

Effective Big Data Management for Business analytics and decision



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

Workshop with Visualization data tool



Introduction

Big Data

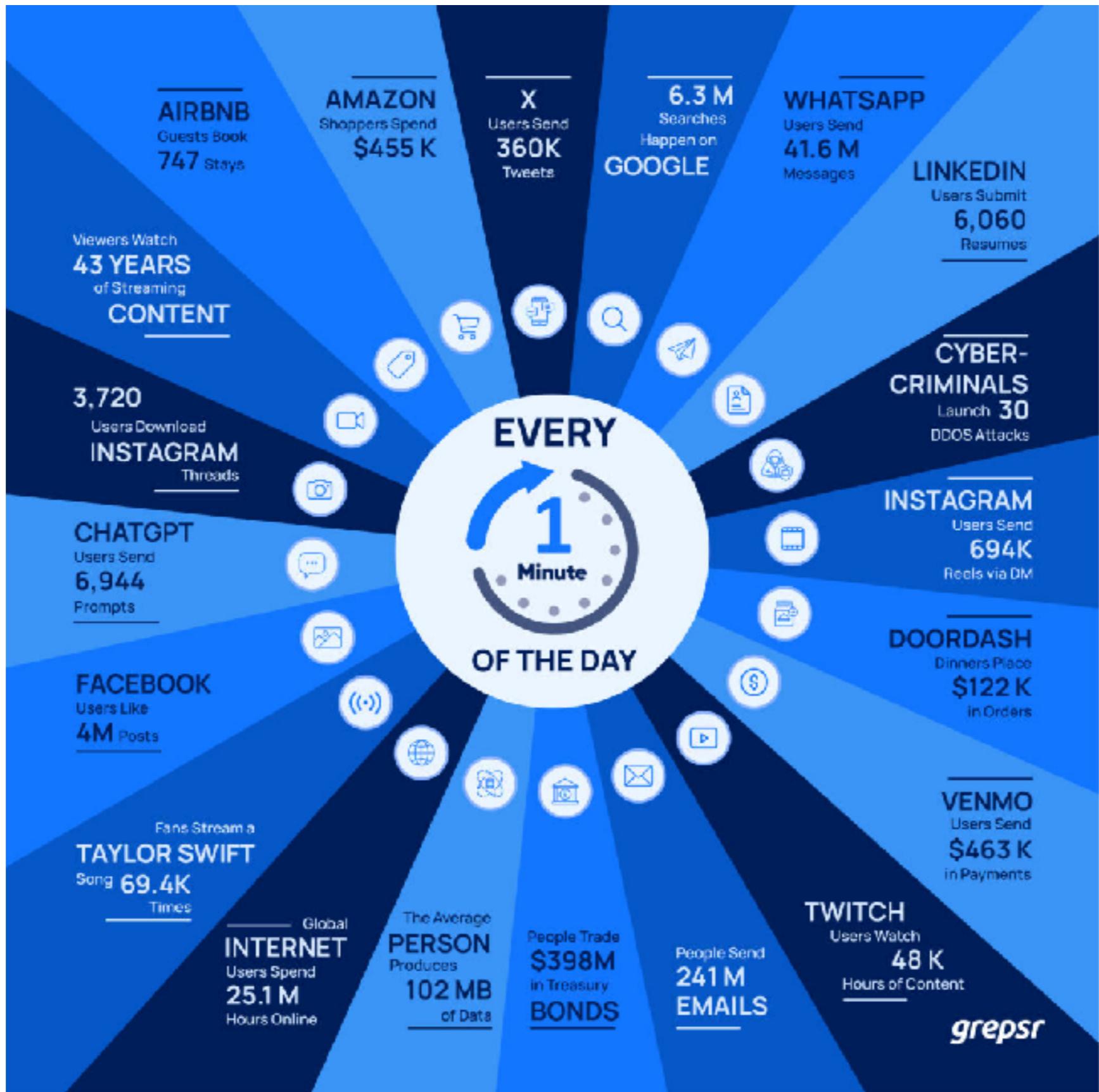
Business
Analytics (BA)

Business
Intelligence (BI)

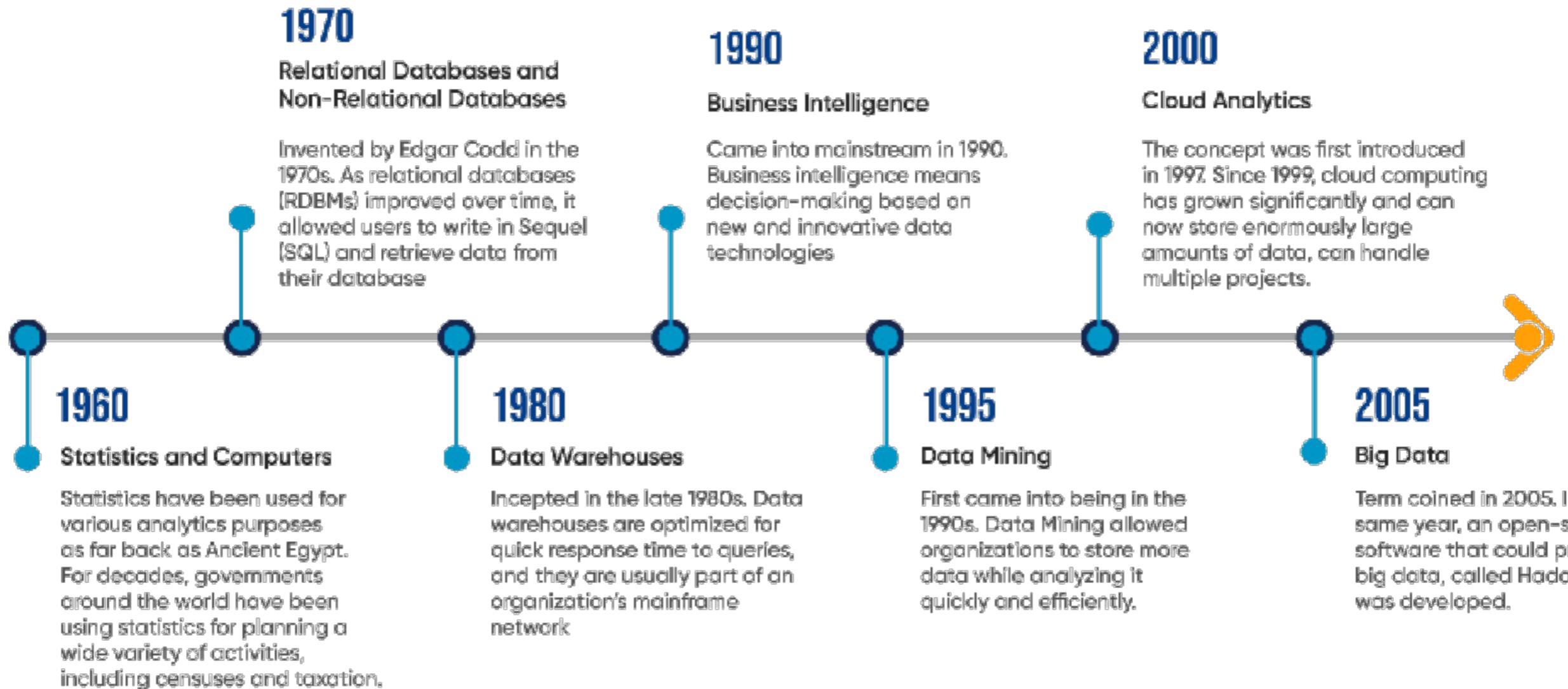


Introduction to **Big** Data



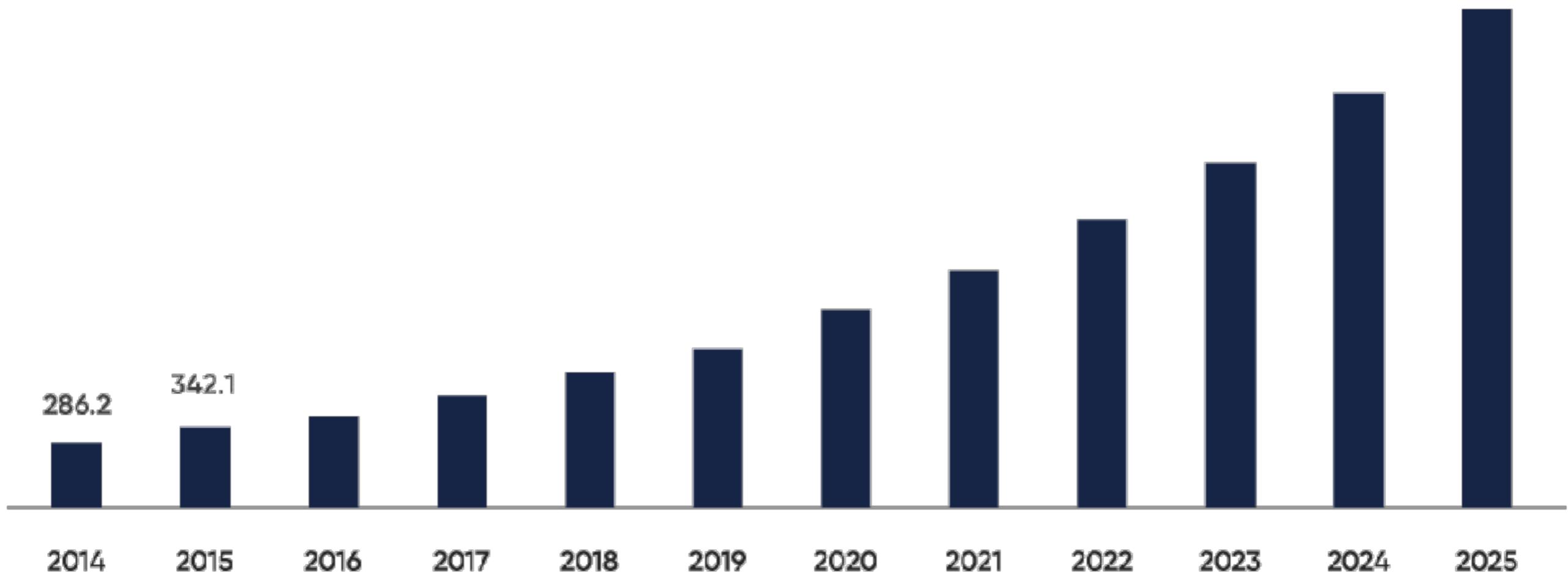


Timeline



Timeline

U.S. Data Analytics outsourcing market size, 2014 - 2025 (USD Million)



Big Data

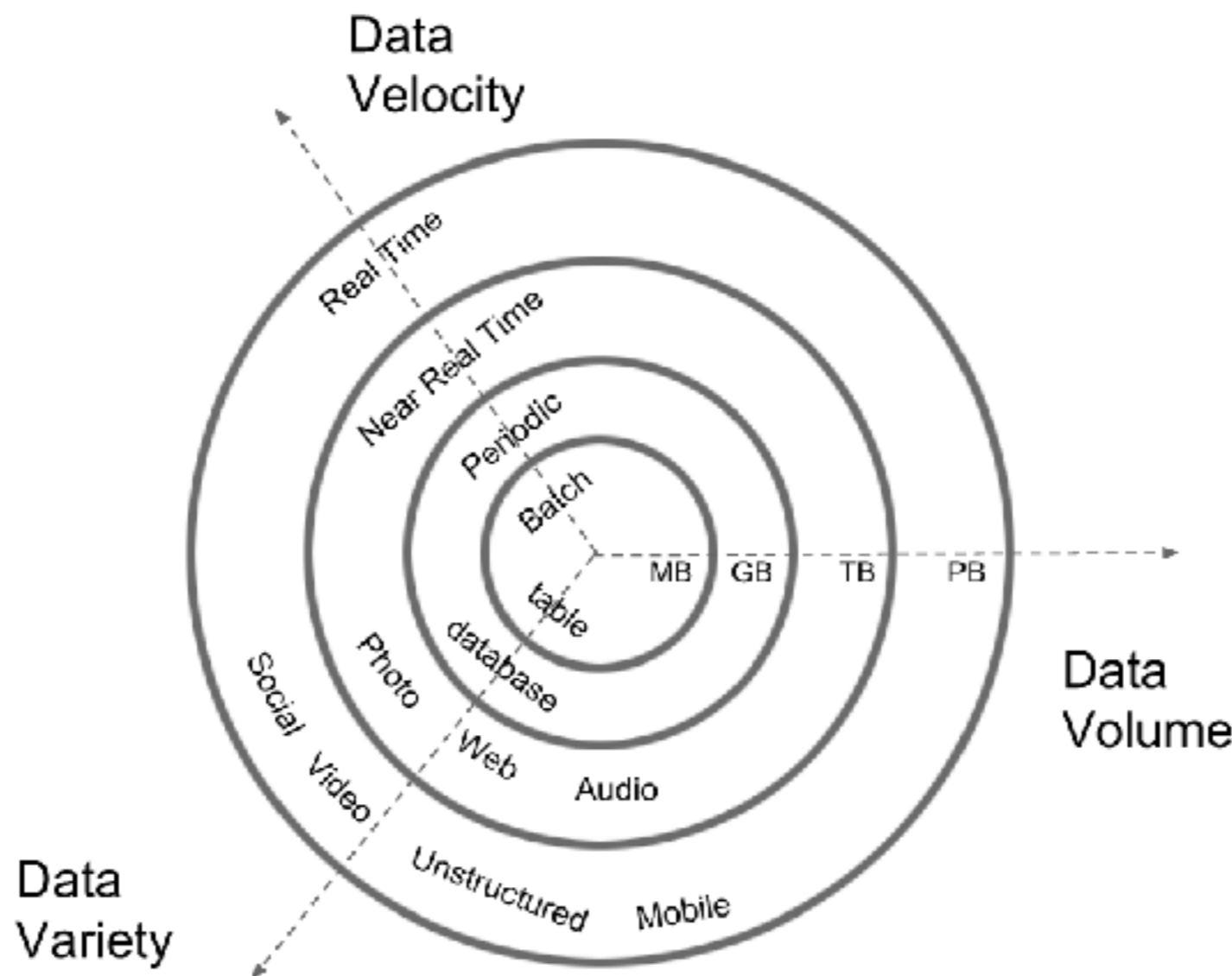
Refer to the vast amount of structured and unstructured data generated at high velocity, variety, and volume

Data is too complex to processed by traditional databases and tools

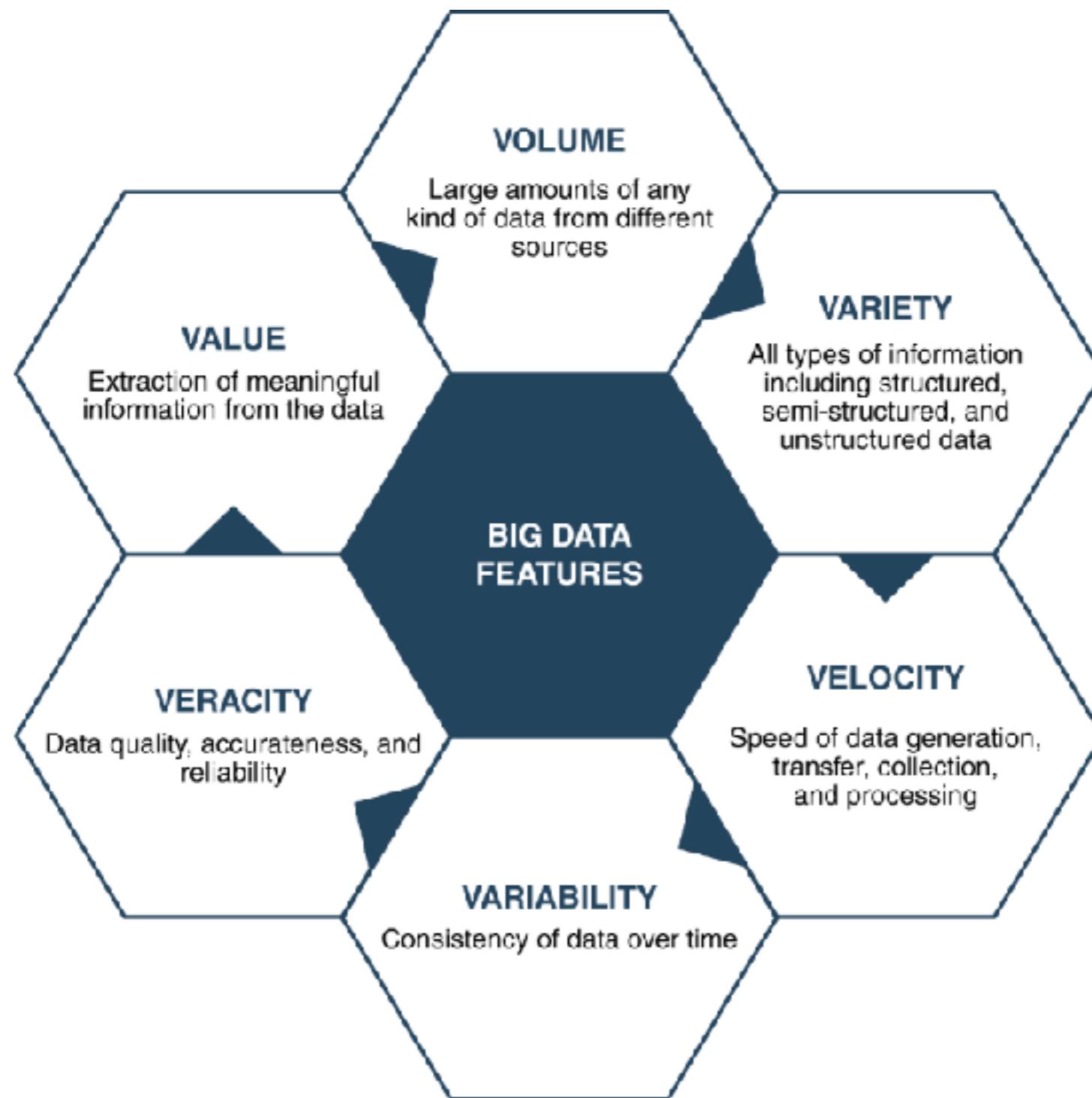


Characteristics for Big Data

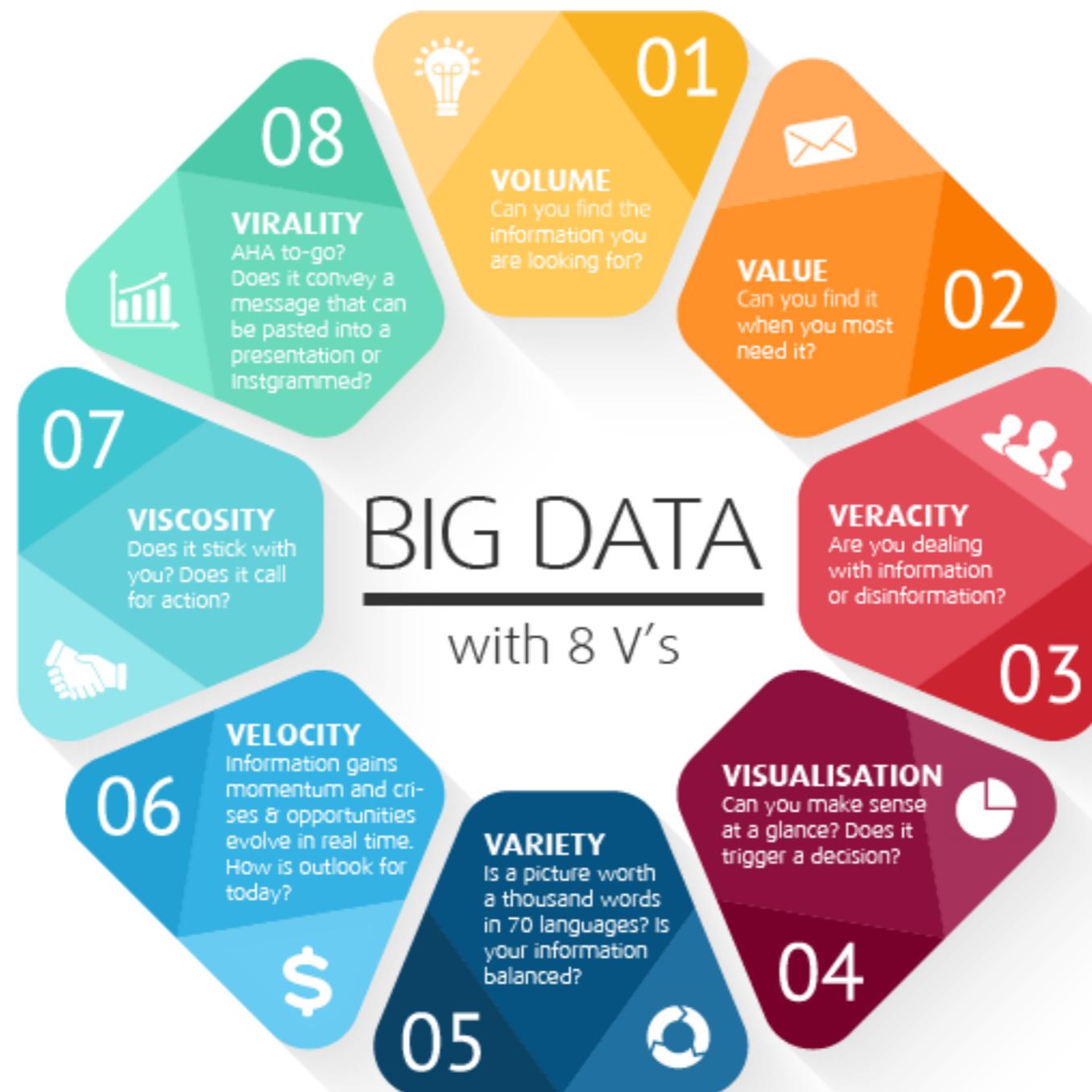
3 V's !!!



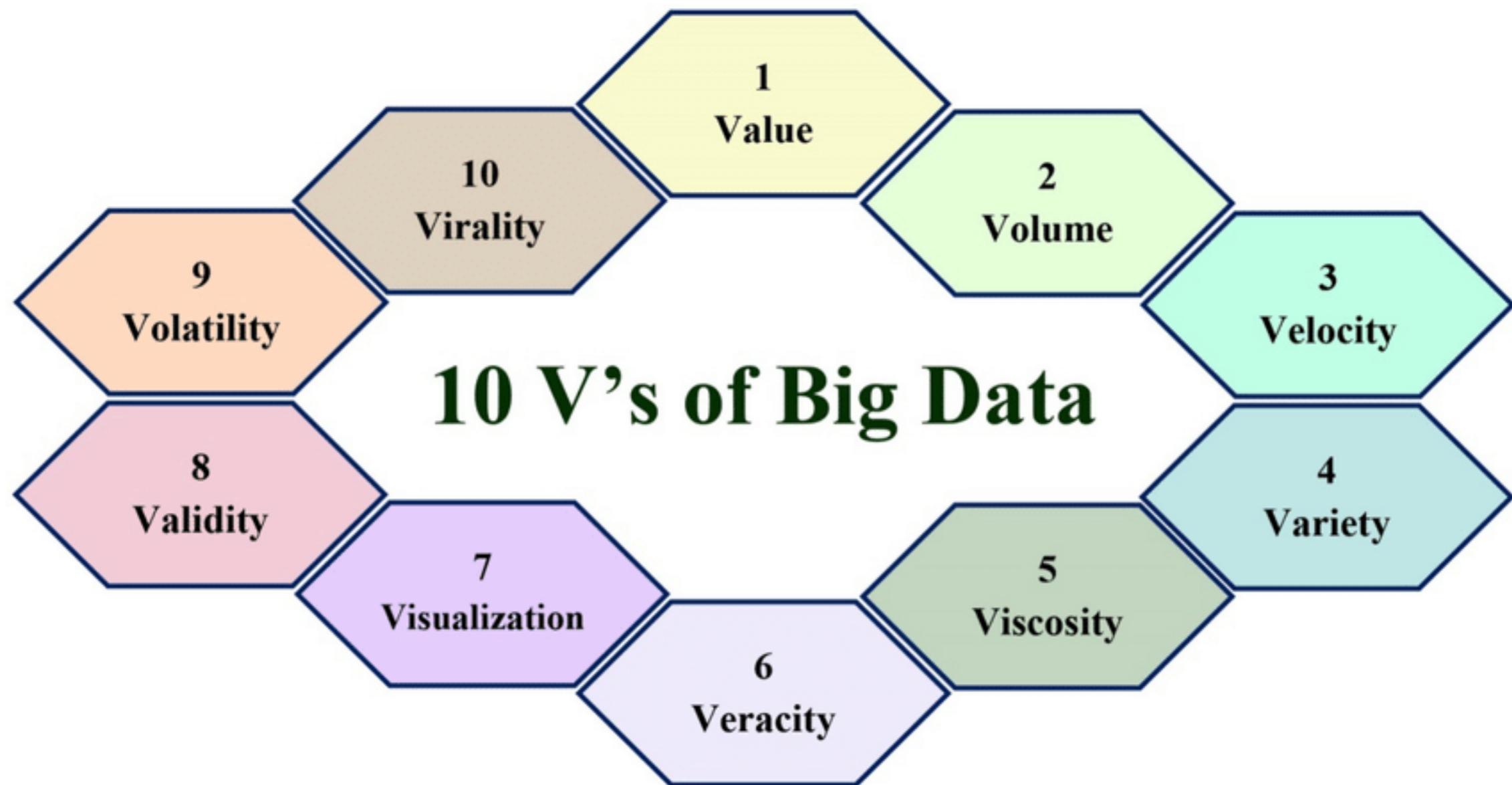
Characteristics for Big Data



Characteristics for Big Data



Characteristics for Big Data

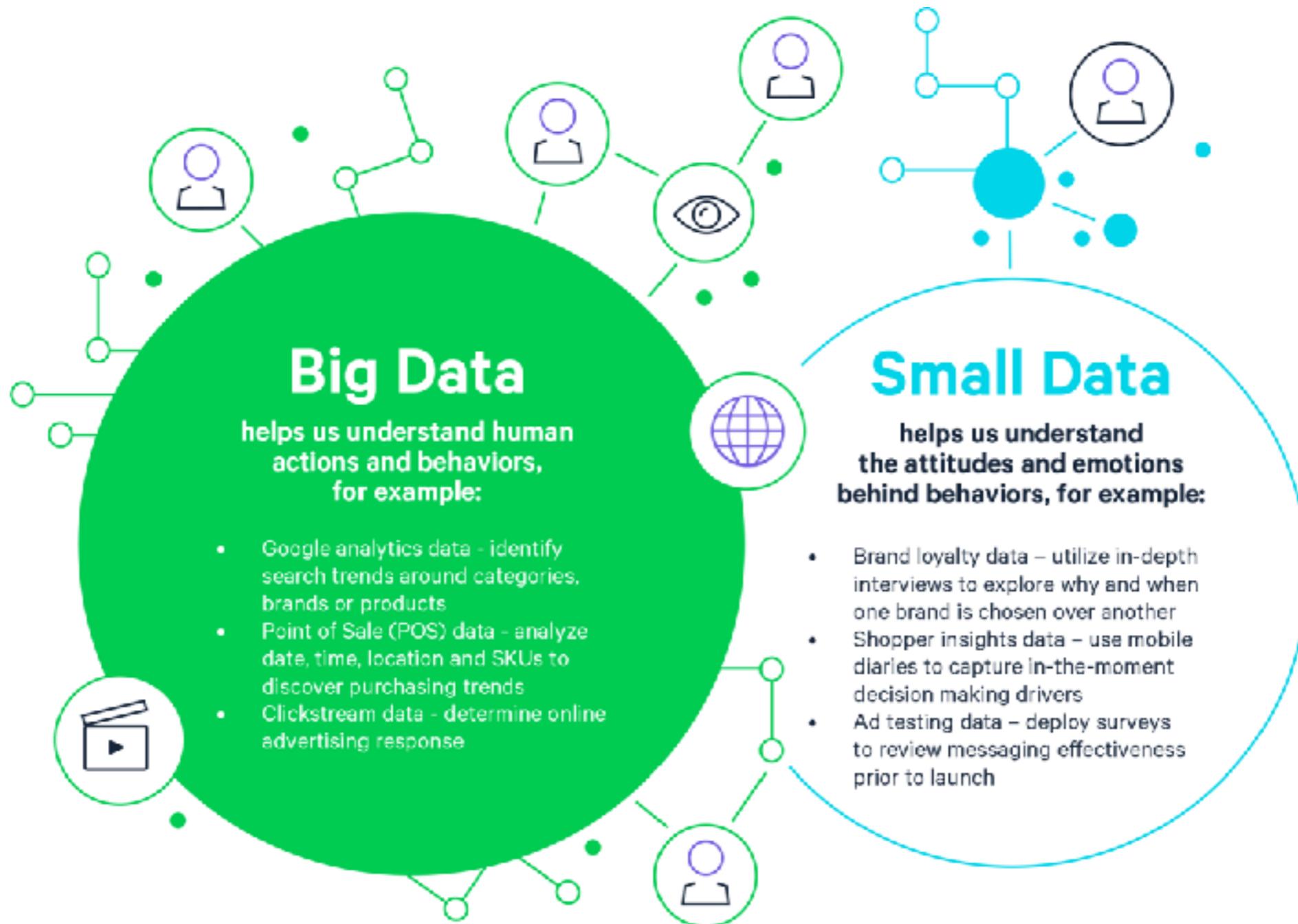


Purpose of Big Data

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



Big Data vs Small Data

Category	Small Data	Big Data
Data Sources (แหล่งข้อมูล)	Transaction แหล่งข้อมูลจากระบบงานต่างๆ	แหล่งข้อมูลนอกเหนือจากระบบงานที่มี เช่น Log, Social Data
Volume (จำนวนข้อมูล)	Megabytes (10^6) Gigabytes (10^9) Terabytes (10^{12})	Terabytes (10^{12}) Petabytes (10^{15}) Exabytes (10^{18}) Zettabytes (10^{21})
Velocity (ความต้องการใช้ข้อมูล)	Batch, Periodic, Near Real	Real Time
Variety (ความหลากหลาย)	Structure Data	Structure Data และ Unstructure Data
Value (คุณค่าที่ได้รับ)	Analysis, Reporting หรือ Business Intelligence	Predicts (ทำนายอนาคต), หรือ Insight ใช้ Data Mining ช่วย
View (การแสดงผล)	แสดงข้อมูลโดยไม่มีการสังเคราะห์ อาจเป็นการนำข้อมูลมาแสดงผล โดยการพิจารณาขานข่ายตามประเภทสินค้า หรือตามภูมิภาค	แสดงข้อมูลที่เกิดจากการสังเคราะห์ เช่น เกิดจากการทำ Data Mining เพื่อหาความเกี่ยวข้อง หาข้อมูลที่ซ่อนอยู่

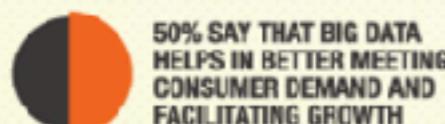
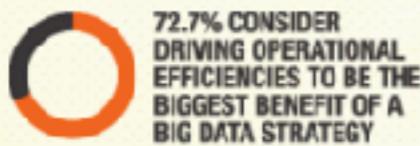
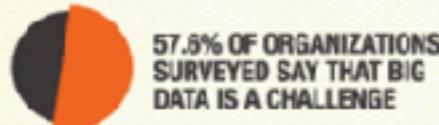


BIG DATA



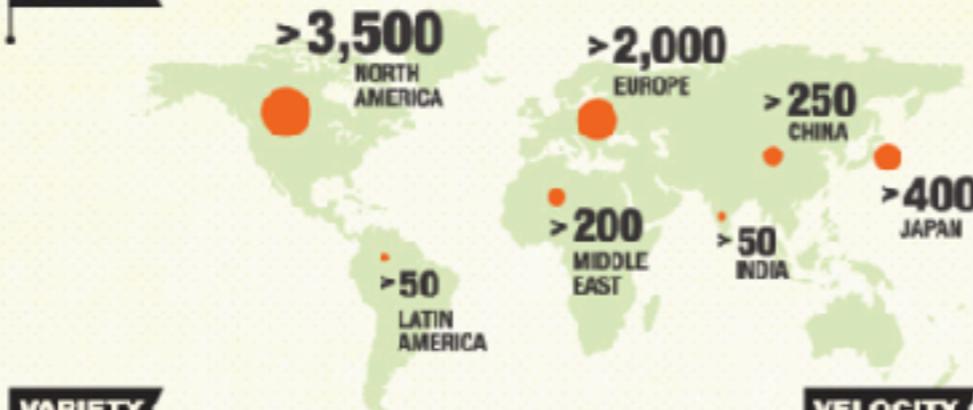
Big Data is data that is too large, complex and dynamic for any conventional data tools to capture, store, manage and analyze.

The right use of Big Data allows analysts to spot trends and gives niche insights that help create value and innovation much faster than conventional methods.



The “three V’s”, i.e the Volume, Variety and Velocity of the data coming in is what creates the challenge.

VOLUME



VARIETY



PEOPLE TO PEOPLE

NETIZENS, VIRTUAL COMMUNITIES, SOCIAL NETWORKS, WEB LOGS...



PEOPLE TO MACHINE

ARCHIVES, MEDICAL DEVICES, DIGITAL TV, E-COMMERCE, SMART CARDS, BANK CARDS, COMPUTERS, MOBILES...



MACHINE TO MACHINE

SENSORS, GPS DEVICES, BAR CODE SCANNERS, SURVEILLANCE CAMERAS, SCIENTIFIC RESEARCH...

VELOCITY



2.9 MILLION EMAILS SENT EVERY SECOND

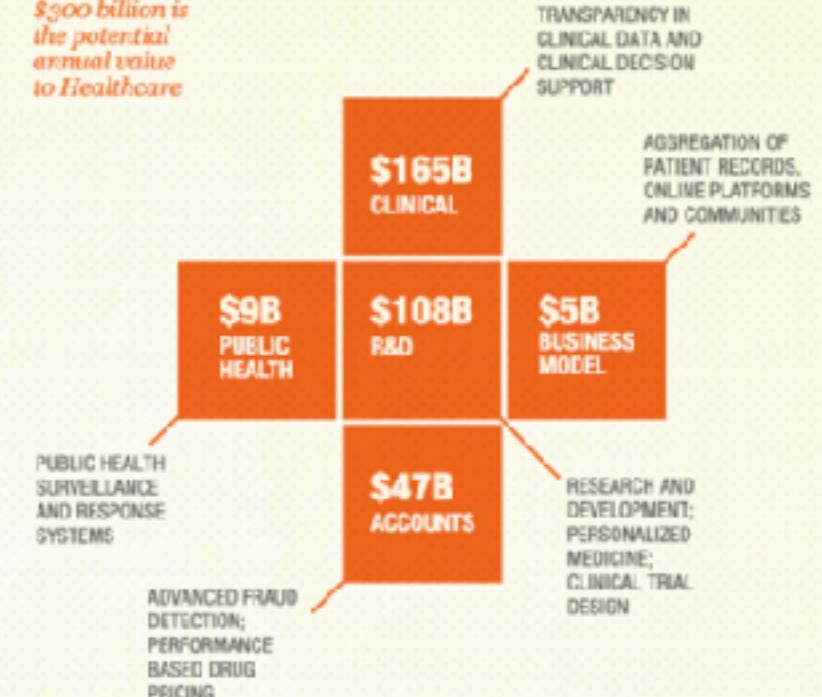


20 HOURS OF VIDEO UPLOADED EVERY MIN



CASE STUDY - Healthcare

\$300 billion is the potential annual value to Healthcare



VALUE



40% PROJECTED GROWTH IN GLOBAL DATA CREATED PER YEAR



The estimated size of the digital universe in 2011 was 1.8 zettabytes. It is predicted that between 2009 and 2020, this will grow 44 fold to 35 zettabytes per year. What is the right data management strategy for you, to successfully utilize Big Data?

Sources - ① Realizing the Rewards of Big Data - Wipro Report ② Big Data: The Next Frontier for Innovation, Competition and Productivity - McKinsey Global Institute Report ③ ComScore, Radicati Group ④ Measuring the Business Impacts of Effective Data - study by University of Texas/Austin ⑤ U.S. Department of Labor.

DO BUSINESS BETTER

WIPRO | OVER 100,000 EMPLOYEES | 54 COUNTRIES | CONSULTING | SYSTEM INTEGRATION | OUTSOURCING

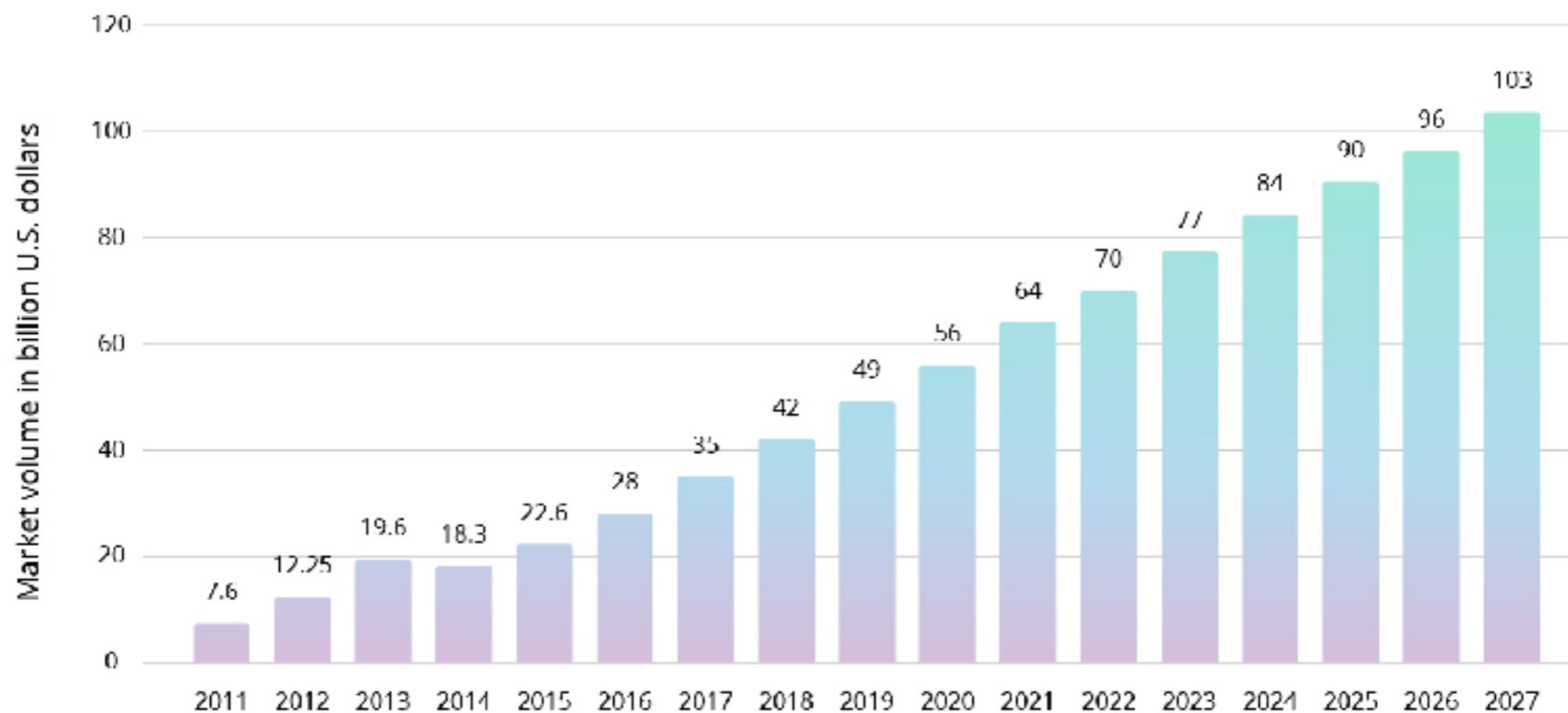


Sharing

17

Big Data Trend

Big data market size revenue forecast worldwide from 2011 to 2027
(in billion U.S. dollars)



<https://www.statista.com/statistics/254266/global-big-data-market-forecast/>



Big Data Challenges

Storage

Data Quality

Analysis

Cost vs Value

Accessibility

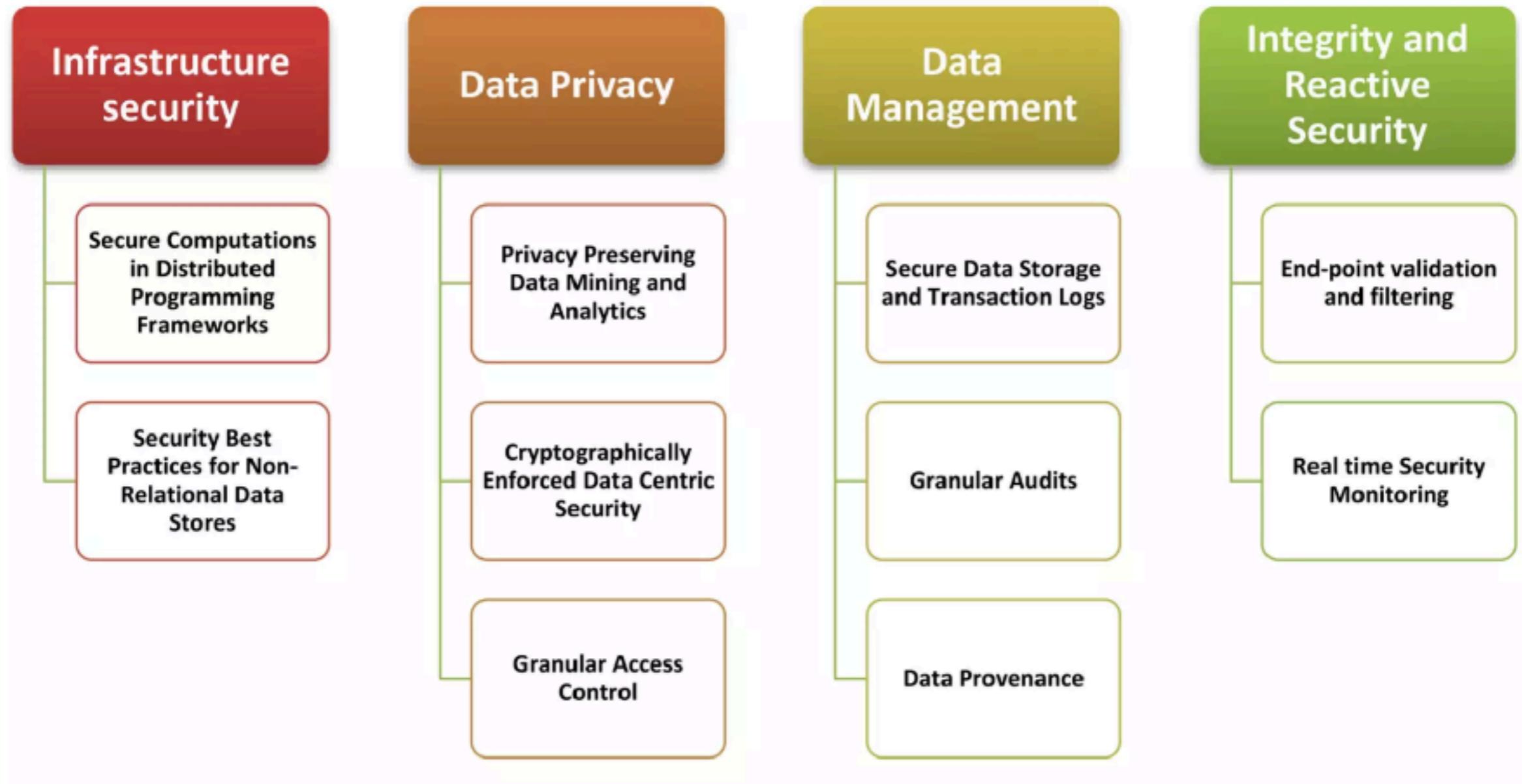
Security



Big Data Challenges



Big Data Security Challenges



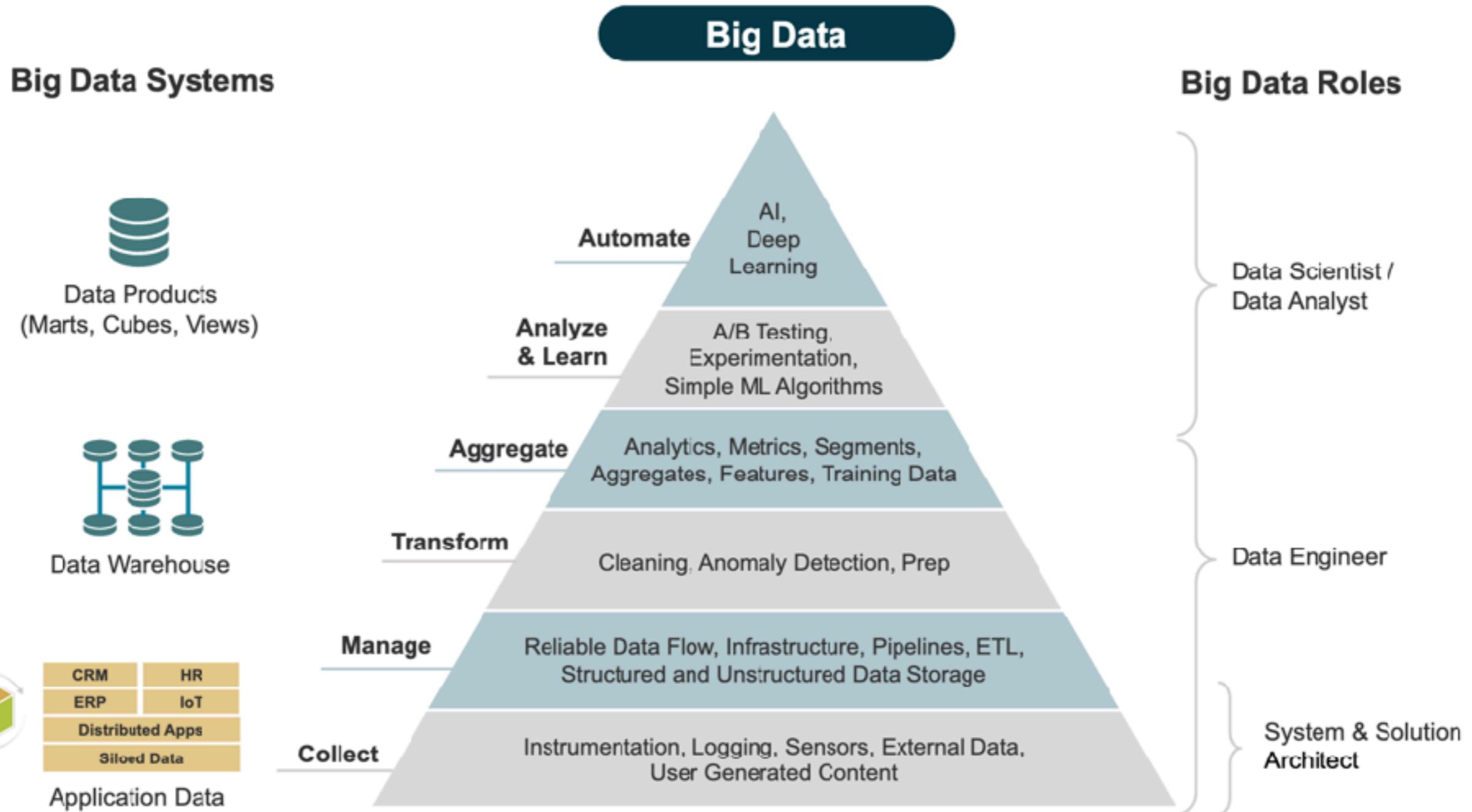
Data Governance



<https://intellias.com/future-big-data-trends/>



All about Big Data



© Scaled Agile, Inc.

<https://scaledagileframework.com/big-data/>

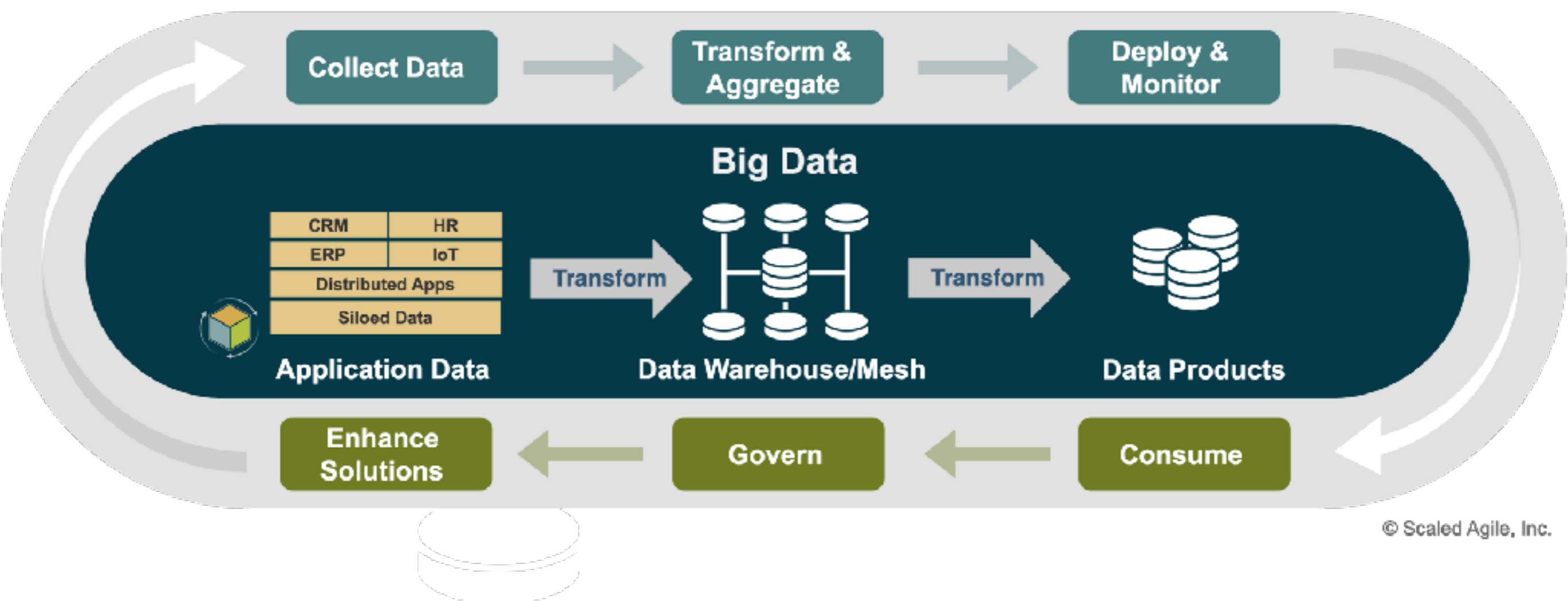


Sharing

© 2020 - 2024 Siam Chamnkit Company Limited. All rights reserved.

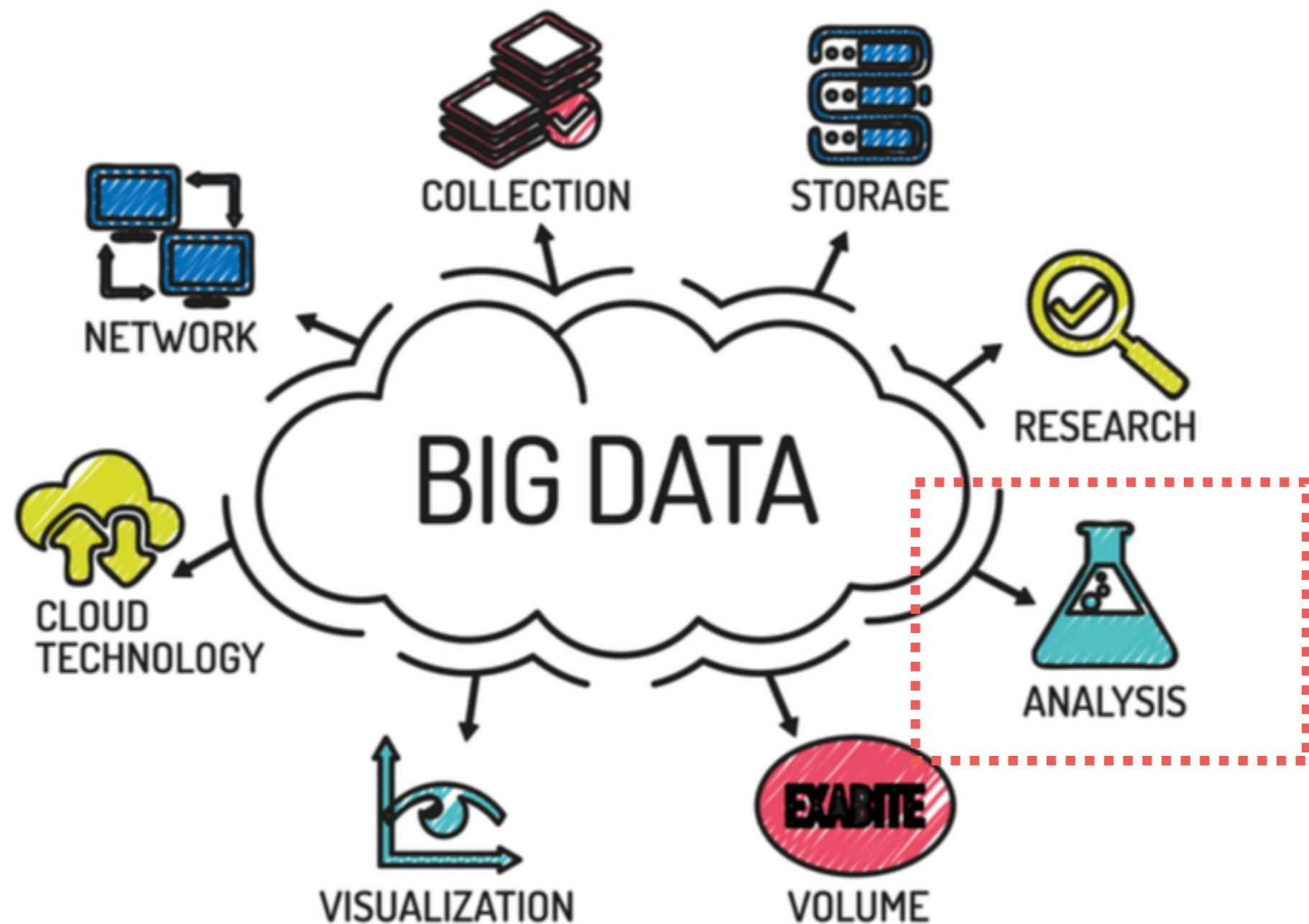
DataOps

Collaboration data management across teams

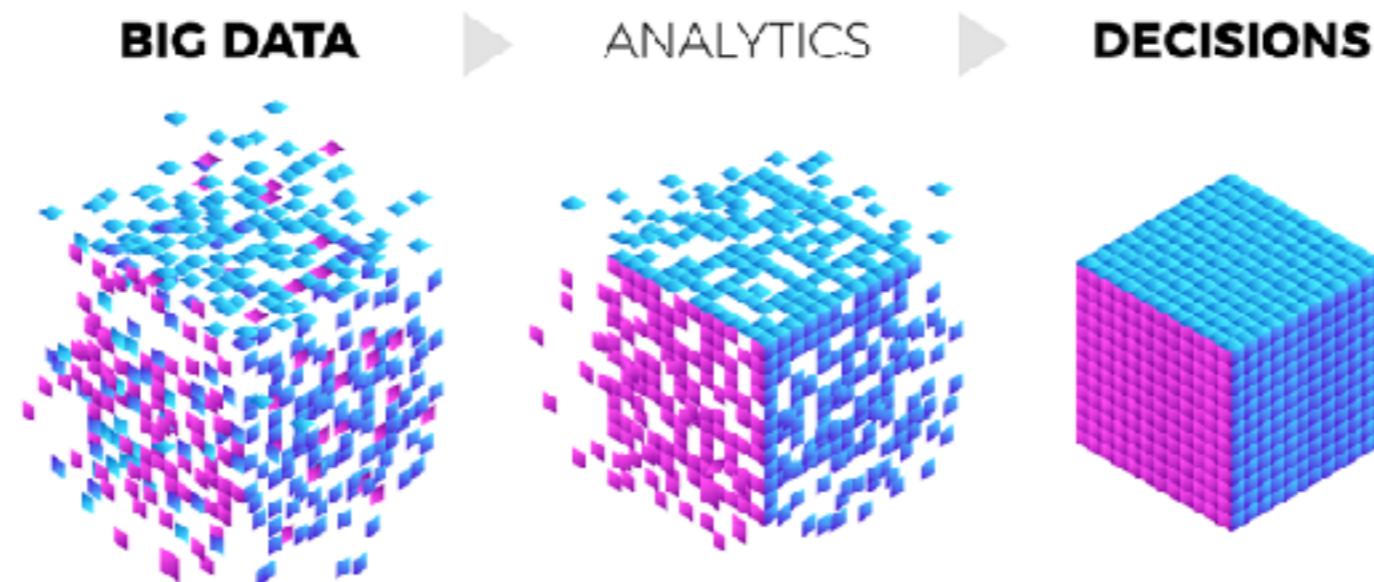


<https://scaledagileframework.com/big-data/>





Rise of Big Data Analytics



stargazr



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



Data-driven Decision



Make confident decisions



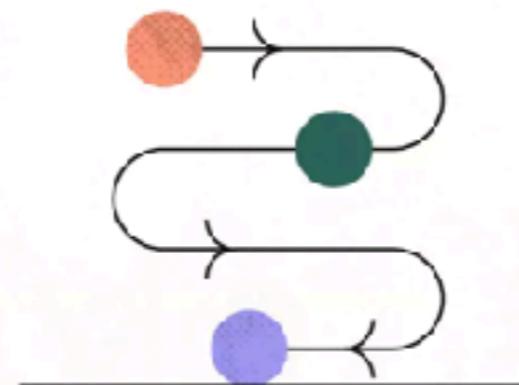
Guard against biases



Find unresolved questions



Set measurable goals



Improve company processes

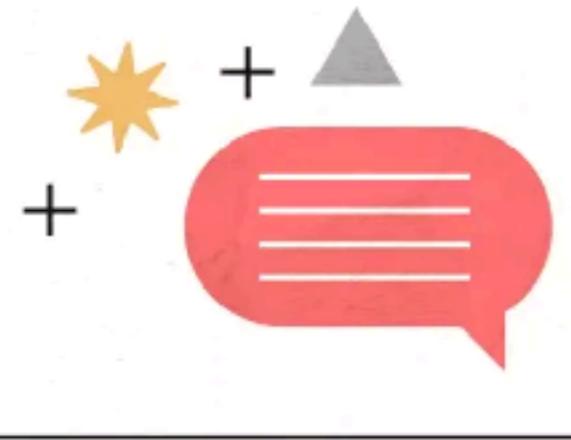
<https://asana.com/resources/data-driven-decision-making>



Sharing

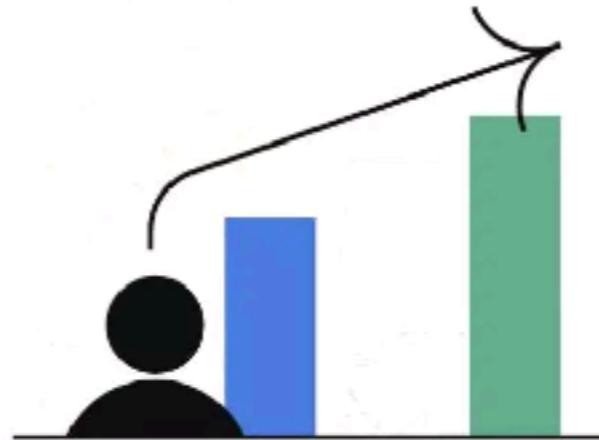
© 2020 - 2024 Siam Chamnankit Company Limited. All rights reserved.

Tips for Data-driven Decision



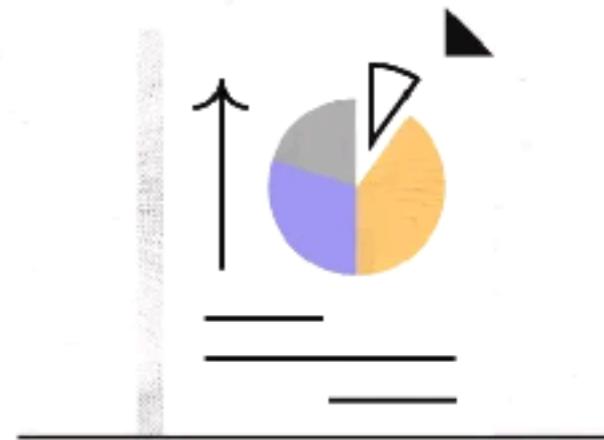
Find the story

When analyzing data, look for patterns and meanings behind the numbers and figures.



Consult the data

Before making any gut decisions, see if the facts align with your feelings.



Learn data visualization

Make sure your data is visually appealing and easy to understand.

<https://asana.com/resources/data-driven-decision-making>



Key Techniques

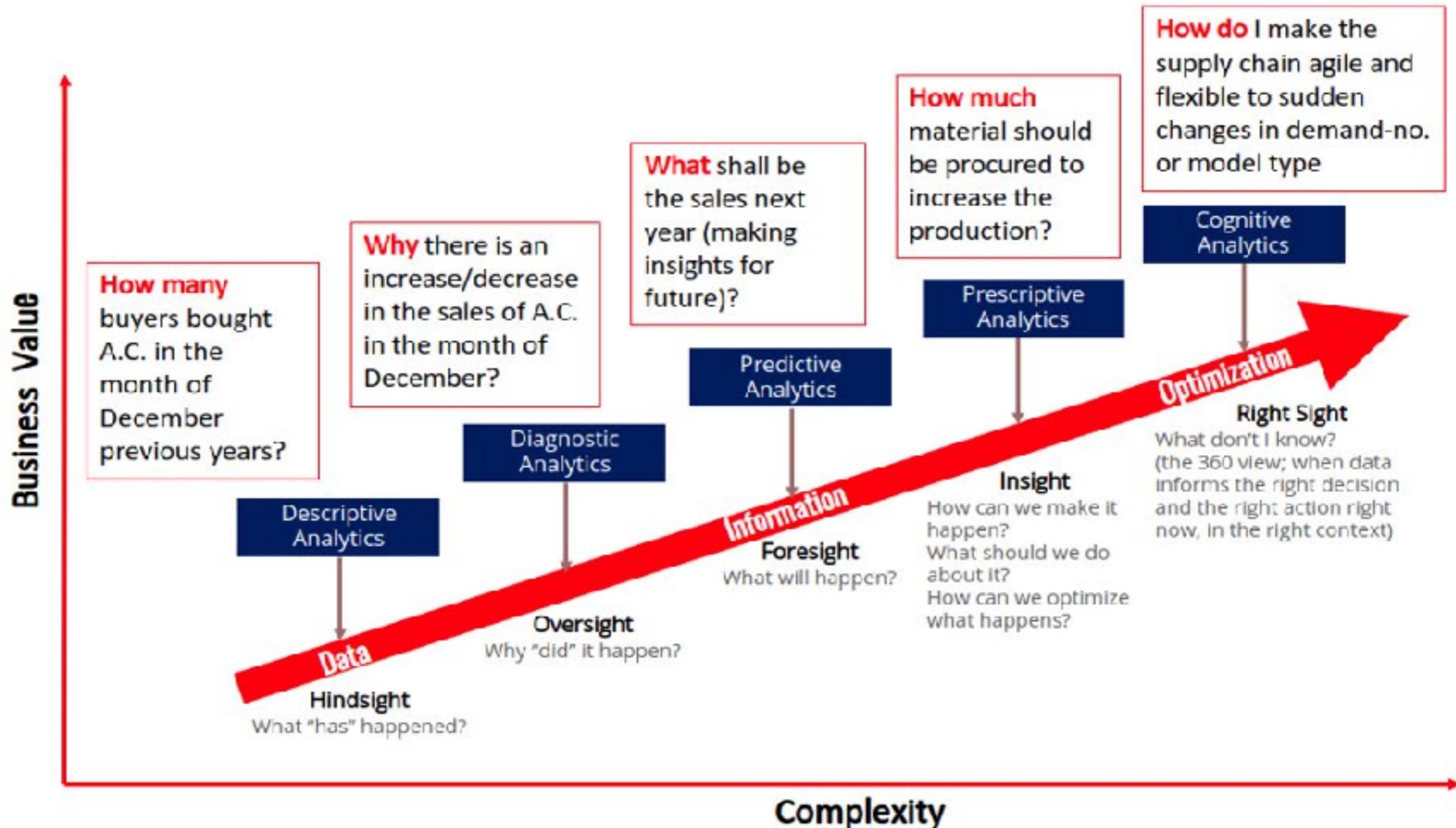
Predictive
Analytics

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



Predictive Analytics



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
Descriptive Analytics	What happened?	Understanding past performance.	Data aggregation, reporting, visualization	Sales reports showing performance by region.
Diagnostic Analytics	Why did it happen?	Identifying causes of past outcomes.	Drill-down analysis, correlation, root cause	Analyzing why sales declined in a certain area.
Predictive Analytics	What will happen?	Forecasting future trends and outcomes.	Machine learning, statistical models	Predicting customer churn for an online store.
Prescriptive Analytics	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.**



Start with Data Sources



Data Sources !!

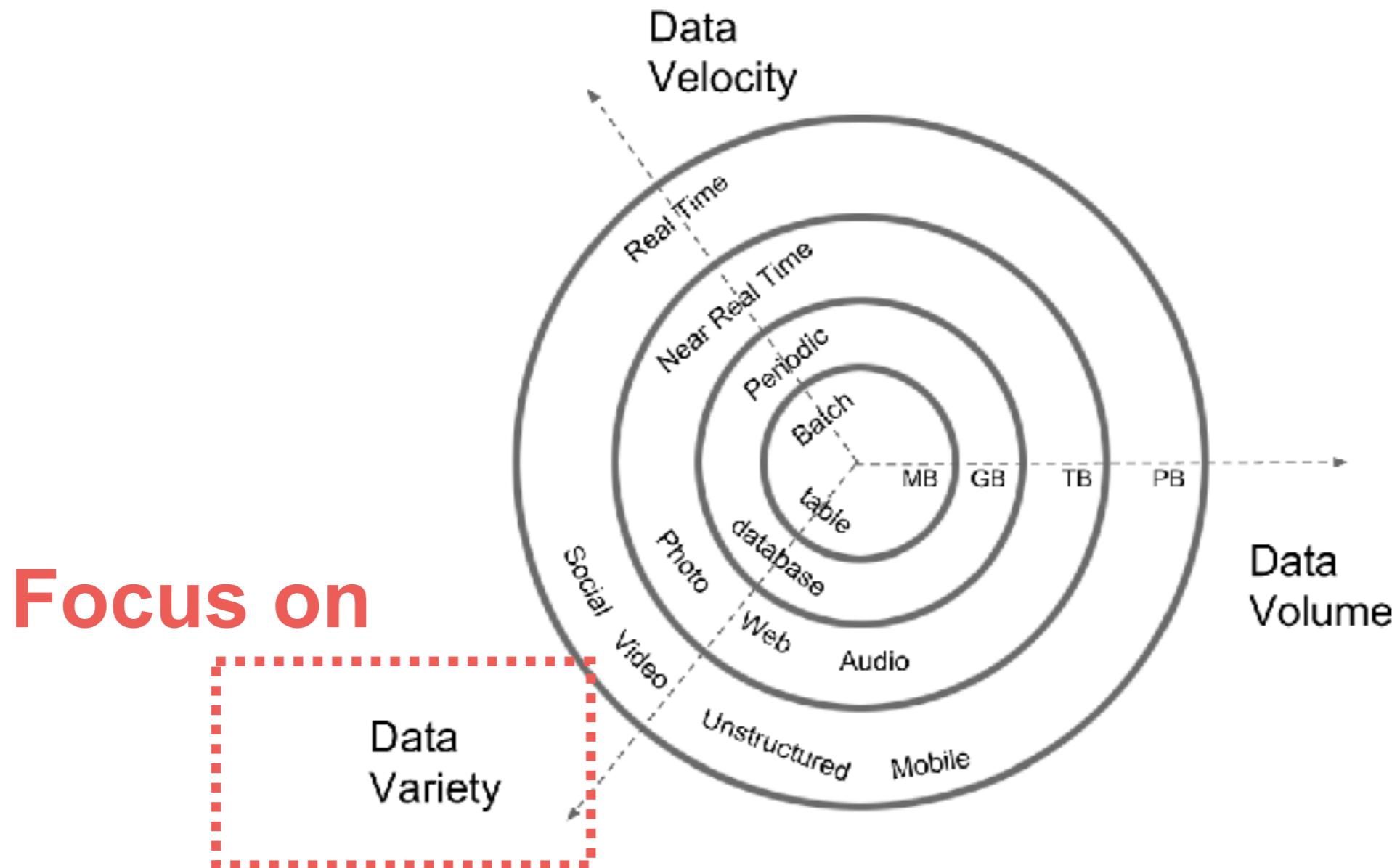


Sharing

© 2020 - 2024 Siam Chamnkit Company Limited. All rights reserved.

Characteristics for Big Data

3 V's !!!

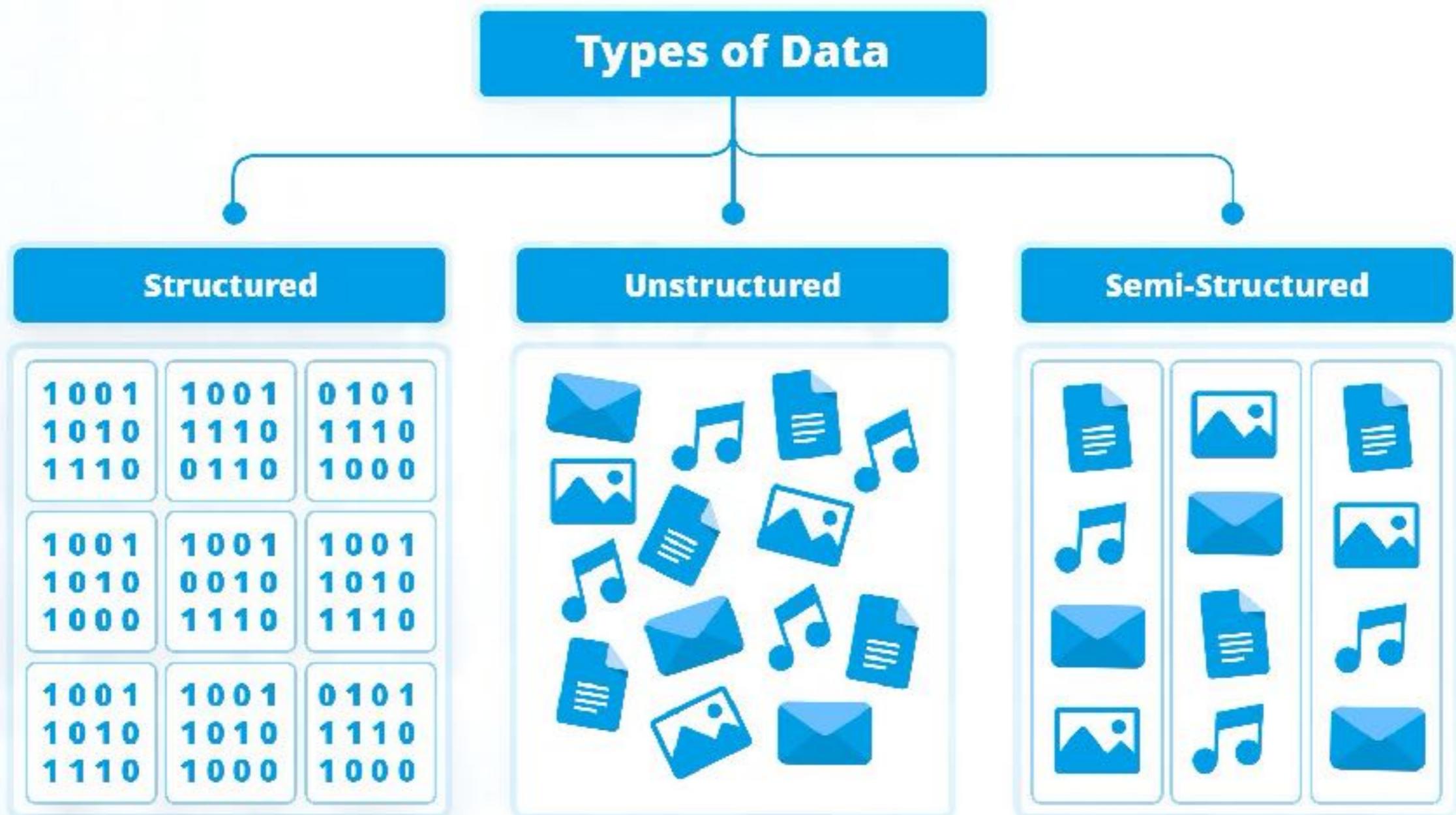


Types of data sources

Structured
Unstructured
Semi-structured



Types of data



Astera
Enabling Data-Driven Innovation

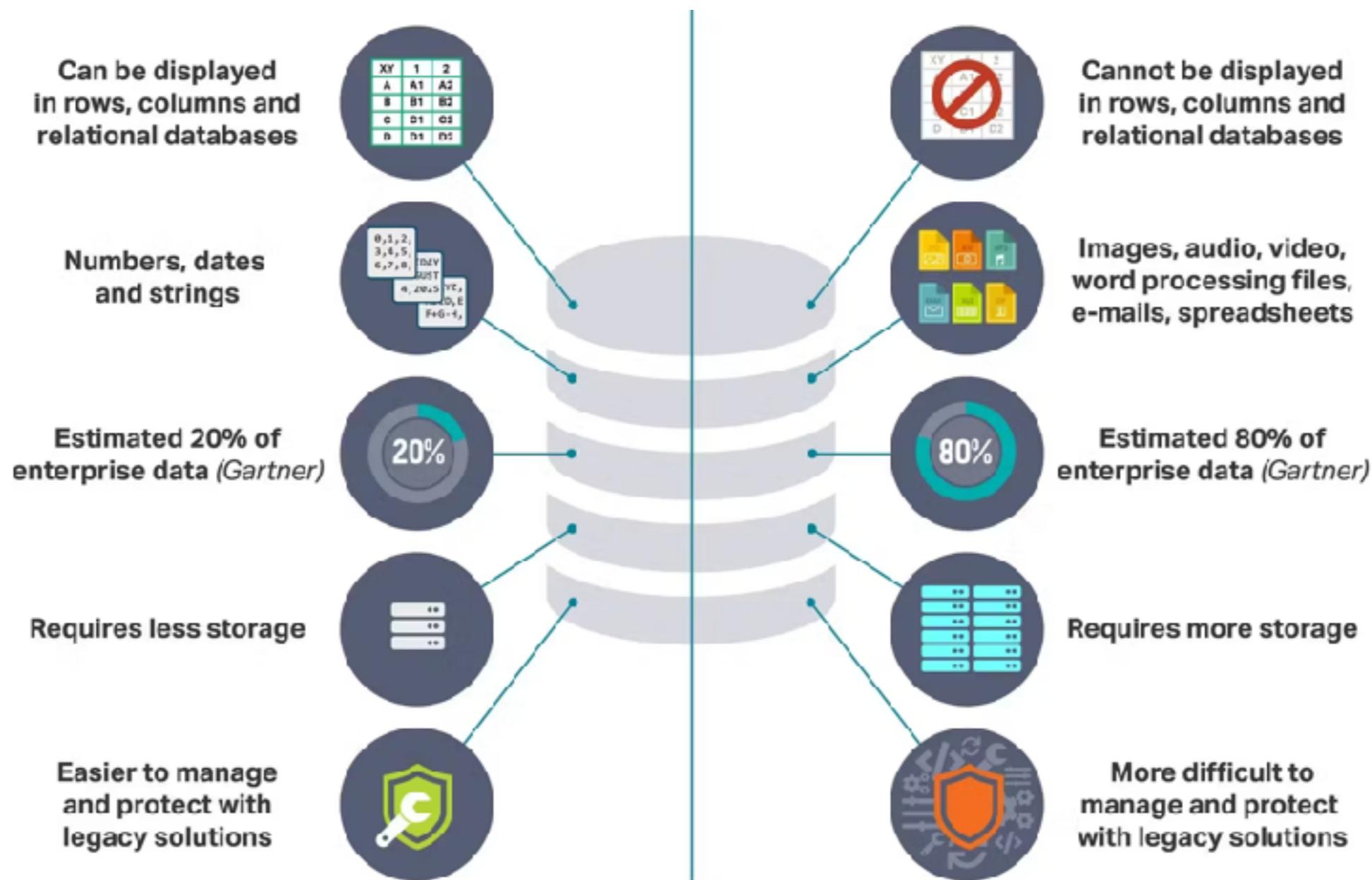


Types of data sources

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



Structured vs Unstructured data

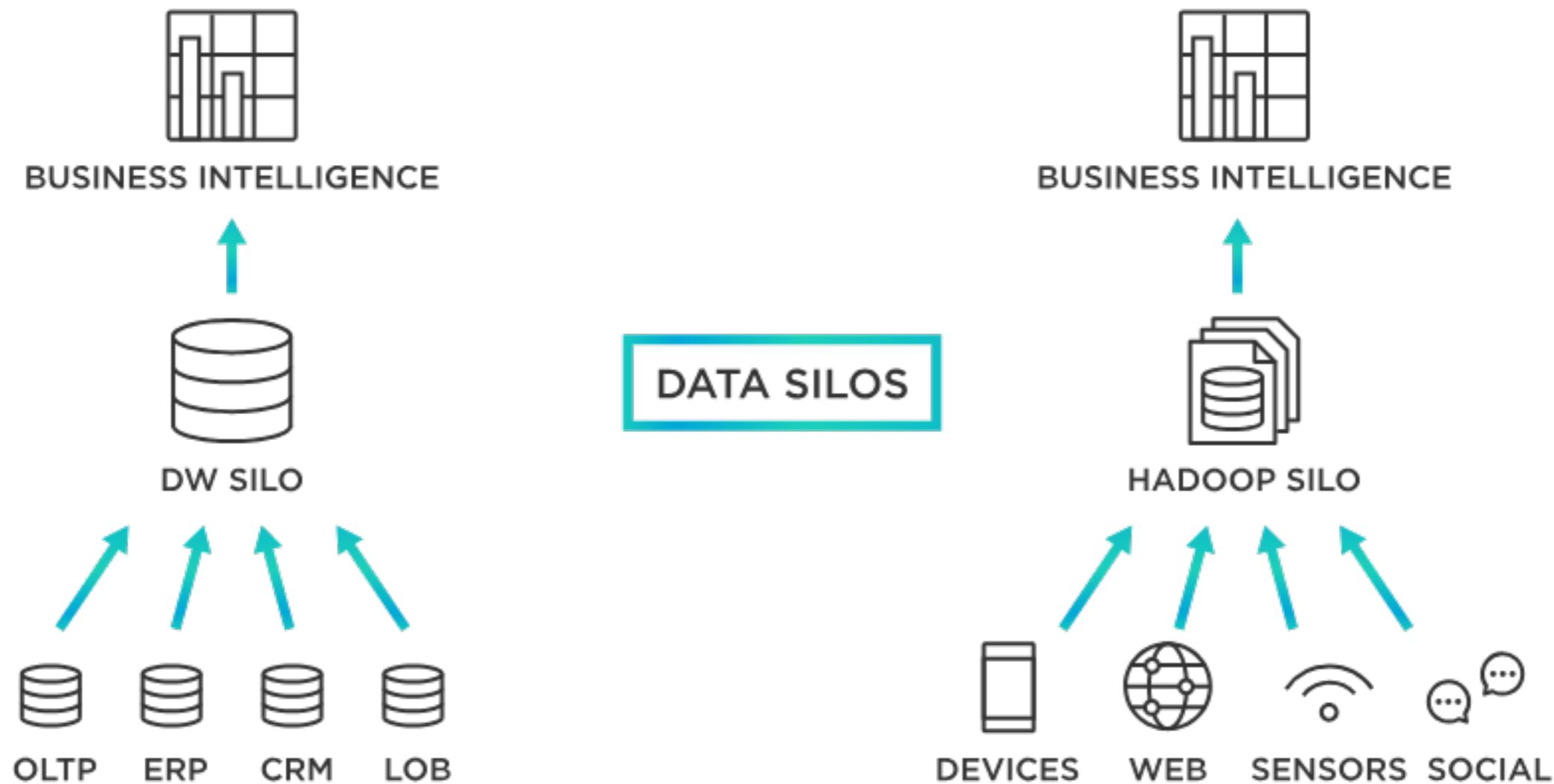


Unstructured Data Challenges

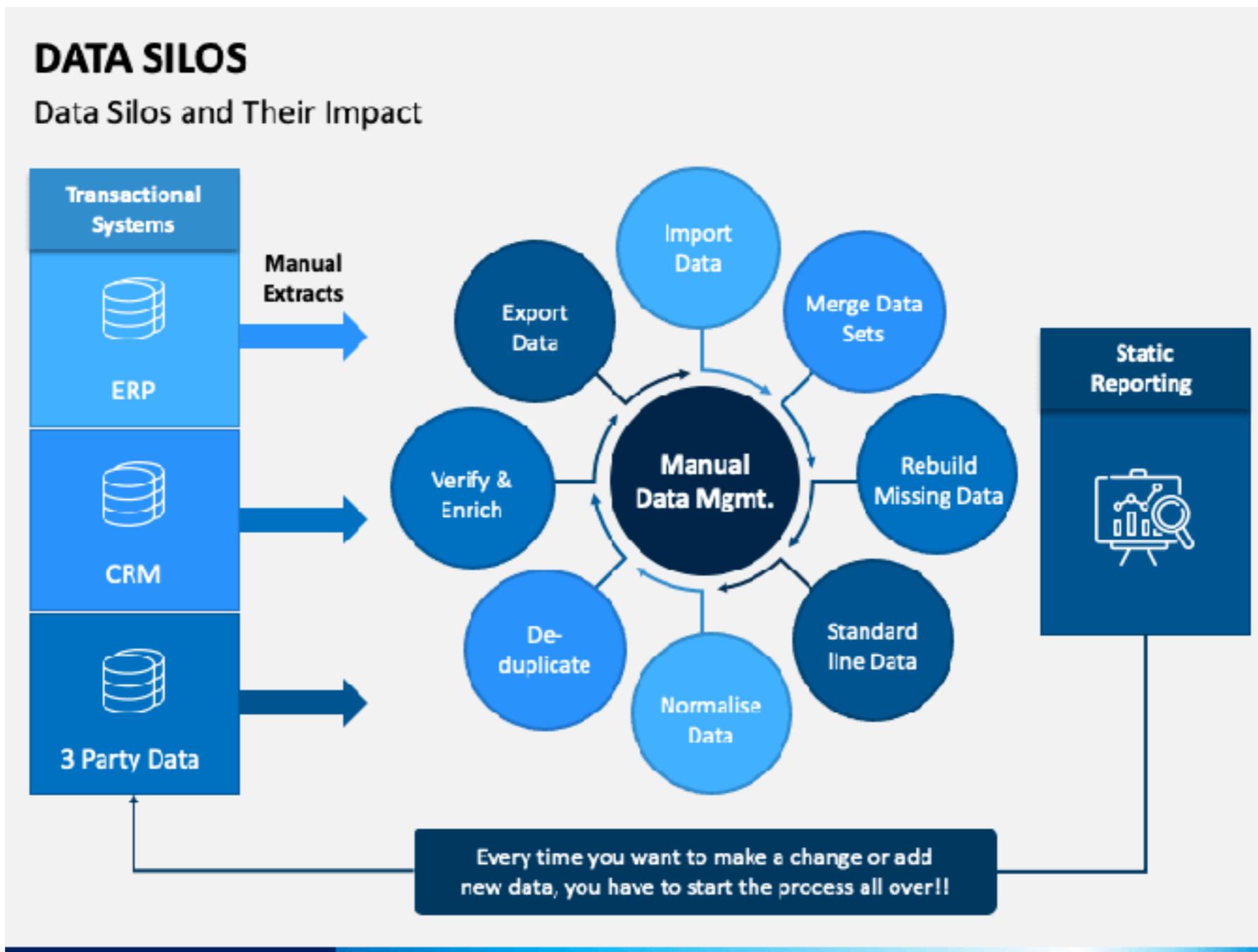
- Inability** to process growing data volumes
- Accessing **siloed** data
- Regulatory non-compliance
- Reduced data **usability**
- Increased vulnerability to cyber attacks



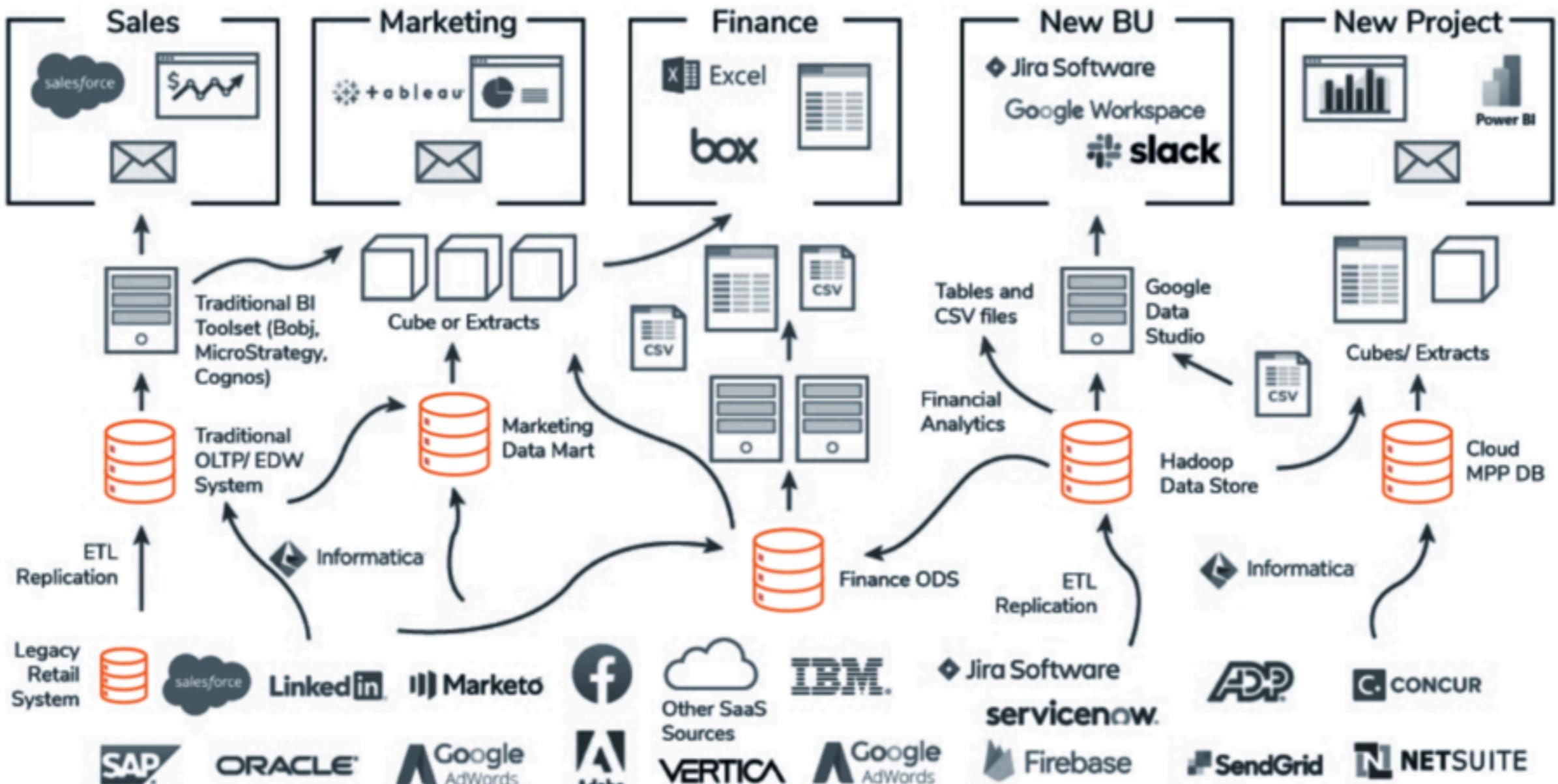
Data siloed ?



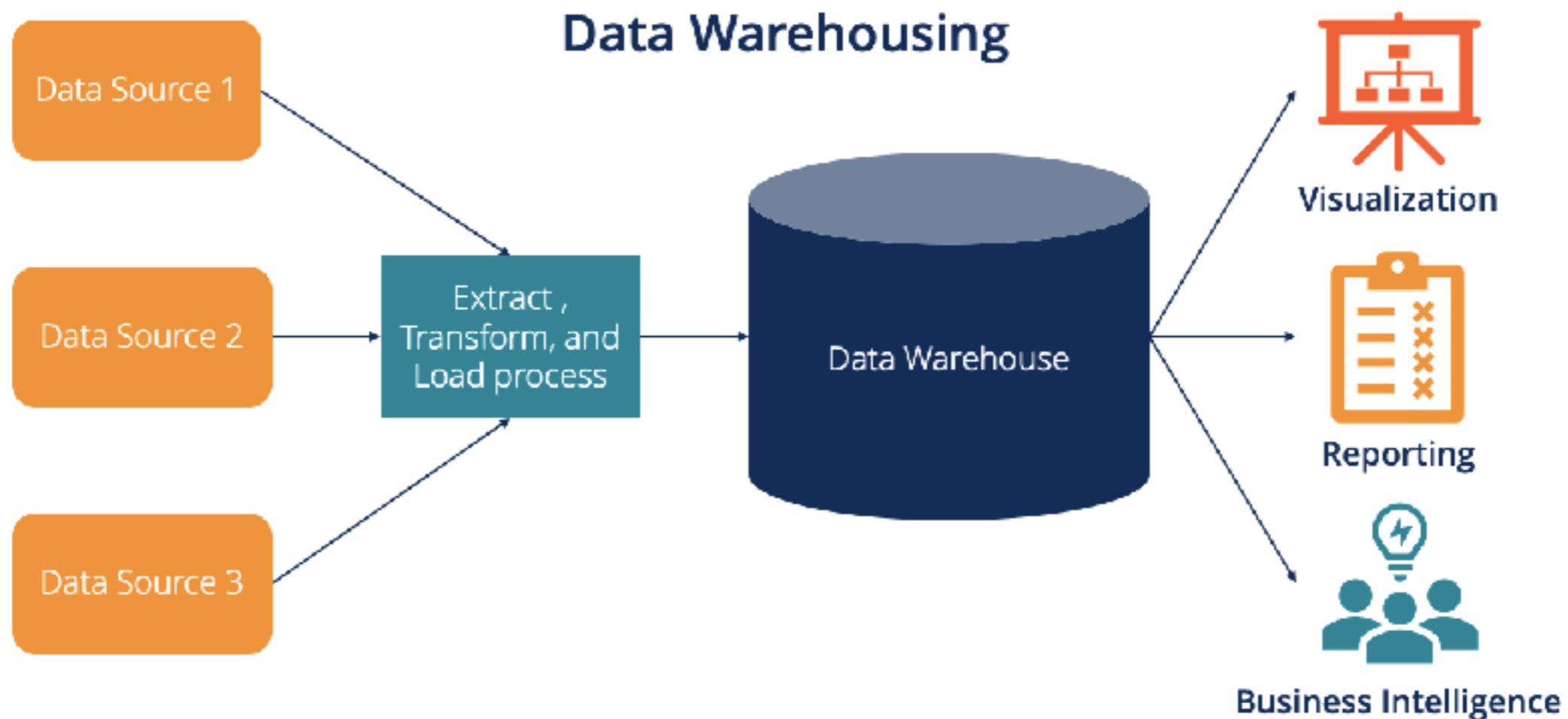
Data siloed ?



Data siloed ?

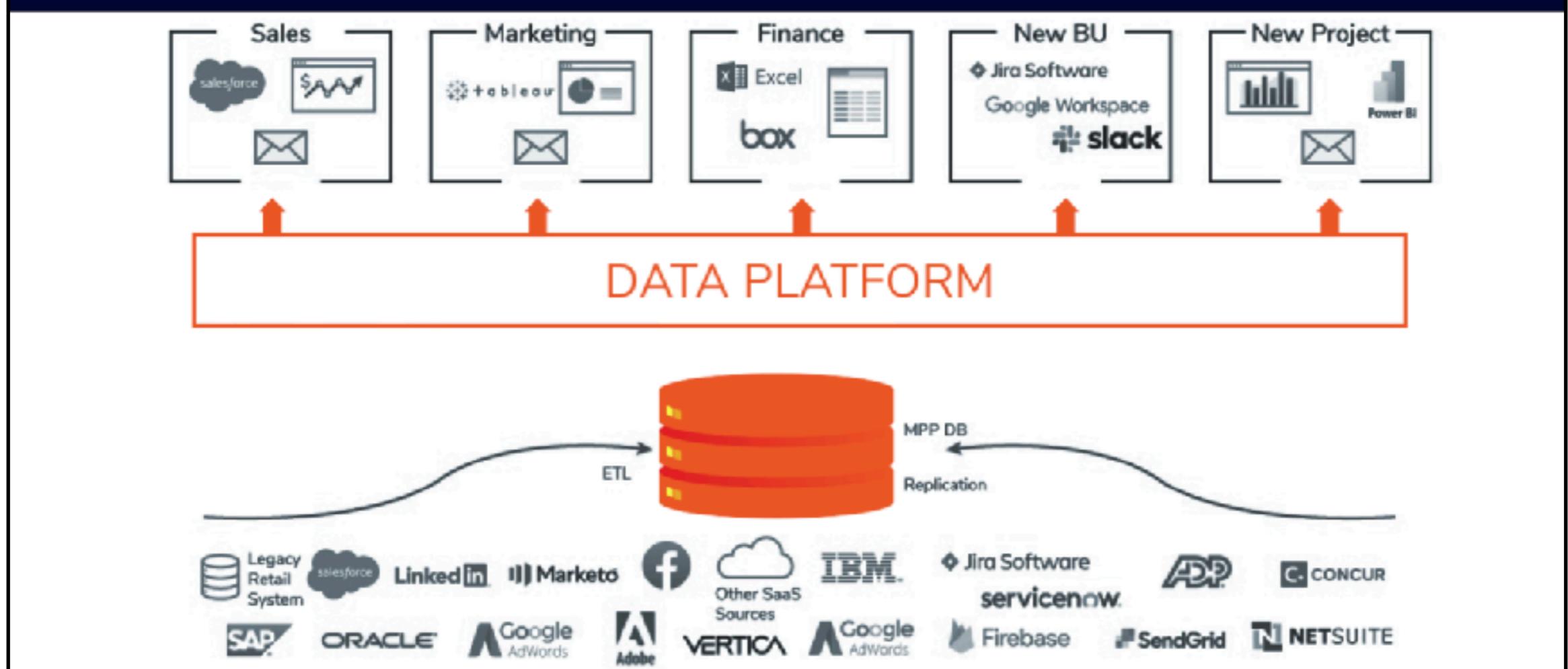


Try to solve with data platform



Data Platform

We clean it up



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**.



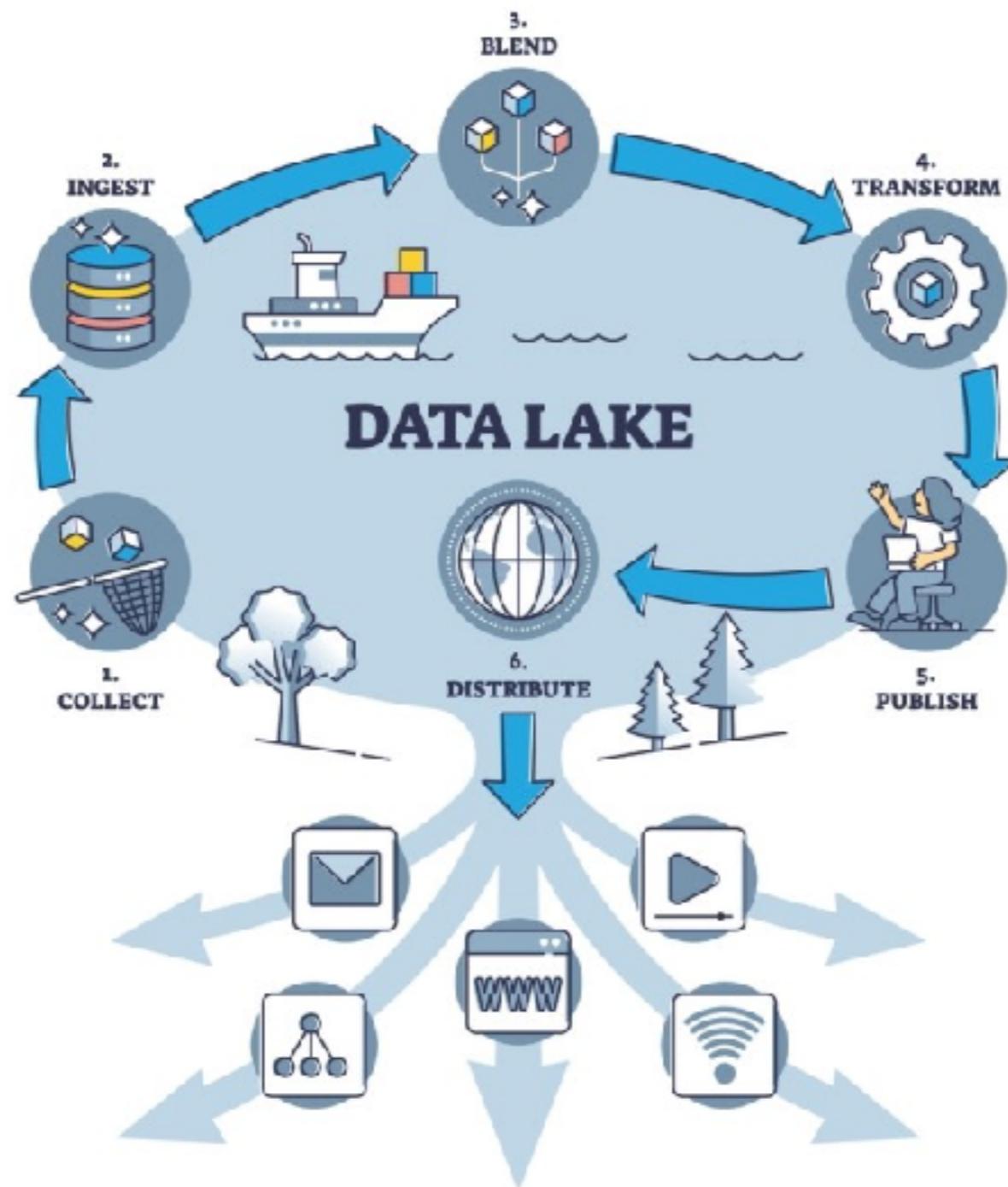
Data Lake

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.



Data Lake

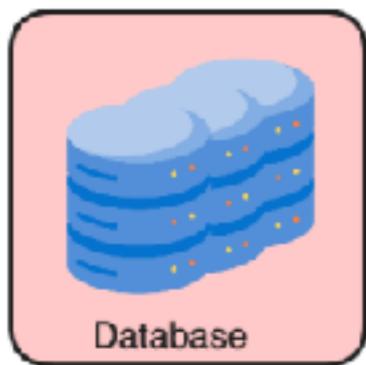


Key differences

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



Different between data types



VS



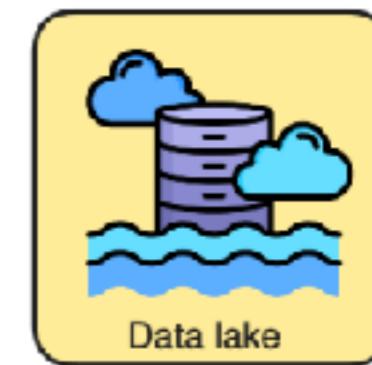
Data Warehouse

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)

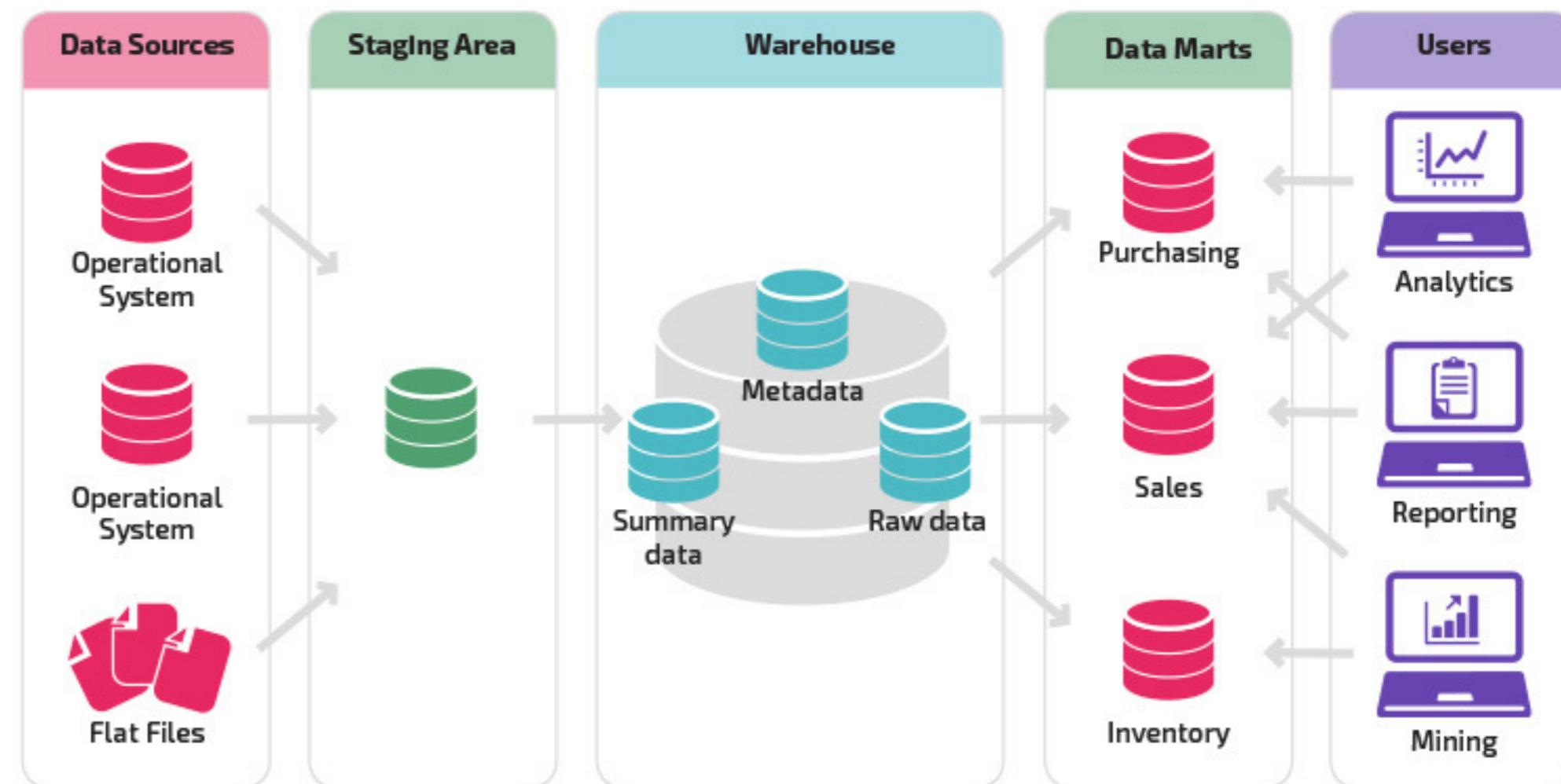
Business intelligence,
historical
analysis(OLAP).

Specific business function
analysis

Big data analytics,
data exploration.



Data Warehouse vs Data Mart



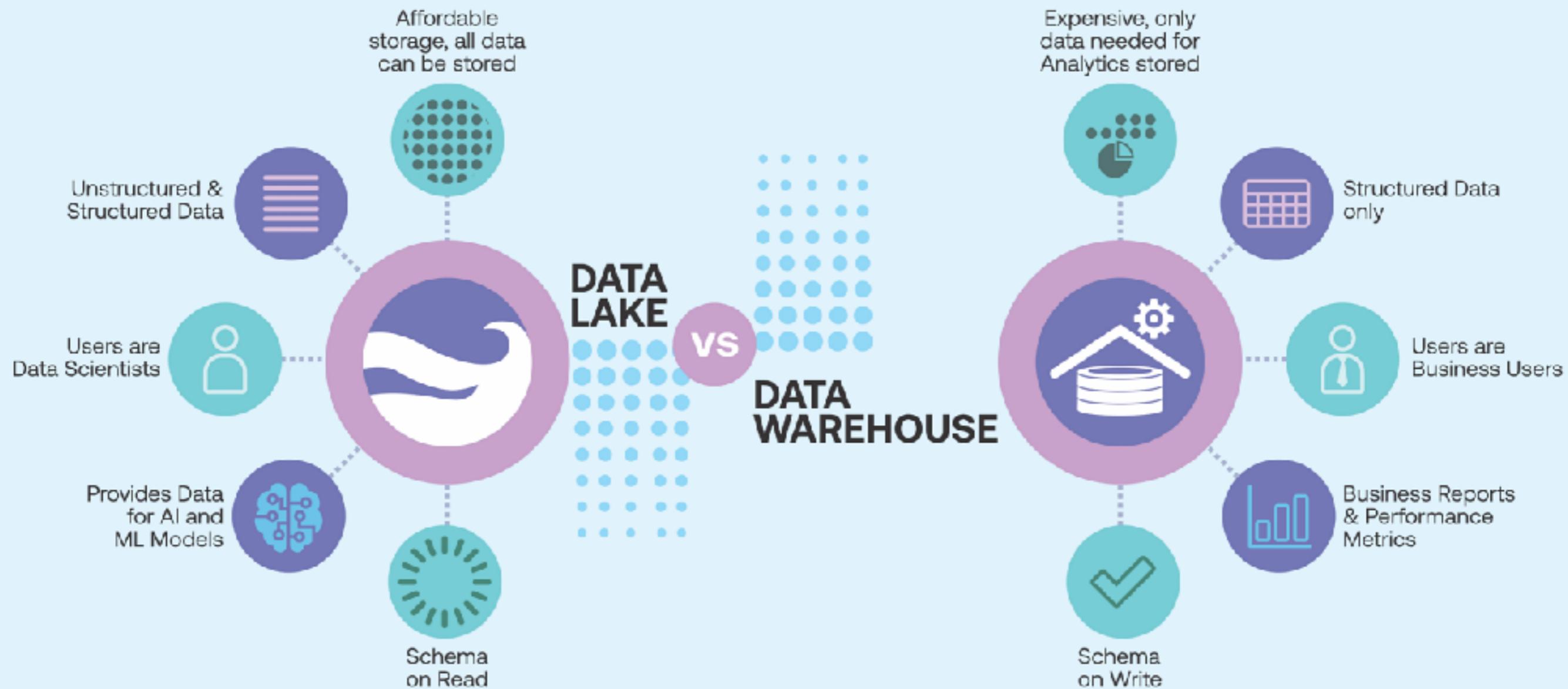
<https://panoply.io/data-warehouse-guide/data-mart-vs-data-warehouse/>



Sharing

© 2020 - 2024 Siam Chamnankit Company Limited. All rights reserved.

Data Warehouse vs Data Lake



Copyright © 2021 www.BryteFlow.com. All Rights Reserved.

<https://panoply.io/data-warehouse-guide/data-mart-vs-data-warehouse/>



Sharing

© 2020 - 2024 Siam Chamnkit Company Limited. All rights reserved.

Data pipeline

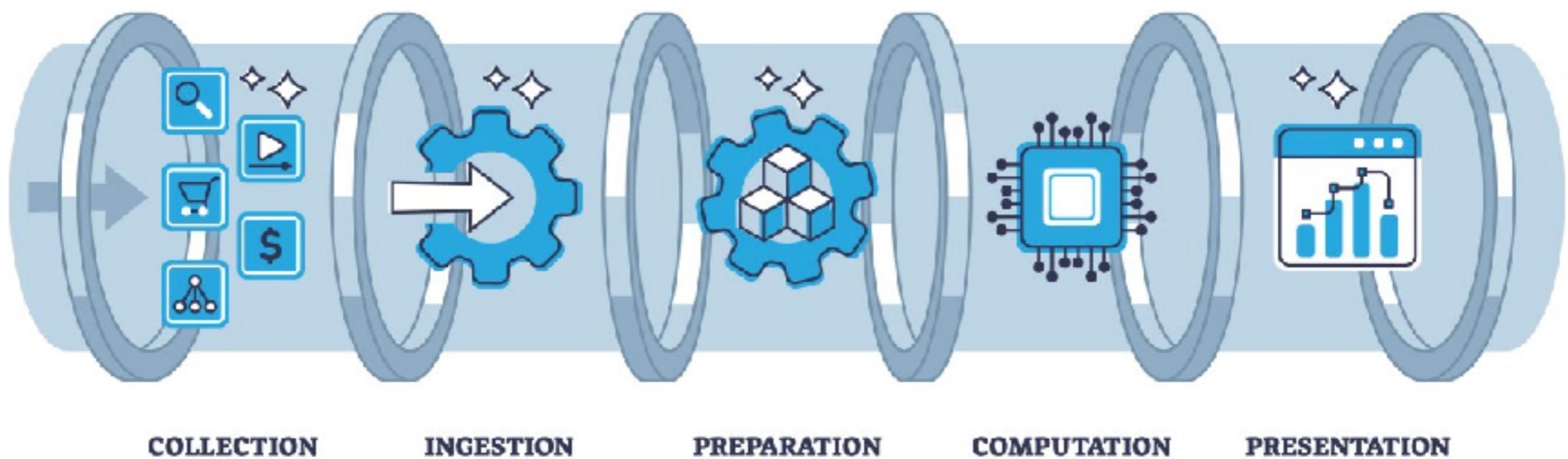


Data pipeline

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.



Data pipeline



Data Ingestion

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

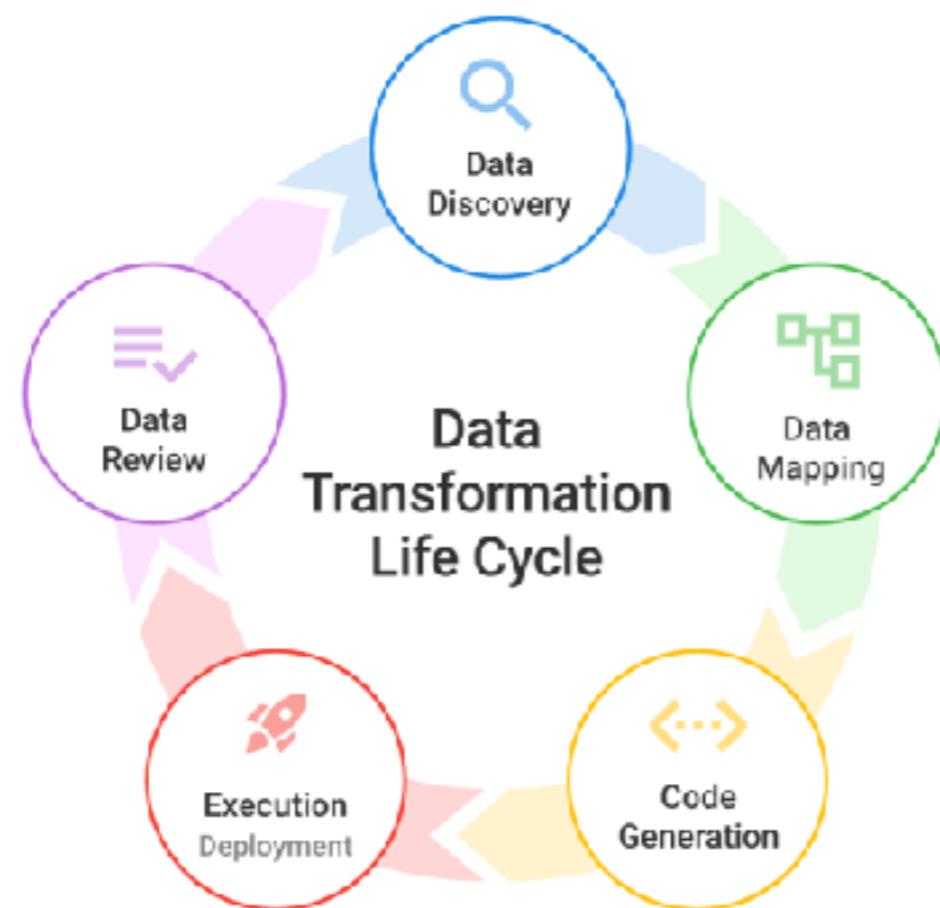
Batch
processing

Real-time
processing



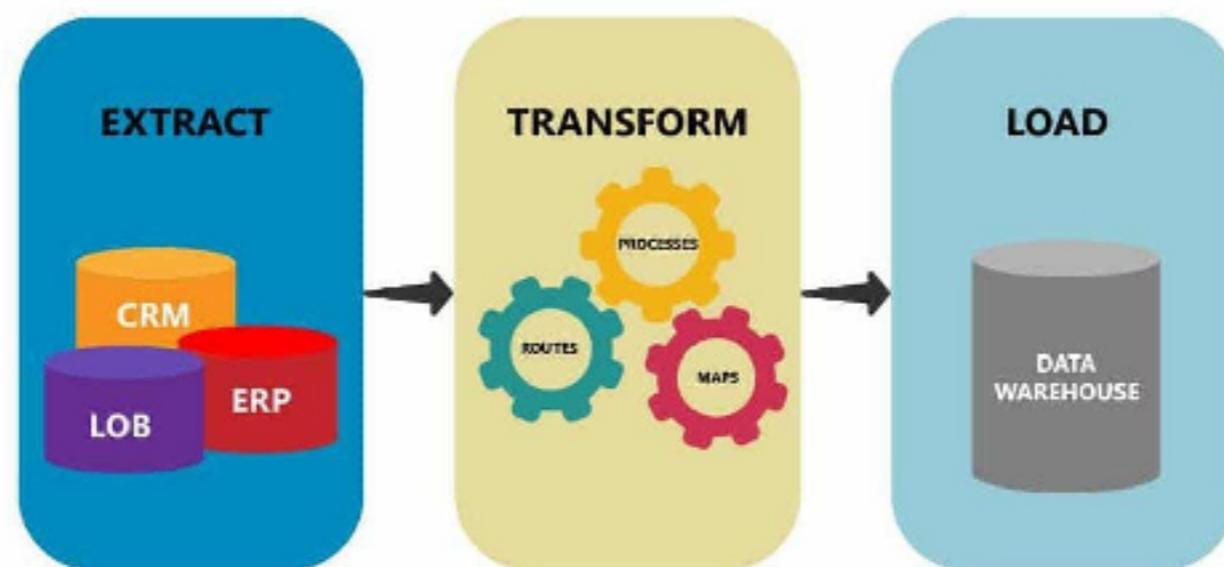
Data Transformation

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



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.



How to cleaning data ?



How to cleaning data ?



Key Techniques with MS Excel

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



Workshop



Big Data Analysis Techniques

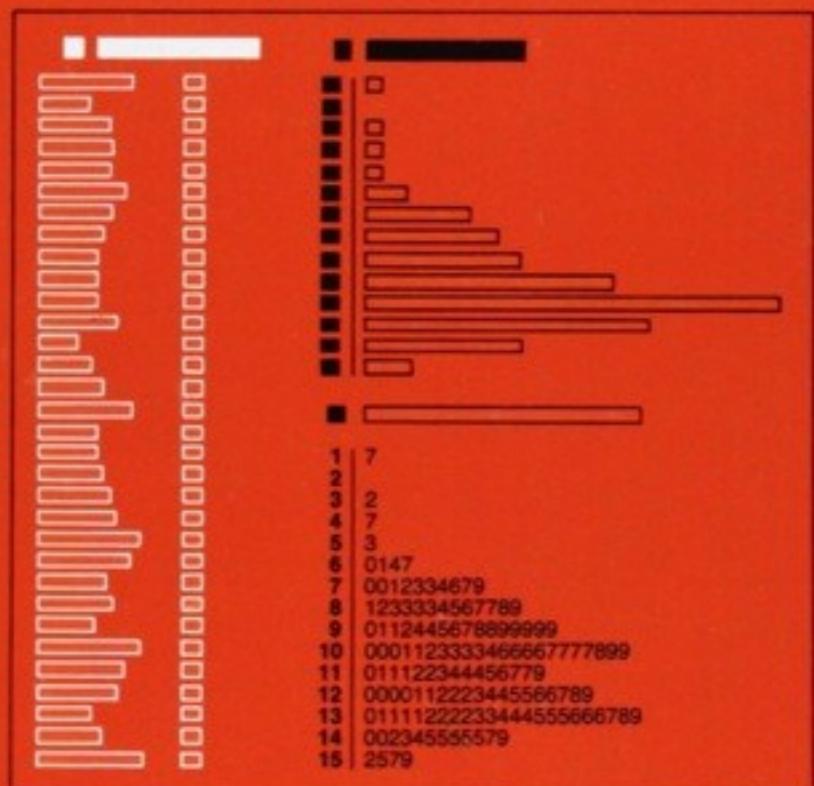


Exploratory Data Analysis (EDA)



John W. Tukey

EXPLORATORY DATA ANALYSIS



Objectives of EDA :

- Suggest hypotheses about the causes of observed phenomena
- Assess assumptions on which statistical inference will be based
- Support the selection of appropriate tools and techniques
- Provide a basis for further data collection through surveys or experiments

John Tukey,
Exploratory Data Analysis
(New York: Pearson, 1977)



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

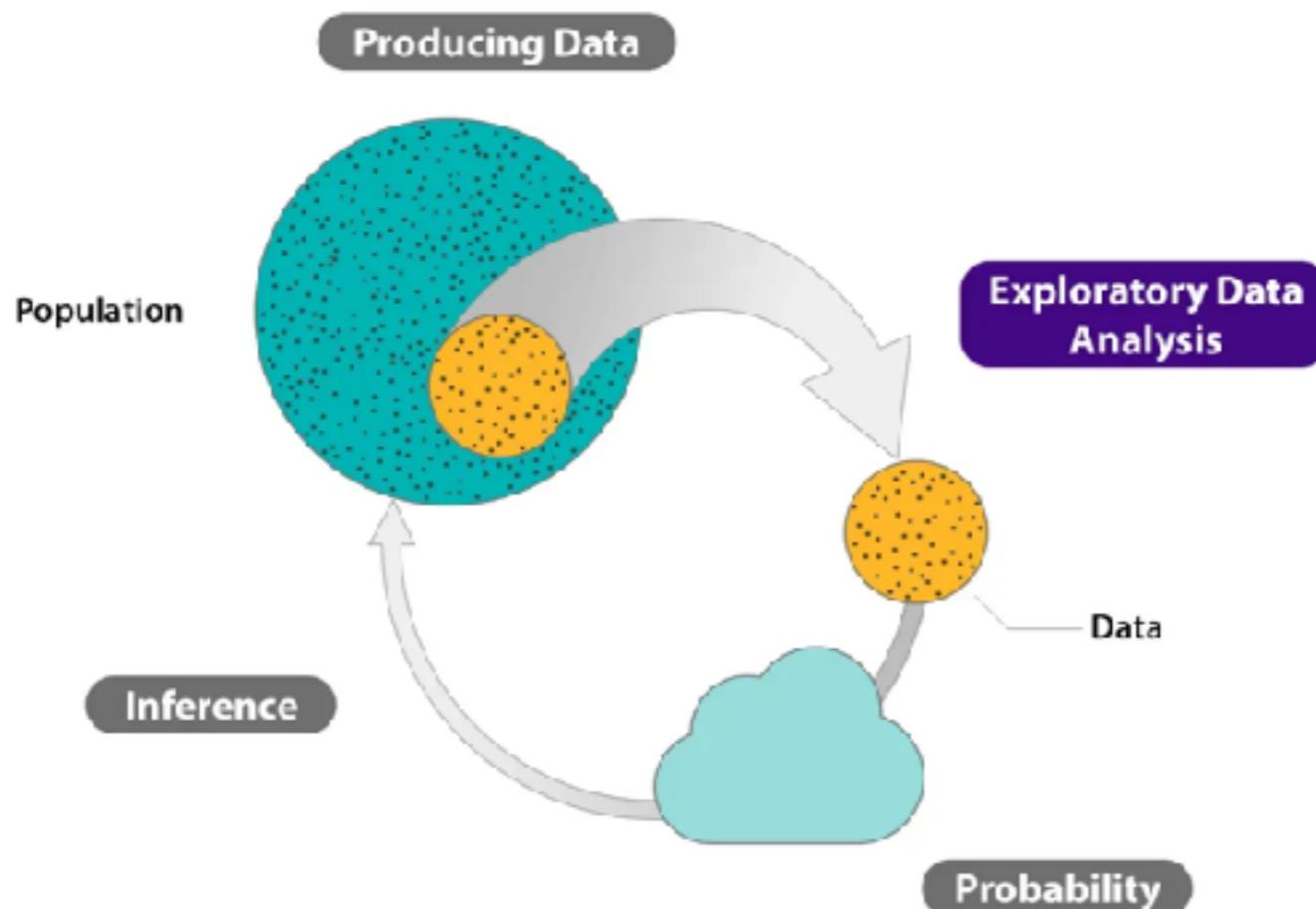
Identify

Missing data/values



Exploratory Data Analysis

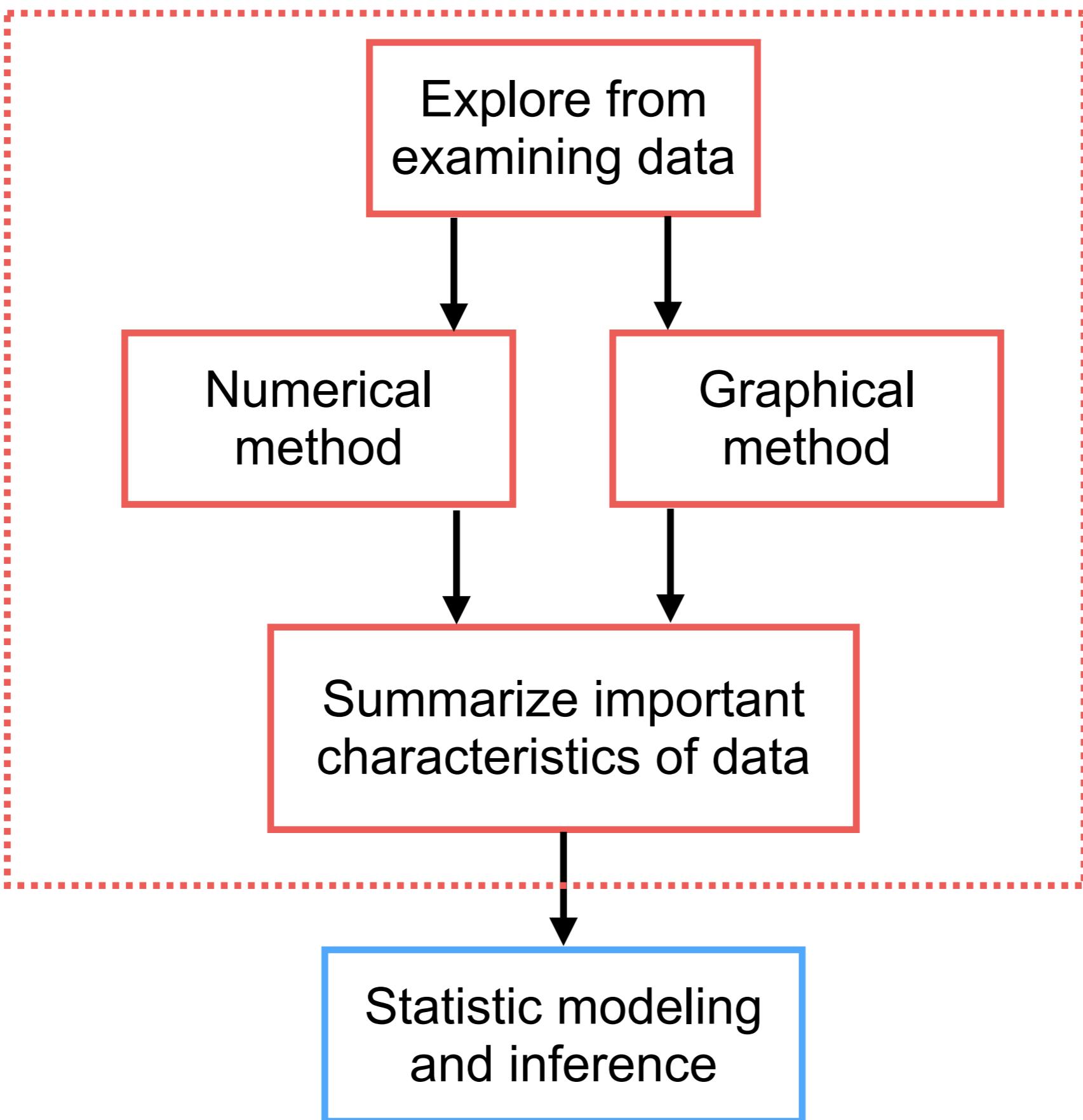
EDA helps in understanding the data, identifying patterns, detecting anomalies, and testing hypotheses.



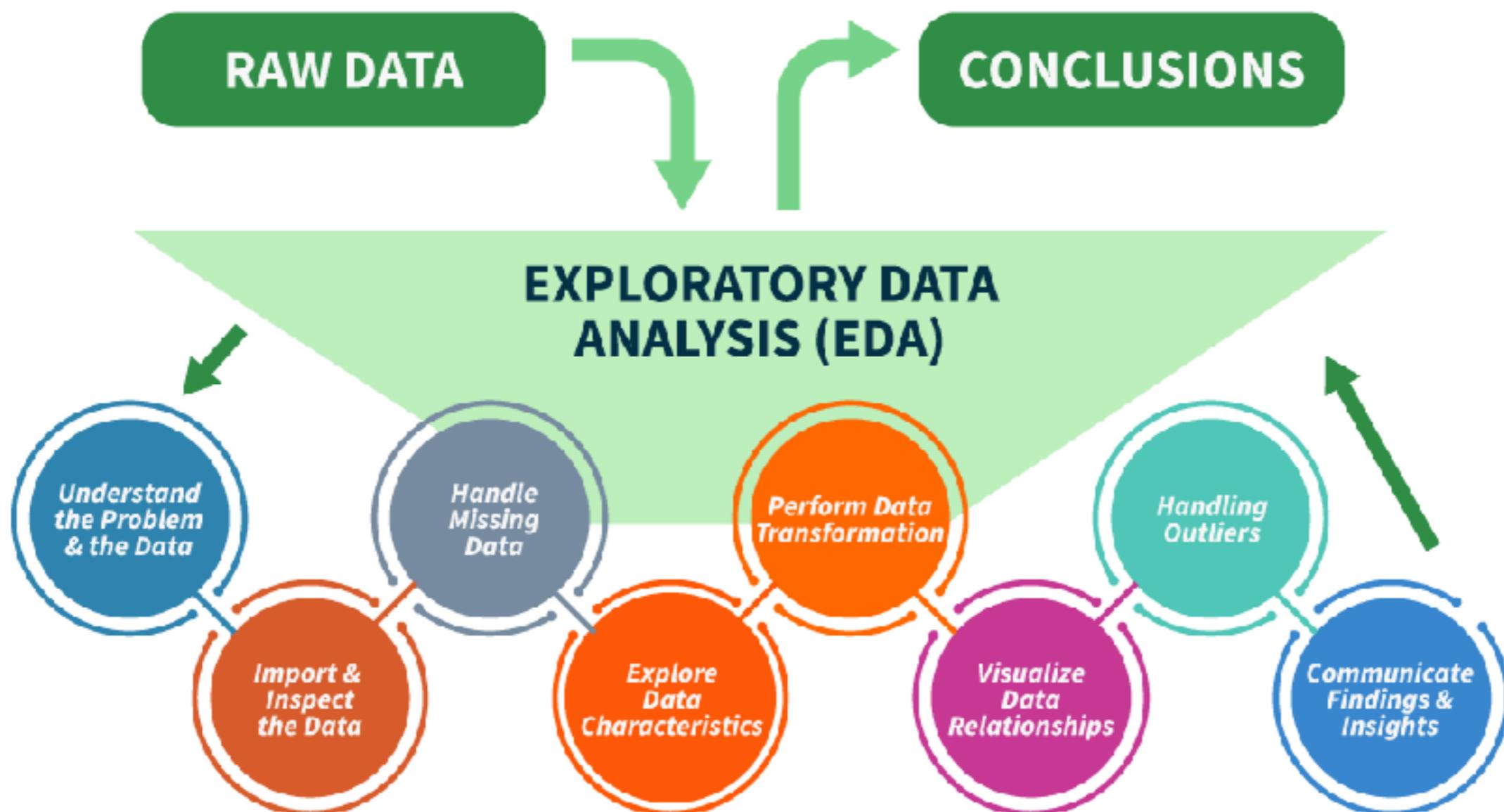
Why EDA ?

- Better understand of data
- Validate assumption about data
- Identify errors and outliers in data
- Interesting trends, patterns and relationships
- Insight and new ideas/hypothesis
- Input for statistic modeling





Steps for Performing Exploratory Data Analysis



Steps

Problem and data understanding

 Data cleaning

 Pattern discovery

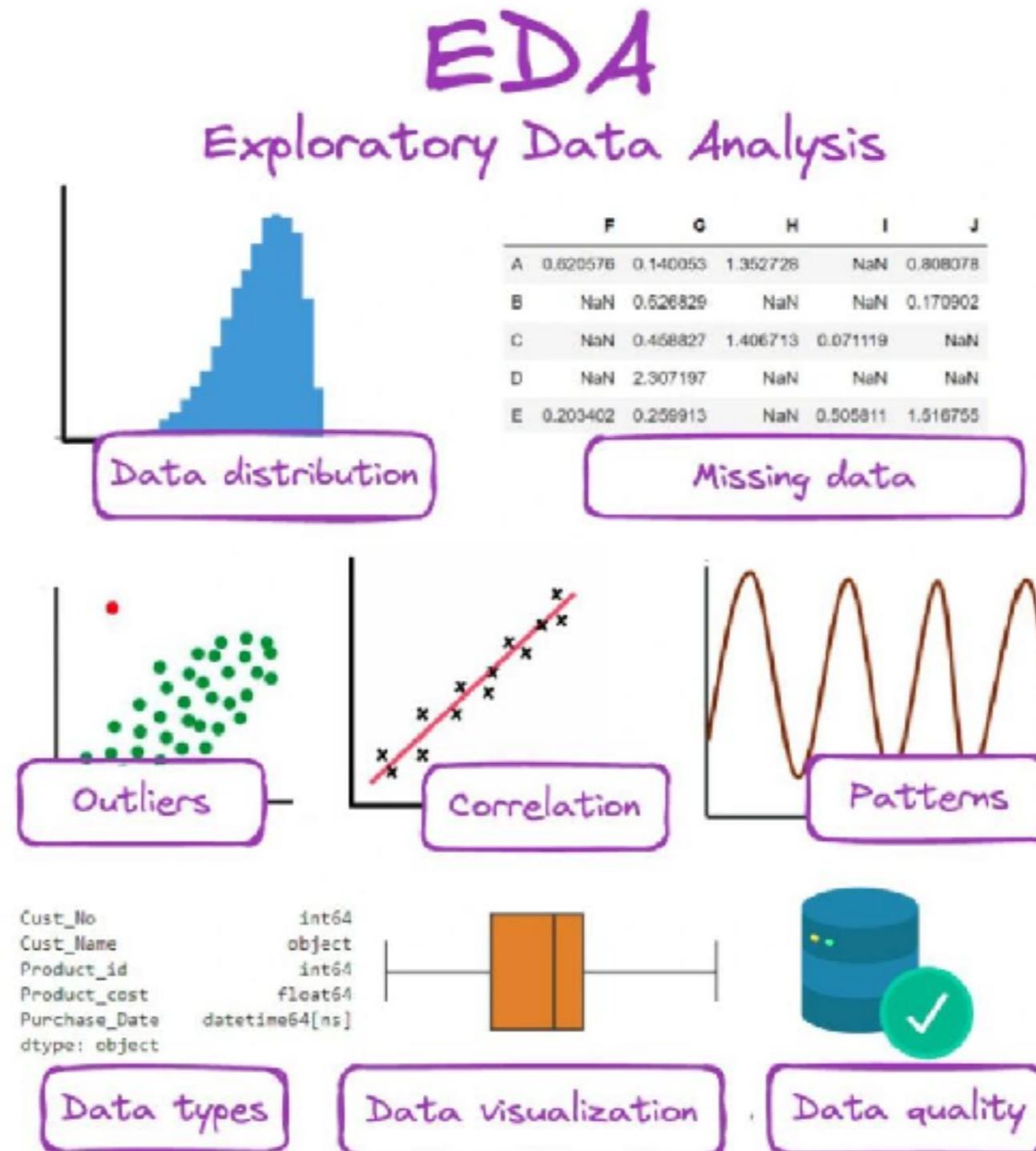
 Data visualization

 Model selection

 Quality control



Workshop with EDA



Exploratory Data Analysis (EDA)



Analysis Process

Iterative process

Data
Transformation

Data
Analysis

Data
Visualization

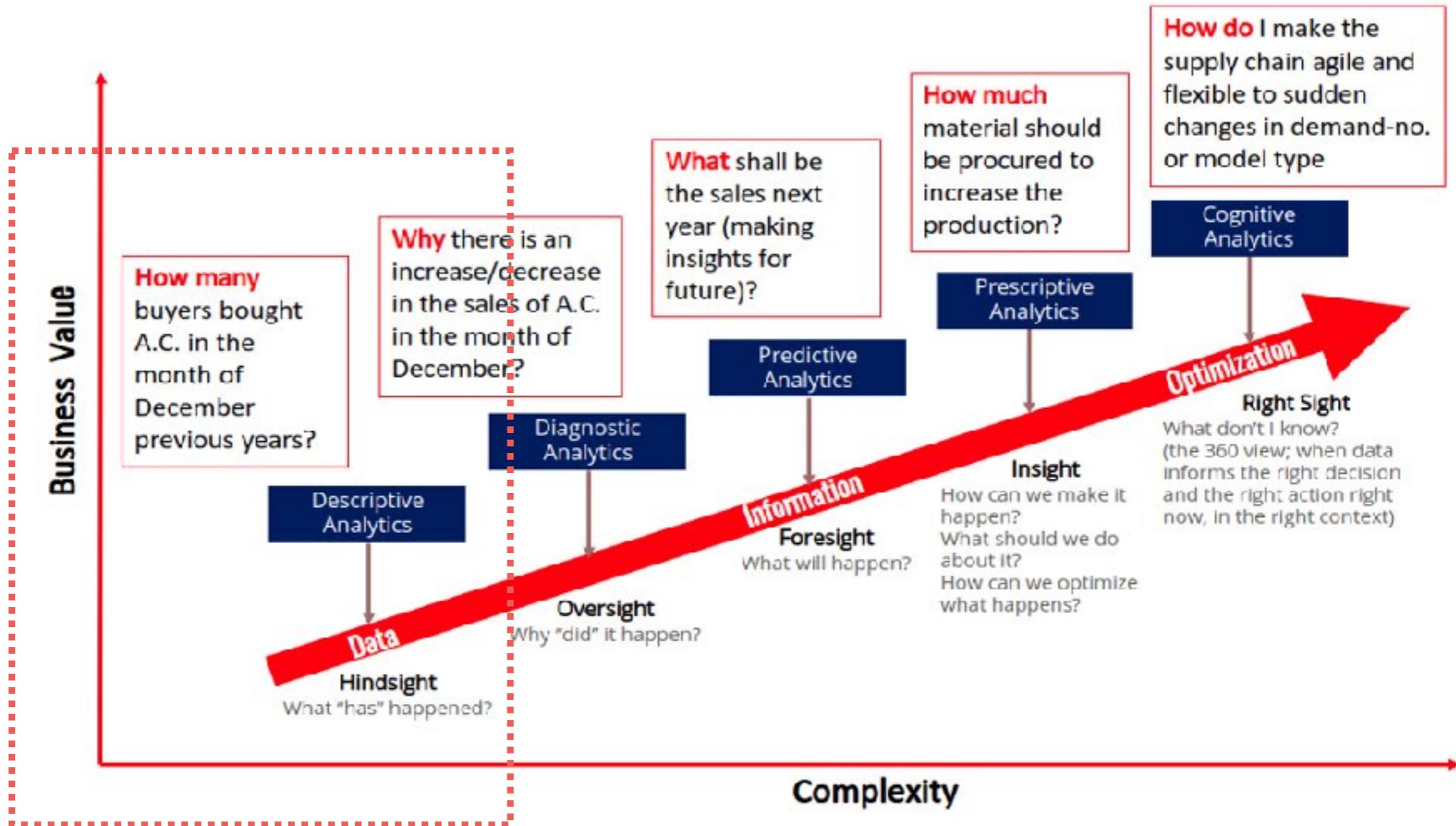
Data manipulation
Data wrangling



Statistical Analysis for data analysis



Key Techniques



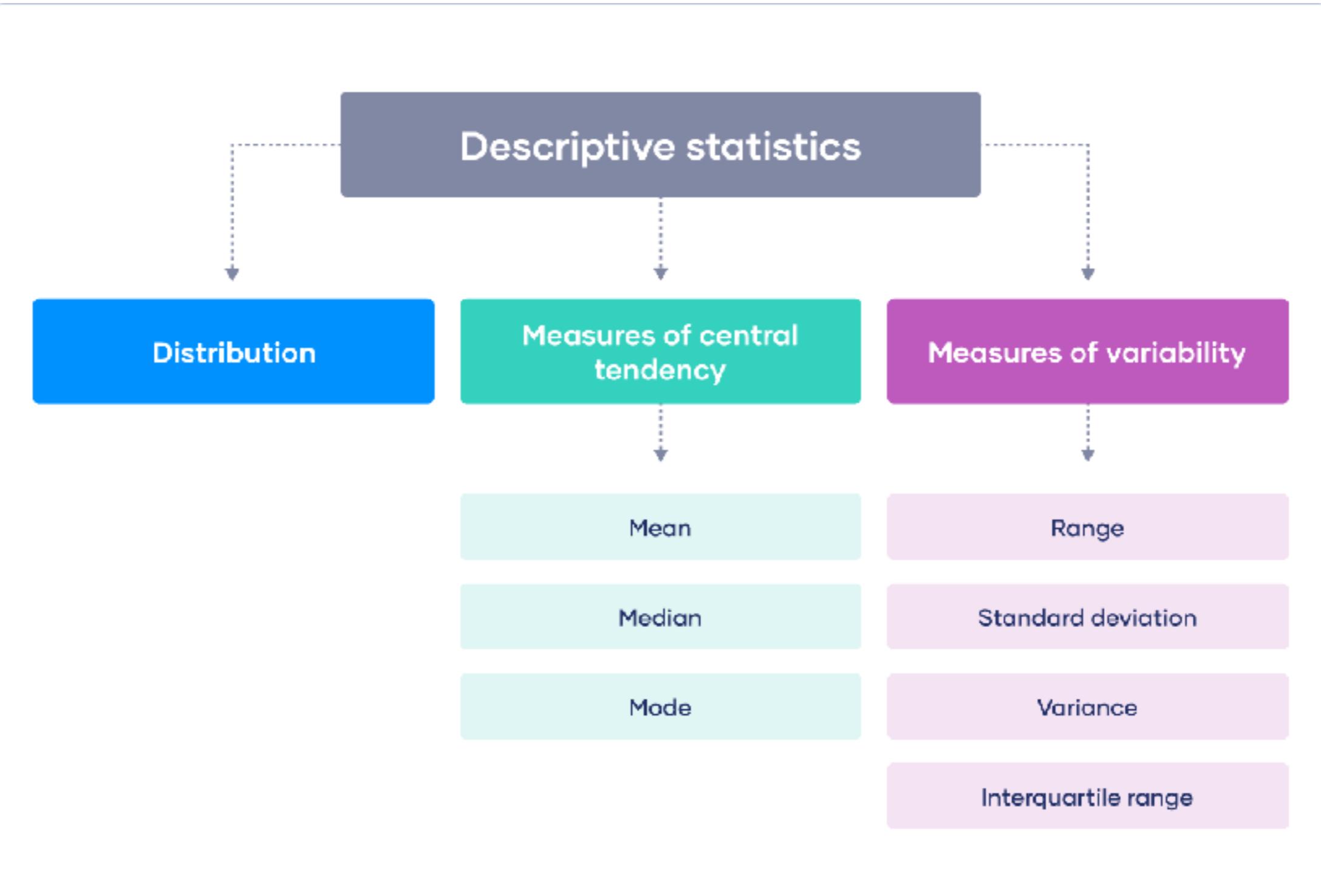
Descriptive statistic

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



Mean

2, 3, 4

1 2 ③

$$2 + 3 + 4 = 9$$

$$\text{MEAN} = 9 \div 3 = 3$$

wikiHow to Find Mean, Median, and Mode

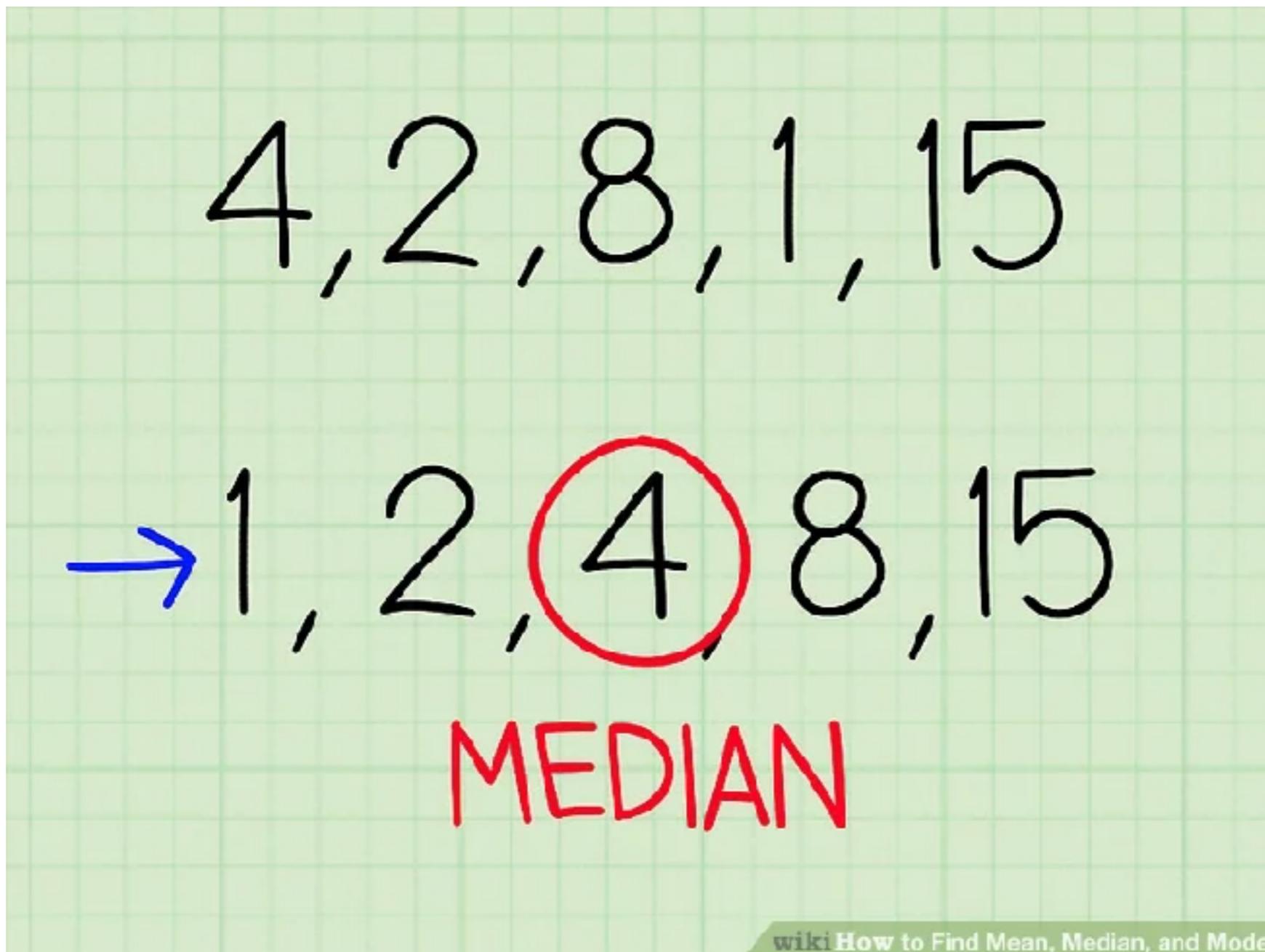
<https://www.wikihow.com/Find-Mean,-Median,-and-Mode>



Sharing

© 2020 - 2024 Siam Chamnankit Company Limited. All rights reserved.

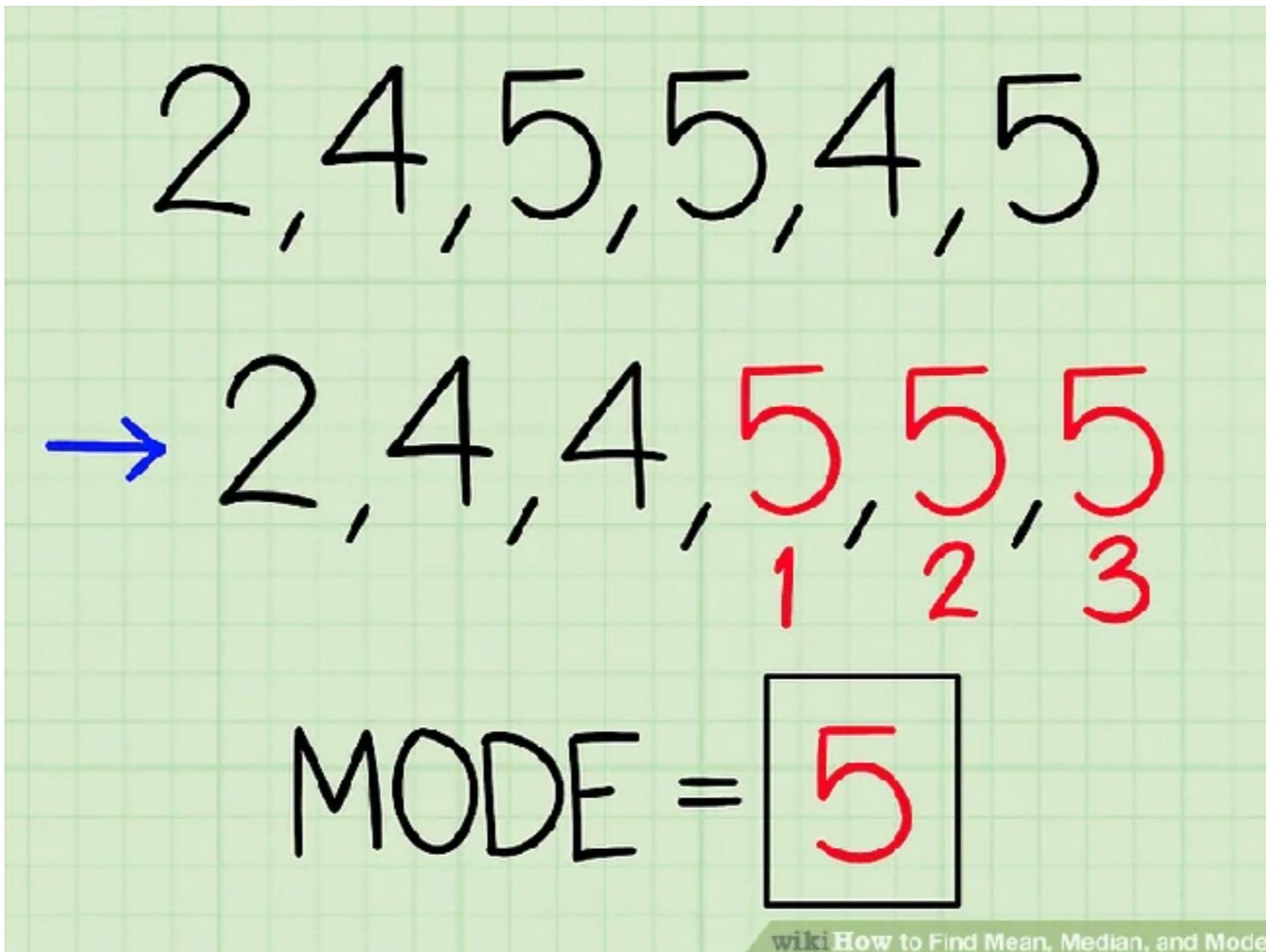
Median



<https://www.wikihow.com/Find-Mean,-Median,-and-Mode>



Mode



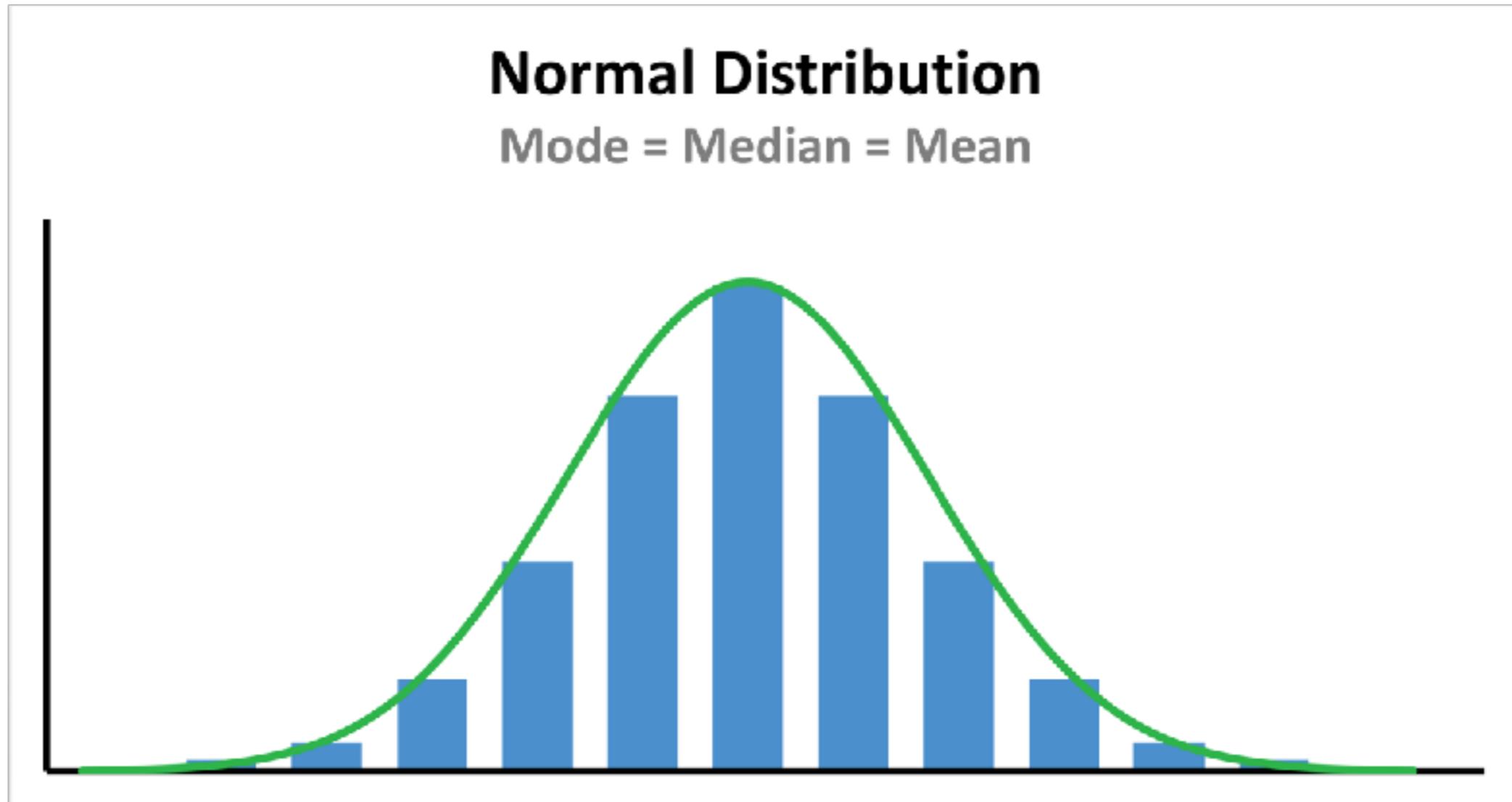
<https://www.wikihow.com/Find-Mean,-Median,-and-Mode>



Sharing

© 2020 - 2024 Siam Chamnankit Company Limited. All rights reserved.

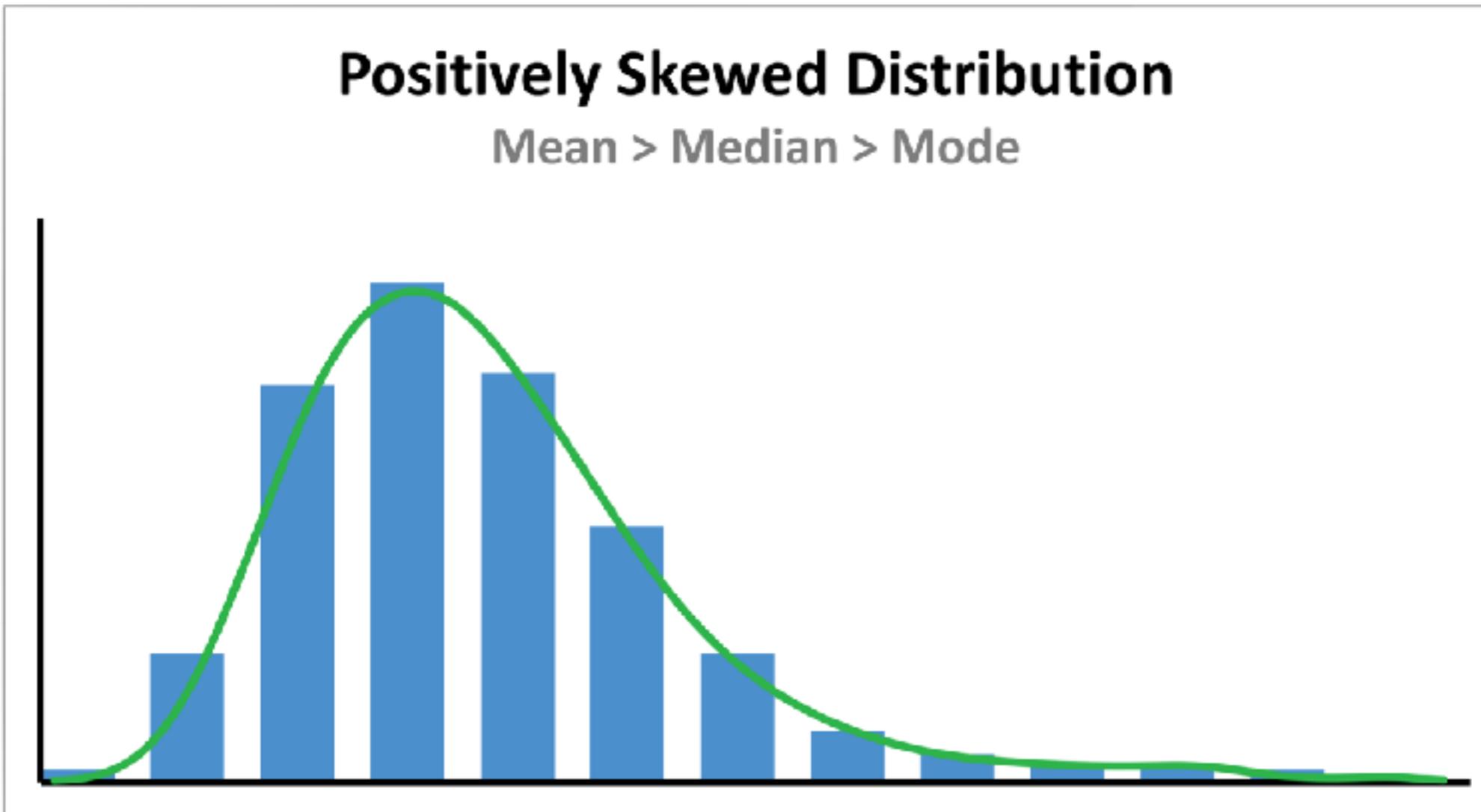
Normal Distribution (Gaussian)



Positive Skewed Distribution

Positively Skewed Distribution

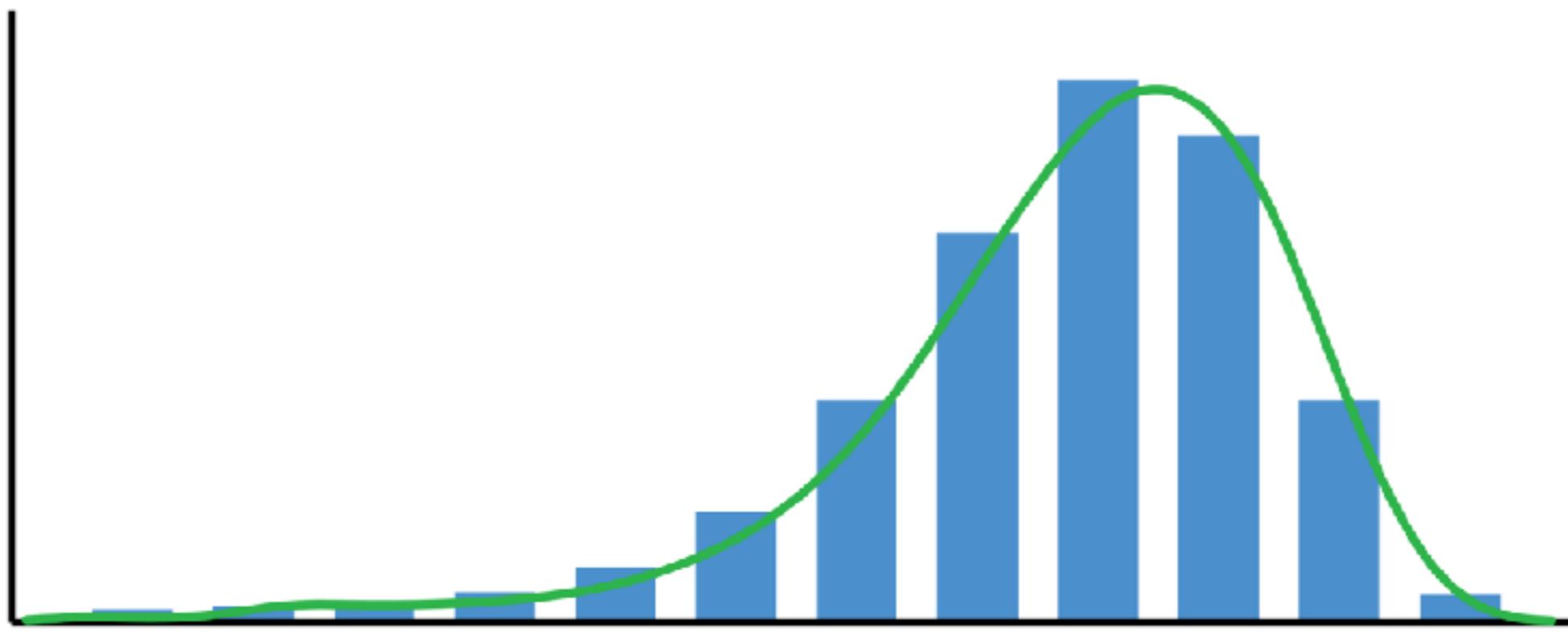
Mean > Median > Mode



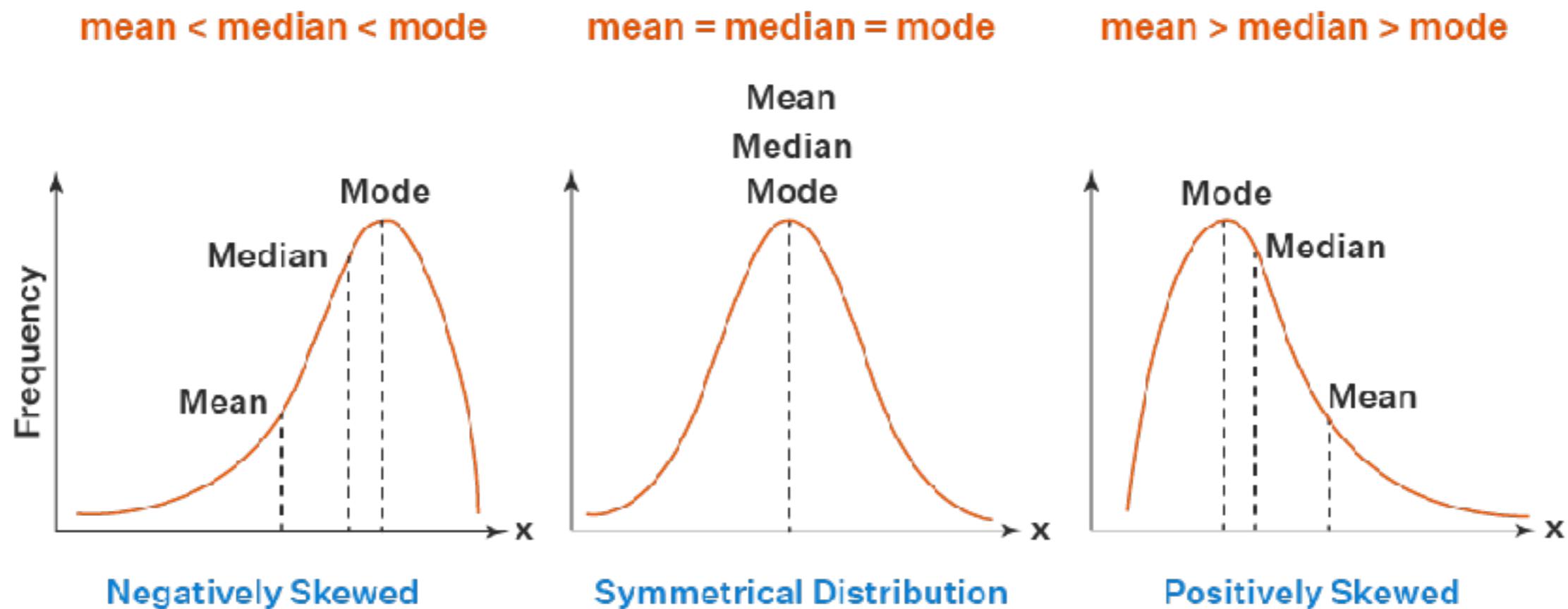
Negative Skewed Distribution

Negatively Skewed Distribution

Mean < Median < Mode



Distribution



Measures of Variability

Range

The difference between the maximum and minimum values

Variance

The average of the squared differences from the mean

Standard Deviation

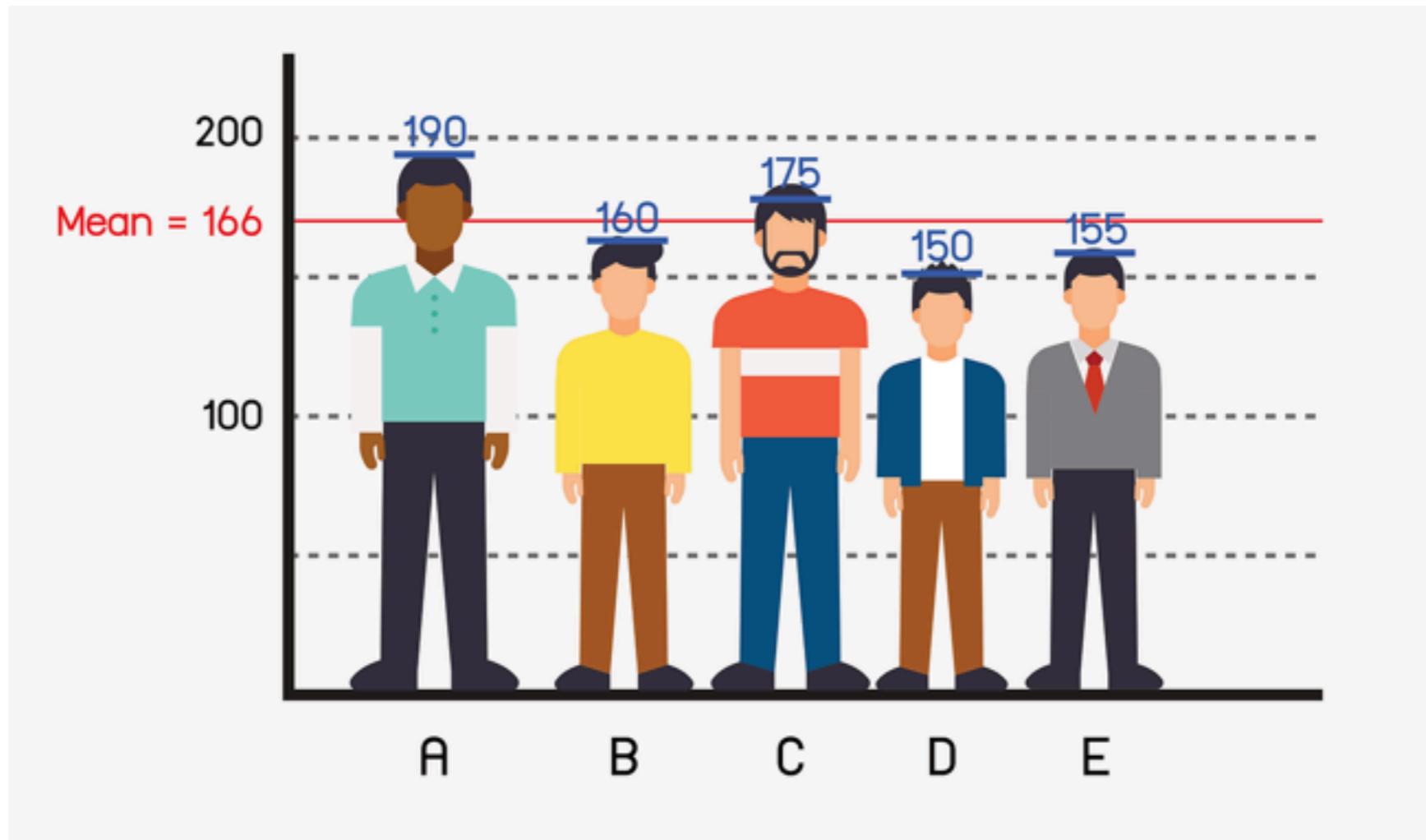
The square root of the variance, showing how much the values deviate from the mean

Interquartile Range (IQR)

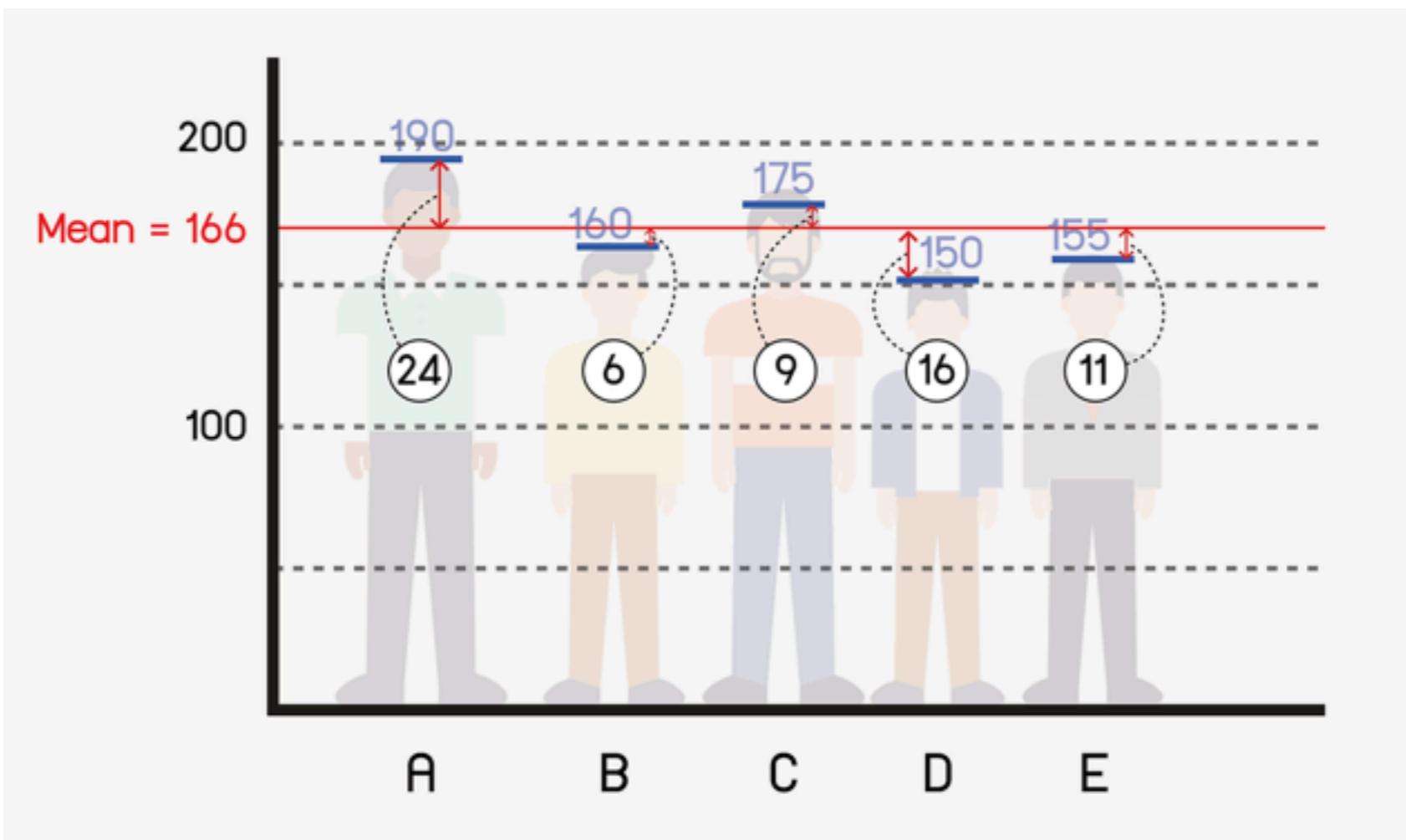
The difference between the first (25th percentile) and third (75th percentile) quartiles



Variance ? (1)



Variance ? (2)



Variance ? (3)

ค่าเฉลี่ย คือ
Mean

$$\frac{190 + 160 + 175 + 150 + 155}{5} = 166$$

ค่าแปรปวน คือ
Variance

$$\frac{24^2 + (-6)^2 + 9^2 + (-16)^2 + (-11)^2}{5} = 214$$

ค่าเบี่ยงเบนมาตรฐาน คือ
Standard Deviation

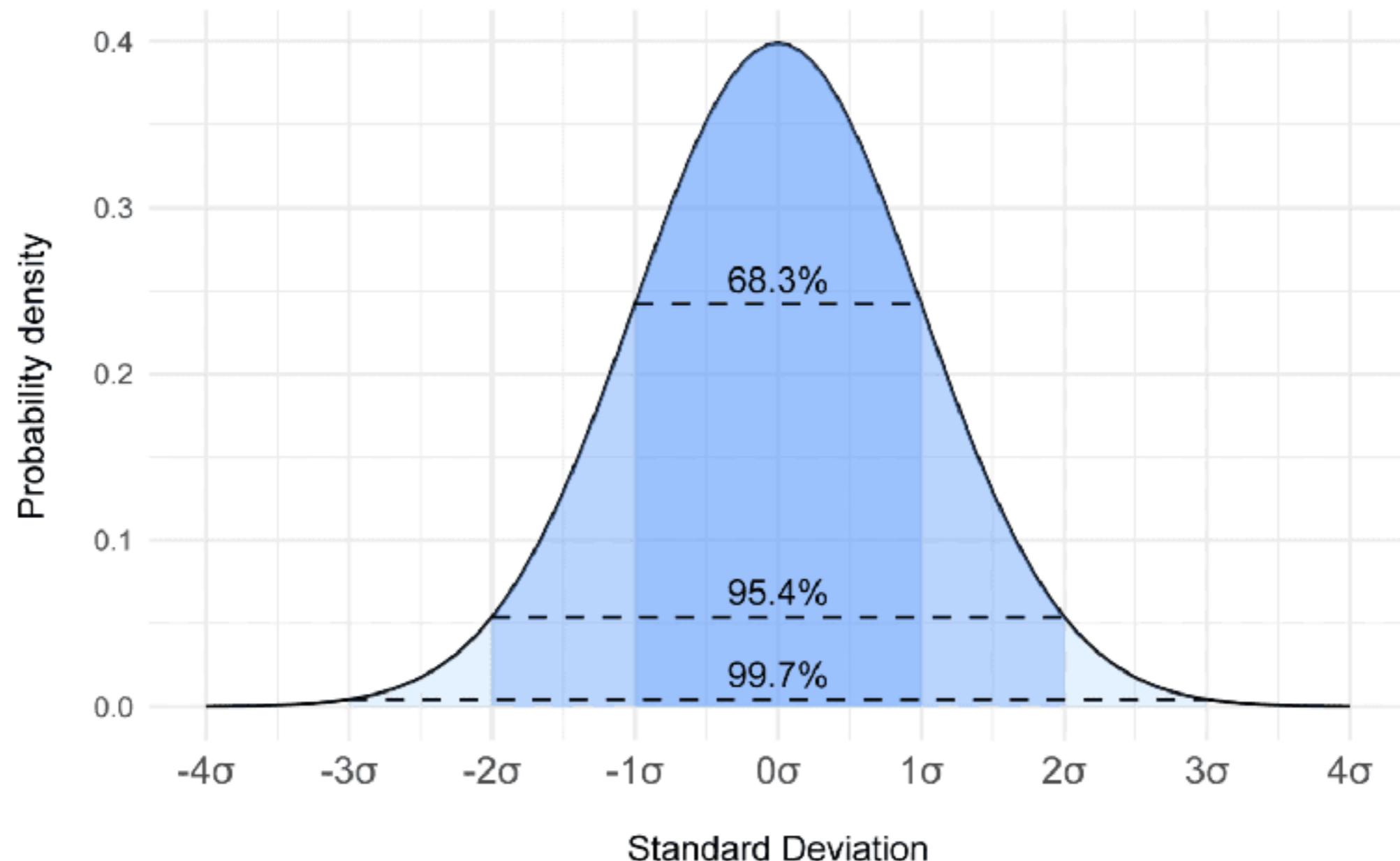
$$\sqrt{214} = 14.63$$



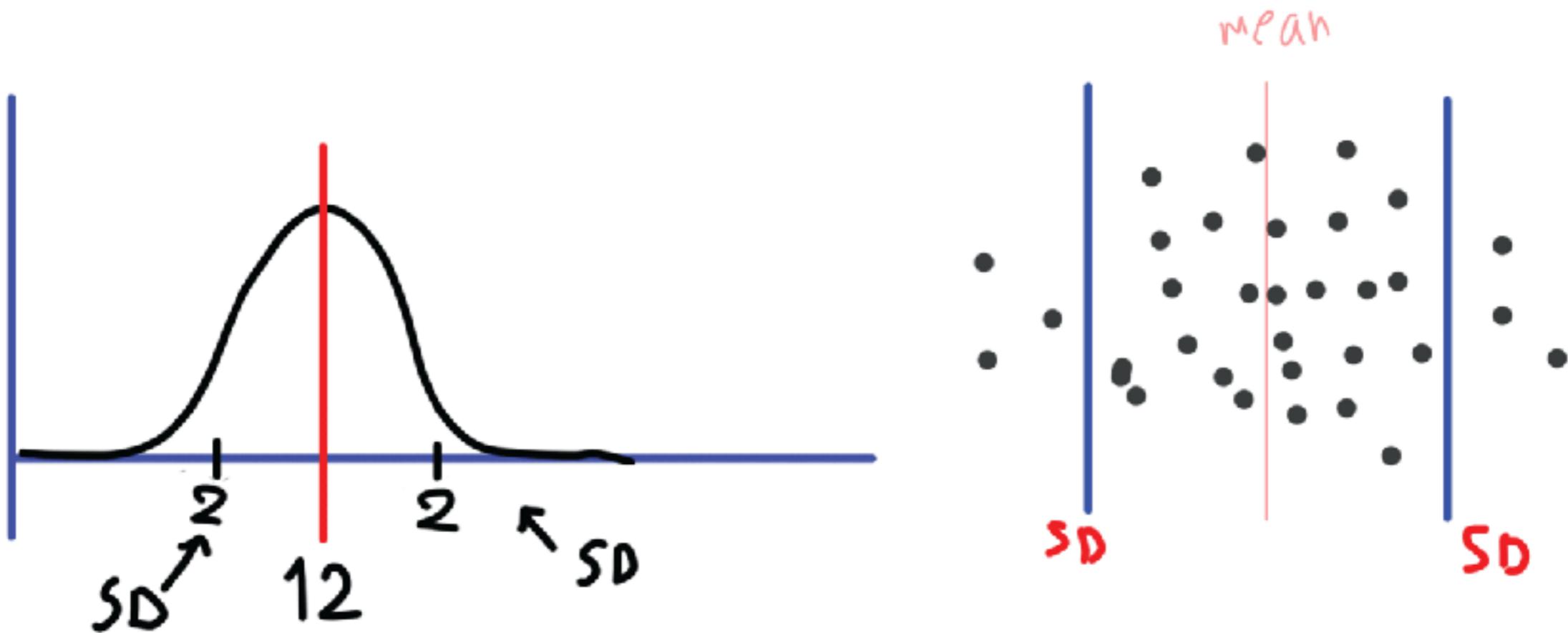
Workshop



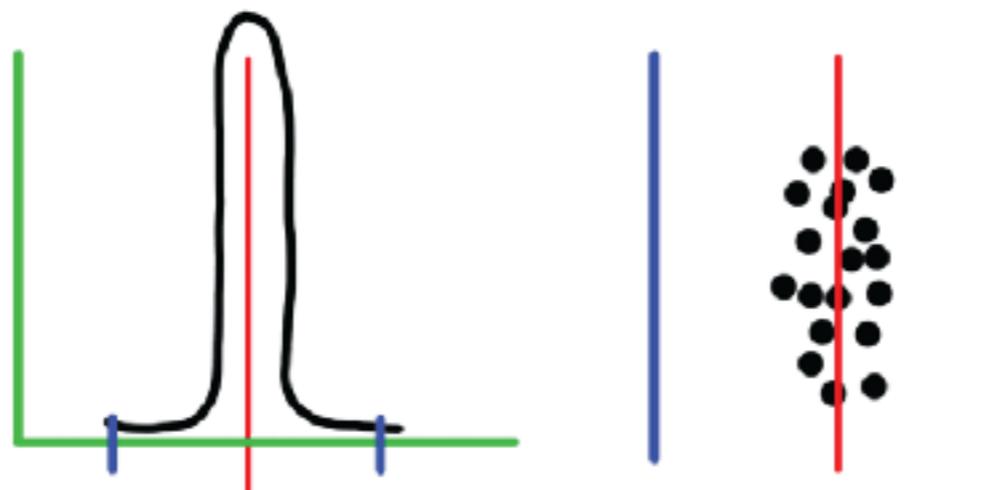
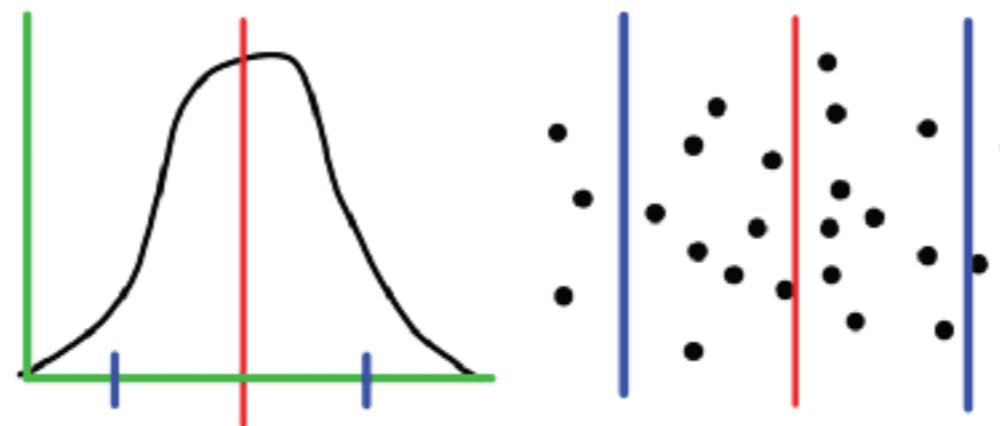
Standard Deviation (1)



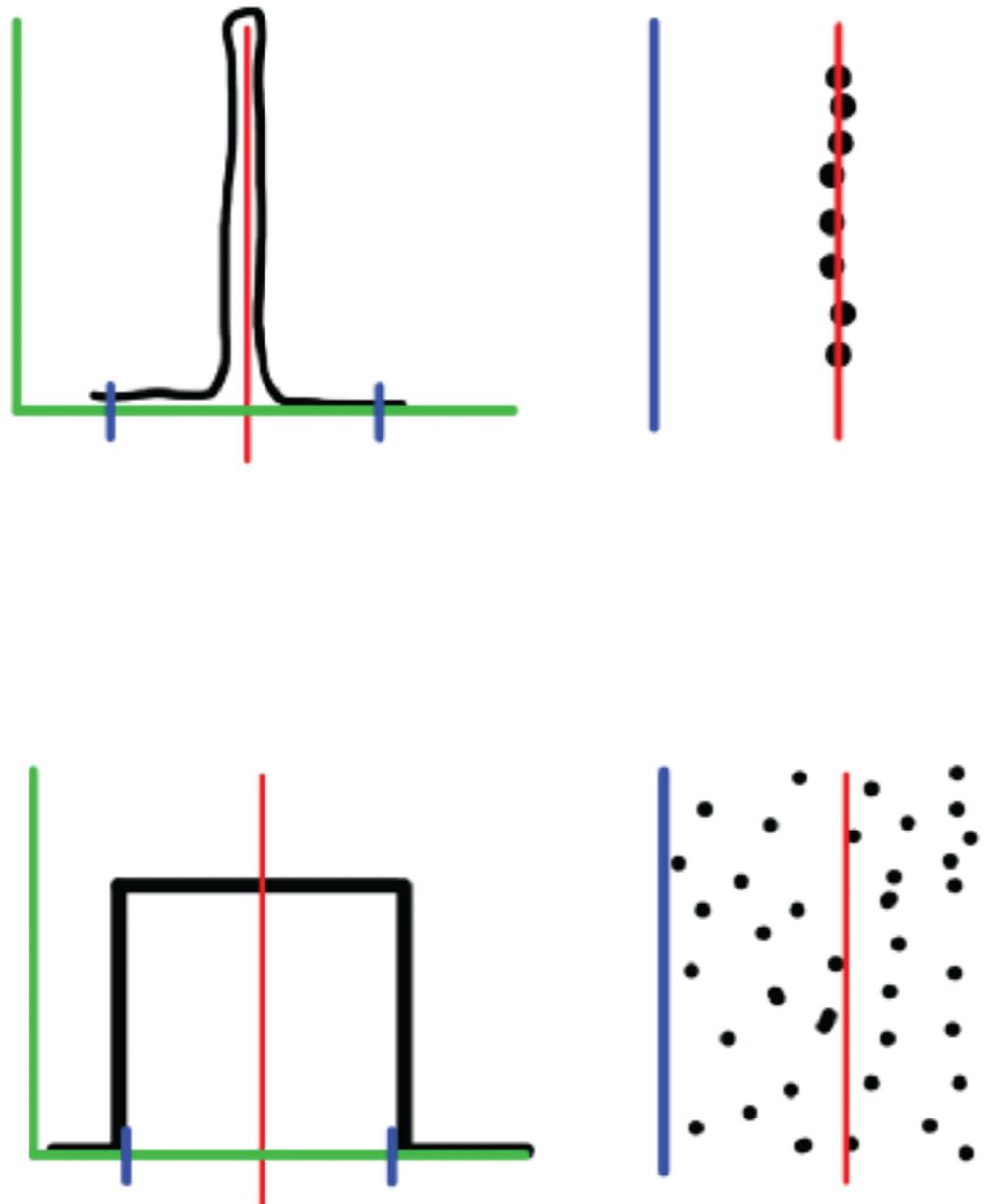
Standard Deviation (2)



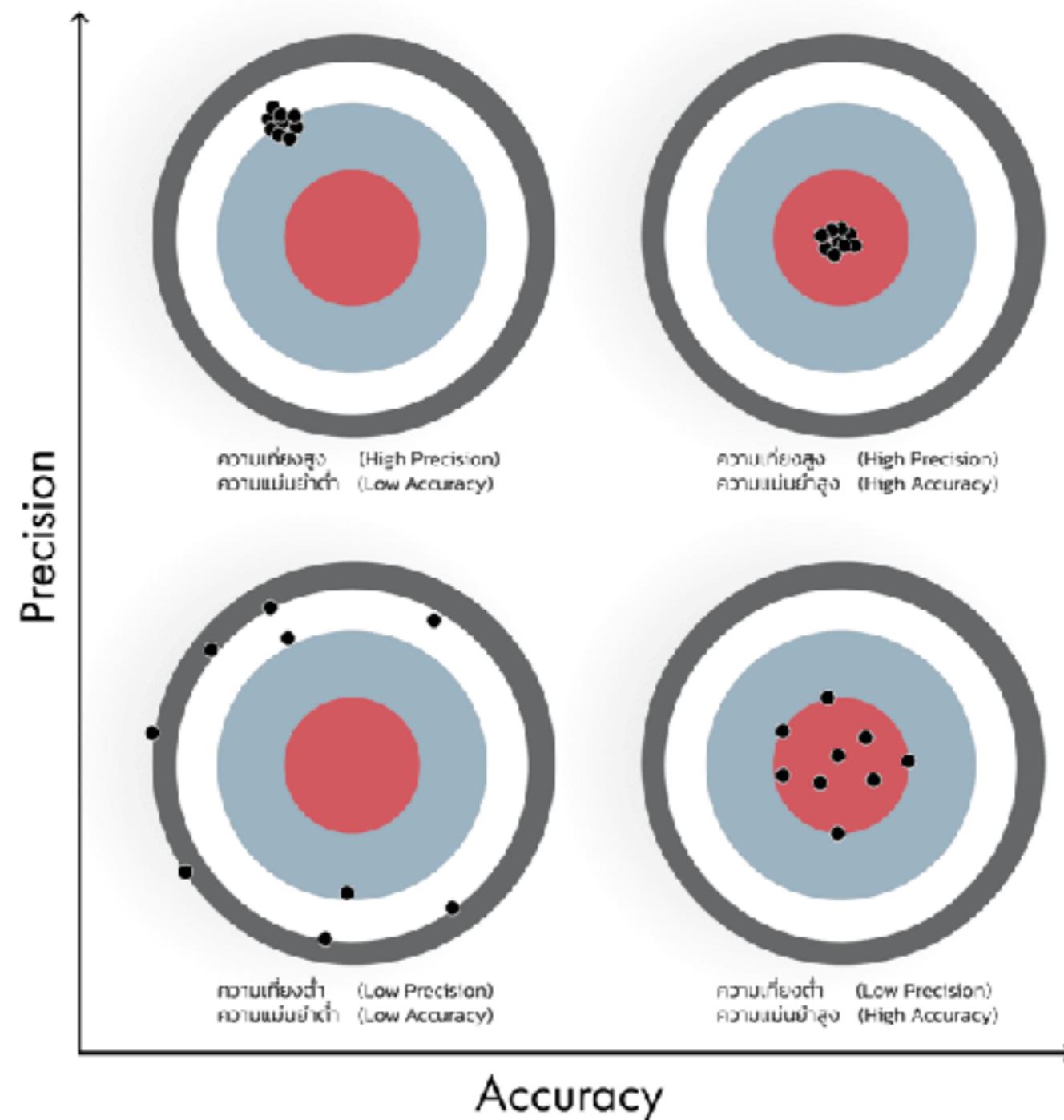
Standard Deviation (3)



Standard Deviation (4)



Standard Deviation and Precision



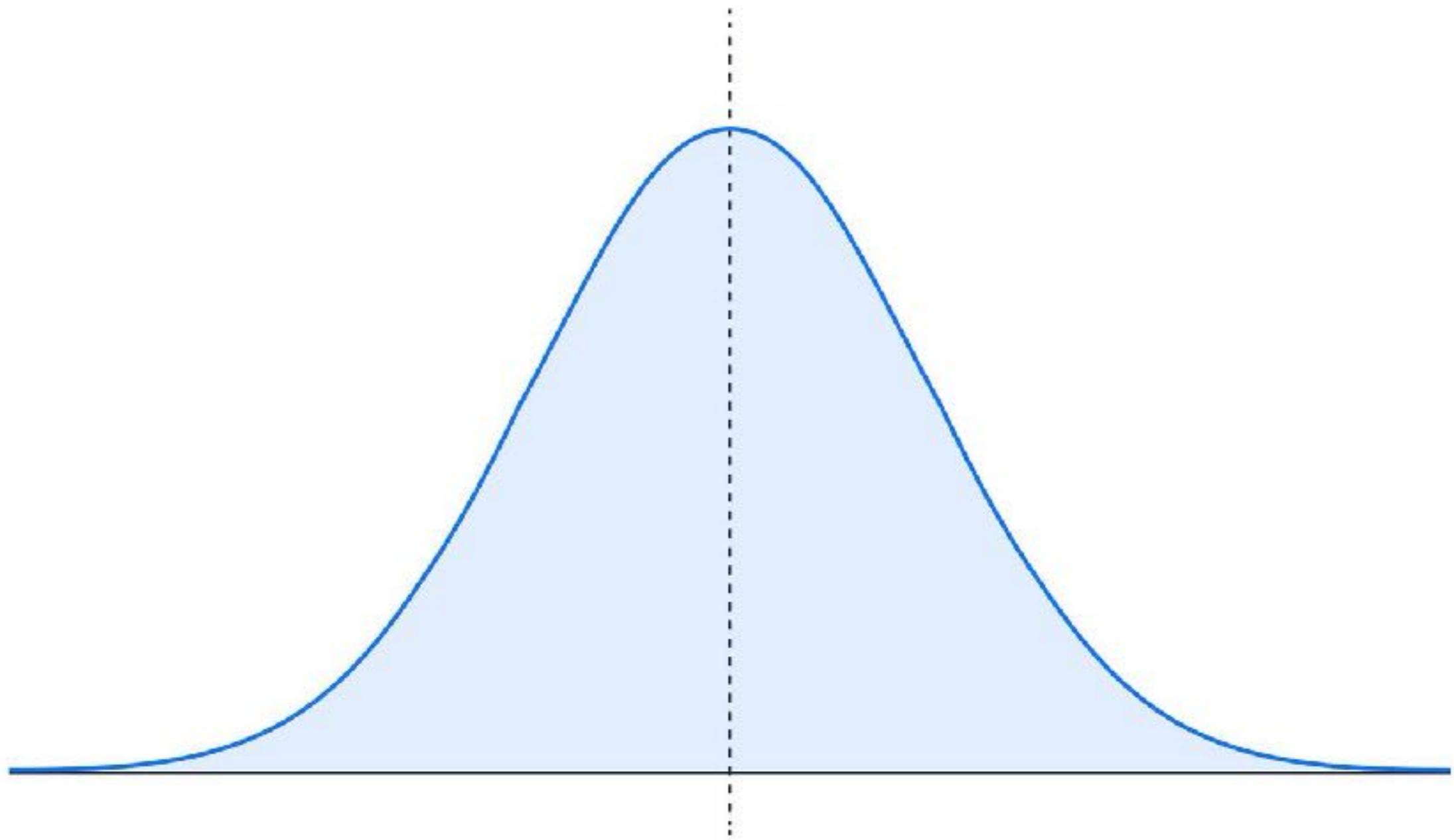
Percentile

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

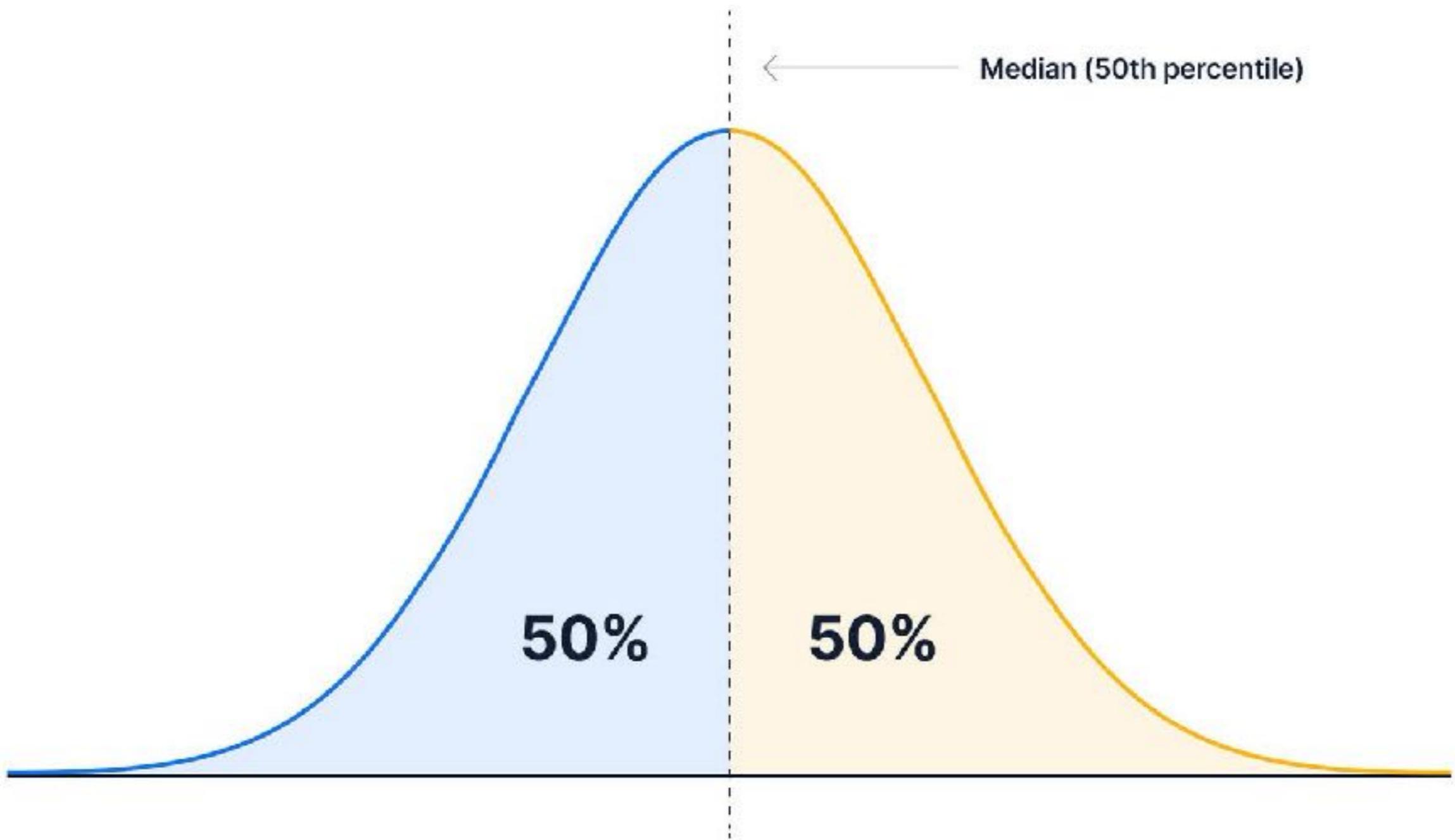
It represents the **percentage of values** in the data that are below a certain point.



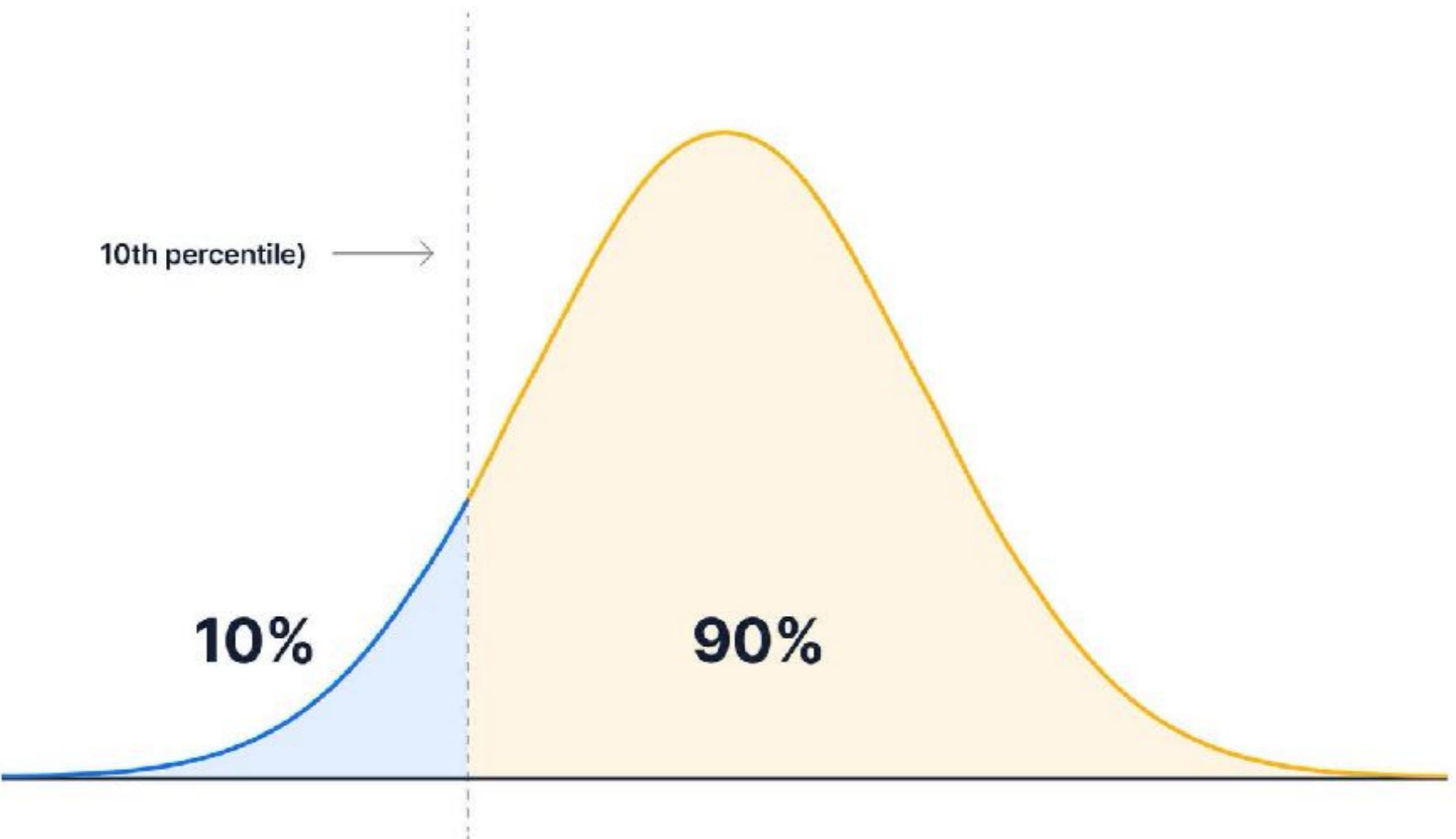
Normal distribution



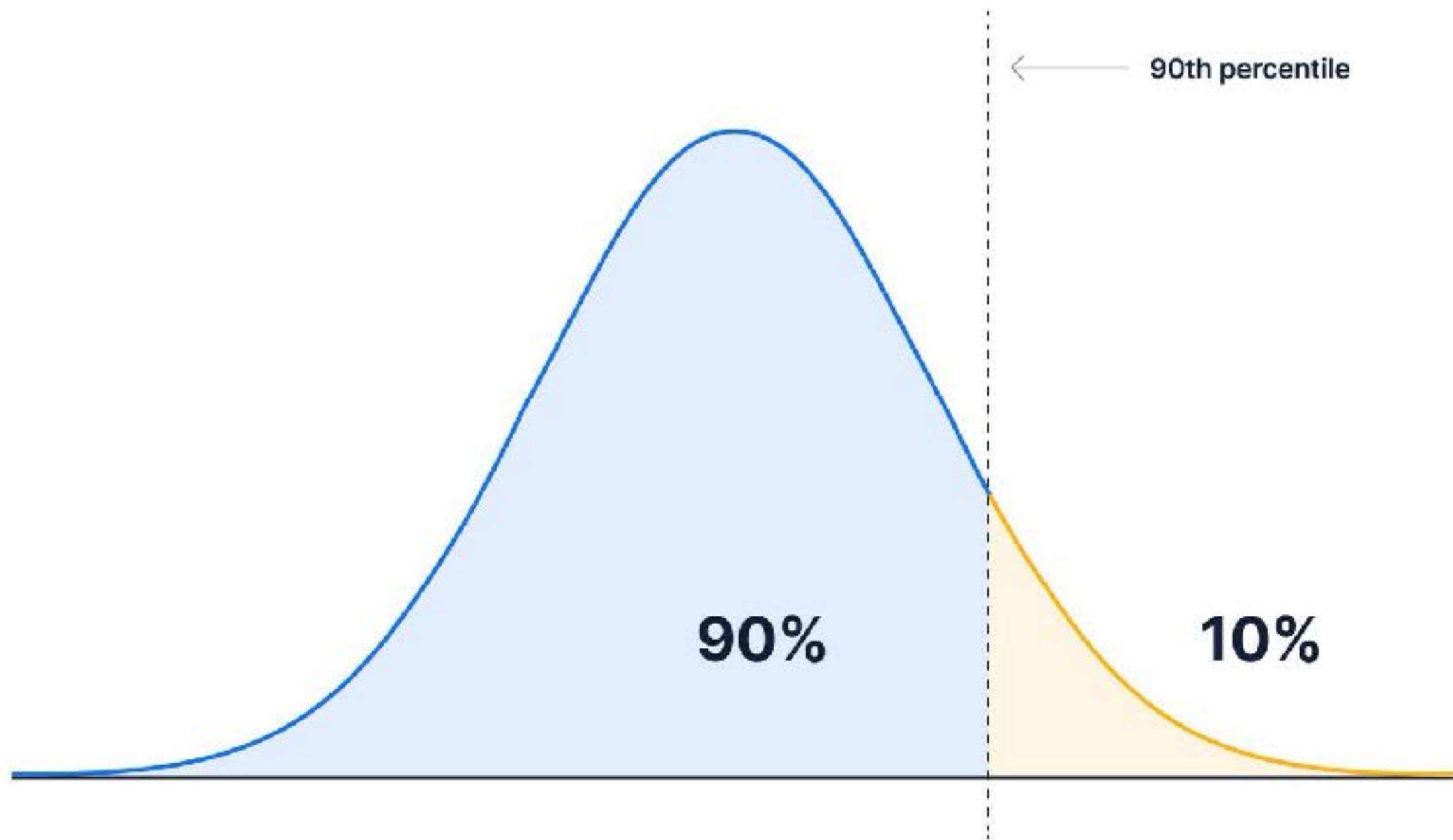
50th percentile



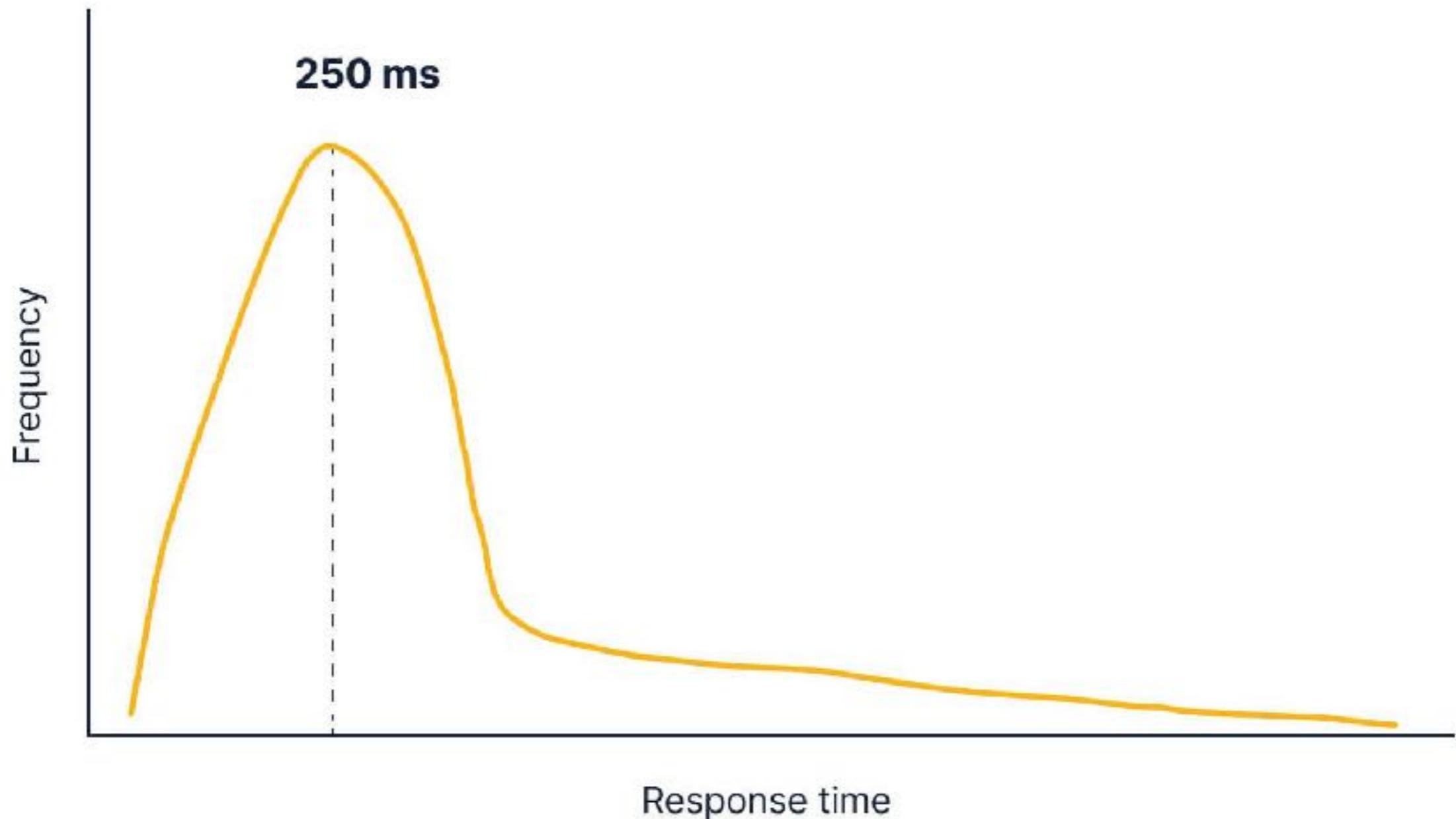
10th percentile



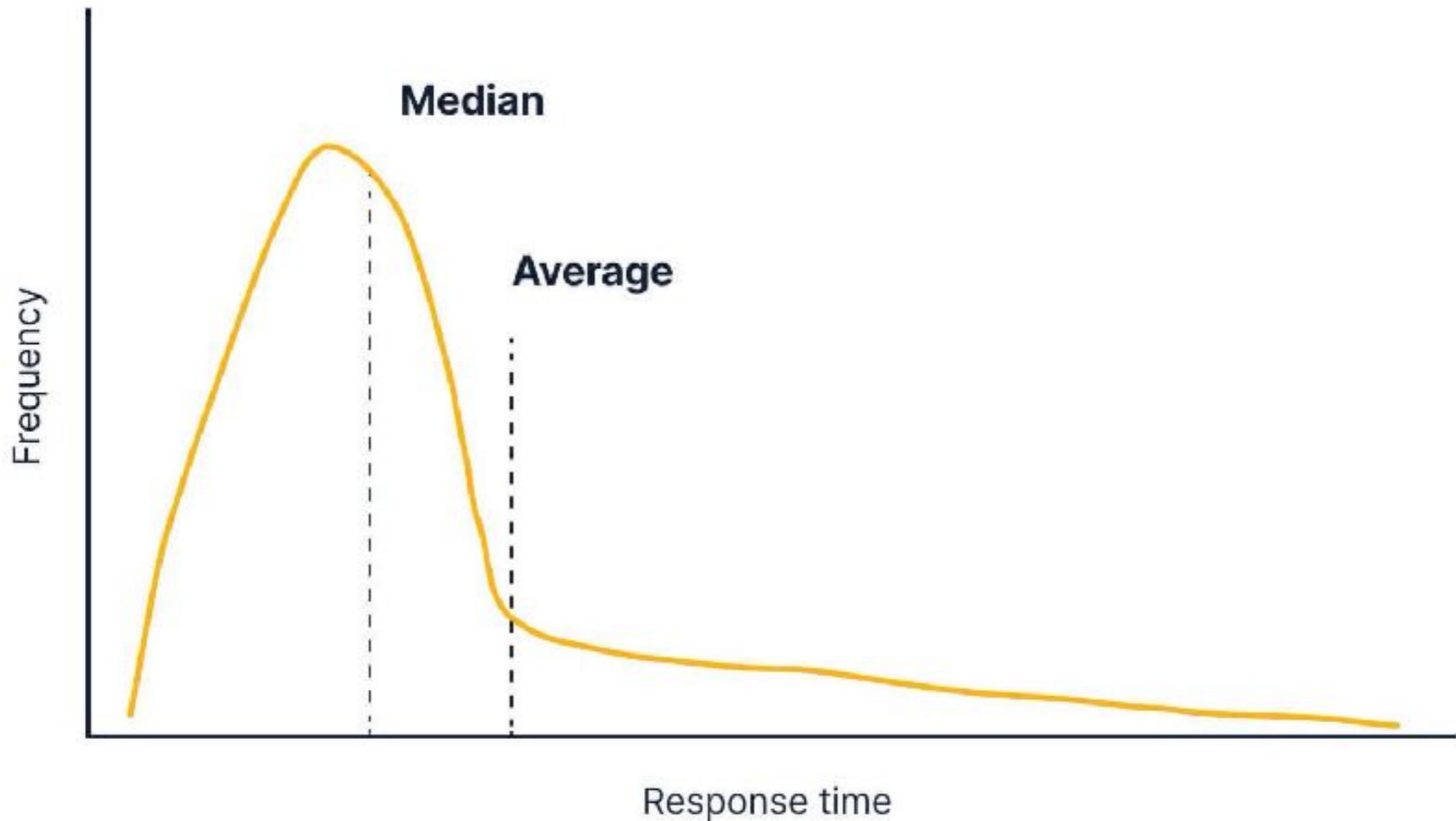
90th percentile



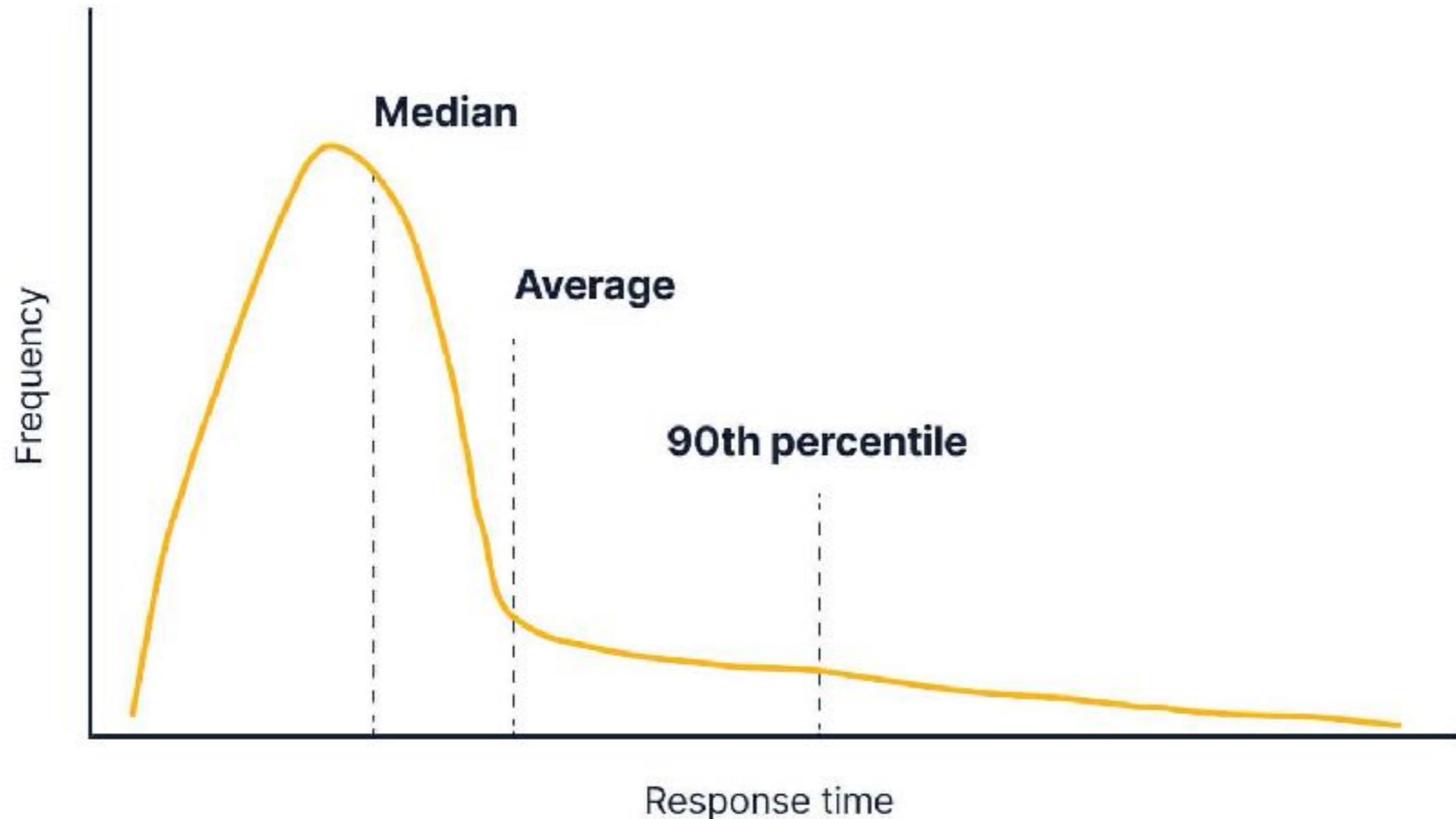
Long tails (1)



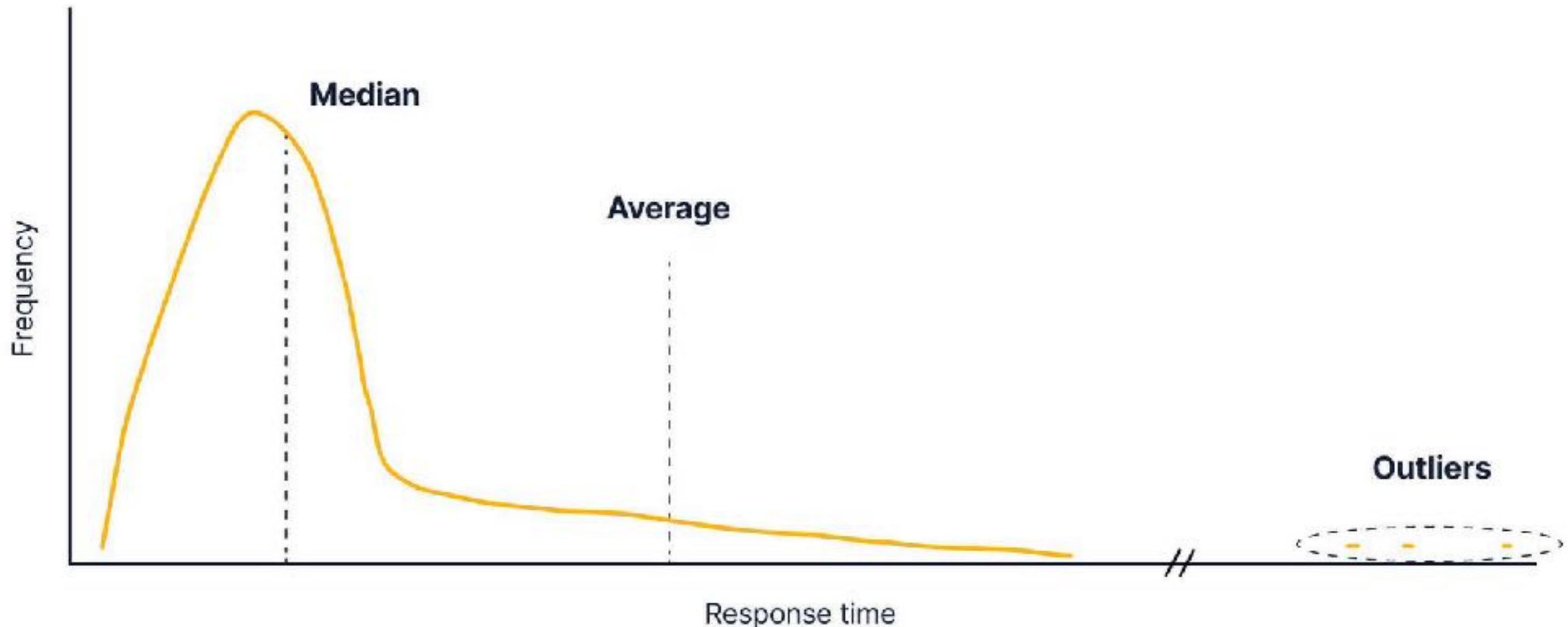
Long tails (2)



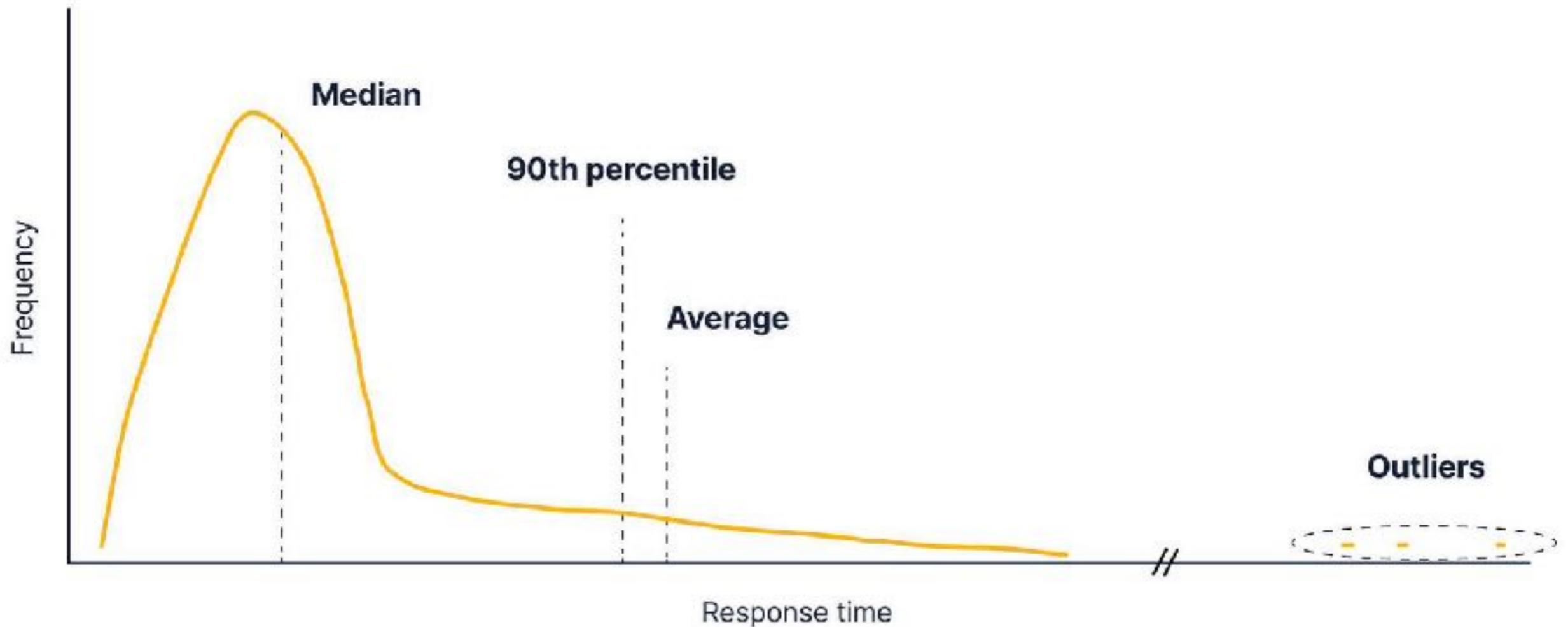
Long tails (2)



Long tails with outliers



Long tails with outliers



Workshop



Business Intelligence (BI) and Visualization



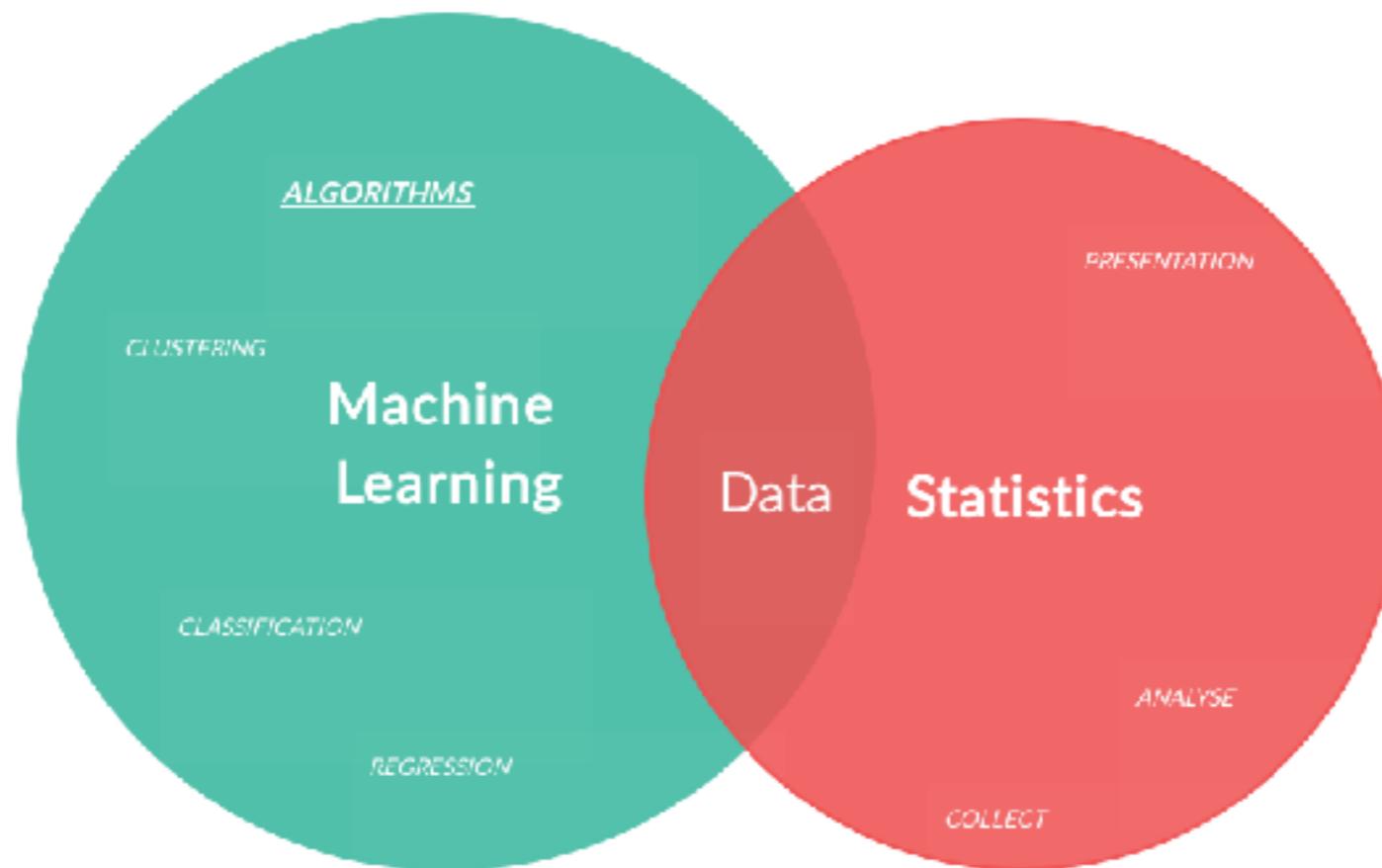
Data visualization

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

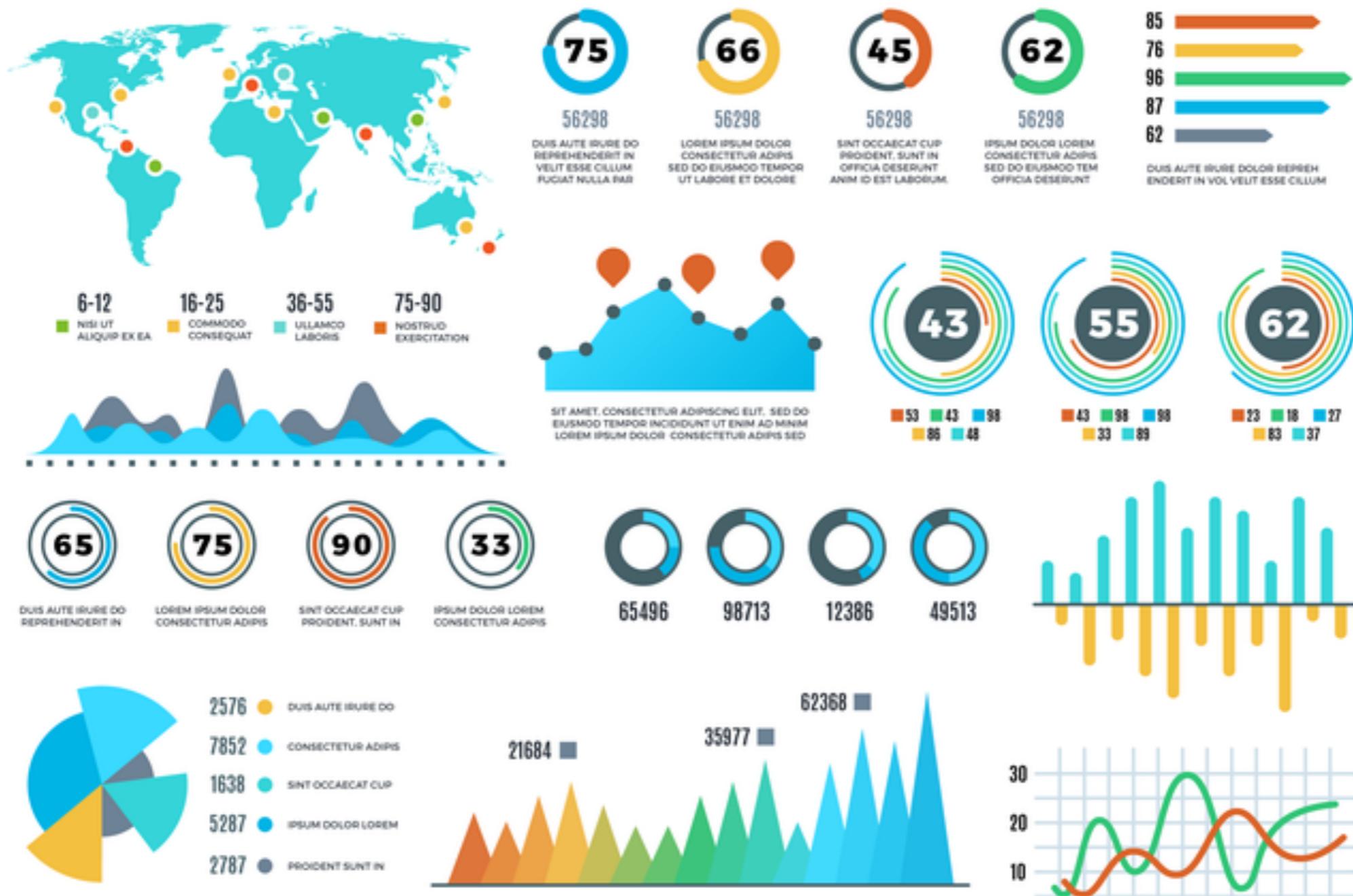


Data visualization

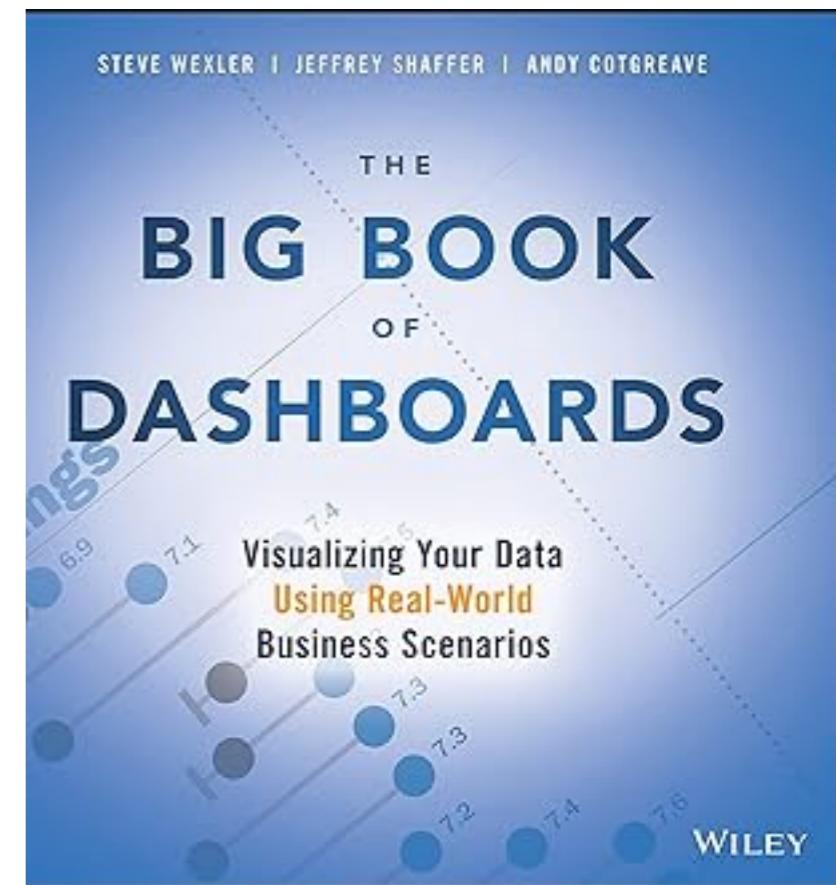
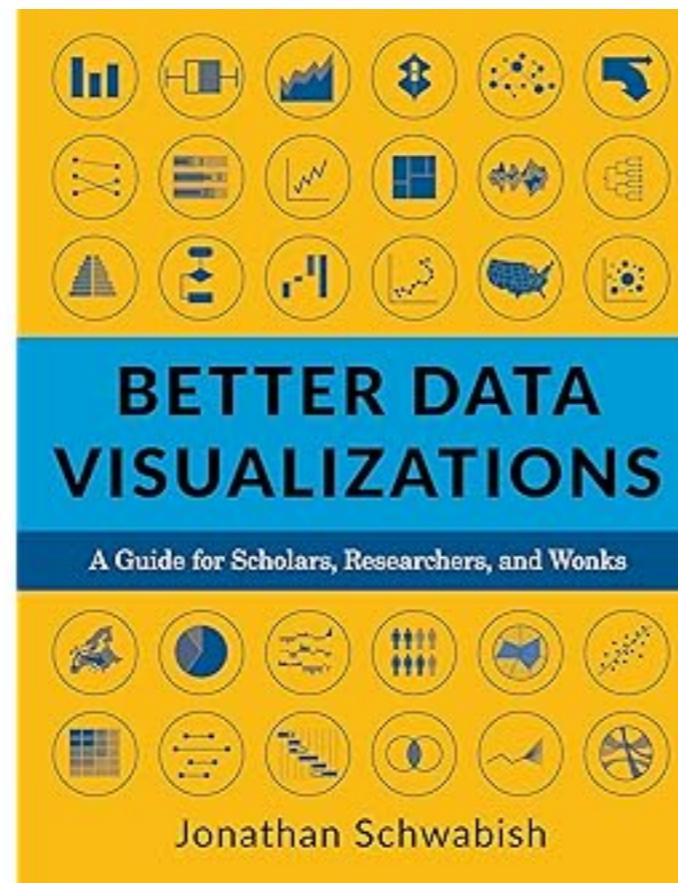
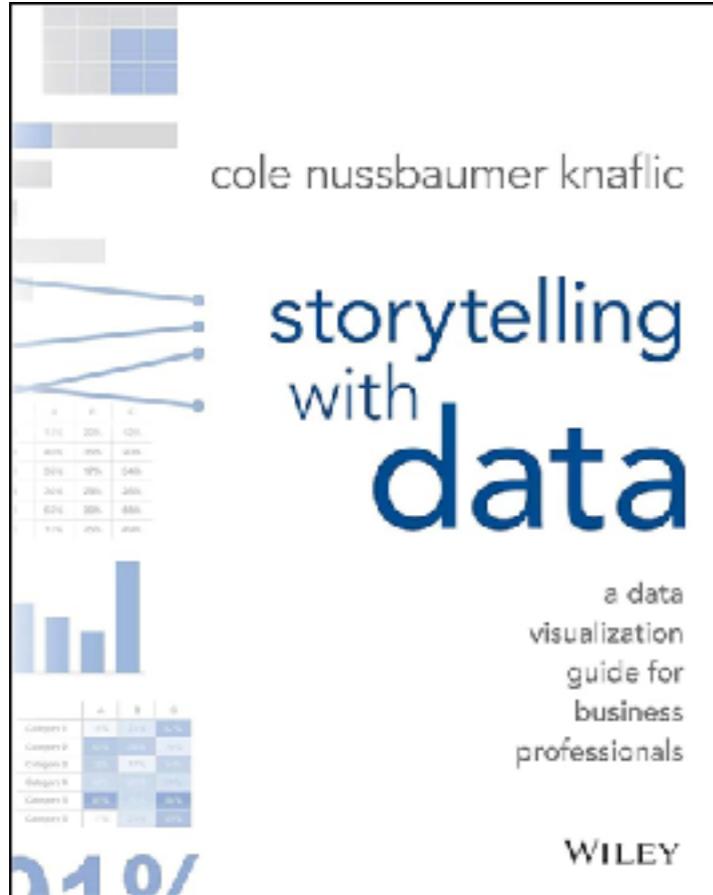
Data visualization is an **important skill** in applied statistics and machine learning.



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 ?

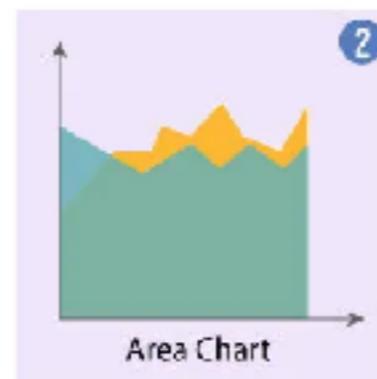
Before choosing a visual chart or graph,
it is important to understand your **audience**



TYPES OF DATA VISUALIZATION CHARTS



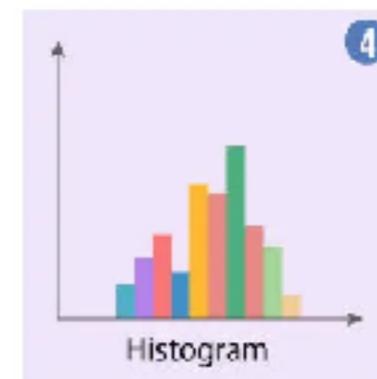
Display trends over time



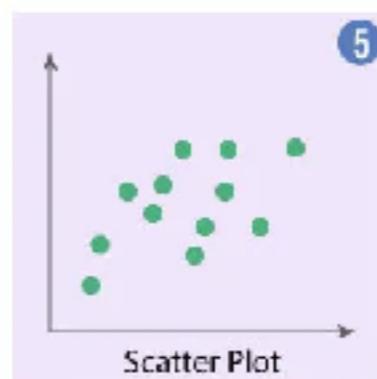
A line chart with areas below the lines filled with colors



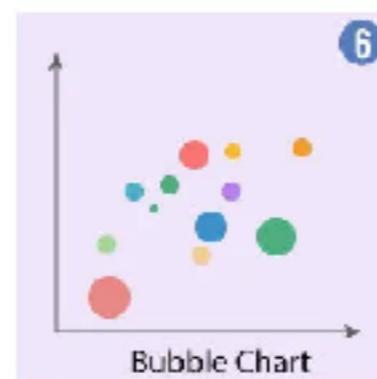
Display trends with multiple variables



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



Workshop



Data is everywhere

Descriptive

Predictive

Prescriptive

Metrics
Historical data

Insights
Modeling

Data products

เกิดอะไรขึ้นในอดีต ?

จะเกิดอะไรขึ้นในอนาคต ?

จะทำให้สิ่งที่
ผู้ใช้งานต้องการ
เกิดขึ้นได้อย่างไร



Understand your data

of rows

of columns

Column
description

Alignment

Styling

Cleaning data

Outlining

Summarize

Visualize



Workshop #1

Power of Data Visualization



Power of Visualization

A	B	C	D	E	F	G	H	
1	Anscombe's Quartet							
2	I	II		III		IV		
3	x1	y1	x2	y2	x3	y3	x4	y4
4	10	8.04	10	9.14	10	7.46	8	6.58
5	8	6.95	8	8.14	8	6.77	8	5.76
6	13	7.58	13	8.74	13	12.74	8	7.71
7	9	8.81	9	8.77	9	7.11	8	8.84
8	11	8.33	11	9.26	11	7.81	8	8.47
9	14	9.96	14	8.1	14	8.84	8	7.04
10	6	7.24	6	6.13	6	6.08	8	5.25
11	4	4.26	4	3.1	4	5.39	19	12.5
12	12	10.84	12	9.13	12	8.15	8	5.56
13	7	4.82	7	7.26	7	6.42	8	7.91
14	5	5.68	5	4.74	5	5.73	8	6.89
15								

<https://www.kaggle.com/datasets/carlmcbrideellis/data-anscombes-quartet/data>



Sharing

© 2020 - 2024 Siam Chamnankit Company Limited. All rights reserved.

Workshop #2

Data Transformation



Structure of Data ?

A	B	C	D	E	F	G	H	I	
1	Country	Year	Type	Count		Country	Year	Case	Population
2	Afghanistan	1999	cases	745		Afghanistan	1999	745	19987071
3	Afghanistan	1999	population	19987071		Afghanistan	2000	2666	20595360
4	Afghanistan	2000	cases	2666		Brazil	1999	37737	172006362
5	Afghanistan	2000	population	20595360		Brazil	2000	80488	174504898
6	Brazil	1999	cases	37737		China	1999	212258	1272915272
7	Brazil	1999	population	172006362		China	2000	213766	1280428583
8	Brazil	2000	cases	80488					
9	Brazil	2000	population	174504898					
10	China	1999	cases	212258					
11	China	1999	population	1272915272					
12	China	2000	cases	213766					
13	China	2000	population	1280428583					
14									
15									
16		1999		2000					
17	Country	Case	Population	Case	Population				
18	Afghanistan	745	19987071	2666	20595360				
19	Brazil	37737	172006362	80488	174504898				
20	China	212258	1272915272	213766	1280428583				
21									

<https://github.com/up1/course-basic-big-data-analytic/blob/main/workshop/data03.xlsx>



Sharing

© 2020 - 2024 Siam Chamnankit Company Limited. All rights reserved.

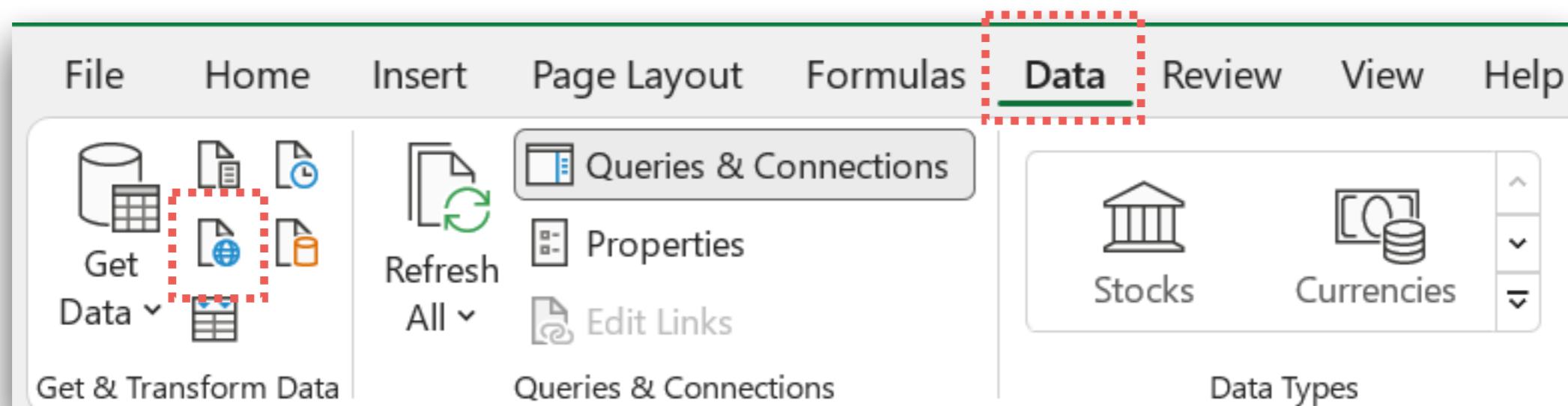
Workshop #3

Import data from web



Import data from web

Goto menu Data



Import data from web

Choose table data and load to Excel

The screenshot shows the Microsoft Power BI Navigator interface. On the left, there is a sidebar with a search bar and a list of tables under 'Select multiple items' and 'Display Options'. A specific table, 'Table 57', is highlighted with a green selection bar at the bottom. The main area displays 'Table View' of 'Table 57' with the following data:

ลำดับ	ชื่อสุก	ปี	ภาคผนวก	บริษัทผู้จัดจำหน่าย
1	1	2562	อาภรณ์เจริญ	บริษัทผู้จัดจำหน่าย
2	1	2556	พีระ..พาร์ค	จังหวัด
3	2	2561	อาภรณ์เจริญ มหาวิทยาลัยเชียงใหม่	จังหวัดเชียงใหม่
4	2	2558	เรือนแพเชียงใหม่	จังหวัด
5	5	2565	อาภรณ์เจริญสาขาแม่	จังหวัดเชียงใหม่
6	2	2557	ไทรโยค..บ้านเชียงใหม่	จังหวัด
7	1	2544	ศูนย์ฯ	ห้างหุ้นส่วนจำกัด บ้านเชียงใหม่แม่
8	5	2560	เชียงใหม่สุขุม 8	จังหวัด
9	5	2559	สำนักพัฒนาฯ ศูนย์ฯ เชียงใหม่	จังหวัดเชียงใหม่
10	3	2557	หมายฟิล์มเมอร์ 4: มหาวิทยาลัยเชียงใหม่	จังหวัด
11	2	2552	อาภรณ์	หมายฟิล์มเมอร์เชียงใหม่
12	7	2558	อาภรณ์เจริญ: มหาวิทยาลัยราชภัฏเชียงใหม่	จังหวัดเชียงใหม่
13	3	2554	หมายฟิล์มเมอร์ 3	จังหวัด
14	9	2558	อุไรสิน เรือง	จังหวัด
15	13	2561	อุไรสิน เรือง ลาก่อนวันครบรอบลาก่อน	จังหวัด
16	5	2556	ไทรโยค..บ้านเชียงใหม่	บริษัท ไทรโยค รับรองเชียงใหม่
17	15	2562	สำนักงานเชียงใหม่	จังหวัดเชียงใหม่

At the bottom right, there are buttons for 'Load', 'Transform Data', and 'Cancel'.



Example Data

Screenshot of Microsoft Excel showing two tables and the 'Queries & Connections' ribbon.

Table 1 (Left):

อันดับ	สูงสุด	ปี	ภาคยนตร์
ชื่อคุณ	สูงสุด	ปี	ภาคยนตร์
1	1	2562	อาเวนเจอร์ส: เมตัลลิก
2	1	2556	พิงก้า..พาราโนയ
3	2	2561	อาเวนเจอร์ส: มหาสงครามล้างจักรวาล
4	2	2558	เร็ว..แรงทะลุนรก 7
5	5	2565	อาทิตย์: วัยหงายสายฟ้า
6	2	2557	ไอไฟยก..แต่งกิ๊ว..เด็กฟู
7	1	2544	สุริโยไท
8	5	2560	เร็ว..แรงทะลุนรก 8
9	5	2559	กัปปั้นอเมริกา: ตึกอีซี่ไรรั่วห่าโลโก
10	3	2557	ทรานส์ฟอร์เมอร์ส 4: มหาวบดิบุคสูญพันธุ์
11	2	2552	อาทิตย์
12	7	2558	อาเวนเจอร์ส: มหาศึกอัลตรอนกล้มโลก
13	3	2554	ทรานส์ฟอร์เมอร์ส 3
14	9	2558	อุราสสัค เวิลต์
15	13	2561	อุราสสัค เวิลต์: อาทิตย์จักรล่มสลาย
16	5	2556	ไอรอนแมน 3
17	15	2562	กัปปั้น นาส์เวล
18	15	2561	อดาแมน เจ้าสมุทร
19	4	2555	ติ อเวนเจอร์ส
20	20	2567	ชีพบด 2+
21	20	2566	ลัปเปหรอ
22	19	2564	ลไปเดอร์แมน: โน เวท โน ชิม
23	19	2562	ลไปเดอร์แมน: ฟาร์ ฟรอม โนม

Table 2 (Right):

บริษัทผู้จัดจำหน่าย	ทำเงิน (ล้านบาท)	อ้างอิง
บริษัทฟูจิดจ้าหน่าย	617.55	ทำเงิน(ล้านบาท)
วอลต์ดิส尼บีบีพิกเชอร์ส	568.55	อ้างอิง
จีทีเอช	420.89	
วอลต์ดิสบีบีพิกเชอร์ส	387.85	
บูไอยิพี	376.23	
วอลต์ดิสบีบีพิกเชอร์ส	330.55	
สหนงค์คลิฟฟ์ม อินเตอร์เนชันแนล	324.5	
บูไอยิพี	321.96	
วอลต์ดิสบีบีพิกเชอร์ส	311.21	
บูไอยิพี	310.23	
ทเวนตี้ทีชานดูริฟอกซ์	303	
วอลต์ดิสบีบีพิกเชอร์ส	294.77	
บูไอยิพี	291.20	
บูไอยิพี	290.64	
บูไอยิพี	288.88	
บัวนา วิสต้า อินเตอร์เนชันแนล	262.92	
วอลต์ดิสบีบีพิกเชอร์ส	260.78	
วอร์เนอร์บราเธอร์สฟิกเจอร์ส	254.09	
บัวนา วิสต้า อินเตอร์เนชันแนล	252.16	
เจมส์ สตูดิโอ	245.93	
ไทยบ้าน สตูดิโอ	245.10	
โซนบีบิกเจอร์สคลิปชิป	239.29	
โซนบีบิกเจอร์สคลิปชิป	238.99	

Queries & Connections:

- Queries: ภาคยนตร์ที่ทำเงินสูงสุดในประเทศไทย (จำนวน 41 rows loaded).



Sharing

© 2020 - 2024 Siam Chamnankit Company Limited. All rights reserved.

Power Query

The screenshot shows the Microsoft Power Query Editor interface. The main area displays a table with four columns: 'Name' (含姓氏), 'Last Name (姓氏)' (姓氏), 'First Name (名前)' (名前), and 'Last Characters' (最后字母). The 'Last Name' column contains various names, while the 'First Name' column has some numerical values and one 'Error' entry. A tooltip at the bottom left indicates a 'DataFormat.Error' issue with the 'First Name' column. The 'Applied Steps' pane on the right lists the following steps:

- Source
- Navigation
- Changed Type
- Inserted Last Characters
- Changed Type1 (highlighted)

Below the table, the status bar shows "3 COLUMNS, 41 ROWS" and "Column profiling based on top 1000 rows". The "Query Settings" pane is visible on the right side.

Name	Last Name (姓氏)	First Name (名前)	Last Characters
กานต์	กานต์	Error	ก
กานต์	กานต์	617.55	62
กานต์	กานต์	568.55	56
กานต์	กานต์	420.89	61
กานต์	กานต์	387.89	58
กานต์	กานต์	376.23	65
กานต์	กานต์	330.55	57
กานต์	กานต์	324.5	44
กานต์	กานต์	321.96	60
กานต์	กานต์	311.21	59
กานต์	กานต์	310.23	57
กานต์	กานต์	303	52
กานต์	กานต์	294.77	58
กานต์	กานต์	291.2	54
กานต์	กานต์	290.64	58



Excel Table

Select any cells and press **Ctrl + t**
Or goto **Insert -> Table**

Sorting

Filter

Add formula

Total

Table design



Excel Table

File Home Insert Page Layout Formulas Data Review View Help Power Pivot Table Design Query

G9 : =IF([@gross]>500, "100M+", "Under 100M")

A	B	C	D	E	F	G	
กับล้าน	ล้านบาท	ปี	รายการ	จำนวนเงินที่หักภาษี	gross	Column1	
จำนวน	ล้านบาท	ปี	รายการ	จำนวนเงินที่หักภาษี	gross	Under 100M	
1	1	2562	ค่าใช้จ่าย: เมืองศึกษา	จำนวนที่หักภาษี	617.55	100M+	
2	1	2562	ค่าใช้จ่าย: เมืองศึกษา	จำนวนที่หักภาษี	508.50	100M+	
3	2	2561	ค่าใช้จ่าย: ค่าเชื้อเชิญและจัดตั้งบ้าน	จำนวนที่หักภาษี	470.84	Under 100M	
4	2	2559	ค่าใช้จ่าย: ค่าเชื้อเชิญและจัดตั้งบ้าน	จำนวนที่หักภาษี	387.65	Under 100M	
5	3	2561	ค่าใช้จ่าย: ค่าเชื้อเชิญและจัดตั้งบ้าน	จำนวนที่หักภาษี	376.20	Under 100M	
6	4	2558	ค่าใช้จ่าย: ค่าเชื้อเชิญและจัดตั้งบ้าน 7	จำนวนที่หักภาษี	340.05	Under 100M	
7	5	2565	ค่าใช้จ่าย: ค่าเชื้อเชิญและจัดตั้งบ้าน	จำนวนที่หักภาษี	324.5	Under 100M	
8	6	2567	ค่าใช้จ่าย: ค่าเชื้อเชิญและจัดตั้งบ้าน	จำนวนที่หักภาษี	321.05	Under 100M	
9	7	1	2544	ค่าใช้จ่าย:	311.21	Under 100M	
10	8	5	2560	ค่าใช้จ่าย: ค่าเชื้อเชิญและจัดตั้งบ้าน 8	จำนวนที่หักภาษี	310.24	Under 100M
11	9	5	2552	ค่าใช้จ่าย: ค่าเชื้อเชิญและจัดตั้งบ้าน	จำนวนที่หักภาษี	303	Under 100M
12	10	4	2567	ค่าใช้จ่าย: ค่าเชื้อเชิญและจัดตั้งบ้าน 9: ค่าเชื้อเชิญและจัดตั้งบ้าน	จำนวนที่หักภาษี	298.88	Under 100M
13	11	2	2552	ค่าใช้จ่าย:	292	Under 100M	
14	12	2	2558	ค่าใช้จ่าย: ค่าเชื้อเชิญและจัดตั้งบ้าน	จำนวนที่หักภาษี	288.88	Under 100M
15	13	3	2551	ค่าใช้จ่าย: ค่าเชื้อเชิญและจัดตั้งบ้าน 3	จำนวนที่หักภาษี	282.02	Under 100M
16	14	4	2564	ค่าใช้จ่าย: ค่าเชื้อเชิญและจัดตั้งบ้าน	จำนวนที่หักภาษี	270.14	Under 100M
17	15	13	2561	ค่าใช้จ่าย: ค่าเชื้อเชิญและจัดตั้งบ้าน	จำนวนที่หักภาษี	262.02	Under 100M
18	16	5	2556	ค่าใช้จ่าย: ค่าเชื้อเชิญและจัดตั้งบ้าน 3	จำนวนที่หักภาษี	260.73	Under 100M
19	17	15	2562	ค่าใช้จ่าย: ค่าเชื้อเชิญและจัดตั้งบ้าน	จำนวนที่หักภาษี	254.04	Under 100M
20	18	15	2561	ค่าใช้จ่าย: ค่าเชื้อเชิญและจัดตั้งบ้าน	จำนวนที่หักภาษี	252.15	Under 100M
21	19	4	2555	ค่าใช้จ่าย: ค่าเชื้อเชิญและจัดตั้งบ้าน	จำนวนที่หักภาษี	245.00	Under 100M
22	20	20	2567	ค่าใช้จ่าย: ค่าเชื้อเชิญและจัดตั้งบ้าน	จำนวนที่หักภาษี	243.1	Under 100M
23	21	20	2566	ค่าใช้จ่าย:	244.24	Under 100M	
24	22	19	2564	ค่าใช้จ่าย: ค่าเชื้อเชิญและจัดตั้งบ้าน 19	จำนวนที่หักภาษี	238.00	Under 100M
25	23	19	2562	ค่าใช้จ่าย: ค่าเชื้อเชิญและจัดตั้งบ้าน 19	จำนวนที่หักภาษี	236.20	Under 100M
26	24	4	2551	ค่าใช้จ่าย: ค่าเชื้อเชิญและจัดตั้งบ้าน 24: ค่าเชื้อเชิญและจัดตั้งบ้าน 2	จำนวนที่หักภาษี	230.42	Under 100M
27	25	22	2567	ค่าใช้จ่าย: ค่าเชื้อเชิญและจัดตั้งบ้าน 25	จำนวนที่หักภาษี	229.06	Under 100M
28	26	2	2550	ค่าใช้จ่าย: ค่าเชื้อเชิญและจัดตั้งบ้าน 26: ค่าเชื้อเชิญและจัดตั้งบ้าน 2	จำนวนที่หักภาษี	216.67	Under 100M
29	27	3	2550	ค่าใช้จ่าย: ค่าเชื้อเชิญและจัดตั้งบ้าน 27: ค่าเชื้อเชิญและจัดตั้งบ้าน 3	จำนวนที่หักภาษี	214.58	Under 100M
30	28	4	2552	2012 หักภาษี	จำนวนที่หักภาษี	213.77	Under 100M
31	29	11	2566	ค่าใช้จ่าย: ค่าเชื้อเชิญและจัดตั้งบ้าน 29	จำนวนที่หักภาษี	213.65	Under 100M
32	30	1	2540	ค่าใช้จ่าย:	206.05	Under 100M	
33	31	13	2557	ค่าใช้จ่าย: ค่าเชื้อเชิญและจัดตั้งบ้าน 31: ค่าเชื้อเชิญและจัดตั้งบ้าน 13	จำนวนที่หักภาษี	201.0	Under 100M
34	32	3	2552	ค่าใช้จ่าย: ค่าเชื้อเชิญและจัดตั้งบ้าน 32: ค่าเชื้อเชิญและจัดตั้งบ้าน 3	จำนวนที่หักภาษี	201.08	Under 100M
35	33	8	2564	ค่าใช้จ่าย: ค่าเชื้อเชิญและจัดตั้งบ้าน 33: ค่าเชื้อเชิญและจัดตั้งบ้าน 8	จำนวนที่หักภาษี	199.04	Under 100M
36	34	23	2561	ค่าใช้จ่าย: ค่าเชื้อเชิญและจัดตั้งบ้าน 34: ค่าเชื้อเชิญและจัดตั้งบ้าน 23	จำนวนที่หักภาษี		



Sharing

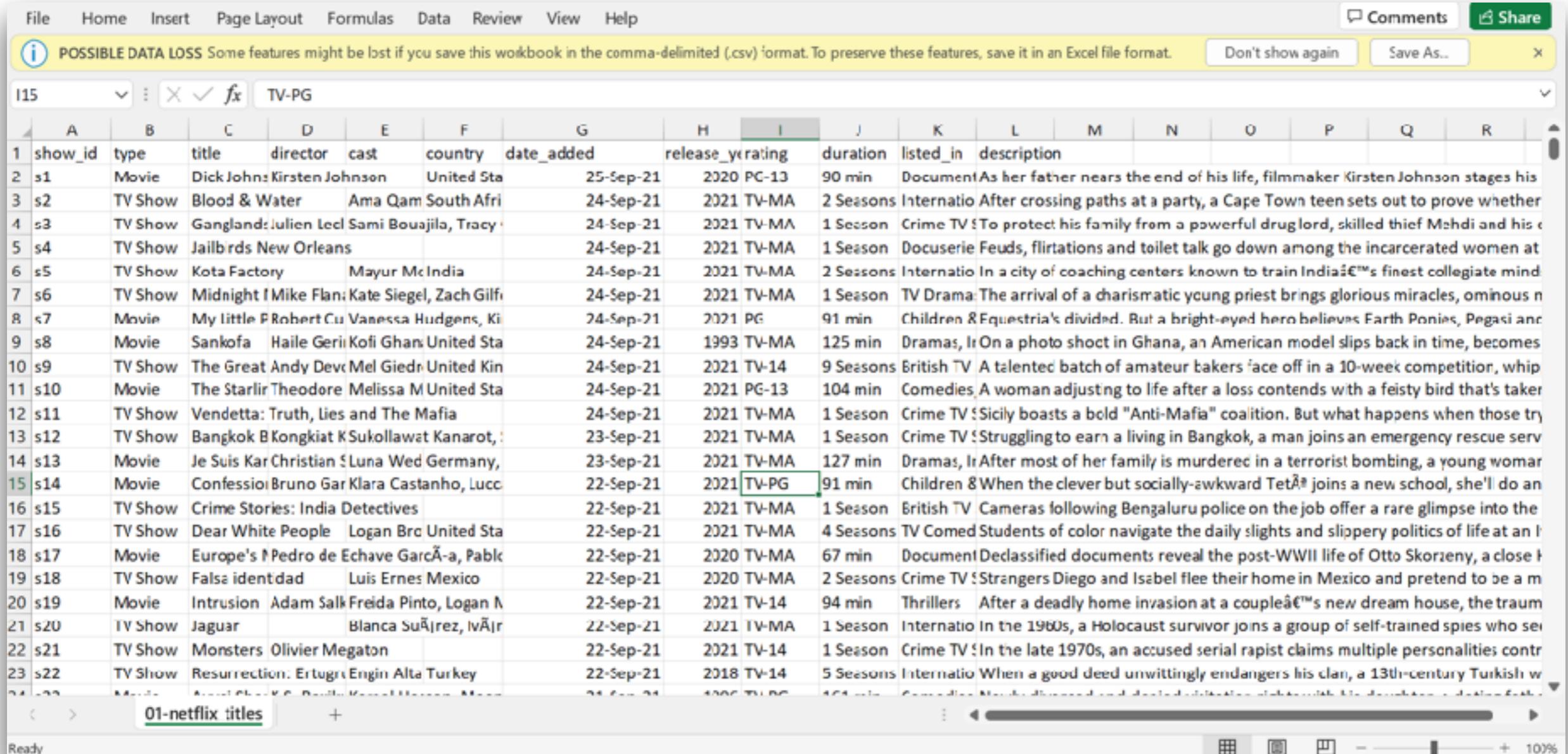
© 2020 - 2024 Siam Chamnankit Company Limited. All rights reserved.

Workshop #3

01-netflix_titles



Open Data in Excel



show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
s1	Movie	Dick Johnson	Kirsten Johnson		United States	25-Sep-21	2020	PG-13	90 min	Documentary	As her father nears the end of his life, filmmaker Kirsten Johnson stages his final film.
s2	TV Show	Blood & Water	Ama Qam	South Africa		24-Sep-21	2021	TV-MA	2 Seasons	International	After crossing paths at a party, a Cape Town teen sets out to prove whether she can change her family's violent past.
s3	TV Show	Gangland	Julien Leclercq	Sami Bouajila, Tracy	France	24-Sep-21	2021	TV-MA	1 Season	Crime TV	To protect his family from a powerful drug lord, skilled thief Mehdi and his crew must infiltrate a gangland prison.
s4	TV Show	Jailbirds	New Orleans		United States	24-Sep-21	2021	TV-MA	1 Season	Documentary	Feuds, flirtations and toilet talk go down among the incarcerated women at a maximum-security prison.
s5	TV Show	Kota Factory		Mayur McIndia	India	24-Sep-21	2021	TV-MA	2 Seasons	International	In a city of coaching centers known to train India's finest collegiate mind, a group of students prepare for their exams.
s6	TV Show	Midnight	Mike Flanagan	Kate Siegel, Zach Gilford	United States	24-Sep-21	2021	TV-MA	1 Season	TV Drama	The arrival of a charismatic young priest brings glorious miracles, ominous new powers and a secret past to a small town.
s7	Movie	My Little Pony	Robert Cuccioli	Vanessa Hudgens, Kiersey Clemons	United States	24-Sep-21	2021	PG	91 min	Children & Family	Equestria's divided. But a bright-eyed hero believes Earth Ponies, Pegasus and Unicorns can work together to save the day.
s8	Movie	Sankofa	Haile Gerima	Kofi Ghan	United States	24-Sep-21	1993	TV-MA	125 min	Dramas	On a photo shoot in Ghana, an American model slips back in time, becomes a child and must navigate a dangerous world.
s9	TV Show	The Great Andy Devine	Mel Giedroyc		United Kingdom	24-Sep-21	2021	TV-14	9 Seasons	British TV	A talented batch of amateur bakers face off in a 10-week competition, whipping up some serious drama along the way.
s10	Movie	The Star	Levi Theodore	Melissa Mays	United States	24-Sep-21	2021	PG-13	104 min	Comedies	A woman adjusting to life after a loss contends with a feisty bird that's taken over her body.
s11	TV Show	Vendetta: Truth, Lies and The Mafia				24-Sep-21	2021	TV-MA	1 Season	Crime TV	Sicily boasts a bold "Anti-Mafia" coalition. But what happens when those trying to bring them down team up?
s12	TV Show	Bangkok B	Kongkiat K	Sukollawat Kanarot, Jirayut	Thailand	23-Sep-21	2021	TV-MA	1 Season	Crime TV	Struggling to earn a living in Bangkok, a man joins an emergency rescue service.
s13	Movie	Je Suis Kar	Christian St	Luna Wed	Germany, France	23-Sep-21	2021	TV-MA	127 min	Dramas	After most of her family is murdered in a terrorist bombing, a young woman must confront her past and find a way to move forward.
s14	Movie	Confession	Bruno Ganz	Klara Castanho, Lucca	Portugal	22-Sep-21	2021	TV-PG	91 min	Children & Family	When the clever but socially-awkward Teté joins a new school, she'll do anything to fit in.
s15	TV Show	Crime Stories: India Detectives				22-Sep-21	2021	TV-MA	1 Season	British TV	Cameras following Bengaluru police on the job offer a rare glimpse into the daily lives of these officers.
s16	TV Show	Dear White People	Logan Browning		United States	22-Sep-21	2021	TV-MA	4 Seasons	TV Comedies	Students of color navigate the daily slights and slippery politics of life at an Ivy League school.
s17	Movie	Europe's Next	Pedro de Echave	García, Pablo	Spain	22-Sep-21	2020	TV-MA	67 min	Documentary	Declassified documents reveal the post-WWII life of Otto Skorzeny, a close friend of Adolf Hitler.
s18	TV Show	Falsa Identidad	Luis Ernesto	Mexico	Mexico	22-Sep-21	2020	TV-MA	2 Seasons	Crime TV	Strangers Diego and Isabel flee their home in Mexico and pretend to be a married couple.
s19	Movie	Intrusion	Adamalko	Freida Pinto, Logan	Netherlands	22-Sep-21	2021	TV-14	94 min	Thrillers	After a deadly home invasion at a couple's new dream house, the trauma continues.
s20	TV Show	Jaguar		Blanca Suárez, Iván	Spain	22-Sep-21	2021	TV-MA	1 Season	International	In the 1960s, a Holocaust survivor joins a group of self-trained spies who seek justice.
s21	TV Show	Monsters	Olivier Megaton			22-Sep-21	2021	TV-14	1 Season	Crime TV	In the late 1970s, an accused serial rapist claims multiple personalities controlled by different parts of his brain.
s22	TV Show	Resurrection: Ertugrul	Engin Altay		Turkey	22-Sep-21	2018	TV-14	5 Seasons	International	When a good deed unwittingly endangers his clan, a 13th-century Turkish warrior must make a difficult choice.

<https://www.kaggle.com/datasets/shivamb/netflix-shows>



Sharing

© 2020 - 2024 Siam Chamnkit Company Limited. All rights reserved.

Step 1

Observe your data !!

of rows

of columns

Column
description



Exploratory Data Analysis (EDA)

Data
transformation

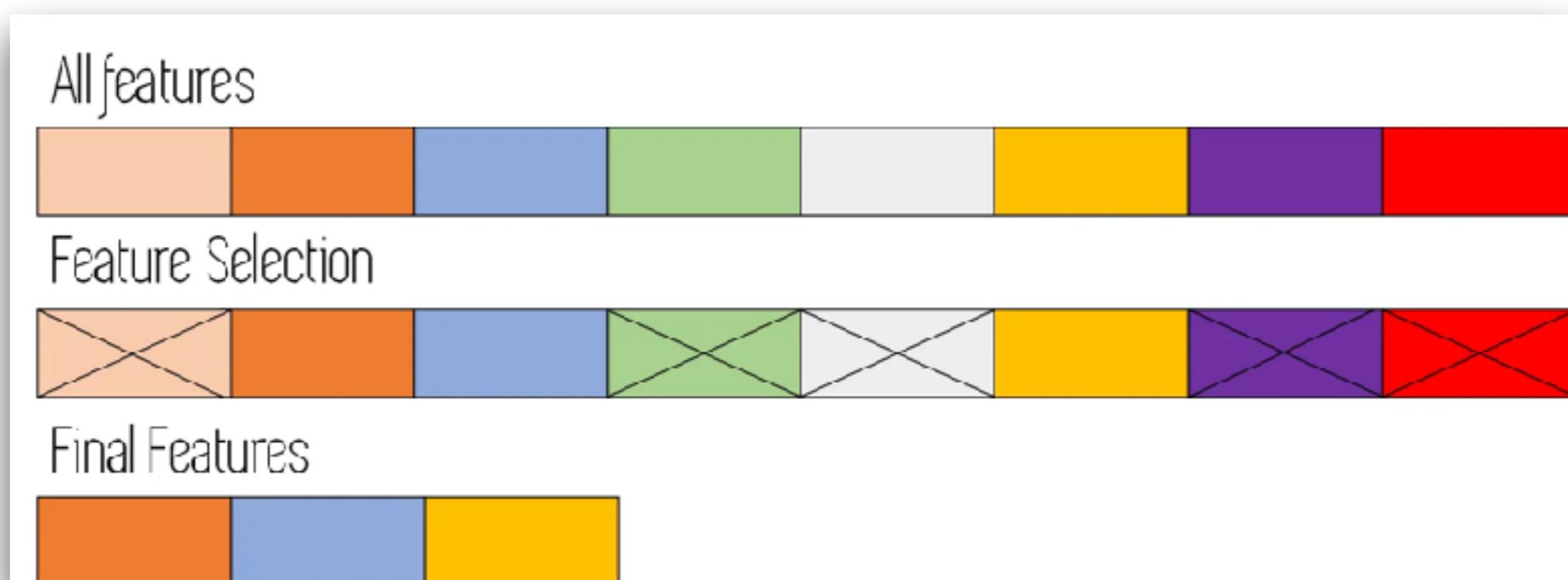
Data
analysis

Data
visualization



Step 2 :: Feature selection

Consists on selecting the best features for our models and algorithms, by taking these insights from the data



Feature selection ?

Title and description columns are low priority



Step 3 :: Start with Question ?



Sample questions ?

Top 10 director has cast more movie, tv show ?

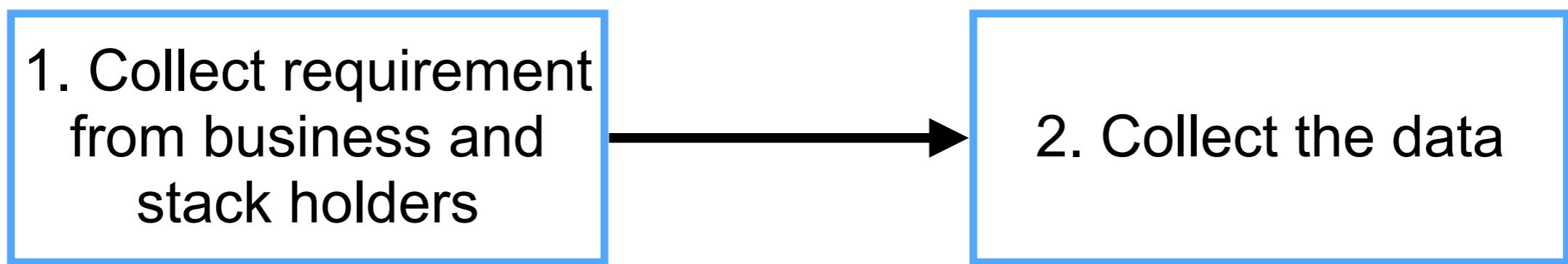
Top 5 country has produced more movies and TV shows ?

Which Genre has more in Movie Type ?

Count the movies & TV Shows before & after the year 2010 ?



Steps in data analytic

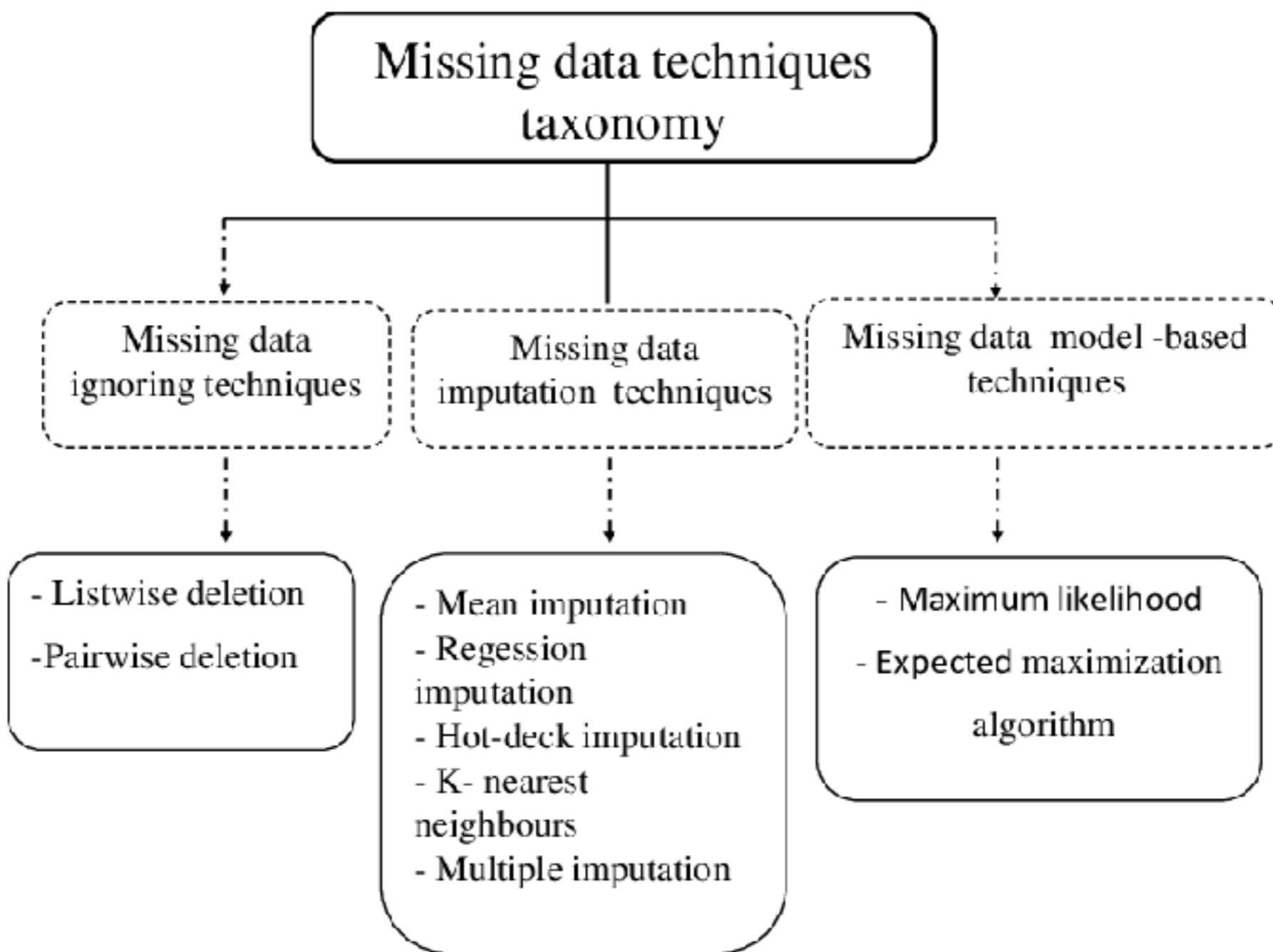


Step 4 :: Clean your data

Empty/missing value
Remove duplication and unwanted data
Outliner data



Missing data techniques



Working with Excel #1

Press CTRL + T

To make it as a table for readability purpose



Working with Excel #2

Look at each column
and check it have **empty values**

COUNTBLANK()

COUNTA()



Working with Excel #2

Check your data with = COUNTA(table[column])

1	Counting Value
2	show_id
3	type
4	title
5	director
6	cast
7	country
8	date_added
9	release_year
10	rating
11	duration
12	listed_in



Working with Excel #3

Duplicated data in each columns

Listed_in

Director

Cast

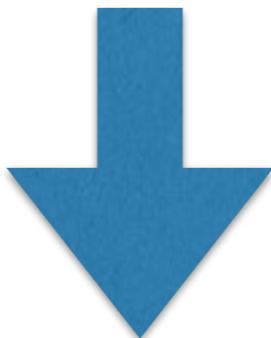
Country



Working with Excel #3

Sample with listed_in column

Crime TV Shows, International TV Shows, TV Action & Adventure



Crime TV
Shows

International TV
Shows

TV Action &
Adventure



Working with Excel #3

Select Data -> Text to columns

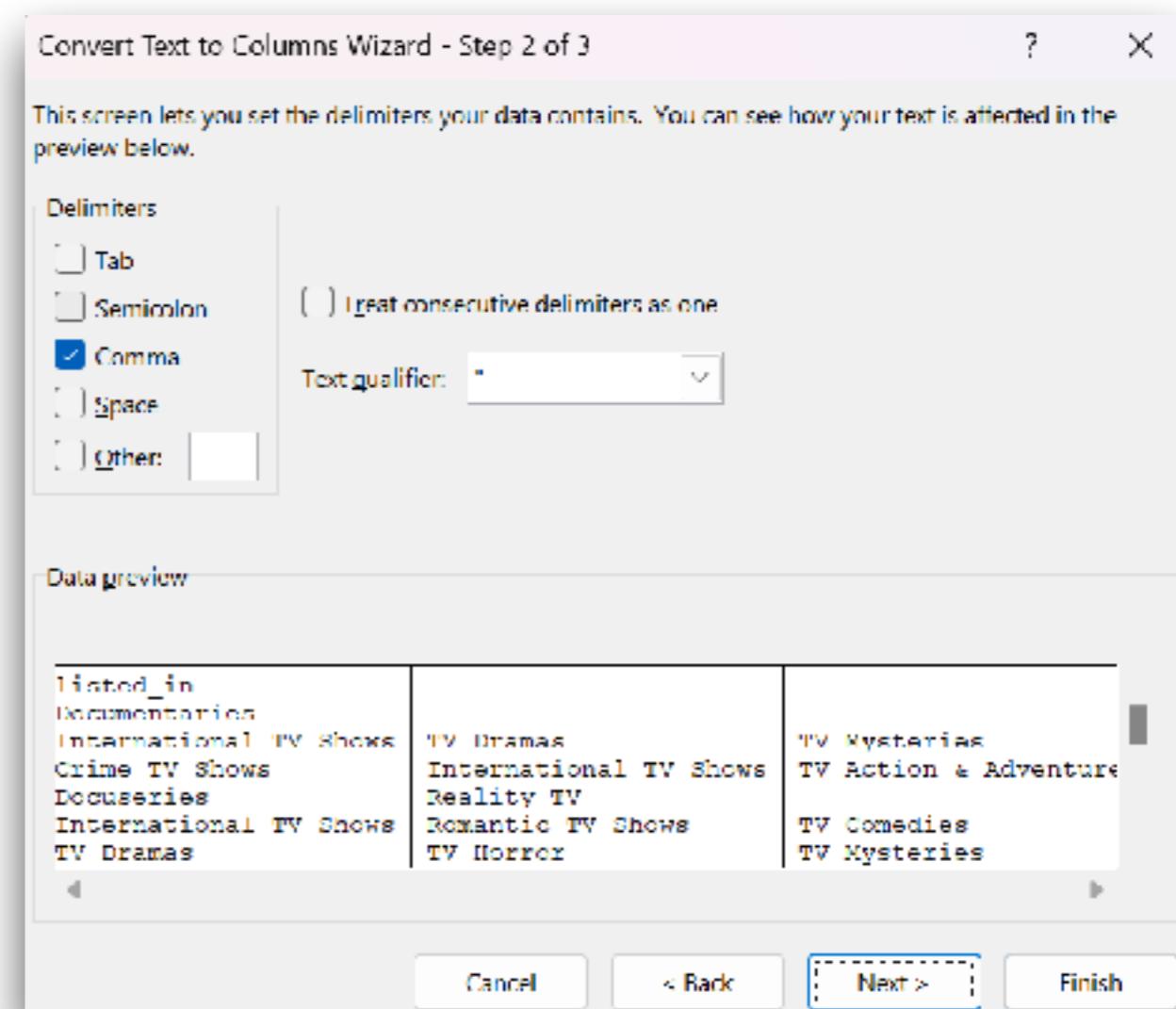
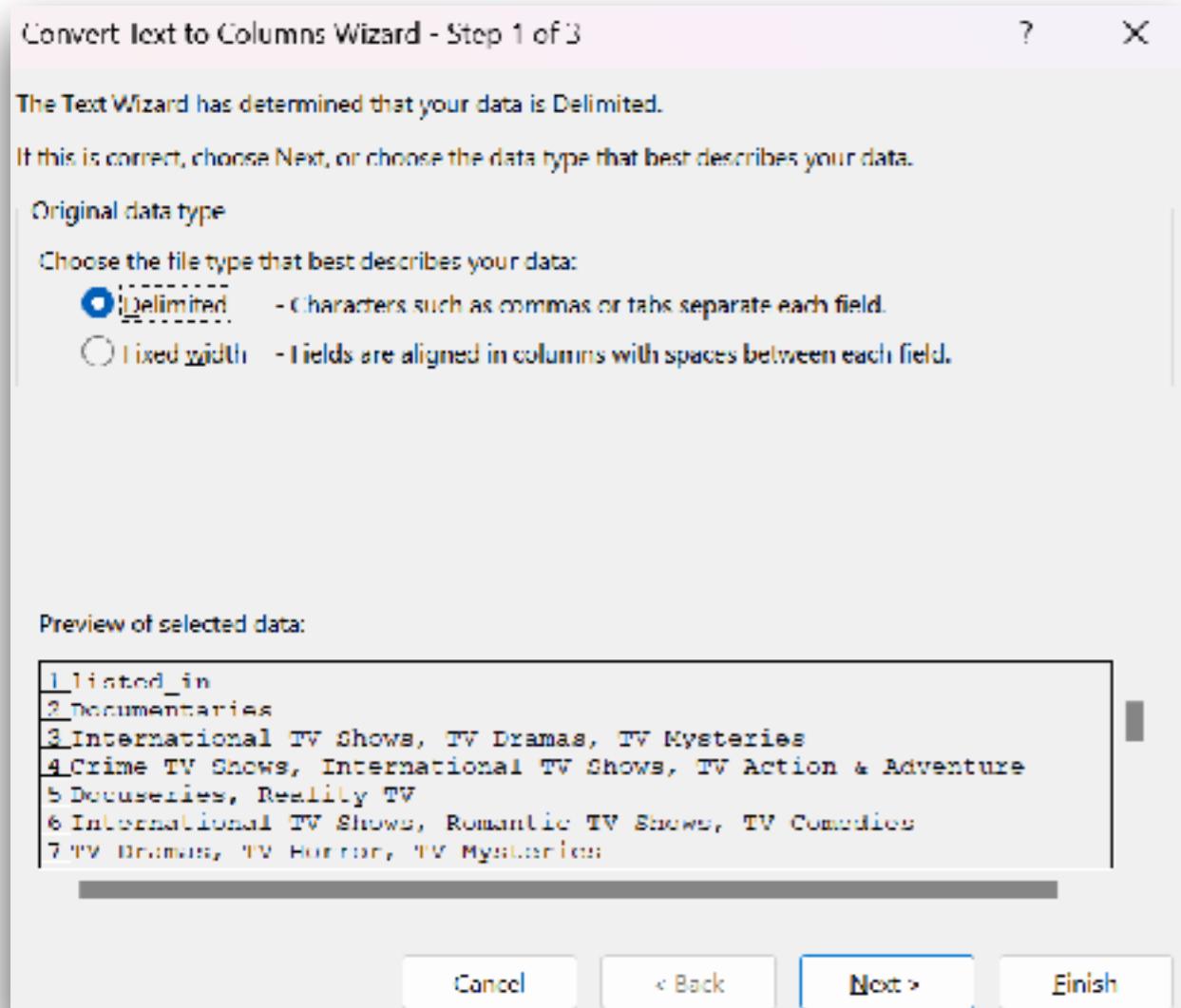
The screenshot shows a Microsoft Excel spreadsheet with a red dashed box highlighting the 'Text to Columns' button in the 'Data Tools' section of the ribbon. A tooltip for 'Text to Columns' is displayed, explaining that it splits a single column of text into multiple columns. It provides examples of separating full names into first and last name columns and splitting by fixed width or delimiter. A 'Tell me more' link is also present.

3	I later	Ama Qam South Afric	24-Sep-21	2021 TV-MA	2 Seasons	International TV Shows, TV Crime TV Shows, International Docuseries, Reality TV
4	Julien Lelio Sami Bouajila, Tracy C		24 Sep 21	2021 TV-MA	1 Season	TV Dramas, TV Horror, TV Children & Family Movies
5	ew Orleans		24-Sep-21	2021 TV-MA	1 Season	Dramas, Independent Movie British TV Shows, Reality TV Comedies, Dramas
6	TV	Mayur Mc India	24-Sep-21	2021 TV-MA	2 Seasons	Crime TV Shows, Docuseries, International TV Dramas, International Movie British TV Shows, Crime TV Shows, Docuseries
7	Mike Flanagan, Kate Siegel, Zach Gilfoyle		24 Sep 21	2021 TV-MA	1 Season	Children & Family Movies, Thrillers
8	Robert Cullum Vanessa Hudgens, Kira Bell		24-Sep-21	2021 PG	91 min	International TV Shows, TV Action & Adventure
9	Haile Gerin Kofi Ghan	United States	24-Sep-21	1993 TV-MA	125 min	International TV Shows, TV Action & Adventure, TV Dramas
10	Andy Devine Mel Giedroyc	United Kingdom	24-Sep-21	2021 TV-14	9 Seasons	International TV Shows, TV Action & Adventure, TV Dramas
11	Theodore T. Melissa M. McCarthy	United States	24-Sep-21	2021 PG-13	104 min	International TV Shows, TV Action & Adventure, TV Dramas
12	Truth, Lies and The Mafia		24-Sep-21	2021 TV-MA	1 Season	International TV Shows, TV Action & Adventure, TV Dramas
13	Kongkiat Kongkollawat Kanarot, S		23-Sep-21	2021 TV-MA	1 Season	International TV Shows, TV Action & Adventure, TV Dramas
14	Christian S. Luna Wed	Germany, France	23-Sep-21	2021 TV-MA	127 min	International TV Shows, TV Action & Adventure, TV Dramas
15	Bruno Garcia Klara Castanho, Lucca		22-Sep-21	2021 TV-PG	91 min	International TV Shows, TV Action & Adventure, TV Dramas
16	ies: India Detectives		22-Sep-21	2021 TV-MA	1 Season	International TV Shows, TV Action & Adventure, TV Dramas
17	the People	Logan Browning United States	22-Sep-21	2021 TV-MA	4 Seasons	International TV Shows, TV Action & Adventure, TV Dramas
18	Pedro de Echave Garcia, Pablo		22-Sep-21	2020 TV-MA	67 min	International TV Shows, TV Action & Adventure, TV Dramas
19	idad	Luis Ernesto Mexico	22-Sep-21	2020 TV-MA	2 Seasons	International TV Shows, TV Action & Adventure, TV Dramas
20	Adam Salky Freida Pinto, Logan Marshall		22-Sep-21	2021 TV-14	94 min	International TV Shows, TV Action & Adventure, TV Dramas
21	Blanca Suárez, Iván		22-Sep-21	2021 TV-MA	1 Season	International TV Shows, TV Action & Adventure, TV Dramas
22	Olivier Megaton		22-Sep-21	2021 TV-14	1 Season	International TV Shows, TV Action & Adventure, TV Dramas
23	on: Ertugrul Engin Altay	Turkey	22-Sep-21	2018 TV-14	5 Seasons	International TV Shows, TV Action & Adventure, TV Dramas



Working with Excel #3

Choose delimiter with comma (,)



Working with Excel #3

Result in Excel file

listed_in	Column1	Column2
Documentaries		
International TV Shows	TV Dramas	TV Mysteries
Crime TV Shows	International TV Shows	TV Action & Adventure
Docuseries	Reality TV	
International TV Shows	Romantic TV Shows	TV Comedies
TV Dramas	TV Horror	TV Mysteries
Children & Family Movies		
Dramas	Independent Movies	International Movies
British TV Shows	Reality TV	
Comedies	Dramas	
Crime TV Shows	Docuseries	International TV Shows
Crime TV Shows	International TV Shows	TV Action & Adventure
Dramas	International Movies	
Children & Family Movies	Comedies	
British TV Shows	Crime TV Shows	Docuseries
TV Comedies	TV Dramas	
Documentaries	International Movies	
Crime TV Shows	Spanish-Language TV Shows	TV Dramas
Thrillers		
International TV Shows	Spanish-Language TV Shows	TV Action & Adventure
Crime TV Shows	Docuseries	International TV Shows



Try by yourself

Listed_in

Director

Cast

Country



Step 5 :: Start analyze the data

Top 10 director has cast more movie ?

Top 5 country has produced more movies and TV shows ?

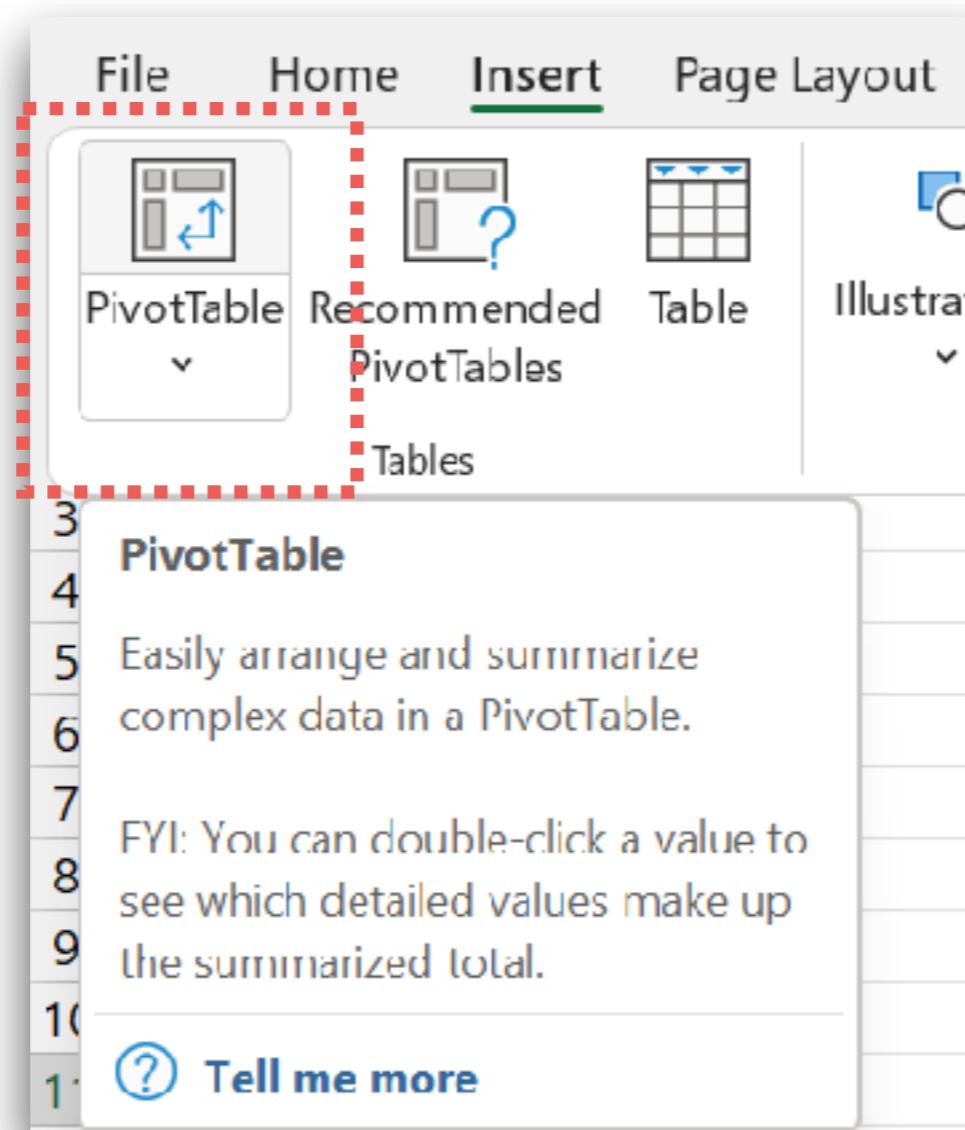
Which Genre has more in Movie Type ?

Count the movies & TV Shows before & after the year 2010 ?



Top 10 director has cast more movie ?

Working with Pivot Table



Top 10 director has cast more movie ?

Choose Columns, Rows and Values

The screenshot shows the 'PivotTable Fields' dialog box from Microsoft Excel. In the 'Choose fields to add to report:' section, 'type' and 'director' are checked. In the 'Drag fields between areas below:' section, 'director' is selected under 'Rows'. On the right, the 'Columns' pane shows 'type' and the 'Values' pane shows 'Count of director'. Red annotations on the left side of the dialog box read '1. Column = type', '2. Rows = director', and '3. Values = director'.

PivotTable Fields

Choose fields to add to report:

Search

type
 title
 director
 cast
 country

Drag fields between areas below:

Filters

1. Column = type

2. Rows = director

3. Values = director

Columns

type

Σ Values

Count of director

Rows

director

Defer Layout Update



Top 10 director has cast more movie ?

Count of director Row Labels	Column Labels		
	Movie	TV Show	Grand Total
A. L. Vijay		2	2
A. Raajdheep		1	1
A. Salaam		1	1
A.R. Murugadoss		2	2
Ã“skar ThÃ³r Axelsson		1	1
Ã€lex Pastor, David Pastor		2	2
Ã‡agan Irmak		1	1
Ãlex de la Iglesia		2	2
Ãlvaro Brechner		1	1
Ãlvaro Delgado-Aparicio L.		1	1
Ãlvaro Longoria, Gerardo Olivares		1	1
Ãngel GÃ³mez HernÃ¡ndez		1	1
Ãngeles ReinÃ©		1	1
Ãsold UggadÃ³ttir		1	1
Aadish Keluskar		1	1
A. S. D. L.		1	1



Try by yourself

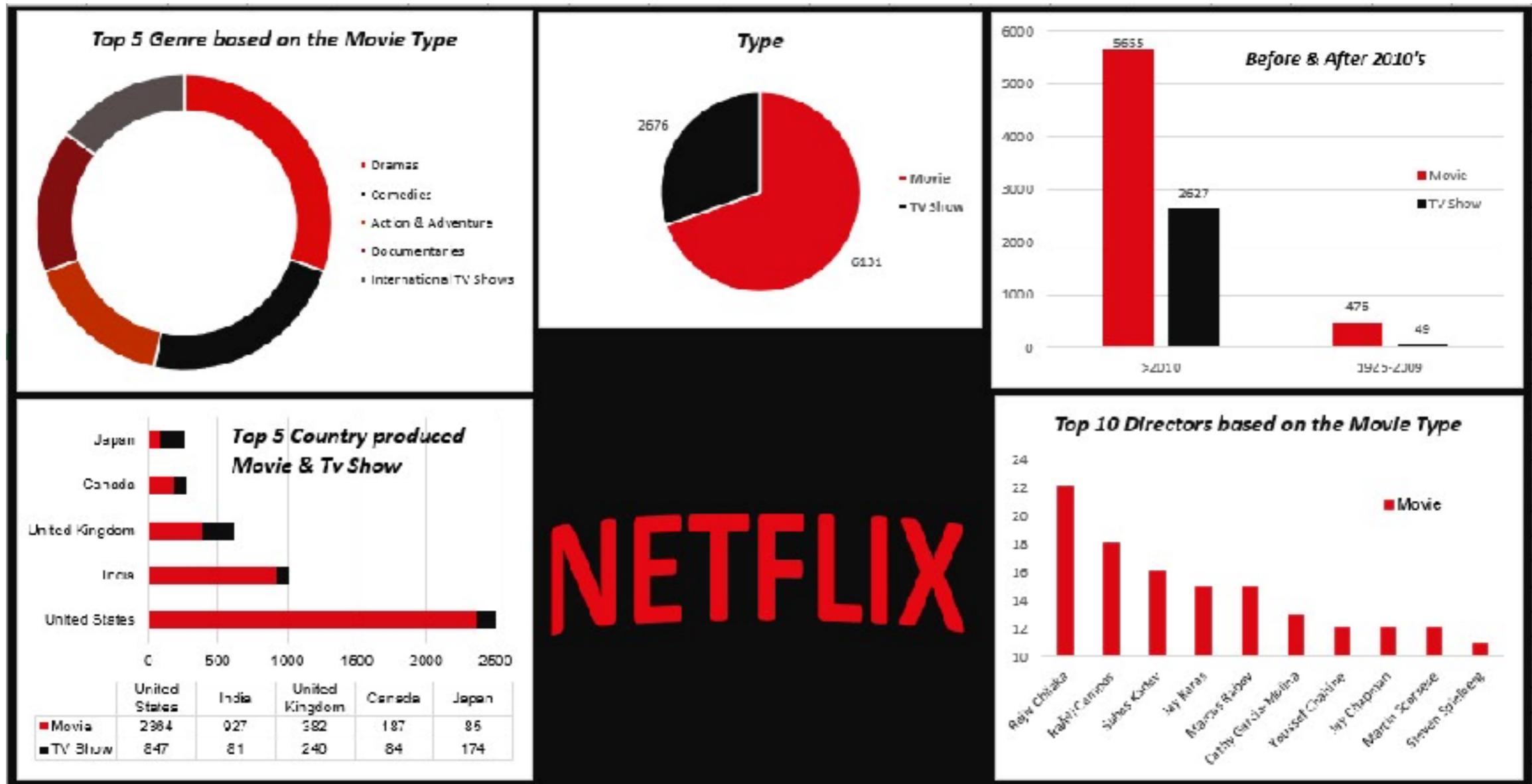
Top 5 country has produced more movies and TV shows ?

Which Genre has more in Movie Type ?

Count the movies & TV Shows before & after the year 2010 ?



Step 6 :: Data Visualization



Step 7 ::

Communicate

the insight/findings to business



Workshop with Power BI Desktop



Power BI



Power BI (Business Intelligence)

Complete reporting solutions !!

Provide development tools and platforms

**Visual drag&drop data exploration
and interactive reporting**

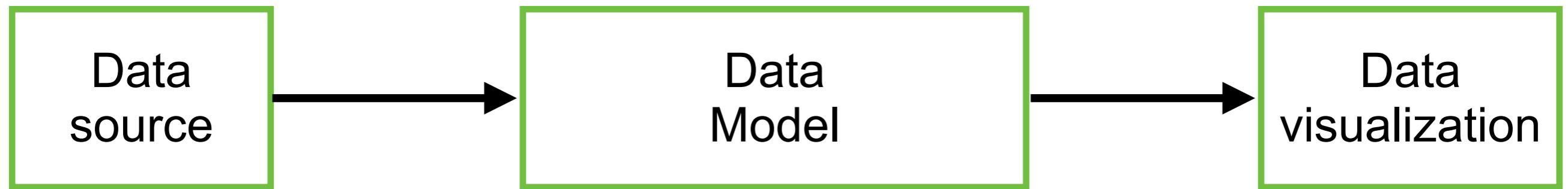
Data preparation

Data visualization

Distribution



Flow of Power BI



Excel
Database
Data Lake
Data Mart
Data Warehouse

Cleansing
Transformation
Relationship
Calculation



Flow of Power BI



Excel, CSV

Database

Data Lake

Data Mart

Data Warehouse

Cleansing
Transformation
Relationship
Calculation

Business Analyst



Data Source Cleansing

Working with Microsoft Excel

Table format

1 Table per sheet

1 Column per 1 content

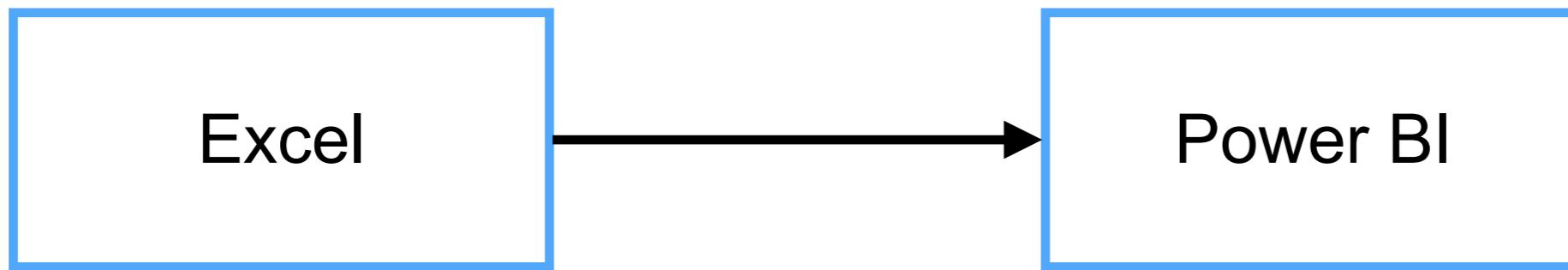
No merge cell

No blank/empty cell

Header/Record



Excel vs Power BI



Data source

Cleansing

Transformation

Power BI

Visualization



Types of Power BI

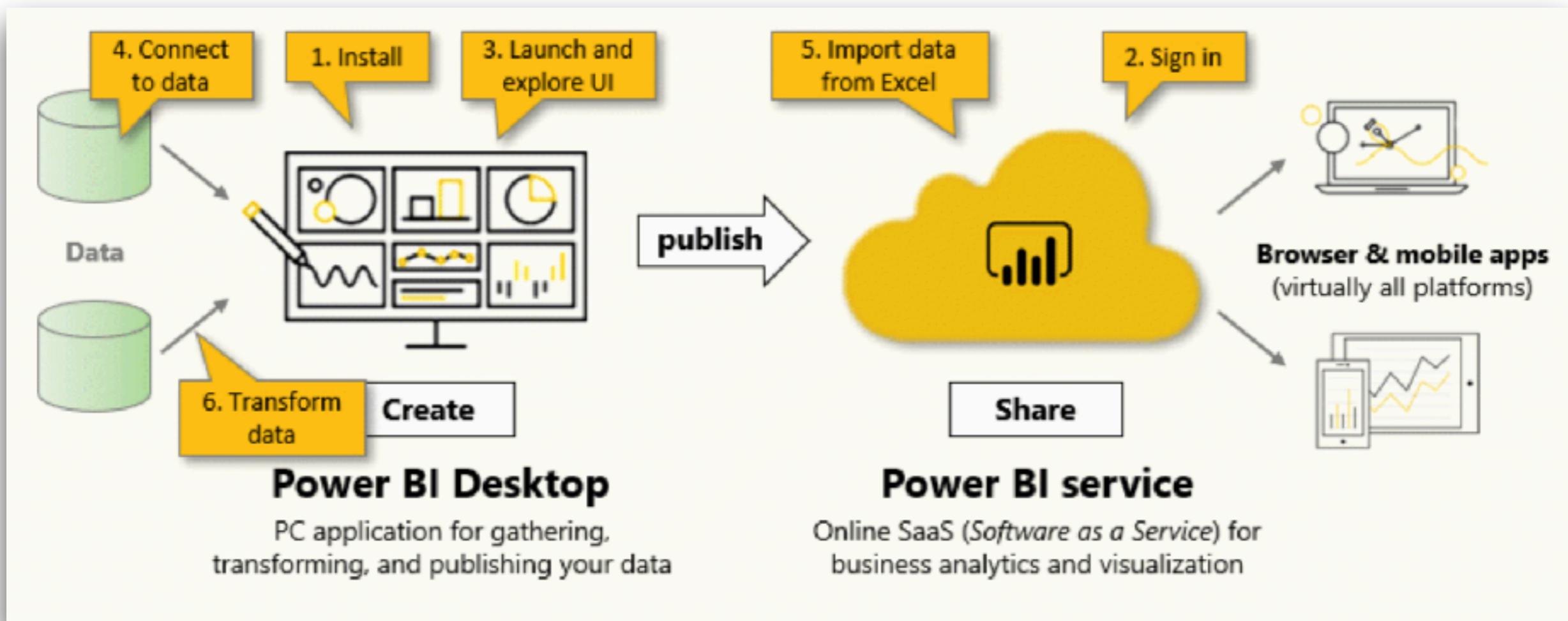
Web app (Power BI Service)

Desktop (Windows only)

Mobile



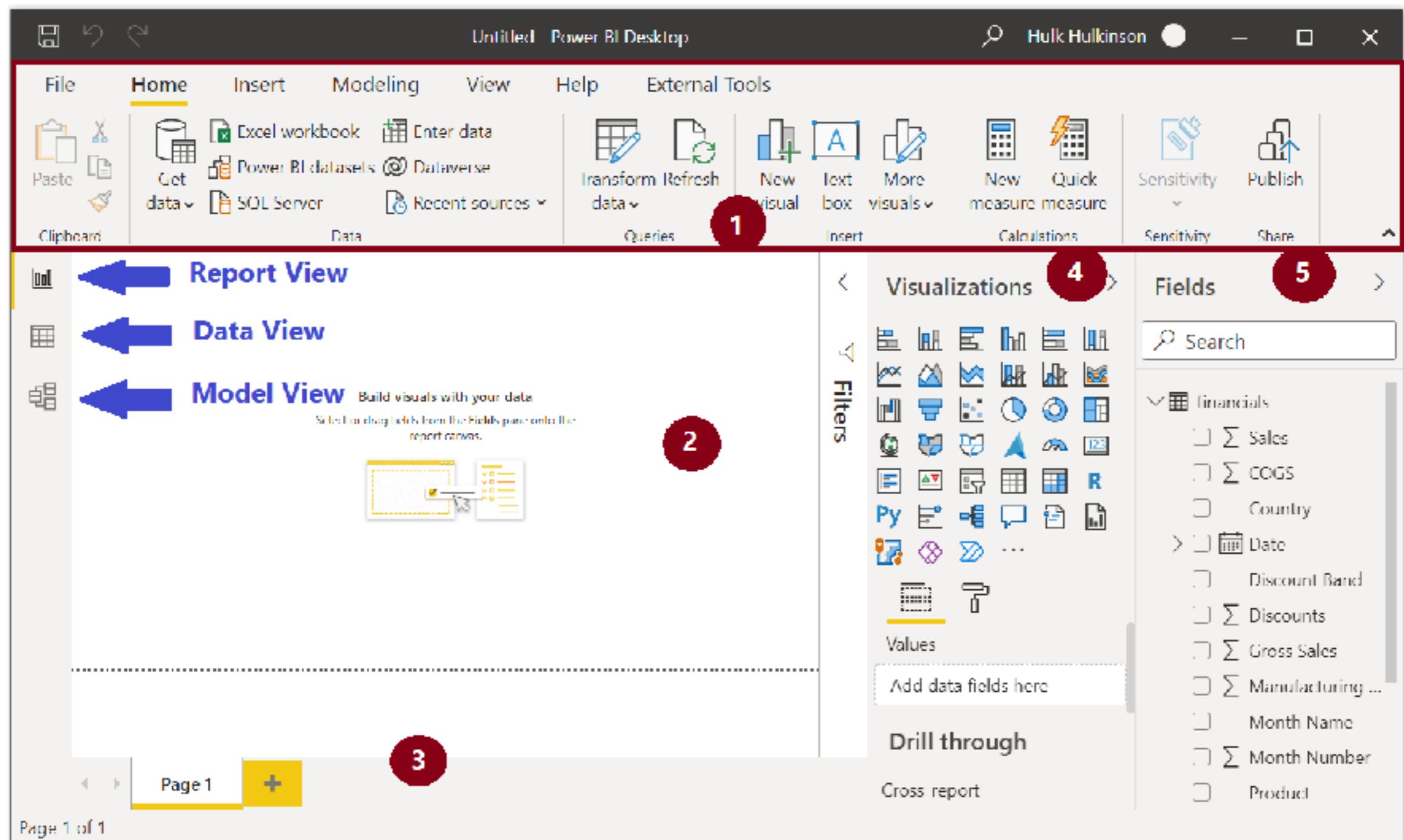
Types of Power BI



<https://app.powerbi.com/>



Power BI Desktop



Sharing

© 2020 - 2024 Siam Chamnankit Company Limited. All rights reserved.

Power BI Desktop

The screenshot shows the Power BI Desktop interface with three main panes: Filters, Visualizations, and Fields.

- Filters:** Contains sections for "Filters on this page" and "Filters on all pages", each with an "Add data fields here" button. It also includes "Drill through" options for "Cross-report" (set to "Off") and "Keep all filters" (set to "On").
- Visualizations:** Displays a grid of visualization icons.
- Fields:** Shows a list of fields under the "financials" category, including:
 - Sales
 - COGS
 - Country
 - Date
 - Discount Band
 - Discounts
 - Gross Sales
 - Manufacturing P...
 - Month Name
 - Month Number
 - Product
 - Profit
 - Sale Price
 - Segment
 - Units Sold
 - Year

A large red arrow points from the Fields pane towards the right side of the screen, indicating the transition to the next slide.



Create Visualization

The screenshot shows the Power BI desktop interface for creating visualizations. On the left, there is a world map visualization titled "Country". The map highlights continents like North America, Europe, and Africa in blue. An orange arrow points from the "Country" label in the visualization pane to the "Country" filter field in the "Filters" pane.

Visualizations

- Filters
- Visualizations
- Fields

Filters

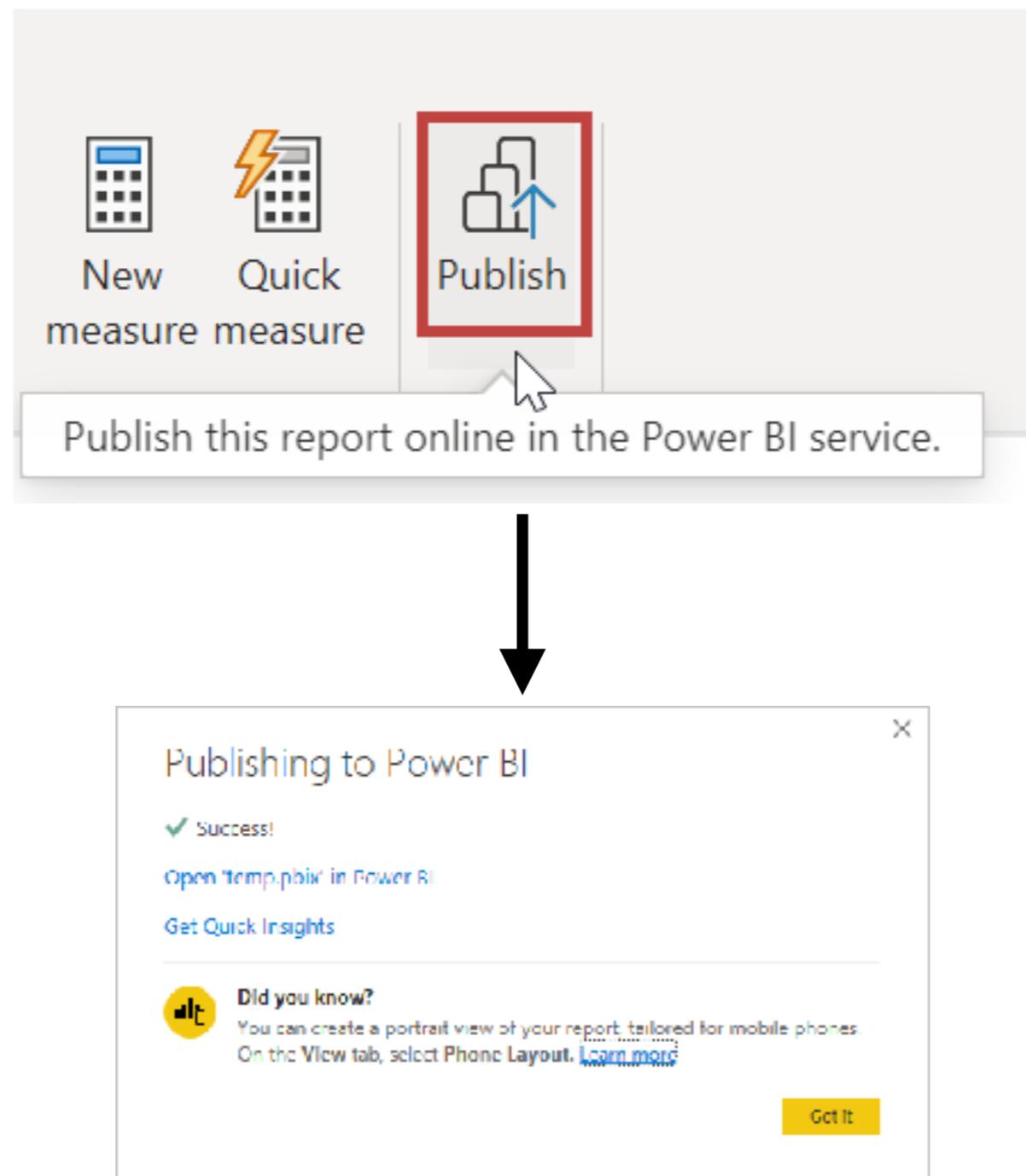
- Search
- Filters on this visual
- Country is (All)
- Add data fields here
- Filters on this page
- Add data fields here
- Filters on all pages
- Add data fields here

Fields

- Search
- financials
 - Σ Sales
 - Σ COGS
 - Country
- Date
- Discount Band
- Σ Discounts
- Σ Gross Sales
- Σ Manufacturing P...
- Month Name
- Σ Month Number
- Product
- Σ Profit
- Σ Sale Price
- Segment
- Σ Units Sold
- Σ Year



Publish a report to web app

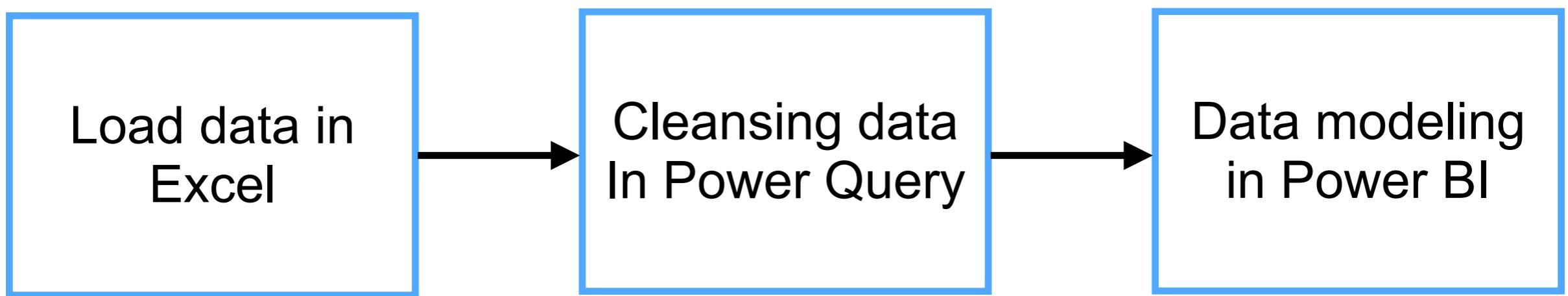


Workshop #4

Working with Power BI



Steps of Workshop



1. Load dataset

01-netflix_titles.csv

File Origin: 65001: Unicode (UTF-8) | Delimiter: Comma | Data Type Detection: Based on first 200 rows

show_id	type	title	director	cast
s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	
s2	TV Show	Blood & Water		Ama Qamata, Khosi Ngema, Gail N...
s3	TV Show	Ganglands	Julien Lederman	Sami Bouajila, Tracy Gotoos, Samu...
s4	TV Show	Jailbirds New Orleans		
s5	TV Show	Kota Factory		Mayur More, Jitendra Kumar, Ranj...
s6	TV Show	Midnight Mass	Mike Hanagan	Kate Siegel, Zach Gilford, Hamish L...
s7	Movie	My Little Pony: A New Generation	Robert Cullen, José Luis Ucha	Vanessa Hudgens, Kimiko Glenn, Is...
s8	Movie	Sankaria	Haile Gerima	Kuli Ghannabi, Oyalumimike Ogund...
s9	TV Show	The Great British Baking Show	Andy Devonshire	Mel Giedroyc, Sue Perkins, Mary B...
s10	Movie	The Starling	Theodore Melfi	Melissa McCarthy, Chris O'Dowd, I...
s11	TV Show	Vendetta: Truth, Lies and The Mafia		
s12	TV Show	Bangkok Breaking	Kongkiat Kornesiri	Sukollawat Kanerat, Sushar Menay...
s13	Movie	Je Suis Karl	Christian Schwochow	Luna Wedler, Jannis Niewöhner, M...
s14	Movie	Confessions of an Invisible Girl	Bruno Garotti	Klara Castanho, Lucca Picon, Júlia...
s15	TV Show	Crime Stories: India Detectives		
s16	TV Show	Dear White People		Logan Browning, Brandon P. Bell, I...
s17	Movie	Europe's Most Dangerous Man: Otto Skorzeny in Spain	Pedro de Echave García, Pablo Azorin Williams	
s18	TV Show	Falsa identidad		Luis Ernesto Franco, Camilo Sodi, S...
s19	Movie	Intrusion	Adam Sally	Ireida Pinto, Logan Marshall-Gree...
s20	TV Show	Jaguar		Ulanca Suárez, Iván Marcos, Óscar...

Extract Table Using Examples | Load | Transform Data | Cancel



2. Data cleansing with Power Query in Excel

Missing data

Change Data
type

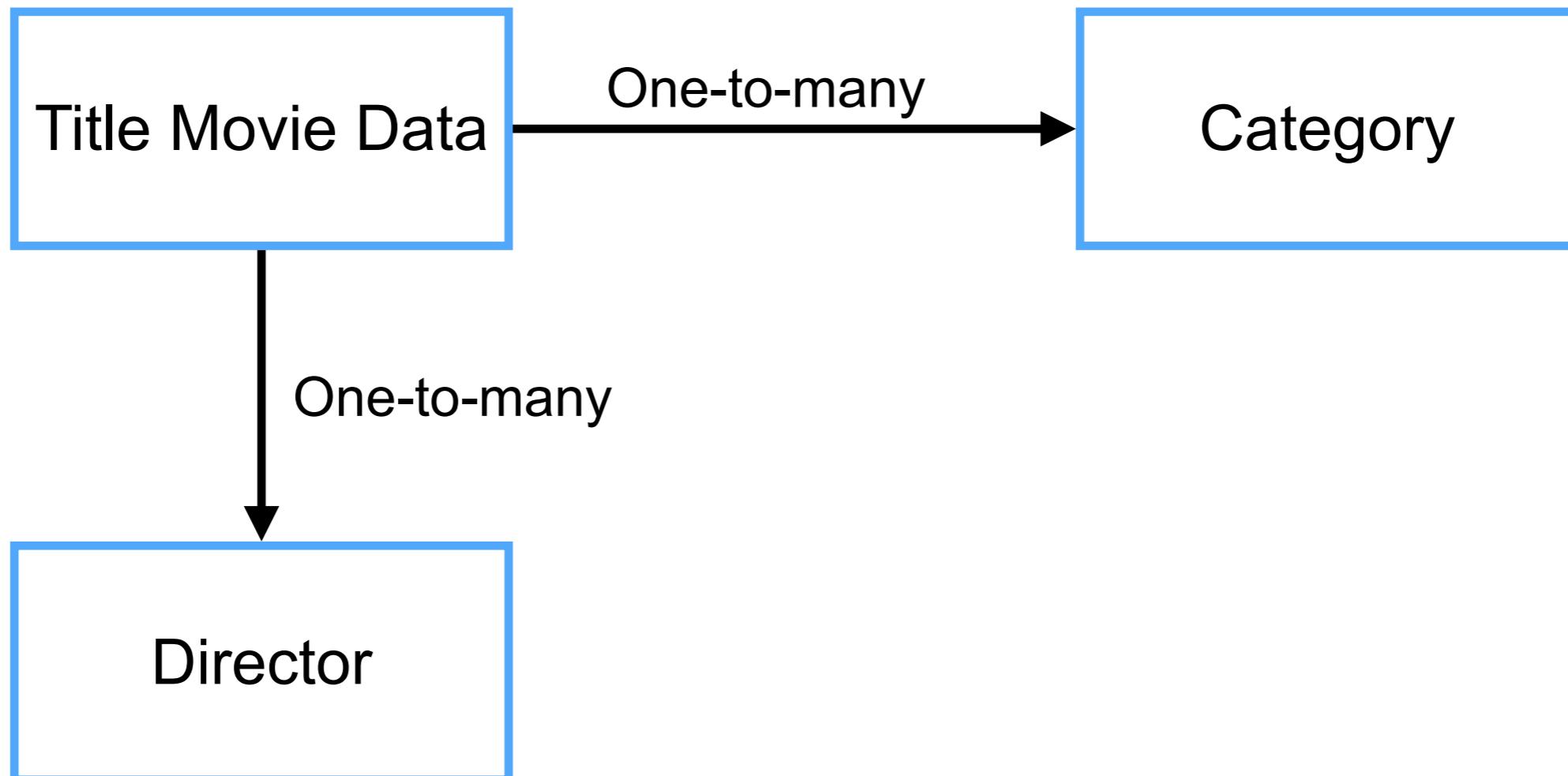
Split data

Trim data

Data model
Relationship



Data Modeling before Analysis



2.1 Split data into new row

The screenshot shows the Microsoft Power BI Data Editor interface. The top navigation bar includes File, Home, Transform, Add Column, and View tabs. The Transform tab is selected, revealing various tools for managing data. A context menu is open over the second column of a table, specifically over the header 'Column2'. The menu options include:

- Copy
- Remove
- Remove Other Columns
- Duplicate Column
- Add Column From Examples...
- Remove Duplicates
- Remove Errors
- Change Type
- Transform
- Replace Values...
- Replace Errors...
- Create Data Type
- Split Column** (selected)
 - By Delimiter...
 - By Number of Characters...
 - By Positions...
 - By Uppercase to Uppercase
 - By Uppercase to Lowercase
 - By Digit to Non-Digit
 - By Non-Digit to Digit
- Group By...
- Hill
- Unpivot Columns
- Unpivot Other Columns
- Unpivot Only Selected Columns
- Rename...
- Move
- Drill Down
- Add as New Query

The table has two columns: 'Column1' and 'Column2'. Column1 contains numerical values from 1 to 21. Column2 contains categorical descriptions. The context menu is positioned over the first few rows of Column2.

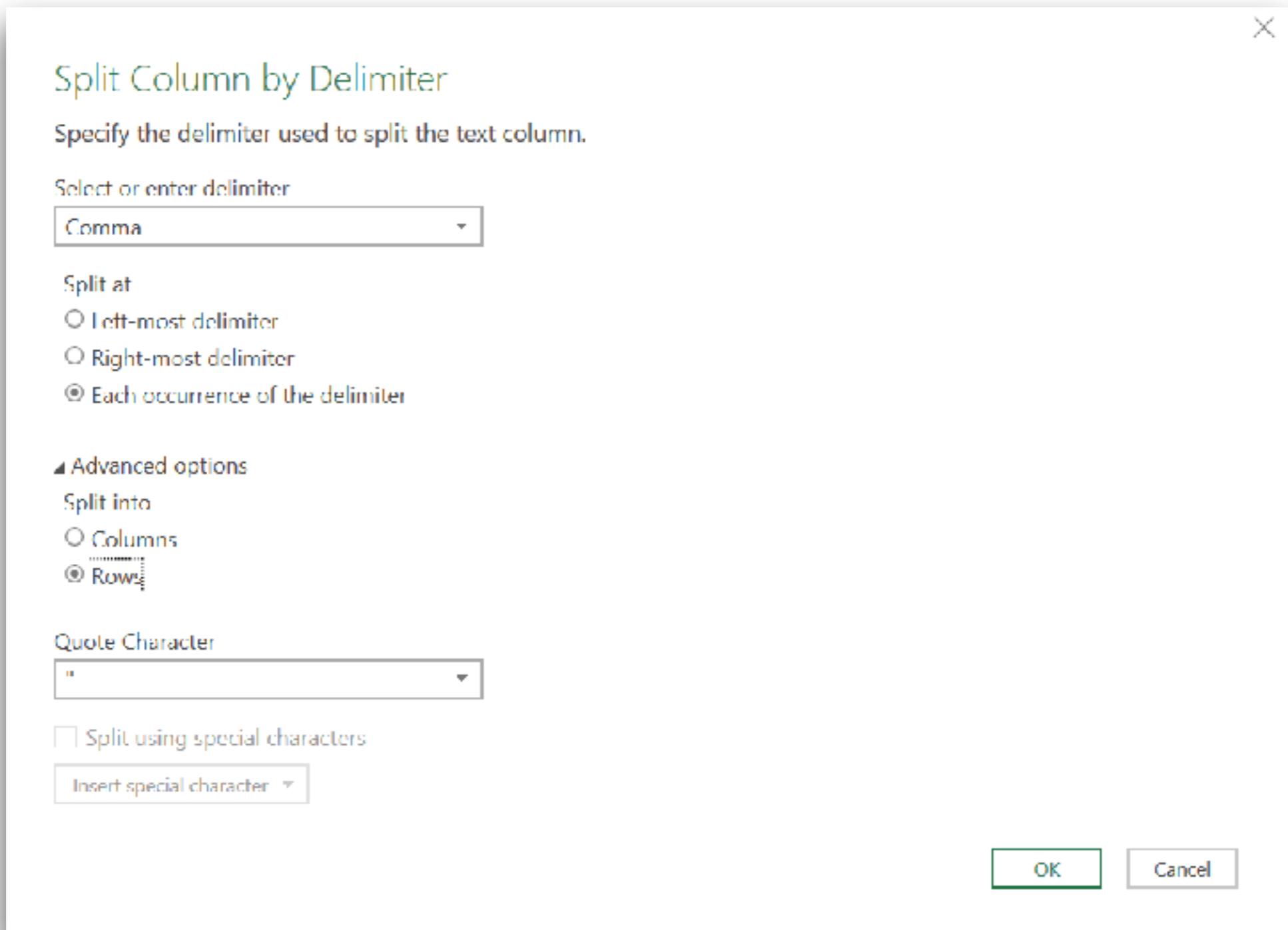
Column1	Column2
1	listed_in
2	Documentaries
3	International TV Shows, IV D
4	Crime TV Shows, Internationa
5	Docuseries, Reality TV
6	International TV Shows, Rom
7	IV Dramas, IV Horror, IV My
8	Children & Family Movies
9	Dramas, Independent Movies
10	British TV Shows, Reality TV
11	Comedies, Dramas
12	Crime TV Shows, Docuseries,
13	Crime TV Shows, International
14	Dramas, International Movie
15	Children & Family Movies, Co
16	British TV Shows, Crime TV Sh
17	TV Comedies, TV Dramas
18	Documentaries, International
19	Crime TV Shows, Spanish Lan
20	Thrillers
21	International TV Shows, Span



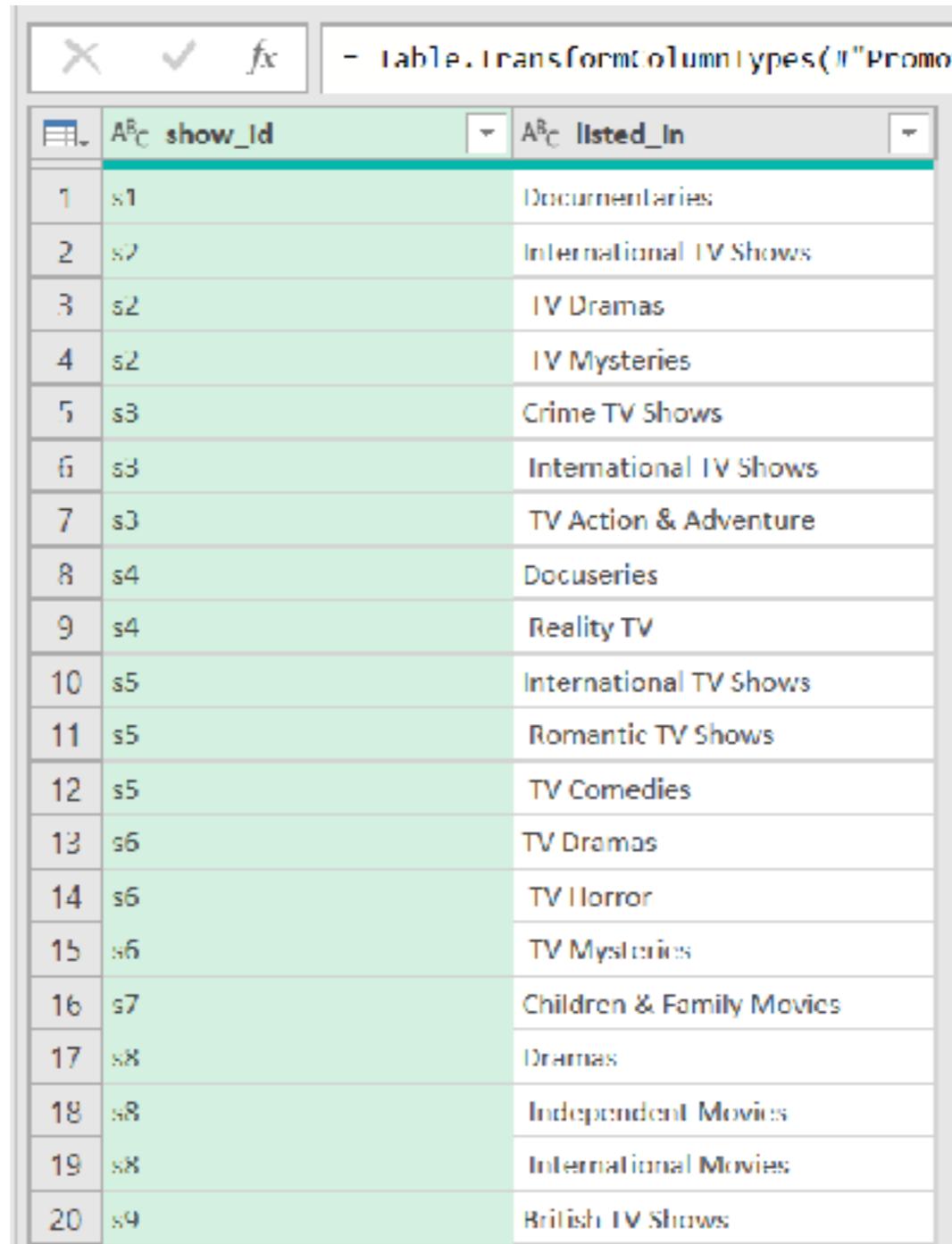
Sharing

© 2020 - 2024 Siam Chamnankit Company Limited. All rights reserved.

2.2 Use delimiter by comma



2.3 New sheet of category



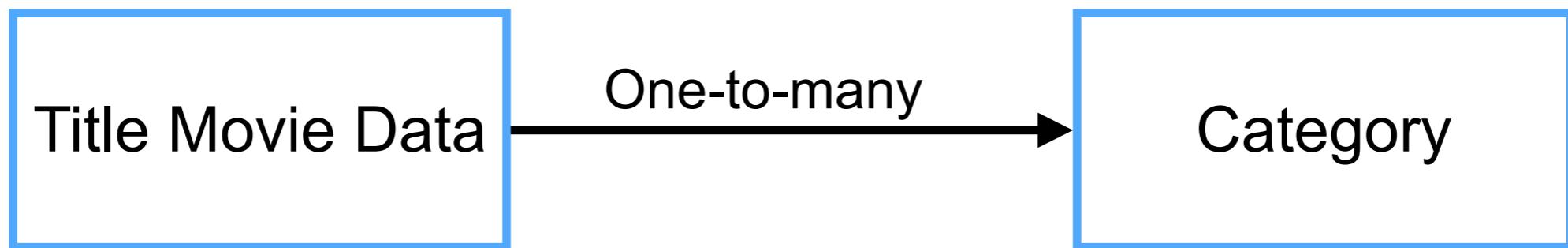
The screenshot shows a Microsoft Excel table with two columns: 'ABC_show_Id' and 'ABC_listed_In'. The table contains 20 rows of data, each consisting of a show ID and its corresponding category. The categories listed are: Documentaries, International TV Shows, TV Dramas, TV Mysteries, Crime TV Shows, International TV Shows, TV Action & Adventure, Docuseries, Reality TV, International TV Shows, Romantic TV Shows, TV Comedies, TV Dramas, TV Horror, TV Mysteries, Children & Family Movies, Dramas, Independent Movies, International Movies, and British TV Shows.

ABC_show_Id	ABC_listed_In
1	s1 Documentaries
2	s2 International TV Shows
3	s2 TV Dramas
4	s2 TV Mysteries
5	s3 Crime TV Shows
6	s3 International TV Shows
7	s3 TV Action & Adventure
8	s4 Docuseries
9	s4 Reality TV
10	s5 International TV Shows
11	s5 Romantic TV Shows
12	s5 TV Comedies
13	s6 TV Dramas
14	s6 TV Horror
15	s6 TV Mysteries
16	s7 Children & Family Movies
17	s8 Dramas
18	s8 Independent Movies
19	s8 International Movies
20	s9 British TV Shows



3. Working with Power BI (1)

Data modeling and relationships



3. Create relationships in Power BI (2)

← New relationship →

list_in	rating	release_year	show_id	Title	Type
Reality TV	TV-PG	2021	s201	Bake Squad	TV Show
Reality TV	TV-MA	2019	s621	Droppin' Cas...	TV Show
Reality TV	TV-PG	2021	s749	Fresh, Fried &...	TV Show

To table

category

list_in	show_id
International ...	s8
International ...	s13
International ...	s17

Cardinality

One to many (1:n)

Cross-filter direction

Single

Make this relationship active

Apply security filter in both directions

Assume referential integrity



4. Create Dashboard in Power BI

Chart

Table

Map



4. Create Dashboard in Power BI

The screenshot shows the Microsoft Power BI desktop application. The ribbon menu at the top includes File, Home (selected), Insert, Modeling, View, Optimize, and Help. The Home tab has sections for Paste, Cut, Copy, Format painter, Get data, Excel, OneLake, SQL Server, Enter data, Data, Transform data, Refresh data, New visual, Text box, More visuals, New visual calculation, New measure, Quick measure, Sensitivity, Publish, and Copilot. The main area displays a bar chart titled "Count of show_id by list_in" with the Y-axis labeled "Count of show_id" ranging from 0 to 3,000 and the X-axis labeled "list_in". The chart shows various categories like "Independent Movies", "Documentaries", "TV Documentaries", "TV Shows", etc., with "Independent Movies" having the highest count. To the right of the chart are three panels: "Filters", "Visualizations", and "Data". The "Filters" panel contains sections for "Filters on this page" and "Filters on all pages", each with a "Add data fields here" button. The "Visualizations" panel shows icons for different visualization types. The "Data" panel shows a search bar and a tree view with nodes like "U1-netflix_titles" and "category". At the bottom, there are navigation buttons for "Page 1" and a "+" sign, along with icons for monitor, smartphone, and other sharing options.

