

# Effective Big Data Management



# Topics (1)

Introduction to Big Data

Basic understanding of business analytic

Data format (structured vs unstructured)

Fundamental of Data management



# Topics (2)

Data analysis techniques

Statistical analysis

Working with data analysis tool

Business intelligence

Visualization data tool



# Introduction

Big Data

Business  
Analytics (BA)

Business  
Intelligence (BI)

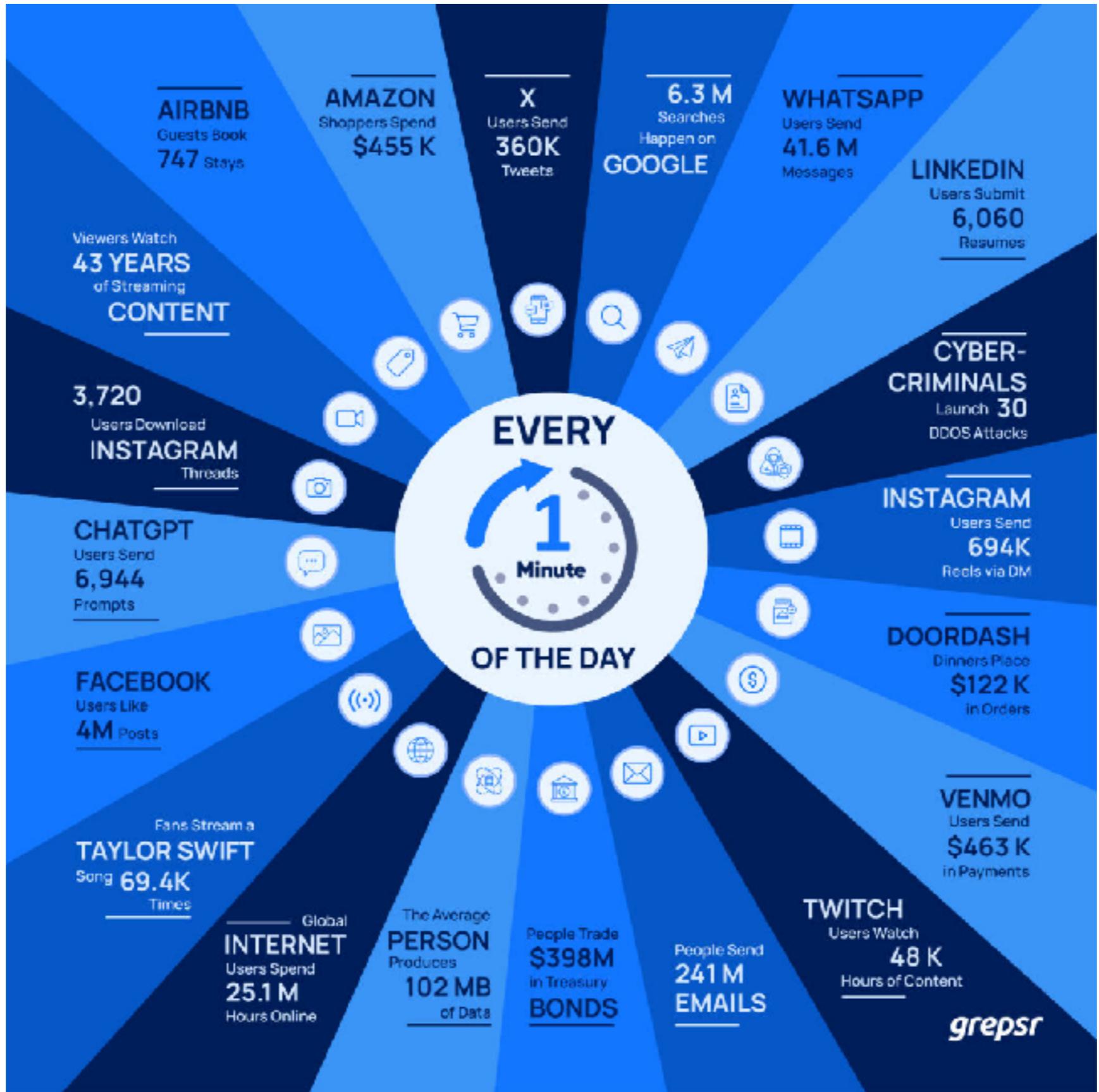
Data-driven Decision

Data management tools and platforms



# Introduction to Big Data







# Big Data

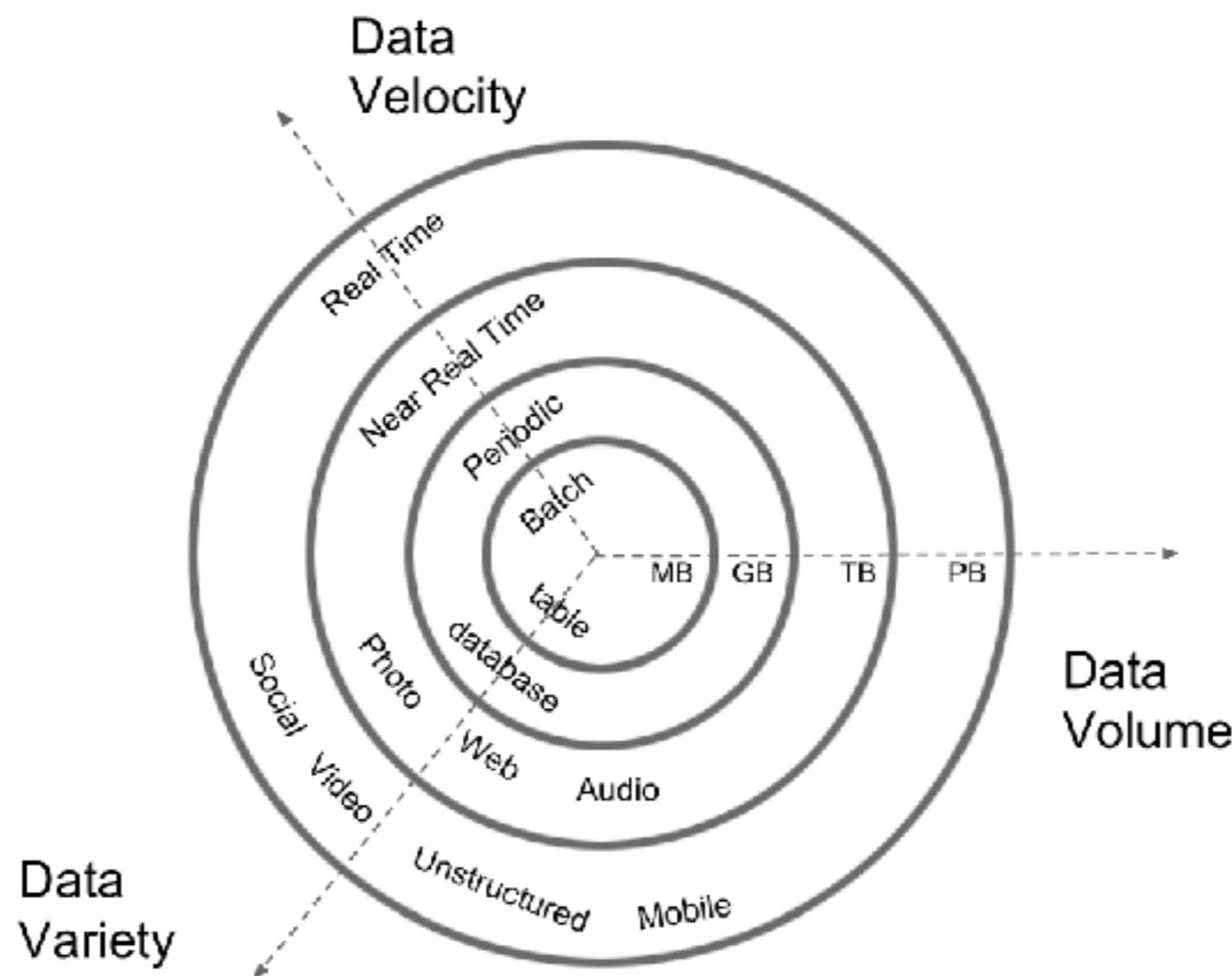
Refer to the vast amount of **structured** and **unstructured** data generated at high **velocity**, **variety**, and **volume**

Data is too complex to processed by traditional databases and tools !!



# Characteristics for Big Data

3 V's !!!



**40 ZETTABYTES**

[43 TRILLION GIGABYTES]

of data will be created by 2020, an increase of 300 times from 2005

**6 BILLION PEOPLE**

have cell phones

WORLD POPULATION: 7 BILLION

## Volume SCALE OF DATA



The New York Stock Exchange captures

**1 TB OF TRADE INFORMATION**

during each trading session



## Velocity ANALYSIS OF STREAMING DATA

By 2015, it is projected there will be

**18.9 BILLION NETWORK CONNECTIONS**

– almost 2.5 connections per person on earth



Modern cars have close to **100 SENSORS** that monitor items such as fuel level and tire pressure

# The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: Volume, Velocity, Variety and Veracity.

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015  
**4.4 MILLION IT JOBS**

will be created globally to support big data, with 1.9 million in the United States



As of 2011, the global size of data in healthcare was estimated to be

**150 EXABYTES**

[161 BILLION GIGABYTES]



## Variety DIFFERENT FORMS OF DATA

**30 BILLION PIECES OF CONTENT**

are shared on Facebook every month



**1 IN 3 BUSINESS LEADERS**

don't trust the information they use to make decisions



**27% OF RESPONDENTS**

in one survey were unsure of how much of their data was inaccurate

## Veracity UNCERTAINTY OF DATA

By 2014, it's anticipated there will be **420 MILLION WEARABLE, WIRELESS HEALTH MONITORS**

**4 BILLION+ HOURS OF VIDEO** are watched on YouTube each month



**400 MILLION TWEETS** are sent per day by about 200 million monthly active users

Fool data quality costs the US economy around

**\$3.1 TRILLION A YEAR**

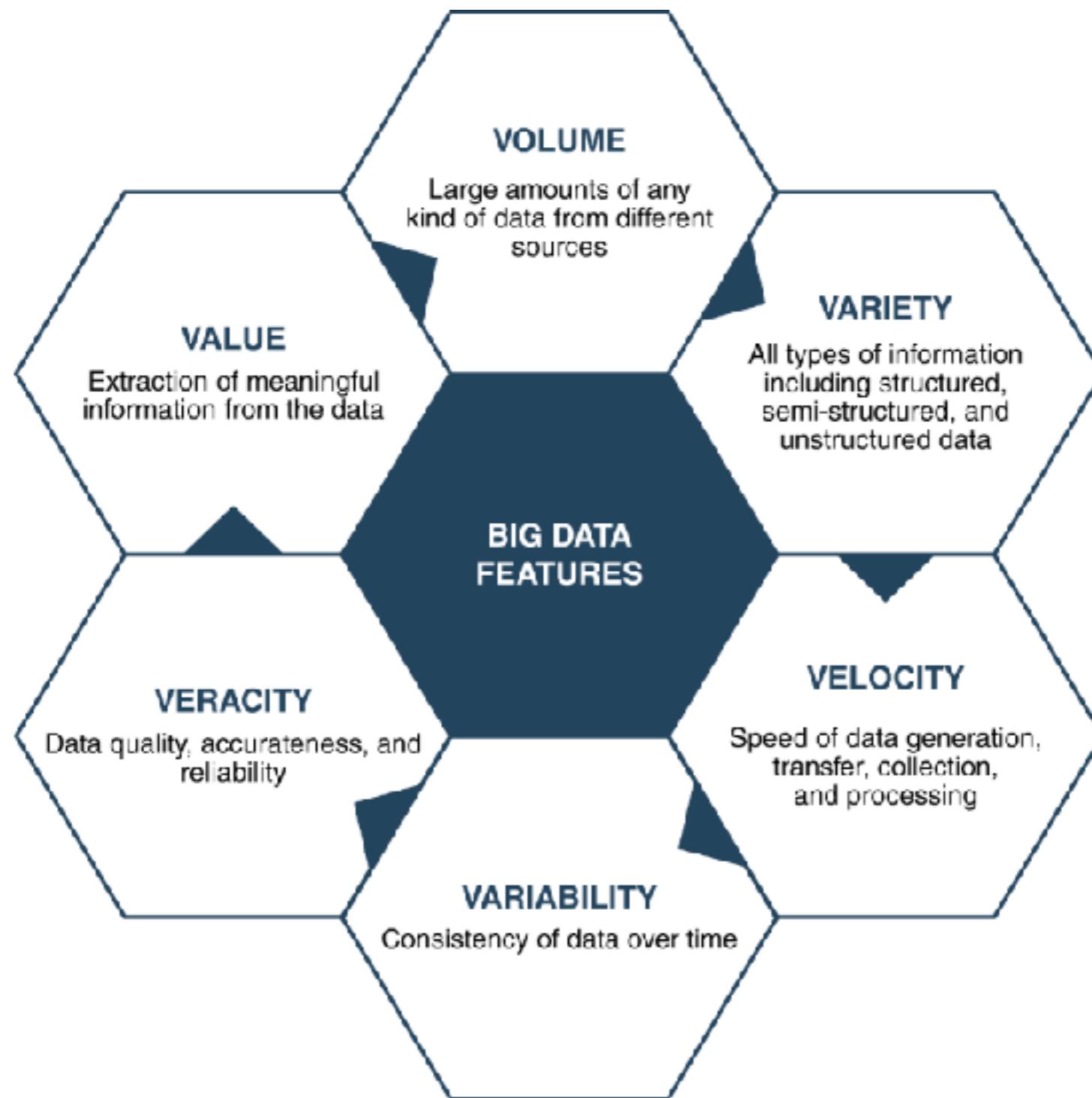


**IBM**

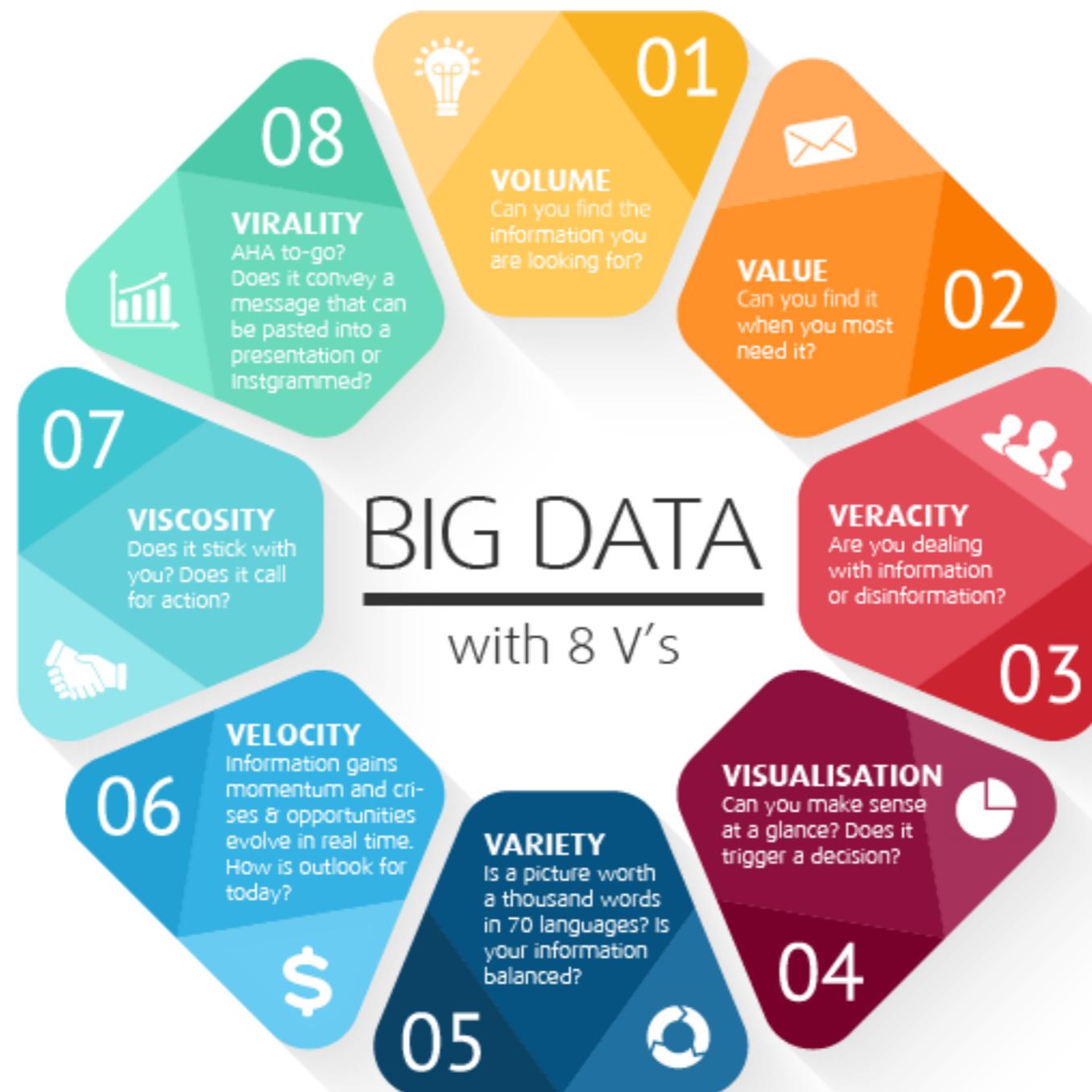
Sources: McKinsey Global Institute, Twitter, Cisco, Barne, EMC, SAS, IBM, MPTEC, QAS



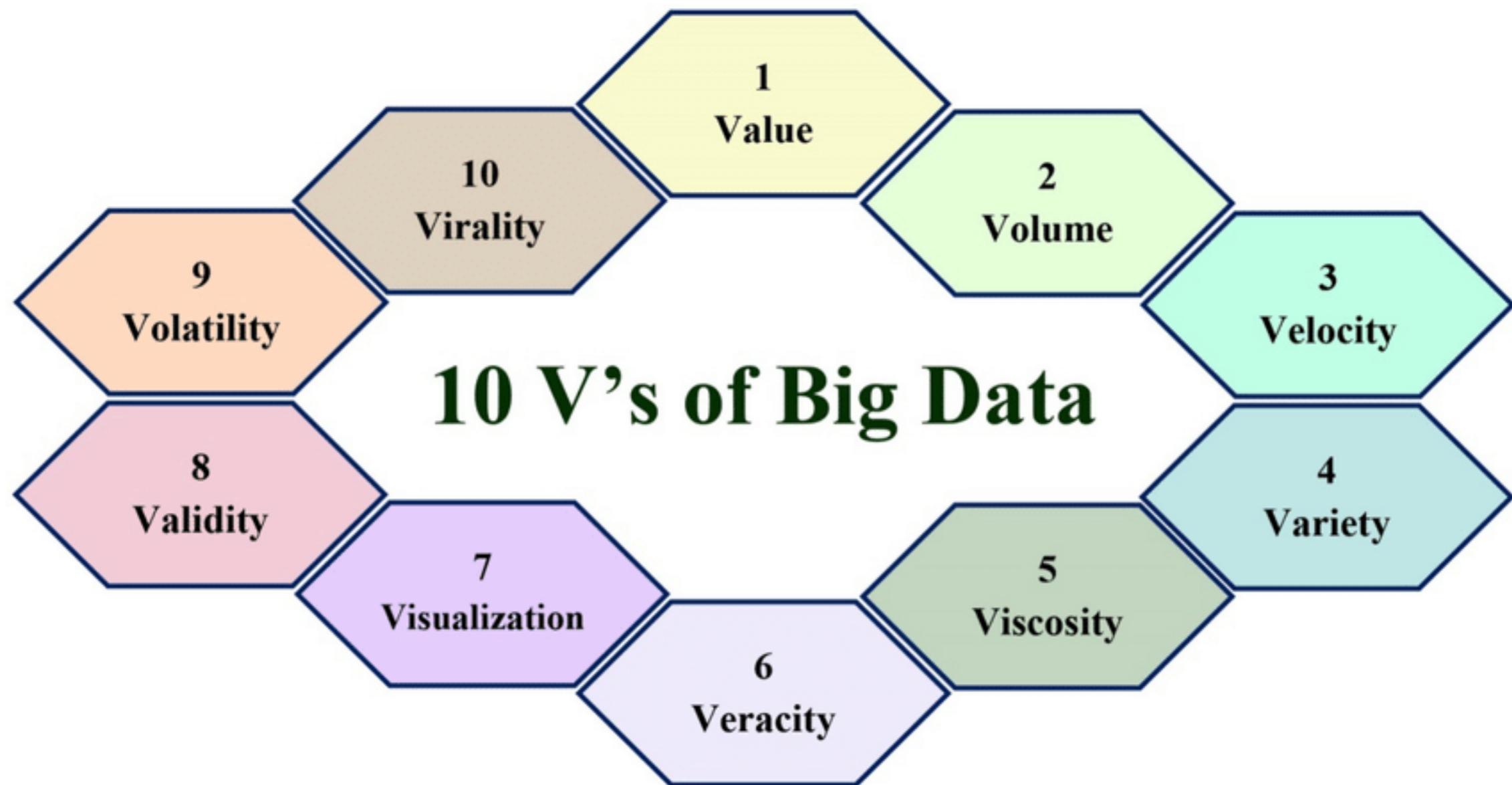
# Characteristics for Big Data



# Characteristics for Big Data



# Characteristics for Big Data

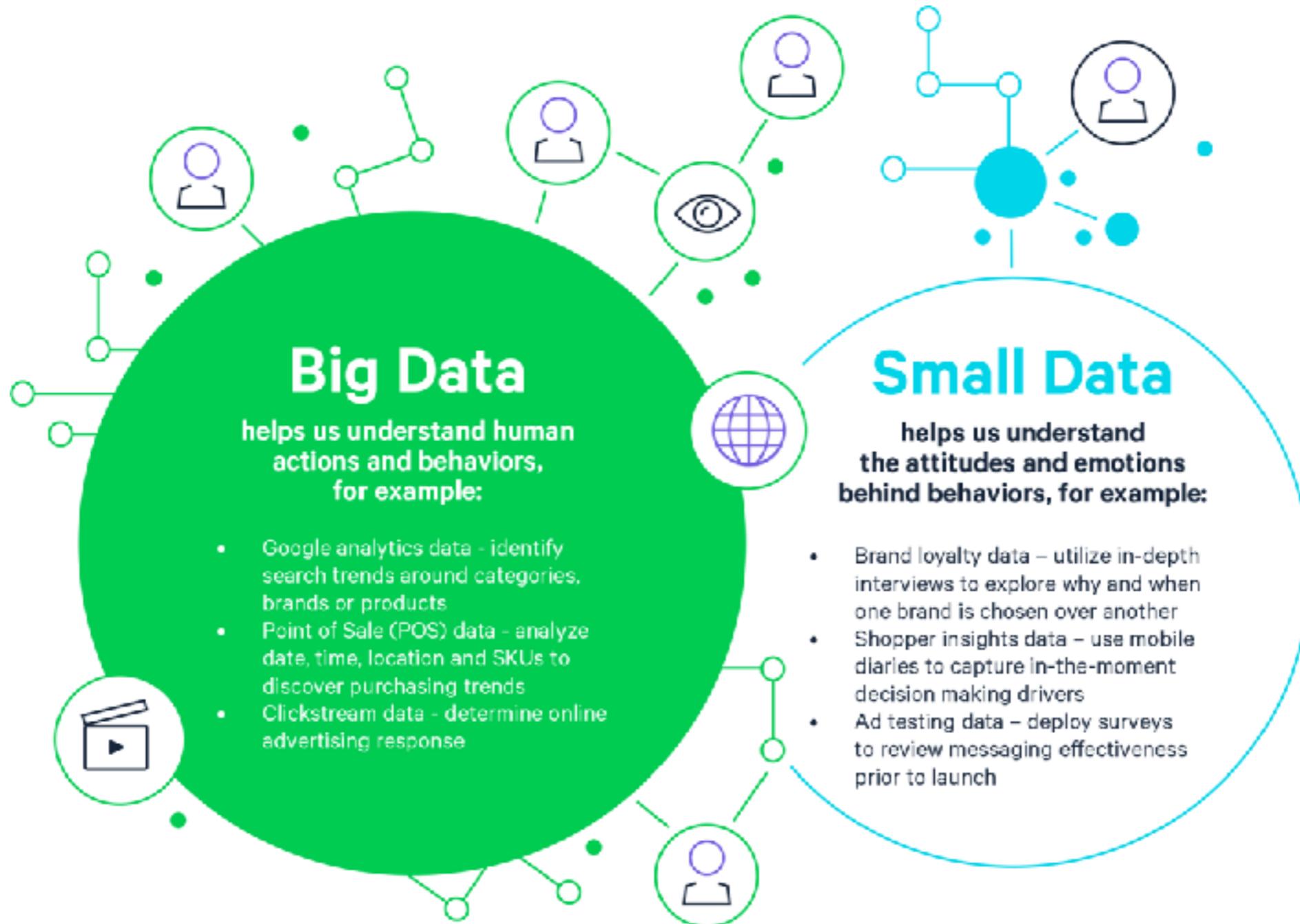


# Purpose of Big Data

To extract **meaningful** information from massive datasets that conventional tools can't manage, using **advanced analytics** and data processing methods.



# Big Data vs Small Data



# **Characteristics for Big Data ?**

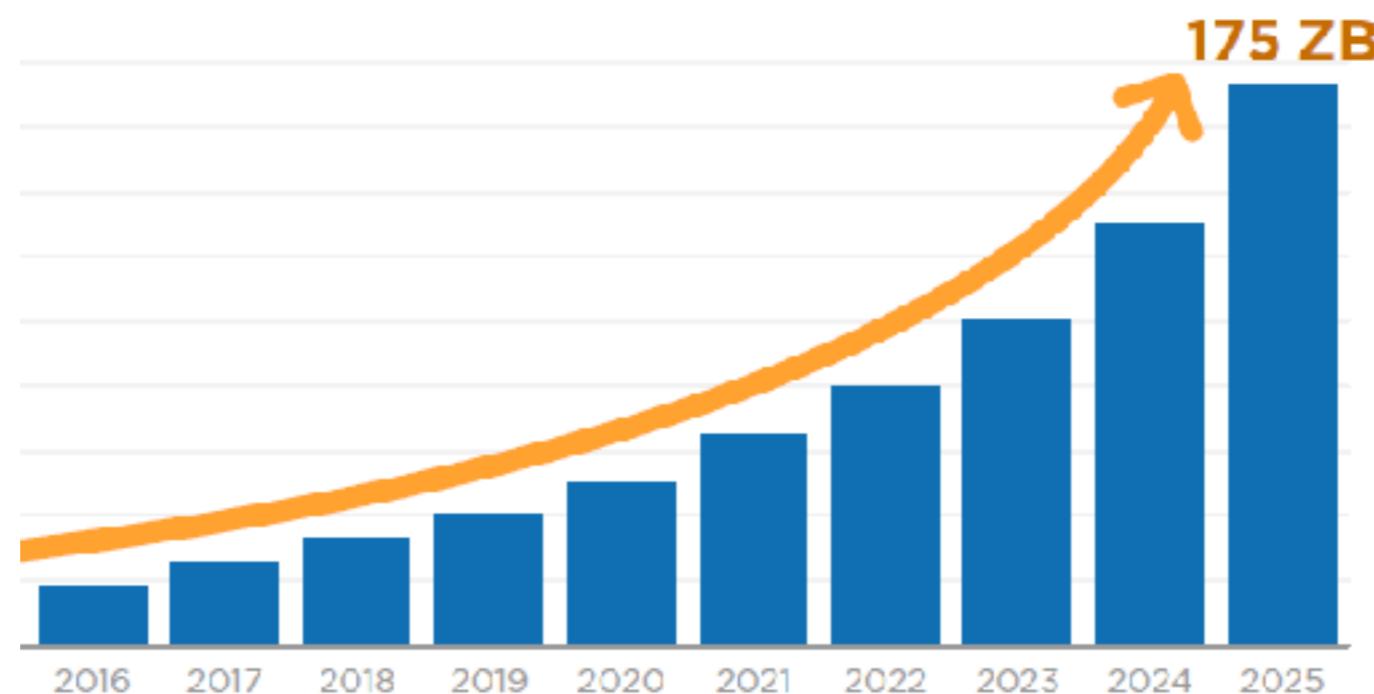


# Volume

Scale of data

Online and offline transactions

Save data in records and tables

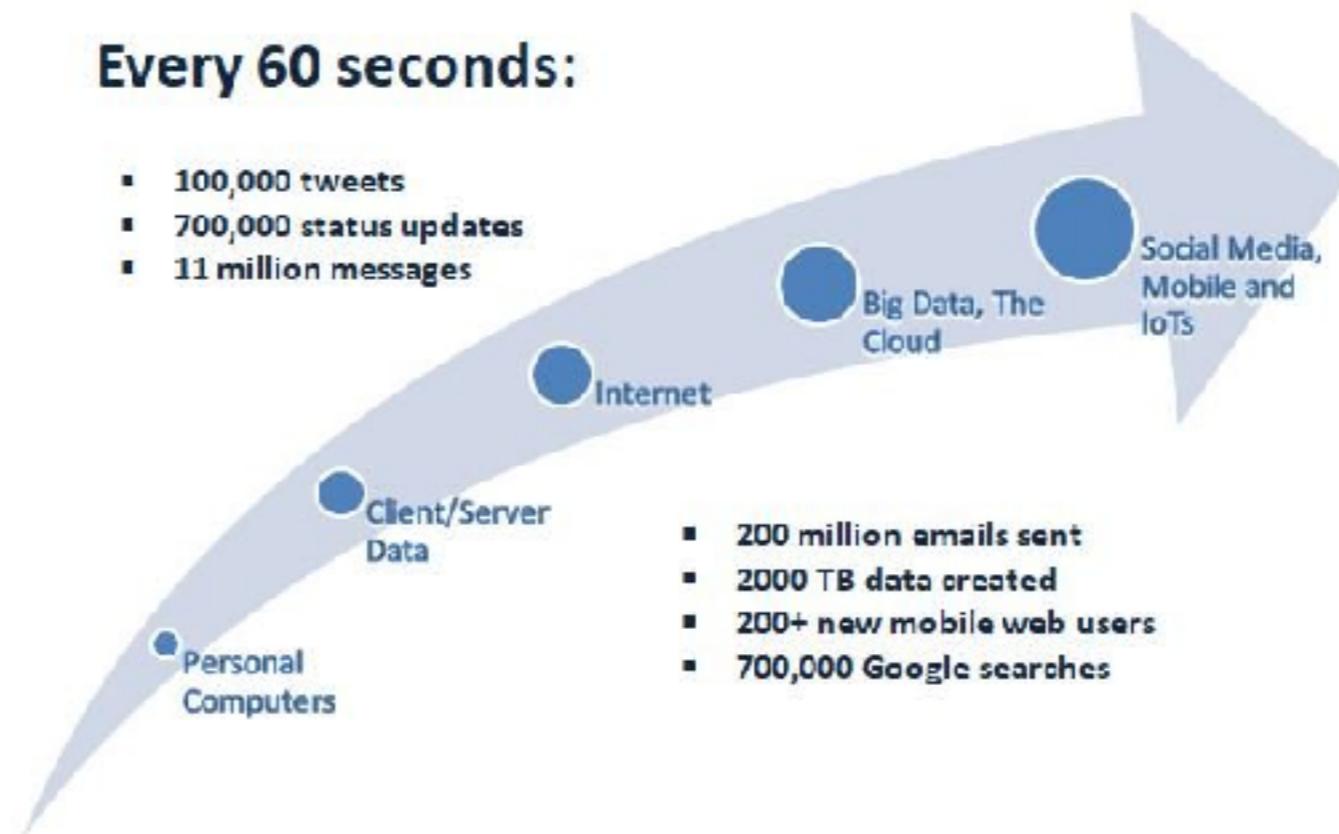


# Velocity

Speed of generating data  
Generated data in realtime

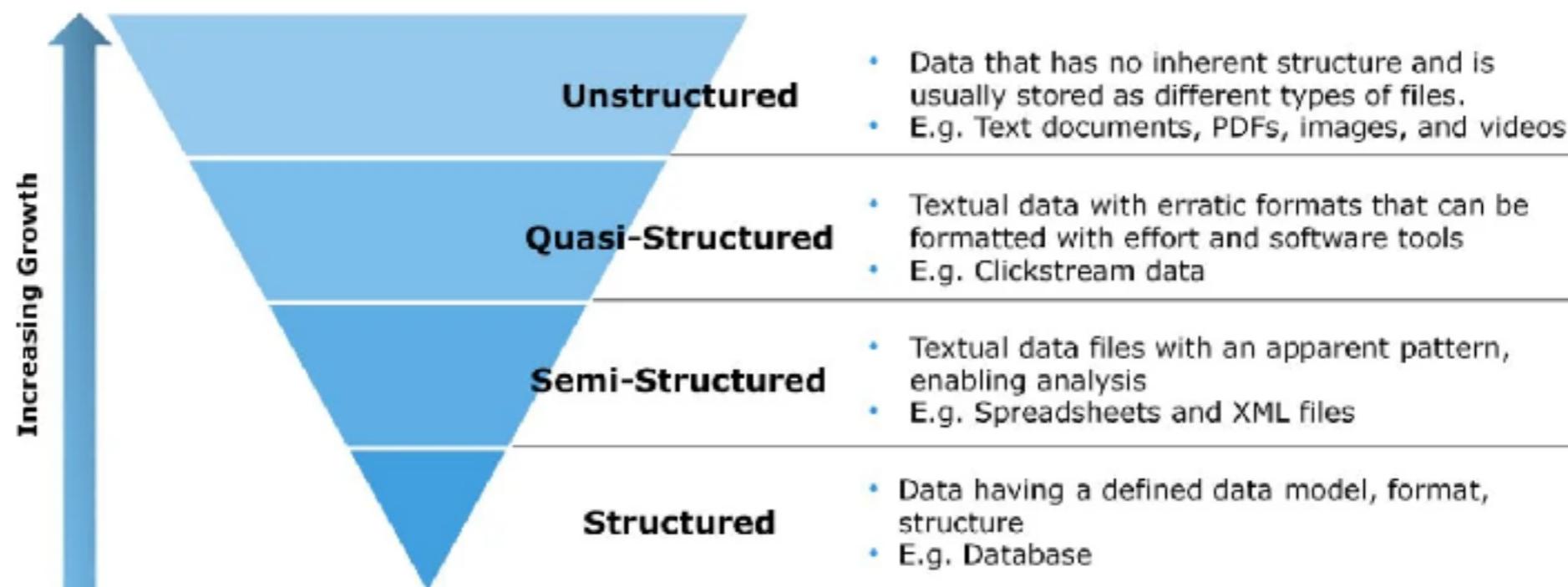
## Every 60 seconds:

- 100,000 tweets
- 700,000 status updates
- 11 million messages

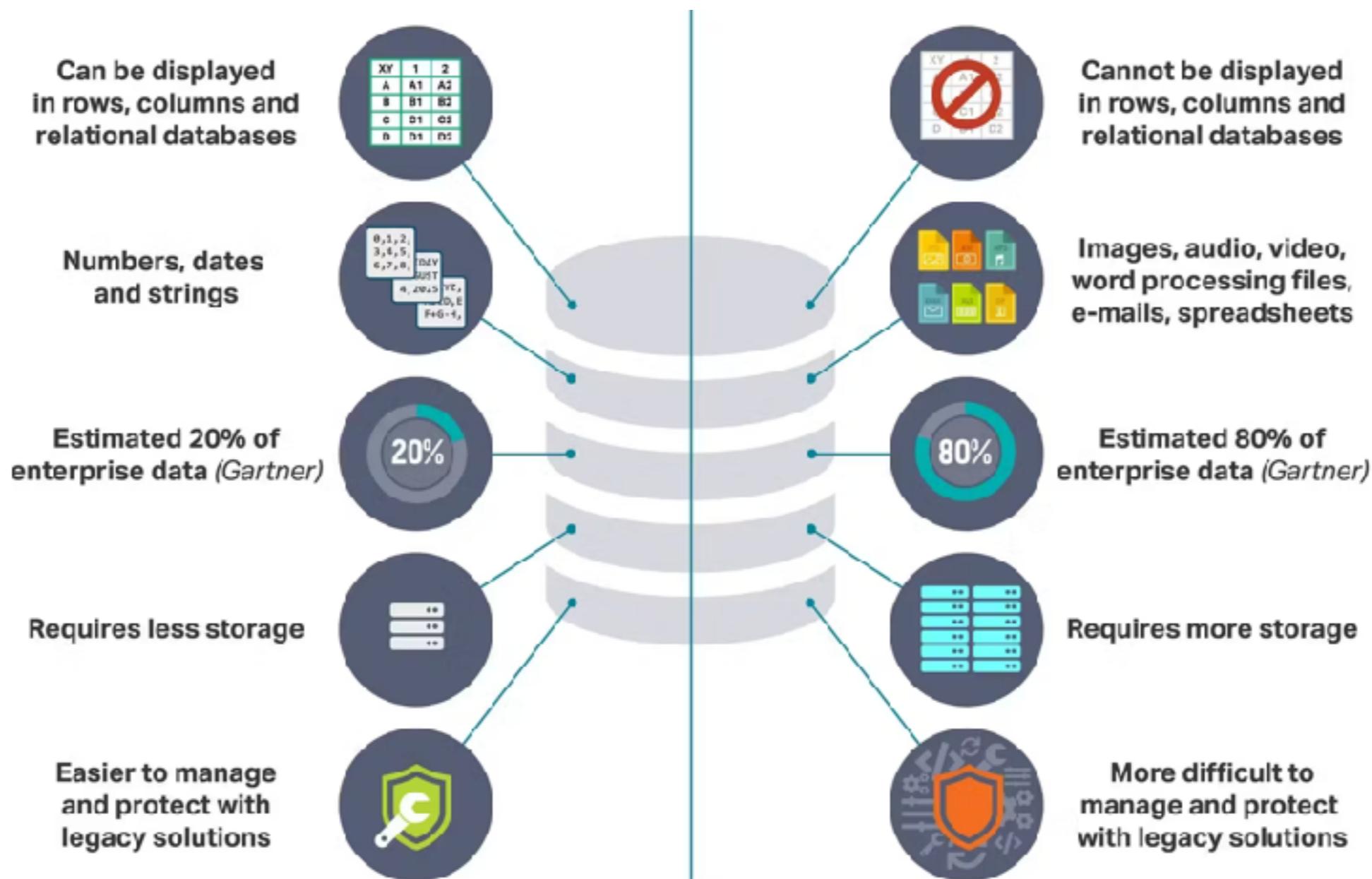


# Variety

## Diversity of data Structure, Unstructured and Semi-structure data



# Structured vs Unstructured data



# Types of data sources

Feature	Structured Data	Unstructured Data	Semi-structured Data
Definition	Data organized in a predefined format with a clear structure	Data lacking a predefined format and organization	Data with some internal structure but lacking a rigid schema
Examples	Relational databases, spreadsheets, CSVs, APIs	Text documents, emails, images, videos, social media posts	JSON files, XML files, log files
Characteristics	Standardized format, easily searchable and analyzable, consistent and reliable	Diverse formats, challenging to process and analyze, rich and diverse information	Flexible format, adaptable to evolving data needs, requires specialized tools
Advantages	Easy to process and analyze, supports efficient data retrieval, suitable for statistical analysis	Rich and diverse information, captures real-world context, valuable for sentiment analysis and trend identification	Adaptable to evolving data needs, flexible and scalable, suitable for real-time applications
Disadvantages	Limited flexibility, unable to capture complex relationships, not suitable for all types of data	Difficult to process and analyze, requires specialized tools, data quality concerns	Limited data integration potential, lack of standardized formats, evolving data structures
Use Cases	Business intelligence, financial transactions, scientific research, data warehousing	Customer feedback analysis, social media monitoring, content analysis, multimedia processing	Real-time analytics, sensor data analysis, web scraping, scientific experiments
Tools and Techniques	SQL databases, spreadsheets, data warehouses, data analytics tools	Natural language processing (NLP), machine learning, sentiment analysis, image recognition	JSON parsers, XML parsers, stream processing tools, data pipelines



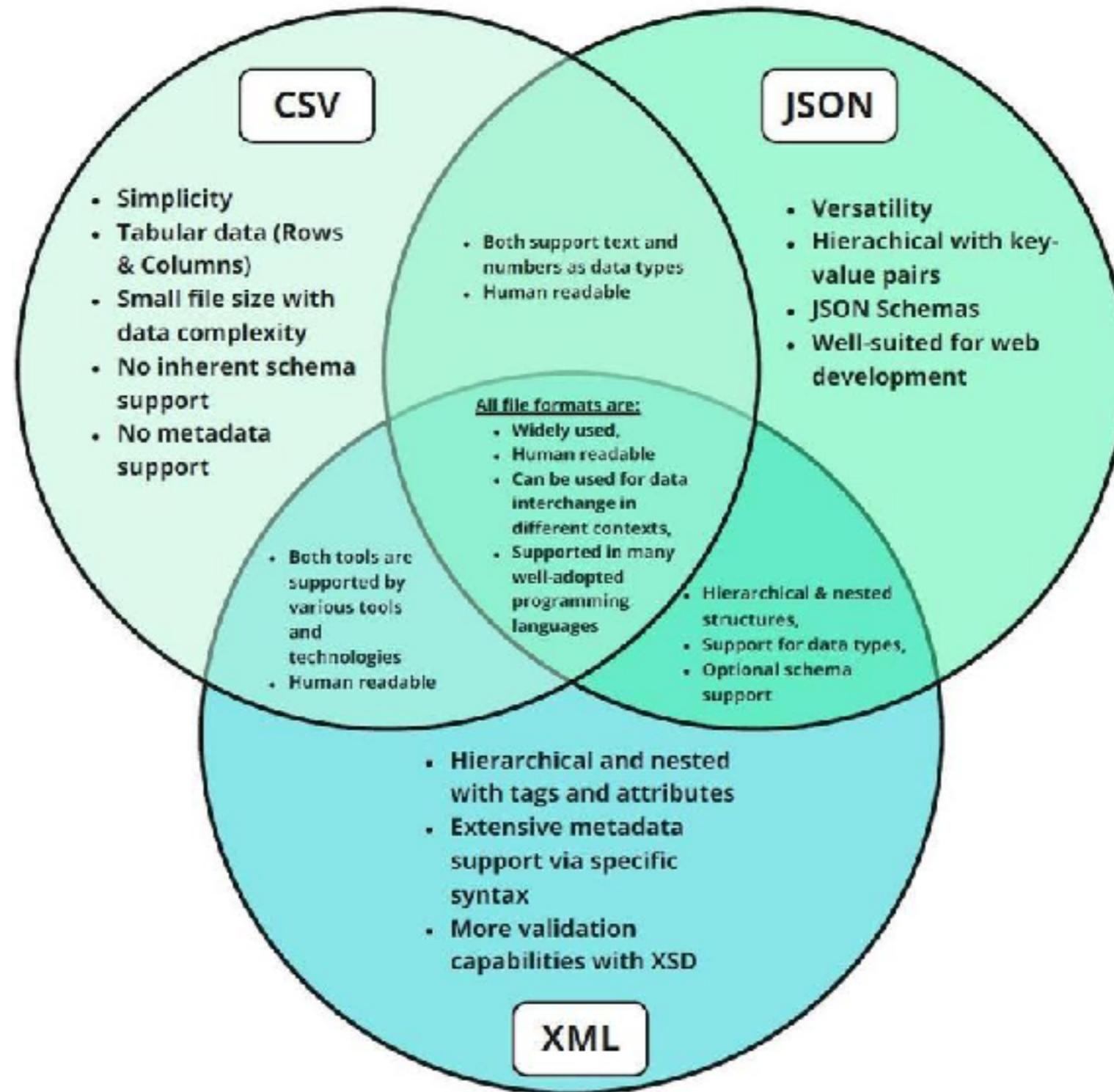
# File formats in Big Data ?

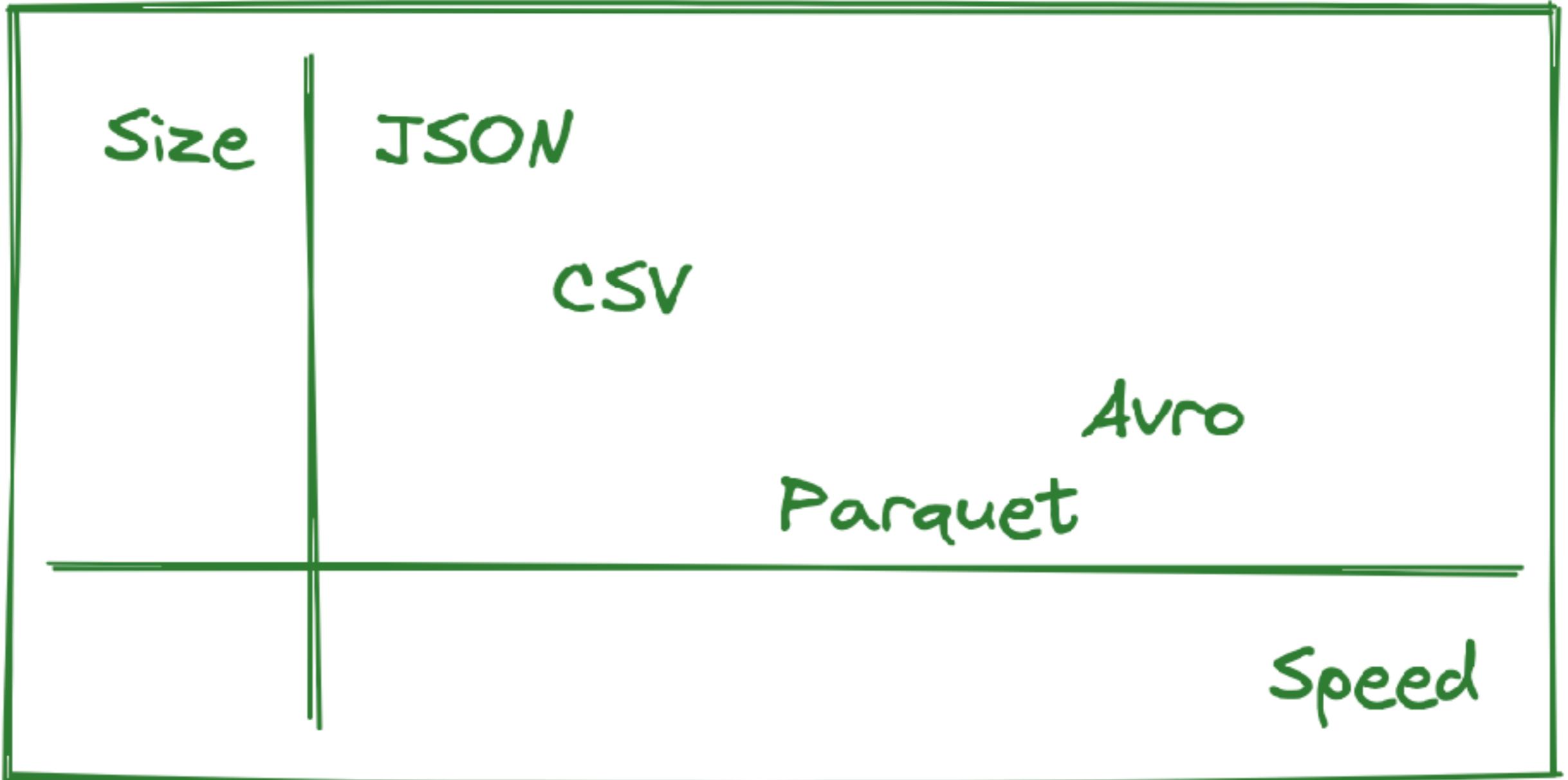
CSV/Excel

XML/JSON

Parquet







<https://datacrump.com/csv-parquet-json-avro/>



# Apache Parquet

Column storages  
High Compression  
Schema enforcement

Predicated pushdown (filter data in storage layer)

Designed for Big data analytic

<https://parquet.apache.org/>



# Veracity

Trustworthiness of data  
Inconsistency and uncertainty in data  
Data messy, low quality and low accuracy !!

Enhance decision making

Effective problem-solving

Data-driven innovation

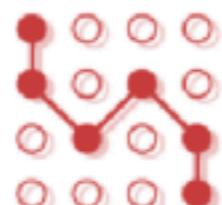
Reduce risk !!



# Sources of Data Veracity



Statistical biases



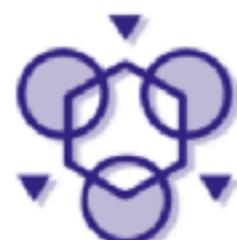
Lack of data lineage



Software bugs



Noise



Abnormalities



Information Security



Untrustworthy data sources



Falsification



Uncertainty and ambiguity of data



Duplication of data



Out of date and obsolete data



Human error

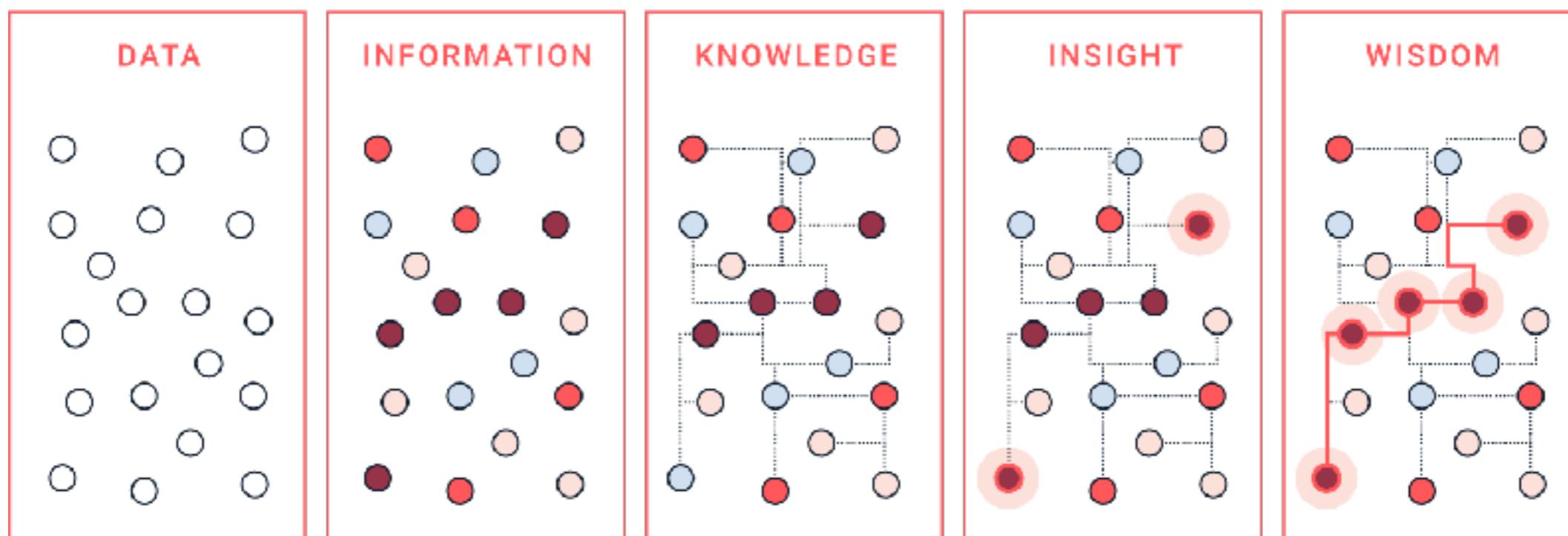
<https://datafloq.com/read/data-veracity-new-key-big-data/>



# Value

Usefulness of data

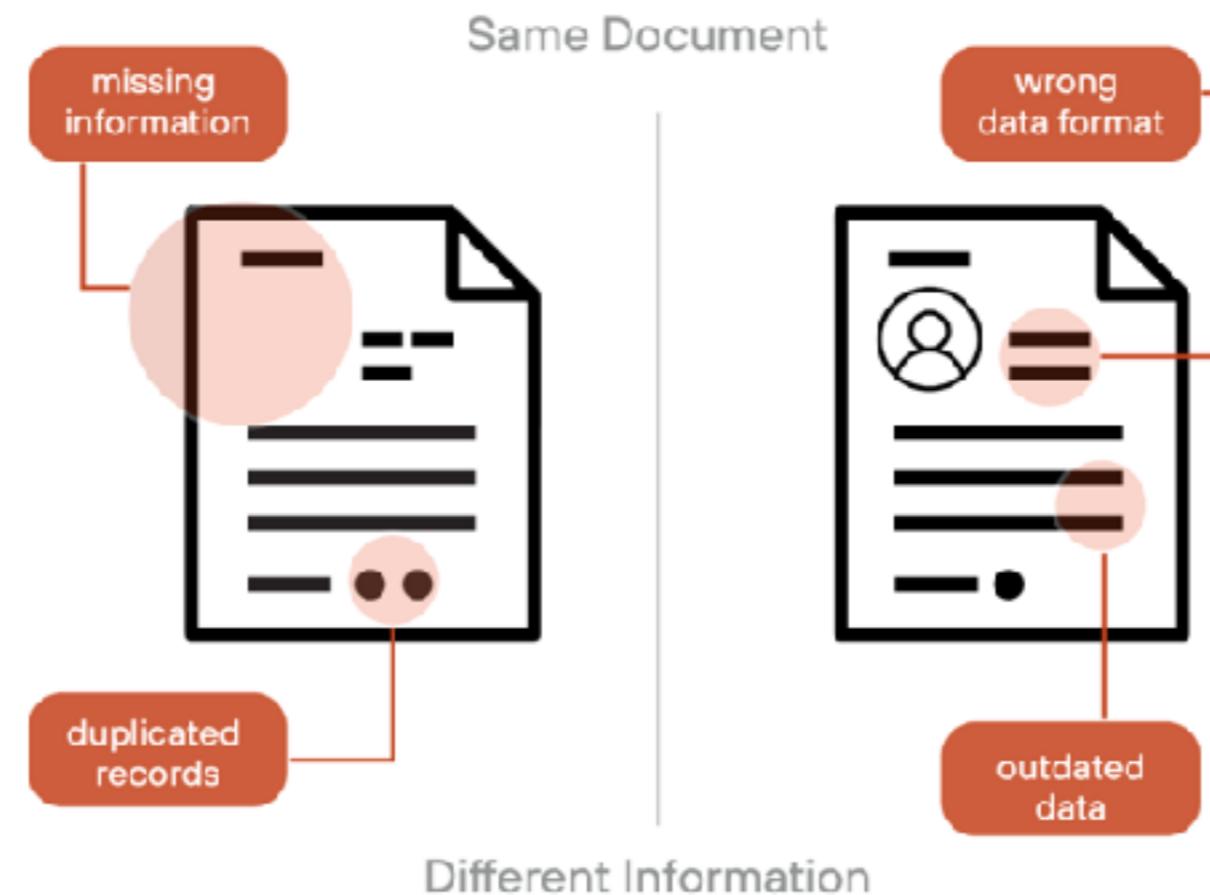
Data must be extracted, transforms to useful data



# Variability

Inconsistency of data

Data changes overtime, rate of data, format, meaning



# Data Quality ?



# **Introduction to Business Analytics (BA)**



# **Business Analytics (BA)**

**Practice** of using data, statistical analysis, and quantitative methods to drive **decision-making** and **improve business outcomes**

It often involves predictive modeling, data mining, and machine learning to **forecast trends** and derive **insights**.



# Purpose of BA

The goal of BA is  
to help businesses make informed,  
data-driven decisions by predicting  
future outcomes or trends



# Key Techniques

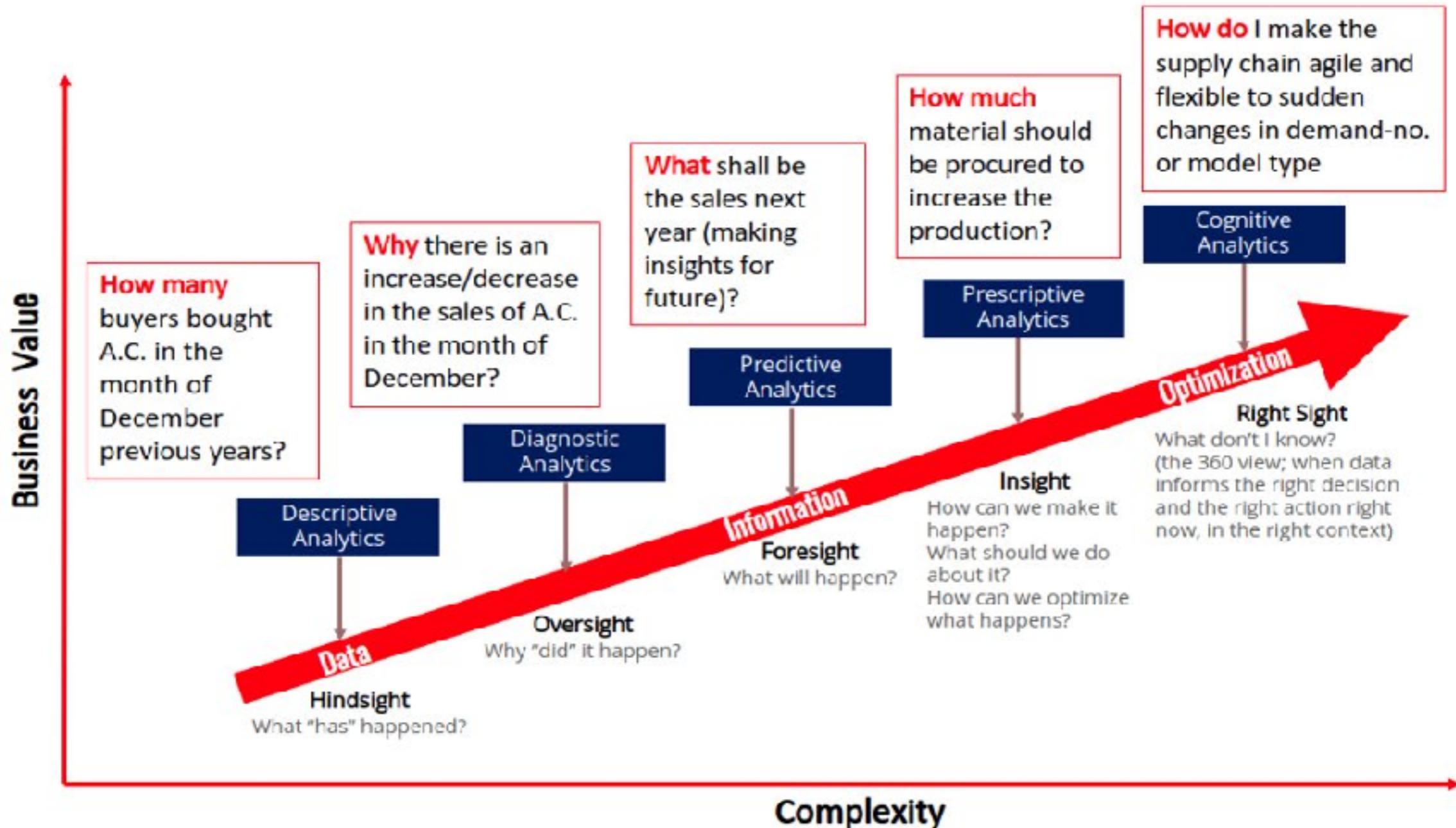
Predictive  
Analytics

Prescriptive  
Analytics

Diagnostic  
Analytics



# Key Techniques



# **Descriptive Analytics**

Descriptive analytics answers the question,

**“What happened?”**

by summarizing historical data to identify patterns  
and trends.



# Descriptive Analytics

It helps businesses understand **past performance** and gain insights from historical data to determine what has occurred over a given period.

Basic data aggregation

Reporting tool

Data visualisation tool

Statistic and trend



# Diagnostic Analytics

Diagnostic analytics answers the question,

**“Why did it happen?”**

by analyzing data to understand  
the root causes of past events.



# Diagnostic Analytics

It digs deeper into descriptive analytics to find the reasons behind specific outcomes or patterns.

Drill-down  
analytics

Correlation  
analysis

Root cause  
analysis

Statistic with  
regression analysis



# Predictive Analytics

Predictive analytics answers the question,

**“What will likely happen in the future?”**

by using historical data, statistical models, and machine learning techniques to forecast future trends.



# Predictive Analytics

It helps businesses anticipate **future outcomes** based on patterns and trends found in historical data.

Statistic  
modeling

Machine learning  
algorithm

Time series  
forecast



# Prescriptive Analytics

Prescriptive analytics answers the question,

**“What should be done?”**

by recommending specific actions  
or strategies to achieve desired outcomes or  
optimize processes.



# Prescriptive Analytics

It goes beyond predicting future outcomes by suggesting the best course of action based on predictions.

Optimization  
model

Simulation

Decision/scenario  
analysis



# Summary

Type	Main Question	Purpose	Techniques	Example
<b>Descriptive Analytics</b>	What happened?	Understanding past performance.	Data aggregation, reporting, visualization	Sales reports showing performance by region.
<b>Diagnostic Analytics</b>	Why did it happen?	Identifying causes of past outcomes.	Drill-down analysis, correlation, root cause	Analyzing why sales declined in a certain area.
<b>Predictive Analytics</b>	What will happen?	Forecasting future trends and outcomes.	Machine learning, statistical models	Predicting customer churn for an online store.
<b>Prescriptive Analytics</b>	What should we do?	Recommending actions to optimize future outcomes.	Optimization, simulation, decision models	Suggesting optimal pricing strategies for sales.



# Summary

## **Descriptive Analytics**

focuses on summarizing what has already happened

## **Diagnostic Analytics**

helps businesses understand why those things happened

## **Predictive Analytics**

forecasts future trends based on historical data

## **Prescriptive Analytics**

provides recommendations to optimize decision-making and future outcomes



# **Introduction to Business Intelligence (BI)**



# Business Intelligence (BI)

Refers to technologies, applications, and practices used for the collection, integration, analysis, and presentation of business data.

Collect

Integrate

Analysis

Visualize



# Purpose of BI

BI is about **reporting and dashboards** that help businesses monitor their current state and past performance by organizing data into actionable insights.



# **Basic of Data Management**



# Data management

Data management refers to the **process** of organizing, storing, maintaining, and retrieving data efficiently and securely.

It plays a crucial role in **ensuring** that data is accurate, available, and usable for decision-making, analysis, and operations



# Keys of data management

Data collection

Data storage

Data Security

Data quality

Data governance

Data Integration

Data backup/  
restore

Data  
accessibility

Data analysis



# Data Storage ?

Database

Data  
Warehouse

Data Mart

Data Lake



# Data Warehouse

A data warehouse is a centralized repository that stores **structured** and processed data from **multiple sources**.

It is optimized for querying and analysis, primarily used for reporting, business intelligence, and decision-making.



# Data Mart

A data mart is a smaller, **more focused subset of a data warehouse**, designed to serve the needs of a **specific business unit or department** (e.g., sales, marketing).

It contains only relevant data for that particular group, **enabling quicker access and more tailored analysis**.



# Data Lake

A data lake is a **vast storage** repository that holds a large amount of raw data in its **native format** (structured, semi-structured, and unstructured) until it is needed for analysis.

It is designed for **storing data** before it is processed and transformed for analysis, offering flexibility and scalability.

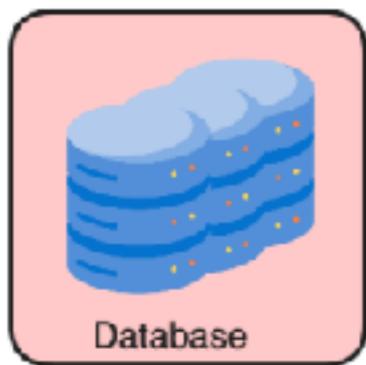


# Key differences

Feature	Data Warehouse	Data Mart	Data Lake
<b>Data Type</b>	Structured	Structured	Structured, semi-structured, and unstructured
<b>Scope</b>	Enterprise-wide (broad)	Department-specific (narrow)	Enterprise-wide (but raw)
<b>Storage Type</b>	Structured (rows/columns)	Structured	Raw format
<b>Data Processing</b>	Pre-processed (ETL)	Pre-processed (ETL)	Raw, processed when queried
<b>Primary Use</b>	Reporting, historical analysis	Department-specific insights	Big data analysis, machine learning
<b>Speed of Query</b>	Fast for complex queries	Faster for localized queries	Slower for complex queries
<b>Cost</b>	Generally more expensive	Less expensive than data warehouse	Cheaper for storage, but processing may incur costs
<b>Schema</b>	Schema-on-write	Schema-on-write	Schema-on-read



# Different between data types



VS



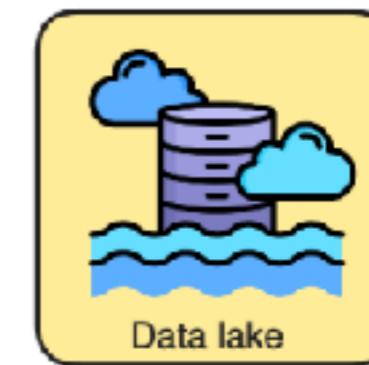
Data Warehouse

VS



Data mart

VS



Data lake

## Scope

Application-specific

Organization-wide,  
structured data.

Department-specific,  
structured data.

Organization-wide,  
any type of data

## Data Type

Structured

Structured

Structured

Structured,  
semi-structured,  
unstructured.

## Structure

Predefined schema

Schema on write

Schema on write (inherited  
from data warehouse)

Schema on read

## Use Case

Operational  
applications(OLTP)

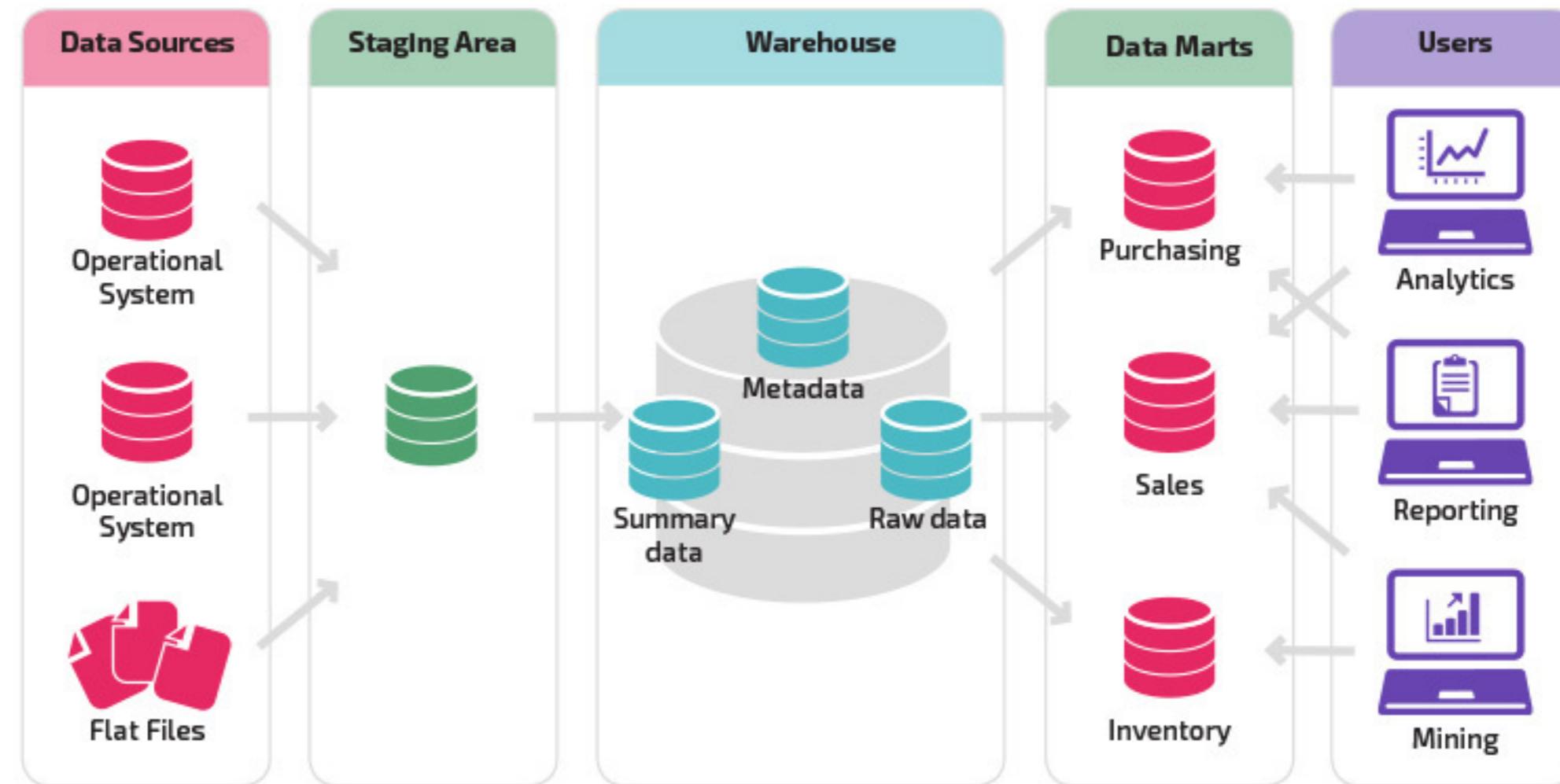
Business intelligence,  
historical  
analysis(OLAP).

Specific business function  
analysis

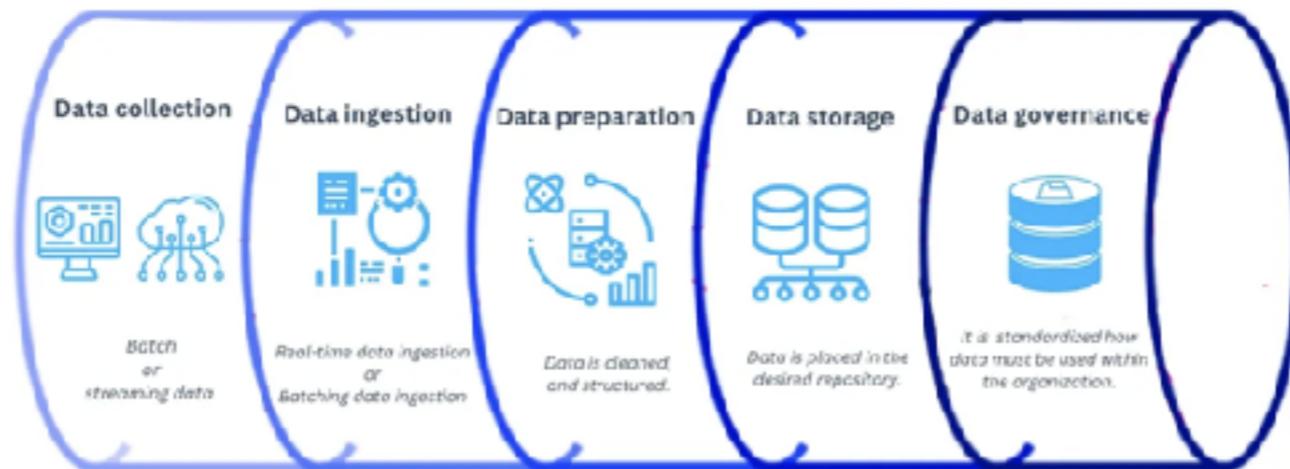
Big data analytics,  
data exploration.



# Data Warehouse vs Data Mart



# Data Pipeline

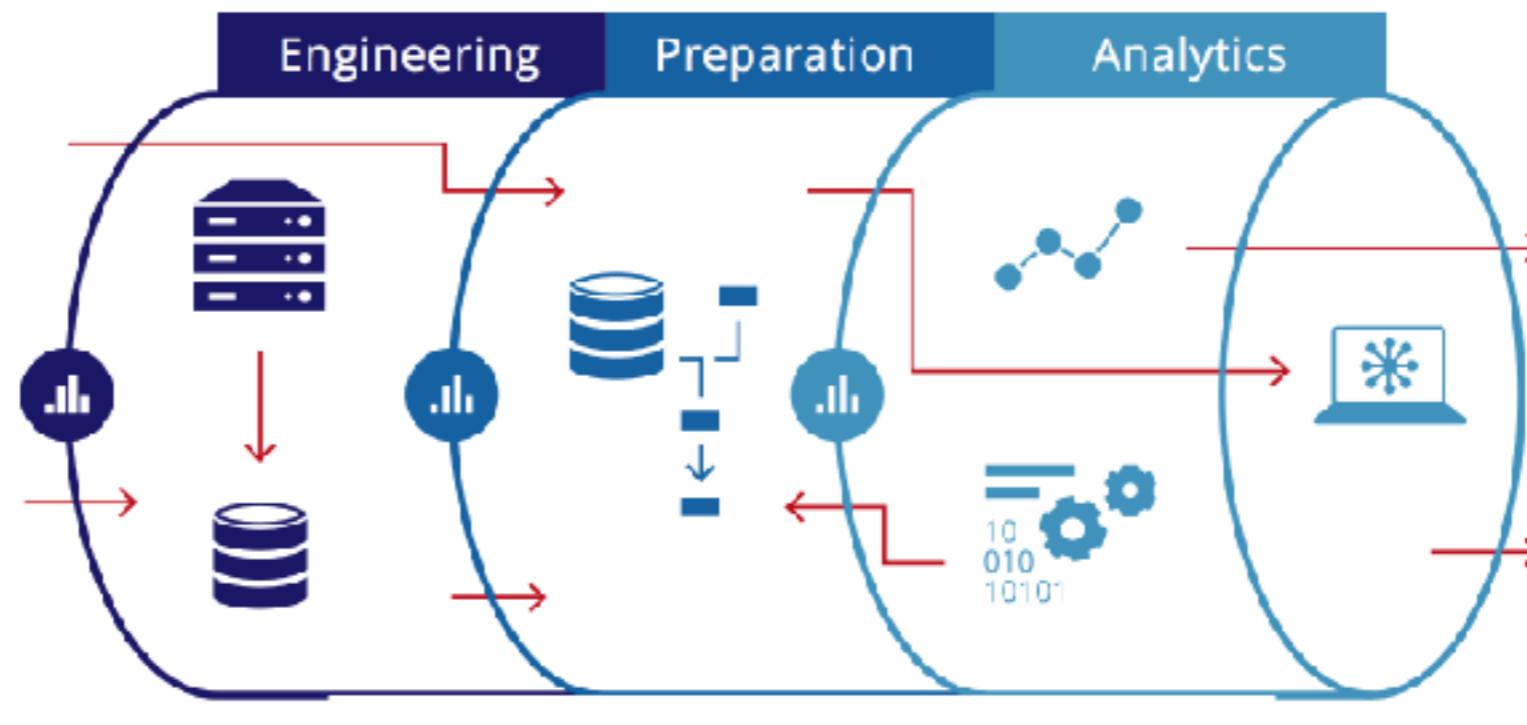


# Data pipeline

A **data pipeline** refers to a series of **processes** and **tools** used to **automate** the flow of data from source systems to destination systems, such as databases, data lakes, or data warehouses.



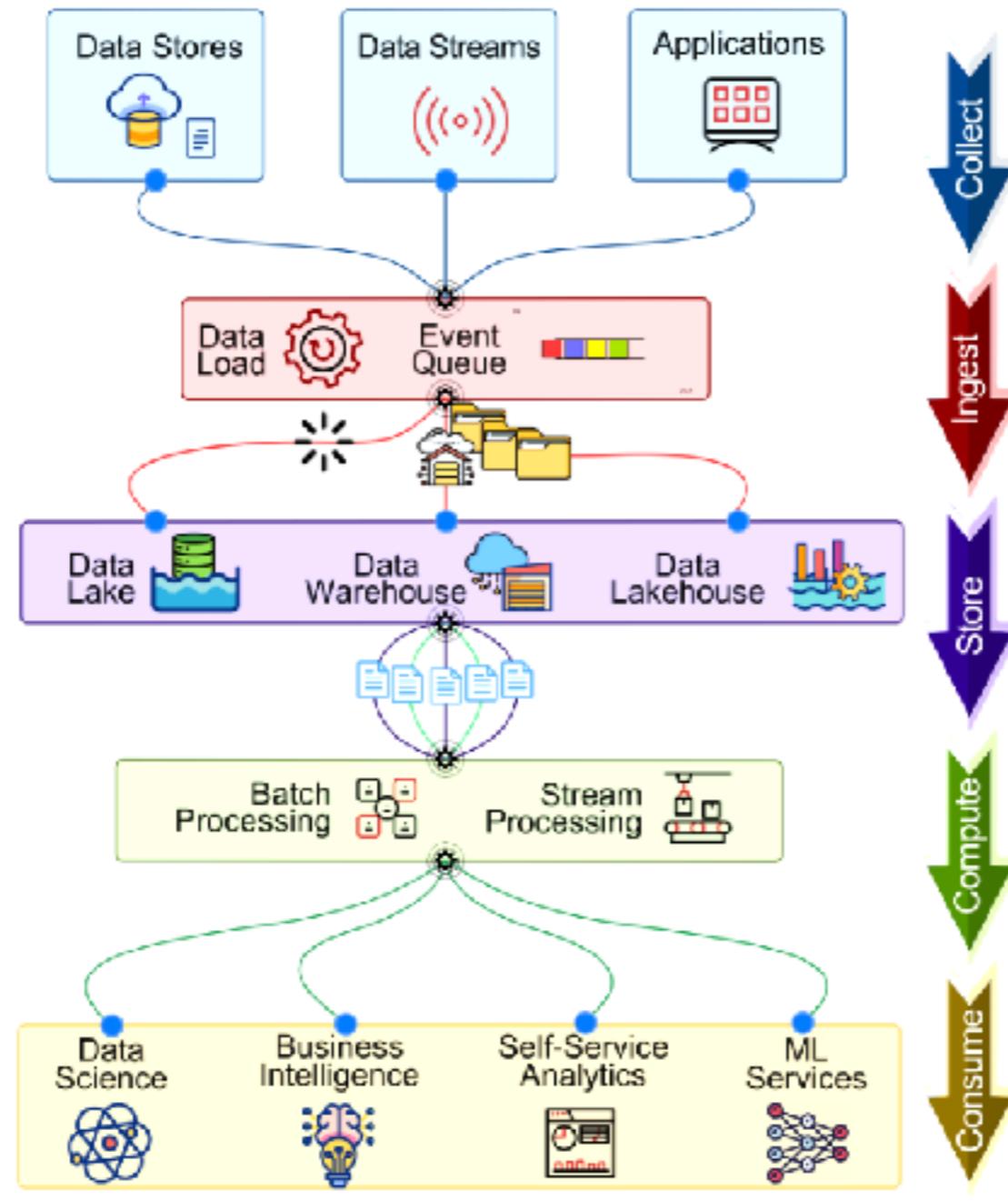
# Building robust data pipelines



Administration   Security   Lifecycle Management   Data Provenance   Dynamic Data Pipeline   Monitoring   Automation



# Data pipeline overview



<https://bytebytogo.com/guides/data-pipelines-overview/>



# Data Ingestion

The process of collecting data from various sources and bringing it into the pipeline.

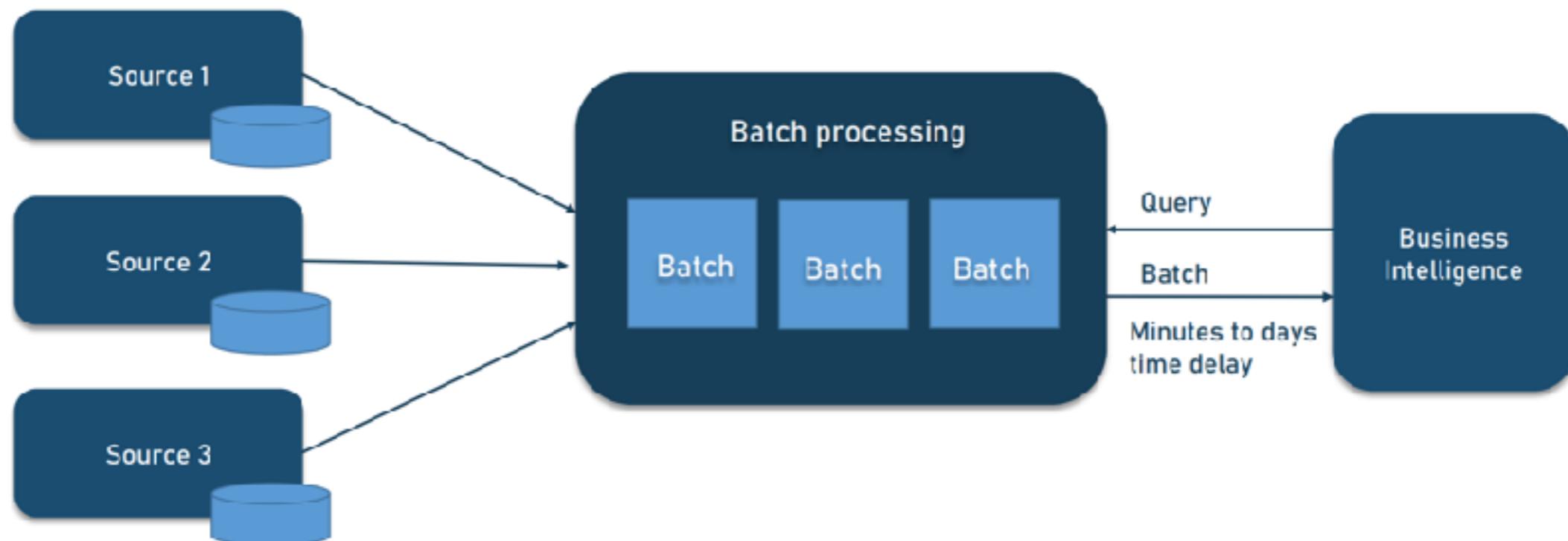
Batch  
processing

Real-time  
processing



# Batch processing pipeline

Working with large dataset at one time or in schedule

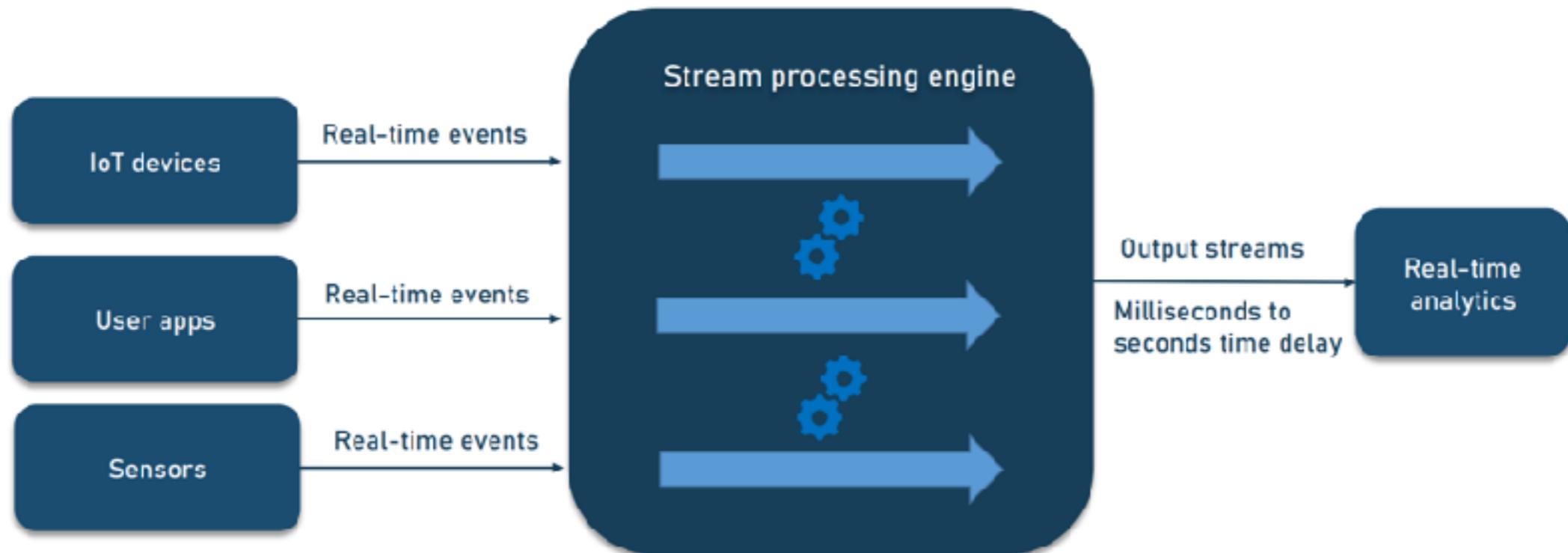


<https://www.altexsoft.com/blog/data-pipeline-components-and-types/>



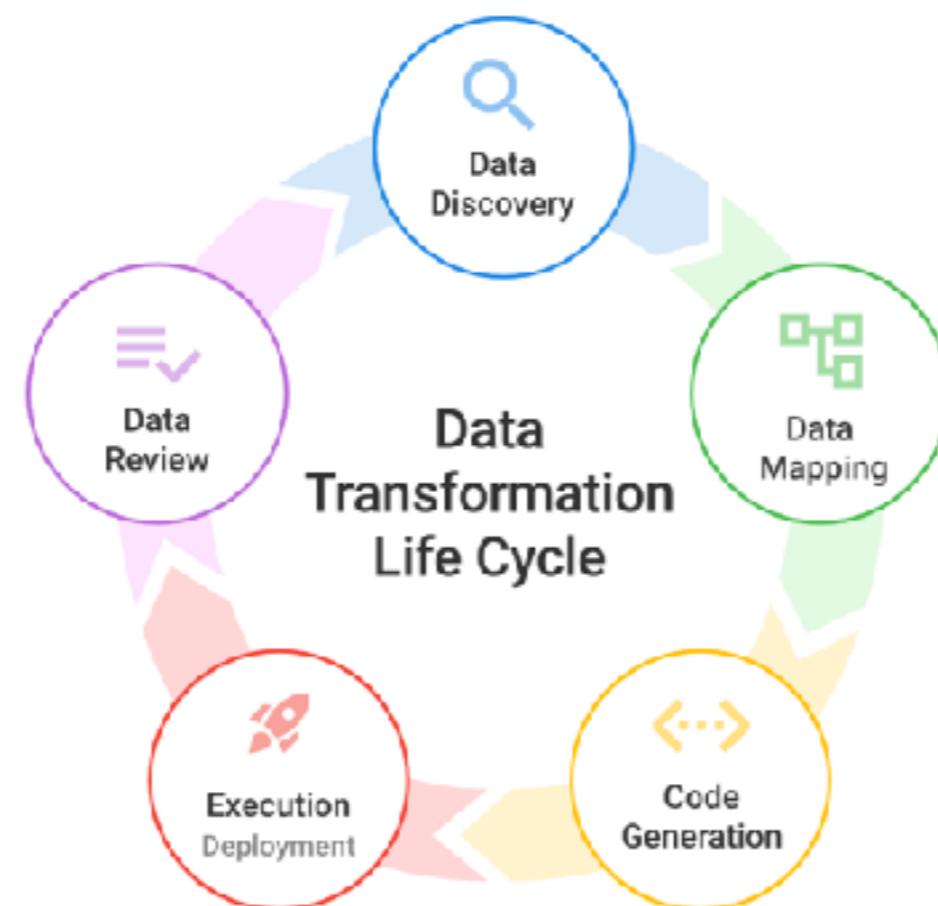
# Streaming data pipeline

Working with realtime data



# Data Transformation

This step involves cleaning, validating, and transforming the data into a standardized format or structure suitable for analysis or storage.



# How to cleaning data ?



# Key Techniques ?

Remove  
duplication

Handle missing  
data

Standardize  
Data Formatting

Remove  
Unwanted  
Characters

Convert Text to  
Columns

Data Validation

Find and  
Replace  
Inconsistent

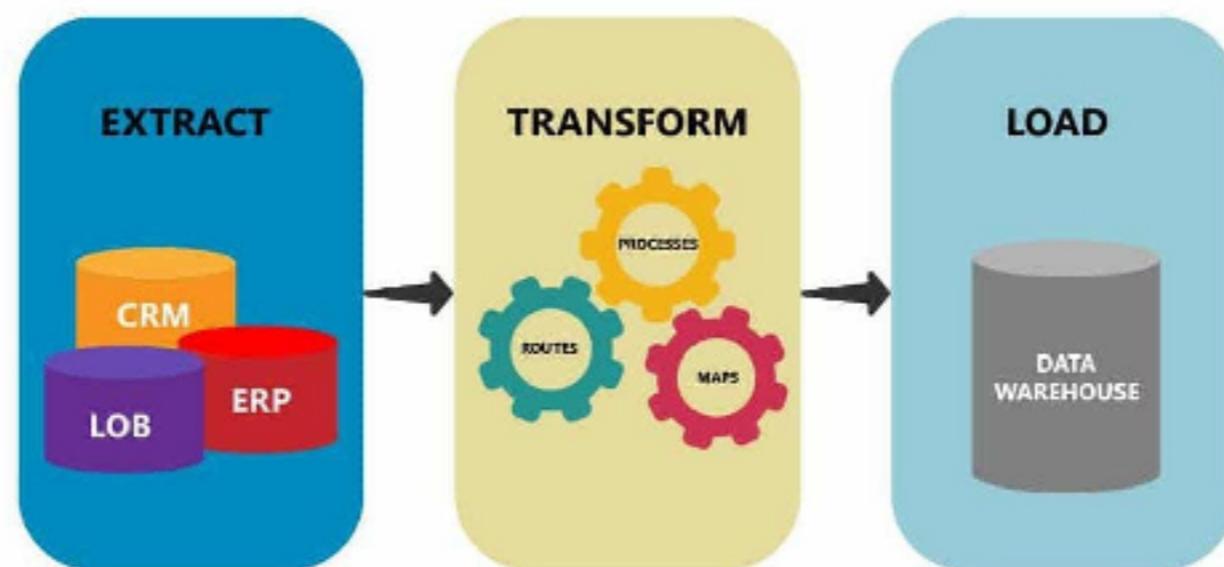
Fix Outliers

Data Cleansing



# Data Loading

The transformed data is loaded into a target destination, which could be a data warehouse, database, or cloud storage platform.

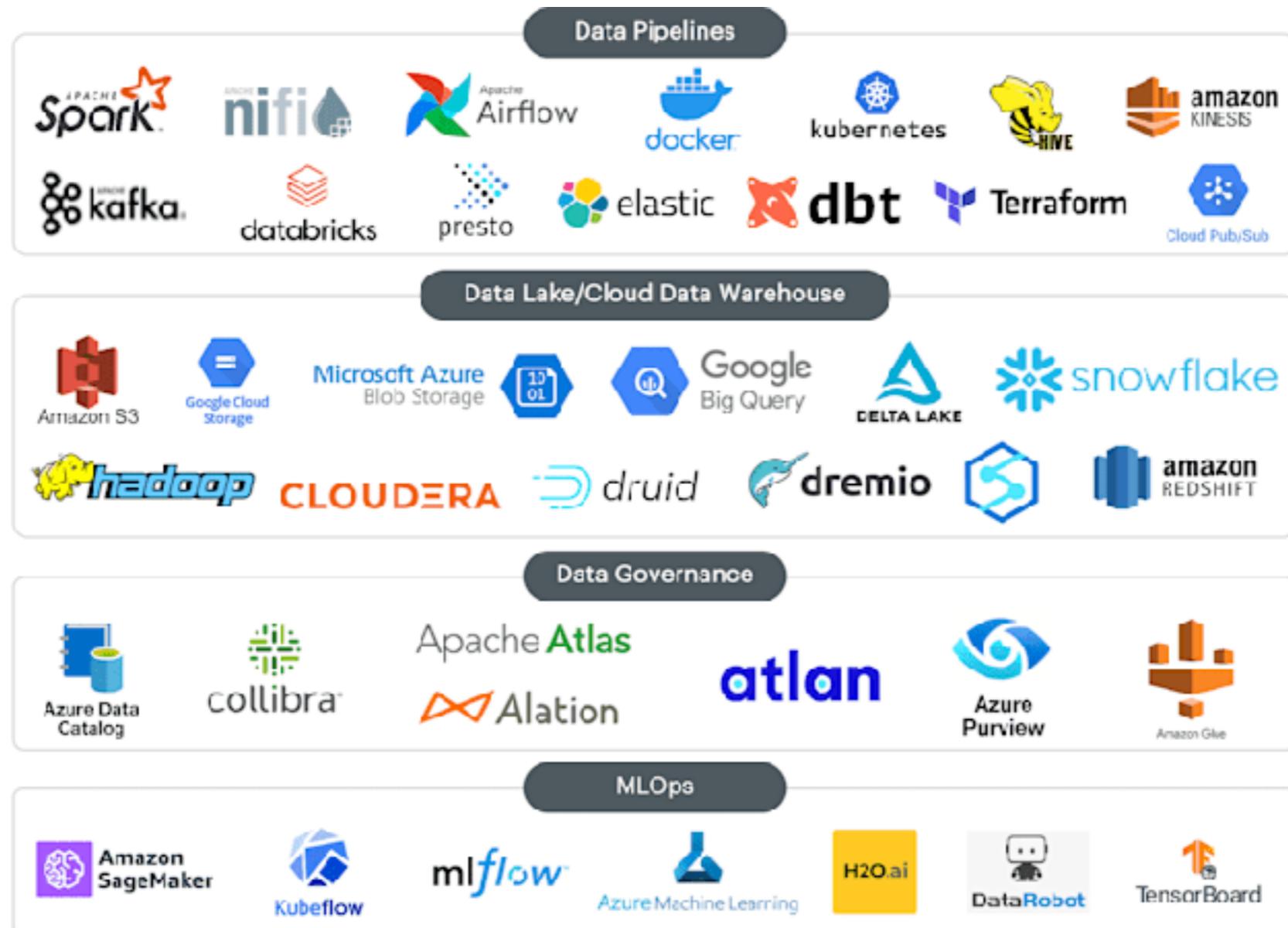


# Orchestration

This component **manages the overall workflow** and scheduling of the pipeline, ensuring the proper execution and coordination of data processing tasks.



# Tools



# Data Analysis Techniques



# Exploratory Data Analysis

Critical step in the **data science process** that involves summarizing the main characteristics of a dataset, often using visual methods, before applying more formal **statistical techniques**.

Summarize

Visualize

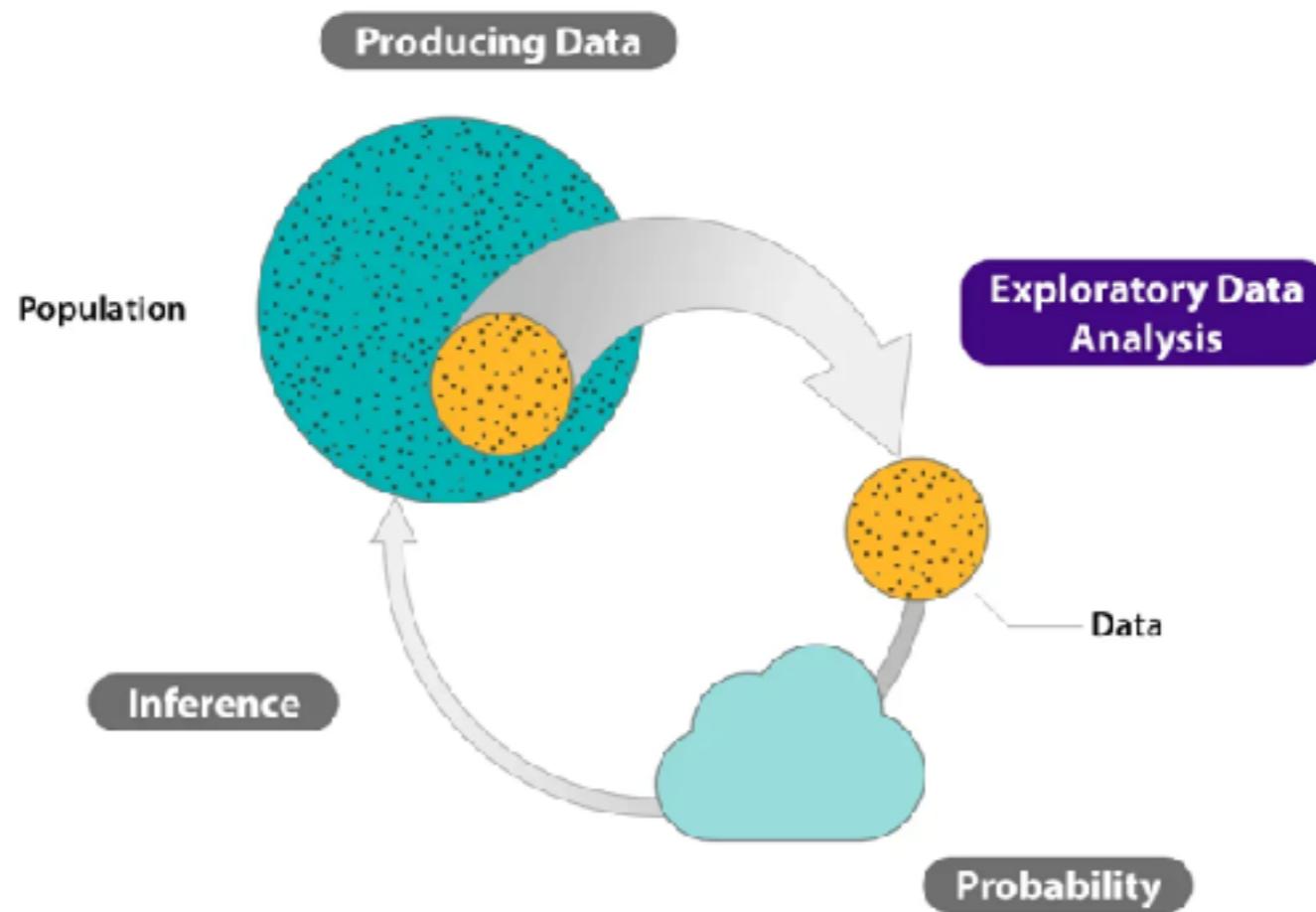
Identify

Missing data/values

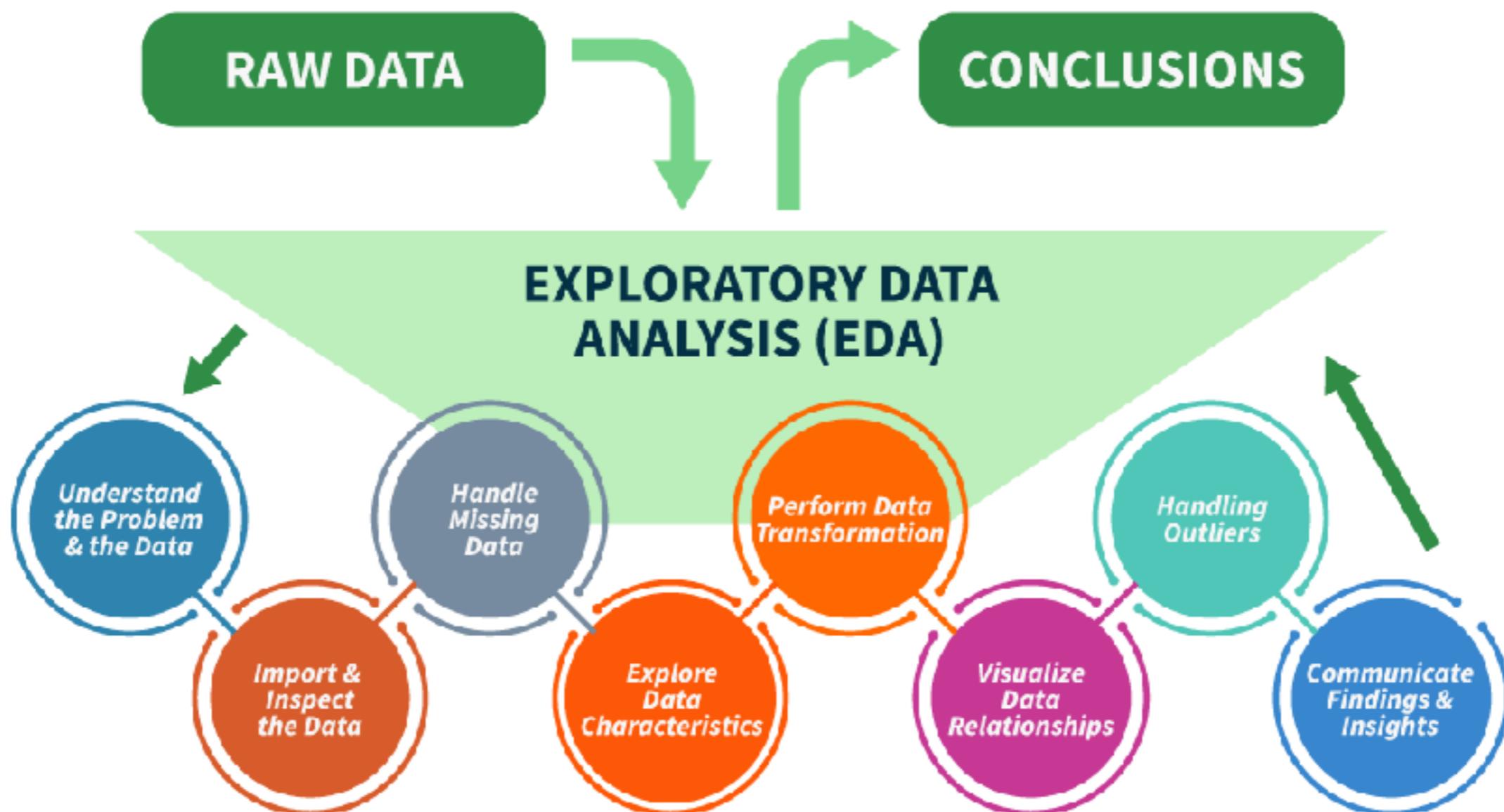


# Exploratory Data Analysis

EDA helps in understanding the data, identifying patterns, detecting anomalies, and testing hypotheses.



# Steps for Performing Exploratory Data Analysis



# Steps

Problem and data understanding

    Data cleaning

    Pattern discovery

    Data visualization

    Model selection

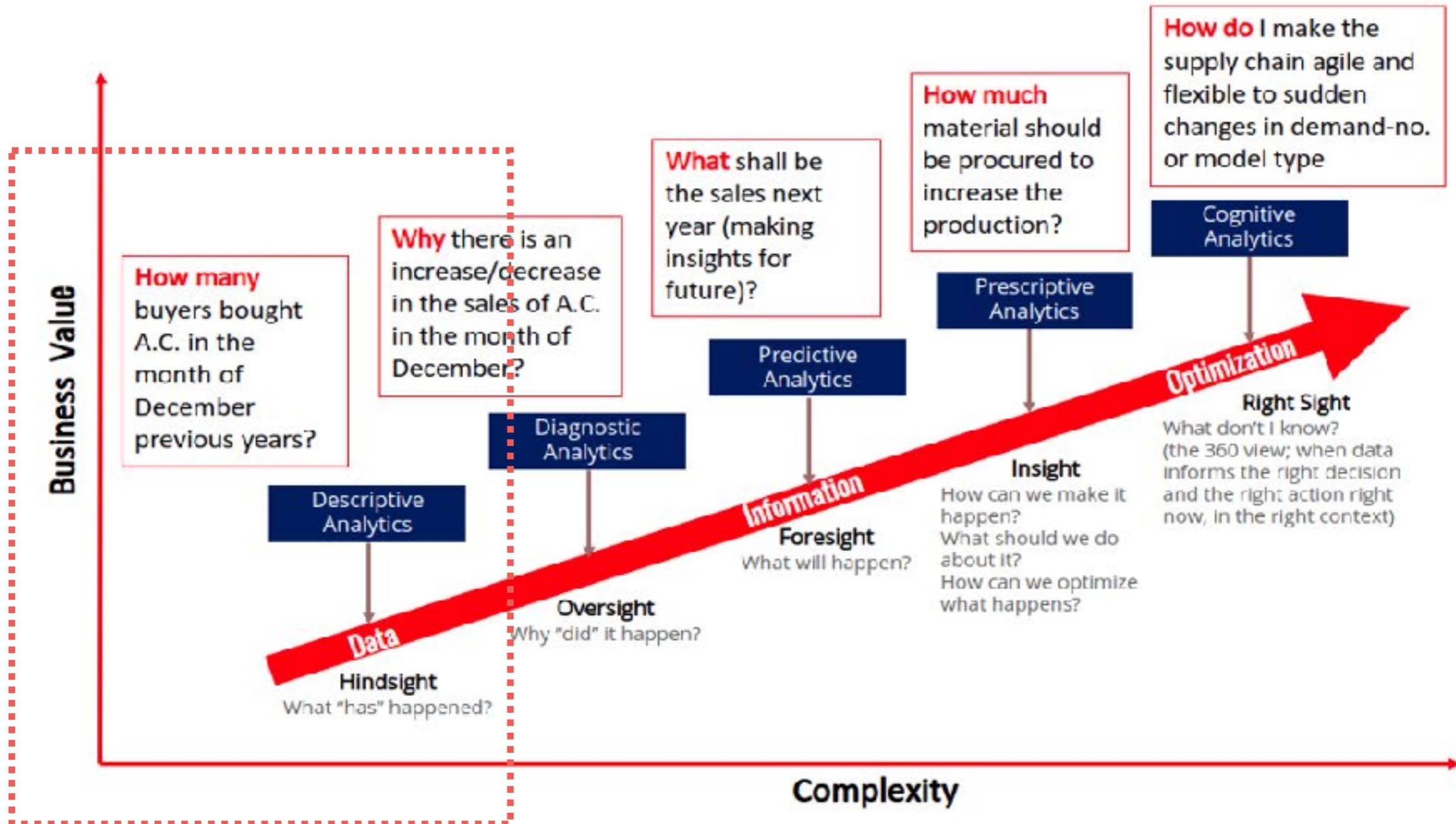
    Quality control



# **Statistical Analysis for data analysis**



# Key Techniques



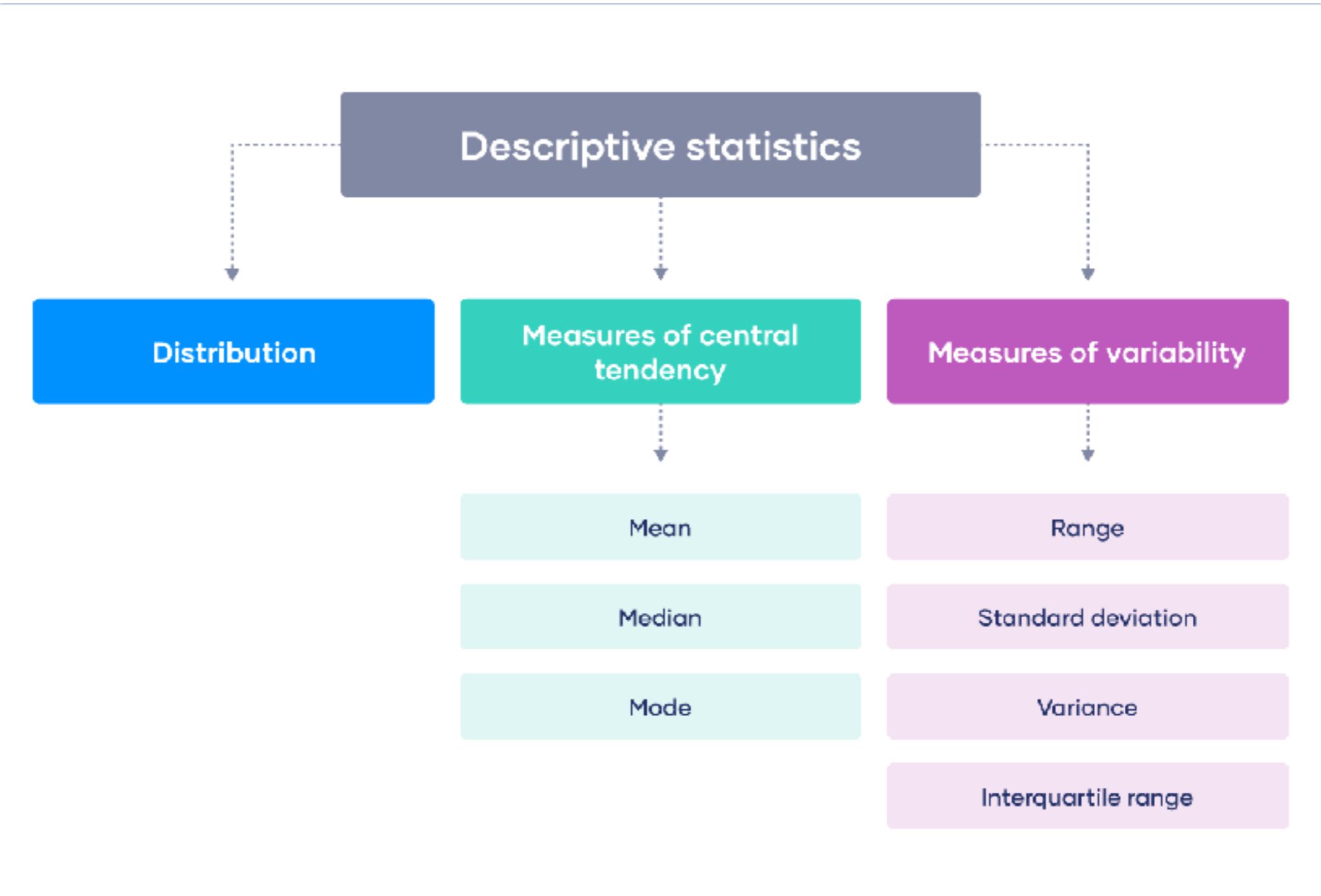
# **Descriptive statistic**

Statistical methods used to **summarize** and describe the **main features** of a dataset.

Help in **understanding the data** by providing a clear overview through numerical measures and visual representations.



# Types of descriptive statistic



# Measures of Central Tendency

## Mean

The average of all data points

## Median

The middle value when data points are arranged in order

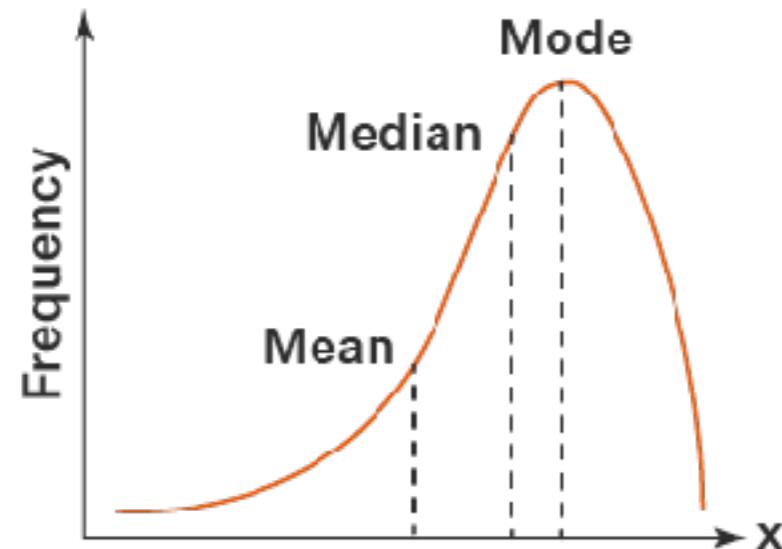
## Mode

The value that appears most frequently in the dataset



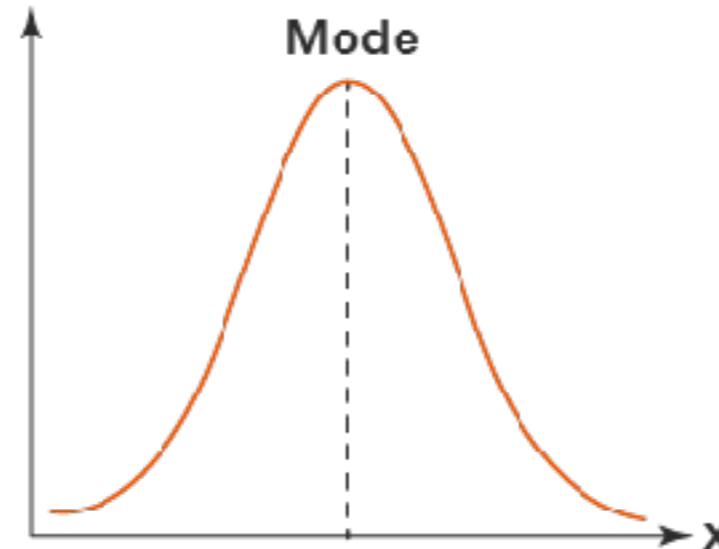
# Distribution

**mean < median < mode**



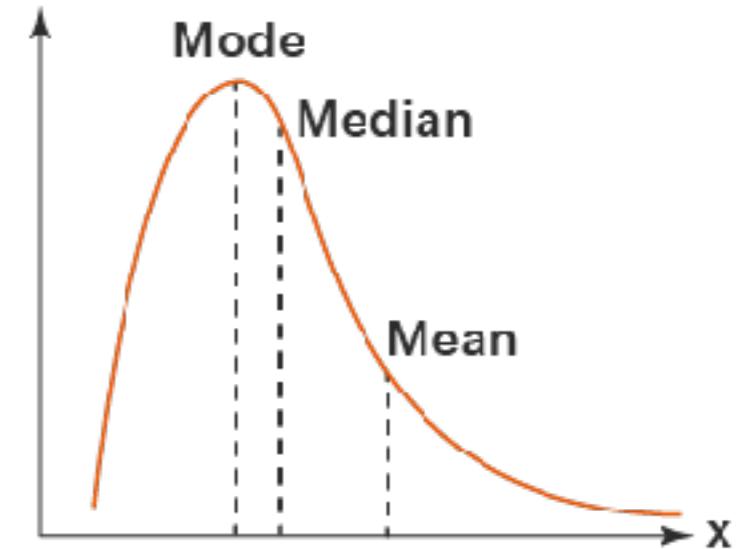
Negatively Skewed

**mean = median = mode**



Symmetrical Distribution

**mean > median > mode**



Positively Skewed



# Measures of Variability

## Range

The difference between the maximum and minimum values

## Variance

The average of the squared differences from the mean

## Standard Deviation

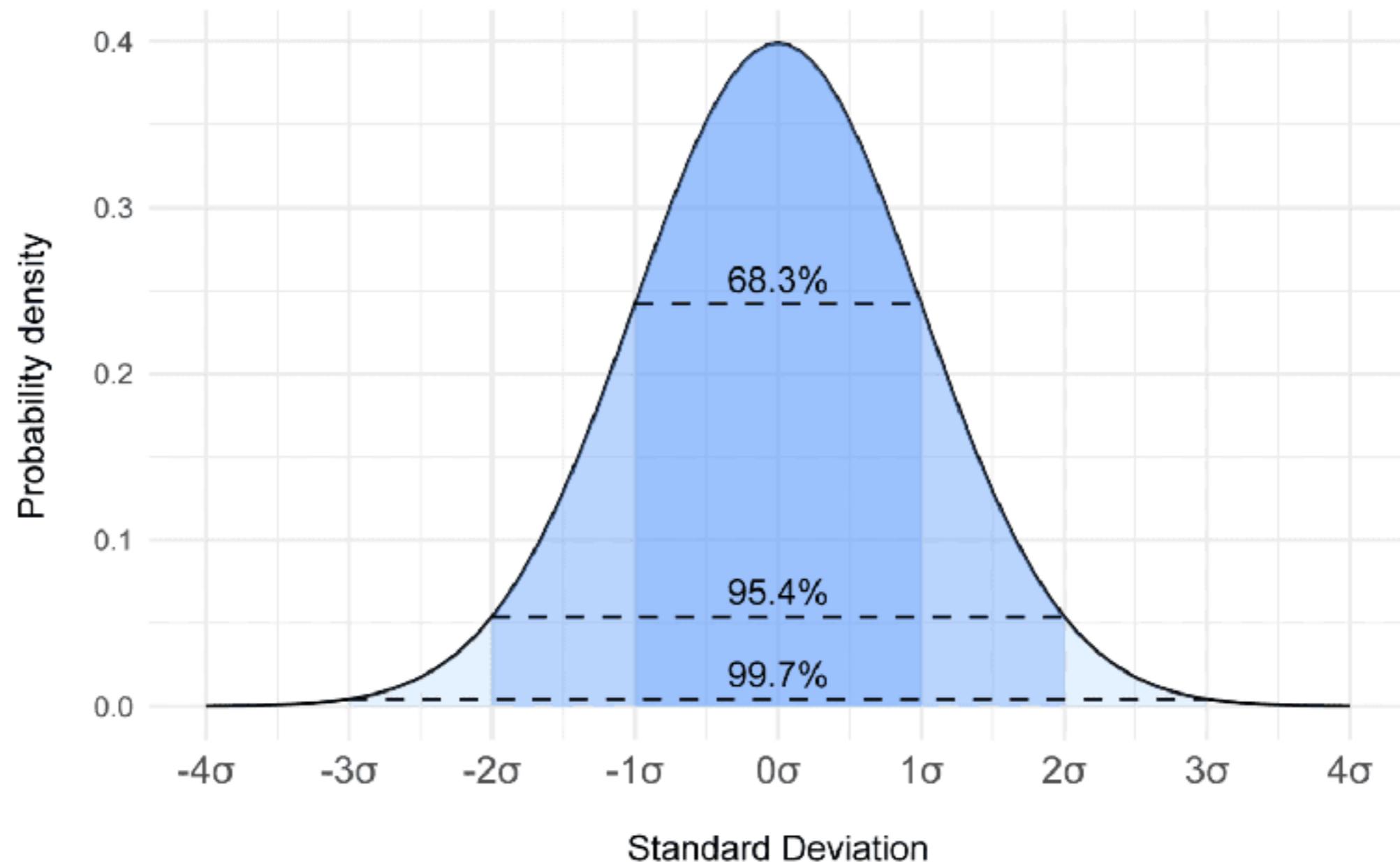
The square root of the variance, showing how much the values deviate from the mean

## Interquartile Range (IQR)

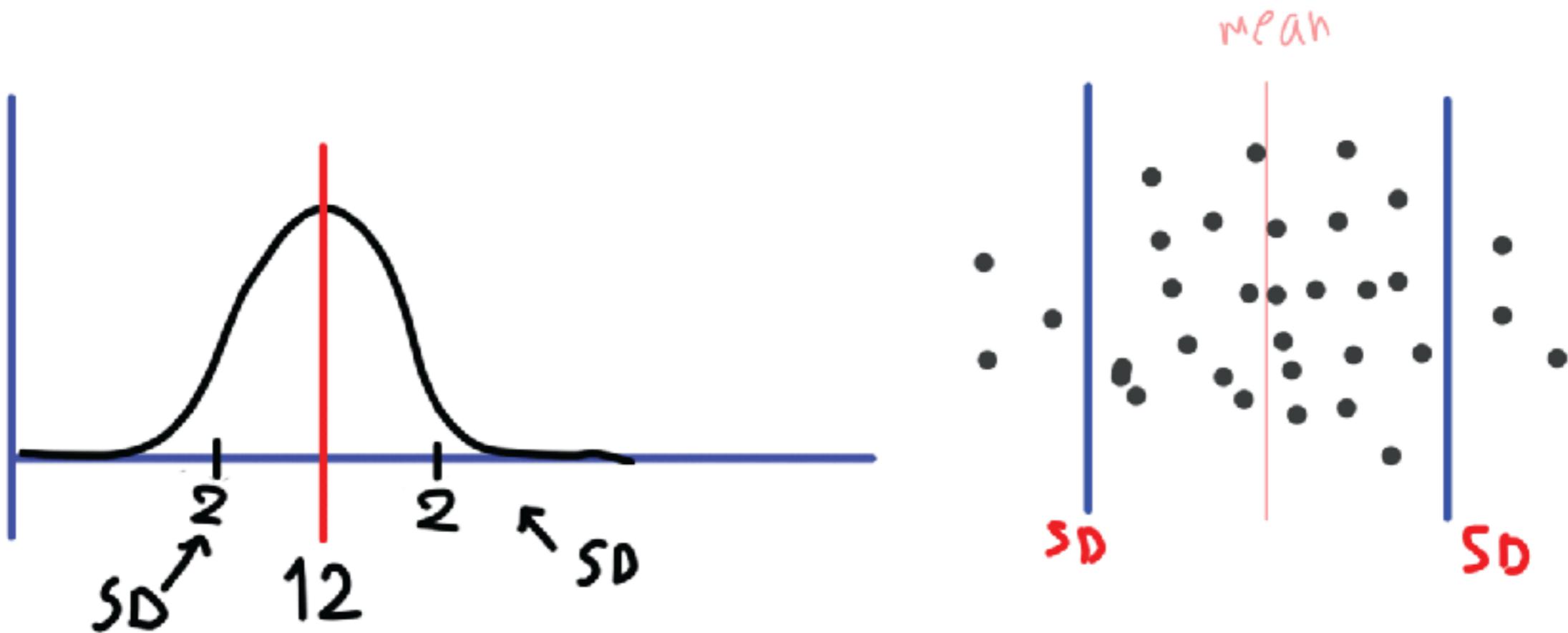
The difference between the first (25th percentile) and third (75th percentile) quartiles



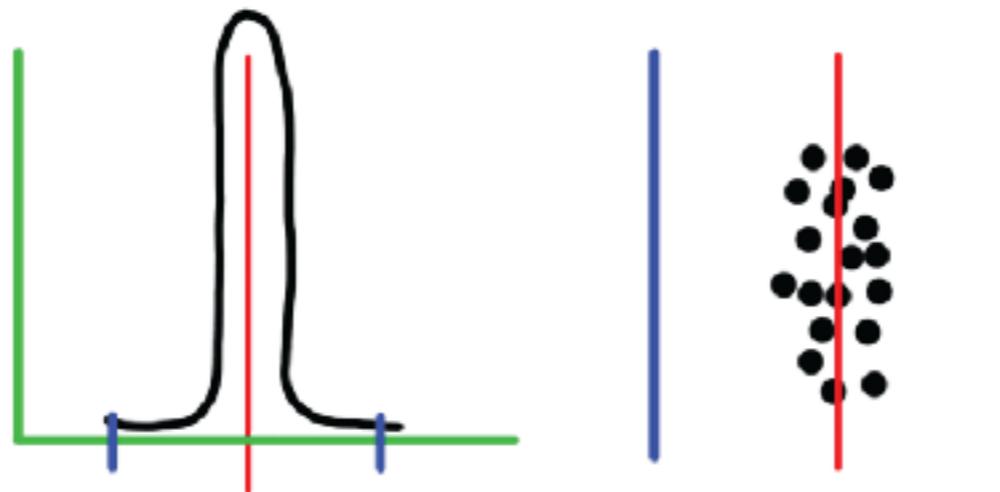
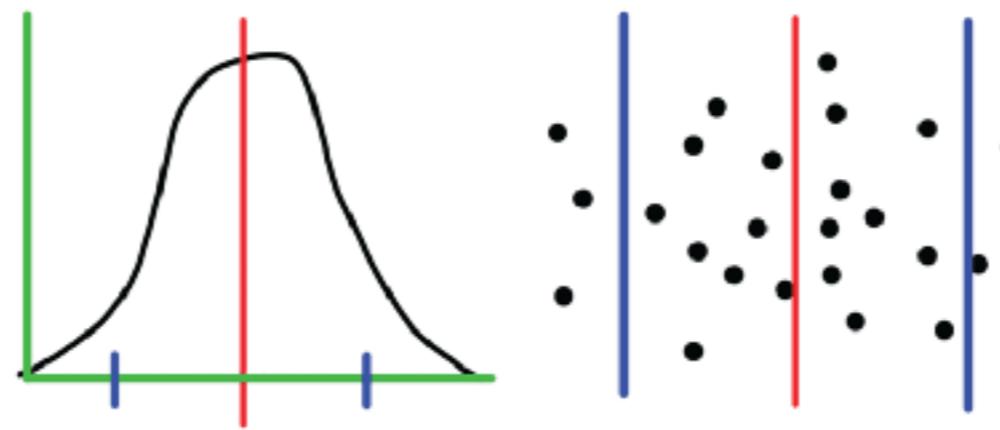
# Standard Deviation (1)



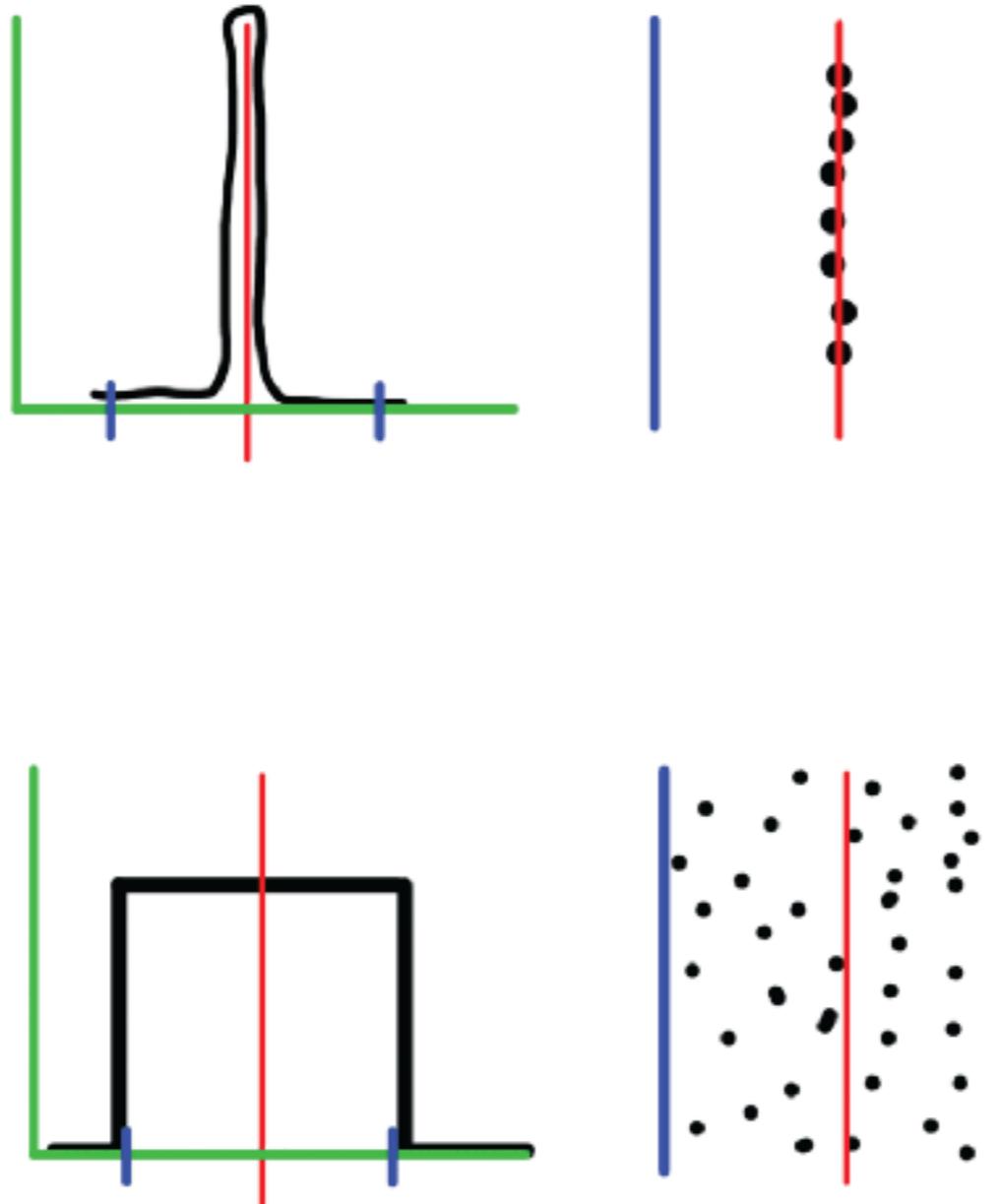
# Standard Deviation (2)



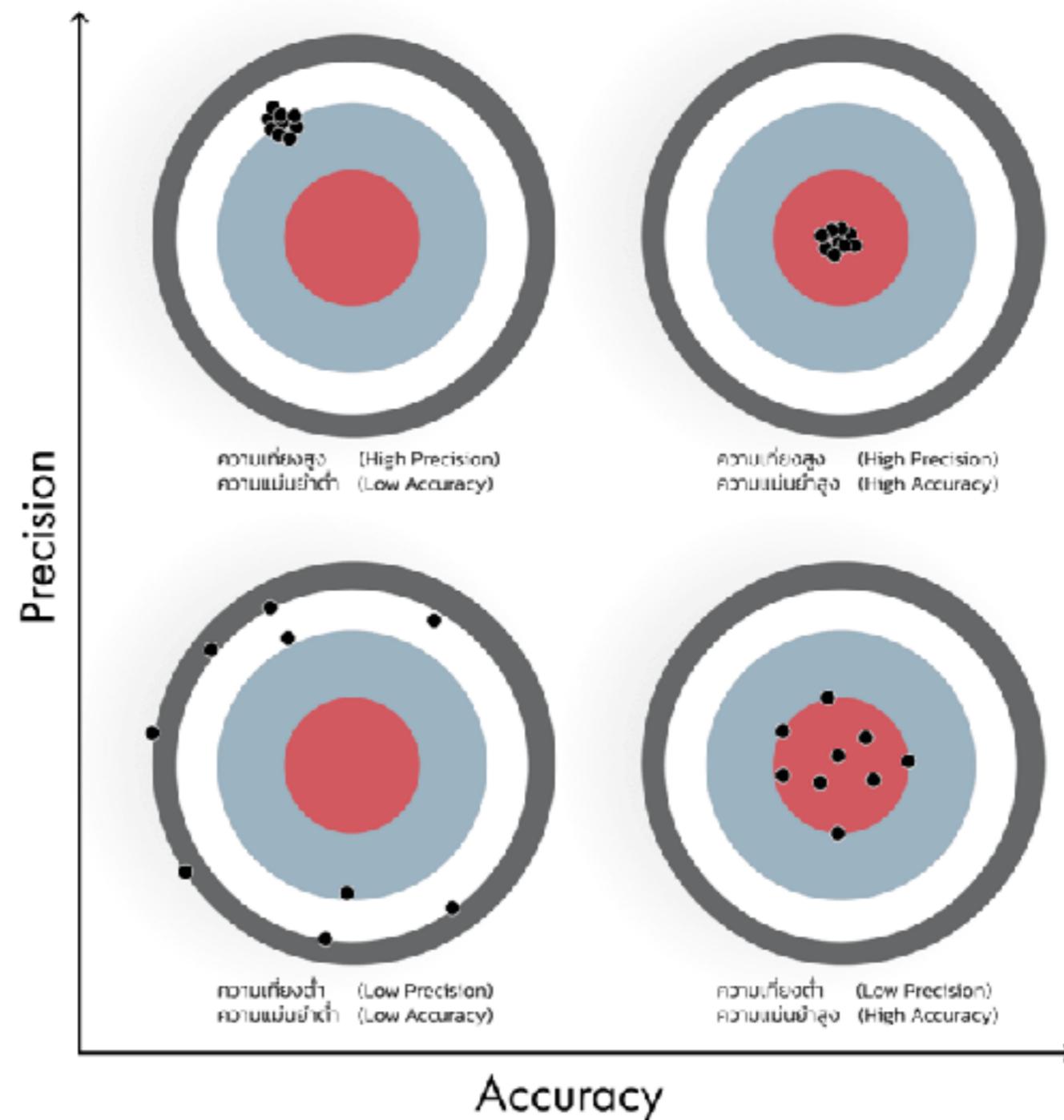
# Standard Deviation (3)



# Standard Deviation (4)



# Standard Deviation and Precision



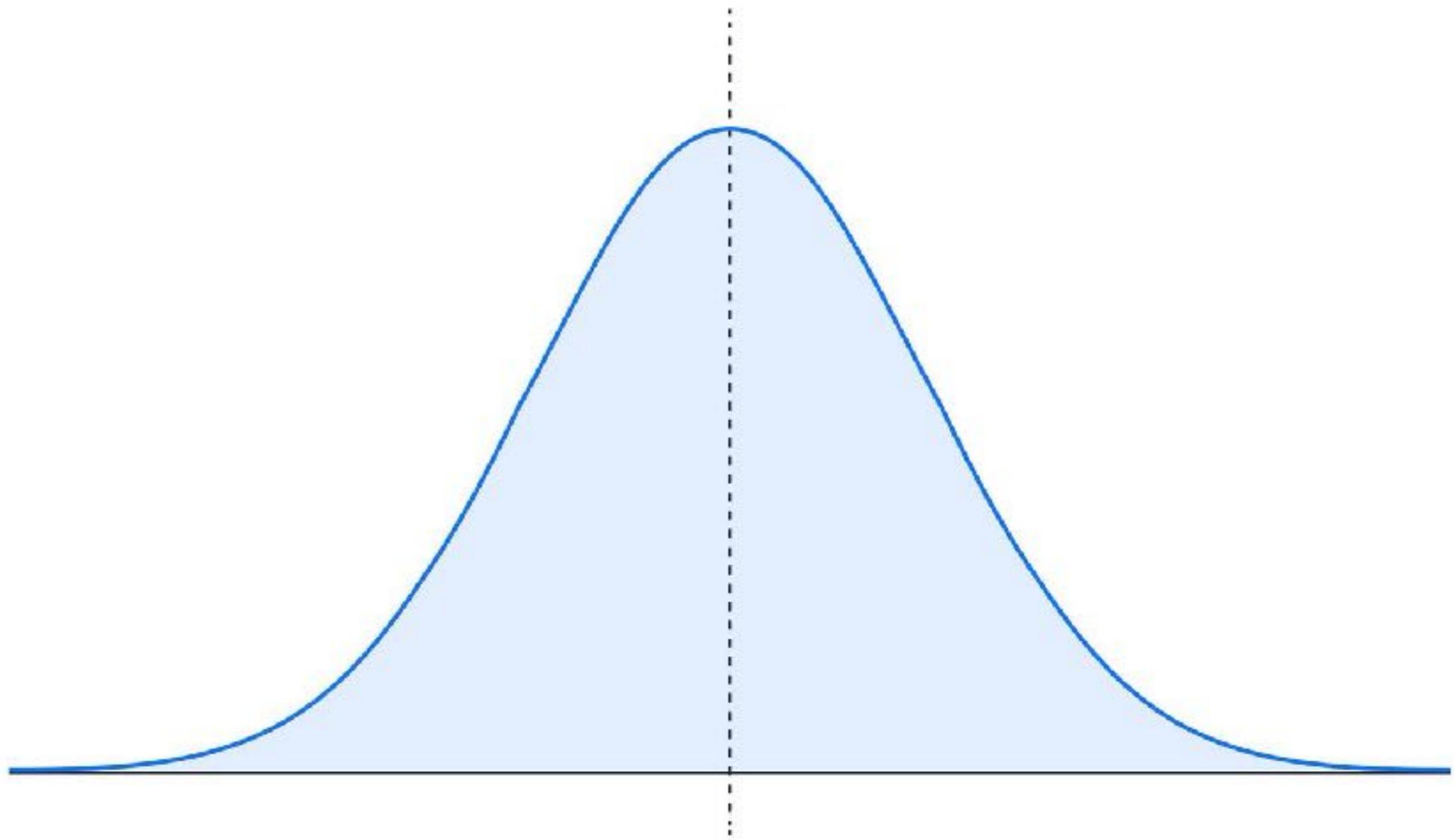
# Percentile

Measure used in statistics to indicate the relative standing of a value within a dataset.

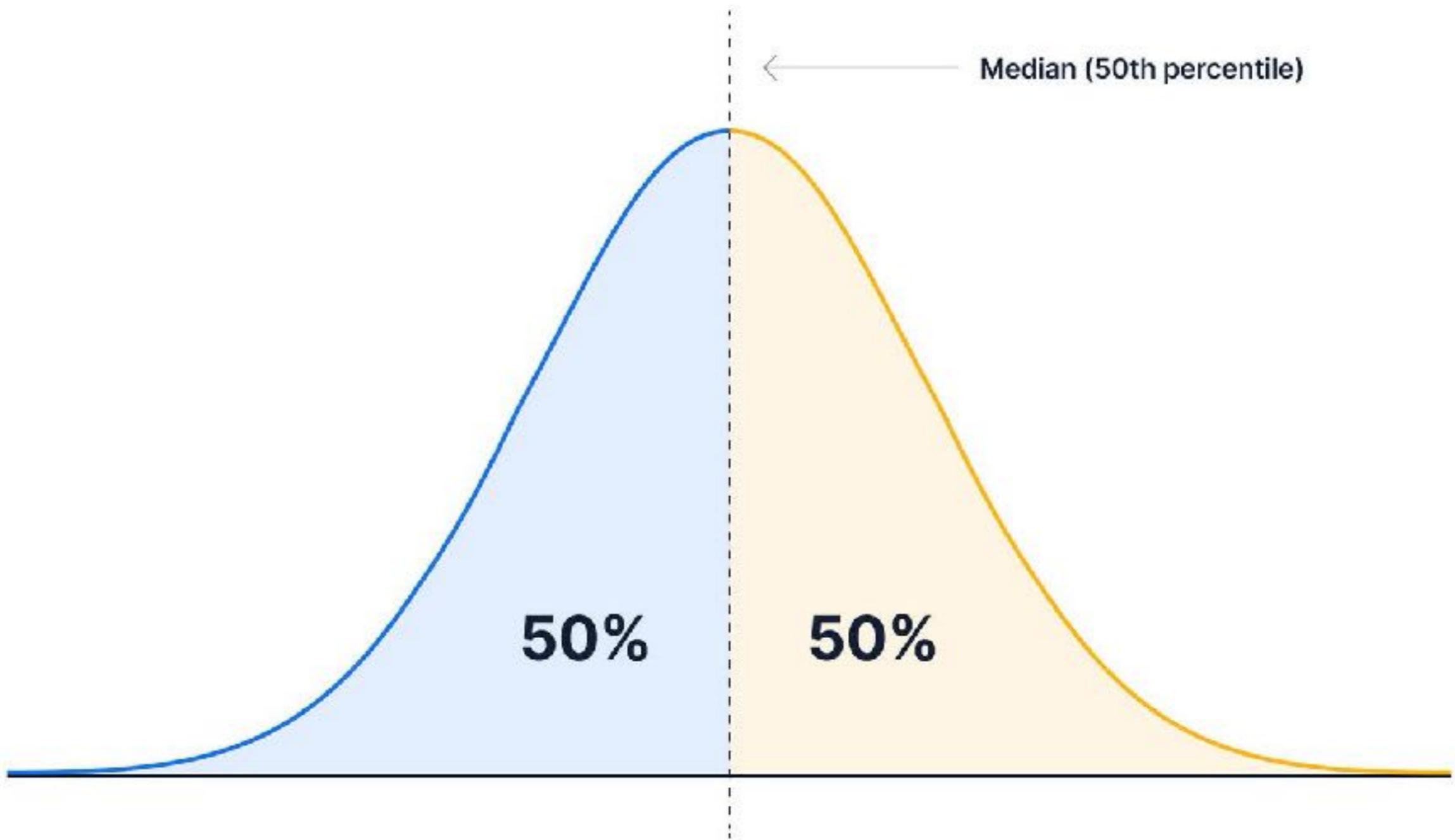
It represents the **percentage of values** in the data that are below a certain point.



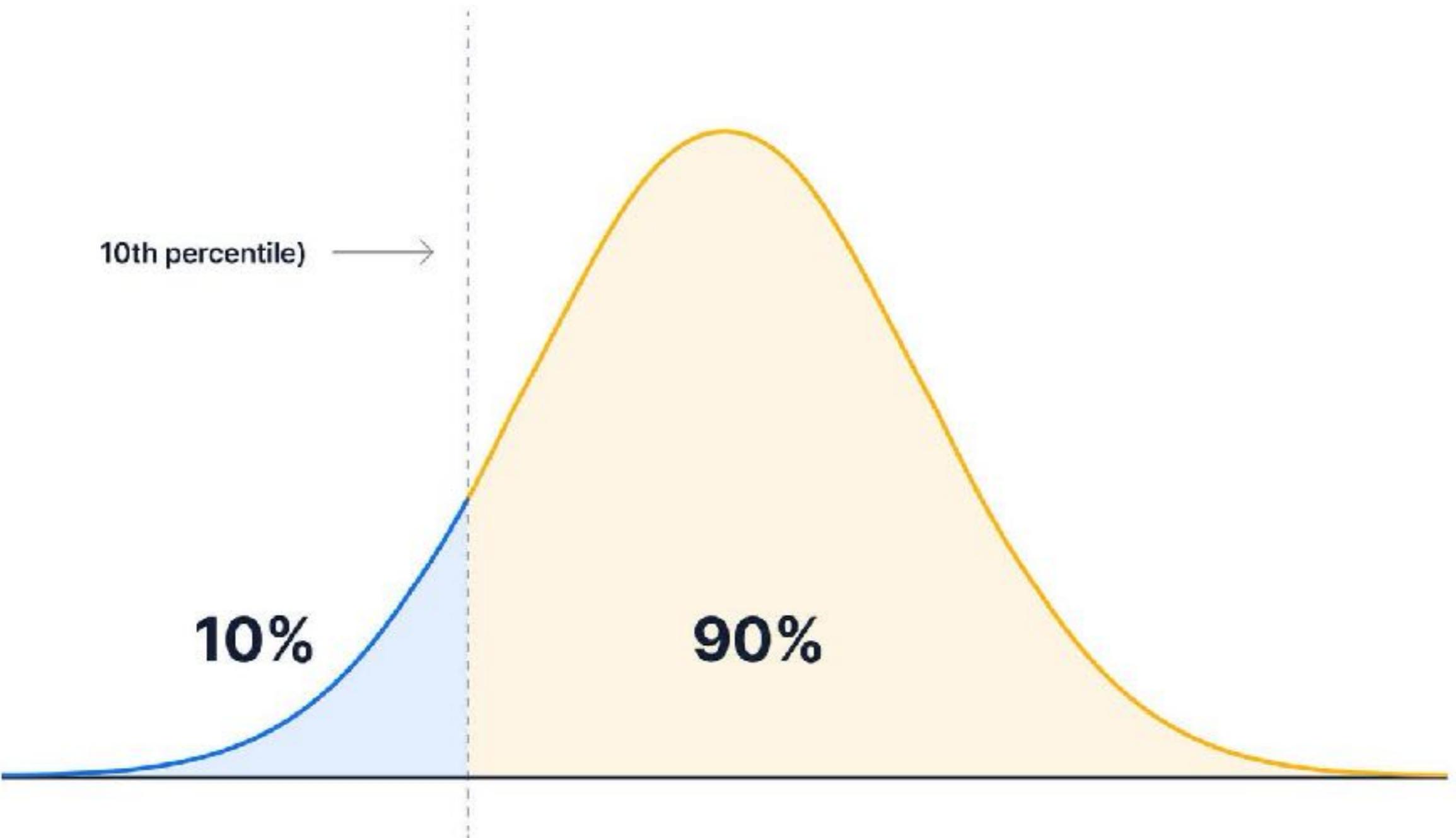
# Normal distribution



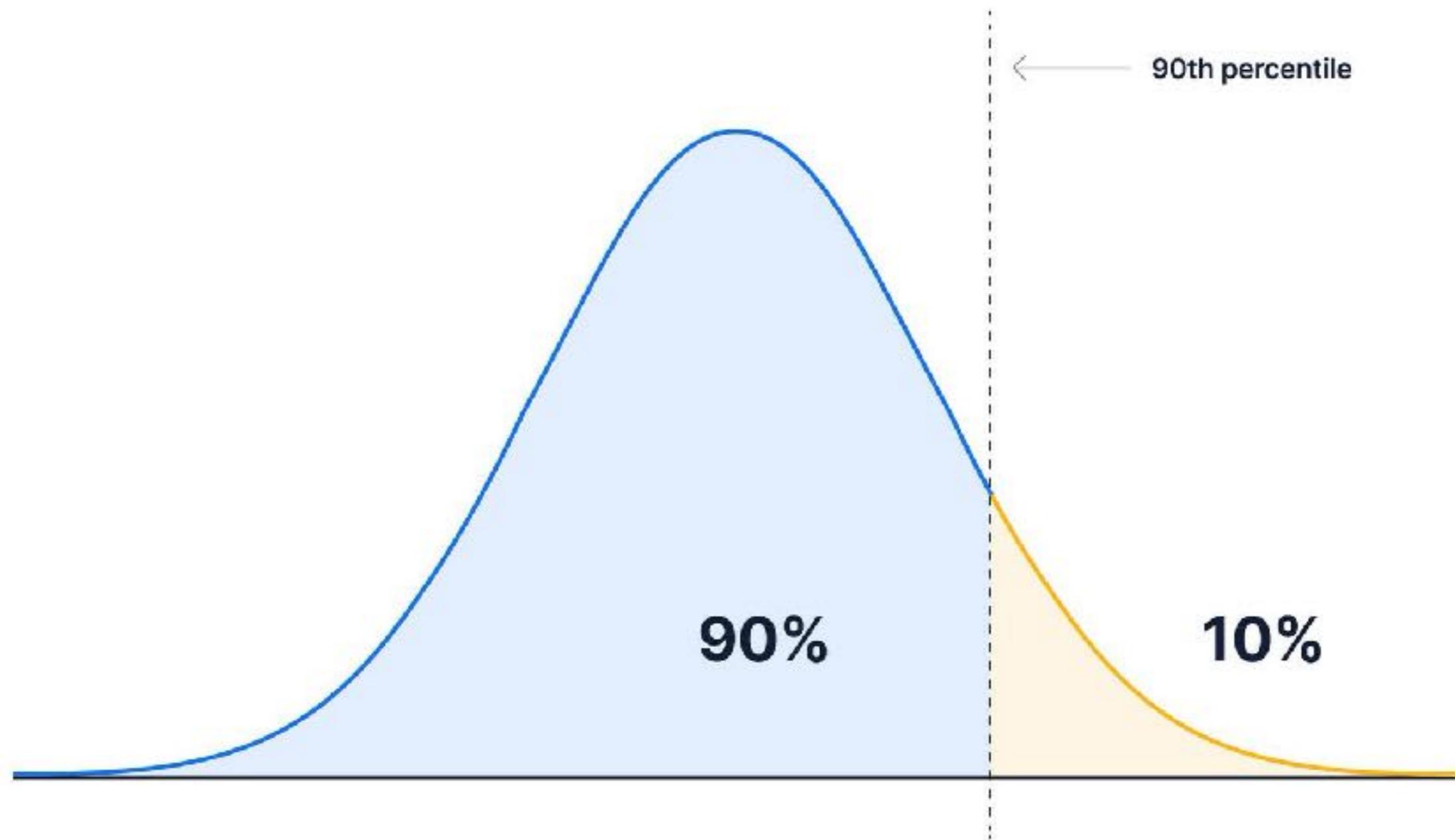
# 50th percentile



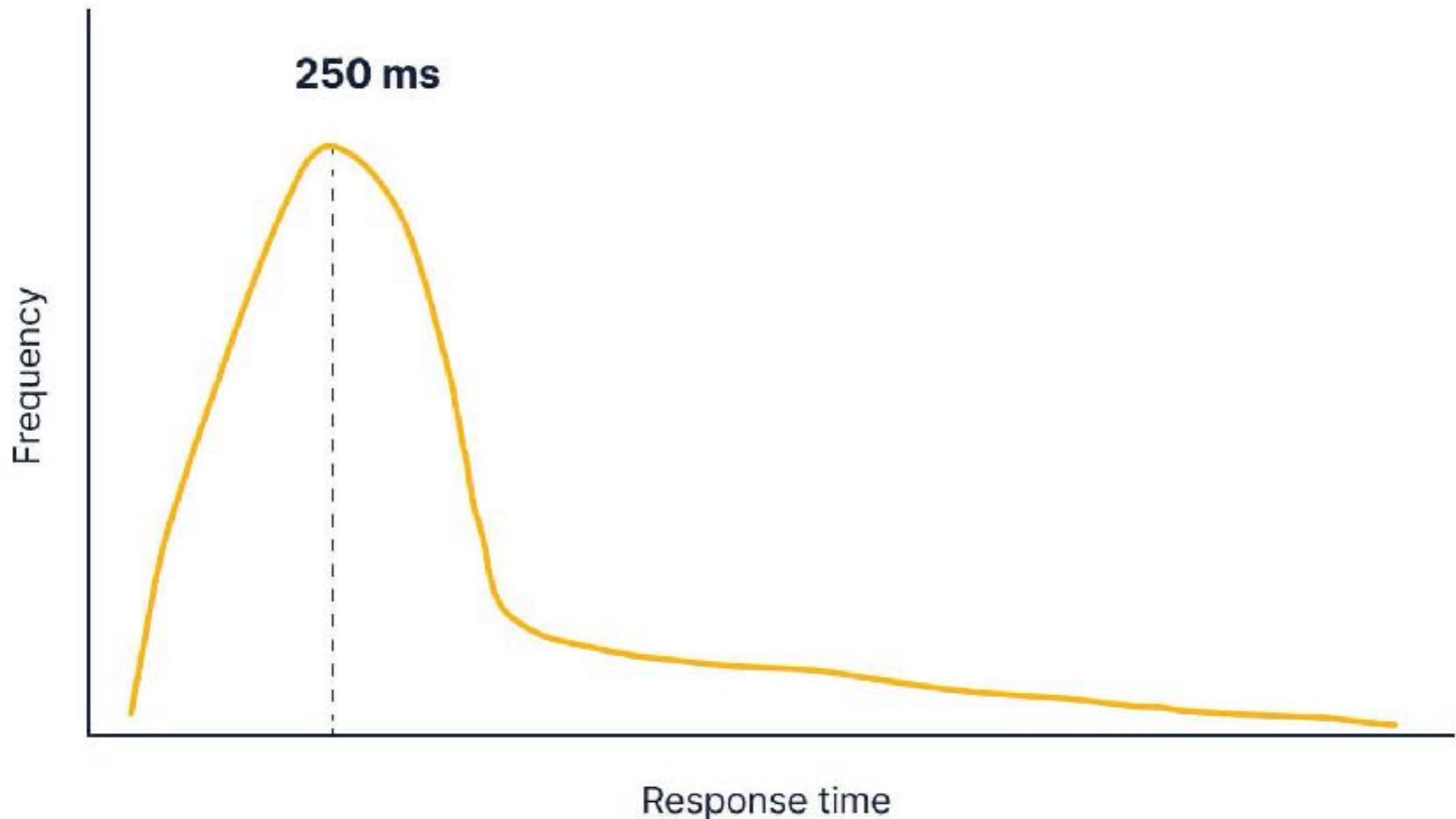
# 10th percentile



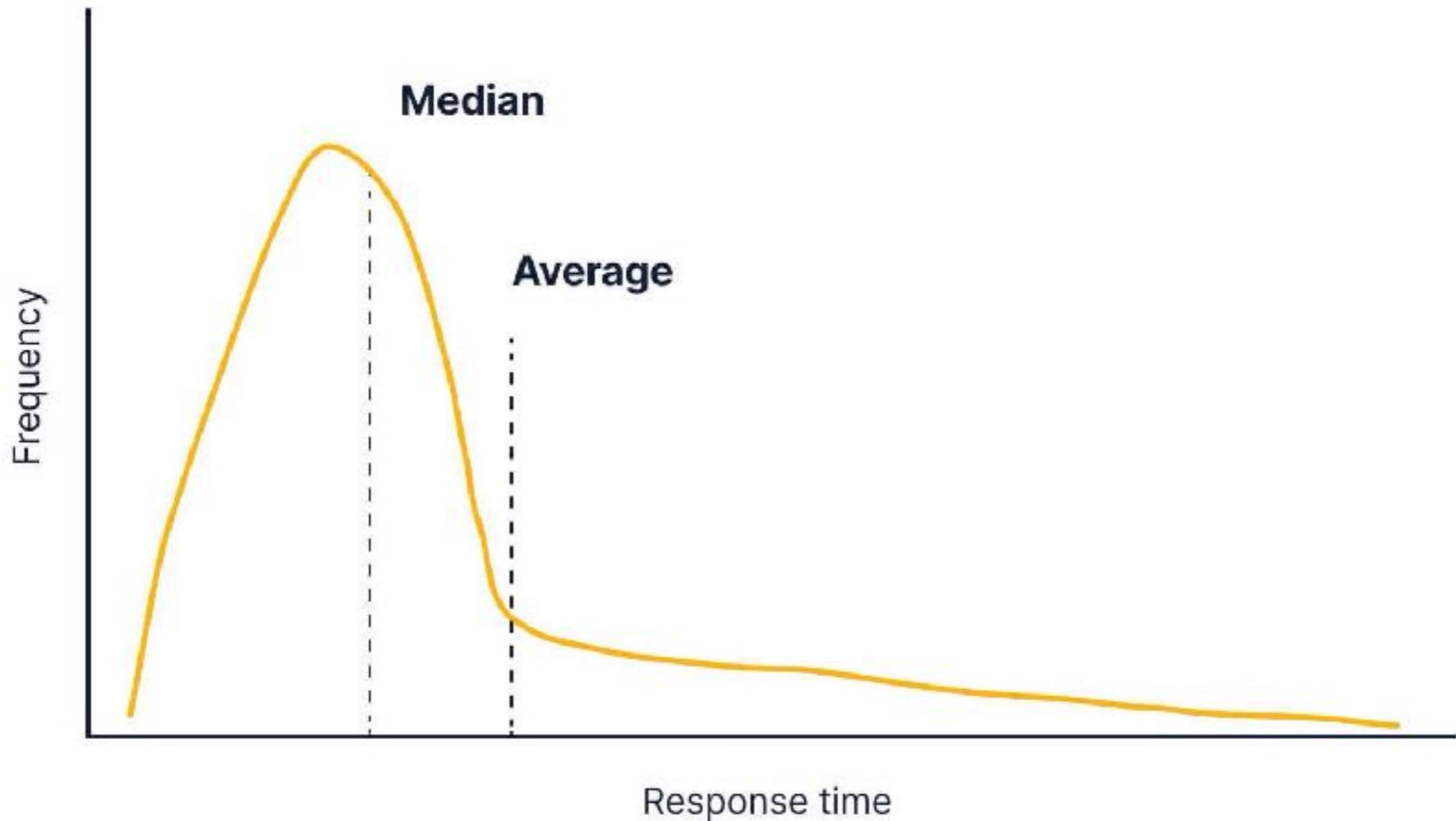
# 90th percentile



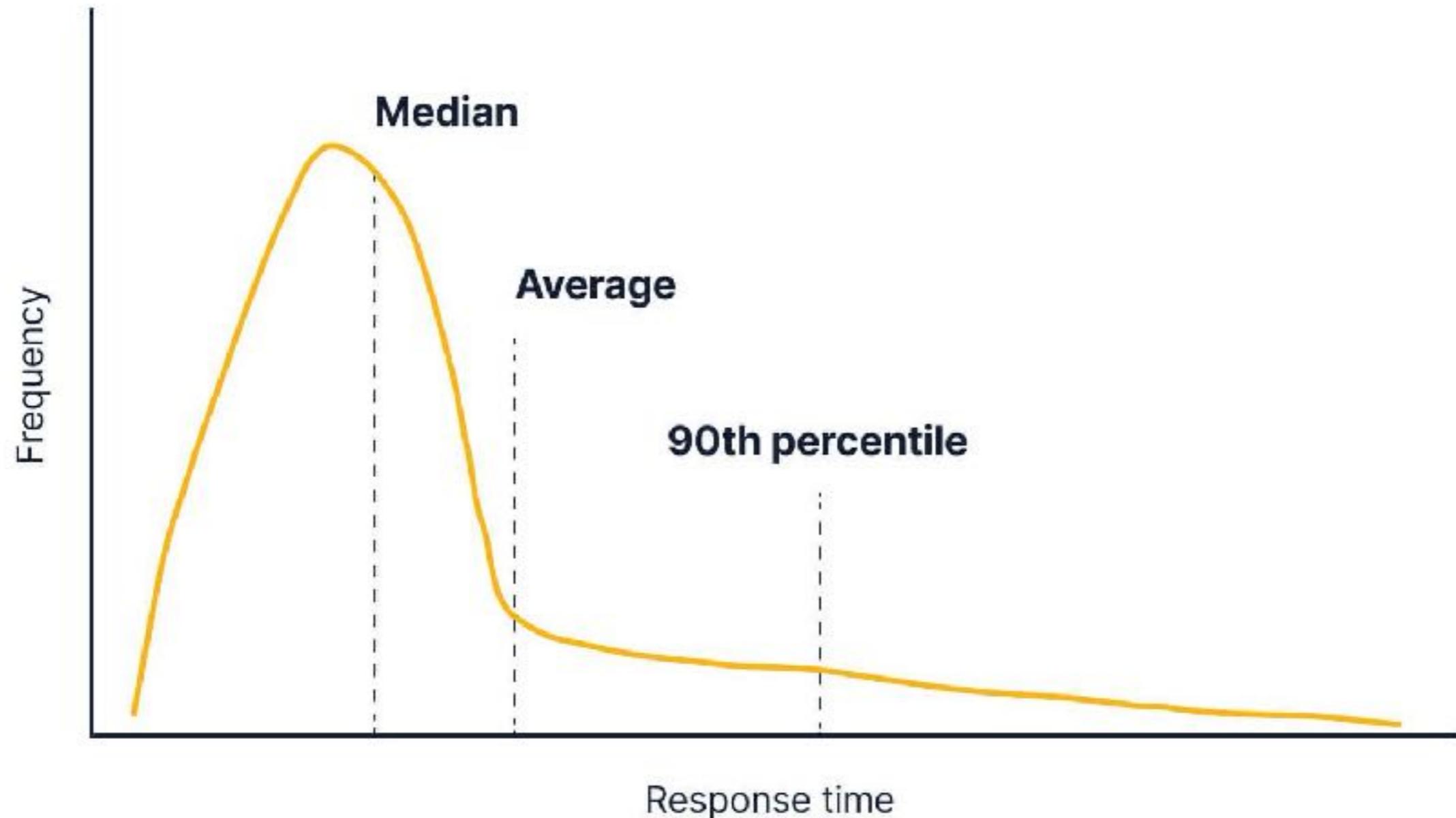
# Long tails (1)



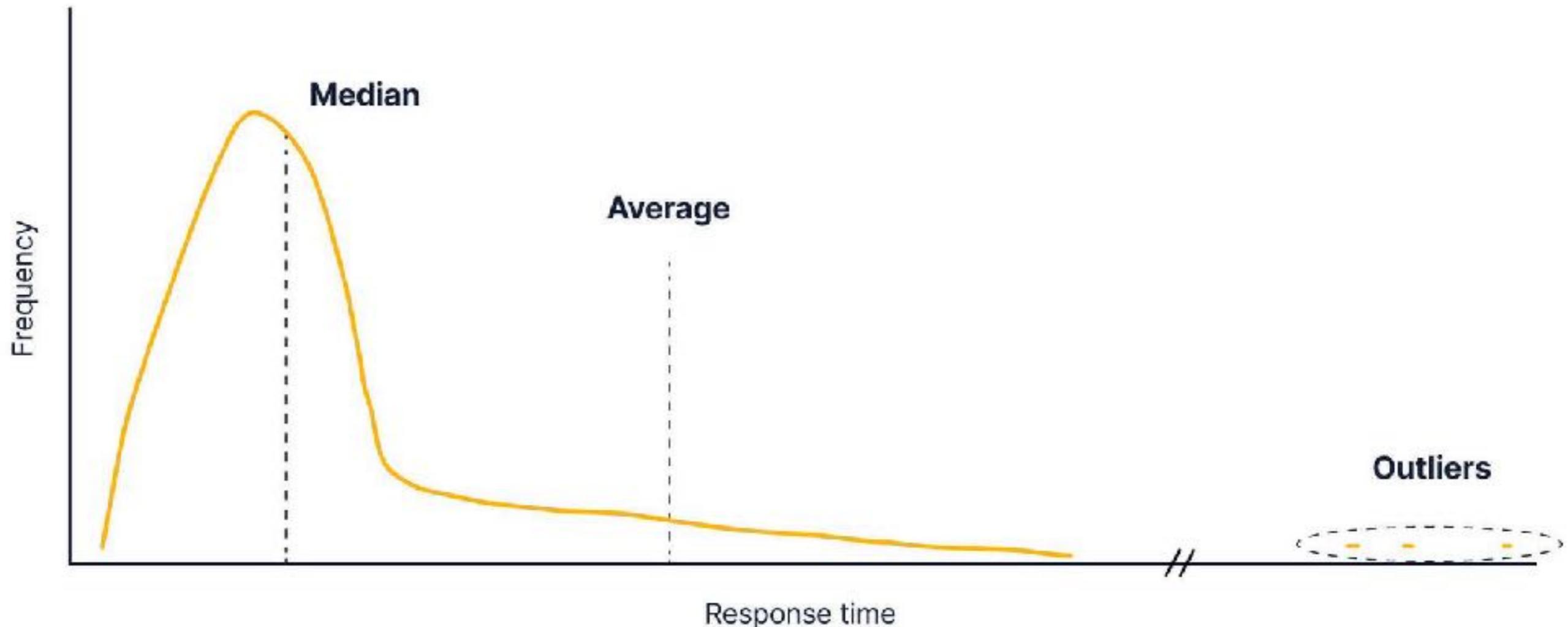
# Long tails (2)



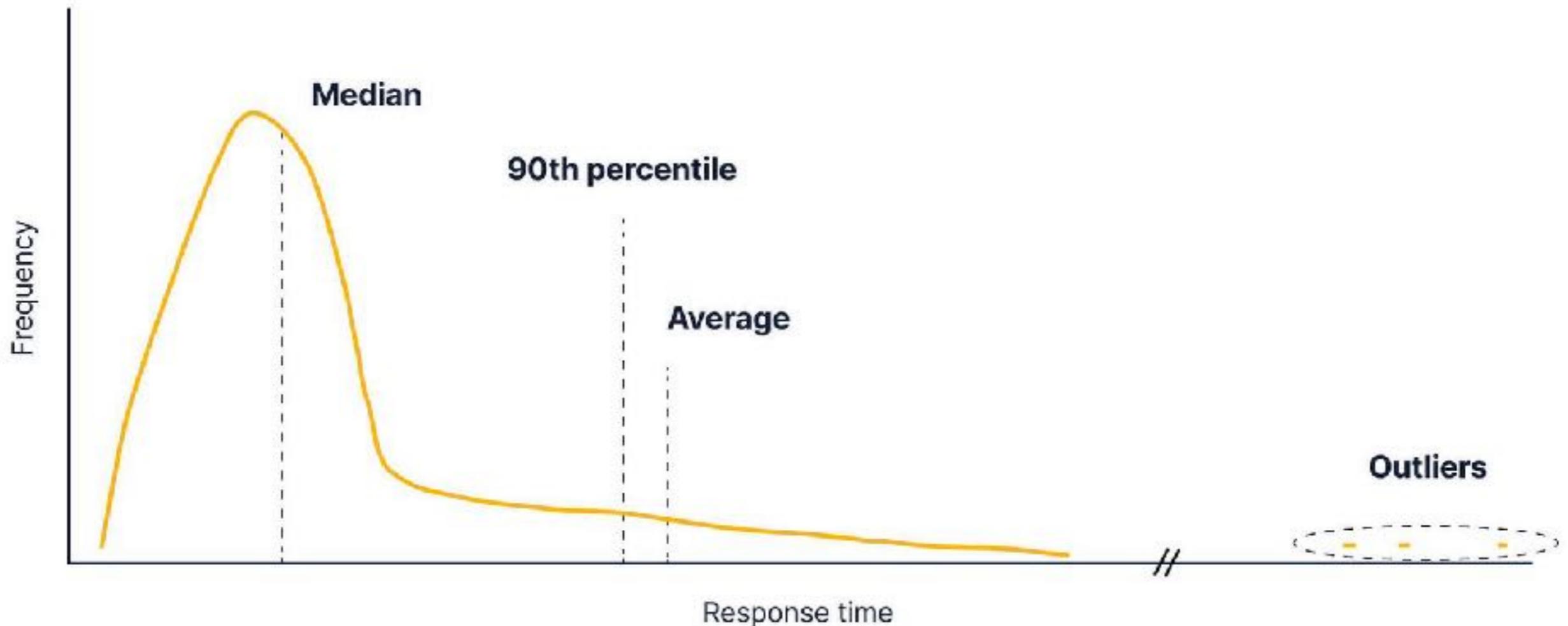
# Long tails (2)



# Long tails with outliers



# Long tails with outliers



# **Business Intelligence (BI) and Visualization**



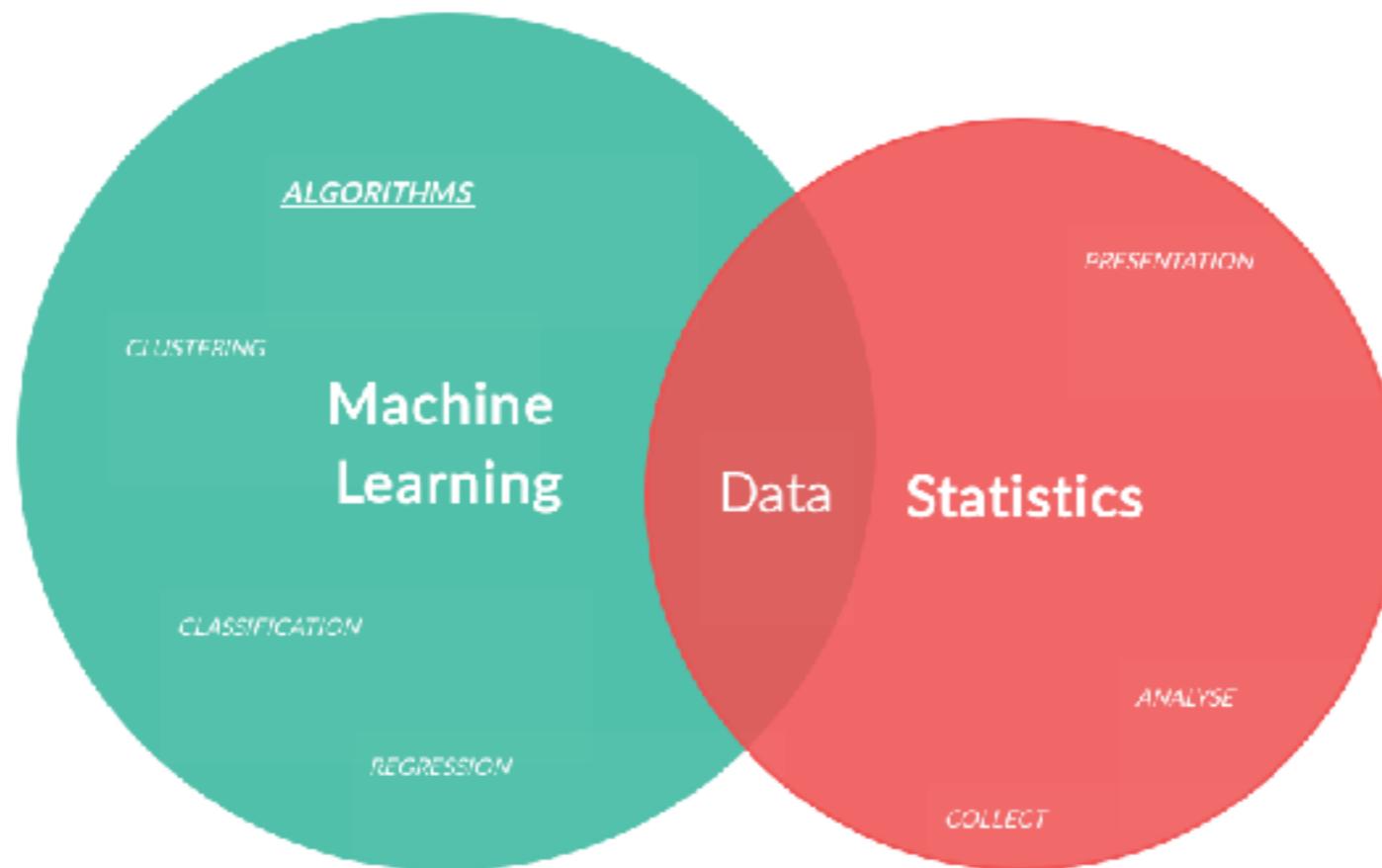
# Data visualization

Data visualization is a method that uses **visuals**,  
both static and interactive,  
to help people **understand**  
the large amount of data being collected.

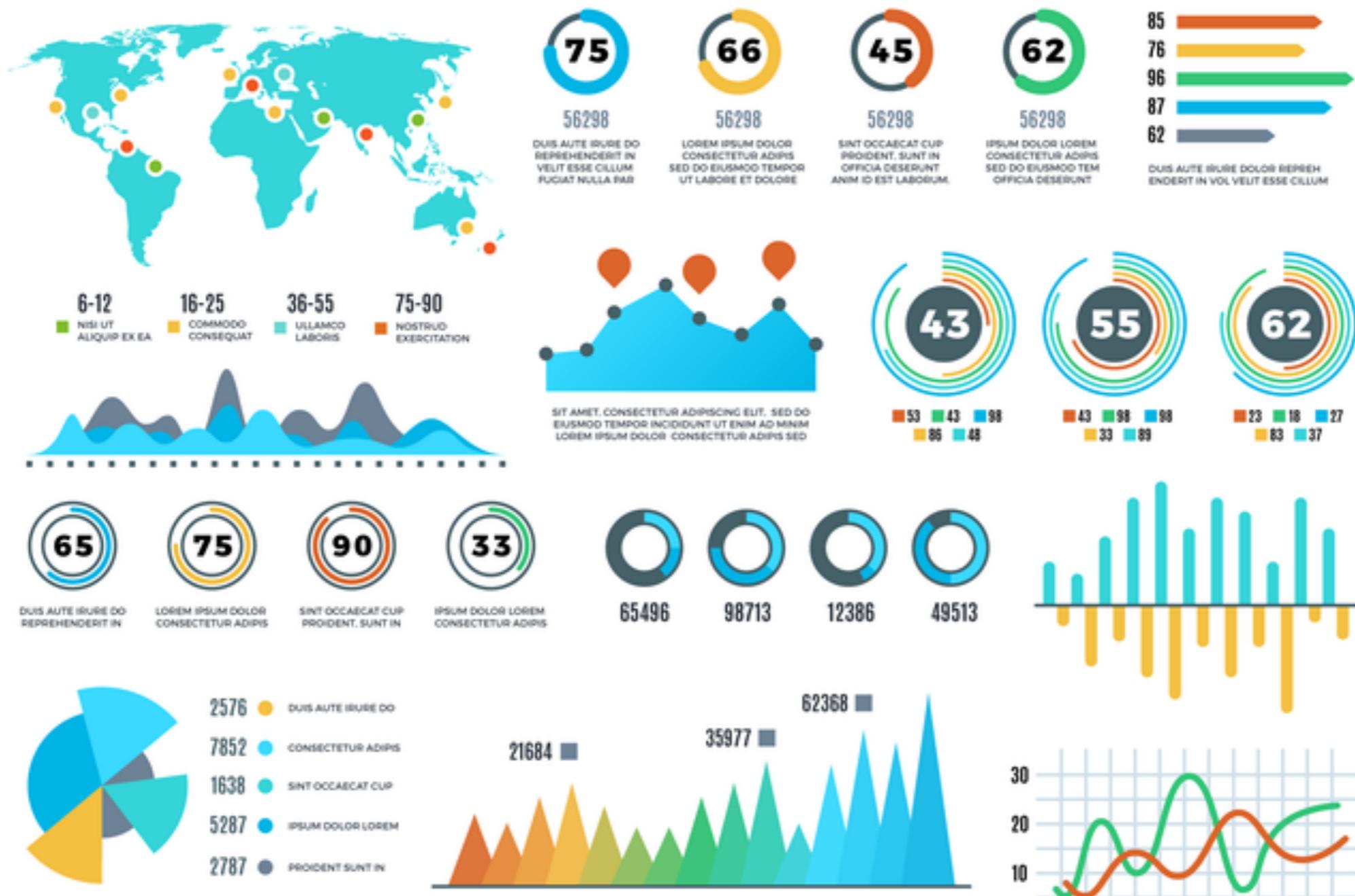


# Data visualization

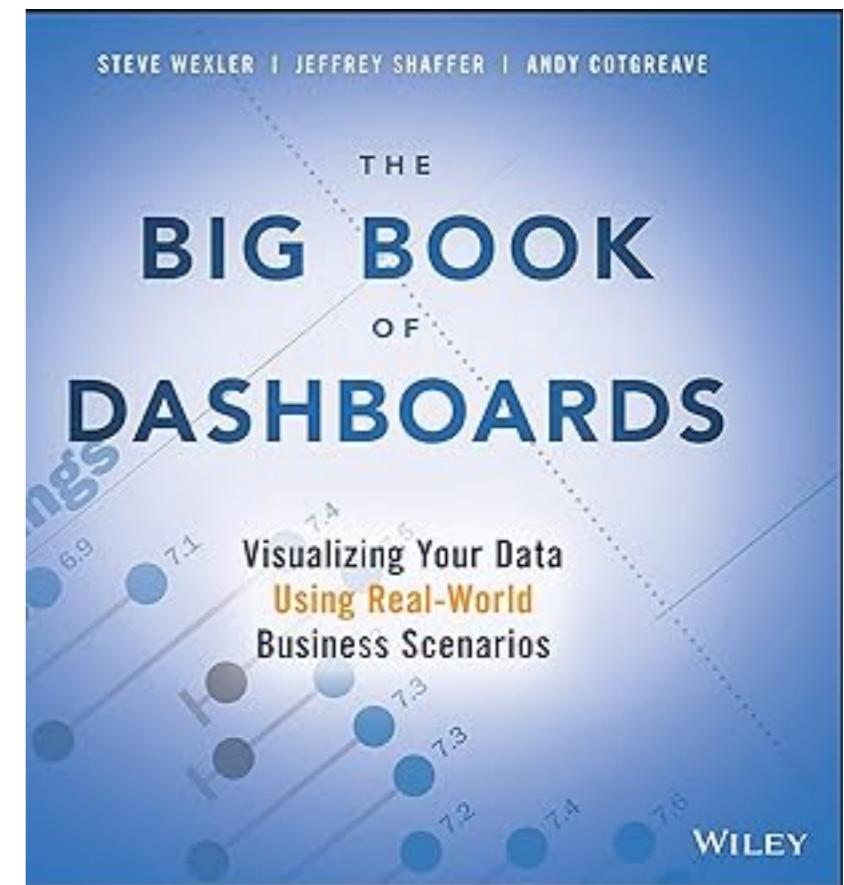
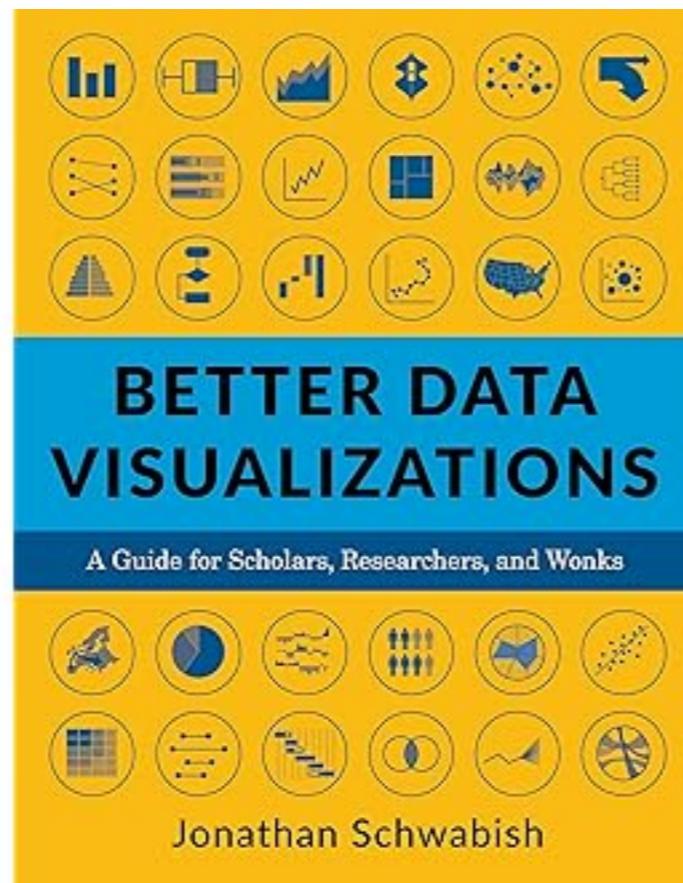
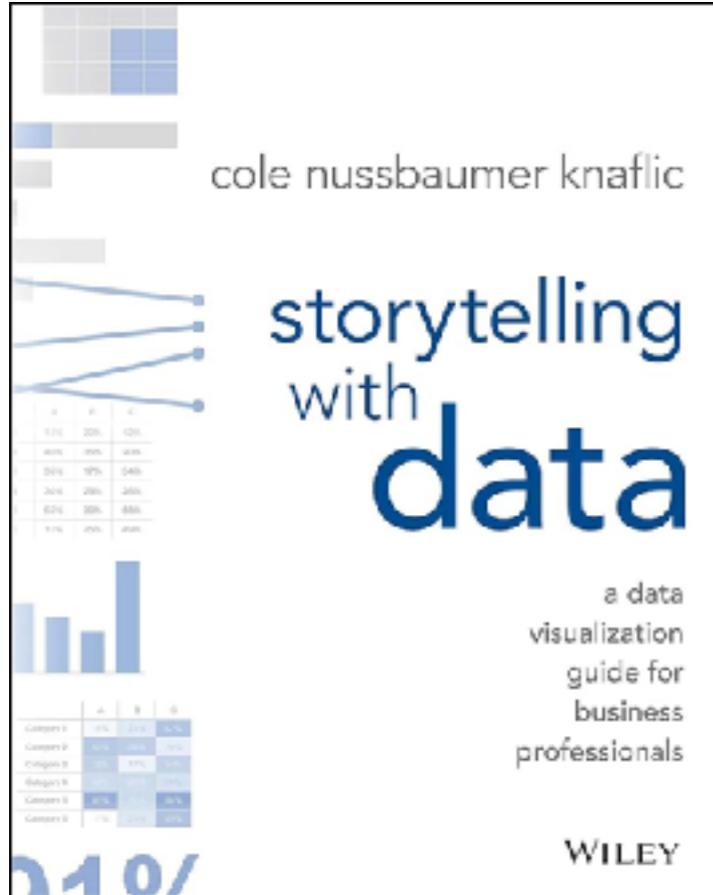
Data visualization is an **important skill** in applied statistics and machine learning.



# Data visualization



# Books



# Data visualization process

1

Collecting data

2

Clean your data

3

Choose a chart type

4

Prepare data

5

Visualize data

6

Presentation



# How to choose a chart type ?

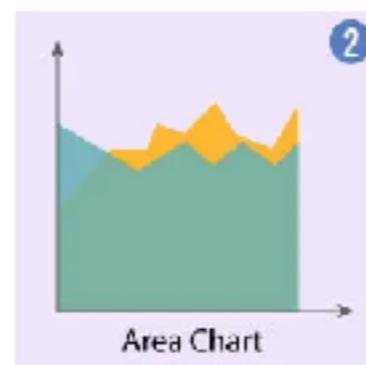
Before choosing a visual chart or graph,  
it is important to understand your **audience**



# TYPES OF DATA VISUALIZATION CHARTS



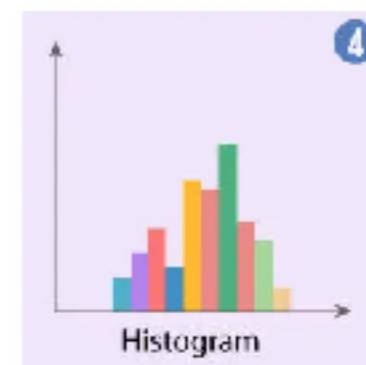
Display trends over time



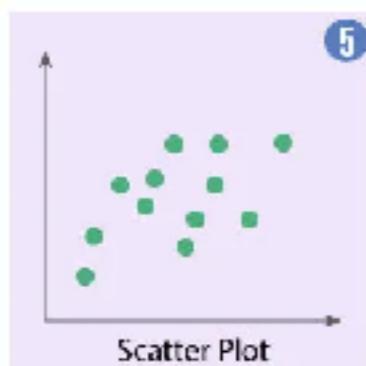
A line chart with areas below the lines filled with colors



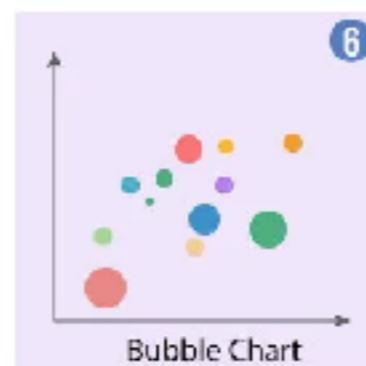
Display trends with multiple variables



Display the shape and spread of continuous dataset samples



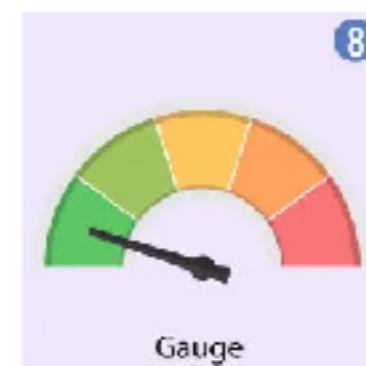
Show correlation in a dataset



Show and compare the relationship between the labelled circles



Show the contribution of data point inside a whole dataset



Visualize the distance between intervals



Show data with location as a variable



Show magnitude of a phenomenon



# Key Concepts in Data Visualization

Data  
Representation

Clarity

Context

Accuracy

Color Theory

Storytelling

Interactivity



# Q/A

