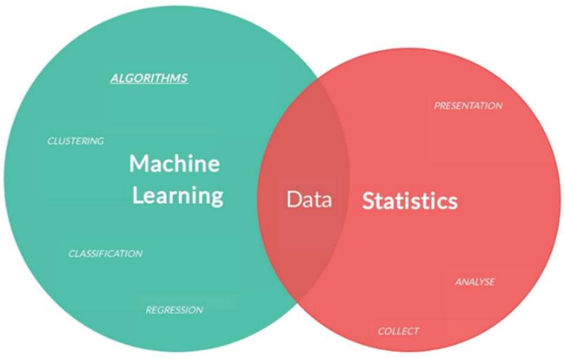
Study Material

Module II: Types of Learning and Its Application [Part 1]

Overview of Statistical Approaches in Machine Learning

In the field of machine learning (ML), statistics plays a pivotal role in extracting meaningful insights from data to make informed decisions. Statistics provides the foundation upon which various ML algorithms are built, enabling the analysis, interpretation, and prediction of complex patterns within datasets.



What is Statistics?

Statistics is the science of collecting, organizing, analyzing, interpreting, and presenting data. It encompasses a wide range of techniques for summarizing data, making inferences, and drawing conclusions. Statistical methods help quantify uncertainty and variability in data, allowing researchers and analysts to make data-driven decisions with confidence.

What is Machine Learning?

Machine learning is a branch of artificial intelligence (AI) that focuses on developing algorithms and models capable of learning from data without being explicitly programmed. ML algorithms learn patterns and relationships from data, which they use to make predictions or decisions. Machine learning encompasses various techniques, including supervised learning, unsupervised learning, and reinforcement learning.

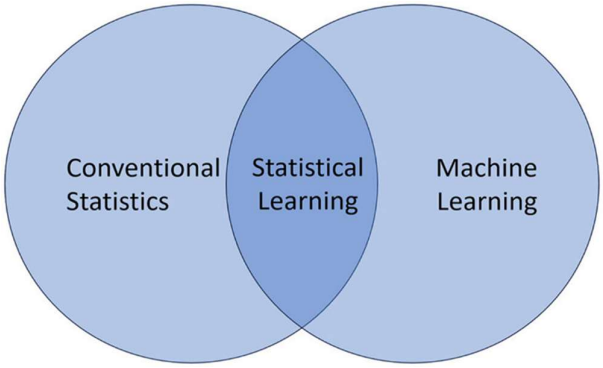


Figure - Statistical learning at the crossroads of statistics and machine learning

Statistical learning serves as a bridge between statistics as a centuries-old branch of mathematics and machine learning as an upcoming sub-discipline of artificial intelligence. Statistical learning uses the principles of both disciplines, trying to detect patterns in the data and make individual predictions based on assumed underlying data models.

Applications of Statistics in Machine Learning

Statistics is a key component of machine learning, with broad applicability in various fields.

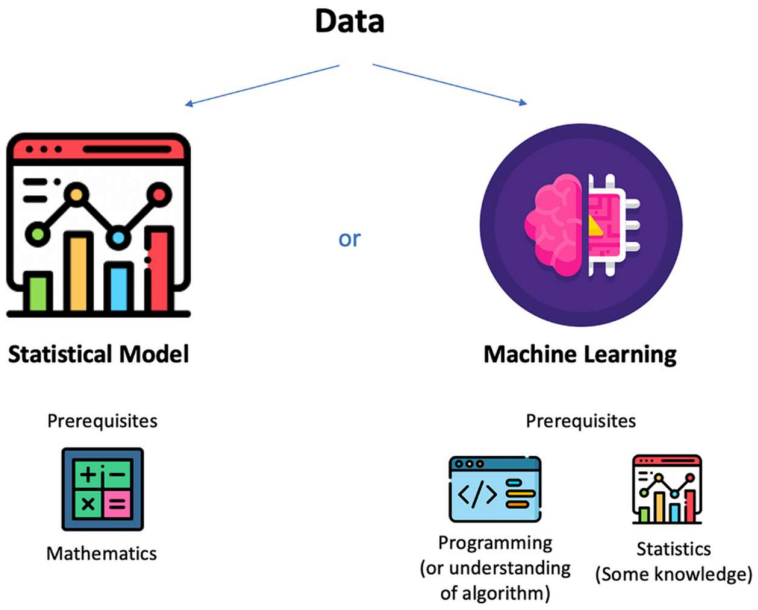
i) Feature engineering relies heavily on statistics to convert geometric features into meaningful predictors for machine learning algorithms.

ii) In image processing tasks like object recognition and segmentation, statistics accurately reflect the shape and structure of objects in images.

iii) Anomaly detection and quality control benefit from statistics by identifying deviations from norms, aiding in the detection of defects in manufacturing processes.

iv) Environmental observation and geospatial mapping leverage statistical analysis to monitor land cover patterns and ecological trends effectively.

Overall, statistics plays a crucial role in machine learning, driving insights and advancements across diverse industries and applications.



Types of Statistics

There are commonly two types of statistics, which are discussed below:

i) **Descriptive Statistics:** Descriptive Statistics helps us simplify and organize big chunks of data. This makes large amounts of data easier to understand.

ii) **Inferential Statistics:** Inferential Statistics is a little different. It uses smaller data to conclude a larger group. It helps us predict and draw conclusions about a population.

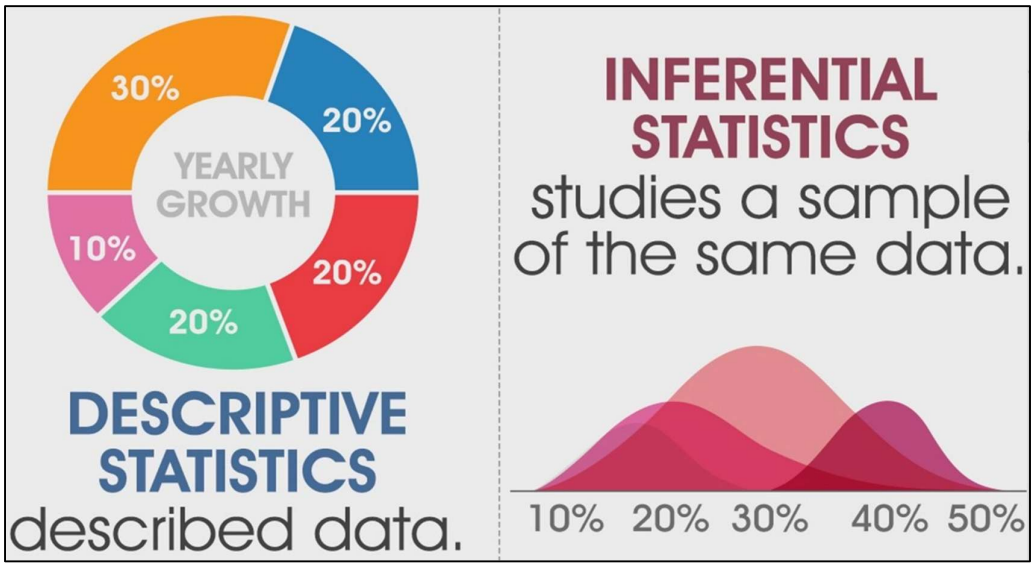


Figure – Types of Statistics

Machine learning turns statistics to predict outcomes and adapt to data.

Machine learning is often said to be "an evolution of statistics" because it builds on statistical concepts to handle larger, more complex data problems with a focus on predictions and automation. In simple terms, machine learning builds on statistics to solve bigger, more complex problems, often focusing more on predictions than explanations.

Roles of Statistics in Machine Learning Statistics

Statistics is the backbone of machine learning. It provides the theoretical foundation for:

i) **Understanding data**: Descriptive stats (mean, median, standard deviation) help summarize and visualize patterns in data.

ii) **Making inferences**: Inferential statistics help assess relationships, significance, and trends.

iii) **Model validation**: Metrics like R^2, p-values, accuracy, and cross-validation scores come from statistical theory.

iv) **Probabilistic modeling**: Many ML algorithms, such as Naive Bayes, Bayesian Networks, are grounded in probability theory, a branch of statistics.

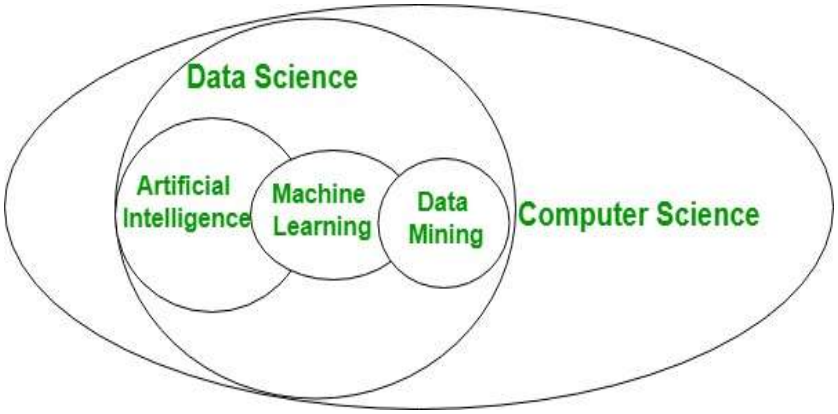
Difference Between Statistics and Machine Learning

Machine Learning	Statistics
Machine Learning is a subset of AI that focuses on designing algorithms that learn from data and improve over time.	**Statistics** is a branch of mathematics that deals with data analysis, interpretation, and presentation.
It makes the most accurate prediction possible and then foresees future events or arranges current materials.	It interfaces the relationship between the variables and finds out the connection between the information points.
The Goal is to develop systems that can make predictions or decisions without explicit programming	The goal is to understand data, identify trends, and make inferences about a population based on a sample.
It is based on a data-driven approach focusing on building predictive models and optimizing performance through iterations	It is based on a theory-driven approach, focusing on understanding the underlying structure of data and deriving conclusions.
It works well with large datasets; performance improves with more data.	It can work with smaller datasets but requires proper sampling to ensure representativeness.
It Uses algorithms like neural networks, decision trees, and clustering.	It uses methods like regression analysis, hypothesis testing, and descriptive statistics.
It often relies on fewer assumptions about the underlying data distribution	It typically involves assumptions about data distribution (e.g., normality, independence).

Statistics is the foundation of machine learning, allowing for the extraction of useful insights from data across multiple domains. Machine learning algorithms can use statistical techniques and methodologies to learn from data, generate predictions, and solve complicated problems successfully. Understanding the significance of statistics in machine learning is critical for practitioners and researchers who want to use the power of data-driven decision-making in their domains.

Data Mining and Machine Learning

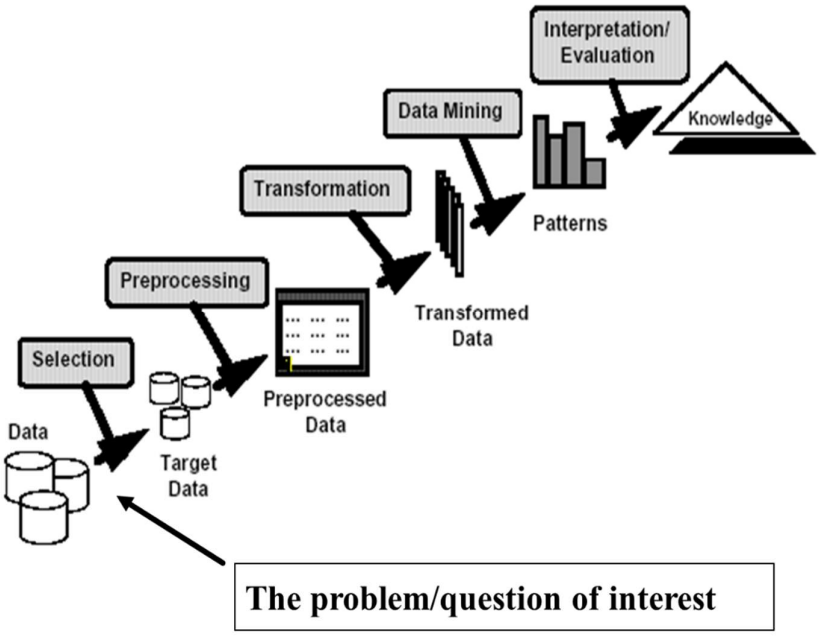
Data mining and machine learning are related fields that both involve learning from data, but they have different focuses.



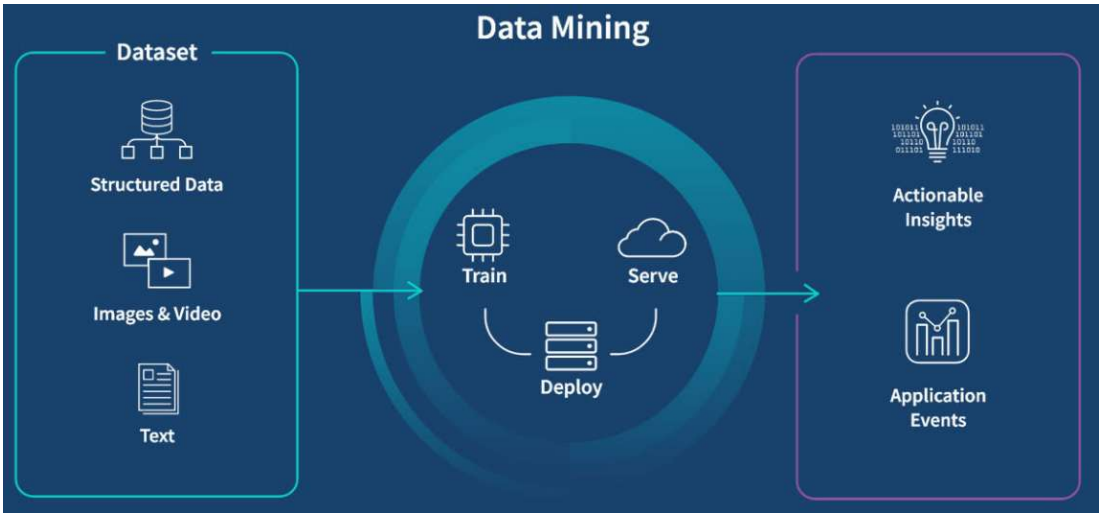
Both data mining and machine learning fall under the aegis of Data Science, which makes sense since they both use data. Both processes are used for solving complex problems, so consequently, many people (erroneously) use the two terms interchangeably. While data gathered from data mining can be used to teach machines, the lines between the two concepts become a bit blurred.

What is Data Mining?

The process of extracting useful information from a huge amount of data is called Data mining. Data mining is a tool that is used by humans to discover new, accurate, and useful patterns in data or meaningful, relevant information for those who need it.



Data mining is the overall process of identifying patterns and extracting useful insights from big data sets. This can be used to evaluate both structured and unstructured data to identify new information and is commonly used to analyze consumer behaviors for marketing and sales teams. For example, data mining methods can be used to observe and predict behaviors, including customer churn, fraud detection, market basket analysis and more.



Data mining can be seen as a subset of data analytics that specifically focuses on extracting hidden patterns and knowledge from data. Historically, a data scientist was required to build, refine, and deploy models. However, with the rise of different automated tools, data analysts can now perform these tasks if the model is not too complex.

Data Mining: Examples and Non-Examples

Data Mining	NOT Data Mining
• Certain names are more prevalent in certain US locations (O'Brien, O'Rurke, O'Reilly... in Boston area) • Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com, etc.)	• Look up phone number in phone directory • Query a Web search engine for information about "Amazon"

Why Mine Data?

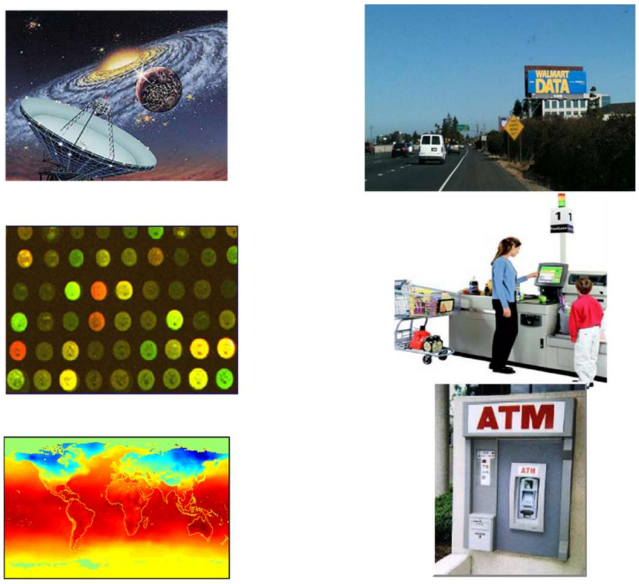
1) Scientific Viewpoint -
   i) A huge amount of data is collected and stored at enormous speeds
      - Remote sensors on a satellite
      - Telescopes scanning the skies
      - Microarrays generating gene expression data
      - Scientific simulations generating terabytes of data

ii) Traditional techniques are infeasible for large data sets

iii) Data mining may help scientists
- in classifying and segmenting data
- in hypothesis formation

2) Commercial Viewpoint – Lots of data is being collected and warehoused

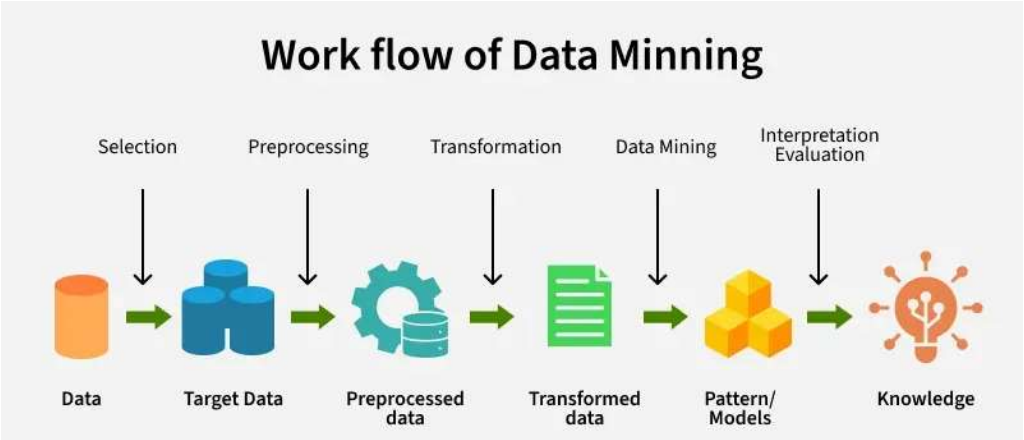
- E-commerce data
- Purchases at departmental or grocery stores
- Credit card transactions
- Computers have become cheaper and more powerful
- Strong competitive pressure, to provide better, customized services for an edge



(a) (b)

Figure – (a) Scientific Viewpoint (b) Commercial Viewpoint

The data mining process involves several steps from data collection to visualization to extract valuable information from large data sets. Data mining techniques can be used to generate descriptions and predictions about a target data set.



Relationship between Data Mining and Machine Learning

Data mining focuses on extracting useful patterns or knowledge from large datasets. It overlaps with machine learning but has a slightly different goal:

i) **Pattern discovery**: Association rules, clustering, and anomaly detection.

ii) **Feature extraction**: Selecting or engineering meaningful variables for ML models.

iii) **Preprocessing**: Cleaning, transformation, and reduction of data complexity.

iv) **Exploratory analysis**: Understanding structure and relationships before model training.

While machine learning is often predictive, data mining is more exploratory and descriptive**,** helping you figure out what to model in the first place.

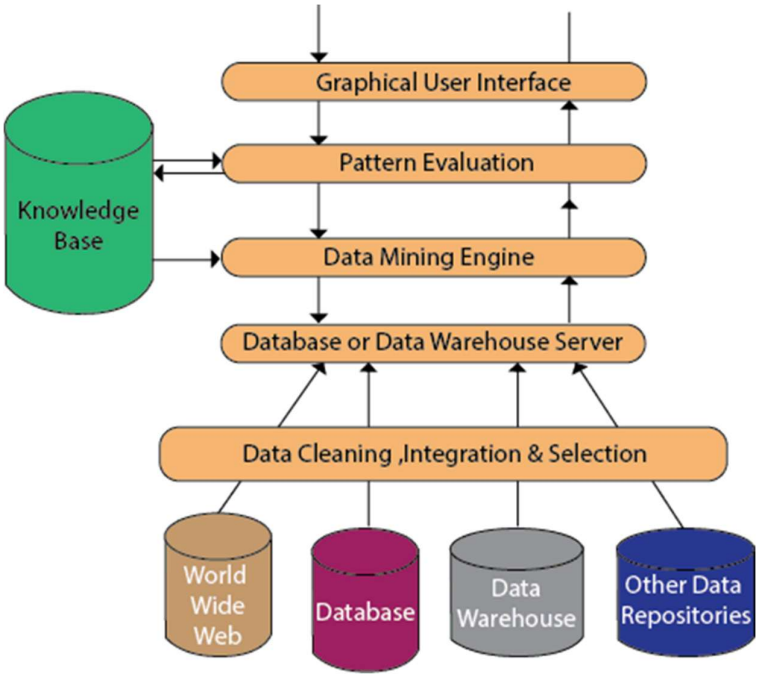


Figure: Architecture of a Typical Data Mining System

Difference Between Data Science and Machine Learning

Data Mining	Machine Learning
Extracting useful information from large amount of data	Introduce algorithm from data as well as from past experience
Used to understand the data flow	Teaches the computer to learn and understand from the data flow
Huge databases with unstructured data	Existing data as well as algorithms
Models can be developed for using data mining technique	Machine learning algorithm can be used in the decision tree, neural networks and some other area of artificial intelligence
Human interference is more in it.	No human effort required after design
It is used in cluster analysis	It is used in web Search, spam filter, fraud detection and computer design
Data mining abstract from the data warehouse	Machine learning reads machine
Data mining is more of research using methods like machine learning	Self-learned and trains system to do the intelligent task
Applied in limited area	Can be used in vast area
Uncovering hidden patterns and insights	Making accurate predictions or decisions based on data
Exploratory and descriptive	Predictive and prescriptive
Historical data	Historical and real-time data
Patterns, relationships, and trends	Predictions, classifications, and recommendations
Clustering, association rule mining, outlier detection	Regression, classification, clustering, deep learning
Data cleaning, transformation, and integration	Data cleaning, transformation, and feature engineering
Strong domain knowledge is often required	Domain knowledge is helpful, but not always necessary
Can be used in a wide range of applications, including business, healthcare, and social science	Primarily used in applications where prediction or decision-making is important, such as finance, manufacturing, and cybersecurity

Data mining is the process of discovering patterns and relationships in large datasets, often using machine learning techniques, with the goal of extracting useful information for decision-making. Machine learning, on the other hand, is a broader field focused on developing algorithms that allow computers to learn from data and make predictions or decisions without explicit programming. In essence, data mining leverages machine learning to uncover hidden knowledge within data, while machine learning uses data to train models for various tasks.