Data Mining and Data Warehousing

Unit 1

Introduction


REAL-WORLD DATA MINING
SOCIAL MEDIA
CUSTOMER INSIGHTONS
USELFUL INSIGHTS
SALES TRANTHATIONS
USEFUL INSIGHTS
AND PRODUCT SUGGESTION

Why Data Mining? (*Motivation*)

- The Explosive Growth of Data: from terabytes to petabytes
- We are drowning in data, but starving for knowledge
- "***Necessity is the mother of invention***"—Data mining—Automated analysis of massive data sets

- **YouTube:**
  - Video Uploads: Over 500 hours of video are uploaded to YouTube every minute around 720,000 hours of video uploaded per day
  - Given YouTube's average video size (1-2 GB for a high-quality HD video), it is estimated that YouTube generates over **50 petabytes (PB)** of data daily.
  - View Data: YouTube users watch over 1 billion hours of content daily generating several petabytes

Facebook (Meta):

- Post Uploads: More than 350 million photos are uploaded to Facebook each day, and 5 million videos are shared daily.
- Data Generated: Each photo can be roughly 1 MB, and videos can range from 5-100 MB, resulting 150-200 TB of data per day from photos alone, while videos generate more than **50 TB of data** per day.

TikTok:

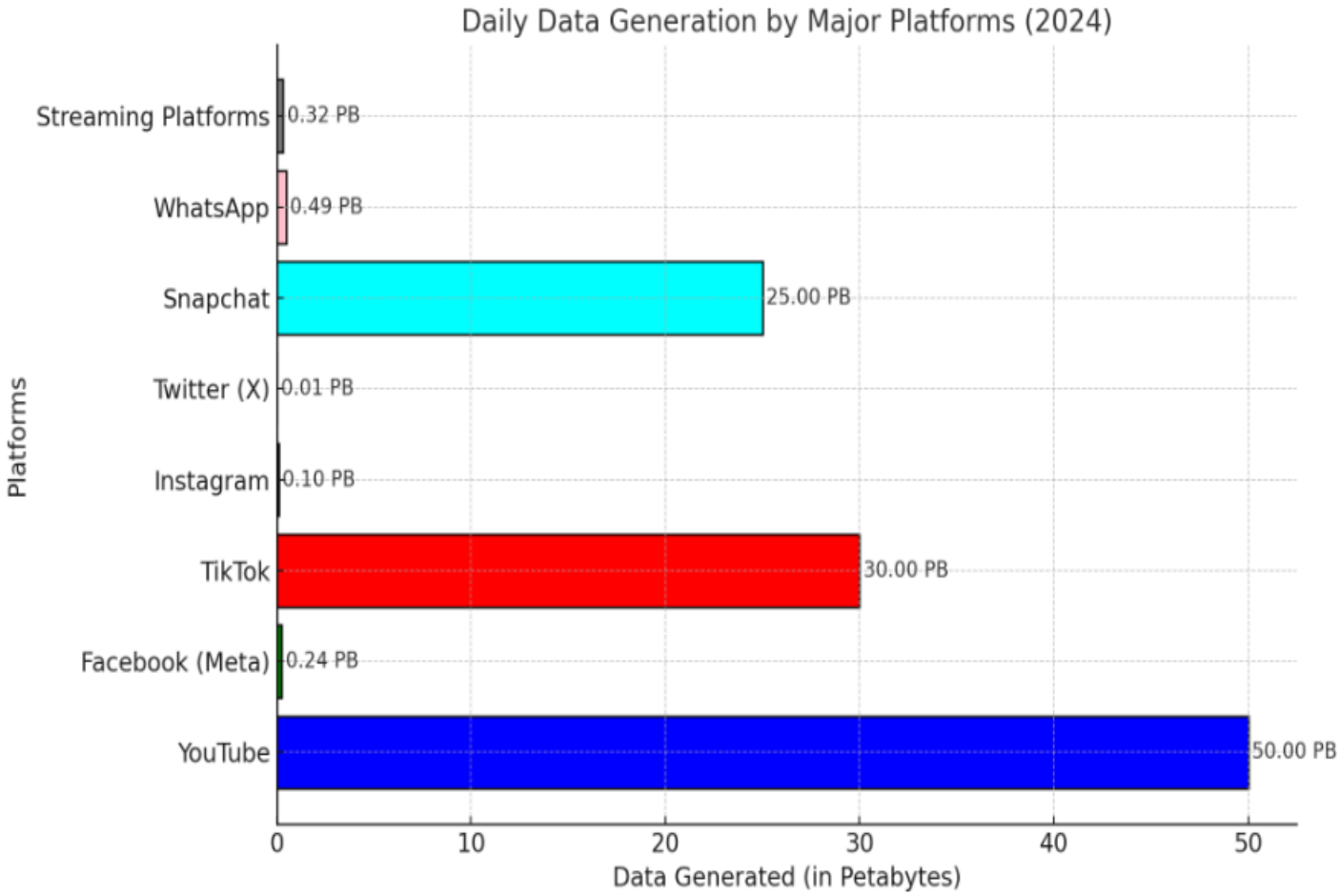
- Video Uploads: As of 2023, over 1 billion videos are watched daily on TikTok. Videos on TikTok are usually around 15 to 60 seconds long and can be anywhere from 20 MB to 100 MB in size, estimated **20 petabytes per day**

Instagram (also part of Meta):

- Post Uploads: Instagram users upload over 100 million photos and videos per day, generated from both photos and videos is estimated to be in the range of 20-50 TB per day.

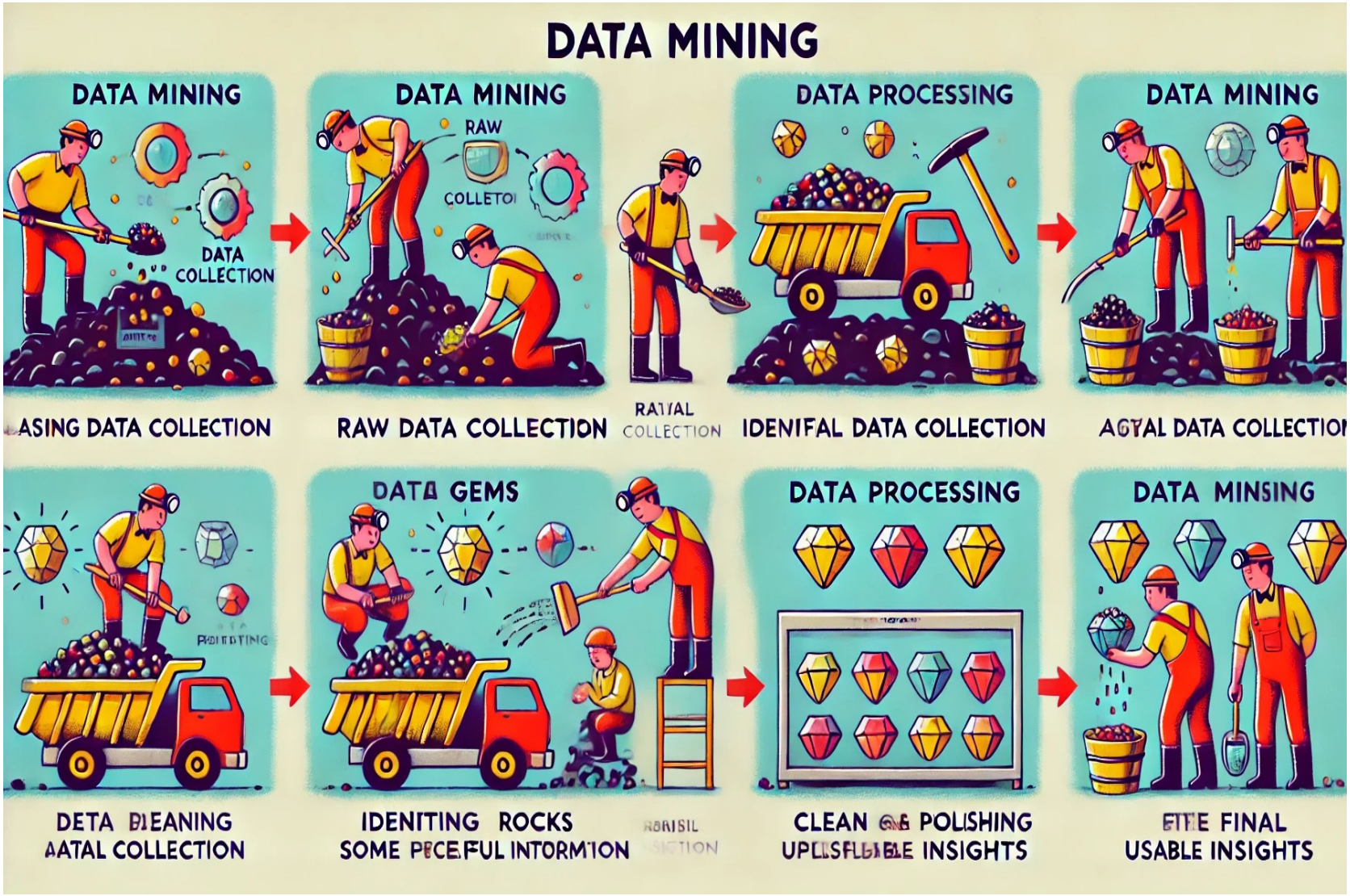
Snapchat:

- Snaps Sent: Snapchat users send over 4 billion snaps daily. Since each snap may include a short video (around 1 MB) or a photo (around 1-2 MB), Snapchat generates several petabytes of data each day


Daily Data Generation by Major Platforms (2024)
Platforms
Streaming Platforms
WhatsApp
Snapchat
Twitter (X)
Instagram
TikTok
Facebook (Meta)
YouTube
0.32 PB
0.49 PB
25.00 PB
0.01 PB
0.10 PB
30.00 PB
0.24 PB
50.00 PB
0
10
20
30
40
50
Data Generated (in Petabytes)

What is Data Mining?

- The process of Discovering meaningful patterns & trends often ***previously unknown,*** from large amount of data, using pattern recognition, statistical and mathematical techniques


DATA MINING
DATA MINING
DATA COLLECTION
DATA MINING
RAW
DATA PROCESSING
DATA MINING
RAW DATA COLLECTION
DATA GEMS
DATA PROCESSING
DATA MINSING
IDENITING ROCKS
CLEAN
POLISHING
INSIGHTS
FINAL
USABLE INSIGHTS


DATA MINING
You
WATCH HISTORY
PERSONALIZED RECOMMENATIONS
WATCH HISTORY
You
You
WATCH HISTORY
YOUTUBE
RECOMMENTIONS
SIMILAL VIDEOS
YOUTUBE
YouTube
YouTube
RECOMMEINATIONS

Why Data Mining?—Potential Applications

- Data analysis and decision support
  - Market analysis and management
    - Market basket analysis, sale techniques, customer feedback on items (Opinion Mining)
  - Risk analysis and management
    - Forecasting, decision support system
  - Fraud detection and detection of unusual patterns (outliers)
- Other Applications
  - Text mining (news group, email, documents) and Web mining
  - Stream data mining (mining from continuous / rapid data
    - Eg Telephone communication pattern, Web Searching, Sensor data
  - Bioinformatics and bio-data analysis


REAL-WORLD DATA MINING
SOCIAL MEDIA
RAW DATA STREAMS
REAL-WORLD DATA MINING
SALES
INSIGHTES
CUSTOMER
SALES
INSIGHTES
REAL-WORLD
INSIGHTS
MARKE
ACTIONABLE DECISIONS
ACTIOLE DECISIONS

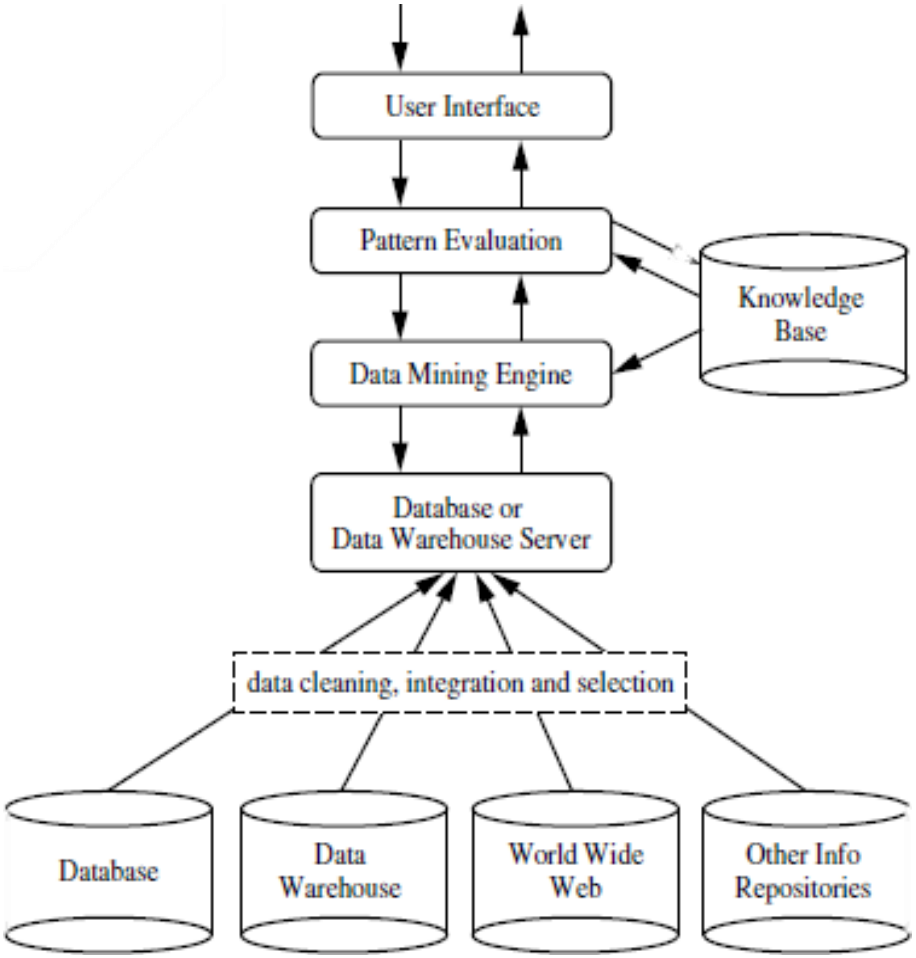
Data Mining: On What Kinds of Data?

- Database-oriented data sets and applications
  - Relational database, data warehouse, transactional database
- Advanced data sets and advanced applications
  - Data streams and sensor data
  - Time-series data
  - Structure data, graphs, social networks and multi-linked data
  - Heterogeneous databases and legacy databases
  - Spatial data and spatiotemporal data
  - Multimedia database
  - Text databases
  - The World-Wide Web

Datamining Architecture

1. **Graphic User Interface:**

   Since the user cannot fully understand the complexity of the data mining process so graphical user interface helps the user to communicate effectively with the data mining system.



Datamining Architecture

2. Pattern Evaluation:

This involves assessing the relevance and usefulness of the patterns in solving the problem at hand. This involves measuring the accuracy, reliability, and validity of the patterns and assessing their potential for use in decision-making. They are responsible for finding interesting patterns in the data and sometimes they also interact with the database servers for producing the result of the user requests. For example, if you have identified a segment of high-value customers, you may evaluate the profitability of targeting this segment with a loyalty program.

3. Data Sources:

The first step in data mining is to identify the relevant data sources that will be used for analysis. Database, World Wide Web(WWW), and data warehouse are parts of data sources. The data in these sources may be in the form of plain text, spreadsheets, or other forms of media like photos or videos. WWW is one of the biggest sources of data. For example, a marketing department might collect data from customer surveys, social media analytics, and web traffic logs.

Datamining Architecture

4.Data Cleaning and Preprocessing:

Before data can be analyzed, it must be cleaned and preprocessed to remove any errors, inconsistencies, or missing values. This involves tasks such as data integration, data transformation, data reduction, and data discretization.

a) **Data Cleaning:** Before data can be analyzed, it must be cleaned and preprocessed to remove any errors, inconsistencies, or missing values. For example, the marketing department might remove duplicate entries, correct misspellings, and fill in missing data with estimates or averages.

b) **Data Integration:** If the data is coming from multiple sources, it must be integrated into a single dataset. For example, the marketing department might merge customer survey responses with web traffic logs to create a more comprehensive dataset.

c) **Data Transformation:** Data may need to be transformed into a different format or structure to be analyzed. For example, the marketing department might convert text data into numerical data using natural language processing techniques.

d) **Data Reduction:** Large datasets may need to be reduced in size to make them easier to analyze. For example, the marketing department might filter out irrelevant data or focus on a specific subset of customers.

e) **Data Discretization:** Continuous data may need to be discretized into discrete categories or bins. For example, the marketing department might group customers into age ranges or income brackets.

Datamining Architecture

5. **Data Warehousing:**
   - The preprocessed data is then stored in a data warehouse or data mart, where it can be easily accessed and analyzed. The data warehouse is typically optimized for querying and analysis, with indexes and other performance-enhancing features. For example, the marketing department might store the preprocessed data in a data warehouse that is optimized for querying and analysis.
6. **Data Mining Engine:**
   - The data mining engine is the heart of the data mining system, responsible for carrying out the actual analysis. This involves using machine learning algorithms, statistical techniques, and other methods to identify patterns and trends in the data. For example, if you are analyzing customer behavior, you may use clustering to identify segments of customers with similar characteristics.
7. **Knowledge Representation:**
   - The final step in the data mining process is to represent the patterns and insights in a form that is understandable and usable by decision-makers. This may involve creating visualizations, reports, or other formats that present the data in a clear and meaningful way.

Data Mining Functionalities

- Multidimensional data concept
- Association analysis on frequent patterns
- Classification and prediction
- Cluster analysis
- Outlier analysis
  - Outlier is an unusual behavior, i.e. Data object that does not comply with the general behavior of the data
  - Useful in fraud detection, rare events analysis
- Trend and evolution analysis

Are All the "Discovered" Patterns Interesting?

- Data mining may generate thousands of patterns: Not all of them are interesting
- Suggested approach: Human-centered, query-based, focused mining
- A pattern is interesting if it is easily understood by humans, valid on new or test data with some degree of certainty, potentially useful, novel, or validates some hypothesis that a user seeks to confirm

Major Issues In Data Mining:

- **Mining different kinds of knowledge in databases.** - The need of different users is not the same. And Different user may be in interested in different kind of knowledge. Therefore it is necessary for data mining to cover broad range of knowledge discovery task.
- **Interactive mining of knowledge at multiple levels of abstraction. -** The data mining process needs to be interactive because it allows users to focus the search for patterns, providing and refining data mining requests based on returned results.
- **Incorporation of background knowledge. -** To guide discovery process and to express the discovered patterns, the background knowledge can be used. Background knowledge may be used to express the discovered patterns not only in concise terms but at multiple level of abstraction.

Major Issues In Data Mining:

- **Data mining query languages and ad hoc data mining. -** Data Mining Query language that allows the user to describe ad hoc mining tasks, should be integrated with a data warehouse query language and optimized for efficient and flexible data mining.
- **Presentation and visualization of data mining results. -** Once the patterns are discovered it needs to be expressed in high level languages, visual representations. This representations should be easily understandable by the users.
- **Handling noisy or incomplete data. -** The data cleaning methods are required that can handle the noise, incomplete objects while mining the data regularities. If data cleaning methods are not there then the accuracy of the discovered patterns will be poor.

Major Issues In Data Mining:

- **Pattern evaluation. -** It refers to interestingness of the problem. The patterns discovered should be interesting because either they represent common knowledge or lack novelty.
- **Efficiency and scalability of data mining algorithms. -** In order to effectively extract the information from huge amount of data in databases, data mining algorithm must be efficient and scalable.

Issues in Data Mining

- **Mining methodology**
  - Mining different kinds of knowledge from diverse data types, e.g., bio, stream, Web
  - Performance: efficiency, effectiveness, and scalability
  - Pattern evaluation: the interestingness problem
  - Handling noise and incomplete data
  - Parallel, distributed and incremental mining methods
  - Integration of the discovered knowledge with existing one: knowledge fusion
- **User interaction**
  - Data mining query languages and ad-hoc mining
  - Expression and visualization of data mining results
  - Interactive mining of knowledge at multiple levels of abstraction
- **Applications and social impacts**
  - Domain-specific data mining & invisible data mining
  - Protection of data security, integrity, and privacy

Types of Database

- Relational database
- Data warehouse
- Transactional database

Data Objects and Attribute Types

- Nominal Attribute
- Binary Attribute
- Ordinal Attribute
- Numeric Attribute
  - Interval Scaled Attribute
  - Ratio Scaled Attribute
- Discrete VS Continuous Data

Nominal Attributes

- Symbols or Name of things
- Each value represent some value of category
- Enumerations
- Eg hair color, marital status

Binary Attributes

- With only two categories (state 0 or 1)
- True / False, Present / Absence
- Eg Patient Smokes, Testing Result (+ve / -ve)
- Binary attribute is symmetric if both of its states are equally valuable (i.e. no preference on which outcome should be coded as 0 or 1 like gender)
- Binary Attribute is asymmetric if the outcomes of the states are not equally important eg HIV test, Corona test)

Ordinal Attributes

- Attribute with possible values that have a meaningful order or ranking among them
- Eg Drinking (Small, Medium, Large)
- Lecturer, Asst. Professor, Professor
- Strongly Agree, Agree, Disagree

Numeric Attributes

- Quantitative
- Interval Scaled Attribute
  - Measured on a scale of equal size units
  - Eg Calendar Date (2002 and 2010 are 8 years apart)
- Ratio Scaled Attribute
  - If a measurement is ratio scaled means a value being multiple (or ratio) of another value
  - Eg Frequency of words in a document

Discrete VS Continuous Attributes

- A discrete attribute has a finite or infinite set of values which may or may not be represented as integer (Eg Age)
- If attribute is not discrete then it is continuous (Eg temperature)

Data Preprocessing

- Why preprocess the data?
- Data cleaning
- Data integration and transformation

Why Data Preprocessing?

- Data in the real world is dirty
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., occupation=“ ”
  - noisy: containing errors or outliers
    - e.g., Salary=“-10”
  - inconsistent: containing discrepancies in codes or names
    - e.g., Age=“42” Birthday=“03/07/1997”
    - e.g., Was rating “1,2,3”, now rating “A, B, C”

Why Is Data Dirty?

- Incomplete data may come from
  - "Not applicable" data value when collected
  - Different considerations between the time when the data was collected and when it is analyzed.
  - Human/hardware/software problems
- Noisy data (incorrect values) may come from
  - Faulty data collection instruments
  - Human or computer error at data entry
  - Errors in data transmission
- Inconsistent data may come from
  - Different data sources
  - Functional dependency violation (e.g., modify some linked data)
- Duplicate records also need data cleaning

Why Is Data Preprocessing Important?

- No quality data, no quality mining results!
  - Quality decisions must be based on quality data
    - e.g., duplicate or missing data may cause incorrect or even misleading statistics.
  - Data warehouse needs consistent integration of quality data
- Data extraction, cleaning, and transformation comprises the majority of the work of building a data warehouse

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 transformation
  - Normalization and aggregation
- Data reduction
  - Obtains reduced representation in volume but produces the same or similar analytical results
- Data discretization
  - Part of data reduction but with particular importance, especially for numerical data

Data Cleaning

- Importance
  - "Data cleaning is one of the three biggest problems in data warehousing"—Ralph Kimball
  - "Data cleaning is the number one problem in data warehousing"—DCI survey
- **Data cleaning tasks**
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data

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 (assuming the tasks in classification—not effective when the percentage 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

Noisy Data

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

Data Transformation

- Smoothing: remove noise from data
- Aggregation: summarization, data cube construction
- Generalization: concept hierarchy climbing
- Normalization: scaled to fall within a small, specified range
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling
- Attribute/feature construction
  - New attributes constructed from the given ones

Data Transformation: Normalization

- **Min-max normalization: to [new_min$_A$, new_max$_A$]**

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,000 is mapped to $\frac{73,600 - 12,000}{98,000 - 12,000}(1.0 - 0) + 0 = 0.716$

- **Z-score normalization (μ: mean, σ: standard deviation):**

$$v' = \frac{v - \mu_A}{\sigma_A}$$

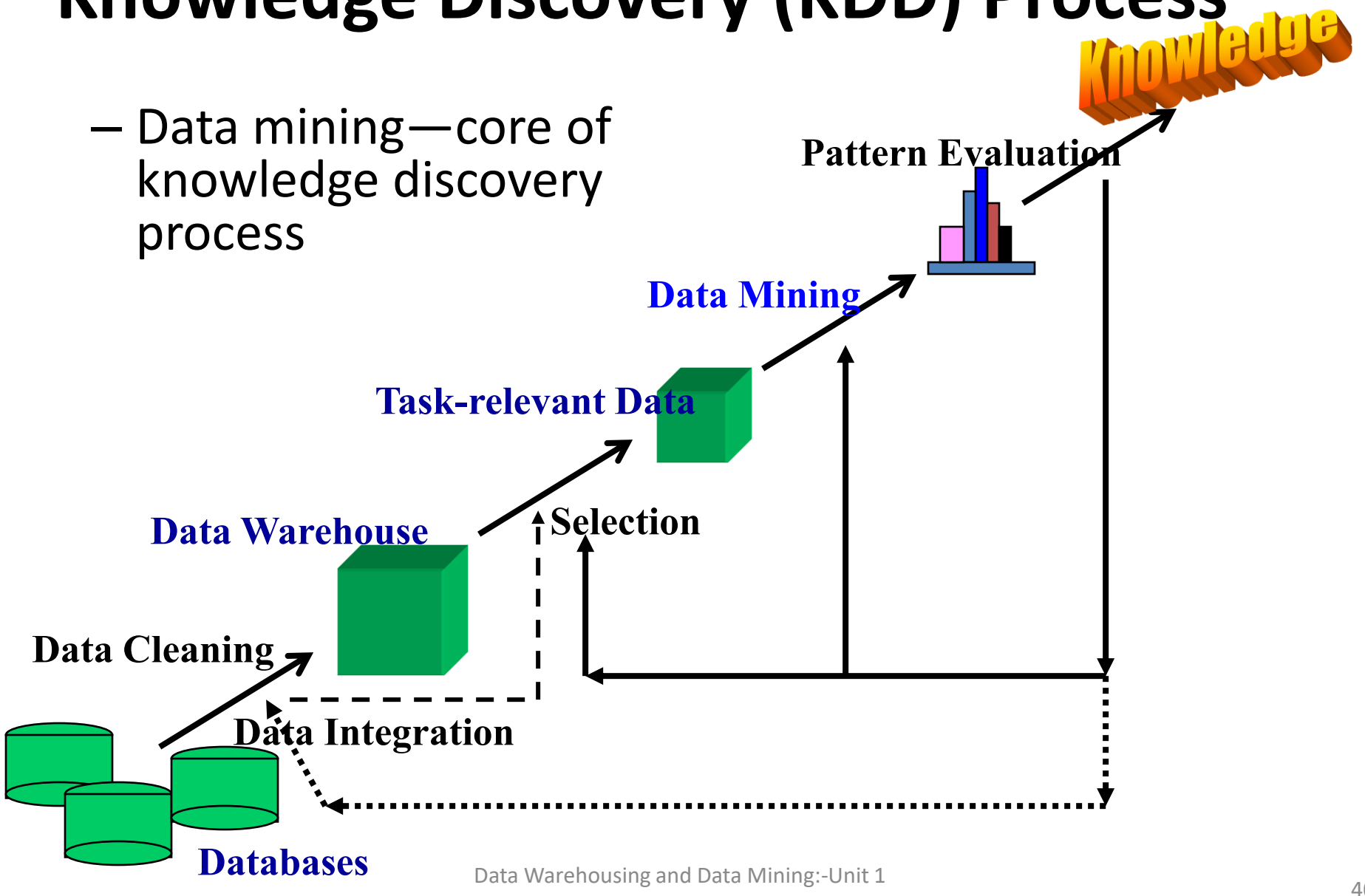
  - Ex. Let μ = 54,000, σ = 16,000. Then $\frac{73,600 - 54,000}{16,000} = 1.225$

- **Normalization by decimal scaling**

$$v' = \frac{v}{10^j}$$ Where j is the smallest integer such that Max(|v'|) < 1

Knowledge Discovery (KDD) Process

– Data mining—core of knowledge discovery process



Key Features of Data Mining

- **Pattern Recognition:** Data Mining identifies patterns and trends within datasets.
- **Predictive Analysis**: It enables businesses to make predictions about future trends.
- **Classification**: Categorizes data into predefined classes or groups.
- **Clustering:** Group similar data points together based on their characteristics.
- **Association Rule Mining:** Discovers relationships and connections between variables.

End of Session