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

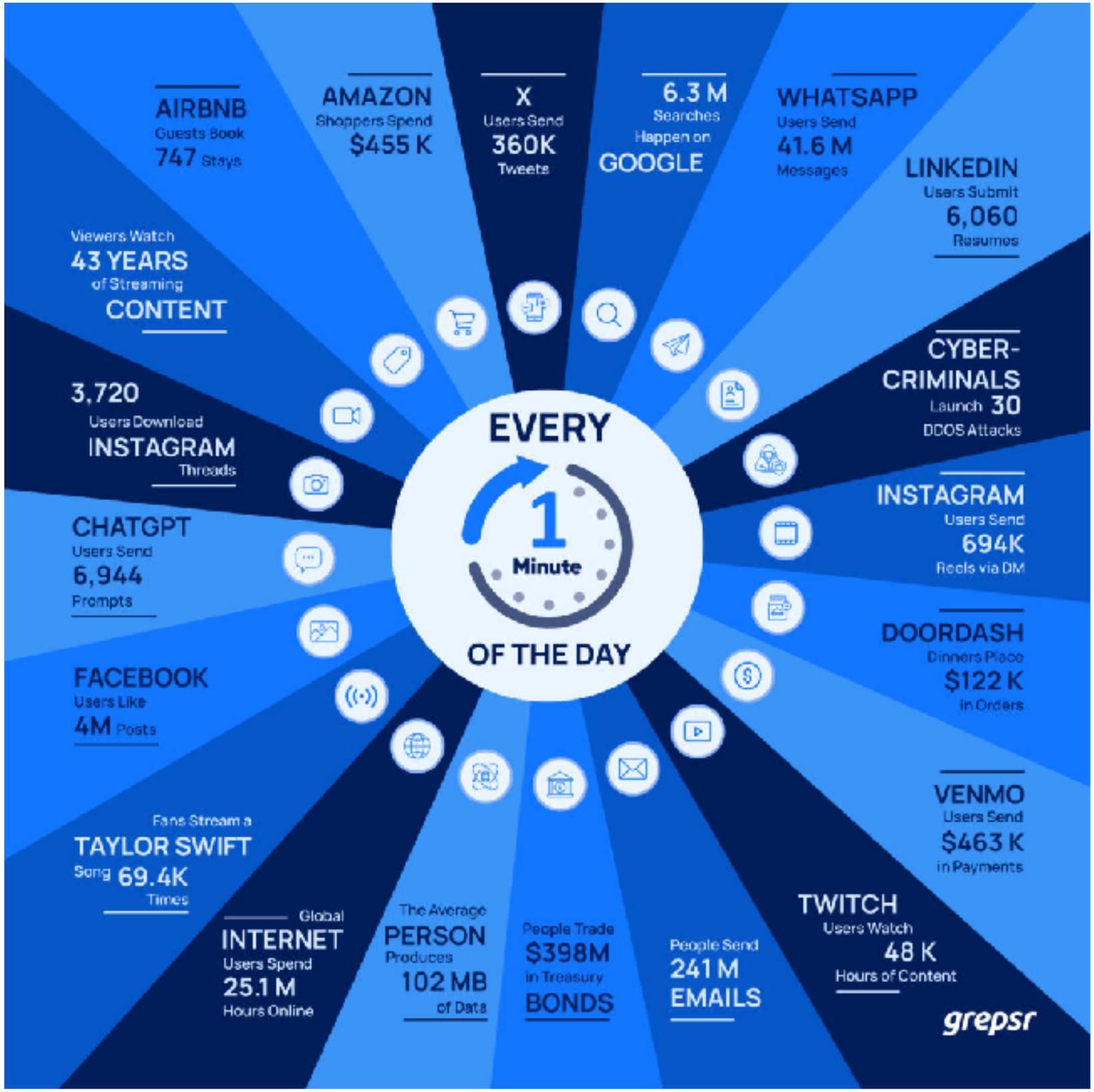
Introduction

Big Data

Business Analytics (BA)

Business Intelligence (BI)

Introduction to Big Data



Timeline

1960

Statistics and Computers

Statistics have been used for various analytics purposes as far back as Ancient Egypt. For decades, governments around the world have been using statistics for planning a wide variety of activities, including censuses and taxation.

1970

Relational Databases and Non-Relational Databases

Invented by Edgar Codd in the 1970s. As relational databases (RDBMs) improved over time, it allowed users to write in Sequel (SQL) and retrieve data from their database

1980

Data Warehouses

Incepted in the late 1980s. Data warehouses are optimized for quick response time to queries, and they are usually part of an organization's mainframe network

1990

Business Intelligence

Came into mainstream in 1990. Business intelligence means decision-making based on new and innovative data technologies

1995

Data Mining

First came into being in the 1990s. Data Mining allowed organizations to store more data while analyzing it quickly and efficiently.

2000

Cloud Analytics

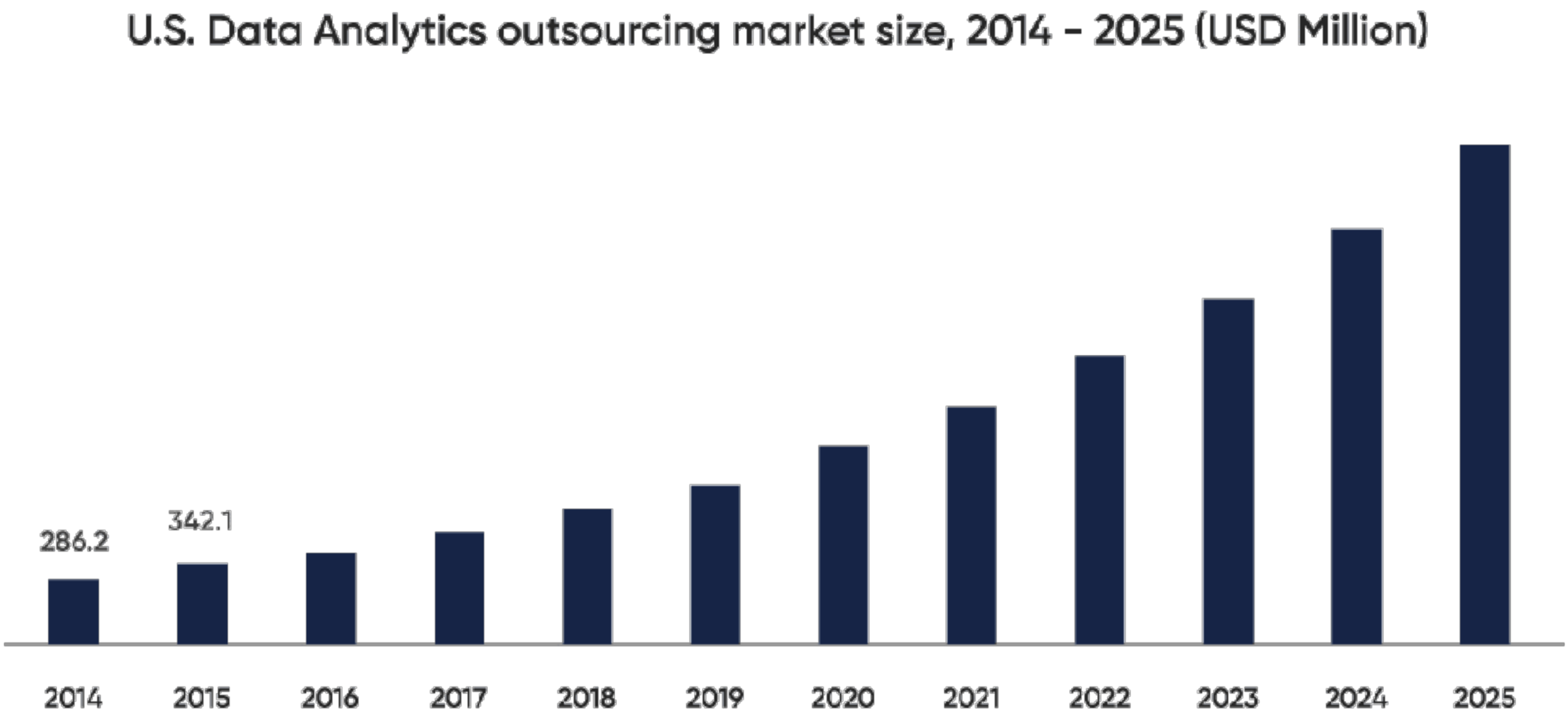
The concept was first introduced in 1997. Since 1999, cloud computing has grown significantly and can now store enormously large amounts of data, can handle multiple projects.

2005

Big Data

Term coined in 2005. In the same year, an open-source software that could process big data, called Hadoop, was developed.

Timeline



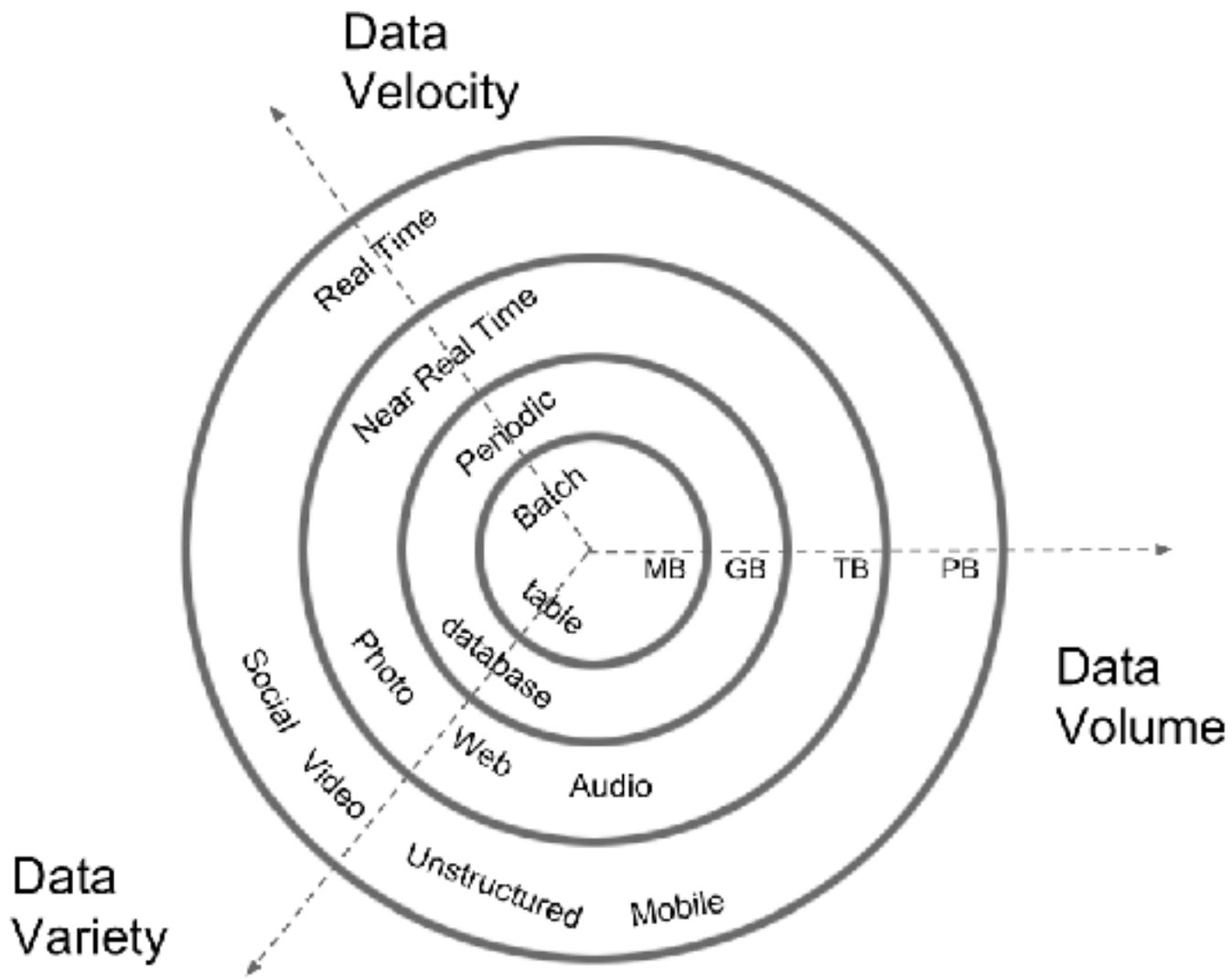
Big Data

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

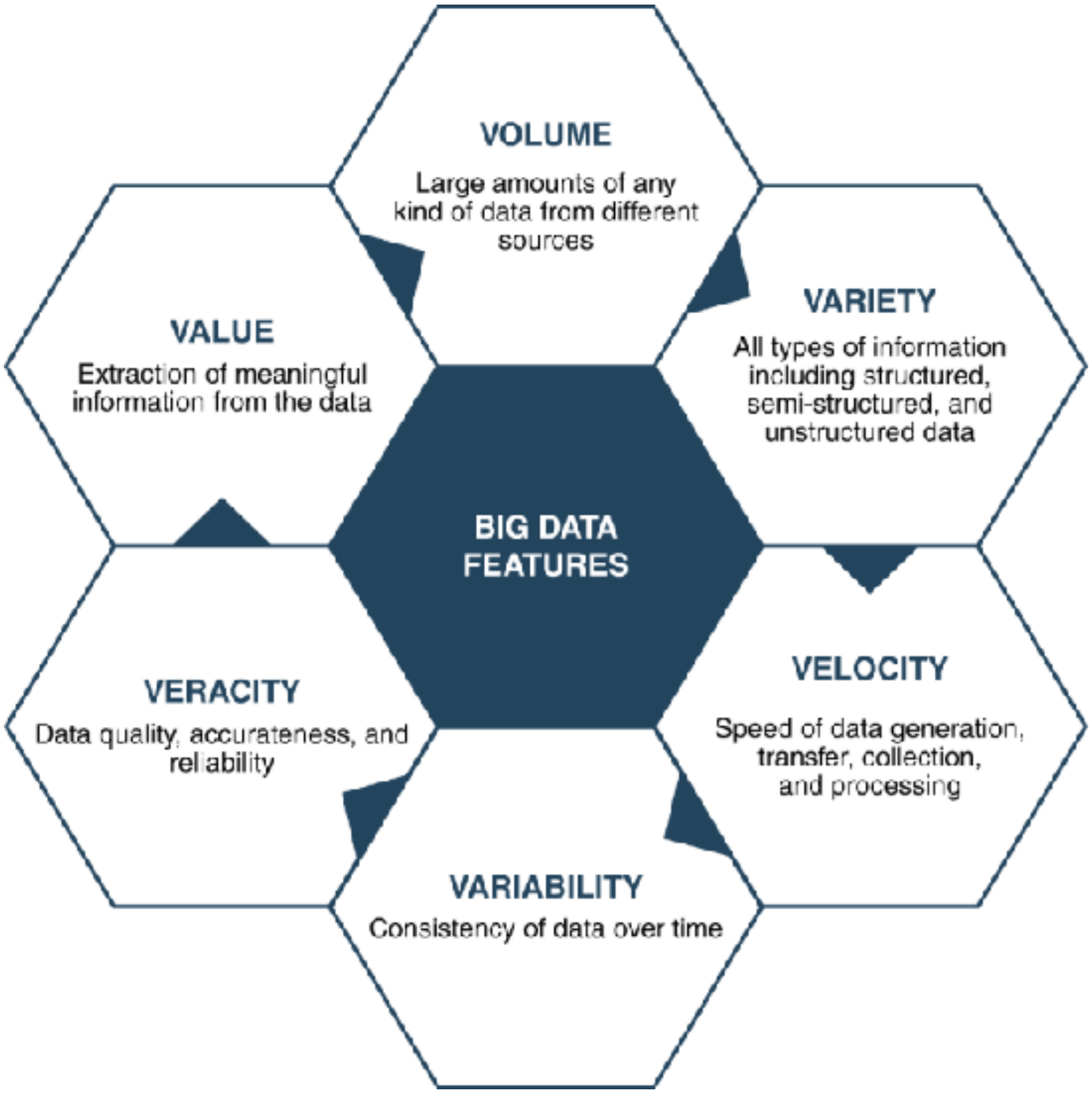
Data is too complex to processed by traditional databases and tools

Characteristics for Big Data

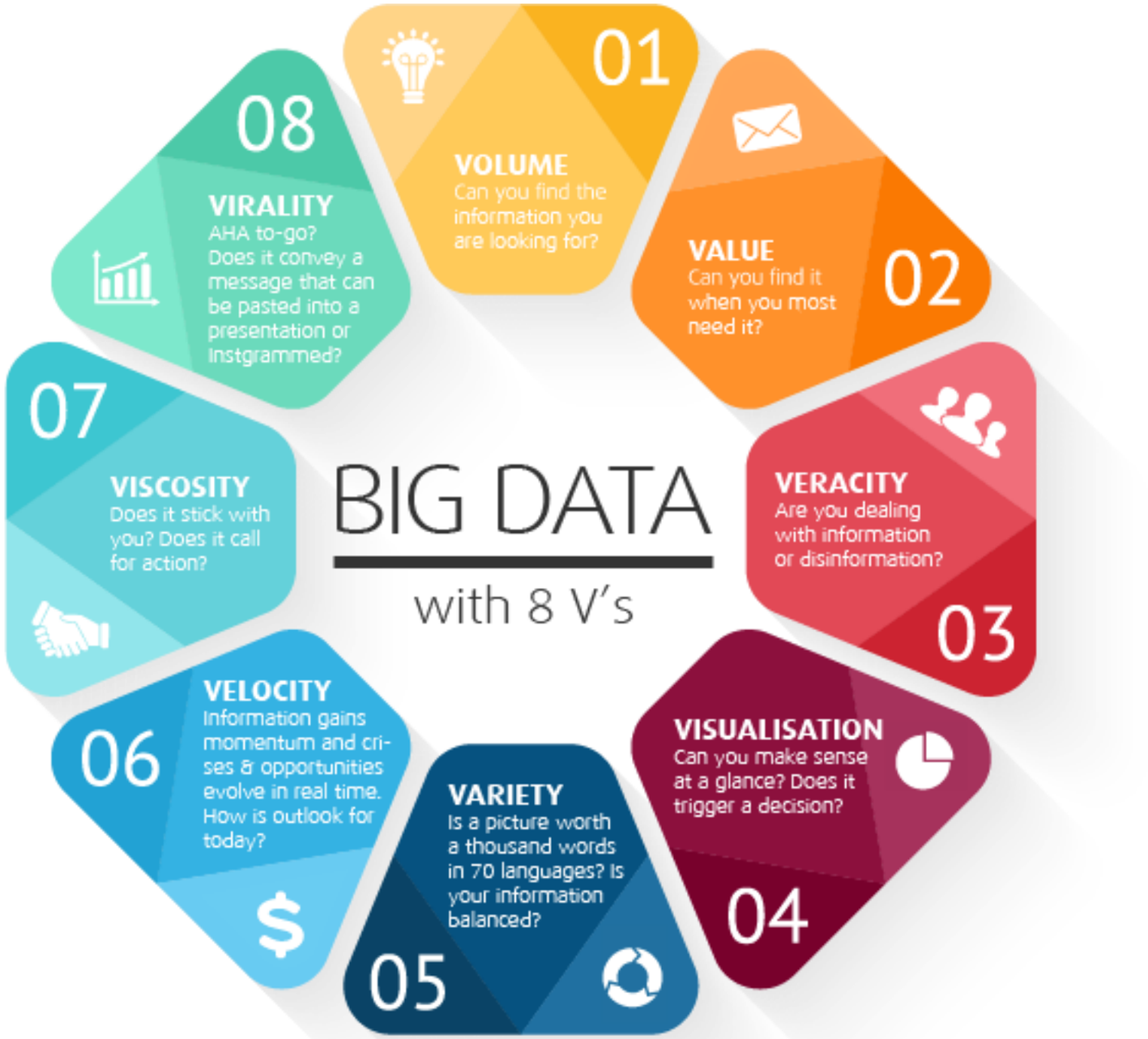
3 V's !!!



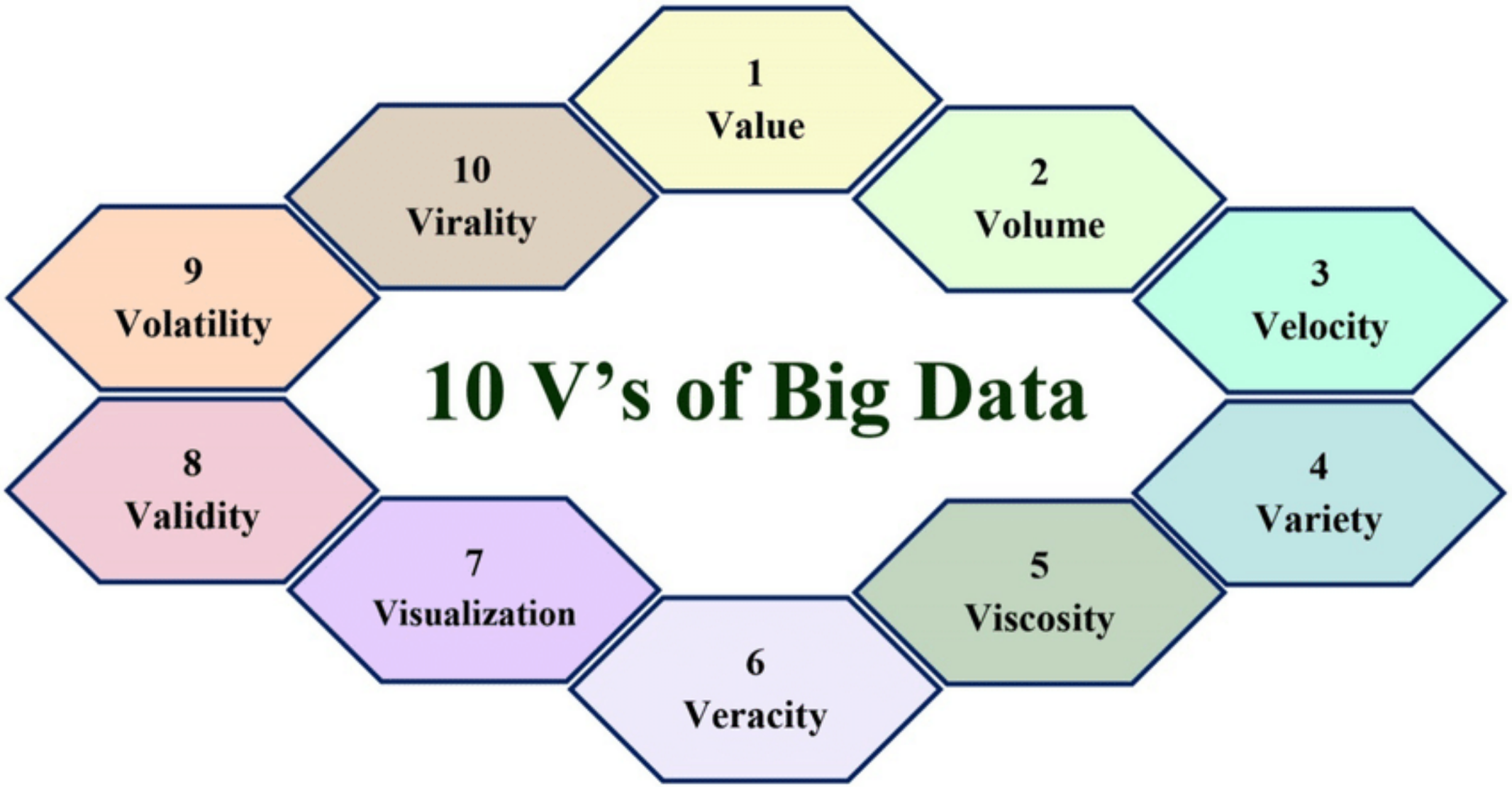
Characteristics for Big Data



Characteristics for Big Data



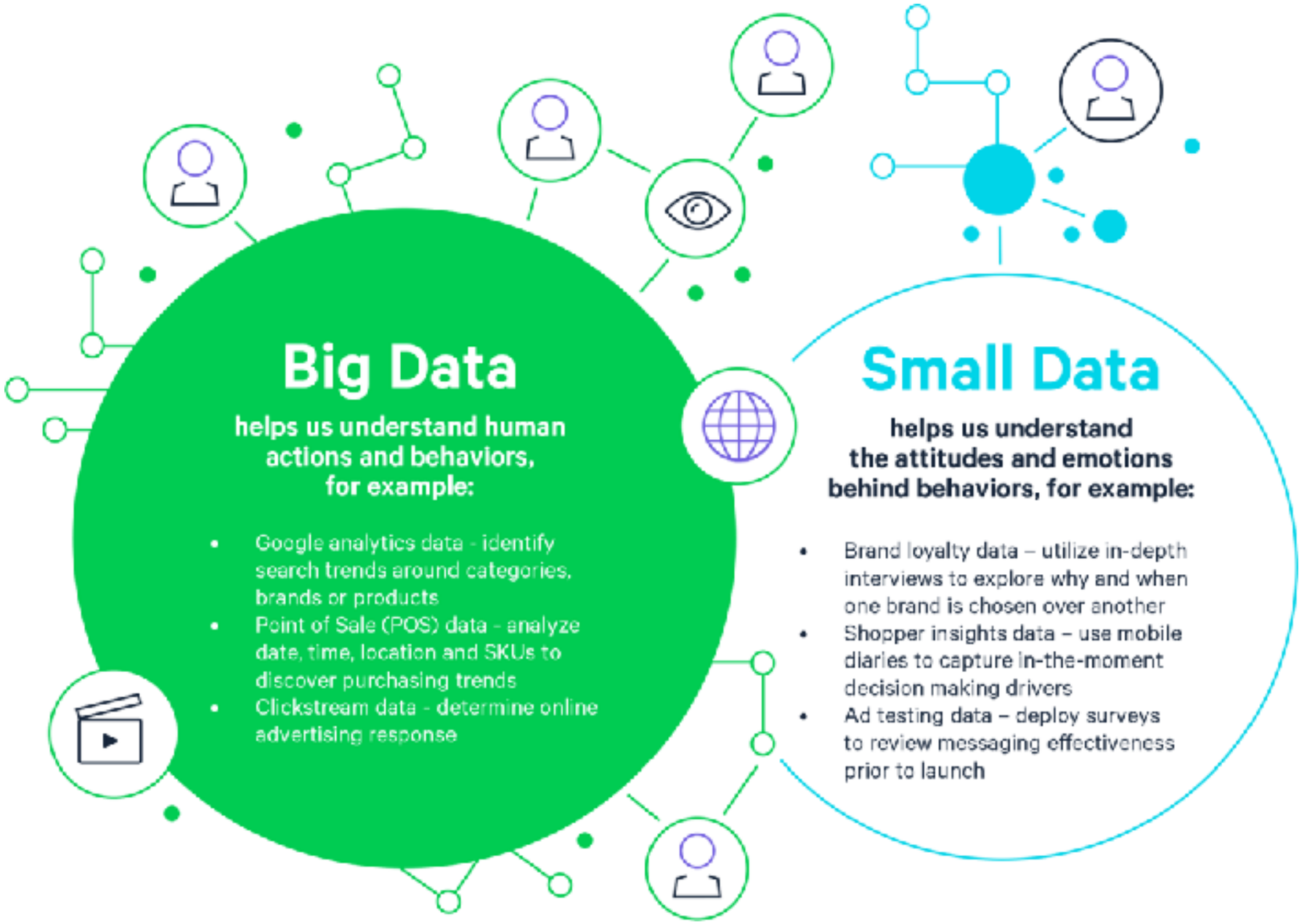
Characteristics for Big Data



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)	แสดงข้อมูลดิบ โดยไม่มีการสังเคราะห์ อาจจะเป็นการนำอมูลมาแสดงผล โดยการพลิกมุม รายงานขายตามประเภทสินค้า หรือตามภูมิภาค	แสดงข้อมูลที่เกิดจากการสังเคราะห์ เช่น เกิดจากการทำ Data Mining เพื่อหาความเกี่ยวข้อง หาข้อมูลที่ซ่อนอยู่

www.wipro.com

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

Amount of Big Data stored across the world (in petabytes)

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

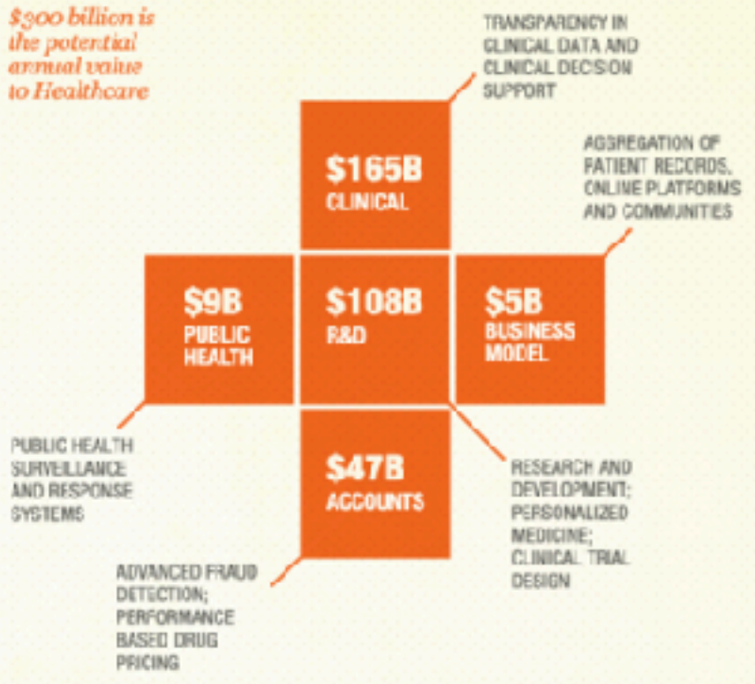
2.9 MILLION EMAILS SENT EVERY SECOND

20 HOURS OF VIDEO UPLOADED EVERY MIN

50 MILLION TWEETS PER DAY

CASE STUDY - Healthcare

$300 billion is the potential annual value to Healthcare

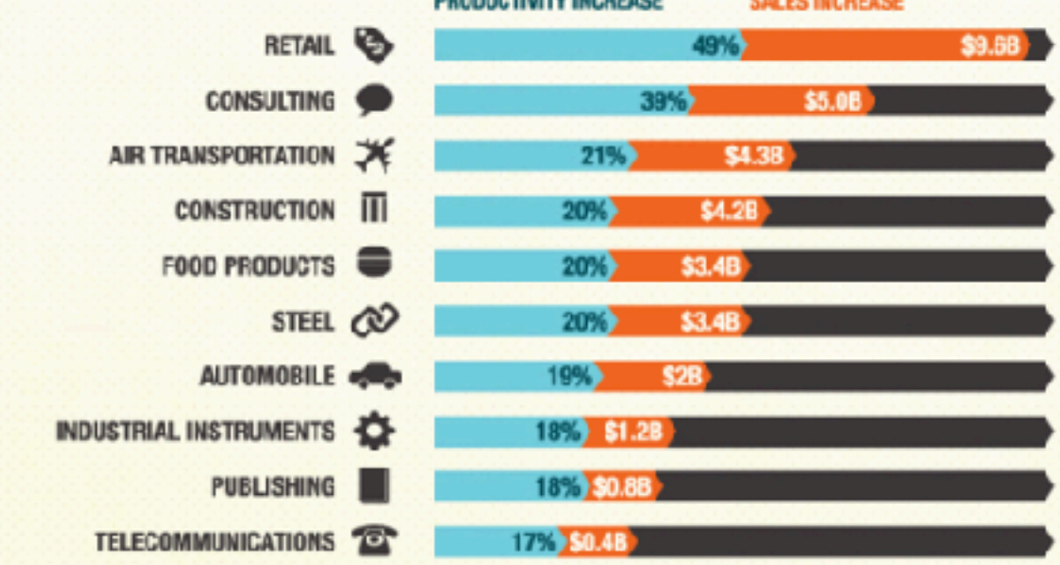


57.6% OF ORGANIZATIONS SURVEYED SAY THAT BIG DATA IS A CHALLENGE

72.7% CONSIDER DRIVING OPERATIONAL EFFICIENCIES TO BE THE BIGGEST BENEFIT OF A BIG DATA STRATEGY

50% SAY THAT BIG DATA HELPS IN BETTER MEETING CONSUMER DEMAND AND FACILITATING GROWTH

VALUE



40% PROJECTED GROWTH IN GLOBAL DATA CREATED PER YEAR

5% PROJECTED GROWTH IN GLOBAL IT SPENDING 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?

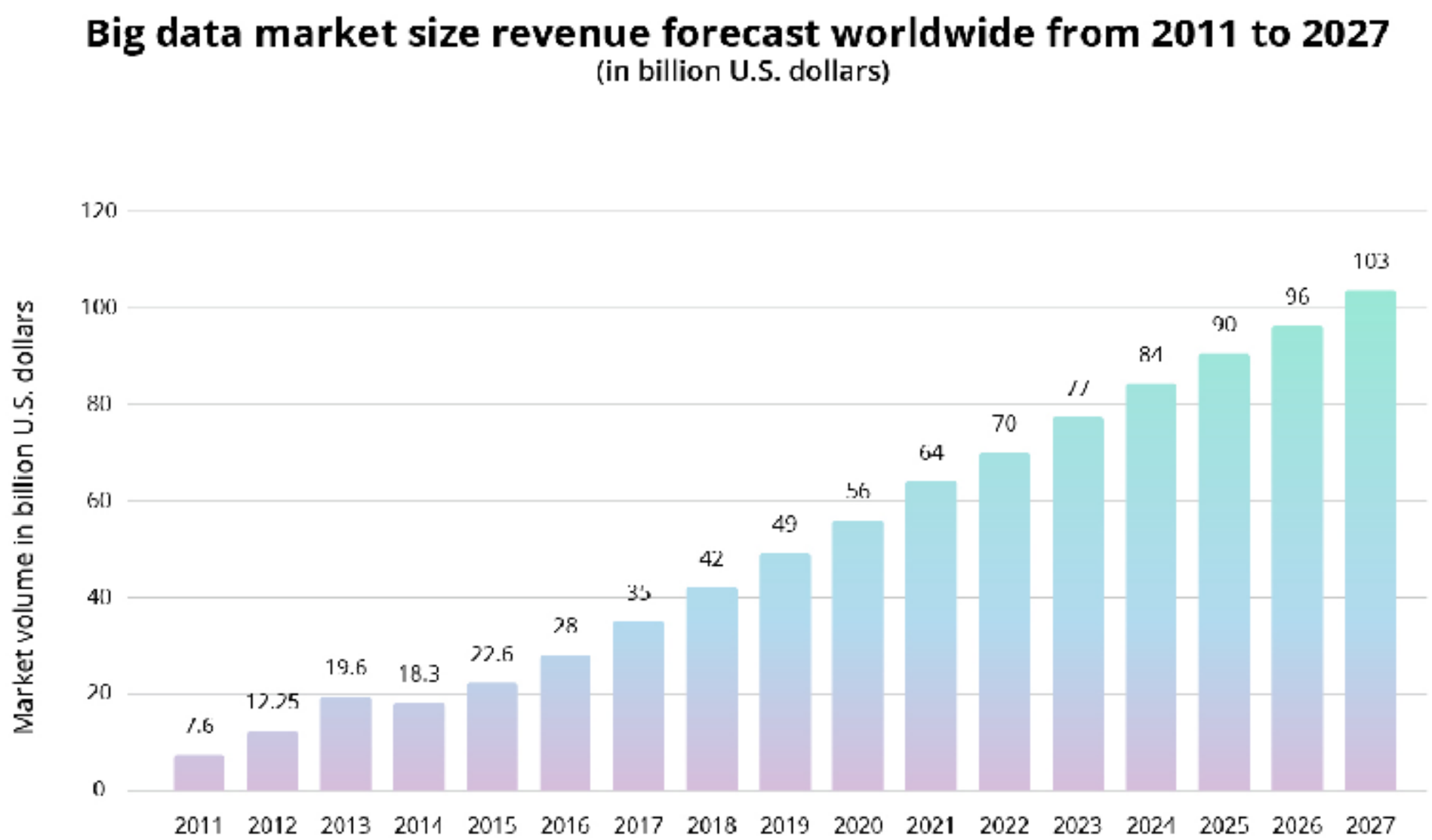
Sources - 1 Reaping the Rewards of Big Data - Wipro Report 2 Big Data: The Next Frontier for innovation, Competition and Productivity - McKinsey Global Institute Report 3 comScore, Radicati Group 4 Measuring the business impacts of effective Data - study by University of Texas, Austin 5 US Department of Labour.

DO BUSINESS BETTER

FY15:$8.1BT | OVER 130,000 EMPLOYEES | 54 COUNTRIES | CONSULTING | SYSTEM INTEGRATION | OUTSOURCING

Visualization: Sreenidhi

Big Data Trend



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

Big Data Challenges

Storage

Data Quality

Analysis

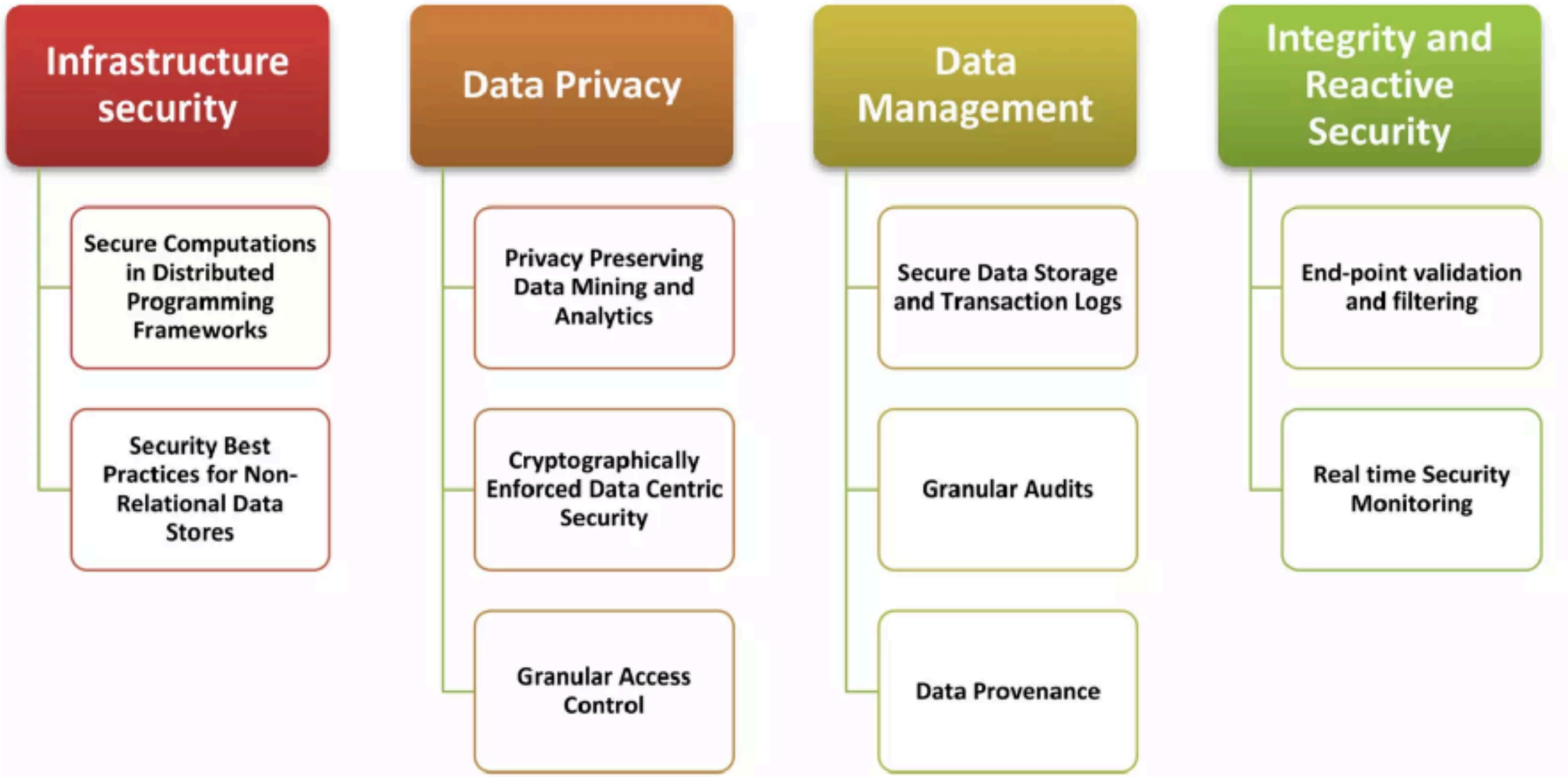
Cost vs Value

Accessibility

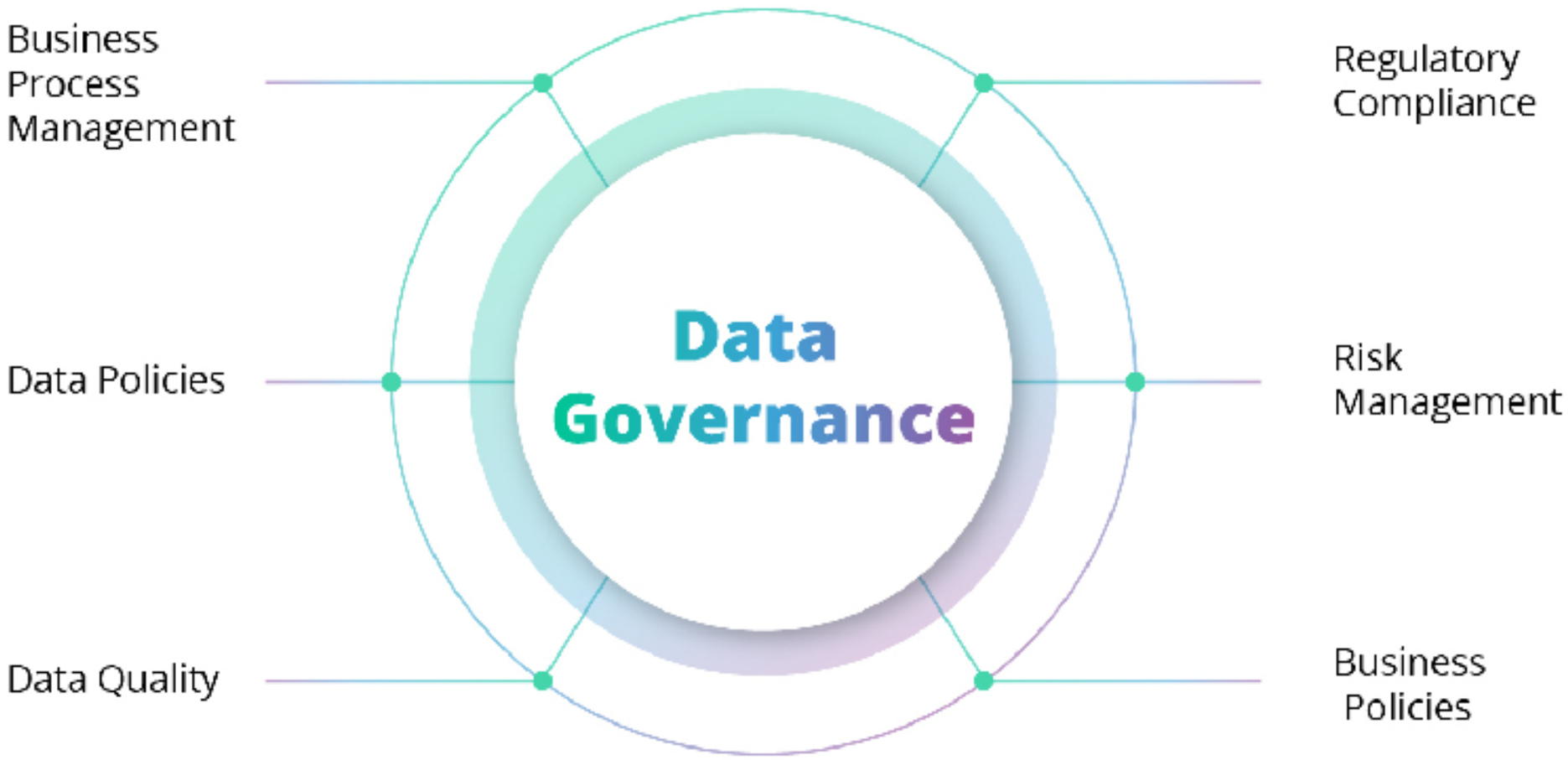
Security

Big Data Challenges

Big Data Security Challenges

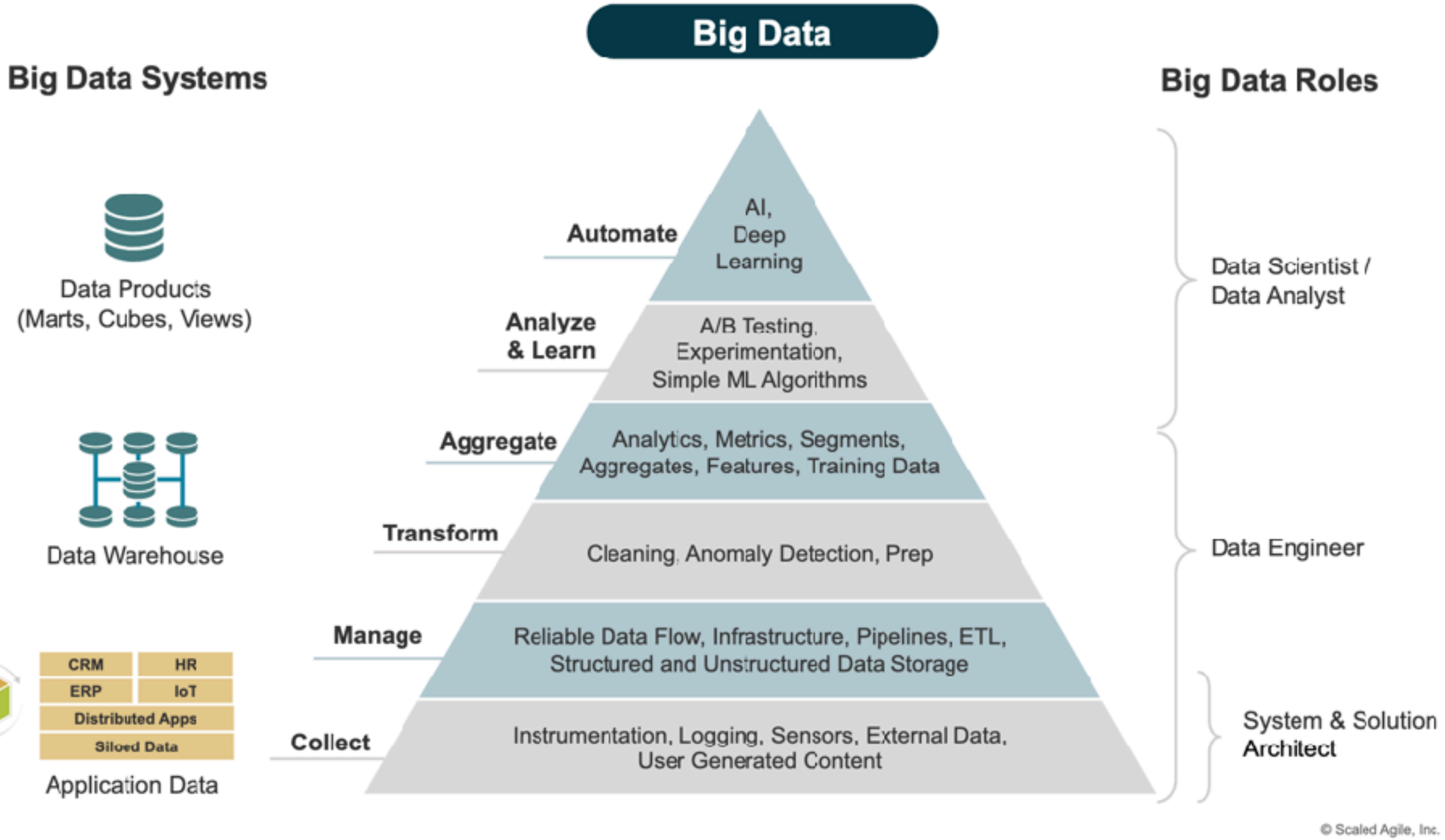


Data Governance



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

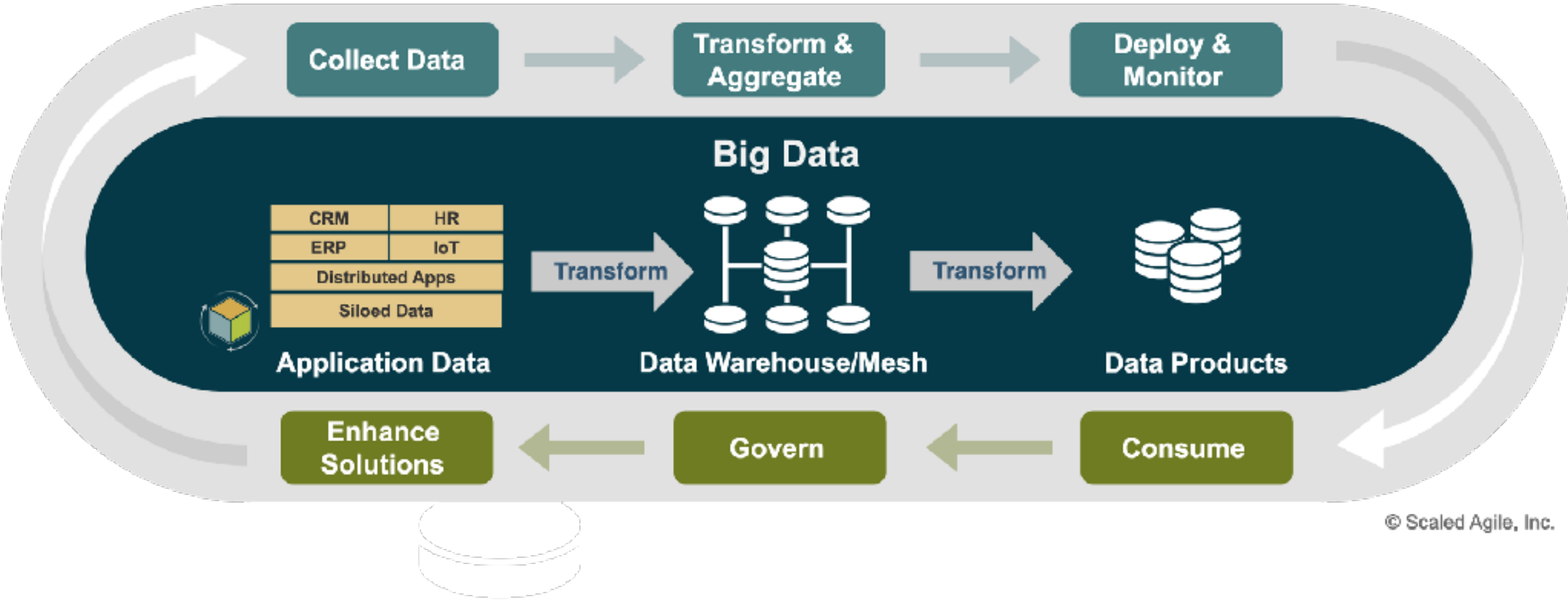
All about Big Data



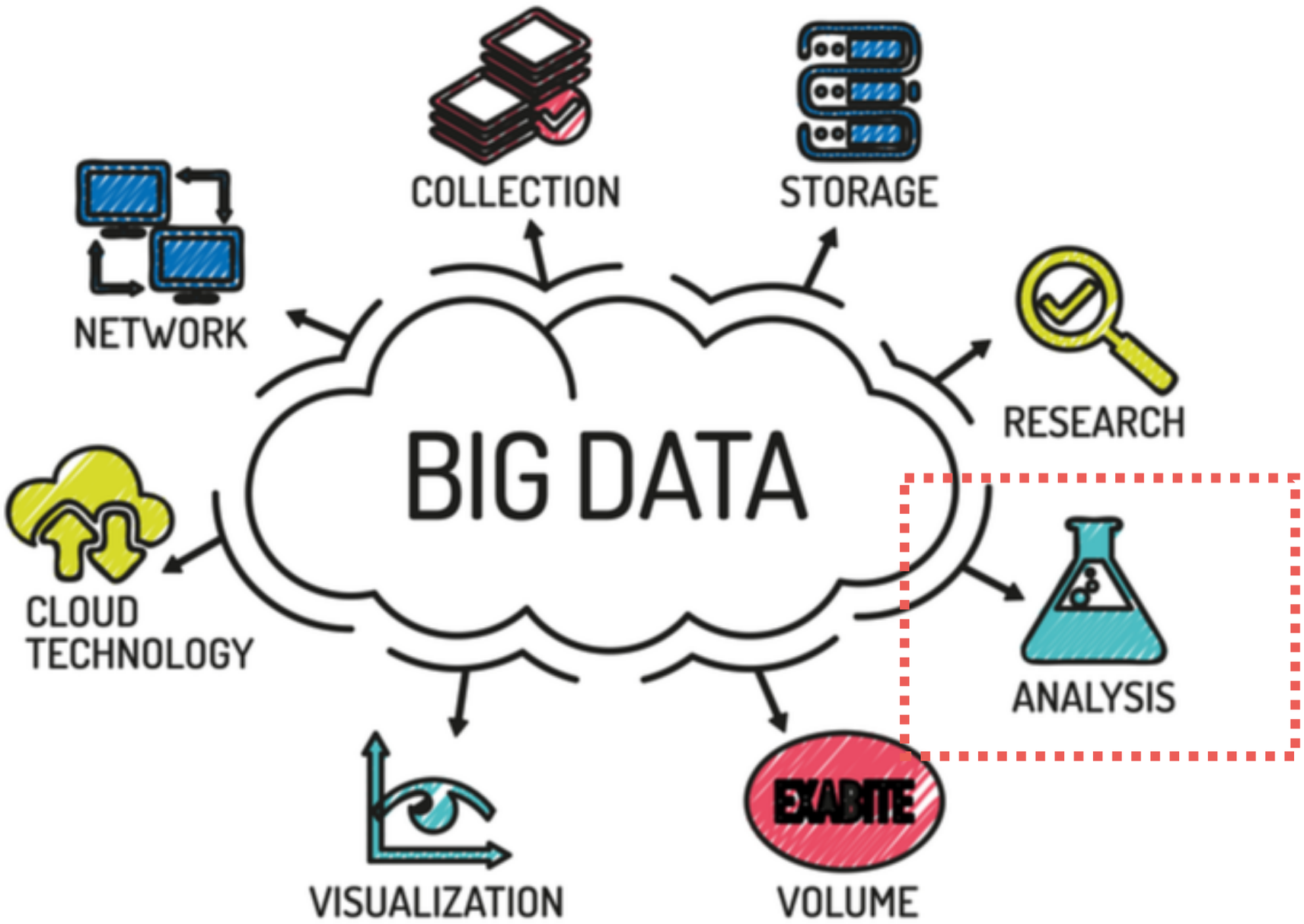
https://scaledagileframework.com/big-data/

DataOps

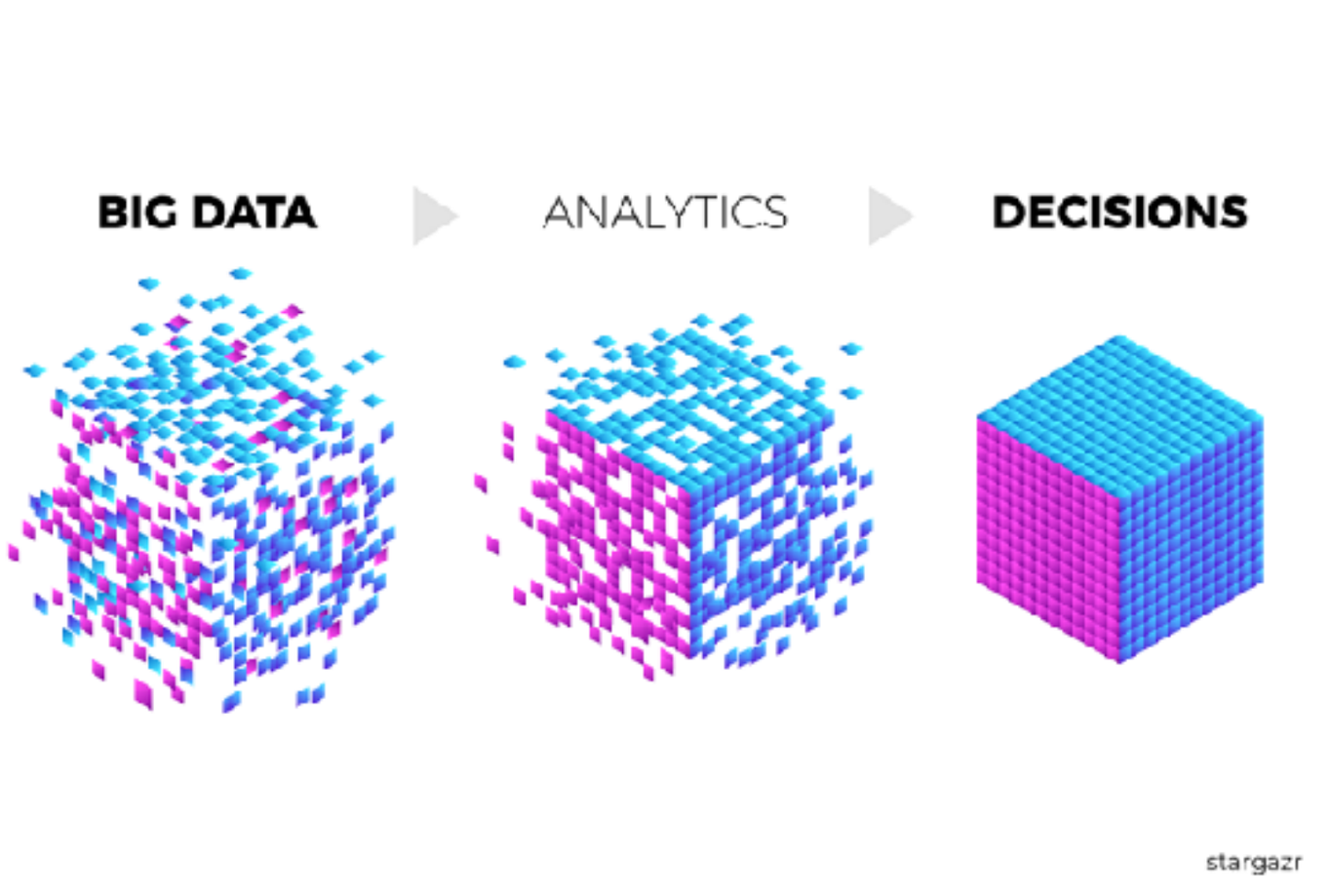
Collaboration data management across teams



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



Rise of Big Data Analytics



Introduction to Business Analytics (BA)

Business Analytics (BA)

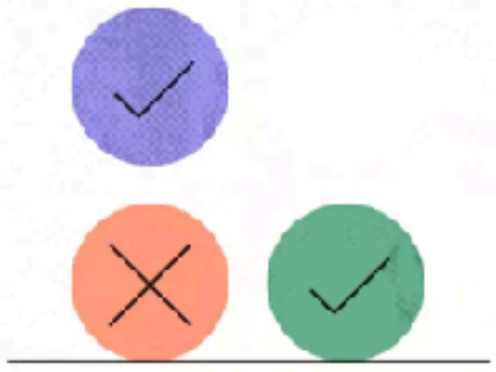
Practice of using data, statistical analysis, and quantitative methods to drive **decision-making** and **improve business outcomes**

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

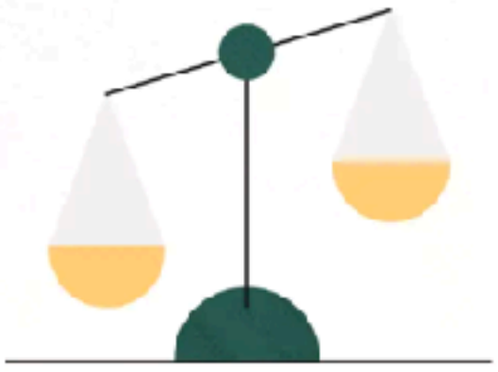
Purpose of BA

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

Data-driven Decision



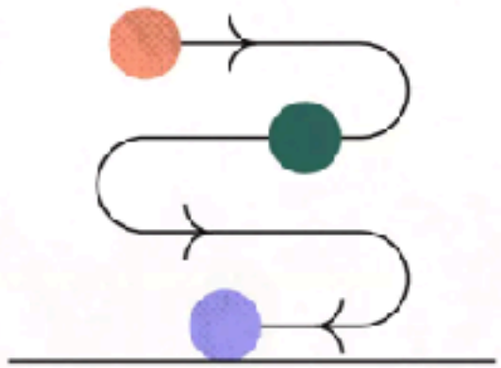
Make confident decisions



Guard against biases

Find unresolved questions

Set measurable goals



Improve company processes

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

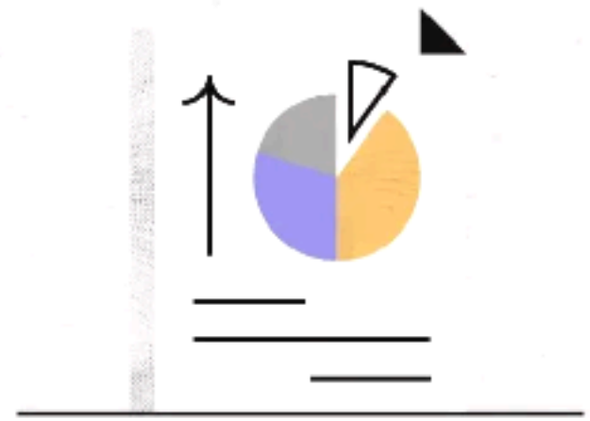
Tips for Data-driven Decision

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

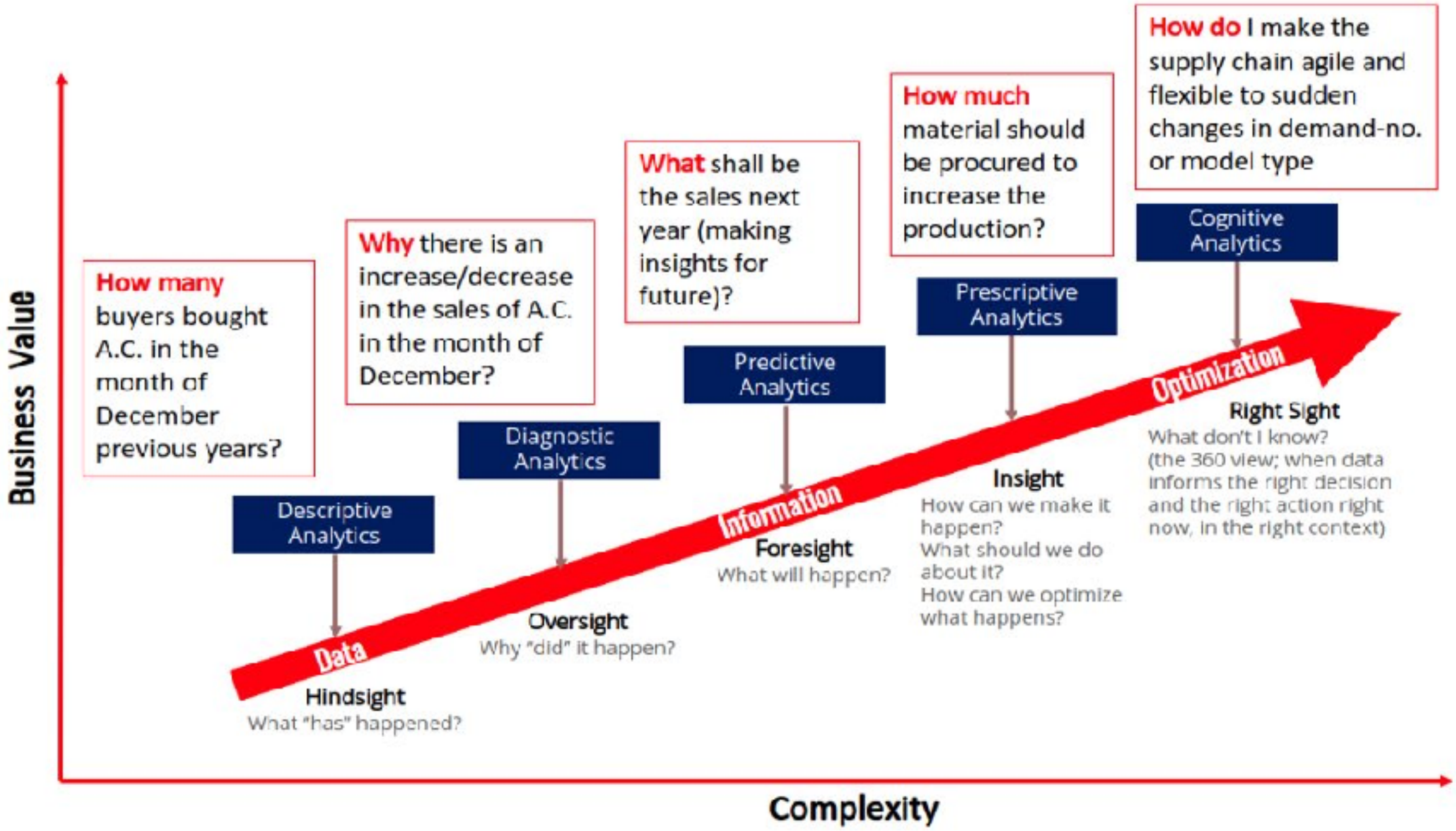
Key Techniques

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?”

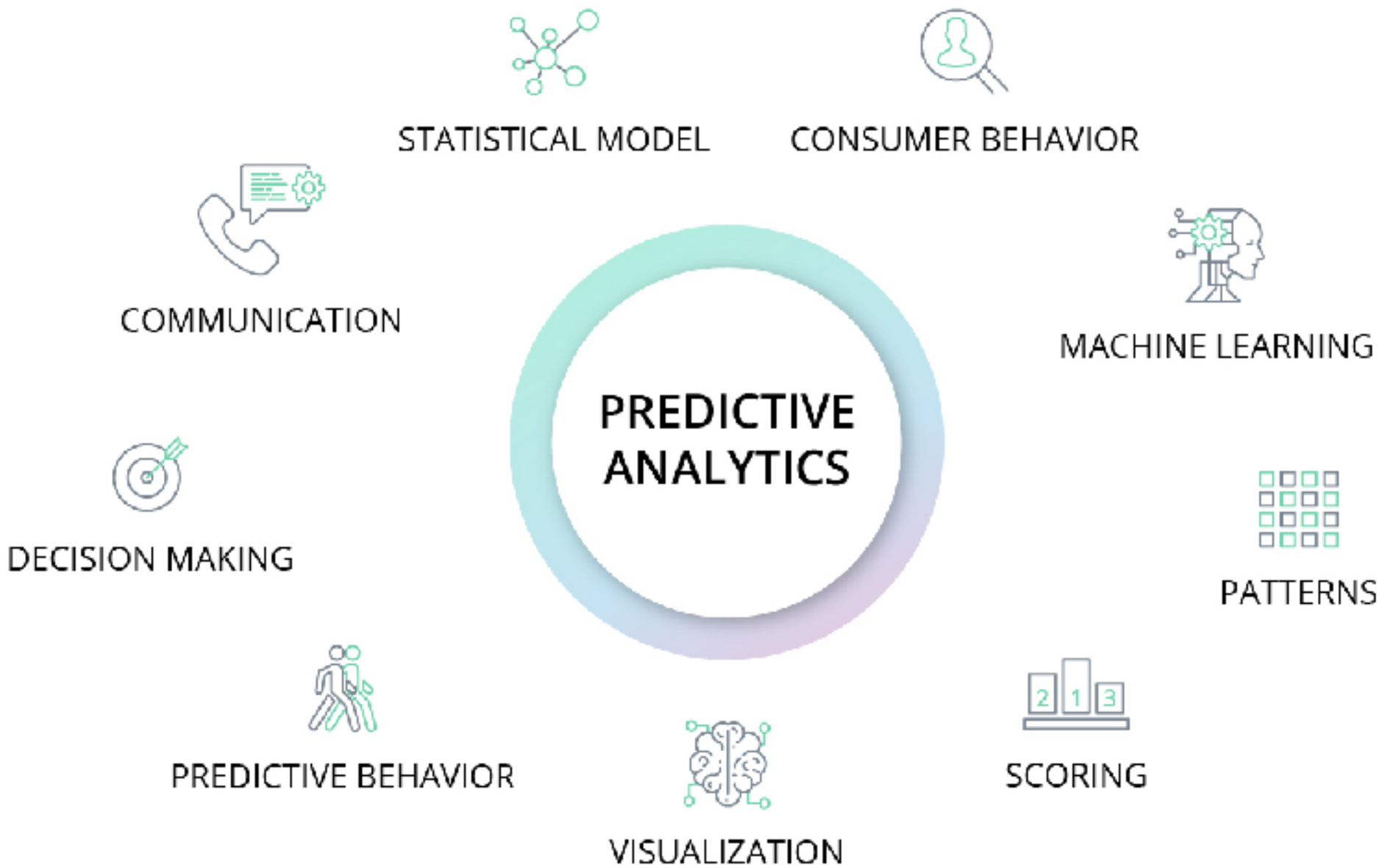
by using historical data, statistical models, and machine learning techniques to forecast future trends.

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

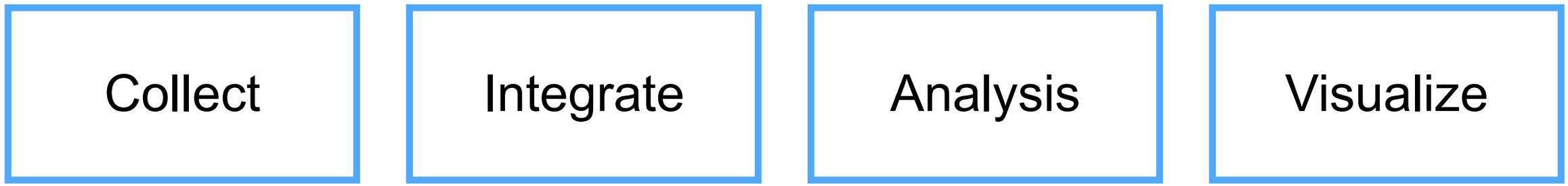
Prescriptive Analytics

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.



Purpose of BI

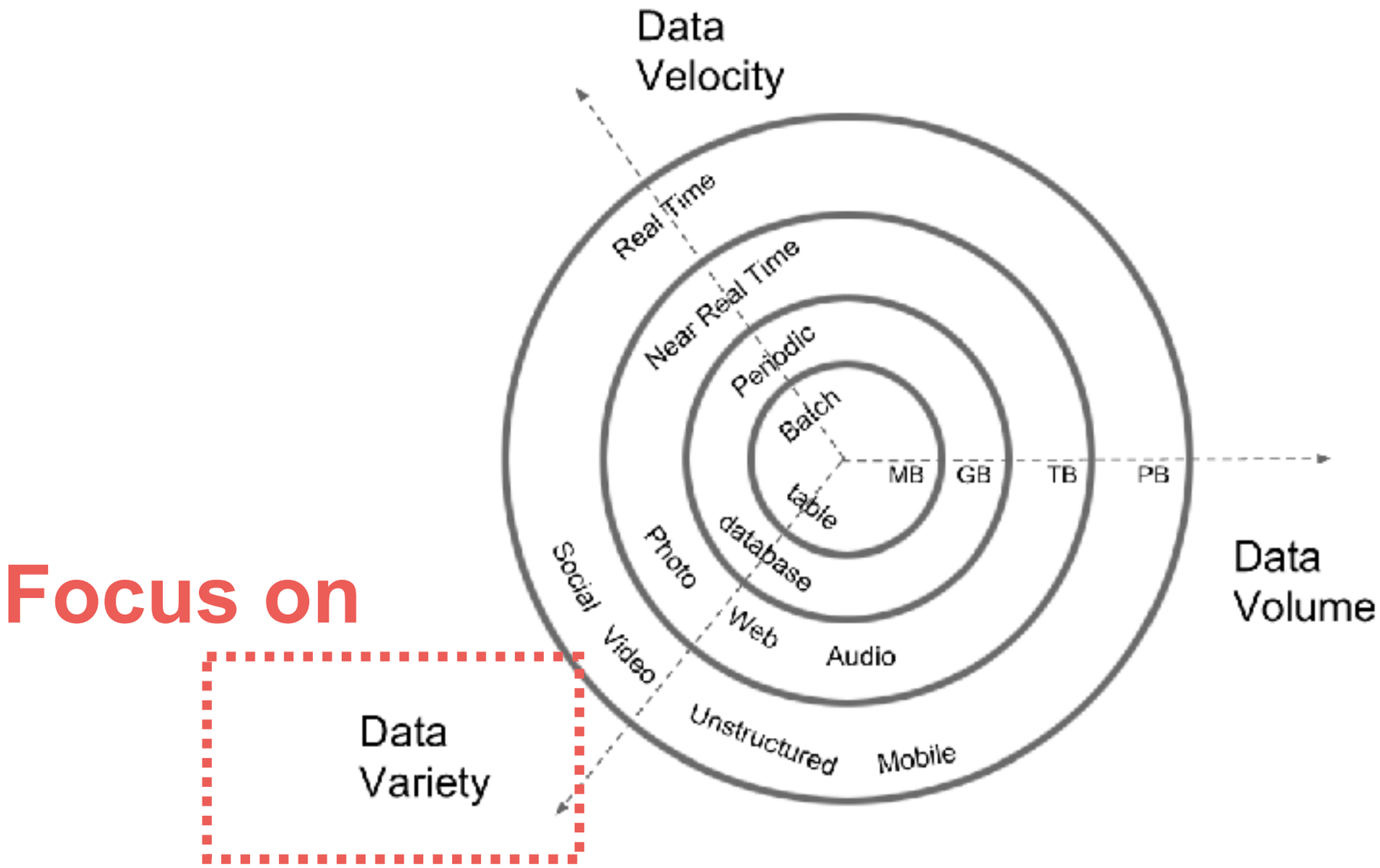
BI is about **reporting and dashboards**
that help businesses monitor their current state
and past performance
by organizing data into actionable insights.

Start with Data Sources

Data Sources !!

Characteristics for Big Data

3 V's !!!

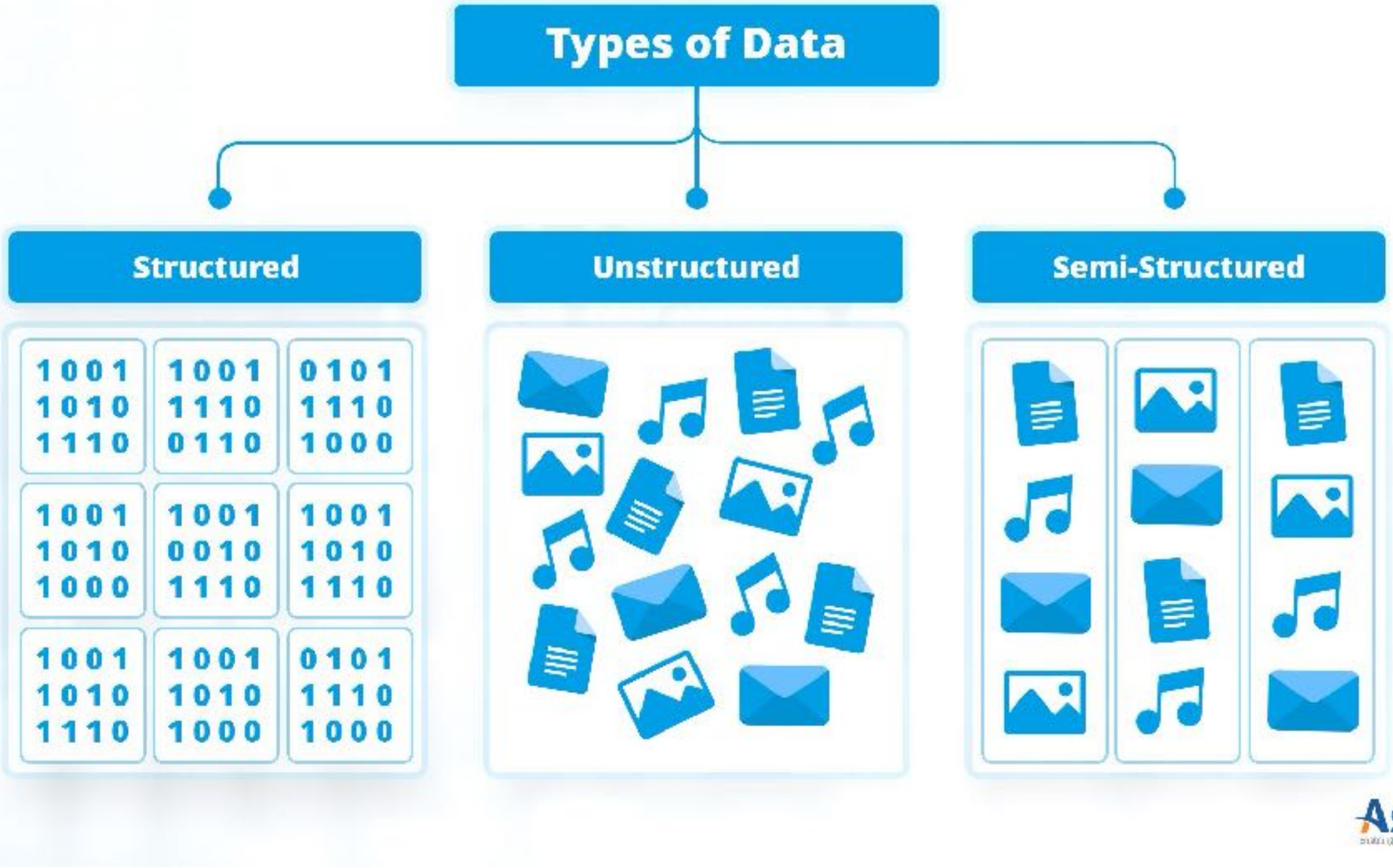


Focus on

Types of data sources

Structured
Unstructured
Semi-structured

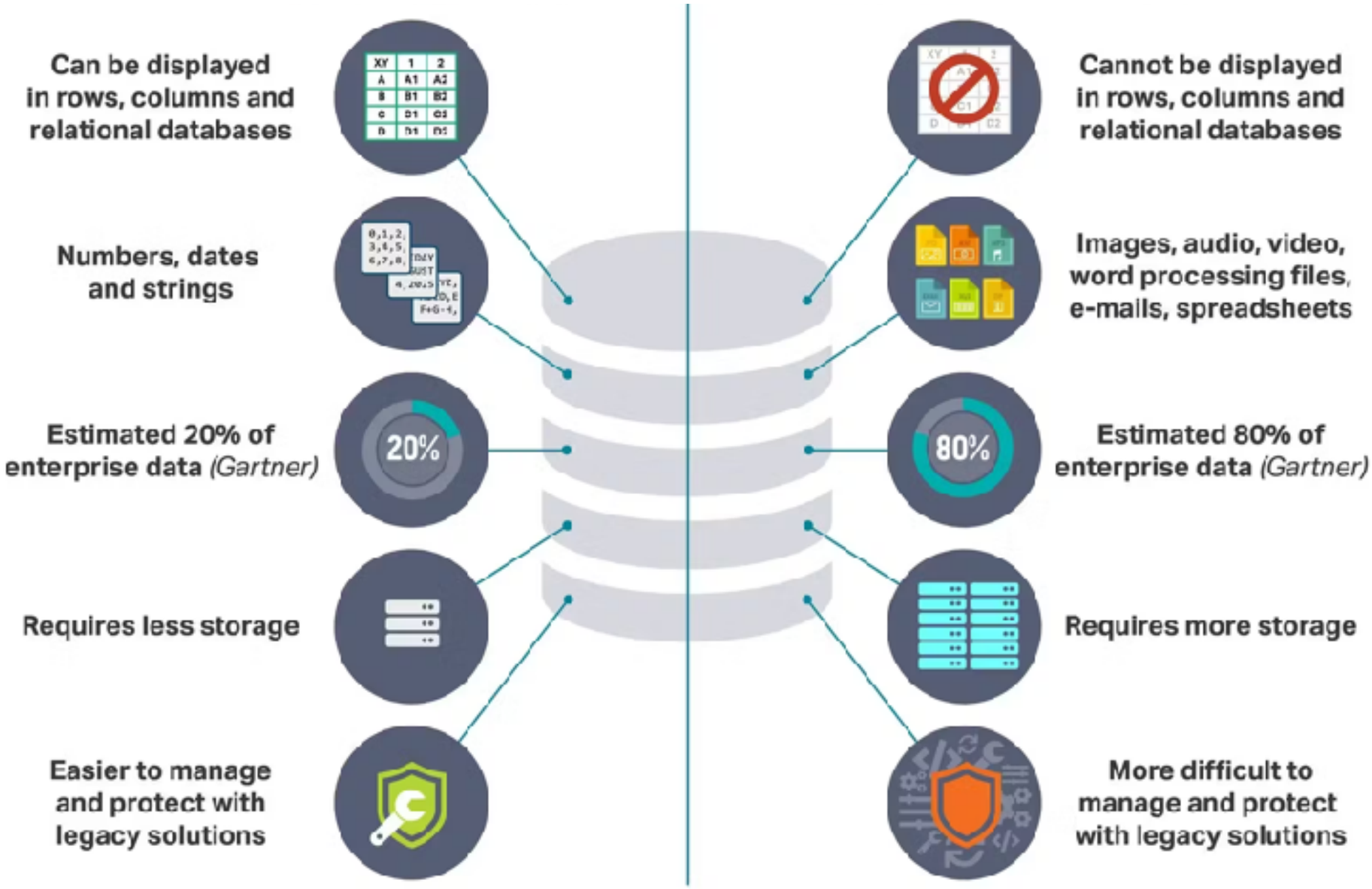
Types of data



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

Structured vs Unstructured data



Unstructured Data Challenges

Inability to process growing data volumes

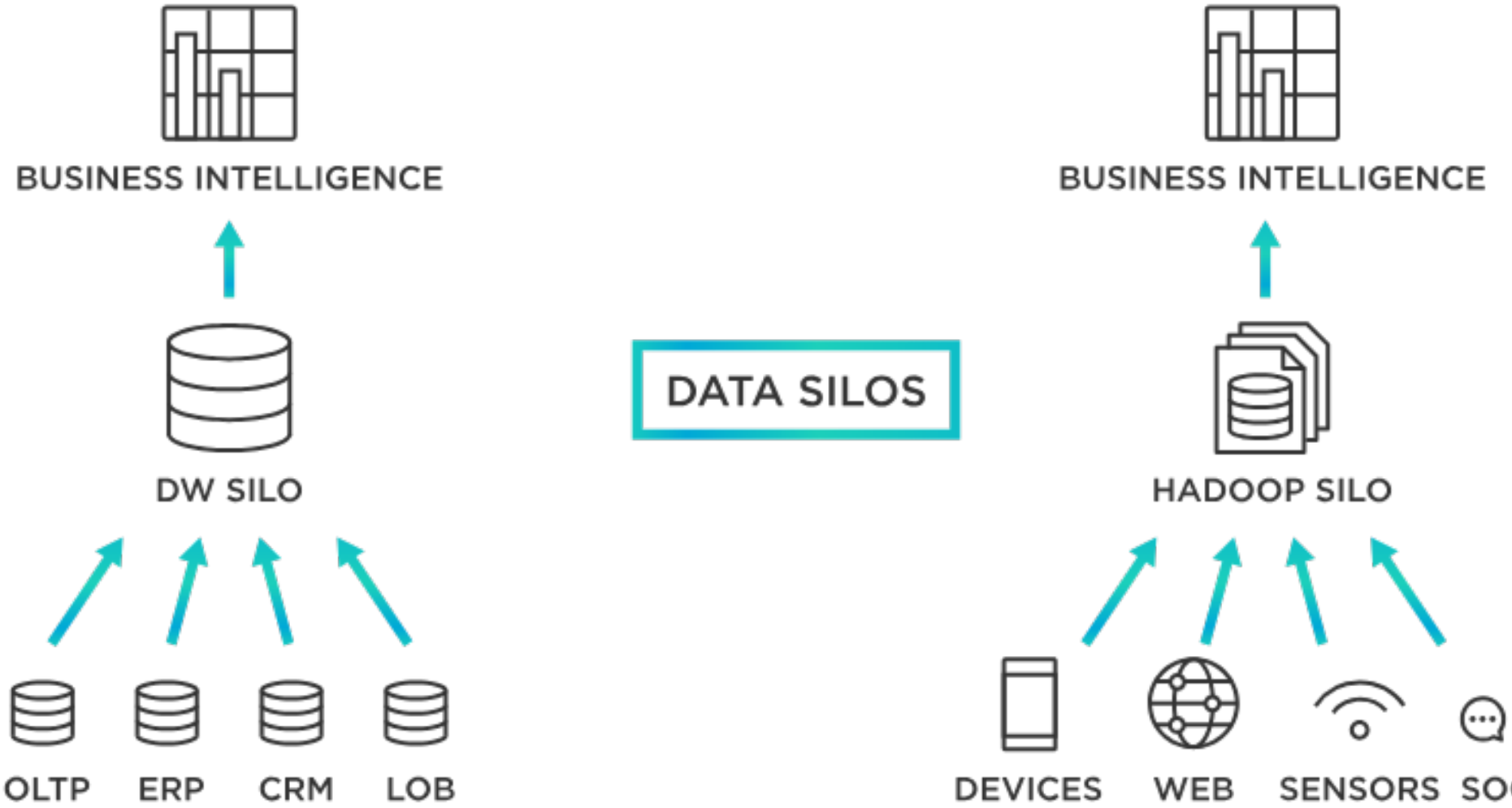
Accessing **siloed** data

Regulatory non-compliance

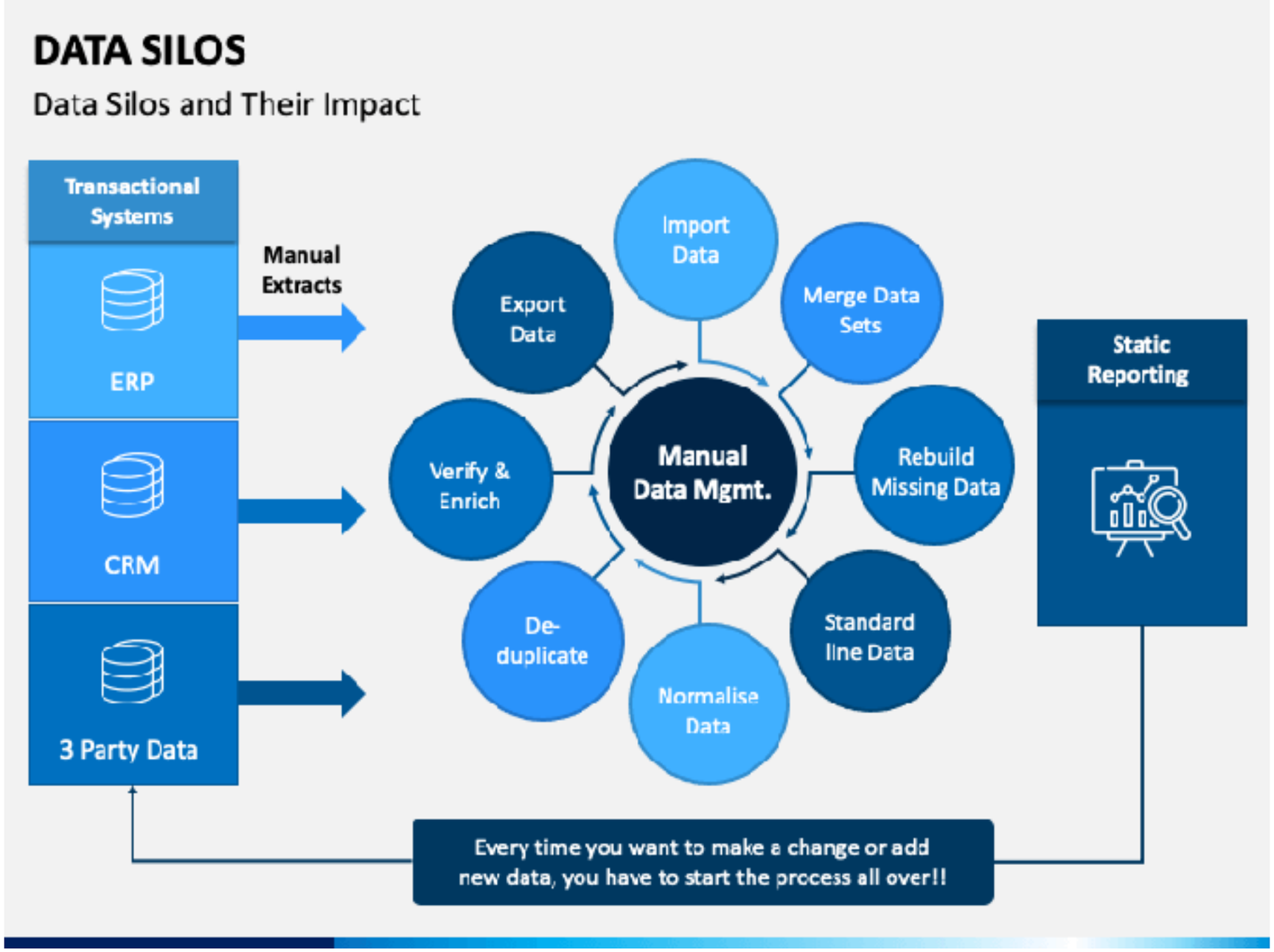
Reduced data **usability**

Increased vulnerability to cyber attacks

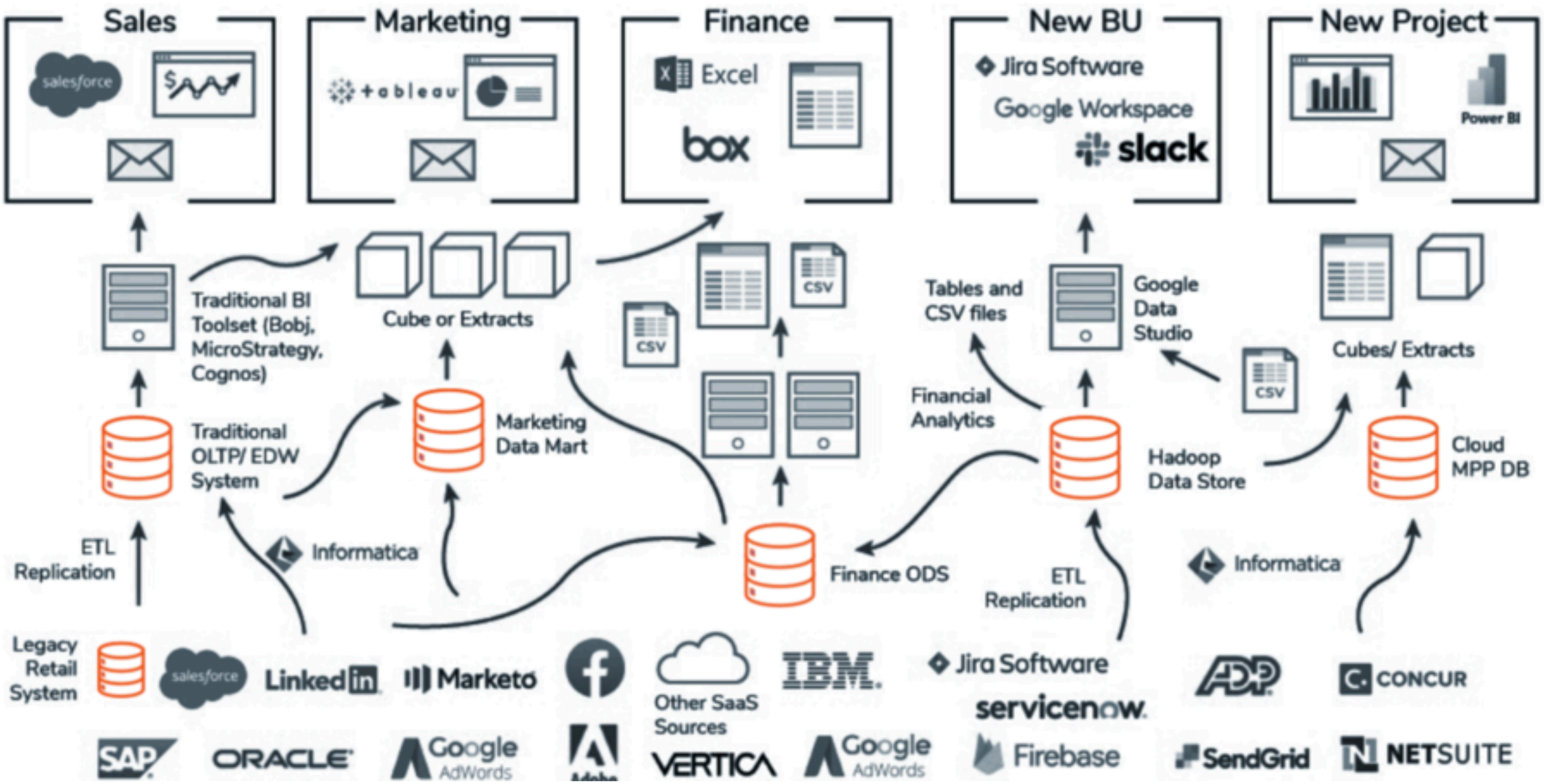
Data siloed ?



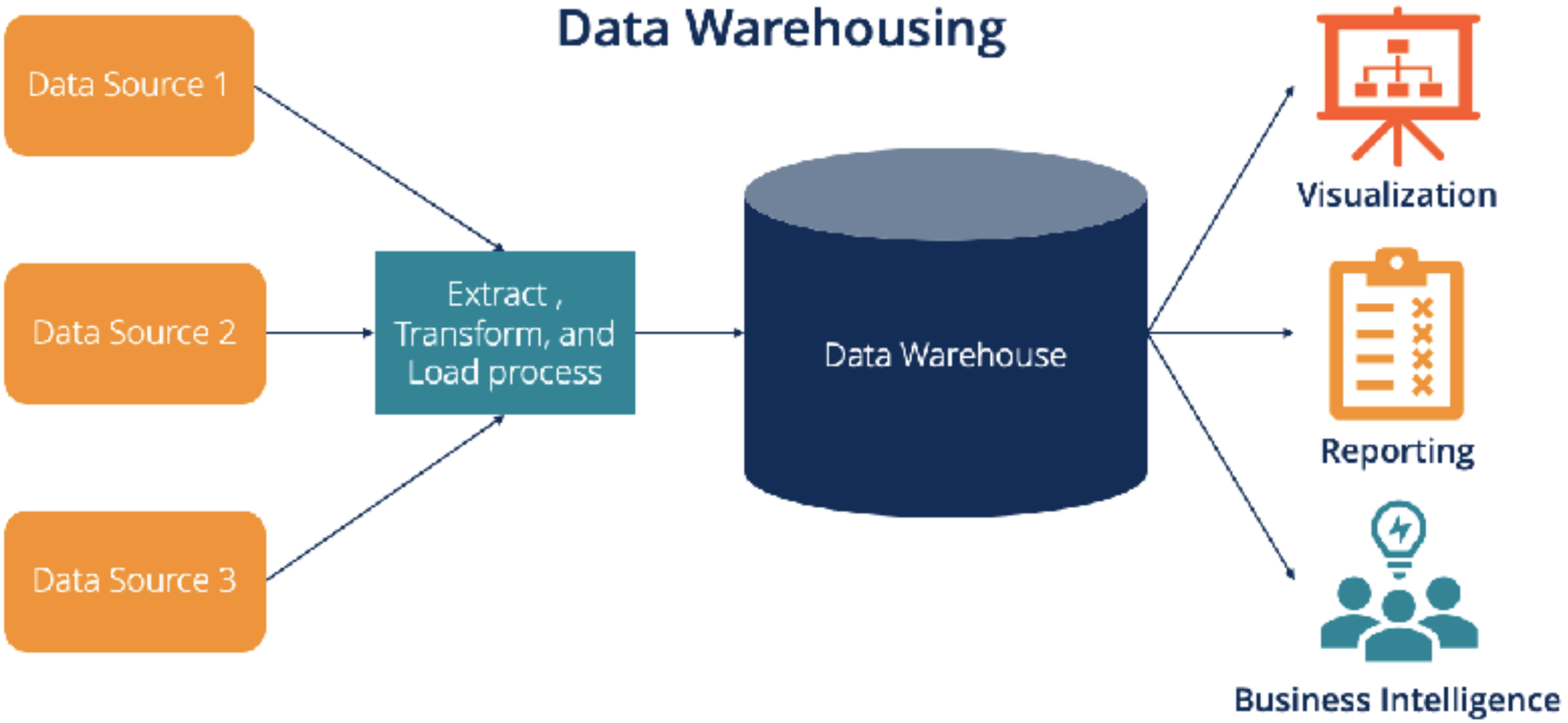
Data siloed ?



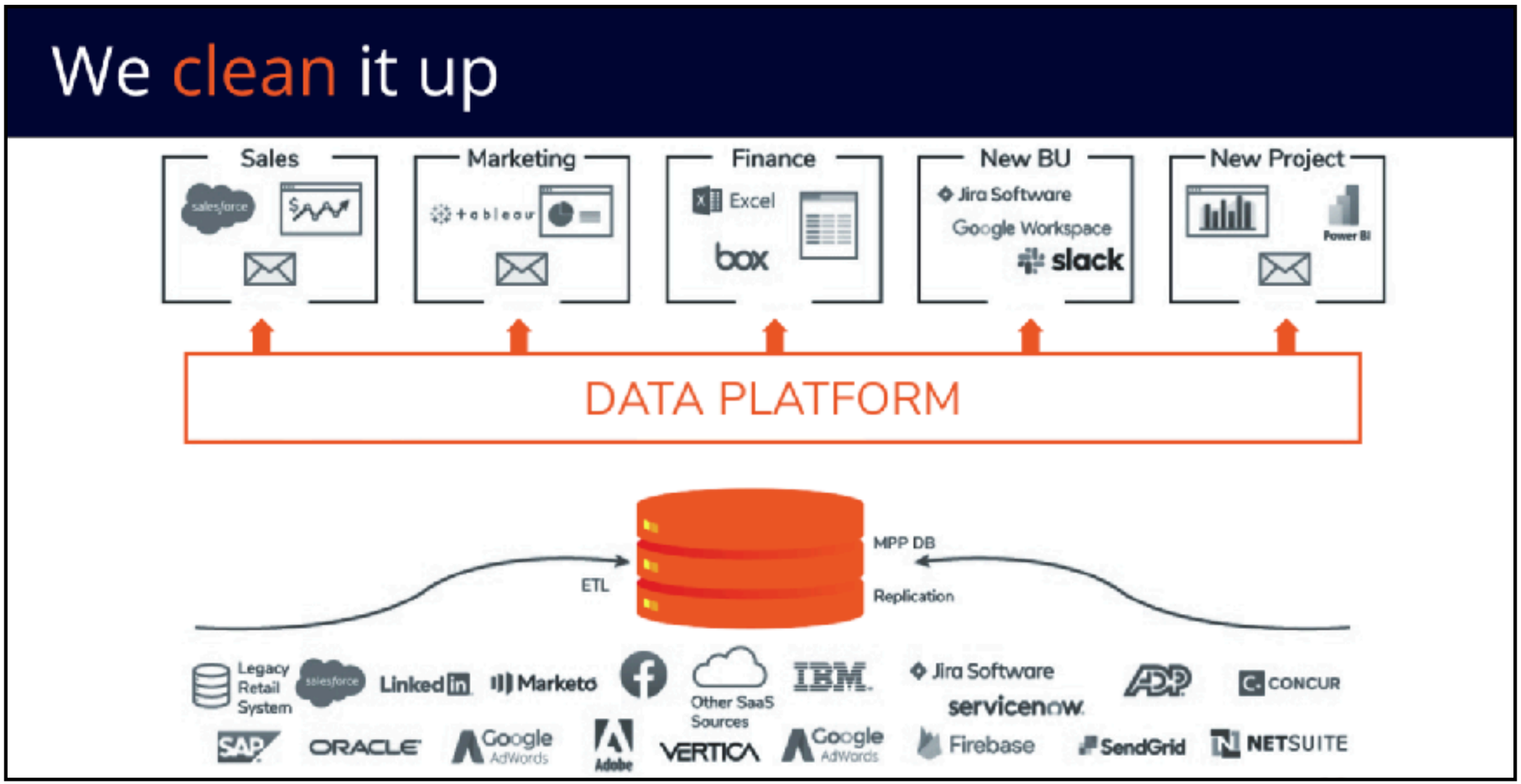
Data siloed ?



Try to solve with data platform



Data Platform



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).

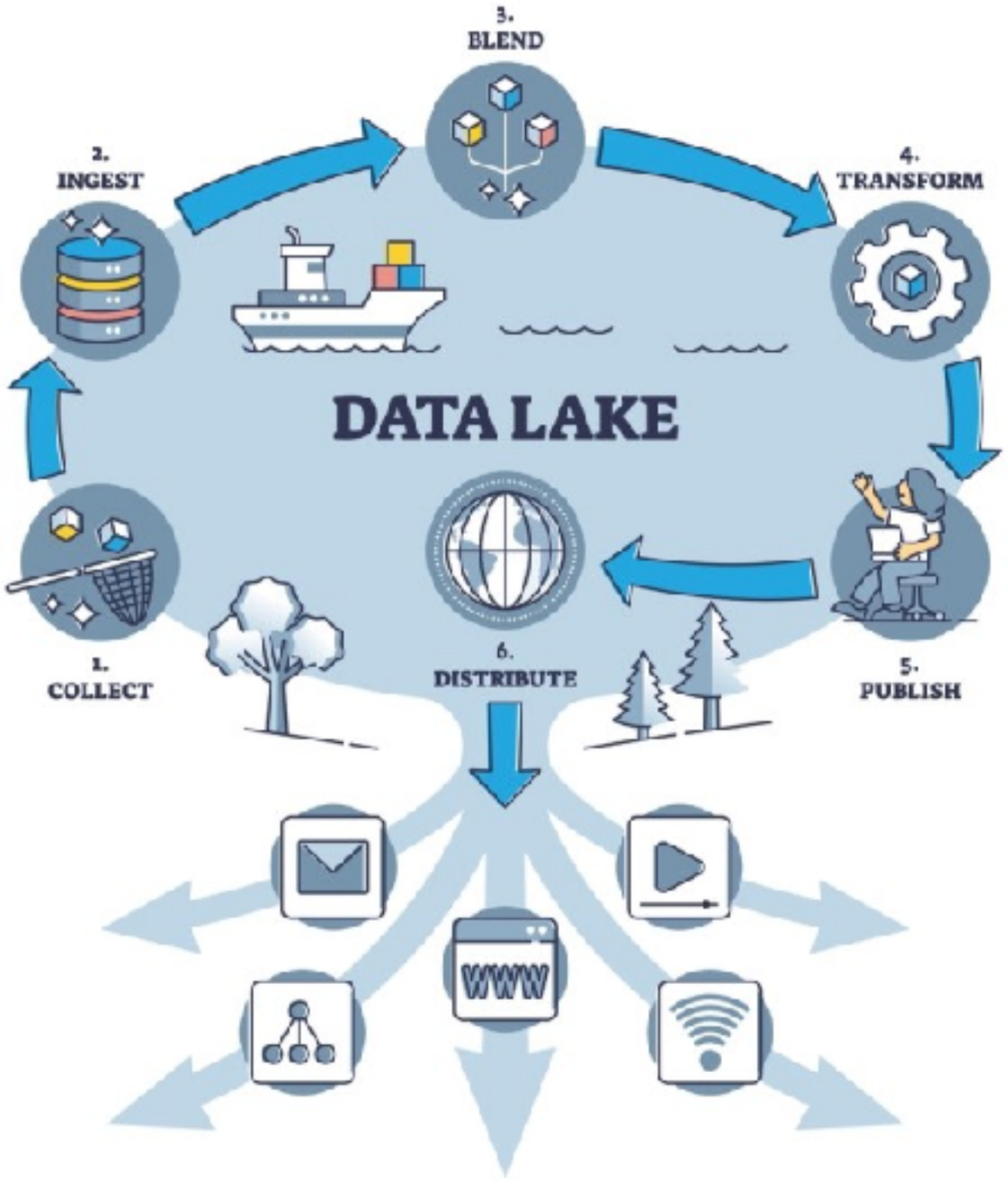
It contains only relevant data for
that particular group, **enabling quicker access** and more tailored analysis.

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.

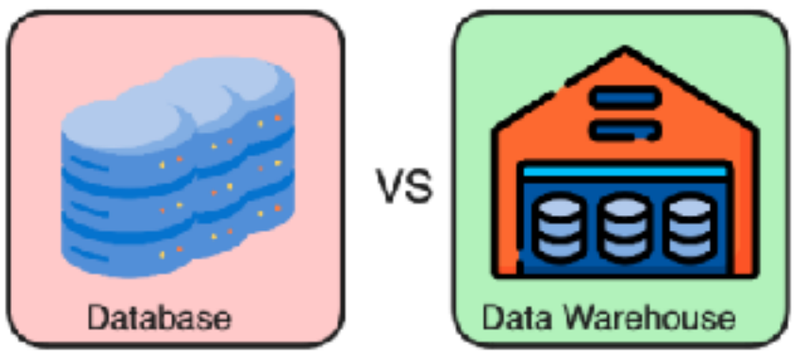
Data Lake



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



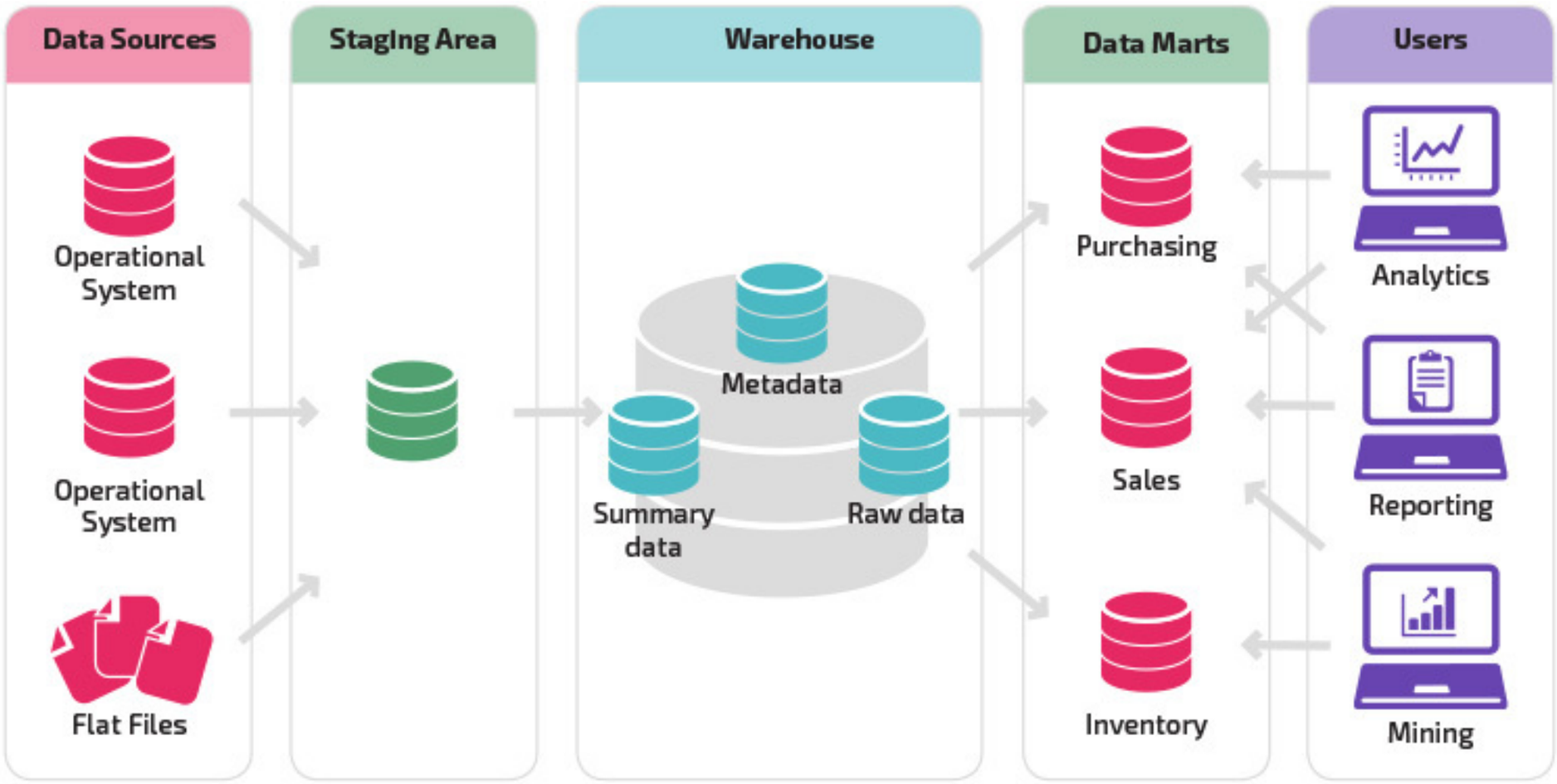
vs

vs

vs

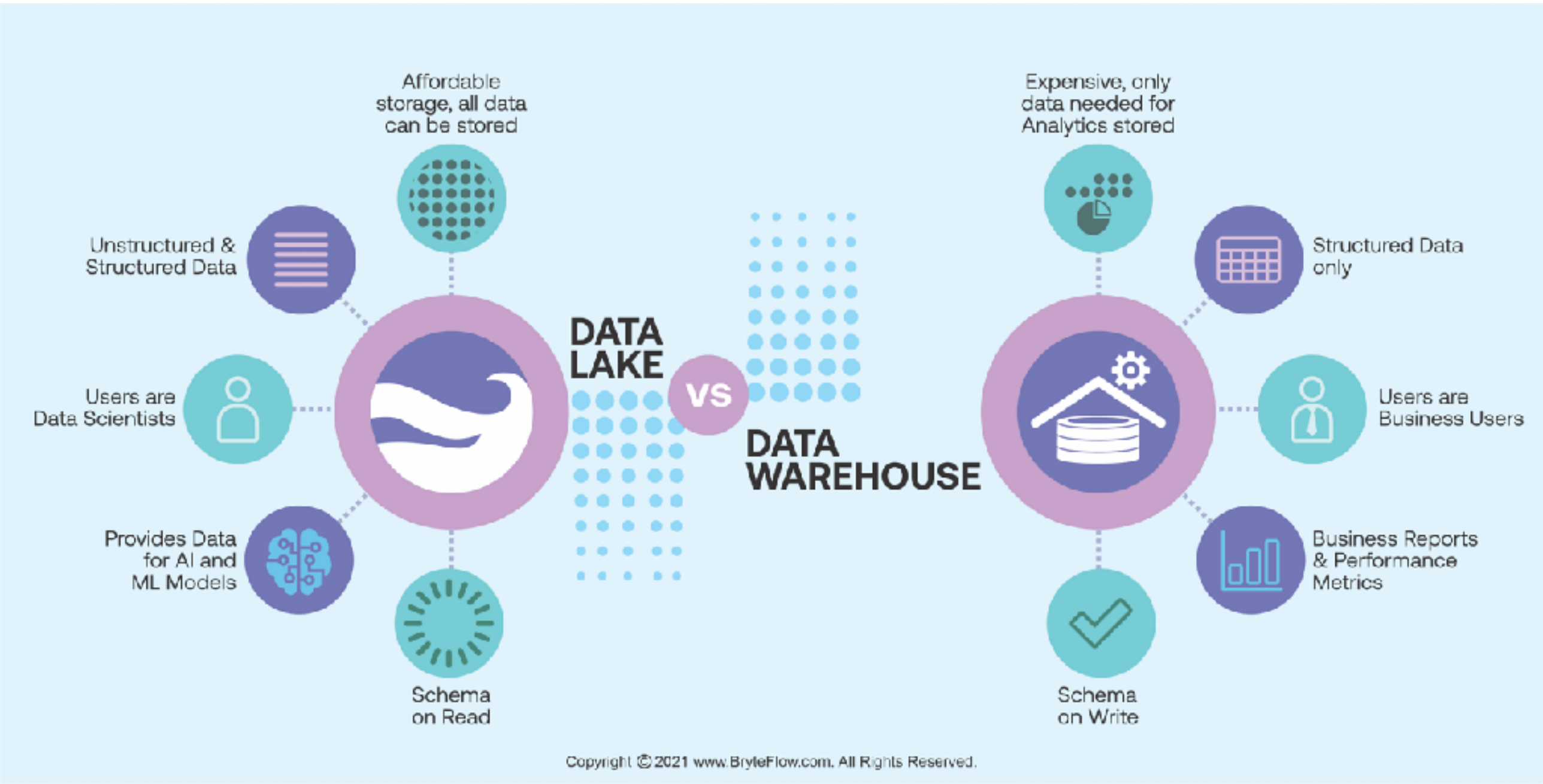
Scope	Application-specific	Organization-wide, structured data.	Department-specific, structured data.	Organization-wide, any type of data
Data Type	Structured	Structured	Structured	Structured, semi-structured, unstructured.
Structure	Predefined schema	Schema on write	Schema on write (inherited from data 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/

Data Warehouse vs Data Lake



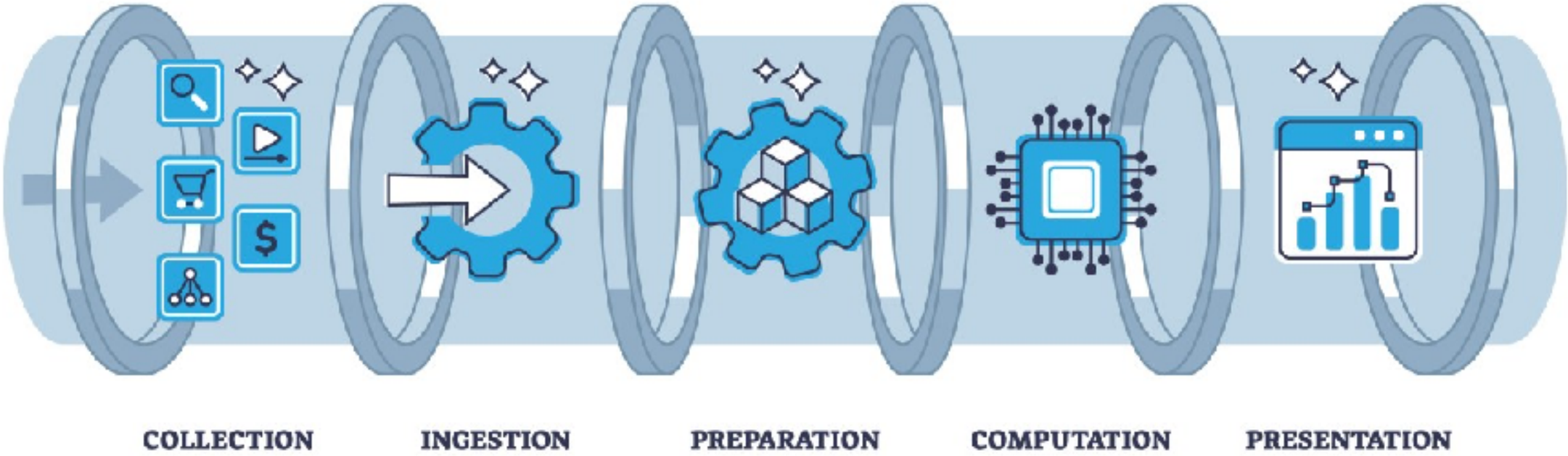
https://panoply.io/data-warehouse-guide/data-mart-vs-data-warehouse/

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

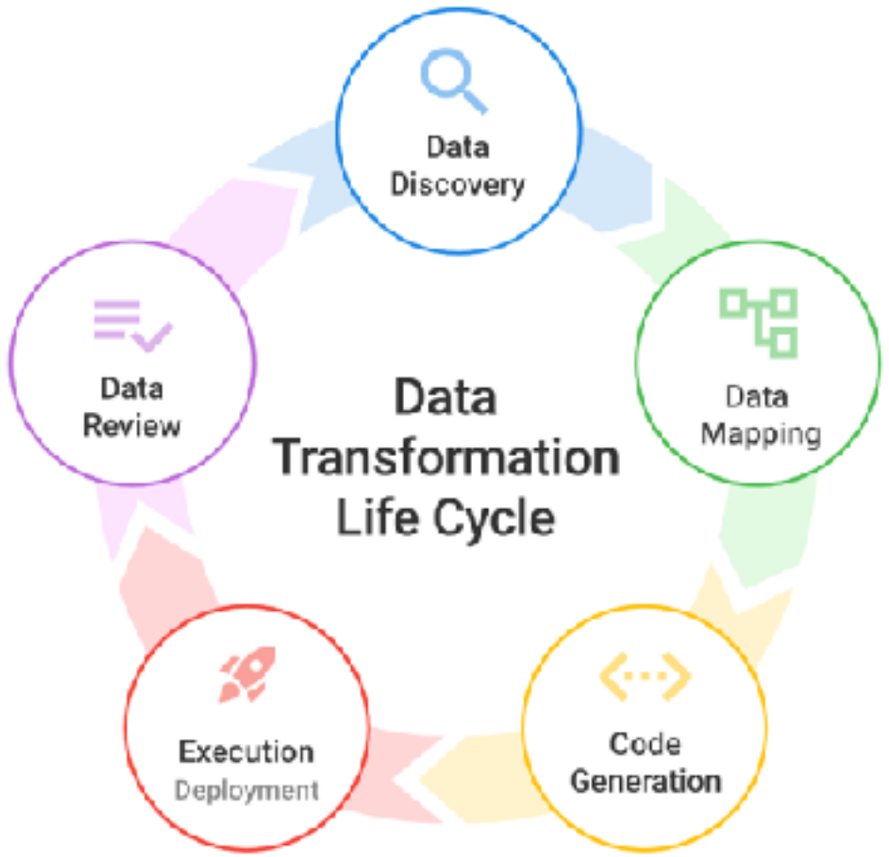
The process of collecting data from various sources and bringing it into the pipeline.

Batch processing

Real-time processing

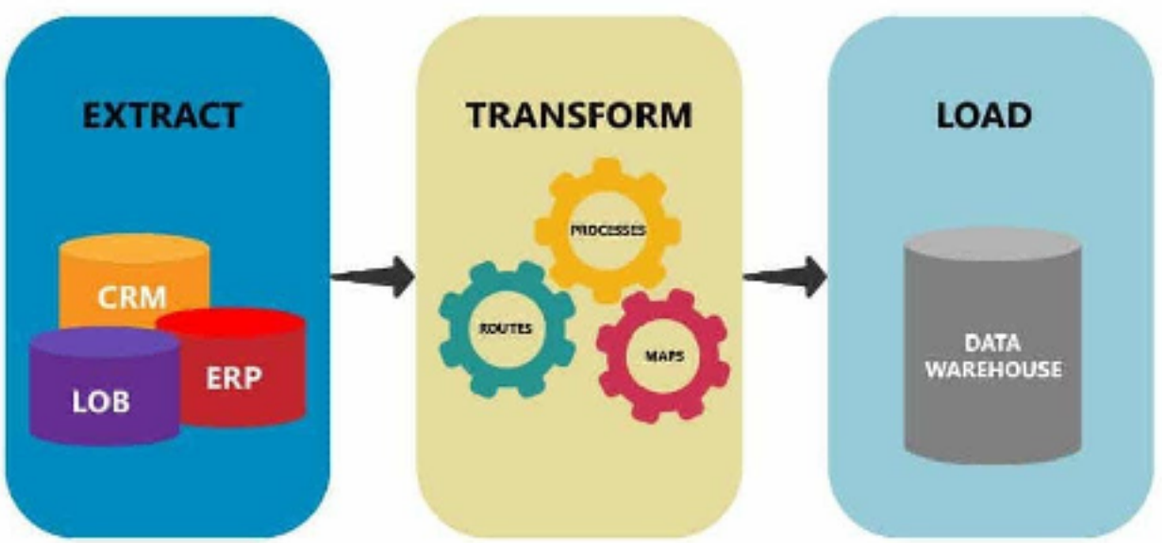
Data Transformation

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



Data Loading

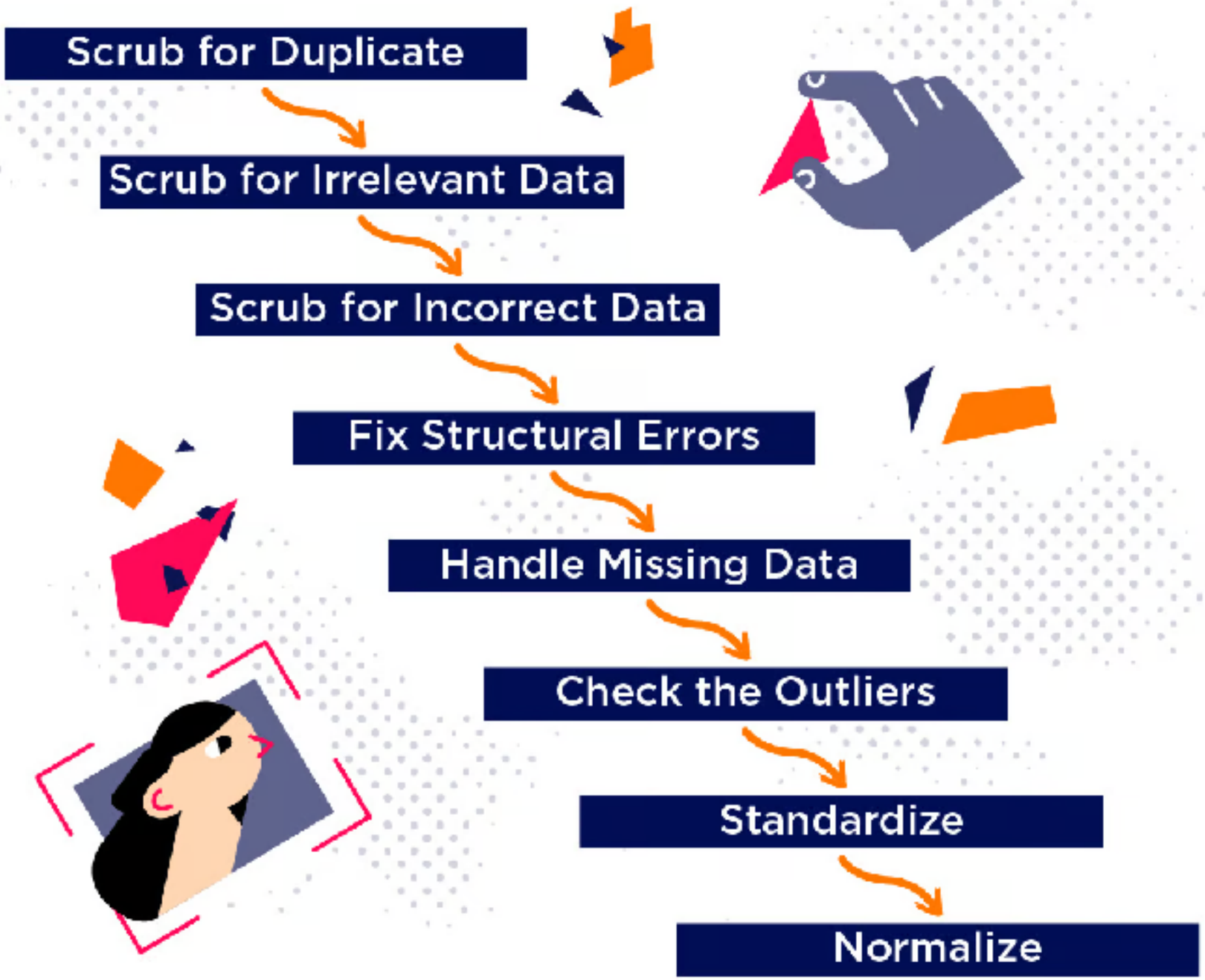
The transformed data is loaded into a target destination, which could be a data warehouse, database, or cloud storage platform.



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 ?

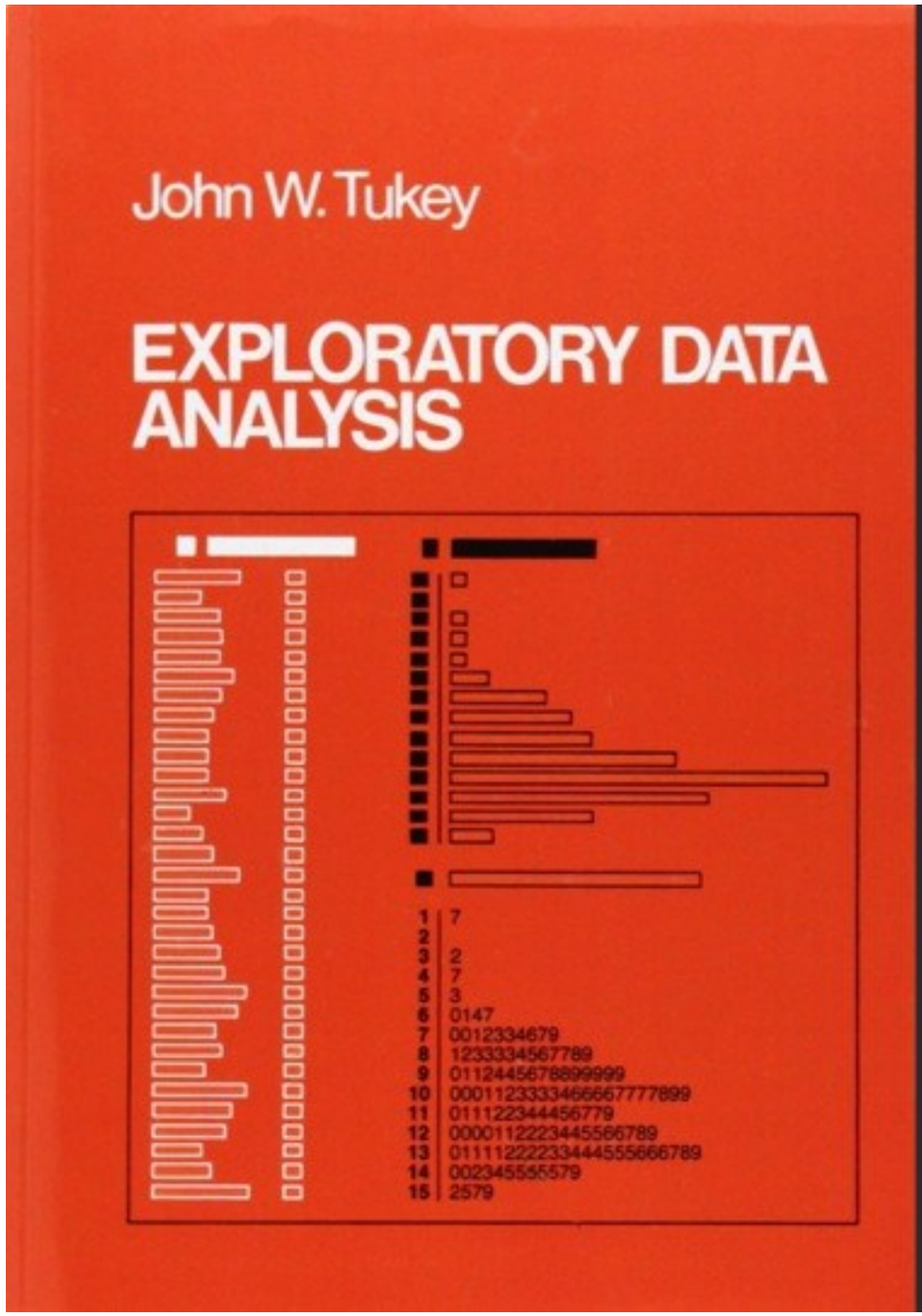
Key Techniques with MS Excel

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)



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)

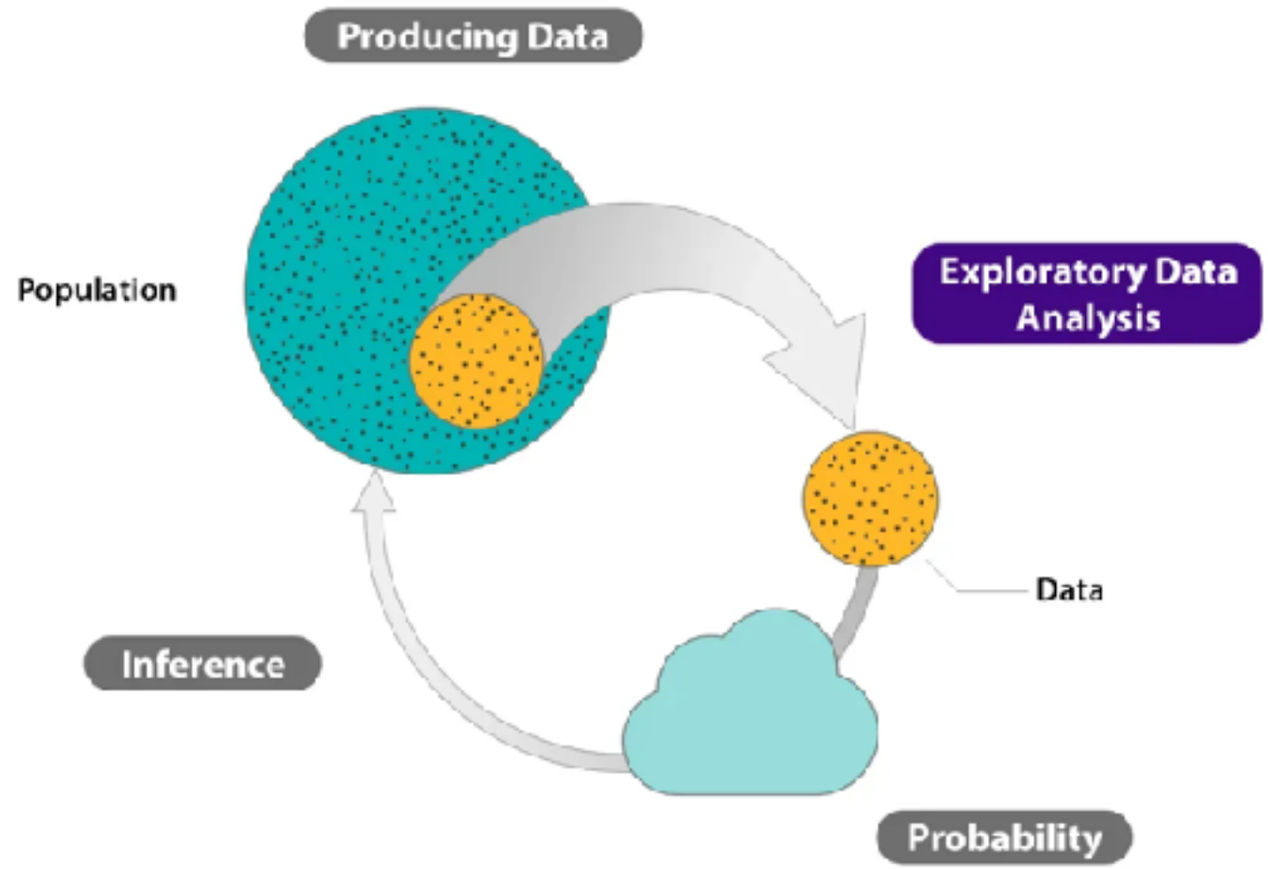
Exploratory Data Analysis

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

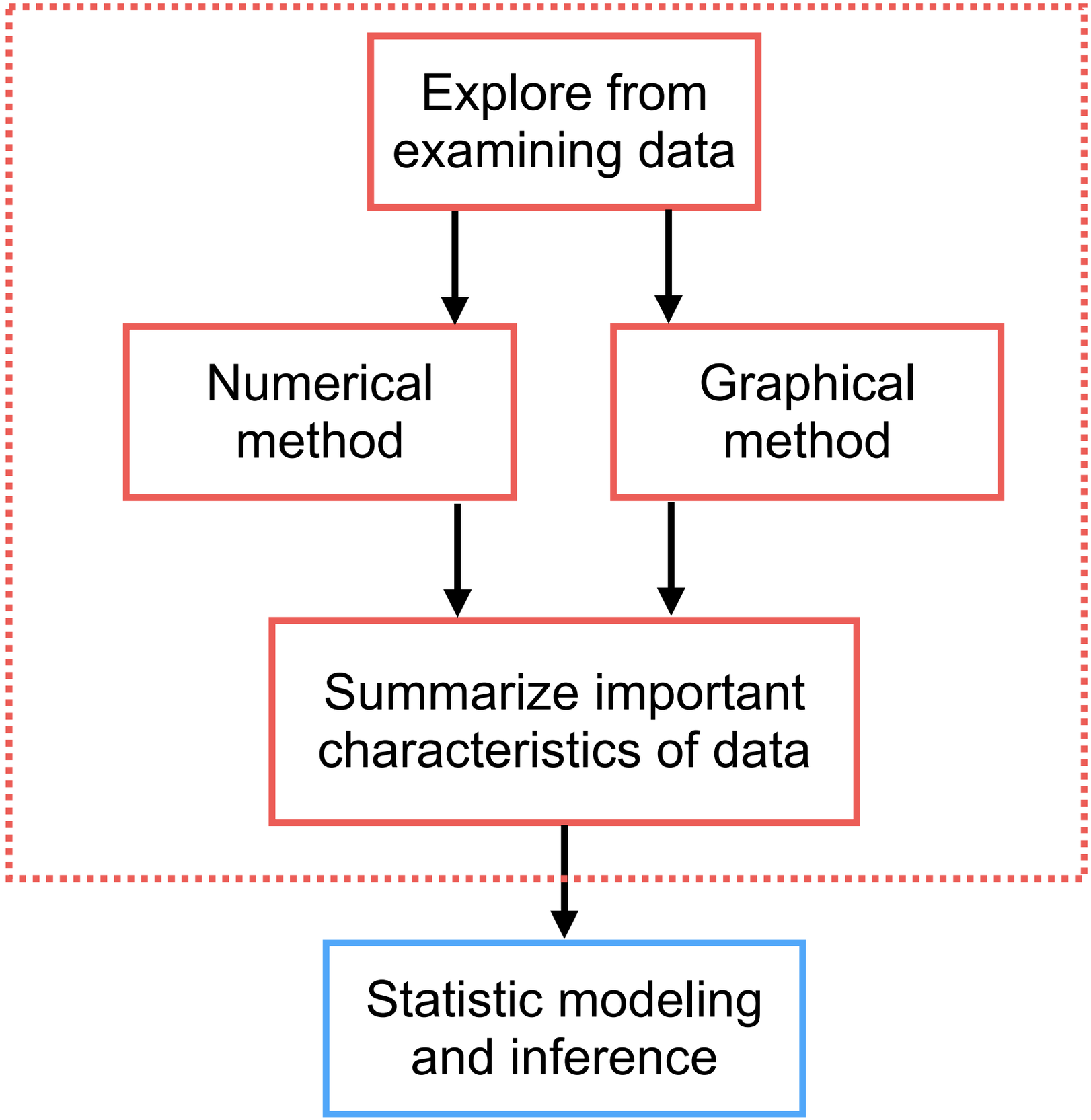
Exploratory Data Analysis

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

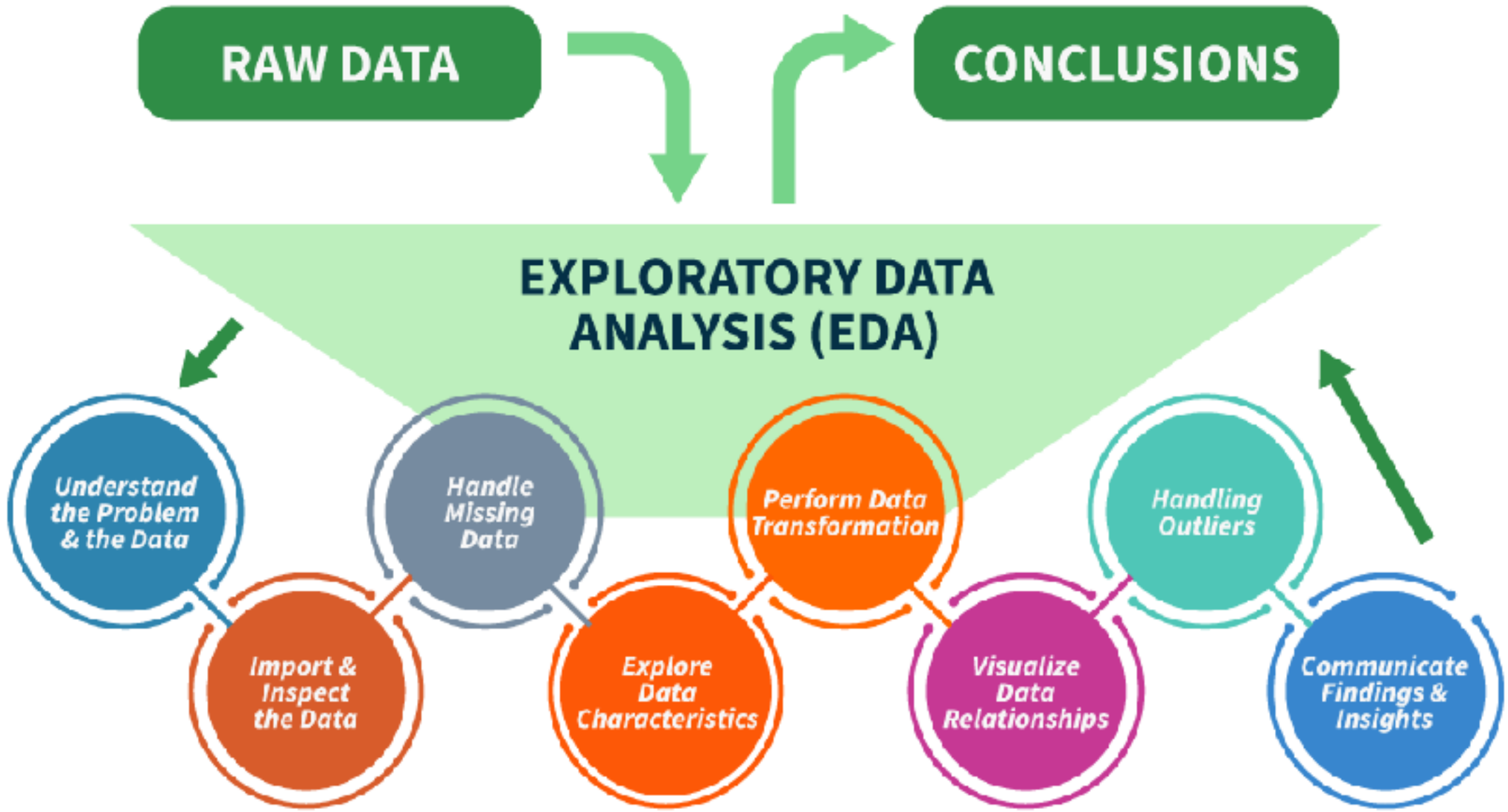


Why EDA ?

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



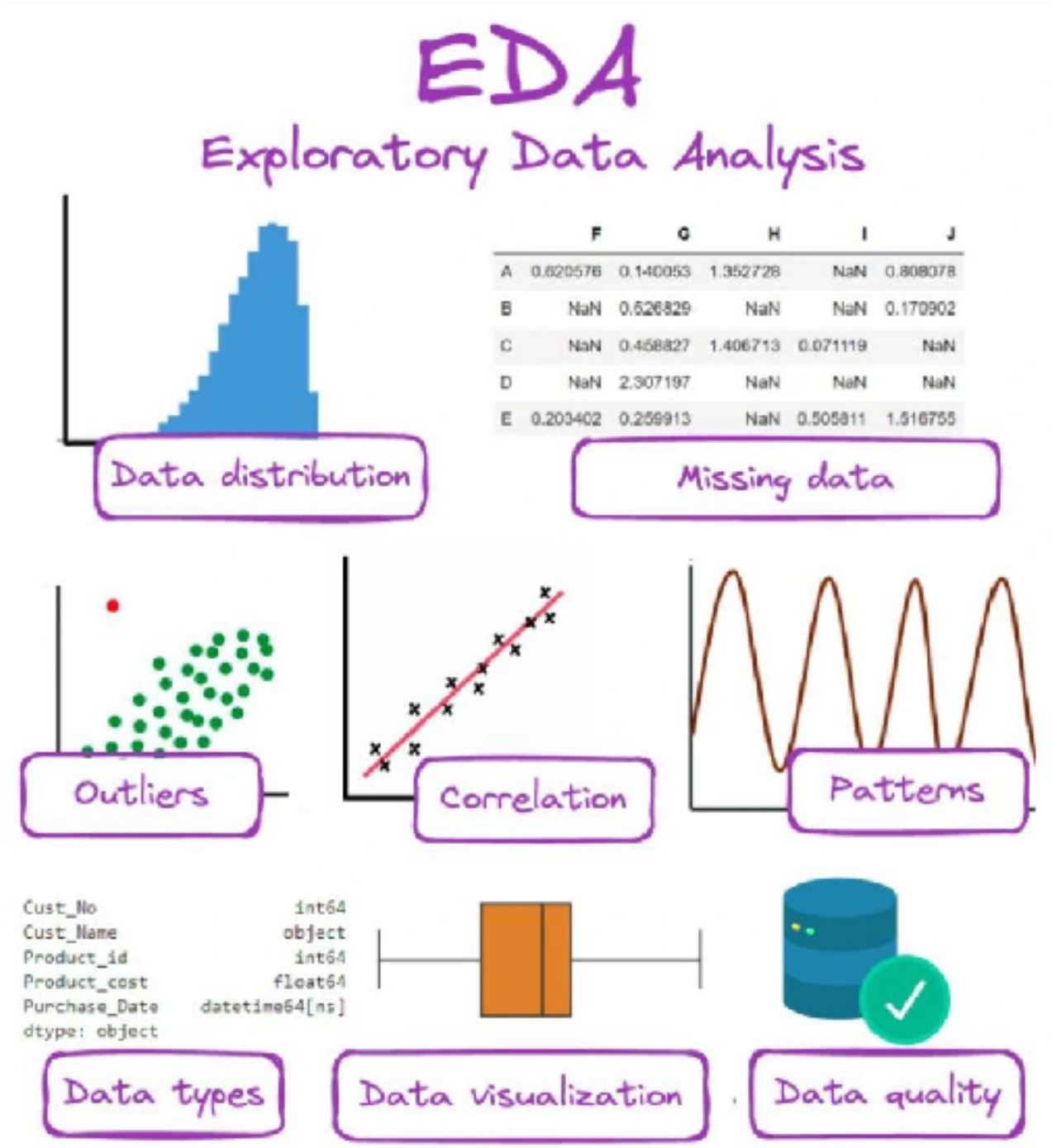
Steps for Performing Exploratory Data Analysis



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

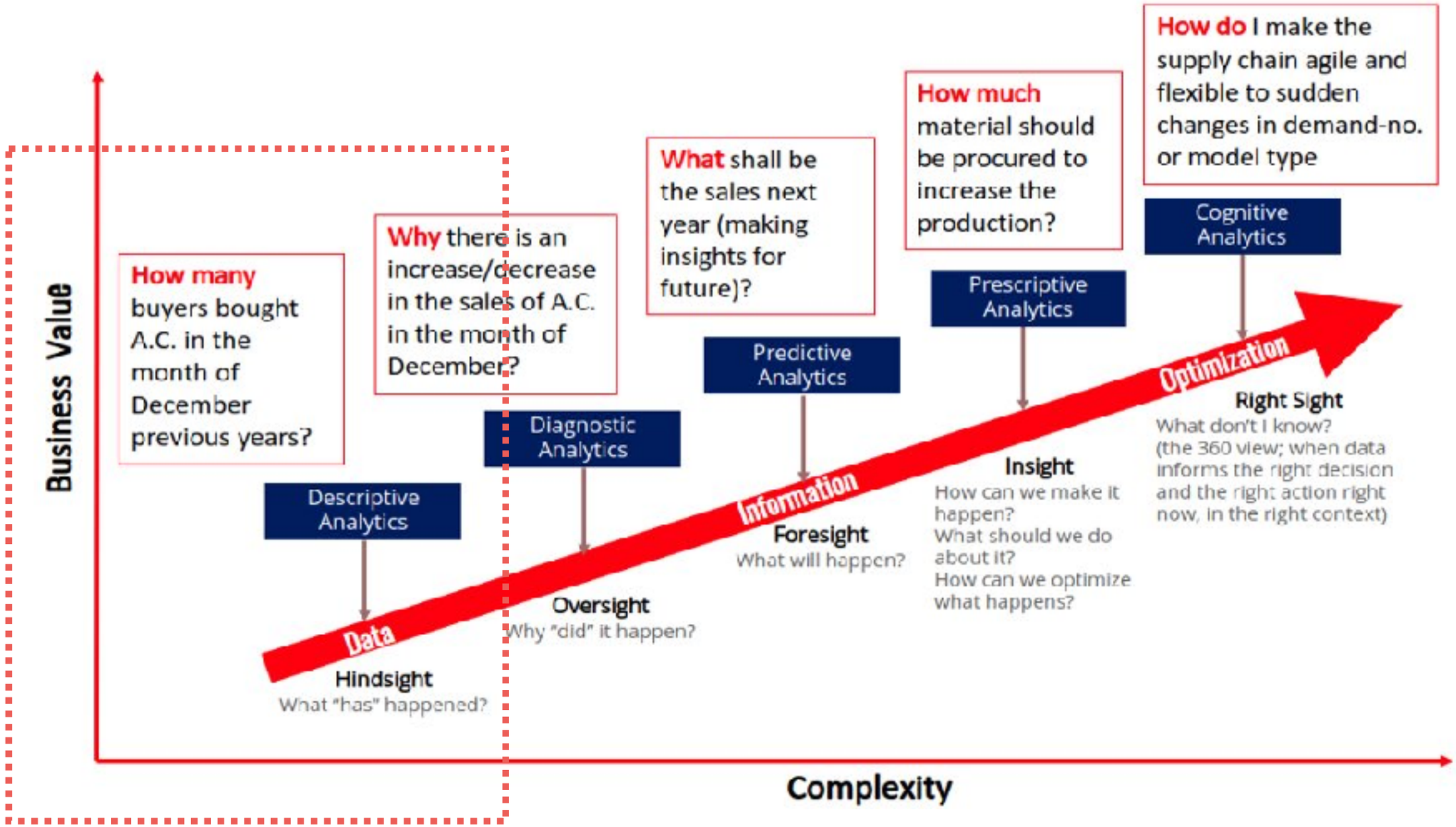
Data manipulation
Data wrangling

Data Analysis

Data Visualization

Statistical Analysis for data analysis

Key Techniques

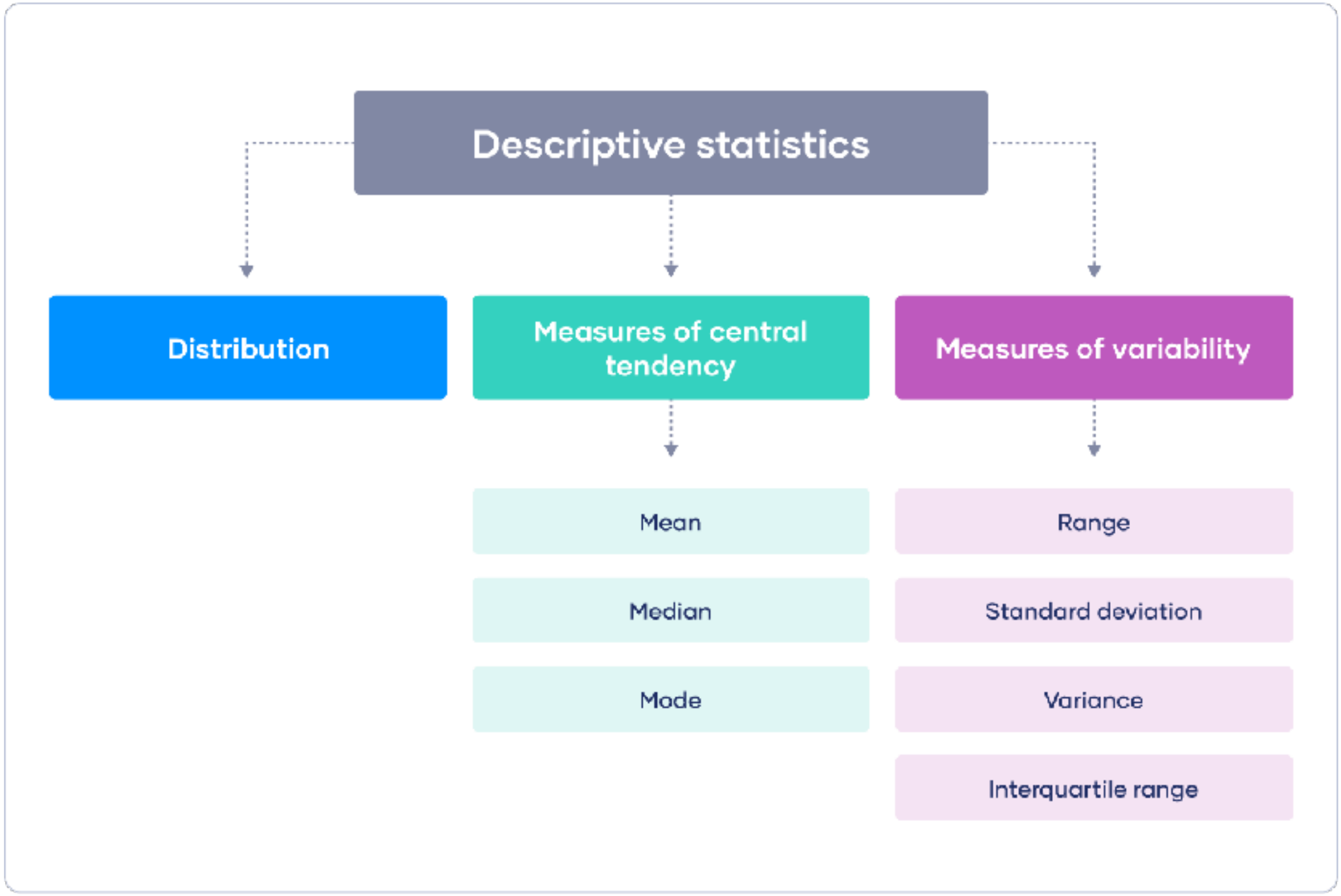


Descriptive statistic

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

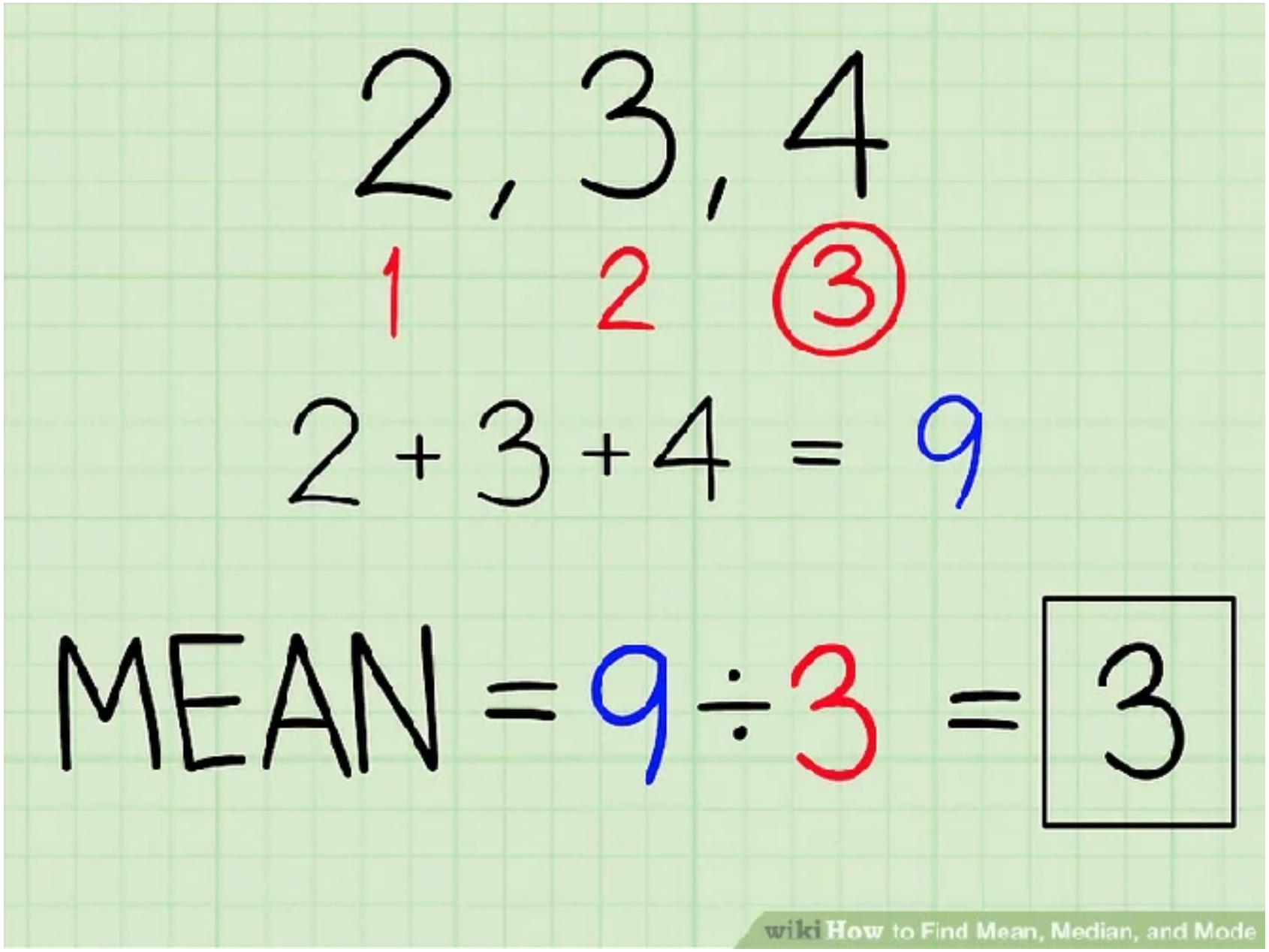
Median

The middle value when data points are arranged in order

Mode

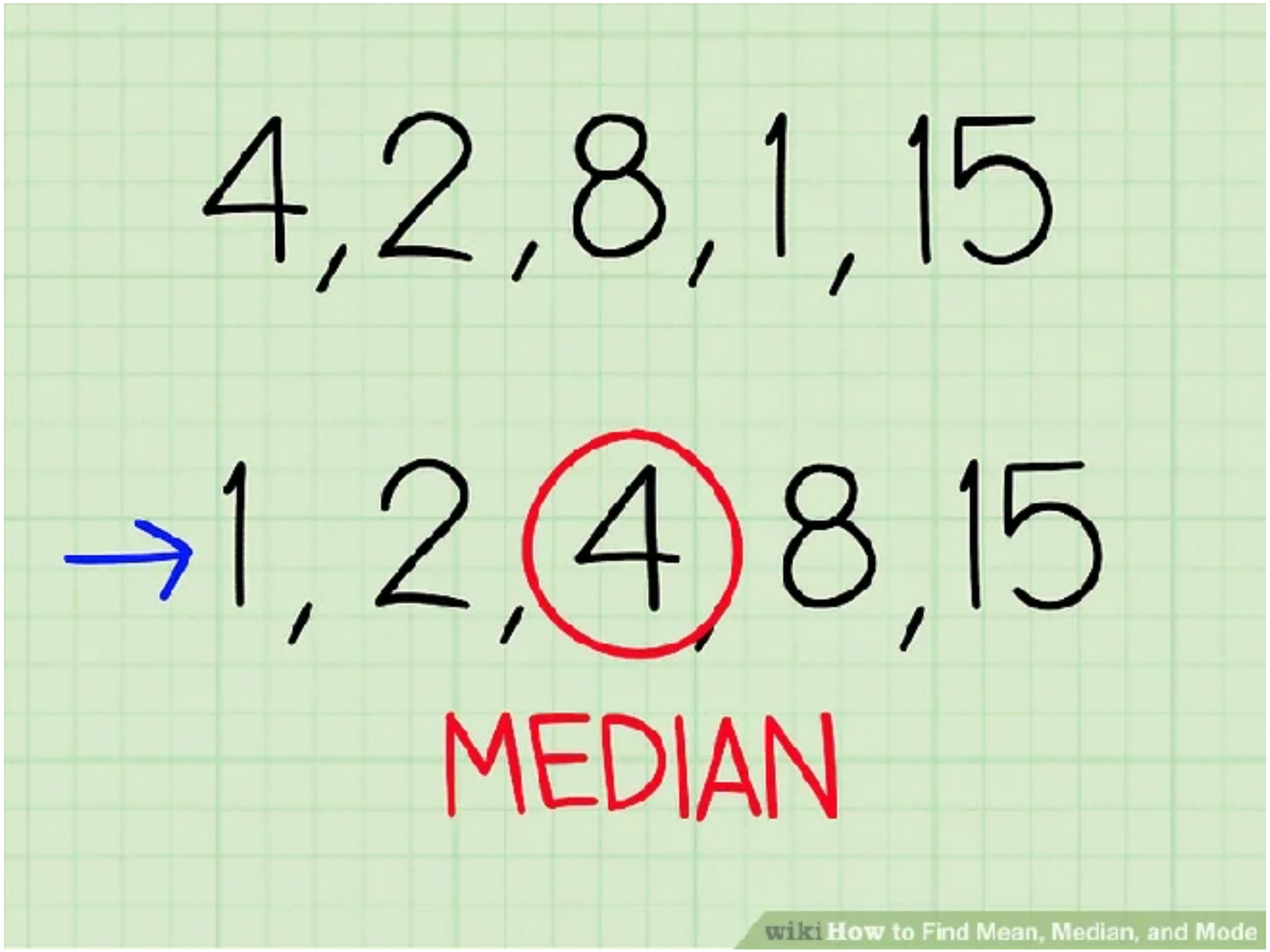
The value that appears most frequently in the dataset

Mean



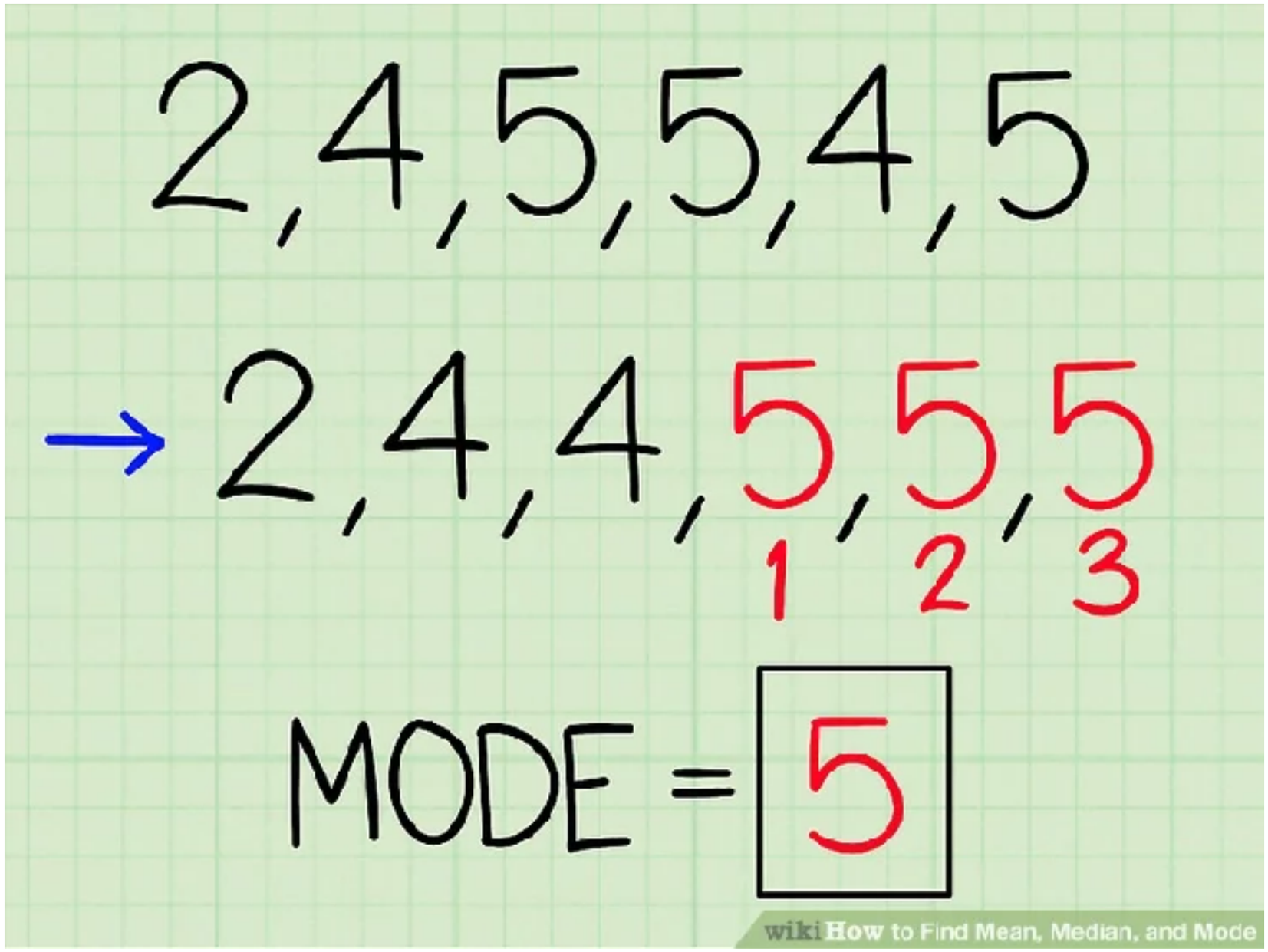
https://www.wikihow.com/Find-Mean,-Median,-and-Mode

Median



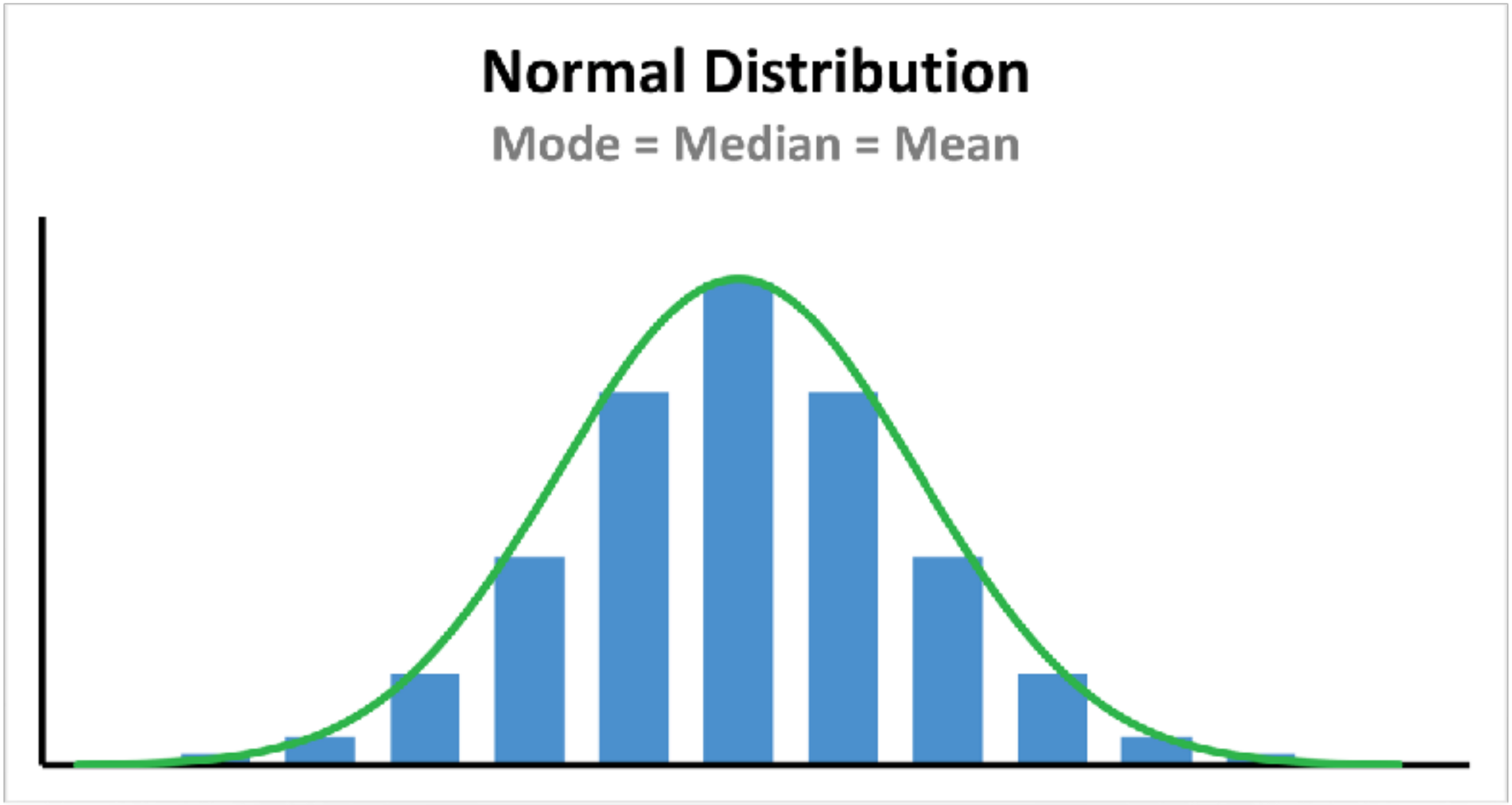
https://www.wikihow.com/Find-Mean,-Median,-and-Mode

Mode

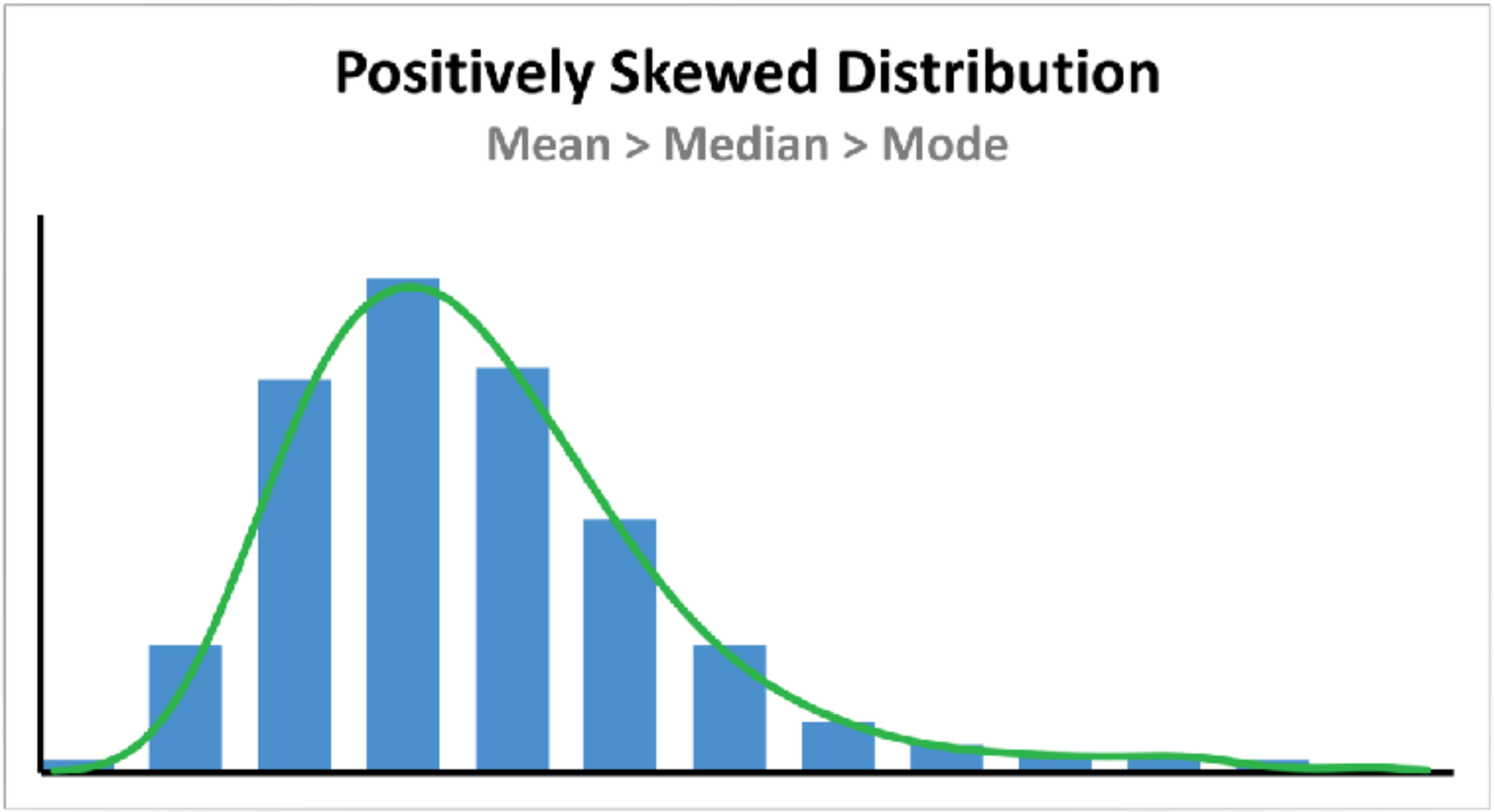


https://www.wikihow.com/Find-Mean,-Median,-and-Mode

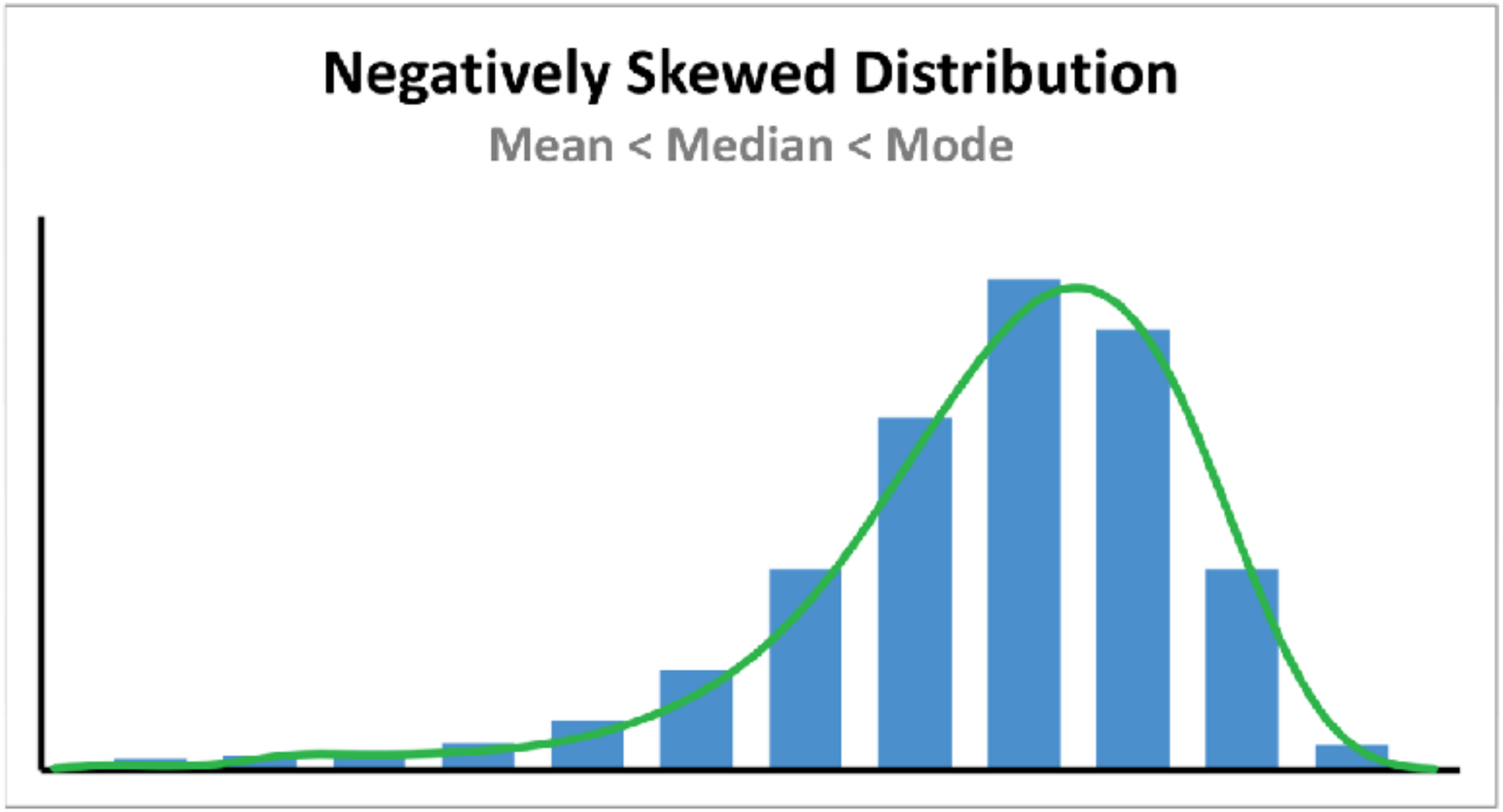
Normal Distribution (Gaussian)



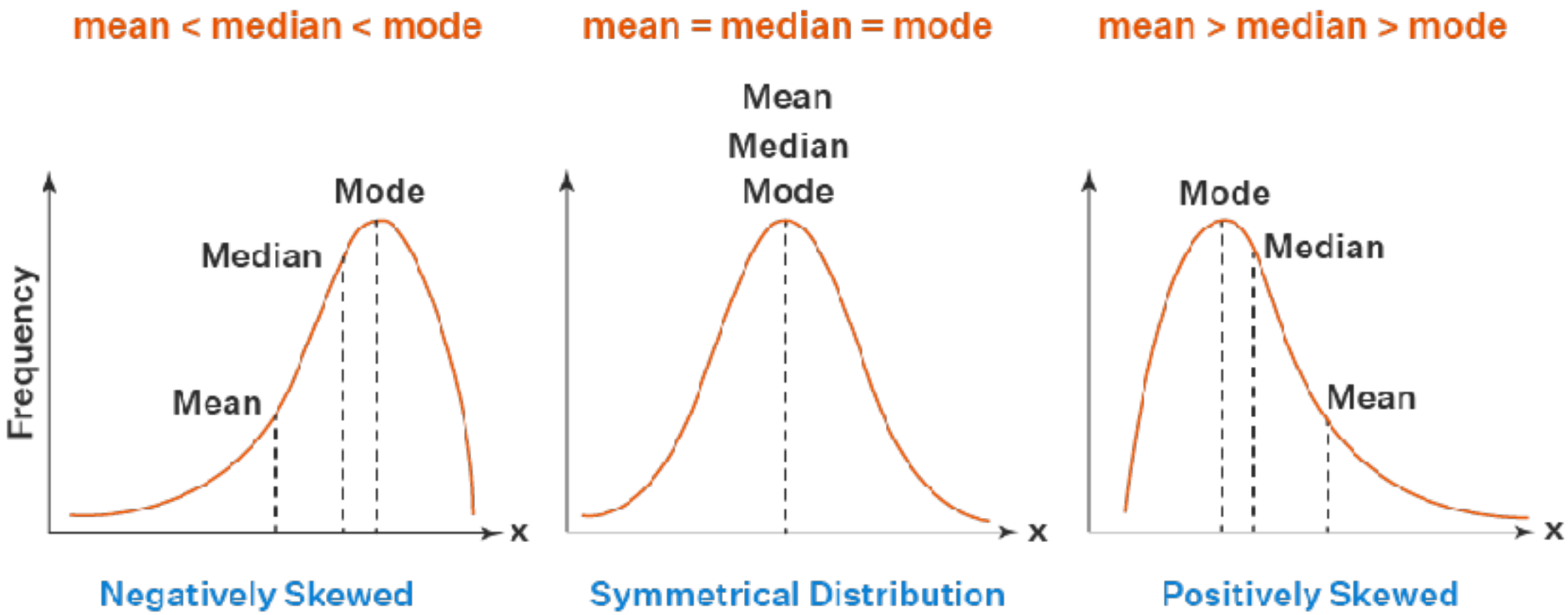
Positive Skewed Distribution



Negative Skewed Distribution



Distribution



Measures of Variability

Range

The difference between the maximum and minimum values

Variance

The average of the squared differences from the mean

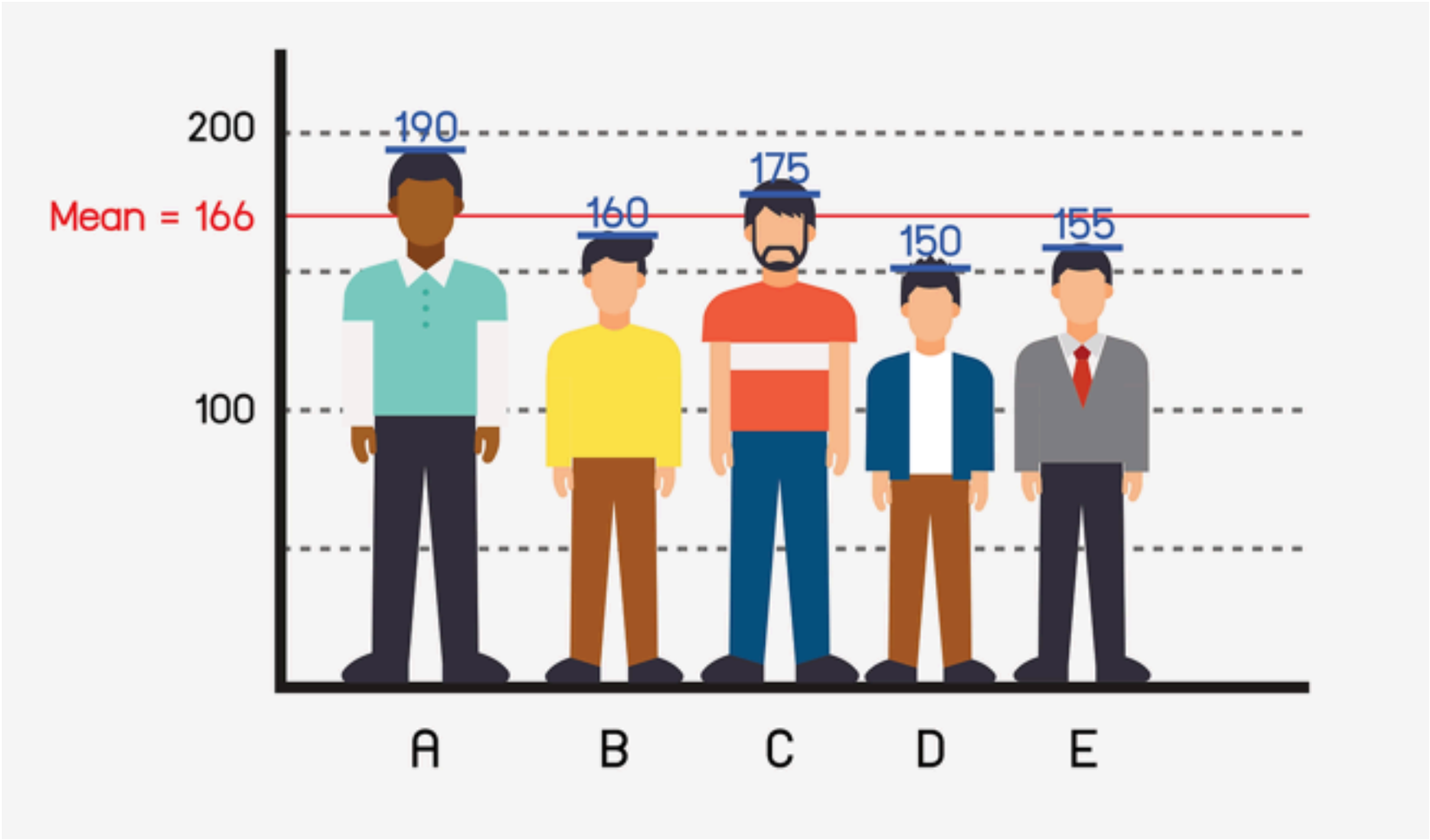
Standard Deviation

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

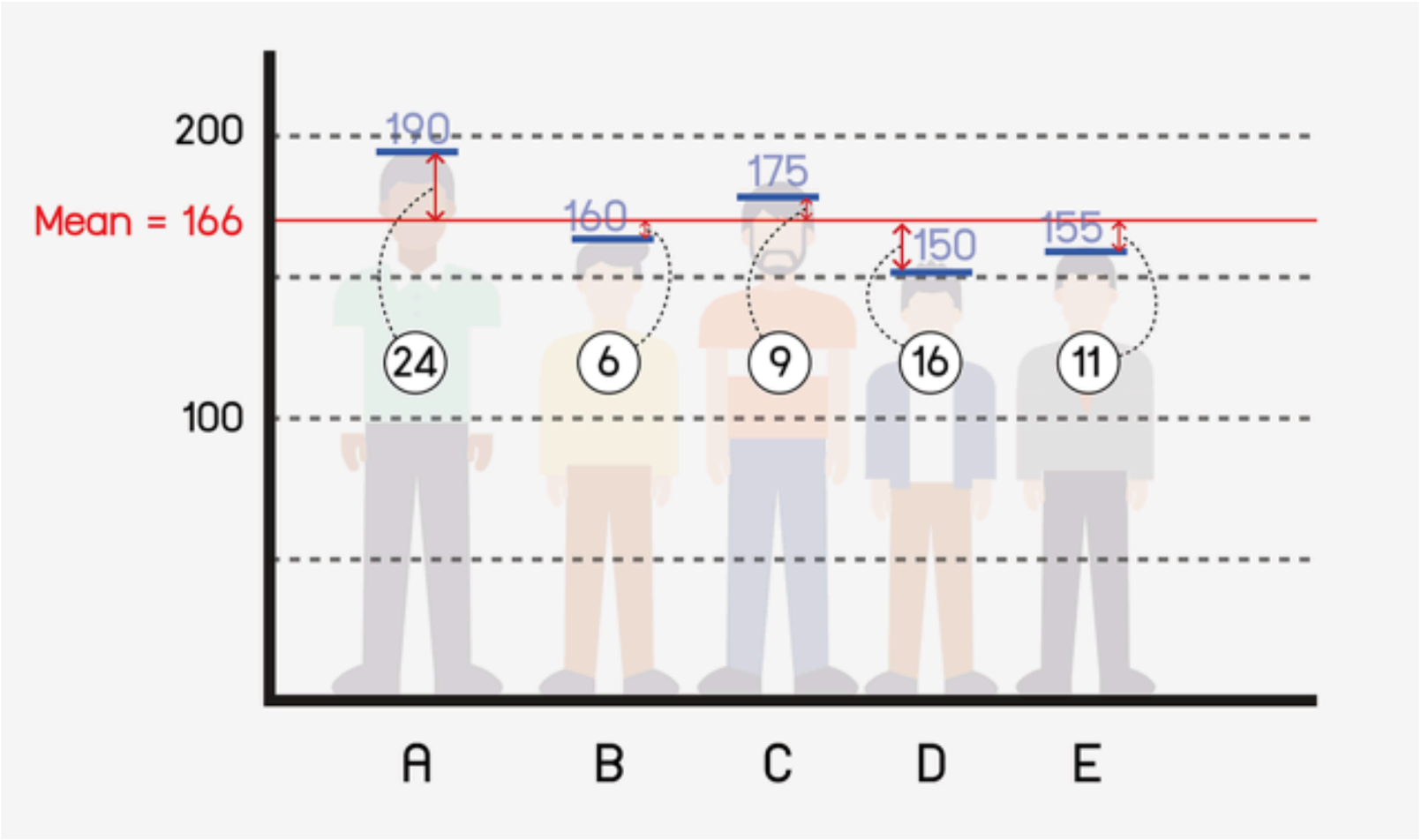
Interquartile Range (IQR)

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

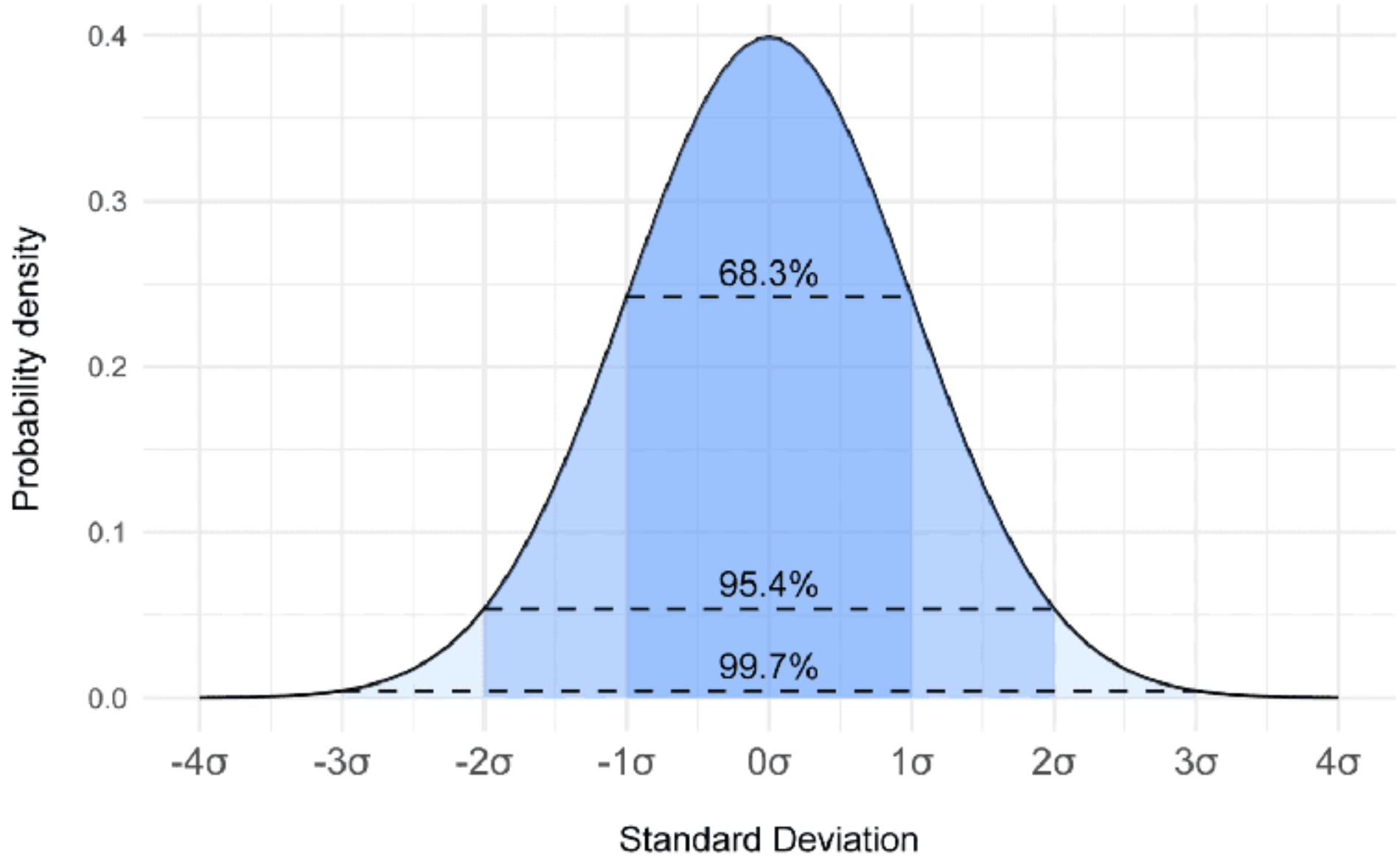
$$\frac{24^2+(-6)^2+9^2+(-16)^2+(-11)^2}{5}=214$$

ค่าเบี่ยงเบนมาตรฐาน คือ
Standard Deviation

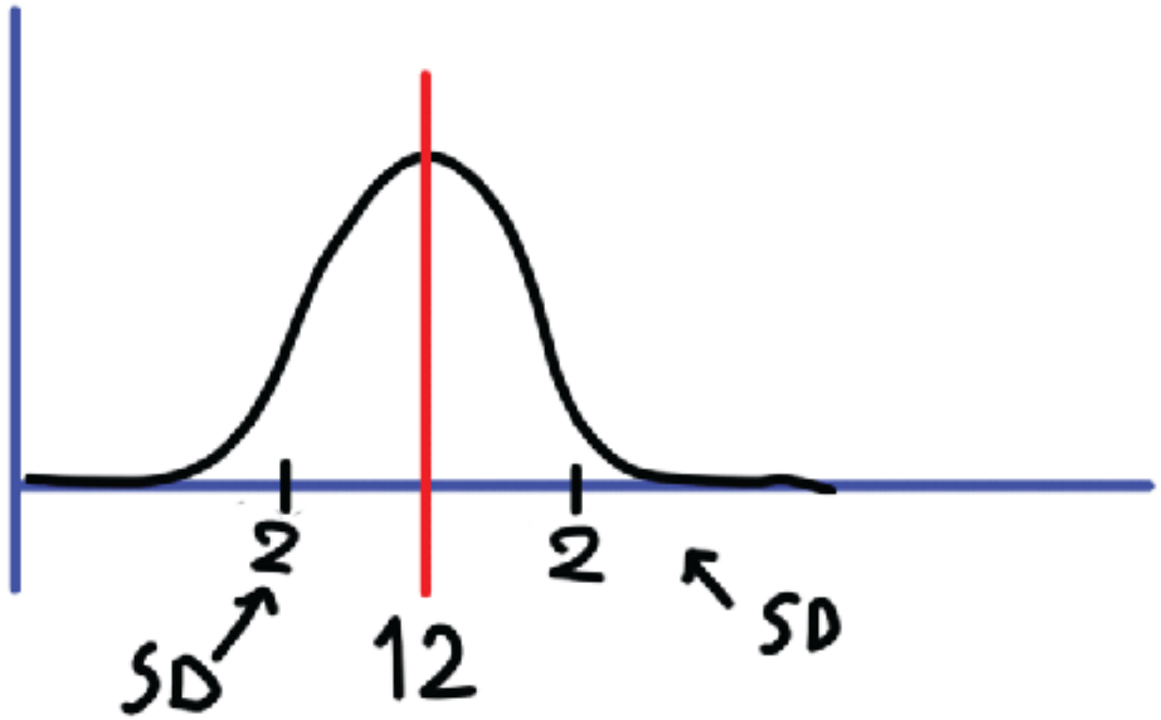
$$\sqrt{214}=14.63$$

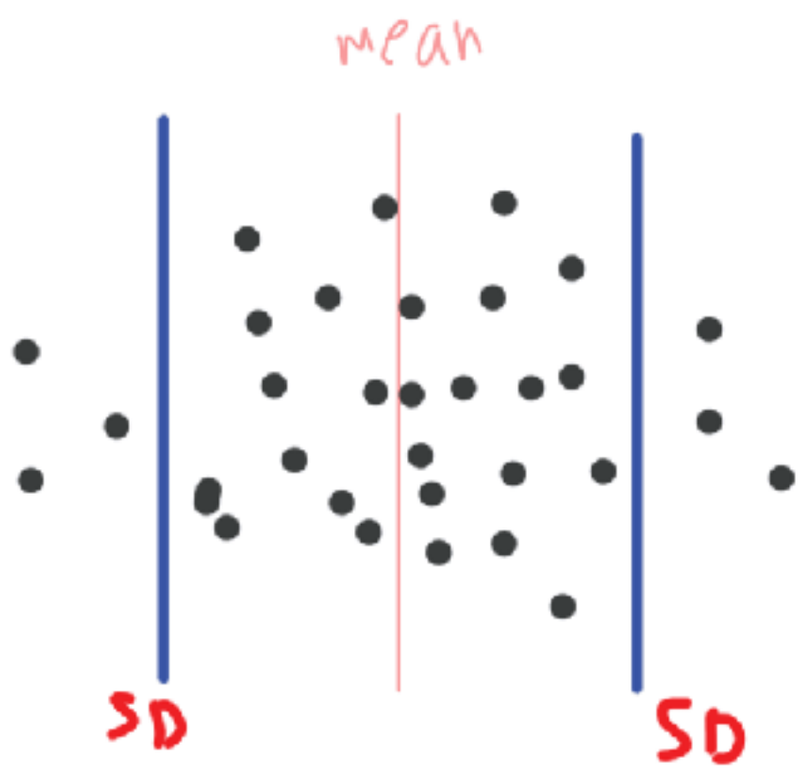
Workshop

Standard Deviation (1)

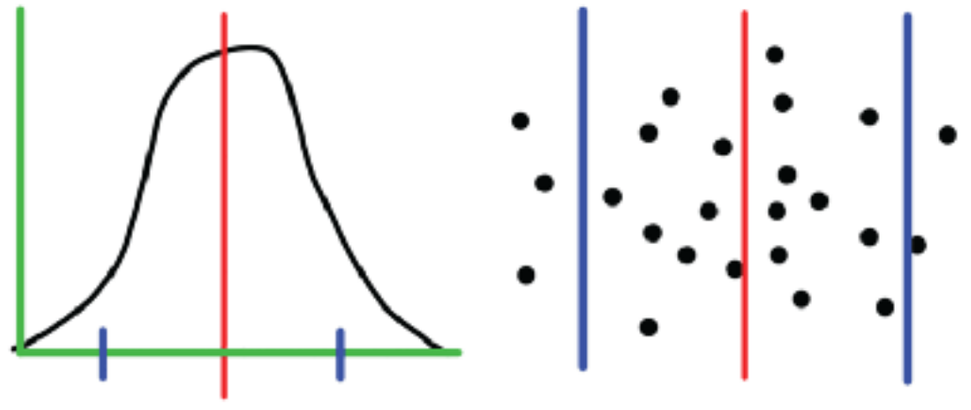


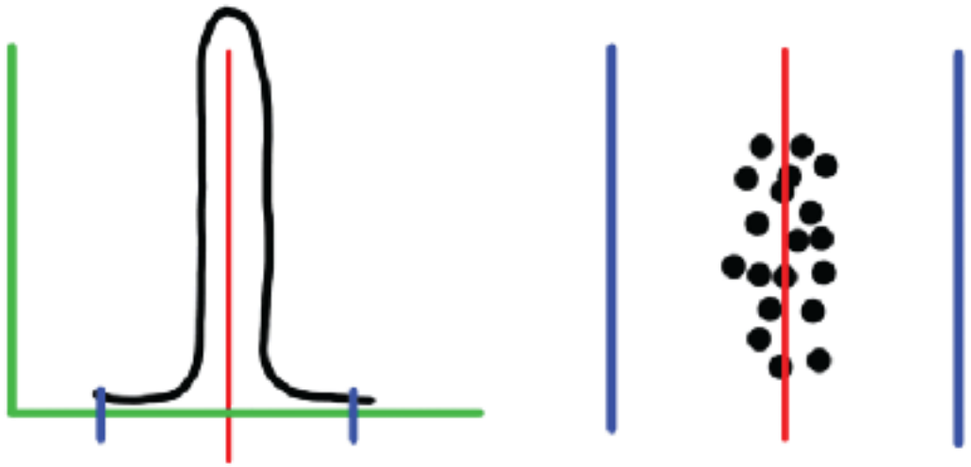
Standard Deviation (2)



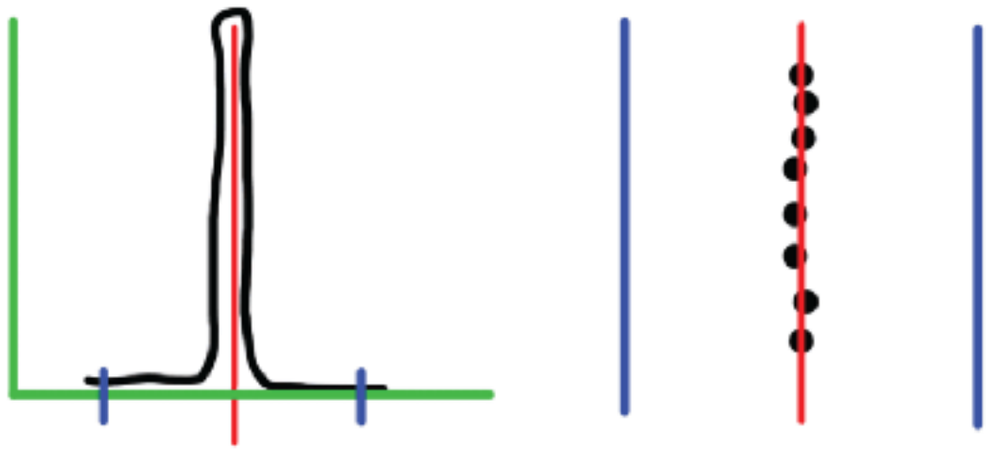


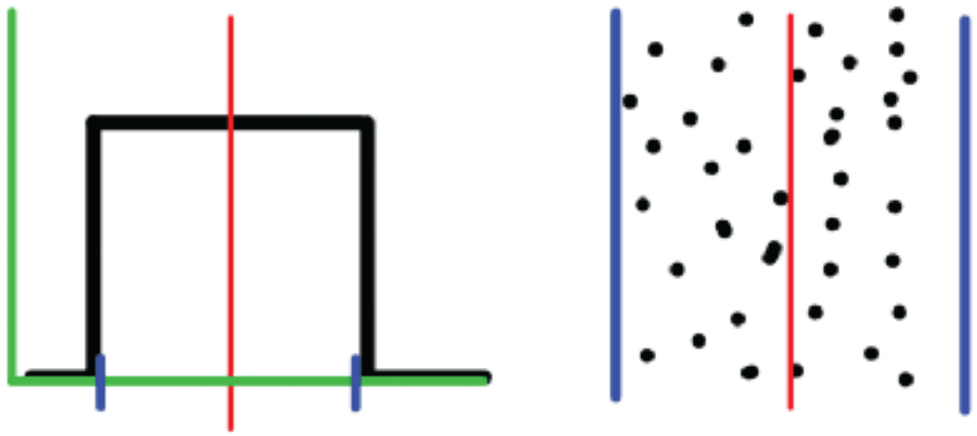
Standard Deviation (3)



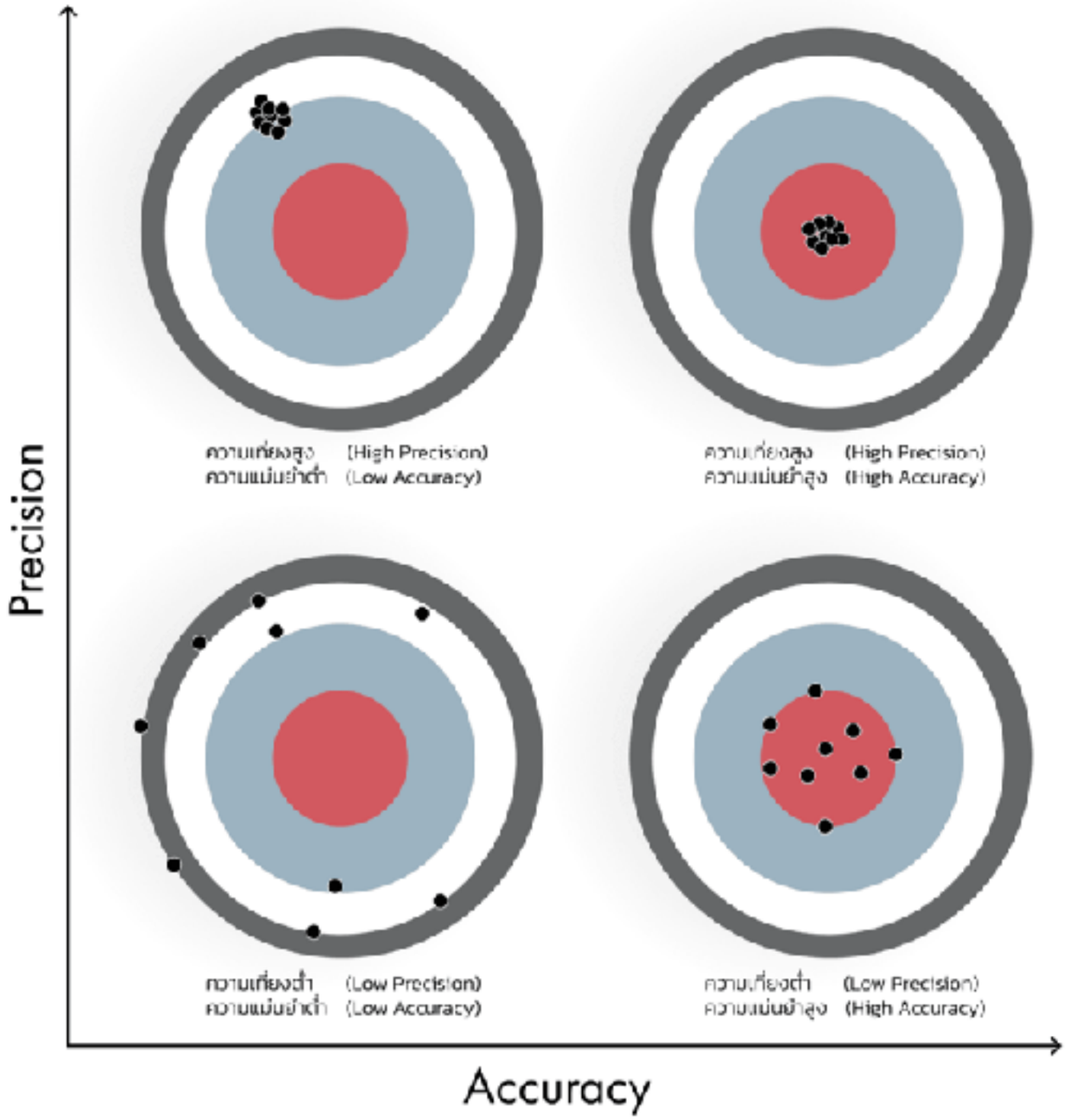


Standard Deviation (4)





Standard Deviation and Precision

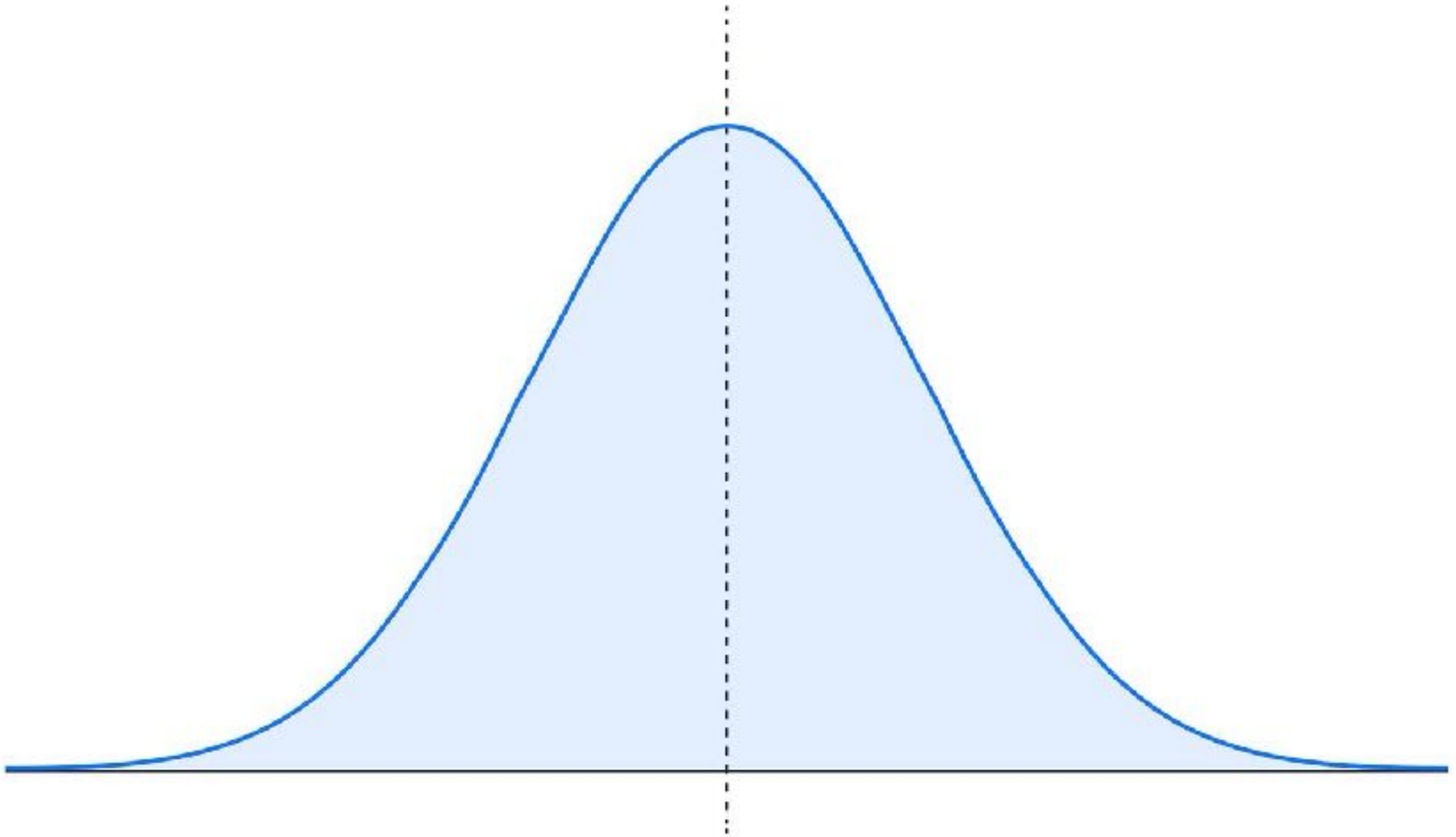


Percentile

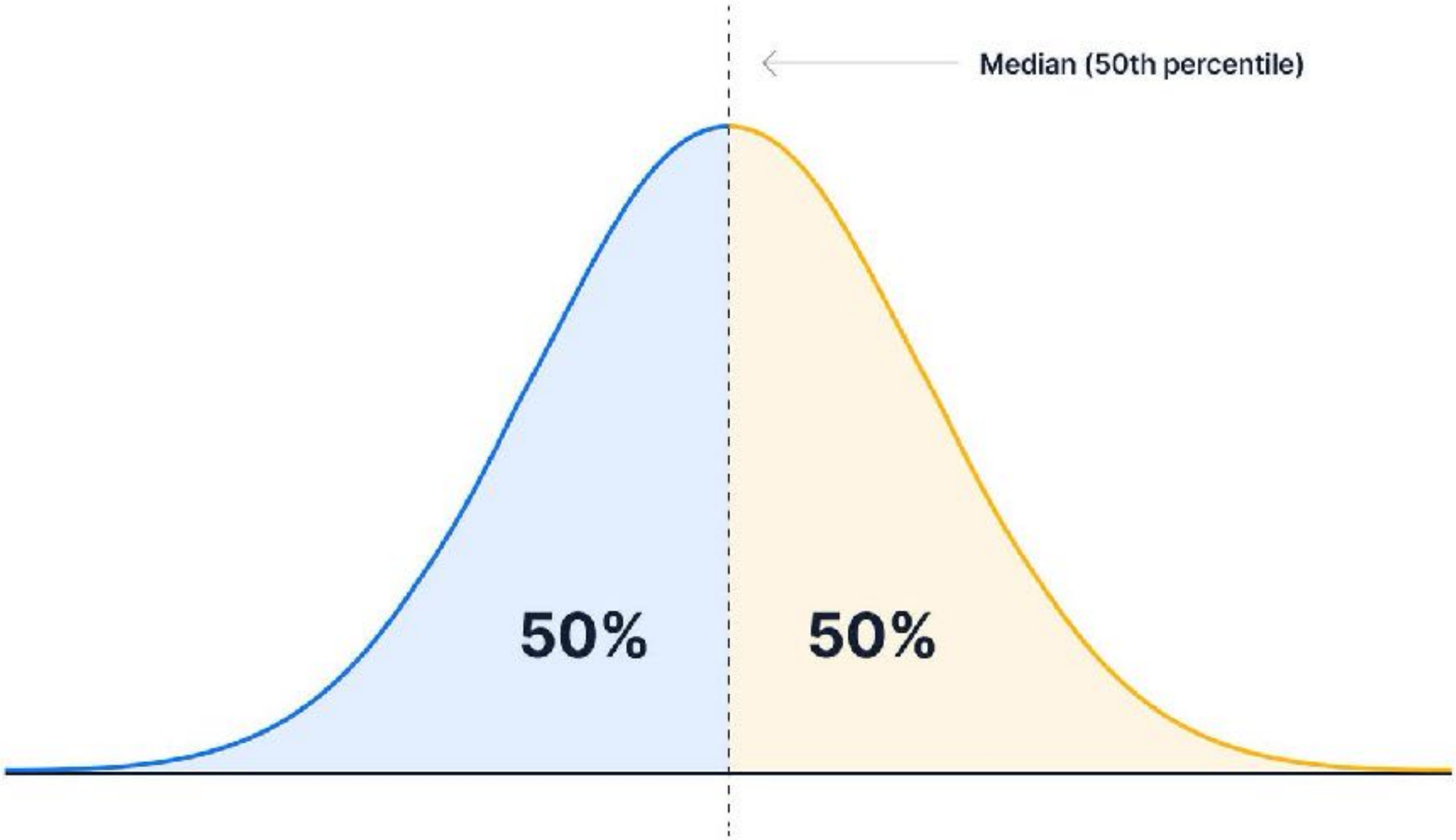
Measure used in statistics to indicate the relative standing of a value within a dataset.

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

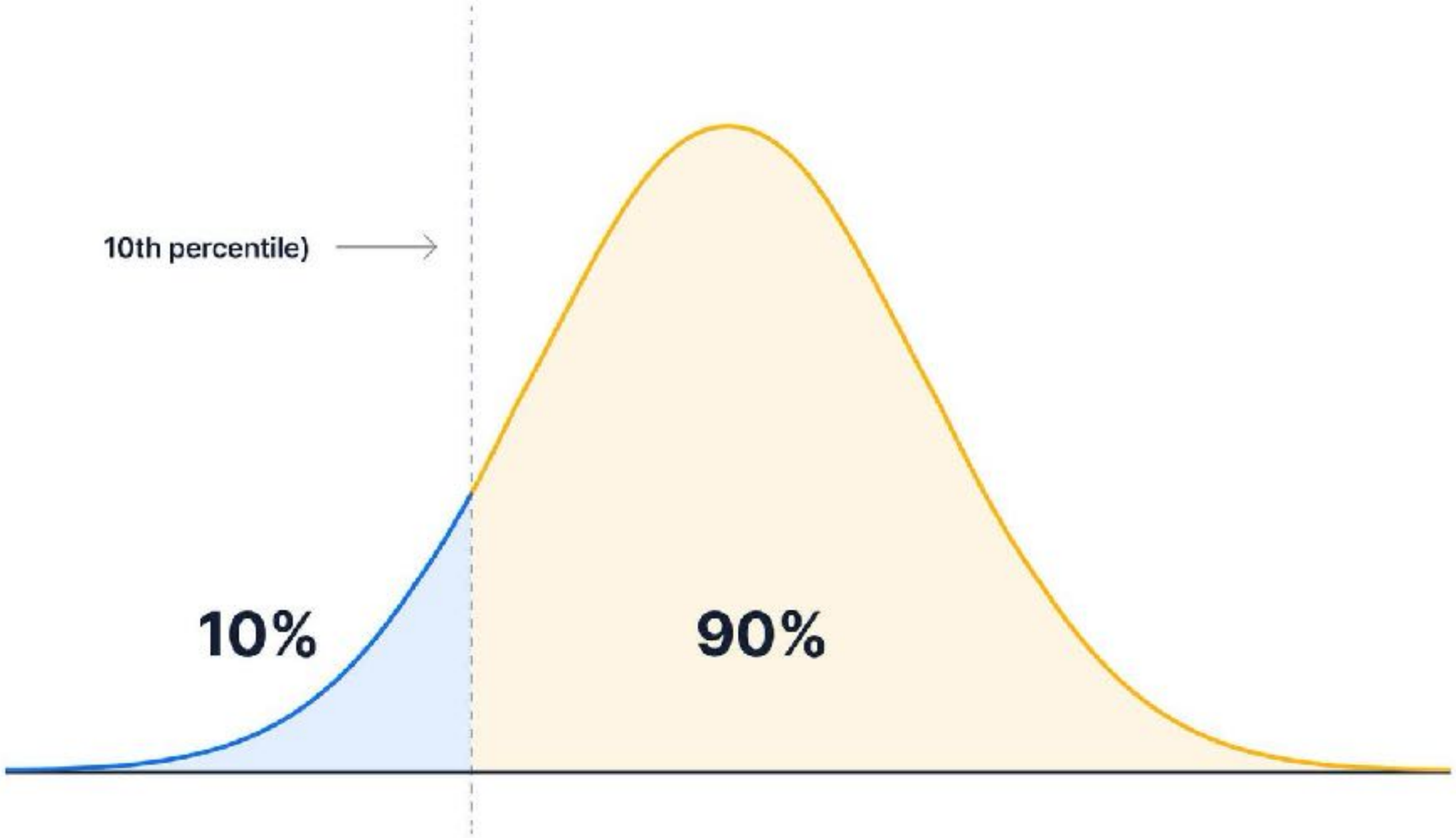
Normal distribution



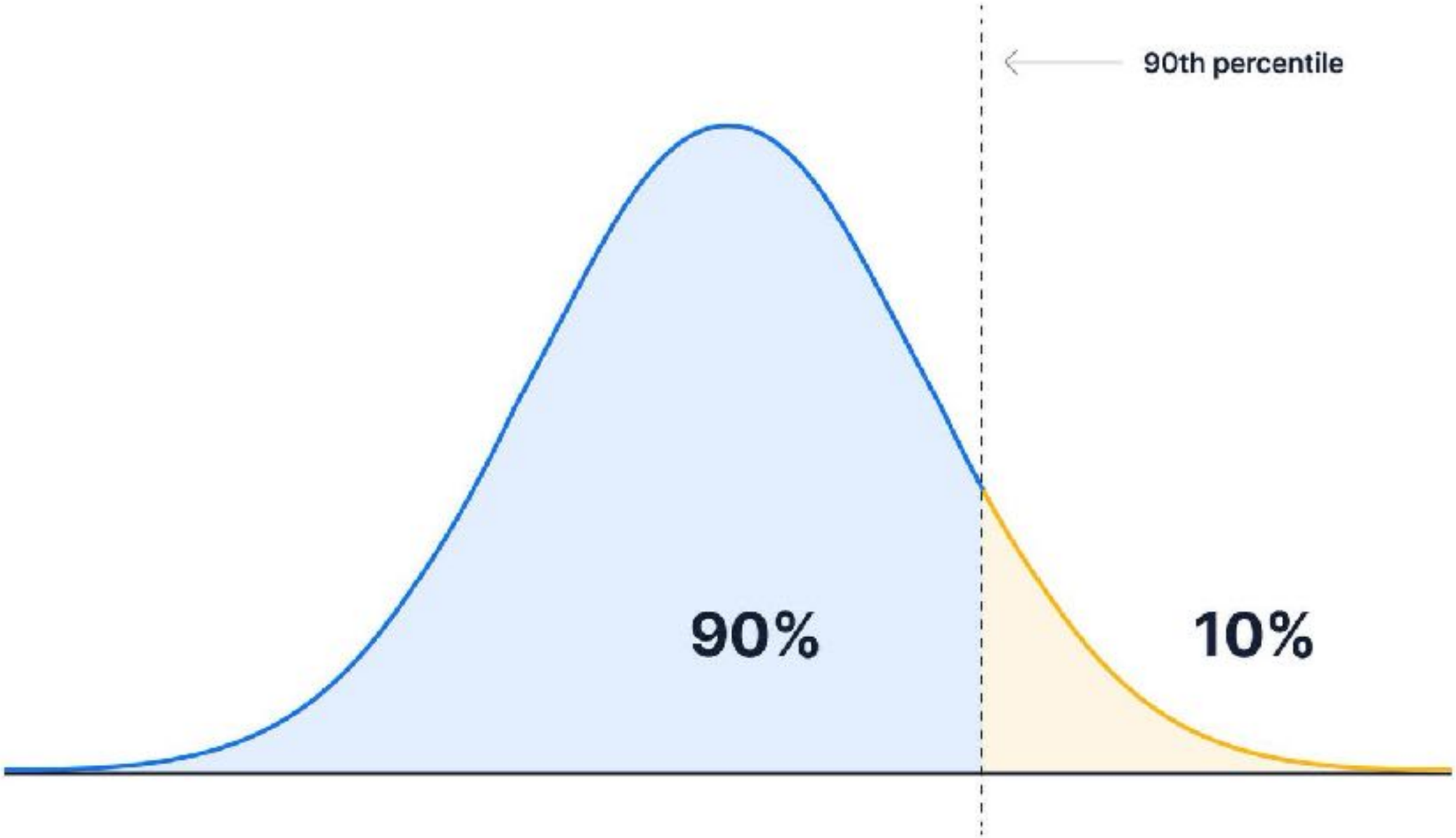
50th percentile



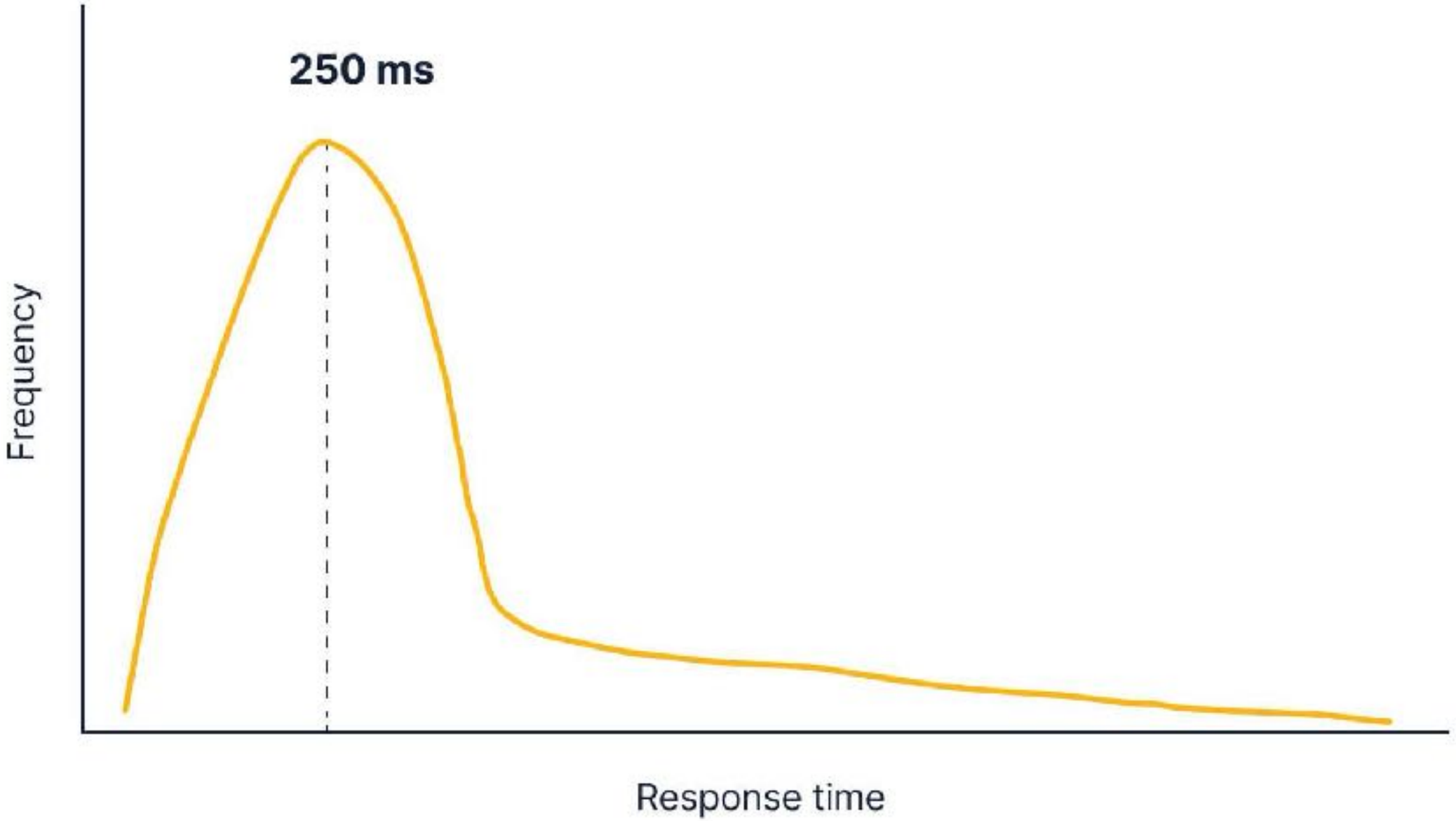
10th percentile



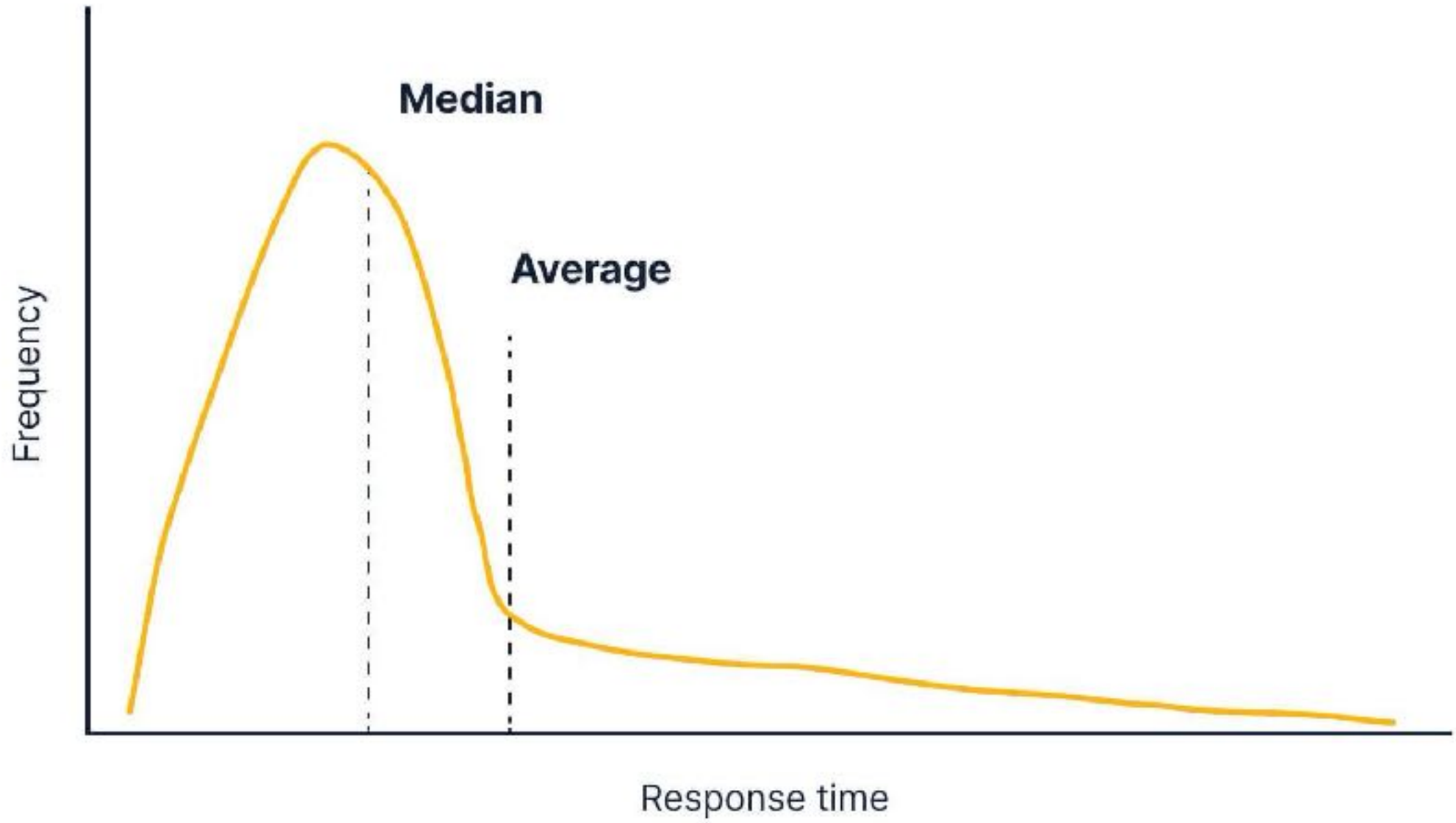
90th percentile



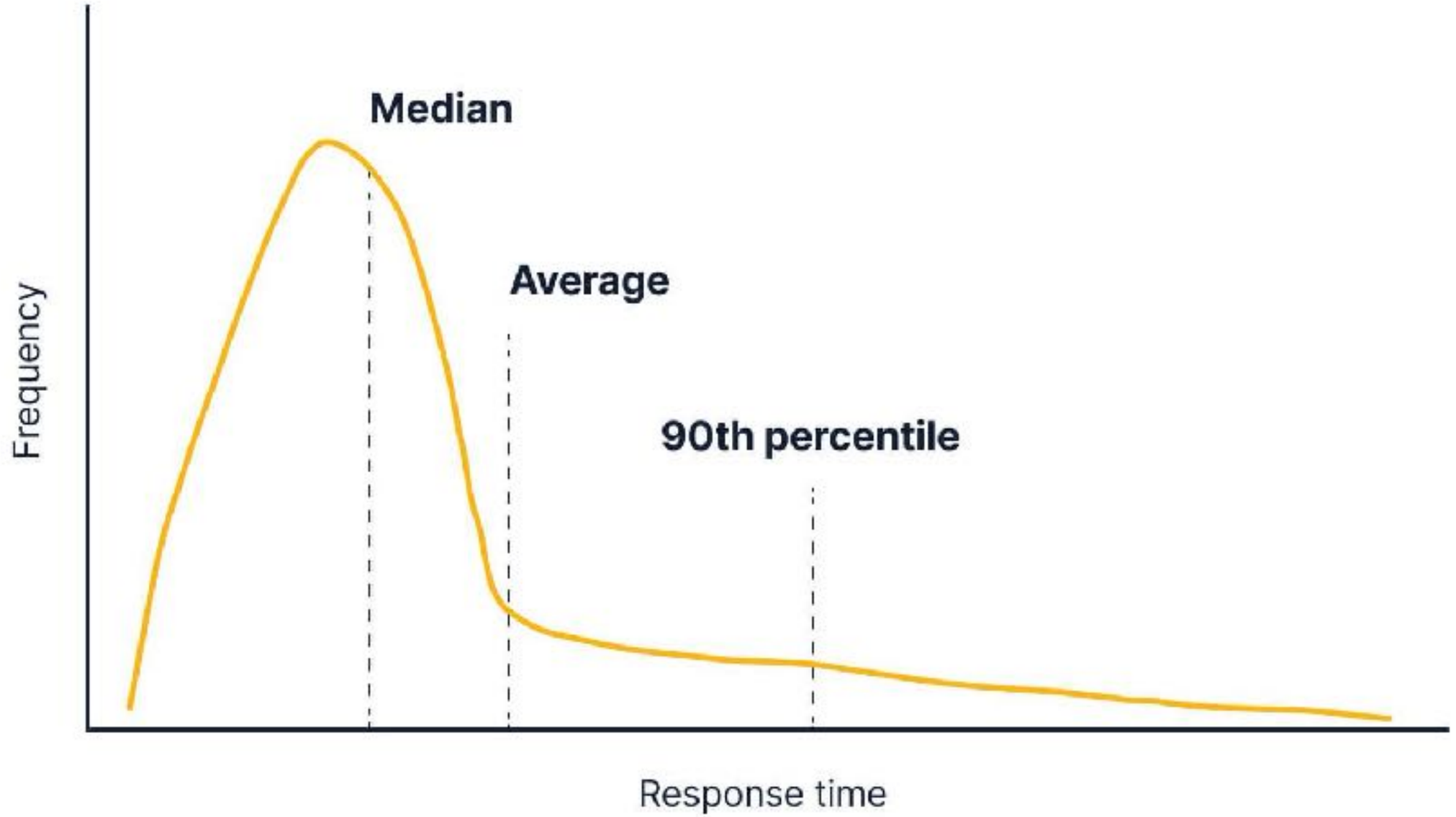
Long tails (1)



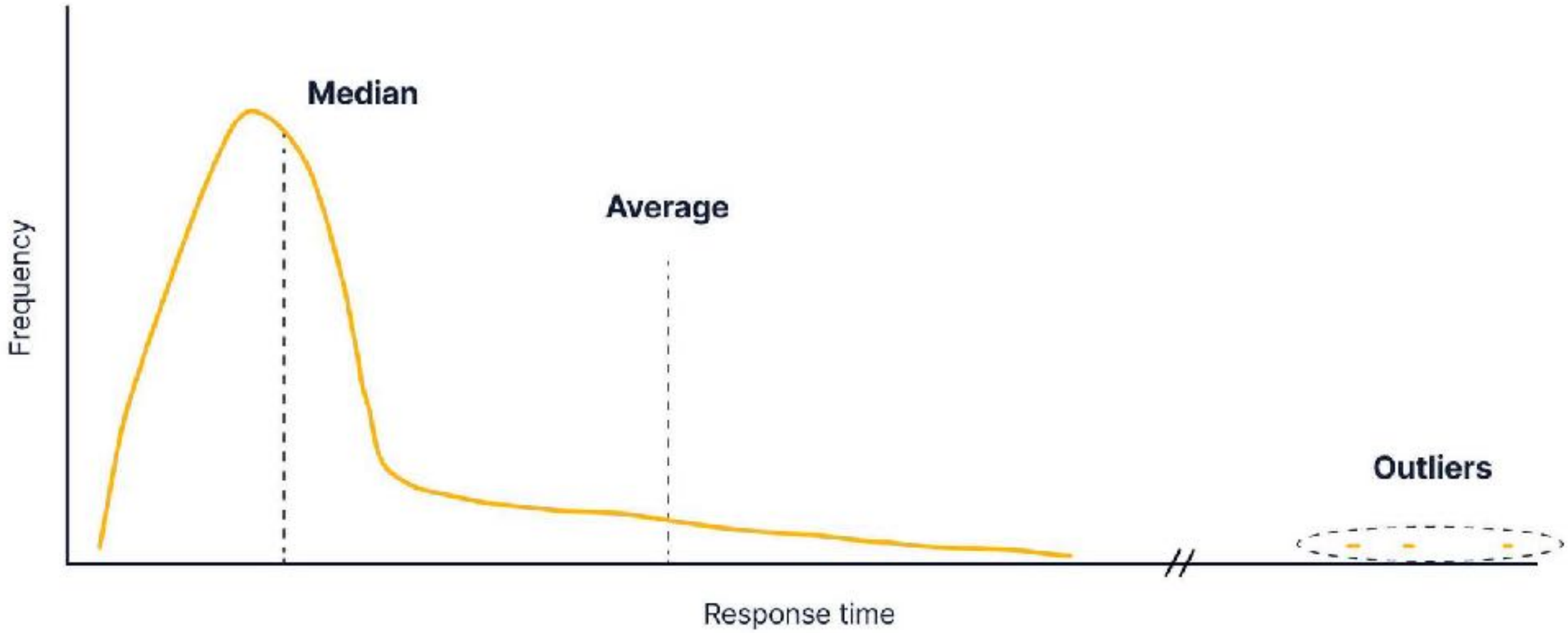
Long tails (2)



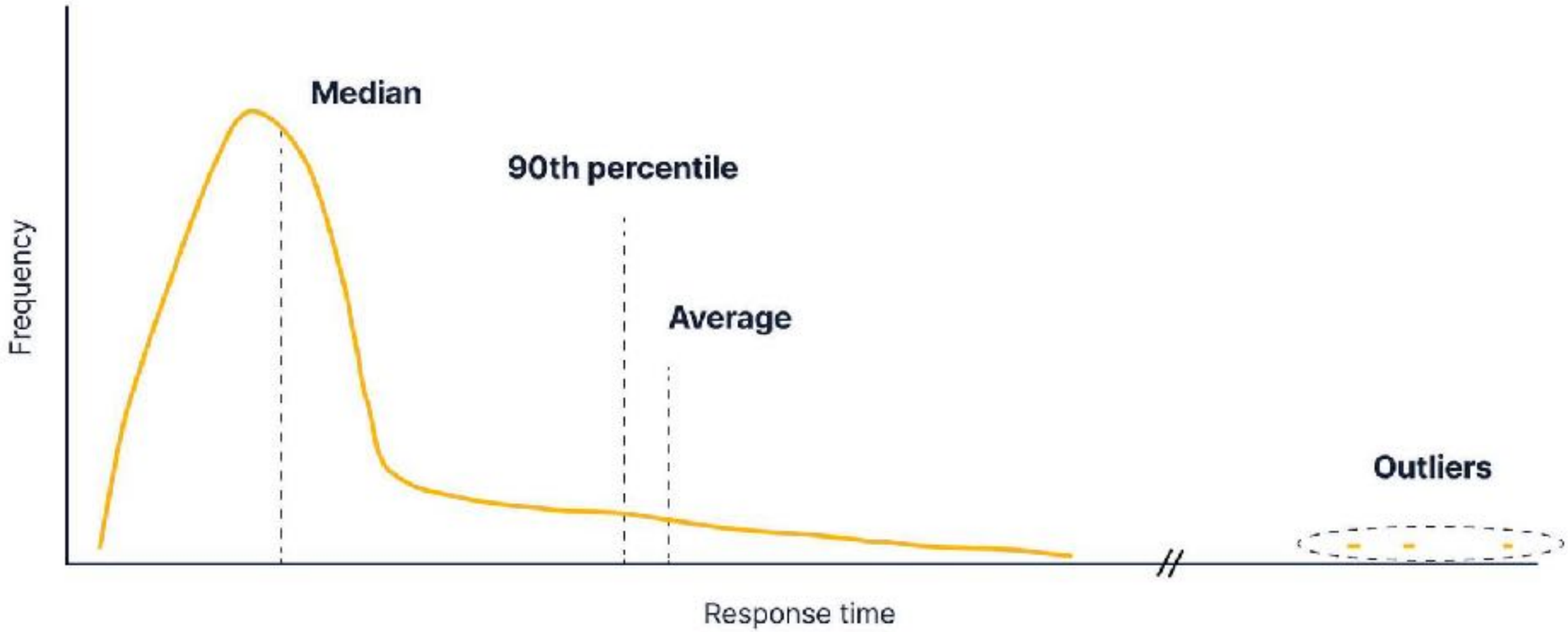
Long tails (2)



Long tails with outliners



Long tails with outliners



Workshop

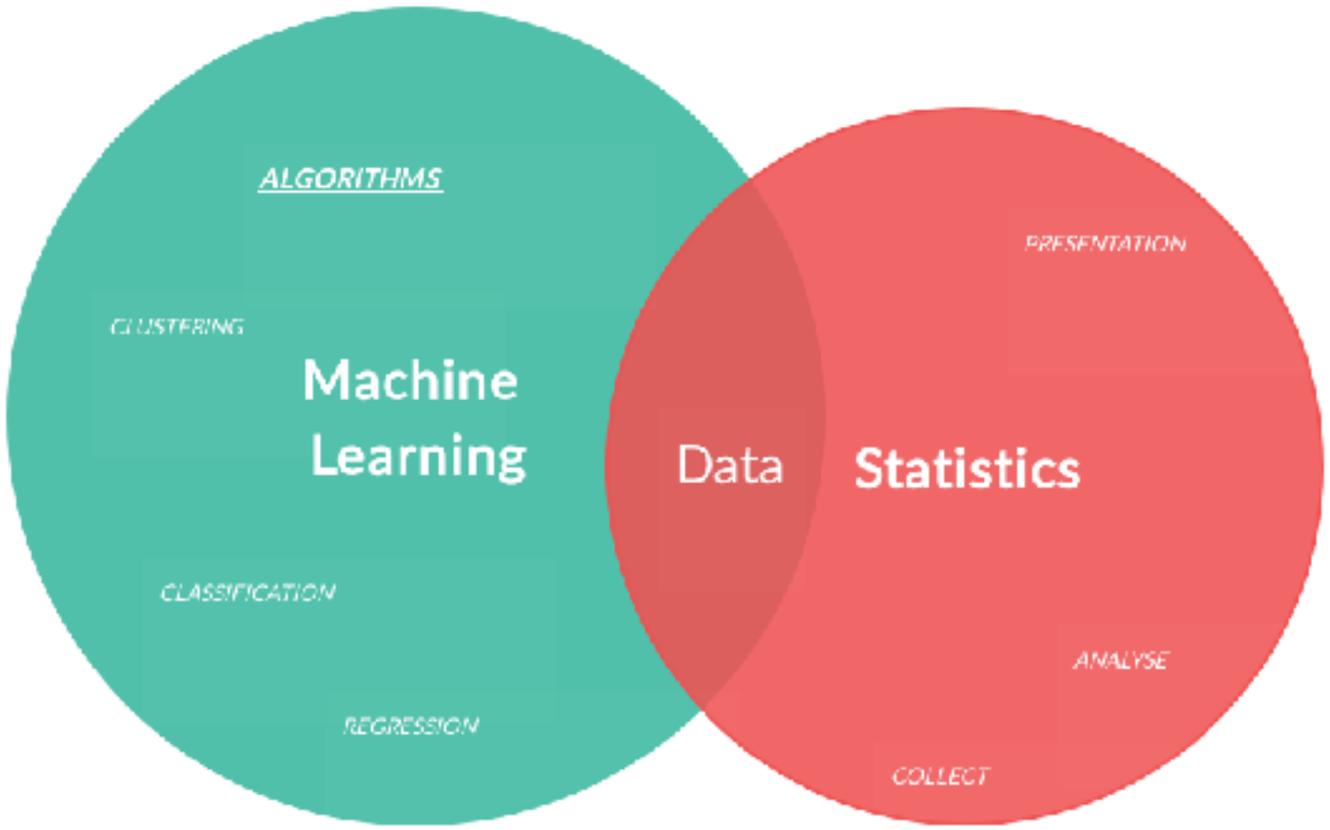
Business Intelligence (BI) and Visualization

Data visualization

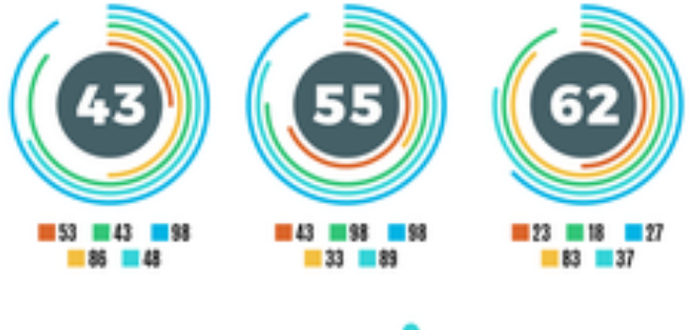
Data visualization is a method that uses **visuals**,
both static and interactive,
to help people **understand**
the large amount of data being collected.

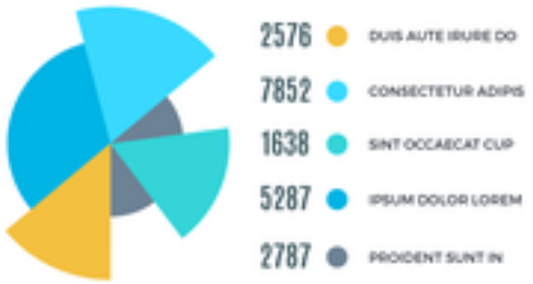
Data visualization

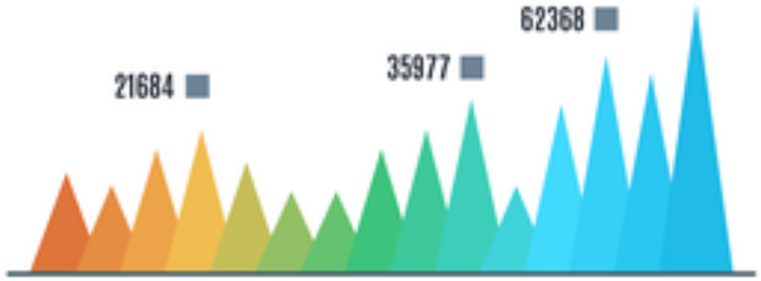
Data visualization is an **important skill** in applied statistics and machine learning.

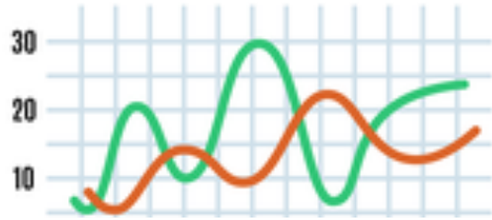


Data visualization

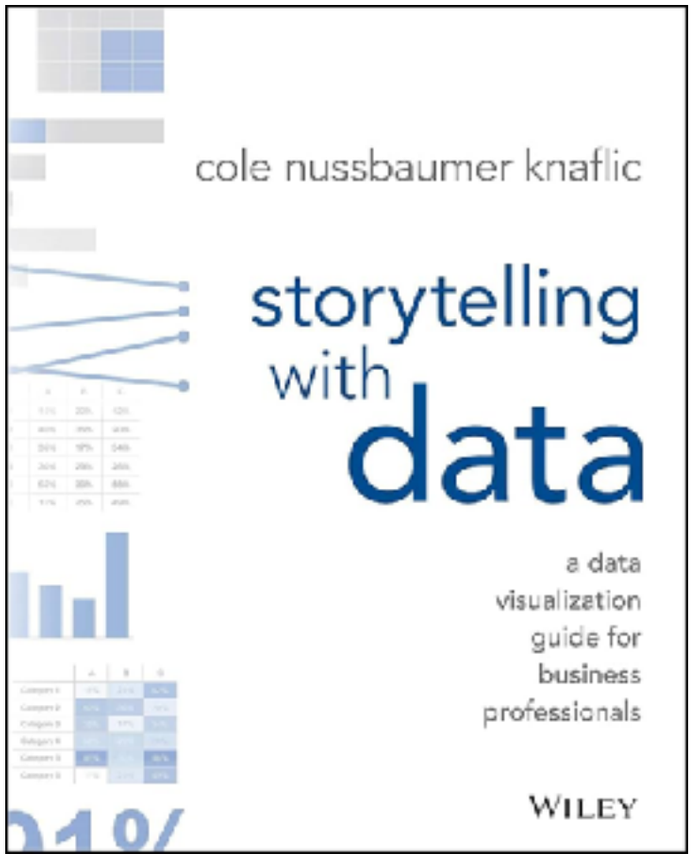


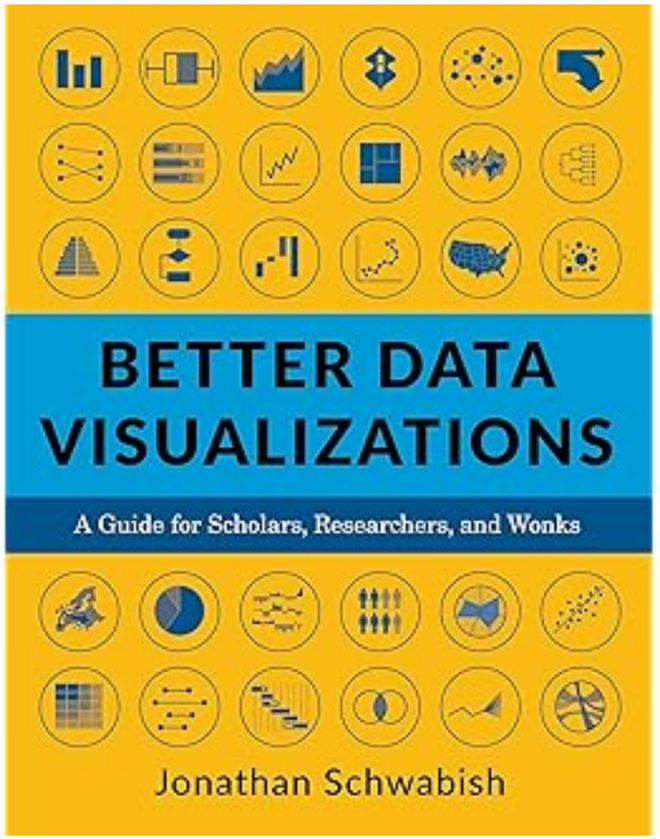


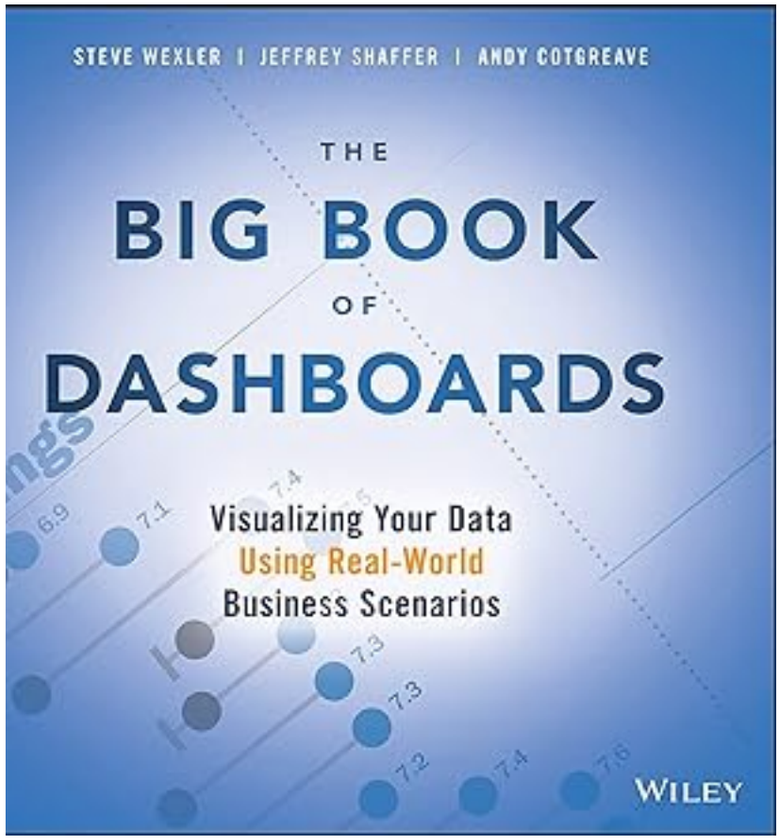




Books







Data visualization process

1
Collecting data

2
Clean you data

3
Choose a chart type

4
Prepare data

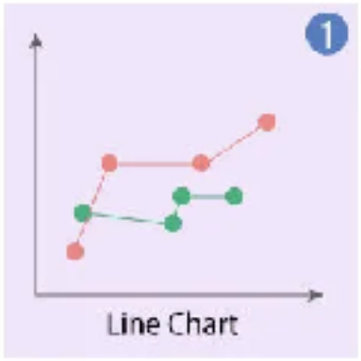
5
Visualize data

6
Presentation

How to choose a chart type ?

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

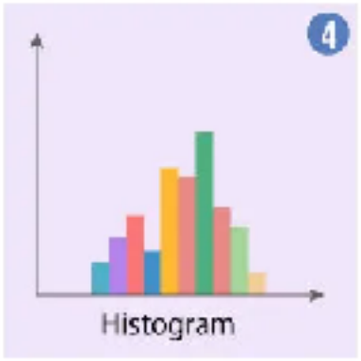
TYPES OF DATA VISUALIZATION CHARTS



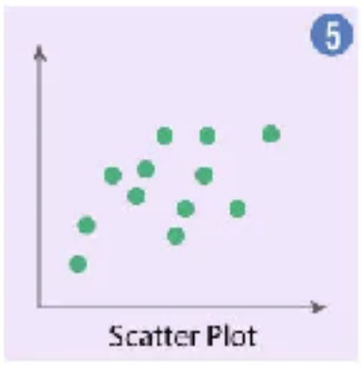
Display trends over time

A line chart with areas below the lines filled with colors

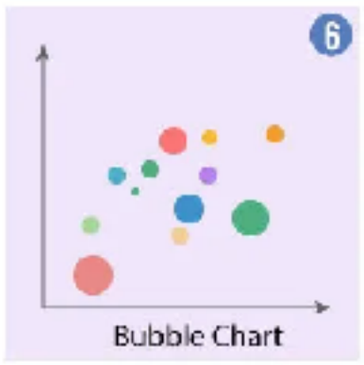
Display trends with multiple variables



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	Insights	Data products
Historical data	Modeling	
เกิดอะไรขึ้นในอดีต ?	จะเกิดอะไรขึ้นในอนาคต ?	จะทำให้สิ่งที่ผู้ใช้งานต้องการเกิดขึ้นได้อย่างไร

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

Workshop #2

Data Transformation

Structure of Data ?

	Country	Year	Type	Count
2	Afghanistan	1999	cases	745
3	Afghanistan	1999	population	19987071
4	Afghanistan	2000	cases	2666
5	Afghanistan	2000	population	20595360
6	Brazil	1999	cases	37737
7	Brazil	1999	population	172006362
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

Country	Year	Case	Population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

	1999		2000	
Country	Case	Population	Case	Population
Afghanistan	745	19987071	2666	20595360
Brazil	37737	172006362	80488	174504898
China	212258	1272915272	213766	1280428583

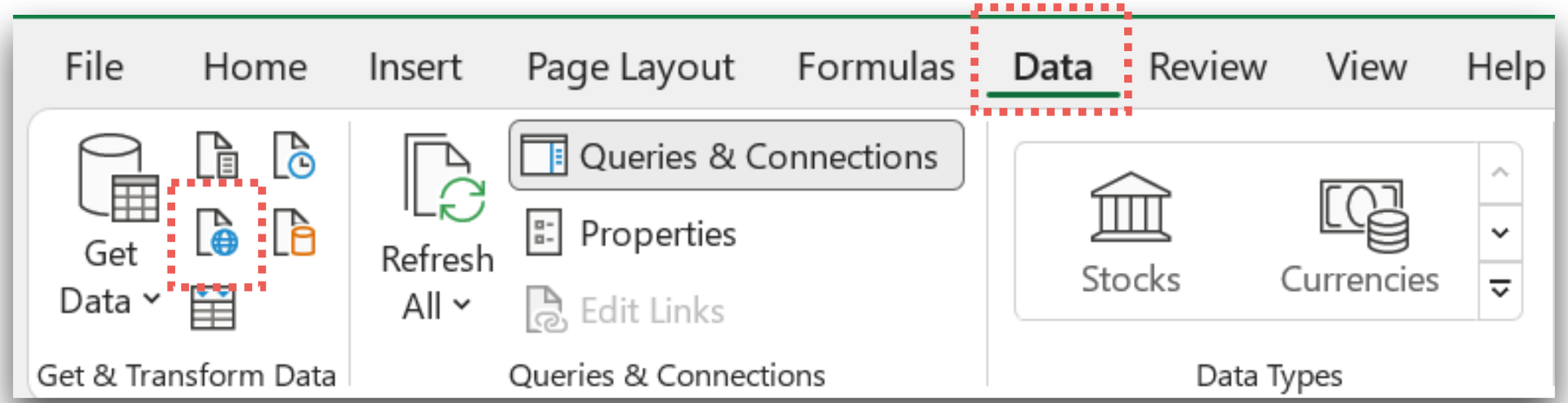
https://github.com/up1/course-basic-big-data-analytic/blob/main/workshop/data03.xlsx

Workshop #3

Import data from web

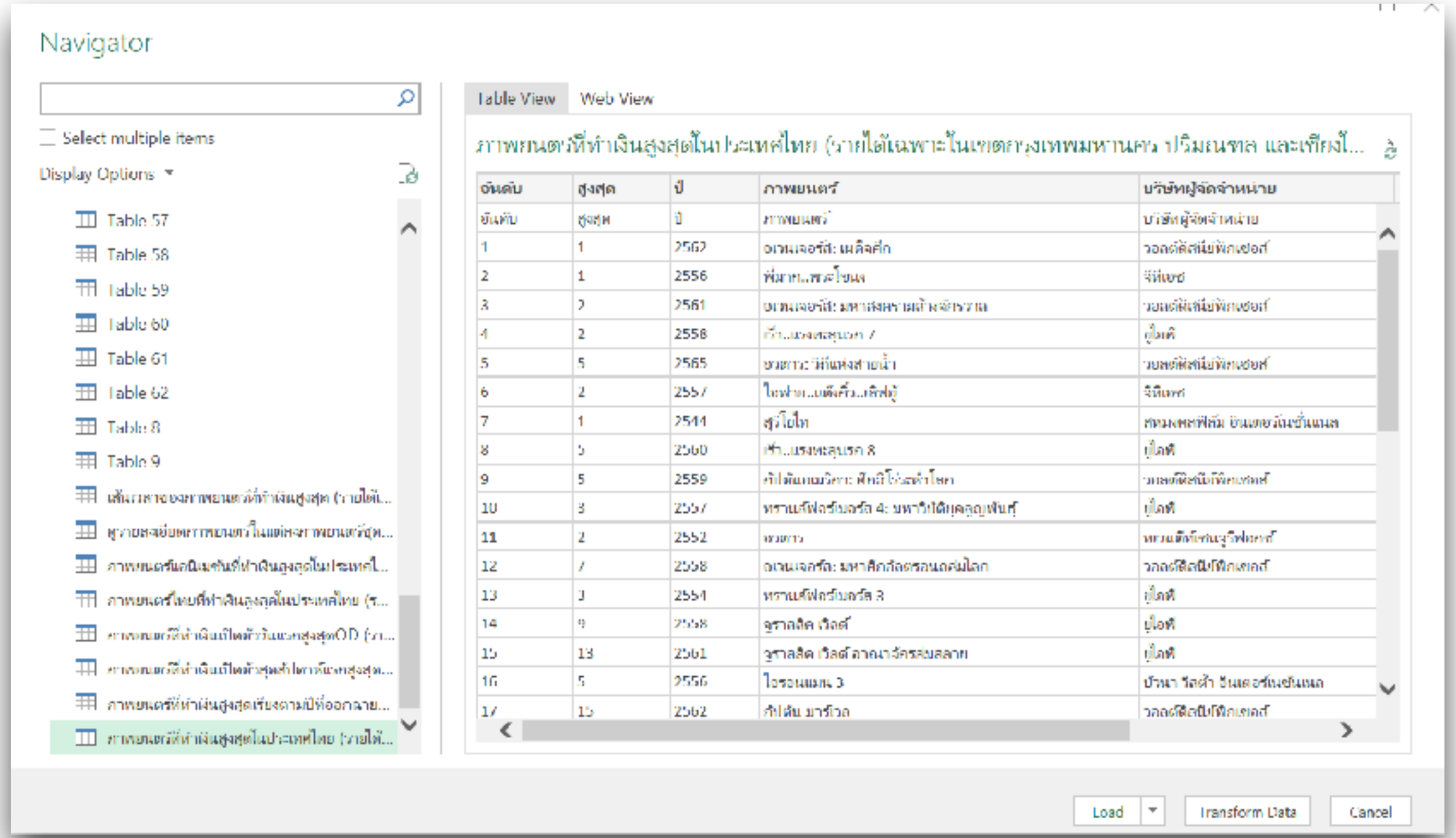
Import data from web

Goto menu **Data**



Import data from web

Choose table data and load to Excel



Example Data

	A	B	C	D	E	F	G
1	อันดับ	สูงสุด	ปี	ภาพยนตร์	บริษัทผู้จัดจำหน่าย	ทำเงิน (ล้านบาท)	อ้างอิ
2	อันดับ	สูงสุด	ปี	ภาพยนตร์	บริษัทผู้จัดจำหน่าย	ทำเงิน(ล้านบาท)	อ้างอิ
3	1	1	2562	อเวนเจอร์ส: เผด็จศึก	วอลต์ดิสนีย์พิคเชอส์	617.55	
4	2	1	2556	พี่มาก..พระโขนง	จีทีเอช	568.55	
5	3	2	2561	อเวนเจอร์ส: มหาสงครามล้างจักรวาล	วอลต์ดิสนีย์พิคเชอส์	420.89	
6	4	2	2558	เร็ว..แรงทะลุนรก 7	ยูไอพี	387.85	
7	5	5	2565	อวตาร: วิถีแห่งสายน้ำ	วอลต์ดิสนีย์พิคเชอส์	376.23	
8	6	2	2557	ไอฟาย..แต๋งกิ้ว..เสี่ยฟู่	จีทีเอช	330.55	
9	7	1	2544	สุริโยไท	สหมงคลฟิล์ม อินเตอร์เนชั่นแนล	324.5	
10	8	5	2560	เร็ว..แรงทะลุนรก 8	ยูไอพี	321.96	
11	9	5	2559	คัปตันอเมริกา: ศึกฮีโร่ระห่ำโลก	วอลต์ดิสนีย์พิคเชอส์	311.21	
12	10	3	2557	ทรานส์ฟอร์เมอร์ส 4: มหาวิบัติยุคสูญพันธุ์	ยูไอพี	310.23	
13	11	2	2552	อวตาร	ทเวนตีทเซนจูรีฟอกซ์	303	
14	12	7	2558	อเวนเจอร์ส: มหาศึกอัลตรอนถล่มโลก	วอลต์ดิสนีย์พิคเชอส์	294.77	
15	13	3	2554	ทรานส์ฟอร์เมอร์ส 3	ยูไอพี	291.20	
16	14	9	2558	จูราสสิค เวิลด์	ยูไอพี	290.64	
17	15	13	2561	จูราสสิค เวิลด์ อาณาจักรล่มสลาย	ยูไอพี	288.88	
18	16	5	2556	ไอรอนแมน 3	บัวนา วิสต้า อินเตอร์เนชันเนล	262.92	
19	17	15	2562	กัปตัน มาร์เวล	วอลต์ดิสนีย์พิคเชอส์	260.78	
20	18	15	2561	อควาแมน เจ้าสมุทร	วอร์เนอร์บราเธอร์สพิคเจอร์ส	254.09	
21	19	4	2555	ดิ อเวนเจอร์ส	บัวนา วิสต้า อินเตอร์เนชันเนล	252.16	
22	20	20	2567	ซี่หยด 2 †	เอ็ม สตูดิโอ	245.93	
23	21	20	2566	สัปเหร่อ	ไทบ้าน สตูดิโอ	245.10	
24	22	19	2564	สไปเดอร์แมน: โน เวย์ โฮม	โซนี่พิคเจอร์สรีลิสชิง	239.29	
25	23	19	2562	สไปเดอร์แมน: ฟาร์ ฟรอม โฮม	โซนี่พิคเจอร์สรีลิสชิง	238.99	

ภาพยนตร์ที่ทำเงินสูงสุดในประเทศ | Sheet1 | +

Ready

Queries & Connections

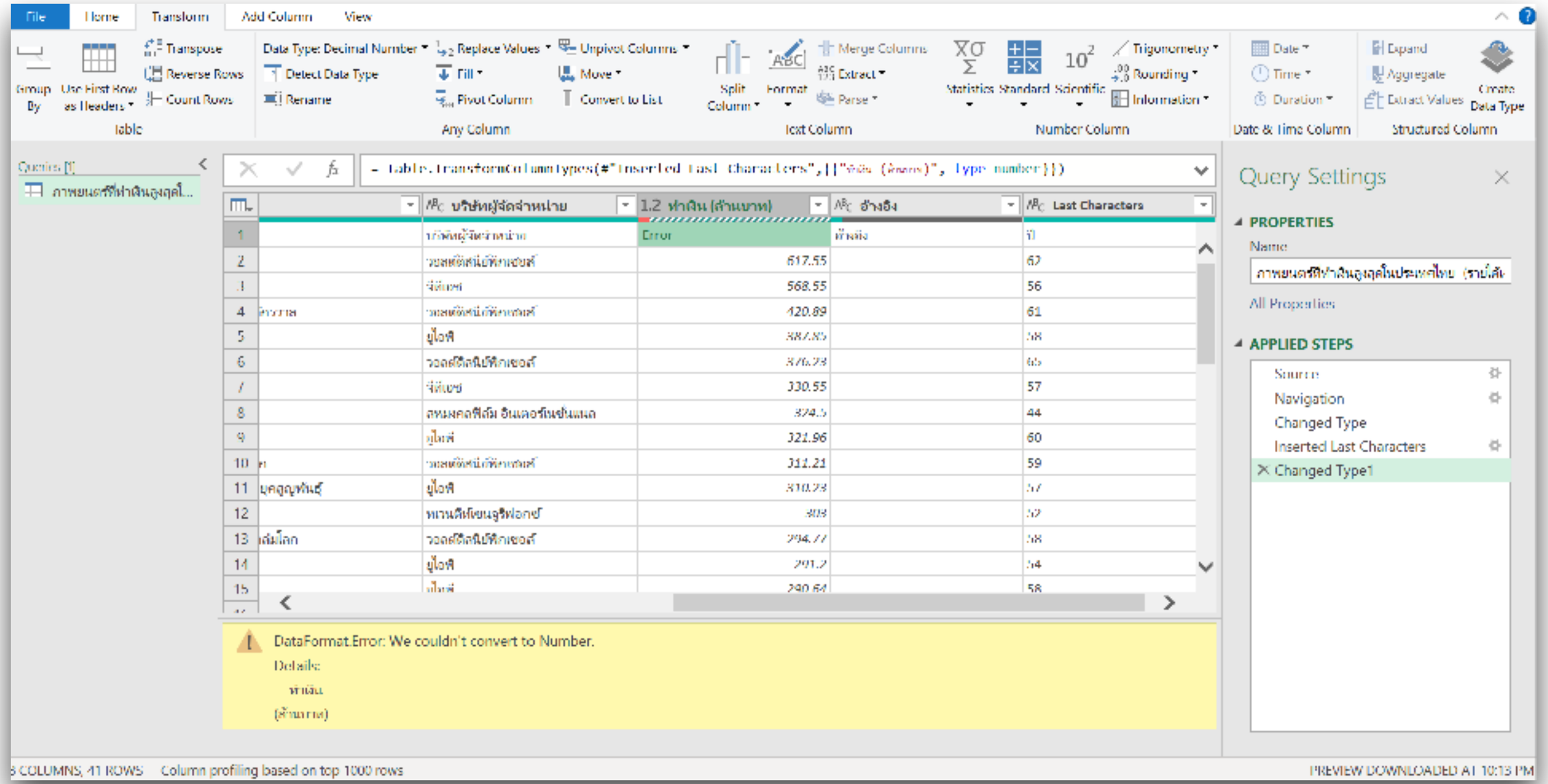
Queries | Connections

1 query

ภาพยนตร์ที่ทำเงินสูงสุดในประเทศไทย (ราย...

41 rows loaded.

Power Query



Excel Table

Select any cells and press **Ctrl + t**

Or goto **Insert -> Table**

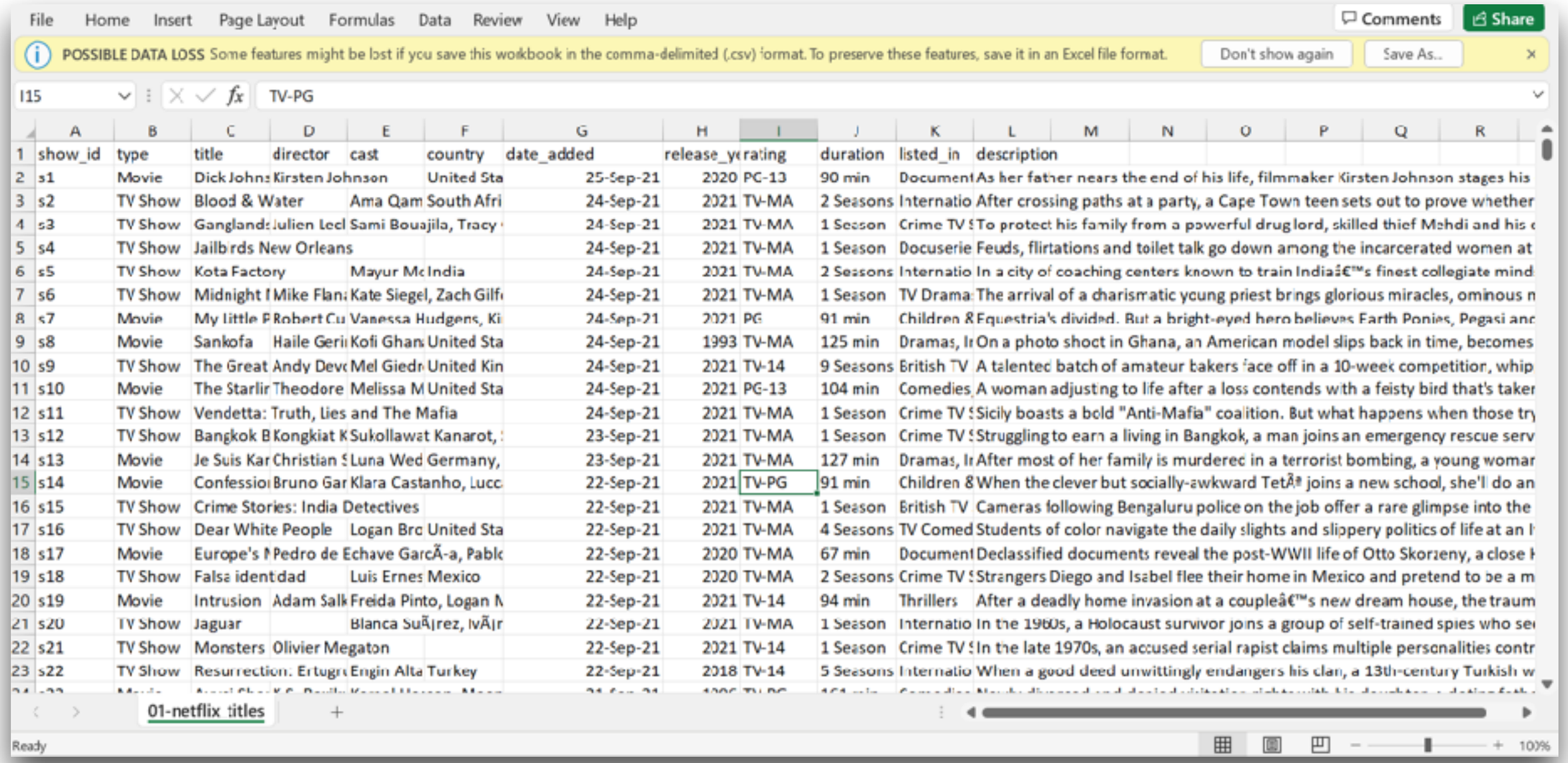
Sorting	Filter	Add formula
Total	Table design	

Excel Table

Workshop #3

01-netflix_titles

Open Data in Excel



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

Step 1

Observe your data !!

# of rows	# of columns	Column description

Exploratory Data Analysis (EDA)

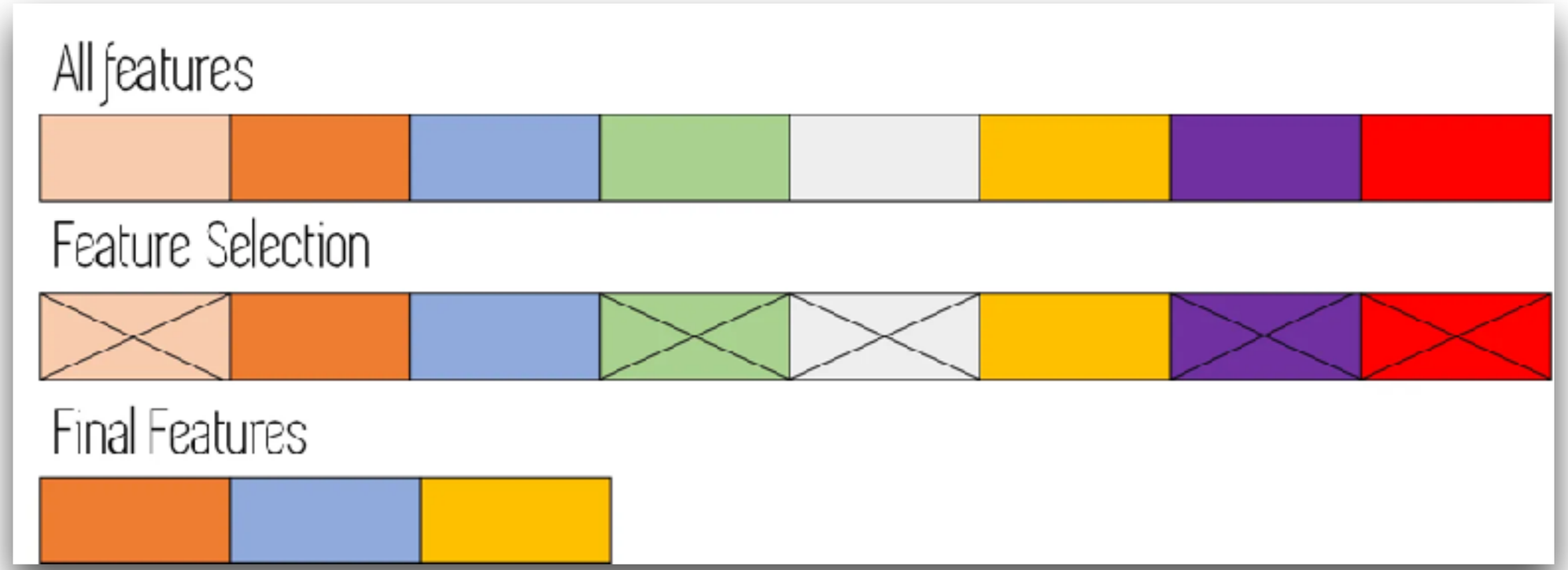
Data transformation

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 ?

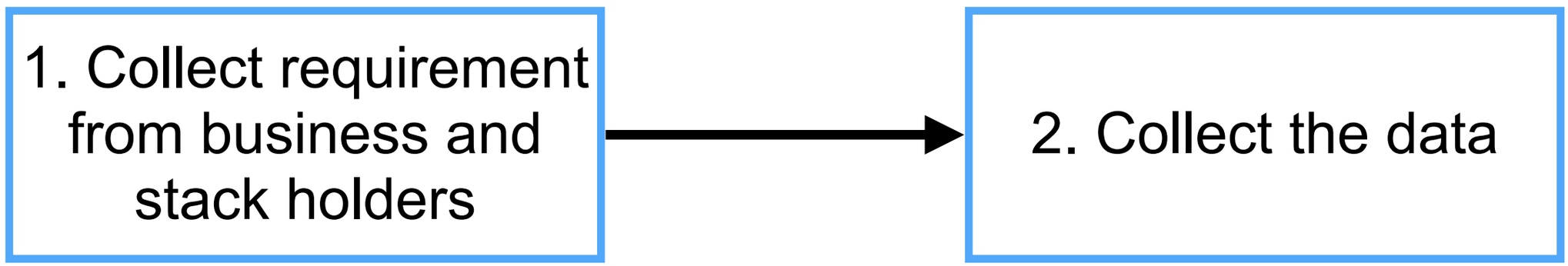
Top 10 director has cast more movie, tv show ?

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 ?

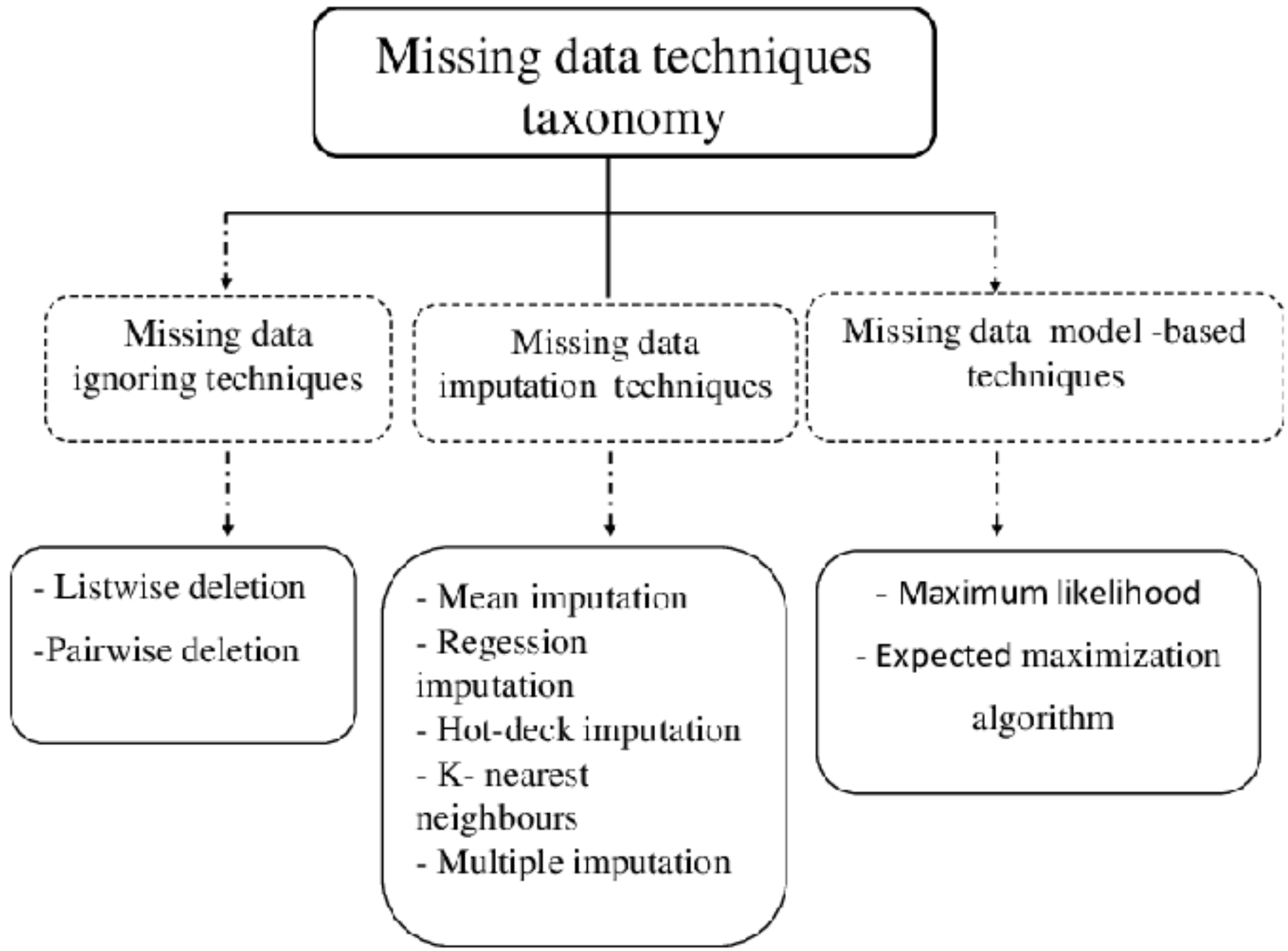
Steps in data analytic



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	8807
3	type	8807
4	title	8807
5	director	6173
6	cast	7982
7	country	7976
8	date_added	8797
9	release_year	8807
10	rating	8803
11	duration	8804
12	listed_in	8807

Working with Excel #3

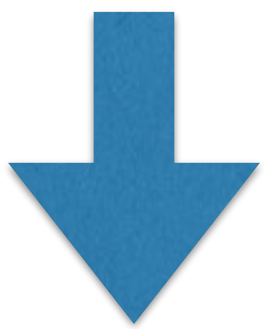
Duplicated data in each columns

Listed_in	Director
Cast	Country

Working with Excel #3

Sample with **listed_in** column

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



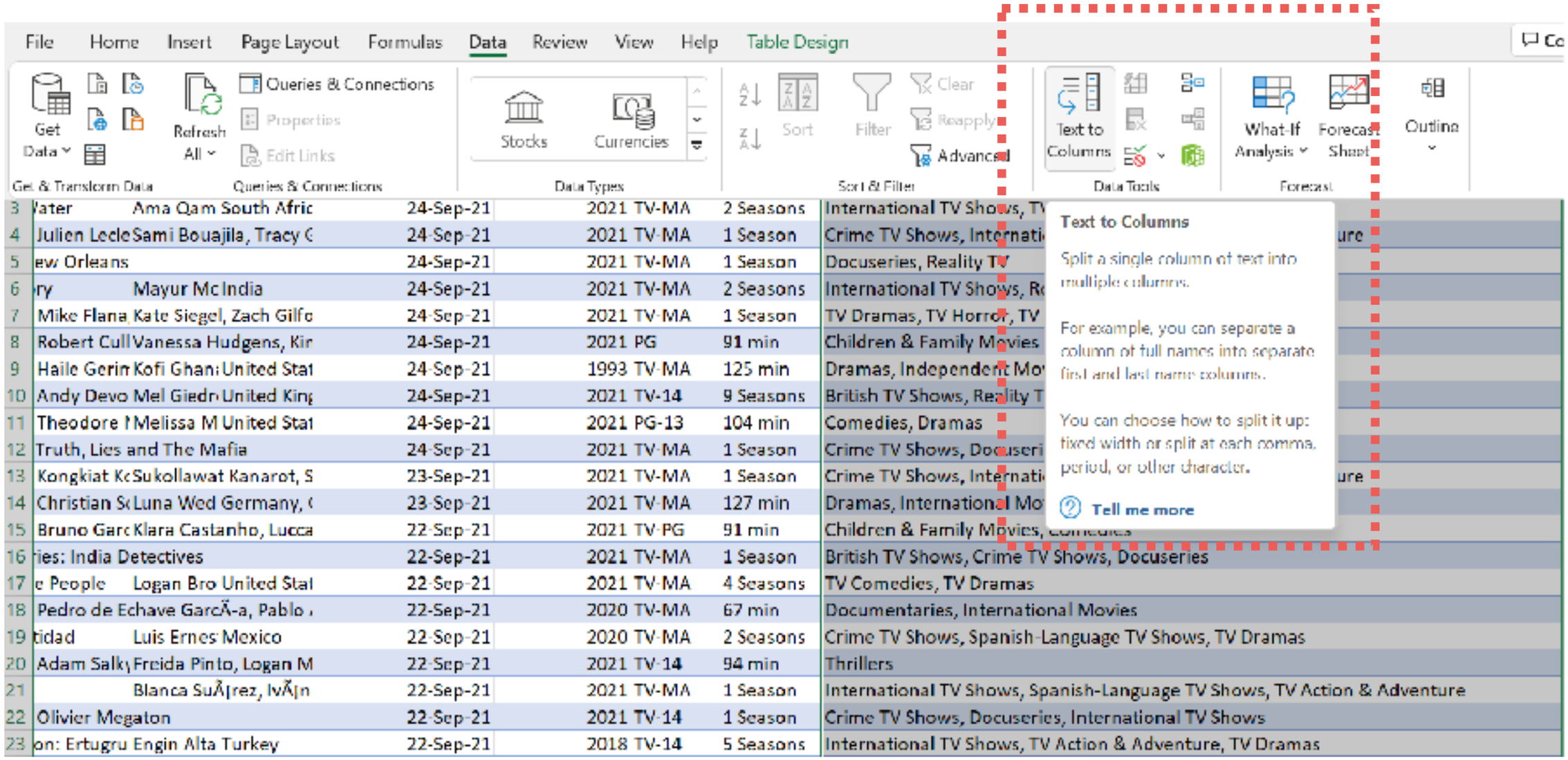
Crime TV Shows

International TV Shows

TV Action & Adventure

Working with Excel #3

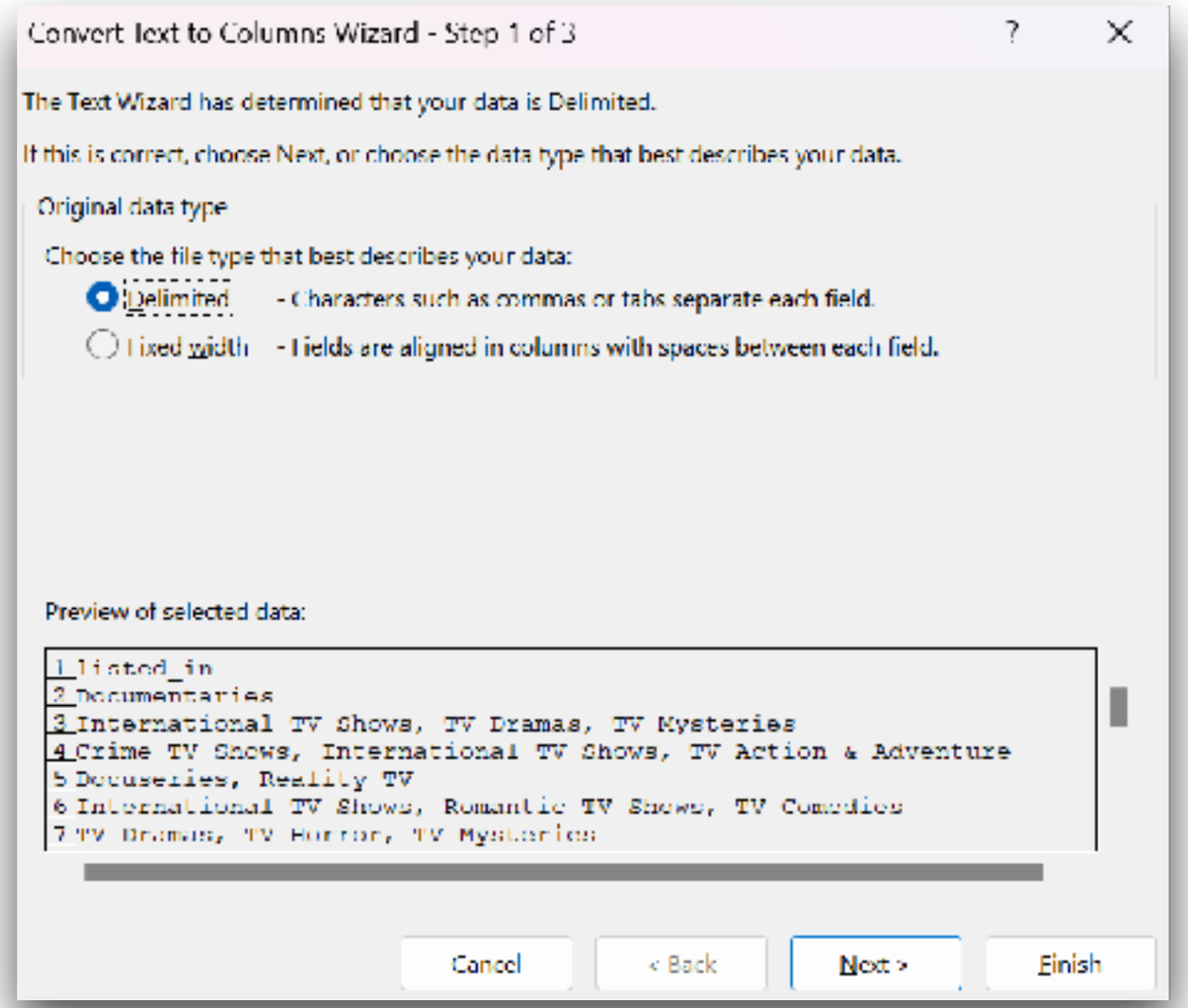
Select **Data -> Text to columns**

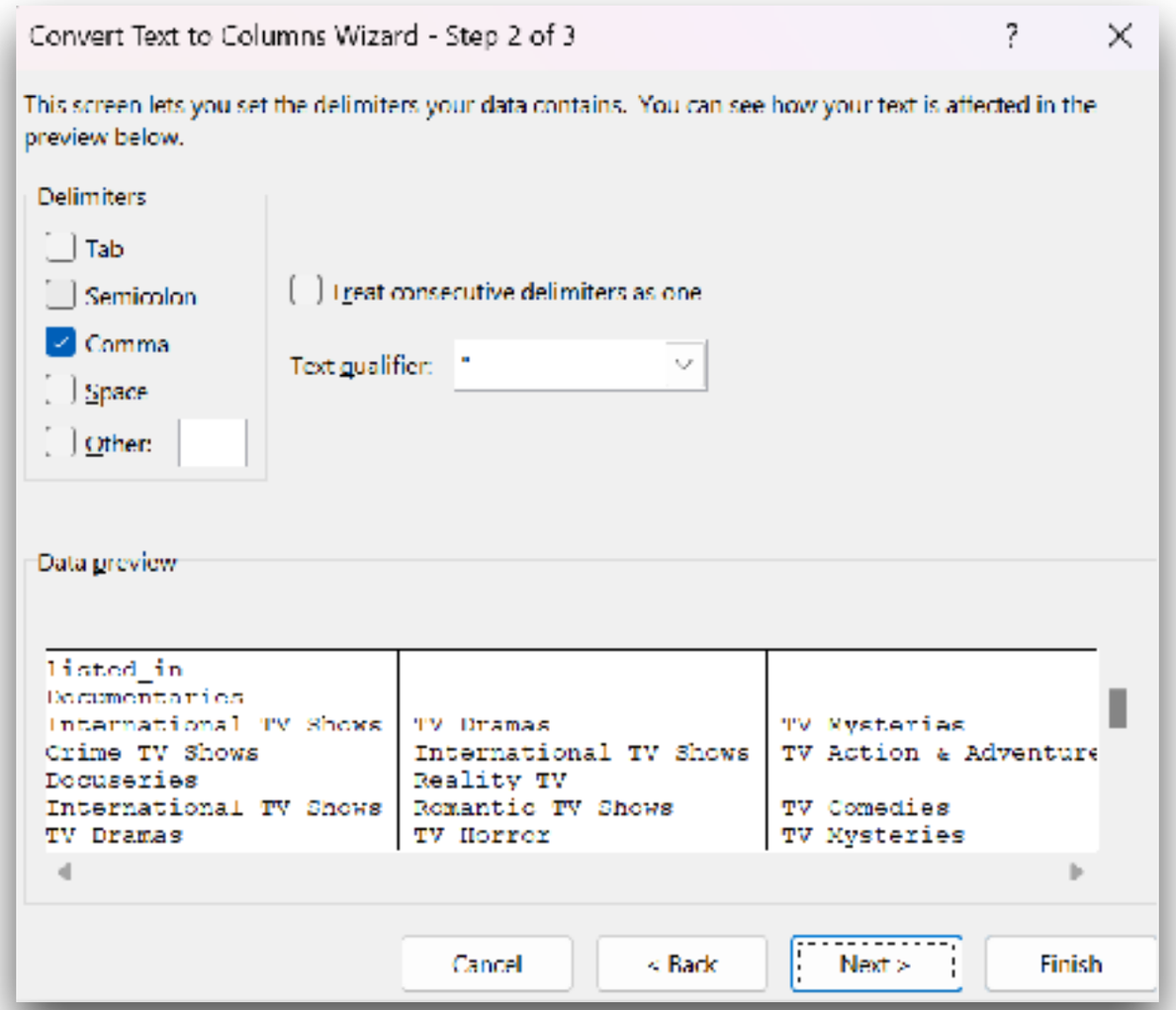


3	Vater	Ama Qam South Afric	24-Sep-21	2021	TV-MA	2 Seasons	International TV Shows, TV
4	Julien Lecle	Sami Bouajila, Tracy C	24-Sep-21	2021	TV-MA	1 Season	Crime TV Shows, Internati
5	ew Orleans		24-Sep-21	2021	TV-MA	1 Season	Docuseries, Reality TV
6	ry	Mayur Mc India	24-Sep-21	2021	TV-MA	2 Seasons	International TV Shows, R
7	Mike Flana	Kate Siegel, Zach Gilfo	24-Sep-21	2021	TV-MA	1 Season	TV Dramas, TV Horror, TV
8	Robert Cull	Vanessa Hudgens, Kir	24-Sep-21	2021	PG	91 min	Children & Family Movies
9	Haile Gerin	Kofi Ghan United Stat	24-Sep-21	1993	TV-MA	125 min	Dramas, Independent Mo
10	Andy Devo	Mel Giedr United King	24-Sep-21	2021	TV-14	9 Seasons	British TV Shows, Reality T
11	Theodore I	Melissa M United Stat	24-Sep-21	2021	PG-13	104 min	Comedies, Dramas
12	Truth, Lies and The Mafia		24-Sep-21	2021	TV-MA	1 Season	Crime TV Shows, Docuseri
13	Kongkiat Kc	Sukollawat Kanarot, S	23-Sep-21	2021	TV-MA	1 Season	Crime TV Shows, Internati
14	Christian Sc	Luna Wed Germany, (	23-Sep-21	2021	TV-MA	127 min	Dramas, International Mo
15	Bruno Garc	Klara Castanho, Lucca	22-Sep-21	2021	TV-PG	91 min	Children & Family Movies, Comedies
16	ies: India	Detectives	22-Sep-21	2021	TV-MA	1 Season	British TV Shows, Crime TV Shows, Docuseries
17	e People	Logan Bro United Stat	22-Sep-21	2021	TV-MA	4 Seasons	TV Comedies, TV Dramas
18	Pedro de	Echave GarcÃ-a, Pablo	22-Sep-21	2020	TV-MA	67 min	Documentaries, International Movies
19	tidad	Luis Ernes Mexico	22-Sep-21	2020	TV-MA	2 Seasons	Crime TV Shows, Spanish-Language TV Shows, TV Dramas
20	Adam Salky	Freida Pinto, Logan M	22-Sep-21	2021	TV-14	94 min	Thrillers
21		Blanca SuÃ¡rez, IvÃ¡n	22-Sep-21	2021	TV-MA	1 Season	International TV Shows, Spanish-Language TV Shows, TV Action & Adventure
22	Olivier Megaton		22-Sep-21	2021	TV-14	1 Season	Crime TV Shows, Docuseries, International TV Shows
23	on: Ertugru	Engin Alta 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

	K	L	M
	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 ?

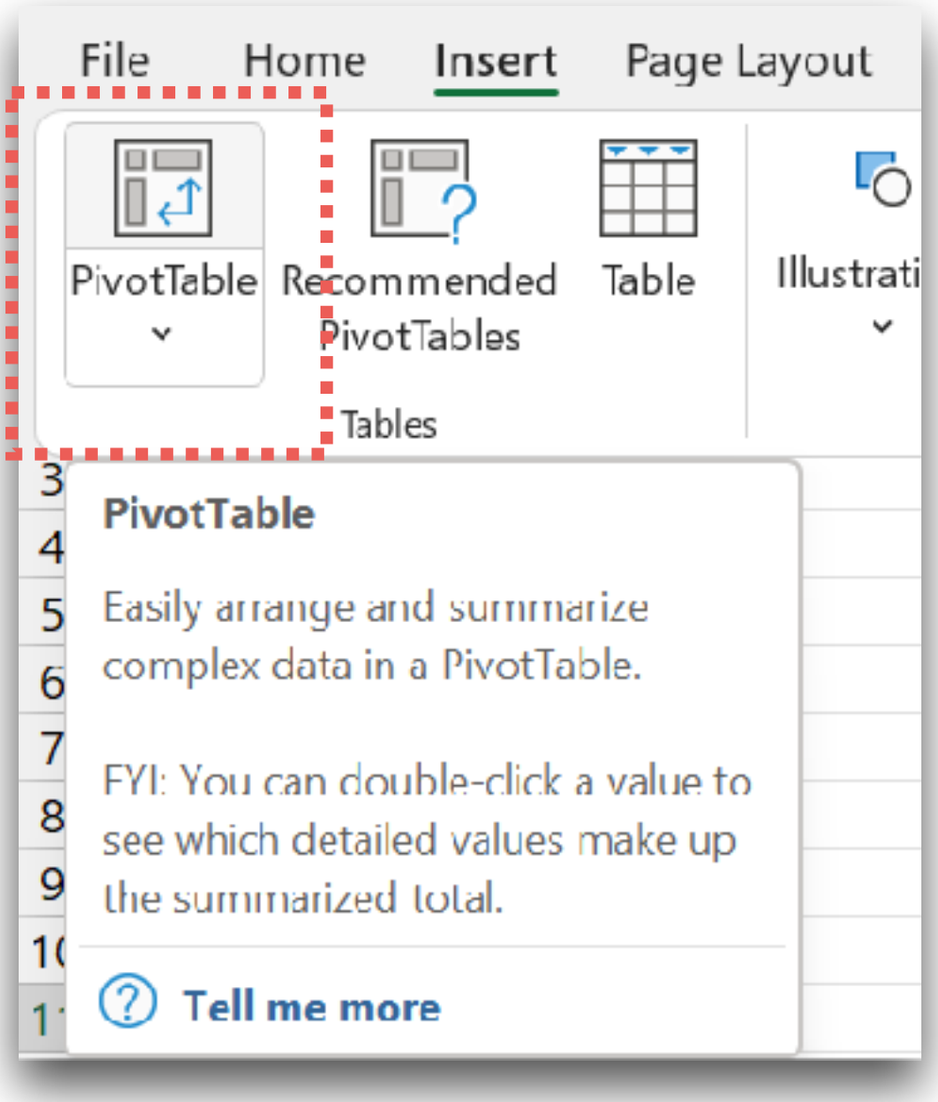
Top 5 country has produced more movies and TV shows ?

Which Genre has more in Movie Type ?

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

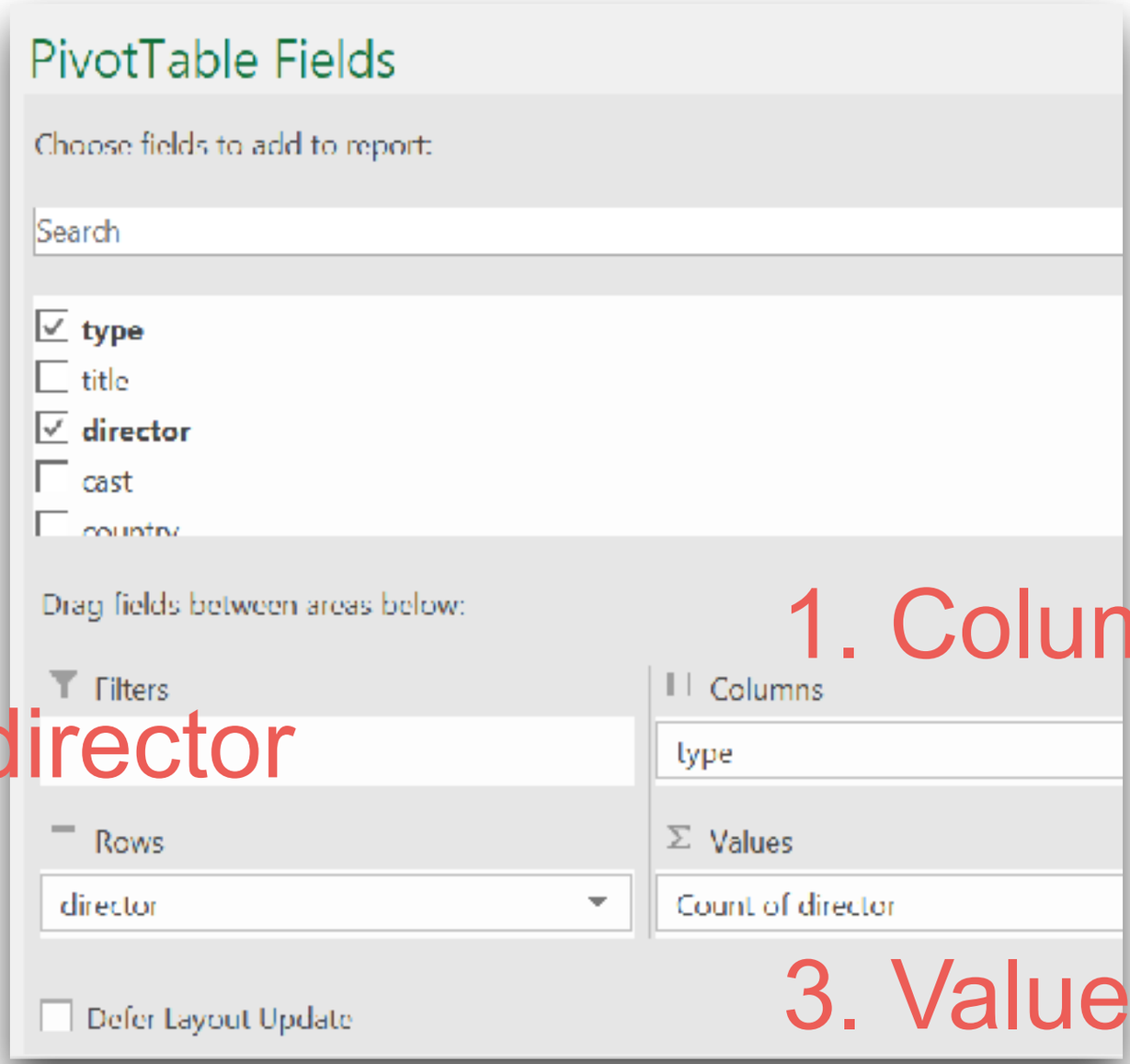
Top 10 director has cast more movie ?

Working with **Pivot Table**



Top 10 director has cast more movie ?

Choose Columns, Rows and Values



2. Rows = director

1. Column = type

3. Values = director

Top 10 director has cast more movie ?

Count of director	Column Labels		
Row 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

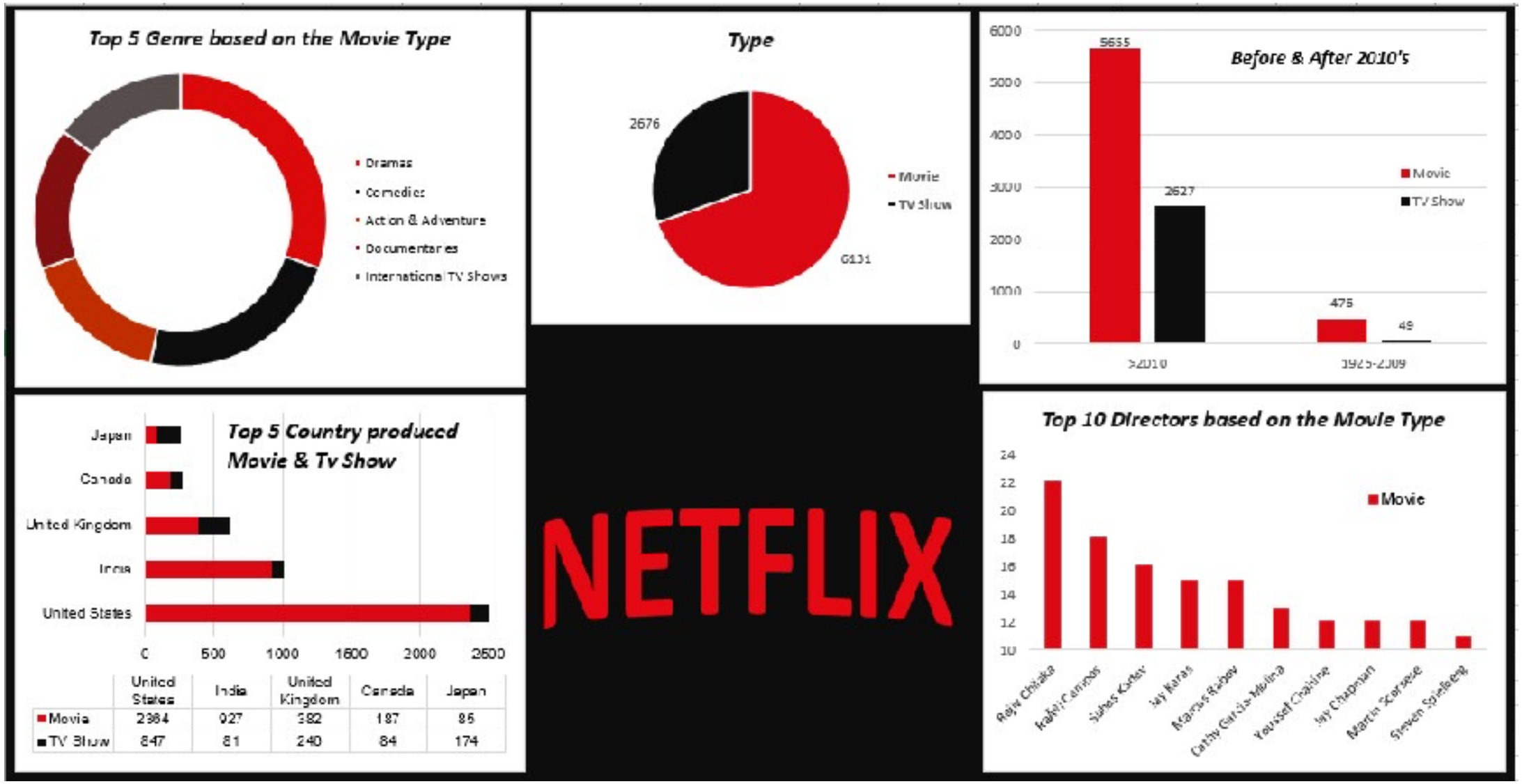
Try by yourself

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 (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

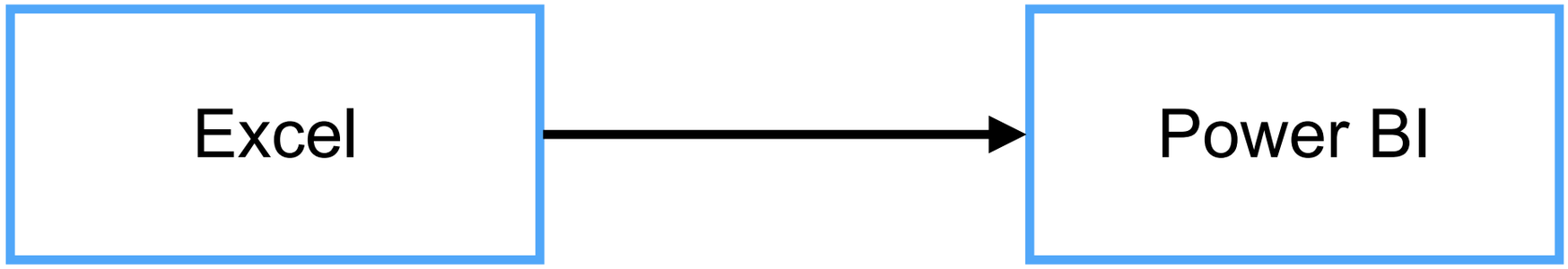
1 Column per 1 content

No merge cell

No blank/empty cell

Header/Record

Excel vs Power BI



Data source

Cleansing

Transformation

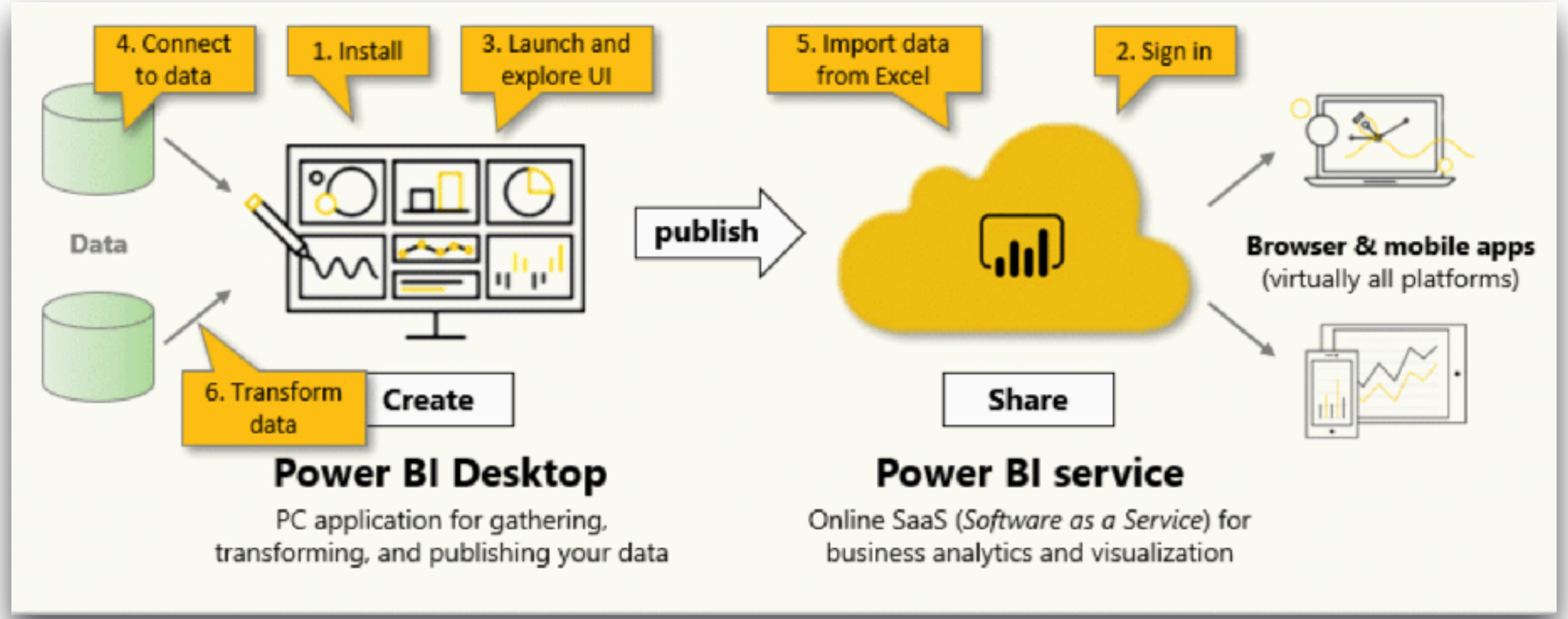
Visualization

Sharing

Types of Power BI

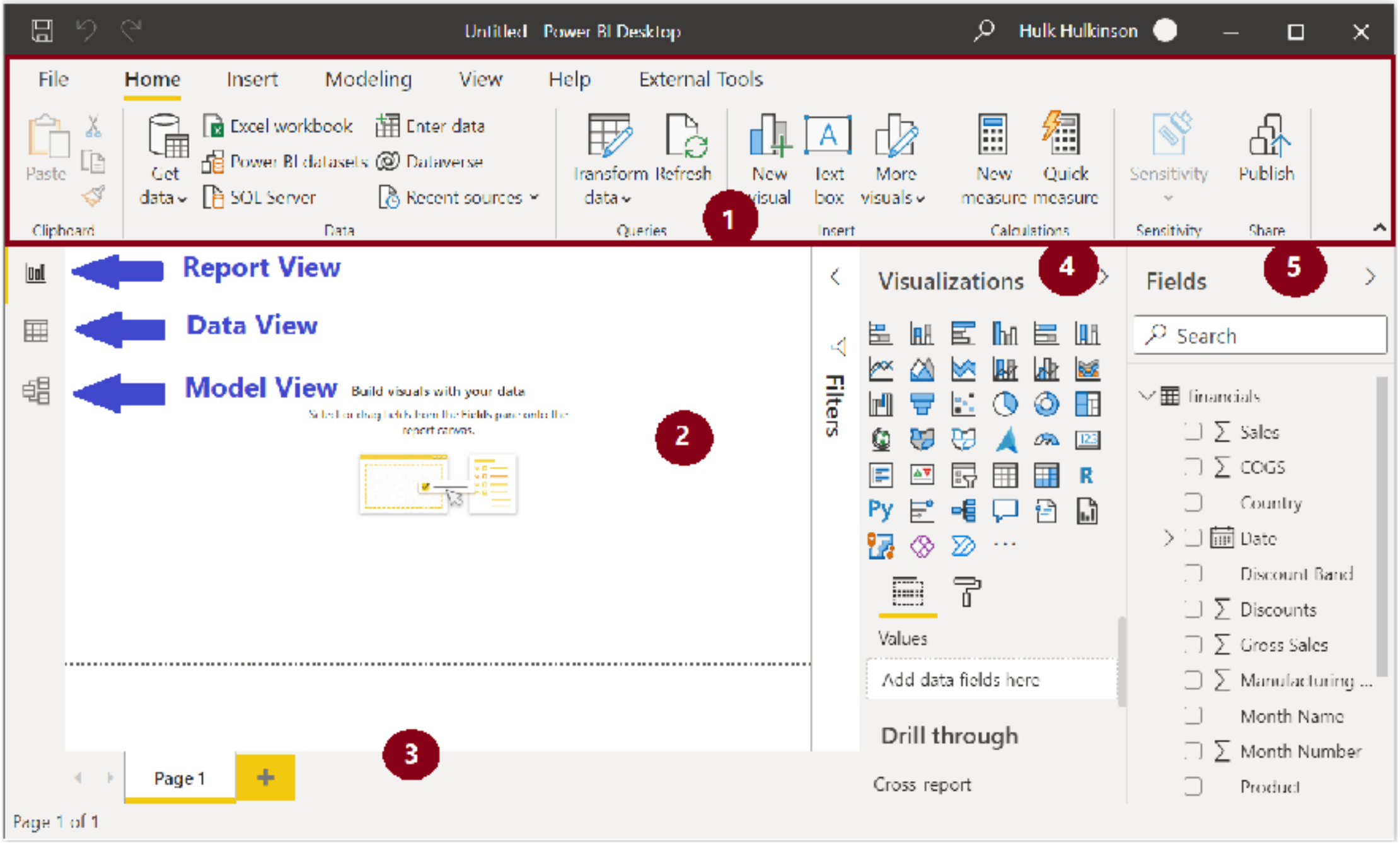
Web app (Power BI Service)
Desktop (Windows only)
Mobile

Types of Power BI

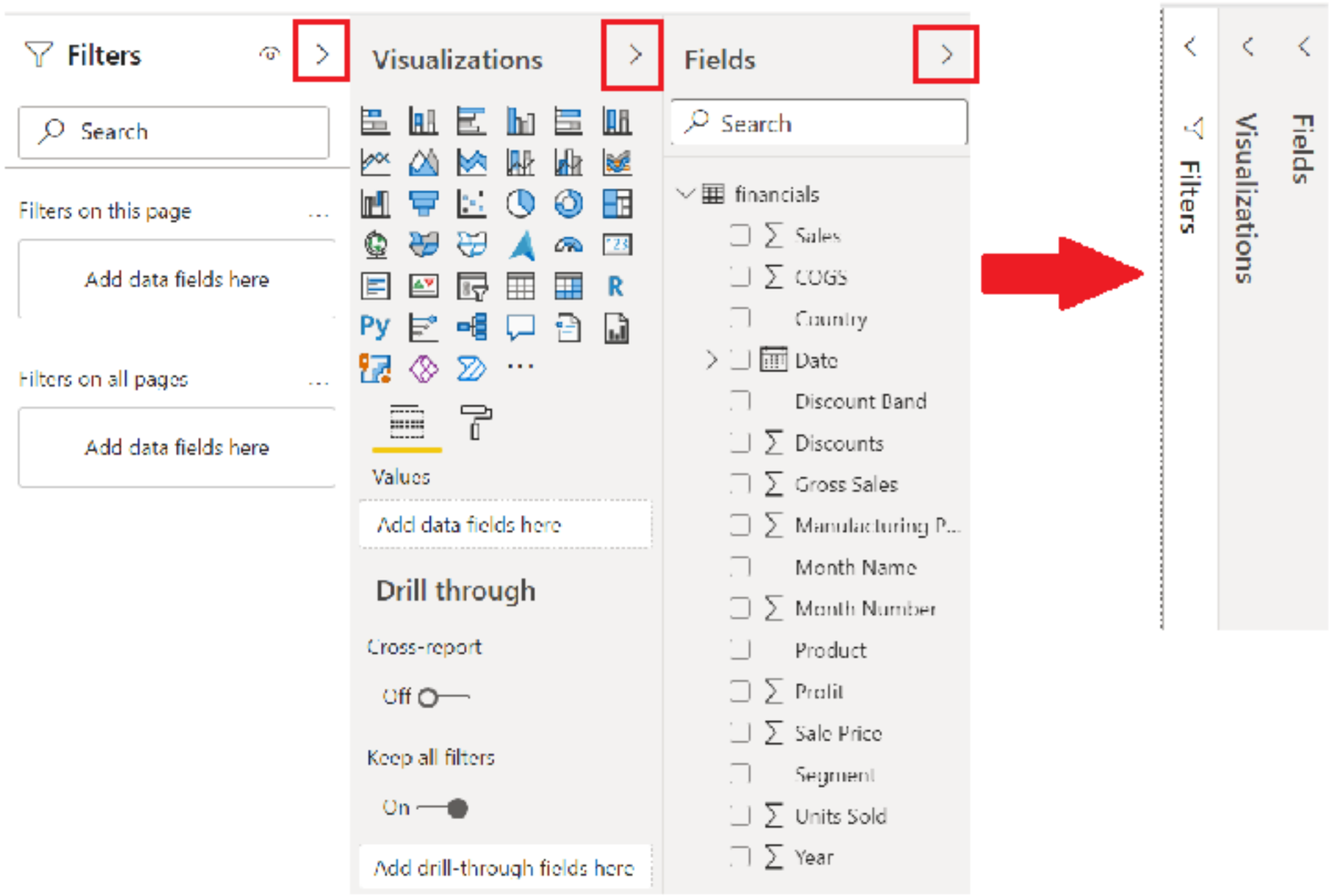


https://app.powerbi.com/

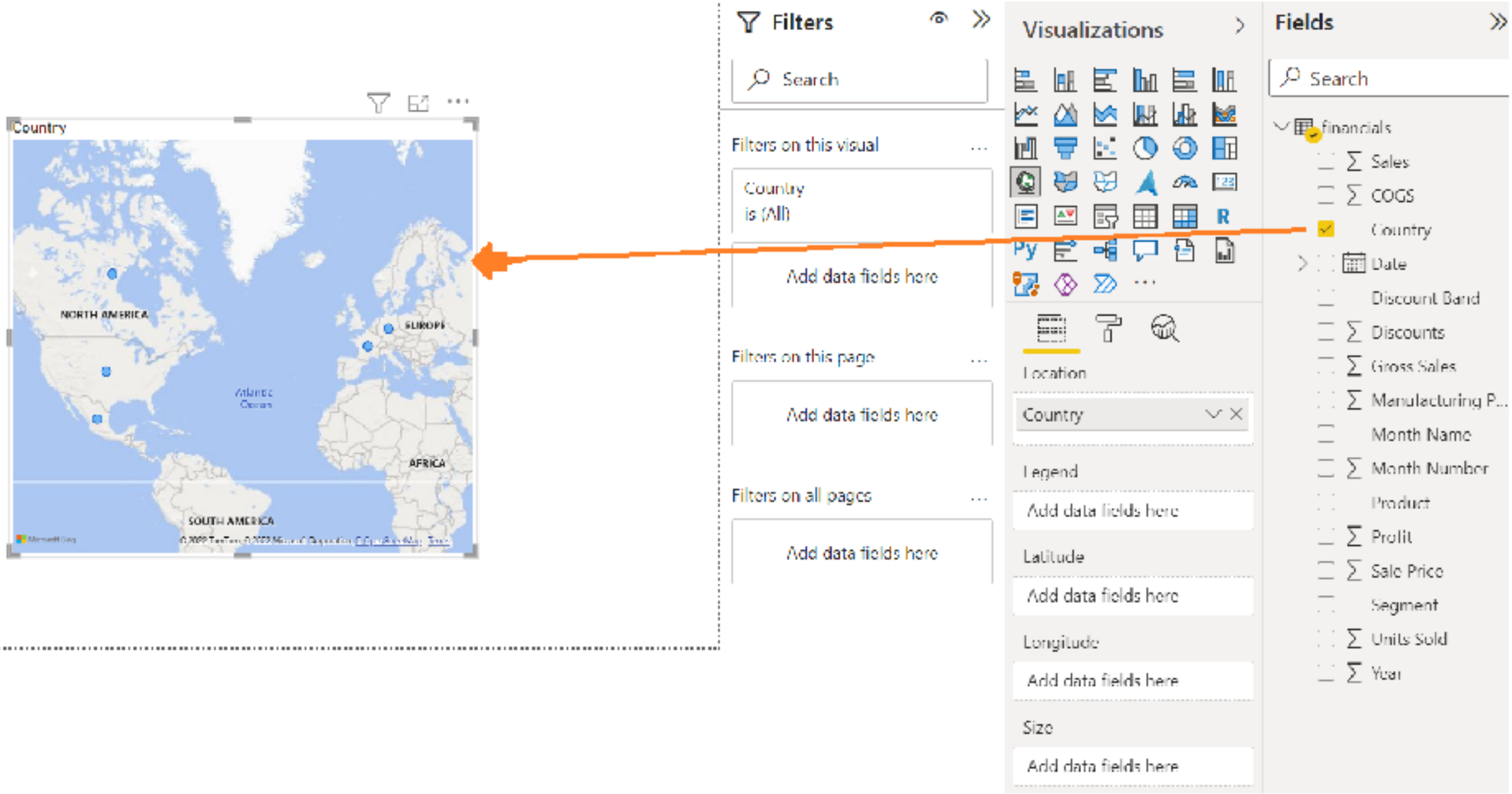
Power BI Desktop



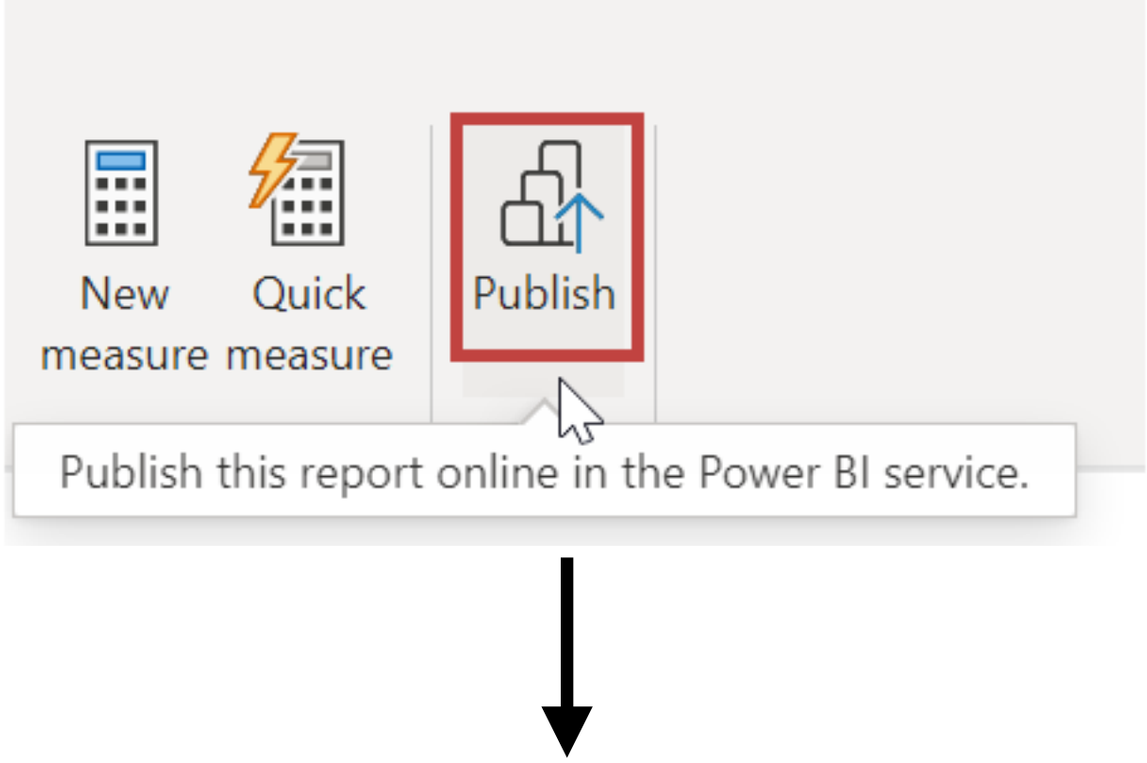
Power BI Desktop

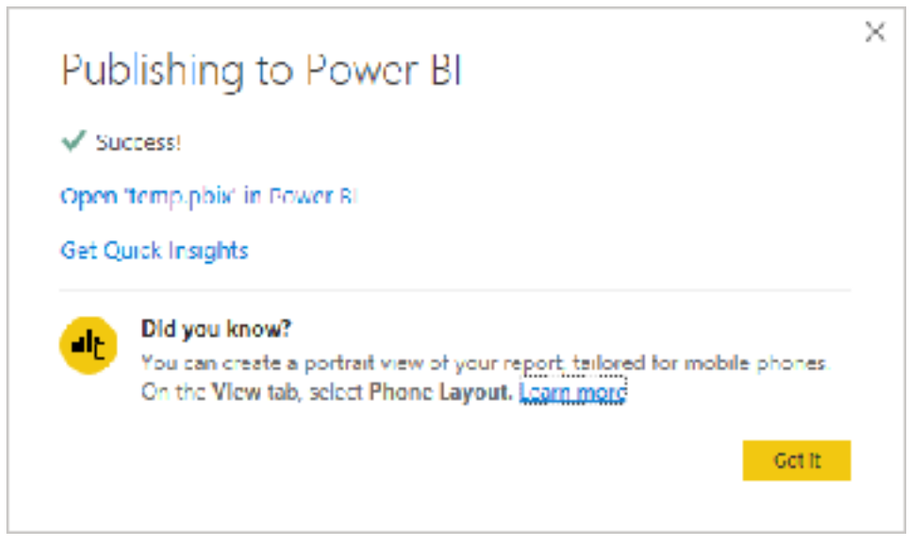


Create Visualization



Publish a report to web app

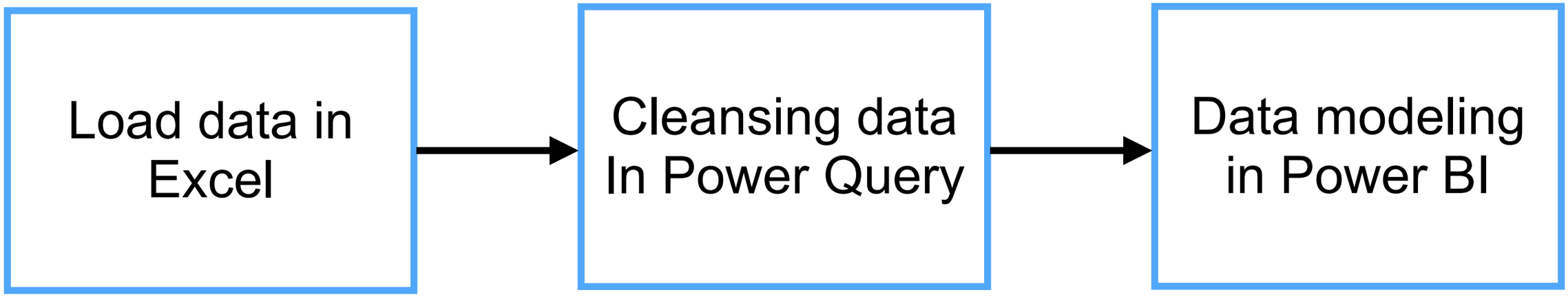




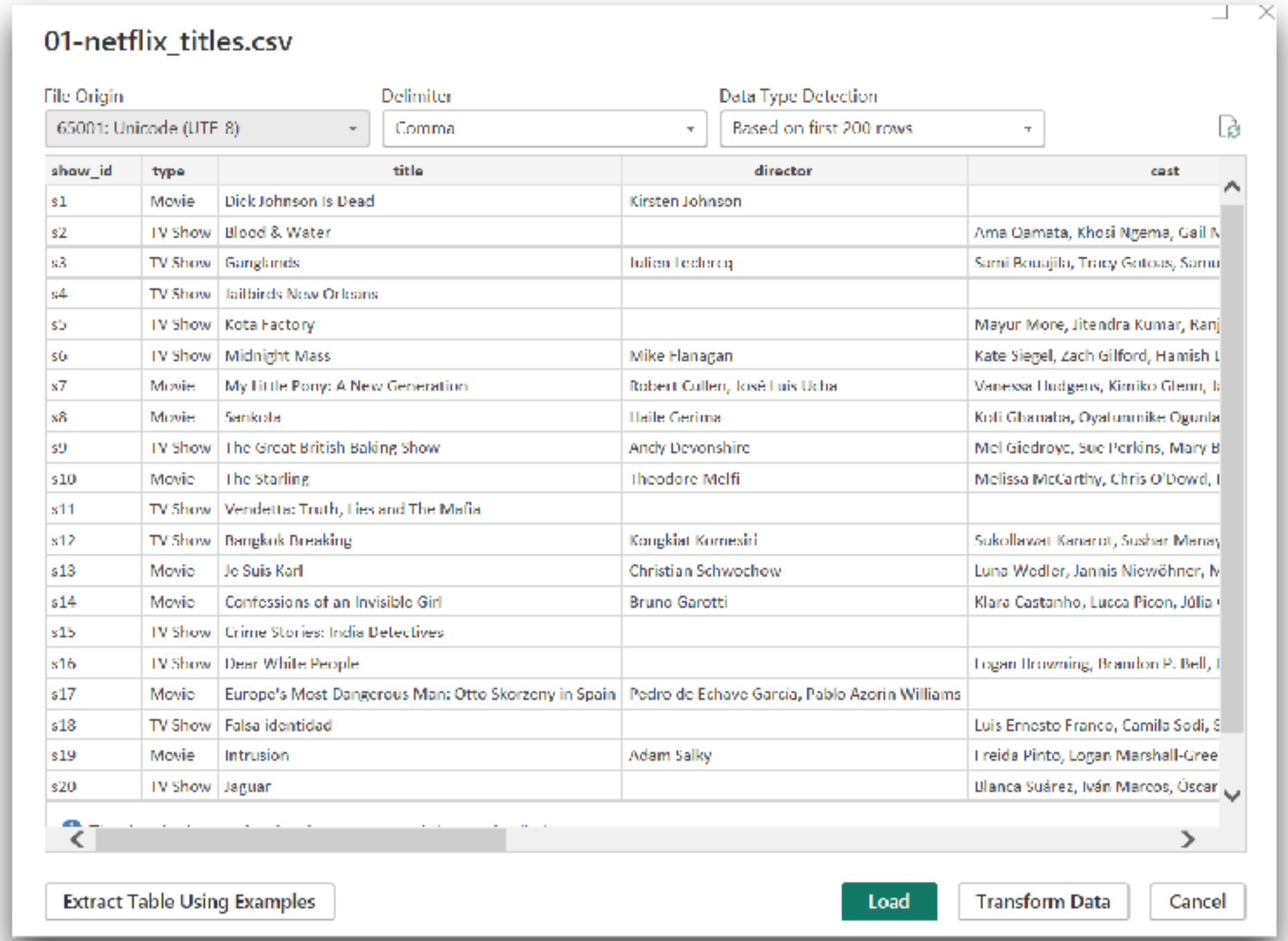
Workshop #4

Working with Power BI

Steps of Workshop



1. Load dataset



show_id	type	title	director	cast
s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	
s2	TV Show	Blood & Water		Ama Qamata, Khosi Ngema, Gail M
s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samu
s4	TV Show	Jailbirds New Orleans		
s5	TV Show	Kota Factory		Mayur More, Jitendra Kumar, Ranj
s6	TV Show	Midnight Mass	Mike Flanagan	Kate Siegel, Zach Gilford, Hamish L
s7	Movie	My Little Pony: A New Generation	Robert Cullen, José Luis Ucha	Vanessa Hudgens, Kimiko Glenn, J
s8	Movie	Sankofa	Haile Gerima	Kofi Ghanaba, Oyafunmike Ogunla
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,
s11	TV Show	Vendetta: Truth, Lies and The Mafia		
s12	TV Show	Bangkok Breaking	Kongkiat Komesiri	Sukollawat Kanarot, Susher Manay
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,
s17	Movie	Europe's Most Dangerous Man: Otto Skorzeny in Spain	Pedro de Echave Garcia, Pablo Azorin Williams	
s18	TV Show	Falsa identidad		Luis Ernesto Franco, Camila Sodi, S
s19	Movie	Intrusion	Adam Salky	Freida Pinto, Logan Marshall-Gree
s20	TV Show	Jaguar		Blanca Suárez, Iván Marcos, Óscar

2. Data cleansing with Power Query in Excel

Missing data

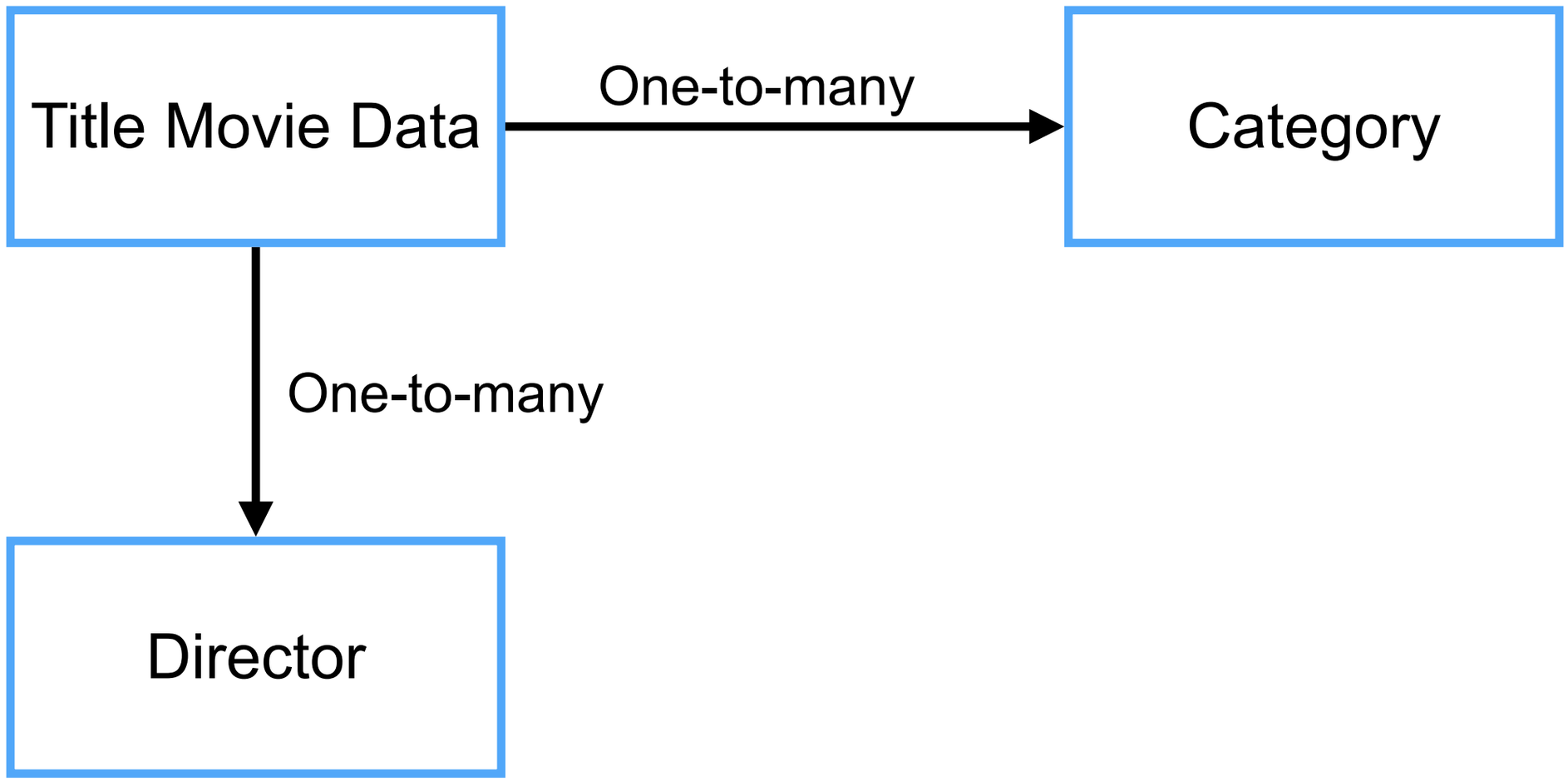
Change Data type

Split data

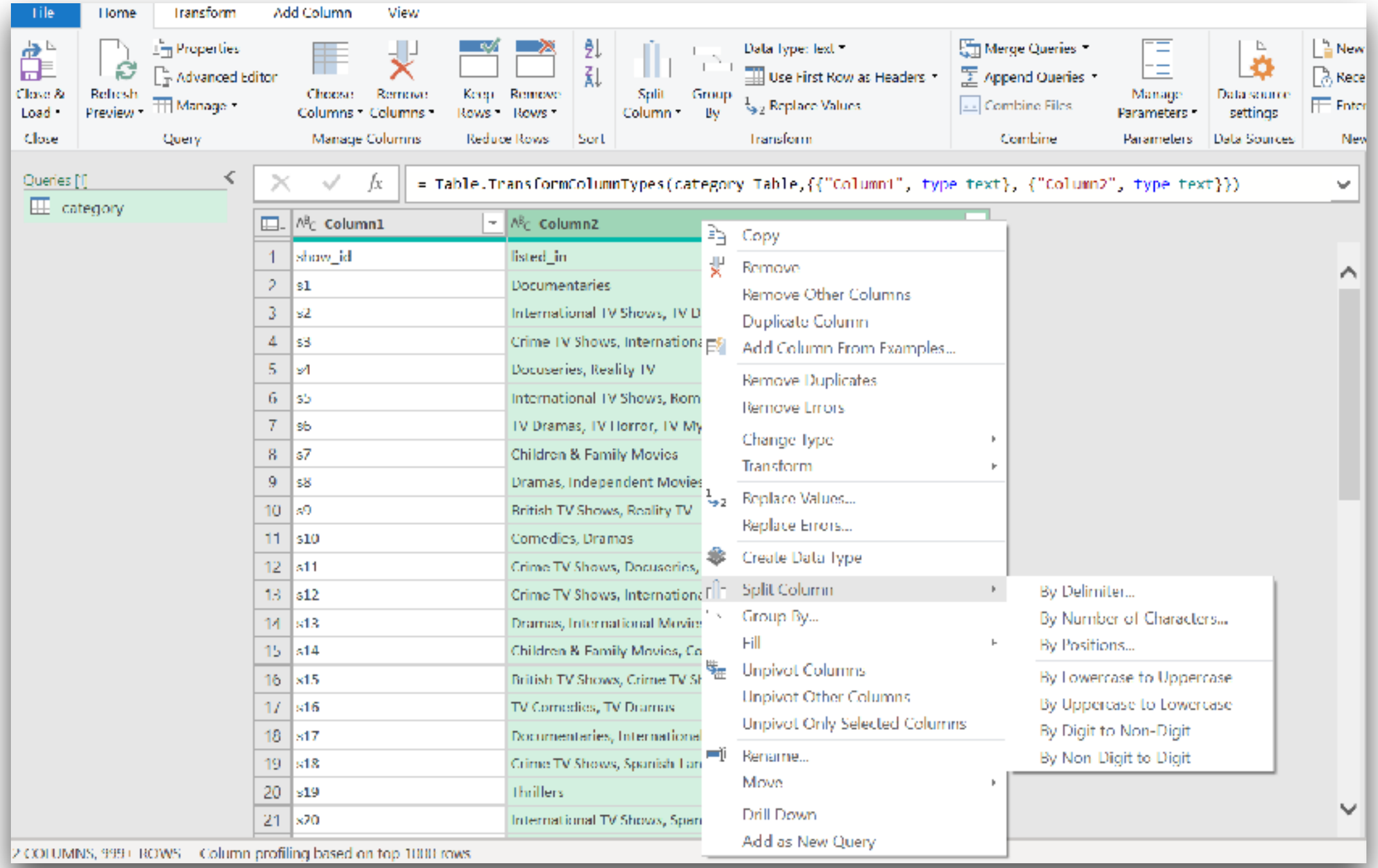
Trim data

Data model Relationship

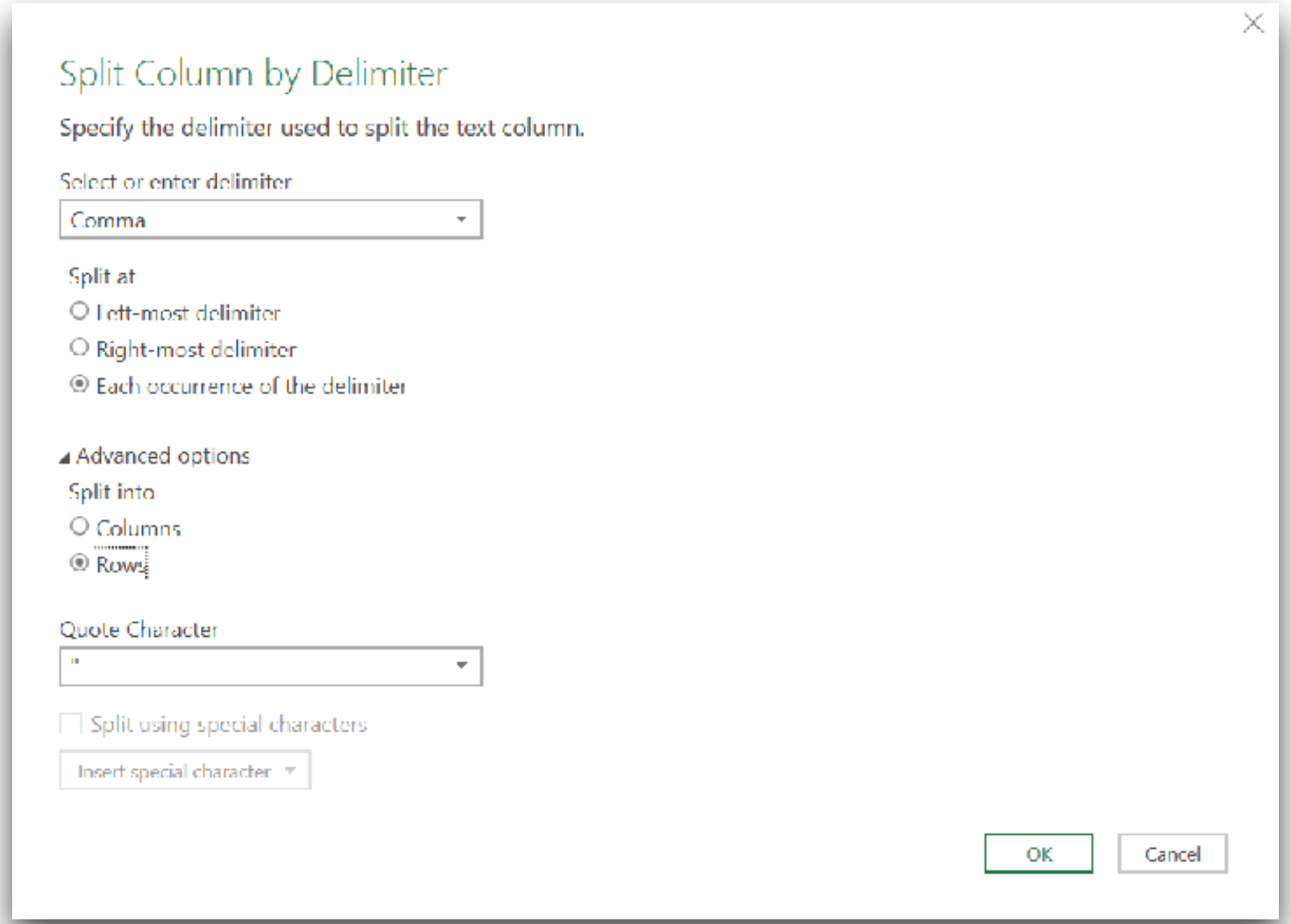
Data Modeling before Analysis



2.1 Split data into new row



2.2 Use delimiter by comma



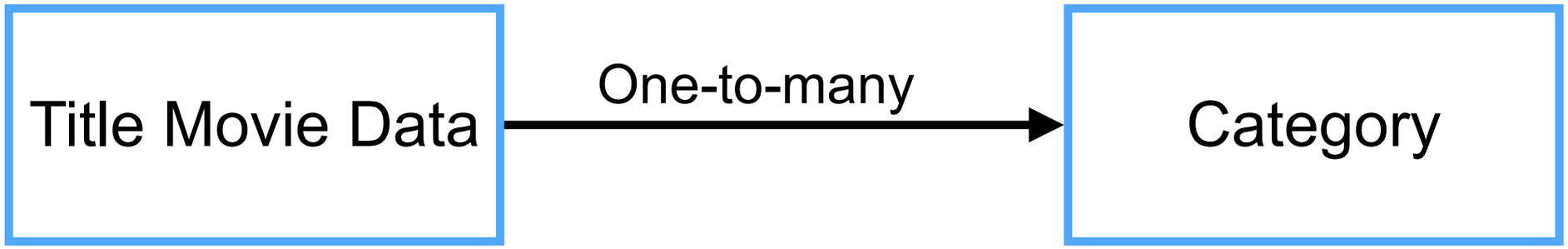
2.3 New sheet of category

= Table.TransformColumnTypes(#"Promot

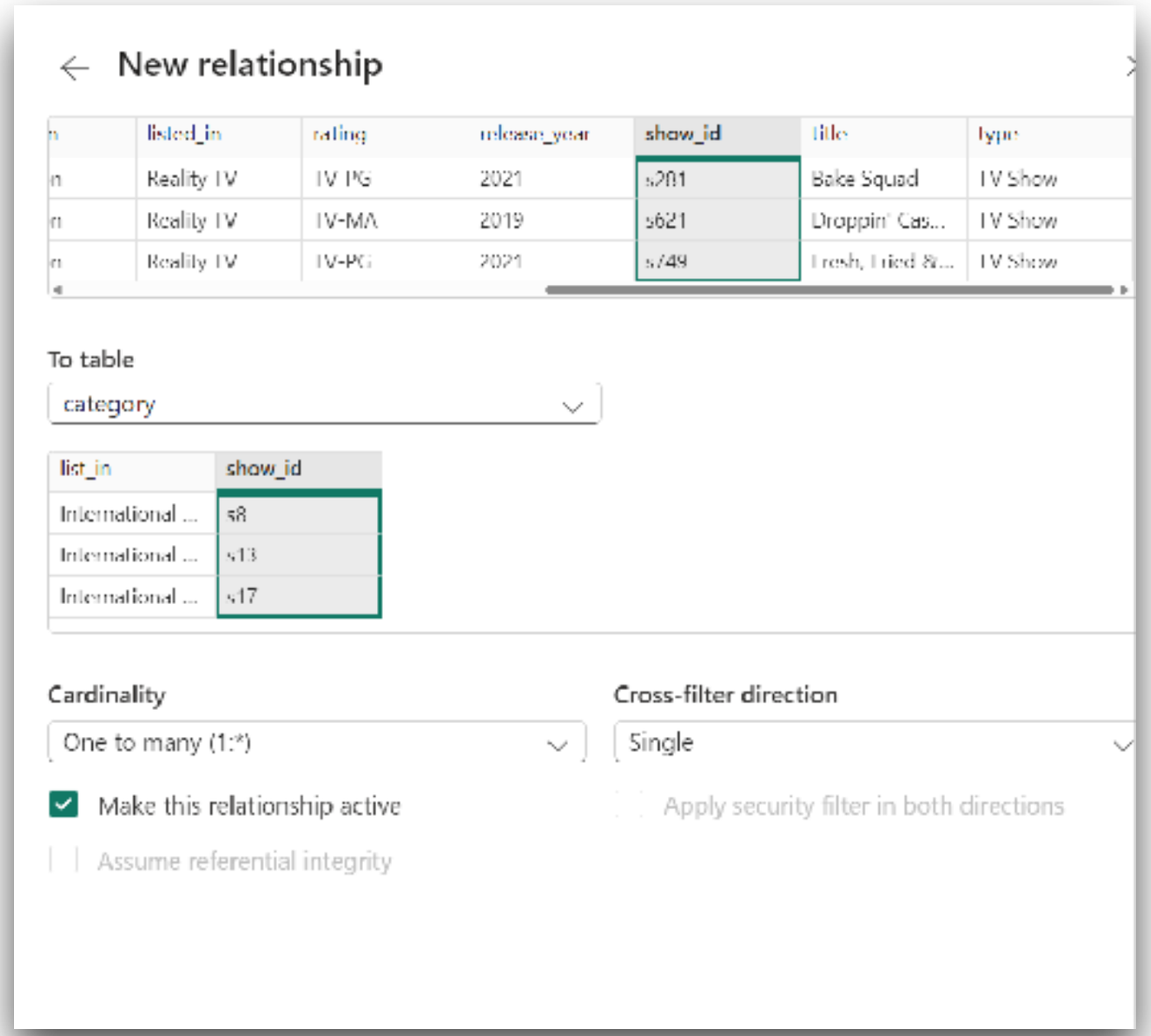
	show_id	listed_in
1	s1	Documentaries
2	s2	International TV Shows
3	s2	TV Dramas
4	s2	TV Mysteries
5	s3	Crime TV Shows
6	s3	International TV Shows
7	s3	TV Action & Adventure
8	s4	Docuseries
9	s4	Reality TV
10	s5	International TV Shows
11	s5	Romantic TV Shows
12	s5	TV Comedies
13	s6	TV Dramas
14	s6	TV 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)



4. Create Dashboard in Power BI

Chart

Table

Map

4. Create Dashboard in Power BI