CS699
Lecture 1
Introduction

CS699

- Our focus is “data mining” not “data warehousing”
- Data mining is an important component of data analysis.
- Will discuss
  - Data preprocessing
  - Basic data mining algorithms
  - How to evaluate data mining models and data mining results
  - How to perform data mining using software tools
- A good data mining web site: kdnuggets.com

CS699

- Prerequisites:
  - CIS students: CS546 and CS669
  - CS students: CS579
- Math requirements
  - Math is a tool to describe algorithms
  - Mostly basic algebra (not linear algebra) and basic probabilities and statistics
  - A little bit of calculus
  - You will have to do calculations using a calculator (which has a “log” function)

CS699

- You will practice data mining with Weka and Oracle.
- These software are used for assignments.
- Weka:
  - Free
  - Easy to learn and easy to use
  - Has a large number of data mining algorithms
  - You will use it immediately
  - Also used for class project

CS699

- Oracle data mining: takes time to learn
- You will learn how to use them with assignments
- Oracle:
  – Will use preconfigured virtual machine
  – VM runs on Linux
  – But, you will rarely use Linux (actually don’t need to use it)
  – You will use SQL Developer for data mining

CS699

- JMP Pro
  - JMP Pro is a statistical analysis software.
  - It also has some data mining algorithms.
  - Available through the license BU has.
  - Will be used for some assignments.

CS699

- Class project:
  - Classification
  - You can use any tools for data preprocessing.
  - You will use Weka for building and testing classifier models.

Why Data Mining?

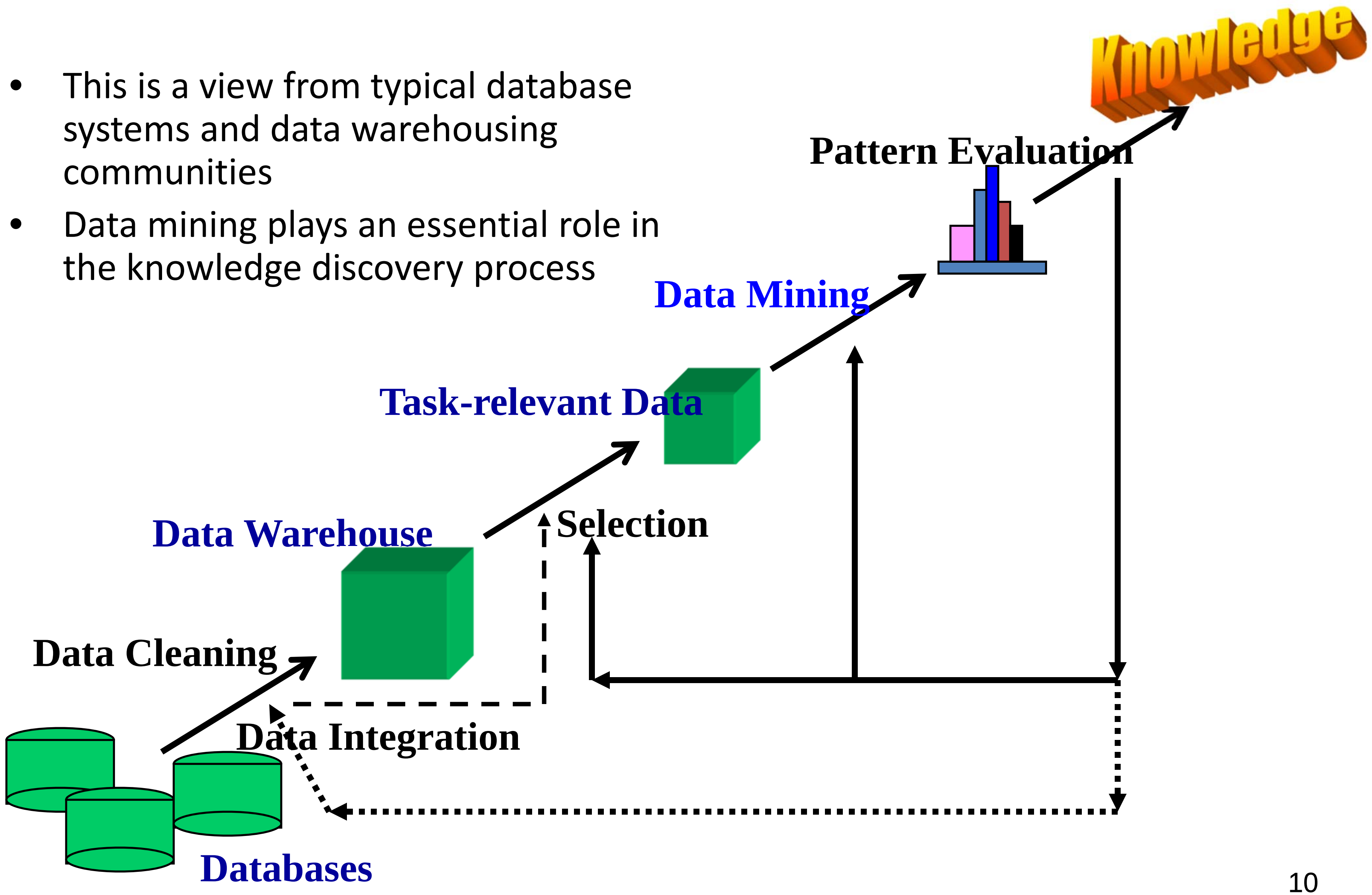
- The Explosive Growth of Data: from terabytes to petabytes
  - Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized society
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, digital cameras, YouTube, social network
- <u>We are drowning in data, but starving for knowledge!</u>
- “Necessity is the mother of invention”—Data mining—Automated analysis of massive data sets

What Is Data Mining?

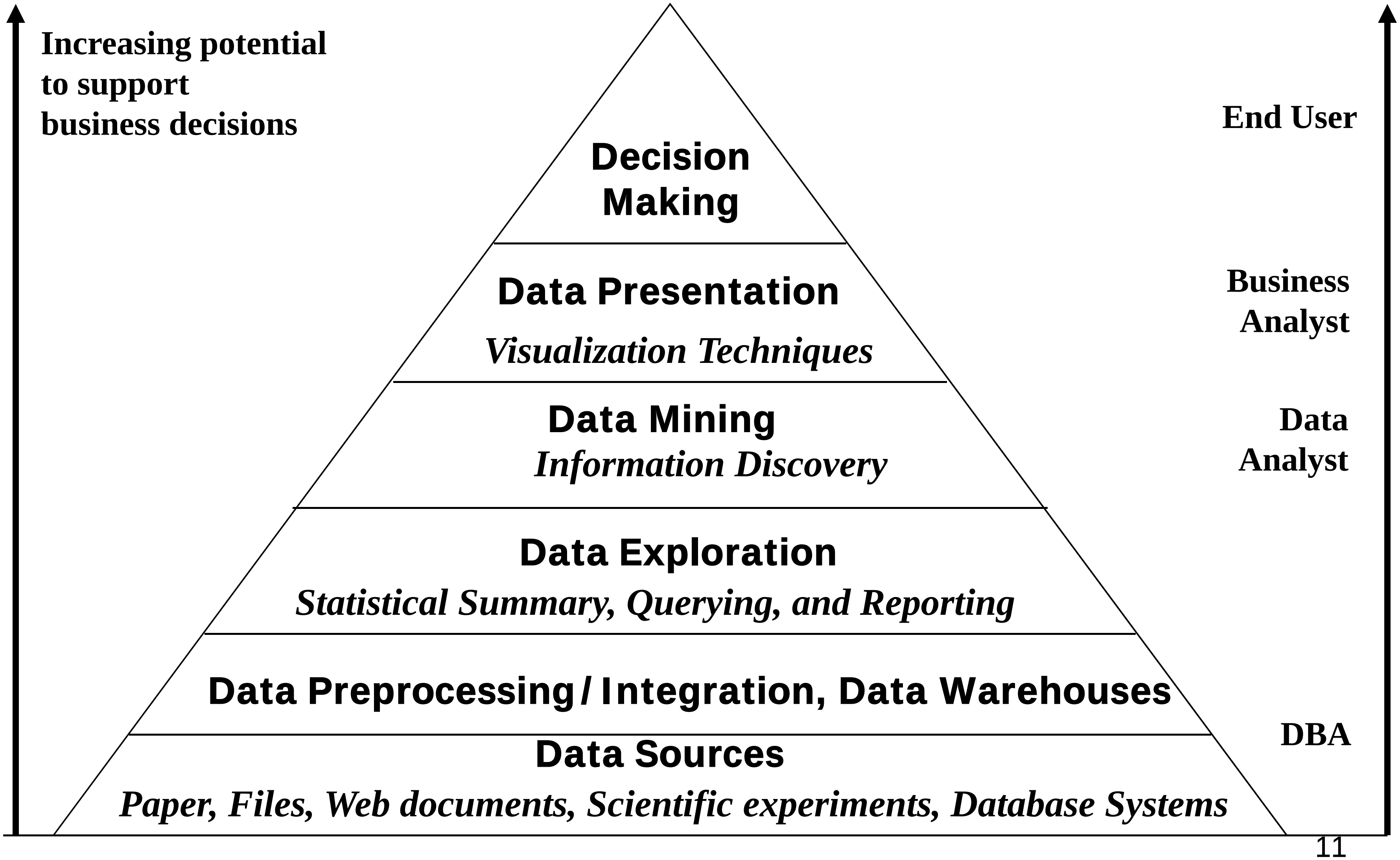
- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- Watch out: Is everything "data mining"?
  - Simple search and query processing
  - (Deductive) expert systems

Knowledge Discovery (KDD) Process

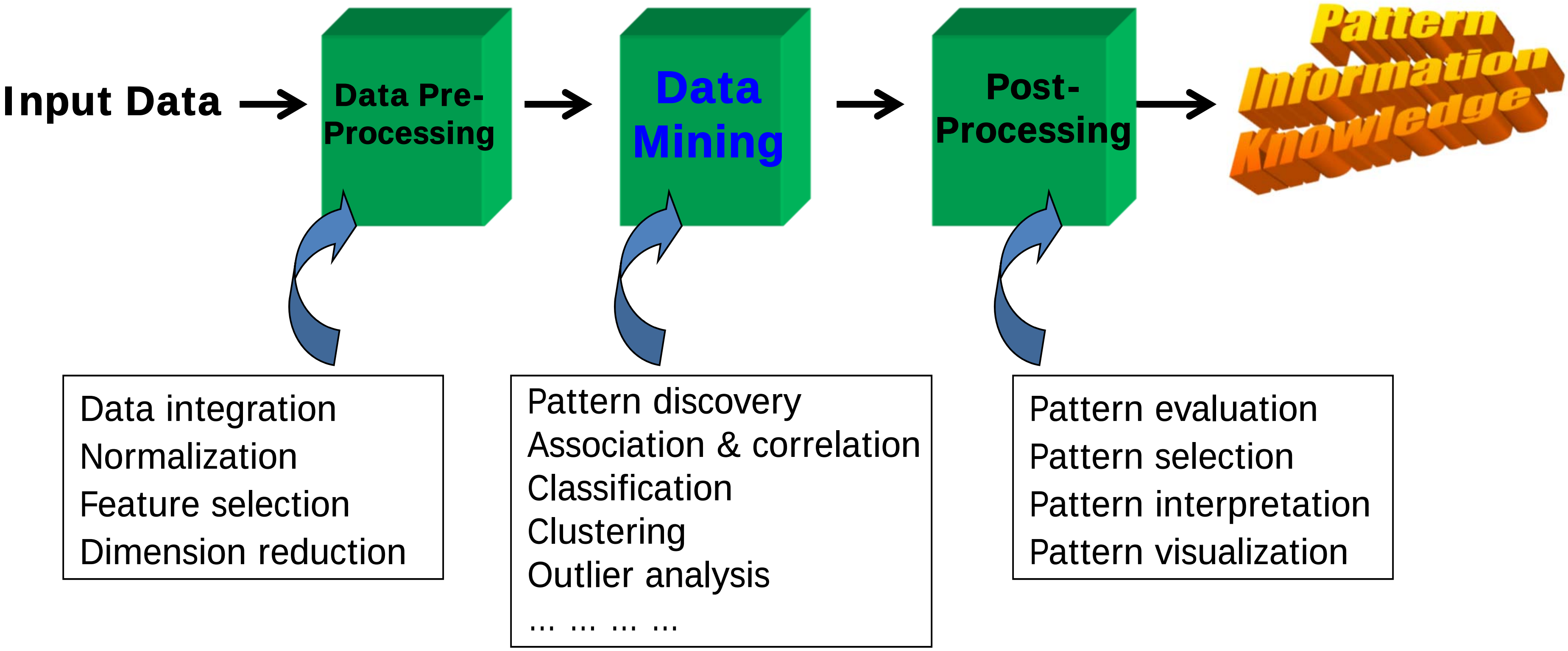
- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process



Data Mining in Business Intelligence



A Typical View from ML and Statistics



Input Data → Data Pre-Processing → Data Mining → Post-Processing → Pattern Information Knowledge

Data integration
Normalization
Feature selection
Dimension reduction

Pattern discovery
Association & correlation
Classification
Clustering
Outlier analysis
… … … …

Pattern evaluation
Pattern selection
Pattern interpretation
Pattern visualization

- This is a view from typical machine learning and statistics communities

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, temporal data, sequence data (incl. bio-sequences)
  - Structure data, graphs, social networks and multi-linked data
  - Object-relational databases
  - Heterogeneous databases and legacy databases
  - Spatial data and spatiotemporal data
  - Multimedia database
  - Text databases
  - The World-Wide Web

Data Types

- Categorical (or nominal) vs. numeric data:

Categorical

OID	Age	Income	Buy?
1	Young	Low	Y
2	Young	High	Y
3	Old	Low	N
4	Middle	Low	Y
5	Middle	High	N
6	Old	Low	N
7	Young	High	N
8	Old	High	Y
9	Old	High	Y
10	Young	Low	N

Numeric

OID	Age	Height	Weight
1	15	60	180
2	8	48	115
3	32	72	153
4	27	65	145
5	17	58	189
6	56	70	150
7	72	56	163
8	22	63	172
9	42	71	139
10	39	68	150

Association and Correlation Analysis

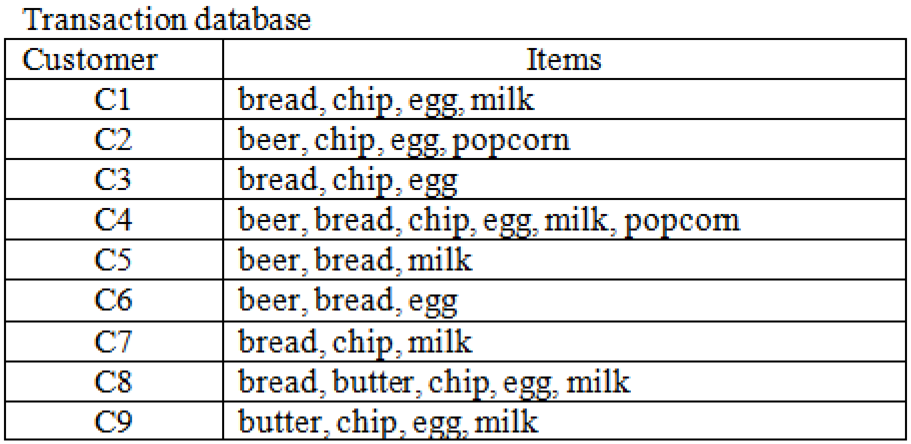
- Frequent patterns (or frequent itemsets)
  - What items are frequently purchased together in a grocery store?
  - Mine all *frequent* itemsets and then all *strong* rules.
  - An itemset is *frequent*,

    if its support is >= predefined threshold, *minimum support*
  - A rule is written as: <left hand side> => <right hand side>
  - Example of a rule: {milk, butter} => {cheese, egg}
  - A rule is *strong*,

    if its confidence is >= predefined threshold, *minimum confidence*

Association and Correlation Analysis

- Frequent patterns (or frequent itemsets)
  – What items are frequently purchased together in a grocery store?



Transaction database

Customer	Items
C1	bread, chip, egg, milk
C2	beer, chip, egg, popcorn
C3	bread, chip, egg
C4	beer, bread, chip, egg, milk, popcorn
C5	beer, bread, milk
C6	beer, bread, egg
C7	bread, chip, milk
C8	bread, butter, chip, egg, milk
C9	butter, chip, egg, milk

A frequent itemset : chip, milk, and egg are frequently purchased together

A rule: Whenever a customer buys chip, he/she is likely to buy milk and egg.

This rule is written as:

$\{chip\} \Rightarrow \{milk, egg\}$

Support = 44.4% (or 4/9),

Confidence = 57.1% (or 4/7)

Association and Correlation Analysis

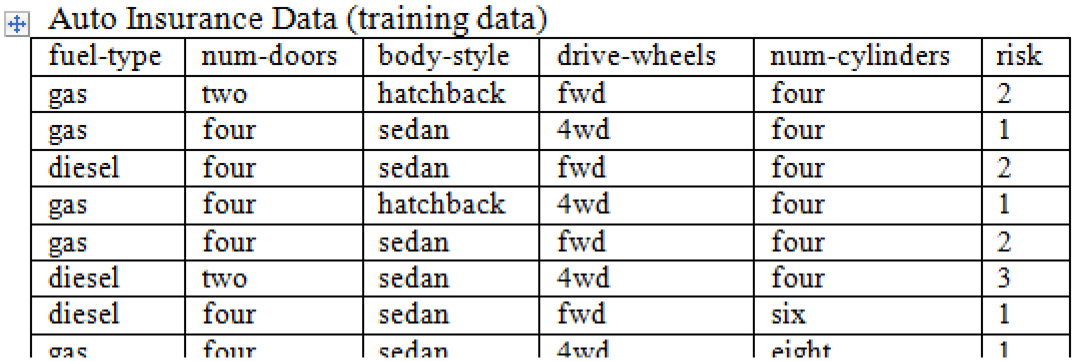
- Association, correlation vs. causality
  - Are strongly associated items also strongly correlated?
  - If two items are strongly correlated, is there a causal relationship?
- How to mine such patterns and rules efficiently in large datasets?
- Association rules can also be used for classification or clustering.

Classification

- Classification and label prediction
  - Construct models (functions) based on some training examples
  - Describe and distinguish classes or concepts for future prediction
    - E.g., classify countries based on (climate), or classify cars based on (gas mileage)
  - Predict unknown class label (or class attribute)
- Typical methods
  - Decision trees, naïve Bayesian classification, support vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression, …
- Typical applications:
  - Credit card fraud detection, classifying stars, diseases, web-pages, …
- Also called supervised learning

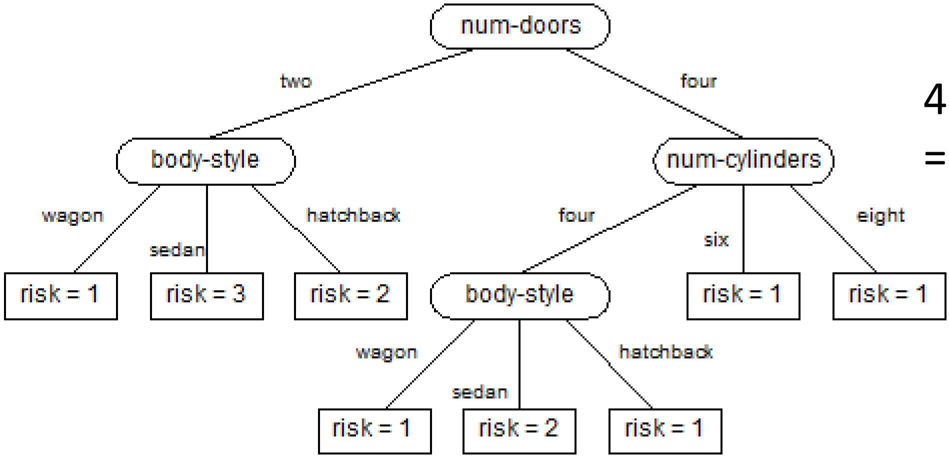
Classification

- Example (decision tree)



Auto Insurance Data (training data)

fuel-type	num-doors	body-style	drive-wheels	num-cylinders	risk
gas	two	hatchback	fwd	four	2
gas	four	sedan	4wd	four	1
diesel	four	sedan	fwd	four	2
gas	four	hatchback	4wd	four	1
gas	four	sedan	fwd	four	2
diesel	two	sedan	4wd	four	3
diesel	four	sedan	fwd	six	1
gas	four	sedan	4wd	eight	1



Classify a car with unknown class label (risk):

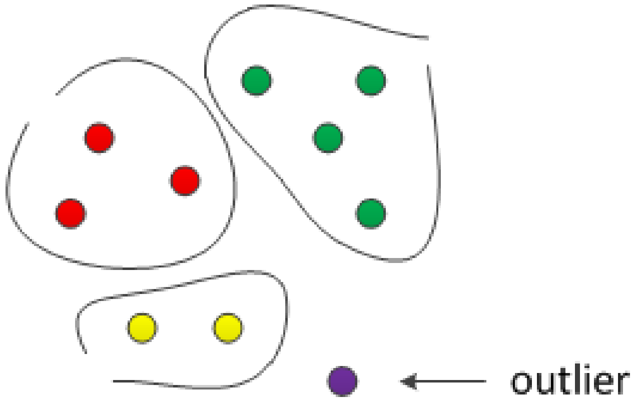
4-door, 4-cylinder, wagon.
==> risk = 1

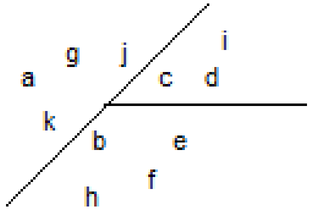
Cluster Analysis

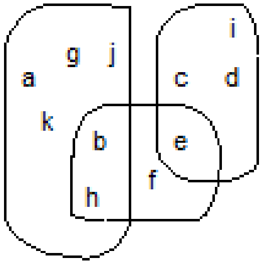
- Unsupervised learning (i.e., there is no class label)
- Group data to form new categories (i.e., clusters), e.g., cluster houses to find distribution patterns
- Principle: Maximizing intra-class similarity & minimizing interclass similarity
- Many methods and applications
- Example application: Divide a set of customers into clusters in such a way that customers belonging to the same cluster share common properties.

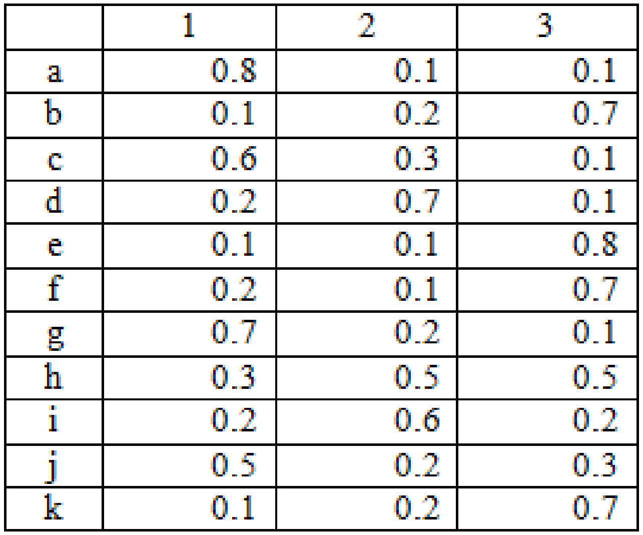
Cluster Analysis

- Examples:





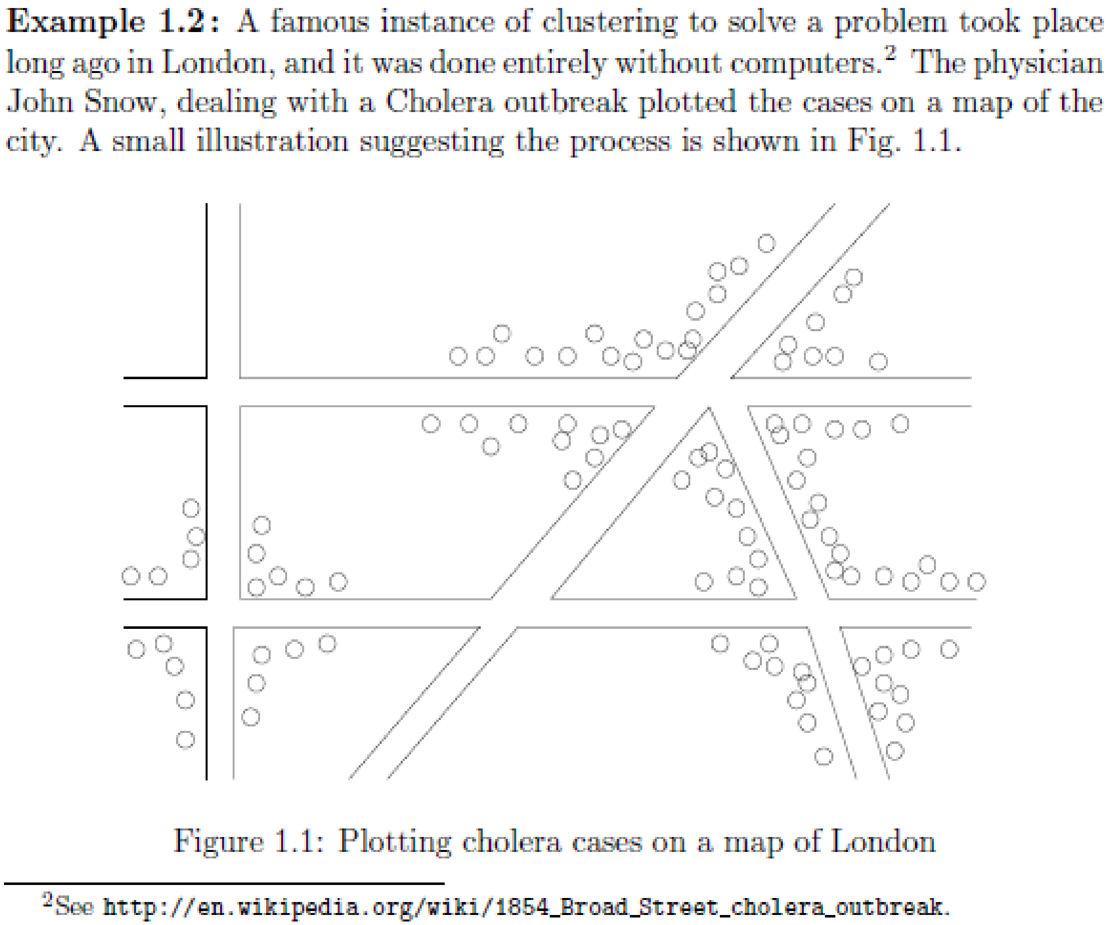




	1	2	3
a	0.8	0.1	0.1
b	0.1	0.2	0.7
c	0.6	0.3	0.1
d	0.2	0.7	0.1
e	0.1	0.1	0.8
f	0.2	0.1	0.7
g	0.7	0.2	0.1
h	0.3	0.5	0.5
i	0.2	0.6	0.2
j	0.5	0.2	0.3
k	0.1	0.2	0.7

Cluster Analysis

- London cholera epidemic (Source: J. Leskovec, A. Rajaraman, and J.D. Ullman, “Mining of Massive Datasets,” 2014, page 3.)



Example 1.2 : A famous instance of clustering to solve a problem took place long ago in London, and it was done entirely without computers.[2] The physician John Snow, dealing with a Cholera outbreak plotted the cases on a map of the city. A small illustration suggesting the process is shown in Fig. 1.1.

Figure 1.1: Plotting cholera cases on a map of London

[2] See `http://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak`.

Outlier Analysis

- Outlier: A data object that does not comply with the general behavior of the data
- Noise or exception? — One person’s garbage could be another person’s treasure
- Methods: byproduct of clustering or regression analysis, …
- Useful in fraud detection, rare events analysis

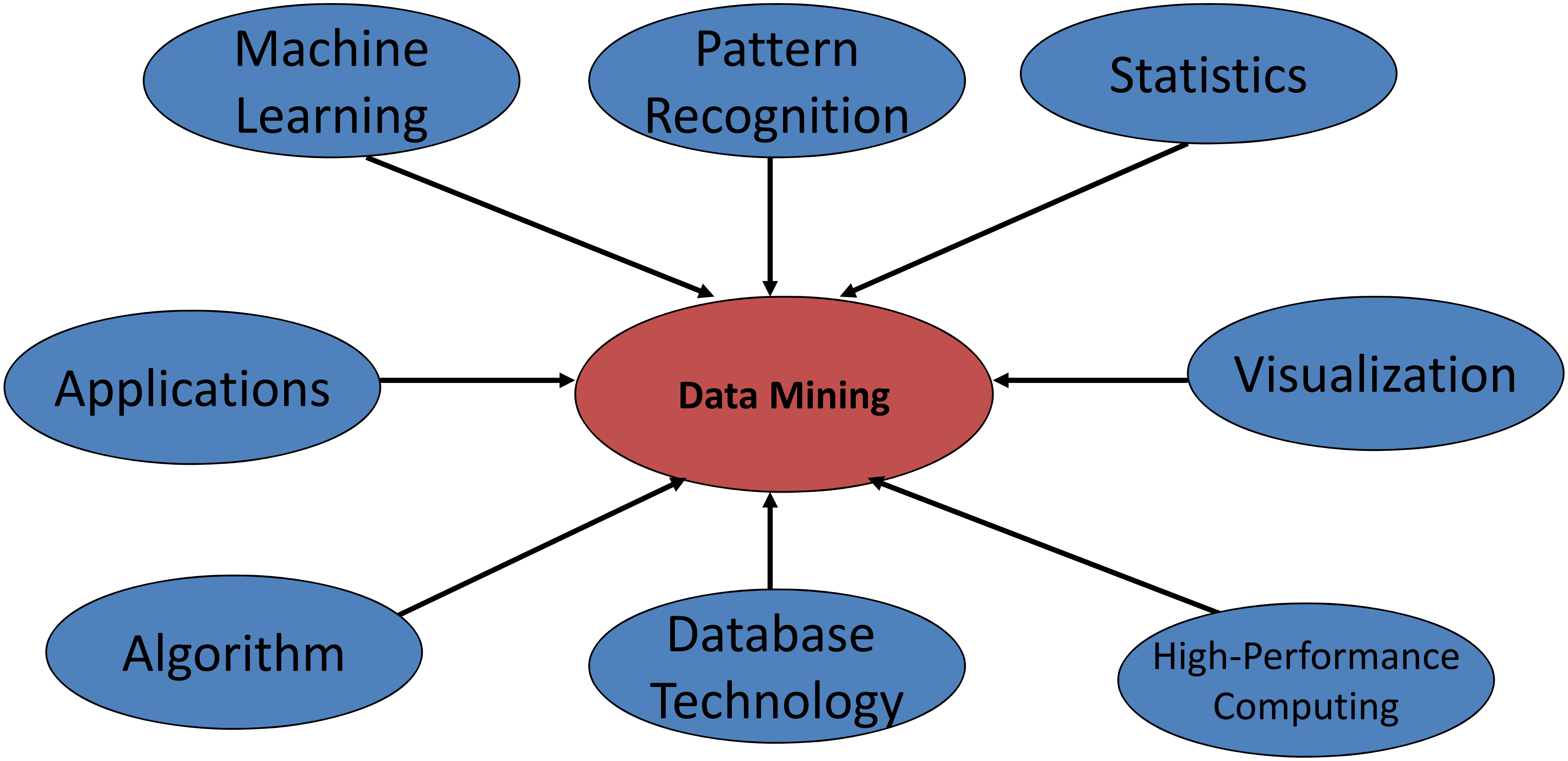
Sequential Pattern, Trend and Evolution Analysis

- Trend, time-series, and deviation analysis: e.g., regression and value prediction
- Sequential pattern mining
  - e.g., first buy digital camera, then buy large SD memory cards
- Periodicity analysis
- Biological sequence analysis

Evaluation of Knowledge

- Are all mined knowledge interesting?
  - One can mine tremendous amount of “patterns” and knowledge
  - Some may fit only certain dimension space (time, location, …)
  - Some may not be representative, may be transient, …
- A pattern is interesting if
  - easily understood
  - valid on new data or test data with some degree of certainty
  - potentially useful
  - novel
- Objective measures (e.g., support and confidence of an association rule)
- Subjective measures (e.g., expected/unexpected, actionable)

Technologies Used in Data Mining



Applications of Data Mining

- Web page analysis: from web page classification, clustering to PageRank & HITS algorithms
- Collaborative analysis & recommender systems
- Basket data analysis to targeted marketing
- Biological and medical data analysis: classification, cluster analysis (microarray data analysis), biological sequence analysis, biological network analysis
- Data mining and software engineering
- From major dedicated data mining systems/tools (e.g., SAS, MS SQL-Server Analysis Manager, Oracle Data Mining Tools) to invisible data mining

Major Issues in Data Mining

- Mining Methodology
- User Interaction
- Efficiency and Scalability
- Diversity of data types
- Data mining and society

What is a Data Warehouse?

- Defined in many different ways, but not rigorously.
  - A decision support database that is maintained separately from the organization’s operational database
  - Support information processing by providing a solid platform of consolidated, historical data for analysis.
- “A data warehouse is a <u>subject-oriented</u>, <u>integrated</u>, <u>time-variant</u>, and <u>nonvolatile</u> collection of data in support of management’s decision-making process.”—W. H. Inmon
- Data warehousing:
  - The process of constructing and using data warehouses

Data Warehouse—Subject-Oriented

- Organized around major subjects, such as customer, product, sales
- Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing
- Provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process

Data Warehouse—Integrated

- Constructed by integrating multiple, heterogeneous data sources
  - relational databases, flat files, on-line transaction records
- Data cleaning and data integration techniques are applied.
  - Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources
    - E.g., Hotel price: currency, tax, breakfast covered, etc.
  - When data is moved to the warehouse, it is converted.

Data Warehouse—Time Variant

- The time horizon for the data warehouse is significantly longer than that of operational systems
  - Operational database: current value data
  - Data warehouse data: provide information from a historical perspective (e.g., past 5-10 years)
- Every key structure in the data warehouse
  - Contains an element of time, explicitly or implicitly
  - But the key of operational data may or may not contain “time element”

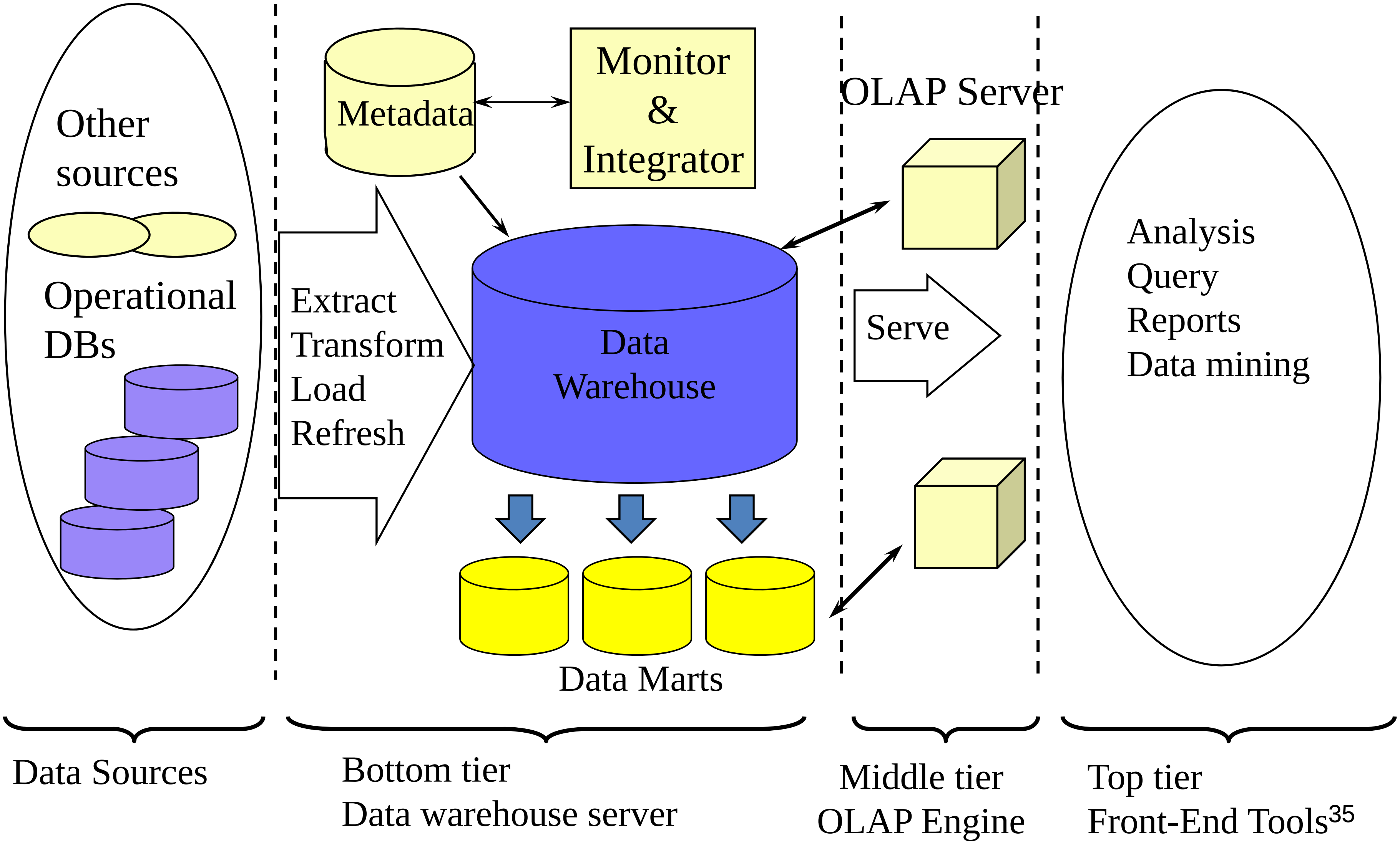
Data Warehouse—Nonvolatile

- A physically separate store of data transformed from the operational environment
- Operational update of data does not occur in the data warehouse environment
  - Does not require transaction processing, recovery, and concurrency control mechanisms
  - Requires only two operations in data accessing:
    - *initial loading of data* and *access of data*

OLTP vs. OLAP

	OLTP	**OLAP**
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response

Data Warehouse: A Three-Tier Architecture



Three Data Warehouse Models

- Enterprise warehouse
  - collects all of the information about subjects spanning the entire organization
- Data Mart
  - a subset of corporate-wide data that is of value to a specific groups of users. Its scope is confined to specific, selected groups, such as marketing data mart
    - Independent vs. dependent (directly from warehouse) data mart
- Virtual warehouse
  - A set of views over operational databases
  - Only some of the possible summary views may be materialized

Extraction, Transformation, and Loading (ETL)

- **Data extraction**
  - get data from multiple, heterogeneous, and external sources
- **Data cleaning**
  - detect errors in the data and rectify them when possible
- **Data transformation**
  - convert data from legacy or host format to warehouse format
- **Load**
  - sort, summarize, consolidate, compute views, check integrity, and build indicies and partitions
- **Refresh**
  - propagate the updates from the data sources to the warehouse

Metadata Repository

- **Meta data** is the data defining warehouse objects. It stores:
- Description of the structure of the data warehouse
  - schema, view, dimensions, hierarchies, derived data defn, data mart locations and contents
- Operational meta-data
  - data lineage (history of migrated data and transformation path), currency of data (active, archived, or purged), monitoring information (warehouse usage statistics, error reports, audit trails)
- The algorithms used for summarization
- The mapping from operational environment to the data warehouse
- Data related to system performance
  - warehouse schema, view and derived data definitions
- Business data
  - business terms and definitions, ownership of data, charging policies

References

- Han, J., Kamber, M., Pei, J., “Data mining: concepts and techniques,” 3rd Ed., Morgan Kaufmann, 2012
- http://www.cs.illinois.edu/~hanj/bk3/