

Artificial Intelligence, Machine Learning, and Data Science

Samatrix Consulting Pvt Ltd

Course Introduction

Course Objective

- Learn the concepts of
 - Data Science
 - Data Science Processes
 - Machine Learning
 - Artificial Intelligence

Data Science – Learning Objective

- Introduction to Data Science:
 - Defining Data Science and Big Data,
 - Benefits and Uses of Data Science and Big Data,
 - Facets of Data, Structured Data, Unstructured Data, Natural Language, Machine generated Data, Graph based or Network Data, Audio, Image, Video, Streaming data,
 - Data Science Process,
 - Big data ecosystem and data science, Distributed file systems, Distributed programming framework, data integration framework, machine learning framework, No SQL Databases, scheduling tools, benchmarking tools, system deployments

Data Science Process – Learning Objective

- Data Science Processes:
 - Six steps of data science processes, define research goals, data retrieval, cleansing data, correct errors as early as possible, integrating – combine data from different sources, transforming data,
 - Exploratory data analysis,
 - Data modeling, model and variable selection, model execution,
 - Model diagnostic and model comparison, presentation and automation.

Machine Learning – Learning Objective

- Introduction to Machine Learning:
 - What is Machine Learning, Learning from Data, History of Machine Learning, Big Data for Machine Learning, Leveraging Machine Learning,
 - Descriptive vs Predictive Analytics,
 - Machine Learning and Statistics
 - Artificial Intelligence and Machine Learning,
 - Types of Machine Learning – Supervised, Unsupervised, Semi-supervised, Reinforcement Learning,
 - Types of Machine Learning Algorithms,
 - Classification vs Regression Problem,
 - Bayesian,
 - Clustering
 - Decision Tree
 - Dimensionality Reduction,
 - Neural Network and Deep Learning,
 - Training machine learning systems

AI – Learning Objective

- Introduction to AI:
 - What is AI,
 - Turing test
 - cognitive modelling approach
 - law of thoughts
 - the relational agent approach
 - the underlying assumptions about intelligence
 - techniques required to solve AI problems
 - level of details required to model human intelligence
 - history of AI

Learning Pedagogy

- The course has been split into Sections, Subsections and Units
- Quiz after every subsection that every participant should practice
- Graded and timed test after every section
- Midterm during midway of the course
- End-term after the course
- Assignments and Projects during the course. Please submit on time

Introduction to Data Science

Introduction to Data Science

- One of the biggest challenges that every company, irrespective of the size, across all the industries faces is managing and analyzing the data.
- The ability to manage and analyze has provided the organizations a competitive edge over their competitors.
- Every business struggle with finding a pragmatic approach to capture the relevant information about their products, services, customers, and suppliers.
- In the era of globalization and global supply chains, the markets have become very complicated.
- In the pursuit of gaining a competitive edge with customers, the companies continue to innovate and develop more new products and find innovative ways to reach their customer in the global market

Introduction to Big Data

- To gain insights about the market and customer preference, every company collects the data from a variety of sources such as documents, customer service records, pictures, videos, sensors, social media, and clickstream data generated from website interaction.
- The availability of newer and more powerful mobile devices has created the potential for developing new data sources.
- To manage the intersection of all the different types of data, traditional data management techniques such as Relational Database Management System (RDBMS) are not sufficient.
- Even though the RDBMS has been regarded as one-size-fits-all-solution, the requirement of handling data that varies in type and timeliness, have created the necessity of managing data differently.
- Big Data is the latest trend that has emerged to handle these complexities.

Data Science and Big Data

- Big data helps organizations gather, store, manage, manipulate the large and complex data at the right speed, at the right time to gain the right insights.
- Data Science provides methods to analyze the data and extract meaningful insights out of the data.
- The relationship between big data and data science is the same as the relationship between crude oil and oil refinery.
- Even though the data science and big data have evolved from statistics and data management techniques, they are considered to be an independent field of study.

Characteristics of Big Data

- We can define big data as the data source that has at least the three shared characteristics, often referred to as three Vs
 - Volume: An extremely high volume of data
 - Velocity: Extremely high velocity of data
 - Variety: An extremely high variety of data
- In addition to the three characteristics, the fourth V, veracity, extremely high accuracy of data, make the big data different from the traditional data management systems.
- Every aspect of big data that include data capture, storage, search, curation, share, transfer, and visualization, require specialized techniques.

Tools and Techniques

- Data science finds its roots in statistics.
- It helps organizations deal with the massive data that is produced today.
- Data science requires a knowledge in computer science and statistics.
- A typical job description of a data scientist includes the ability to work with big data and experience in machine learning, computing, algorithm building.
- These job descriptions require the ability to use Hadoop, map reduce, pig, hive, Scala, Spark, R, Python, and Java among others.
- This job description is different from that of a statistician.
- During this course, we will slowly introduce these tools.
- However, our main focus would be on Python. Python has emerged the best language for data science because of the availability of data science specific libraries.
- Every popular NoSQL database offers a python specific API.
- So, the popularity of python has been steadily increasing in the world of data science

Benefits and Uses of Data Science and Big Data

Data Science in Industries

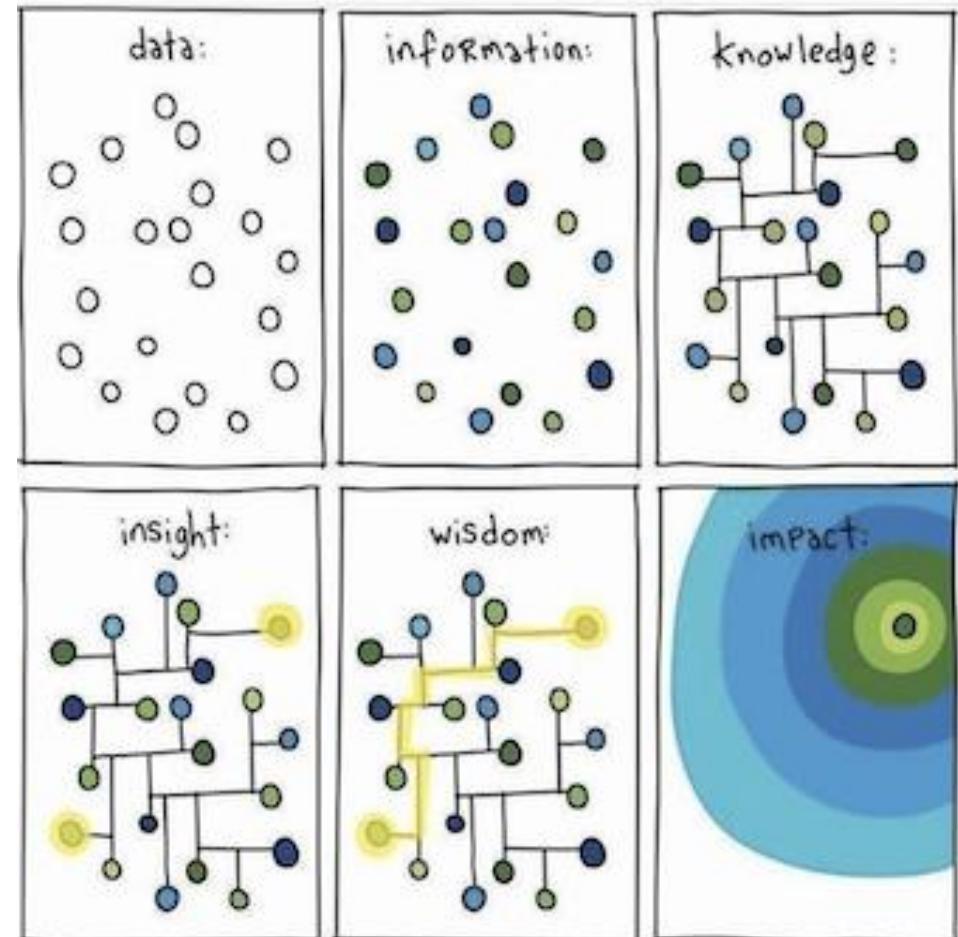
- Today, data science and big data have their footprints in every commercial and non-commercial organization.
- Commercial companies in every industry segment use data science and big data to gain insights about their customers, staff, processes, and products.
- Many companies such as Amazon use the data science and big data techniques in marketing such as targeted marketing, digital advertising, and recommendation systems for cross selling and up selling.
- Organizations use data science for general customer relationship management to analyze the ever-changing customer preferences and behavior to manage attrition and maximize the expected customer value.

Data Science in Industries

- The financial institutions use data science to manage credit risk through credit scoring, predict the stock market, detect fraud, and manage the workforce.
- The human resource department in the organization uses people analytics to screen candidates, monitor the mood of the employees, reduce employee attrition, and improve employee engagement and productivity.
- Many retailers such as Walmart and Amazon use data science tools and techniques throughout their business that includes marketing and global supply chain management.

Data Science For Business

- The primary objective of the course is to help you view business opportunities from data science perspective.
- How the data can be used to gather information and knowledge.
- You should be able to transform data into actionable insights.
- Based on the insights, by using wisdom take decision and actions to create an impact.



Case Study – Hurricane Frances

- Consider an example from a New York Times story from 2004:

“Hurricane Frances was on its way, barreling across the Caribbean, threatening a direct hit on Florida’s Atlantic coast. Residents made for higher ground, but far away, in Bentonville, Ark., executives at Wal-Mart Stores decided that the situation offered a great opportunity for one of their newest data-driven weapons ... predictive technology. A week ahead of the storm’s landfall, Linda M. Dillman, Wal-Mart’s chief information officer, pressed her staff to come up with forecasts based on what had happened when Hurricane Charley struck several weeks earlier. Backed by the trillions of bytes’ worth of shopper history that is stored in Wal-Mart’s data warehouse, she felt that the company could ‘start predicting what’s going to happen, instead of waiting for it to happen,’ as she put it. (Hays, 2004)”

Case Study – Hurricane Frances

- Why in the case of natural calamity, the data-driven prediction might be useful.
- People, who are in the path of a hurricane, might be interested to buy bottled water.
- But this is an obvious point, why we need data science tools and techniques to discover this fact.
- Wal-Mart executives might be interested in predicting how the hurricane will impact the sales so that the company can make the necessary arrangements.
- There could be some coincidence whereas the sales of a particular newly released movie CD went up during the week but the impact on sales was nationwide not just the areas impacted by the hurricane.
- Ms. Dillman is referring to more useful information than some general patterns.

Case Study – Hurricane Frances

- Using data science, the data analyst team could discover the patterns that were not obvious.
- By analyzing the huge volume of Wal-Mart data from prior similar situations, the analyst could identify the surge in unusual local demand for few products and rush stocks ahead of hurricane's landfill.
- In the actual scenario, the same thing happened. The New York Times (Hays, 2004) reported that: “*... the experts mined the data and found that the stores would indeed need certain products—and not just the usual flashlights.*
- ‘*We didn’t know in the past that strawberry PopTarts increase in sales, like seven times their normal sales rate, ahead of a hurricane,’ Ms. Dillman said in a recent interview. ‘And the pre-hurricane top-selling item was beer.’”*

Facets of Data

Variety of Data Types

- Variety is one of the basic principles of big data.
- It is also one of the four characteristics of big data.
- The data scientist should be able to manage a variety of data types.
- The information from various sources of data from bank transactions to tweets to images to videos should be integrated for analysis and data management.

Variety of Data Types

- Based on the business problem, you may come across different facets of data.
- You would require different data management tools and techniques to analyze and extract results for each flavor of data.
- Certain situations such as monitoring traffic data require real-time data management and analysis techniques whereas other situations such as data analysis to determine unsuspected patterns require massive historic data collection.
- In certain situations, we need to integrate data from a variety of sources for our analysis.

Main Categories of Data

- The main categories of data are
 - Structured
 - Unstructured
 - Natural language
 - Machine-generated
 - Graph-based
 - Audio, video, and images
 - Streaming

Why Understanding Data Type is Important

- As soon as a new project is assigned, it is always tempting to jump into the exploration of statistical and machine learning models to get results faster before applying data science.
- However, without understanding the data, you would waste your time and energy in implementing the solutions that are not suitable for the given data type.
- Whenever you are assigned, you should spend time in analyzing the data according to the different categories of data.

Structured Data

- Structured data has defined length and format.
- Number, dates, strings (such as name and address, etc.), are some of the examples of structured data.
- Structured data is usually stored in a database and can be queried using a structured query language (SQL).
- Traditionally the companies have been collecting data from sources such as customer relationship management (CRM) data, operational enterprise resource planning (ERP) data, and financial data.
- Structured data is about 20% of the data that we currently have in the overall system.

Structured Data - Example

Indicator ID	Dimension List	Timeframe	Numeric Value	Missing Value Flag	Confidence Interval
214390830	Total (Age-adjusted)	2008	74.6%		73.8%
214390833	Aged 18-44 years	2008	59.4%		58.0%
214390831	Aged 18-24 years	2008	37.4%		34.6%
214390832	Aged 25-44 years	2008	66.9%		65.5%
214390836	Aged 45-64 years	2008	88.6%		87.7%
214390834	Aged 45-54 years	2008	86.3%		85.1%
214390835	Aged 55-64 years	2008	91.5%		90.4%
214390840	Aged 65 years and over	2008	94.6%		93.8%
214390837	Aged 65-74 years	2008	93.6%		92.4%
214390838	Aged 75-84 years	2008	95.6%		94.4%
214390839	Aged 85 years and over	2008	96.0%		94.0%
214390841	Male (Age-adjusted)	2008	72.2%		71.1%
214390842	Female (Age-adjusted)	2008	76.8%		75.9%
214390843	White only (Age-adjusted)	2008	73.8%		72.9%
214390844	Black or African American only (Age-adjusted)	2008	77.0%		75.0%
214390845	American Indian or Alaska Native only (Age-adjusted)	2008	66.5%		57.1%
214390846	Asian only (Age-adjusted)	2008	80.5%		77.7%
214390847	Native Hawaiian or Other Pacific Islander only (Age-adjusted)	2008	DSU		
214390848	2 or more races (Age-adjusted)	2008	75.6%		69.6%

Structured Data - Pros

- The structured data is highly organized. It can be easily understood by the machine language. The data stored in relational database can be easily and quickly searched and manipulated.
- The business users can use structured data relatively easily. They need not understand various data types and relationships among them. It is easy to develop self-service tools for the business user.
- The relational database has been around for a long time now. Several advanced tools have been developed and tested to manage structured data. It offers Data managers a variety of advanced tools and techniques.

Structured Data - Cons

- The world is not made up of structured data. The natural data is unstructured. The structured data has a predefined structure and can be used only for the intended purpose. Due to this the flexibility and use cases are limited
- The structured data is stored in data warehouses and relational databases. The schema structure of both is very rigid. Any change in data needs would require updates in all the structured data. This results in massive expenditure in terms of time and resources.

Unstructured Data

- Unstructured data does not follow any specified format, rather it is stored in its native format.
- Unstructured data is not processed until it is used.
- That is why unstructured data is also known as schema-on-read.
- Unstructured data is the most prevalent data.
- It is everywhere.
- Approximately 80% of the data is unstructured.

Unstructured Data

- Unstructured data has no pre-defined model so it is difficult to deconstruct and cannot be organized in relational databases.
- Instead, we use non-relational, or NoSQL databases to manage unstructured data.
- We can also use a data lake to manage data in a raw and unstructured format.

Examples of Unstructured Data

- Satellite images, weather data, remote sensing images, Google Earth, etc.
- Scientific data include seismic imagery, atmospheric data, etc.
- Photographs and video including security, surveillance, and traffic video
- Radar or sonar data including vehicular, meteorological, and oceanographic seismic profiles.
- Social media data generated from social media platforms such as YouTube, Facebook, Twitter, LinkedIn, and Flickr
- Mobile data including text messages and location information
- E-mail, survey results, documents, and logs

Unstructured Data - Pros

- Unstructured data can provide a much deeper understanding of the customer behavior and intent whereas the structured data provide the bird eye-view of the customer.
 - For example, data science techniques help understand customer buying habits, timings, spending patterns, sentiments towards certain products and services.
- Unstructured data collected from sensors provide real-time data and helps in predictive analytics.
 - For example, sensors installed on a machine can detect the anomalies in the manufacturing process and adjust itself to avoid a costly breakdown.
- Since the unstructured data is stored in native format and is not defined until needed, the purpose of the data is adaptable. Due to which data can be used for a wider range of use cases.
- Since the data is not defined at the time of accumulation, the data can be collected quickly and easily
- Organizations prefer to store the unstructured data in a cloud data lake, which is a cost-effective and scalable solution.

Unstructured Data - Cons

- One of the biggest drawbacks of unstructured stats is the requirement of data science expertise.
- A normal user cannot prepare and analyze the data because the data is unorganized and unstructured
- In order to manipulate the unstructured data, the data manager needs the specialized tools that are still in their infancy.

Structured vs Unstructured Data

	Structured Data	Unstructured Data
Who	Self Service Access	Requires Data Science Expertise
What	Only Select Data Types	Variety of Data Types
When	Schema-on-write	Schema-on-read
Where	Commonly stored in Data Warehouse	Commonly stored in data lakes
How	Predefined Format	Native Format

Structured vs Unstructured Data

Structured Data

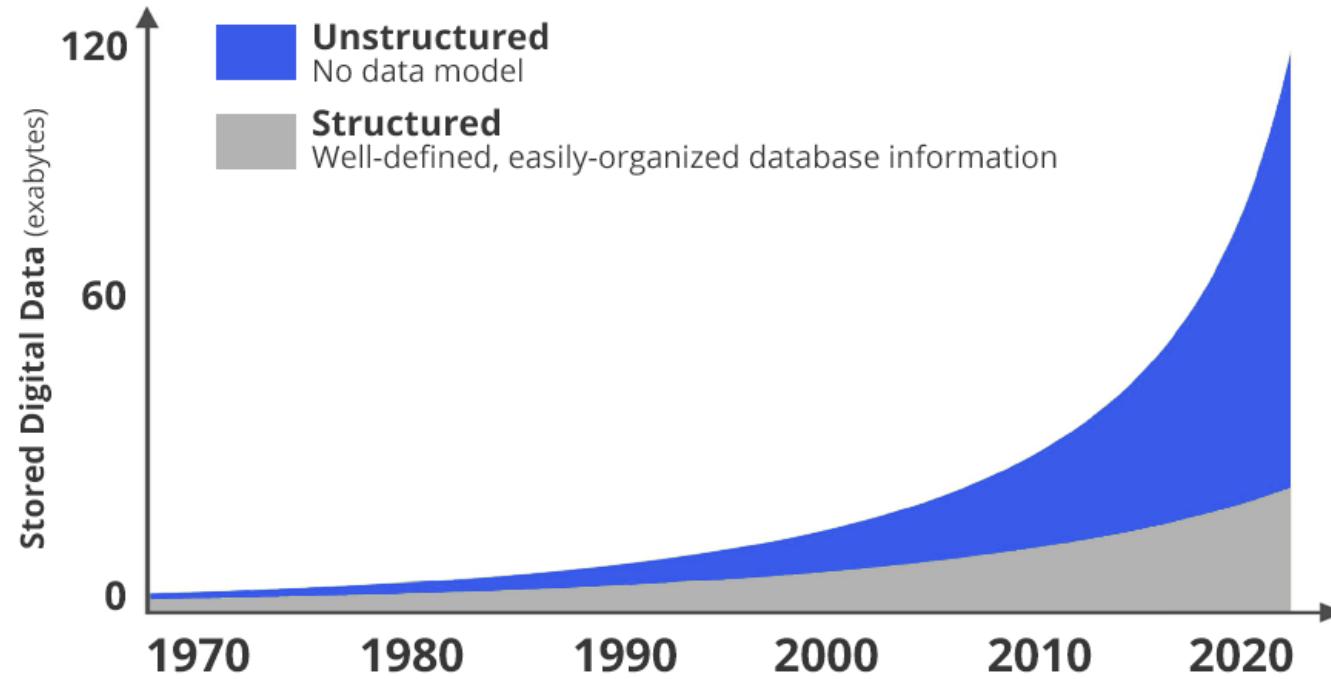


0.103	0.176	0.387	0.300	0.379
0.333	0.384	0.564	0.587	0.857
0.421	0.309	0.654	0.729	0.228
0.266	0.750	1.056	0.936	0.911
0.225	0.326	0.643	0.337	0.721
0.187	0.586	0.529	0.340	0.829
0.153	0.485	0.560	0.428	0.628

Unstructured Data



Structured vs Unstructured Data



Natural Language

- According to the theories from neuropsychology, philosophy of language, and linguistic, a natural language is any language that has evolved in humans due to use and repetition without conscious planning.
- Speech and singing are different forms of natural language.

Human Language

- Human language is highly complex and diverse.
- The human can express himself in infinite ways and there are hundreds of languages across the globe.
- Every language has its own grammar and rules.
- Moreover, every language has its own regional accent.
- Many people mix words from different languages and use abbreviated words while speaking and writing.

Human Machine Interaction

- Researchers from computer science and computational linguistics have been working for many decades to fill the gap between human communication and computer understanding.
- Recently due to increased interest in human-machine interaction, big data, computing power, and enhanced algorithms, significant advancements have taken place.
- Humans speak, write, and understand languages such as English, Hindi, French, and German.
- The native language of computers is machine code or machine language that constitutes millions of zeros and ones.

Natural Language Processing

- Natural language processing uses natural language rules to convert the unstructured data into a computer recognizable form.
- Natural language processing helps the computers understand, interpret, and manipulate the human's natural language and communicate with humans in their own language.
- Teaching machines to understand human communication is a challenging task.
- Today if you say, "Alexa play party songs". Alexa will play a song for you and in its reply tell you the name of the song.

Natural Language Processing

- This interaction has been made possible by natural language processing.
- In this interaction, as soon as the device hears your voice, it is activated.
- It understands your unspoken intent, executes the action, and provides you feedback in well-formed human language within five seconds.
- Natural language processing helps machines analyze staggered text and speech data that is generated every day on social media platforms or medical records and add useful numeric structure for downstream applications such as speech recognition or text analytics.

Machine Generated Data

- A big component of data is created by machine without any human intervention.
- For example, every time a Boeing 787 flies, it generates half a terabyte of data.
 - Every part of the plane generates data and constantly update the inflight crew and ground staff about its status. This machine-generated data comes from various sensors installed on several parts of the plane.
- Machine-generated data could be both structured as well as unstructured

Machine Generated Data

- Internet of Things (IoT):
 - These devices collect data, connect to other devices or networks, and execute services on their own.
 - These smart devices are used everywhere: home, cars, cities, remote areas, the sky, the ocean. They all are connected and generate data.
- Sensor data:
 - Radiofrequency ID (RFID) tags have become a popular technology.
 - RFID uses a small chip to track the objects from distance.
 - The RFID can track the containers as they move from one location to another in the supply chain.
 - Companies have been using RFID technology for supply chain management and inventory control.
 - Our smartphones contain sensors such as GPS that can capture location data to understand consumer behavior.

Machine Generated Data

- **Weblog data:**
 - The servers, applications, and networks create a log of their activities.
 - That creates a massive amount of useful data.
 - This data can be used for proactive server or application maintenance activities or predict security breach
- **Point of sale data:**
 - On the billing counter, as soon as the cashier scans the bar code and the data associated with the product is generated.
 - With many people shopping so many products, a huge amount of data is generated.
- **Financial Data:**
 - Every financial transaction has a digital footprint now.
 - Whether the transaction is related to the banking or stock market, data is generated.
 - Some of the data is machine-generated whereas some data is human-generated.

Machine Generated Unstructured Data

- Satellite images, weather data, remote sensing images, Google Earth, etc.
- Scientific data include seismic imagery, atmospheric data, etc.
- Photographs and video including security, surveillance, and traffic video
- Radar or sonar data including vehicular, meteorological, and oceanographic seismic profiles.

Human Generated Data

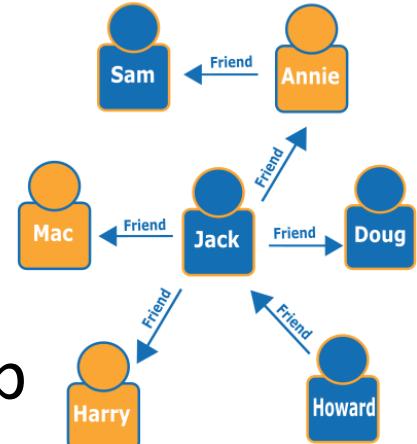
- While interacting with computers, we generate data that is known as human-generated data. This data type includes email, messages, documents, presentations, spreadsheets, audio and video, and images. We create and share this day every day.
- Human-generated data is one of the fastest-growing data. It contains highly valuable and relevant information that is critical for data analysis.
- The human-generated data may not be very big on its own. But considering millions of other users generating such data, the size is astronomical. This data type also has a real-time component that can be used to understand the patterns and predict outcomes.

Human Generated Data - Examples

- Input Data:
 - We input a lot of information into a computer.
 - For example, while filling an online registration or application form, we input a name, age, and address.
 - This data is used to understand basic consumer behavior.
- Click-stream data:
 - Every time you click a link on a website, click-stream data is generated.
 - This data is used to understand the online behavior of an online visitor and buying patterns.
- Gaming-related data:
 - While playing games, every move you make, generates gaming-related data.
 - This data helps understand how the end-user moves throughout a game.
 - This is used to optimize the existing games and develop new games

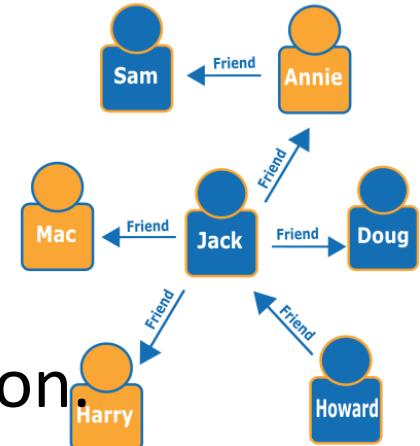
Graph Based or Network Data

- In graph theory, the graph represents a pair-wise relationship between objects.
- Graph databases are used to store and navigate relationships.
- The value of the graph depends upon the relationships. To store and navigate the graphical data, the graph databases use nodes, edges, and properties.
- Data entries are stored in nodes and the relationship between the entities is stored in edges.
- The graph database is queried using specialized query languages such as SPARQL.



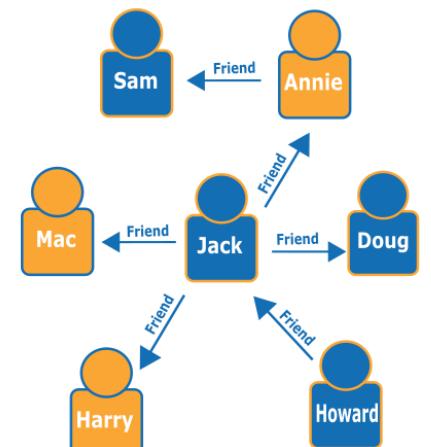
Graph Based or Network Data

- An edge consists of a start node, end node, type, and direction.
- The edge is also used to describe the parent-child relationships, ownerships, and actions.
- A node can have an unlimited number and kind of relationships.
- Using the graph database, a graph can either traverse along with specific edge type or an entire graph.
- The relationships between the nodes are persisted in the database.
- They are not calculated at the query time.
- So, the traversing in the graph database is very fast.



Graph Based or Network Data – Use Cases

- The use cases such as social networking, recommendation engines, and fraud detection use the graph databases extensively.
- In these use cases, there is a need to build a relationship between the data and the ability to query these relationships quickly
- Refer to the following picture that shows an example of a social network graph.
- If you have people (nodes) and relationships among them (edge), you can easily find the “friends of friends” of a particular person.



Graph Based or Network Data – Use Cases

- Fraud Detection:
 - Graph databases are used for fraud detection.
 - Graph relationships can be used for real-time processing of financial and purchase transactions.
 - For example, if a person is using the same email address and credit card as recorded in a known fraud case, you can detect easily using graph databases.
 - You can easily detect the cases where multiple people are associated with one personal email address or cases where multiple people, who are residing at a different physical address, are sharing some IP addresses.
- Recommendation Engines:
 - Using a graph database, you can easily store the relationship between several categories such as customer interests, friends, and purchase history.
 - Using these relationships, you can make product recommendations based on the purchase history of other users with similar interests.

Audio, Images and Videos

- The advancement in multimedia technology, high-speed internet, and smart devices are changing the digital data landscape.
- A huge amount of digital data in the form of text, audio-video, and images is generated every second.
- With technology, the trend is on the increase.
- It is important for any organization to equip itself to address the needs of data analytics for multimedia content.
- Audio, images, video, and other multimedia data analytics have a challenging task for data scientists.
- Tasks such as recognizing an object in an image or video look trivial for humans but are challenging for the machines.

Streaming Data

- The term “streaming” means a continuous data stream that is never-ending and that does not have a beginning or an end.
- The streaming data can be utilized without downloading.
- Streaming data has become an important part of our day to day life.
- We can gather real-time data from online gaming, social media, eCommerce, GPS, and IoT devices.
- In order to gain a competitive advantage, every company wants to take a lead in extracting accurate customer insights from streaming data.
- Real-time data processing data science technologies such as Kafka and Kinesis are used to collect and analyze the data in real-time.

Data Science Processes

Six Steps

- In order to be successful in a data science project at a low cost, it is important to follow a structured approach.
- A typical data science structured approach involves six steps that every project should undertake before taking up any data science project.



1. Setting the research goal

- Before starting any data science project, the first step is to identify the problem statement.
- Along with the problem statement, you should understand why there is a problem and how the solution will help meet the final business objectives.
- This step will help you understand the data that you would require, where would you find it, and the required architecture of the system.

1. Setting the research goal

- For example, if you got a data science assignment from the sales department of a company. Before getting into a solution space you need to understand the problem. You can do so by asking the right questions to understand the sales process and customer.
 1. Who the customers are?
 2. What are the company's unique selling points?
 3. How to predict customer buying behavior?
 4. Which segments are working well and which ones are not and why?
 5. What are the opportunity losses and why?
- Such a question will help you define the problem and understand the business and the data flow across the organization.

2. Retrieve data

- For a data science project, you would be handling a huge amount of data.
- You need to collect lots of data from lots of sources.
- Available data may not meet all your requirements and may not match your problem statement.
- The data may be available within your organization as well as other organizations also.
- The required data for your project should be available at the given sources.
- You also need to ensure that you have the required access to retrieve data from the sources.
- The data quality is also important.
- Before retrieving the data, you need to ensure how trustworthy the data is and what is the data quality.

3. Data preparation

- Raw data that is available at multiple sources may have some inconsistencies that need to be taken care of before using it for your data science project. Examples of issues that a data scientist face are
 - Some important values might be missing
 - Some values might be corrupted. It may have invalid entries
 - Due to time zone differences, the user data may have inconsistencies in the date-time data
 - Due to inconsistencies in capturing data, some data may have date range error.
- The data from the different sources and having different formats and structures need to be integrated.
- Several data science techniques have certain requirements of data format.
- They require data in a separate format than the data that is available.
- So, some data conversion is necessary.

3. Data preparation

- The following steps are often required.
 - Data cleansing: remove false value and inconsistencies from the data source
 - Data integration: enrich data by combining information from multiple sources
 - Data transformation: change the data to a suitable format that can be used in your model

4. Data exploration

- In this phase, you build a deeper understanding of the data.
- You try to understand each variable and relationship among the variables in the data set.
- You access the data distribution and presence of outliers (if any).
- Descriptive statistics, data visualization, and sample modeling are methods used for data exploration.
- We often call this step as Exploratory Data Analysis (EDA).

5. Data modeling

- The models, domain knowledge, and insights from the data exploration phase are used to answer the research questions.
- In this phase, you select a model from statistics, machine learning, etc.
- It is an iterative process where you select variables from the model, execute the model and perform model diagnostics.

6. Presentation and automation

- In this final phase, the outcome of the data science process is implemented in an information system or business process to meet the initial objective.
- In this step, you may also present your findings to your client or business.
- You may present data in various forms such as presentations and research reports.
- If the outcome of the process is required in any other project, you may also automate the execution of the process.

Big Data Ecosystem and Data Science

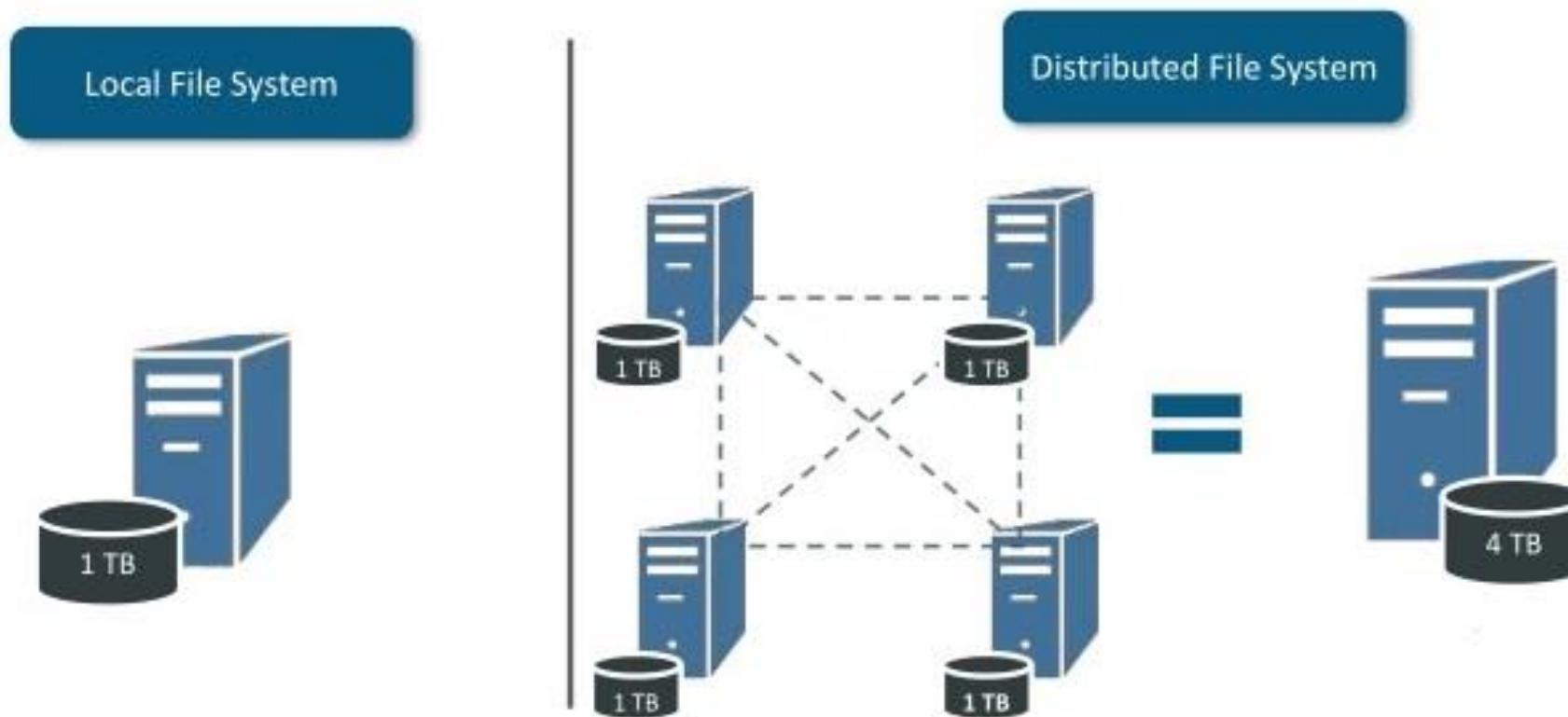
Introduction

- A data scientist can use many big data tools and frameworks.
- But this is an evolving field.
- New technologies appear rapidly.
- For a project, the data scientist uses many technologies but not all the technologies.
- Based on functionalities and goals, we can group the bigdata ecosystem into the technologies.

Distributed File System

- A distributed file system is similar to a normal file system.
- In the case of a distributed file system, many individual computers or servers are networked together across geographies as if they are a single file system.
- The distributed computing environment shares resources ranging from memory to network and storage.
- On the distributed file system, you can perform almost all the actions such as storing, reading, deleting, and securing the files, that you can do on the single file system.

Distributed File System



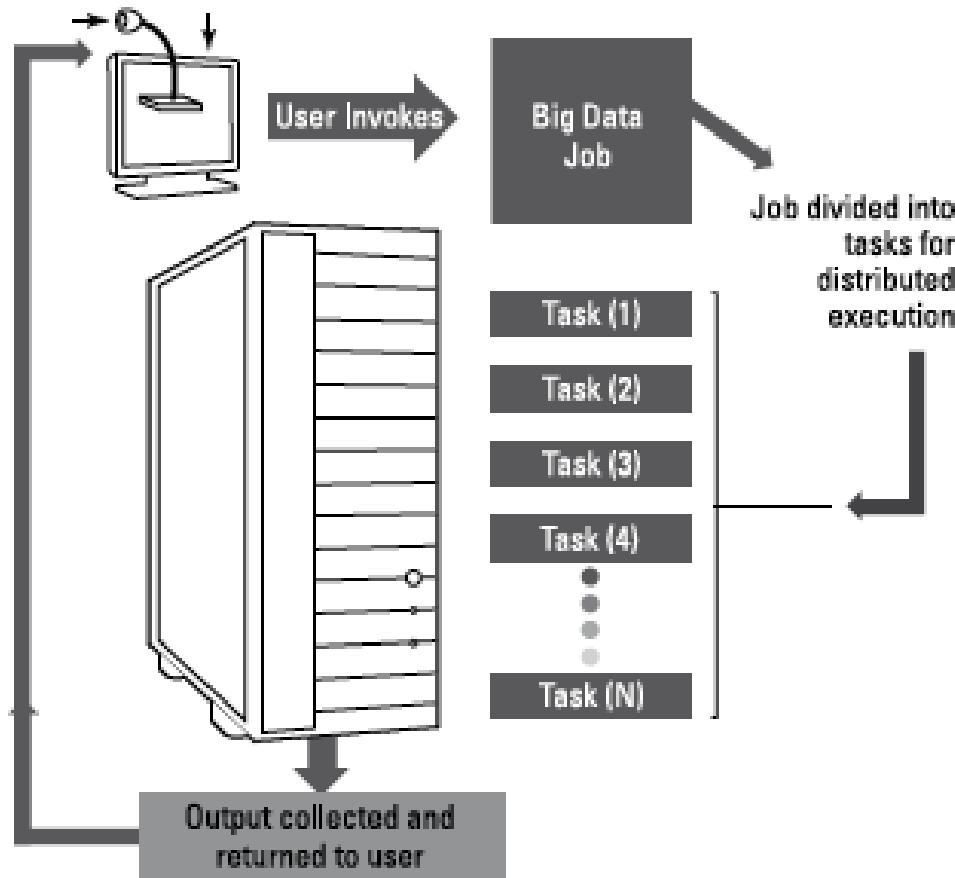
Distributed File System

- The advantages of the distributed file systems are as follows:
 - A file larger than the computer disk can be stored
 - The files are replicated automatically across networked servers for redundancy or parallel operations. Such complexities are hidden from the users
 - The distributed file systems can be scaled easily without adding the memory or storage restrictions

Distributed File System

- Earlier, to achieve the scale, we used to move everything to another server with more memory, storage, and better CPU.
- This process is known as vertical scaling.
- However, with a distributed file system, you can add another small server.
- The principle of horizontal scaling has made the scaling potential virtually limitless.
- Hadoop File System (HDFS) is the best-known distributed file system.
- HDFS is an open-source implementation of the Google File System.
- HDFC is the most popular distributed file system among users.
- Other commonly used distributed file systems are: Red Hat Cluster File System, Ceph File System, and Tachyon File Systems

Distributed Programming Framework



Distributed Programming Framework

- Once the data is stored in the distributed file system, the data scientist would use it for analyzing and solving business problems.
- To enable parallel processing of data, code is transferred to nodes. The tasks are performed on the node where the data is stored.
- In the case of distributed hard disk, we do not move the data to the program, rather we move the program to the data.
- That is why while working with a normal general-purpose programming language such as C, Python, or Java, the data scientist has to deal with the complexities of distributed programming such as restarting jobs that have failed and tracking the results from different sub-processes.
- The distributed programming frameworks such as Apache MapReduce, Apache Pig, Apache Spark, Apache Twill, Apache Hama are few open-source frameworks that help work with the distributed data and deal with the challenges the distributed file system carries.

Data Integration Framework



Data Integration Framework

- Data integration is an important step for any data science project.
- The data integration framework combines data from various sources and data in various formats before presenting meaningful and valuable information to the users.
- In the traditional data warehouse environments, the Extract, Transform, and Load (ETL) technologies have been used.
- However, the elements of the big data platform manage the data differently from the traditional relational database.
- Scalability and high performance are the basic requirements for managing structured and unstructured data.
- Every component of the big data ecosystem ranging from Hadoop to NoSQL database has its own approach for extracting, transforming, and loading the data.
- Apache Airflow, Apache Kafka, Apache Flume, Chukwa, Scribe, Talend Open Studio are some of the examples of open source data integration frameworks.

Machine Learning Framework

- After collecting the data in the required formats, the data scientists focus on extracting the insights.
- In this phase, they use machine learning, statistics, and applied mathematics.
- To deal with high volume and complex data, we need specialized machine learning frameworks and libraries.
- A machine learning framework is an interface, library, or tool that helps the data scientists build machine learning models without getting into the underlying mathematical and statistical algorithms.

Machine Learning Framework

The most popular machine learning frameworks are as follows:

- **Tensorflow**

- is a python library that is provided by Google.
- It is an open-source python library that is used for numerical computation using data flow graphs.

- **Keras**

- is python based open-source neural-network library that can run on the top of TensorFlow, Theano, R, Microsoft Cognitive Toolkit, or PlaidML.
- Keras is capable of conducting fast experimentation with deep neural networks.
- It is a user-friendly, modular, and extensible.

Machine Learning Framework

- **Scikit-learn**
 - is one of the most popular machine-learning frameworks.
 - Scikit-learn can easily implement supervised and unsupervised learning algorithms.
 - The library has modules for classification, regressions, and clustering algorithms.
 - Scikit-learn can interoperate easily with Python numerical and scientific libraries: NumPy and SciPy.
- **Apache Spark MLlib**
 - interoperates well with NumPy in Python and R libraries.
 - You can also use HDFC, HBase, or any other Hadoop data source that makes MLlib easy to plug into the Hadoop workflows.
 - MLlib leverages iterative computation and contains high-quality algorithms.
 - Due to which MLlib can provide better results as compared to one-pass approximations used on MapReduce.

Machine Learning Framework

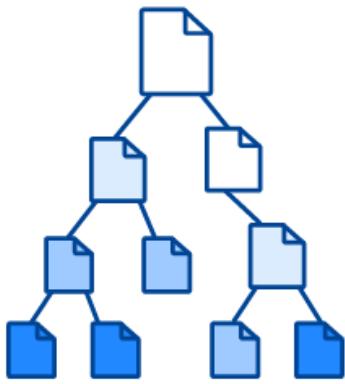
- In addition to the frameworks given above, there are other popular machine learning frameworks such as Azure ML Studio, Google Cloud ML Engine, Torch, and Amazon Machine Learning frameworks.

NoSQL Database

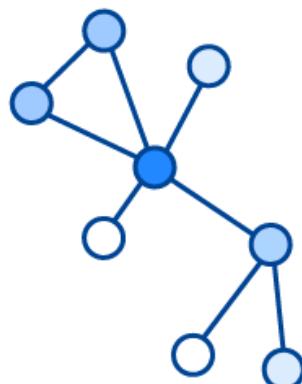
- To store a huge amount of data, the data scientist needs specialized software that can manage and query this data.
- Traditionally, relational databases such as Oracle SQL, MySQL, and Sybase IQ to manipulate and retrieve the data.
- With the emergence of new data types especially streaming, graphs, and unstructured datasets, the traditional databases cannot scale well.
- However, NoSQL databases can allow the endless growth of the data.
- The term “NoSQL” stands for “non SQL” or “not only SQL”.
- NoSQL databases can store data in a format that is different from relational tables.
- NoSQL databases can store the relational data in a more effective way than in a relational table because in the case of a NoSQL database there is no need of splitting the data between the tables.

NoSQL Database

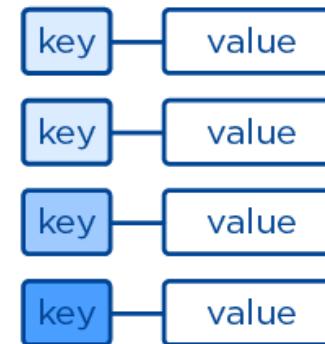
Document



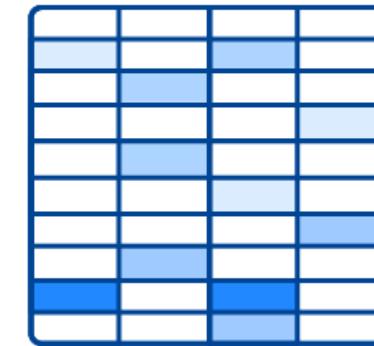
Graph



Key-Value



Wide-column



NoSQL Database

Over time, four major types of NoSQL databases have emerged

- Document databases

Relational

ID	first_name	last_name	cell	city	year_of_birth	location_x	location_y
1	'Mary'	'Jones'	'516-555-2048'	'Long Island'	1986	'-73.9876'	'40.7574'

ID	user_id	profession
10	1	'Developer'
11	1	'Engineer'

ID	user_id	name	version
20	1	'MyApp'	1.0.4
21	1	'DocFinder'	2.5.7

ID	user_id	make	year
30	1	'Bentley'	1973
31	1	'Rolls Royce'	1965

MongoDB

```
{  
  first_name: "Mary",  
  last_name: "Jones",  
  cell: "516-555-2048",  
  city: "Long Island",  
  year_of_birth: 1986,  
  location: {  
    type: "Point",  
    coordinates: [-73.9876, 40.7574]  
  },  
  profession: ["Developer", "Engineer"],  
  apps: [  
    { name: "MyApp",  
      version: 1.0.4 },  
    { name: "DocFinder",  
      version: 2.5.7 }  
  ],  
  cars: [  
    { make: "Bentley",  
      year: 1973 },  
    { make: "Rolls Royce",  
      year: 1965 }  
  ]  
}
```

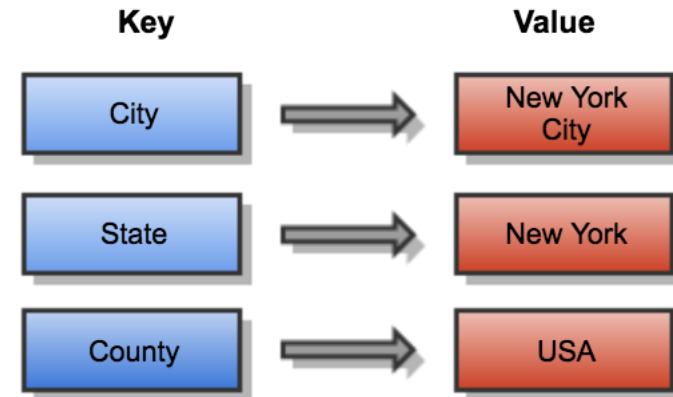
NoSQL Database

- **Document databases**
 - can store data in document format, which is similar to JSON (JavaScript Object Notation) objects.
 - Each document has pairs of fields and values.
 - The document databases can take a variety of value types such as strings, numbers, booleans, arrays, or objects.
 - The structure aligns well with objects that the developers use in their code.
 - Due to a variety of field-value types, the document databases are termed as general-purpose databases and they can be used for a wide range of use cases.
 - **MongoDB** is the most popular NoSQL document database.

NoSQL Database

- **Key-value databases**

- are simpler databases.
- Each item in this database contains keys and values.
- By referencing the key, the value can be easily retrieved.
- Key-value databases are useful when you need to store a large amount of data without using complex queries to retrieve it.
- Storing user preferences or caching are common use cases of key-value databases.
- **Redis** and **DynamoDB** are examples of key-value databases.



NoSQL Database

- **Wide-column stores**

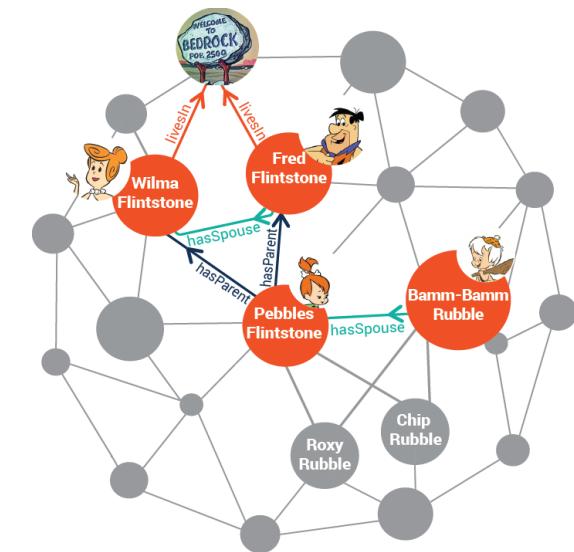
- use tables, rows, and dynamic columns to store the data.
- It is more flexible than the relational databases because each row can have different columns.
- In the wide column store, the data is stored in columns instead of rows as in a conventional relational database management system (RDBMS).
- The names and format of the columns can vary from row to row in the same table
- Wide-column stores can also be considered to be two-dimensional key-value databases.
- We use wide-column stores when we need to store a large amount of data and we can predict the query pattern.
- This database can be used to store the Internet of Things (IoT) and user profile data.
- **Cassandra** and **Hbase** are examples of wide-column stores.

The diagram illustrates a wide-column store database structure. At the top level is a **super column family** named "company". Inside "company" are two **row keys**: "1" and "2". Each row key contains three **column families**: "name", "address", and "website". The "name" column family has three columns: "city", "state", and "street num". The "address" column family has three columns: "city", "state", and "street". The "website" column family has three columns: "protocol", "domain", and "subdomain". The data values are as follows:

row key	name			address			website		
	city	state	street num	city	state	street	protocol	domain	subdomain
1	DataX	San Francisco	135	Arlington	Virginia	Kearny St	https	datax.com	www
2	Process-One	California	3500	Wilson St			https	process1.com	www

NoSQL Database

- **Graph databases**
 - use nodes and edges to store data.
 - We store the information about people, places, and things in the nodes whereas we store the information about the relationships between the nodes in the edges.
 - We use the graph database when we need to traverse the relationship to find patterns such as social networks, fraud detection, and recommendation engines.
 - Neo4j and JanusGraph are the most popular graph databases.



Scheduling Tools

- For any real implementation, we always have limited resources.
- On a busy server or cluster, often an application has to wait for some of its requests to be completed.
- The scheduling tools such as YARN, allocate resources to applications according to the pre-defined policies.
- Scheduling is a complex task and there is no one “best” policy.
- YARN provides a choice of schedulers and configurable policies.

Scheduling Tools

- Three schedulers, FIFO, Capacity, and Fair schedulers, are available in YARN.
 - **FIFO scheduler** first places all the application in a queue and run them in the order of their submission (first in first out)
 - **Capacity schedulers** allow the small job to start as soon as it is submitted. It means that the larger job waits longer when compared to FIFO scheduler
 - **Fair scheduler** dynamically balances resources between all the running jobs and hence there is no need to reserve a capacity

Benchmarking Tools

- The benchmarking tools are used to optimize the big data installation.
- Before any big data installation, the performance of each tool in the installation is measured.
- The performance metrics are compared with other tools which are known as big data benchmarking.
- Using an optimized infrastructure can make a big cost difference.
- If you can optimize to reduce 10% clusters of servers, you can save the cost of 10 servers. In the majority of organizations, the benchmarking tools are not in the job scope of the data scientists.
- These activities are carried out by IT infrastructure teams.
- Yahoo Streaming benchmark, BigBench, TPC-DS are recent approaches to big data benchmarking.

System Deployment

- Setting up the big data infrastructure is a big task.
- The system deployment tools help the deployment of new applications into the big data cluster.
- The system deployment tools automate the installation and configuration of big data components.
- In many organizations, system deployment is not the core task of data scientists.

Thanks

Samatrix Consulting Pvt Ltd

Artificial Intelligence, Machine Learning, and Data Science

Samatrix Consulting Pvt Ltd

Introduction to Machine Learning

What is Machine Learning

- Machine learning is one of the most important technical approaches to AI and the basis of many recent advances and commercial applications of AI.
- Machine learning is a form of AI that enables a system to learn from data rather than through explicit programming.
- Machine learning uses a variety of algorithms that iteratively learn from data to improve, describe data, and predict outcomes.
- Machine learning helps construct computer systems that automatically improve through experience.
- Machine learning algorithms train a system by showing it examples of desired input-output behaviour than to program it manually by anticipating the desired response for all possible inputs.

Machine Learning vs Expert Systems

- Modern machine learning is a statistical process that starts with a body of data and tries to derive a rule or procedure that explains the data or can predict future data.
- This approach—learning from data—contrasts with the older “expert system” approach to AI, in which programmers sit down with human domain experts to learn the rules and criteria used to make decisions and translate those rules into software code.
- An expert system aims to emulate the principles used by human experts, whereas machine learning relies on statistical methods to find a decision procedure that works well in practice.

Machine Learning Usages

- Machine learning has progressed dramatically over the past two decades, from laboratory curiosity to a practical technology in widespread commercial use.
- Within artificial intelligence (AI), machine learning has emerged as the method of choice for developing practical software for computer vision, speech recognition, natural language processing, robot control, and other applications.

Machine Learning Process

- The practitioner of machine learning divides the historical data set into a training set and a test set.
- Then the practitioner chooses a machine learning model or mathematical structure.
- The model consists of a range of possible decision-making rules with adjustable parameters.
- The practitioner also defines an **objective function** or **loss function** that is used to evaluate the quality of the solution that results from the choice of parameters.
- The objective function will reward the model for closely matching the training set and use of simpler rules.

Training Machine Learning Models

- Training the model is the process of adjusting parameters to maximize the objective function.
- Once a model has been trained, the practitioner can use the test set to evaluate the accuracy and effectiveness of the model.
- The goal of machine learning is to create a trained model that will generalize—it will be accurate not only on examples in the training set, but also on future cases that it has never seen before.
- While many of these models can achieve better-than-human performance on narrow tasks such as image labelling, even the best models can fail in unpredictable ways.

Training Machine Learning Models

- Another challenge in using machine learning is that it is typically not possible to extract or generate a straightforward explanation for why a particular trained model is effective.
- Because trained models have a very large number of adjustable parameters—often hundreds of millions or more—training may yield a model that "works," in the sense of matching the data, but is not necessarily the simplest model that works.

Learning From Data

- Machine learning helps model to train on data sets before the deployment.
- Some machine learning models are online whereas some models are offline.
- The online models adapt continuously as new data is analysed.
- Offline machine learning models do not change once they are deployed.
- The iterative process of online models helps model improvise the types of associations between data elements.
- Human may overlook the patterns and associations due to complexity and size of the models.
- Once the model is trained, it can be used in real time to learn from data

History of Machine Learning

History of Machine Learning

- AI and machine learning algorithms are old fields.
- The field of AI dates back to 1950s. Arthur Lee Samuels, an IBM researcher, developed a self-learning program for playing checking, which was one of the earliest machine learning programs.
- He published the paper in the IBM Journal of Research and Development in 1959.
- In last few years due to focus on distributed computing models and cheaper compute and storage, there has been a surge in the fields of AI and machine learning.
- A huge amount of money has been invested in start-up software companies, which has led to major advancements and commercial solutions

Key Market Enablers

- The six key market enablers are
 - **Processors:** Modern processors are more powerful and denser. The density to performance ratio has improved significantly
 - **Storage:** The cost of managing and storing large dataset has reduced significantly. New storage innovations have enabled faster performance. The ability to analyse vastly larger data sets have also improved
 - **Distributed compute processing:** The ability to analyse the complex data in record time has also improved due to the ability of distributed compute processing across cluster of computers

Key Market Enablers

- **Commercial data:** More commercial data such as weather data, social media data, and medical sets data is available as cloud services and well-defined Application Programming Interfaces (APIs).
- **Open-source communities:** Machine learning algorithms have been made available through open-source communities with large user bases. Therefore, there are more resources, frameworks, and libraries that have made development easier.
- **Visualization:** Visualization has gotten more consumable. You don't need to be a data scientist to interpret results, making use of machine learning broader within many industries.

Big Data

Big Data

- Any kind of data that has at least one of the following four shared characteristics. They are also known as 4 Vs
 - Extremely large **Volumes** of data
 - The ability to move that data at a high **Velocity** of speed
 - An ever-expanding **Variety** of data sources
 - **Veracity** so that data sources truly represent truth
- Big data incorporates all data, including structured, unstructured, and semi-structured data from email, social media, text streams, images, and machine sensors

Big Data

- To gain right insights using analytics on big data, we need appropriate technology that can gather, store, manage, and manipulate vast amounts data at the right speed and at the right time.
- The evolution of computing technology along with hybrid cloud architectures have enabled the management of immense volumes of data that could have only been handled by supercomputers at great expense.

Big Data For Machine Learning

- The big data helps improve the accuracy of machine learning models substantially.
- The low volume of data may lead to misinterpreting a trend or missing an emerging pattern.
- Big data can be very useful for training machine learning models.
- An organization does not have to have big data to use machine learning models.

Big Data For Machine Learning

- But the availability of big data can help improve the accuracy of the models.
- Using big data, the data can be virtualized so that it can be stored in the most efficient and cost-effective manner.
- With the help of big data, the data can be stored on premise or in the cloud.
- The improvements in network speeds and reliability have helped manage massive amount of data at the acceptable speed.

Big Data For Machine Learning

- The big data technologies include data virtualization, parallel processing, distributed file systems, in-memory databases, containerization, and micro-services.
- This combination of technology advances can help organizations address significant business problems.
- Businesses always had large amount of data for decades.
- But the ability to use the richness of data source to gain actionable insights from data was not available.
- Using machine learning models and big data technologies, the organizations can use the data to gain useful insights, anticipate the future, and be prepared for disruption.

Leveraging Machine Learning

Advanced Analytics

- The role of analytics has been changing in organization's operational processes has been changing for past 30 years.
- The companies have progressed in analytics maturity levels ranging from descriptive analytics to predictive analytics to machine learning and cognitive computing
- Companies have been using analytics to understand both the current status of their business and how they can learn from the past to anticipate the future.
- They can analyse how various actions and events can impact the outcomes. The knowledge from this analysis can help predict future.

Advanced Analytics

- Data scientists and business analysts can make predictions using analytical models that are based on historical data.
- In business environment, unknown factors can impact future outcomes significantly.
- The companies focus on building predictive models that can react and change with the changes in the business environment.

Types of Advanced Analytics

- There are two types of advanced analytics
 - Descriptive analytics
 - Predictive analytics

Descriptive Analytics

- Descriptive analytics helps analysts understand the current reality in the business.
- You need to understand the context for historical data in order to understand the current reality of where the business is today.
- Using this approach, the organizations can answer questions such as which product styles are selling better this quarter as compared to last quarter, and which regions are exhibiting the highest/lowest growth.

Predictive Analytics

- Predictive analytics helps anticipate changes using the patterns and anomalies within the data.
- Using predictive analytics models, the analyst integrates various related data sources to predict outcome.
- Predictive analytics uses machine learning algorithms to gain insights.
- The predictive analytics tools require new data that can reflect the business changes.
- The addition of new data helps improve the business's ability to anticipate subtle changes in customer preferences, price erosion, market changes, and other factors that will impact the future of business outcomes.

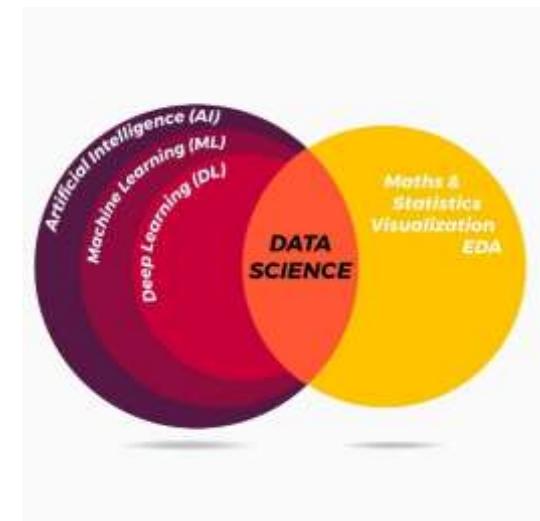
Predictive Analytics

- With a predictive model, you predict future. For example, you can answer the following types of questions
 - How improved web experience can entice a customer to buy frequently?
 - How a stock or a portfolio will perform based on factors such as international news and internal financial factors?
 - Which combination of drugs will provide the best outcome for this cancer patient based on the specific characteristics of the tumor and genetic sequencing?

Machine Learning and Statistics

Machine Learning and Statistics

- Statistics is the discipline that concerns the collection, organization, displaying, analysis, interpretation and presentation of data. Two main statistical methods used in data analysis are:
 - **Descriptive statistics** summarizes data from a sample using indexes such as the mean or standard deviation
 - **Inferential statistics** draws conclusions from data that are subject to random variation (e.g., observational errors, sampling variation)



Machine Learning and Statistics

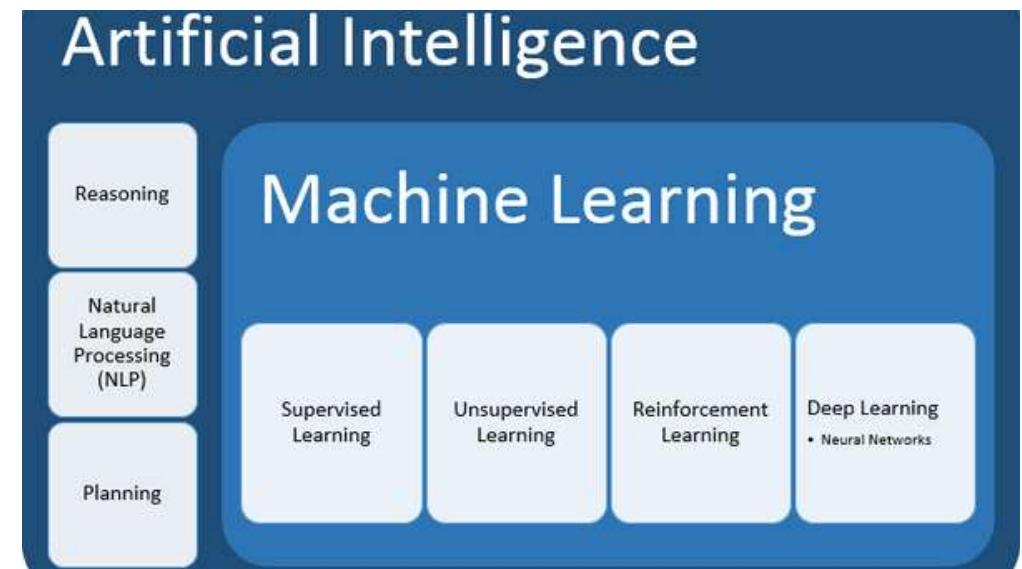


- Machine learning models leverage statistical algorithms to predict analytics.
- The discipline of statistics, data mining, and machine learning have a role in understanding data and characteristics of the data sets.
- They are used in finding relationships and patterns in the data.
- Hence there is a significant overlap.
- The tools and techniques of statistics and machine learning are used to solve business problems
- Many machine learning algorithms are rooted in classical statistical analysis. Data scientists combine the expertise in statistics and machine learning to use all disciplines in collaboration.

AI and Machine Learning

AI and Machine Learning

- AI can be used to describe systems that can “think.”
- For example, thermostats that learn your preference or applications that can identify people and what they are doing in photographs can be thought of as AI systems
- There are four main subsets of AI. Reasoning, natural language processing (NLP), planning, and machine learning



Reasoning System

- **Reasoning system** is a software system that generates conclusions from available knowledge and data using logical techniques.
- If the data is incomplete, the reasoning helps fill in the blanks using connected data.
- For example, given that the system has enough data and the following question is asked, “What is a safe internal temperature for eating a drumstick?” The system can tell the answer is 165 degrees.

Reasoning System

- The logic chain would be as follows:
 - A drumstick refers to a chicken leg as opposed to a part of a musical instrument
 - A chicken leg contains dark chicken meat
 - The dark chicken meat requires temperature of 165 degrees to cook
 - So, the answer is 165 degree.
- In this example the system was not explicitly trained to predict the safe internal temperature of chicken drumstick.
- The system used the existing data and knowledge to fill in data gaps.

Natural Language Processing

- **Natural Language Processing (NLP)** is the ability of a computer program to understand human language both written text and human speech.
- The idea of giving computers the ability to process human language is as old as the idea of the computer themselves.
- The goal of the NLP is to get computers perform tasks involving human language.

Natural Language Processing

- It includes tasks like enabling human-machine communication, improving human-human communication, or simply doing useful processing of text or speech.
- There are two main reasons why we want our computers to process the natural languages: first, to communicate with humans, and second, to acquire information from written language

Natural Language Processing

- There are over a trillion of pages on web.
- Almost all of them in human language.
- Computer program that wants to do knowledge acquisition, needs to understand ambiguous, messy language that human use.
- We use information seeking tasks such as text classification, information retrieval, and information extraction.
- We use language models to address the tasks.
- Language models predict the probability distribution of language expressions.

Planning

- **Planning** is about how an agent achieves its goals.
- To achieve anything but the simplest goals, an agent must reason about its future.
- Because an agent does not usually achieve its goals in one step, what it should do at any time depends on what it will do in the future.
- What it will do in the future depends on the state it is in, which, in turn, depends on what it has done in the past.

Planning

- Automated planning is the ability of the intelligent system to act autonomously and flexibly to construct the sequence of action to reach the final goal.
- Rather than a pre-programmed decision-making process that goes from A to B to C to reach a final output, automated planning is complex and requires a system to adapt based on the context surrounding the given challenge.

Types of Machine Learning

Types of Machine Learning

- Learning is the ability of an agent to improve its behavior based on experience. This could mean the following:
 - The range of behaviors is expanded; the agent can do more.
 - The accuracy on tasks is improved; the agent can do things better.
 - The speed is improved; the agent can do things faster.

Types of Machine Learning

- There are four main types of learning
 - Supervised learning
 - Unsupervised learning
 - Self-supervised learning
 - Reinforcement learning

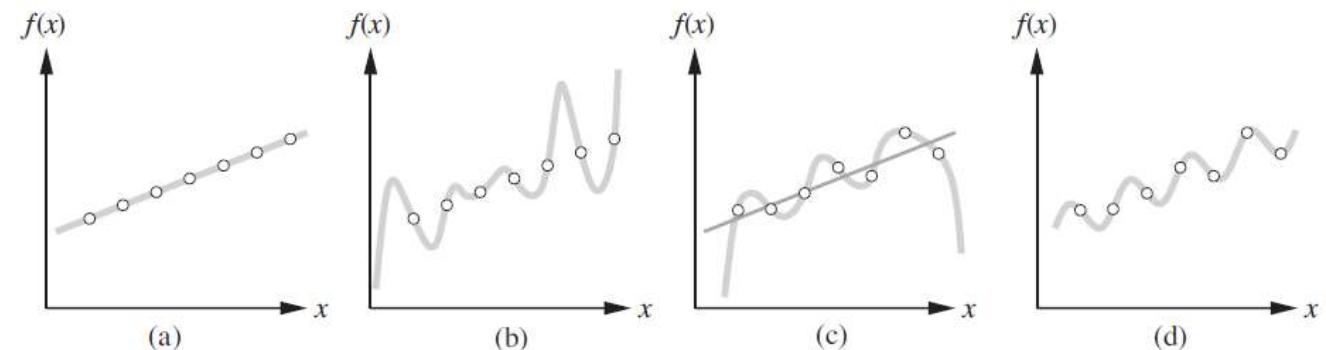
Supervised Learning

- In **predictive** or supervised learning, the agent observes some example input-output pairs and learn a function that map input and output.

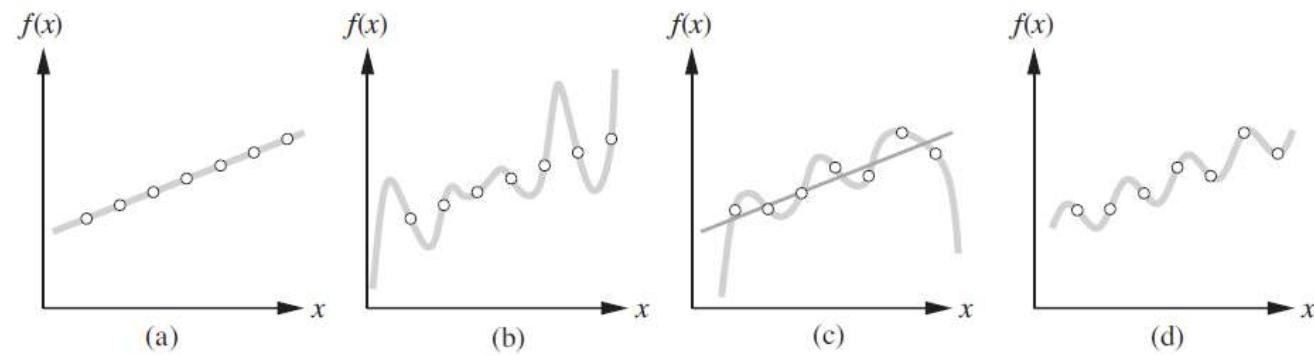
Given a training set of N example input-output pairs

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_N, y_N),$$

Where each y_j was generated by unknown function $y = f(x)$, discover a function h that approximates the true function f .



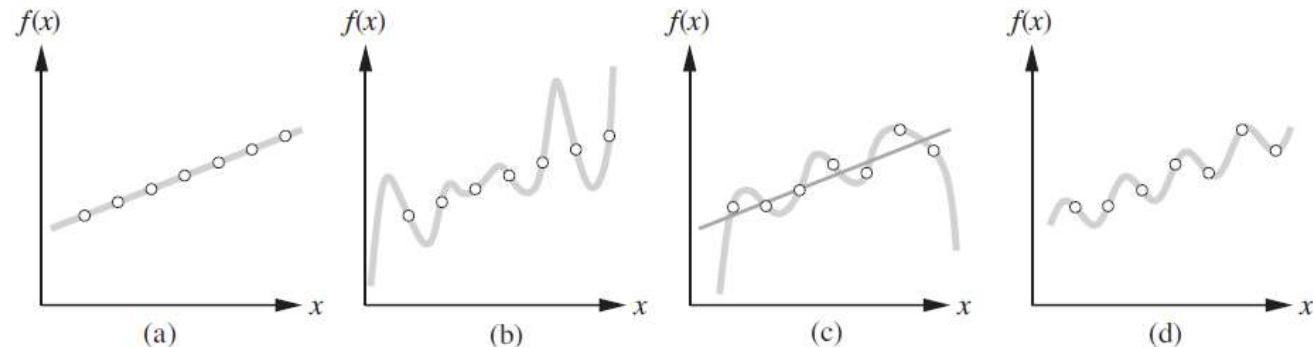
Supervised Learning



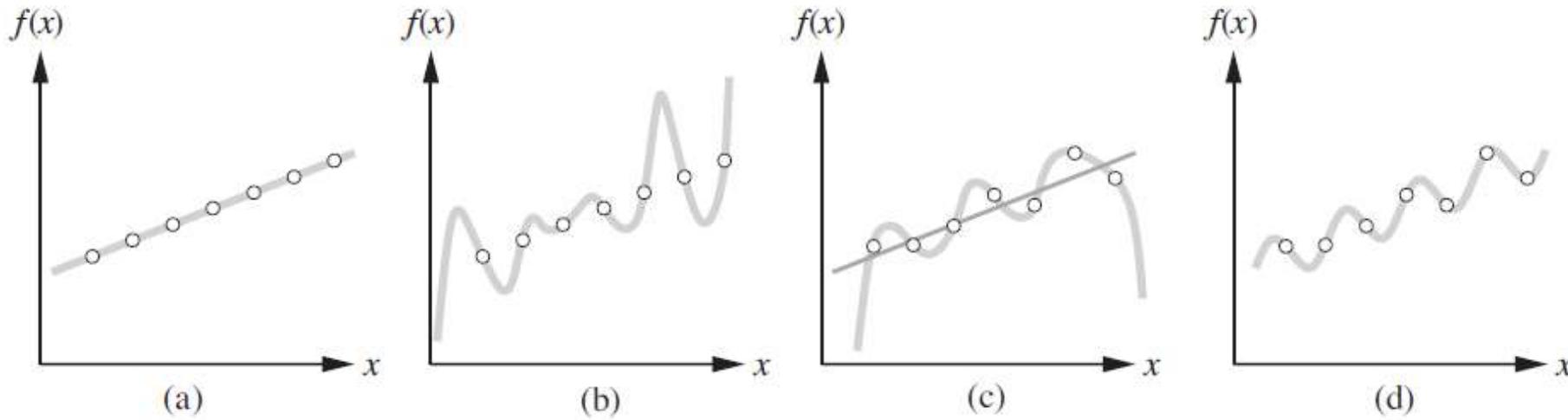
- Here x and y can be of any value not necessarily be the numbers.
- Each training input x_j can also represent the height and weight of a person.
- These are called **features**, **attributes**, or **covariates**.
- The x_j can also be a complex structured object such as an image, a sentence, an email message, a time series, a molecular shape, a graph etc.
- The function h is a **hypothesis**.
- Learning is the search through the space of possible hypothesis for one that will perform well, even on new examples that are beyond the **training set**.
- To measure the accuracy of the training set, we give the **test set** of examples that are separate from the training set.
- We say that the hypothesis **generalizes** well, if it correctly predicts the value of y for new examples.

Supervised Learning

- When the output y is one among the finite set of **categorical** or **nominal** values (such as sunny, cloudy, or rainy), the learning problem is called **classification or pattern recognition**.
- When y is a number (such as income level), the learning problem is called **regression**.

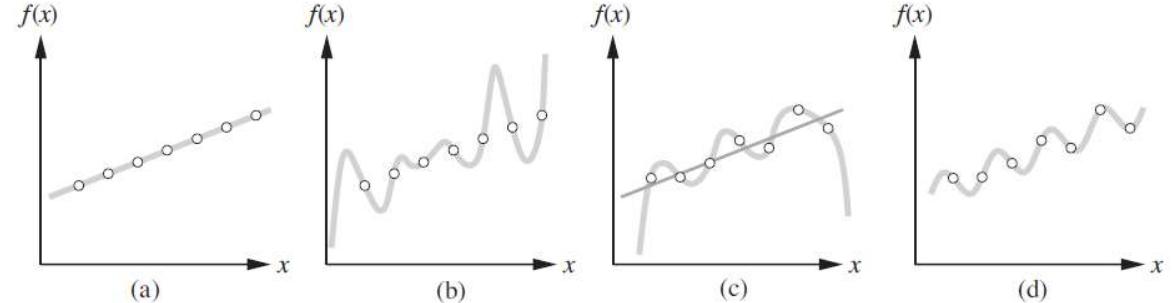


Supervised Learning



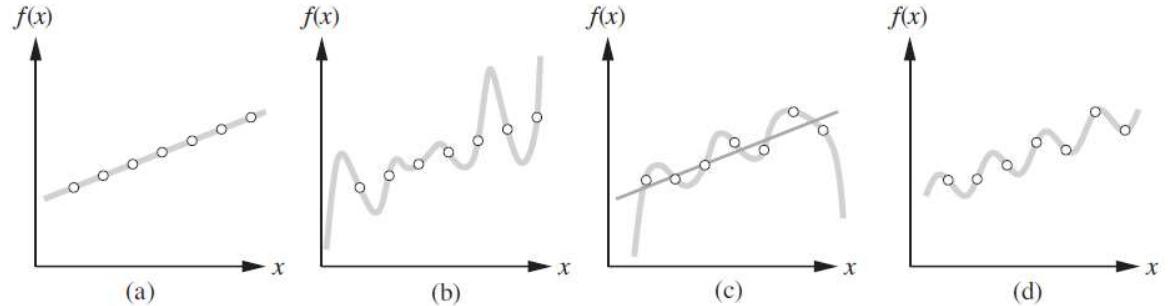
- Fig shows the example of fitting a function of a single variable to some data points. (a) Example $(x, f(x))$ pairs and a consistent, linear hypothesis. (b) A consistent, degree-7 polynomial hypothesis for the same data set. (c) A different data set, which admits an exact degree-6 polynomial fit or an approximate linear fit. (d) A simple, exact sinusoidal fit to the same data set.

Supervised Learning



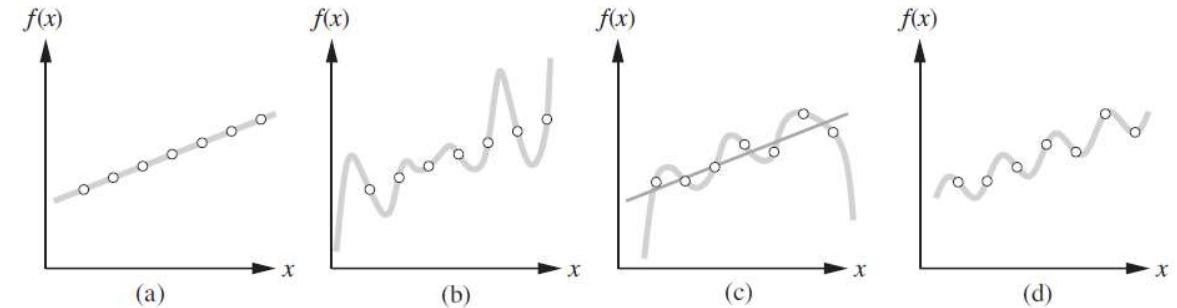
- The fig shows the example of fitting a function of a single variable to some data points.
- The example of the points in the (x, y) plane where $y = f(x)$.
- We do not know what f is, but we will approximate it with a function h selected from **hypothesis space**.
- Fig (a) shows some data with an exact fit by a straight line (the polynomial $0.4x + 3$).
- The line is called **consistent** hypothesis because it agrees with all the data.

Supervised Learning



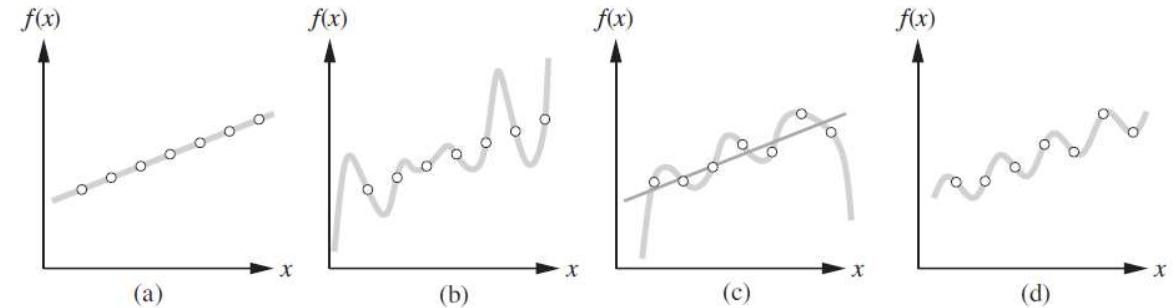
- Figure (b) shows a high degree polynomial that is also consistent with the same data.
- This illustrates the fundamental problem: *how do we choose from among multiple consistent hypothesis?*
- One answer is to prefer the simplest hypothesis consistent with the data.
- This principle is known as **Ockham's razor**, after 14th century English philosopher William of Ockham, who used it to argue against all sorts of complications.
- Since degree-1 polynomial is simpler than degree-7 polynomial, so (a) should be preferred over (b).

Supervised Learning



- Figure (c) shows a second data set.
- There is no consistent straight line for this data set.
- It requires degree-6 polynomial for an exact fit.
- There are just 7 data points, so a polynomial with 7 data points does not find any pattern in the data, so we do not expect it to generalize well.
- A straight line that is not consistent with any of the data points, may generalize well for the unseen values of x .

Supervised Learning



- *In general, there is a trade-off between complex hypothesis that fit the training data well and simpler hypothesis that may generalize better.*
- In figure (d), we expand the hypothesis space to allow polynomials over both x and $\sin(x)$, and find that the data in (c) can be fitted better by a simpler function of the form $ax + b + c\sin(x)$. This shows the importance of hypothesis space.
- The learning problem is **realizable** if the hypothesis space contains the true function. Unfortunately, we cannot always tell whether the given learning problem is realizable, because the true function is not known

Unsupervised Learning

- The second main type of machine learning is **descriptive** or unsupervised learning approach.
- Here we are only given a training set of N features or inputs x_1, x_2, \dots, x_N .
- We are not interested in prediction, because we do not have associated response variable y .
- Rather the goal is to identify the interesting things about the measurements on x_1, x_2, \dots, x_N .
- Is there an informative way to visualize the data?
- Can we discover subgroups among the variables or among the observations?

Unsupervised Learning

- Unsupervised learning refers to a diverse set of techniques to answer such questions.
- The goal of the unsupervised learning is to find “interesting patterns” in the data.
- This is also known as **knowledge discovery**.
- The purpose of unsupervised learning is data visualization, data compression, or data denoising, or to better understand the correlation present in the data at hand.
- Two most common unsupervised learning types are **principal component analysis**, a tool used for data visualization or data pre-processing before supervised techniques are applied and **clustering**, that consists of dividing the dataset into clusters of similar examples.
- Unsupervised learning is the bread and butter of data analytics, and it's often a necessary step in better understanding a dataset before attempting to solve the supervised learning problem.

Self - supervised Learning

- This is a specific instance of supervised learning, but it is different enough that it deserves its own category.
- Self-supervised learning is supervised learning without human-annotated label or response variable y .
- You can think of it as supervised learning without any human in the loop.
- There are still labels or response variables involved (because the learning must be supervised by something), but they are generated from the input data, typically using a heuristic algorithm. Input data can be labelled by finding and exploiting the relations (or correlations) between different input signals.

Self – supervised Learning

- For instance, **autoencoders** are a well-known instance of self-supervised learning, that learns to copy its input to its output.
- The purpose of the autoencoder is to reconstruct its inputs by minimizing the difference between the input and the output instead of predicting the target value Y given inputs X.
- Therefore, autoencoders do not require labelled inputs to enable learning.
- In the same way, trying to predict the next frame in a video, given past frames, or the next word in a text, given previous words, are instance of self-supervised learning (**temporally supervised learning**, in this case: supervision comes from future input data).

Reinforcement Learning

- In reinforcement learning, an agent receives information about its environment and learns to choose actions that will maximize some **rewards or reinforcement**.
- In a reinforcement learning problem, a robot can act in a world, receiving rewards and punishments and determining from these what it should do.
- Reinforcement learning differs from other types of supervised learning because the system isn't trained with the sample data set.
- Rather, the system learns through trial and error. Therefore, a sequence of successful decisions will result in the process being “reinforced” because it best solves the problem at hand.

Reinforcement Learning

- Consider, for example, the problem of learning to play chess.
- A supervised learning agent needs to be told the correct move for each position it encounters, but such feedback is seldom available.
- In the absence of feedback from a teacher, an agent can learn a transition model for its own moves and can perhaps learn to predict the opponent's moves, but without some feedback about what is good and what is bad, the agent will have no grounds for deciding which move to make.
- The agent needs to know that something good has happened when it (accidentally) checkmates the opponent, and that something bad has happened when it is checkmated—or vice versa, if the game is suicide chess.

Reinforcement Learning

- This kind of feedback is called a reward, or reinforcement.
- In games like chess, the reinforcement is received only at the end of the game.
- In other environments, the rewards come more frequently.
- In ping-pong, each point scored can be considered a reward; when learning to crawl, any forward motion is an achievement.

Reinforcement Learning

- One of the most common applications of reinforcement learning is in robotics or game playing.
- Take the example of the need to train a robot to navigate a set of stairs.
- The robot changes its approach to navigating the terrain based on the outcome of its actions.
- When the robot falls, the data is recalibrated, so the steps are navigated differently until the robot is trained by trial and error to understand how to climb stairs.
- In other words, the robot learns based on a successful sequence of actions.

Reinforcement Learning

- Reinforcement learning is also the algorithm that is being used for self-driving cars.
- In many ways, training a self-driving car is incredibly complex because there are so many potential obstacles.
- If all the cars on the road were autonomous, trial and error would be easier to overcome.
- However, in the real world, human drivers can often be unpredictable.
- Even with this complex scenario, the algorithm can be optimized over time to find ways to adapt to the state where actions are rewarded.
- One of the easiest ways to think about reinforcement learning is the way an animal is trained to take actions based on rewards.
- If the dog gets a treat every time he sits on command, he will take this action each time.

Type of Machine Learning Algorithms

Type of Machine Learning Algorithms

- Selecting the right machine learning algorithms is part art and part science.
- Two data scientists can use the two different machine learning algorithms solving the same business problem using the same data sets.
- Hence, understanding different machine learning algorithms help the data scientists select the best types of algorithms to solve a given business problem

Regression vs Classification

- Variables can be characterised as either *quantitative* or *qualitative* (also known as categorical).
- The quantitative variables take on numerical values. Example includes a person's age, height, or income.
- The qualitative variables take on values in one of K different classes or categories.
- Example of qualitative variables include a person's gender (male or female), the brand of the product purchased (brand X, Y, or Z), whether the person defaults on a debt (yes or no), or a cancer diagnosis (Acute Myelogenous Leukemia, Acute Lymphoblastic Leukemia, or No Leukemia).

Regression vs Classification

- The problems with quantitative response are referred as **regression** problems.
- Those involved in qualitative response are referred as **classification** problems.
- However, the distinction is always not very crisp.
- Least square linear regression is used with a quantitative response.
- Logistic regression is typically used for qualitative (two-class or binary) response.

Regression vs Classification

- We tend to select the learning model on the basis whether the response is quantitative or qualitative.
- However, whether the predictors are quantitative or qualitative is considered less important.

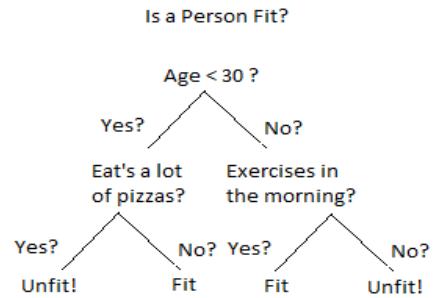
Bayesian

- Bayesian algorithms help the data analysts encode their prior beliefs about what the model look like, independent of the data set.
- These algorithms are especially useful when you do not have massive amount of data to train the model confidently.
- The Bayesian algorithm would be helpful if you have prior knowledge to some part of the model and you can code that directly.
- For example, if you want to model a medical imaging diagnosis system that looks for lung disorder.
- If a published journal study estimates the probability of different lung disorders based on lifestyle, those probabilities can be encoded into the model.

Clustering

- In clustering objects with similar parameters are grouped together in a cluster.
- All objects in a cluster are more like each other than objects in other cluster.
- The clustering is a type of unsupervised learning because the data is not labelled.
- The clustering algorithm interprets the parameter that make up each item and then groups them accordingly.

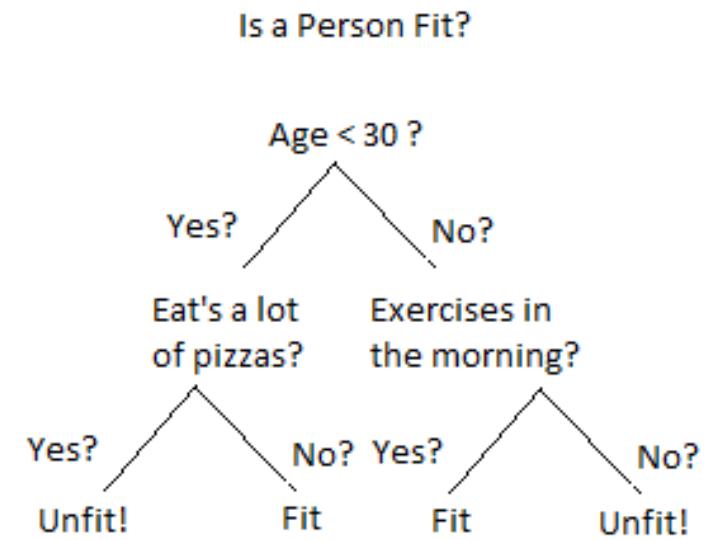
Decision Tree



- Decision tree algorithms uses a tree-like graph or branching structure to illustrate the event outcomes, resource costs, and utility.
- The decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether the object is cat or a dog), each branch represents the outcome of the test.
- Each leaf node represents a possible outcome.
- The paths from root to leaf represent classification rule.
- Percentages are assigned to nodes based on the likelihood of the outcome occurring.

Decision Tree

- Decision tree algorithms are one of the most widely used supervised learning methods.
- Tree based algorithms empower predictive models with high accuracy, stability, and ease of interpretation.
- They can easily solve both classification and regression problems.
- Decision Tree algorithms are referred to as CART (Classification and Regression Trees).



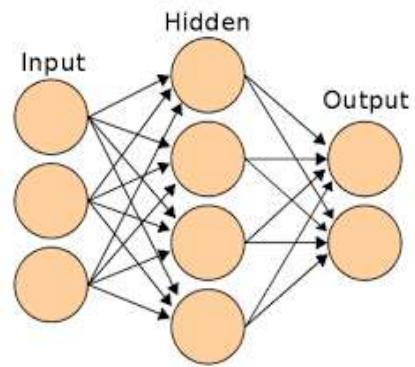
Dimensionality Reduction

- Dimensionality reduction helps systems remove data that's not useful for analysis.
- This group of algorithms is used to remove redundant data, outliers, and other non-useful data.
- Dimensionality reduction can be helpful when analyzing data from sensors and other Internet of Things (IoT) use cases.

Dimensionality Reduction

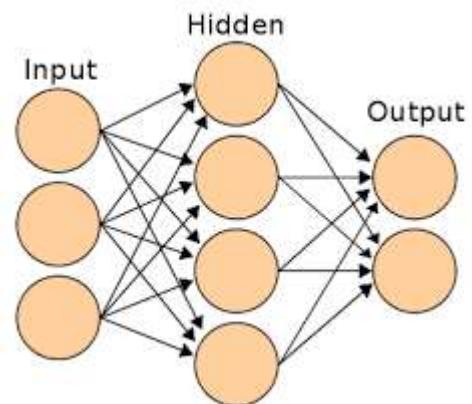
- In IoT systems, there might be thousands of data points simply telling you that a sensor is turned on.
- Storing and analyzing that “on” data is not helpful and will occupy important storage space.
- In addition, by removing this redundant data, the performance of a machine learning system will improve.
- Finally, dimensionality reduction will also help analysts visualize the data.

Neural Network and Deep Learning



- A neural network attempts to mimic the way a human brain approaches problem and uses layers of interconnected units to learn and infer relationships based on observed data.
- A neural network can have several connected layers.
- When there is more than one hidden layer in a neural network, it is sometimes called deep learning.
- Neural network models can adjust and learn as data changes.
- Neural networks are often used when data is unlabeled or unstructured.
- One of the key use cases for neural networks is computer vision.

Neural Network and Deep Learning



- Deep learning is being leveraged today in a variety of applications.
- Self-driving cars use deep learning to help the vehicle understand the environment around the car.
- As the cameras capture images of the surrounding environment, deep learning algorithms interpret the unstructured data to help the system make near real-time decisions.
- Likewise, deep learning is embedded in applications that radiologists use to help interpret medical images

Training Machine Learning Systems

Training Machine Learning Systems

- The process of developing and refining the model is an iterative process.
- The steps include selecting the correct algorithm, training, and testing a system.
- Training is a critical step in machine learning process.

Training Machine Learning Systems

- To train your machine learning systems, you need to know the inputs (for example: Income level, size of the house, location of the house and so on), you should know your desired goal (the price of the house in the area).
- However, the mathematical function to transform the row data to the price of the house is unknown.
- As the learning algorithm is exposed to more and more data, the model will become more accurate in predicting the price of the house.

Steps

- There are three steps in training the machine learning algorithm:
 - Representation
 - Evaluation
 - Optimization

Representation

- The machine learning algorithm creates a model to transform the input data into the desired results.
- As the model is exposed to more data, it will learn the relationship between the raw data and point out data points that are strong predictors for the desired output

Evaluation

- The algorithm creates multiple models.
- Once the models are ready, we need to evaluate and score the models based and find out which model produces the most accurate predictions.
- After the model is operationalize, it is will be exposed to unknown data.
- As a result, we need to ensure that the model is generalized and not overfit to training data.

Optimization

- After the algorithm creates and scores various models, we need to select the best performing model and algorithm.
- As the algorithm is exposed to more diverse sets of input data, we need to select the most generalized model.

Training Process

- For the training process, it is very important to have enough data so that you can also test the model.
- The first pass of the training process provides mixed results.
- That means either you need to refine the model or provide more data.

Thanks

Samatrix Consulting Pvt Ltd

Artificial Intelligence, Machine Learning, and Data Science

Samatrix Consulting Pvt Ltd

Introduction to Artificial Intelligence

Quote

“There are three kinds of intelligence: one kind understands things for itself, the other appreciates what others can understand, the third understands neither for itself nor through others. This first kind is excellent, the second good, and the third kind useless.”

-Niccolo Machiavelli

(1469 – 1527), Italian diplomat, politician, historian, philosopher, humanist, writer, playwright and poet of the Renaissance period

Intelligence

- **Intelligence** is studied in human beings, non-human animals, and plants.
- By Collins English dictionary, intelligence is the ability to think, reason, and understand instead of doing things automatically or by instinct.
- For thousands of years, we have been trying to understand how we think and what makes some brains smarter than others.

Artificial Intelligence

- The field of **artificial intelligence** goes one step further.
- It is about how to build an intelligent computer that can-do things, which humans can do better.
- Artificial intelligence emphasizes perception, reasoning, and action.
- The field of the study indicates the similarity with Philosophy.
- Philosophy uses logic and reasoning to analyze the nature of human thoughts, the essence of the world, and how an individual connects with the world.
- Philosophy is the study of the fields of knowledge that humanity has understood poorly.
- As the implementations of real-time applications of AI increases, the domain of Philosophy will be an empty set.

Artificial Intelligence

- The objective of AI is not to develop a chess-playing computer.
- Instead, the vision is to develop a robot, which sits in front of us as an opponent, perceives the chessboard, and makes the right chess moves in the physical world.
- Such implementations will push the definition of AI to new dimensions.
- AI is relatively a new field.
- The work on AI started after World War II.
- The name artificial intelligence was coined in summer 1956 when John McCarthy called the famous Dartmouth Conference.
- This conference started AI as a field.

What is AI

Define AI

- There is no single universally accepted definition of AI.
- Some practitioners define AI as a computerized system that exhibits behavior that is commonly thought of as requiring intelligence.
- Whereas other practitioners define AI as a system that is capable of rationally solving complex problems or taking appropriate actions to achieve its goals in whatever real-world circumstances it encounters?
- According to Patrick Henry Winston, “Artificial Intelligence is the study of the computations that make it perceive, reason, and act”
- Whereas Stuart Russell and Peter Norvig used different taxonomy.

Purpose of AI

- The purpose of AI is to build a system that can
 - Think like humans (e.g., cognitive design and neural networks)
 - Act like humans (pass Turing Test)
 - Think rationally (logic solvers, reasoning, and optimization)
 - Act rationally (intelligent software agents)
- Traditionally all the four approaches have been studied and followed for a very long time.
- The human-centered approach involves understanding the human behavior whereas rational approach involves mathematics and engineering.

Turing Test

- Alan Turing developed the Turing Test in 1950.
- It tests the ability of a machine to exhibit intelligent behavior that is equivalent to or better than that of a human.
- During the test, a human interrogator poses questions to a human and a computer. All the participants separate from each other.
- The whole conversation takes place in written format instead of an oral one utilizing a keyboard and a monitor.
- The computer passes the test if the interrogator cannot tell whether the response is coming from the human or the machine.

Turing Test

- To pass the test, the computer needs the following capabilities:
 - **Natural language processing** enables the computer to understand and process the language and speech;
 - **Knowledge representation** to encode knowledge, actions, beliefs and other mental states in an artificial system;
 - **Automated reasoning** helps computers apply logical reasoning to solve the problem, answer the questions, and draw new conclusions;
 - **Machine learning** helps computers learn automatically, adapt to new circumstances and improve from experience without external intervention;

Turing Test

- One of the variations of the Turing Test is the **Total Turing Test**.
 - Using this test the interrogator can test the perceptual abilities of the subject using **computer vision**, and the ability to manipulate the objects and move about using **robotics**.
- The Turing Test is relevant 70 years later also.
- However, the focus of the researchers has been more on studying the underlying principles of intelligence than duplicating a test machine.

Cognitive Modelling Approach

- Cognitive science is the scientific study of the human mind and intelligence. If we want to build a system that can think like a human, we need to understand how human thinks. The cognitive models can be:
- **Predictive** which describes how people react in different scenarios
- Prescriptive which describes the limitations in cognition and the ways in which the limitations can be overcome

Cognitive Modelling Approach

- According to Craik [1943]:

“If the organism carries a ‘small-scale model’ of external reality and of its own possible actions within it head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilize the knowledge of past events in dealing with the present and future, and in every way to react in a much fuller, safer, and more competent manner to the emergencies which face it.” (p. 61)
- Before building cognitive models, we need to understand how the human brain works.
- Applying the understanding of the brain, we can decode it into a computer program.
- If the input-output behavior of the computer program matches the behavior of the human, we can say that the computer is able to mimic the human in certain instances.

The Law of Thoughts

- The logical thinking has been since the creation of the human species. Greek Philosopher, Aristotle, was the first to analyze and codify the process called **logic**. He identified a type of reasoning called a syllogism. A remarkable example of the same is as follows.
 - All humans are mortal (stated)
 - All Greeks are humans (stated)
 - All Greeks are mortal (result)

The Law of Thoughts

- In his famous syllogisms, Aristotle provided patterns for argument structures.
- These patterns always give correct conclusions when they are provided with a specific premise.
- In these syllogisms, the "law of thoughts" plays one of the most significant roles.
- These laws of thought are so simple and basic that one can argue that they should not be said.
- However, they are fundamental to clear and sound thinking.

The Law of Thoughts

- The laws are known as the **Law of Identity**, **Law of Contradiction**, and **Law of Extended Middle**.
 - **Law of Identity** asserts about the absolute truth. It states that if any statement is true then it is true. This law asserts that everything is itself. It cannot be anything else. A river cannot be a cloud; A mountain cannot be an ocean.
 - **Law of Contradiction** asserts that no statement can be both true and false. For example, Jim can be the father of Rick and the son of Mike. But Jim cannot be biologically father and son of the same person. He can be one or another. That is the truth of life.
 - **Law of Excluded-Middle** asserts that a statement is either true or false. There is no middle ground. If a statement is half true, then it is false. A half-true is a lie.

The Law of Thoughts

- Logicians developed all the notations for all kinds of objects in the world and relationships among them.
- By 1965, the programs could solve any problems described in logical notation.
- The logicist tradition in AI hopes to create intelligent systems using such programs.
- However, there are two obstacles to this approach.
- First, it is not possible to state informal knowledge in formal terms when knowledge is less than 100% certain.
- Secondly solving problems in principle is different from solving the problem in practice.

The Rational Agent Approach

- Here agent refers to a computer. An **agent** acts. The word agent comes from Latin *agree* to do. Even though all the computer programs give some output, but they are expected to do more than giving the output. They should
 - operate autonomously
 - identify the environment they are operating in
 - continue to work for a long period of time
 - adapt to changes
 - Identify their goals and pursue them
- A rational agent acts to achieve the best outcome.
- When there is uncertainty, acts to achieve the best expected outcome.
- Acting rationally means to achieve one's goals given one's beliefs.

The Rational Agent Approach

- In the “law of thoughts” approach the focus was on a correct inference.
- Rational agents can also make correct inference, but correct inference is not always rational.
- For example, in certain situations, there is no correct thing to do but sometimes must be done.
- On the other hand, there are situations when one should act rationally without involving inference.
- For example, a pulling backhand from a hot object is a reflex action that is done without deliberation.

The Rational Agent Approach

- The rational agent approach in artificial intelligence has two advantages.
- First, this approach is more general than the law of thoughts because the correct inference is one of the useful mechanisms for achieving rationality.
- It is not a necessary mechanism.
- Second, the rational agent-based approach is amenable to scientific development than the approach based on human behaviour or human thoughts because the standard of rationality is clearly defined and completely general.
- On another hand, human behaviour is adapted to a specific environment and defined by the sum of what humans do.

Artificial and Natural Intelligence

Artificial vs Natural Intelligence

- The term “artificial intelligence” could be confusing.
- Artificial Intelligence could be interpreted as opposite to “Real Intelligence”
- You can distinguish between real and fake. Fake is non-real.
- You can also distinguish natural vs artificial. Natural means occurring in nature whereas artificial means made by humans.
- Intelligence is different. Fake intelligence is not possible.
- If an agent behaves intelligently, it is intelligent. Acting intelligently is being intelligent.
- Artificial intelligence, whenever it will be achieved, will be real intelligence that is artificially created.
- Human is a natural intelligent agent.

Artificial vs Natural Intelligence

- Some people may argue that worm and insects are also intelligent.
- Some others will argue that dogs, monkeys, or horses are intelligent.
- However, the class of organizations is more intelligent than human beings.
- Ant colony as an organization is an intelligent agent.
- Individual ant may not be very intelligent, but ant colony can act very intelligently.
- The colony can discover food, exploit it, and adapt to changing circumstances very effectively.
- Each ant may not be very intelligent, but the ant colony can act more intelligent than any individual ant.

Artificial vs Natural Intelligence

- Similarly, the companies can develop, manufacture, and distribute products where the sum of the skills required is much more than any individual could master.
- Human society viewed as an agent is arguably the most intelligent agent known.
- The main sources of human intelligence are as follows:
- Biology: Humans are adaptable and can survive in any habitat
- Culture: Language, tools, concepts, and the wisdom that is passed from parents and teachers to children
- Life-long learning: Human learn life-long and accumulate knowledge and skills

AI and Other Disciplines

- The fields of knowledge such as philosophy, neurobiology, evolutionary biology, psychology, economics, political science, sociology, anthropology, and control engineering have been studying intelligence for a very long time.
- Comparatively, AI is a very young discipline.
- AI could also be described as “synthetic psychology,” “experimental philosophy,” or “computational epistemology” – epistemology is the study of knowledge.
- AI is about the study of nature and intelligence. But uses more powerful experimental tools than other older disciplines.

AI and Other Disciplines

- Other disciplines such as philosophy, psychology, economics, and sociology observe only the external behaviour of intelligent systems.
- AI researchers experiment with executable models of intelligent behaviour.
- The AI models can be inspected, redesigned, and experimented within a complete and rigorous way.
- Using computers, AI researchers can construct the models whereas the philosophers can only theorize.
- AI researchers can also experiment with these models as opposed to just discussing their abstract properties.
- AI theories can be empirically grounded in implementation.

Thinking vs Flying Machines

- Let's consider an analogy between the development of flying machines and development of the thinking machines.
- There are several ways to understand flying.
- One of the ways is to dissect the flying animals and hypothesize their common structural features as necessary fundamental characteristics of any flying agent.
- The examination of birds, bats, and insects would suggest that flying involves the flapping of wings.
- The researcher can test the hypothesis by strapping feathers to one's arms, flapping, and jumping in the air, as Icarus did.

Thinking vs Flying Machines

- Alternatively, the researcher can understand the principles of flying without restricting the natural occurrence of flying.
- The approach involves the construction of artifacts that do not behave like flying animals. But it flies.
- The second method has provided useful tools namely airplanes and the understanding of the principles underlying flying, namely aerodynamics.

Thinking vs Flying Machines

- AI takes an approach that is analogous to the aerodynamics.
- AI researchers focus on testing general hypotheses about the nature of intelligence by building machines that are intelligent.
- The machine need not mimic human intelligence.
- The question "Can airplanes really fly?" is analogous to "Can computers really think?"

Thinking vs Flying Machines

- AI comes under the umbrella of cognitive science.
- Cognitive science links various disciplines that study cognition and reasoning, from psychology to linguistics to anthropology to neuroscience.
- AI provides the tools to build intelligence rather than studying the external behaviour of intelligent agents or dissecting the inner workings of intelligent systems.

AI Problems

AI Problems

- What kind of problems AI can solve?
- Earlier work focused on formal tasks such as game playing and theorem solving.
- People who can play games and solve theorem are known to be intelligent.
- Instead of this, it was believed that computers could perform well in game-playing and theorem solving by exploring many solution paths and selecting the best one.
- But later, it turned out to be false because no computer was fast enough to face the combinational explosion that is part of every problem.

AI Problems

- Another early foray of AI was solving our day to day problems such as how to get to work in the morning.
- This was called common sense reasoning.
- This includes the reasoning about physical objects and their relationship with other objects and reasoning about actions and their consequences. Only simple tasks were taken.
- There was no attempt to solve the problem with a large amount of data and knowledge in one domain.
- As the techniques of handling a large amount of data and knowledge were developed, complex problems were attempted.

Perception Problems

- A lot of research was done in areas such as perception (vision and speech) and natural language processing.
- New solutions were developed to address the problems in specialized domains such as medical diagnosis and chemical analysis.
- Perception of the world around is critical for the survival not only of human beings but also of animals.
- Animals even though have lesser intelligence than human but have higher visual perception than current machines.
- Perception problems are difficult to solve because they include analog signals.
- The signals are noisy, and they focus on too many things at the same time.

Language Problems

- One important characteristic that distinguishes humans from animals is the ability to use language to communicate.
- Understanding the spoken language is more difficult than the written language.
- The problems related to written language are referred to as natural language understanding are also difficult to solve.
- To understand a sentence and the unstated assumptions about a topic, one should have a good understanding of the vocabulary and grammar of the language as well as the topic itself.

Solving Specialized vs Mundane Tasks

- After acquiring knowledge in specialised tasks humans can perform specialised tasks in addition to the mundane tasks in the fields of engineering design, scientific discovery, medical diagnosis, and financial planning.
- The problems in these fields also fall in the aegis of artificial intelligence.
- A person who has the skills and knowledge to perform a task builds the necessary skills in a standard order.
- First, he learns perceptual, linguistic, and common-sense skills. Later, he acquires expert skills such as engineering, medicine, or finance.
- So, people believed that mundane skills are easier and more amenable to computerized duplication than specialized skills.

Solving Specialized vs Mundane Tasks

- But this assumption was not right.
- Even though specialized skills require knowledge that all human beings do not have but they often require less knowledge than many mundane skills.
- Not only that the specialized skills are easier to encode in a computer program than the mundane skills.
- As a result, the AI is flourishing most in solving practical problems related to domain that require specialized skills only.
- It is important to understand the following four topics
 - The underlying assumptions about intelligence
 - Techniques required to solve AI problems
 - Level of details required to model human intelligence
 - Successfully building an intelligent problem

The underlying assumptions

- The physical symbol system hypothesis (PSSH) lies at the heart of the research of artificial intelligence.
- In their Turing award-winning paper, Newell and Simon formulated PSSH.
- It states that “a physical symbol system [such as a digital computer, for example] has the necessary and sufficient means for intelligent action”.
- The hypothesis implies that when we provide computers with the appropriate symbol-processing programs, they will be capable of general intelligent action.
- It also implies as Newell and Simon wrote, that “the symbolic behaviour of man arises because he has the characteristics of a physical symbol system.”
- Efficient computing machine with increased processing power, have helped conduct experimental investigations of the physical symbol system hypothesis.
- In such investigations, a computer program conducts a test for a task that requires intelligence.
- The program may not be able to perform all the tasks at this time, but in the future, more sophisticated programs would be able to perform such tasks.

The underlying assumptions

- The physical symbol system hypothesis has been successful in game-playing and visual perception.
- However, the subsymbolic processes (such as neural networks) have been more successful in solving visual perception problems.
- Even though the conflict between subsymbolic processes and the physical subsystem hypothesis is under debate, the success of subsymbolic systems cannot be considered the evidence against the Physical symbol system hypothesis because there is more than one way to accomplish any task.
- The physical symbol system hypothesis is important for artificial intelligence due to two reasons.
- First, this theory gives insights into the nature of human intelligence.
- So, it is very important for the field of Psychology.
- Second, it forms the belief that it is possible to build computer programs and machines that can perform tasks either as efficiently as or better than humans can perform.

Required AI Technique

- The artificial intelligence problems are hard, unique, and span a broad spectrum. AI techniques provide appropriate solutions for this variety of problems. These AI techniques should possess certain properties
- One of the outcomes of research during several past decades is that **intelligence requires knowledge**. Some of the properties of knowledge are:
 - It is huge and complicated.
 - It is difficult to characterize correctly.
 - It changes continuously.
 - It differs from data by being organized in a way that corresponds to its application.

Required AI Technique

- An AI technique exploits knowledge that should be represented so that
 - Knowledge captures generalization.
 - The individual situation needs not to be handled separately however the situation that shares important property could be grouped together.
 - If the knowledge is not grouped together, an enormous amount of memory and computing power would be required.
 - Knowledge without generalization sometimes is called “data” instead of knowledge.
 - People who provide knowledge should be able to understand it. However, many programs capture the data automatically.
 - It can be easily updated to address the errors and changes in real-world scenarios.
 - It can be widely used in many situations even if it is not complete or not accurate.
- It is possible to solve AI problems without using AI techniques even though it is not recommended. It is also possible to solve the non-AI problem by using AI techniques. These are good things for problems that possess the same characteristics as AI problems do.

Level of the model

- Before start producing an AI project, we need to ask the following questions. AI projects have been motivated by these goals
- What is our goal in trying to produce programs that do the intelligent things that people do?
- Are we trying to produce programs that do the tasks the same way that people do?
- OR
- Are we trying to produce programs that simply do the tasks the easiest way that is possible?
- The efforts to build AI programs can be divided into two classes. Programs in the first class try to solve problems that be easily solved by a computer and AI techniques are usually not required. An example of this class of program is the Elementary Perceiver and Memorizer (EPAM) which memorized garbage syllables. For a computer it is easy to memorise the pairs of nonsense syllables. We can simply input them. But this task is hard for people.

Level of the model

- The second class of problems attempts to solve problems that are non-trivial for a computer and use AI techniques. We wish to model human performance on these:
 1. To test psychological theories of human performance. Ex. PARRY [Colby, 1975] – a program to simulate the conversational behaviour of a paranoid person.
 2. To enable computers to understand human reasoning – for example, programs that answer questions based upon newspaper articles indicating human behaviour.
 3. To enable people to understand computer reasoning. Some people are reluctant to accept computer results unless they understand the mechanisms involved in arriving at the results.
 4. To exploit the knowledge gained by people who are best at gathering information. This persuaded the earlier workers to simulate human behaviour in the SB part of AISB simulated behaviour. Examples of this type of approach led to GPS (General Problem Solver)

Criteria for the success

- For any scientific or engineering research project, we often ask – How will we know if we have succeeded?
- We ask the same question in the case of artificial intelligence. How we will know if we have constructed an intelligent machine?
- Turing Test as proposed by Alan Turing can help determine whether the machine can think.
- Some people believe that a computer never will be able to pass the Turing Test.
- If we are willing to settle for an incomplete imitation of humans, is it possible to measure the achievement of AI in restricted domains?

Criteria for the success

- Sometimes it is possible to measure the achievement of a program.
- For example, a program can achieve a chess rating in the same way the human player achieves.
- It is also possible to compare the time required by the program to complete the task to the time required by a human to complete the same/similar task.
- So, we can construct a computer that meets some performance standards for a task.
- It does not mean that the program will accomplish the task in the best possible way.
- It means there is at least one way of doing at least part of the task.
- When we design an AI program, we should specify the success criteria for that program in its restricted domain.

History of AI

Upto Year 1900



- The quest for Artificial Intelligence begins with dreams. Since ancient times, people have been talking about machines with human capabilities. These machines have been told in many tales and portrayed in many portraits and sculptures.
- According to ancient Greek mythology, as retold by Ovid in his Metamorphoses, Pygmalion, a sculptor, makes an ivory statue, named Galatea. The art represents his ideal of womanhood. He falls in love with his creation. To answer his prayer, the goddess Venus brings the statue to life.

Upto Year 1900

The Greek mathematician Archytas of Tarentum conceived a mechanical bird. The ancient Greek philosopher Aristotle (384-322 BCE) talked about abolition of the slavery through automation. He wrote in *The Politics*:

"For suppose that every tool we had could perform its task, either at our bidding or itself perceiving the need, and if-like the statues made by Daedalus or the tripods of Hephaestus, of which the poet [that is, Homer] says that "self-moved they enter the assembly of the gods" - shuttles in a loom could fly to and fro and a plucker [the tool used to pluck the strings] play a lyre of their own accord, then master-craftsmen would have no need of servants nor masters of slaves."

Aristotle

The Politics

Upto Year 1900

- Around 1495, Leonardo Da Vinci sketched the designs of a humanoid robots in the form of a medieval knight. The knight was able to sit up, move its arms and head, and open its jaws. It is not clear whether Leonardo Da Vinci or his associates tried to create the robot, but the seeds of an artificial intelligence-mechanized human were sown.



Upto Year 1900

- In 1651, Thomas Hobbes, the patriarch of artificial intelligence, published his book *Leviathan* about the social contract and the ideal state. In the book, Hobbes propounded that it might be possible to build an artificial animal.

"For seeing life is but a motion of limbs, the beginning whereof is in some principal part within, why may we not say that all automata (engines that move themselves by springs and wheels as doth a watch) have an artificial life? For what is the heart, but a spring; and the nerves, but so many strings; and the joints, but so many wheels, giving motion to the whole body..."

Thomas Hobbes
Leviathan

Upto Year 1900

- The science historian George Dyson referred to Hobbes as the "patriarch of artificial intelligence"
- In 1738, the French inventor and engineer, Jacques de Vaucanson created a mechanical duck that could quack, flap its wings, paddle, drink water, eat, and digest grains.
- In 1840, Lady Ada Lovelace foresaw part of artificial intelligence. She focused on symbols and logic. But she did not have any leaning towards the psychological aim of AI. Her focus was completely technological. She described various components of modern programming: stored programs, looping, conditionals, comments, and bugs. According to her AI was possible. But how to achieve was not clear.

From 1900-1956

- By the 1940s, a generation of computer scientists, mathematicians, and philosophers started working on the concept of artificial intelligence (AI).
- One of the pioneers of AI was Vannevar Bush in his work “As We May Think” which talked about a future world in which man-made machines will start to think.
- Alan Turing, a young British, who explored the possibilities of the mathematical possibility of artificial intelligence.
- In his famous 1950 paper “Computing Machinery and Intelligence” wrote about machines that can simulate human beings and do intelligent things. He raised the possibility that machines might be programmed to learn from experience like a young child.

From 1900-1956

- In 1943, Warren McCulloch and Walter Pitts suggested a model of artificial neurons in which each neuron is characterized as being “on” or “off”. They showed, for example, that any computable function could be computed by some network of connected neurons, and that all the logical connectives (and, or, not, etc.) could be implemented by simple net structures.
- In 1949, Donald Hebb demonstrated a simple updating rule for modifying the connection strengths between neurons. His rule, now called Hebbian learning, remains an influential model to this day
- In 1950, two undergraduate students at Harvard, Marvin Minsky, and Dean Edmonds built the first neural network computer.

1956 – Birth of AI

- John McCarthy organized a two-month workshop, Dartmouth Summer Research Project on Artificial Intelligence (DSRPAI), at Dartmouth in the summer of 1956. The proposal states:
 - *"We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer"*

1956 – Birth of AI

- No new breakthroughs could achieve from the Dartmouth workshop.
- But the workshop introduced all the major figures in the field to each other.
- These people worked in the field of AI for the next 20 years with their students and colleagues at MIT, CMU, Stanford, and IBM.

1957 – 1990

- AI flourishes from 1957 to 1974. Computers became faster, cheaper, and accessible. Computers could also store more information. During this time, the machine learning algorithms also improved. People started understanding better which algorithms they could apply to solve their problems.
- Newell and Simon developed General Problem Solve (GPS) whereas Joseph Weizenbaum developed ELIZA.
- GPS and ELIZA were promising demonstrations towards the goals of problem-solving and interpretation of spoken language.
- These successes convinced government agencies such as the Defence Advanced Research Projects Agency (DARPA) to fund AI research at several institutions.
- The government looked for a machine that could transcribe and translate the spoken language. Even though the optimism was high, and expectations were even higher, there was still a long way to go.
- In 1970 Marvin Minsky told Life Magazine, “from three to eight years we will have a machine with the general intelligence of an average human being.” However, a lot of work was required to achieve the end goals of natural language processing, abstract thinking, and self-recognition.

1957 – 1990

- AI faced several obstacles. The biggest obstacle was the lack of computational power, which is required to achieve anything substantial.
- Computers were primitive and they could not store enough data or process it fast.
- Hans Moravec, a doctoral student of McCarthy at the time, stated that “computers were still millions of times too weak to exhibit intelligence.”
- This led to dwindling funds and hence the research in AI came to a slow roll for ten years.

1957 – 1990

- In the 1980s, the expansion of the algorithmic toolkit and boost of funds reignited AI. John Hopfield and David Rumelhart proposed “deep learning” techniques.
- These techniques helped computers to learn user experience.
- Edward Feigenbaum developed **expert systems**.
- These systems mimicked the decision-making process of a human being. The expert systems were widely used in the industry.
- From 1982-1990, the Japanese government invested \$400 million dollars in expert system and other AI related projects, as part of the First-Generation Computer Project (FGCP).
- The goal was to revolutionize computer processing, implement logic programming, and improve Artificial Intelligence.
- The FGCP inspired several talented young engineers and scientists but most of the goals were not achieved. Slowly the funding of the FGCP ceased, and AI fell out of the limelight.

1990 Onwards

- Even though the government funding ceased, AI prospered during the 1990s and 2000s.
- During this period several landmark goals of artificial intelligence were achieved.
- In 1997, IBM's Deep Blue, a chess-playing computer program, defeated Gary Kasparov, reigning world chess champion and grandmaster during a highly publicized match.
- The match served as a huge step towards an artificially intelligent decision-making program.
- In the same year, Dragon Systems developed speech recognition software that was implemented on Windows.
- The speech recognition software was another great step towards the direction of the spoken language interpretation efforts.
- Cynthia Breazeal developed a robot, Kismet, that could recognize and display emotions.

1990 Onwards

- The memory and speed of computers double every year. So, the fundamental problem of computational power and storage was no longer a problem.
- The substantial improvement in computational power helped IBM's Deep Blue beat Gary Kasparov and Google's Alpha Go beat Chinese Go champion, Ke Jie. This helps explain the roller coaster of AI research.
- The AI capabilities saturate to the level of existing computational power. As the computation power increases with Moore's Law, the AI capabilities improve again.
- In the age of “big data”, we have the capacity to collect a huge sum of information, which is very difficult for a person to process.
- Even if the algorithms do not improve much, big data and massive computing allow artificial intelligence to learn through brute force.
- This has helped in the fruitful implementation of AI in several industries such as technology, banking, marketing, healthcare, drug discovery, and entertainment.

Thanks

Samatrix Consulting Pvt Ltd

Artificial Intelligence, Machine Learning, and Data Science

Samatrix Consulting Pvt Ltd

Data Science Processes

Structured Approach

- By following a structured approach, you can maximize the chances of success of a data science and a machine learning project at a low cost.
- The structured approach helps the project manager assign the roles and responsibilities of the project team effectively.

Six Steps

- A typical machine learning and data analytics project consists of six steps. The main steps and actions that needs to be taken are as follows.
 - Set the research goal:- The first step is to set a research goal. Every stakeholder understands the what, how, and why of the project. This step also results into a project charter.
 - Data retrieval:- The second step is data retrieval that includes finding the suitable data and getting access to the data from various sources. At the end of the step, we may receive the raw data that needs polishing and transformation.

Six Steps

- Data preparation:- After receiving the raw data, the next step would be to prepare it. The data preparation includes data transformation so that the raw data can be appropriately used in your model. In this step, you detect and correct different kinds of errors in the data, combine data from different sources, and transform it. After the successful completion of this step, you can proceed to the data visualization and data modeling steps.
- Data exploration: - This step helps gain a deeper understanding of the data. You discover the patterns, correlations, and deviations using the visual and descriptive techniques. The insights from this step helps you start modeling.
- Data Model building: - In this step, you gain the insights or make the predictions as stated in the project charter.
- Presenting results and automating the analysis: - This is the last step of the data science and machine learning process. This step helps implement the process changes and automation for better business decisions.

Six Steps

- These steps help achieve the project goals and improve the project success ratio while reducing the project cost.
- This process ensures that you have a well-defined research plan, a good understanding of the business question, and clear deliverables before you start the project.
- In the first step, you focus on getting the high-quality data.
- If you are able to get the high-quality data, the chances of building a high performing data model is high.
- If the data is not of high quality, the chances of building a high performing data model is low because garbage in equals garbage out.

Structured Approach - Benefits

- The structured approach helps you focus on prototype mode where you try multiple models and focus on business value instead of focusing on program speed or coding standards.
- The structured approach also helps the project manager assign the project team on different aspects of the project while focusing on the overall goals.
- The tasks such as extracting and uploading the data to different databases, designing the data scheme that works not only for your applications but also for other projects inside your company, tracking the statistical and data-mining techniques, and expertizing the presentation tools need specialized skills.

Agile Methodology

- Every project does not follow these steps.
- The process followed by a project team is subjected to the preference of the lead data scientist, the company, and the nature of the project.
- Other project models such as **agile** project model is an alternative to the sequential process with iterations.
- Agile project model has gained more ground in the IT department and the companies.
- Hence, the data science communities have also been adopting agile methodologies.

Step 1: Define research goals and
create project charter

Define Research Goals

- The project starts with understanding of what, why, and how of the project.
- What is the goal of the project?
- What are the expectations from the project team?
- Why the senior management is investing in the project?
- How the senior management would get the return on investment (ROI) of the project?
- Whether the project is a part of a bigger strategic picture or it is “lone wolf” project.

Project Charter

- The goal of the first phase is answer three questions (what, why, and how) and help the project team and the stakeholders understand what needs to be done and everybody agrees on an action plan.
- The outcome of the step should be a clear research goal, a good understanding of the context, well-defined deliverables, and an action plan with timelines.
- This information is also included in the project charter.
- The important tools and techniques adopted during this step are: expert judgement, facilitated workshops, product analysis, alternative generation, questionnaires and surveys, observations, prototypes, benchmarking, and document analysis.
- This stage is guided by a senior project manager or data scientist because this stage requires strong people skills and business acumen.

Goal and Scope of the Project

- The goal and scope of the project should be clearly documented and signed off by the stakeholders.
- Understanding the project goal and the scope is very important for the success of the project.
- The data scientist need to ask questions continuously until he clearly understands the business expectations and deliverables of the projects.
- He should be clear about how the project fits in the bigger organization objectives, how the project deliverables are going to be used, how the project outcome will change the business.

Goal and Scope of the Project

- After spending months and critical resources, if you realize that you misunderstood the question, you will be frustrated.
- So, this is an important phase.
- Any project manager or data scientist or machine learning engineer should be taken this stage lightly.
- Even though the team is highly motivated and technically competitive, many projects fail to meet their objective, if this stage is not completed appropriately.

Create a Project Charter

- The project charter brings all the stakeholders on the same page and to avoid scope creep, the data scientist should document the problem statement and deliverables, and get the approval from the clients.
- All the information is documented in the project charter.
- The project charter is mandatory for a mid-size to large project.
- The project charter should document the following:
 - The research goals
 - The project mission
 - The project requirements
 - The project scope and deliverables
 - The work breakdown structure (WBS)
 - Resource requirement in terms of man-hours and other item of cost
 - Prototype or proof of concepts (if any)
 - Project schedule
- The project client may need this information to estimate the project cost, resource requirement, and the data required for the project.

Step 2 – Data Retrieval

Data Retrieval

- During the data retrieval step, the required data is retrieved.
- On many occasions, the data scientist needs to design the data collection process himself.
- On several other occasions, the data is already available within the company.
- If required, the company can also buy the required data if the data is not available in the company.
- These days, high quality data is also available free of cost for public and commercial use.
- The data scientists also look for such data sources.

Data Retrieval

- The data is available in many forms.
- The objective of the project team is to acquire all the required data which may be a difficult task.
- Even if, you are able to acquire the data, you need to pre-process the data before you can use it.

Acquire Internal Data

- First of all, you should access the relevance and quality of data that is available within the company.
- Many companies adopt a robust data management policy.
- The quality and pre-processed data is readily available in such companies.
- The data is stored in official data repositories such as databases, data marts, data warehouses, and data lakes.

Acquire Internal Data

- The databases are used to store the data.
- Data warehouse is used for reading and analyzing the data.
- A data mart is geared towards serving the needs of a specific business unit.
- The preprocessed data is stored in the data warehouses and data marts whereas raw data is stored in the data lakes.
- In addition to data repositories, data also resides in Excel files on the desktop of a domain expert.

Acquire Internal Data

- Sometimes, the required data is not easily available in the company.
- As the company grows, the data scatters around many places.
- The knowledge of the data also disperses as people change positions and leave the company.
- The metadata is not maintained properly. In such situation, the task of the data scientist becomes challenging.

Acquire Internal Data

- Sometimes, getting access to the data is also challenging.
- Many organizations adopt strict data access policies because they understand the value and sensitivity of the data.
- Due to this, they implement digital and physical barriers and access is given on need-to-know basis.
- Hence, in many organizations, getting access to the data may take time and involve company politics.

Acquire External Data

- If the required data is not available within the organization, the data scientist can look for external data sources.
- Many companies such as Nielsen and GFK specialize in collecting valuable data from various industries.
- Even though the data is very valuable, more and more organizations and governments share their data for free for commercial use.
- Often, this data is of excellent quality.
- The valuable information such as number of accidents or amount of drug abuse in a certain region and its demographics.
- Some of the reputed open-data providers are given below:

Acquire External Data

Open Data Site	Description
Data.gov	US Government's Open Data
https://open-data.europa.eu	European Commission's Open Data
Freebase.org	Open Database that retrieves the information from sites like Wikipedia, MusicBrains, and the SEC archive
Data.worldbank.org	Open Database from the World Bank
Aiddata.org	Open Database for the international development
Open.fda.gov	From the US Food and Drug Administrator

Check the Data Quality at Source

- During the lifecycle of a machine learning or a data science project, the project team spends a good amount of time (up to 80% of the time) in data correction and cleansing.
- The first time when you inspect the data is during the data retrieval stage.
- Most of the errors in data are easy to spot during this stage.
- If the errors are ignored at this stage, you will end up spending good amount of time in solving data issues at a later stage.
- The data is investigated during three phases of the project lifecycle: data retrieval, data preparation, and data exploratory phases.
- There is a difference in the goal and the depth of investigation during each phase.

Check the Data Quality at Source

- During the data retrieval phase, you check to see if the data retrieved is same as data in the source document and its data types.
- During the data preparation phase, more elaborate data checking is done.
- If you are able to check the data thoroughly during the data retrieval phase, you will find errors in the source document.
- You should focus on the content of the variables.
- You should remove typo errors, and convert the data to the common format among the data sets.

Check the Data Quality at Source

- For example, you might correct USQ to USA and Indio to India.
- During the exploratory phase, you focus on the statistical data analysis.
- You assume the data to be clean and look at the statistical properties such as distribution, correlations, and outliers.
- On several occasions, you need to iterate over these processes.
- For example, the root cause of an outlier that you may find in the exploratory phase, could be data entry error. Hence during the process, the data quality improves.

Step 3: Cleansing, Integrating, and Transforming Data

Cleansing, Integrating, and Transforming Data

- During this stage, you sanitize and prepare the data so that it can be used in modeling and reporting phase.
- This is an important step so that performance of the model can be maximized and the strange errors can be minimized.
- Your model may need the data in a specific format.
- So, the data transformation is also required.
- You should strive to correct the data errors as early as possible during the process.

Cleansing, Integrating, and Transforming Data

- During this stage, you sanitize and prepare the data so that it can be used in modeling and reporting phase.
- This is an important step so that performance of the model can be maximized and the strange errors can be minimized.
- Your model may need the data in a specific format.
- So, the data transformation is also required.
- You should strive to correct the data errors as early as possible during the process.

Sub Processes

- Data Cleansing
 - Data entry errors
 - Redundant whitespace
 - Impossible Values
 - Outliers
 - Missing Values
 - Deviation from a code book
 - Different units of measurement
 - Different levels of aggregation

Sub Processes

- Combine data from various data sources
 - Joining Table
 - Appending Table
 - Using Views
- Data Transformation
 - Reduce the number of variables
 - Creating dummies
 - Extrapolating Data
 - Data Aggregation
 - Derived measures

Cleansing Data

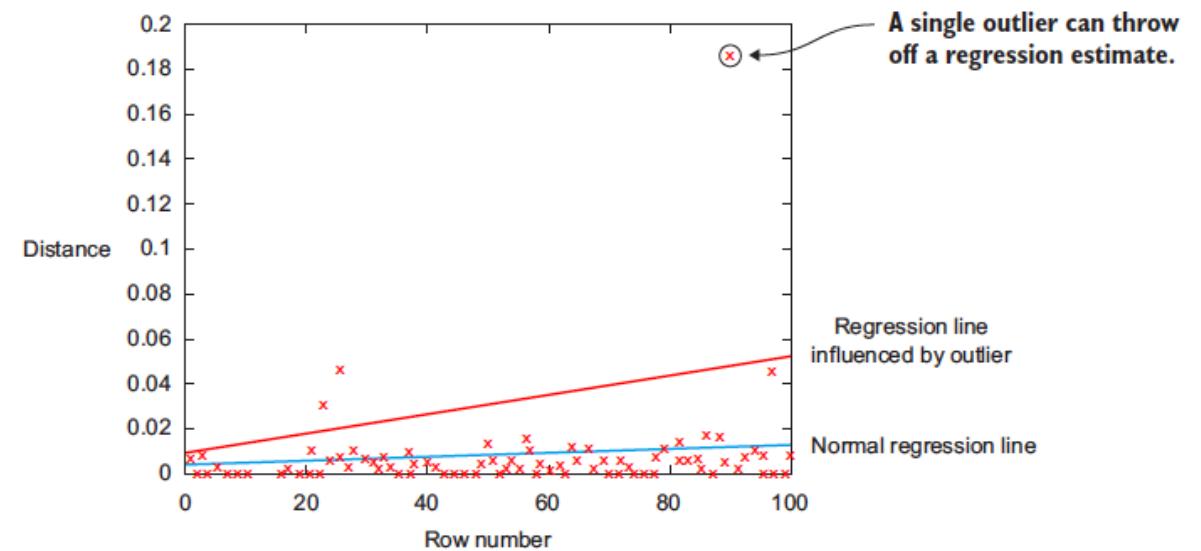
- This process focused on removing the errors from the data so that the data becomes a true and consistent representation of the processes from which it originates.
- At least two types of errors exist in the data.
- The first one is known as interpretation error.
- These are due to values in the data such as saying that a person's age is greater than 300 years.
- Second type of error is due to inconsistencies between the data source and your companies standardized values.

Cleansing Data

- For example, the data source has value “Female” whereas the company standard value is “F” even though they represent the same thing.
- Another example is that the data source has financial values in Dollars whereas the company standard table has financial values in Rupee.
- Sometimes, to identify the data errors, we use advance methods such as simple modeling.
- For example, we used regression to identify the data points that seem out of place.

Cleansing Data

- The regression method can also be used to understand the data and detect the influence of each observations on the regression line.
- On various occasion, we may find a single observation that has too much influence.
- Such observations may represent errors or a valid data point.
- Such advanced methods are rarely used at the data cleansing stage but they are useful tools.



Data Entry Errors

- The data collection and data entry processes are prone to errors.
- Due to human errors such as typos or lack of concentration, such error occurs.
- The machine generated data is also prone to errors.
- Transmission errors or bugs in the extract are some of the common cause of errors in machine generated data.
- One of the methods of check the manual data entry errors is to tabulate the data with counts.
- For example, if a variable in the dataset can take only two values: “Good” and “Bad”.
- You can create a frequency table to check if only those two values are present. In the table below, the values “Goood” and “Bda” point out wrong entries. At least in 20 cases
- Such errors can be fixed manually or programmatically.

Data Entry Errors

Value	Count
Good	1678763
Bad	1274648
Goood	15
Bda	5

Redundant Whitespace

- Whitespaces at the end of the string are hard detect and results in wrong results from the data analysis.
- During the ETL process, the whitespace may lead into errors in the dataset.
- For example there is a difference between “AB ” and “AB”.
- During the data cleaning process, most of the programming languages can easily remove the leading and trailing whitespace.
- In Python, the leading and trailing whitespaces can be removed by using `strip()` function.

Capital Letter Mismatch

- The capital letter mismatches in data are common.
- For most programming languages there is difference between “India” and “india”.
- Using Python you can solve this problem by applying a function, `.lower()`, that returns both strings in lowercase.
`“India”.lower() == “india”.lower()`
- should result in true.

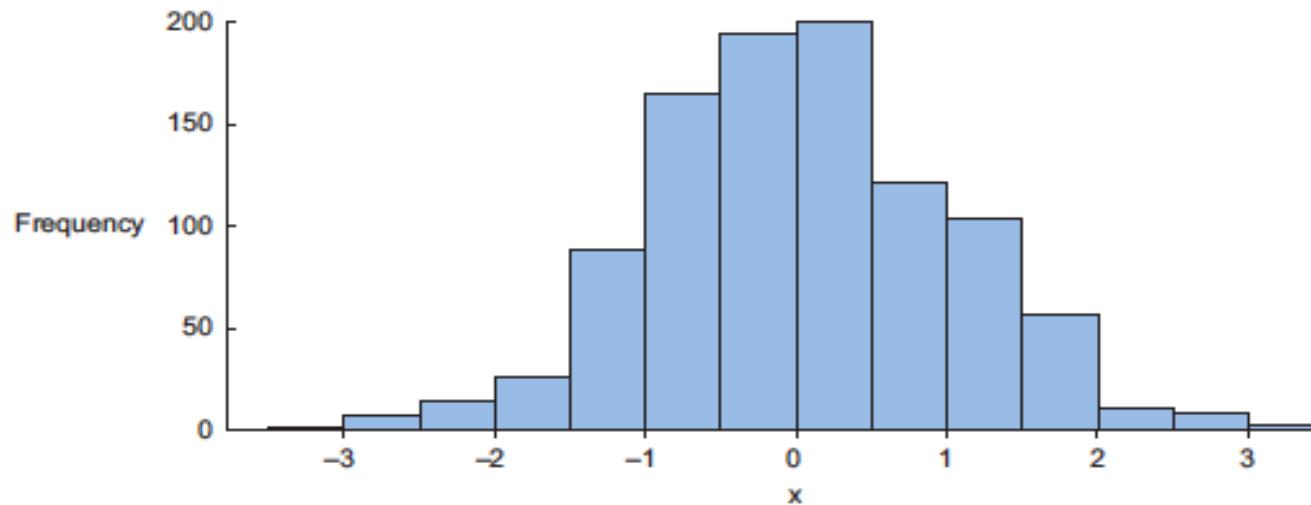
Impossible Values

- Another valuable type of data check is sanity check of impossible values.
- In this case, you need to check whether there is any physically or theoretically impossible values in the data set.
- People taller than 3 meters or people with an age of 299 years are examples are impossible values.
- You can do the sanity check with rules: $\text{check} = 0 \leq \text{age} \leq 120$

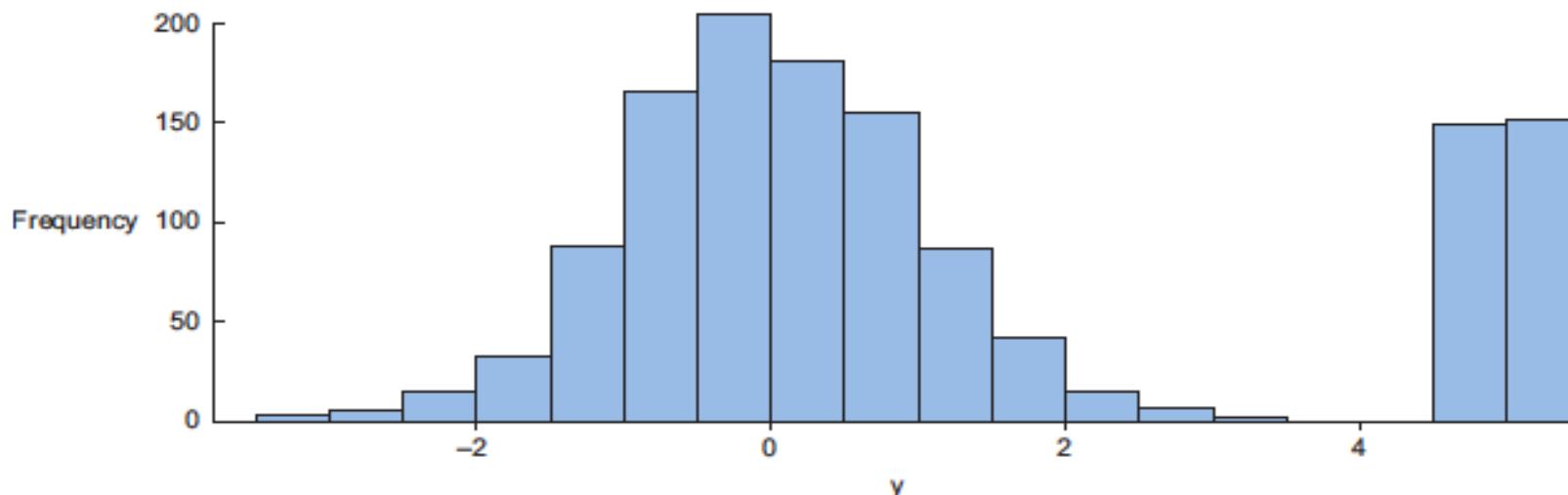
Outliers

- The observation that seems to be distant from other observations is known as outliers.
- Outliers follow different generative process or logic than rest of the observation.
- You can easily find the outliers using a plot or a table with the minimum and maximum values.
- The first figure below shows no outliers whereas the second figure shows the presence of the outliers.

Expected distribution



Distribution with outliers



Outliers

- The first plot shows that the most of the observations are around the mean value of the distribution.
- The frequency of observation decreases as they are away from the mean value.
- For a normal distribution, the high values in the right hand side of the bottom graph shows the presence of the outliers.
- Outliers can influence the results of your data model.
- So, you need to investigate them first.

Missing Values

- In our dataset, you may find several missing values.
- They are not always wrong, still you need to handle them.
- Several data modeling techniques cannot handle missing values effectively.
- Sometimes, they may be a result of faulty data collection process or an error during ETL process.
- The common techniques of fixing the missing values are given below.
- The data scientists and machine learning engineers use them according to their data and requirements of the model.

Missing Values

Technique	Advantage	Disadvantage
Omit the values	Easy to perform	You lose information from observation
Set value to null	Easy to perform	All the data modeling techniques cannot handle the null values
Insert a static value such as 0 or the mean	Easy to perform Information from other variables in the observation is not lost	May result into the false estimations from the model
Insert values based on estimated or theoretical distribution	Model is not disturbed	Harder to execute You make data assumptions
Modeling the value (nondependent)	Model is not disturbed	Can lead to too much confidence in the model Can artificially raise dependence among the variables Harder to execute You make data assumptions

Unit of Measurements

- While integrating two different data sets, you should check the respective unit of measurement.
- For example: the distance between two cities in the world can be provided by different data providers.
- One dataset may contain the distance in miles where other dataset may contain the distance in kilometers.
- You can use the simple conversion formula to fix such issues.

Correct Errors as early as possible

- Many key organization decisions depend on the insights from the data.
- Hence organizations spend good amount of money on the data retrieval process.
- But the data retrieval process is a difficult, expensive, and error prone task.
- Hence, as soon as the data is retrieved, it should be corrected due to following reasons.

Correct Errors as early as possible

- Spotting the data anomalies is not easy for everyone. The organization may end up in taking costly and wrong decisions, if they are taken based models that are using the incorrect data
- If the data cleansing is not done at an early stage, all the downstream projects have to clean the data that is a costly and time-consuming process
- Data errors may be due to a fault in the business process that needs to be corrected else it may result in the loss of revenue for the organization
- Data errors may be due to defects in the equipment such as defective sensors
- Data errors may be due to bugs in the software that may be critical for the company.

Correct Errors as early as possible

- Ideally, the data errors should be fixed as soon as it is captured.
- But the data scientists do not control every data sources.
- They may point out the data errors but it is up to the owner of the data to fix the issue.
- If you cannot fix the data at the source, you have to handle it inside your data model and the code.
- It is always recommended to keep a copy of the original data.
- On several occasions, when you start data cleaning process, you may make mistakes.
- You may insert the wrong values, or delete an outlier that contained useful information, or alter the data due to misinterpretation of the data.

Correct Errors as early as possible

- If you have a copy of the original data, you can start again.
- The real time data is manipulated at the time of arrival.
- Hence it is not possible to keep a copy of the original data.
- In that case you can use data between certain timeframe for tweaking before you start using the data.
- Cleansing the individual data is not the most difficult thing.
- However, combining the data from different sources is the real challenge.

Combine Data from Different Sources

- In this step we focus on integrating the data that we get from different sources.
- This data varies in size, type, and structure.
- It is available in several formats ranging from database and Excel files to text documents.
- In this section for the sake of brevity, we will focus on data in table structure.
- In order to combine data from two different sources, you can perform two operations: joining and appending or stacking.
- In the joining operation, we enrich an observation from one table with information from another table.
- In the second operation, appending or stacking, we add the observations of one table to those of another table.
- During these operations, you can either create a new physical table or a new virtual table, called view. The views do not consume much of the disk space.

Joining Tables

- Joining tables help you combine two corresponding observations in two tables and enrich a single observation.
- For example, if one table contains the information about the purchases of a customer.
- Another table contains the information about the region where the customer lives.
- You can join the tables to combine the information and use it for your model.

Joining Tables

- In order to join table, you need to use the columns or variables that represent the same observation in both the tables such as customer name, customer id, or social security number.
- These common fields are known as keys.
- When the keys store unique and not null information, they are called Primary Keys.
- In the picture above, you can see that Client variable is common in two tables.
- Using the join, the region variable has been included with the purchase information.

Client	Item	Month
John Doe	Coca-Cola	January
Jackie Qi	Pepsi-Cola	January

Client	Region
John Doe	NY
Jackie Qi	NC

Client	Item	Month	Region
John Doe	Coca-Cola	January	NY
Jackie Qi	Pepsi-Cola	January	NC

Appending Tables

- In the appending operation, you can append or stack observations from one table to another table.
- In the example below, one table contains the data from January whereas another table contains the data from February.
- As a result of appending table operation, you get a larger table with data from January as well as February.
- This operation is known as “Union” in set theory as well as SQL.

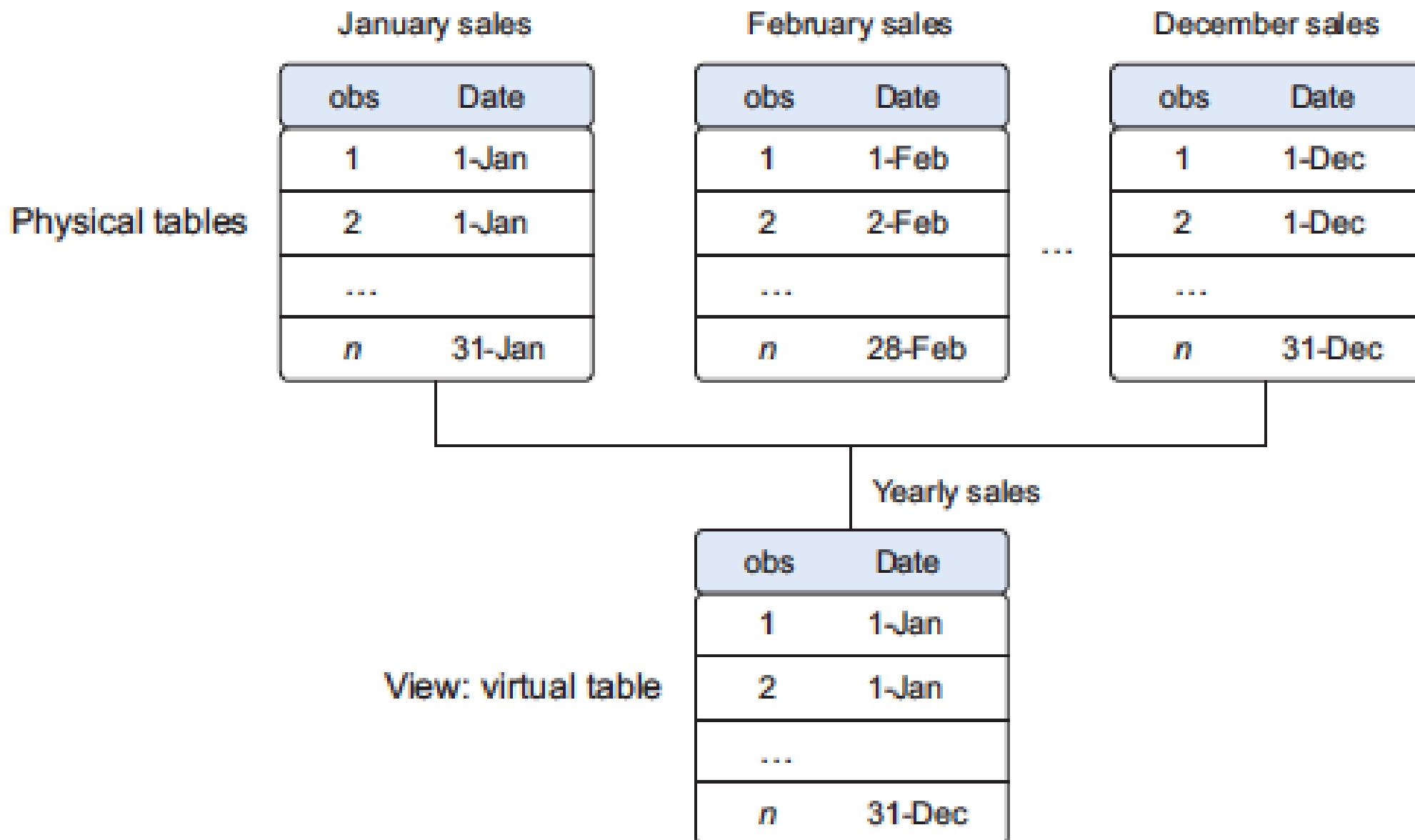
Client	Item	Month
John Doe	Coca-Cola	January
Jackie Qi	Pepsi-Cola	January

Client	Item	Month
John Doe	Zero-Cola	February
Jackie Qi	Maxi-Cola	February

Client	Item	Month
John Doe	Coca-Cola	January
Jackie Qi	Pepsi-Cola	January
John Doe	Zero-Cola	February
Jackie Qi	Maxi-Cola	February

Using Views

- We use views to avoid the duplication of data.
- In the previous example, we took the monthly data from two tables and combined both the data in a new physical table.
- Due to this the data was duplicated and more storage space was required.
- If each table contains terabyte of data, the duplication of data would be an issue.
- In such situations you can use the views.
- The view creates a virtual layer that combines the tables for you.
- The figure below shows how the sales data from the different months is combined virtually into a yearly sales table instead of duplicating the data.
- The views use more processing power than a pre-calculate table because, every time the view is queried, the join that creates the view is recreated.



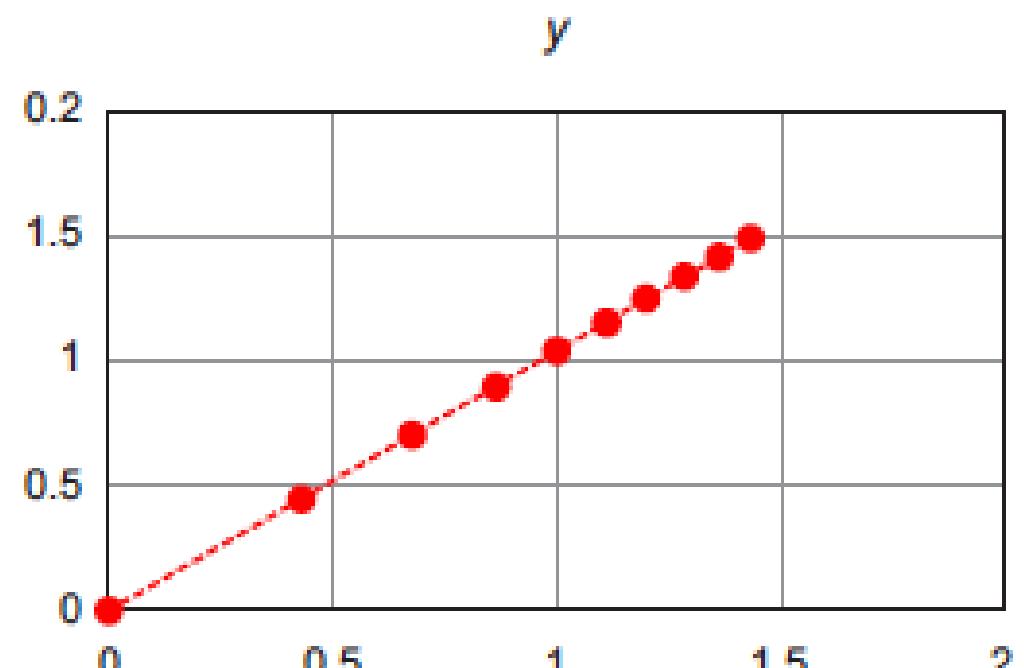
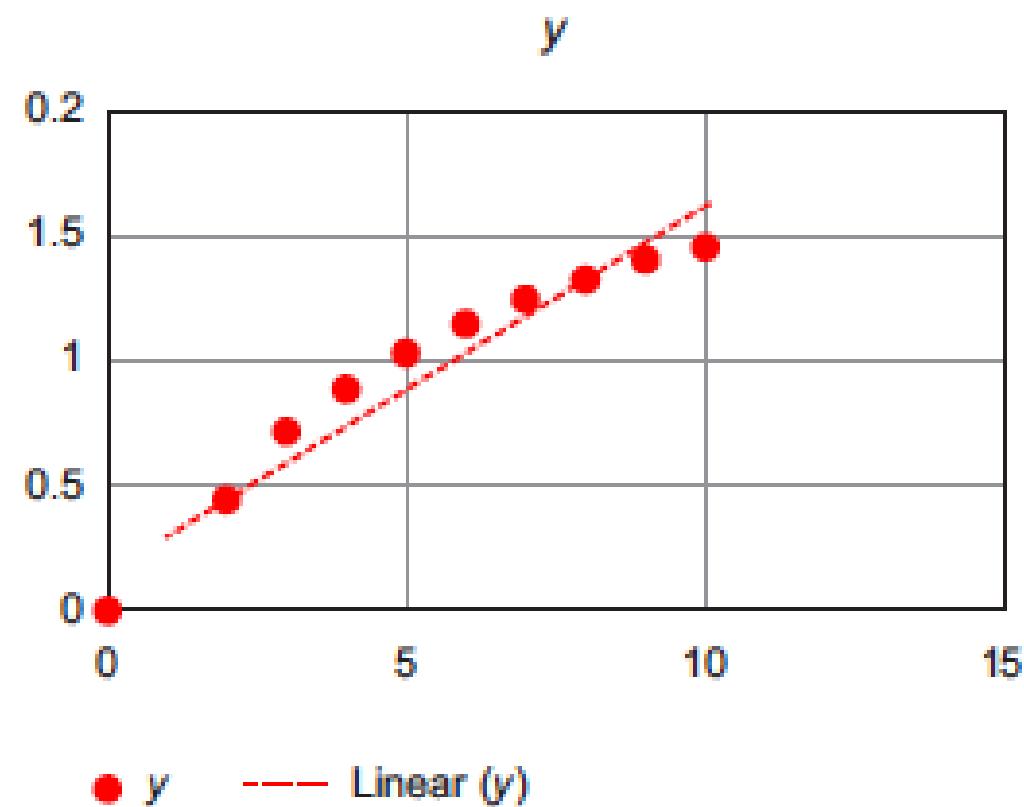
Transforming Data

- Certain models need the data in a certain format.
- After cleansing and integrating steps, you focus on transforming the data.
- The objective of this step is to ensure that the data takes the required shape for the data modeling.

Data Transformation

- On various occasions, the relationship between input and output variable is not linear.
- For example, the relationship $y = ae^{bx}$
- Log of the equation simplifies the estimation problem significantly.
- Combining two variables can also simplifies the estimation problem.

x	1	2	3	4	5	6	7	8	9	10
$\log(x)$	0.00	0.43	0.68	0.86	1.00	1.11	1.21	1.29	1.37	1.43
y	0.00	0.44	0.69	0.87	1.02	1.11	1.24	1.32	1.38	1.46



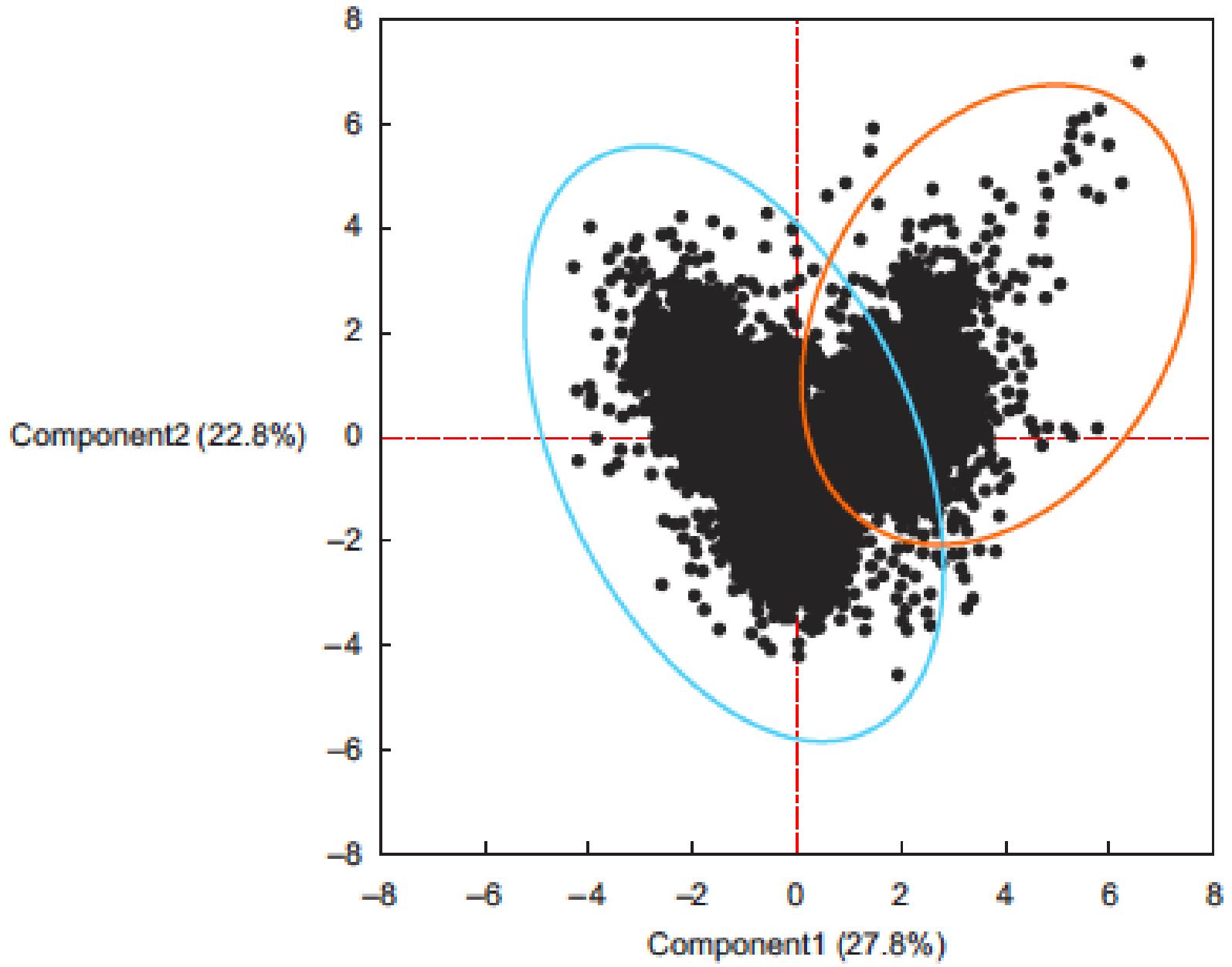
● y —— Linear (y)

Reduce the number of variables

- Too many input variables in a model, make the model complex and difficult to handle.
- You need to reduce the number of variables without loosing the information.
- You can do so by reducing the number of variables that do not add new information to the model.
- You can use the special methods to reduce the number of variables but retain the maximum amount of data.
- One such method, known as principal component analysis, is given below.

Reduce the number of variables

- In this method we can see that two variables account for 50.6% of variations within data set (component1 = 27.8% + component2 = 22.8%).
- These variables, called “component1” and “component2,” are both combinations of the original variables.
- They’re the principal components of the underlying data structure.



Dummy Variables

- Dummy variable work with categorical variables in which the different values have no real numerical relationship with each other.
- Dummy variables take only two values: true (1) or false (0).
- For example, if we take the value of male and female in the application form as 0 and 1, it does not mean that male has zero influence among all the input variable and female has one.
- In such case, we make insert two dummy variables one for male and another one for female.
- For the male dummy variable, we insert the value 1 for the rows corresponding to the male and 0 otherwise.
- For the female dummy variable, we insert the value 1 for the rows corresponding to the female and 0 otherwise

Customer	Year	Gender	Sales
1	2015	F	10
2	2015	M	8
1	2016	F	11
3	2016	M	12
4	2017	F	14
3	2017	M	13

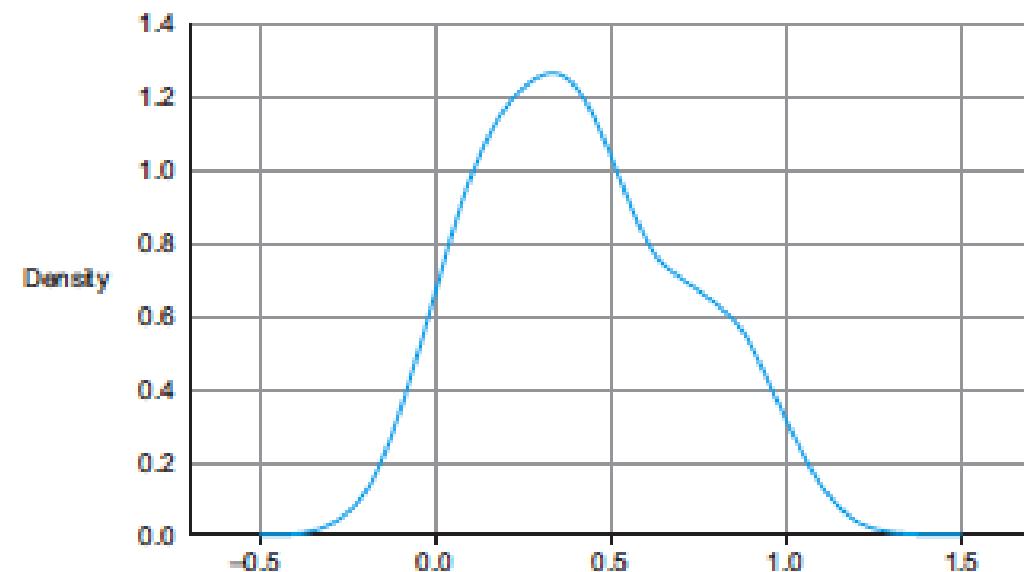
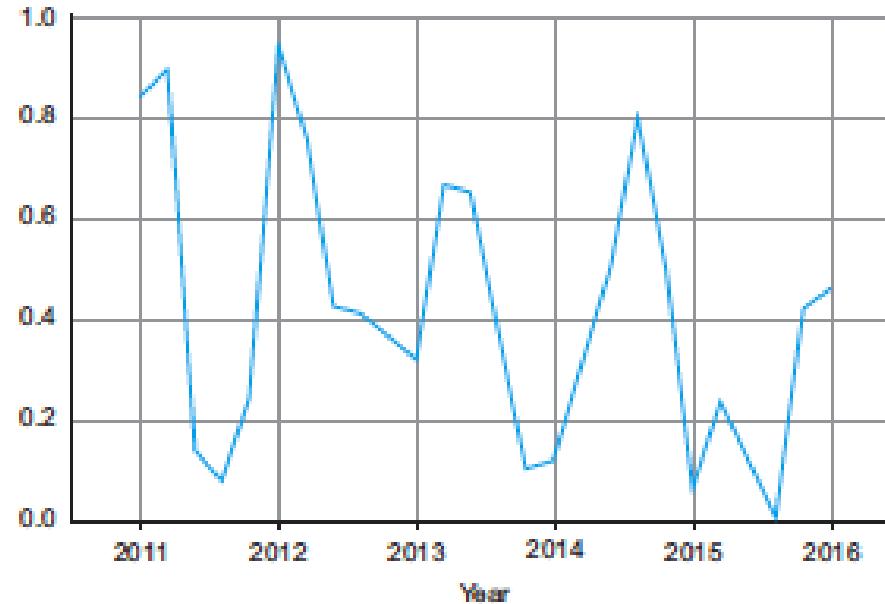
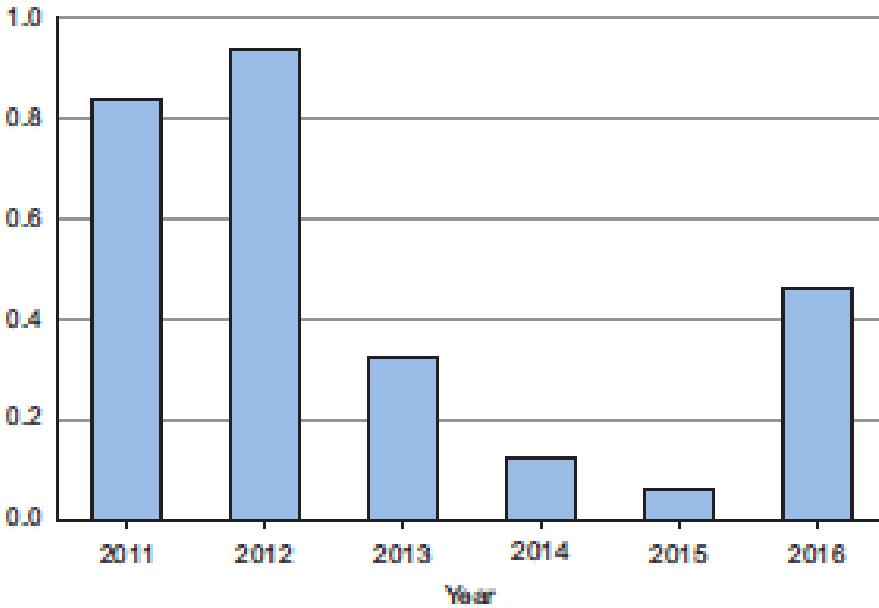


Customer	Year	Sales	Male	Female
1	2015	10	0	1
1	2016	11	0	1
2	2015	8	1	0
3	2016	12	1	0
3	2017	13	1	0
4	2017	14	0	1

Step 4: Exploratory Data Analysis

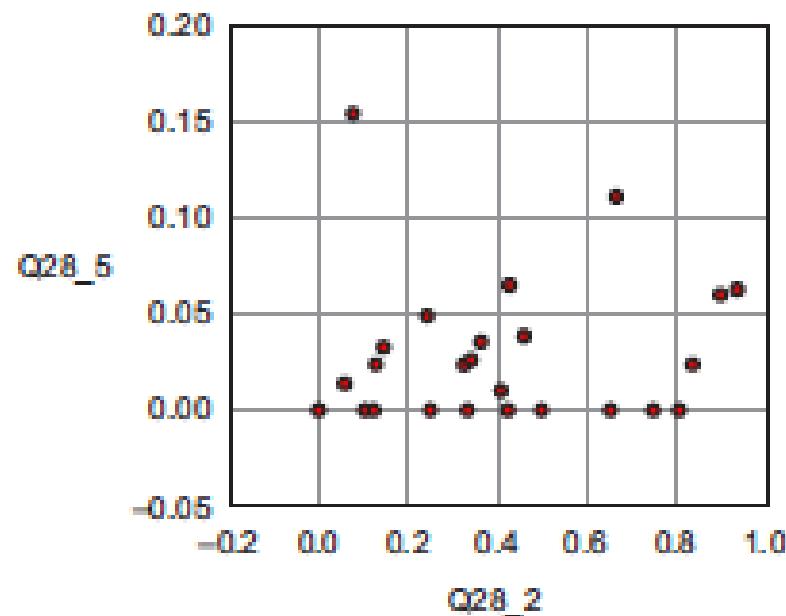
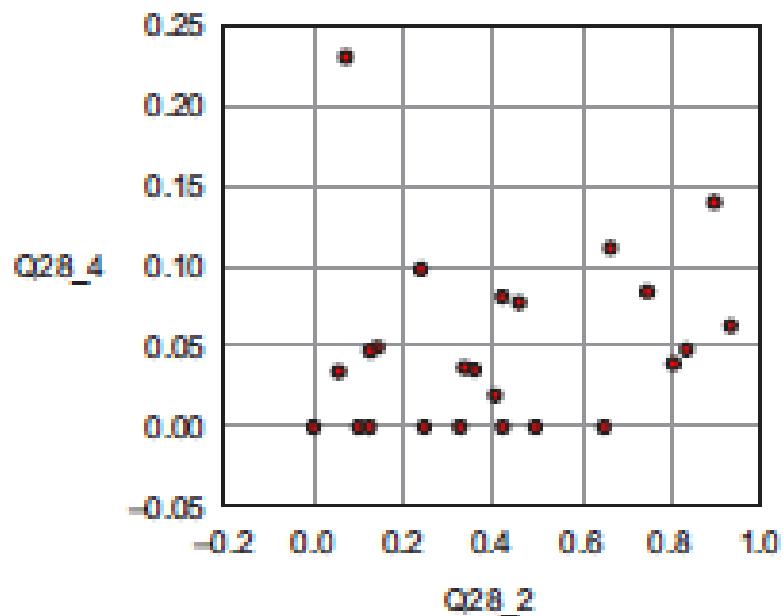
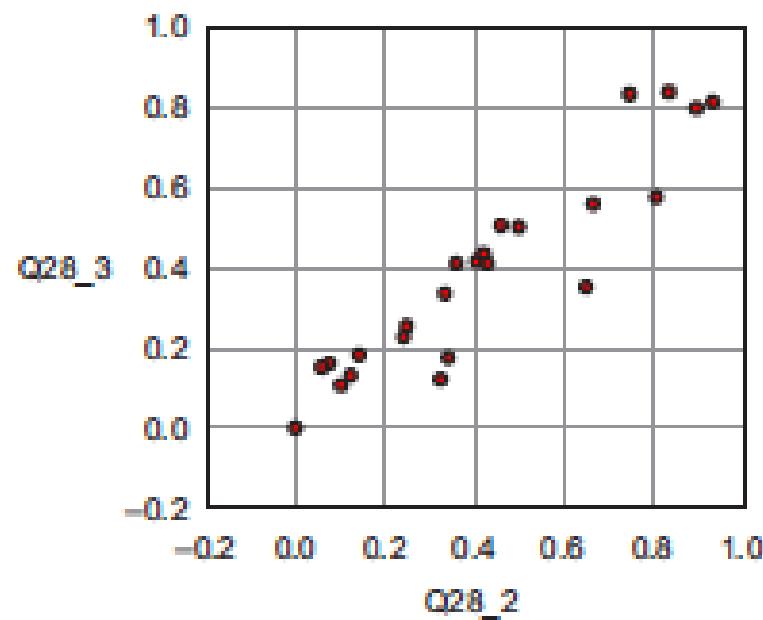
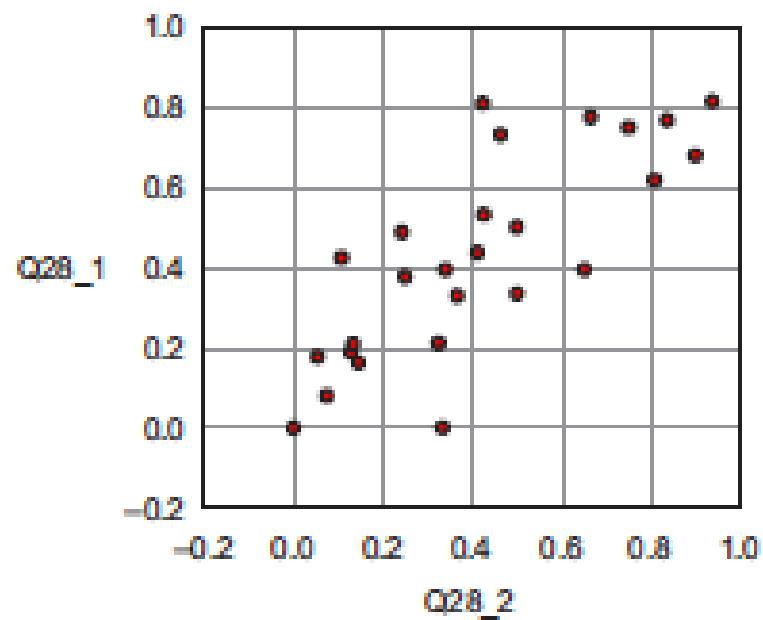
Exploratory Data Analysis (EDA)

- During the exploratory data analysis (EDA), you take a deep dive into the data.
- You can easily gain the information about the data from the graphs.
- Therefore, you use graphical techniques to understand the data and relationship among the variables.
- We explore the data in this step.
- However, we still try to find if there is any anomaly left in the data even after the data cleansing and transformation steps.
- If you discover them in this step, you should go back and fix them before moving to next step.
- You can use several visualization techniques that includes bar chart, line plots, and distributions as given below.



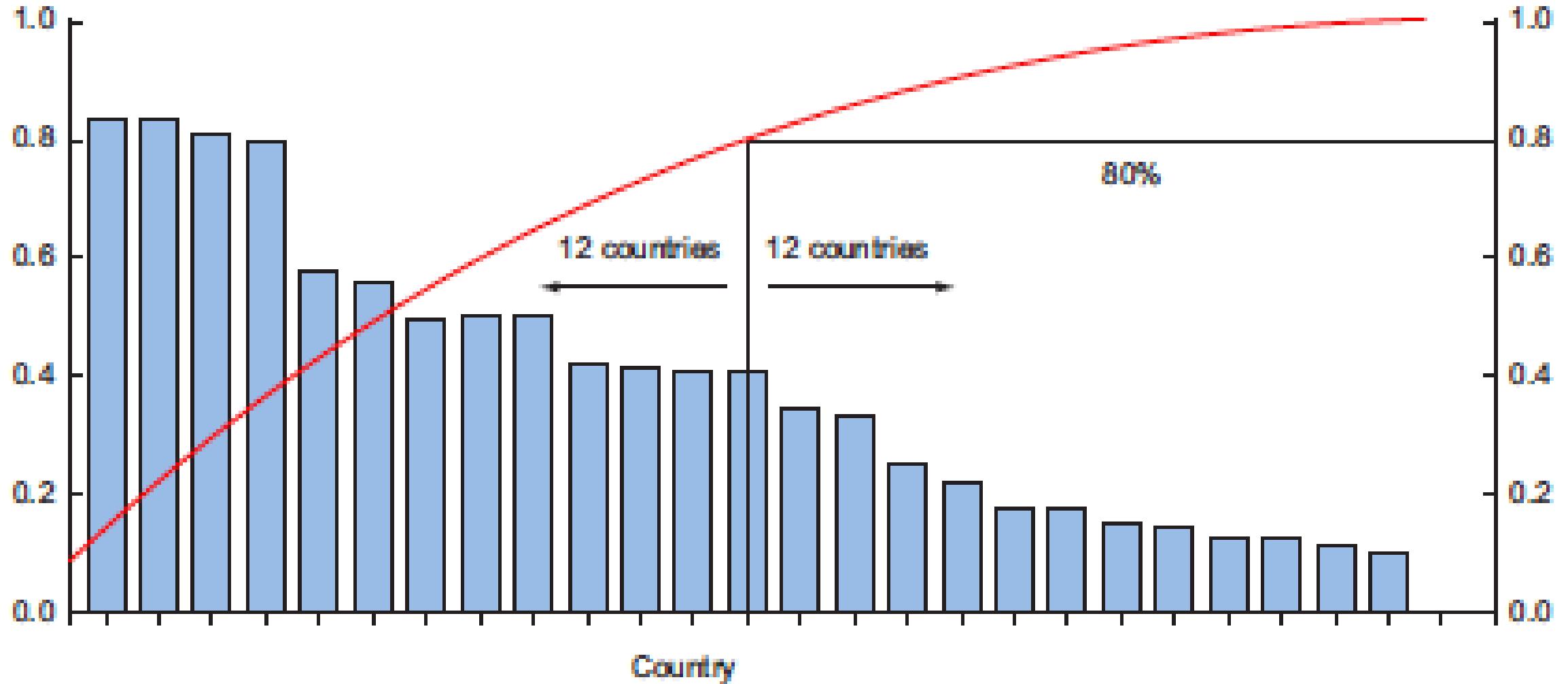
Exploratory Data Analysis (EDA)

- Sometimes, you can compose a composite graph from the simple graphs to gain better insights from the data.
- The graph below helps you understand the relationships among various variables and their structure.
- You can also make animated and interactive graphs.



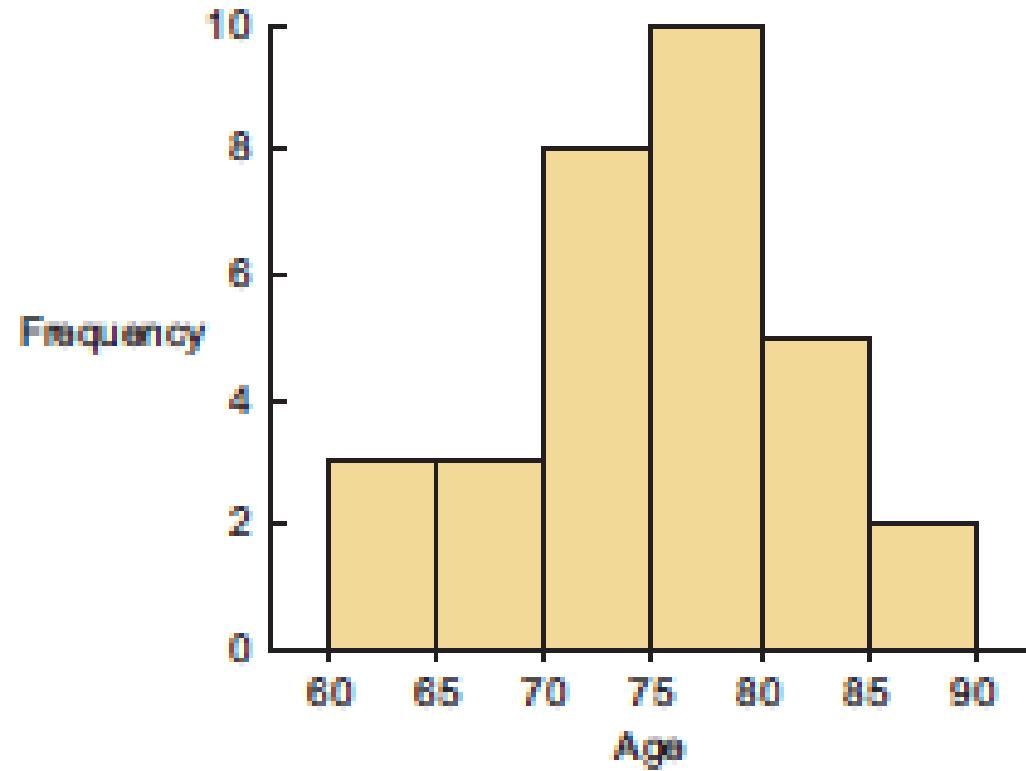
Exploratory Data Analysis (EDA)

- You can also overlay several plots.
- In the example below, we have combined several graphs into a Pareto chart or 80:20 diagram.
- The Pareto diagram below is a combination of the values and cumulative distribution.
- We can find out from the diagram below that the first 50% of the countries contain approximately 80% of the total amount.
- If this data represents sales of a multi-national company, we can conclude that 80% of the sales is from 50% of the countries.



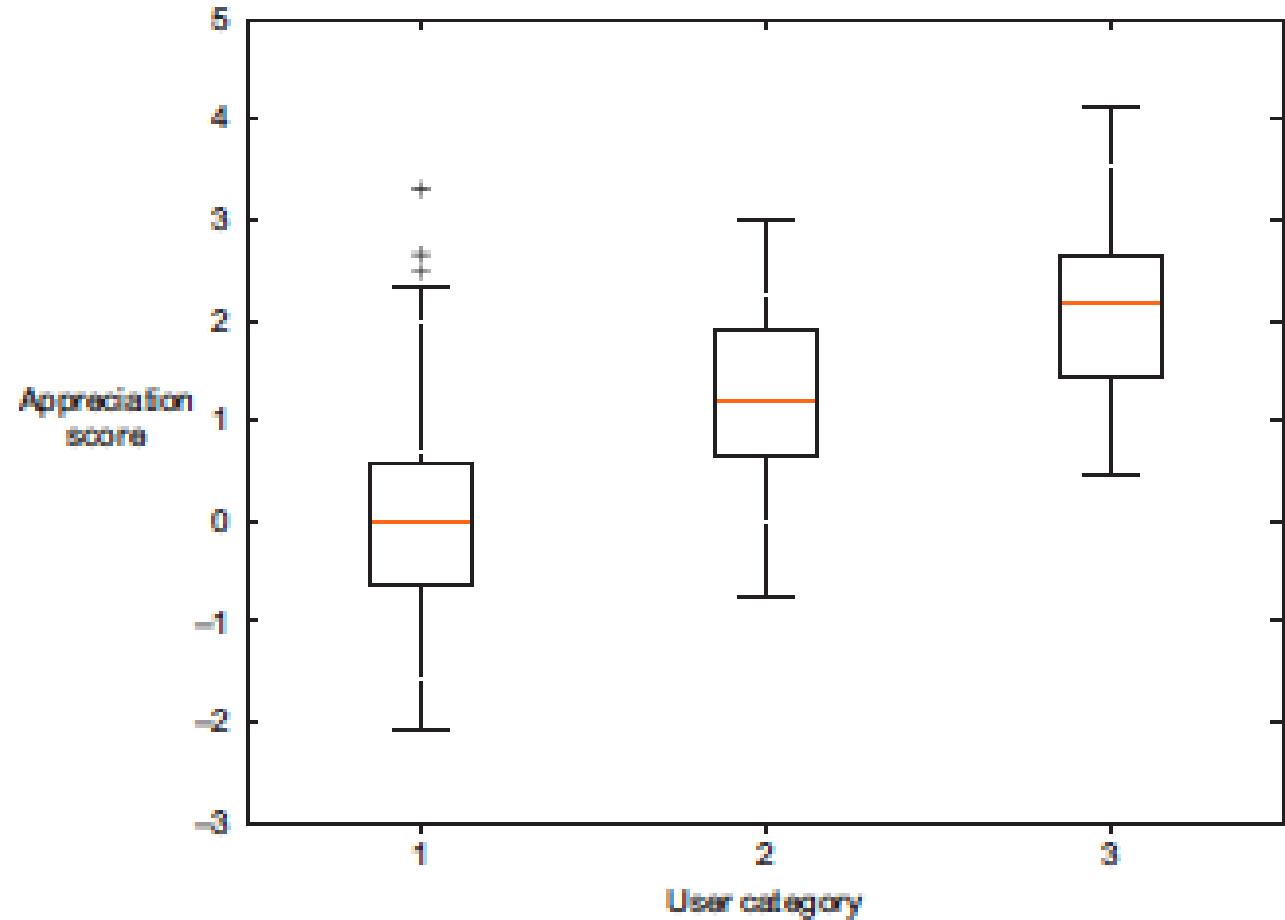
Exploratory Data Analysis (EDA)

- Boxplot and histogram are one of the most important graphs.
- In a histogram a variable is cut into discrete categories and the number of occurrences in each category are summed up and shown in the graph.



Exploratory Data Analysis (EDA)

- The boxplot, on the other hand, doesn't show how many observations are present but does offer an impression of the distribution within categories.
- It can show the maximum, minimum, median, and other characterizing measures at the same time.



EDA

- So far, we have covered the visual techniques only in practice we also use other techniques such as tabulation, clustering, and other modeling techniques in this step.
- Sometimes we also build simple models in this step.

Step 5: Data Modeling

Data Modeling

- By now, you have clean data in place and you have developed a good understanding of the data.
- Now you can focus on building models with the goal of making better predictions, classifying objects, or achieving any other research goal.
- This step is more focused than the exploratory data analysis step because you understand the data, requirement, and expected outcomes better.
- You will be using machine learning, statistical modeling, and data mining techniques while building the models.
- Data modeling is an iterative process.

Data Modeling

- Data modeling is an iterative process.
- You can either use classic statistical techniques or recent machine learning and deep learning models.
- You can also use both the techniques according to the nature of the data and research objectives.
- The data modeling consists of the following steps:
 - Selection of a modeling technique and variables to enter in the model
 - Execution of the model
 - Diagnosis and model comparison

Model and Variable Selection

- In this step, you select the variable that you want to include in your model and the data modeling techniques.
- With the help of exploratory data analysis, you should get the understanding the required variables for your model.
- For selecting the modeling technique, you can use your judgement based on data types, variables, and project objective.
- You need to consider the performance of the model and the project requirements. Other factors that you need to consider are
 - Whether the model would be moved to production. If yes, how the model would be implemented
 - How the maintenance of the model would be done
 - Whether the model is explainable

Model Execution

- Once you have decided upon your model, now you have to write a code to implement it.
- Python has libraries such as StatsModels or Scikit-learn.
- Using these packages, you can easily and quickly model several of the most popular techniques.
- As you can see in the following code, it's fairly easy to use linear regression with StatsModels or Scikit-learn.
- The following listing shows the execution of a linear prediction model.
- The linear regression model tries to fit a line while minimizing the distance to each point.

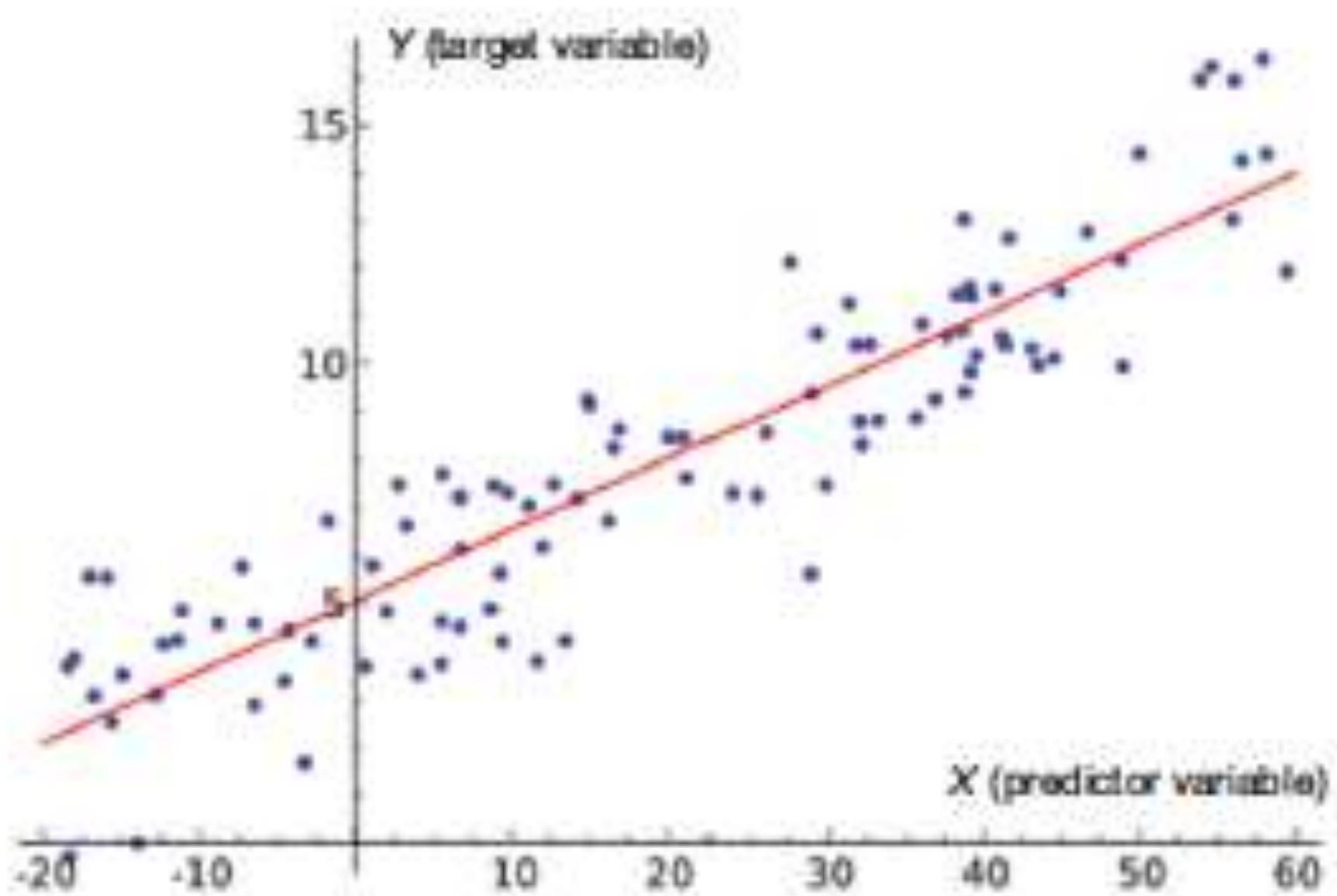
```
import statsmodels.api as sm  
import numpy as np  
predictors = np.random.random(1000).reshape(500, 2)  
target = predictors.dot(np.array([0.4, 0.6])) + np.random.random(500)  
lmRegModel = sm.OLS(target, predictors)  
result = lmRegModel.fit()  
result.summary()
```

Imports required Python modules.

Fits linear regression on data.

Shows model fit statistics.

Creates random data for predictors (x-values) and semi-random data for the target (y-values) of the model. We use predictors as input to create the target so we infer a correlation here.



Model Execution

- The `results.summary()` outputs the table as given below

Dep. Variable:	y	R-squared:	0.893
Model:	OLS	Adj. R-squared:	0.893
Method:	Least Squares	F-statistic:	2088.
Date:	Fri, 30 Oct 2015	Prob (F-statistic):	7.13e-243
Time:	12:44:31	Log-Likelihood:	-178.74
No. Observations:	500	AIC:	357.5
Df Residuals:	498	BIC:	365.9
Df Model:	2		
Covariance Type:	nonrobust		

Model fit: higher is better but too high is suspicious.

	coef	std err	t	P> t	95.0% Conf. Int.
x1	0.7658	0.040	19.130	0.000	0.687 0.844
x2	1.1252	0.039	28.603	0.000	1.048 1.202

p-value to show whether a predictor variable has a significant influence on the target. Lower is better and <0.05 is often considered "significant."

Omnibus:	34.269	Durbin-Watson:	1.943
Prob(Omnibus):	0.000	Jarque-Bera (JB):	13.480
Skew:	-0.125	Prob(JB):	0.00118
Kurtosis:	2.235	Cond. No.	2.51

Linear equation coefficients.
 $y = 0.7658x_1 + 1.1252x_2$.

Model Fit

- We use the R-squared or adjusted R-squared to analyze the model fit.
- R-squared explains the degree to which your input variables explain the variation of your output / predicted variable.
- So, if R-square is 0.8, it means 80% of the variation in the output variable is explained by the input variables.
- So, in simple terms, higher the R squared, the more variation is explained by your input variables and hence better is your model.

Model Fit

- However, the problem with R-squared is that it will either stay the same or increase with addition of more variables, even if they do not have any relationship with the output variables.
- This is where “Adjusted R square” comes to help.
- Adjusted R-square penalizes you for adding variables which do not improve your existing model.
- Hence, if you are building Linear regression on multiple variables, it is always suggested that you use Adjusted R-squared to judge goodness of model.
- In case you only have one input variable, R-square and Adjusted R squared would be exactly same.

Model Fit

- Typically, the more non-significant variables you add into the model, the gap in R-squared and Adjusted R-squared increases.
- A model gets complex when many variables (or features) are introduced.
- You don't need a complex model if a simple model is available, so the adjusted R-squared punishes you for overcomplicating.
- At any rate, 0.893 is high, and it should be because we created the data by keeping the relationship in mind.
- Rules of thumb exist, but for models in businesses, models above 0.85 are often considered good.
- If you want to win a competition you need in the high 90s. For research however, often very low model fits (<0.2 even) are found.

Predictor variables have a coefficient

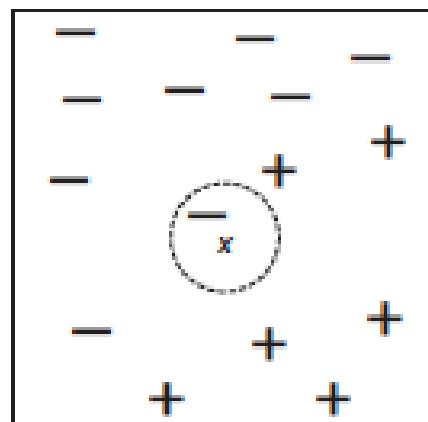
- Coefficients are the numbers by which the variables in an equation are multiplied.
- For example, in the equation $y = 0.7658x_1 + 1.1252x_2$, the variables x_1 and x_2 are multiplied by 0.7658 and 1.1252, respectively, so the coefficients are 0.7658 and 1.1252.
- The size and sign of a coefficient in an equation affect its graph. In a simple linear equation (contains only one x variable), the coefficient is the slope of the line.
- For a linear model this is easy to interpret. In our example if you increase x_1 by 1, it will change y by 0.7658 if x_2 is held constant.

Predictor Significance

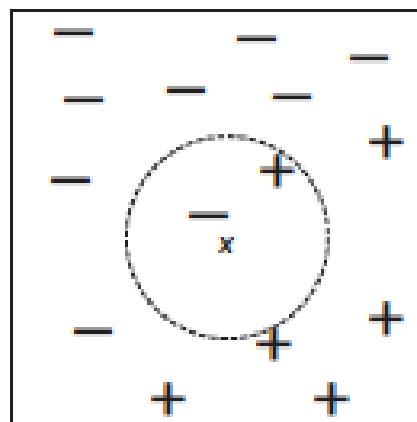
- Coefficients provide useful information about the relationship between predictor and response.
- But sometimes we need evidence to show that the influence is there.
- We need p-value for this.
- If the p-value of a coefficient is less than the chosen significance level, such as 0.05, the relationship between the predictor and the response is statistically significant.

K-Nearest Neighbour

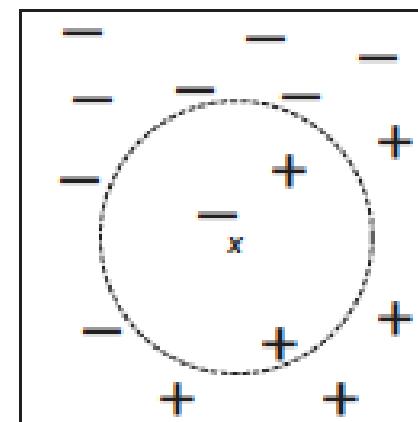
- We use linear regression to predict a value. But we use classification models to classify the observations.
- One of the best-known classification methods is k-nearest neighbors.
- The k-nearest neighbors model looks at labeled point that are nearby an unlabeled point and try to predict the label of the unlabeled point.



(a) 1-nearest neighbor



(b) 2-nearest neighbor



(c) 3-nearest neighbor

```
from sklearn import neighbors
predictors = np.random.random(1000).reshape(500, 2)
target = np.around(predictors.dot(np.array([0.4, 0.6])) +
                   np.random.random(500))
clf = neighbors.KNeighborsClassifier(n_neighbors=10)
knn = clf.fit(predictors, target)
knn.score(predictors, target)
```

Imports modules.

Creates random predictor data and semi-random target data based on predictor data.

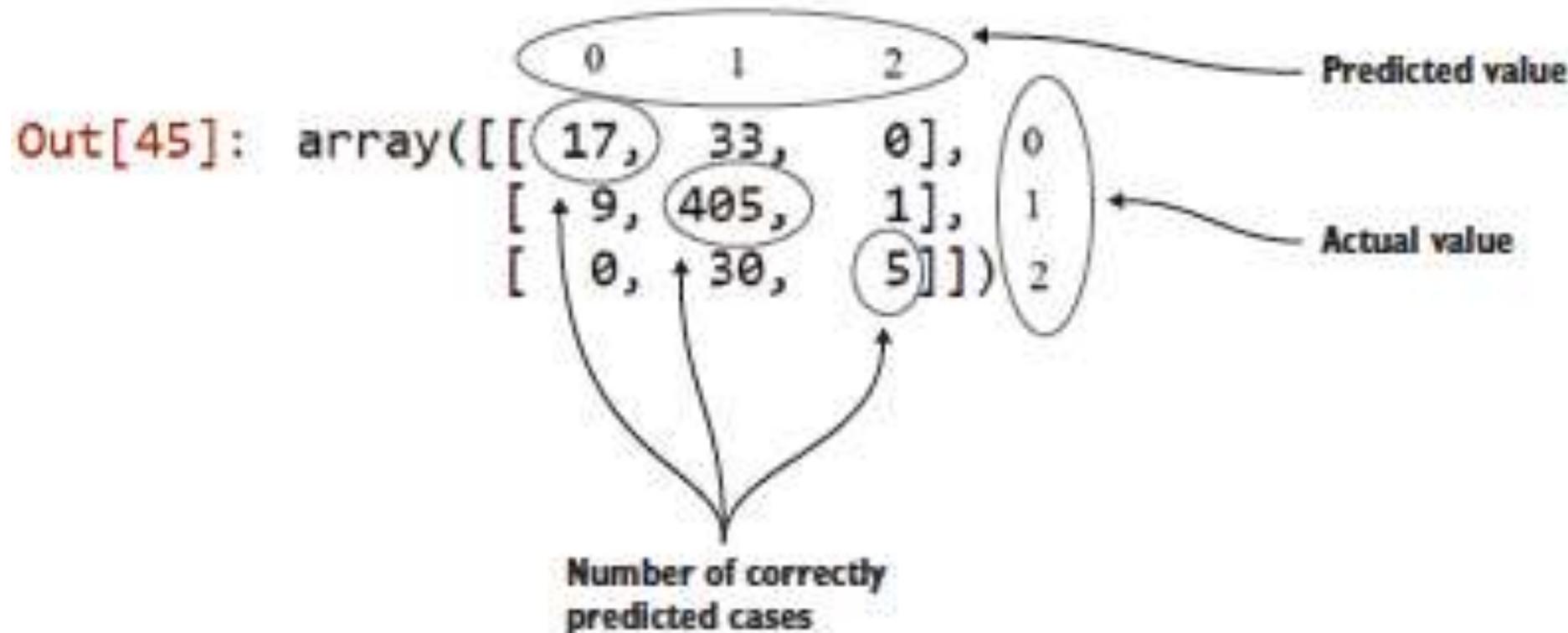
Fits 10-nearest neighbors model.

Gets model fit score: what percent of the classification was correct?

K-Nearest Neighbour

- In this case also, we construct random correlated data and hence 85% of the cases were correctly classified. `knn.score()` returns the model accuracy.
- However, to score the model, first we apply the model to predict the values
- `prediction = knn.predict(predictors)`
- Now we can use the prediction and compare it to the real thing using a confusion matrix.
- `metrics.confusion_matrix(target, prediction)`
- We get 3×3 matrix as given below .

```
In [45]: metrics.confusion_matrix(target,prediction)
```



Confusion Matrix

- **Confusion matrix:** it shows how many cases were correctly classified and incorrectly classified by comparing the prediction with the real values.
- The confusion matrix shows we have correctly predicted 17+405+5 cases

Model Diagnostics and Model Comparison

- In this process, you will build multiple models from which you will choose the best model based on multiple criteria.
- You can split the data between training and testing data.
- A fraction of data is used to train the model on the data set.
- The remaining data, that is unseen to the model, is then used to evaluate the performance of the model.
- Your model should work on the unseen data.
- You can use several error measures to evaluating and comparing the performance of the models.

Model Diagnostics and Model Comparison

- Mean Square Error (MSE) is one of the most commonly used measure. The formula for Mean Square Error is given below

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - Y_i)^2$$

- Mean square error is a simple measure.
- It checks for every prediction how far it was from the truth, square this error, and add up the error of every prediction.

Model Diagnostics and Model Comparison

- The figure below compares the performance of two models to predict the order size from the price.
- The first model is $size = 3 \times price$ and the second model is $size = 10$.
- To estimate the models, we use 800 randomly chosen observations out of 1,000 (or 80%), without showing the other 20% of data to the model.
- Once the model is trained, we predict the values for the other 20% of the variables based on those for which we already know the true value, and calculate the model error with an error measure.
- Then we choose the model with the lowest error. In this example we chose model 1 because it has the lowest total error.

	<i>n</i>	Size	Price	Predicted model 1	Predicted model 2	Error model 1	Error model 2
80% train	1	10	3				
	2	15	5				
	3	18	6				
	4	14	5				
					
	800	9	3				
	801	12	4	12	10	0	2
	802	13	4	12	10	1	3
	...						
	999	21	7	21	10	0	11
20% test	1000	10	4	12	10	-2	0
				Total	586.1	110.225	

Model Diagnostics and Model Comparison

- Many models make strong assumptions, such as independence of the inputs, and you have to verify that these assumptions are indeed met. This is called **model diagnostics**.

Step 6: Presenting findings and building applications on top of them

Presenting findings and building applications on top of them

- After you've successfully analyzed the data and built a well-performing model, you're ready to present your findings to the world. This is an exciting part; all your hours of hard work have paid off and you can explain what you found to the stakeholders.
- Sometimes people get so excited about your work that you'll need to repeat it over and over again because they value the predictions of your models or the insights that you produced. For this reason, you need to automate your models. This doesn't always mean that you have to redo all of your analysis all the time. Sometimes it's sufficient
- that you implement only the model scoring; other times you might build an application that automatically updates reports, Excel spreadsheets, or PowerPoint presentations.
- The last stage of the data science process is where your soft skills will be most useful, and yes, they're extremely important.

Thanks

Samatrix Consulting Pvt Ltd



Session 2
Unit No 5
Semester 1

?

Format the below table to make the data more readable?

Employee ID	Department	Employee Designation	Year of Joining	Salary before Tax (INR)	Income Tax (INR)	Salary after Tax (INR)
00001	Accounts	Asst. Manager	2000	568404.14	56840.4143376	
00002	Accounts	CFO	2002	1009845.3166925	100984.5317	
00003	Accounts	Executive	2007	890693.91	89069.39068	
00004	Admin	Senior Executive	2001	638137.7679	63813.77679	
00005	Admin	Associate	2003	562819.90540	56281.9905397	
00006	Admin	Manager	2006	267118.6132	26711.8613191	
00007	Admin	Executive	2008	225425.04	22542.50385	
00008	HR	Executive	1999	673606.608156	67360.66082	
00009	HR	Manager	2004	497269.0034	49726.90034	
00010	HR	Senior Executive	2005	2030657.718178	203065.7718	

2 Format the below table to make the data more readable?

Employee ID	Department	Employee Designation	Year of Joining	Salary before Tax (INR)	Income Tax (INR)	Salary after Tax (INR)
00001	Accounts	Asst. Manager	2000	568,404	56,840	511,564
00002	Accounts	CFO	2002	1,009,845	100,985	908,861
00003	Accounts	Executive	2007	890,694	89,069	801,625
00004	Admin	Senior Executive	2001	638,138	63,814	574,324
00005	Admin	Associate	2003	562,820	56,282	506,538
00006	Admin	Manager	2006	267,119	26,712	240,407
00007	Admin	Executive	2008	225,425	22,543	202,883
00008	Admin	Manager	2006	267,119	26,712	240,407
00009	Admin	Executive	2008	225,425	22,543	202,883
00010	HR	Executive	1999	673,607	67,361	606,246
00011	HR	Manager	2004	497,269	49,727	447,542
00012	HR	Executive	1999	673,607	67,361	606,246
00013	HR	Manager	2004	497,269	49,727	447,542
00014	HR	Senior Executive	2005	2,030,658	203,066	1,827,592

2 How to align RHS/LHS/MHS?

The screenshot shows a Microsoft Excel interface with the 'Home' tab selected. The ribbon menu includes File, Home, Insert, Page Layout, Formulas, Data, Review, View, Developer, Help, and Power Pivot. The 'Font' group on the ribbon has 'Calibri' and '11' selected. The 'Alignment' group has several icons: 'Horizontal Alignment' (centered), 'Vertical Alignment' (top), 'Wrap Text', 'Merge & Center', and 'Number' (General). A red circle highlights the 'Horizontal Alignment' icon. Red arrows point from this icon to the 'Align Left' and 'Align Right' buttons on the ribbon, with the text 'Align Left/Right' written in red.

Below the ribbon, a table is displayed with columns A through I. Row E contains values: 2000, 2008, 1999, 2004, 2005. Row F contains values: 225425.04, 673606.608156, 197269.0034, 2030657.718178. Row G contains values: 22542.50385, 67360.66082, 49726.90034, 203065.7718. Row H contains values: 202882.53469, 606245.9473, 447542.10305, 1827591.946.

Below the table, the text 'Conditional Formatting, Paste Special :-' is shown. A second table is displayed with columns Employee ID, Department, Employee Designation, Year of Joining, Salary before Tax (INR), Income Tax (INR), and Salary after Tax (INR). The data rows are:

Employee ID	Department	Employee Designation	Year of Joining	Salary before Tax (INR)	Income Tax (INR)	Salary after Tax (INR)
00001	Accounts	Asst. Manager	2000	568,404	56,840	511,564
00002	Accounts	CFO	2002	1,009,845	100,985	908,861
00003	Accounts	Executive	2007	890,694	89,069	801,625
00004	Admin	Senior Executive	2001	638,138	63,814	574,324
00005	Admin	Associate	2003	562,820	56,282	506,538
00006	Admin	Manager	2006	267,119	26,712	240,407
00007	Admin	Executive	2008	225,425	22,543	202,883
00008	Admin	Manager	2006	267,119	26,712	240,407
00009	Admin	Executive	2008	225,425	22,543	202,883
00010	HR	Executive	1999	673,607	67,361	606,246
00011	HR	Manager	2004	497,269	49,727	447,542
00012	HR	Executive	1999	673,607	67,361	606,246
00013	HR	Manager	2004	497,269	49,727	447,542
00014	HR	Senior Executive	2005	2,030,658	203,066	1,827,592

Red arrows and circles highlight specific cells in the second table, particularly in the first few rows, likely illustrating how to use conditional formatting or paste special features.

2 Why Comma insert?

Home Insert Page Layout Formulas Data Review View Developer Help Pov

Calibri 11 A A Wrap Text General

B I U A Merge & Center % ,

Font Alignment Number

Font Size: 11, Font Style: Normal, Alignment: Center, Number Format: General

203065.771817787

B	C	D	E	F	G	H
	Admin	Executive	2008	225425.04	22542.50385	20288
	HR	Executive	1999	673606.608156	67360.66082	6062
	HR	Manager	2004	497269.0034	49726.90034	44754
	HR	Senior Executive	2005	2030657.718178	203065.7718	1827

Comma style

litional Formatting, Paste Special :-

Employee ID	Department	Employee Designation	Year of Joining	Salary before Tax (INR)	Income Tax (INR)	Salary after Tax
	Accounts	Asst. Manager	2000	568,404	56,840	
	Accounts	CFO	2002	1,009,845	100,985	
	Accounts	Executive	2007	890,694	89,069	
	Admin	Senior Executive	2001	638,138	63,814	
	Admin	Associate	2003	562,820	56,282	
	Admin	Manager	2006	267,119	26,712	
	Admin	Executive	2008	225,425	22,543	
	Admin	Manager	2006	267,119	26,712	
	Admin	Executive	2008	225,425	22,543	
	HR	Executive	1999	673,607	67,361	
	HR	Manager	2004	497,269	49,727	
	HR	Executive	1999	673,607	67,361	
	HR	Manager	2004	497,269	49,727	
	HR	Senior Executive	2005	2,030,658	203,066	1,

2 How Decimal show more Precise Value?

The screenshot shows the Microsoft Excel ribbon with the 'Home' tab selected. In the 'Number' group of the ribbon, there is a dropdown menu set to 'General'. Below this, there is a button labeled '0.00' with arrows pointing left and right, indicating it's a spinner for decimal places. The main worksheet area displays a table of employee data. The last row of the table has the formula `=203065.771817787` in cell G10, and the result `203065.7718` is displayed in cell H10. The entire table is highlighted with a green border.

A	B	C	D	E	F	G	H
00007		Admin	Executive	2008	225425.04	22542.50385	202882.53469
00008		HR	Executive	1999	673606.608156	67360.66082	606245.9473
00009		HR	Manager	2004	497269.0034	49726.90034	447542.10305
00010		HR	Senior Executive	2005	2030657.718178	203065.7718	1827591.946

Conditional Formatting, Paste Special :-

Show decimal to precise value

Employee ID	Department	Employee Designation	Year of Joining	Salary before Tax (INR)	Income Tax (INR)	Salary after Tax (INR)
00001	Accounts	Asst. Manager	2000	568,404	56,840	511,564
00002	Accounts	CFO	2002	1,009,845	100,985	908,861
00003	Accounts	Executive	2007	890,694	89,069	801,625
00004	Admin	Senior Executive	2001	638,138	63,814	574,324
00005	Admin	Associate	2003	562,820	56,282	506,538
00006	Admin	Manager	2006	267,119	26,712	240,407
00007	Admin	Executive	2008	225,425	22,543	202,883
00008	Admin	Manager	2006	267,119	26,712	240,407
00009	Admin	Executive	2008	225,425	22,543	202,883
00010	HR	Executive	1999	673,607	67,361	606,246
00011	HR	Manager	2004	497,269	49,727	447,542
00012	HR	Executive	1999	673,607	67,361	606,246
00013	HR	Manager	2004	497,269	49,727	447,542
00014	HR	Senior Executive	2005	2,030,658	203,066	1,827,592

Inputs(x)

FUNCTIONS

Output-f(x)

34	67	68
67	To Sum a range of cell	
68		

SUM

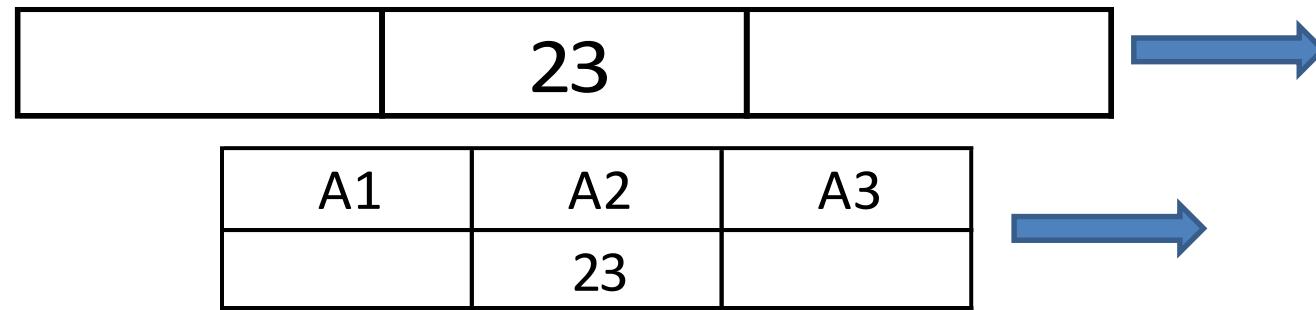
Every Command starting with “=” in Excel

=SUM(Number 1, Number 2, Number 3.....)

34	67	68
67	To count cells which contain numerical value	
68		

COUNT

=COUNT(Value1,Value 2, Value 3.....)



34	67	68
67	Average of cells which contain numbers	
68		

AVERAGE

=AVERAGE(Number 1, Number 2, Number 3.....)

34	67	68
67	To find the max value in range of cell	
68		

MAXIMUM

=MAX(Number 1, Number 2, Number 3.....)

34	67	68
67	To find the min value in range of cell	
68		

MINIMUM

=MIN(Number 1, Number 2, Number 3.....)

34	67	68
67	To count the values based on one criteria.	
68		

Countif

=Countif(Range,criteria...)

Criteria- Less than 30

Country	Covid Cases per min
India	43
Pakistan	12
Russia	30
America	37
China	4

=COUNTIF(B1:B6,"<30")

The screenshot shows a Microsoft Excel spreadsheet. The formula bar at the top contains the formula =COUNTIF(B1:B6,"<30"). The main area displays a table with columns 'Country' and 'Covid Cases per min'. The table has 6 rows, with the first row being the header. The data is as follows:

Country	Covid Cases per min
India	43
Pakistan	12
Russia	30
America	37
China	4

The formula bar also shows other cells like F8, X, Y, and Z. The status bar at the bottom right indicates "Criteria less than 30" and the number "2". The ribbon menu is visible at the top.

34	67	68
67	To count the values based on multiple criteria.	
68		

Countifs

=COUNTIFS(criteria_range1, criteria1, [criteria_range2, criteria2]...)

Criteria- India, Greater than 15

Country	Covid Cases per min
India	43
Pakistan	12
Russia	30
America	37
China	4
India	54

=COUNTIFS(A1:A7,"India",B1:B7,>15")

The screenshot shows a Microsoft Excel spreadsheet with a table of COVID-19 cases. The table has two columns: 'Country' and 'Covid Cases per min'. The data includes India (43), Pakistan (12), Russia (30), America (37), China (4), and India again (54). A formula bar at the top shows the formula =COUNTIFS(A1:A7,"India",B1:B7,>15"). Below the table, a large blue oval highlights the text 'Criteria India' and 'Greater than 15'. Inside this oval, the number '2' is written, indicating the count of countries that meet both criteria. A small green rectangular box is also visible near the bottom right of the oval.

A	B
Country	Covid Cases per min
India	43
Pakistan	12
Russia	30
America	37
China	4
India	54

34	67	68
67	To sum values based on one criteria.	
68		

SUMIF

=Sumif(Range,criteria,[Sum_range]...)

Criteria- Less than 30

Country	Covid Cases per min
India	43
Pakistan	12
Russia	30
America	37
China	4

=SUMIF(B1:B6,"<30")

The screenshot shows a Microsoft Excel spreadsheet. The table in the main area has columns A and B. Column A is labeled "Country" and column B is labeled "Covid Cases per min". The data rows are: India (43), Pakistan (12), Russia (30), America (37), and China (4). The formula bar at the top contains the formula =SUMIF(B1:B6,"<30"). The ribbon menu is visible, showing the "Font" and "Alignment" tabs. A blue arrow points from the text "Criteria less than 30" to the "<30" part of the formula. The number 16 is displayed in the bottom right corner.

A	B
1 Country	Covid Cases per min
2 India	43
3 Pakistan	12
4 Russia	30
5 America	37
6 China	4

Criteria less than 30

16

34	67	68
67	To sum values based on multiple criteria.	
68		

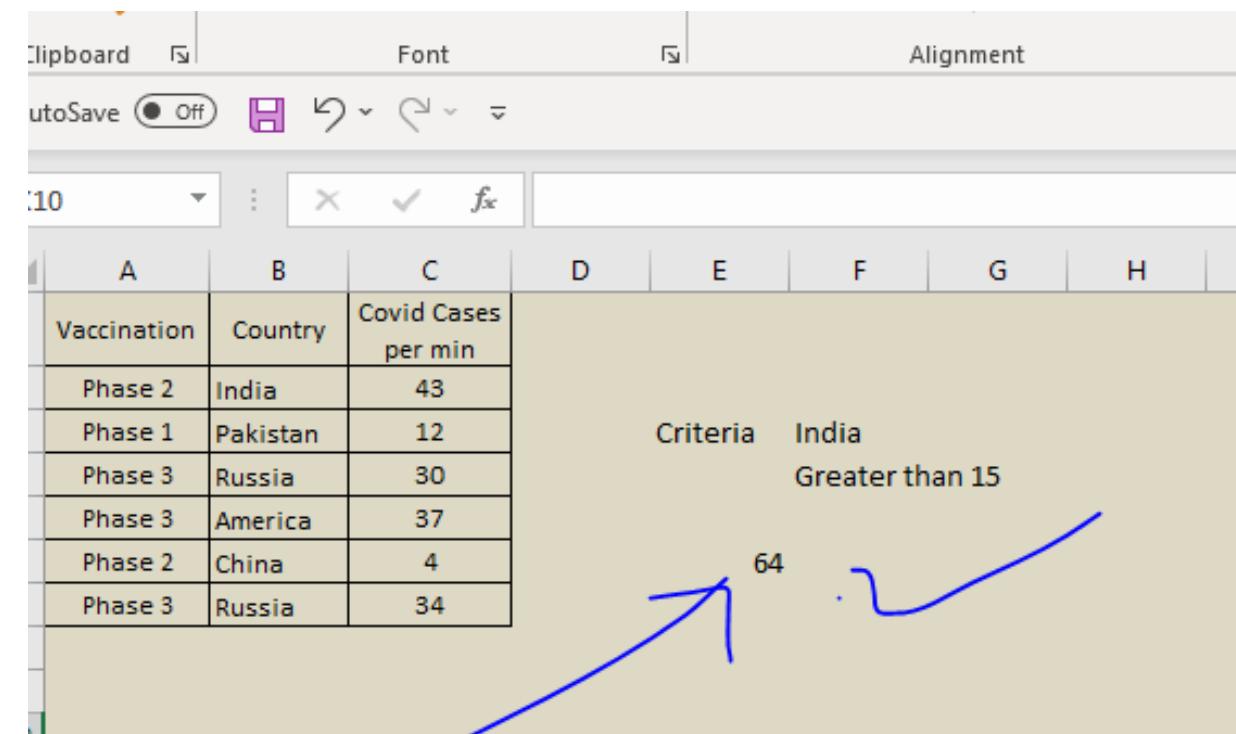
Sumifs

=Sumifs(sum_range,criteria_range1,...)

Criteria- India, Greater than 15

Vaccination	Country	Covid Cases per min
Phase 2	India	43
Phase 1	Pakistan	12
Phase 3	Russia	30
Phase 3	America	37
Phase 2	China	4
Phase 3	Russia	34

=SUMIFS(C1:C7,A1:A7,"Phase 3",B1:B7,"Russia")





3/4/5
?

1. Calculate the Sum, Count, Average, Maximum and Minimum in the space provided below. The formulae should be draggable horizontally.
Over-arching principle :Use one formula to populate the entire first table

RECORD FOR TECHNOLOGY & STEEL COMPANIES

Currency: INR

3)



Date	TATA STEEL	RELIANCE	JSW
1-May-12	295	1245	128
2-May-12	296	1350	129
3-May-12	298	1400	132
4-May-12	305	1100	138
5-May-12	318	950	145
6-May-12	310	1100	142
7-May-12	305	1250	141
8-May-12	300	1275	140

Descriptives			
Total			
Count			
Average			
Maximum			
Minimum			

4)

A	B	C
1	State	City
2	Rajasthan	Jaipur
3	Rajasthan	Kota
4	MPs	Bhopal
5	MPss	Gwalior
6	MPss	Bhopal
7	Haryana12	Gurgaon
8	Haryana123	Faridabad



- Counts the number of cells that are equal to 10?
- Counts the number of cells that are greater than or equal to 8?
- Counts the number of cells that are not equal to 8?
- Count the number of cells that are equal to 4 or 5?
- Counts the number of cells that contain exactly MPs?
- Counts the number of cells that contain exactly MP + 1 character?
- Counts the number of cells that contain exactly MP + x series of zero or more characters?
- Counts the number of cells that contain MP in any way?
- Counts the number of cells that contain text(C-Column)?
- Count the number of rows that contain MPs and Bhopal?

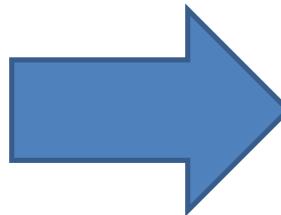
5)

	A	B	C
1	State	City	University
2	Rajasthan	Jaipur	10
3	Rajasthan	Kota	2
4	MPs	Bhopal	8
5	MPss	Gwalior	5
6	MPss	Bhopal	7
7	Haryana12	Gurgaon	8
8	Haryana123	Faridabad	4



1. Sum the value that are equal to 10?
2. Sum the value that are greater than or equal to 8?
3. Sum the value that are not equal to 8?
4. Sum the value that are equal to 4 or 5?
5. Sum the value that contain exactly MPs?
6. Sum the value that contain exactly MP+1 character?
7. Sum the value that contain exactly x series of zero or more characters +an?
8. Sum the value that contain Haryana12 and Gurgaon?

Magic Function





MANGO	Market Price	Qty(Kg)
Agra	150	2
Jhansi	140	3
Noida	170	4
Gurgaon	290	5
Mumbai	320	6
Kanpur	140	7



MANGO	Market Price	Qty(Kg)	Value1	Value2
Agra	150	2	300	300
Jhansi	140	3	420	280
Noida	170	4	680	340
Gurgaon	290	5	1450	580
Mumbai	320	6	1920	640
Kanpur	140	7	980	280

Functions:

DATE("year","month","day")

NOW()

TODAY()



TEXT TO COLUMNS and DATA CLEANING FUNCTIONS :

Excel - Module 1_v1.xlsx - Excel

The screenshot shows the Microsoft Excel ribbon with the "Data" tab selected. A blue circle highlights the "Text to Columns" button in the "Data Tools" group. A blue arrow points from the "Text to Columns" button down to the "Convert Text to Columns Wizard - Step 1 of 3" dialog box.

Convert Text to Columns Wizard - Step 1 of 3

The Text Wizard has determined that your data is Delimited.
If this is correct, choose Next, or choose the data type that best describes your data.

Original data type
Choose the file type that best describes your data:

Delimited - Characters such as commas or tabs separate each field.
 Fixed width - Fields are aligned in columns with spaces between each field.

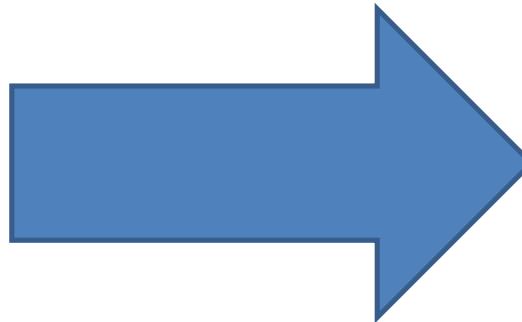
Preview of selected data:

21	Employee ID
22	HR_00001
23	Ac_00002
24	Ad_00003
25	Ma_00004

Cancel < Back **Next >** Finish

TEXT TO COLUMNS and DATA CLEANING FUNCTIONS :

Name
Bäte, OliverGE
Baumgartner, ThomasGE
Bender, MichaelBT
Beyer v. Morgenstern, IngoGE
Beyer v. Morgenstern, IngoGE
Goland, AnthonyDC
Kanarek, LarryDC
Lal, ShyamBT
Latoff, RobertCL
Mendonca, LennySF
Orr, GordonGC



First name	Last name

FUNCTIONS 2

Return the specified number of
characters from the start of a text string

LEFT

=LEFT(text,[Num_Chars])

Return the specified number of
characters from the end of a text string

RIGHT

=RIGHT(text,[Num_Chars])

Return the characters from the middle
of a text string, given a starting position
& length.

MID

=MID(text, start_num, [num_chars])

Return the number of characters in a text string.

LEN

=LEN(text)

Bäte, OliverGE

Convert the text string into proper case
that means, First letter is upper case &
other letters is lower case.

PROPER

=proper(text)

Return the characters into “UPPER CASE” / “lower case”

UPPER/LOWER

=upper(text)

=lower(text)

- 1) Mallika Gandhi
- 2) Mallika Gandhi
- 3) Mallika Gandhi

- Remove the extra spaces in data
- TRIM will remove extra spaces from text.
- It will leave only single spaces between words
- At the start or end of the text/word, no space character.



TRIM

=TRIM(text)

Character:

- At the start or end of the text/word
- If there is a space between words
- If there are extra spaces at the start or end of the text
- If there are extra spaces between words
- If there are extra spaces at the start or end of the text



Q-Please use Data cleaning functions to fill up the table below

Code	Actual	Name	First name	Last name
13217200000 Std. Cost Of Goods Sold	-	Bäte, OliverGE		
13217200250 Cost of Goods Sold Other Material	-	Baumgartner, ThomasGE		
13217200265 Cost of Goods Sold Laundry Overhead	-	Bender, MichaelBT		
13217200270 Cost of Goods Sold Refrig. Mat	-	Beyer v. Morgenstern, IngoGE		
13217000005 Gas & Elec	-	Beyer v. Morgenstern, IngoGE		
13217000006 Purchase Discount	-	Goland, AnthonyDC		
13217201160 Sub Advertising Maytag	-	Kanarek, LarryDC		
13217010010 Cycle Count Adj	-	Lal, ShyamBT		
13217010060 Scrap Steel Sales	-	Latoff, RobertCL		
13217112000 Rate Variance	-	Mendonca, LennySF		
13219110000 Depreciation Exp	44,870	Orr, GordonGC		
13219111000 Depreciation CL .	23,875	Ostrowski, KennethAT		
13218111000 Maintenance Materials	1	Pinkus, GarySF		
13218112000 Maintenance P&E	83	Rall, WilhelmGE		
13218223000 Cleaning And Sundries	414	Schrader, JürgenGE		
13218226000 Waste Removal	3,945	Simensen, Simen VierSC		

LOGICAL FUNCTIONS

- Conditions is met or not, check it.
- Check if value is True or false.



IF

=IF(logical test, [value_if_true], [value_if_false])

- Check if all conditions are true.



AND

=AND(logical value1, logical value2....))

- Check if any of the conditions are true.



OR

=OR(logical1, logical2....))



Q)-Mr. Sherlock Holmes, a first year MBA student, has just received his results of his semester, which is shown in the table below

Subject name	Score obtained by Sherlock Holmes	Maximum score
Module 1	70	90
Module 2	80	100
Module 3	55	80
Module 4	84	95
Module 5	79	100
Module 6	64	100
Module 7	88	90
Module 8	70	100
Module 9	91	100



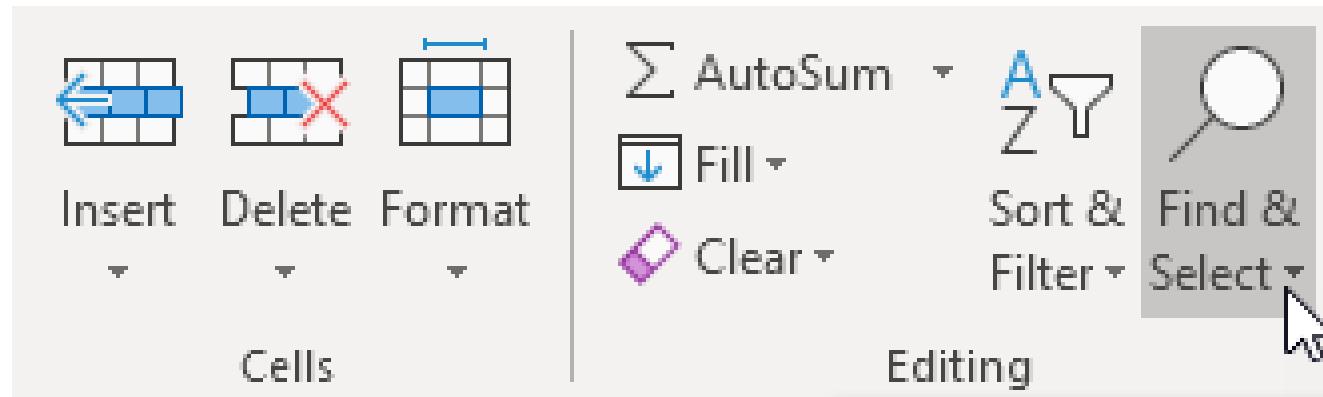
Unable to decide whether his scores are good or not, he asks his senior, Mr James Bond to help him out. Bond takes out a high-tech, fancy digital device (called the iScore) that asks a set of questions and based on the answer, provides a text message as shown in the table below. What would be the response of the iScore when Sherlock Holmes enters his scores ?

Unable to decide whether his scores are good or not, he asks his senior, Mr James Bond to help him out. Bond takes out a high-tech, fancy digital device (called the iScore) that asks a set of questions and based on the answer, provides a text message as shown in the table below. What would be the response of the iScore when Sherlock Holmes enters his scores ?

Criteria	Message displayed if criteria met	Message displayed if criteria not met	What should be the response based on Sherlock Holmes score?
Score in module 3 is greater than 75	"Well done"	"NA"	
Score in module 7 is greater than 80 and maximum score is less than 100	"Great success!!!"	"NA"	
Score in both module 4 and module 5 is greater than 80	"Super"	"NA"	
Score in either module 1 or module 2 is less than 75	"Unimpressive"	"NA"	
Score in both module 5 and module 6 is less than 70	"You're doomed!!!"	"NA"	

Find & Select Function

Find & Select Function



Find & Select Function

The screenshot shows a Microsoft Excel spreadsheet with the following characteristics:

- Clipboard:** Cut, Copy, Format Painter.
- Font:** Calibri, 11pt, Bold (B), Italic (I), Underline (U).
- Alignment:** Wrap Text, Merge & Center.
- Number:** General, %, , ., .00, .00.
- Styles:** Conditional Formatting, Table Styles, Cell Styles.
- Cells:** Insert, Delete, Format.
- Editing:** AutoSum, Fill, Clear, Sort & Filter, Find & Select.

The spreadsheet contains a single row of data from A1 to T1, with the formula bar showing "Ajay". The data consists of 11 columns of names:

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
Ajay																			
Raju																			
Mallika																			
Ajay																			
last																			
mail																			
Ajay																			
sanjay																			
raju																			
raha																			
haskdjl																			

Find & Select Function

The screenshot shows a Microsoft Excel spreadsheet with a table of names. The table has columns labeled A through T and rows labeled 1 through 11. The first row contains the names 'Ajay', 'Raju', 'Mallika', 'Ajay', 'last', 'mail', 'Ajay', 'sanjay', 'raju', 'raha', and 'haskdjl'. The second row contains 'Ajay', 'Raju', 'Mallika', 'Ajay', 'last', 'mail', 'Ajay', 'sanjay', 'raju', 'raha', and 'haskdjl'. This pattern repeats for the remaining rows. The cell containing 'Ajay' in the first row is highlighted with a green border. The formula bar at the top shows the text 'Ajay'. The ribbon menu is visible at the top, with the 'Editing' tab selected. A blue circle highlights the first row of the table, and a blue arrow points from the highlighted cell in the first row towards the formula bar.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
Ajay	Ajay																		
Raju																			
Mallika																			
Ajay																			
last																			
mail																			
Ajay																			
sanjay																			
raju																			
raha																			
haskdjl																			

Find & Select Function

The screenshot shows a Microsoft Excel spreadsheet with data in columns A through T. A green selection box highlights a range of cells from A1 to J10. The formula bar at the top contains the text "Ajay". The "Find and Replace" dialog box is open in the foreground, with the "Find" tab selected. The "Find what:" field contains "Ajay" and the "Replace with:" field contains "Sunny". The "Find Next" button is highlighted with a blue border. The "Replace All" button is located below the "Find" tab. The "Replace" and "Find All" buttons are also visible. The "Close" button is located at the bottom right of the dialog box.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
Ajay																			
Raju																			
Mallika																			
Ajay																			
last																			
mail																			
Ajay																			
sanjay																			
raju																			
raha																			
haskdj																			

Find & Select Function

Sunny Sunny Sunny Sunny Sunny Sunny Sunny Sunny Sunny

Raju Raju Raju Raju Raju Raju Raju Raju Raju

Mallika Mallika Mallika Mallika Mallika Mallika Mallika Mallika Mallika

Sunny Sunny Sunny Sunny Sunny Sunny Sunny Sunny

last last last last last last last last

mail mail mail mail mail mail mail mail

Sunny Sunny Sunny Sunny Sunny Sunny Sunny Sunny

sanjay sanjay sanjay sanjay sanjay sanjay sanjay sanjay

raju raju raju raju raju raju raju raju

raha raha raha raha raha raha raha raha

haskdjil haskdjil haskdjil haskdjil haskdjil haskdjil haskdjil haskdjil haskdjil

Microsoft Excel Find and Replace

All done. We made 27 replacements.

OK

Replace All Replace Find All Find Next Close

Enter certain values into a cell.

Data Validation Function

Sunny Pathak Share

F3 : X ✓ fx 100

A B C D E F G H I J K L M N O P Q R

1
2
3 Will the Covid Vaccination work well?
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22

Sheet1

The screenshot shows the Microsoft Excel ribbon with the "Data" tab selected. A blue arrow points from the "Data" tab down to the "Data Tools" section of the ribbon, which contains "Text to Columns", "Flash Fill", "Remove Duplicates", "Relationships", "Data Validation", and "What-If Analysis". The "Data Validation" button is highlighted with a red circle. A yellow Data Validation dialog box is open over the worksheet, showing the "Settings" tab. The validation criteria are set to "Allow: Whole number" and "Data: between". The "Minimum:" field is set to 10 and the "Maximum:" field is set to 100. The "OK" button is visible at the bottom right of the dialog.

C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
Will the Covid Vaccination work well?														

Data Validation

Settings Input Message Error Alert

Show input message when cell is selected

When cell is selected, show this input message:

Title:
Covid Patient

Input message:
Please tell me the %age of success of vaccine.

Clear All OK Cancel

From Text Sources Connections Query Recent Sources All Edit Links Advanced Columns Data Validation Analysis Sheet Subtotal Get External Data Get & Transform Connections Sort & Filter Data Tools Forecast Outline

3 : X ✓ fx 60

A B C D E F G H I J K L M N O P Q R

Will the Covid Vaccination work well?

Data Validation ? x

Settings Input Message Error Alert

Show error alert after invalid data is entered

When user enters invalid data, show this error alert:

Style: Stop Title: Not a valid %age

Error message: Are you Scared me?

X Clear All OK Cancel

Are you Scared me?

File Home Insert Page Layout Formulas Data Review View Developer Power Pivot Tell me what you want to do... Sunny Pathak Share

From Access From Web From Text From Other Sources Existing Connections New Query Show Queries From Table Recent Sources Get External Data Get & Transform Connections Refresh All Properties Edit Links Sort A Z Z A Sort Filter Clear Reapply Advanced Text to Columns Flash Fill Consolidate Remove Duplicates Relationships What-If Analysis Forecast Sheet Subtotal Group Ungroup Subtotal Outline

A B C D E F G H I J K L M N O P Q R

1
2
3 Will the Covid Vaccination work well? 61
4
5
6 Covid Patient Please tell me the Not a valid %age X
7 Are you Scared me? X Retry Cancel Help

61

Covid Patient
Please tell me the

Not a valid %age

Are you Scared me?

Retry Cancel Help

From Web From Other Sources Existing Connections New Query Recent Sources Refresh All Properties Edit Links Sort Filter Reapply Advanced Text to Columns Remove Duplicates Relationships What-If Analysis Forecast Sheet Ungroup Subtotal Get External Data Get & Transform Connections Sort & Filter Data Tools Forecast Outline

A B C D E F G H I J K L M N O P Q R

1
2
3 Will the Covid Vaccination work well? 70 Covid Patient Please tell me the Not a valid %age X Thank God? Retry Cancel Help

Sheet1 +

When you think that Customer select an item from the list we use drop down function.

Drop Down Function

File Home Insert Page Layout Formulas Data Review View Developer Power Pivot Tell me what you want to do... Sunny Pathak Share

From Access From Web From Text From Other Sources Existing Connections New Query Recent Sources Refresh All Connections Get External Data Get & Transform Sort & Filter Data Tools Forecast Outline

Connections Properties Edit Links Sort Filter Advanced Text to Columns Flash Fill Consolidate Remove Duplicates Relationships What-If Analysis Forecast Sheet Relationships Data Validation

I19

A B C D E F G H I J K L M N O P Q R

1
2
3 Will the Covid Vaccination work well? 60
4
5
6 City
7 Bhopal
8 Jaipur
9 Agra
10 Jhansi
11 Noida
12 Gurgaon
13 Jodhpur
14 Dehradun
15 Delhi
16 Mathura
17
18
19
20
21
22

Sheet1

100%

The screenshot shows a Microsoft Excel spreadsheet titled "Book1 - Excel (Product Activation Failed)". The "Data" tab is selected. In cell E6, the text "City" is highlighted with a blue circle. A vertical blue line extends from this circle down to a green rectangular box in cell I19. A horizontal blue line extends from the right side of the circle to the right side of the green box. A blue checkmark is drawn at the bottom left corner of the green box. The formula bar shows the address "I19". The column headers A through R are visible, along with row numbers 1 through 22. The cell I19 contains a blank green box.

File Home Insert Page Layout Formulas Data Review View Developer Power Pivot Tell me what you want to do... Sunny Pathak Share

From Access From Web From Text From Other Sources Get External Data Existing Connections New Query Recent Sources Refresh All Get & Transform Connections Sort & Filter Text to Columns Flash Fill Consolidate Remove Duplicates Relationships What-If Analysis Forecast Sheet Subtotal Group Ungroup Forecast Outline

G8 : X ✓ fx Jaipur

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1																		
2																		
3																		
4																		
5																		
6																		
7																		
8																		
9																		
10																		
11																		
12																		
13																		
14																		
15																		
16																		
17																		
18																		
19																		
20																		
21																		
22																		

Will the Covid Vaccination work well?

City

Bhopal

Jaipur

Agra

Jhansi

Noida

Gurgaon

Jodhpur

Dehradun

Delhi

Mathura

Data Validation

Settings Input Message Error Alert

Validation criteria

Allow: List ✓ Ignore blank ✓ In-cell dropdown

Data: between

Source: =\\$E\$7:\\$E\$16

Apply these changes to all other cells with the same settings

Clear All OK Cancel

Sheet1 +

Ready

100%

BOOK1 - Excel (Product Activation Failed)

Sunny Pathak Share

File Home Insert Page Layout Formulas Data Review View Developer Power Pivot Tell me what you want to do...

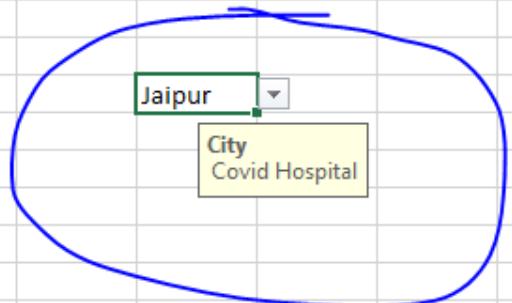
From Access From Web From Text From Other Sources Existing Connections New Query Recent Sources Refresh All Properties Edit Links Sort A-Z Sort Z-A Filter Clear Reapply Advanced Text to Columns Flash Fill Consolidate Remove Duplicates Relationships What-If Analysis Forecast Sheet Group Ungroup Subtotal Outline

G8 : X ✓ fx Jaipur

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1																		
2																		
3					Will the Covid Vaccination work well?		60											
4																		
5																		
6					City													
7					Bhopal													
8					Jaipur													
9					Agra													
10					Jhansi													
11					Noida													
12					Gurgaon													
13					Jodhpur													
14					Dehradun													
15					Delhi													
16					Mathura													
17																		
18																		
19																		
20																		
21																		
22																		

Jaipur

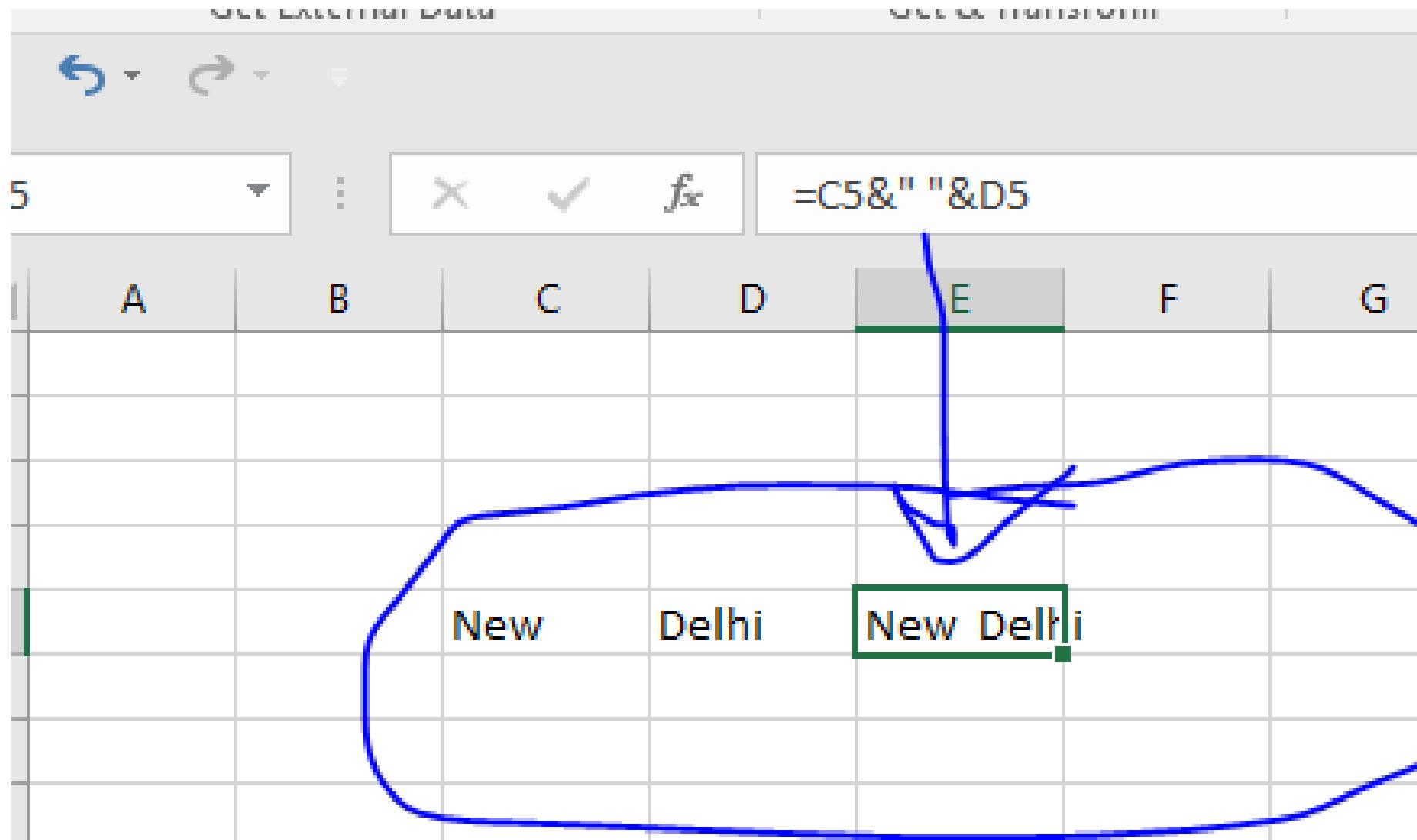
City Covid Hospital



Text Function

Join String

To join strings, use the & operator.



Compare Text Exact Function

Case Sensitive

Case Insensitive

The screenshot shows an Excel spreadsheet with data in columns C and D. The formula bar displays the formula `=EXACT(C4,D4)`. A blue oval highlights this formula. A green box highlights the range F3:F10, which contains the results of the EXACT function for each row. The results are:

	C	D	F
1			
2			
3	Mallika Gandhi	mallika Gandhi	FALSE
4	Navneet Tiwari	NavnEet Tiwari	FALSE
5	Sunny Pathak	Sunny pathak	FALSE
6	Vishal Jain	Vishal Jain	TRUE
7	Bhavya Sharma	Bhavya ShArma	FALSE
8	Chandan Garg	Chandan Garg	TRUE
9	Sanchit Sharma	Sanchit sharma	FALSE
10	Deepika Tiwari	Deepika Tiwari	TRUE
11			
12			
13			
14			
15			
16			

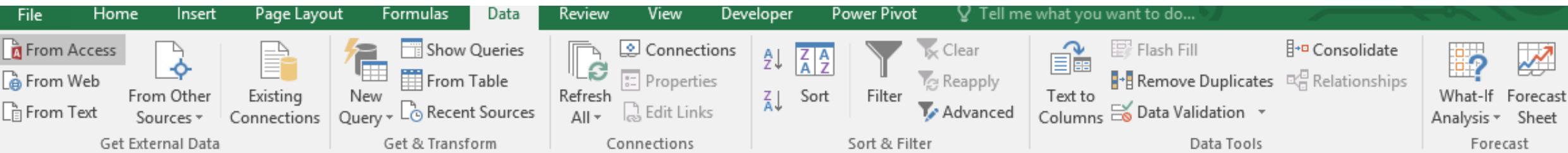
The VLOOKUP (Vertical lookup) function looks for a value in the leftmost column of a table, and then returns a value in the same row from another column you specify.

Vlook-up Function

	A	B	C	D	E	F	G	H	I	J	K
1	Employee Id	Employee Name									
2	S001										
3	S002										
4	S003										
5	S004										
6	S002										
7	S003										
8	S004										
9	S003										
10	S004										
11	S001										
12	S002										
13	S001										
14											
15											
16											
17											
18											
19											
..											

A screenshot of a Microsoft Excel spreadsheet. The formula bar at the top shows the formula =VLOOKUP(A2,\$F\$3:\$H\$6,3,0). The main area contains two tables. The first table (rows 1-6) has columns A (Employee Id) and B (Employee Name). The second table (rows 3-6) has columns E (Employee Id), F (Department), and G (Employee Name). A blue oval highlights the formula in the formula bar, and a blue arrow points from the formula to the lookup value in cell A2.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Employee Id	Employee Name												
2	S001	Navneet Tiwari												
3	S002				S001	Software	Navneet Tiwari							
4	S003				S002	IT	Mallika Gandhi							
5	S004				S003	Data Analyst	Bhavya Sharma							
6	S002				S004	Manager	Deepika Sharma							
7	S003													
8	S004													
9	S003													
10	S004													
11	S001													
12	S002													
13	S001													
14														
15														
16														
17														
18														
19														
20														



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Employee Id	Employee Name													
2	S001	Navneet Tiwari				Employee Id	Department	Employee Name							
3	S002	Mallika Gandhi				S001	Software	Navneet Tiwari							
4	S003	Bhavya Sharma				S002	IT	Mallika Gandhi							
5	S004	Deepika Sharma				S003	Data Analyst	Bhavya Sharma							
6	S002	Mallika Gandhi				S004	Manager	Deepika Sharma							
7	S003	Bhavya Sharma													
8	S004	Deepika Sharma													
9	S003	Bhavya Sharma													
10	S004	Deepika Sharma													
11	S001	Navneet Tiwari													
12	S002	Mallika Gandhi													
13	S001	Navneet Tiwari													
14															
15															
16															
17															
18															

DATA HANDLING PART 1

Category	Product	Sales	Ship to Country	Quarter
Condiments	French's	748	South Africa	Q1
Meat/Poultry	Shelton Poultry Produce	753	Portugal	Q2
Meat/Poultry	Banquet Chicken	755	China	Q1
Condiments	Mang Tomas	757	France	Q4
Meat/Poultry	Banquet Chicken	757	Australia	Q4
Grain/Cereals	Heart to Heart	772	Brazil	Q4
Condiments	Trappey's Hot Sauce	773	Portugal	Q2
Grain/Cereals	Sultana Bran	781	Portugal	Q3
Meat/Poultry	Toledano Wings	784	Argentina	Q3
Wines and Beer	Heiniken	785	Brazil	Q4
Beverages	Sjora	786	Ireland	Q3
Confections	Fantails	789	South Africa	Q2
Confections	Bassett's Sour Squirms	800	Ireland	Q1
Seafood	Chicken of the Sea	800	China	Q3
Grain/Cereals	Reese's Puffs	802	Portugal	Q1
Condiments	Homepride	803	France	Q4
Seafood	Something Fishy	804	South Africa	Q2
Wines and Beer	Ecco Domani	807	Japan	Q1
Dairy Products	Rose's Marmalade	808	France	Q1
Beverages	Bovril	811	Argentina	Q3
Grain/Cereals	Granola	814	Japan	Q4
Beverages	Horlicks	815	France	Q3

Insert a Pivot Table | Drag fields | Sort | Filter | Change
Summary Calculation | Two-dimensional Pivot Table

PIVOT TABLE

Pivot tables are one of Excel's most powerful features.
It allows you to extract the significance from a large, detailed data set.

Category	Product	Sales	Ship to Country	Quarter
Condiments	French's	748	South Africa	Q1
Meat/Poultry	Shelton Poultry Produce	753	Portugal	Q2
Meat/Poultry	Banquet Chicken	755	China	Q1
Condiments	Mang Tomas	757	France	Q4
Meat/Poultry	Banquet Chicken	757	Australia	Q4
Grain/Cereals	Heart to Heart	772	Brazil	Q4
Condiments	Trappey's Hot Sauce	773	Portugal	Q2
Grain/Cereals	Sultana Bran	781	Portugal	Q3
Meat/Poultry	Toledano Wings	784	Argentina	Q3
Wines and Beer	Heiniken	785	Brazil	Q4
Beverages	Sjora	786	Ireland	Q3
Confections	Fantails	789	South Africa	Q2
Confections	Bassett's Sour Squirms	800	Ireland	Q1
Seafood	Chicken of the Sea	800	China	Q3
Grain/Cereals	Reese's Puffs	802	Portugal	Q1
Condiments	Homepride	803	France	Q4
Seafood	Something Fishy	804	South Africa	Q2
Wines and Beer	Ecco Domani	807	Japan	Q1
Dairy Products	Rose's Marmalade	808	France	Q1
Beverages	Bovril	811	Argentina	Q3
Grain/Cereals	Granola	814	Japan	Q4
Beverages	Horlicks	815	France	Q3



- Q1. What are the Total Sales for each category**
- Q2. What are the Top 5 selling products in the Beverages category**
- Q3. How did each product category do in each quarter**
- Q4. What is the Average, Maximum and Minimum Sales for the Seafood category in the 3rd and 4th quarters?**
- Q5. Which are the Top 2 countries that buy the highest selling products in Q1?**