OLAP stands for Online Analytical Processing

OLAP stands for **Online Analytical Processing**. It is a category of data processing that allows users to **analyze multidimensional data interactively** from multiple perspectives.

OLAP helps you **ask complex questions about large amounts of data** and get answers **fast** — great for business reporting, forecasting, budgeting, and decision-making.

Key Features:

- **Multidimensional Analysis**: Data is organized in dimensions (e.g., time, geography, product).
- **Fast Query Performance**: Optimized for rapid querying and reporting.
- **Aggregations**: Performs operations like **sum**, **average**, **count**, etc., across dimensions.
- **Slicing and Dicing**: Lets users "slice" data (e.g., see sales by region) and "dice" it (e.g., view by region and product).
- **Drill-Down and Roll-Up**: Navigate from summary data to details (drill-down) or aggregate data (roll-up).

OLAP Cube:

The core of OLAP is the **OLAP cube**, a data structure that lets you view data across multiple dimensions. Imagine a cube where each side represents a different dimension (like **time**, **location**, **product**), and each cell holds a metric (like **sales**).

OLAP vs OLTP:

Feature	**OLAP**	**OLTP**
	Online Analytical Processing	**Online Transaction Processing**
Purpose	Analytics	Transaction processing
Data Type	Historical, summarized	Real-time, detailed
Speed	Fast for reads/queries	Fast for inserts/updates
Examples	Business intelligence, dashboards	Banking systems, e-commerce orders

Types of OLAP:

1. **MOLAP (Multidimensional OLAP)** – Pre-computed cubes, very fast.
2. **ROLAP (Relational OLAP)** – Uses relational databases, flexible but slower.
3. **HOLAP (Hybrid OLAP)** – Combines MOLAP and ROLAP for performance and scalability.

Characterization

- **Definition:** Characterization is the process of summarizing the general features of a target class of data. You're basically building a high-level description of what a particular group looks like.
- **Goal:** Find *common patterns and trends* within one class.
- **Example:** Suppose you look at data about "graduate students." Characterization might tell you:
    - 70% are between 22–28 years old
    - 60% live within 5 km of campus
    - Average GPA is 3.2

Discrimination

- **Definition:** Discrimination is the process of comparing features of one class with those of other classes. It's about highlighting differences between groups.
- **Goal:** Identify *contrasting characteristics* between two or more classes.
- **Example:** Compare "graduate students" vs. "undergraduate students." Discrimination might reveal:
    - Graduate students are older on average (24 vs. 20)
    - Graduate students study more hours per week (40 vs. 25)
- Here, you're not just describing; you're *contrasting*.

Characterization:

- **Age**: Graduate students tend to be older, with a median age around 25, while undergraduates are younger, with a median age around 20.
- **Study Hours**: Graduate students study significantly more hours on average (around 46 hours per week) compared to undergraduates, who only study about 20 hours weekly.

Discrimination:

- **Age**: Graduate students are, on average, 5 years older than undergraduates.
- **Study Hours**: Graduate students spend approximately 26 more hours per week studying than undergraduates.

.**Rule**: **buys(X, “computer”) → buys(X, “software”) [Support = 1%, Confidence = 50%]**

Support (1%)

- **Definition**: Support refers to how frequently a particular itemset or rule occurs in the dataset. It's the percentage of transactions in which both items in the rule (computer and software, in this case) appear together.
- **Interpretation**: **Support = 1%** means that **1% of all transactions** in the dataset involve both buying a computer and buying software. This gives an idea of how common the rule is across the entire dataset.

Confidence (50%)

- **Definition**: Confidence measures the likelihood that the rule holds true, given the presence of the antecedent (the first part of the rule, i.e., buying a computer).
- **Interpretation**: **Confidence = 50%** means that **50% of the transactions that involve buying a computer** also involve buying software. So, if someone buys a computer, there is a 50% chance they will also buy software.

In summary:

- **Support = 1%**: 1% of all transactions involve both a computer and software.
- **Confidence = 50%**: Among the transactions where a computer is bought, 50% also involve buying software.

Are All Patterns Interesting? A data mining system can generate thousands or millions of patterns, but **not all** are interesting. Only a small fraction of them will be meaningful to the user.

- **What Makes a Pattern Interesting?** A pattern is considered interesting if it is:
  - **Easily understood** by humans.
  - **Valid** on new or test data with some certainty.
  - **Potentially useful** to the user.
  - **Novel**, meaning it provides new information.
  - It can also **validate a hypothesis** that the user is trying to confirm.

- **Objective Measures of Interestingness:** These measures are based on the structure and statistics of the patterns.
  - **Support**
  - **Confidence**
  - **Accuracy**: The percentage of data correctly classified by a rule.
  - **Coverage**: The percentage of data to which a rule applies.
  - **Understandability**: Measures the complexity or length of a rule or pattern (e.g., in terms of bits).

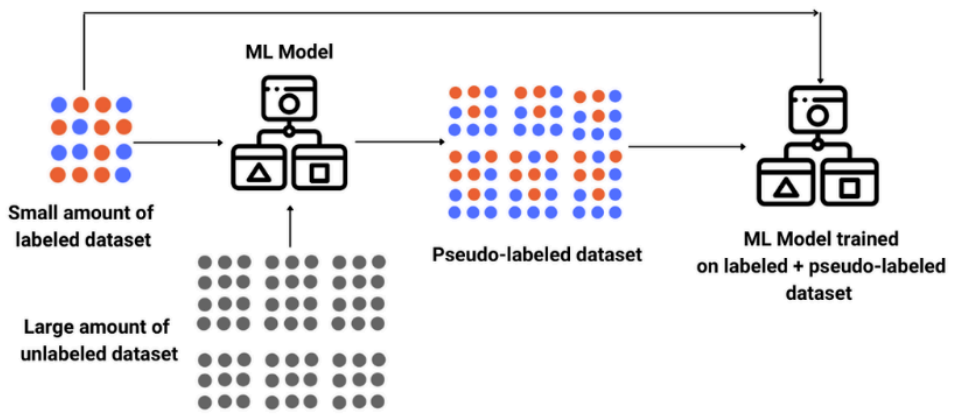
- **Subjective Measures of Interestingness:** These are based on the user's beliefs and needs:
    - A pattern is interesting if it is **unexpected** (contradicts the user's belief) or offers **strategic** information that the user can act on.
    - For instance, **actionable patterns** like "a large earthquake often follows small quakes" are useful if they can be acted upon to save lives.
    - Patterns may also be interesting if they **confirm a hypothesis** or align with the user's expectations.

Three Key Questions:

- **Are All Patterns Interesting? No**. Only a small subset of generated patterns is actually of interest to a given user.
- **Can a Data Mining System Generate All of the Interesting Patterns? No**. It is often inefficient to generate all possible patterns. User-provided constraints and interestingness measures help focus the search for patterns.
- **Can a Data Mining System Generate Only Interesting Patterns? Ideally, yes**, but it remains a challenging optimization problem. Data mining systems often generate patterns that need to be filtered based on their interestingness.

Inferential Statistics is a way of using data from a smaller sample to make predictions or conclusions about a larger population. It helps us account for **uncertainty and randomness** in the data, so we can make reasonable guesses or decisions about things we can't directly measure.

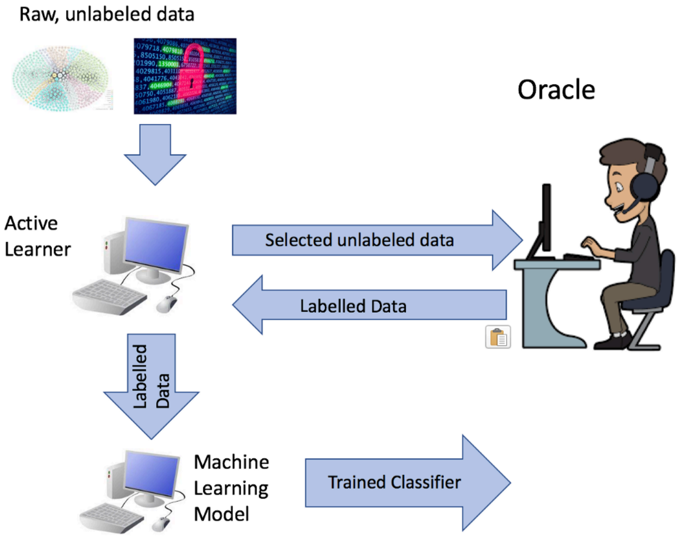
Semi Supervised Learning


ML Model
Small amount of labeled dataset
Large amount of unlabeled dataset
Pseudo-labeled dataset
ML Model trained on labeled + pseudo-labeled dataset

Active Learning

1. Initial Model Training
2. Uncertainty Sampling
3. Query the Oracle
4. Model Retraining
5. Iterative Process

Active Machine Learning



Data Mining Issues

1. Mining Methodology

This is about improving the techniques used to mine data. Key areas here include:

- **New types of knowledge**: Researchers are exploring various ways to find patterns in data, like trends, correlations, and classifications, to make mining more effective.

- **Multidimensional data mining**: Data can be analyzed across multiple dimensions or aspects (like time, location, etc.). A good example is analyzing sales data by location, product, and time.

- **Interdisciplinary approaches**: Data mining can benefit from combining knowledge from other fields like natural language processing (NLP) or software engineering to enhance how we mine data.

- **Dealing with uncertainty**: Data is often messy—full of noise, missing information, or errors. Handling these issues is crucial to ensure that the patterns found are valid and accurate.

- **Pattern evaluation**: Not every pattern discovered is useful. Researchers are working on methods to assess which patterns are genuinely interesting based on user needs.

2. User Interaction

The user plays a crucial role in guiding data mining processes. This includes:

- **Interactive mining**: Users should be able to interact with data mining systems to explore data dynamically, refine searches, and adjust how the data is being mined.

- **Incorporating background knowledge**: Users can provide specific domain knowledge (like business rules or constraints) to guide the mining process.

- **Ad hoc data mining**: Users should have the flexibility to create customized queries or data mining tasks, like asking the system to search for specific patterns.

- **Visualization**: The results of data mining need to be presented in a way that is easy for users to understand. Visual tools help make the results clear and actionable.

3. Efficiency and Scalability

As the amount of data increases, it's important that data mining systems can handle large datasets quickly and efficiently:

- **Efficient algorithms**: Data mining methods need to be fast and scalable to deal with large datasets, especially in real-time.

- **Parallel and distributed mining**: To process huge amounts of data, mining can be done in parallel across multiple computers, which speeds up the process.

- **Incremental mining**: Instead of starting the mining process from scratch every time new data is added, incremental mining updates the existing findings as new data comes in.

4. Diversity of Data Types

Data mining systems need to handle various types of data, such as:

- **Complex data**: Data can be structured (like databases), semi-structured (like XML files), or unstructured (like text or images). Mining techniques need to handle all these formats.

- **Dynamic data**: Some data sources are constantly changing, such as social media feeds or real-time sensor data. These need to be handled differently from static data.

- **Global and interconnected data**: Data is often spread across different locations or sources, especially on the internet. Mining systems must be able to work with interconnected, diverse datasets like web data or social networks.

5. Data Mining and Society

Data mining has a major impact on society, both positive and negative. Some concerns include:

- **Social impact**: While data mining can bring great benefits (like helping businesses improve customer experiences), it also raises ethical issues, such as misuse of data and privacy concerns.

- **Privacy-preserving data mining**: It's essential to ensure that personal information is not compromised when data mining is used. Researchers are working on methods to protect privacy while still allowing for effective mining.

- **Invisible data mining**: Data mining is often hidden from users. For example, online shopping sites use data mining to recommend products without users knowing it. This raises questions about how much data we should collect and how transparent the process should be.