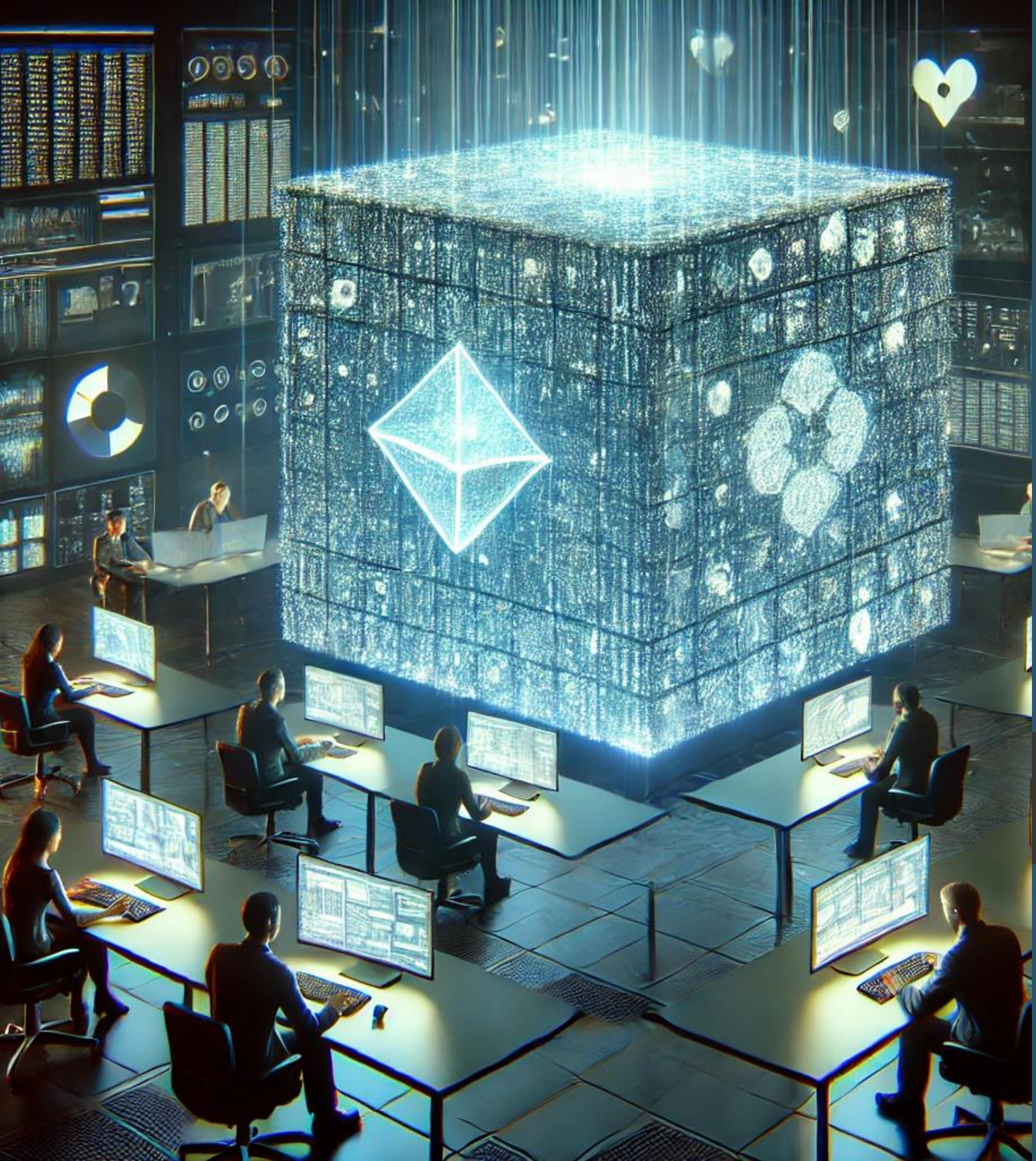


Data Mining and Text Analytics 2025/2026 Lecture_01

Data Mining Introduction

Prof Alessandro Bruno

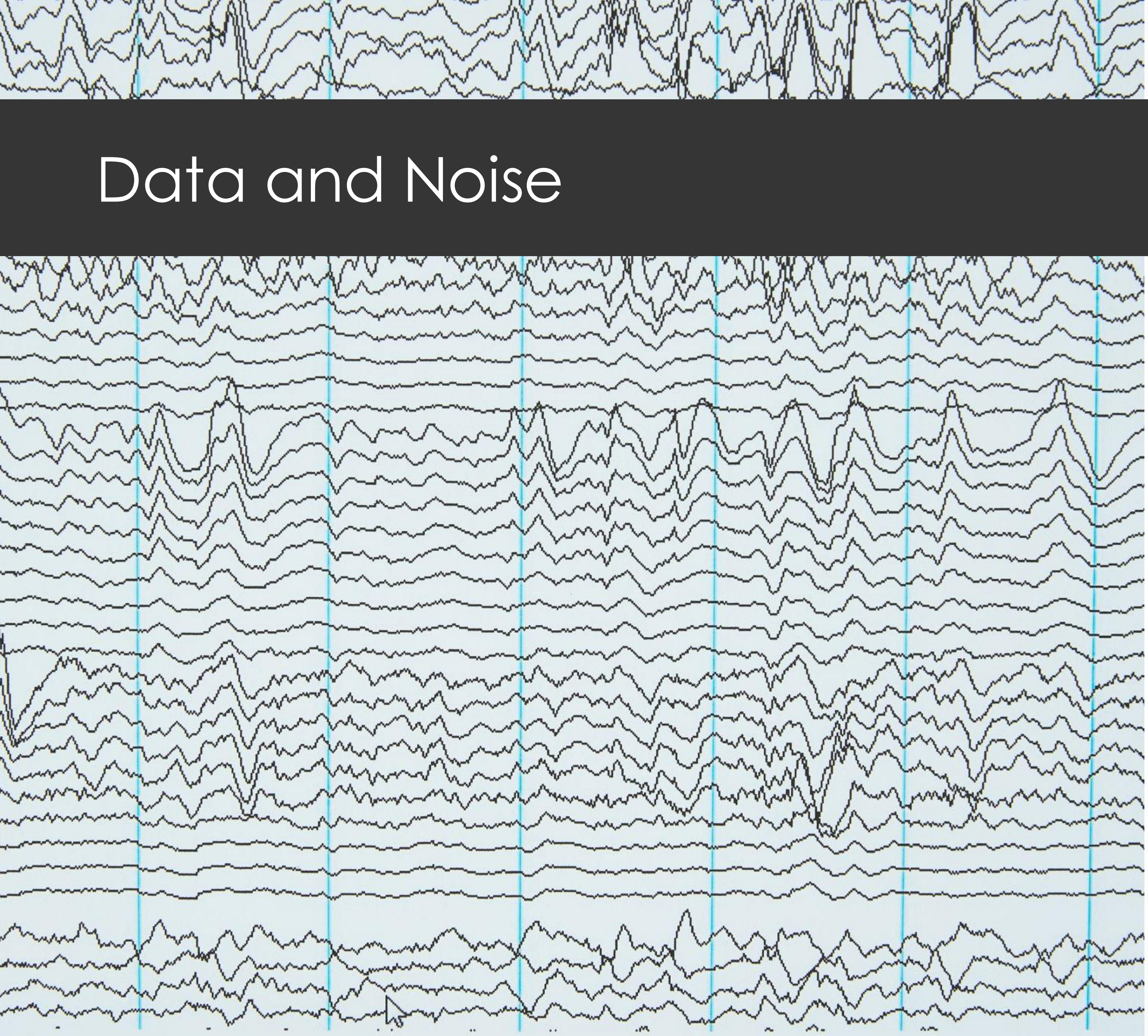
IULM University – Faculty of Communication

Postgraduate Programme in Artificial Intelligence for Business and Society

Outline

- Data and Noise
- A formal definition of Data Mining
- Data Mining as the Evolution of Information Technology
- Knowledge Discovery Process
- What kind of data can be mined?
- What kinds of Patterns can be mined?
- Mining Frequent Patterns, Associations, and Correlations

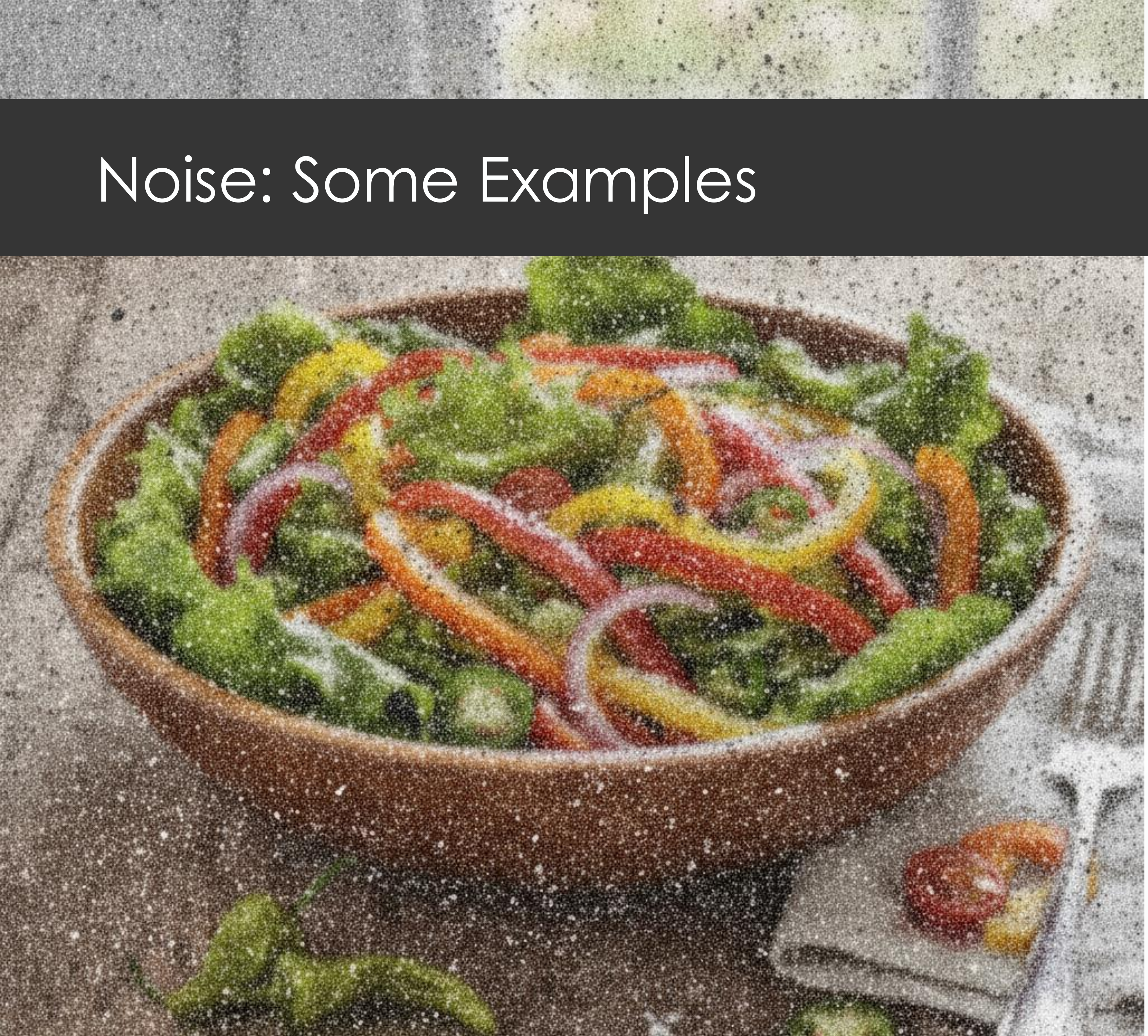
Data and Noise



"To find signals in data, we must learn to reduce the noise - not just the noise that resides in the data, but also the noise that resides in us (bias). It is nearly impossible for noisy minds to perceive anything but noise in data."

Stephen Few, Signal: Understanding What Matters in a World of Noise

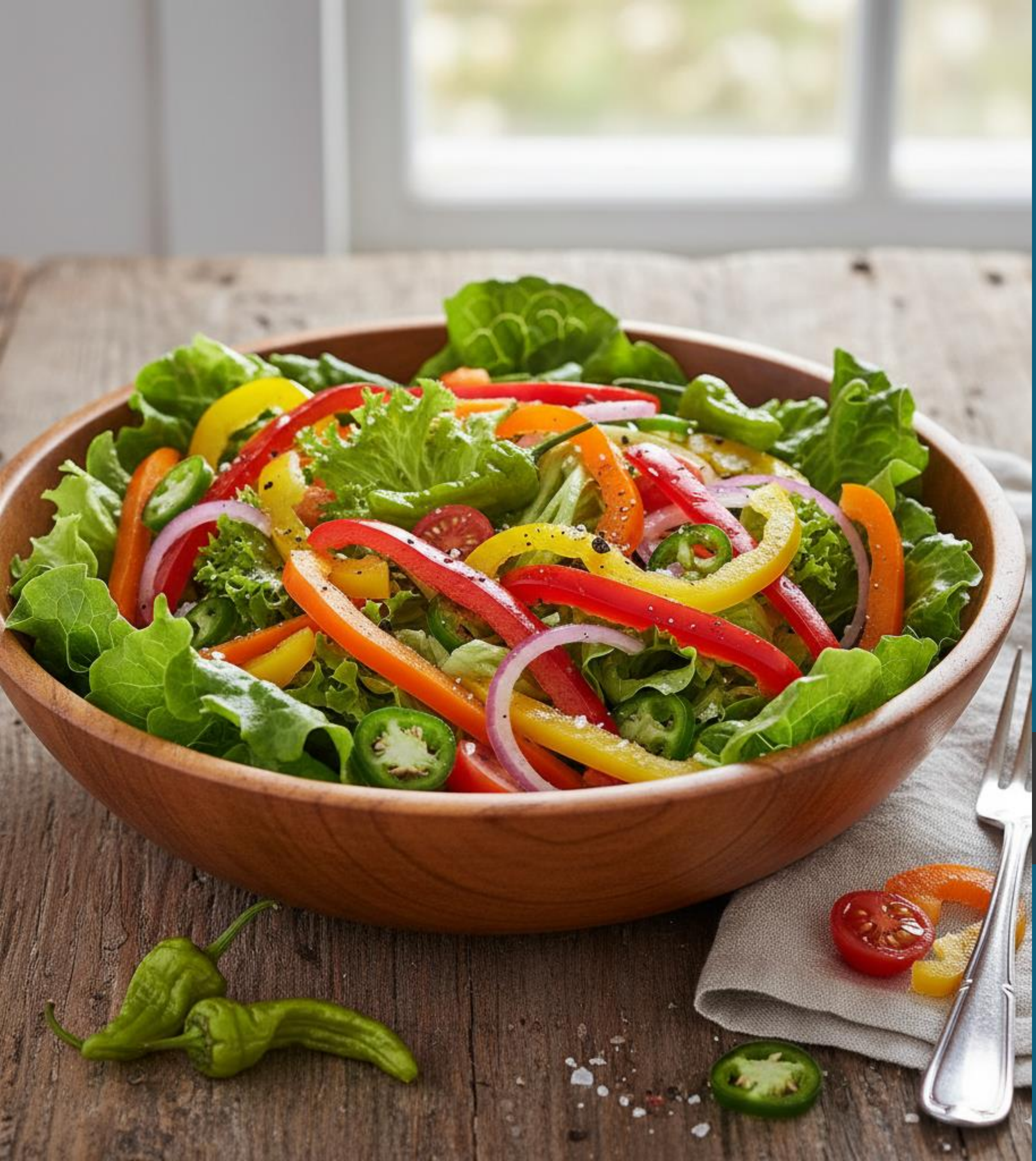
Noise: Some Examples



- Let's have a look at the given picture.
- How would you describe its content?

Image Denoising Techniques to remove additive noise

- The scientific literature offers a wide plethora of techniques to help enhance the image quality in terms of pixel content.
- Let's apply a common image filtering technique named "median filter" and see the outcome.



Data without Noise

Here is the original picture

Data Mining. Why?

- What is data mining for?
- Imagine digging a hole while you look for something precious.
- You will go through excavation, rock and stone residuals, and other elements that might not exactly be what you are seeking.
- You need some tools and techniques to find your way to your goal: INFORMATION
- Before extracting INFORMATION, you need to put together some little pieces or components that, once combined, will foster the composition of the INFORMATION itself.
- Those little pieces are called PATTERNS.
- PATTERNS are an ensemble of elements that are in data and express "sequences", "motifs", "codes", and so on.

Patterns in Stock Price

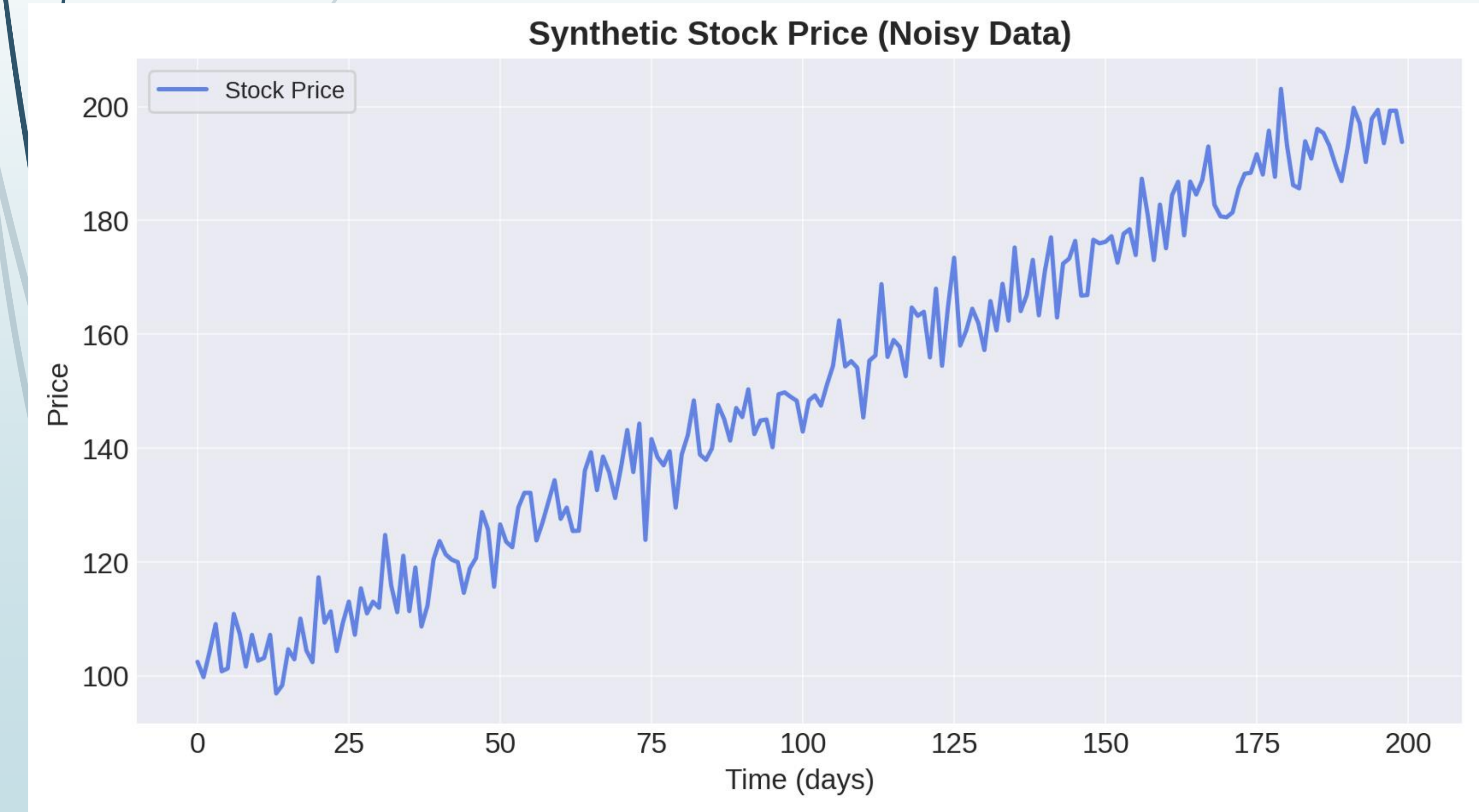
- **Stock price** analysis is a critical topic. **Fluctuations** in stock price might reveal events of interest to the **market**.
- Here is an example with a stock price series.

$$P_t = 100 + 0.5t + \epsilon_t$$

- P_t is the stock price at t time
- ϵ_t is an additive factor
- Try now draw the trend of the stock price series.

Trends of the Stock Price

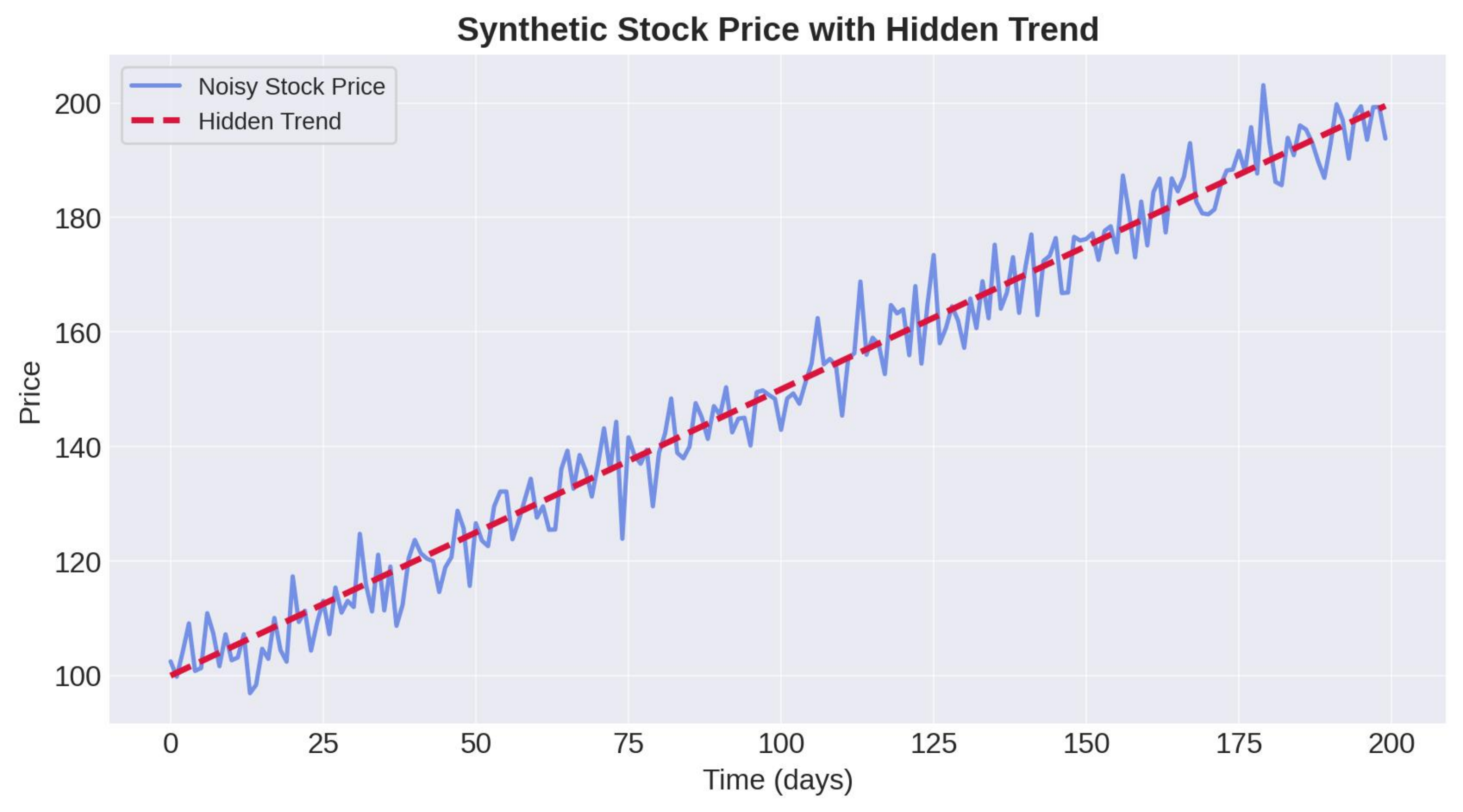
The so-called zig-zag effect is shown on the diagram



$$P_t = 100 + 0.5t + \epsilon_t$$

Trends and hidden patterns

- A pattern can be easily observed by using average function on the stock trend values.
- The red-dotted line shows the "hidden trend" emphasizing the linearly increasing function



Data Mining. Some examples.

- To find correlations in financial and business intelligence data
- Hospitals can spot trends and anomalies in their patient records
- Search engines can do better ranking and ad placement
- Environmental and public health agencies can spot patterns and abnormalities in their data.
- Monitoring of the energy consumption of household appliances
- Pattern analysis in bioinformatics and pharmaceutical data
- Spotting trends in blogs, Twitter, and many more
- The big question is: **how should we analyze the data?**

What is data mining

- Starting off with a **formal definition**:

"***Data mining***, *also popularly referred to as* ***Knowledge Discovery from Data*** *(****KDD****), is the automated or convenient extraction of* *<u>patterns</u>* *representing knowledge implicitly stored or captured in large* *<u>databases</u>*, *data* *<u>warehouses</u>*, *the Web, other massive information repositories, or data streams*"[1]

[1] "Data Mining. Concepts and Techniques". Jiawei Han, Micheline Kamber, Jian Pei

Data Mining

Informally, Data Mining is the process of using a computer program to find patterns or relationships in data.

Looking for combinations of symptoms that are reliable indicators of a given disease

Looking for products that customers tend to purchase together.

Grouping data showing some sort of correlations.

1

What do patterns reveal in data?

2

A pattern refers to anything that repeatedly exhibits a consistent form or organization

3

Patterns show data correlations

- A sequence of signals manifest periodically
- Some items exhibit a linear progression

4

Patterns can be found in structured and unstructured data

Patterns

Structured vs Semi-Structure vs Unstructured Data Content (formal perspective)

- **Structured Data**: Data with high degree of organisation (in a spreadsheet-like manner).
- **Semi-structured Data**: Data with some degree of organisation (forms)
- **Unstructured Data**: Data with no predefined organisation (a text file without pre-define scheme).

Structure vs Unstructured Data

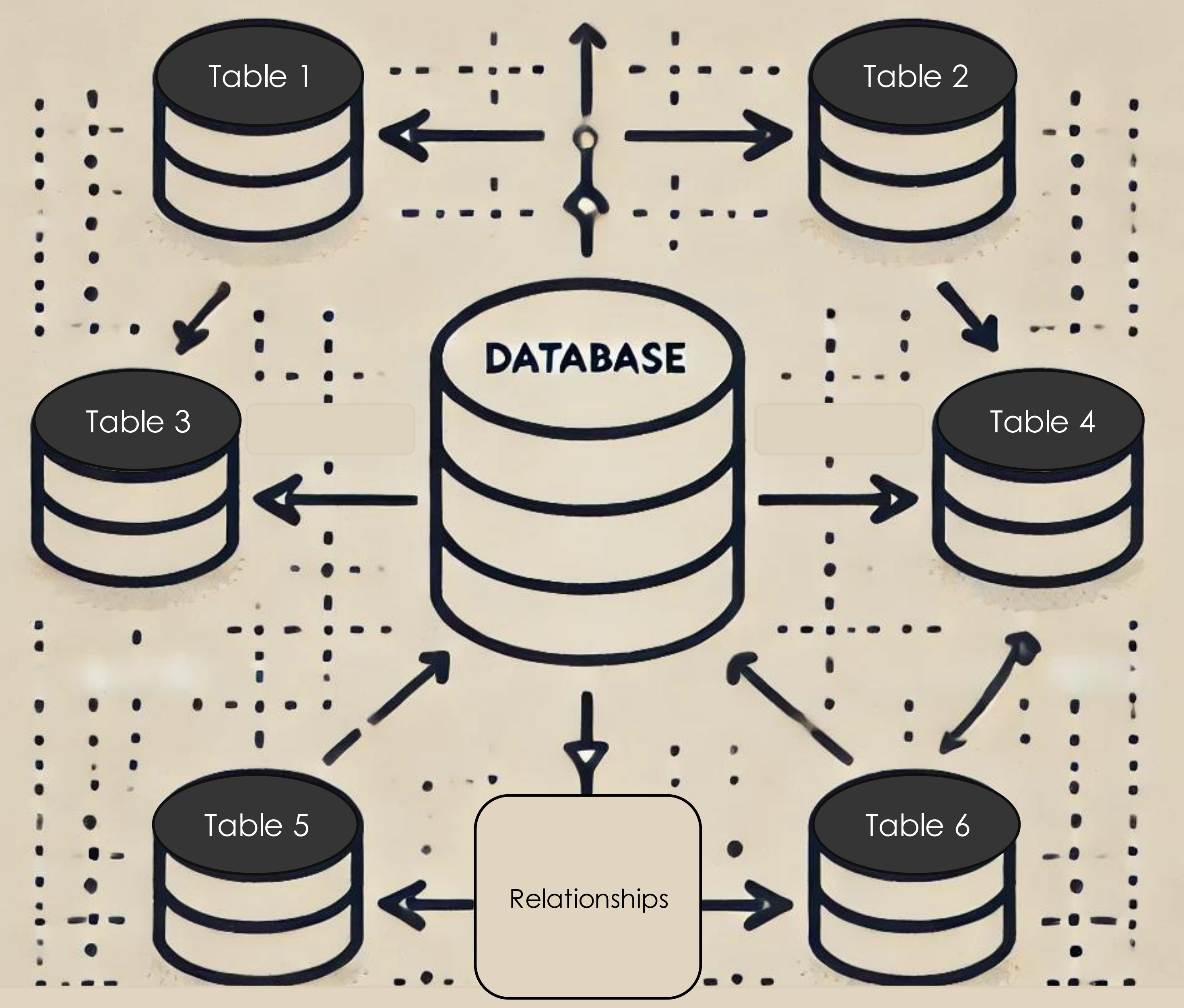
Structured Data	Semi-structured Data	Unstructured Data
Excel spreadsheets; Comma-separated value file; Relational Database Tables.	HTML files; Json files; XML files.	Images (.jpeg or .png files); videos (.mp4); sound files (.wav, .mp3 files); PDF files; Word Files, Plain text files.
Around 20% of worldwide data is structured.	Characterised by hierarchical structure.	Most of data that is created today is unstructured. (Tweets, Facebook posts, social media comments).

Data Management Hierarchy

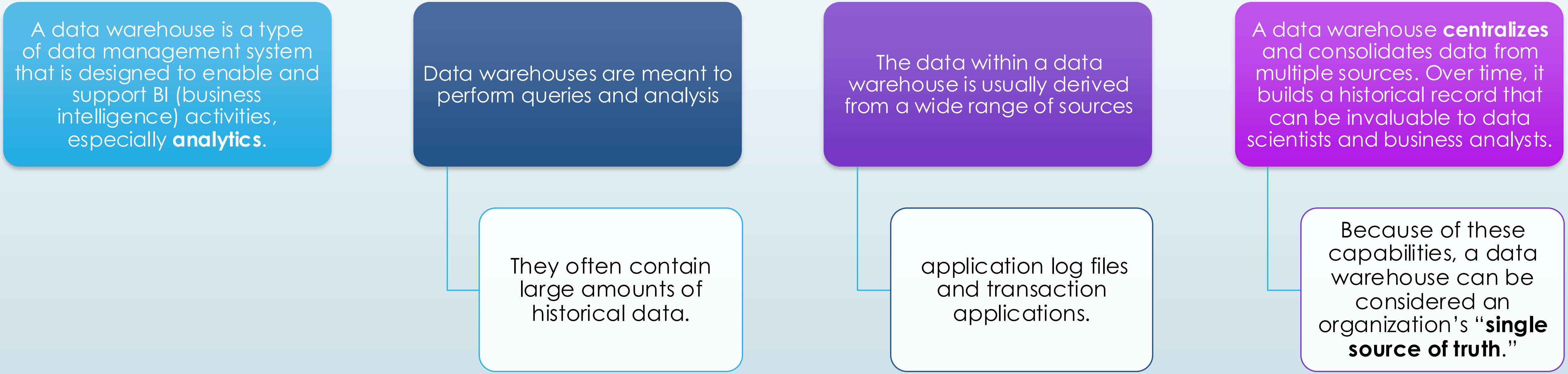
- Database
- Data Warehouse
- Data Lake

Databases

A database can be an organized collection of structured information, or data, typically stored electronically in a computer system. A database is usually controlled by a database management system (DBMS).[2]



Data Warehouses



A data warehouse is a type of data management system that is designed to enable and support BI (business intelligence) activities, especially **analytics**.

Data warehouses are meant to perform queries and analysis

They often contain large amounts of historical data.

The data within a data warehouse is usually derived from a wide range of sources

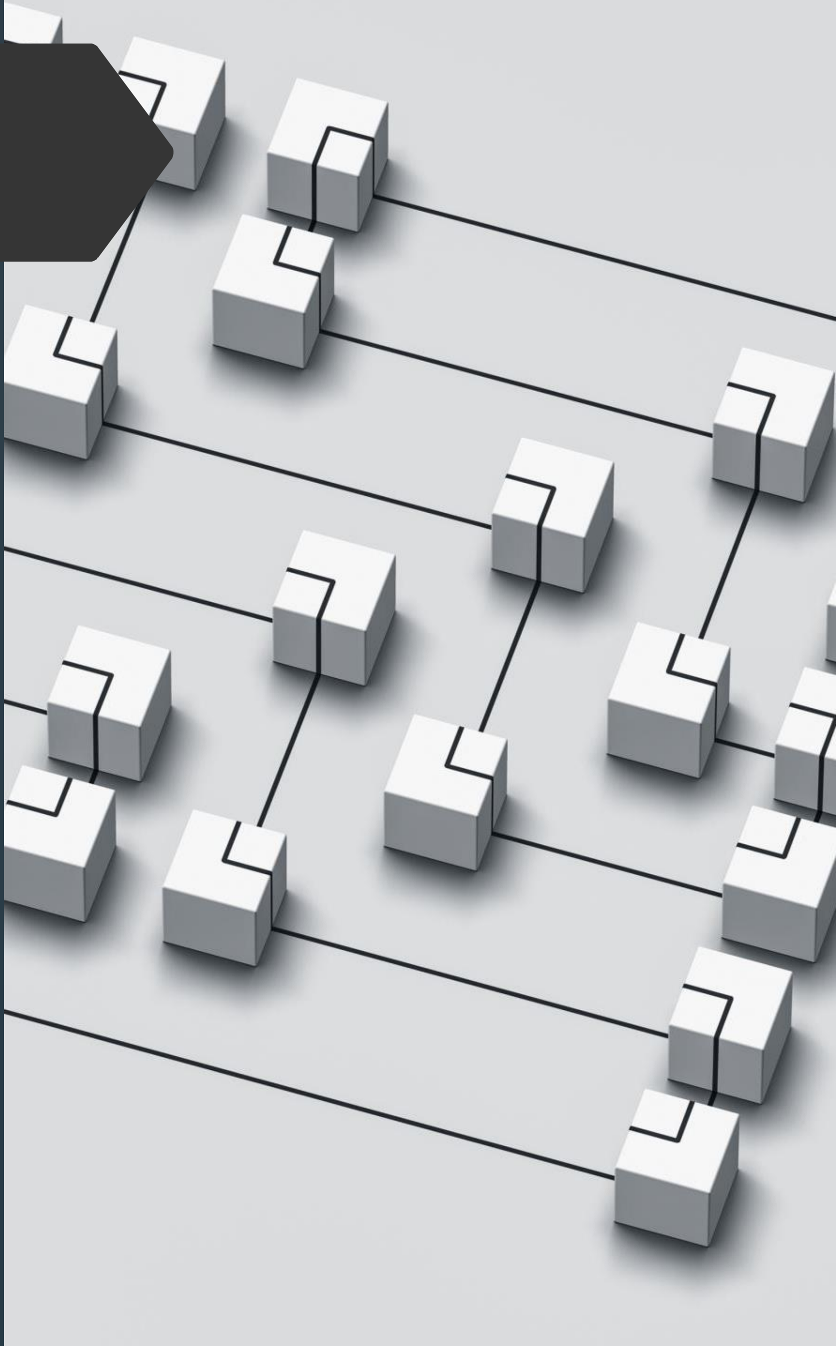
application log files and transaction applications.

A data warehouse **centralizes** and consolidates data from multiple sources. Over time, it builds a historical record that can be invaluable to data scientists and business analysts.

Because of these capabilities, a data warehouse can be considered an organization's "**single source of truth**."

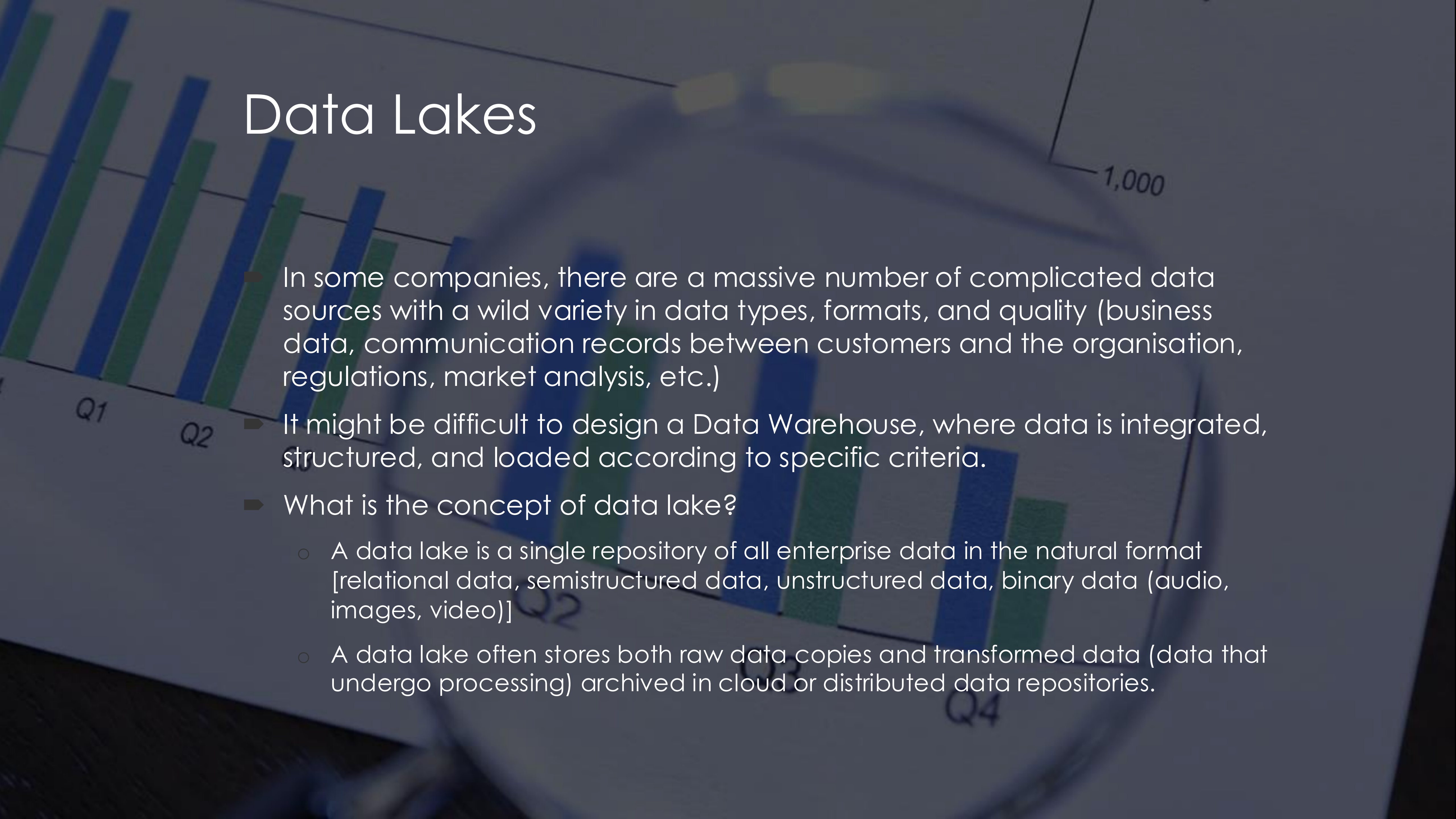
Data Warehouses

- **Enterprise Data Warehouse**: It collects all information about subjects spanning the entire organisation. It typically can range in size from hundreds of gigabytes to terabytes or beyond. It supports corporate-wide data integration (cross functional in its scope).
- **Data Mart**: A data mart contains a subset of corporate-wide data that is of value to a specific group of users, such as those within a business department. The scope is confined to specific subjects.
  - A markteing data mart may confine its subjects to customer, item, marketing channels and sales.
  - A risk control data mart may focus on customer credit, risk, and different type of frauds.



Data Lakes

- In some companies, there are a massive number of complicated data sources with a wild variety in data types, formats, and quality (business data, communication records between customers and the organisation, regulations, market analysis, etc.)
- It might be difficult to design a Data Warehouse, where data is integrated, structured, and loaded according to specific criteria.
- What is the concept of data lake?
  - A data lake is a single repository of all enterprise data in the natural format [relational data, semistructured data, unstructured data, binary data (audio, images, video)]
  - A data lake often stores both raw data copies and transformed data (data that undergo processing) archived in cloud or distributed data repositories.



Databases (DBs), Data Warehouses (DWs), and Data Lakes (DLs)

- Databases are for fast, transactional processing of structured data.
- Data Warehouses are for analytical querying of structured data with a top-down perspective (goal-oriented)
- Data Lakes store raw, unstructured, or semi-structured data for large-scale analytics.

Real Scenarios with DBs DWs and DLs

- **Database Transactional Processing**
- An e-commerce website employs a relational database (MySQL) to store and deal with customer orders, inventory, and user accounts. When a customer places an order, the database processes the transaction in real time, updating product quantities and recording order details.

- **Data Warehouse - Analytical Querying**
- A retail company uses Amazon Redshift as a data warehouse to analyze sales trends over the last five years. The data warehouse aggregates sales data from different stores, allowing the business to run complex queries to understand customer behavior, seasonal trends, and profitability.

- **Data Lake – Dealing with Raw, Unstructured Data**
- A video streaming service uses Amazon S3 as a data lake to store large volumes of raw video files, user interaction logs, and metadata. Data scientists access this raw data to run machine learning algorithms that recommend content to users based on their viewing history and preferences.

Data and Information Technology Evolution

Evolution of Information Technology

“Necessity, who is the mother of invention.” – Plato

Huge growth of available data volume

Computerization of our society

Fast development of powerful data collection and storage tools.

Some figures on Data Growth

720,000 hours of video are uploaded every day to YouTube

1.3 billion images are shared on Instagram daily

500 million tweets are posted daily

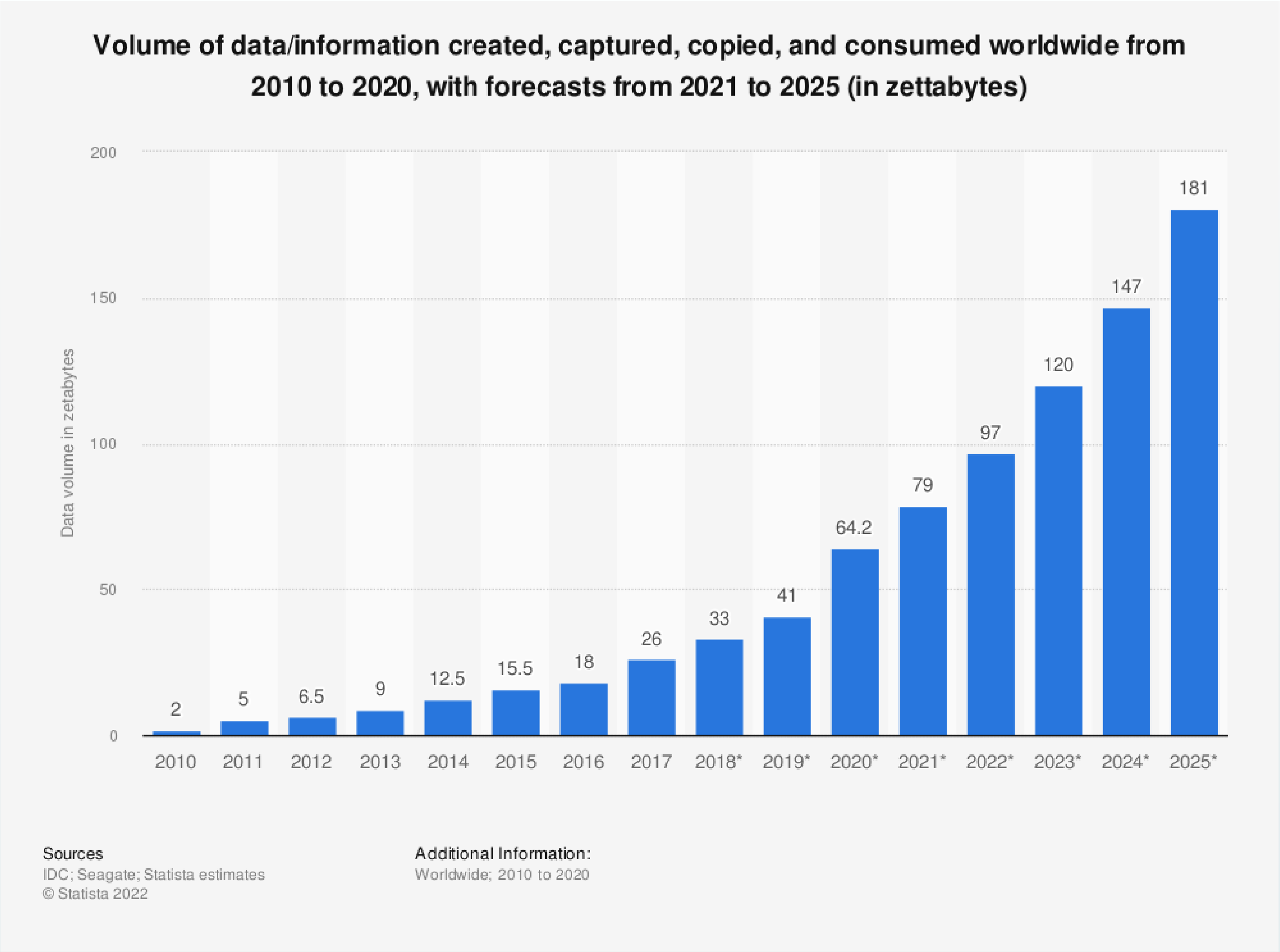
More than 300 million photos get uploaded to Facebook per day

Every minute there are 510,000 comments posted and 293,000 statuses updated

Data Age

Y-axis: Data Volume in Zettabytes

1 Zettabyte = 1 billion Terabytes

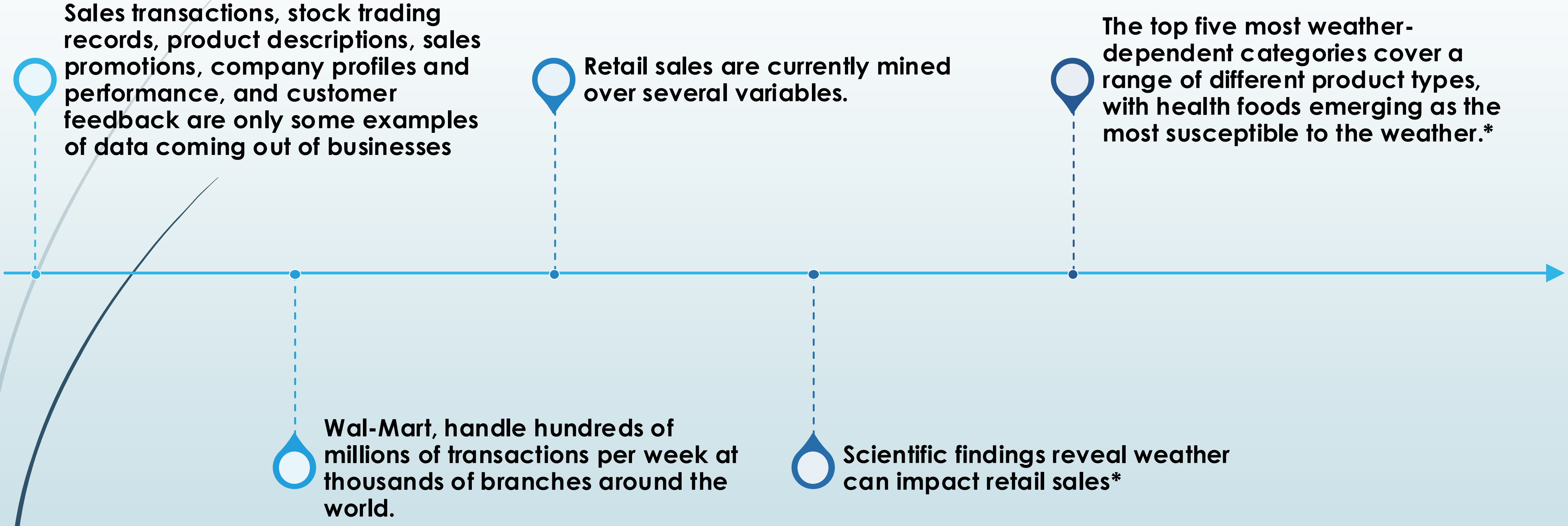


Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025 (in zettabytes)

Sources
IDC; Seagate; Statista estimates
© Statista 2022

Additional Information:
Worldwide; 2010 to 2020

Zooming in on Business Data



Sales transactions, stock trading records, product descriptions, sales promotions, company profiles and performance, and customer feedback are only some examples of data coming out of businesses

Wal-Mart, handle hundreds of millions of transactions per week at thousands of branches around the world.

Retail sales are currently mined over several variables.

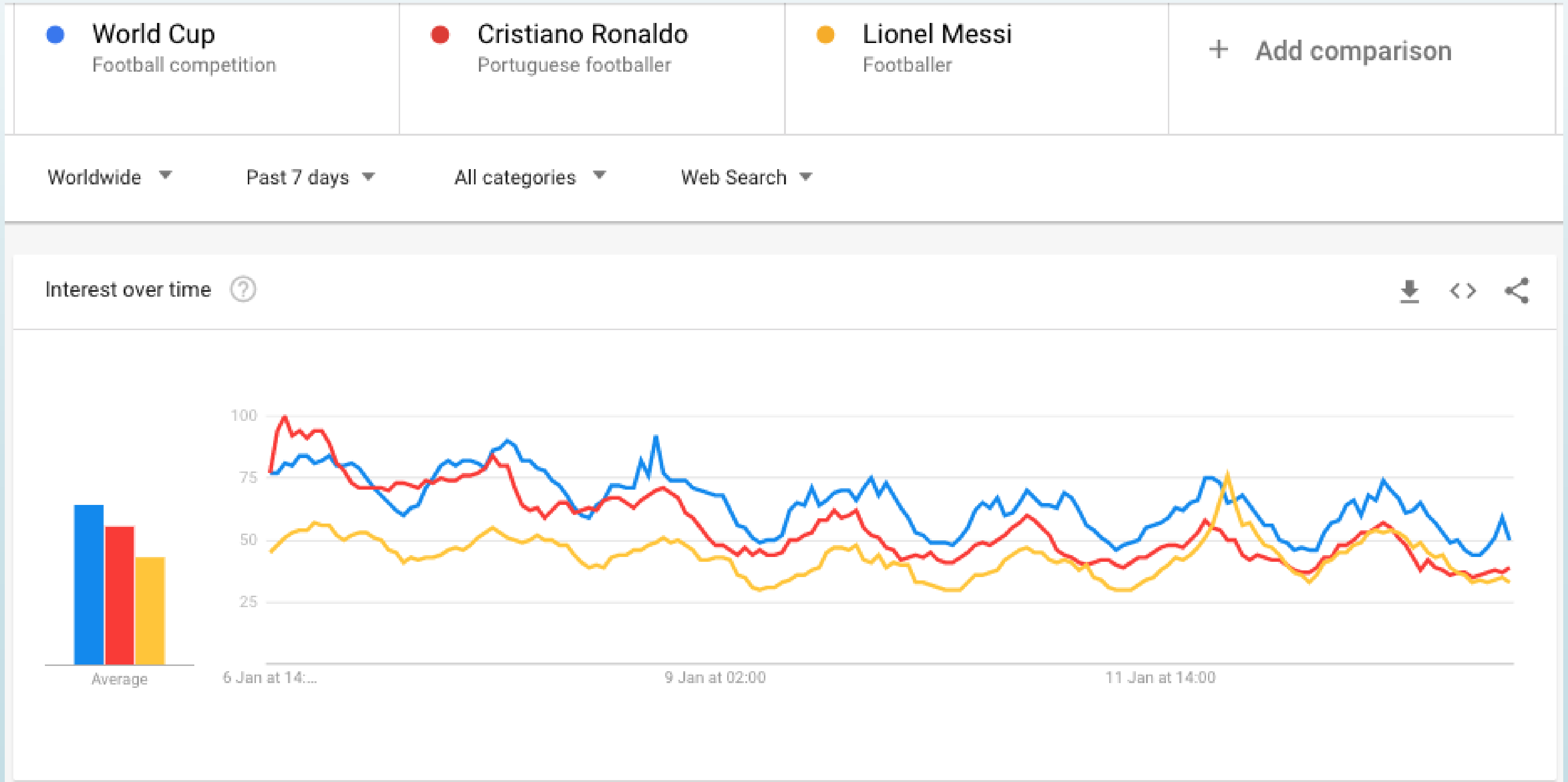
Scientific findings reveal weather can impact retail sales*

The top five most weather-dependent categories cover a range of different product types, with health foods emerging as the most susceptible to the weather.*

*Rose, N., Dolega, L. It's the Weather: Quantifying the Impact of Weather on Retail Sales. *Appl. Spatial Analysis* **15**, 189–214 (2022). https://doi.org/10.1007/s12061-021-09397-0

Exploring queries over time is a good starting point

- Extracting stats out of queries can tell about what the current and past trends are.
- Give a go at **Google Trends** https://trends.google.com/trends/?geo=IT
- Users can easily compare different trends using keywords and see percentages over time.



🔵 World Cup
Football competition

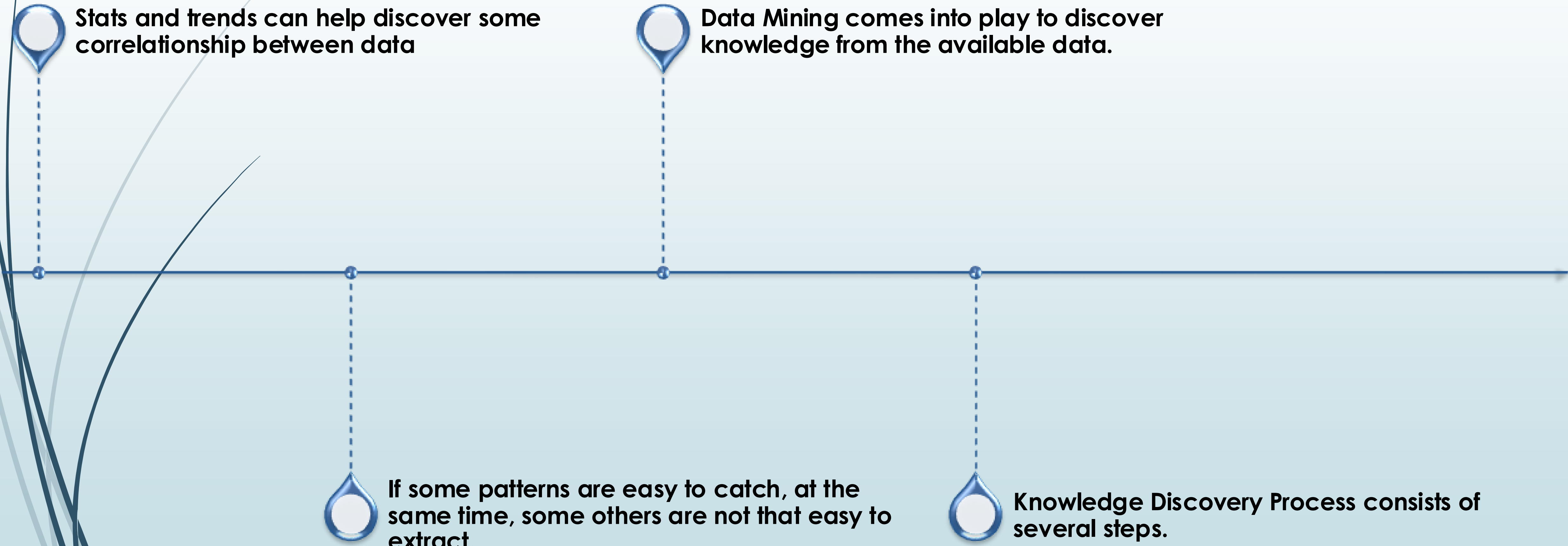
🔴 Cristiano Ronaldo
Portuguese footballer

🟡 Lionel Messi
Footballer

Worldwide ▾ Past 7 days ▾ All categories ▾ Web Search ▾

Interest over time

Stats and Trends

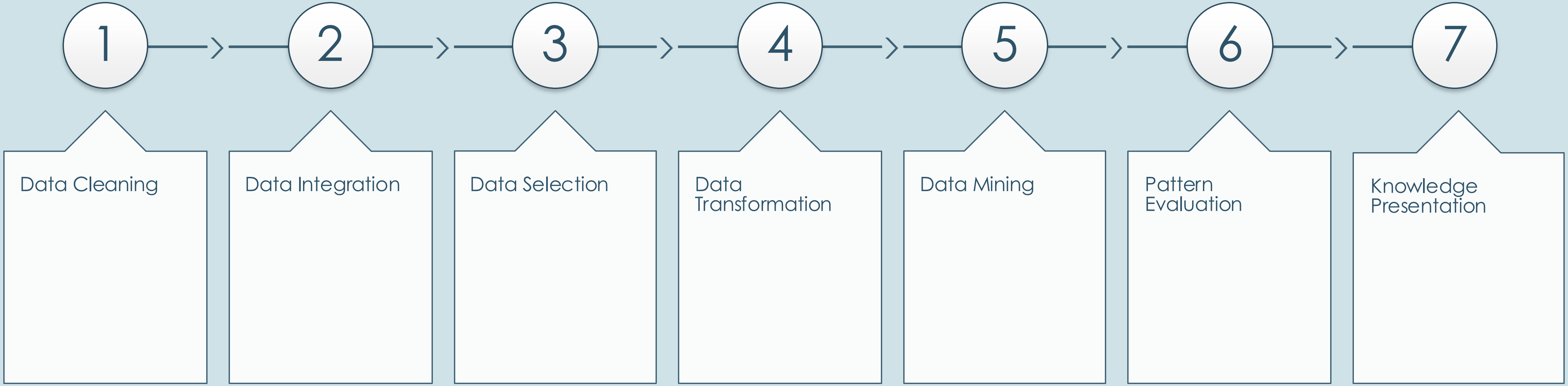


Stats and trends can help discover some correlationship between data

If some patterns are easy to catch, at the same time, some others are not that easy to extract.

Data Mining comes into play to discover knowledge from the available data.

Knowledge Discovery Process consists of several steps.



1 Data Cleaning

2 Data Integration

3 Data Selection

4 Data Transformation

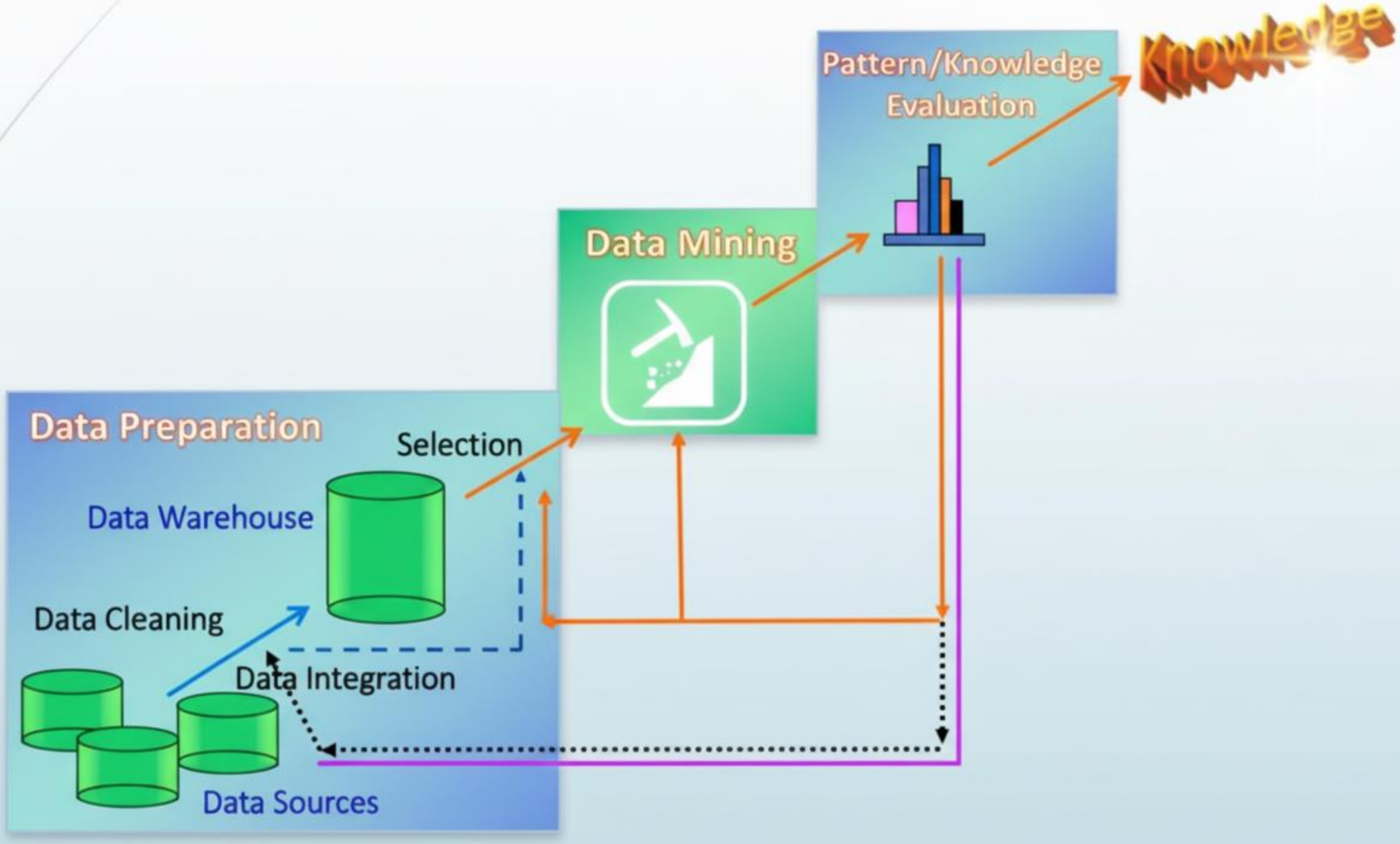
5 Data Mining

6 Pattern Evaluation

7 Knowledge Presentation

Knowledge Discovery Process

Knowledge Discovery Process



KDD steps and explanation

Data Cleaning

- (Removing noise and inconsistent data)

Data Integration

- (where multiple data sources may be integrated)

Data Transformation

- (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)

KDD steps and explanation

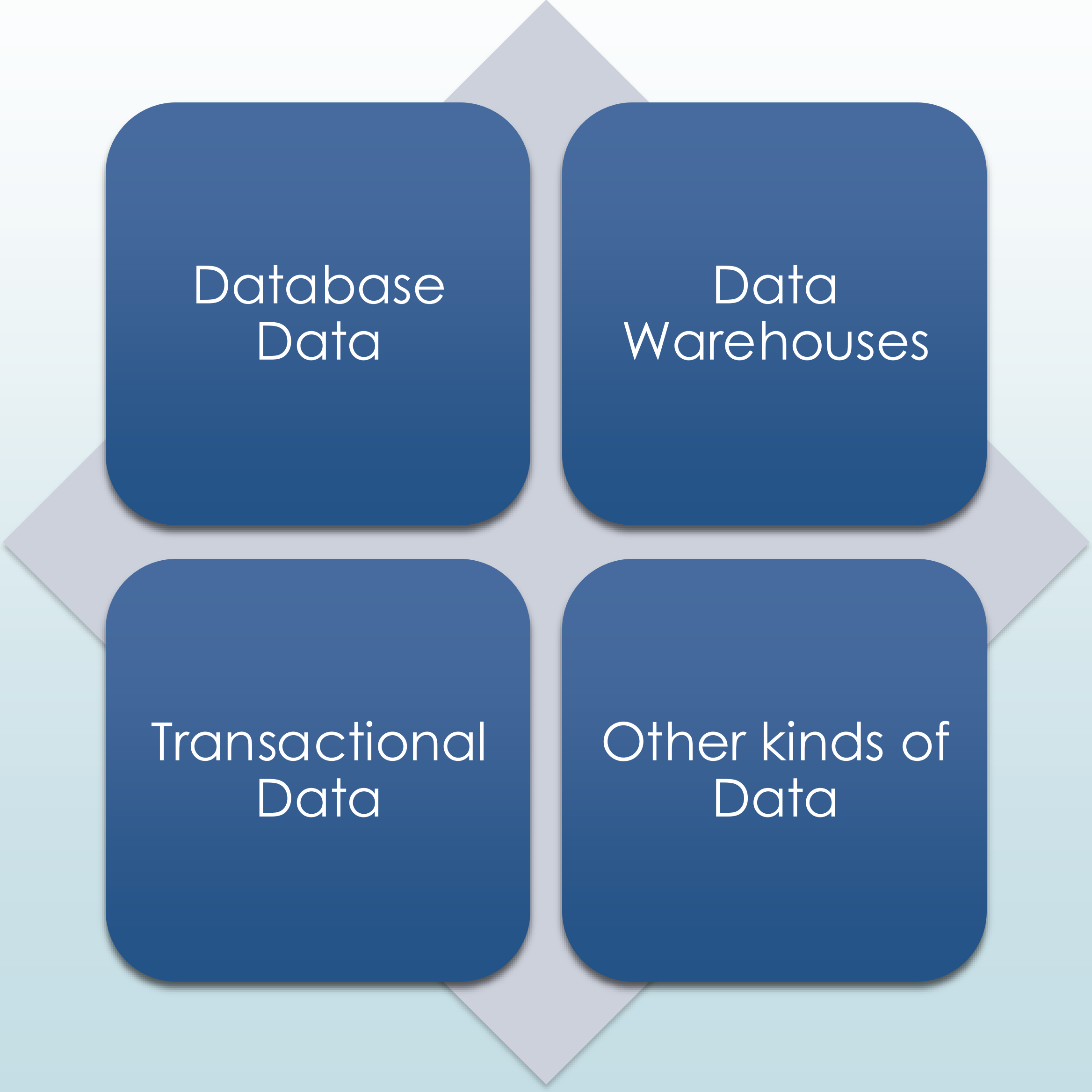
Data Mining	• (Intelligent methods are applied to extract patterns or construct models)
Pattern/Model Evaluation	• (Identifying interesting patterns or models representing knowledge based on specific measures)
Knowledge Presentation	• (Visualisation and Knowledge Representation are used to show mined knowledge to users)

Descriptive vs Predicitive Data Mining

Descriptive Mining deals with properties of the dataset of interest

Predictive Mining performs induction on the dataset to make prediction

What kinds of data can be mined?



Database Data

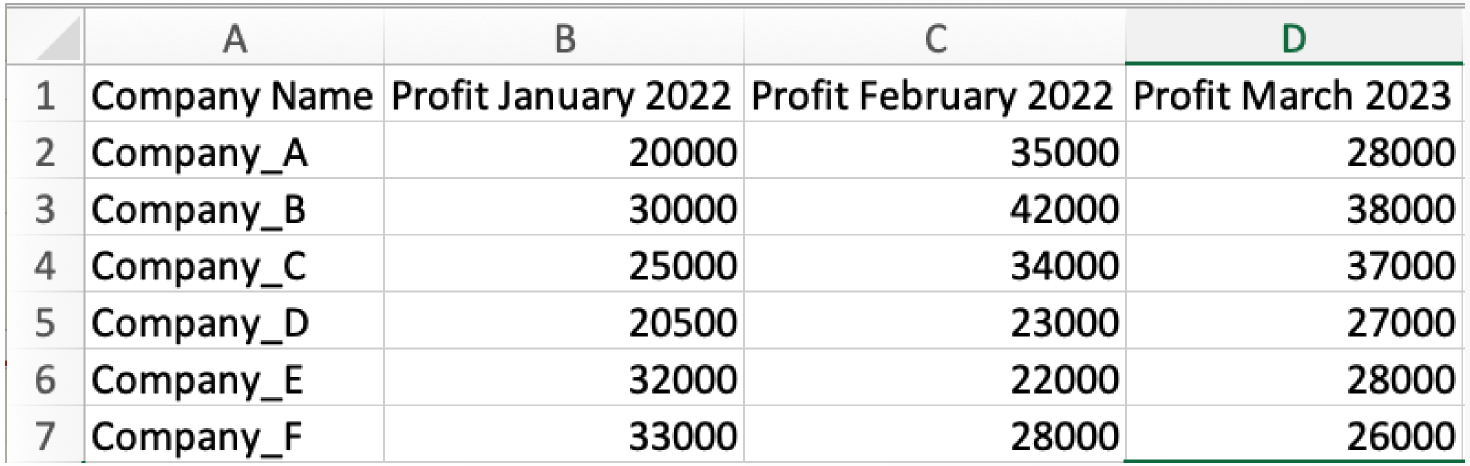
Data Warehouses

Transactional Data

Other kinds of Data

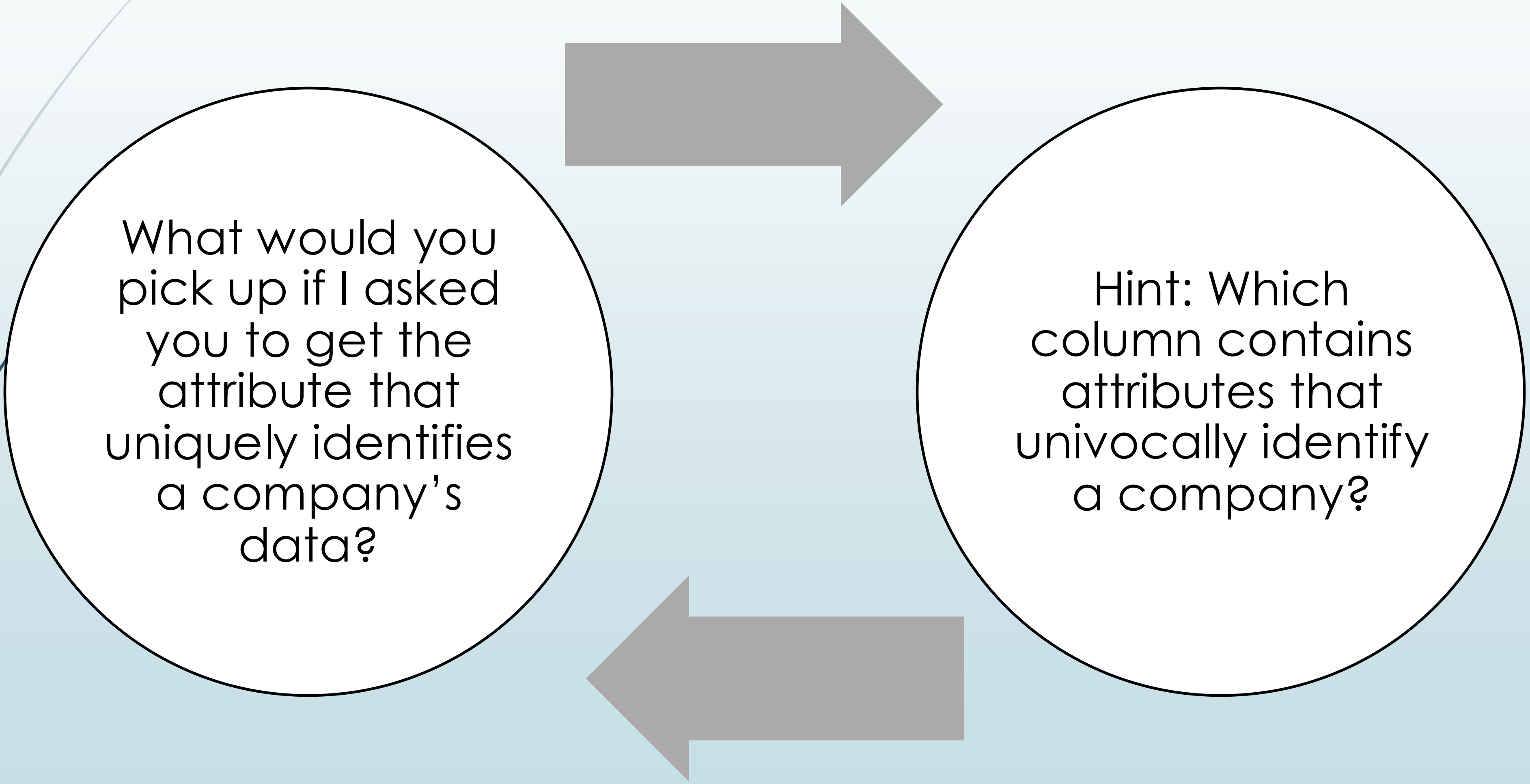
Database Data

- A database system, also called a **database management system (DBMS)**, consists of a collection of interrelated data, known as a database, and a set of software programs to manage and access the data.
- Imagine having something arranged similarly to spreadsheet files with columns and rows showing how data are handled.



	A	B	C	D
1	Company Name	Profit January 2022	Profit February 2022	Profit March 2023
2	Company_A	20000	35000	28000
3	Company_B	30000	42000	38000
4	Company_C	25000	34000	37000
5	Company_D	20500	23000	27000
6	Company_E	32000	22000	28000
7	Company_F	33000	28000	26000

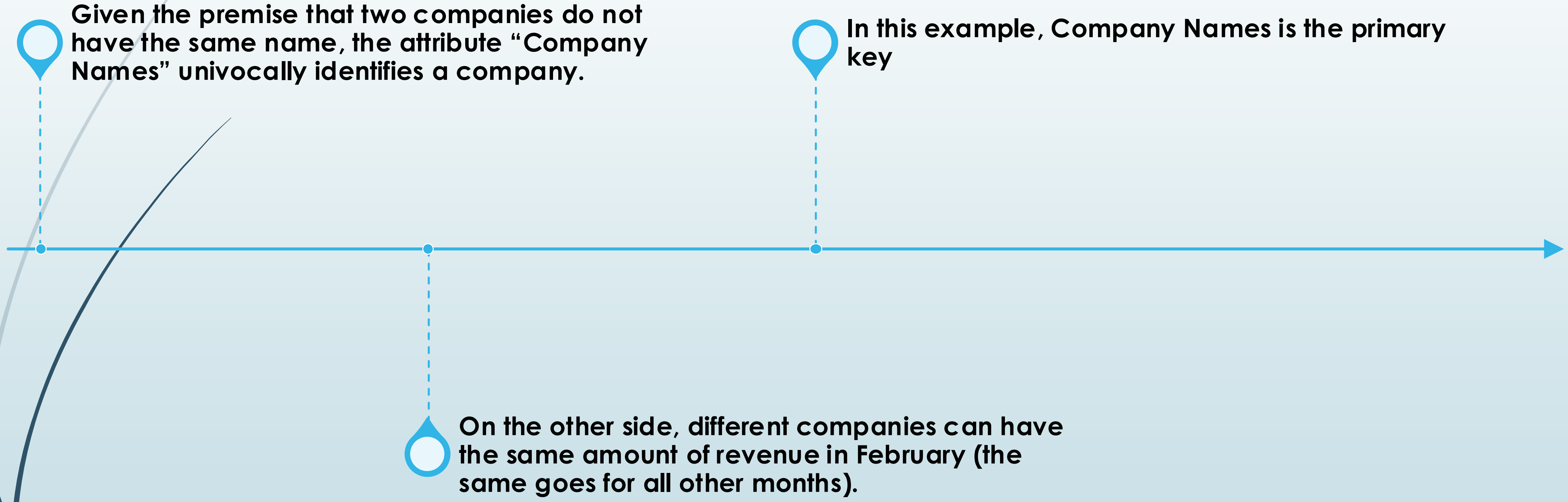
Database Data



What would you pick up if I asked you to get the attribute that uniquely identifies a company's data?

Hint: Which column contains attributes that univocally identify a company?

Answer

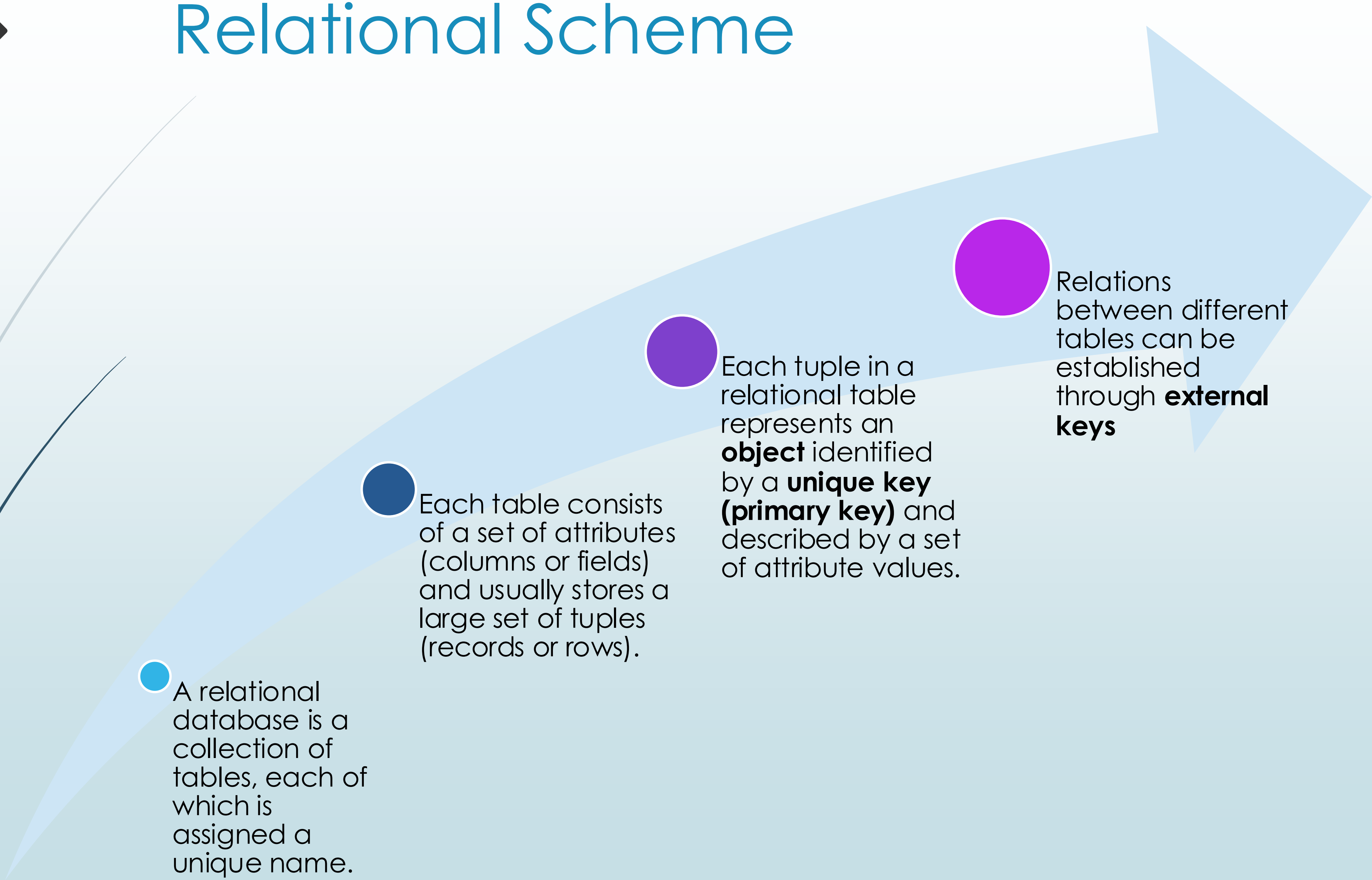


Given the premise that two companies do not have the same name, the attribute “Company Names” univocally identifies a company.

On the other side, different companies can have the same amount of revenue in February (the same goes for all other months).

In this example, Company Names is the primary key

Relational Scheme



A relational database is a collection of tables, each of which is assigned a unique name.

Each table consists of a set of attributes (columns or fields) and usually stores a large set of tuples (records or rows).

Each tuple in a relational table represents an **object** identified by a **unique key (primary key)** and described by a set of attribute values.

Relations between different tables can be established through **external keys**

Relational Database for AllElectronics (example)

Consider AllElectronics being a fictious company for practical examples in this Unit.

The company is described by the following relation tables: **customer**, **item**, **employee**, and **branch**.

Database Logic Scheme

- When one deals with Databases, it is recommended having a look at the so-called Logic Schemes, which includes the following:
- **Tables (Entities)**: Data points in the system.
- **Attributes (Fields)**: Properties or characteristics of the entities.
- **Relationships**: how entities are related to each other (e.g., one-to-one, one-to-many).
- **Constraints**: Rules that govern how data can be entered or linked (e.g., primary keys, foreign keys, unique constraints).

AllElectronics relation tables

- Here are the **logic schemes** related to *AllElectronics* relation tables:
- ***customer*** (*cust ID, name, address, age, occupation, annual income, credit information, category*, . . .)
- ***item*** (*item ID, brand, category, type, price, place made, supplier, cost*, . . .)
- ***employee*** (*empl ID, name, category, group, salary, commission*, . . .)
- ***branch*** (*branch ID, name, address*, . . .)
- ***purchases*** (*trans ID, cust ID, empl ID, date, time, method paid, amount*)
- ***items sold*** (*trans ID, item ID, qty*)
- ***works_at*** (*empl ID, branch ID*)

- In simple words, each of the above lines corresponds to a table with the within-brackets attributes.

Relational Scheme for a relational database, AllElectronics

Access to Relational Databases

Structured Query-based Languages like **SQL** allow accessing DBs (Databases) in order to run some queries.

An SQL query allows retrieving information out of databases.

Join, **Selection**, **Projection** are the most popular relational-based query instructions

Example: Show me a list of all items that have been sold since January 2022"

It should be noted that relational languages also use aggregate functions such as sum, avg (average), count, max (maximum), and min (minimum).

Example: Show me the total sales of the last month grouped by branch

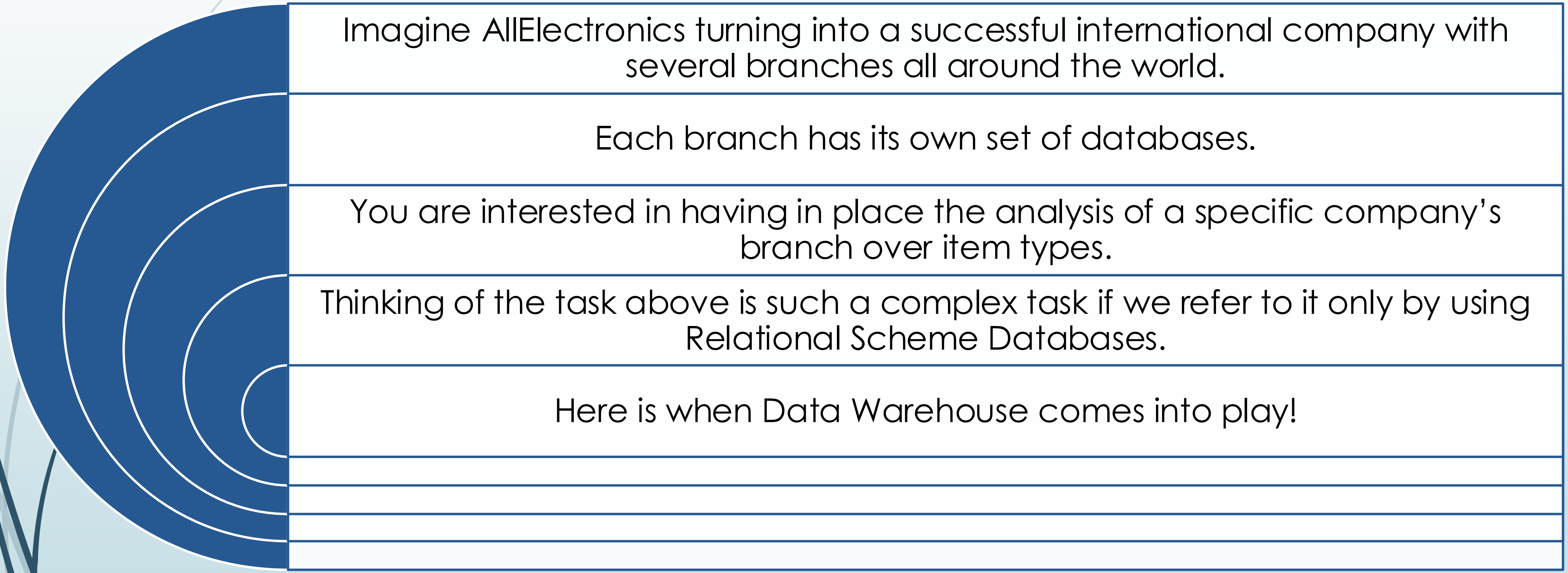
Quiz time!

https://forms.gle/sHYeh1GbGaajVwco6

What about Data Mining Relational Databases?

- Using data mining to search for data **patterns**
- Data Mining can analyse customer data to predict the credit risk of new customers based on income, age, previous credit information.
- Data **deviation** is another essential task accomplished by Data Mining.
- Items with sales far from those expected in comparison with the previous year
- Deviations like that can be further explored
- Data Mining may discover there has been a change in packaging.

Data Warehouses



Imagine AllElectronics turning into a successful international company with several branches all around the world.

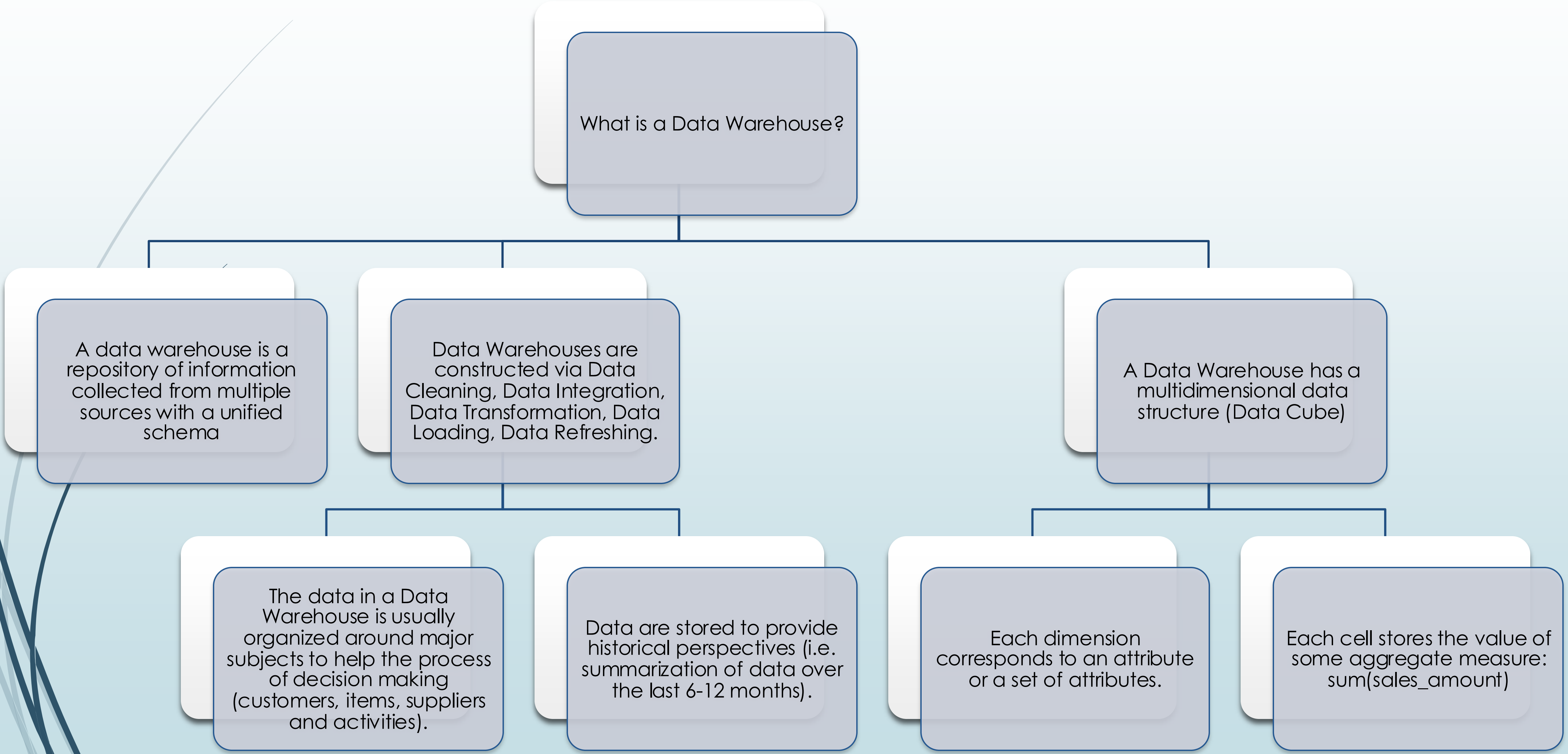
Each branch has its own set of databases.

You are interested in having in place the analysis of a specific company’s branch over item types.

Thinking of the task above is such a complex task if we refer to it only by using Relational Scheme Databases.

Here is when Data Warehouse comes into play!

Data Warehouses



- What is a Data Warehouse?
  - A data warehouse is a repository of information collected from multiple sources with a unified schema
  - Data Warehouses are constructed via Data Cleaning, Data Integration, Data Transformation, Data Loading, Data Refreshing.
    - The data in a Data Warehouse is usually organized around major subjects to help the process of decision making (customers, items, suppliers and activities).
    - Data are stored to provide historical perspectives (i.e. summarization of data over the last 6-12 months).
  - A Data Warehouse has a multidimensional data structure (Data Cube)
    - Each dimension corresponds to an attribute or a set of attributes.
    - Each cell stores the value of some aggregate measure: sum(sales_amount)

Data Warehouse – Drill-down, Roll-Up

- A Data Warehouse allows drill-down and roll-up, corresponding to different views of data.
- Drill-down dissects data by zooming in on certain attributes
- On the other side, Roll-up allows encompassing attributes into a more compact representation.
- Data warehouses provide **online analytical processing** (**OLAP**) tools for the interactive analysis of multidimensional data of varied granularities
- Example:
  - *AllElectronics*' Sales are stored according to item (types), time (quarters), address (cities)

Address (cities)

Time (quarters)

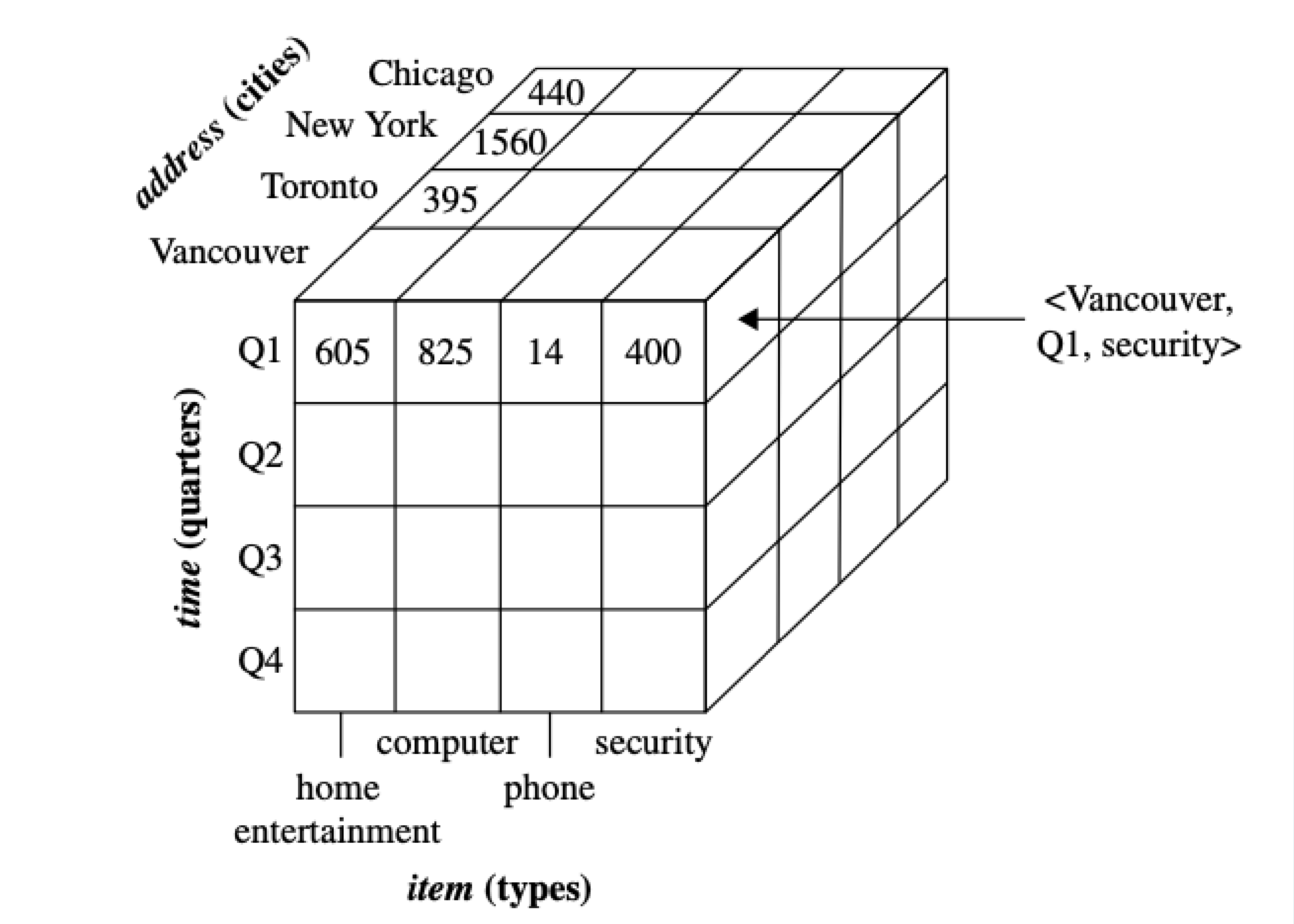
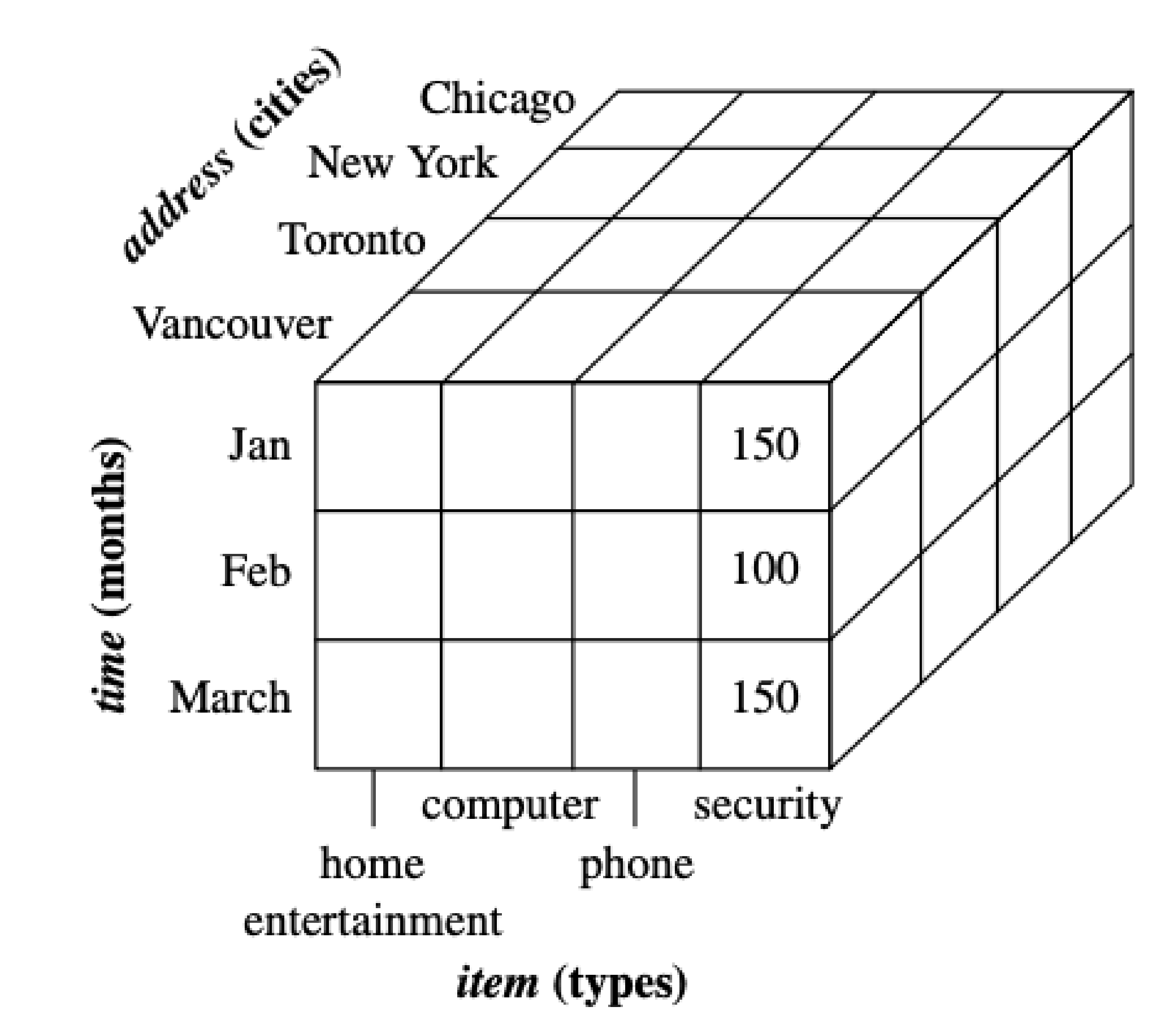


Figure 1.7 take from “Data Mining. Concepts and Techniques”.

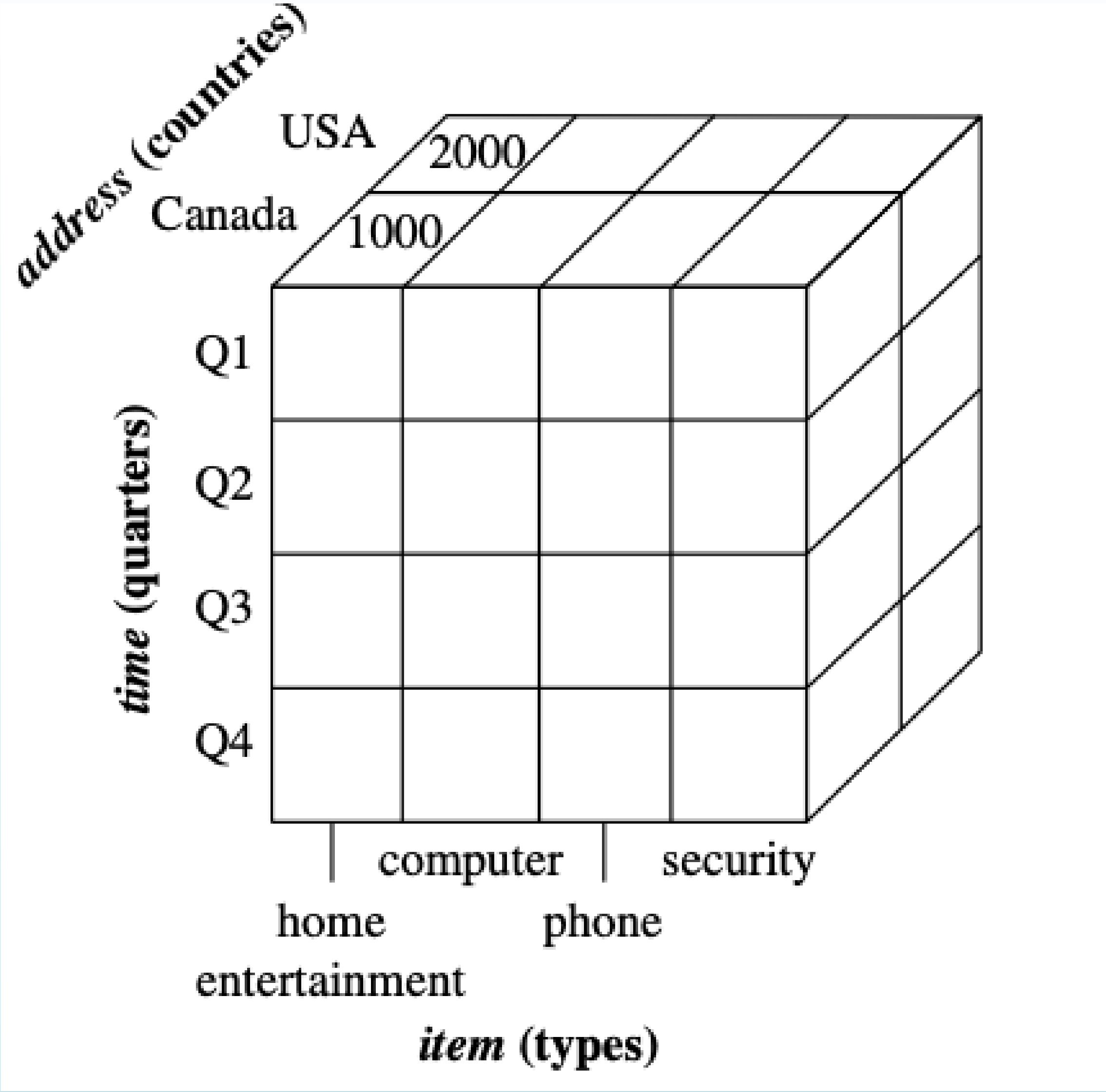
Drill-down on Time data for Q1

Time (months)



Roll-up on address

Address (Countries)

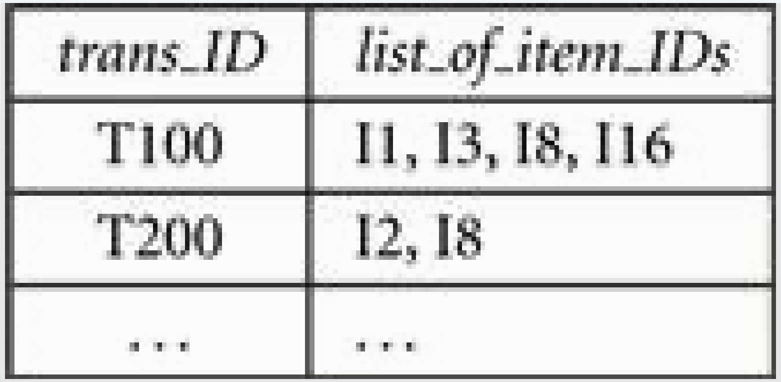


Transactional Data

- Each record in a transactional database **captures a transaction**, such as a customer’s purchase, a flight booking, or a user’s clicks on a web page.
- It generally includes a **transaction ID** and a list of items making up the transaction (the items purchased in the transaction).
- It should be noted that **Transactional Databases** can contain other **tables** related to purchased items (item **description**, **information about** the shop or salesperson).
- An example of transactional database for *AllElectronics* is now given

AllElectronics Transactional DB

- Example: Zooming in on AllElectronics Transactional DB it can be noticed a column for transaction ID and a second one for the list of item IDs.



trans_ID	*list_of_item_IDs*
T100	I1, I3, I8, I16
T200	I2, I8
...	...

- The information provided by transactional DB like this allows bundling items into a market bucket. That is a pragmatic example for for boosting sales.

*Jiawei Han. "Data Mining"

Other Kinds of Data

- Time—related or sequence data (stock exchange; historical records; time series; biological sequence data)
- Hypertext and multimedia data (text; image; video and audio data)
- Graph and networked data (Social and Information networks)

- Stock Exchange can be mined to retrieve trends for investment strategies
- Computer Network data streams can be mined to check anomalies in the network (Intrusion detection systems)
- Images can be mined with machine learning methods to classify objects in them. Semantic labels are then assigned to categorise an image dataset.

What Patterns can be mined?

Classes or Concepts

Associations and Correlations

Classification and Regression

Cluster Analysis

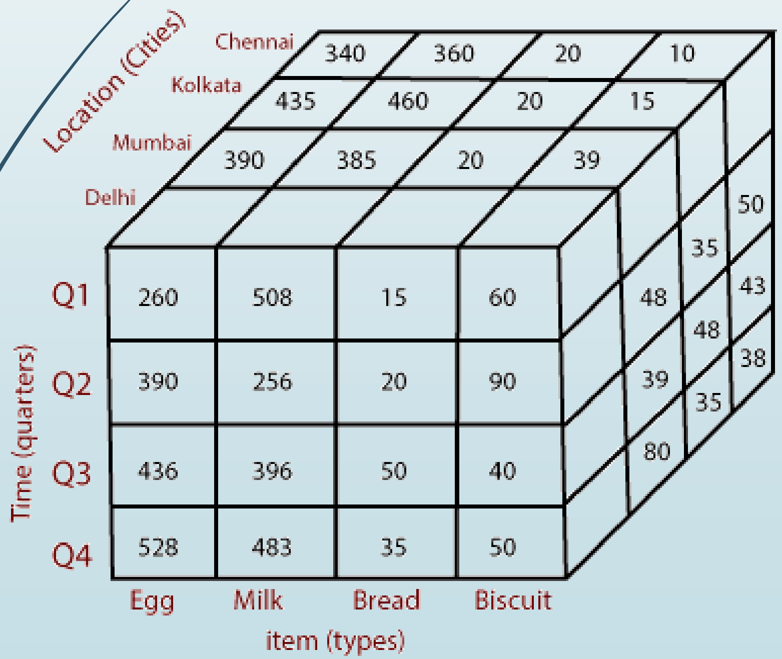
Outlier Analysis

Class and Concepts

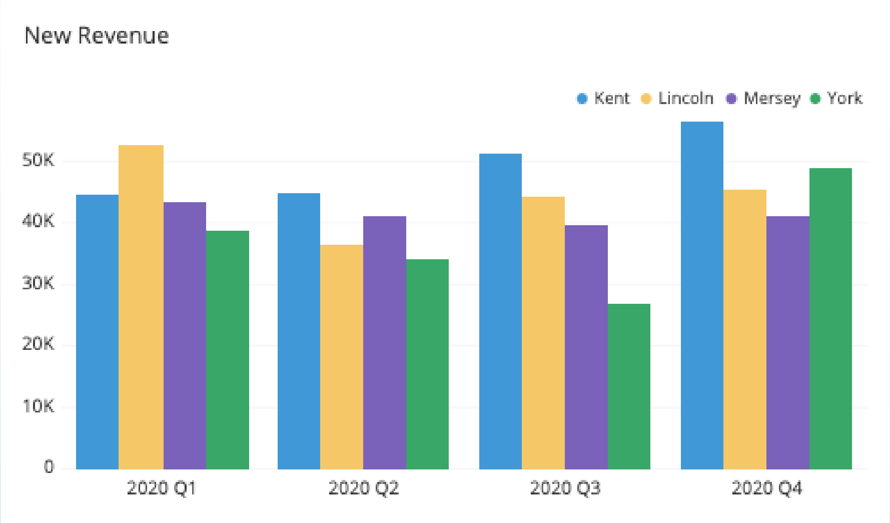
- Data Entries can be associated with classes or concepts.
- AllElectronics store sells several items that can be grouped into categories according to their cost (BigSpender, budgetSpenders).
- Individual classes and concepts are summarised with detailed terms.
- Such descriptions are called "**concepts**" or "**class**".
- These description can be derived using data characterisation or by using data discrimination
- Data **Characterisation** is a **summarization** of the general characteristics of a target class of data (characteristics of software tools with sales incremented by 10% over the last year).
- Data **Discrimination** is a **comparison** of the general features of the target data class objects

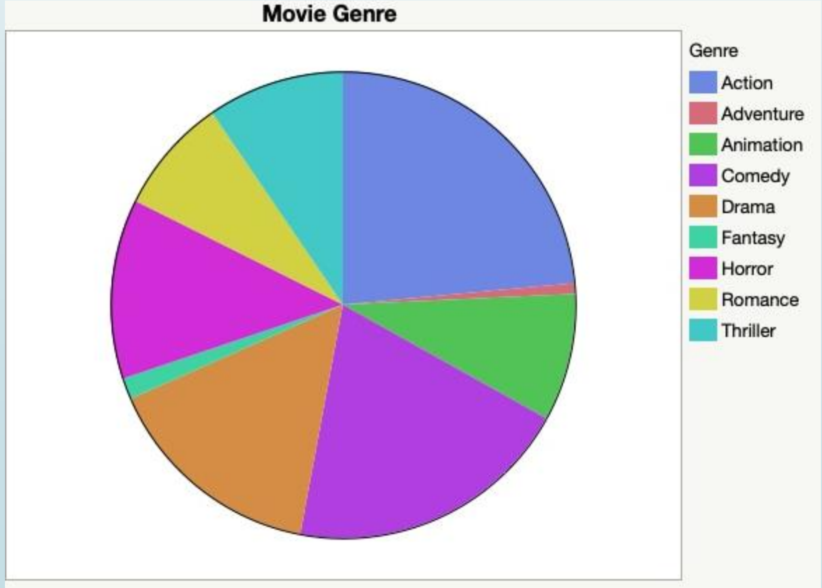
Characterisation & Discrimination

- Data Characterisation and Discrimination output examples:
  - pie charts, bar charts, curves, multidimensional data cubes, and multidimensional tables



	Egg	Milk	Bread	Biscuit
Q1	260	508	15	60
Q2	390	256	20	90
Q3	436	396	50	40
Q4	528	483	35	50





Examples

- **Data Characterisation Task:**
- *Summarize the characteristics of customers who spend more than $5000 a year at AllElectronics.*
- Result: A general profile of these customers (e.g. they are in the range 40 through 40 years of age, employed, excellent credit ratings)
- **Data Discrimination Task:**
- *A customer relationship manager at AllElectronics wants to compare two groups of customers according to their purchased items. In greater detail, the manager wants to compare the customers who regularly buy items and those who rarely shop the same items.*
- Result: A comparative profile of these customers. For instance, 80% of the customers who frequently purchase computer products are in the range 20 through 40 years of age, they have an university education. 60% of customers who rarely purchase the same items have no university degree and are older/younger than the first customers' group.

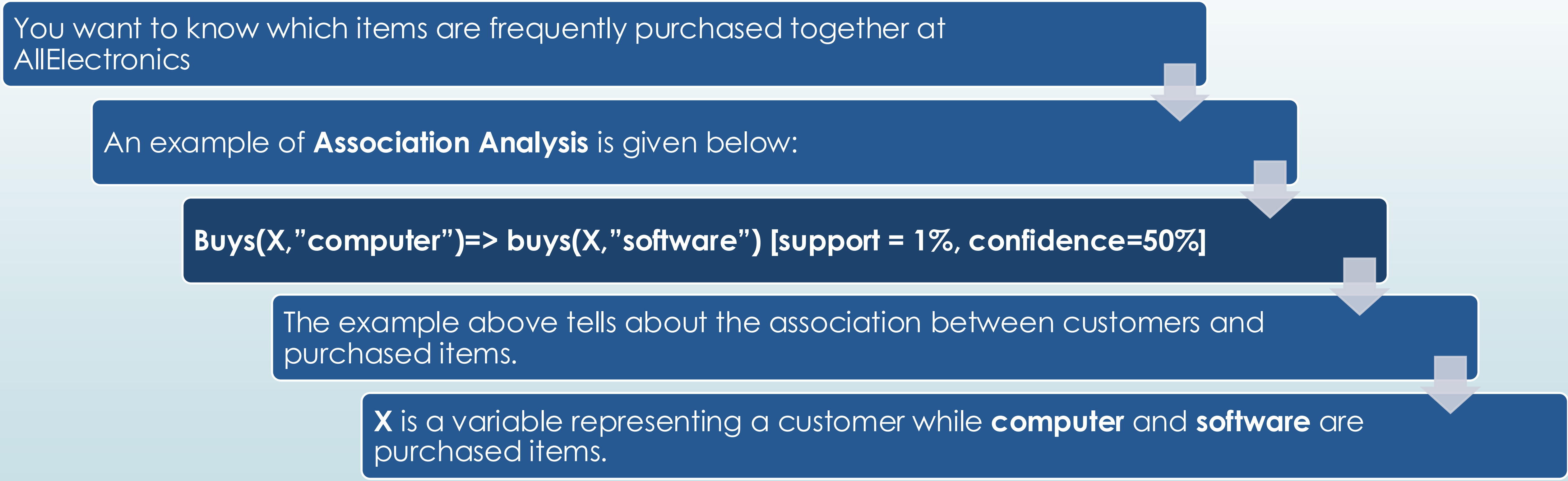
Frequent Patterns

Patterns in data can be found in the form of items that are frequently present in databases.

Set of Items often appear regularly together in a transactional dataset (milk and bread are usually purchased together).

Mining frequent patterns allows extracting associations and correlations within data.

Example: Association Analysis



You want to know which items are frequently purchased together at AllElectronics

An example of **Association Analysis** is given below:

Buys(X,”computer”)=> buys(X,”software”) [support = 1%, confidence=50%]

The example above tells about the association between customers and purchased items.

X is a variable representing a customer while **computer** and **software** are purchased items.

Buys(X,”computer”)=> buys(X,”software”) [support = 1%, confidence = 50%]

Support = 1% means that 1% of all transactions analysed show that computer and software are purchased together.

50% of **confidence** means that if a customer buys a computer there is a 50% chance that they will purchase software as well.

Associations are usually taken in consideration only if the confidence is above a certain **threshold**.

Example: Association Analysis

Classification for Predictive Analysis

What is classification?

- Classification is the process of finding a model or a function that distinguishes data classes or concepts.
- In order to run classification tasks, a model needs to trained over the so-called training set (the items in the training set are with known labels)
- Running a classification task means to make predictions on new input data by using a model that inferred knowledge from the training set.

Classification can be carried out using several approaches:

- IF THEN ELSE (rule-based classification)
- Decision Tree (A network with nodes representing several tests)
- Neural Network (Training sets are needed to infer knowledge from labelled data).

Regression

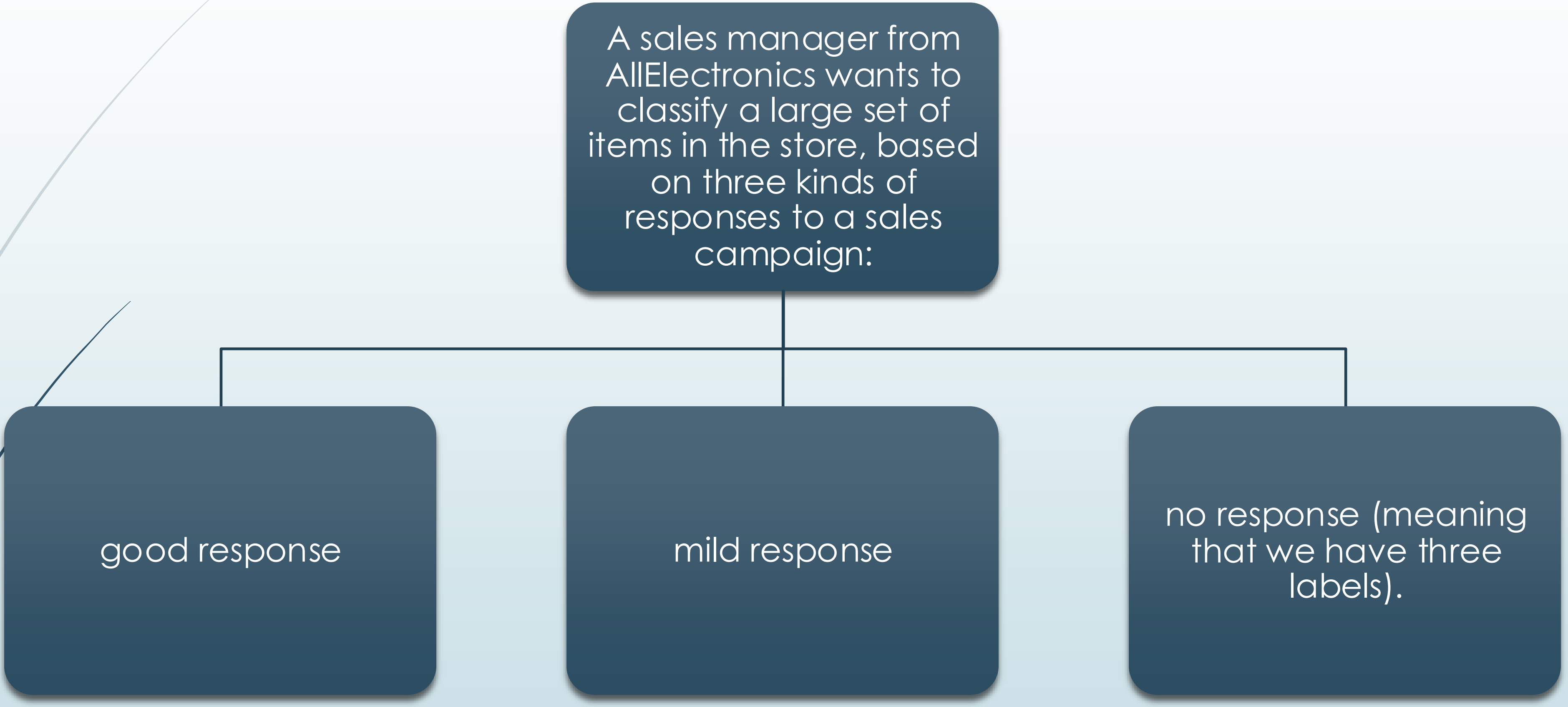
Whereas Classification returns discrete values (labels), regression functions return values in a continuous range.

Regression is used to predict missing or unavailable numerical data values rather than class labels.

Regression analysis is a statistical methodology that is most often used for numeric prediction. It also deals with the identification of **distribution trends**.

Classification needs labelled data, Regression ingests numbers to make statistical predictions (for instance, predictions on future trends based on past data sequences).

Example: Classification Vs Regression



A sales manager from AllElectronics wants to classify a large set of items in the store, based on three kinds of responses to a sales campaign:

good response

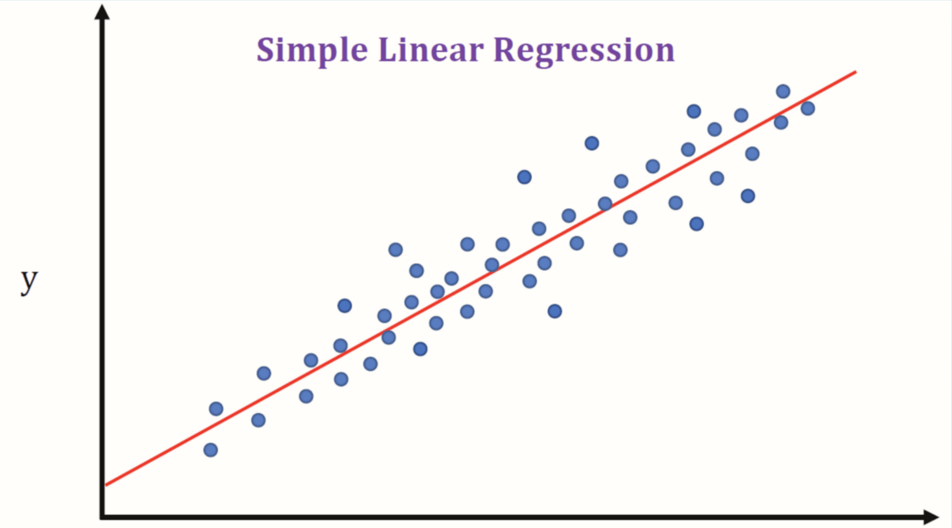
mild response

no response (meaning that we have three labels).

In this case, a classification system will return one out of the given classes

Example: Classification vs Regression

- Suppose you are asked to forecast the possible revenue for an upcoming sale of a specific item.
- What you are supposed to do is to show forecasts across different months
- In cases like this, classification does not suit the purpose!
- **Regression provides a curve in a diagram.**



Cluster Analysis

Unlike classification and regression, cluster analysis groups data samples without consulting labels.

In some case scenarios, labelled data are not available.

Cluster analysis can be used to generate labelled data.

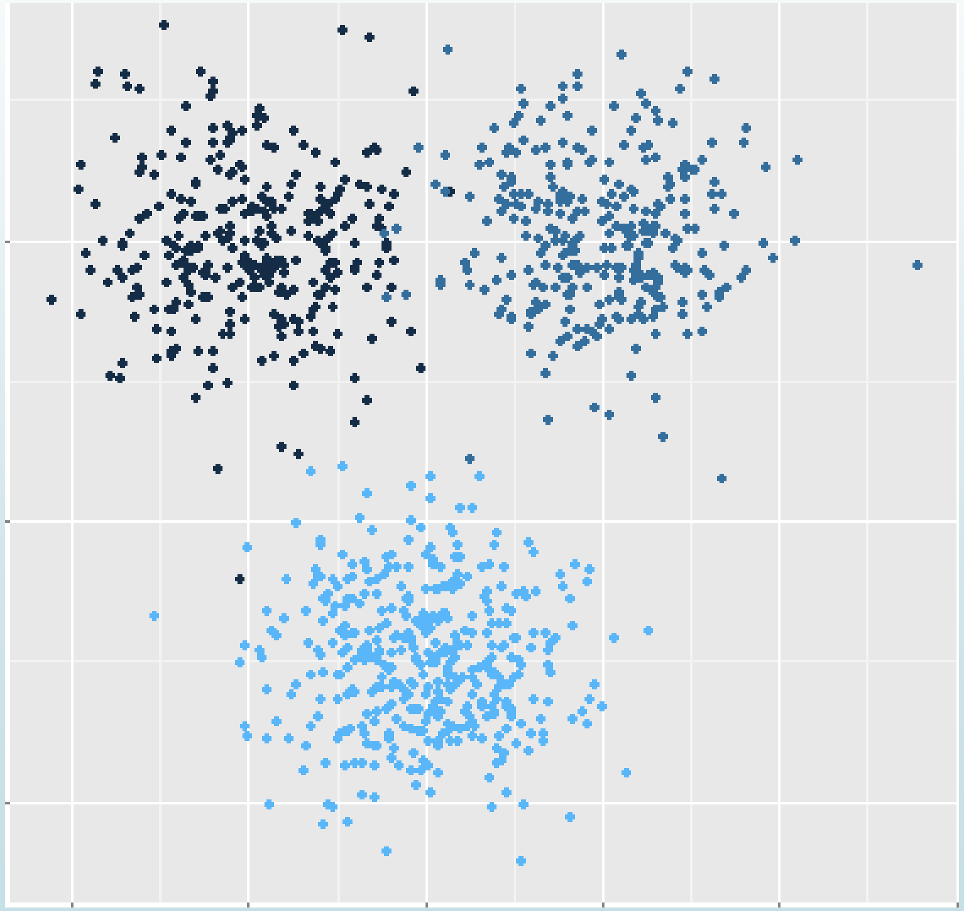
Example: Suppose you are asked to identify subpopulation groups from the customers of your company. For instance, you might be asked to divide the company's customers into three different groups according to customers' locations.

Cluster Analysis

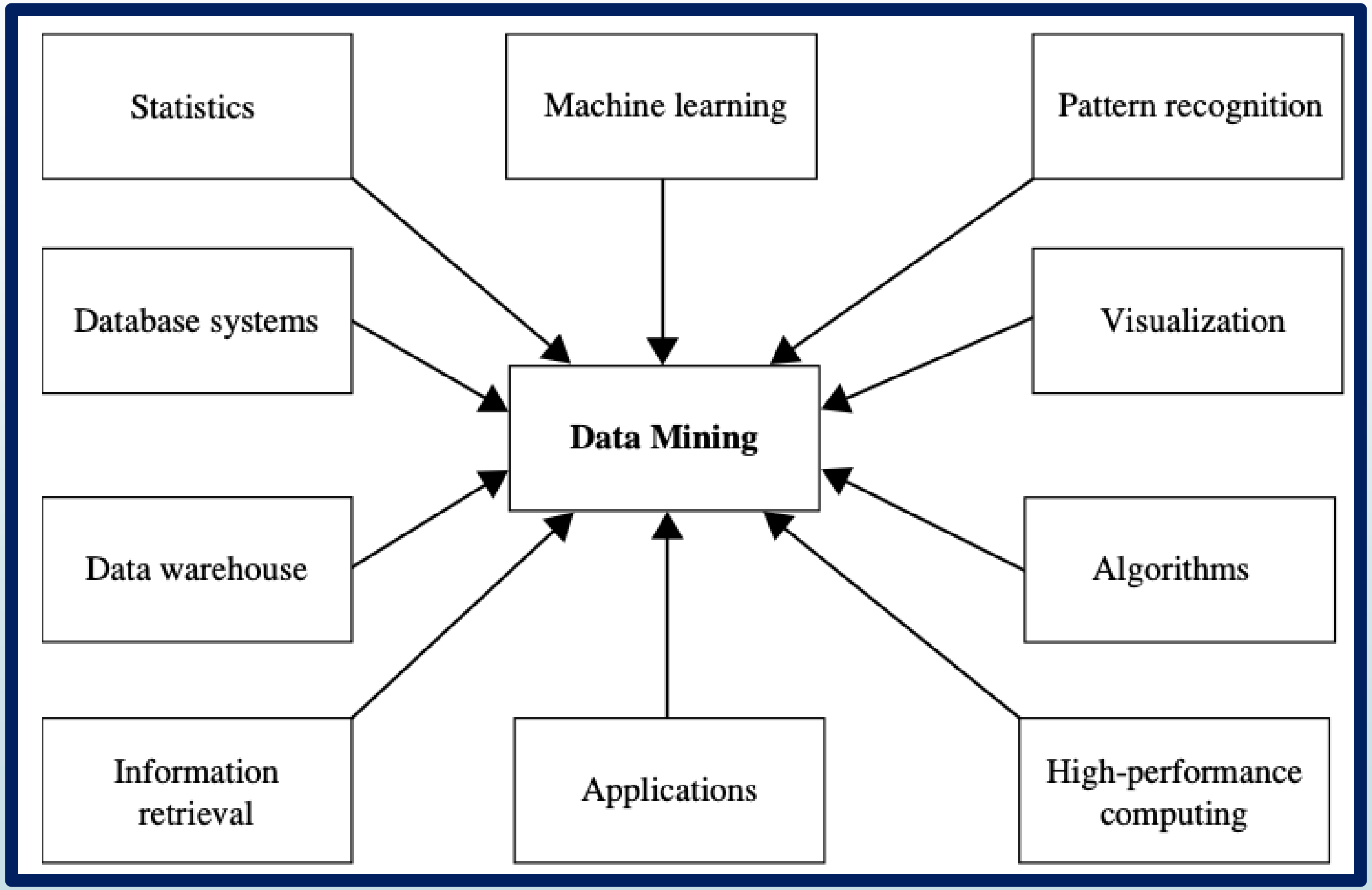
● Customers from Italy

● Customers from Spain

● Customers from Portugal



What Technologies are used?



Statistics

Data Mining has an inherent connection with statistics:

Statistic studies collections, analysis, interpretation of data

A statistical model is a **Set of Mathematical Functions** that describe the behaviour of variables in a dataset.

The behaviour is described with random variables and associated probability distributions.

Statistics can be used to represent a data noise model

Statistics can be used to summarise collections of data

Statistics

Inferential Statistics models data to draw inferences about the population under investigation

A Statistical Hypothesis Test is paramount to the process of statistical decision on top of experimental patterns

A result is “Statistically Significant” whether it does not happen by chance

There are several tests allowing to determine whether a result is Statistically Significant

Quiz time!

https://forms.gle/L75iyiRbUqvwnnWKA