SAKARYA
ÜNİVERSİTESİ

BIG DATA

TOO BIG TO IGNORE

SÜMEYYE KAYNAK

OUTLINE

Big Data Analysis Techniques

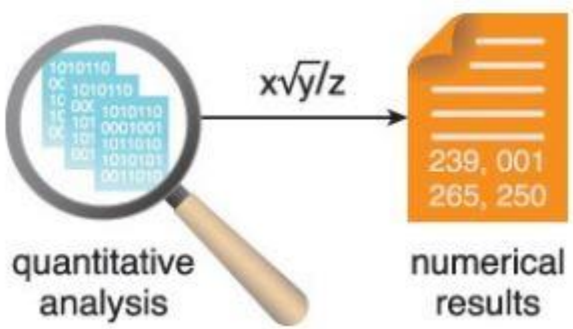
BASIC TYPE OF DATA ANALYSIS

- 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.

QUANTITATIVE ANALYSIS



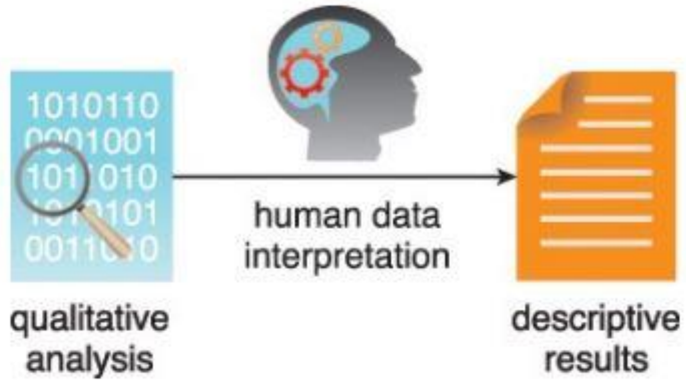
Quantitative analysis results are absolute in nature and can therefore be used for numerical comparisons.

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.

QUALITATIVE ANALYSIS



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

- 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.

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).

STATISTICAL ANALYSIS

- Statistical analysis uses statistical methods based on mathematical formulas as a means for analyzing data.
- Statistical analysis is most often quantitative but can also be qualitative.
- This type of analysis is 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.

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.

A/B TESTING

- Although A/B testing 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.

A/B TESTING

- 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.

A/B TESTING

- 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.

BUY NOW
BUY NOW
Email A
Email B

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

A/B TESTING

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

- Correlation is 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.
- We may be further interested in discovering 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.

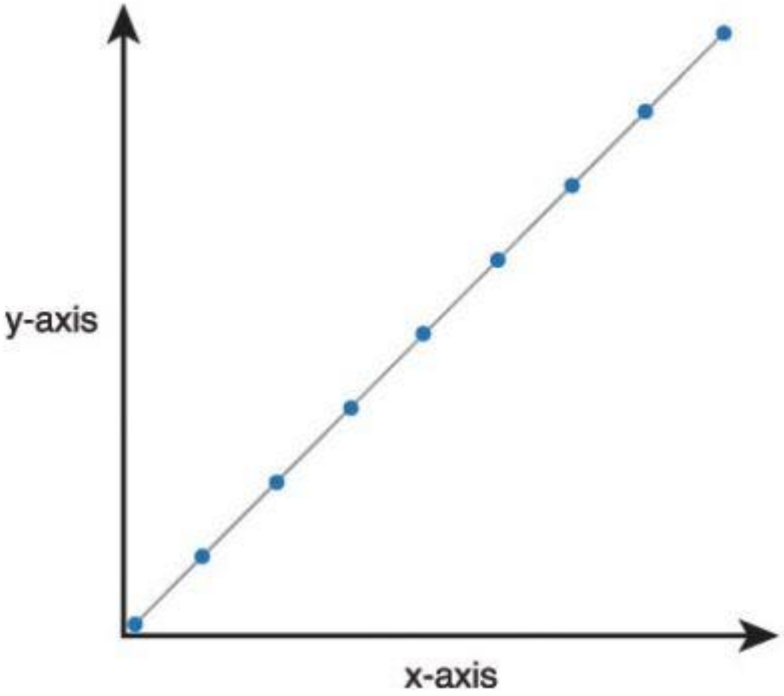
CORRELATION

- The use of correlation helps to develop an understanding of a dataset and find relationships that can assist in explaining a phenomenon.
- Correlation is therefore 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.

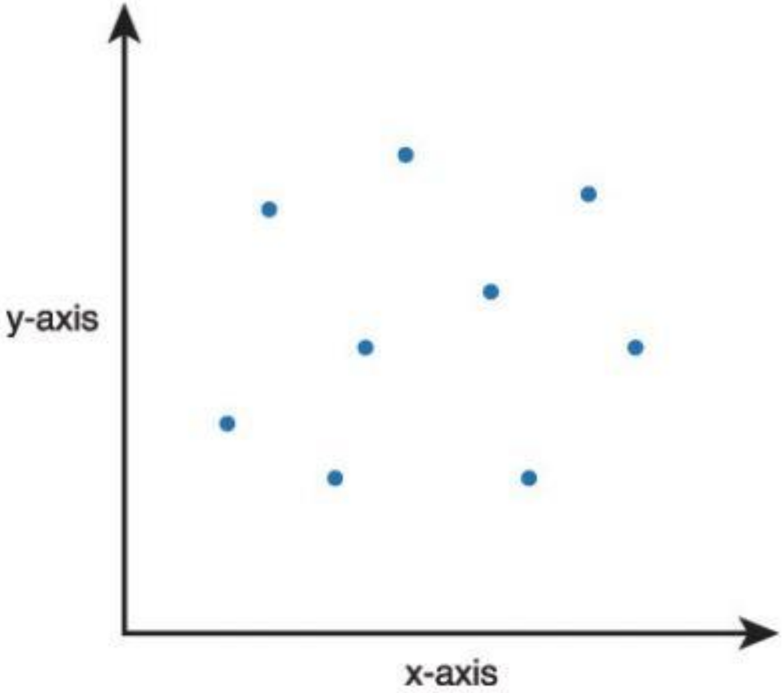
CORRELATION

- 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.

CORRELATION

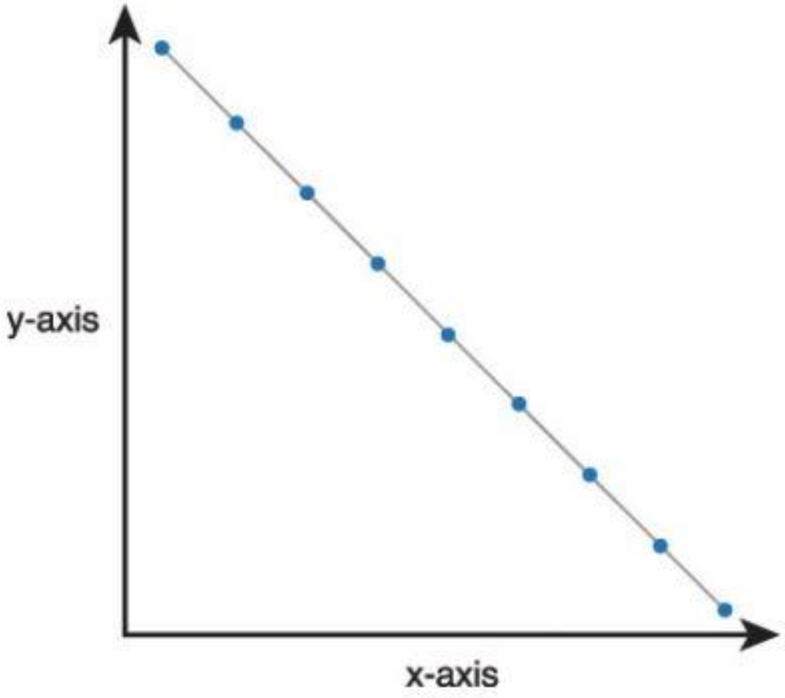


When one variable increases, the other also increases and vice versa.



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

CORRELATION



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

CORRELATION

- 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.

CORRELATION

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.

REGRESSION

- 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?

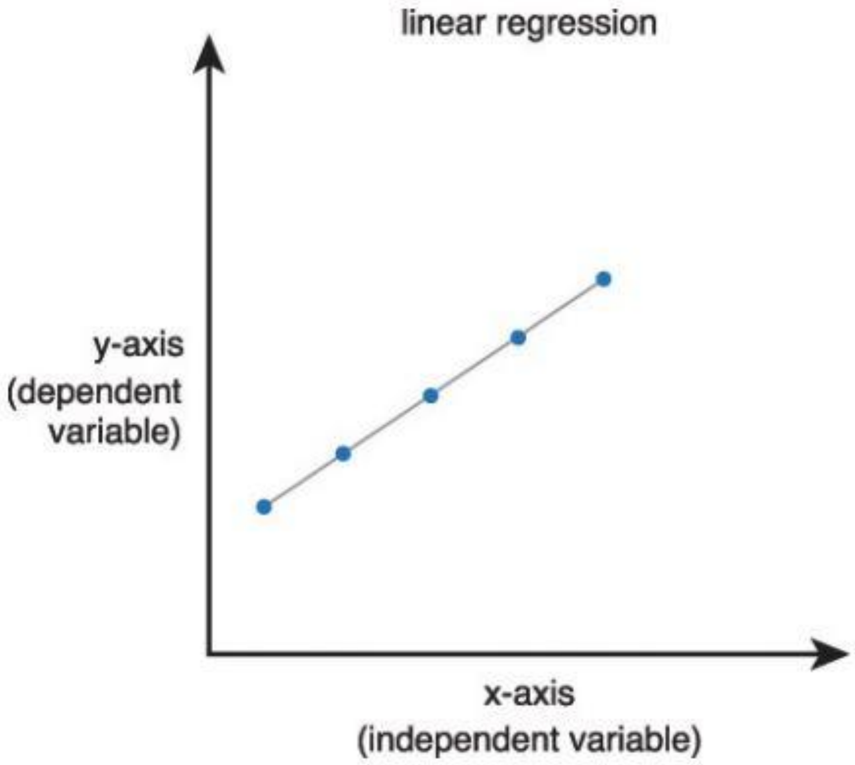
REGRESSION

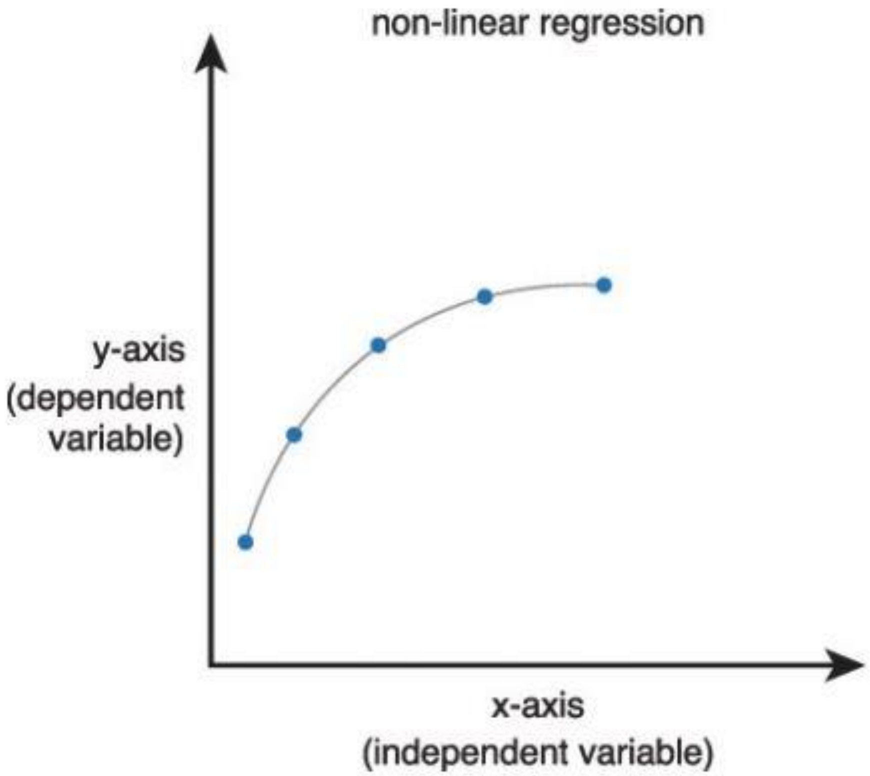
- For example, in order to determine how much extra stock each ice cream store needs to have, the analysts apply regression by feeding in the values of temperature readings.
- These values are based on the weather forecast as an independent variable and the number of ice creams sold as the dependent variable.
- What the analysts discover is that 15% of additional stock is required for every 5-degree increase in temperature.

REGRESSION

- 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 AND NON-LINEAR REGRESSION





REGRESSION

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?

REGRESSION- CORRELATION

- Regression and correlation have a number of important differences.
- 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- CORRELATION

- Regression, on the other hand, is applicable to variables that have previously been identified as dependent and independent variables and implies that there is a degree of causation between the variables.
- The causation may be direct or indirect.

REGRESSION- CORRELATION

- 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

Types of machine learning techniques:

- Classification
- Clustering
- Outlier Detection
- Filtering

CLASSIFICATION

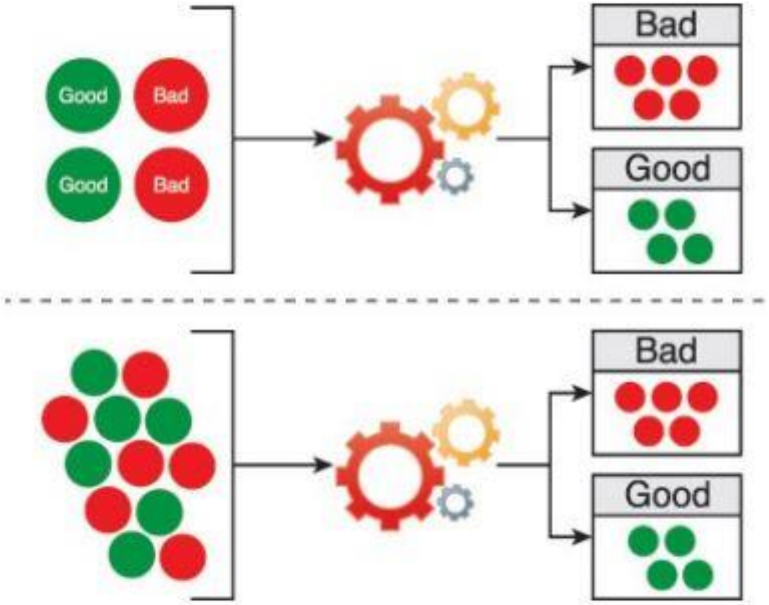
- Classification is a supervised learning technique by which data is classified into relevant, previously learned categories.
- It consists of two steps:

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.

CLASSIFICATION

- A common application of this technique is for the filtering of email spam.
- 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.
- The machine is then fed unlabeled data, which it classifies itself.

CLASSIFICATION



- 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.

CLASSIFICATION

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

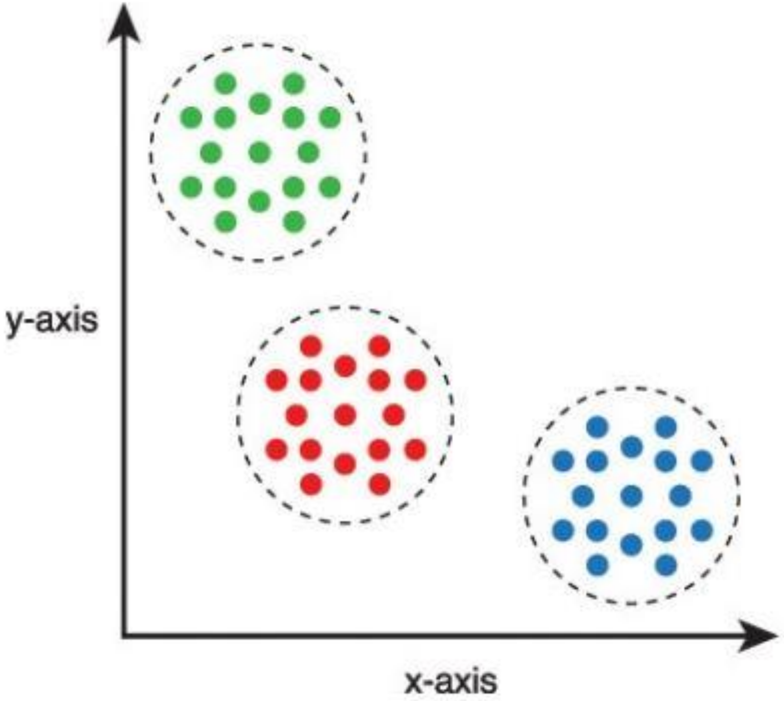
- Clustering is an unsupervised learning technique by which data is divided into different groups so that the data in each group has similar properties.
- There is no prior learning of categories required. Instead, categories are implicitly generated based on the data groupings.
- How the data is grouped depends on the type of algorithm used.
- Each algorithm uses a different technique to identify clusters.

CLUSTERING

- 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

- 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 summarizes the results of clustering.

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.

CLUSTERING

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?

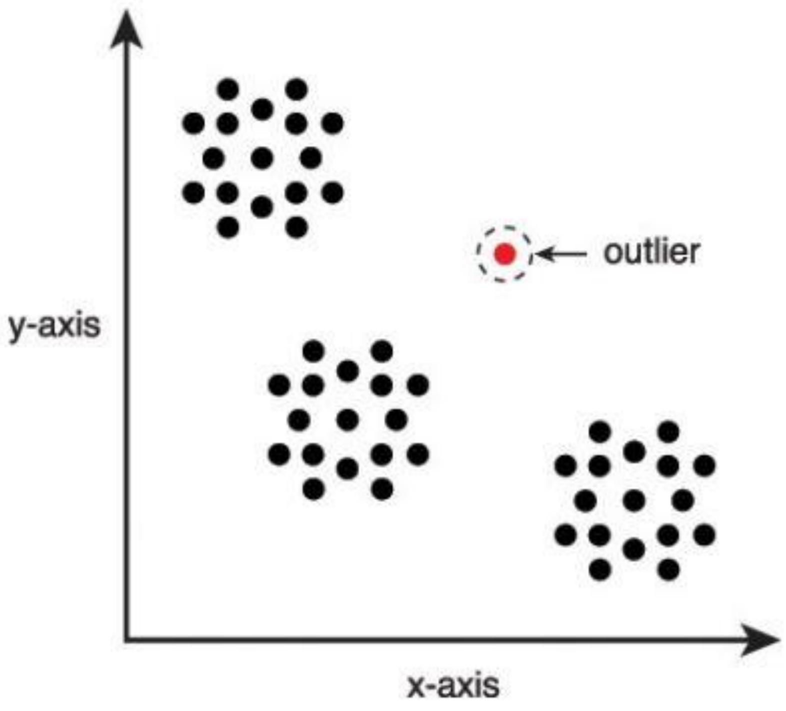
OUTLIER DETECTION

- 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.
- This machine learning technique is used to identify anomalies, abnormalities and deviations that can be advantageous, such as opportunities, or unfavorable, such as risks.

OUTLIER DETECTION

- Outlier detection 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 for outlier detection include fraud detection, medical diagnosis, network data analysis and sensor data analysis.

OUTLIER DETECTION



- 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.

OUTLIER DETECTION

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.

FILTERING

- A common medium by which filtering is implemented is via the use of a recommender system.
- 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.

FILTERING

- 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.

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.

FILTERING

- 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.

FILTERING

- 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.

FILTERING

Sample questions can include:

1. How can only the news articles that a user is interested in be displayed?
2. Which holiday destinations can be recommended based on the travel history of a vacationer?
3. Which other new users can be suggested as friends based on the current profile of a person?

SEMANTIC ANALYSIS

- 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

- Natural language processing is 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.
- For 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.
- In general, the more learning data the computer has, the more correctly it can decipher human text and speech.

NATURAL LANGUAGE PROCESSING

- 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 generally much more 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.

TEXT ANALYTICS

- As a continuation of the preceding NLP example, the transcribed textual data is further analyzed using text analytics to extract meaningful information about the common reasons behind customer discontent.

TEXT ANALYTICS

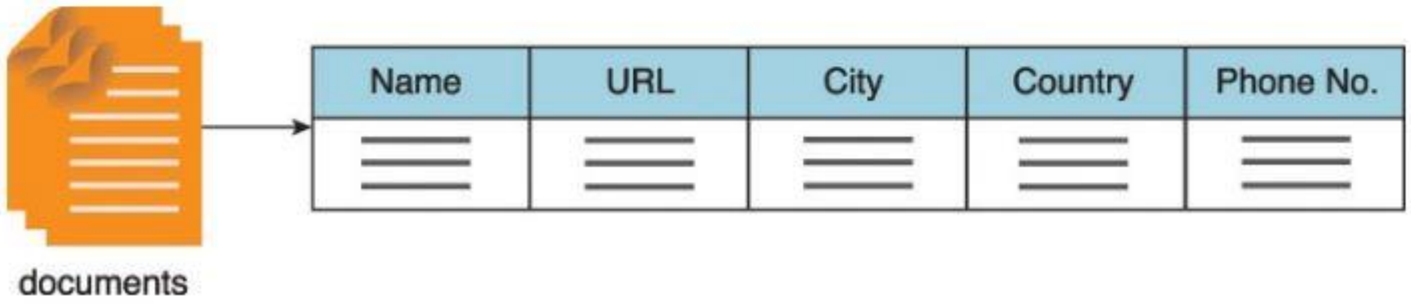
- The basic tenet of text analytics is to turn unstructured text into data that can be searched and analyzed.
- As the amount of digitized documents, emails, social media posts and log files increases, businesses have an increasing need to leverage any value that can be extracted from these forms of 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.

TEXT ANALYTICS

Text analytics generally involves two steps:

1. Parsing text within documents to extract
2. Categorization of documents using these extracted entities and facts.

TEXT ANALYTICS



Name	URL	City	Country	Phone No.

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.

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

- Sentiment analysis is a specialized form of text analysis that focuses on determining the bias or emotions of individuals.
- This form of analysis determines the attitude of the author of the text by analyzing the text within the context of the natural language.
- Sentiment analysis 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, and spotting new trends.

SENTIMENT ANALYSIS

- 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.

SENTIMENT ANALYSIS

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.

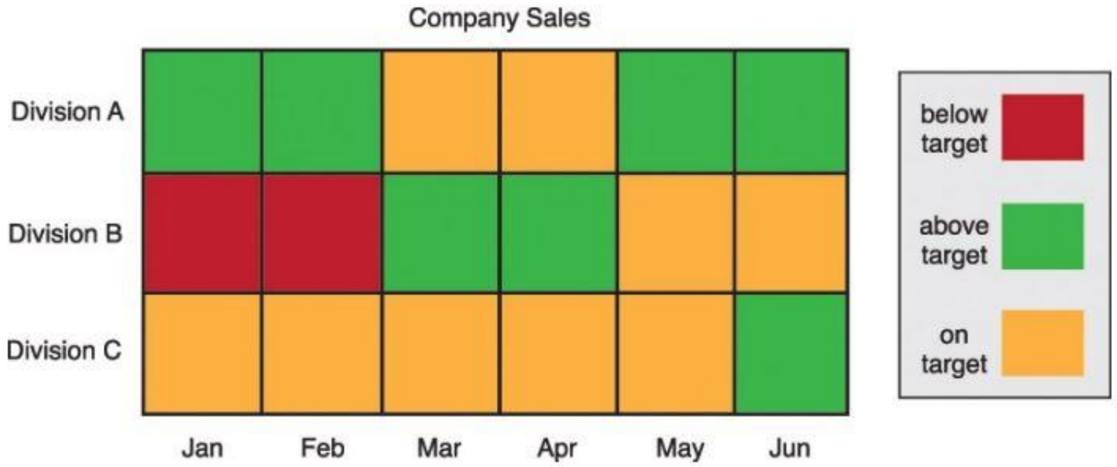
HEAT MAPS

- Heat maps are an effective visual analysis technique for expressing patterns, data compositions via part-whole relations and geographic distributions of data.
- They also 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.

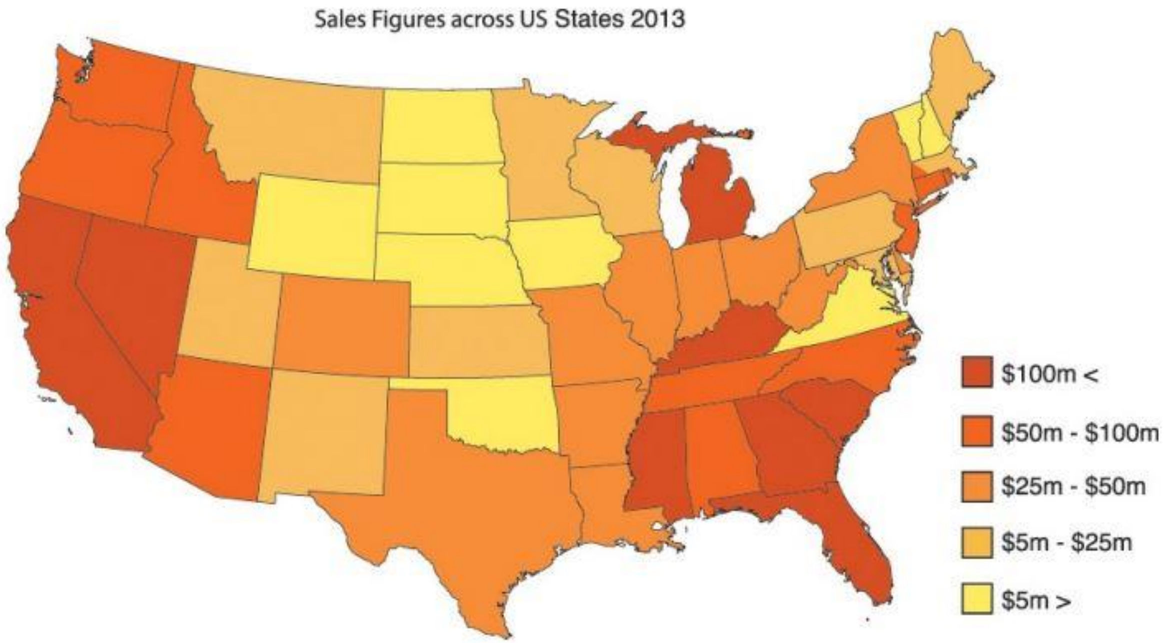
HEAT MAPS

- 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.

HEAT MAPS



	Jan	Feb	Mar	Apr	May	Jun
Division A	🟩	🟩	🟧	🟧	🟩	🟩
Division B	🟥	🟥	🟩	🟩	🟧	🟧
Division C	🟧	🟧	🟧	🟧	🟧	🟩



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.

HEAT MAPS

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?

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.

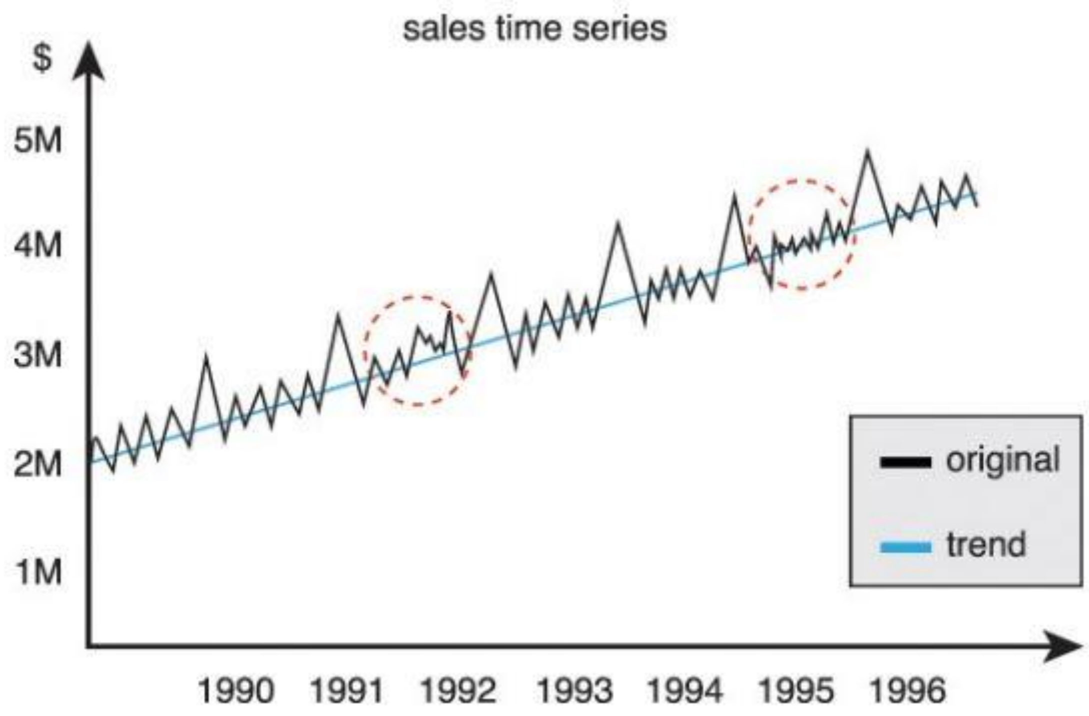
TIME SERIES PLOTS

- 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 PLOTS

- 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.

TIME SERIES PLOTS



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 GRAPHS

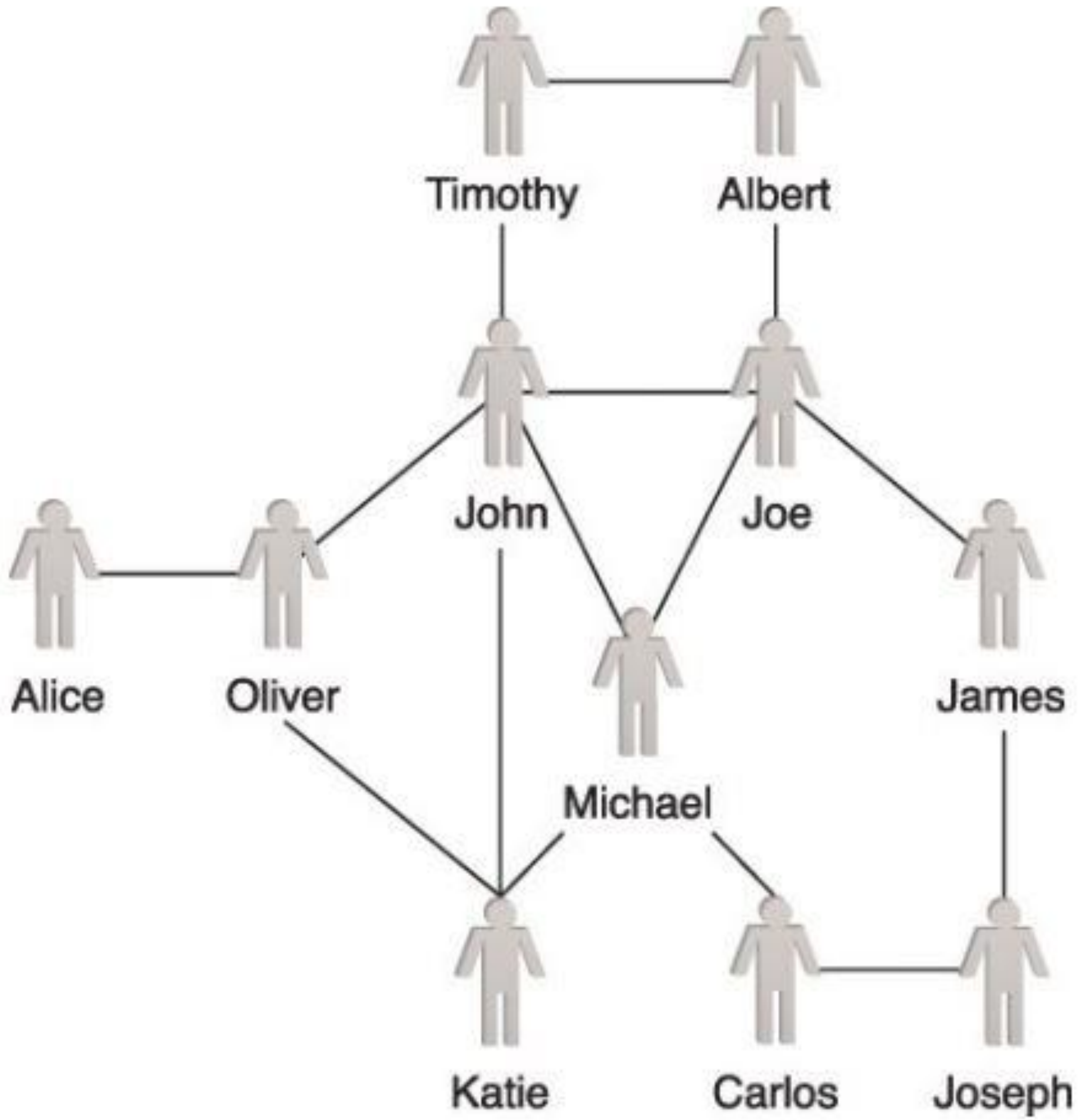
Network analysis is a technique that focuses on analyzing relationships between entities within the network.

It involves plotting entities as nodes and connections as edges between nodes.

There are specialized variations of network analysis, including:

- route optimization
- social network analysis
- spread prediction, such as the spread of a contagious disease

NETWORK GRAPHS



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 MAPPING

- 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 locationbased data, such as sensor and social media data, spatial data can be analyzed to gain location insights.

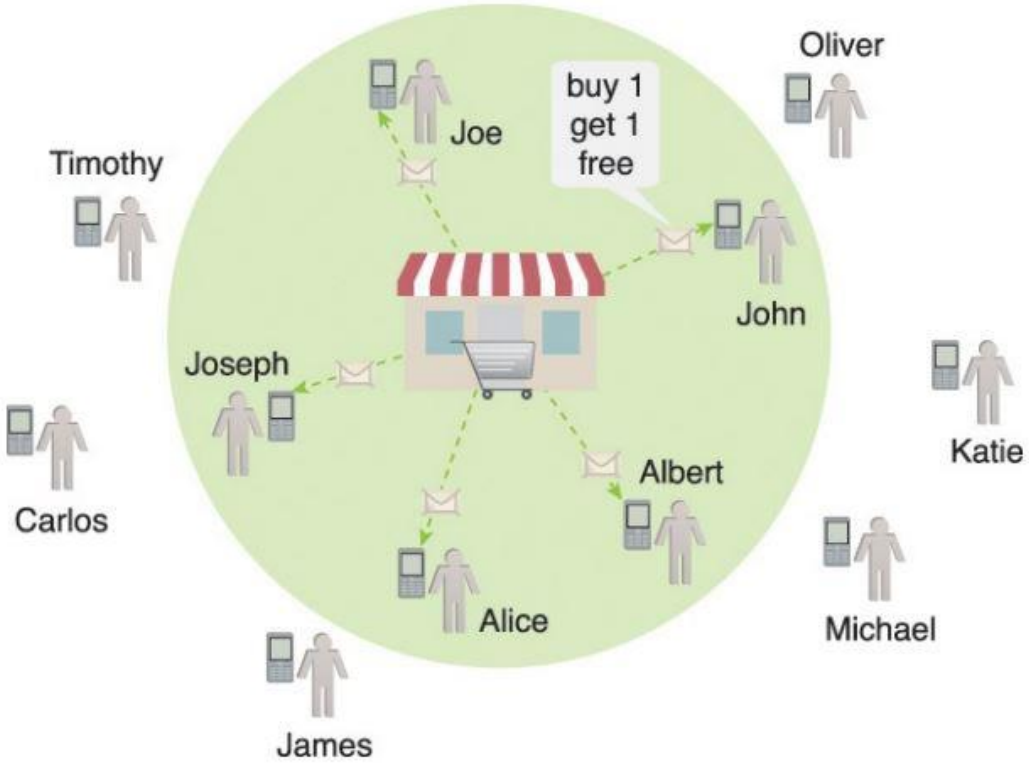
SPATIAL DATA MAPPING

- 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.

SPATIAL DATA MAPPING

- Applications of spatial data analysis include operations and logistic optimization, environmental sciences and 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 MAPPING



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?