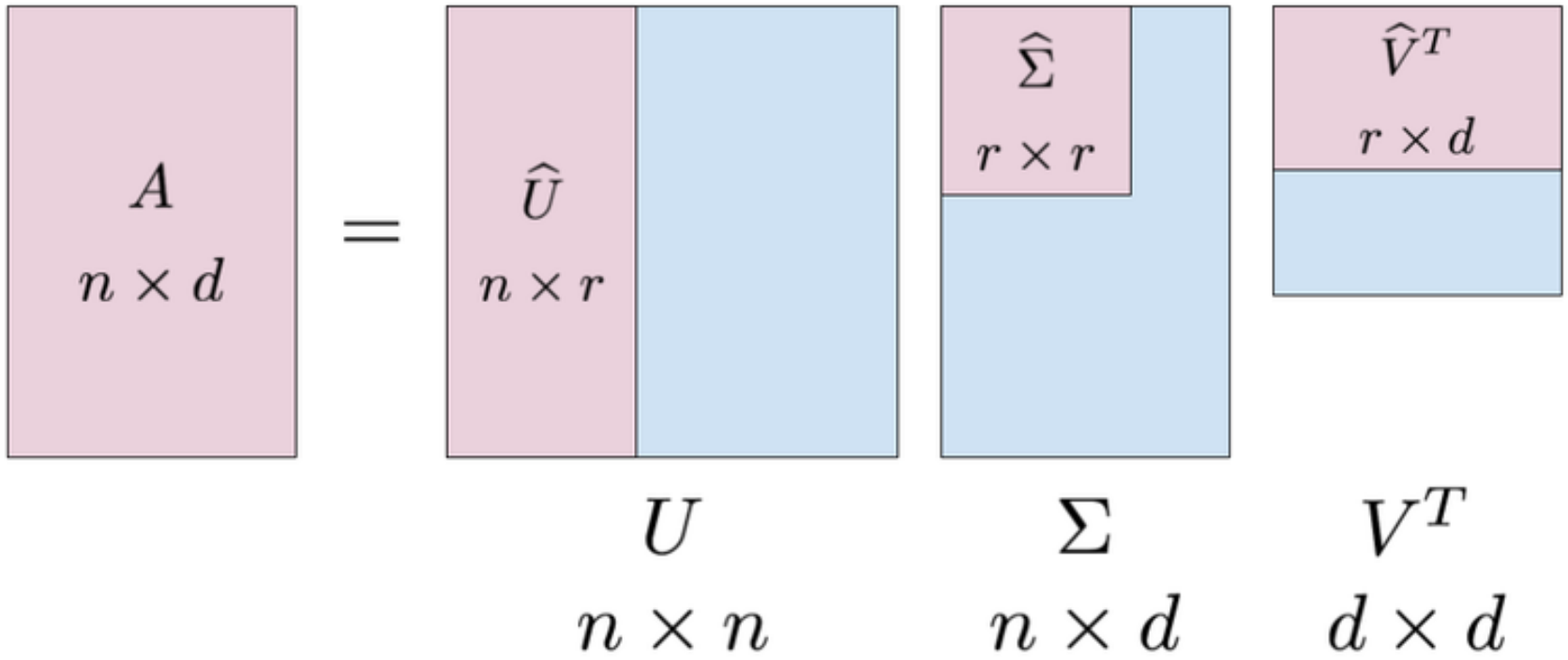
Dimensionality Reduction

Principal Component Analysis (PCA)



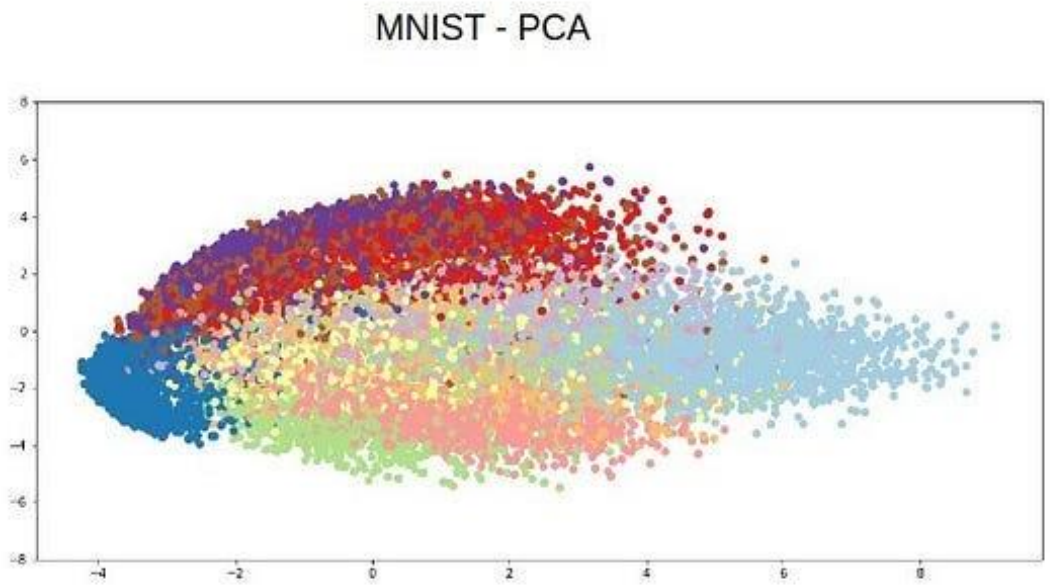
Dimensionality Reduction

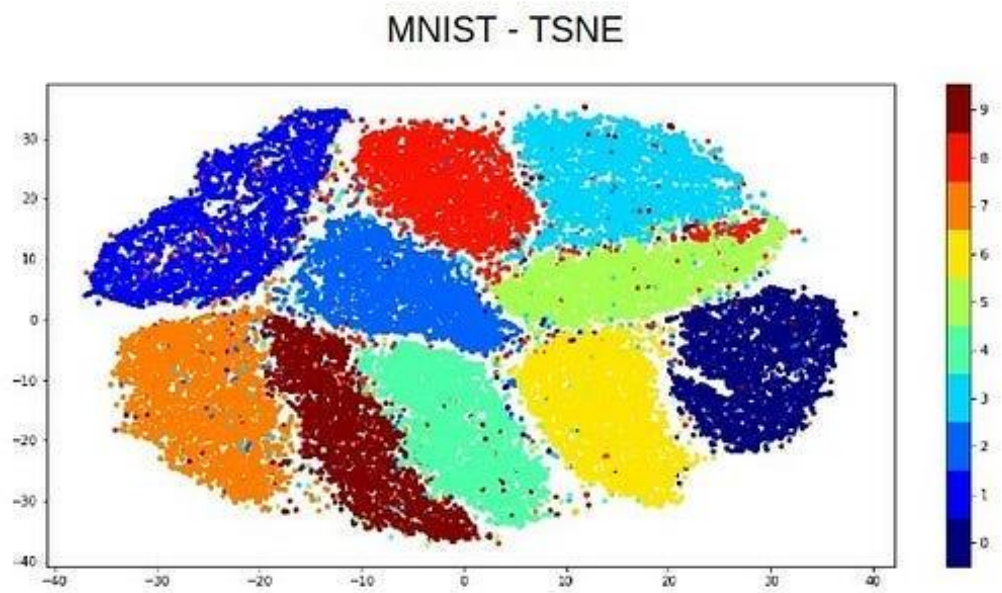
Principal Component Analysis (PCA)



Dimensionality Reduction

t-distributed Stochastic Neighbor Embedding (t-SNE)





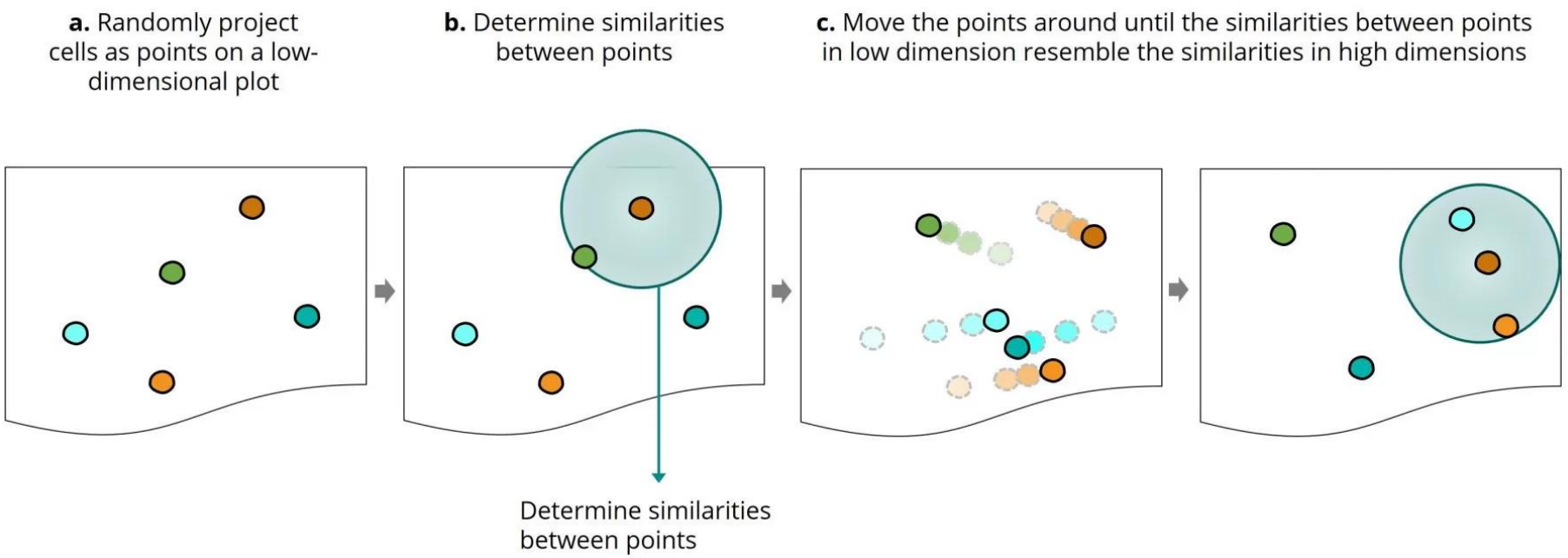
Dimensionality Reduction

t-distributed Stochastic Neighbor Embedding (t-SNE)

	PCA	t-SNE
Purpose	Dimensionality reduction	Data visualisation in low dimensions (typically 2D)
Type of algorithm	Deterministic	Stochastic
Unique solution?	Yes	No
Projection type	Linear	Non-linear
Type of approach	Global	Local/Global*
Interpretation	PCA is just a rotation of axes	Subjective interpretation

Dimensionality Reduction

t-distributed Stochastic Neighbor Embedding (t-SNE)



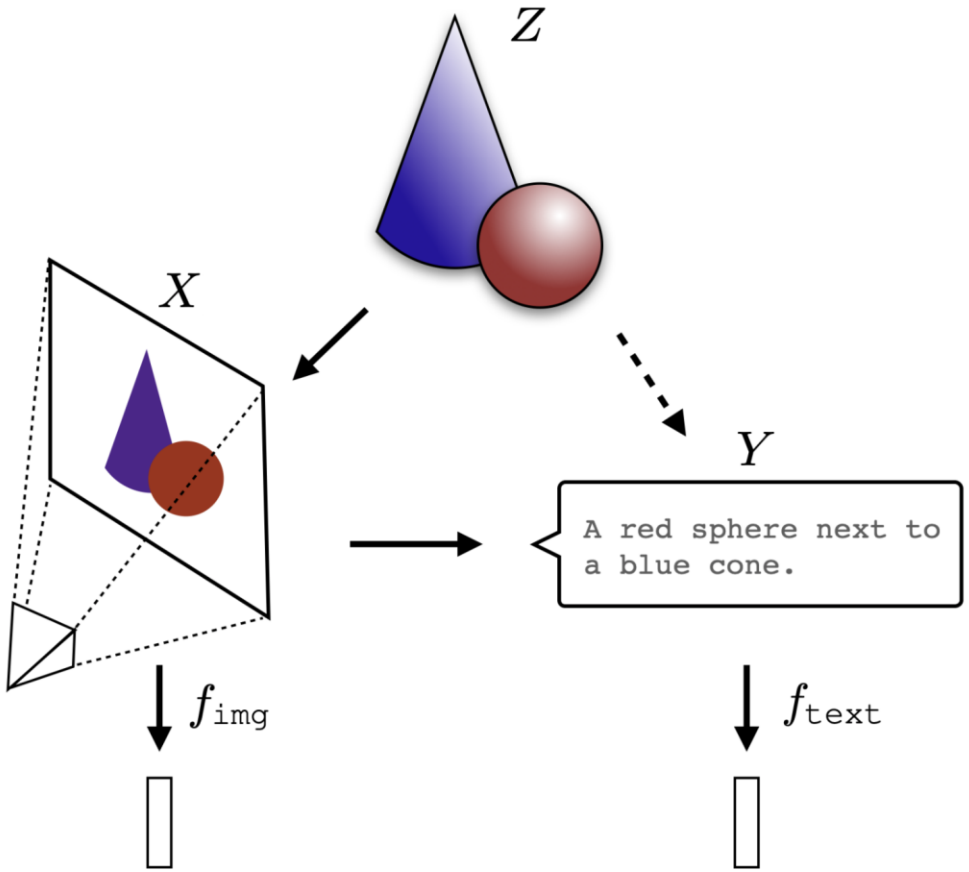
Dimensionality Reduction

Deep Representation Learning



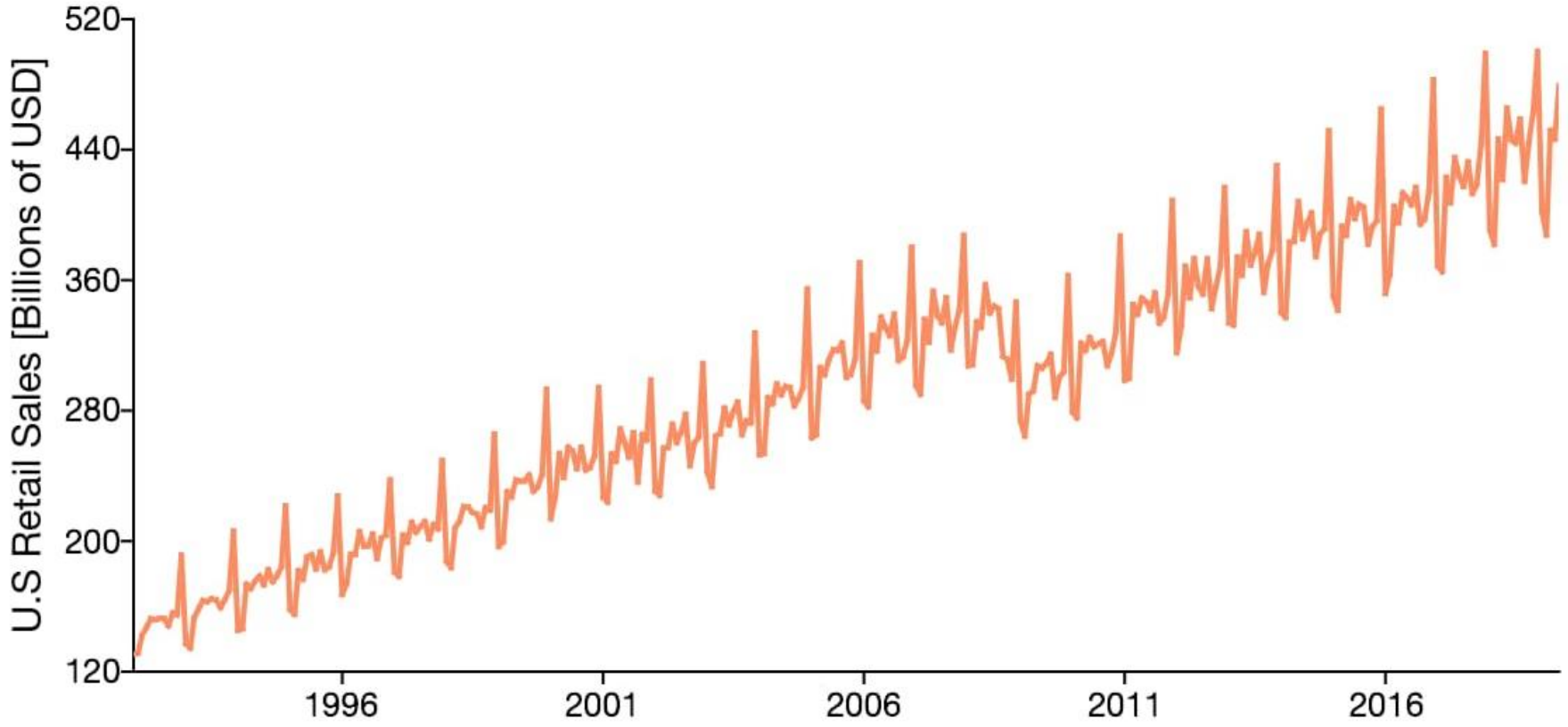
Dimensionality Reduction

Deep Representation Learning

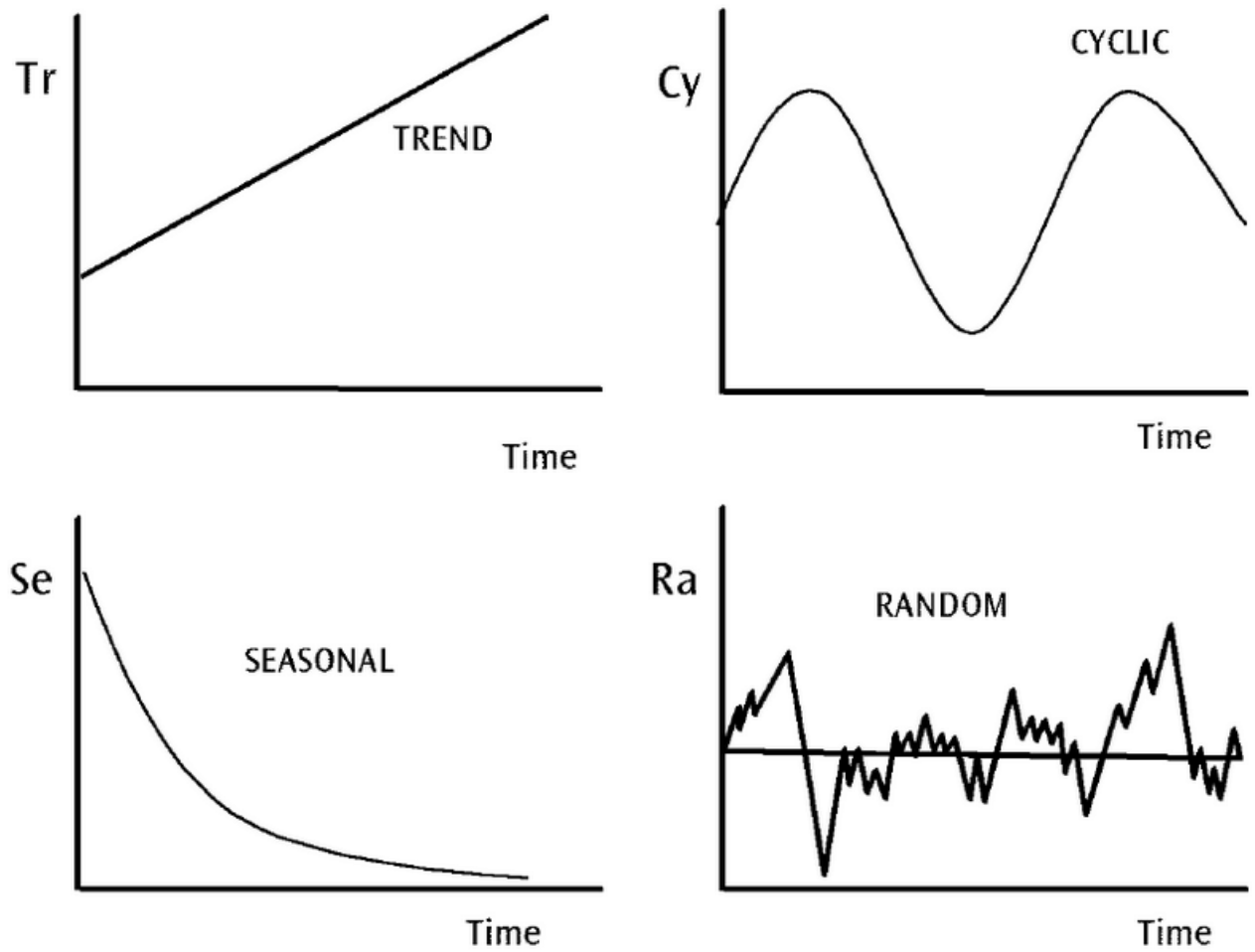


The Platonic Representation Hypothesis. ICML Position Paper, 2024.

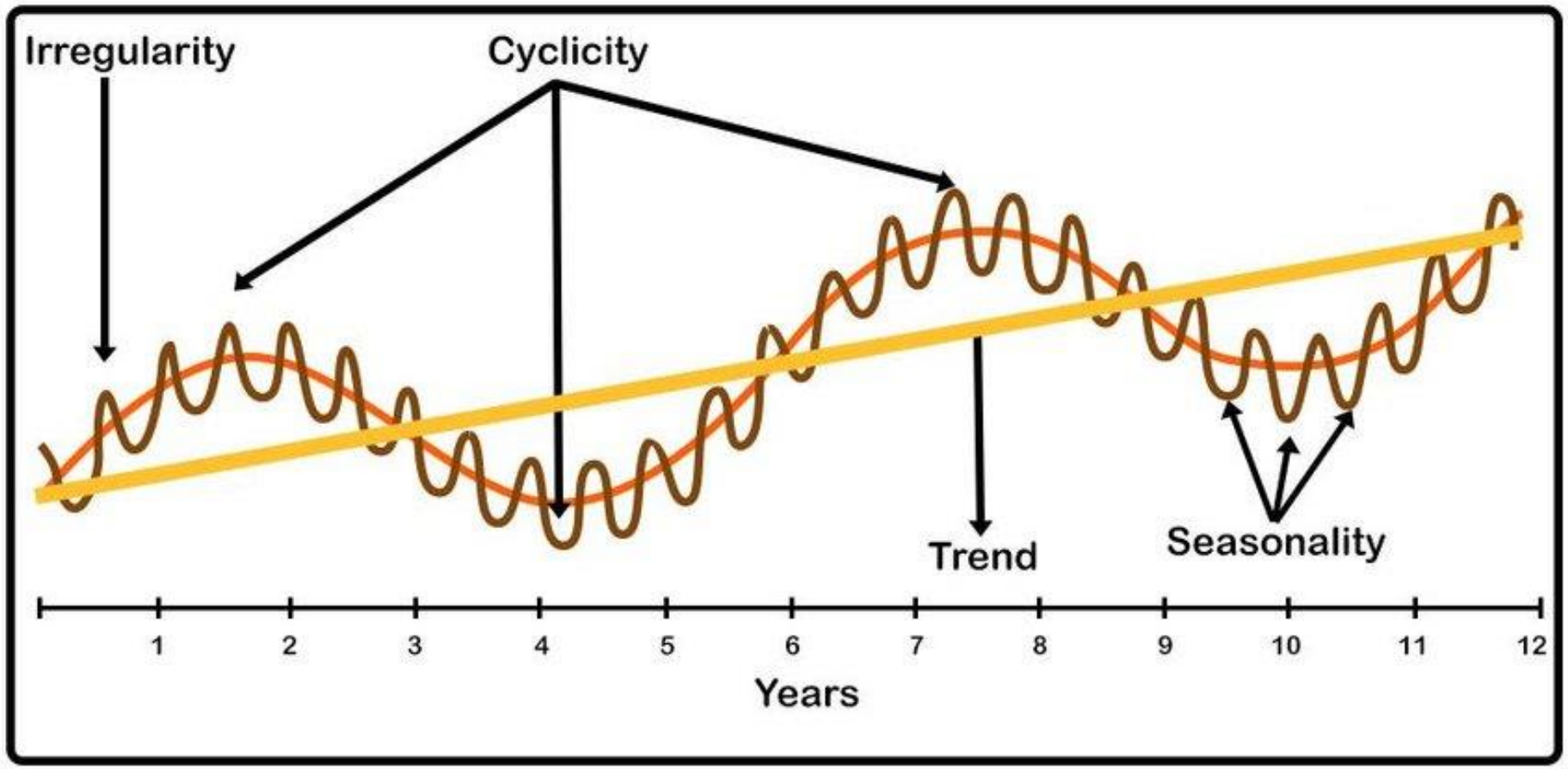
Time-series


U.S Retail Sales [Billions of USD]
520
440
360
280
200
120
1996
2001
2006
2011
2016

Time-series Components

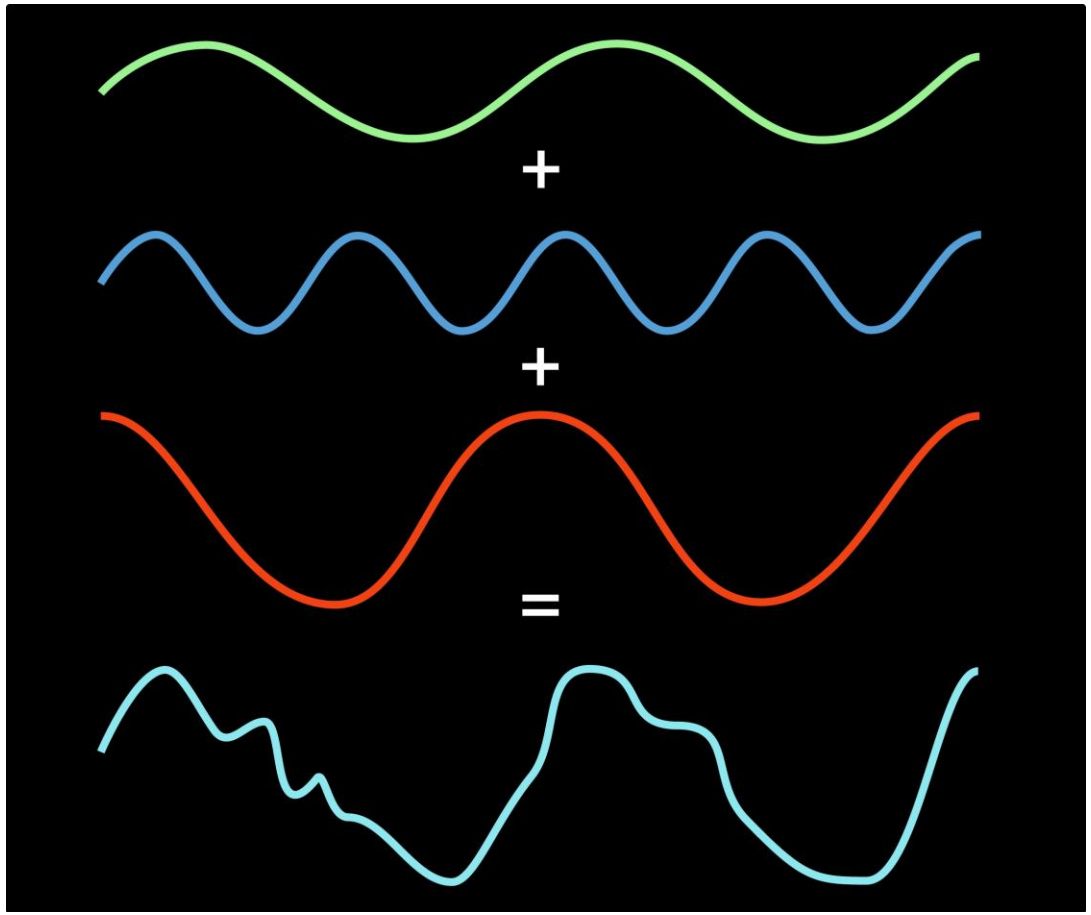


Time-series Components



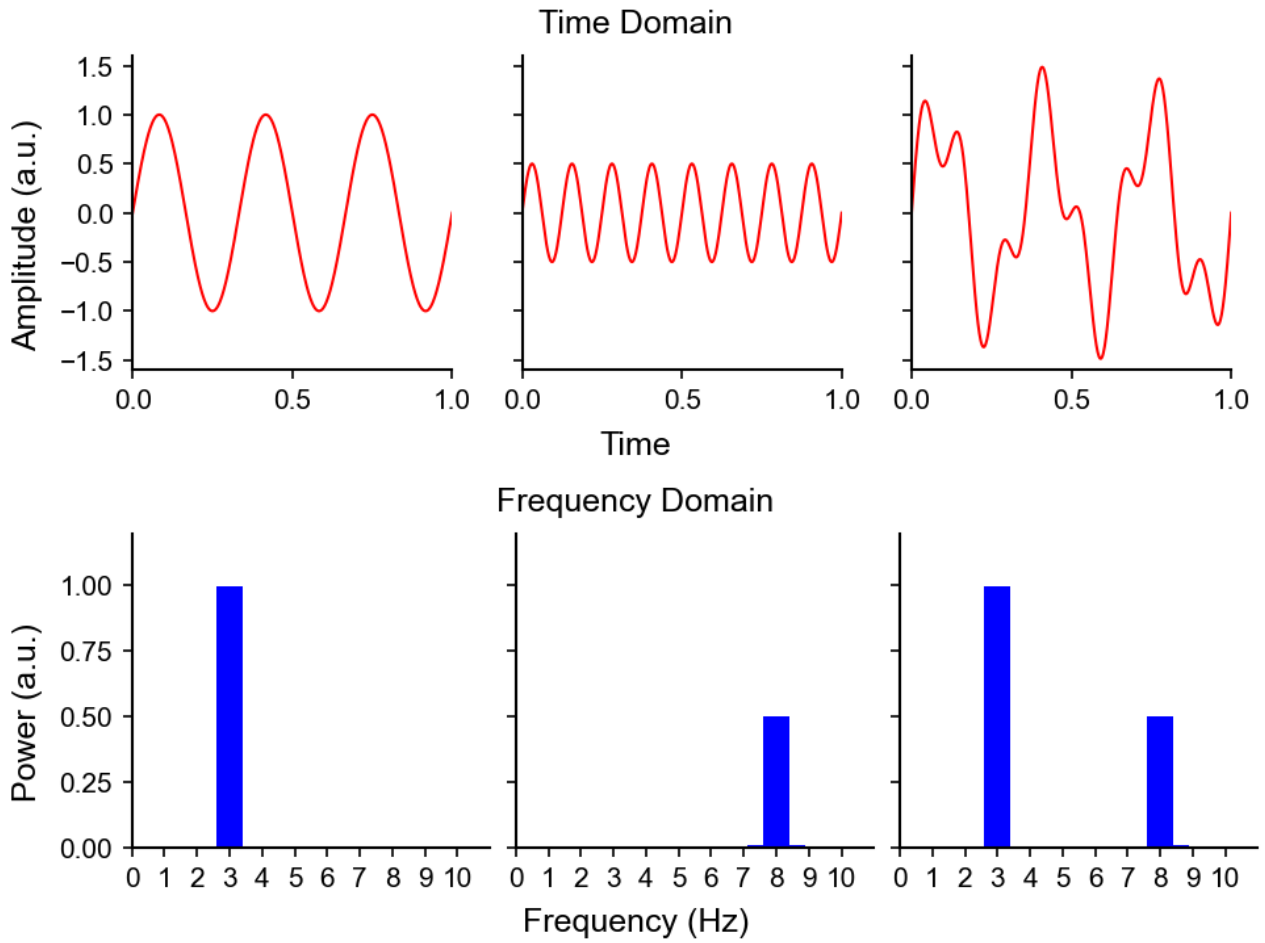
Fourier Decomposition

Time-series ≈ Sum of sinusoidal functions

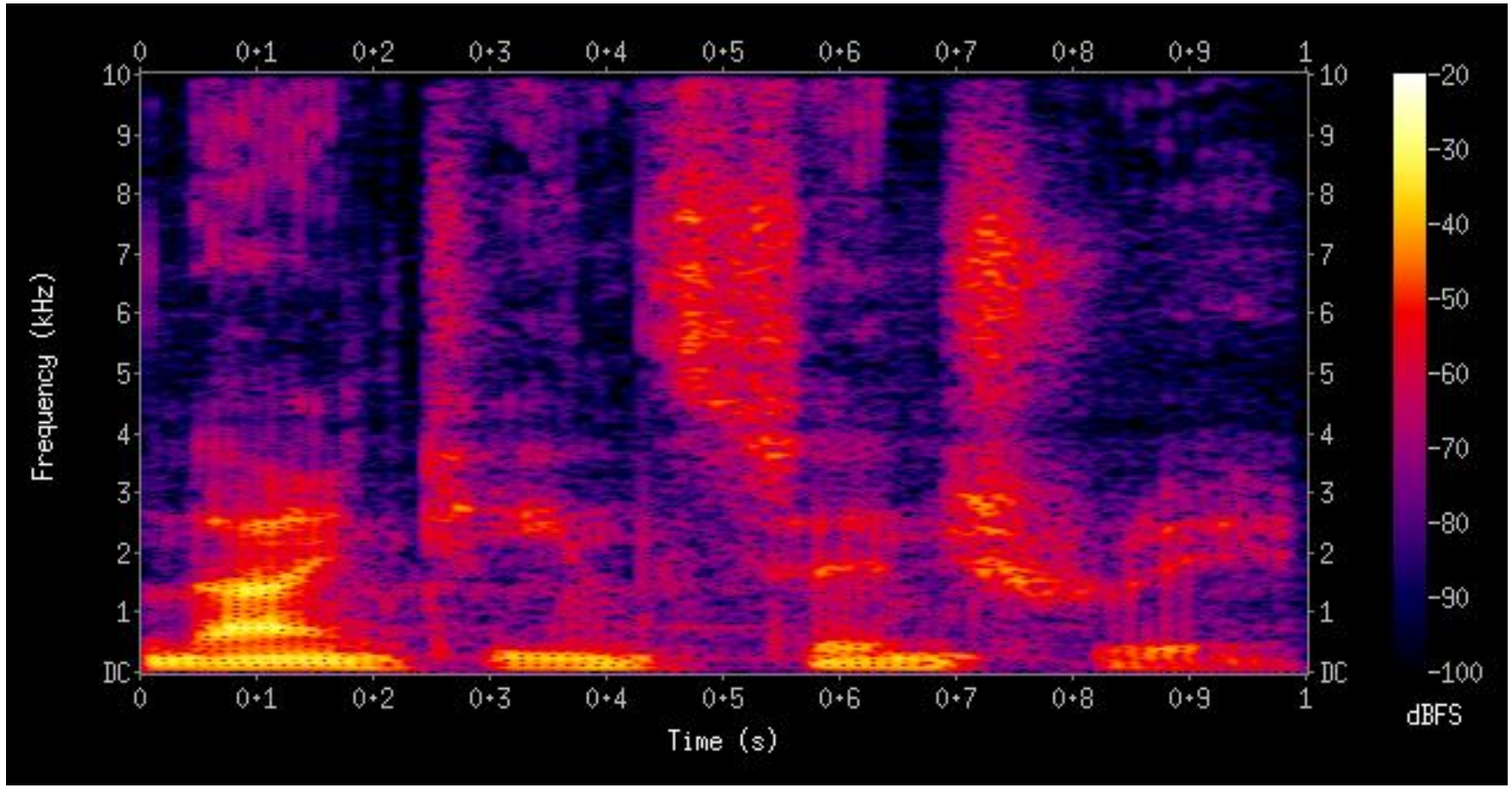


Fourier Decomposition

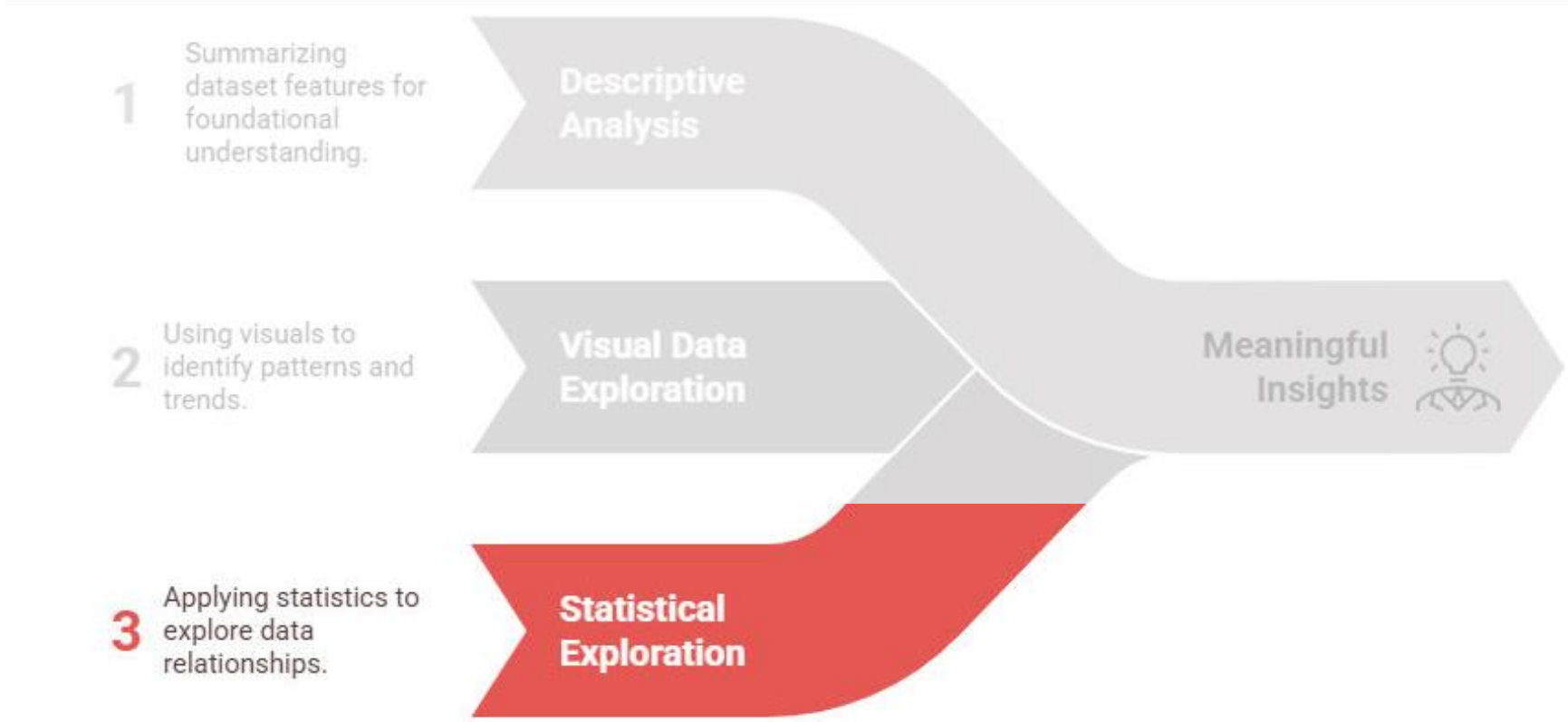
Time & frequency domains



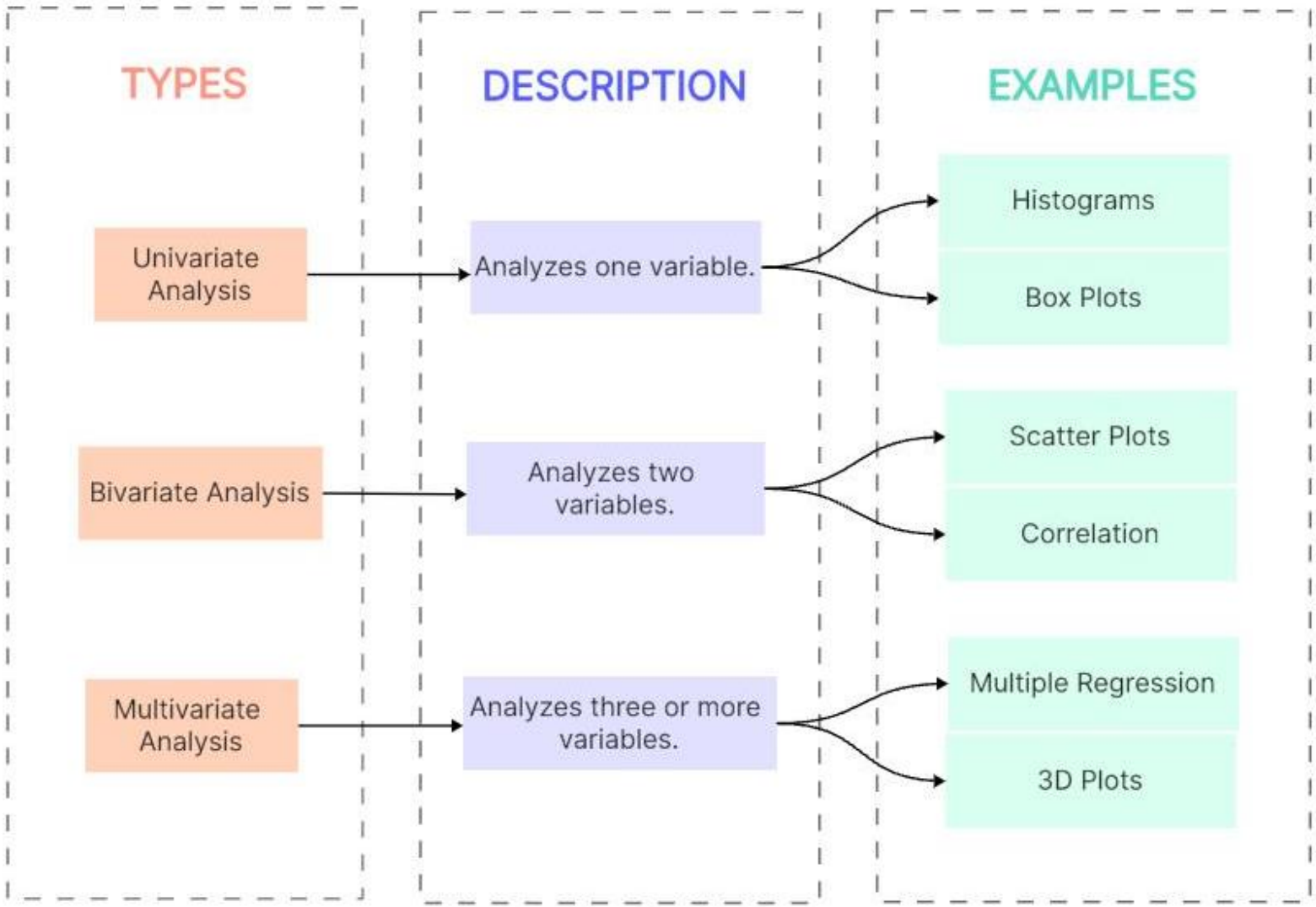
Audio Spectrogram



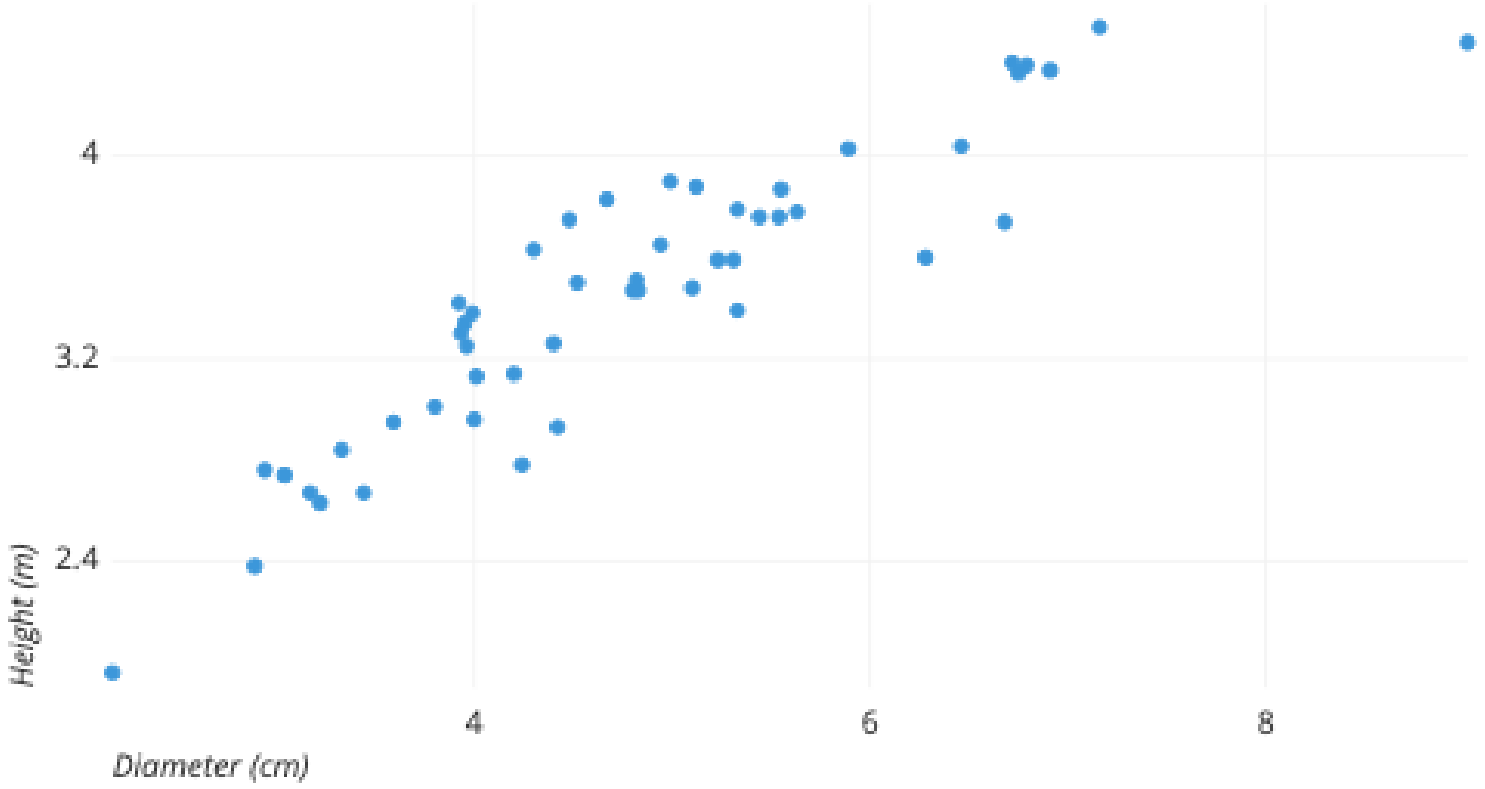
Outline: Data Exploration



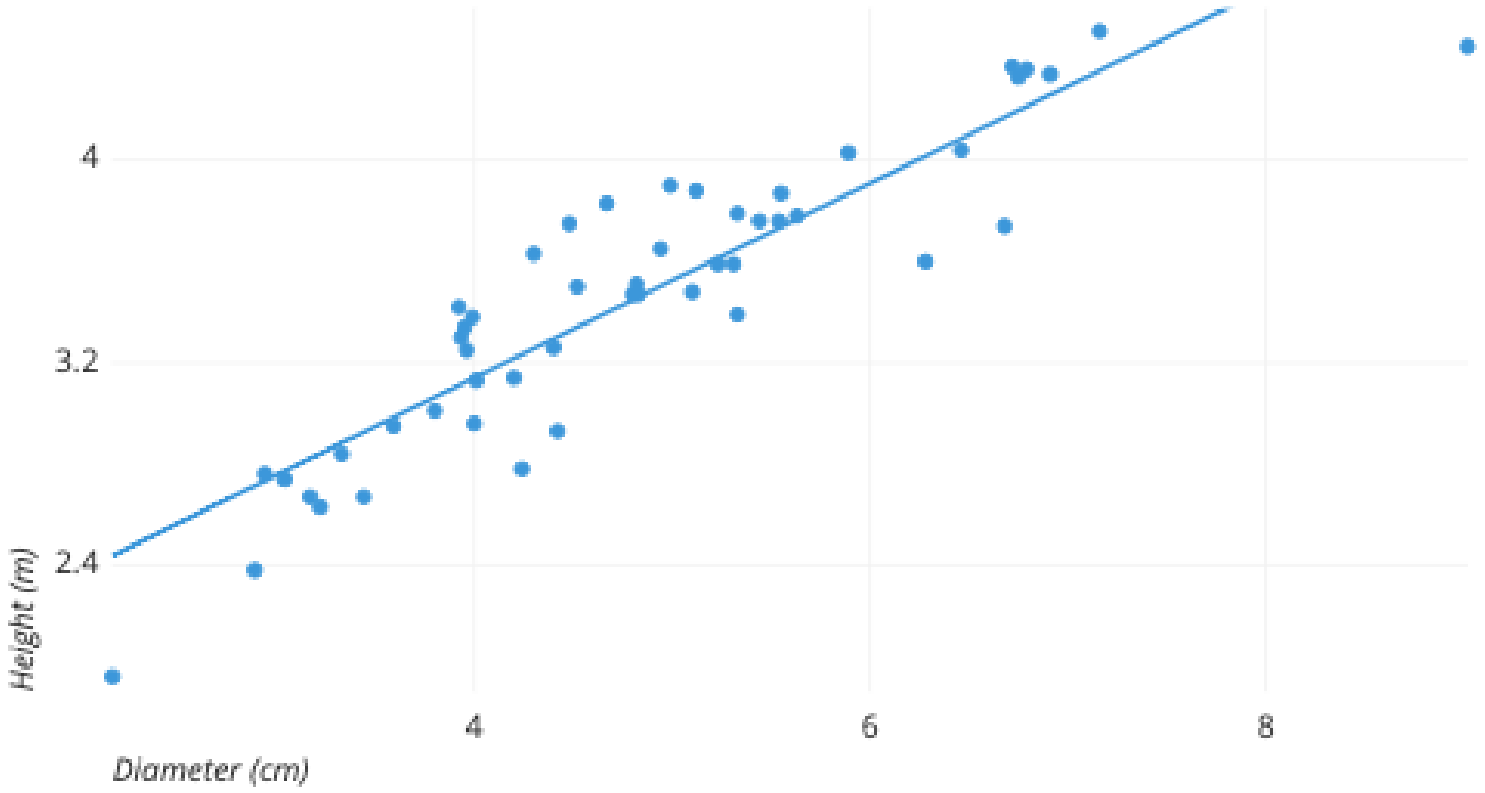
Univariate vs. Multivariate Analysis



Scatter Plots

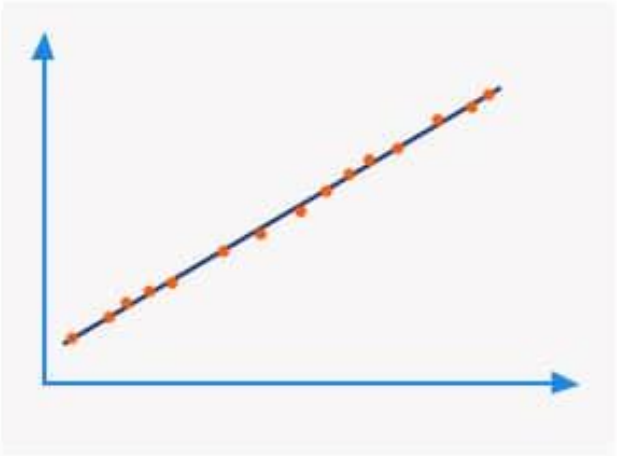


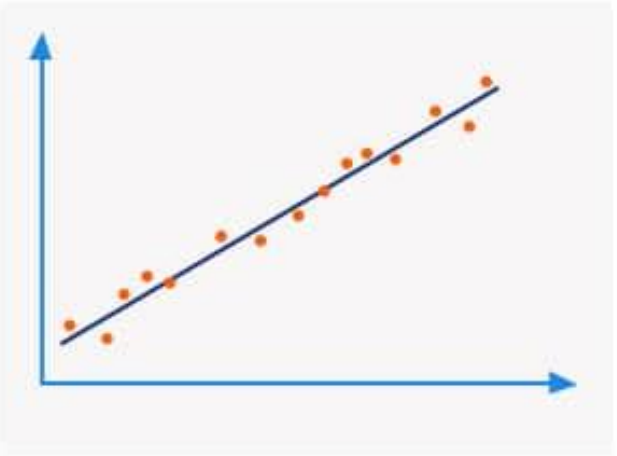
Scatter Plots w/ Trend Lines



(Pearson) Correlation

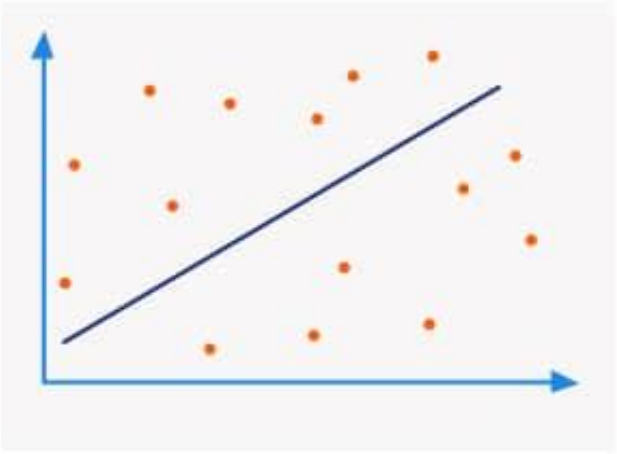
1.
Large positive correlation





2.
Medium positive correlation

4.
Weak / no correlation





3.
Small negative correlation

(Pearson) Correlation

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

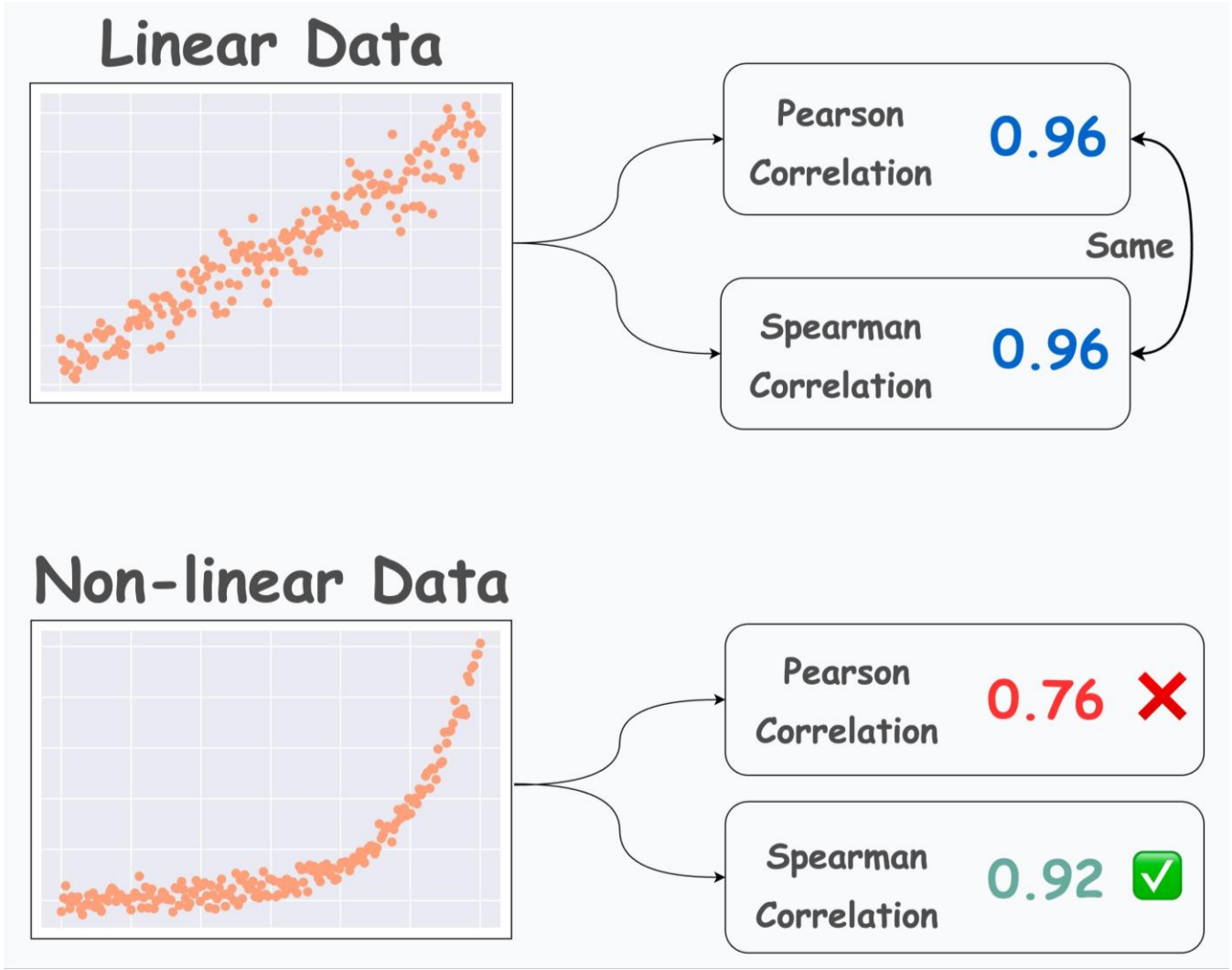
x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

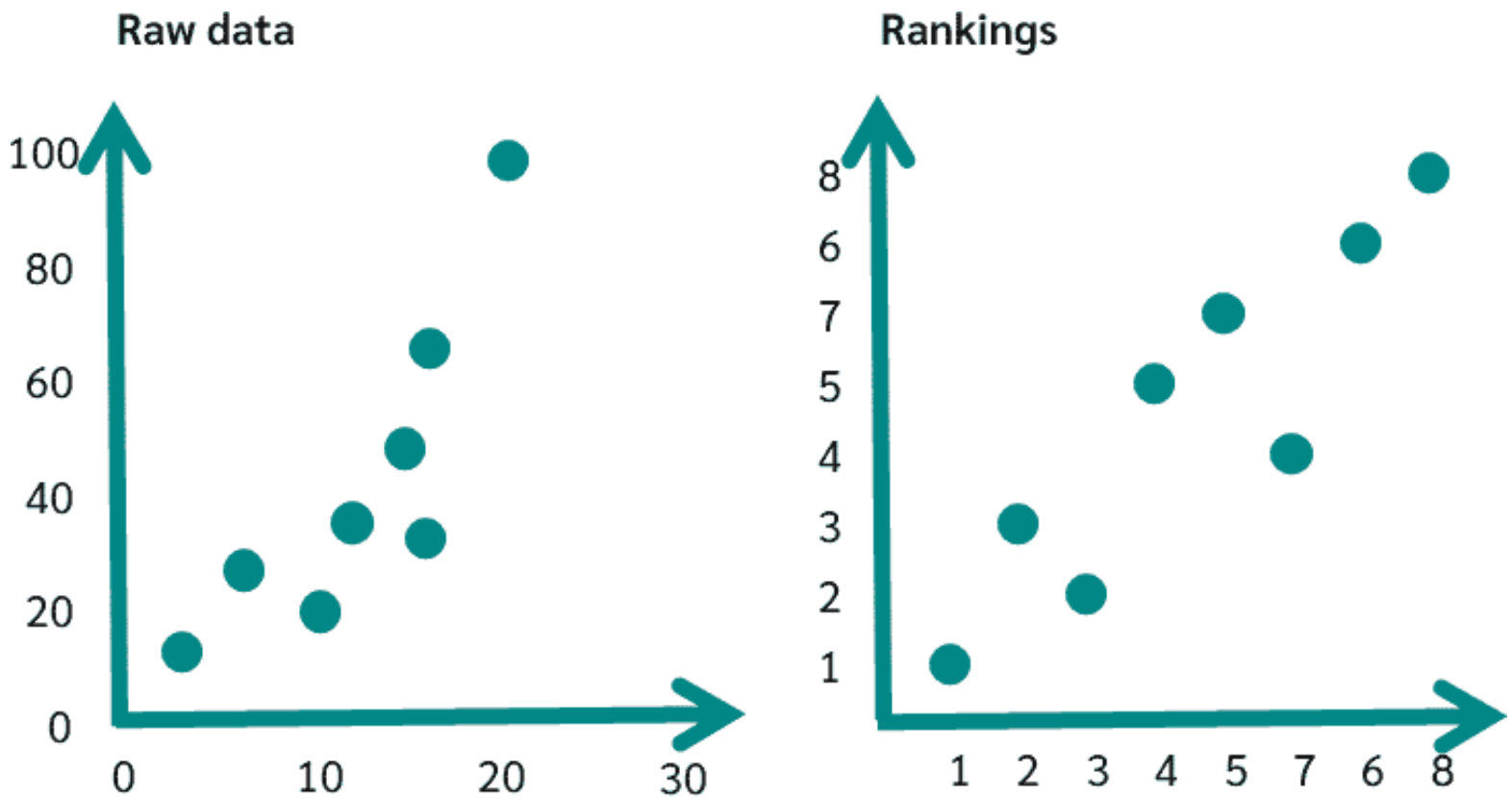
y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

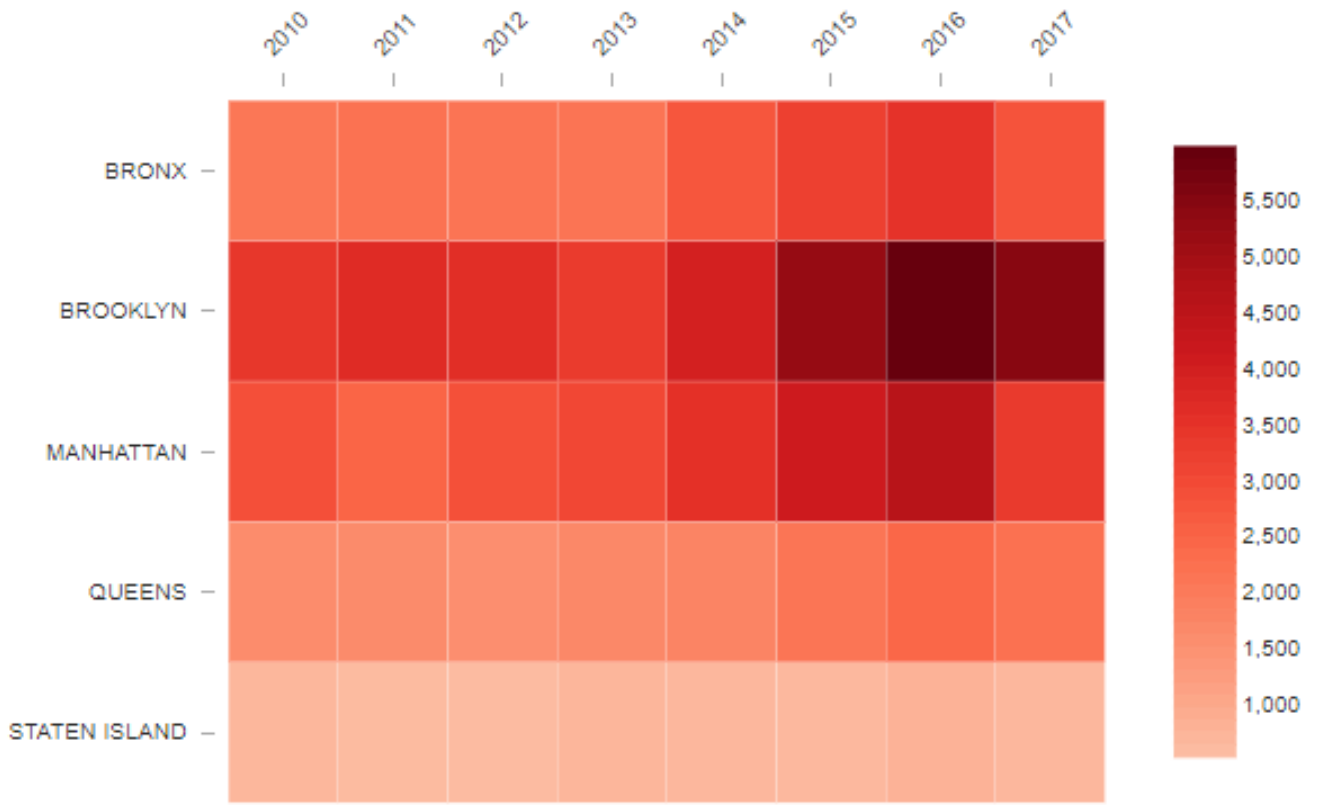
Spearman's Rank Correlation



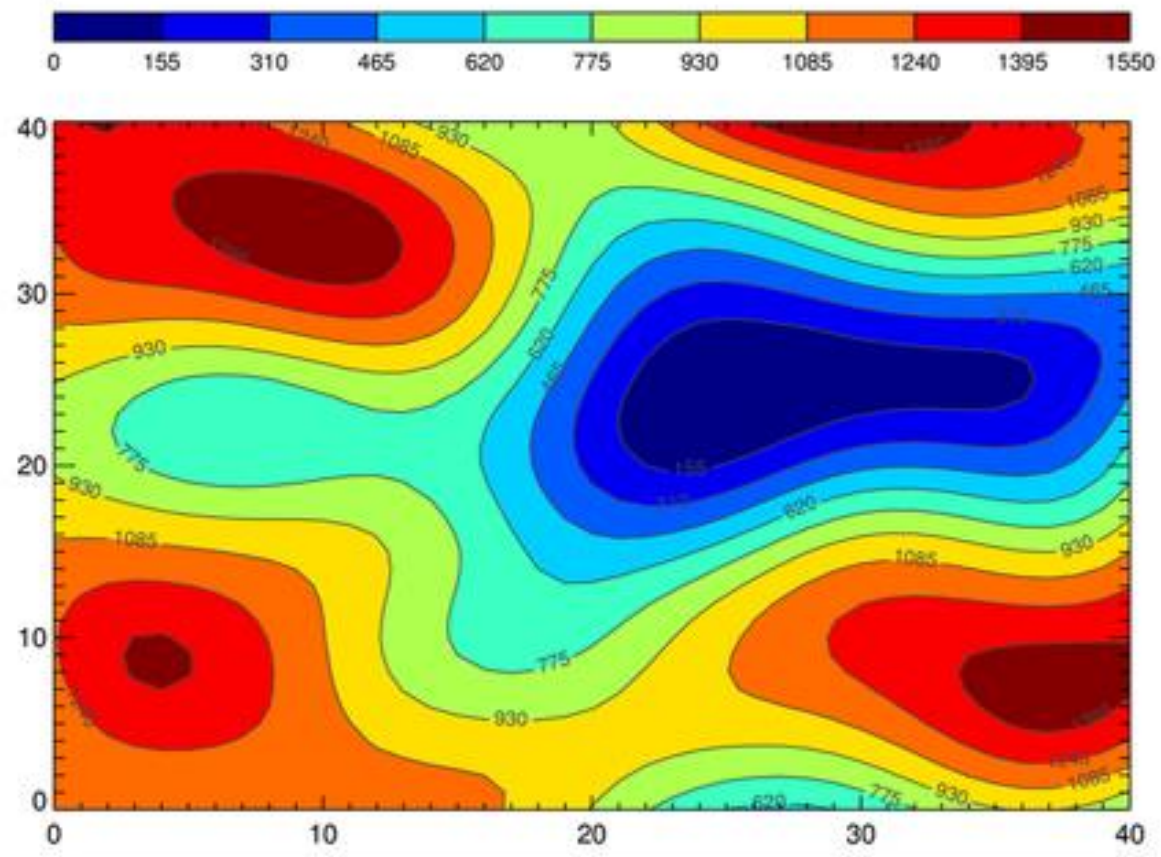
Spearman’s Rank Correlation



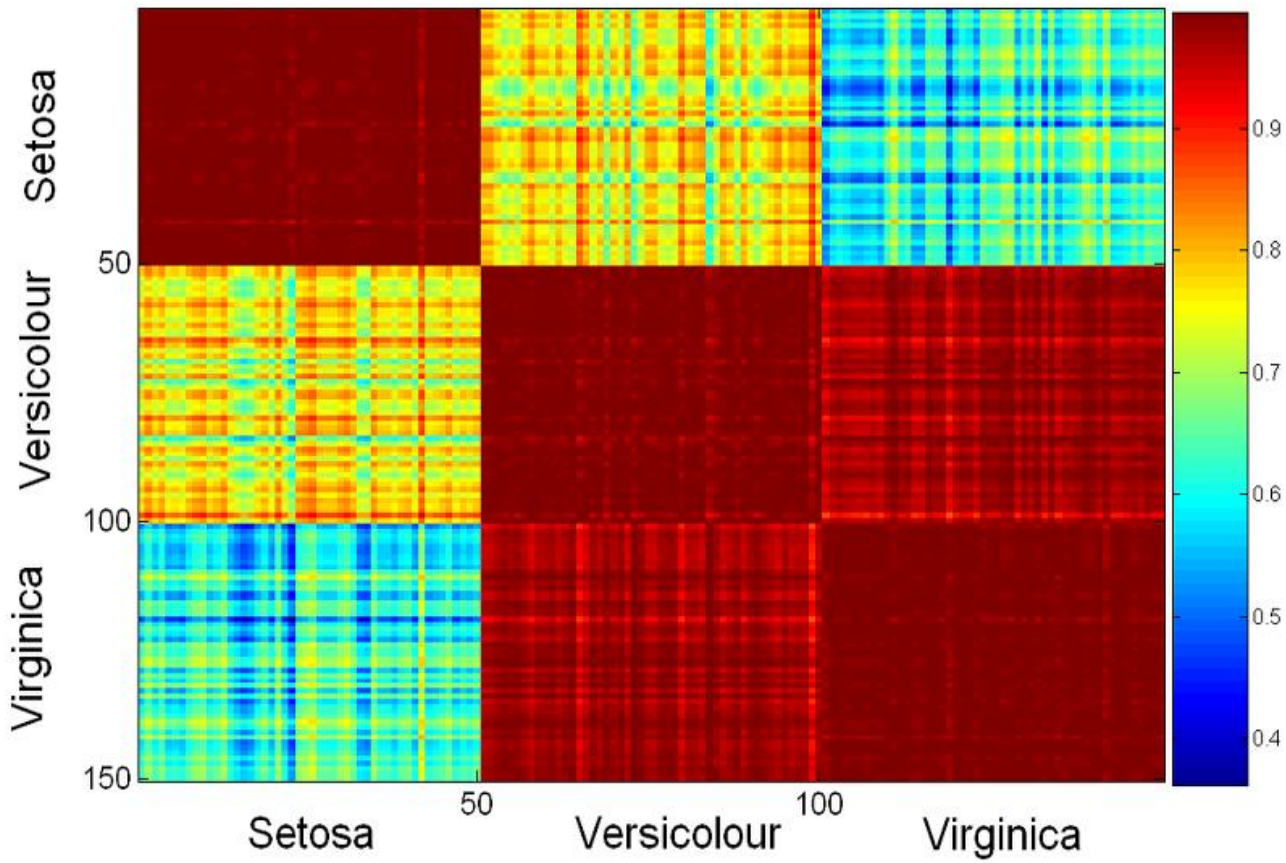
Heatmaps


2010
2011
2012
2013
2014
2015
2016
2017
BRONX
BROOKLYN
MANHATTAN
QUEENS
STATEN ISLAND
5,500
5,000
4,500
4,000
3,500
3,000
2,500
2,000
1,500
1,000

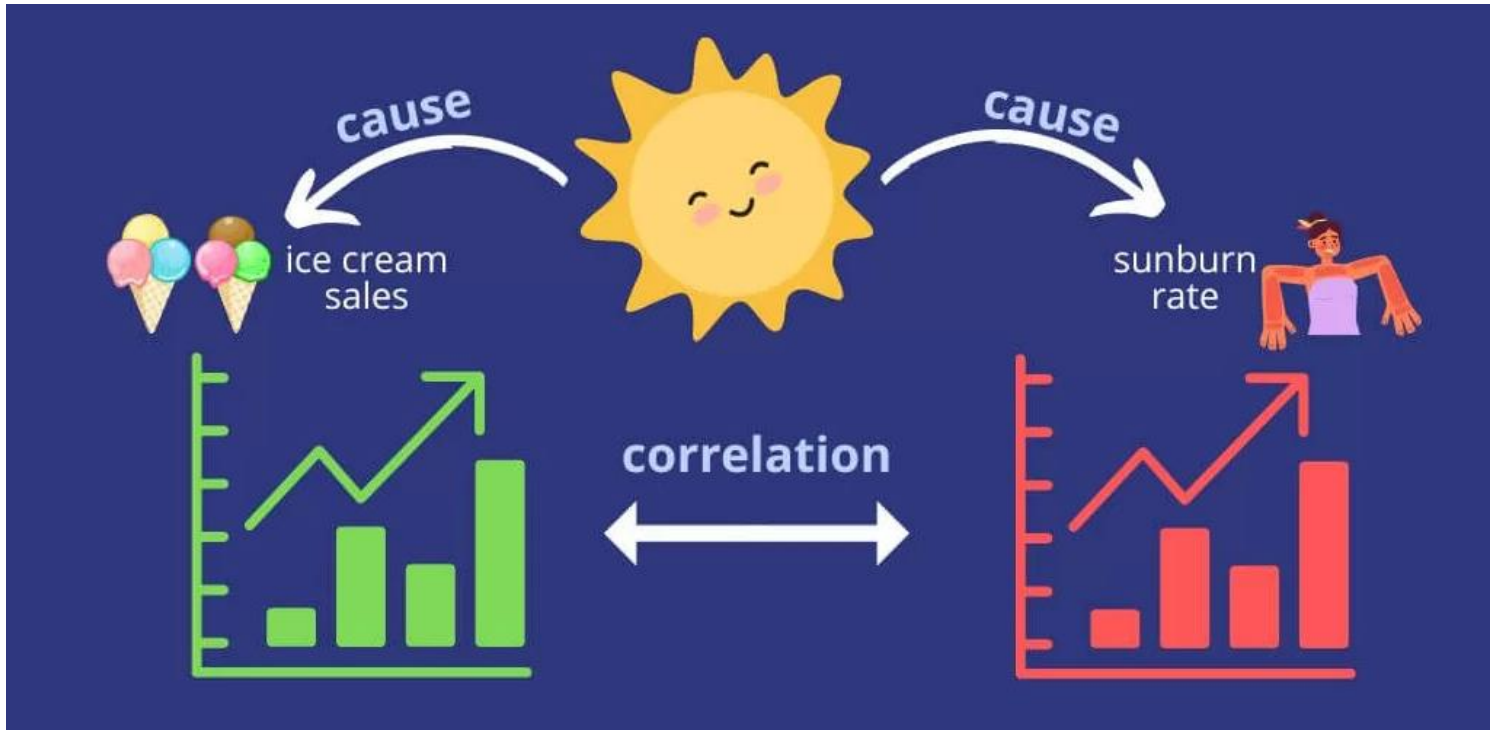
Contour Plots



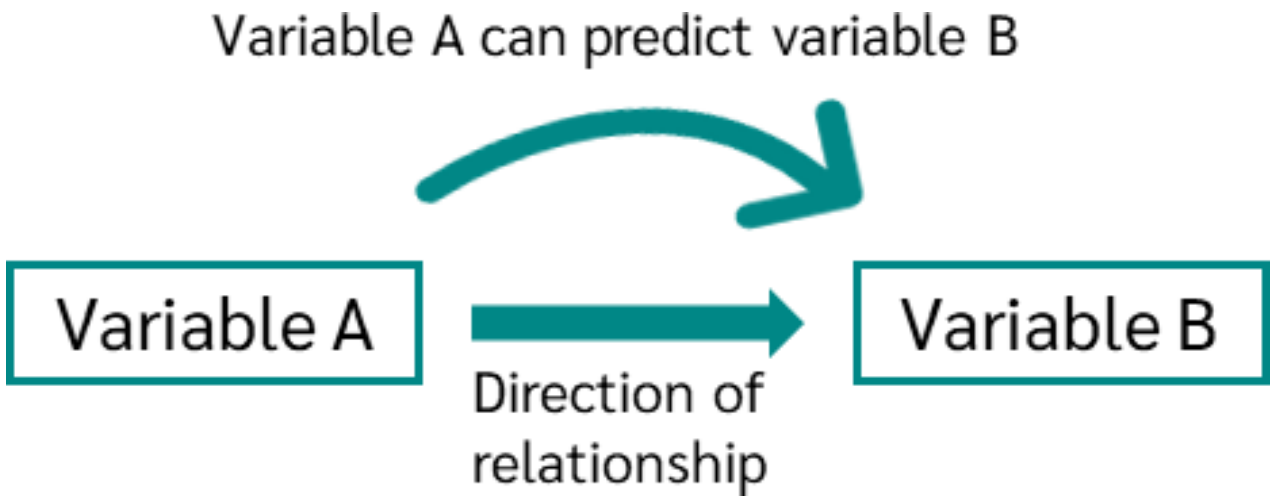
Matrix Plots



Correlation ≠ Causation

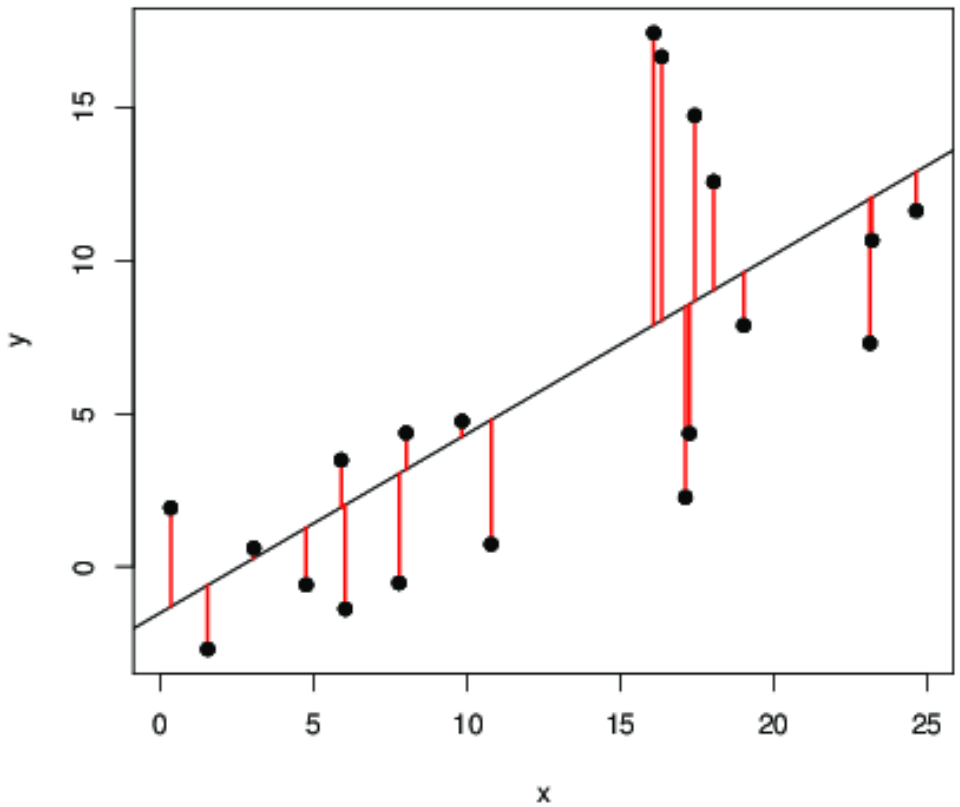


Correlation ≠ Regression

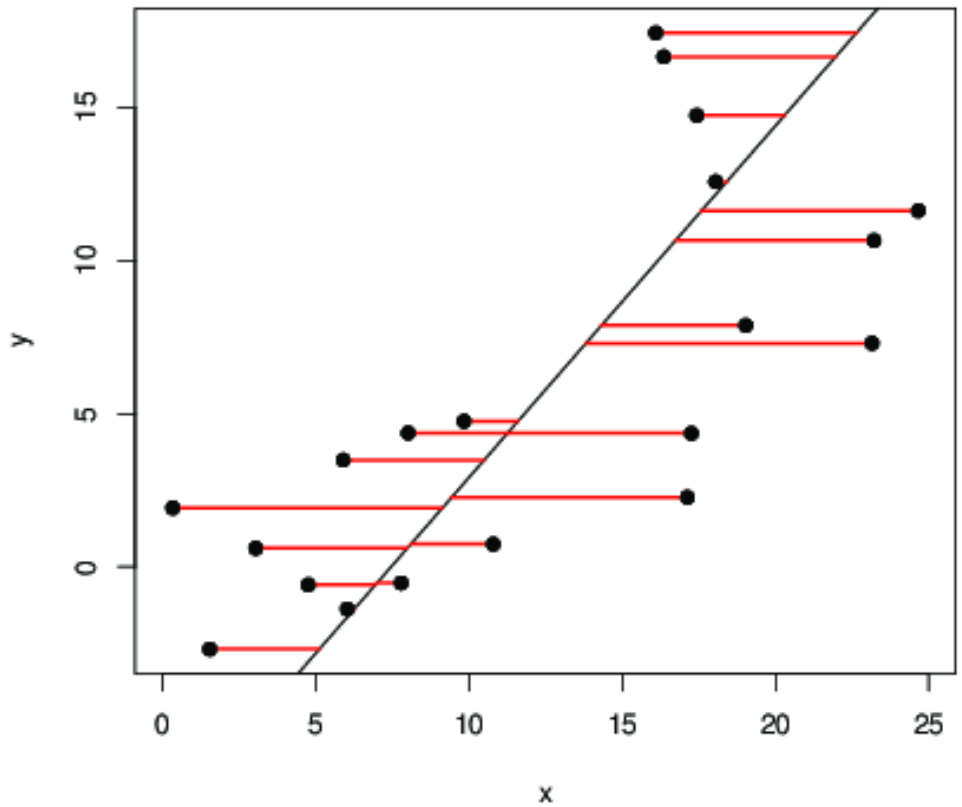


Correlation ≠ Regression

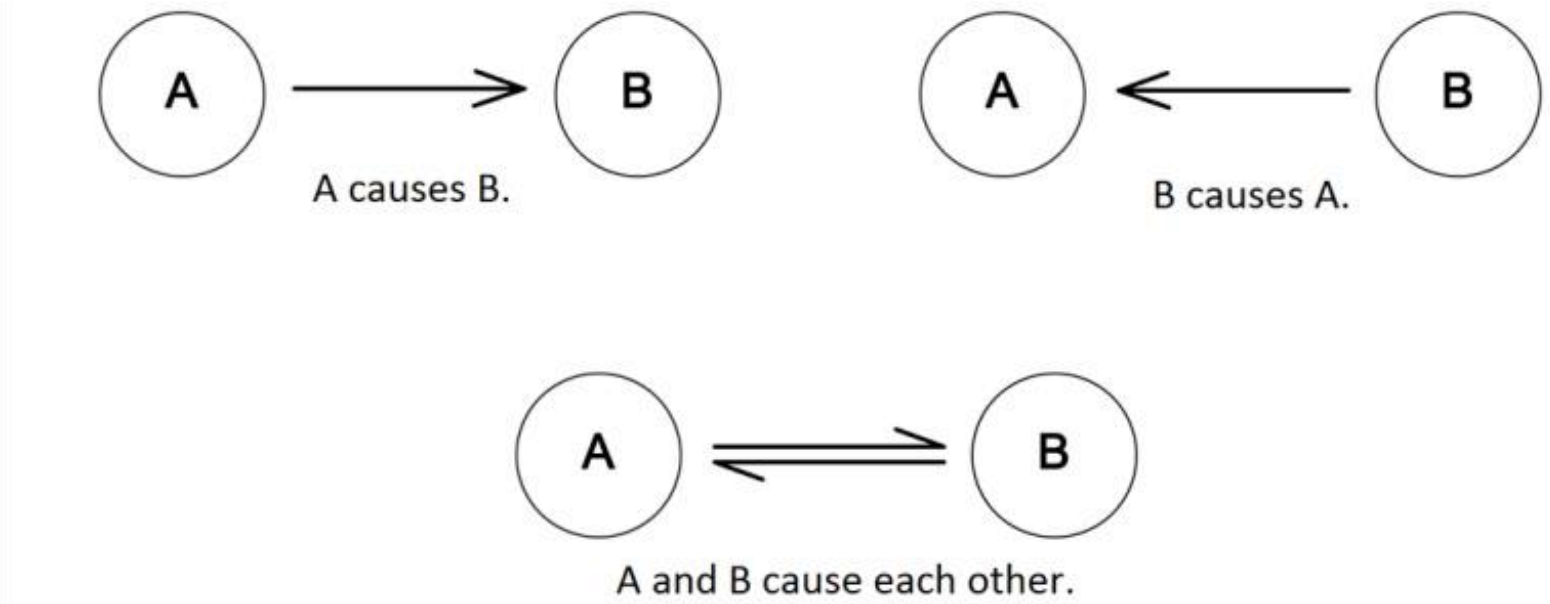
$$y = Ax + b + \epsilon$$



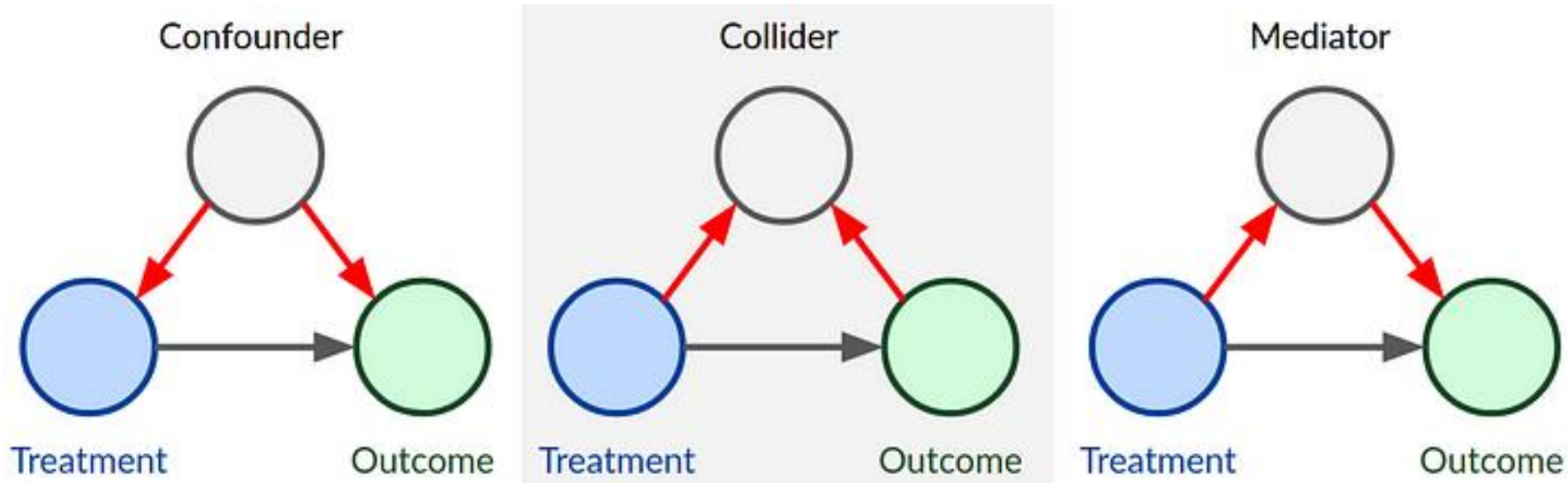
$$x = Ay + b + \epsilon$$



Causal Graph



Causal Graph



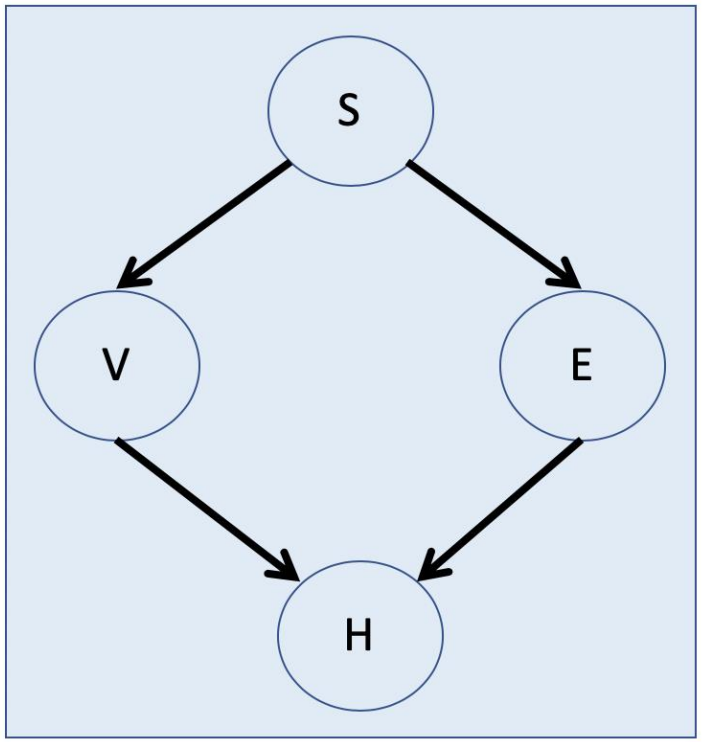
Causal Discovery

Data / Agent Experience

S	V	E	H
0	0	1	0
1	1	0	1
…	…	…	…

Causal Discovery

Causal Graph



Summary: Data Exploration

