FUNDAMENTALS OF DATA SCIENCE

Module 3

Lisha Varghese
AP/MCA

MODULE 3

Data Analytics Lifecycle : Discovery-Data preparation-Model planning-Model execution-Communicate results-Operationalize

Introduction to Big Data Analysis Techniques- Need for Analysis Types of Analysis -Quantitative Analysis -Qualitative Analysis Data Mining Statistical Analysis -Machine Learning - Semantic Analysis - Visual Analysis

DATA ANALYTICS

- Data science projects are exploratory and require a governed process to ensure thoroughness without stifling exploration.
- Breaking down complex problems into manageable phases aids in clarity and focus.
- Rushing into data collection and analysis without proper planning can lead to misalignment with objectives, necessitating a return to initial phases or project cancellation.
- Documenting the process enhances rigor, credibility, and repeatability, fostering future adoption and continuity within teams.

DATA ANALYTICS LIFECYCLE

Discovery

Data preparation

Model planning

Model execution

Communicate results

Operationalize

DATA ANALYTICS LIFECYCLE

- The Data Analytics Lifecycle is designed specifically for Big Data problems and data science projects.
- The lifecycle has six phases, and project work can occur in several phases at once.
- For most phases in the lifecycle, the movement can be either forward or backward.
- This iterative depiction of the lifecycle is intended to more closely portray a real project, in which aspects of the project move forward and may return to earlier stages as new information is uncovered and team members learn more about various stages of th e project.
- This enables participants to move iteratively through the process and drive toward operationalizing the project work.

- Each plays a critical part in a successful analytics project.
- Although seven roles are listed, fewer or more people can accomplish the work depending on the scope of the project, the organizational structure, and the skills of the participants.
- For example, on a small, versatile team, these seven roles may be fulfilled by only 3 people, but a very large project may require 20 or more people.
- The seven roles follow.

- **Business User:** Someone who understands the domain area and usually benefits from the results. This person can consult and advise the project team on the context of the project, the value of the results, and how the outputs will be operationalized. Usually a business analyst, line manager, or deep subject matter expert in the project domain fulfills this role.
- **Project Sponsor**: Responsible for the genesis of the project. Provides the impetus and requirements for the project and defines the core business problem. Generally provides the funding and gauges the degree of value from the final outputs of the working team. This person sets the priorities for the project and clarifies the desired outputs.
- **Project Manager**: Ensures that key milestones and objectives are met on time and at the expected quality.
- **Business Intelligence Analyst:** Provides business domain expertise based on a deep understanding of the data, key performance indicators (KPIs), key metrics, and business intelligence from a reporting perspective. Business Intelligence Analysts generally create dashboards and reports and have knowledge of the data feeds and sources.

- **Database Administrator (DBA)**: Provisions and configures the database environment to support the analytics needs of the working team. These responsibilities may include providing access to key databases or tables and ensuring the appropriate security levels are in place related to the data repositories.
- **Data Engineer**: Leverages deep technical skills to assist with tuning SQL queries for data management and data extraction, and provides support for data ingestion into the analytic sandbox.
- The DBA sets up and configures the databases to be used, the data engineer executes the actual data extractions and performs substantial data manipulation to facilitate the analytics.
- The data engineer works closely with the data scientist to help shape data in the right ways for analyses.
- **Data Scientist**: Provides subject matter expertise for analytical techniques, data modeling, and applying valid analytical techniques to given business problems.
- Ensures overall analytics objectives are met.
- Designs and executes analytical methods and approaches with the data available to the project.

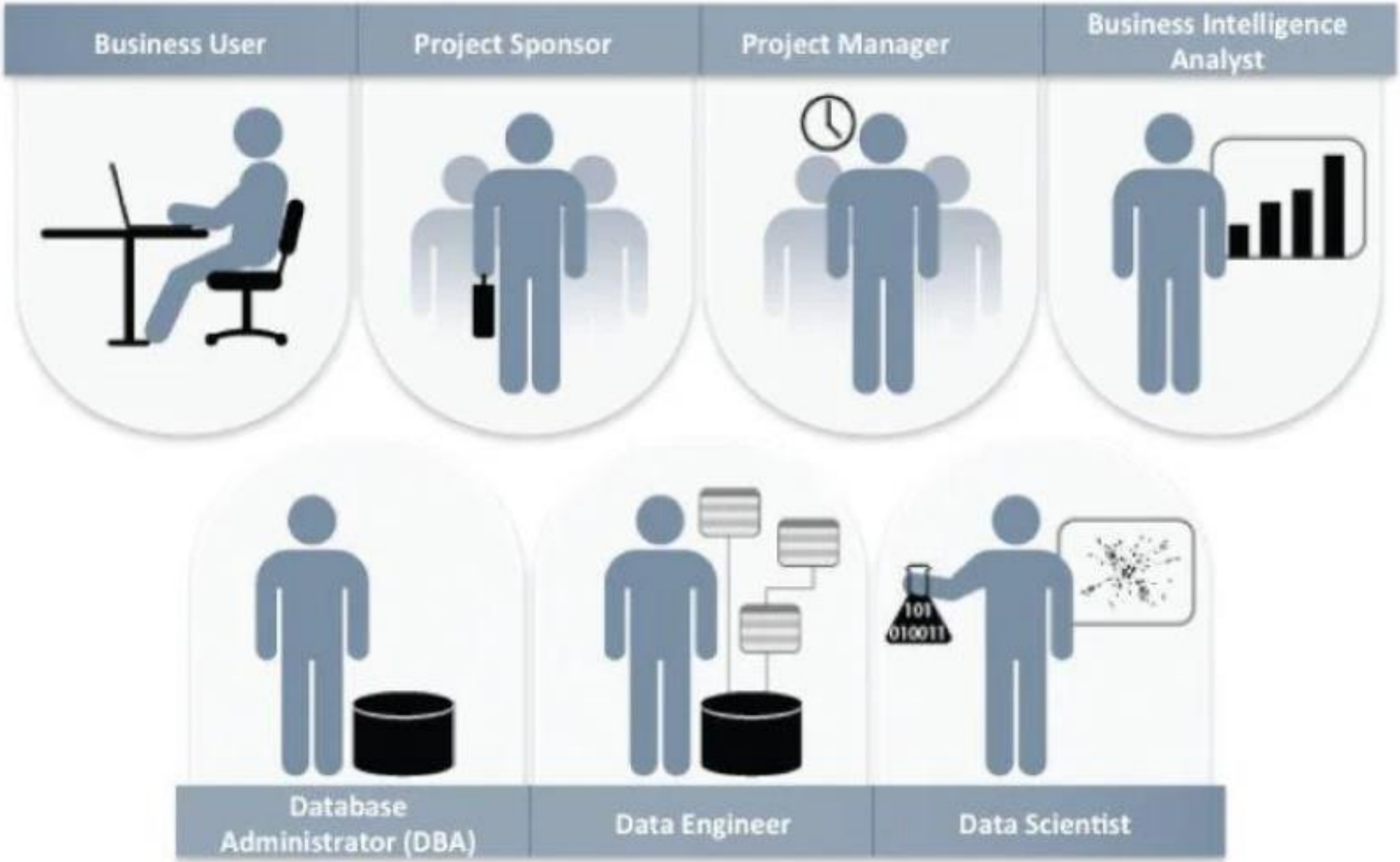


Figure 2.1 Key roles for a successful analytics project

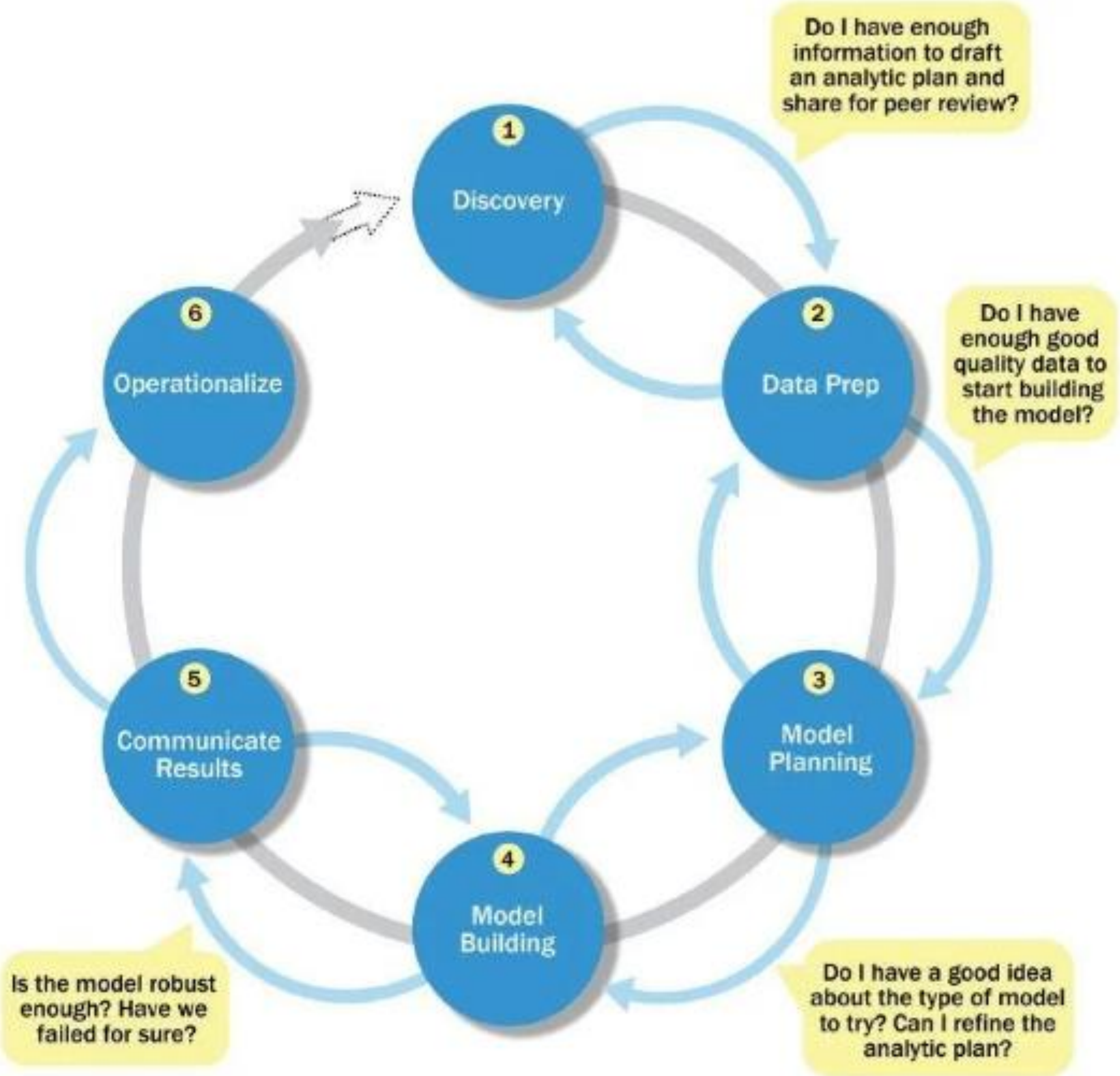

Do I have enough information to draft an analytic plan and share for peer review?
1
Discovery
2
Data Prep
Do I have enough good quality data to start building the model?
3
Model Planning
Do I have a good idea about the type of model to try? Can I refine the analytic plan?
4
Model Building
Is the model robust enough? Have we failed for sure?
5
Communicate Results
6
Operationalize

Figure 2.2 Overview of Data Analytics Lifecycle

PHASES OF THE DATA ANALYTICS LIFECYCLE

Phase 1—Discovery:

- Team learns the business domain, including relevant history such as whether the organization or business unit has attempted similar projects in the past from which they can learn.
- The team assesses the resources available to support the project in terms of people, technology, time, and data.
- Important activities in this phase include framing the business problem as an analytics challenge that can be addressed in subsequent phases and formulating initial hypotheses (IHs) to test and begin learning the data.

In short,

- ❑ the data science team must learn and investigate the problem,
- ❑ develop context and understanding,
  and learn about the data sources needed and available for the project.
- ❑ In addition, the team formulates initial hypotheses that can later be tested with data.

Phase 2—Data preparation:

- Phase 2 requires the presence of an analytic sandbox, in which the team can work with data and perform analytics for the duration of the project.
- The team needs to execute extract, load, and transform (ELT) or extract, transform and load (ETL) to get data into the sandbox.
- The ELT and ETL are sometimes abbreviated as ETLT. Data should be transformed in the ETLT process so the team can work with it and analyze it.
- In this phase, the team also needs to familiarize itself with the data thoroughly and take steps to condition the data

PHASES OF THE DATA ANALYTICS LIFECYCLE (CONTD...)

Phase 3—Model planning:

- Team determines the methods, techniques, and workflow it intends to follow for the subsequent model building phase.
- The team explores the data to learn about the relationships between variables and subsequently selects key variables and the most suitable models.

Phase 4—Model building:

- Team develops datasets for testing, training, and production purposes.
- In addition, in this phase the team builds and executes models based on the work done in the model planning phase.
- The team also considers whether its existing tools will suffice for running the models, or if it will need a more robust environment for executing models and workflows (for example, fast hardware and parallel processing, if applicable).

PHASES OF THE DATA ANALYTICS LIFECYCLE (CONTD…)

Phase 5—Communicate results:

- In Phase 5, the team, in collaboration with major stakeholders, determines if the results of the project are a success or a failure based on the criteria developed in Phase 1. The team should identify key findings, quantify the business value, and develop a narrative to summarize and convey findings to stakeholders.

Phase 6—Operationalize:

- In Phase 6, the team delivers final reports, briefings, code, and technical documents. In addition, the team may run a pilot project to implement the models in a production environment.

PHASE 1 TOOLS

Data Visualization and Exploration:

- Tableau
- Power BI
- QlikView

Web Analytics:

- Google Analytics

Data Exploration, Manipulation, and Basic Analysis:

1. Python Libraries:
    - pandas
    - NumPy

Other Data Exploration Tools:

- Alteryx
- Hadoop
- OpenRefine

PHASE 2- TOOLS :DATA PREPARATION:

1. **Apache Spark:**
   - Efficient data cleaning, transformation, and integration.
2. **Alteryx:**
   - Provides a user-friendly interface for cleaning and transforming data.
3. **Trifacta**:
   - Efficiently wrangles and cleans data.
4. **Apache Pig:**
   - Used for data transformation and analysis.
5. **Apache Hadoop:**
   - Provides distributed storage capabilities.

PHASE 3- TOOLS :MODEL PLANNING

1. **Jupyter Notebooks**: Widely used for interactive data analysis and code development.
2. **RStudio**: An IDE for R, commonly used for statistical modeling and analysis.
3. **KNIME**: Open-source platform for visually creating data flows and planning model workflows.
4. **Lucidchart**: Diagramming tool for creating flowcharts and visualizing the model structure.
5. **IBM SPSS Modeler**: Comprehensive analytics platform for designing and planning predictive models.

PHASE 4- TOOLS :MODEL BUILDING

Commercial Tools:

1. **SAS Enterprise Miner**: Runs predictive and descriptive models, designed for enterprise-level computing.
2. **SPSS Modeler (IBM SPSS Modeler)**: Offers GUI-based data exploration and analysis.
3. **Matlab**: High-level language for data analytics, algorithms, and exploration.
4. **Alpine Miner**: GUI for developing analytic workflows, interacting with Big Data tools.
5. **STATISTICA and Mathematica**: Popular data mining and analytics tools.

PHASE 4- TOOLS :MODEL BUILDING

Free/Open Source Tools:

1. **R and PL/R**: Executes R commands in PostgreSQL, providing scalability.
2. **Octave**: Free software for computational modeling, akin to Matlab.
3. **WEKA**: Free data mining software with an analytic workbench.
4. **Python**: Programming language with machine learning toolkits (scikit-learn, numpy, scipy, pandas).

PHASE 5- TOOLS :COMMUNICATE RESULTS

1. **QlikView** - Provides interactive and associative data visualizations for effective communication.
2. **Microsoft Excel** - Widely used for creating spreadsheets, charts, and simple visualizations.
3. **Google Data Studio** - Allows the creation of interactive and customizable reports and dashboards.
4. **Jupyter Notebooks** - Useful for creating and sharing documents containing live code, equations, visualizations, and narrative text.
5. **Markdown** - Often used for creating documentation, reports, and presentations with a simple and readable format.

PHASE 6- TOOLS :OPERATIONALIZE

1. **Apache Airflow**: Open-source platform for programmatically authoring, scheduling, and monitoring workflows.
2. **Luigi**: Python module for building complex pipelines of batch jobs.
3. **Docker**: Containerization platform facilitating packaging, distributing, and running applications.
4. **Kubernetes**: Container orchestration platform automating deployment, scaling, and management of containerized applications.
5. **AWS Lambda**: Serverless computing service enabling running code without provisioning or managing servers.

SUMMARY

a six-phase approach to managing analytical projects:

1. **Discovery**
2. **Data Preparation**
3. **Model Planning**
4. **Model Building**
5. **Communicate Results**
6. **Operationalize**

INTRODUCTION TO BIG DATA ANALYSIS TECHNIQUES-

Need for Analysis

Types of Analysis

- **Quantitative Analysis**
- **Qualitative Analysis**
- **Data Mining**
- **Statistical Analysis**
- **Machine Learning**
- **Semantic Analysis**
- **Visual Analysis**

BIG DATA ANALYSIS

- blends traditional statistical data analysis approaches with computational ones.
- Statistical sampling from a population is ideal when the entire dataset is available, and this condition is typical of traditional batch processing scenarios.
- Big Data can shift batch processing to realtime processing due to the need to make sense of **streaming data**.
- With streaming data, the dataset accumulates over time, and the data is time-ordered.
- Streaming data places an emphasis on timely processing, for analytic results have a shelf-life.
- Whether it is the recognition of an upsell opportunity that presents itself due to the current context of a customer, or the detection of anomalous conditions in an industrial setting that require intervention to protect equipment or ensure product quality, time is of the essence, and freshness of the analytic result is essential.

NEED FOR ANALYSIS

Why is big data analytics important?

- Big data analytics helps **organizations harness their data** and **use it to identify new opportunities**.
- That, in turn**, leads to smarter business moves, more efficient operations, higher profits and happier customers**
- Businesses that use big data with advanced analytics gain value in many ways, such as:
- **Reducing cost:** Big data technologies like cloud-based analytics can significantly reduce costs when it comes to storing large amounts of data (for example, a data lake).
- Plus, big data analytics helps organizations find more efficient ways of doing business.

- **Making faster, better decisions**: The speed of in-memory analytics-combined with the ability to analyze new sources of data, such as streaming data from lot helps businesses analyze information immediately and make fast, informed decisions.

- **Developing and marketing new products and services**: Being able to gauge(measure) customer needs and customer satisfaction through analyties, empowers businesses to give customers what they want, when they want it.

- With big data analytics, more companies have an opportunity to develop innovative new products to meet customers changing needs.

- Organizations can use big data analytics systems and software to make data-driven decisions that can improve business-related outcomes.

- The benefits may include more effective marketing, new revenue opportunities, customer personalization and improved operational efficiency.

- With an effective strategy, these benefits can provide competitive advantages over rivals(competitors).

HOW?

An organization will operate its Big Data analysis engine at two speeds:

- processing streaming data as it arrives
- performing batch analysis of this data as it **accumulates to look for patterns and trends**.

The symbol used to represent data analysis.

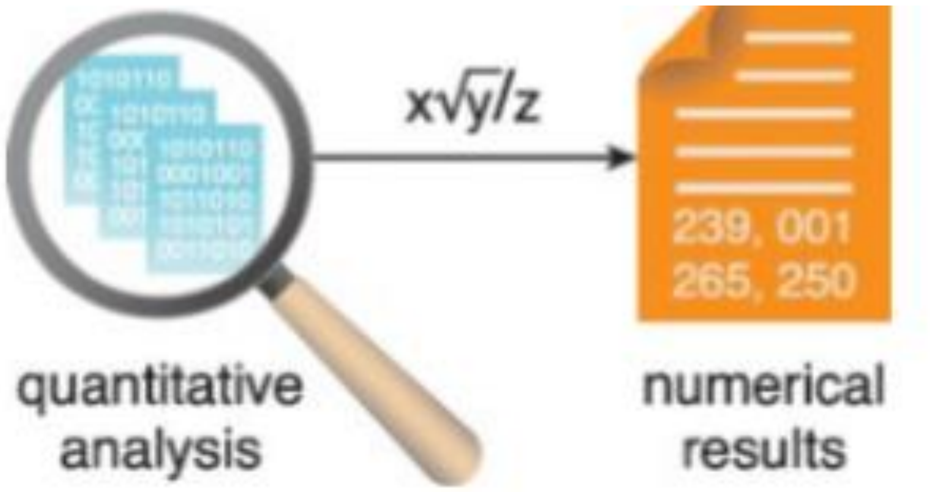
BASIC TYPES OF DATA ANALYSIS

- Quantitative analysis
- Qualitative analysis
- Data mining
- Statistical analysis
- Machine learning
- Semantic analysis
- Visual analysis

QUANTITATIVE ANALYSIS

- Quantitative analysis is a data analysis technique that focuses on **quantifying the patterns and correlations** found in the data.
- Based on statistical practices, this technique involves analyzing a large number of observations from a dataset.
- Since the sample size is large, the results can be applied in a generalized manner to the entire dataset.

Figure depicts the fact that quantitative analysis produces numerical results.



The output of quantitative analysis is numerical in nature.

Quantitative analysis results are absolute in nature and can therefore be used for numerical comparisons.

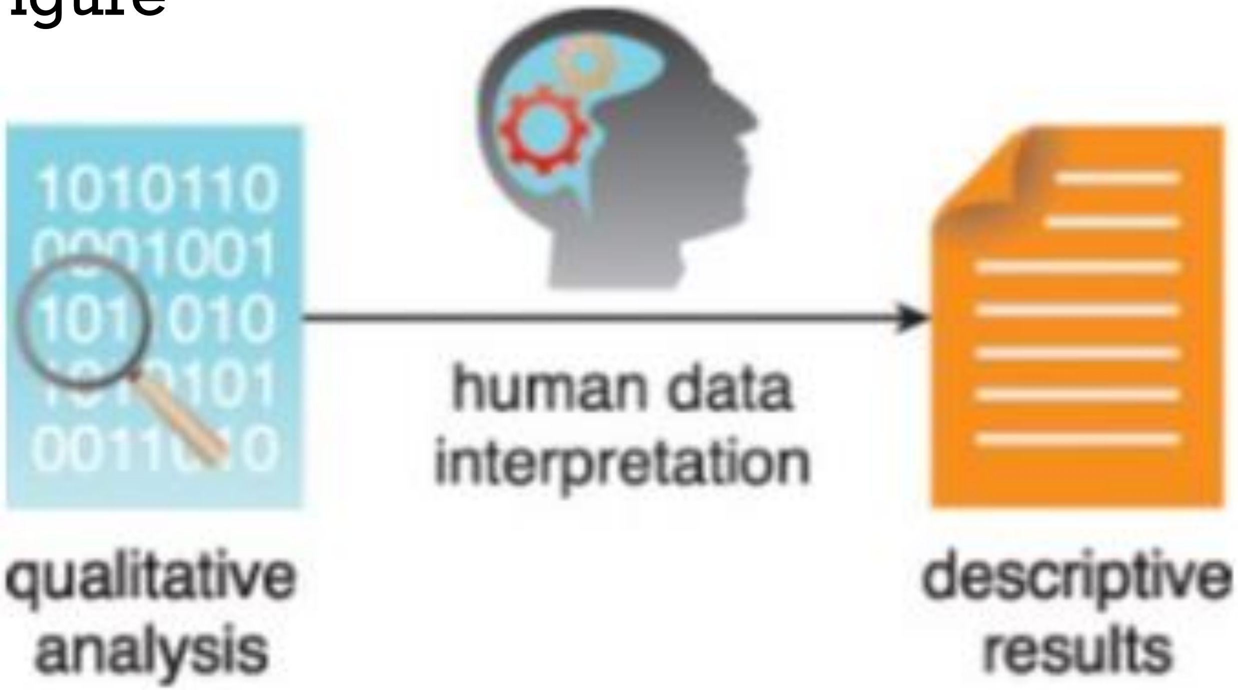
For example

A quantitative analysis of ice cream sales may discover that a 5 degree increase in temperature increases ice cream sales by 15%.

QUALITATIVE ANALYSIS

- Qualitative analysis is a data analysis technique that focuses on describing various data qualities using words.
- It involves analyzing a smaller sample in greater depth compared to quantitative data analysis.
- These analysis results cannot be generalized to an entire dataset due to the small sample size.
- They also cannot be measured numerically or used for numerical comparisons.
- For example, an analysis of ice cream sales may reveal that May’s sales figures were not as high as June’s. The analysis results state only that the figures were “not as high as,” and do not provide a numerical difference.

The output of qualitative analysis is a description of the relationship using words as shown in Figure



Qualitative results are descriptive in nature and not generalizable to the entire dataset.

DATA MINING

- Data mining, also known as data discovery,
- is a specialized form of data analysis that targets large datasets.
- In relation to Big Data analysis, data mining generally refers to automated, software-based techniques that sift through massive datasets to identify patterns and trends.
- Specifically, it involves extracting hidden or unknown patterns in the data with the intention of identifying previously unknown patterns.
- Data mining forms the basis for predictive analytics and business intelligence (BI).
- The symbol used to represent data mining is shown in Figure.

4 The symbol used to represent data mining.

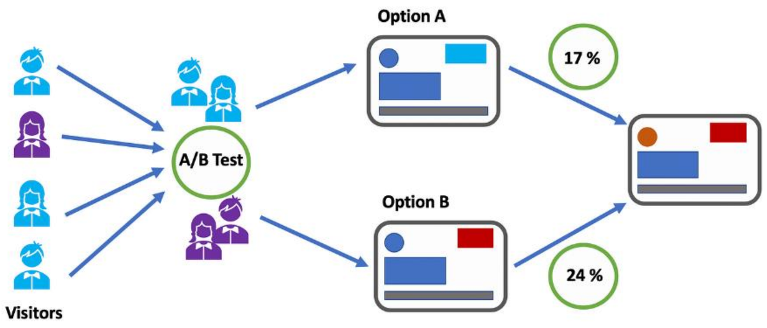
STATISTICAL ANALYSIS

- uses statistical methods based on mathematical formulas as a means for analyzing data.
- is most often quantitative, but can also be qualitative.
- commonly used to describe datasets via summarization, such as providing the mean, median, or mode of statistics associated with the dataset.
- It can also be used to infer patterns and relationships within the dataset, such as regression and correlation.

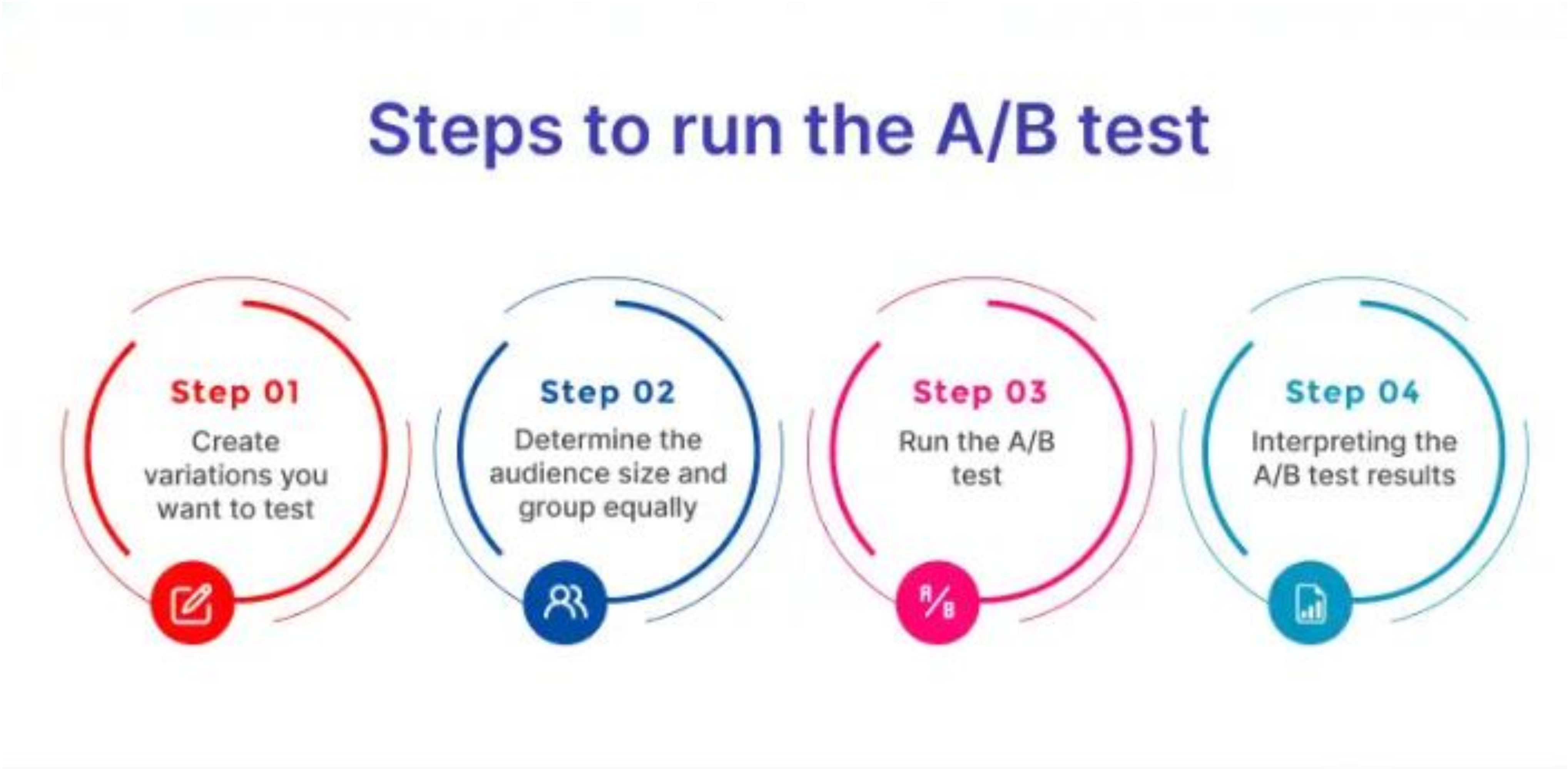
TYPES OF STATISTICAL ANALYSIS

- A/B Testing
- Correlation
- Regression

A/B TESTING



- A/B testing, also known as split or bucket testing
- compares two versions of an element to determine which version is superior based on a pre-defined metric.
- The element can be a range of things.
- For example, it can be content, such as a Web page, or an offer for a product or service, such as deals on electronic items.
- The **current version** of the element is called the **control version**, whereas the modified version is called the treatment.
- Both versions are **subjected to an experiment** simultaneously.
- The **observations are recorded** to determine which version is more successful.
- can be implemented in almost any domain, it is most often used in **marketing**.
- Generally, the objective is to gauge human behavior with the goal of increasing sales.



Steps to run the A/B test

Step 01
Create variations you want to test

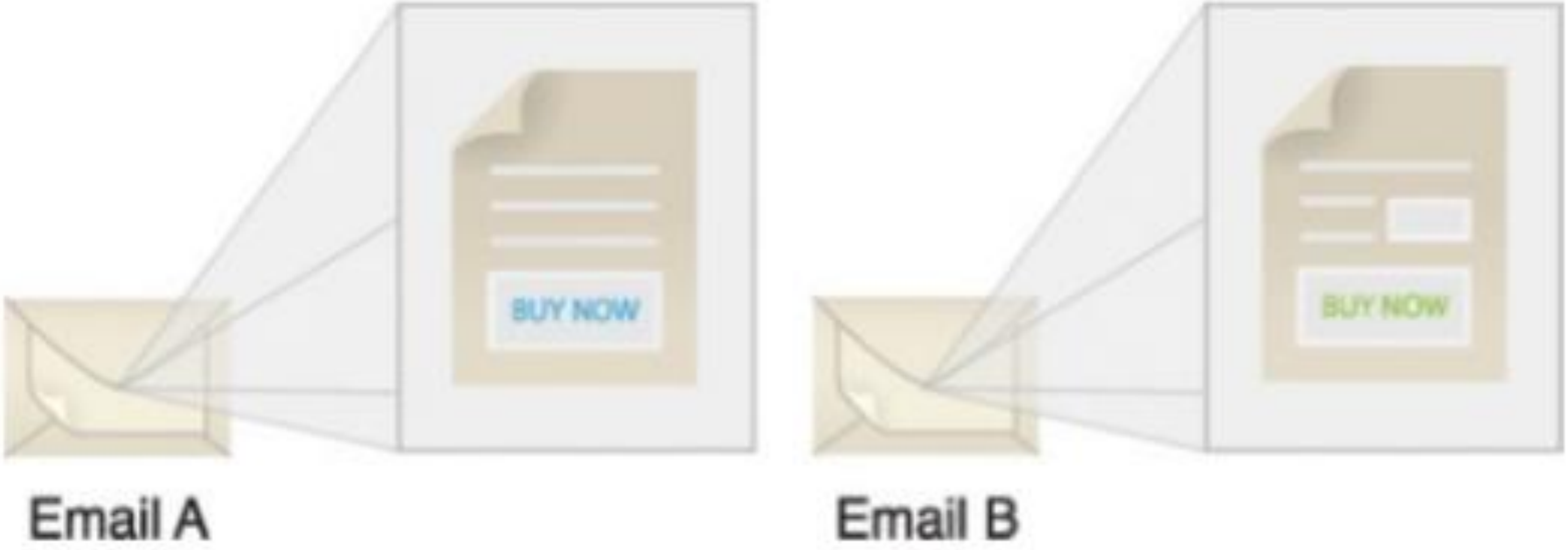
Step 02
Determine the audience size and group equally

Step 03
Run the A/B test

Step 04
Interpreting the A/B test results

EXAMPLE

- For example, in order to determine the best possible layout for an ice cream ad on Company A's Web site, two different versions of the ad are used.
- Version A is an existing ad (the control) while Version B has had its layout slightly altered (the treatment).
- Both versions are then simultaneously shown to different users:
  - Version A to Group A
  - Version B to Group B
- The analysis of the results reveals that Version B of the ad resulted in more sales as compared to Version A.
- In other areas such as the scientific domains, the objective may simply be to observe which version works better in order to improve a process or product.
- Figure provides an example of A/B testing on two different email versions sent simultaneously.



Two different email versions are sent out simultaneously as part of a marketing campaign to see which version brings in more prospective customers.

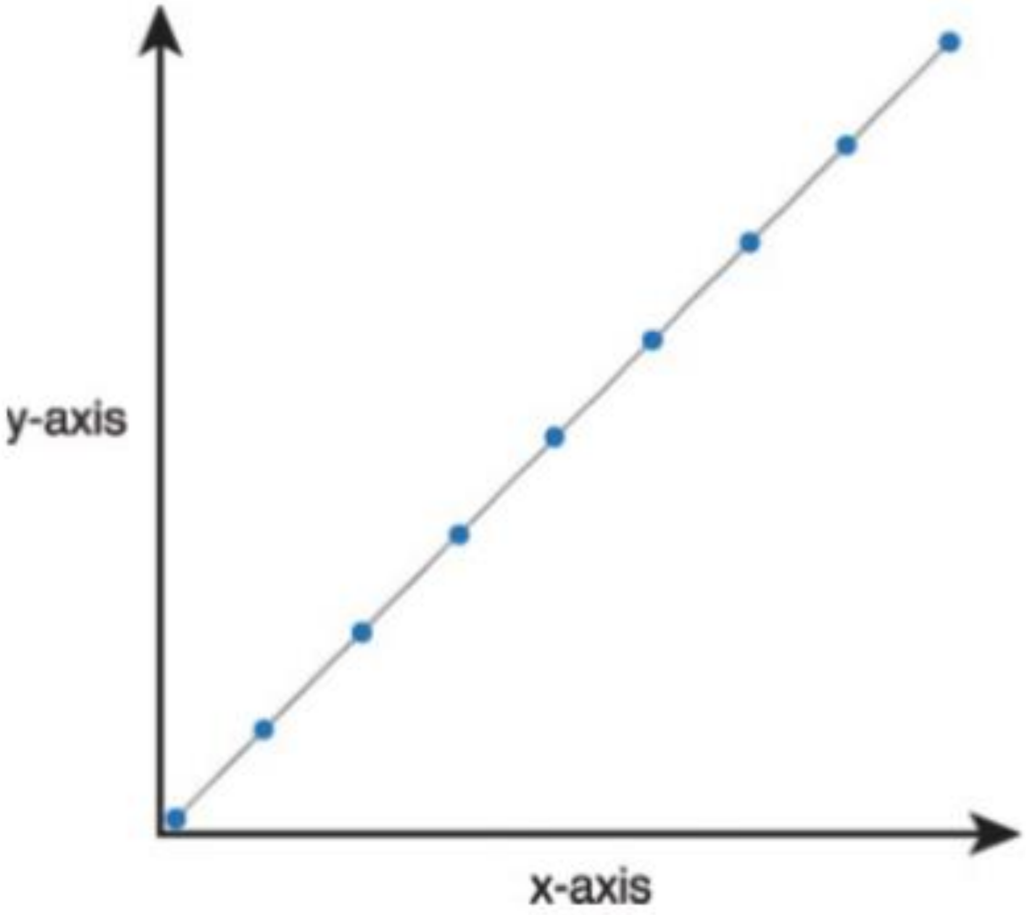
Sample questions can include:

- □ Is the new version of a drug better than the old one?
- □ Do customers respond better to advertisements delivered by email or postal mail?
- □ Is the newly designed homepage of the Web site generating more user traffic?

CORRELATION

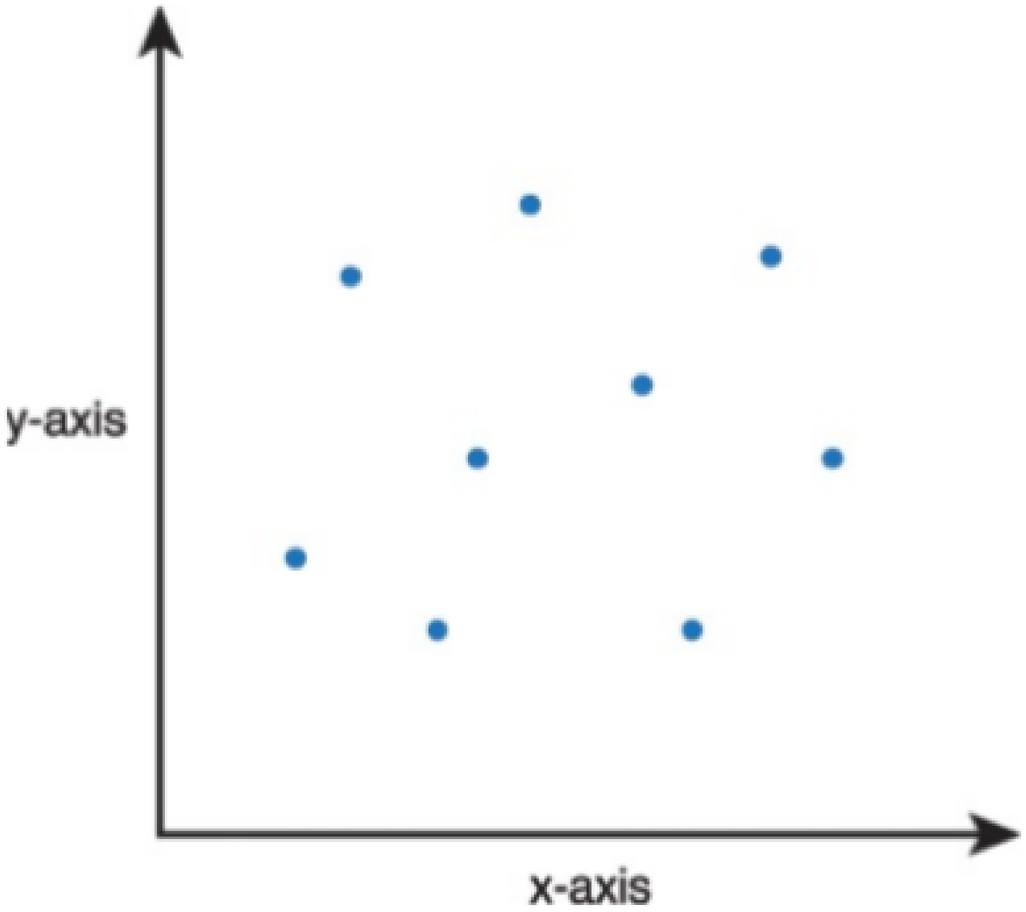
- An analysis technique used to determine whether two variables are related to each other.
- If they are found to be related, the next step is to determine what their relationship is.
- For example, the value of Variable A increases whenever the value of Variable B increases.
- how closely Variables A and B are related, which means we may also want to analyze the extent to which Variable B increases in relation to Variable A's increase.
- correlation helps to develop an understanding of a dataset and find relationships that can assist in explaining a phenomenon.
- commonly used for data mining where the identification of relationships between variables in a dataset leads to the discovery of patterns and anomalies.
- This can reveal the nature of the dataset or the cause of a phenomenon.

- When two variables are considered to be correlated they are aligned based on a linear relationship.
- This means that when one variable changes, the other variable also changes proportionally and constantly.
- Correlation is expressed as a decimal number between –1 to +1, which is known as the correlation coefficient.
- The degree of relationship changes from being strong to weak when moving from –1 to 0 or +1 to 0.



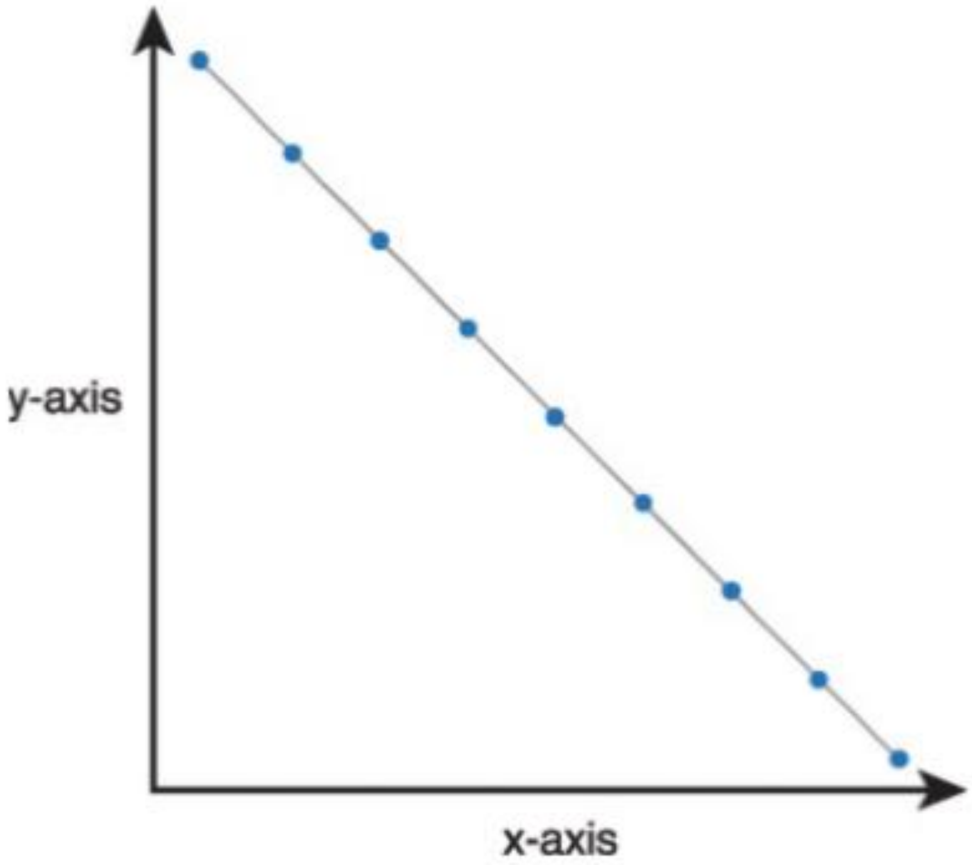
When one variable increases, the other also increases and vice versa.

Figure shows a correlation of +1, which suggests that there is a strong positive relationship between the two variables.



When one variable increases, the other may stay the same, or increase or decrease arbitrarily.

Figure shows a correlation of 0, which suggests that there is no relationship at all between the two variables



When one variable increases, the other decreases and vice versa.

In Figure, a slope of –1 suggests that there is a strong negative relationship between the two variables

EXAMPLE

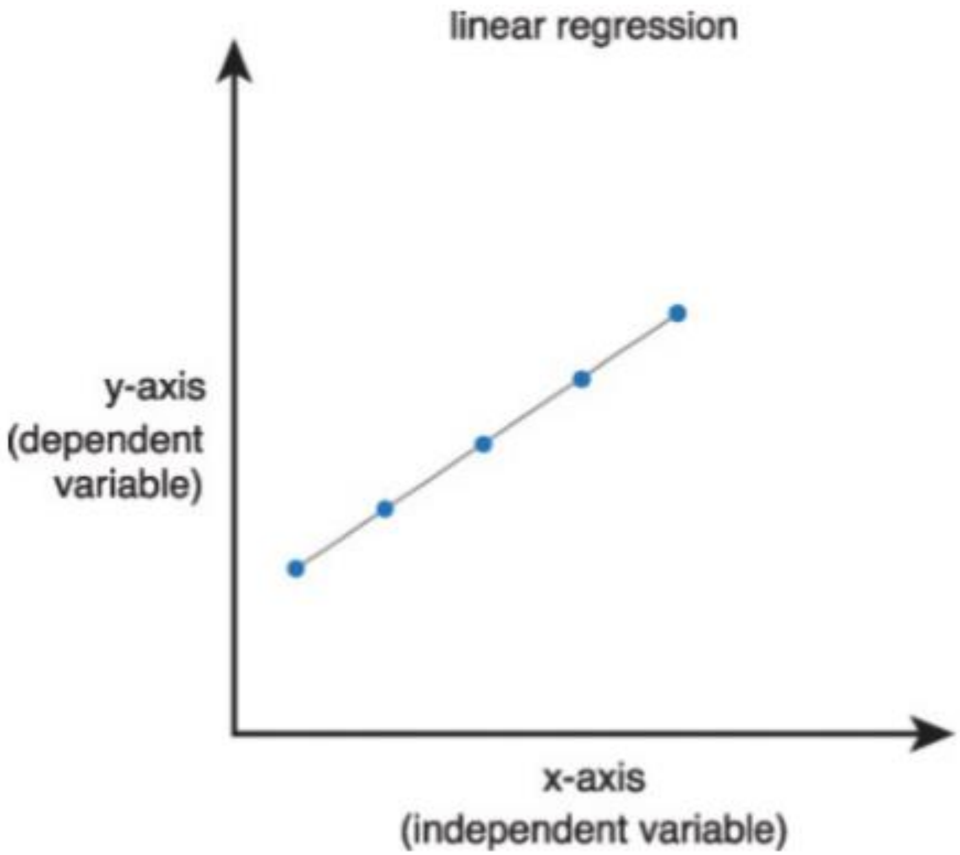
- For example, managers believe that ice cream stores need to stock more ice cream for hot days, but don’t know how much extra to stock.
- To determine if a relationship actually exists between temperature and ice cream sales, the analysts first apply correlation to the number of ice creams sold and the recorded temperature readings.
- A value of +0.75 suggests that there exists a strong relationship between the two. This relationship indicates that as temperature increases, more ice creams are sold.
- Further sample questions addressed by correlation can include:
- Does distance from the sea affect the temperature of a city?
- Do students who perform well at elementary school perform equally well at high school?
- To what extent is obesity linked with overeating?

REGRESSION

- The analysis technique of regression explores how a dependent variable is related to an independent variable within a dataset.
- As a sample scenario, regression could help determine the type of relationship that exists between temperature, the independent variable, and crop yield, the dependent variable.
- Applying this technique helps determine how the value of the dependent variable changes in relation to changes in the value of the independent variable.
- When the independent variable increases, for example, does the dependent variable also increase? If yes, is the increase in a linear or non-linear proportion?
- For example, in order to determine how much extra stock each ice cream store needs to have, the analysts apply regression by feeding in the values of temperature readings. These values are based on the weather forecast as an independent variable and the number of ice creams sold as the dependent variable.
- What the analysts discover is that 15% of additional stock is required for every 5-degree increase in temperature.

- More than one independent variable can be tested at the same time. However, in such cases, only one independent variable may change, while others are kept constant.
- Regression can help enable a better understanding of what a phenomenon is and why it occurred.
- It can also be used to make predictions about the values of the dependent variable.
- Linear regression represents a constant rate of change, as shown in Figure

LINEAR REGRESSION



. Linear regression

Linear regression represents a constant rate of change, as shown in Figure

NON-LINEAR REGRESSION

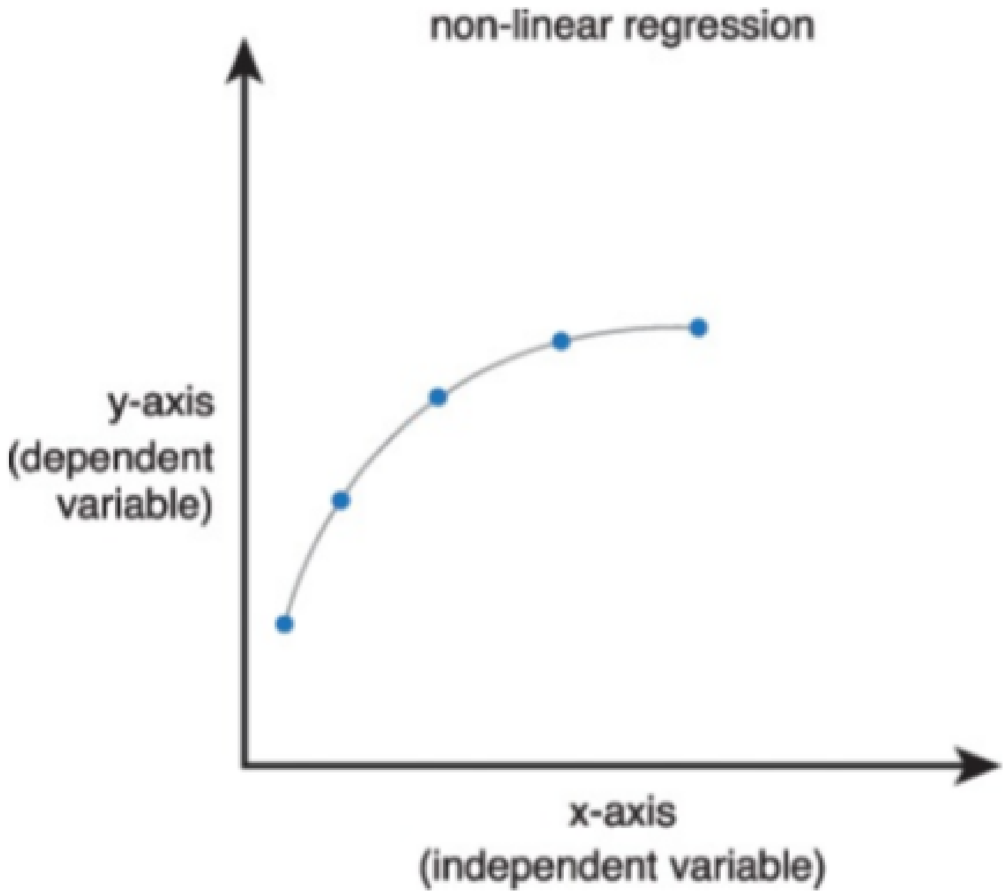
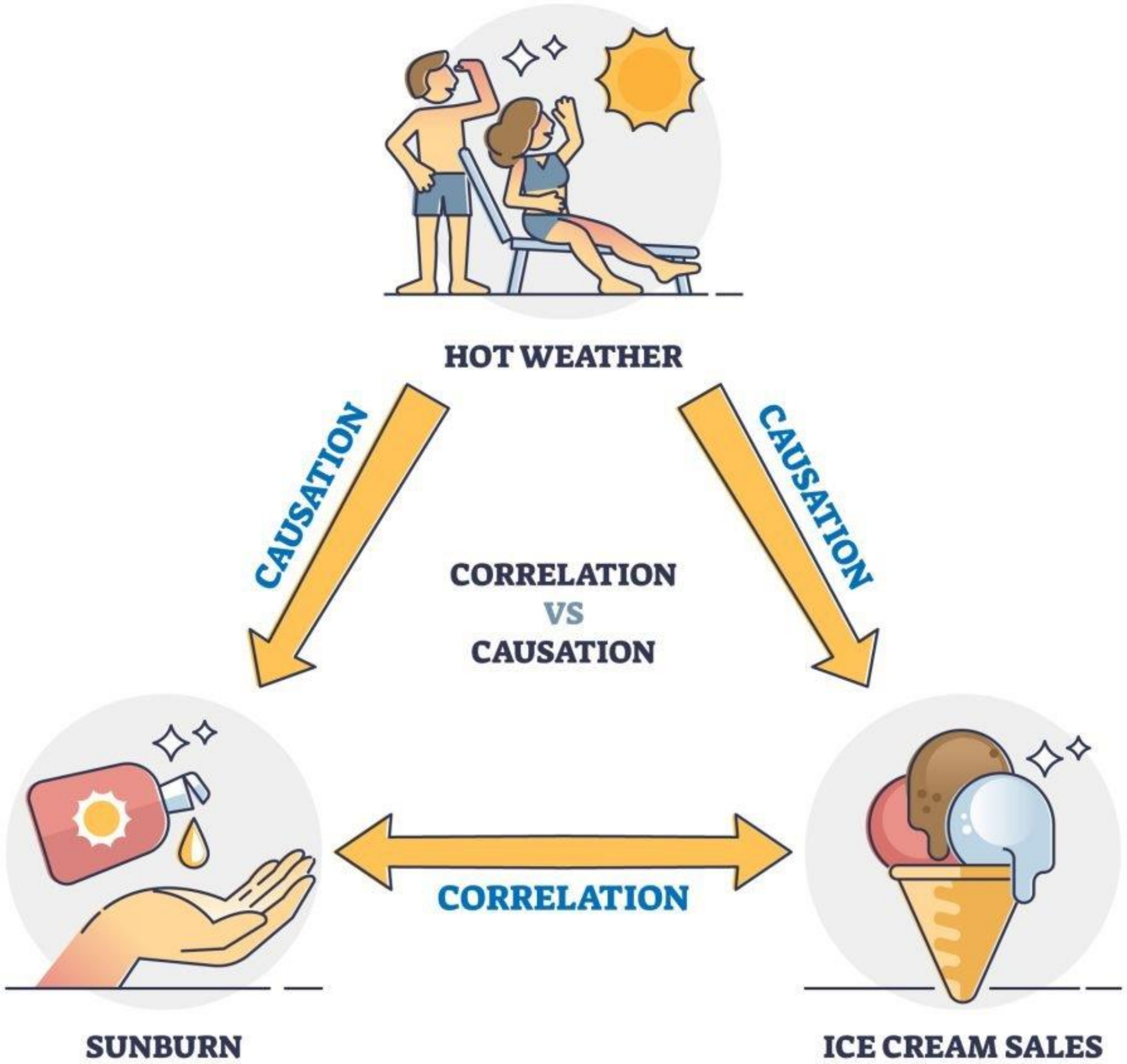


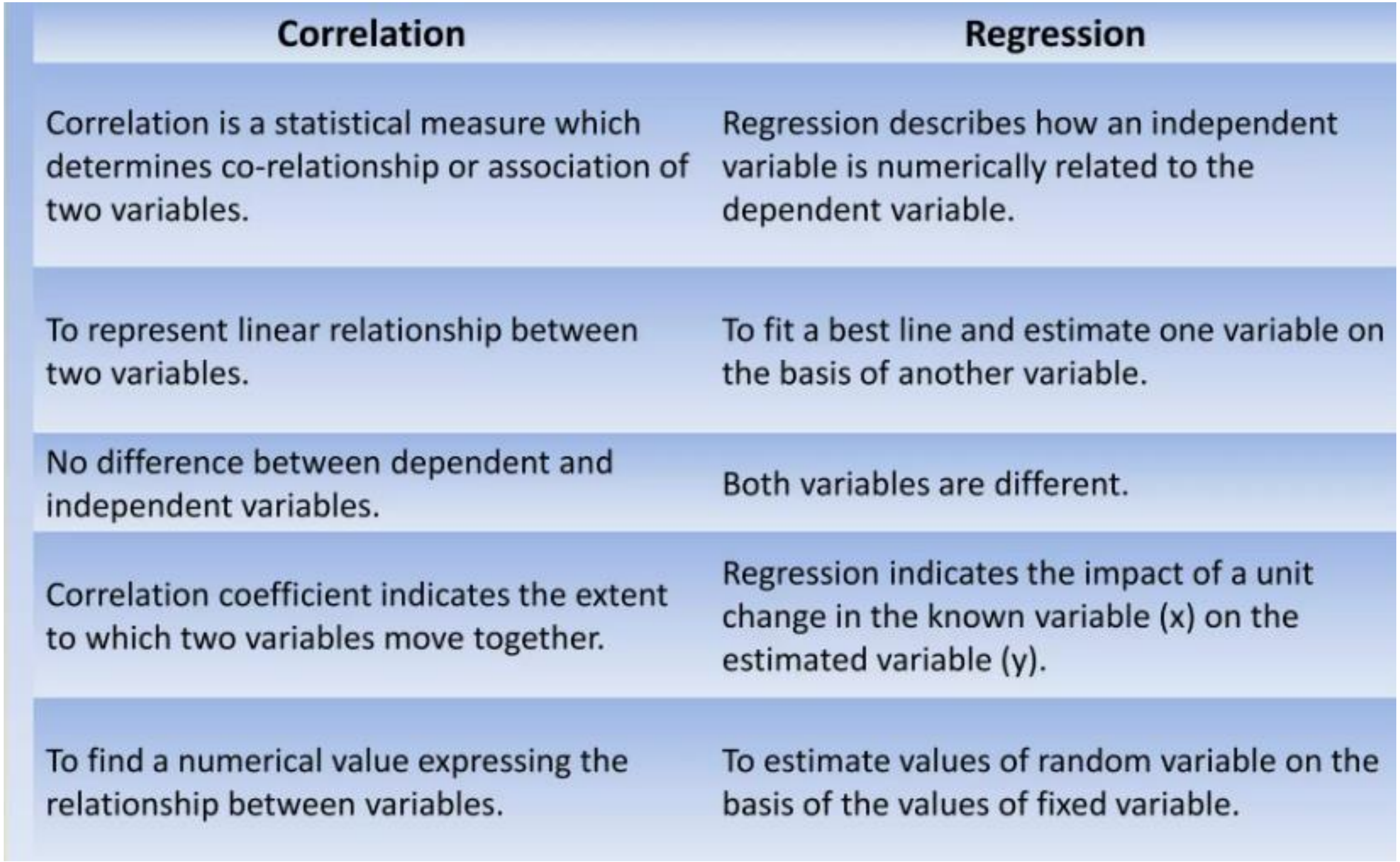
Figure Non-linear regression

Non-linear regression represents a variable rate of change, as shown in Figure

SAMPLE QUESTIONS CAN INCLUDE:

- What will be the temperature of a city that is 250 miles away from the sea?
- What will be the grades of a student studying at a high school based on their primary school grades?
- What are the chances that a person will be obese based on the amount of their food intake?


HOT WEATHER
CAUSATION
CAUSATION
CORRELATION
VS
CAUSATION
CORRELATION
SUNBURN
ICE CREAM SALES



Correlation	Regression
Correlation is a statistical measure which determines co-relationship or association of two variables.	Regression describes how an independent variable is numerically related to the dependent variable.
To represent linear relationship between two variables.	To fit a best line and estimate one variable on the basis of another variable.
No difference between dependent and independent variables.	Both variables are different.
Correlation coefficient indicates the extent to which two variables move together.	Regression indicates the impact of a unit change in the known variable (x) on the estimated variable (y).
To find a numerical value expressing the relationship between variables.	To estimate values of random variable on the basis of the values of fixed variable.

REGRESSION VS CORRELATION

- Correlation
  - does not imply causation.
  - The change in the value of one variable may not be responsible for the change in the value of the second variable, although both may change at the same rate. This can occur due to an unknown third variable, known as the confounding factor.
  - Correlation assumes that both variables are independent.
- Regression
  - is applicable to variables that have previously been identified as dependent and independent variables
  - implies that there is a degree of causation between the variables.
  - The causation may be direct or indirect.
- Within Big Data, correlation can first be applied to discover if a relationship exists. Regression can then be applied to further explore the relationship and predict the values of the dependent variable, based on the known values of the independent variable.

MACHINE LEARNING

- Humans are good at spotting patterns and relationships within data.
- Unfortunately, we cannot process large amounts of data very quickly.
- Machines, on the other hand, are very adept at processing large amounts of data quickly, but only if they know how.
- If human knowledge can be combined with the processing speed of machines, machines will be able to process large amounts of data without requiring much human intervention. This is the basic concept of machine learning.

TYPES OF MACHINE LEARNING TECHNIQUES

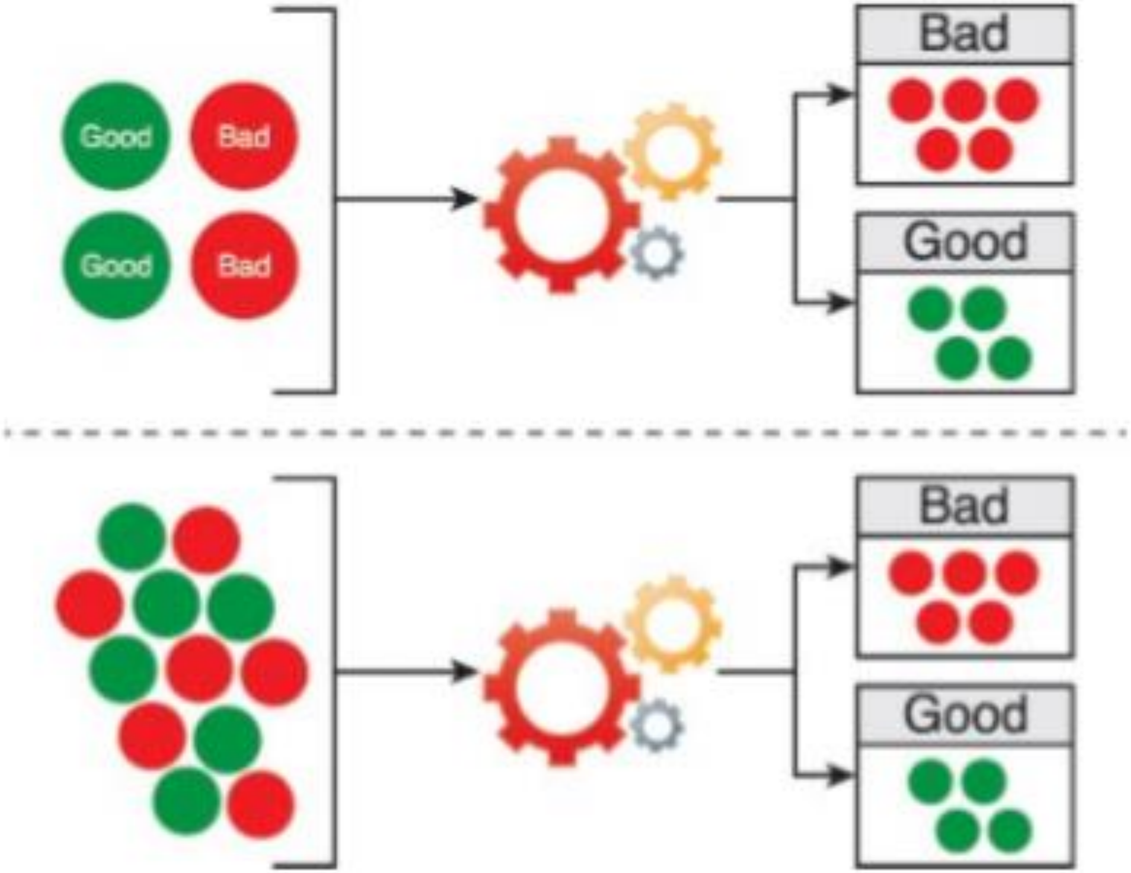
- Classification
- Clustering
- Outlier Detection
- Filtering

CLASSIFICATION (SUPERVISED MACHINE LEARNING)

- Classification is a supervised learning technique by which data is classified into relevant, previously learned categories.
- It consists of two steps:

1. The system is fed training data that is already categorized or labeled, so that it can develop an understanding of the different categories.
2. The system is fed unknown but similar data for classification and based on the understanding it developed from the training data, the algorithm will classify the unlabeled data.

- A common application of this technique is for the filtering of email spam.
- Note that classification can be performed for two or more categories.
- In a simplified classification process, the machine is fed labeled data during training that builds its understanding of the classification, as shown in Figure.
- The machine is then fed unlabeled data, which it classifies itself.



Machine learning can be used to automatically classify datasets.

EXAMPLE

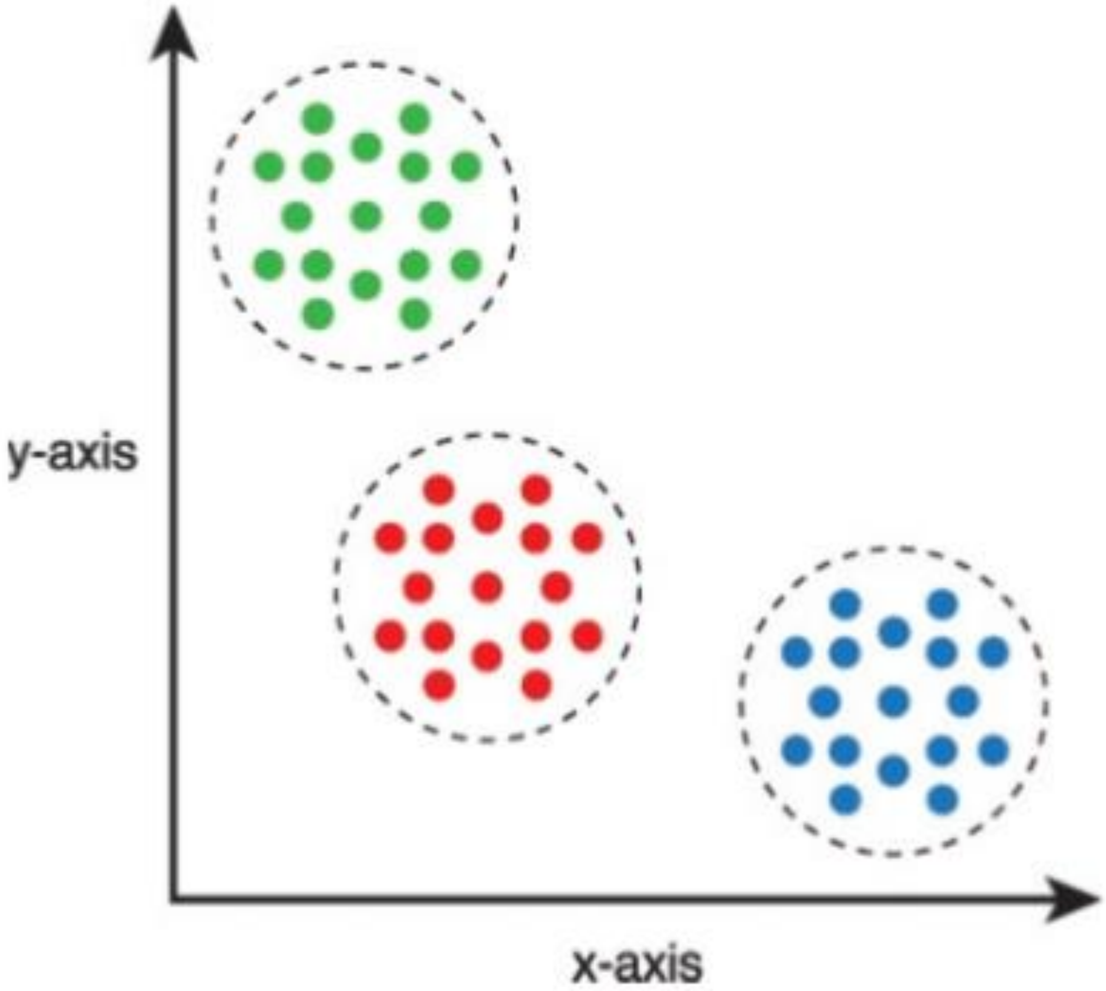
- For example, a bank wants to find out which of its customers is likely to default on loan payments.
- Based on historic data, a training dataset is compiled that contains labeled examples of customers that have or have not previously defaulted.
- This training data is fed to a classification algorithm that is used to develop an understanding of “good” and “bad” customers.
- Finally, new untagged customer data is fed in order to find out whether a given customer belongs to the defaulting category.

SAMPLE QUESTIONS CAN INCLUDE:

- Should an applicant’s credit card application be accepted or rejected based on other accepted or rejected applications?
- Is a tomato a fruit or a vegetable based on the known examples of fruit and vegetables?
- Do the medical test results for the patient indicate a risk for a heart attack?

CLUSTERING (UNSUPERVISED MACHINE LEARNING)

- Clustering is an unsupervised learning technique by which data is divided into different groups so that the data in each group has similar properties.
- There is no prior learning of categories required. Instead, categories are implicitly generated based on the data groupings.
- How the data is grouped depends on the type of algorithm used. Each algorithm uses a different technique to identify clusters.
- Clustering is generally used in data mining to get an understanding of the properties of a given dataset.
- After developing this understanding, classification can be used to make better predictions about similar but new or unseen data.
- Clustering can be applied to the categorization of unknown documents and to personalized marketing campaigns by grouping together customers with similar behavior.
- A scatter graph provides a visual representation of clusters in Figure



A scatter graph summarizes the results of clustering.

FOR EXAMPLE

A bank wants to introduce its existing customers to a range of new financial products based on the customer profiles it has on record. The analysts categorize customers into multiple groups using clustering. Each group is then introduced to one or more financial products most suitable to the characteristics of the overall profile of the group.

SAMPLE QUESTIONS CAN INCLUDE:

- How many different species of trees exist based on the similarity between trees?
- How many groups of customers exist based upon similar purchase history?
- What are the different groups of viruses based on their characteristics?

OUTLIER DETECTION

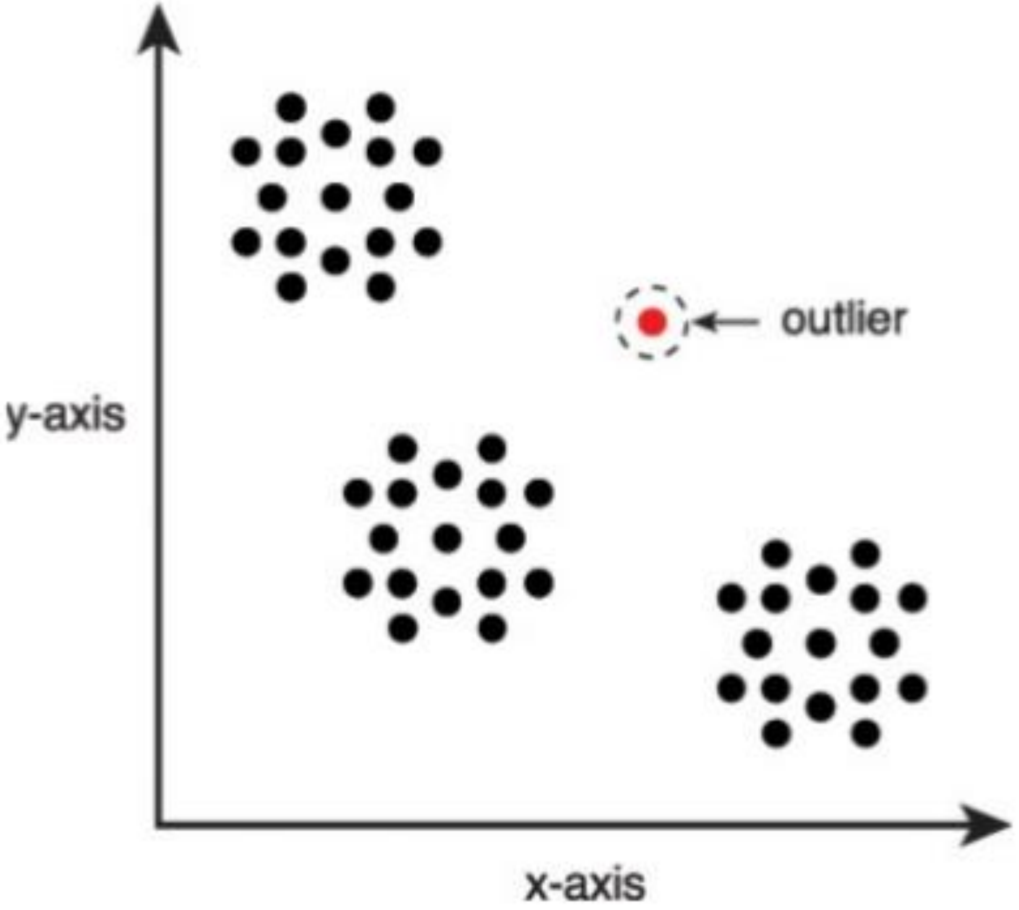
- is the process of finding data that is significantly different from or inconsistent with the rest of the data within a given dataset.
- is used to identify anomalies, abnormalities and deviations that can be advantageous, such as opportunities, or unfavorable, such as risks.
- is closely related to the concept of classification and clustering, although its algorithms focus on finding abnormal values.
- It can be based on either supervised or unsupervised learning.

APPLICATIONS

Applications for outlier detection include

- fraud detection
- medical diagnosis
- network data analysis
- sensor data analysis

A scatter graph visually highlights data points that are outliers, as shown in Figure



A scatter graph highlights an outlier.

FOR EXAMPLE

In order to find out whether or not a transaction is likely to be fraudulent, the bank's IT team builds a system employing an outlier detection technique that is based on supervised learning.

A set of known fraudulent transactions is first fed into the outlier detection algorithm.

After training the system, unknown transactions are then fed into the outlier detection algorithm to predict if they are fraudulent or not.

SAMPLE QUESTIONS CAN INCLUDE:

Is an athlete using performance enhancing drugs?

Are there any wrongly identified fruits and vegetables in the training dataset used for a classification task?

Is there a particular strain of virus that does not respond to medication?

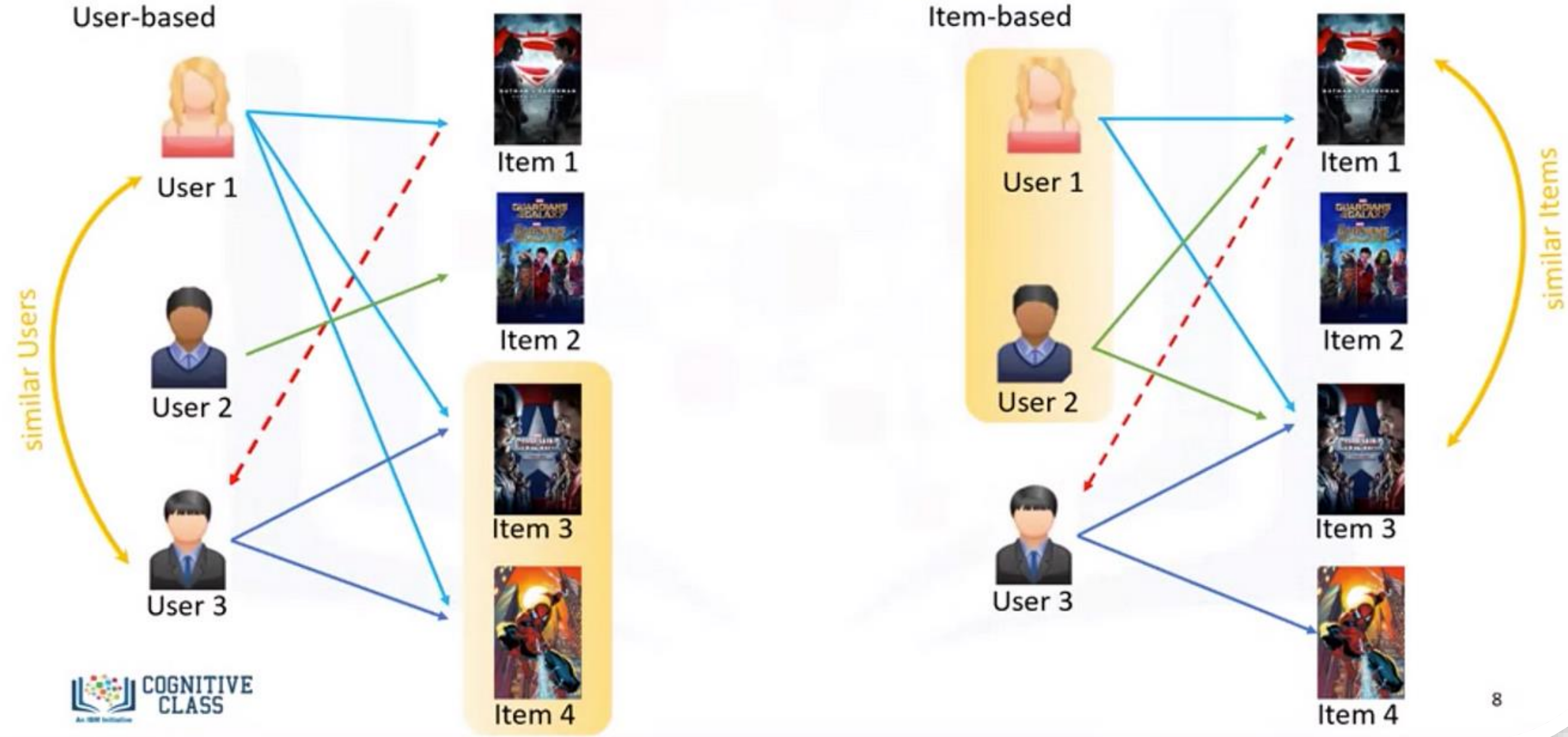
FILTERING

- Filtering is the automated process of finding relevant items from a pool of items.
- Items can be filtered either based on a user’s own behavior or by matching the behavior of multiple users.
- Filtering is generally applied via the following two approaches:
  - ❑collaborative filtering
  - ❑content-based filtering
- A common medium by which filtering is implemented is via the use of a recommender system.

COLLABORATIVE FILTERING

- Collaborative filtering is an item filtering technique based on the collaboration, or merging, of a user's past behavior with the behaviors of others.
- A target user's past behavior, including their likes, ratings, purchase history and more, is collaborated with the behavior of similar users.
- Based on the similarity of the users' behavior, items are filtered for the target user.
- Collaborative filtering is solely based on the similarity between users' behavior.
- It requires a large amount of user behavior data in order to accurately filter items.
- It is an example of the application of the law of large numbers.

Collaborative filtering

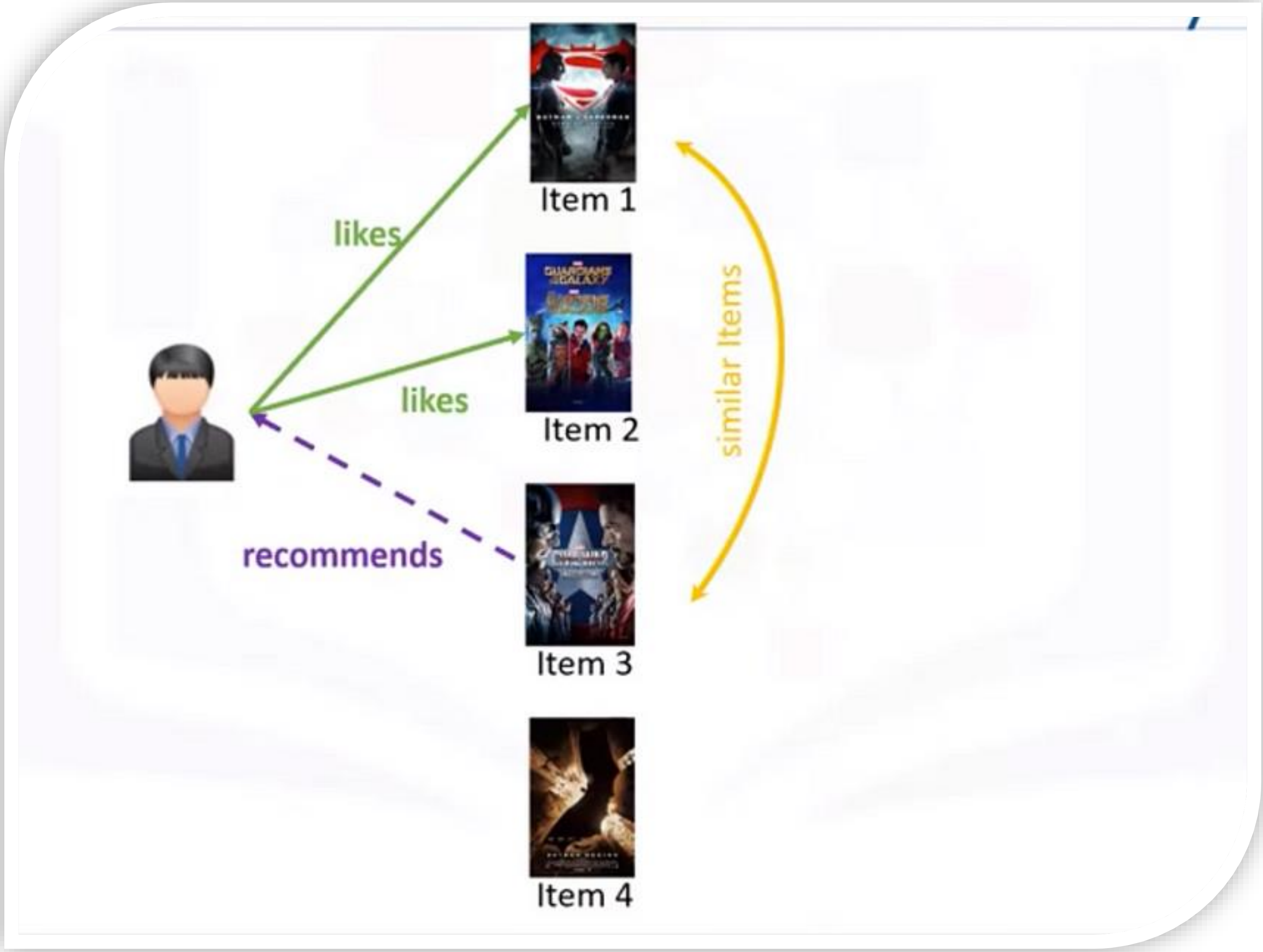


COGNITIVE CLASS

CONTENT-BASED FILTERING

- **Content-based filtering** is an item filtering technique focused on the similarity between users and items.
- A user profile is created based on that user's past behavior, for example, their likes, ratings and purchase history.
- The similarities identified between the user profile and the attributes of various items lead to items being filtered for the user.
- Contrary to collaborative filtering, content-based filtering is solely dedicated to individual user preferences and does not require data about other users.
- A recommender system predicts user preferences and generates suggestions for the user accordingly.
- Suggestions commonly pertain to recommending items, such as movies, books, Web pages and people

- A recommender system typically uses either collaborative filtering or content-based filtering to generate suggestions.
- It may also be based on a hybrid of both collaborative filtering and content-based filtering to fine-tune the accuracy and effectiveness of generated suggestions.
- For example, in order to realize cross-selling opportunities, the bank builds a recommender system that uses content-based filtering.
- Based on matches found between financial products purchased by customers and the properties of similar financial products, the recommender system automates suggestions for potential financial products that customers may also be interested in.


Item 1
likes
likes
Item 2
similar Items
recommends
Item 3
Item 4

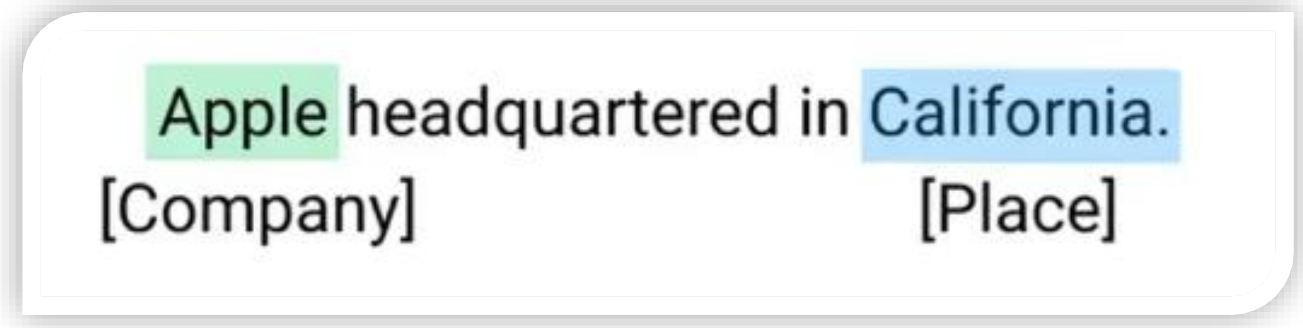
SAMPLE QUESTIONS CAN INCLUDE:

- How can only the news articles that a user is interested in be displayed?

- Which holiday destinations can be recommended based on the travel history of a vacationer?

- Which other new users can be suggested as friends based on the current profile of a person?

SEMANTIC ANALYSIS



Apple headquartered in California.

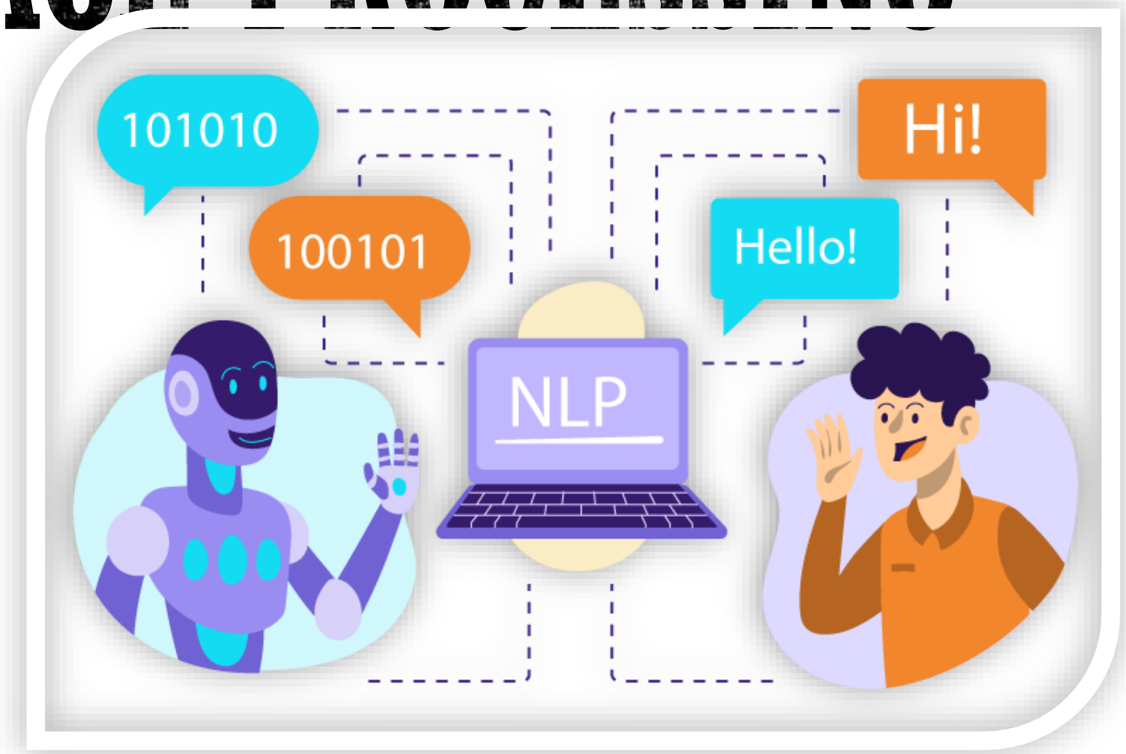
[Company] [Place]

- A fragment of text or speech data can carry different meanings in different contexts,
  - whereas a complete sentence may retain its meaning,
  - even if structured in different ways.
- In order for the machines to extract valuable information, text and speech data needs to be understood by the machines in the same way as humans do.
- Semantic analysis represents practices for extracting meaningful information from textual and speech data.

TYPES OF SEMANTIC ANALYSIS

- Natural Language Processing
- Text Analytics
- Sentiment Analysis

NATURAL LANGUAGE PROCESSING



- A computer's ability to comprehend human speech and text as naturally understood by humans.

- This allows computers to perform a variety of useful tasks, such as full-text searches.

EXAMPLE

- In order to increase the quality of customer care, the ice cream company employs natural language processing to transcribe customer calls into textual data that are then mined for commonly recurring reasons of customer dissatisfaction.
- Instead of hard-coding the required learning rules, either supervised or unsupervised machine learning is applied to develop the computer's understanding of the natural language.
- In general, the more learning data the computer has, the more correctly it can decipher human text and speech.
- Natural language processing includes both text and speech recognition.
- For speech recognition, the system attempts to comprehend the speech and then performs an action, such as transcribing text.

SAMPLE QUESTIONS CAN INCLUDE

- How can an automated phone exchange system that can recognize the correct department extension as dictated verbally by the caller be developed?
- How can grammatical mistakes be automatically identified?
- How can a system that can correctly understand different accents of English language be designed?

TEXT ANALYTICS

- **Unstructured text is difficult to analyze and search** in comparison to structured text.

- Text analytics is the specialized analysis of text through the application of data mining, machine learning and natural language processing techniques to extract value out of unstructured text.

- Text analytics essentially provides the ability to discover text rather than just search it.

- Useful insights from text-based data can be gained by helping businesses develop an understanding of the information that is contained within a large body of text.

TEXT ANALYTICS

- As a continuation of the preceding example,
  - the transcribed textual data after NLP is further analyzed using text analytics to extract meaningful information about the common reasons behind customer discontent.
- The basic tenet of text analytics is to turn unstructured text into data that can be searched and analyzed.
- As the amount of digitized documents, emails, social media posts and log files increases, businesses have an increasing need to leverage any value that can be extracted from these forms of semi-structured and unstructured data.
- Solely analyzing operational (structured) data may cause businesses to miss out on cost-saving or business expansion opportunities, especially those that are customer-focused.
- Applications include document classification and search, as well as building a 360-degree view of a customer by extracting information from a CRM system.

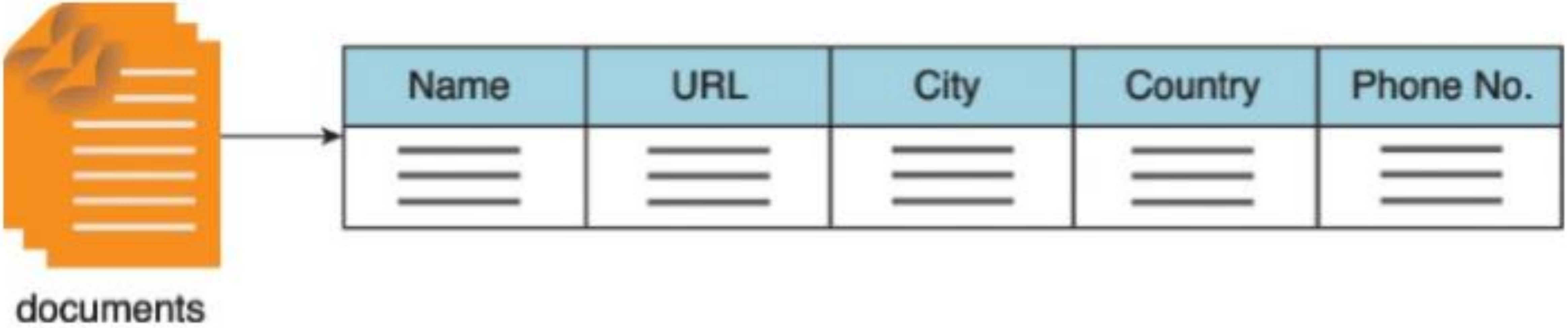
TEXT ANALYTICS GENERALLY INVOLVES TWO STEPS:

1. **Parsing text within documents to extract**:

   - ❑ Named Entities – person, group, place, company
   - ❑ Pattern-Based Entities – social security number, zip code
   - ❑ Concepts – an abstract representation of an entity
   - ❑ Facts – relationship between entities

2. **Categorization of documents using these extracted entities and facts.**

   - ❑ The extracted information can be used to perform a context-specific search on entities, based on the type of relationship that exists between the entities.
   - ❑ Figure shows a simplified representation of text analysis.



documents

Name	URL	City	Country	Phone No.

Entities are extracted from text files using semantic rules and structured so that they can be searched.

SAMPLE QUESTIONS CAN INCLUDE:

- How can I categorize Web sites based on the content of their Web pages?
- How can I find the books that contain content that is relevant to the topic that I am studying?
- How can I identify contracts that contain confidential company information?

SENTIMENT ANALYSIS

- A specialized form of text analysis that focuses on determining the bias or emotions of individuals.
- determines the attitude of the author of the text by analyzing the text within the context of the natural language.
- not only provides information about how individuals feel, but also the intensity of their feeling.
- This information can then be integrated into the decision-making process.
- Common applications for sentiment analysis include
  - identifying customer satisfaction or dissatisfaction early
  - gauging product success or failure
  - spotting new trends.

FOR EXAMPLE

- An ice cream company would like to learn about which of its ice cream flavors are most liked by children.
- Sales data alone does not provide this information because the children that consume the ice cream are not necessarily the purchasers of the ice cream.
- Sentiment analysis is applied to archived customer feedback left on the ice cream company's Web site to extract information specifically regarding children's preferences for certain ice cream flavors over other flavors.

SAMPLE QUESTIONS CAN INCLUDE:

- How can customer reactions to the new packaging of the product be gauged?
- Which contestant is a likely winner of a singing contest?
- Can customer churn be measured by social media comments?

VISUAL ANALYSIS

- Visual analysis is a form of data analysis that involves the graphic representation of data to enable or enhance its visual perception.
- Based on the premise that humans can understand and draw conclusions from graphics more quickly than from text, visual analysis acts as a discovery tool in the field of Big Data.
- The objective is to use graphic representations to develop a deeper understanding of the data being analyzed.
- Specifically, it helps identify and highlight hidden patterns, correlations and anomalies.
- Visual analysis is also directly related to exploratory data analysis as it encourages the formulation of questions from different angles.

TYPES OF VISUAL ANALYSIS

- ❑Heat Maps
- ❑Time Series Plots
- ❑Network Graphs
- ❑Spatial Data Mapping

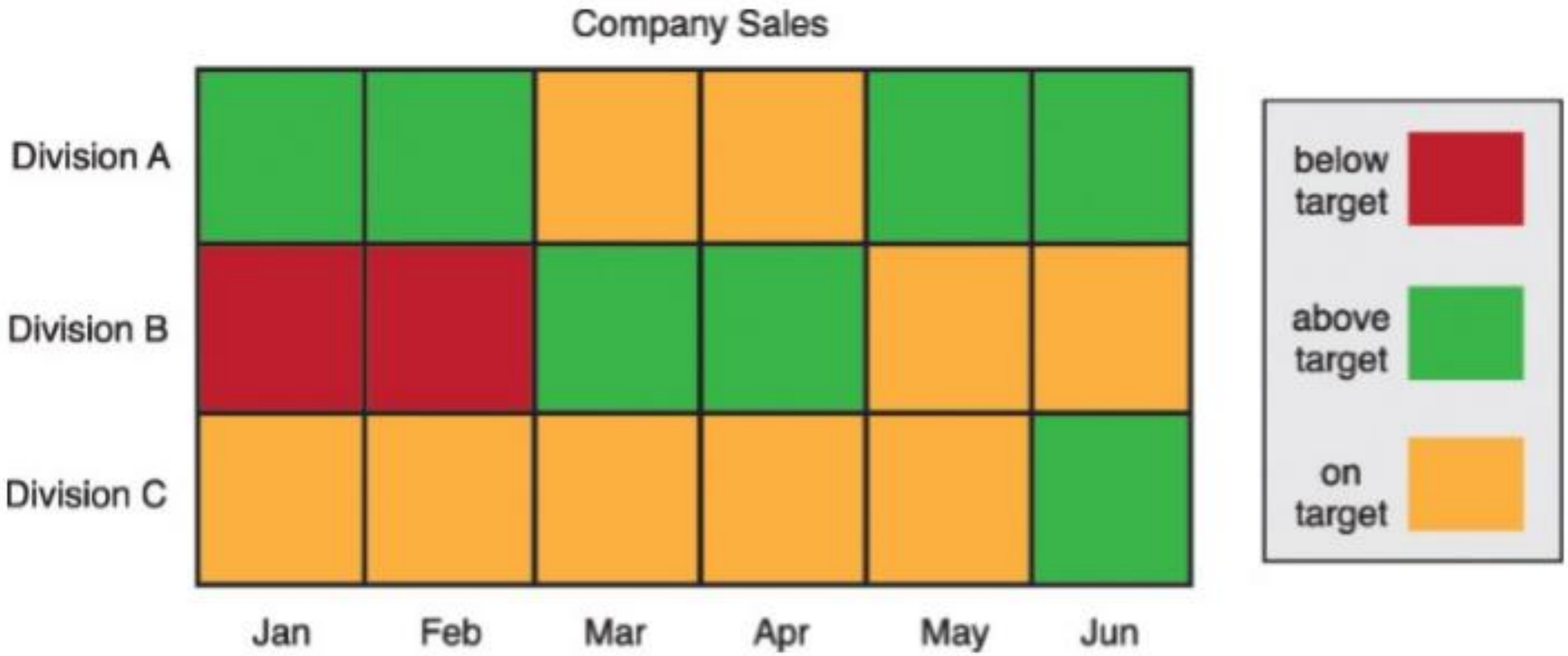
HEAT MAPS

- An effective visual analysis technique for expressing patterns, data compositions via part-whole relations and geographic distributions of data.
- facilitate the identification of areas of interest and the discovery of extreme (high/low) values within a dataset.
- For example, in order to identify the top- and worst-selling regions for ice cream sales, the ice cream sales data is plotted using a heat map.
- Green is used to highlight the best performing regions, while red is used to highlight worst performing regions.

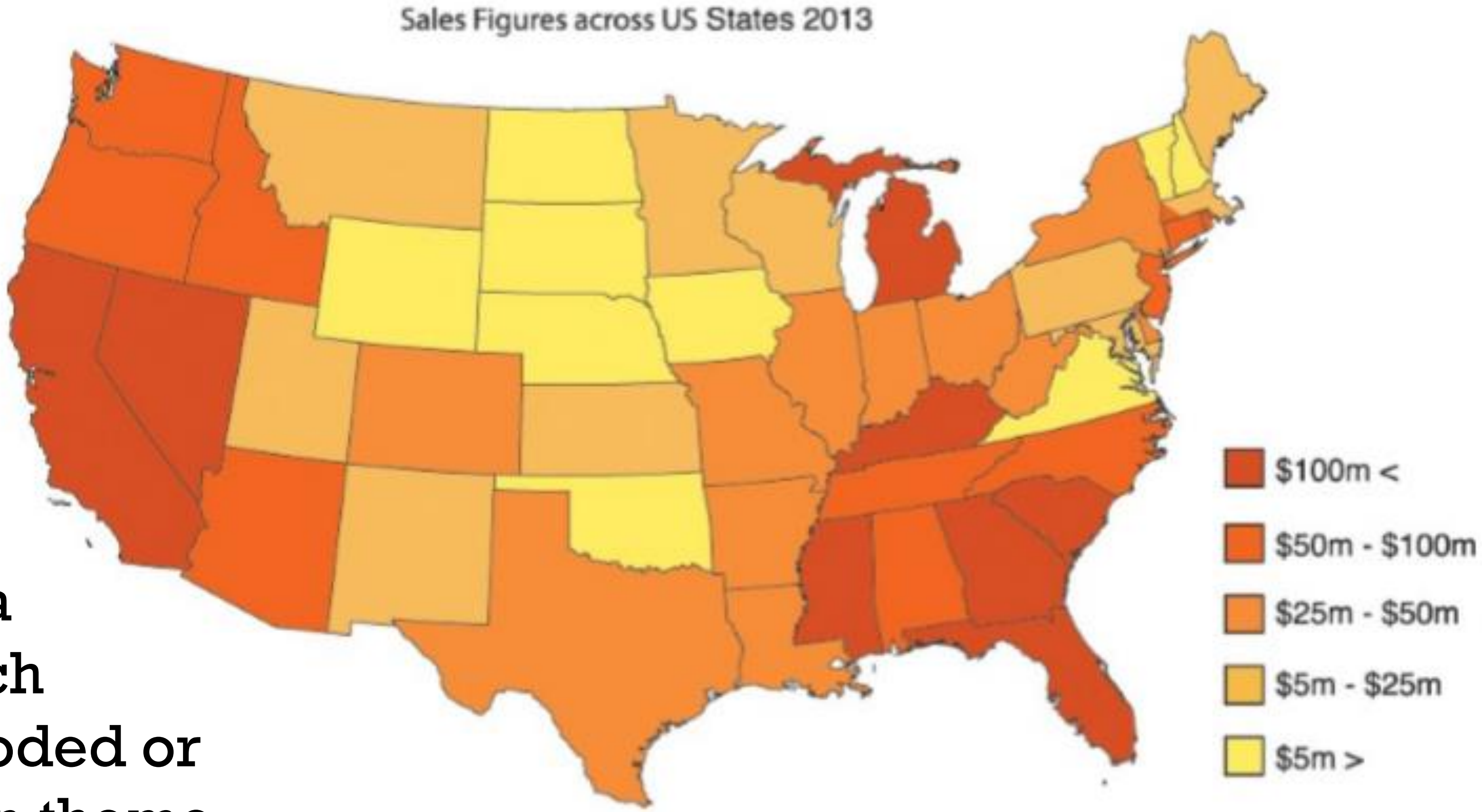
- The heat map itself is a visual, color-coded representation of data values.
- Each value is given a color according to its type or the range that it falls under.

- For example, a heat map may assign the values of 0–3 to the color red, 4–6 to amber and 7–10 to green.

- A heat map can be in the form of a chart or a map.
- A chart represents a matrix of values in which each cell is color-coded according to the value, as shown in Figure
- It can also represent hierarchical values by using color-coded nested rectangles.



This chart heat map depicts the sales of three divisions within a company over a period of six months.



A heat map of the US sales figures from 2013.

In Figure, a map represents a geographic measure by which different regions are color-coded or shaded according to a certain theme. Instead of coloring or shading the whole region, the map may be superimposed by a layer made up of collections of colored/shaded points relating to various regions, or colored/shaded shapes representing

SAMPLE QUESTIONS CAN INCLUDE:

- How can I visually identify any patterns related to carbon emissions across a large number of cities around the world?
- How can I see if there are any patterns of different types of cancers in relation to different ethnicities?
- How can I analyze soccer players according to their strengths and weaknesses?

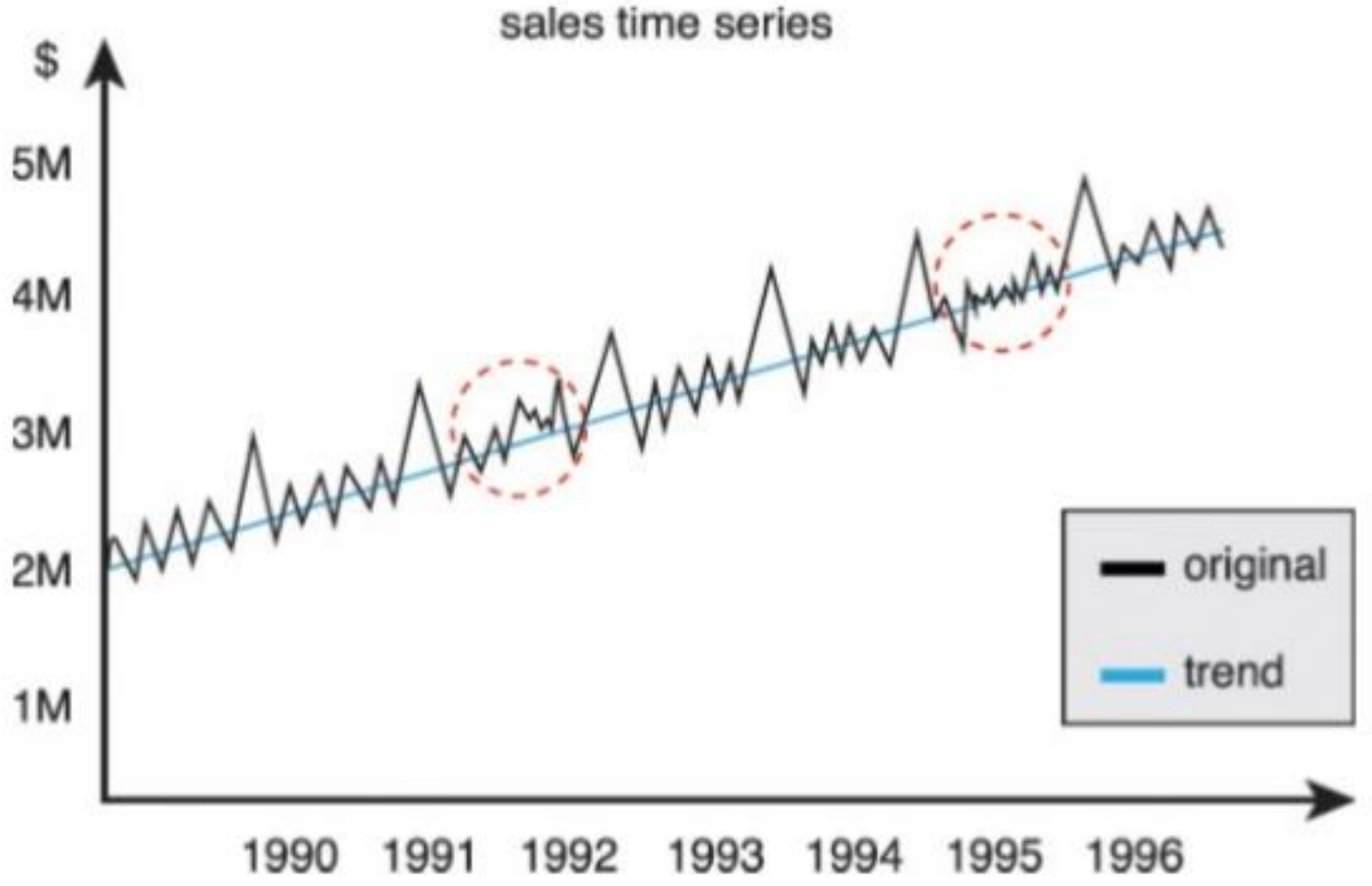
TIME SERIES PLOTS

- Time series plots allow the analysis of data that is recorded over periodic intervals of time.
- This type of analysis makes use of time series, which is a time-ordered collection of values recorded over regular time intervals.
- An example is a time series that contains sales figures that are recorded at the end of each month.
- Time series analysis helps to uncover patterns within data that are time-dependent.
- Once identified, the pattern can be extrapolated for future predictions.

FOR EXAMPLE

- To identify seasonal sales patterns, monthly ice cream sales figures are plotted as a time series, which further helps to forecast sales figures for the next season.
- Time series analyses are usually used for forecasting by identifying long-term trends, seasonal periodic patterns and irregular short-term variations in the dataset.
- Unlike other types of analyses, time series analysis always includes time as a comparison variable, and the data collected is always time-dependent.
- A time series plot is generally expressed using a line chart, with time plotted on the x-axis and the recorded data value plotted on the y-axis, as shown in Figure

LINE CHART



A line chart depicts a sales time series from 1990 to 1996.

- The time series presented in Figure spans seven years.
- The evenly spaced peaks toward the end of each year show seasonal periodic patterns, for example Christmas sales.
- The dotted red circles represent short-term irregular variations.
- The blue line shows an upward trend, indicating an increase in sales.

Sample questions can include:

- How much yield should the farmer expect based on historical yield data?
- What is the expected increase in population in the next 5 years?
- Is the current decrease in sales a one-off occurrence or does it occur regularly?

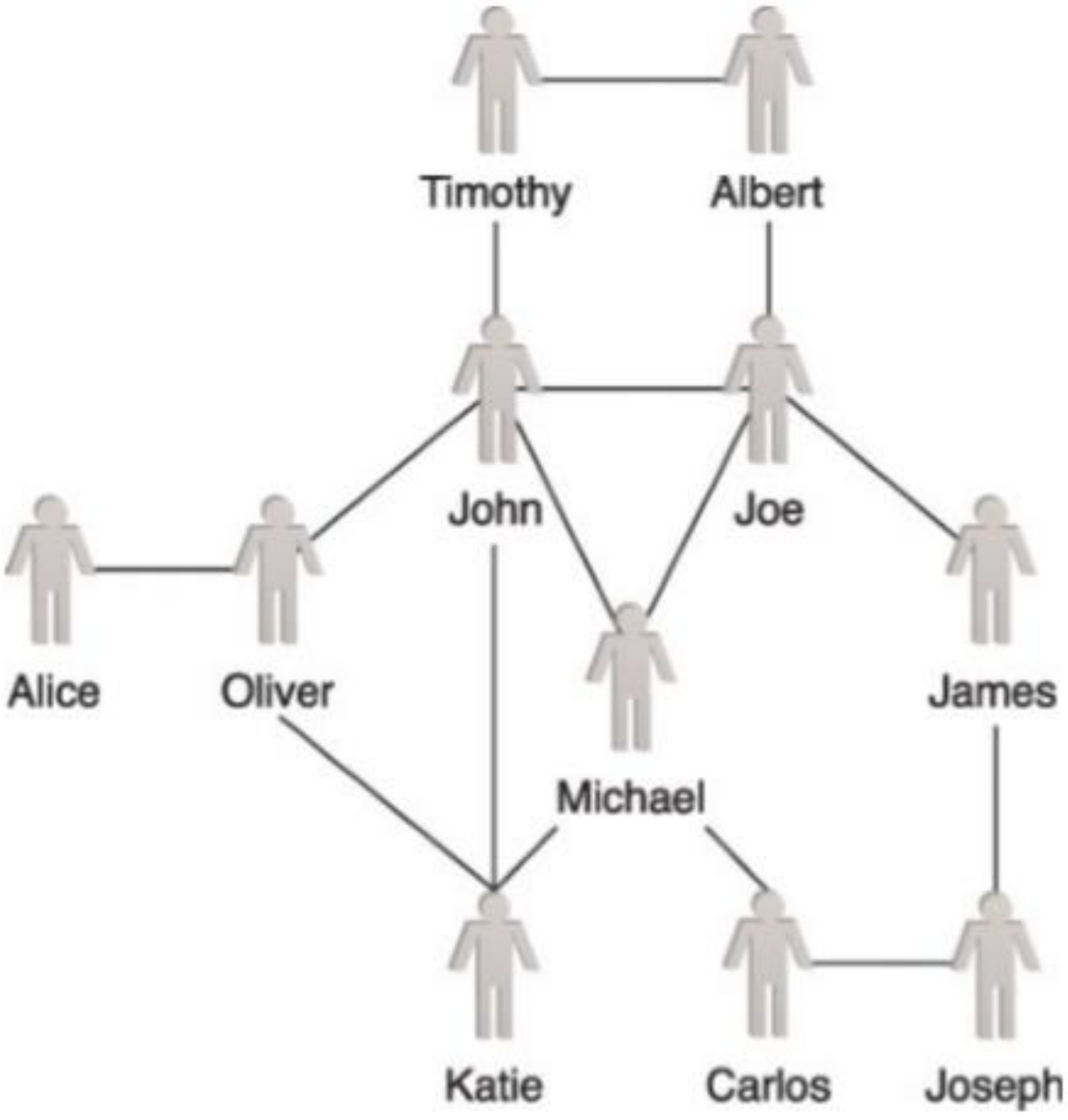
NETWORK GRAPHS

- Within the context of visual analysis, a network graph depicts an interconnected collection of entities.
- An entity can be a person, a group, or some other business domain object such as a product.
- Entities may be connected with one another directly or indirectly.
- Some connections may only be one-way, so that traversal in the reverse direction is not possible.
- Network analysis is a technique that focuses on analyzing relationships between entities within the network.
- It involves plotting entities as nodes and connections as edges between nodes.

NETWORK ANALYSIS...CONTD.

- There are specialized variations of network analysis, including:
  - ❑route optimization
  - ❑social network analysis
  - ❑spread prediction, such as the spread of a contagious disease
- The following is a simple example based on ice cream sales for the application of network analysis for route optimization.
- Some ice cream store managers are complaining about the time it takes for delivery trucks to drive between the central warehouse and stores in remote areas.
- On hotter days, ice cream delivered from the central warehouse to the remote stores melts and cannot be sold.
- Network analysis is used to find the shortest routes between the central warehouse and the remote stores in order to minimize the durations of deliveries.

Consider the social network graph in Figure for a simple example of social network analysis:
John has many friends, whereas Alice only has one friend.
The results of a social network analysis reveal that Alice will most likely befriend John and Katie, since they have a common friend named Oliver.



An example of a social network graph.

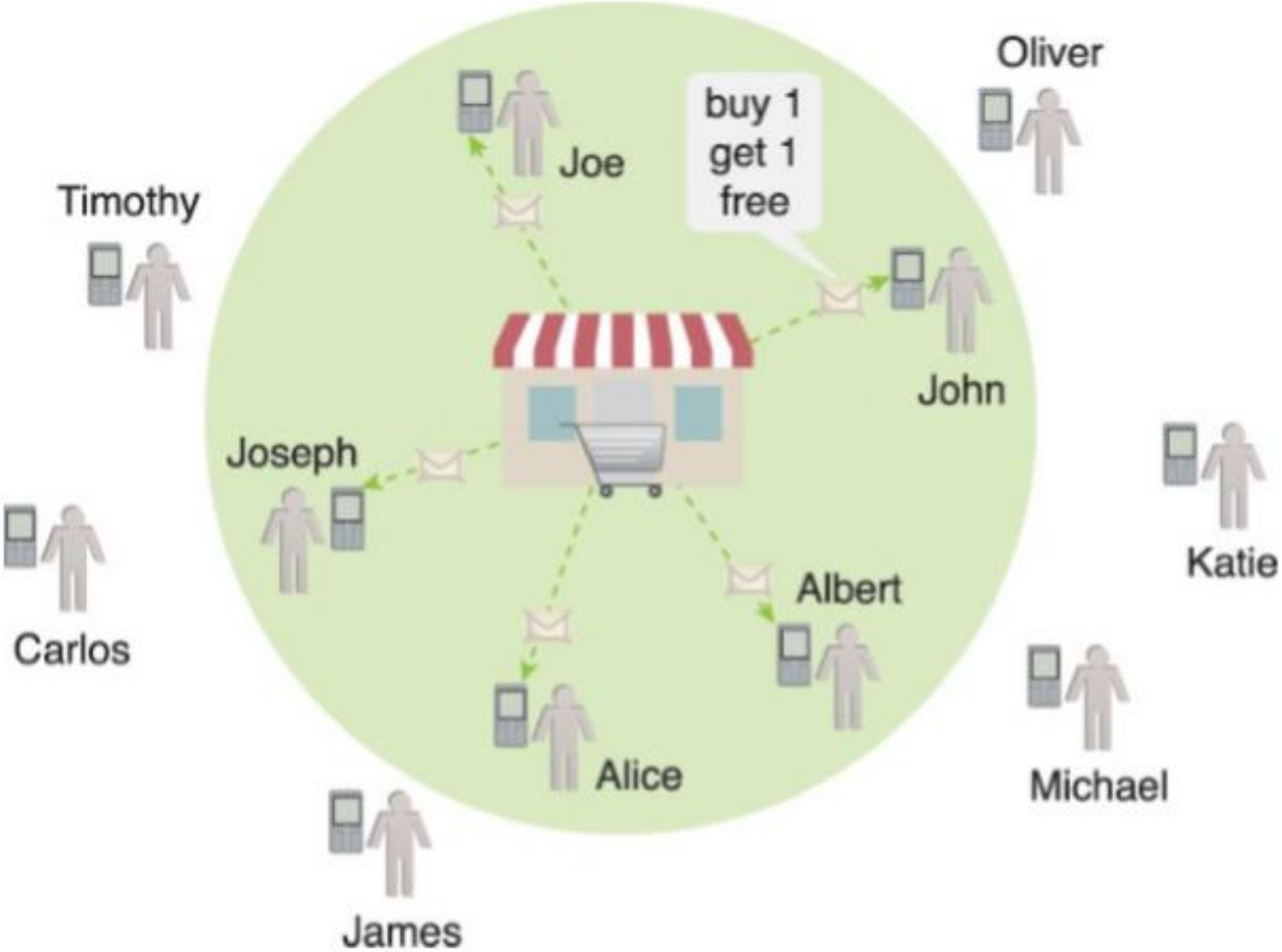
SAMPLE QUESTIONS MAY INCLUDE:

- How can I identify influencers within a large group of users?
- Are two individuals related to each other via a long chain of ancestry?
- How can I identify interaction patterns among a very large number of protein-toprotein interactions?

SPATIAL DATA MAPPING

- Spatial or geospatial data is commonly used to identify the geographic location of individual entities that can then be mapped.
- Spatial data analysis is focused on analyzing location-based data in order to find different geographic relationships and patterns between entities.
- Spatial data is manipulated through a Geographic Information System (GIS) that plots spatial data on a map generally using its longitude and latitude coordinates.
- The GIS provides tooling that enables interactive exploration of the spatial data, for example measuring the distance between two points, or defining a region around a point as a circle with a defined distance-based radius.
- With the ever-increasing availability of location based data, such as sensor and social media data, spatial data can be analyzed to gain location insights.
- For example, as part of a corporate expansion, more ice cream stores are planned to open. There is a requirement that no two stores can be within a distance of 5 kilometers of each other to prevent the stores from competing with each other.
- Spatial data is used to plot existing store locations and to then identify optimal locations for new stores at least 5 kilometers away from existing stores.

- Applications of spatial data analysis include
  - ❑operations and logistic optimization
  - ❑environmental sciences
  - ❑infrastructure planning.
- Data used as input for spatial data analysis can either contain exact locations, such as longitude and latitude, or the information required to calculate locations, such as zip codes or IP addresses.
- Spatial data analysis can be used to determine the number of entities that fall within a certain radius of another entity.
- For example, a supermarket is using spatial analysis for targeted marketing, as shown in Figure.
- Locations are extracted from the users' social media messages, and personalized offers are delivered in realtime based on the proximity of the user to the store.



Spatial data analysis can be used for targeted marketing.

SAMPLE QUESTIONS CAN INCLUDE

- How many houses will be affected due to a road widening project?
- How far do customers have to commute in order to get to a supermarket?
- Where are the high and low concentrations of a particular mineral based on readings taken from a number of sample locations within an area?

THANK YOU