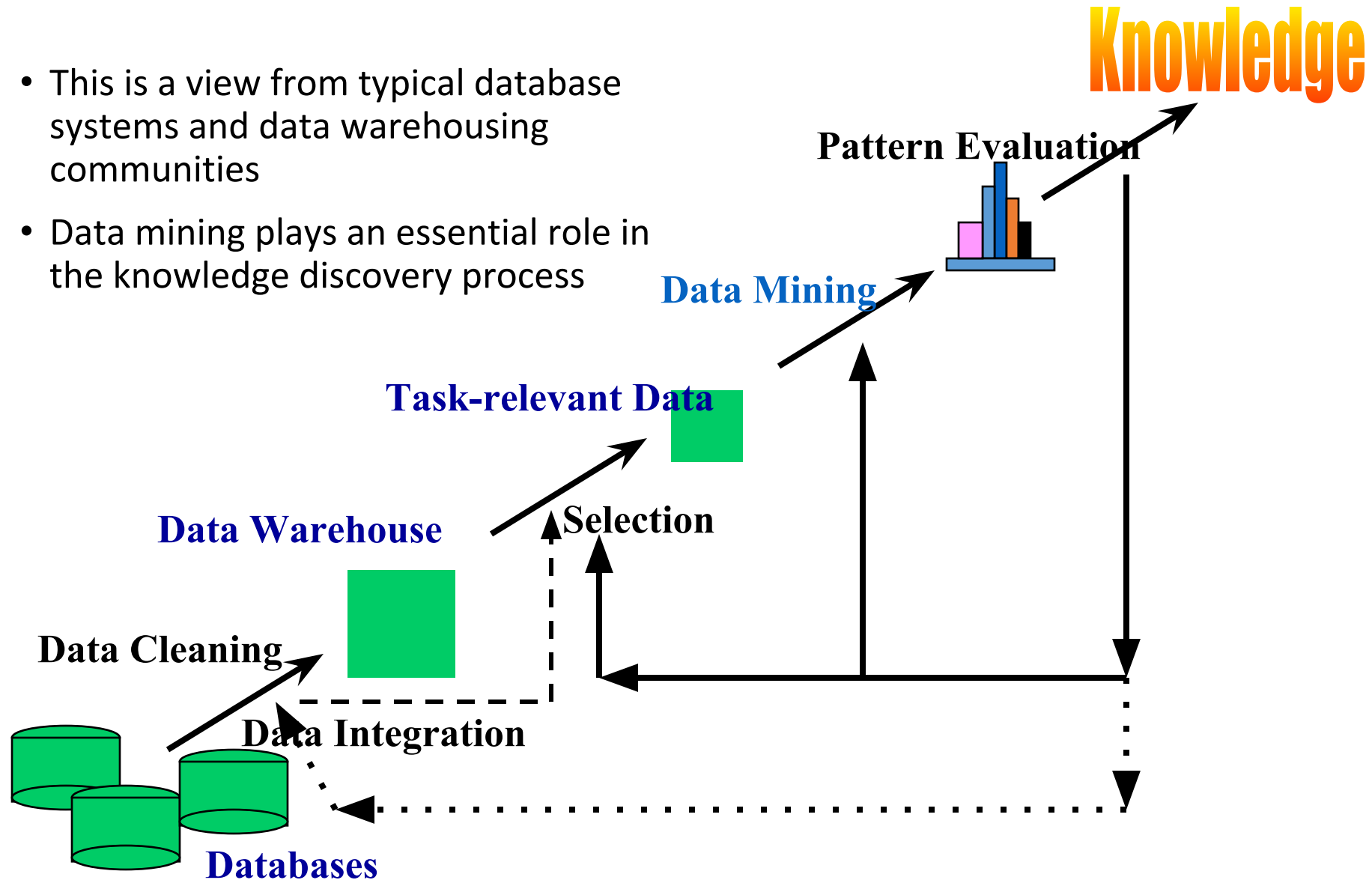


# **Chapter 2**

## **Data Understanding and Pre-processing**

# Knowledge Discovery (KDD) Process

- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process



# Knowledge Discovery (KDD) Process

- **Data Cleaning** (to remove noise or irrelevant data),
- **Data Integration** (where multiple data sources may be combined)
- **Data Selection** (where data relevant to the analysis task are retrieved from the database),
- **Data Transformation** (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance) ,
- **Data Mining** (an essential process where intelligent methods are applied in order to extract data patterns),
- **Pattern Evaluation** (to identify the truly interesting patterns representing knowledge based on some interestingness measures), and
- **Knowledge Presentation** (where visualization and knowledge representation techniques are used to present the mined knowledge to the user).

# Data Objects

- Data sets are made up of data objects.
- A **data object** represents an entity.
- Examples:
  - sales database: customers, store items, sales
  - medical database: patients, treatments
  - university database: students, professors, courses
- Also called *samples* , *examples*, *instances*, *data points*, *objects*, *tuples*.
- Data objects are described by **attributes**.
- Database rows -> data objects; columns -> attributes.

# Attributes

- **Attribute (or dimensions, features, variables):** a data field, representing a characteristic or feature of a data object.
  - *E.g., customer\_ID, name, address*
- **Types:**
  - Nominal
  - Binary
  - Numeric: quantitative
    - Interval-scaled
    - Ratio-scaled

# Attribute Types

- **Nominal:** categories, states, or “names of things”
  - *Hair\_color* = {auburn, black, blond, brown, grey, red, white}
  - marital status, occupation, ID numbers, zip codes
- **Binary**
  - Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important
    - e.g., gender
  - Asymmetric binary: outcomes not equally important.
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., HIV positive)
- **Ordinal**
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - *Size* = {small, medium, large}, grades, army rankings

# Numeric Attribute Types

- Quantity (integer or real-valued)
- **Interval**
  - Measured on a scale of **equal-sized units**
  - Values have order
    - E.g., *temperature in C° or F°, calendar dates*
  - No true zero-point
- **Ratio**
  - Inherent **zero-point**
  - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K° is twice as high as 5 K°).
    - e.g., *temperature in Kelvin, length, counts, monetary quantities*

# Discrete vs. Continuous Attributes

- **Discrete Attribute**

- Has only a finite or countably infinite set of values
  - E.g., zip codes, profession, or the set of words in a collection of documents
- Sometimes, represented as integer variables
- Note: Binary attributes are a special case of discrete attributes

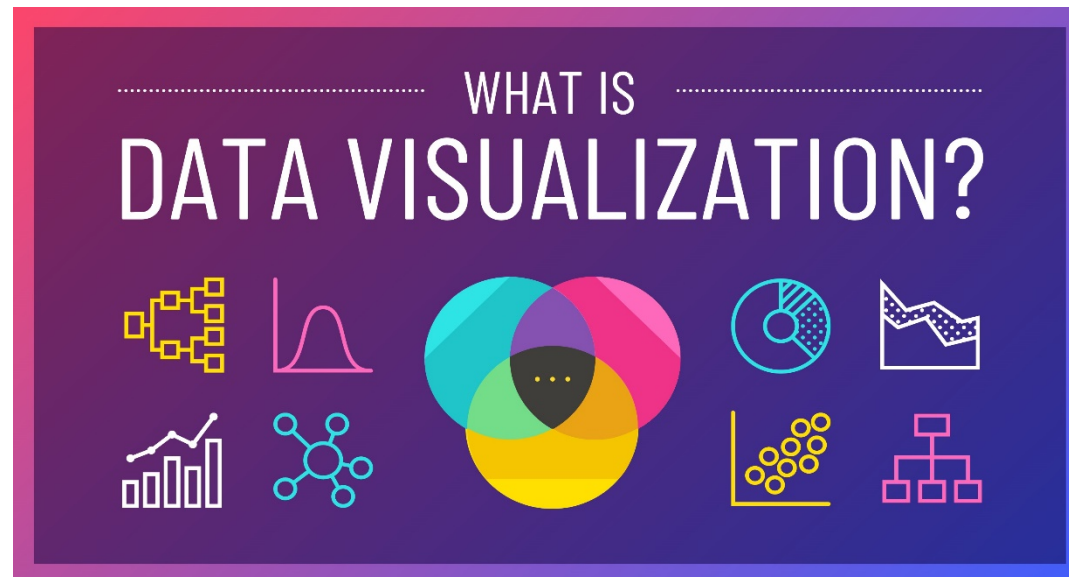
- **Continuous Attribute**

- Has real numbers as attribute values
  - E.g., temperature, height, or weight
- Practically, real values can only be measured and represented using a finite number of digits
- Continuous attributes are typically represented as floating-point variables



# Data Visualization

- Data visualization is the graphical representation of information and data. By using [visual elements like charts, graphs, and maps](#), data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.
- In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions.
- Data visualization is applied in practically every field of knowledge. Scientists in various disciplines use computer techniques to model complex events and visualize phenomena that cannot be observed directly, such as weather patterns, medical conditions or mathematical relationships.



# Data Visualization

- Why data visualization?
  - Gain insight into an information space by mapping data onto graphical primitives
  - Provide qualitative overview of large data sets
  - Search for patterns, trends, structure, irregularities, relationships among data
  - Help find interesting regions and suitable parameters for further quantitative analysis
  - Provide a visual proof of computer representations derived
- Categorization of visualization methods:
  - Pixel-oriented visualization techniques
  - Geometric projection visualization techniques
  - Icon-based visualization techniques
  - Hierarchical visualization techniques
  - Visualizing complex data and relations

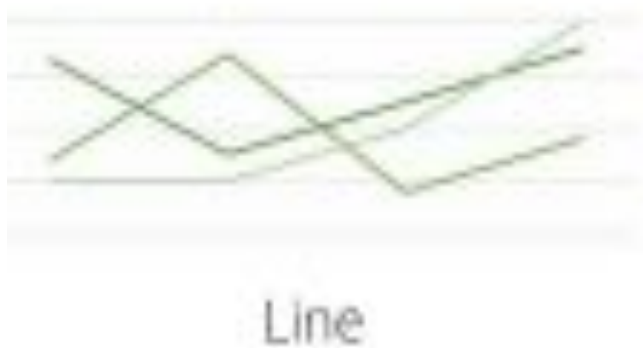
# Data Visualization

- Data visualization provides an important suite of tools and techniques for gaining a qualitative understanding.
- The basic techniques are the following plots:
  1. Charts
  2. Plots
  3. Maps
  4. Diagrams & Matrices

# Data Visualization-Charts

## Charts

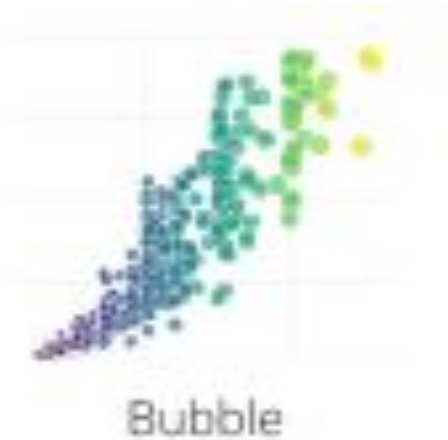
The easiest way to show the development of one or several data sets is a chart. Charts vary from bar and line charts that show the relationship between elements over time to pie charts that demonstrate the components or proportions between the elements of one whole.



# Data Visualization-Plots

## Plots

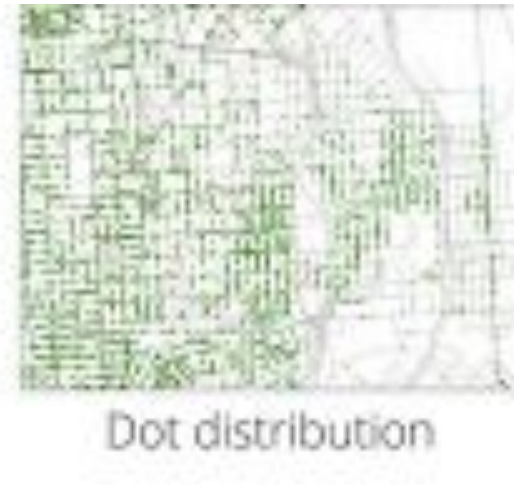
Plots allow to distribute two or more data sets over a 2D or even 3D space to show the relationship between these sets and the parameters on the plot. Plots also vary. Scatter and bubble plots are some of the most widely-used visualizations. When it comes to big data, analysts often use more complex box plots that help visualize the relationship between large volumes of data.



# Data Visualization-Maps

## Maps

Maps are popular ways to visualize data used in different industries. They allow to locate elements on relevant objects and areas — geographical maps, building plans, website layouts, etc. Among the most popular map visualizations are heat maps, dot distribution maps, cartograms.



# Data Visualization-Plots

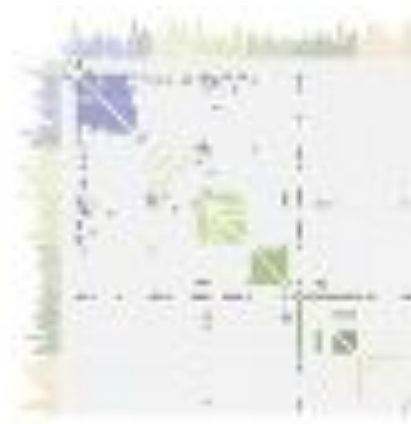
## **Diagrams and matrices**

Diagrams are usually used to demonstrate complex data relationships and links and include various types of data on one visualization. They can be hierarchical, multidimensional, tree-like.

Matrix is one of the advanced data visualization techniques that help determine the correlation between multiple constantly updating (steaming) data sets.



Tree



Matrix

# Data Visualization-Importance

- Today, data visualization has become a rapidly evolving blend of science and art that is certain to change the corporate landscape over the next few years.
- They go on to note, “Because of the way the human brain processes information, using charts or graphs to visualize large amounts of complex data is easier than poring over spreadsheets or reports. ...
- Data visualization can also: Identify areas that need attention or improvement; clarify which factors influence customer behavior; help you understand which products to place where; [and] predict sales volumes.”



# Data Visualization-Importance

- **1. Recognizing patterns and trends.** Balliett writes, “A spreadsheet of raw data just doesn’t cut it if you’re looking for real insights. That’s where graphs, charts, maps, and other types of data viz can help. When your data is visualized accurately and thoughtfully, you’ll no doubt begin to notice trends and patterns instantly.”

# Data Visualization-Importance

- **2. Asking the right questions.** Over the years, I have repeatedly argued better questions result in better analysis. Balliett asserts once insights, trends, and patterns are visually apparent, “you’re better equipped to ask the kinds of questions that will make a real difference for the future of your business — questions that the most adept business leaders ask on a daily basis. But they wouldn’t even know what questions to ask without powerful data visualization to point them in the right direction.”

# Data Visualization-Importance

- **3. Making better plans for the future.** Good questions lead to good plans. Balliett writes, “Now that you’re asking the right questions, you’re empowered to make data-driven decisions about the growth and direction of your business.” Every business leader knows the importance of making good decisions. As Bain analysts, Michael C. Mankins and Lori Sherer ([@lorisherer](#)), explain, “The best way to understand any company’s operations is to view them as a series of decisions. ... We know from extensive research that decisions matter — a lot. Companies that make better decisions, make them faster and execute them more effectively than rivals nearly always turn in better financial performance. Not surprisingly, companies that employ advanced analytics to improve decision making and execution have the results to show for it.”


# Data Visualization-Importance

- **4. Improving customer experience.** Balliett notes, “The Experian report found that an impressive 98 percent of those surveyed agreed that data helps them better the customer experience. ... Data about customer demand and behavior — including what they look for in a product like yours, and how they use your product when they buy it — can be incredibly powerful. It helps you recognize where you’re succeeding and where you can do more. It allows you to give your past and future customers exactly what they’re looking for.”

# Data Visualization-Importance

- **5. Earning your audience's trust.** A nineteenth century aphorism often incorrectly attributed to British Prime Minister Benjamin Disraeli is, "There are three kinds of lies: lies, damned lies, and statistics." The point is, gaining trust, even with data, is not always an easy thing to do. Balliett writes, "Sharing data that supports your claims makes them all the more powerful. And in the process, it builds trust with customers and potential customers. That's because it makes your company look more transparent and proves that your claims are evidence-based."
- The right data visualization will not only strongly underscore your message but ensure your audience that the numbers are sound. The importance of trust in business can't be over-emphasized. Although the following video from Egencia focuses on travel, it makes some excellent points about how data visualization can help data come to life.

# Data Preprocessing

- Data Preprocessing: An Overview 
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization
- Summary

# Data Quality: Why Preprocess the Data?


- Measures for data quality: A multidimensional view
  - Accuracy: correct or wrong, accurate or not
  - Completeness: not recorded, unavailable, ...
  - Consistency: some modified but some not, dangling, ...
  - Timeliness: timely update?
  - Believability: how trustable the data are correct?
  - Interpretability: how easily the data can be understood?

# Major Tasks in Data Preprocessing

- **Data cleaning**
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data integration**
  - Integration of multiple databases, data cubes, or files
- **Data reduction**
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- **Data transformation and data discretization**
  - Normalization
  - Concept hierarchy generation



# Data Preprocessing

- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning 
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization
- Summary

# Data Cleaning

- Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., *Occupation*=" " (missing data)
  - noisy: containing noise, errors, or outliers
    - e.g., *Salary*="−10" (an error)
  - inconsistent: containing discrepancies in codes or names, e.g.,
    - *Age*="42", *Birthday*="03/07/2010"
    - Was rating "1, 2, 3", now rating "A, B, C"
    - discrepancy between duplicate records
  - Intentional (e.g., *disguised missing* data)
    - Jan. 1 as everyone's birthday?

# Incomplete (Missing) Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data
- Missing data may need to be inferred

# How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
- Fill in the missing value manually: tedious + infeasible?
- Fill in it automatically with
  - a global constant : e.g., “unknown”, a new class?!
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: smarter
  - the most probable value: inference-based such as Bayesian formula or decision tree

# Noisy Data

- **Noise**: random error or variance in a measured variable
- **Incorrect attribute values** may be due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention
- **Other data problems** which require data cleaning
  - duplicate records
  - incomplete data
  - inconsistent data


# How to Handle Noisy Data?

- Binning
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Regression
  - smooth by fitting the data into regression functions
- Clustering
  - detect and remove outliers
- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)

# Data Cleaning as a Process

- Data discrepancy detection
  - Use metadata (e.g., domain, range, dependency, distribution)
  - Check field overloading
  - Check uniqueness rule, consecutive rule and null rule
  - Use commercial tools
    - Data scrubbing: use simple domain knowledge (e.g., postal code, spell-check) to detect errors and make corrections
    - Data auditing: by analyzing data to discover rules and relationship to detect violators (e.g., correlation and clustering to find outliers)
- Data migration and integration
  - Data migration tools: allow transformations to be specified
  - ETL (Extraction/Transformation/Loading) tools: allow users to specify transformations through a graphical user interface
- Integration of the two processes
  - Iterative and interactive (e.g., Potter's Wheels)

# Data Preprocessing

- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration 
- Data Reduction
- Data Transformation and Data Discretization
- Summary




# Data Integration

- **Data integration:**
  - Combines data from multiple sources into a coherent store
- Schema integration: e.g.,  $A.cust-id \equiv B.cust-\#$ 
  - Integrate metadata from different sources
- **Entity identification problem:**
  - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton
- Detecting and resolving data value conflicts
  - For the same real world entity, attribute values from different sources are different
  - Possible reasons: different representations, different scales, e.g., metric vs. British units

# Handling Redundancy in Data Integration

- Redundant data occur often when integration of multiple databases
  - *Object identification*: The same attribute or object may have different names in different databases
  - *Derivable data*: One attribute may be a “derived” attribute in another table, e.g., annual revenue
- Redundant attributes may be able to be detected by *correlation analysis* and *covariance analysis*
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

# Data Preprocessing

- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction 
- Data Transformation and Data Discretization
- Summary

# Data Reduction Strategies

- **Data reduction:** Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results
- Why data reduction? — A database/data warehouse may store terabytes of data. Complex data analysis may take a very long time to run on the complete data set.
- Data reduction strategies
  - **Dimensionality reduction**, e.g., remove unimportant attributes
    - Wavelet transforms
    - Principal Components Analysis (PCA)
    - Feature subset selection, feature creation
  - **Numerosity reduction** (some simply call it: Data Reduction)
    - Regression and Log-Linear Models
    - Histograms, clustering, sampling
    - Data cube aggregation
  - **Data compression**

# What Is Data Mining?

- Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events. Data mining is also known as Knowledge Discovery in Data (KDD).



# What Is Data Mining?

- Data mining refers to extracting or “mining” knowledge from large amounts of data
- the mining of gold from rocks or sand is referred to as gold mining rather than rock or sand mining. Thus, data mining" should have been more appropriately named knowledge mining from data", which is unfortunately somewhat long. Knowledge mining", a shorter term, may not react the emphasis on mining from large amounts of data. Nevertheless, mining is a vivid term characterizing the process that and a small set of precious nuggets from a great deal of raw material

# Example: Medical Data Mining

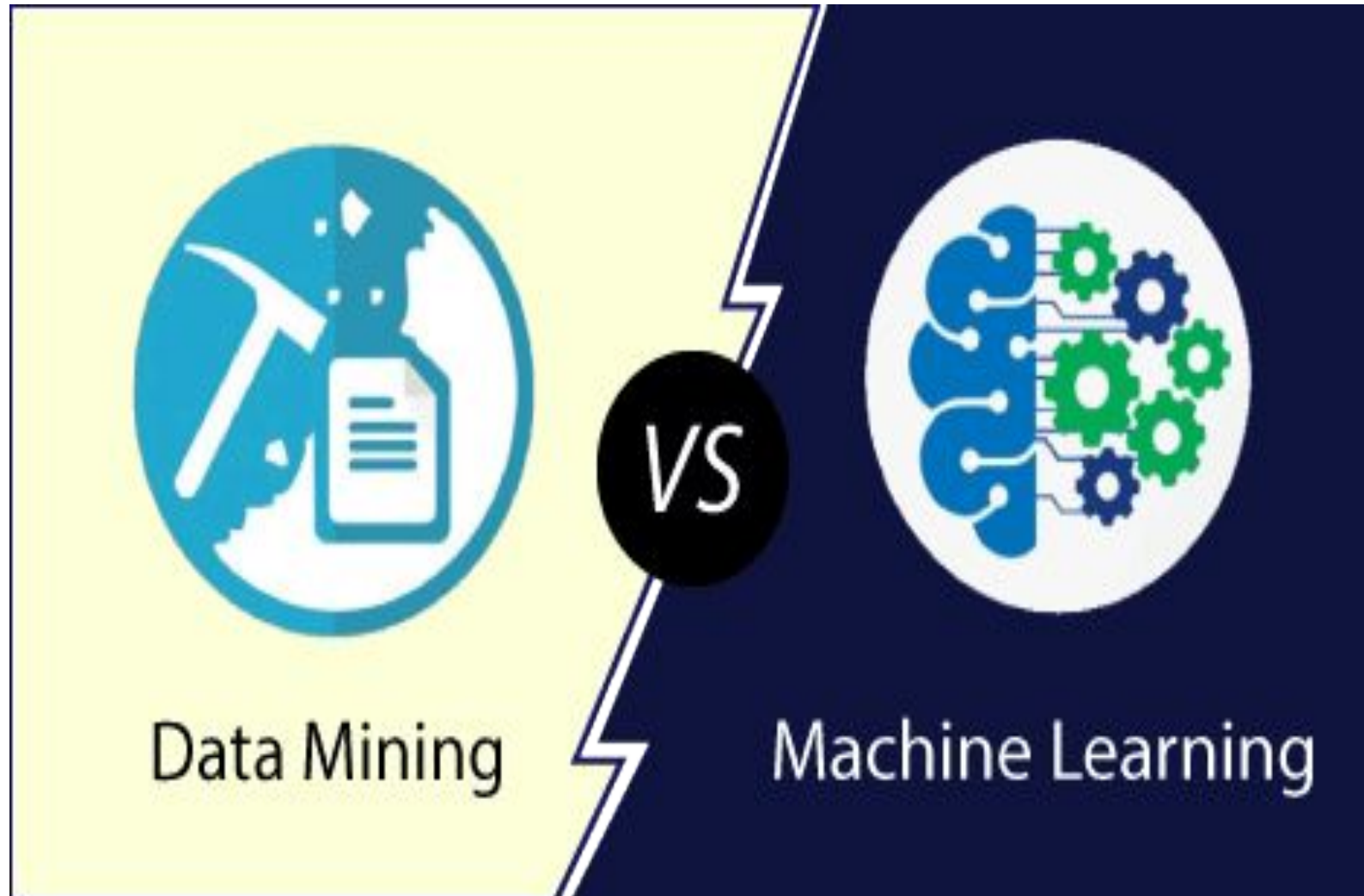
- Health care & medical data mining – often adopted such a view in statistics and machine learning
- Preprocessing of the data (including feature extraction and dimension reduction)
- Classification or/and clustering processes
- Post-processing for presentation

# Data Mining Tools

- Top 10 Data Mining Tools
1. [MonkeyLearn](#) | No-code text mining tools
  2. [RapidMiner](#) | Drag and drop workflows or data mining in Python
  3. [Oracle Data Mining](#) | Predictive data mining models
  4. [IBM SPSS Modeler](#) | A predictive analytics platform for data scientists
  5. [Weka](#) | Open-source software for data mining
  6. [Knime](#) | Pre-built components for data mining projects
  7. [H2O](#) | Open-source library offering data mining in Python
  8. [Orange](#) | Open-source data mining toolbox
  9. [Apache Mahout](#) | Ideal for complex and large-scale data mining
  10. [SAS Enterprise Miner](#) | Solve business problems with data mining



# Data Mining Vs. Machine Learning



# Data Mining Vs. Machine Learning

Factors	Data Mining	Machine Learning
<b>Origin</b>	Traditional databases with unstructured data.	It has an existing algorithm and data.
<b>Meaning</b>	Extracting information from a huge amount of data.	Introduce new Information from data as well as previous experience.
<b>History</b>	In 1930, it was known as knowledge discovery in databases(KDD).	The first program, i.e., Samuel's checker playing program, was established in 1950.
<b>Responsibility</b>	Data Mining is used to obtain the rules from the existing data.	Machine learning teaches the computer, how to learn and comprehend the rules.
<b>Abstraction</b>	Data mining abstract from the data warehouse.	Machine learning reads machine.
<b>Applications</b>	In compare to machine learning, data mining can produce outcomes on the lesser volume of data. It is also used in cluster analysis.	It needs a large amount of data to obtain accurate results. It has various applications, used in web search, spam filter, credit scoring, computer design, etc.
<b>Nature</b>	It involves human interference more towards the manual.	It is automated, once designed and implemented, there is no need for human effort.
<b>Techniques involve</b>	Data mining is more of research using a technique like a machine learning.	It is a self-learned and train system to do the task precisely.
<b>Scope</b>	Applied in the limited fields.	It can be used in a vast area.