Al-Mustaqbal University
College of Sciences
Intelligent Medical System Department

جامعة المستقبل
AL MUSTAQBAL UNIVERSITY

كلية العلوم
قسم الانظمة الطبية الذكية

Lecture: (3)

Data Mining Overview

Subject: Clinical Data Mining
Level: Four
Lecturer: Dr. Maytham Nabeel Meqdad

Study Year: 2025-2026

Data Mining Overview

What Is Data Mining?

Data mining uses advanced algorithms and computing techniques to sift through large volumes of raw data, uncovering patterns and extracting valuable insights. Organizations leverage data mining to understand their customers better, enhance marketing strategies, increase sales, and cut costs effectively. By relying on solid data collection, warehousing, and processing, data mining transforms disparate data points into actionable intelligence, playing a crucial role in modern decision-making processes across various sectors.

Key Takeaways

- Data mining involves analyzing large datasets to identify patterns and extract valuable insights, enhancing business strategies like marketing and fraud detection.
- The data mining process consists of several critical steps, including understanding the business problem, preparing data, building models, and implementing change based on insights.
- Various data mining techniques, such as classification, clustering, and predictive analysis, help in transforming raw data into actionable intelligence.
- Data mining has broad applications across industries, including sales, marketing, manufacturing, fraud detection, and human resources, helping organizations improve efficiency and decision-making.
- While data mining can offer significant advantages by uncovering hidden trends, it also poses challenges such as complexity and potential privacy violations, as seen in the Facebook-Cambridge Analytica scandal.

Understanding the Mechanics of Data Mining

Data mining involves exploring and analyzing large blocks of information to glean meaningful patterns and trends. It's used in credit risk management, fraud detection, spam filtering, and as a market research tool to uncover group sentiments and opinions.

The data mining process breaks down into four steps:

1. Data is collected and loaded into data warehouses on-site or on a cloud service.
2. Business analysts, management teams, and information technology professionals access the data and determine how they want to organize it.
3. Custom application software sorts and organizes the data.
4. The end user presents the data in an easy-to-share format, such as a graph or table.

Exploring Data Warehousing and Mining Tools

Data mining programs analyze relationships and patterns in data based on user requests. It organizes information into classes.

For example, a restaurant may want to use data mining to determine which specials it should offer and on what days. The data can be organized into classes based on when customers visit and what they order. Data miners also identify clusters, associations, and patterns to understand trends in consumer behavior.

Warehousing is an important aspect of data mining. Warehousing centralizes an organization's data in one database, enabling specific user analysis and usage.

Fast Fact

Cloud data warehouse solutions use the space and power of a cloud provider to store data. This allows smaller companies to leverage digital solutions for storage, security, and analytics.

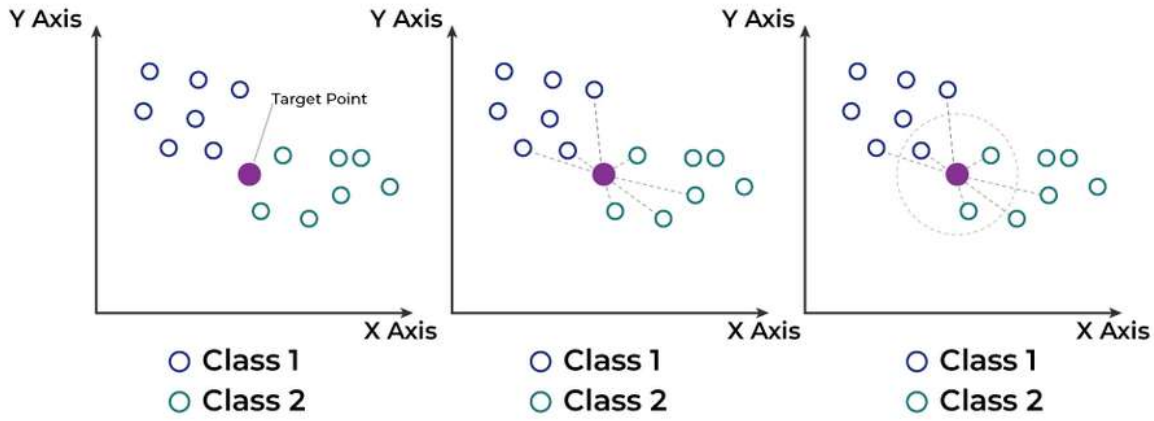
Key Techniques in Data Mining

Data mining uses algorithms and various other techniques to convert large collections of data into useful output. The most popular types of data mining techniques include association rules, classification, clustering, decision trees, K-Nearest Neighbor, neural networks, and predictive analysis.

- **Association rules**, also referred to as market basket analysis, search for relationships between variables. This relationship in itself creates additional value within the data set as it strives to link pieces of data. For example, association rules would search a company's sales history to see which products are most commonly purchased together; with this information, stores can plan, promote, and forecast.
- **Classification** uses predefined classes to assign to objects. These classes describe the characteristics of items or represent what the data points have in common with each other. This data mining technique allows the underlying data to be more neatly categorized and summarized across similar features or product lines.
- **Clustering** is similar to classification. However, clustering identifies similarities between objects, then groups those items based on what makes them different from other items. While classification may result in groups such as "shampoo," "conditioner," "soap," and "toothpaste," clustering may identify groups such as "hair care" and "dental health."
- **Decision trees** are used to classify or predict an outcome based on a set list of criteria or decisions. A decision tree is used to ask for the input of a series of cascading questions that sort the dataset based on the responses given. Sometimes depicted as a tree-like

visual, a decision tree allows for specific direction and user input when drilling deeper into the data.

- **K-Nearest Neighbor (KNN)** is an algorithm that classifies data based on its proximity to other data. The basis for KNN is rooted in the assumption that data points that are close to each other are more similar to each other than other bits of data. This non-parametric, supervised technique is used to predict the features of a group based on individual data points.



- **Neural networks** process data through the use of nodes. These nodes are comprised of inputs, weights, and an output. Data is mapped through supervised learning, similar to how the human brain is interconnected. This model can be programmed to give threshold values to determine a model's accuracy.
- **Predictive analysis** strives to leverage historical information to build graphical or mathematical models to forecast future outcomes. Overlapping with regression analysis, this technique aims to support an unknown figure in the future based on current data on hand.

Step-by-Step Guide to the Data Mining Process

To be most effective, data analysts generally follow a certain flow of tasks along the data mining process. Without this structure, an analyst may encounter an issue in the middle of their analysis that could have easily been prevented had they prepared for it earlier. The data mining process is usually broken into the following steps.

Step 1: Understand the Business

Before any data is touched, extracted, cleaned, or analyzed, it is important to understand the underlying entity and the project at hand. What are the goals the company is trying to achieve by mining data? What is their current business situation? What are the findings of a SWOT

analysis? Before looking at any data, the mining process starts by understanding what will define success at the end of the process.

Step 2: Understand the Data

After defining the business problem, consider data sources, security, storage, collection methods, and potential analysis outcomes. This step also includes determining the limits of the data, storage, security, and collection and assessing how these constraints will affect the data mining process.

Step 3: Prepare the Data

Data is gathered, uploaded, extracted, or calculated. It is then cleaned, standardized, scrubbed for outliers, assessed for mistakes, and checked for reasonableness. During this stage of data mining, the data may also be checked for size, as an oversized collection of information may unnecessarily slow computations and analysis.

Step 4: Build the Model

With a clean data set in hand, it's time to crunch the numbers. Data scientists use the types of data mining above to search for relationships, trends, associations, or sequential patterns. Data can be used in predictive models to see how past information might lead to future outcomes.

Step 5: Evaluate the Results

The data mining process ends by evaluating the findings of the data models. The outcomes from the analysis may be aggregated, interpreted, and presented to decision-makers who have largely been excluded from the data mining process to this point. In this step, organizations can choose to make decisions based on the findings.

Step 6: Implement Change and Monitor

The data mining process concludes with management taking steps in response to the findings of the analysis. The company may decide the information was not strong enough or the findings were not relevant, or the company may strategically pivot based on findings. In either case, management reviews the ultimate impacts of the business and recreates future data mining loops by identifying new business problems or opportunities.

Important

Different data mining processing models will have different steps, though the general process is usually pretty similar. For example, the Knowledge Discovery Databases model has nine steps, the CRISP-DM model has six steps, and the SEMMA process model has five steps.

Practical Applications of Data Mining Across Industries

In today's age of information, almost any department, industry, sector, or company can make use of data mining.

Sales

Data mining encourages smarter, more efficient use of capital to drive revenue growth. Consider the point-of-sale register at your favorite local coffee shop. For every sale, that coffeehouse collects the time a purchase was made and what products were sold. Using this information, the shop can strategically craft its product line.

Marketing

Once the coffeehouse knows its ideal line-up, it's time to implement the changes. However, to make its marketing efforts more effective, the store can use data mining to understand where its clients see ads, what demographics to target, where to place digital ads, and what marketing strategies most resonate with customers. This involves tailoring marketing campaigns, promotions, and cross-sell offers based on data mining insights.

Manufacturing

For companies that produce their own goods, data mining plays an integral part in analyzing how much each raw material costs, what materials are being used most efficiently, how time is spent along the manufacturing process, and what bottlenecks negatively impact the process. Data mining helps ensure the flow of goods is uninterrupted.

Fraud Detection

The heart of data mining is finding patterns, trends, and correlations that link data points together. Therefore, a company can use data mining to identify outliers or correlations that should not exist. For example, a company may analyze its cash flow and find a recurring transaction to an unknown account. If this is unexpected, the company may wish to investigate whether funds are being mismanaged.

Human Resources

Human resources departments often have a wide range of data available for processing, including data on retention, promotions, salary ranges, company benefits, use of those benefits, and employee satisfaction surveys. Data mining can correlate this data to get a better understanding of why employees leave and what entices new hires.

Customer Service

Customer satisfaction may be caused (or destroyed) by many events or interactions. For a shipping company, a customer might be unhappy with delivery times, quality, or communication. The same customer may be frustrated with long telephone wait times or slow e-mail responses. Data mining gathers operational information about customer interactions and summarizes the findings to pinpoint weak points and highlight what the company is doing right.

Weighing the Pros and Cons of Data Mining

Pros of Data Mining

- It drives profitability and efficiency
- It can be applied to any type of data and business problem
- It can reveal hidden information and trends

Cons of Data Mining

- It is complex
- Results and benefits are not guaranteed
- It can be expensive

Pros Explained

- **Profitability and efficiency**: Data mining ensures a company is collecting and analyzing reliable data. It is often a more rigid, structured process that formally identifies a problem, gathers data related to the problem, and strives to formulate a solution. Therefore, data mining helps a business become more profitable, more efficient, or operationally stronger.
- **Wide applications**: Data mining can look very different across applications, but the overall process can be used with almost any new or legacy application. Essentially, any type of data can be gathered and analyzed, and almost every business problem that relies on quantifiable evidence can be tackled using data mining.
- **Hidden information and trends**: The end goal of data mining is to take raw bits of information and determine if there is cohesion or correlation among the data. This benefit

of data mining allows a company to create value with the information they have on hand that would otherwise not be overly apparent. Though data models can be complex, they can also yield fascinating results, unearth hidden trends, and suggest unique strategies.

Cons Explained

- **Complexity**: The complexity of data mining is one of its greatest disadvantages. Data analytics often requires technical skill sets and certain software tools. Smaller companies may find this to be a barrier to entry too difficult to overcome.
- **No guarantees**: Data mining doesn't always mean guaranteed results. A company may perform statistical analysis, make conclusions based on strong data, implement changes, and not reap any benefits. This may be due to inaccurate findings, market changes, model errors, or inappropriate data populations. Data mining can only guide decisions and not ensure outcomes.
- **High cost**: There is also a cost component to data mining. Data tools may require costly subscriptions, and some data may be expensive to obtain. Security and privacy concerns can be pacified, though additional IT infrastructure may be costly as well. Data mining may also be most effective when using huge data sets; however, these data sets must be stored and require heavy computational power to analyze.

Fast Fact

Even large companies or government agencies have challenges with data mining. Consider the FDA's white paper on data mining that outlines the challenges of bad information, duplicate data, underreporting, or overreporting.

The Impact of Data Mining on Social Media

One of the most lucrative applications of data mining has been undertaken by social media companies. Platforms like Facebook, TikTok, Instagram, and X (formerly Twitter) gather reams of data about their users based on their online activities.

That data can be used to make inferences about their preferences. Advertisers can target their messages to the people who appear to be most likely to respond positively.

Data mining on social media has become a big point of contention, with several investigative reports and exposés showing just how intrusive mining users' data can be. The main issue is users often agree to terms without knowing how their data is collected or sold.

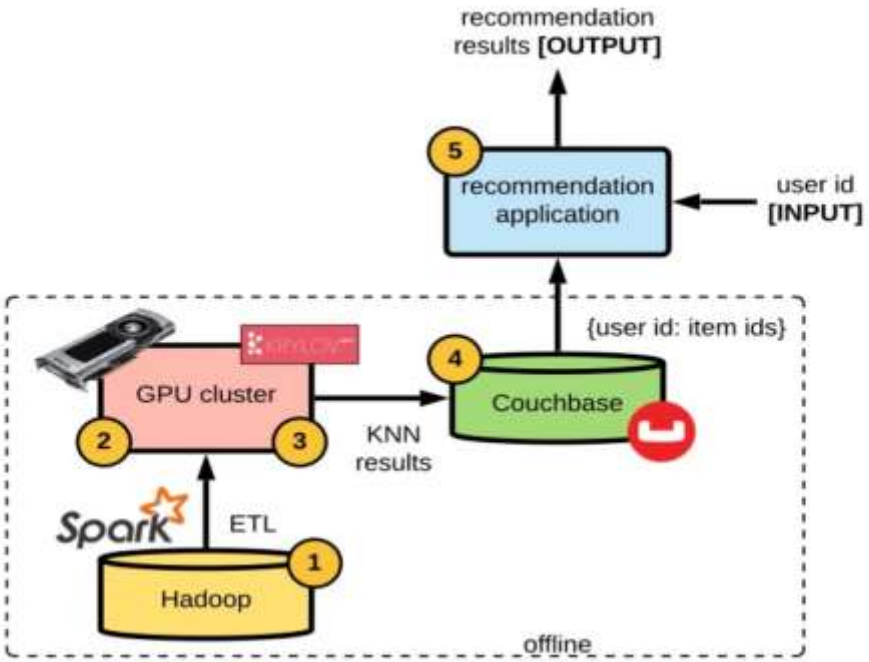
Examples of Data Mining

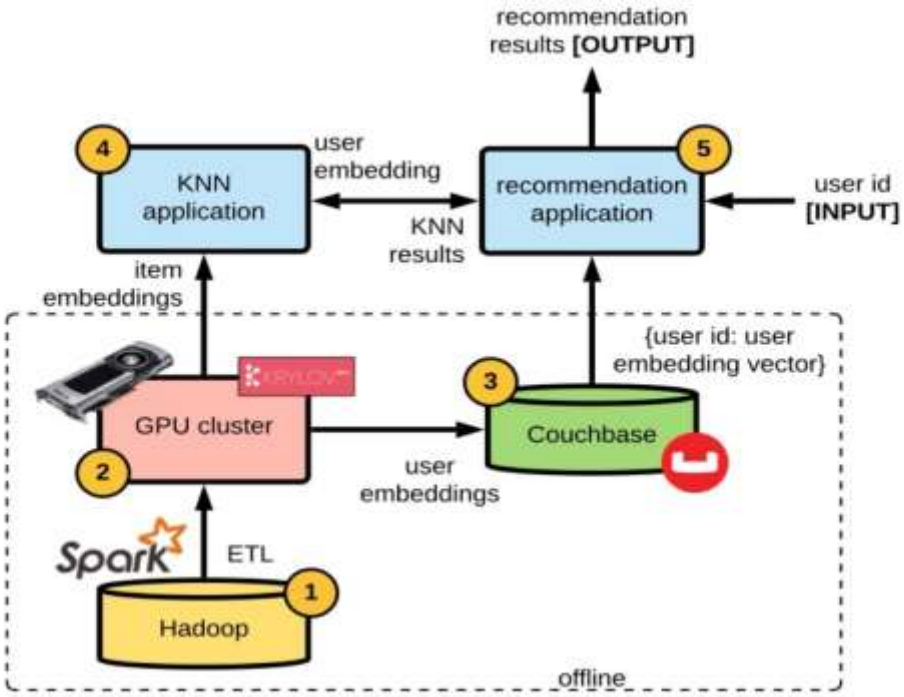
Data mining can be used for good, or it can be used illicitly. Here is an example of both.

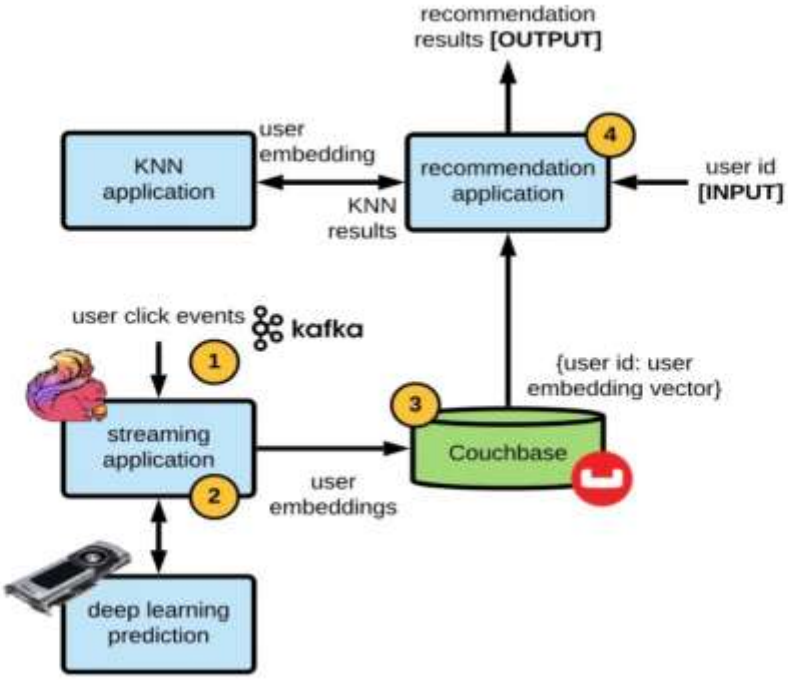
eBay and e-Commerce

eBay collects countless bits of information every day from sellers and buyers. The company uses data mining to attribute relationships between products, assess desired price ranges, analyze prior purchase patterns, and form product categories.3

eBay. "Building a Deep Learning Based Retrieval System for Personalized Recommendations."




recommendation
results [OUTPUT]
4
KNN
application
user
embedding
KNN
results
5
recommendation
application
user id
[INPUT]
item
embeddings
{user id: user
embedding vector}
GPU cluster
2
3
Couchbase
user
embeddings
Spark
ETL
1
Hadoop
offline


recommendation
results [OUTPUT]
KNN
application
user
embedding
KNN
results
4
recommendation
application
user id
[INPUT]
user click events
kafka
1
{user id: user
embedding vector}
streaming
application
2
3
Couchbase
user
embeddings
deep learning
prediction

eBay outlines the recommendation process as:

1. Raw item metadata and user historical data are aggregated.
2. Scripts are run on a trained model to generate and predict the item and user.
3. A KNN search is performed.
4. The results are written to a database.
5. The real-time recommendation takes the user ID, calls the database results, and displays them to the user.3

Facebook-Cambridge Analytica Scandal

A cautionary example of data mining is the Facebook-Cambridge Analytica data scandal. During the 2010s, the British consulting firm Cambridge Analytica Ltd. collected personal data from millions of Facebook users. This information was later analyzed for use in the 2016 presidential campaigns of Ted Cruz and Donald Trump. It is suspected that Cambridge Analytica interfered with other notable events such as the Brexit referendum.4

In light of this inappropriate data mining and misuse of user data, Facebook agreed to pay $100 million for misleading investors about its use of consumer data. The Securities and Exchange Commission claimed Facebook discovered the misuse in 2015 but did not correct its disclosures for more than two years.

What Are the Types of Data Mining?

There are two main types of data mining: predictive data mining and descriptive data mining. Predictive data mining extracts data that may be helpful in determining an outcome. Description data mining informs users of a given outcome.

How Is Data Mining Done?

Data mining relies on big data and advanced computing processes, including machine learning and other forms of artificial intelligence (AI). The goal is to find patterns that can lead to inferences or predictions from large and unstructured data sets.

What Is Another Term for Data Mining?

Data mining also goes by the less-used term "knowledge discovery in data," or KDD.

Where Is Data Mining Used?

Data mining applications have been designed to take on just about any endeavor that relies on big data. Companies in the financial sector look for patterns in the markets. Governments try to identify potential security threats. Corporations, especially online and social media companies, use data mining to create profitable advertising and marketing campaigns that target specific sets of users.

The Bottom Line

Modern businesses have the ability to gather information on their customers, products, manufacturing lines, employees, and storefronts. These random pieces of information may not tell a story, but the use of data mining techniques, applications, and tools helps piece together information.

The ultimate goal of the data mining process is to compile data, analyze the results, and execute operational strategies based on data mining results.

References

[1] eBay https://innovation.ebayinc.com/stories/building-a-deep-learning-based-retrieval-system-for-personalized-recommendations/

[2] .Investopedia https://www.investopedia.com/terms/d/datamining.asp

[3] Han and M. Kamber, " Data Mining Tools and Technique s", Morgan Kaufmann Publishers.

[4] .M.H. Dunham, " Data Mining Introductory and Adv anced Topics", Pear son Education