Data science: study of data to understand the world

Term	Definition	Video where the term is introduced
Algorithms	A set of step-by-step instructions to solve a problem or complete a task.	What is Data Science?
Model	A representation of the relationships and patterns found in data to make predictions or analyze complex systems retaining essential elements needed for analysis.	What is Data Science?
Outliers	When a data point or points occur significantly outside of most of the other data in a data set, potentially indicating anomalies, errors, or unique phenomena that could impact statistical analysis or modeling.	What is Data Science?
Quantitative analysis	A systematic approach using mathematical and statistical analysis is used to interpret numerical data.	Many Paths to Data Science
Structured data	Data is organized and formatted into a predictable schema, usually related tables with rows and columns.	What is Data Science?
Unstructured data	Unorganized data that lacks a predefined data model or organization makes it harder to analyze using traditional methods. This data type often includes text, images, videos, and other content that doesn't fit neatly into rows and columns like structured data.	What is Data Science?

Data format
CSV
.csv
Delimited text files, each value is separated by a delimiter
Comma or tab separated values.

XLSX

- Secure file and can't save malicious code
- Can open multiple sheets

.XML

- Readable by both human and machines
- Platform / Programming Language independent

- Designed to sending data by internet

Portable document format
PDF

- Can be viewed the same way on any device

JSON

- Text-based pen standard designed for transmitting data over the web.
- Compatible with a wide range of browsers
- Best tools for sharing data even audio and video
- API can return JSON

Term	Definition	Video where the term is introduced
Comma-separated values (CSV) / Tab-separated values (TSV)	Commonly used format for storing tabular data as plain text where either the comma or the tab separates each value.	Working on Different File Formats
Data file types	A computer file configuration is designed to store data in a specific way.	Working on Different File Formats
Data format	How data is encoded so it can be stored within a data file type.	Working on Different File Formats
Data visualization	A visual way, such as a graph, of representing data in a readily understandable way makes it easier to see trends in the data.	Data Science Topics and Algorithms
Delimited text file	A plain text file where a specific character separates the data values.	Working on Different File Formats
Extensible Markup Language (XML)	A language designed to structure, store, and enable data exchange between various technologies.	Working on Different File Formats

Hadoop	An open-source framework designed to store and process large datasets across clusters of computers.	What Makes Someone a Data Scientist
JavaScript Object Notation (JSON)	A data format compatible with various programming languages for two applications to exchange structured data.	Working on Different File Formats
Jupyter notebooks	A computational environment that allows users to create and share documents containing code, equations, visualizations, and explanatory text. See Python notebooks.	Data Science Skills & Big Data
Nearest neighbor	A machine learning algorithm that predicts a target variable based on its similarity to other values in the dataset.	Working on Different File Formats
Neural networks	A computational model used in deep learning that mimics the structure and functioning of the human brain’s neural pathways. It takes an input, processes it using previous learning, and produces an output.	A Day in the Life of a Data Scientist
Pandas	An open-source Python library that provides tools for working with structured data is often used for data manipulation and analysis.	Data Science Skills & Big Data
Python notebooks	Also known as a “Jupyter” notebook, this computational environment allows users to create and share documents containing code, equations, visualizations, and explanatory text.	Data Science Skills & Big Data
R	An open-source programming language used for statistical computing, data analysis, and data visualization.	Data Science Skills & Big Data
Recommendation engine	A computer program that analyzes user input, such as behaviors or preferences, and makes personalized recommendations based on that analysis.	A Day in the Life of a Data Scientist
Regression	A statistical model that shows a relationship between one or more predictor variables with a response variable.	Data Science Topics and Algorithms

Tabular data	Data that is organized into rows and columns.	A Day in the Life of a Data Scientist
XLSX	The Microsoft Excel spreadsheet file format.	Working on Different File Formats

Cloud computing

- It is a delivery of on-demand computing resources over the internet on a pay-for-use basis
- 5 characteristic
    - On-demand self-service
        - Access to processing, storage, and network
    - Board network access
        - Resources access via the internet
    - Resource pooling
        - Shared resources dynamically assigned
    - Rapid elasticity
        - Automatically scales resources access
    - Measured service
        - Only pay for what you use or reserve
- 3 deployment models
    - Public
    - Private
    - hybrid
- 3 service models
    - Infrastructure as a service (IaaS)
    - Platform as a Service (PaaS)
    - Software as a service (SaaS) / On demand software

The V’s of big data

- Velocity
    - The speed which data accumulates
- Volumn
    - The scale of the data
- Variety
    - Diversity of the data
- Veracity
    - Quality and origin of data
- Value
    - Ability and need to turn data into value

Apache Hadoop

- A collection of tools that provides distributed storage and processing of big data
- Distributed storage and processing

Apache Hive

- A data warehouse for data query and analysis
- Provides large data set management

Apache spark

- A distributed analytics framework for complex real- time data analytics
- Data processing engine

Hadoop Distributed file system(HDFS)

- Storage system for big data that runs on multiple commodity hardware connected through a network

Data mining

1. Goal set
    a. Identity key questions
2. Select data
    a. Identity data sources
3. Preprocess
    a. Clean the data
4. Transform
    a. Determine storage needs
5. Data mine
    a. Determine methods and analyze
6. Evaluate
    a. Assess outcomes, share results

Term	**Definition**	**Video where the term is introduced**
Analytics	The process of examining data to draw conclusions and make informed decisions is a fundamental aspect of data science, involving statistical analysis and data-driven insights.	Data Scientists at New York University
Big Data	Vast amounts of structured, semi-structured, and unstructured data are characterized by its volume, velocity, variety, and value, which, when analyzed, can provide competitive advantages and drive digital transformations.	How Big Data is Driving Digital Transformation

Big Data Cluster	A distributed computing environment comprising thousands or tens of thousands of interconnected computers that collectively store and process large datasets.	What is Hadoop?
Broad Network Access	The ability to access cloud resources via standard mechanisms and platforms such as mobile devices, laptops, and workstations over networks.	Introduction to Cloud
Chief Data Officer (CDO)	An emerging role responsible for overseeing data-related initiatives, governance, and strategies, ensuring that data plays a central role in digital transformation efforts.	How Big Data is Driving Digital Transformation
Chief Information Officer (CIO)	An executive is responsible for managing an organization's information technology and computer systems, contributing to technology-related aspects of digital transformation.	How Big Data is Driving Digital Transformation
Cloud Computing	The delivery of on-demand computing resources, including networks, servers, storage, applications, services, and data centers, over the Internet on a pay-for-use basis.	Introduction to Cloud
Cloud Deployment Models	Categories that indicate where cloud infrastructure resides, who manages it, and how cloud resources and services are made available to users, including public, private, and hybrid models.	Introduction to Cloud
Cloud Service Models	Models based on the layers of a computing stack, including Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS), represent different cloud computing offerings.	Introduction to Cloud
Commodity Hardware	Standard, off-the-shelf hardware components are used in a big data cluster, offering cost-effective solutions for storage and processing without relying on specialized hardware.	What is Hadoop?

Data Algorithms	Computational procedures and mathematical models used to process and analyze data made accessible in the cloud for data scientists to deploy on large datasets efficiently.	Cloud for Data Science
Data Replication	A strategy in which data is duplicated across multiple nodes in a cluster to ensure data durability and availability, reducing the risk of data loss due to hardware failures.	What is Hadoop?
Data Science	An interdisciplinary field that involves extracting insights and knowledge from data using various techniques, including programming, statistics, and analytical tools.	Data Scientists at New York University
Deep Learning	A subset of machine learning that involves artificial neural networks inspired by the human brain, capable of learning and making complex decisions from data on their own.	Data Scientists at New York University
Digital Change	The integration of digital technology into business processes and operations leads to improvements and innovations in how organizations operate and deliver value to customers.	How Big Data is Driving Digital Transformation
Digital Transformation	A strategic and cultural organizational change driven by data science, especially Big Data, to integrate digital technology across all areas of the organization, resulting in fundamental operational and value delivery changes.	How Big Data is Driving Digital Transformation
Distributed Data	The practice of dividing data into smaller chunks and distributing them across multiple computers within a cluster enables parallel processing for data analysis.	What is Hadoop?
Hadoop	A distributed storage and processing framework used for handling and analyzing large datasets, particularly well-suited for big data analytics and data science applications.	Data Scientists at New York University

Hadoop Distributed File System (HDFS)	A storage system within the Hadoop framework that partitions and distributes files across multiple nodes, facilitating parallel data access and fault tolerance.	What is Hadoop?
Infrastructure as a Service (IaaS)	A cloud service model that provides access to computing infrastructure, including servers, storage, and networking, without the need for users to manage or operate them.	Introduction to Cloud
Java-Based Framework	Hadoop is implemented in Java, an open-source, high-level programming language, providing the foundation for building distributed storage and processing solutions.	Big Data Processing Tools: Hadoop, HDFS, Hive, and Spark
Map Process	The initial step in Hadoop’s MapReduce programming model, where data is processed in parallel on individual cluster nodes, often used for data transformation tasks.	What is Hadoop?
Measured Service	A characteristic where users are billed for cloud resources based on their actual usage, with resource utilization transparently monitored, measured, and reported.	Introduction to Cloud
On-Demand Self-Service	The capability for users to access and provision cloud resources such as processing power, storage, and networking using simple interfaces without human interaction with service providers.	Introduction to Cloud
Rapid Elasticity	The ability to quickly scale cloud resources up or down based on demand, allowing users to access more resources when needed and release them when not in use.	Introduction to Cloud
Reduce Process	The second step in Hadoop's MapReduce model is where results from the mapping process are aggregated and processed further to produce the final output, typically used for analysis.	What is Hadoop?

Replication	The act of creating copies of data pieces within a big data cluster enhances fault tolerance and ensures data availability in case of hardware or node failures.	What is Hadoop?
Resource Pooling	A cloud characteristic where computing resources are shared and dynamically assigned to multiple consumers, promoting economies of scale and cost-efficiency.	Introduction to Cloud
Skills Network Labs (SN Labs)	Learning resources provided by IBM, including tools like Jupyter Notebooks and Spark clusters, are available to learners for cloud data science projects and skill development.	Cloud for Data Science
Spilling to Disk	A technique used in memory-constrained situations where data is temporarily written to disk storage when memory resources are exhausted, ensuring uninterrupted processing.	Big Data Processing Tools: Hadoop, HDFS, Hive, and Spark
STEM Classes	Science, Technology, Engineering, and Mathematics (STEM) courses typically taught in high schools prepare students for technical careers, including data science.	Data Scientists at New York University
Variety	The diversity of data types, including structured and unstructured data from various sources such as text, images, video, and more, posing data management challenges.	Foundations of Big Data
Velocity	The speed at which data accumulates and is generated, often in real-time or near-real-time, drives the need for rapid data processing and analytics.	Foundations of Big Data
Veracity	The quality and accuracy of data, ensuring that it conforms to facts and is consistent, complete, and free from ambiguity, impacts data reliability and trustworthiness.	Foundations of Big Data
Video Tracking System	A system used to capture and analyze video data from games, enabling in-depth analysis of player movements and game dynamics, contributing to data-driven decision-making in sports.	How Big Data is Driving Digital Transformation

Volume	The scale of data generated and stored is driven by increased data sources, higher-resolution sensors, and scalable infrastructure.	Foundations of Big Data
V's of Big Data	A set of characteristics common across Big Data definitions, including Velocity, Volume, Variety, Veracity, and Value, highlighting the rapid generation, scale, diversity, quality, and value of data.	Foundations of Big Data

Generative AI
Producing new similar data

- generative adversarial networks (GANs)
- Variational auto-encoders(VAEs)

Term	Definition	Video where the term is introduced
Artificial Neural Networks	Collections of small computing units (neurons) that process data and learn to make decisions over time.	Artificial Intelligence and Data Science
Bayesian Analysis	A statistical technique that uses Bayes' theorem to update probabilities based on new evidence.	Applications of Machine Learning
Business Insights	Accurate insights and reports generated by generative AI can be updated as data evolves, enhancing decision-making and uncovering hidden patterns.	Generative AI and Data Science
Cluster Analysis	The process of grouping similar data points together based on certain features or attributes.	Neural Networks and Deep Learning
Coding Automation	Using generative AI to automatically generate and test software code for constructing analytical models, freeing data scientists to focus on higher-level tasks.	Generative AI and Data Science

Data Mining	The process of automatically searching and analyzing data to discover patterns and insights that were previously unknown.	Artificial Intelligence and Data Science
Decision Trees	A type of machine learning algorithm used for decision-making by creating a tree-like structure of decisions.	Applications of Machine Learning
Deep Learning Models	Includes Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) that create new data instances by learning patterns from large datasets.	Generative AI and Data Science
Five V's of Big Data	Characteristics used to describe big data: Velocity, volume, variety, veracity, and value.	Neural Networks and Deep Learning
Generative AI	A subset of AI that focuses on creating new data, such as images, music, text, or code, rather than just analyzing existing data.	Generative AI and Data Science
Market Basket Analysis	Analyzing which goods tend to be bought together is often used for marketing insights.	Neural Networks and Deep Learning
Naive Bayes	A simple probabilistic classification algorithm based on Bayes' theorem.	Applications of Machine Learning
Natural Language Processing (NLP)	A field of AI that enables machines to understand, generate, and interact with human language, revolutionizing content creation and chatbots.	Generative AI and Data Science
Precision vs. Recall	Metrics are used to evaluate the performance of classification models.	Applications of Machine Learning
Predictive Analytics	Using machine learning techniques to predict future outcomes or events.	Neural Networks and Deep Learning
Synthetic Data	Artificially generated data with properties similar to real data, used by data scientists to augment their datasets and improve model training.	Generative AI and Data Science

Term	**Definition**	**Video where the term is introduced**

Arithmetic Models	Data science often uses Mathematical models to analyze data and predict outcomes.	Old problems, new data science solutions
Case study	In-depth analysis of an instance of a chosen subject to draw insights that inform theory, practice, or decision-making.	Old problems, new data science solutions
Data mining	Extracting information from raw data, such as making decisions, predicting trends, or understanding phenomena.	How Data Science is Saving Lives
Data Science	The field involves collecting, analyzing, and interpreting data to extract valuable insights and make informed decisions.	Old problems, new data science solutions
Data Strategy	A plan that outlines how an organization will collect, manage, and use data to achieve its goals.	Old problems, new data science solutions
Predictive analytics	Using data, algorithms, models, and machine learning to make predictions.	How Data Science is Saving Lives

Term	**Definition**	**Video where the term is introduced**
Adobe Spark	A suite of software tools that allow users to create and share visual content such as graphics, web pages, and videos.	Recruiting for Data Science
Analytical skills	The ability to analyze information systematically, logically, and organized.	Recruiting for Data Science
Chief information officer (CIO)	A business executive is responsible for an organization’s information technology systems and tech-related initiatives.	How Can Someone Become a Data Scientist
Computational thinking	Breaking problems into smaller parts and using algorithms, logic, and abstraction to develop solutions. Often used but not limited to computer science.	How Can Someone Become a Data Scientist
Data clusters	A group of similar, related data points distinct from other clusters.	How Can Someone Become a Data Scientist

Executive summary	Usually occurring at the beginning of a research paper, this section summarizes the important parts of the paper, including its key findings.	The Report Structure
High-performing computing (HPC) cluster	A computing technology that uses a system of networked computers designed to solve complex and computationally intensive problems in traditional environments.	How Can Someone Become a Data Scientist
Mathematical computing	The use of computers to calculate, simulate, and model mathematical problems.	Importance of Mathematics and Statistics for Data Science
Matrices	Plural for matric, matrices are a rectangular (tabular) array of numbers often used in mathematics, statistics, and computer science.	Recruiting for Data Science
Stata	A software package used for statistical analysis.	Recruiting for Data Science
Statistical distributions	A way of describing the likelihood of different outcomes based on a dataset. The “bell curve” is a common statistical distribution.	How Can Someone Become a Data Scientist
Structured Query Language (SQL)	A language used for managing data in a relational database.	Importance of Mathematics and Statistics for Data Science
TCP/IP network	A network that uses the TCP/IP protocol to communicate between connected devices on that network. The Internet uses TCP/IP.	How Can Someone Become a Data Scientist

Structured data

- Well organized in formats that can be stored in databases and lends itself to standard data analysis methods and tools

Semi-structured data

- Data that is somewhat organized and relies on meta tags for grouping and hierarchy

Unstructured data

- Data that is not conventionally organized in the form of rows and columns in a particular format

Term	Definition	Video where the term is introduced
Comma-separated values (CSVs)	Delimited text files where the delimiter is a comma. Used to store structured data.	Understanding Different Types of File Formats
Delimited text file formats	Text files are used to store data where each line or row has values separated by a delimiter. A delimiter is a sequence of one or more characters specifying the boundary between values. Common delimiters include comma, tab, colon, vertical bar, and space.	Understanding Different Types of File Formats
NoSQL databases	Databases are designed to store and manage unstructured data and provide analysis tools for examining this type of data.	Types of Data
Online Transaction Processing (OLTP) Systems	Systems that focus on handling business transactions and storing structured data.	Types of Data
Relational databases	Databases are designed to store structured data with well-defined schemas and support standard data analysis methods and tools.	Types of Data
Sensors	Devices such as Global Positioning Systems (GPS) and Radio Frequency Identification (RFID) tags generate structured data.	Types of Data
Spreadsheets	Software applications like Excel and Google Spreadsheets are used for organizing and analyzing structured data.	Types of Data
SQL Databases	Databases that use Structured Query Language (SQL) for defining, manipulating, and querying data in structured formats.	Types of Data
Tab-separated values (TSVs)	Delimited text files where the delimiter is a tab. Used as an alternative to CSV when literal commas are present in text data.	Understanding Different Types of File Formats

Databases

- Collection of data for input, storage, search, retrieval, and modification of data

DBMS (databases management program)\

- Set of programs for creating and maintaining the database and storing modifying and extracting from the database

Relational

RDBMS, data is organized into tabular format with rows and columns, following structure of schema

Use sql

Non-relational database

- Response to volume diversity and speed at which data is being generated today
- Built for speed flexibility and scale
- Can be stored in a schema-less form
- Widely used for processing

Advantages of relational database

- Joining table
- Reduce redundancy
- Easy to backup
- Easy to make change when using the data
- ACID
    - Aomicity
    - Consistency
    - Isolation
    - durability

Relational databases are well suited for

- Data warehouse
- IoT solutions

Limitation of RDBMS

- Does not work good for unstructured and semi-structured data
- Migration needs identical schemas

NoSQL(not only SQL)

- There's 4 common type
    - Key-value store
        - Collection of key and value pairs
    - Document based
        - Stores within a single document
        - Enables ad hoc queries

        - Preferred for medical records
    - Column based
        - Stored in cells grouped as columns of data instead of rows
        - A logical grouping of columns is referred to as a column family
    - Graph based
        - Graphical model
        - Good for connected data

Advantages of NoSQL

- Handle all type of data either structured or unstructured, or semi-structured
- Run as distributed system
- Cost-effective scale out

ETL

- Extract
- Transform
- Load

Term	Definition	Video where the term is introduced
ACID-compliance	Ensuring data accuracy and consistency through Atomicity, Consistency, Isolation, and Durability (ACID) in database transactions.	Relational Database Management System
Cloud-based Integration Platform as a Service (iPaaS)	Cloud-hosted integration platforms that offer integration services through virtual private clouds or hybrid cloud models, providing scalability and flexibility.	Data Integration Platforms
Column-based Database	A type of NoSQL database that organizes data in cells grouped as columns, often used for systems requiring high write request volume and storage of time-series or IoT data.	NoSQL
Data at rest	Data that is stored and not actively in motion, typically residing in a database or storage system for various purposes, including backup.	Considerations for Choice of Data Repository

Data integration	A discipline involving practices, architectural techniques, and tools that enable organizations to ingest, transform, combine, and provision data across various data types, used for purposes such as data consistency, master data management, data sharing, and data migration.	Data Integration Platforms
Data Lake	A data repository for storing large volumes of structured, semi-structured, and unstructured data in its native format, facilitating agile data exploration and analysis.	Data Marts, Data Lakes, ETL, and Data Pipelines
Data mart	A subset of a data warehouse designed for specific business functions or user communities, providing isolated security and performance for focused analytics.	Data Marts, Data Lakes, ETL, and Data Pipelines
Data pipeline	A comprehensive data movement process that covers the entire journey of data from source systems to destination systems, which includes data integration as a key component.	Data Integration Platforms
Data repository	A general term referring to data that has been collected, organized, and isolated for business operations or data analysis. It can include databases, data warehouses, and big data stores.	Data Collection and Organization
Data warehouse	A central repository that consolidates data from various sources through the Extract, Transform, and Load (ETL) process, making it accessible for analytics and business intelligence.	Data Collection and Organization
Document-based Database	A type of NoSQL database that stores each record and its associated data within a single document, allowing flexible indexing, ad hoc queries, and analytics over collections of documents.	NoSQL
ETL process	The Extract, Transform, and Load process for data integration involves extracting data from various sources, transforming it into a usable format, and loading it into a repository.	Data Marts, Data Lakes, ETL, and Data Pipelines

Graph-based Database	A type of NoSQL database that uses a graphical model to represent and store data, ideal for visualizing, analyzing, and discovering connections between interconnected data points.	NoSQL
Key-value store	A type of NoSQL database where data is stored as key-value pairs, with the key serving as a unique identifier and the value containing data, which can be simple or complex.	NoSQL
Portability	The capability of data integration tools to be used in various environments, including single-cloud, multi-cloud, or hybrid-cloud scenarios, provides flexibility in deployment options.	Data Integration Platforms
Pre-built connectors	Cataloged connectors and adapters that simplify connecting and building integration flows with diverse data sources like databases, flat files, social media, APIs, CRM, and ERP applications.	Data Integration Platforms
Relational databases (RDBMSes)	Databases that organize data into a tabular format with rows and columns, following a well-defined structure and schema.	Data Collection and Organization
Scalability	The ability of a data repository to grow and expand its capacity to handle increasing data volumes and workload demands over time.	Considerations for Choice of Data Repository
Schema	The predefined structure that describes the organization and format of data within a database, indicating the types of data allowed and their relationships.	Considerations for Choice of Data Repository
Streaming data	Data that is continuously generated and transmitted in real-time requires specialized handling and processing to capture and analyze.	Considerations for Choice of Data Repository
Use cases for relational databases	Applications such as Online Transaction Processing (OLTP), Data Warehouses (OLAP), and IoT solutions where relational databases excel.	Relational Database Management System
Vendor lock-in	A situation where a user becomes dependent on a specific vendor’s technologies and solutions, making it challenging to switch to other platforms.	Data Integration Platforms

Data management
- Collecting persistent and retrieving data securely and cost -effectively

Data integration and transformation
- ETL
- Transformation: process of transforming ht values, structures and format of data

Data visualization
- Graphical representation of data and info

Code Asset management
- Version control
- Collaboration

Data integration tools:
- Apache airflow
- Kube flow
- Spark SQL

Dta management tools
- mysql
- mongodb

Data visualization tools:
- Pixie dust
- Hue
- Kibana
- Apache superset

Model deployment
- Apache prediction IO
- Seldom
- TensorFlow serving

Model monitoring and assessment tools
- modelDB
- prometheus

Code asset management
- Git
- Github

Data asset management
- Apache atlas
- Egeria
- kylo

Commercial tools for data science

Commercial data integration tools:
- Informatica
- Talend
- Watson studio desktop

Commercial data visualization

- Tableau
- IBM cognos analytics

Model building tools for commercial

- SPSS
- SAS

MOdel deployment tools

- SPSS

Fully integrated visual tools and platforms

- Watson studio
- Azure machine learning
- H2O driverless AI

Python library:

1. scientific Computing
    a. Pandas
    b. numpy
2. Visualization
    a. Matplotlib
    b. seaborn
3. ML and DL
    a. Scikit-learn
    b. Keras
    c. Tensorflow
    d. pytorch

REST APIs

- Representational
- State
- Transfer
- APIs
- Allow you to communicate through internet

Kernel

- A computational engine that executes the code contained in a notebook file
- Launches when notebook opens

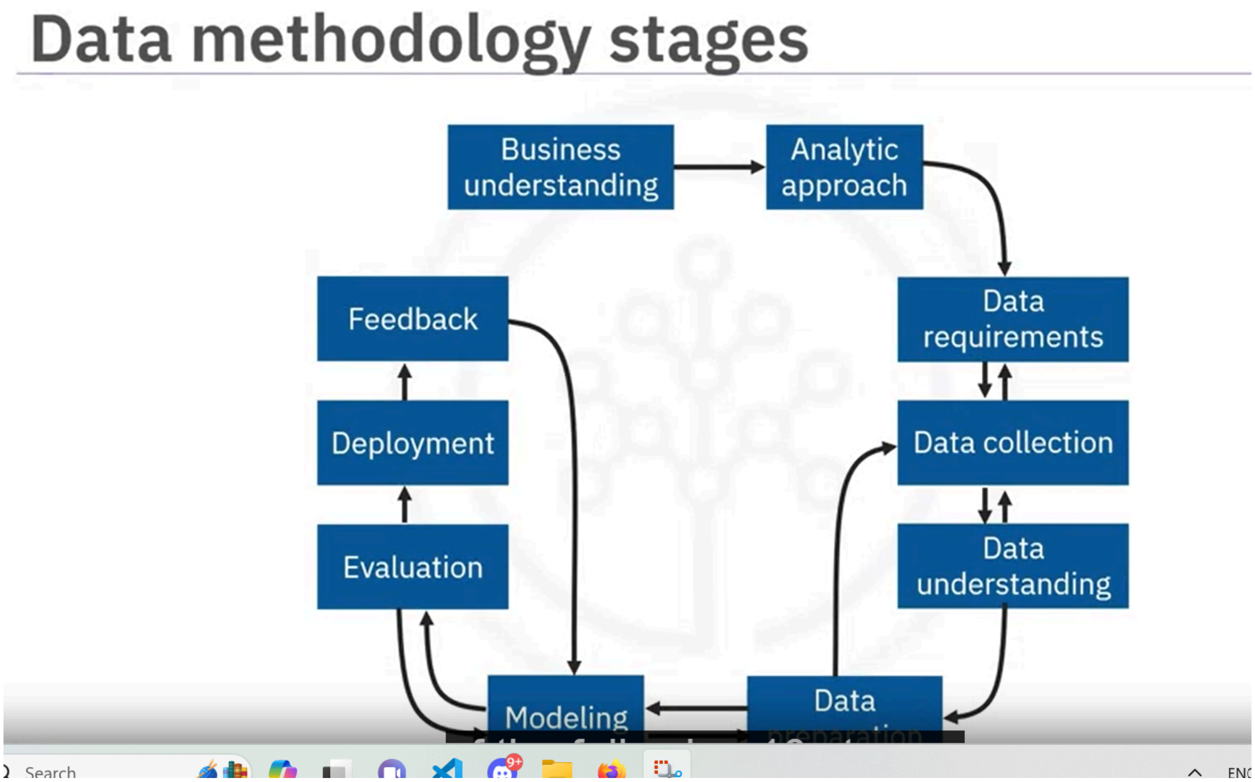
Jupyter architecture

2 process model

Kernel	client
Executes code	interface(browser)

```
User -> browser -> notebook server -> kernel
                              |
                         notebookFile
```

Anaconda



Data methodology stages

Business understanding

Analytic approach

Data requirements

Data collection

Data understanding

Data preparation

Modeling

Evaluation

Deployment

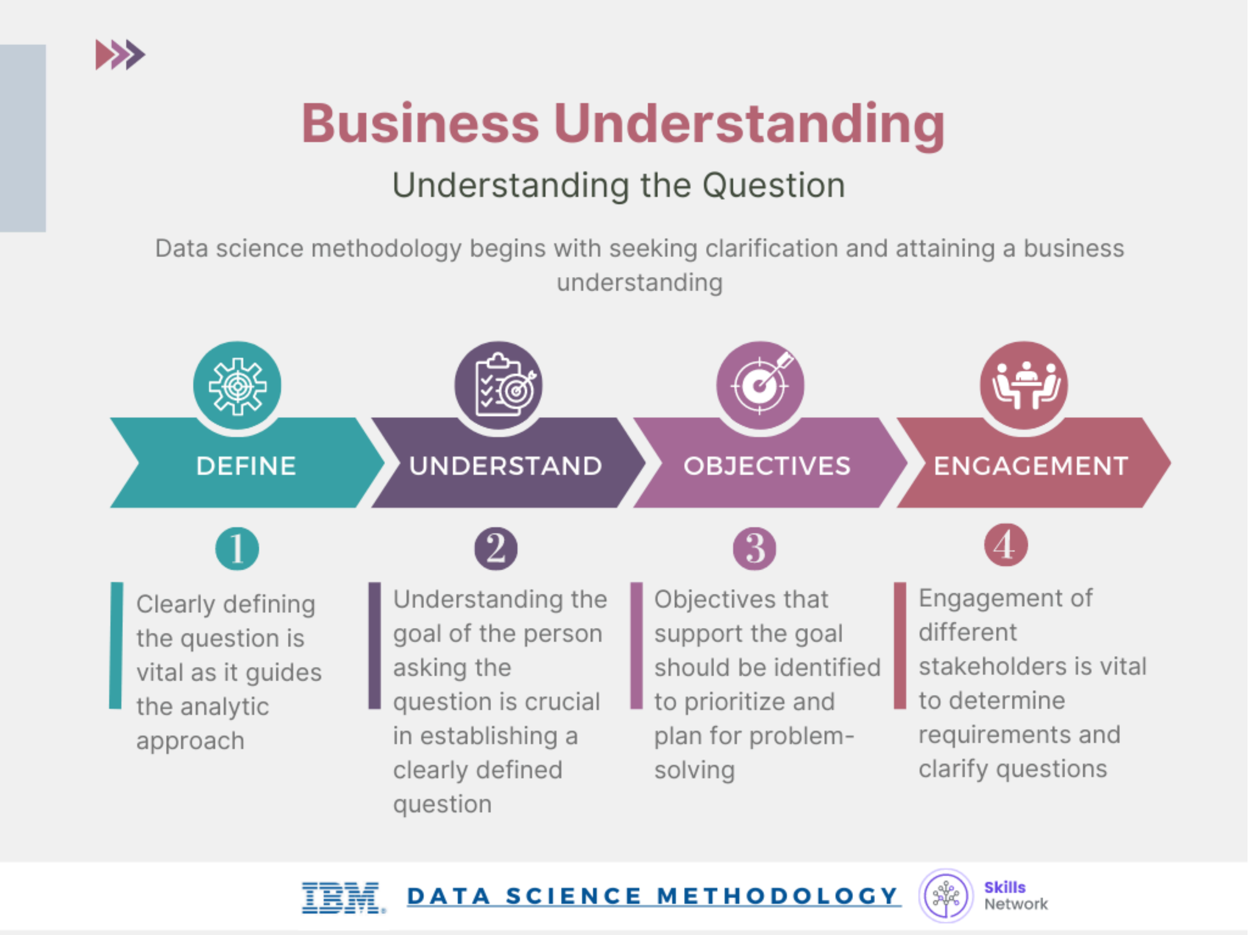
Feedback

Business understanding

- What is the problem that you are trying to solve

Analytic approach

- How can you use data to answer the question?



Business Understanding

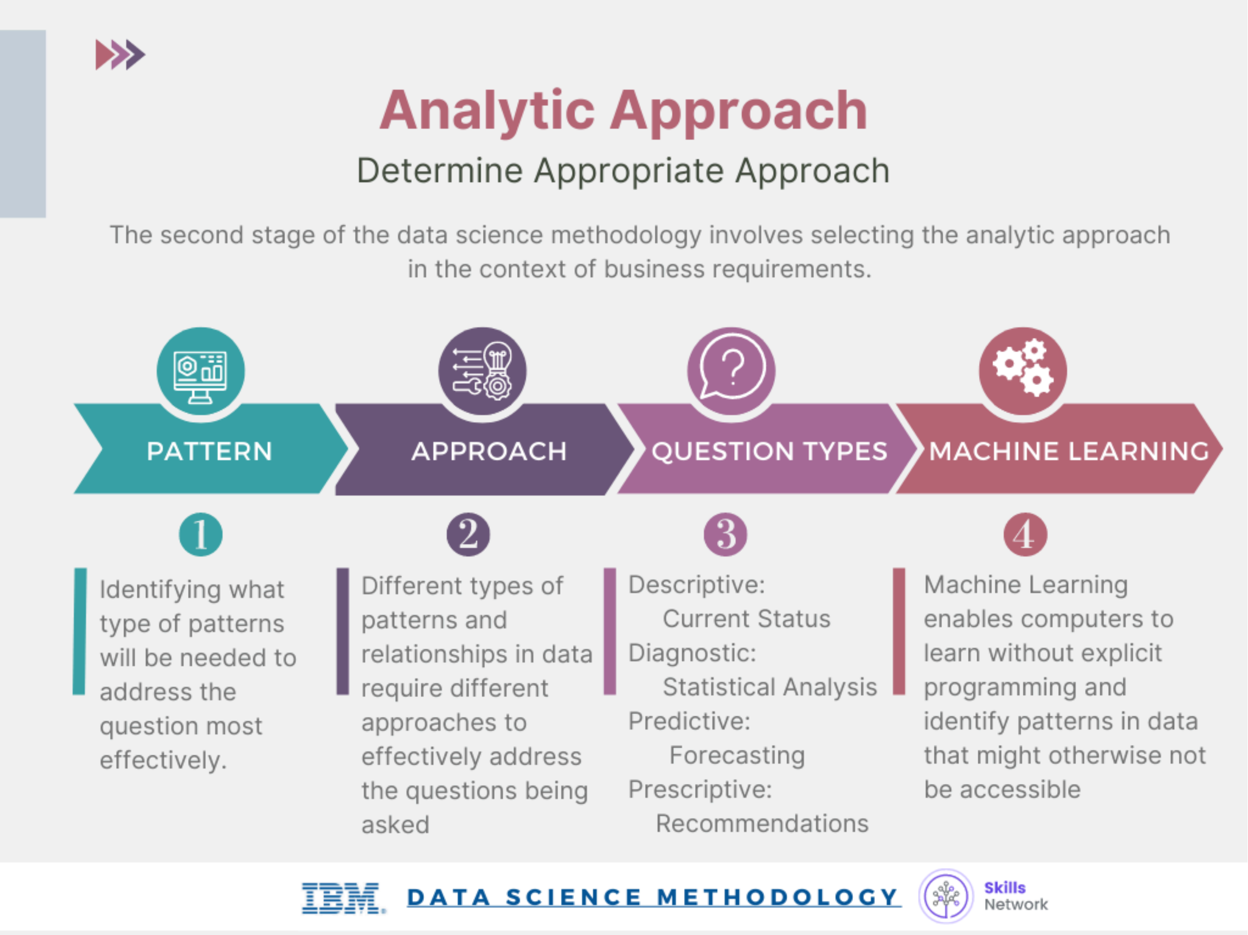
Understanding the Question

Data science methodology begins with seeking clarification and attaining a business understanding

1. DEFINE: Clearly defining the question is vital as it guides the analytic approach
2. UNDERSTAND: Understanding the goal of the person asking the question is crucial in establishing a clearly defined question
3. OBJECTIVES: Objectives that support the goal should be identified to prioritize and plan for problem-solving
4. ENGAGEMENT: Engagement of different stakeholders is vital to determine requirements and clarify questions

IBM DATA SCIENCE METHODOLOGY

Skills Network



Analytic Approach

Determine Appropriate Approach

The second stage of the data science methodology involves selecting the analytic approach in the context of business requirements.

PATTERN

1 Identifying what type of patterns will be needed to address the question most effectively.

APPROACH

2 Different types of patterns and relationships in data require different approaches to effectively address the questions being asked

QUESTION TYPES

3 Descriptive:
Current Status
Diagnostic:
Statistical Analysis
Predictive:
Forecasting
Prescriptive:
Recommendations

MACHINE LEARNING

4 Machine Learning enables computers to learn without explicit programming and identify patterns in data that might otherwise not be accessible

IBM DATA SCIENCE METHODOLOGY Skills Network

Term	Definition
Analytic Approach	The process of selecting the appropriate method or path to address a specific data science question or problem.
Analytics	The systematic analysis of data using statistical, mathematical, and computational techniques to uncover insights, patterns, and trends.
Business Understanding	The initial phase of data science methodology involves seeking clarification and understanding the goals, objectives, and requirements of a given task or problem.
Clustering Association	An approach used to learn about human behavior and identify patterns and associations in data.
Cohort	A group of individuals who share a common characteristic or experience is studied or analyzed as a unit.

Cohort study An observational study where a group of individuals with a specific characteristic or exposure is followed over time to determine the incidence of outcomes or the relationship between exposures and outcomes.

Congestive Heart Failure (CHF) A chronic condition in which the heart cannot pump enough blood to meet the body's needs, resulting in fluid buildup and symptoms such as shortness of breath and fatigue.

CRISP-DM Cross-Industry Standard Process for Data Mining is a widely used methodology for data mining and analytics projects encompassing six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

Data analysis The process of inspecting, cleaning, transforming, and modeling data to discover useful information, draw conclusions, and support decision-making.

Data cleansing The process of identifying and correcting or removing errors, inconsistencies, or inaccuracies in a dataset to improve its quality and reliability

Data science An interdisciplinary field that combines scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data.

Data science methodology A structured approach to solving business problems using data analysis and data-driven insights.

Data scientist A professional using scientific methods, algorithms, and tools to analyze data, extract insights, and develop models or solutions to complex business problems.

Data scientists Professionals with data science and analytics expertise who apply their skills to solve business problems.

Data-Driven Insights Insights derived from analyzing and interpreting data to inform decision-making

Decision tree A supervised machine learning algorithm that uses a tree-like structure of decisions and their possible consequences to make predictions or classify instances.

Decision Tree Classification Model A model that uses a tree-like structure to classify data based on conditions and thresholds provides predicted outcomes and associated probabilities.

Decision Tree Classifier A classification model that uses a decision tree to determine outcomes based on specific conditions and thresholds.

Decision-Tree Model A model used to review scenarios and identify relationships in data, such as the reasons for patient readmissions

Descriptive approach An approach used to show relationships and identify clusters of similar activities based on events and preferences

Descriptive modeling Modeling technique that focuses on describing and summarizing data, often through statistical analysis and visualization, without making predictions or inferences

Domain knowledge Expertise and understanding of a specific subject area or field, including its concepts, principles, and relevant data

Goals and objectives The sought-after outcomes and specific objectives that support the overall goal of the task or problem.

Iteration A single cycle or repetition of a process often involves refining or modifying a solution based on feedback or new information.

Iterative process A process that involves repeating a series of steps or actions to refine and improve a solution or analysis. Each iteration builds upon the previous one.

Leaf The final nodes of a decision tree where data is categorized into specific outcomes.

Machine Learning A field of study that enables computers to learn from data without being explicitly programmed, identifying hidden relationships and trends.

Mean The average value of a set of numbers is calculated by summing all the values and dividing by the total number of values.

Median When arranged in ascending or descending order, the middle value in a set of numbers divides the data into two equal halves.

Model (Conceptual model) A simplified representation or abstraction of a real-world system or phenomenon used to understand, analyze, or predict its behavior.

Model building The process of developing predictive models to gain insights and make informed decisions based on data analysis.

Pairwise comparison (correlation) A statistical technique that measures the strength and direction of the linear relationship between two variables by calculating a correlation coefficient.

Pattern A recurring or noticeable arrangement or sequence in data can provide insights or be used for prediction or classification.

Predictive model A model used to determine probabilities of an action or outcome based on historical data.

Predictors Variables or features in a model that are used to predict or explain the outcome variable or target variable.

Prioritization The process of organizing objectives and tasks based on their importance and impact on the overall goal.

Problem solving The process of addressing challenges and finding solutions to achieve desired outcomes.

Stakeholders Individuals or groups with a vested interest in the data science model's outcome and its practical application, such as solution owners, marketing, application developers, and IT administration.

Standard deviation A measure of the dispersion or variability of a set of values from their mean; It provides information about the spread or distribution of the data.

Statistical analysis Stand deviations are applied to problems that require counts, such as yes/no answers or classification tasks.

Statistics The collection, analysis, interpretation, presentation, and organization of data to understand patterns, relationships, and variability in the data.

Structured data (data model) Data organized and formatted according to a predefined schema or model and is typically stored in databases or spreadsheets.

Text analysis data mining The process of extracting useful information or knowledge from unstructured textual data through techniques such as natural language processing, text mining, and sentiment analysis.

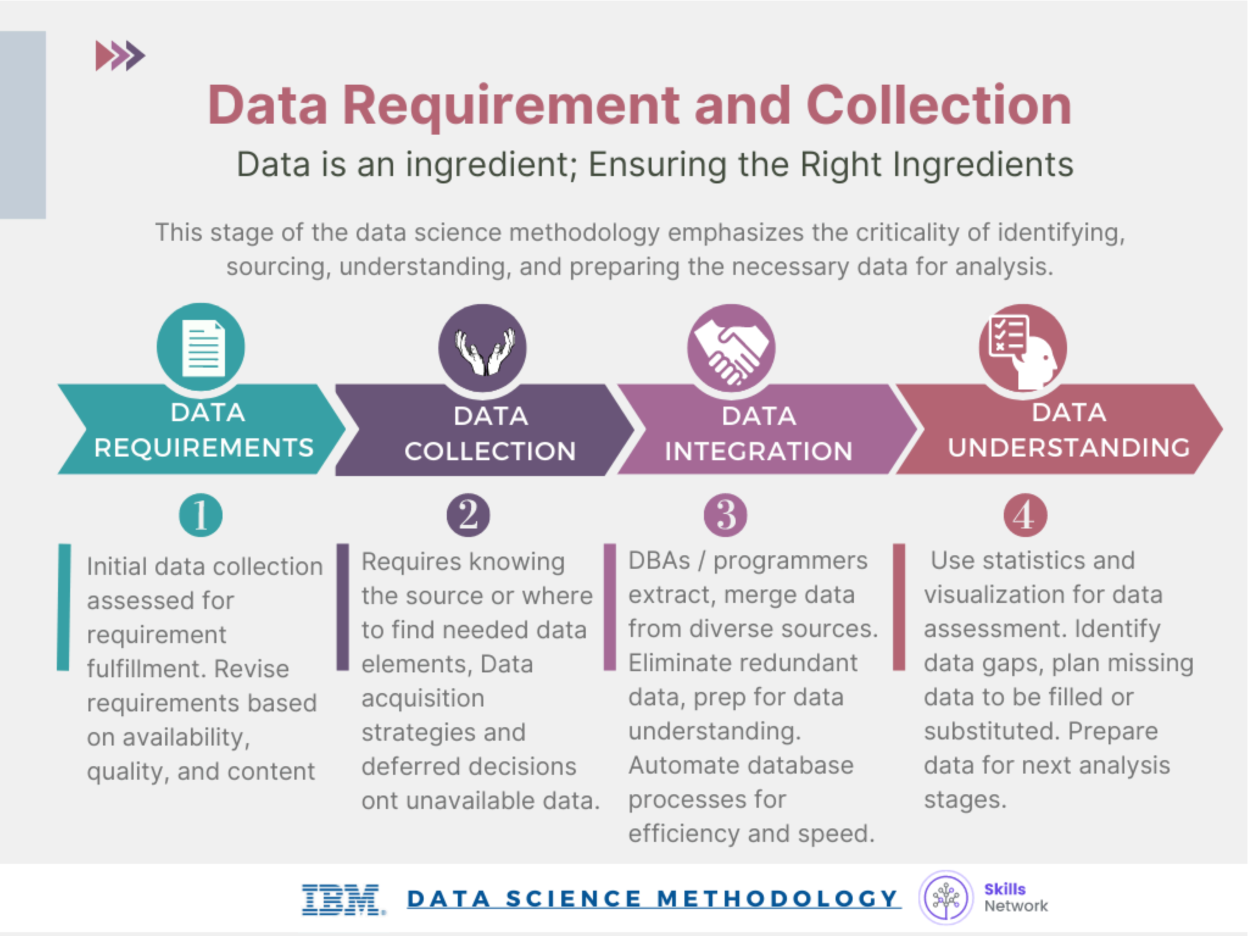
Threshold value The specific value used to split data into groups or categories in a decision tree.

Data requirement

- What are the data requirements?

Data collection

- What occurs during data collection?



Data Requirement and Collection

Data is an ingredient; Ensuring the Right Ingredients

This stage of the data science methodology emphasizes the criticality of identifying, sourcing, understanding, and preparing the necessary data for analysis.

DATA REQUIREMENTS

1 Initial data collection assessed for requirement fulfillment. Revise requirements based on availability, quality, and content

DATA COLLECTION

2 Requires knowing the source or where to find needed data elements, Data acquisition strategies and deferred decisions ont unavailable data.

DATA INTEGRATION

3 DBAs / programmers extract, merge data from diverse sources. Eliminate redundant data, prep for data understanding. Automate database processes for efficiency and speed.

DATA UNDERSTANDING

4 Use statistics and visualization for data assessment. Identify data gaps, plan missing data to be filled or substituted. Prepare data for next analysis stages.

Term	Definition
Analytics team	A group of professionals, including data scientists and analysts, responsible for performing data analysis and modeling.
Data collection	The process of gathering data from various sources, including demographic, clinical, coverage, and pharmaceutical information.
Data integration	The merging of data from multiple sources to remove redundancy and prepare it for further analysis.
Data Preparation	The process of organizing and formatting data to meet the requirements of the modeling technique.
Data Requirements	The identification and definition of the necessary data elements, formats, and sources required for analysis.

Data Understanding A stage where data scientists discuss various ways to manage data effectively, including automating certain processes in the database.

DBAs (Database Administrators) The professionals who are responsible for managing and extracting data from databases.

Decision tree classification A modeling technique that uses a tree-like structure to classify data based on specific conditions and variables.

Demographic information Information about patient characteristics, such as age, gender, and location.

Descriptive statistics Techniques used to analyze and summarize data, providing initial insights and identifying gaps in data.

Intermediate results Partial results obtained from predictive modeling can influence decisions on acquiring additional data.

Patient cohort A group of patients with specific criteria selected for analysis in a study or model.

Predictive modeling The building of models to predict future outcomes based on historical data.

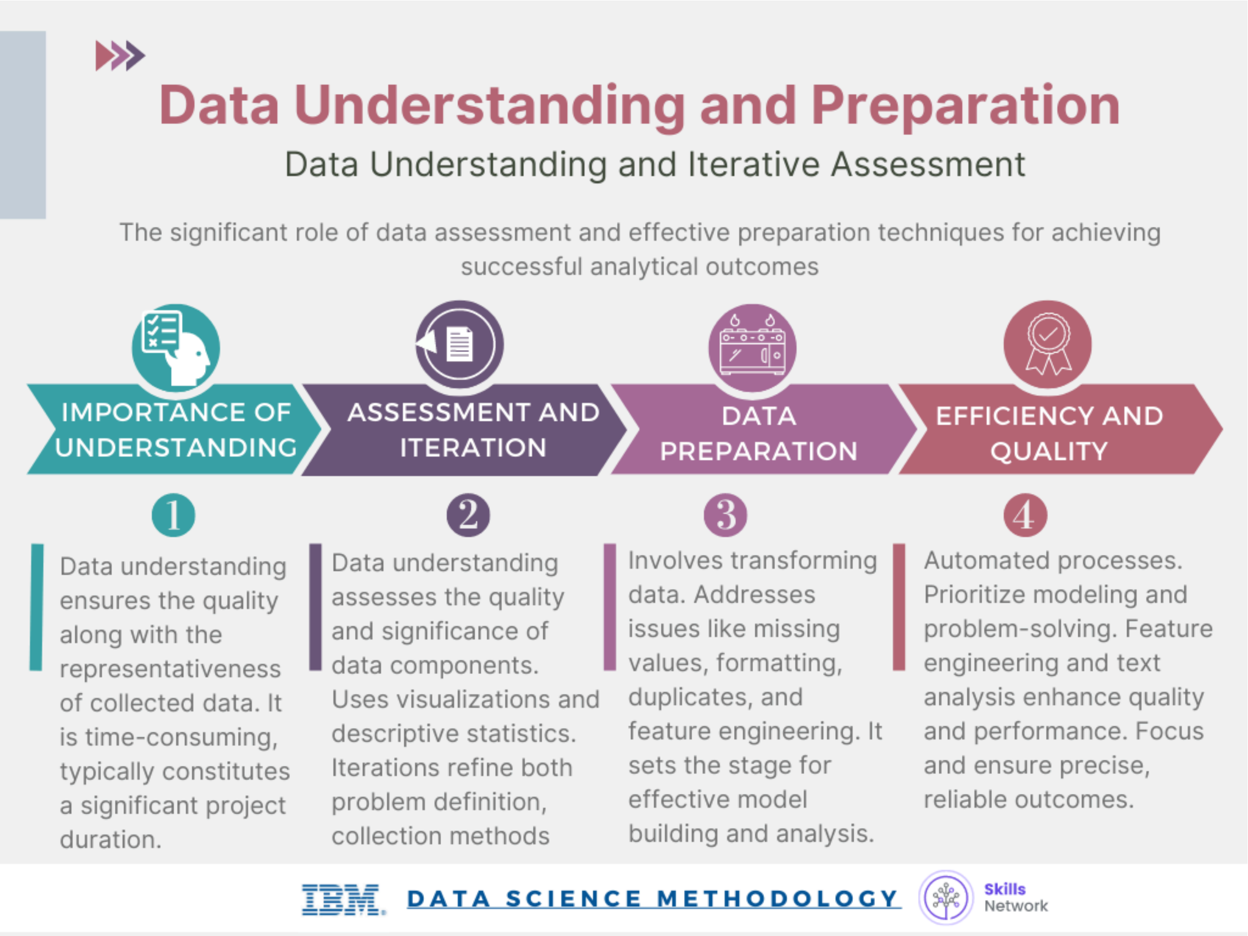
Training set A subset of data used to train or fit a machine learning model; consists of input data and corresponding known or labeled output values.

Unavailable data Data elements are not currently accessible or integrated into the data sources.

Univariate Modeling analysis focused on a single variable or feature at a time, considering its characteristics and relationship to other variables independently.

Unstructured data Data that does not have a predefined structure or format, typically text images, audio, or video, requires special techniques to extract meaning or insights.

Visualization The process of representing data visually to gain insights into its content and quality.



Data Understanding and Preparation

Data Understanding and Iterative Assessment

The significant role of data assessment and effective preparation techniques for achieving successful analytical outcomes

1. Data understanding ensures the quality along with the representativeness of collected data. It is time-consuming, typically constitutes a significant project duration.

2. Data understanding assesses the quality and significance of data components. Uses visualizations and descriptive statistics. Iterations refine both problem definition, collection methods

3. Involves transforming data. Addresses issues like missing values, formatting, duplicates, and feature engineering. It sets the stage for effective model building and analysis.

4. Automated processes. Prioritize modeling and problem-solving. Feature engineering and text analysis enhance quality and performance. Focus and ensure precise, reliable outcomes.

IBM. DATA SCIENCE METHODOLOGY Skills Network

Term	Definition
Automation	Using tools and techniques to streamline data collection and preparation processes.
Data Collection	The phase of gathering and assembling data from various sources.
Data Compilation	The process of organizing and structuring data to create a comprehensive data set.
Data Formatting	The process of standardizing the data to ensure uniformity and ease of analysis.
Data Manipulation	The process of transforming data into a usable format.
Data Preparation	The phase where data is cleaned, transformed, and formatted for further analysis, including feature engineering and text analysis.

Data Preparation The stage where data is transformed and organized to facilitate effective analysis and modeling.

Data Quality Assessment of data integrity and completeness, addressing missing, invalid, or misleading values.

Data Quality Assessment The evaluation of data integrity, accuracy, and completeness.

Data Set A collection of data used for analysis and modeling.

Data Understanding The stage in the data science methodology focused on exploring and analyzing the collected data to ensure that the data is representative of the problem to be solved.

Descriptive Statistics Summary statistics that data scientists use to describe and understand the distribution of variables, such as mean, median, minimum, maximum, and standard deviation.

Feature A characteristic or attribute within the data that helps in solving the problem.

Feature Engineering The process of creating new features or variables based on domain knowledge to improve machine learning algorithms' performance.

Feature Extraction Identifying and selecting relevant features or attributes from the data set.

Interactive Processes Iterative and continuous refinement of the methodology based on insights and feedback from data analysis.

Missing Values Values that are absent or unknown in the dataset, requiring careful handling during data preparation.

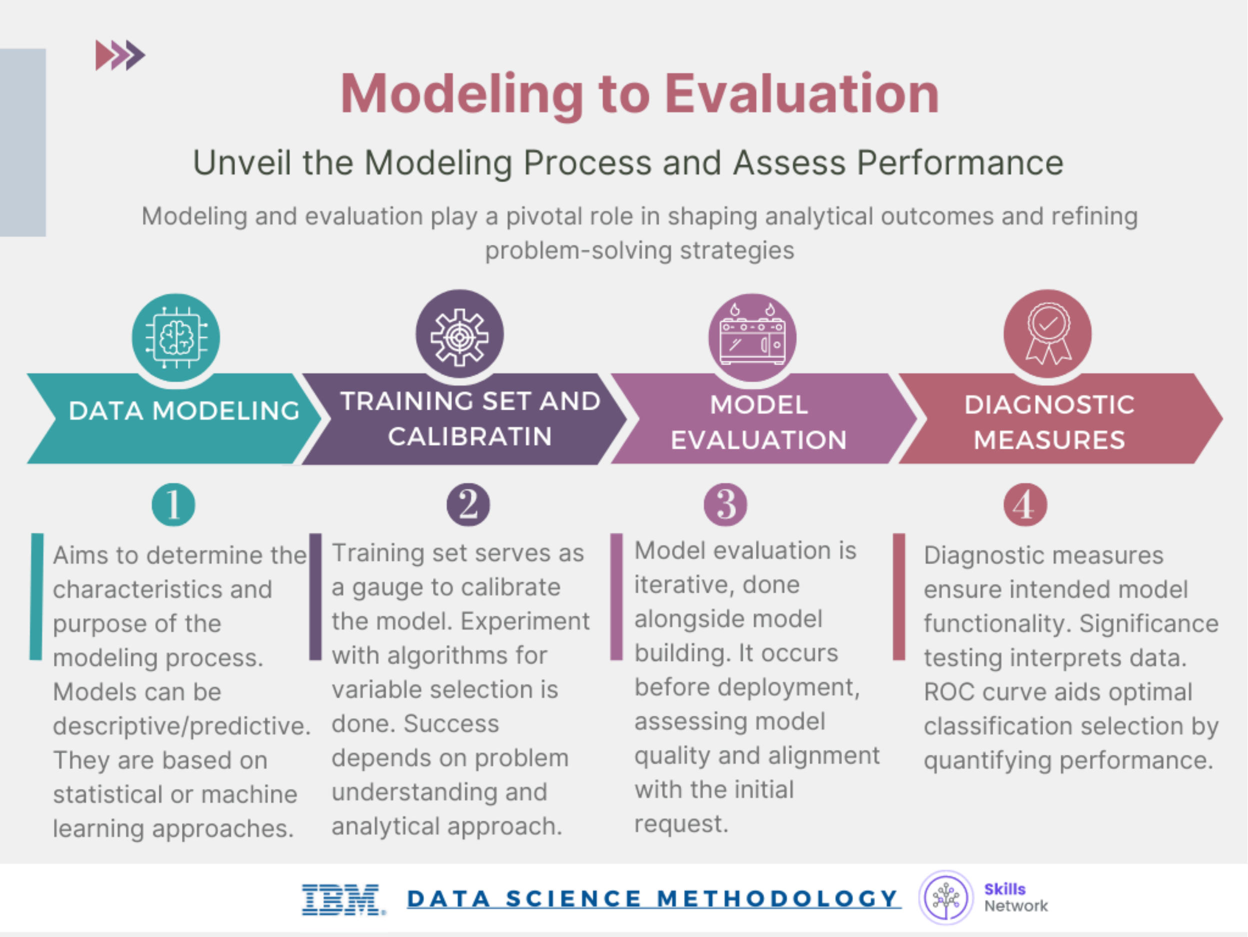
Model Calibration Adjusting model parameters to improve accuracy and alignment with the initial design.

Pairwise Correlations An analysis to determine the relationships and correlations between different variables.

Text Analysis Steps to analyze and manipulate textual data, extracting meaningful information and patterns.

Text Analysis Groupings Creating meaningful groupings and categories from textual data for analysis.

Visualization techniques Methods and tools that data scientists use to create visual representations or graphics that enhance the accessibility and understanding of data patterns, relationships, and insights.



Modeling to Evaluation

Unveil the Modeling Process and Assess Performance

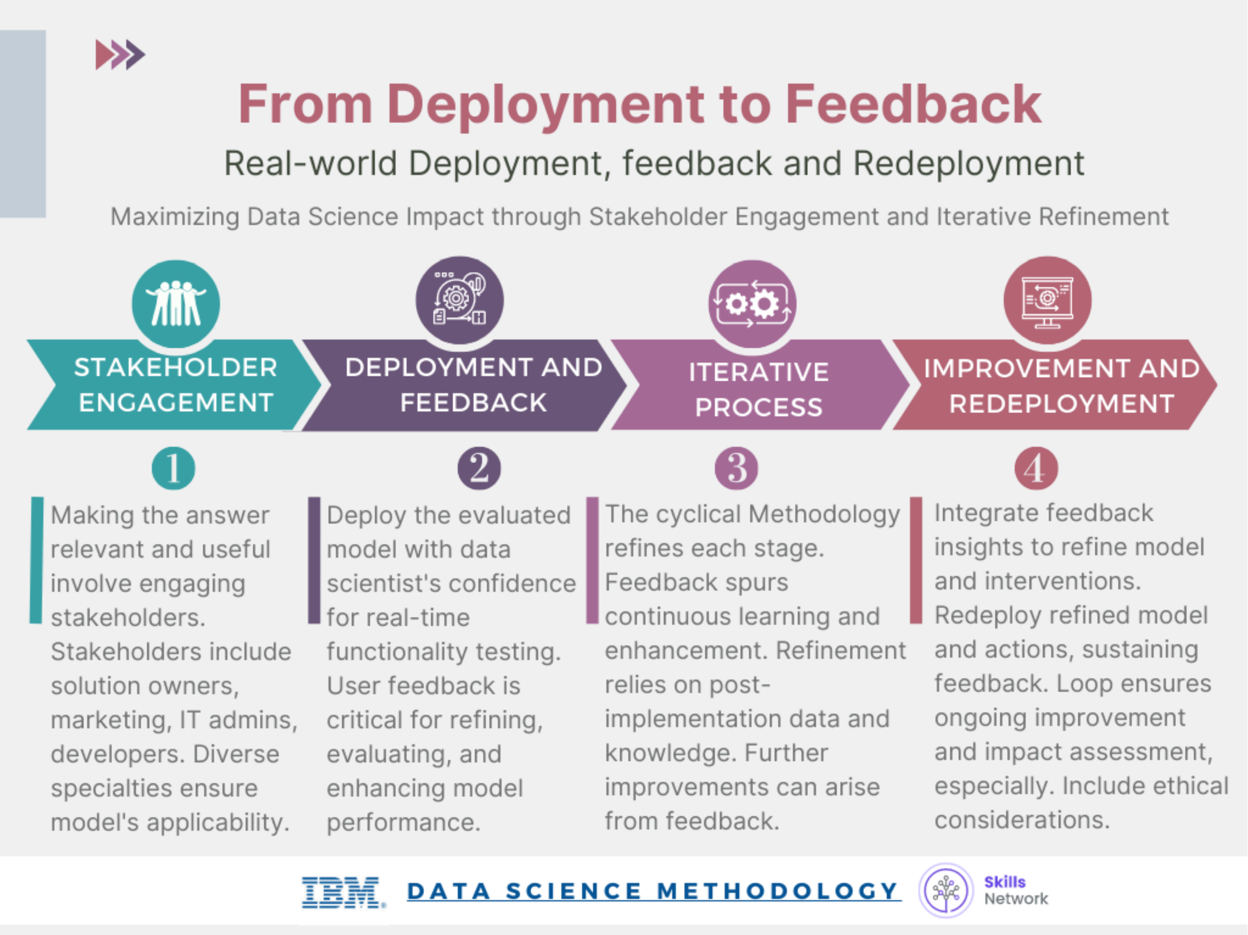
Modeling and evaluation play a pivotal role in shaping analytical outcomes and refining problem-solving strategies

1. DATA MODELING
   Aims to determine the characteristics and purpose of the modeling process. Models can be descriptive/predictive. They are based on statistical or machine learning approaches.

2. TRAINING SET AND CALIBRATIN
   Training set serves as a gauge to calibrate the model. Experiment with algorithms for variable selection is done. Success depends on problem understanding and analytical approach.

3. MODEL EVALUATION
   Model evaluation is iterative, done alongside model building. It occurs before deployment, assessing model quality and alignment with the initial request.

4. DIAGNOSTIC MEASURES
   Diagnostic measures ensure intended model functionality. Significance testing interprets data. ROC curve aids optimal classification selection by quantifying performance.



From Deployment to Feedback

Real-world Deployment, feedback and Redeployment

Maximizing Data Science Impact through Stakeholder Engagement and Iterative Refinement

1. Making the answer relevant and useful involve engaging stakeholders. Stakeholders include solution owners, marketing, IT admins, developers. Diverse specialties ensure model's applicability.

2. Deploy the evaluated model with data scientist's confidence for real-time functionality testing. User feedback is critical for refining, evaluating, and enhancing model performance.

3. The cyclical Methodology refines each stage. Feedback spurs continuous learning and enhancement. Refinement relies on post-implementation data and knowledge. Further improvements can arise from feedback.

4. Integrate feedback insights to refine model and interventions. Redeploy refined model and actions, sustaining feedback. Loop ensures ongoing improvement and impact assessment, especially. Include ethical considerations.

IBM DATA SCIENCE METHODOLOGY Skills Network

Term	Definition
Browser-based application	An application that users access through a web browser, typically on a tablet or other mobile device, to provide easy access to the model's insights.
Cyclical methodology	An iterative approach to the data science process, where each stage informs and refines the subsequent stages.
Data collection refinement	The process of obtaining additional data elements or information to improve the model's performance.
Data science model	The result of data analysis and modeling that provides answers to specific questions or problems.
Feedback	The process of obtaining input and comments from users and stakeholders to refine and improve the data science model.

Model refinement The process of adjusting and improving the data science model based on user feedback and real-world performance.

Redeployment The process of implementing a refined model and intervention actions after incorporating feedback and improvements.

Review process The systematic assessment and evaluation of the data science model's performance and impact.

Solution deployment The process of implementing and integrating the data science model into the business or organizational workflow.

Solution owner The individual or team responsible for overseeing the deployment and management of the data science solution.

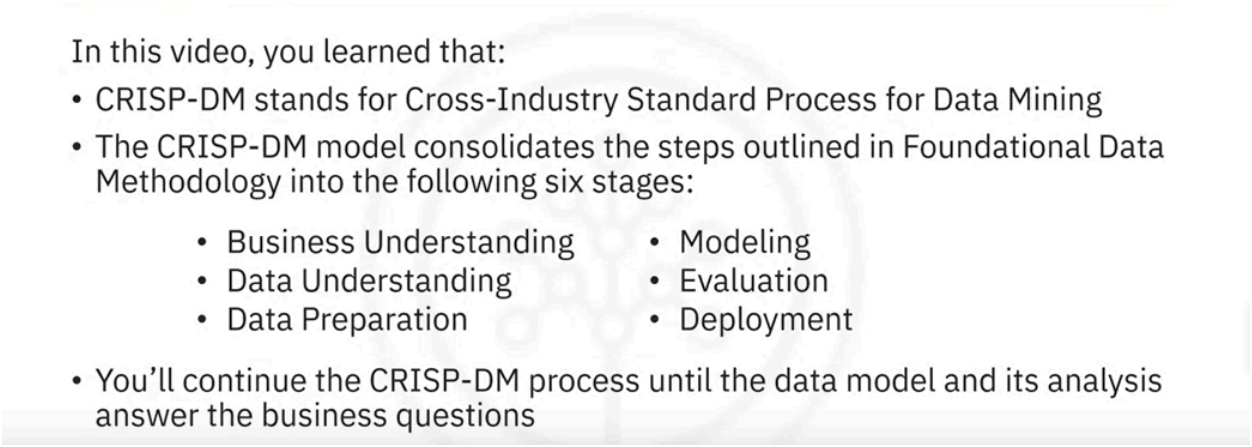
Stakeholders Individuals or groups with a vested interest in the data science model's outcome and its practical application, such as solution owners, marketing, application developers, and IT administration.

Storytelling Storytelling is the art of conveying your message, or ideas through a narrative structure that engages, entertains, and resonates with the audience.

Test environment A controlled setting where the data science model is evaluated and refined before full-scale implementation.

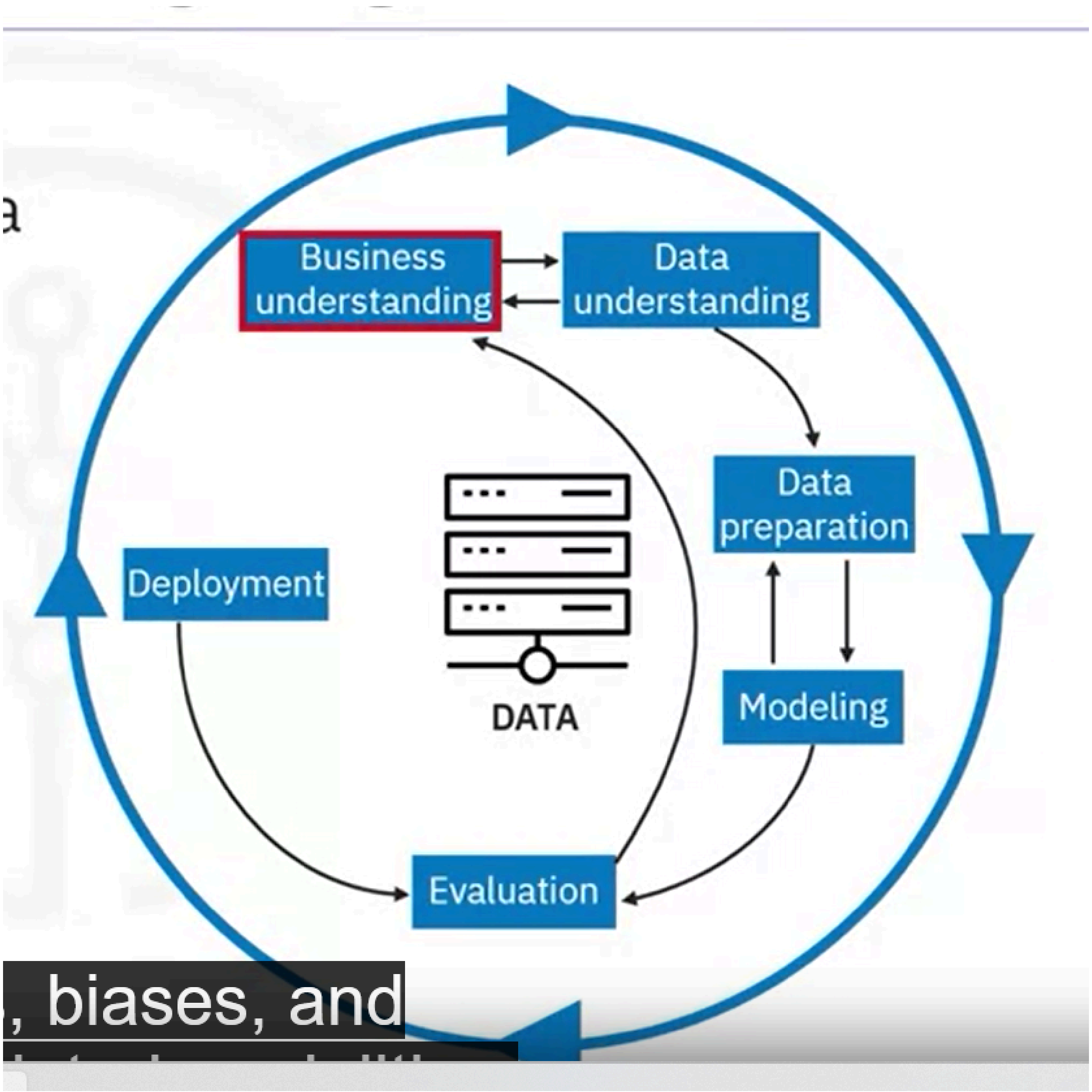
CRISP-DM

- Cross-industry standard process for data mining
- A structured approach to guide data-driven decision making



In this video, you learned that:

- CRISP-DM stands for Cross-Industry Standard Process for Data Mining
- The CRISP-DM model consolidates the steps outlined in Foundational Data Methodology into the following six stages:
  - Business Understanding
  - Data Understanding
  - Data Preparation
  - Modeling
  - Evaluation
  - Deployment
- You'll continue the CRISP-DM process until the data model and its analysis answer the business questions



, biases, and

SQL statement has 2 types : DDL and DML
DDL vs DML
DDL

- Data definition language statement
- Define, change, drop
- Create
- Alter
- Truncate
- drop

DML

- Data manipulation language statement
- Read and modify data
- CRUD
- Insert select update delete

ACID transaction

- Atomic
    - All changes must be performed successfully or not at all
- Consistent
    - Data must be in a consistent state before and after the transaction
- Isolated
    - No other process can change the data while the transaction is running
- Durable
    - The changes made by the transaction must be persist

Topic	Syntax	Description	Example
Create View	`CREATE VIEW view_name AS SELECT column1, column2, ... FROM table_name WHERE condition;`	A `CREATE VIEW` is an alternative way of representing data that exists in one or more tables.	`CREATE VIEW EMPSALARY AS SELECT EMP_ID, F_NAME, L_NAME, B_DATE, SEX, SALARY FROM EMPLOYEES;`
Update a View	`CREATE OR REPLACE VIEW view_name AS SELECT column1, column2, ... FROM table_name WHERE condition;`	The `CREATE OR REPLACE` VIEW command updates a view.	`CREATE OR REPLACE VIEW EMPSALARY AS SELECT EMP_ID, F_NAME, L_NAME, B_DATE, SEX, JOB_TITLE, MIN_SALARY, MAX_SALARY FROM EMPLOYEES, JOBS WHERE EMPLOYEES.JOB_ID = JOBS.JOB_IDENT;`

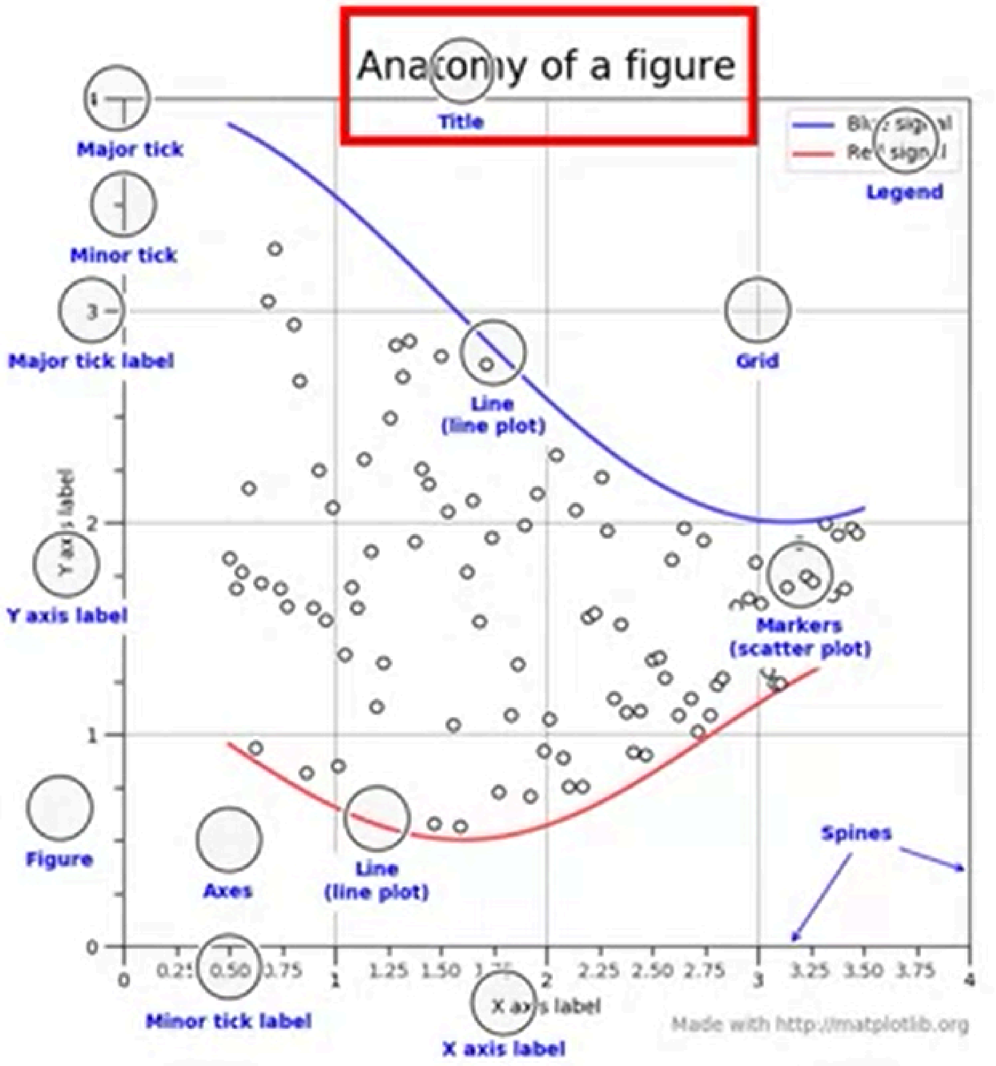
Drop a View	`DROP VIEW view_name;`	Use the `DROP VIEW` statement to remove a view from the database.	`DROP VIEW EMPSALARY;`

Scatter plot: Use case

- Examine the relationship between 2 continuous variables
- Investigate patterns or trends in data
- Detect outliers or unusual observations
- Identity clusters or groups in the data

Matplotlib architecture

- Scripting layer (pyplot)
- Artist layer (artist)
    - Knows how to use renderer to draw on the canvas
    - 2 types of artist objects
        - Primitive: line2d, rectangle, circle and text
        - Composite: axis, tick, axes, and figure
- Backend layer(figureCanvas(where to draw), Renderer(how to draw), Event(mouse click))



Function	Description	Syntax	Example
Plotly Express			
scatter	Create a scatter plot	`px.scatter(dataframe, x=x_column, y=y_column)`	`px.scatter(df, x=age_array, y=income_array)`

line	Create a line plot	px.line(x=x_column, y=y_column,'title')	px.line(x=months_array, y=no_bicycle_sold_array)
bar	Create a bar plot	px.bar(x=x_column, y=y_column,title='title')	px.bar(x=grade_array, y=score_array, title='Pass Percentage')
sunburst	Create a sunbust plot	px.sunburst(dataframe, path=[col1,col2..], values='column',title='title')	px.sunburst(data, path=['Month', 'DestStateName'], values='Flights',title='Flight Distribution Hierarchy')
histogram	Create a histogram	px.histogram(x=x,title="title")	px.histogram(x=heights_array,title="Distribution of Heights")
bubble	Create a bubble chart	px.scatter(dataframe, x=x,y=y,size=size,title="title")	px.scatter(bub_data, x="City", y="Numberofcrimes", size="Numberofcrimes",hover_name="City", title='Crime Statistics')
pie	Create a pie chart	px.pie(values=x,names=y,title="title")	px.pie(values=exp_percent, names=house_holdcategories, title='Household Expenditure')

Plotly Graph Objects

Scatter	Create a scatter	`go.Scatter(x=x, y=y, mode='markers')`	`go.Scatter(x=age_array, y=income_array, mode='markers')`
	Create a line plot	`go.Scatter(x=x, y=y, mode='lines')`	`go.Bar(x=months_array, y=no_bicycle_sold_array,mode='lines')`
add_trace	Add additional traces to an existing figure	`fig.add_trace(trace_object)`	`fig.add_trace(go.Scatter(x=months_array, y=no_bicycle_sold_array))`
update_layout	Update the layout of a figure, such as title, axis labels, and annotations.	`fig.update_layout(layout_object)`	`fig.update_layout(title='Bicycle Sales', xaxis_title='Months', yaxis_title='Number of Bicycles Sold')`

Dash

dash_core_components.Input	Create an input component	`dcc.Input(value='', type='text')`	`dcc.Input(value='Hello', type='text')`
dash_core_components.Graph	Create a graph component	`dcc.Graph(figure=fig)`	`dcc.Graph(figure=fig)`
dash_html_components.Div	Create a div element	`html.Div(children=component_list)`	`html.Div(children=[html.H1('Hello Dash'), html.P('Welcome to Dash')])`

dash_core_components.Dropdown	Create a dropdown component	`dcc.Dropdown(options=options_list, value=default_value)`	`dcc.Dropdown(options=[{'label': 'Option 1', 'value': '1'}, {'label': 'Option 2', 'value': '2'}], value='1')`

Entropy

- How much randomness is in the data
- The lower the entropy, the less uniform the distribution, the purer the node
- Low entropy: 1 drug A 7 drub B
- High entropy: 3 drug A 5 durg B
- Entropy = 0 when 0 draug A all drug B
- Entropy = 1 when 4 drug A 4 drug B

Information gain

- Information that can increase the level of certainty after splitting
- Entropy before split - weighted entropy after split

Decision tree

- Our split must return the highest information gain

Advantages if SVM

- ACCURATE IN HIGH-DIMENSIONAL SPACES
- MEMORY EFFICIENT

Disadvantages

- Prone to overfitting
- No probability estimation
- Small dataset
- Not efficient when computing