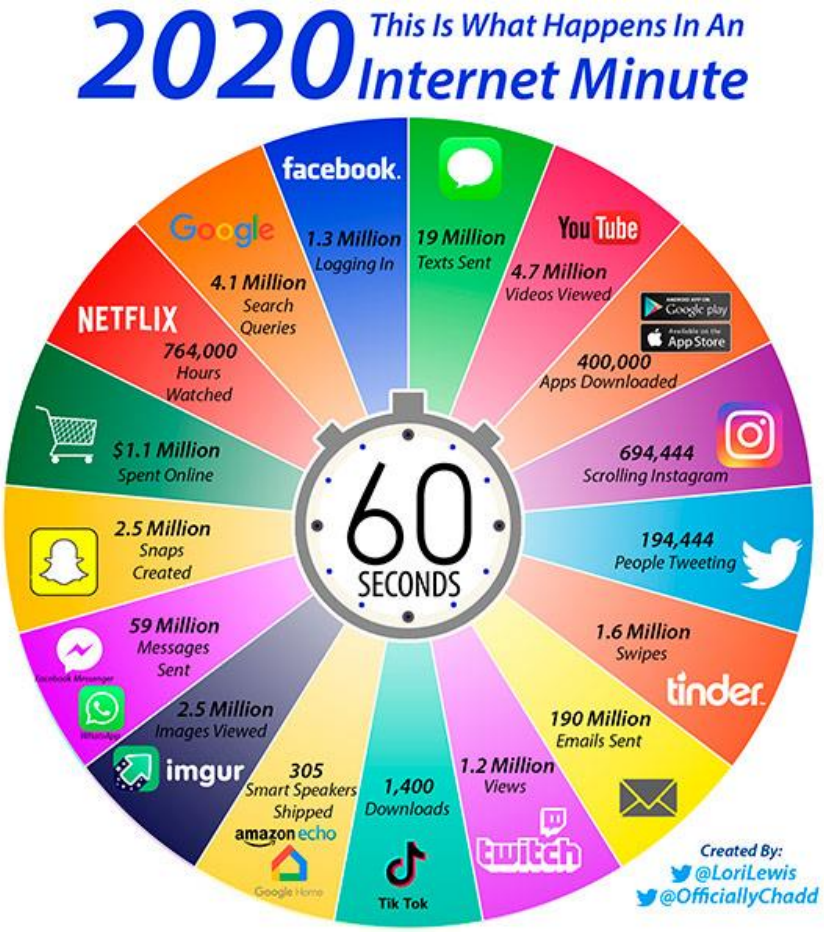

2020 This Is What Happens In An Internet Minute
facebook.
1.3 Million Logging In
19 Million Texts Sent
YouTube
4.7 Million Videos Viewed
400,000 Apps Downloaded
694,444 Scrolling Instagram
194,444 People Tweeting
1.6 Million Swipes
tinder
190 Million Emails Sent
1.2 Million Views
twitch
1,400 Downloads
Tik Tok
305 Smart Speakers Shipped
amazon echo
Google Home
2.5 Million Images Viewed
imgur
59 Million Messages Sent
2.5 Million Snaps Created
$1.1 Million Spent Online
NETFLIX
764,000 Hours Watched
Google
4.1 Million Search Queries
60 SECONDS
Created By:
@LoriLewis
@OfficiallyChadd


2021 This Is What Happens In An Internet Minute
facebook
1.4 Million Scrolling
21.1 Million Texts Sent
YouTube
500 Hours Content Uploaded
414,764 Apps Downloaded
695,000 Stories Shared
200,000 People Tweeting
2 Million Swipes
tinder
197.6 Million Emails Sent
2 Million Views
twitch
5,000 Downloads
Tik Tok
932 Smart Audio Devices Shipped
amazon echo
3 Million Images Viewed
imgur
69 Million Messages Sent
3.4 Million Snaps Created
$1.6 Million Spent Online
NETFLIX
28,000 Subscribers Watching
LinkedIn
9,132 Connections Made
60 SECONDS
Created By:
@LoriLewis
@OfficiallyChadd

Need for Data Mining

- Data perspective
  - ✓Explosive growth
  - ✓Abundant sources
  - ✓Lack of knowledge
  - ✓Availability of computing power
- Application perspective
  - ✓Credit ratings
  - ✓Targeted marketing, CRM
  - ✓Fraud detection
  - ✓Healthcare
  - ✓Web mining
  - ✓IoT, Smart Cities

Alternative names

- Knowledge Discovery in Databases (KDD)
- Knowledge Extraction
- Data/Pattern Analysis
- Business Intelligence

Syllabus

Module	Content
Unit I	**Introduction to Data Mining:** Basic concepts of data mining, Types of Data to be mined, Stages of the Data Mining Process, Data Mining Techniques, Knowledge Discovery in Databases, Data Mining Issues, Applications of Data
Unit II	**Introduction to Data Warehouse:** Data Warehouse and DBMS Architecture of Data Warehouse, Multidimensional data model, Concepts of OLAP and Data Cube, OLAP operations, Dimensional Data Modelling- Star, Snow flake schemas
Unit III	

Syllabus

Module	Content
Unit III	**Data pre-processing:** Need Data pre-processing, Attributes and Data types, Statistical descriptions of Data, Handling missing Data, Data sampling, Data cleaning, Data Integration and transformation, Data reduction-– Curse of Dimensionality, Feature Selection and Feature Engineering, Principle Component Analysis (PCA), Discretization and generating concept hierarchies
Unit IV	**Data Mining Techniques: Association Rule Mining:** Basic idea: item sets, Frequent Item-sets, Association Rule Mining, Generating item sets and rules efficiently, FP growth algorithm

Syllabus

Module	Content
Unit V	**Data Mining Techniques: Classification:** Definition of Classification, Decision tree Induction: Information gain, gain ratio, Gini Index, Issues: Over-fitting, tree pruning methods, missing values, continuous classes, Bayesian Classification: Bayes Theorem, Naïve Bayes classifier, Bayesian Networks, least squares, SVM classifiers, Lazy Learners (or Learning from Your Neighbours)

Syllabus

Module	Content
Unit VI	**Data Mining Techniques: Prediction**: Definition of Prediction, Parametric and Non-Parametric algorithms, Linear Regression Algorithm, Linear Regression Model, OLS, Derivation of Beta coefficients for OLS, OLS Cost function, RMSE, R-Squared Error, Linear Regression Assumptions. Non-linear regression, logistic regression. **Data Mining Techniques: Introduction to Clustering:** Definition of Clustering, Partitioning Methods, Hierarchical Methods, Distance Measures in Algorithmic Methods, Density Based Clustering

Assessment

- **Class Continuous Assessment (CCA) 60 marks**
- Mid Term Examination – 30 MARKS
- Formative Assessment Test 1 – 15 MARKS
- Formative Assessment Test 2 – 15 MARKS

- **Term End Examination : 40 Marks**

Books

1. **Data Mining: Concepts and Techniques, Han and Kamber, 3rd edition**
2. Margaret H. Dunham, S. Sridhar, Data Mining – Introductory and Advanced Topics, Pearson Education
3. Data warehousing: fundamentals for IT professionals 3rd edition , Kimball, Wiley Publication
4. Ian H.Witten, Eibe Frank Data Mining: Practical Machine Learning Tools and Techniques, Elsevier/(Morgan Kauffman), ISBN:9789380501864
5. Introduction to Data Mining (2005) By Pang-Ning Tan, Michael Steinbach, Vipin Kumar Addison Wesley ISBN: 0-321-32136-7

Unit 1
Introduction to Data Mining

Introduction

Why Data Mining?

What Is Data Mining?

A Multi-Dimensional View of Data Mining

What Kind of Data Can Be Mined?

What Kinds of Patterns Can Be Mined?

What Technology Are Used?

What Kind of Applications Are Targeted?

Major Issues in Data Mining

What is Data Mining?

- Discovering interesting data patterns hidden in large datasets
- Data Mining is sorting through data to identify patterns and establish relationships
- A process used by companies to turn raw data into useful information
- Data mining involves use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large datasets

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
- Data mining is the exploration and analysis of large quantities of data to discover meaningful patterns and rules which are hidden in the data. They are different from the retrieval of data.
- Data -> Information -> Knowledge

Anu	98
Parth	99
Soham	100

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

We are drowning in data, but starving for knowledge!

"Necessity is the mother of invention"—Data mining—Automated analysis of massive data sets

Evolution of Sciences

Before 1600, **empirical science**

1600-1950s, **theoretical science**

- Each discipline has grown a *theoretical* component. Theoretical models often motivate experiments and generalize our understanding.

1950s-1990s, **computational science**

- Over the last 50 years, most disciplines have grown a third, *computational* branch (e.g. empirical, theoretical, and computational ecology, or physics, or linguistics.)
- Computational Science traditionally meant simulation. It grew out of our inability to find closed-form solutions for complex mathematical models.

1990-now, **data science**

- The flood of data from new scientific instruments and simulations
- The ability to economically store and manage petabytes of data online
- The Internet and computing Grid that makes all these archives universally accessible
- Scientific info. management, acquisition, organization, query, and visualization tasks scale almost linearly with data volumes. Data mining is a major new challenge!

Jim Gray and Alex Szalay, *The World Wide Telescope: An Archetype for Online Science*, Comm. ACM, 45(11): 50-54, Nov. 2002

Evolution of Database Technology

1960s:

- Data collection, database creation, IMS and network DBMS

1970s:

- Relational data model, relational DBMS implementation

1980s:

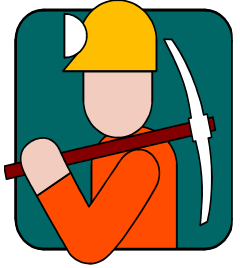
- RDBMS, advanced data models (extended-relational, OO, deductive, etc.)
- Application-oriented DBMS (spatial, scientific, engineering, etc.)

1990s:

- Data mining, data warehousing, multimedia databases, and Web databases

2000s

- Stream data management and mining
- Data mining and its applications
- Web technology (XML, data integration) and global information systems



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
- Data mining: a misnomer?

Alternative names

- Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

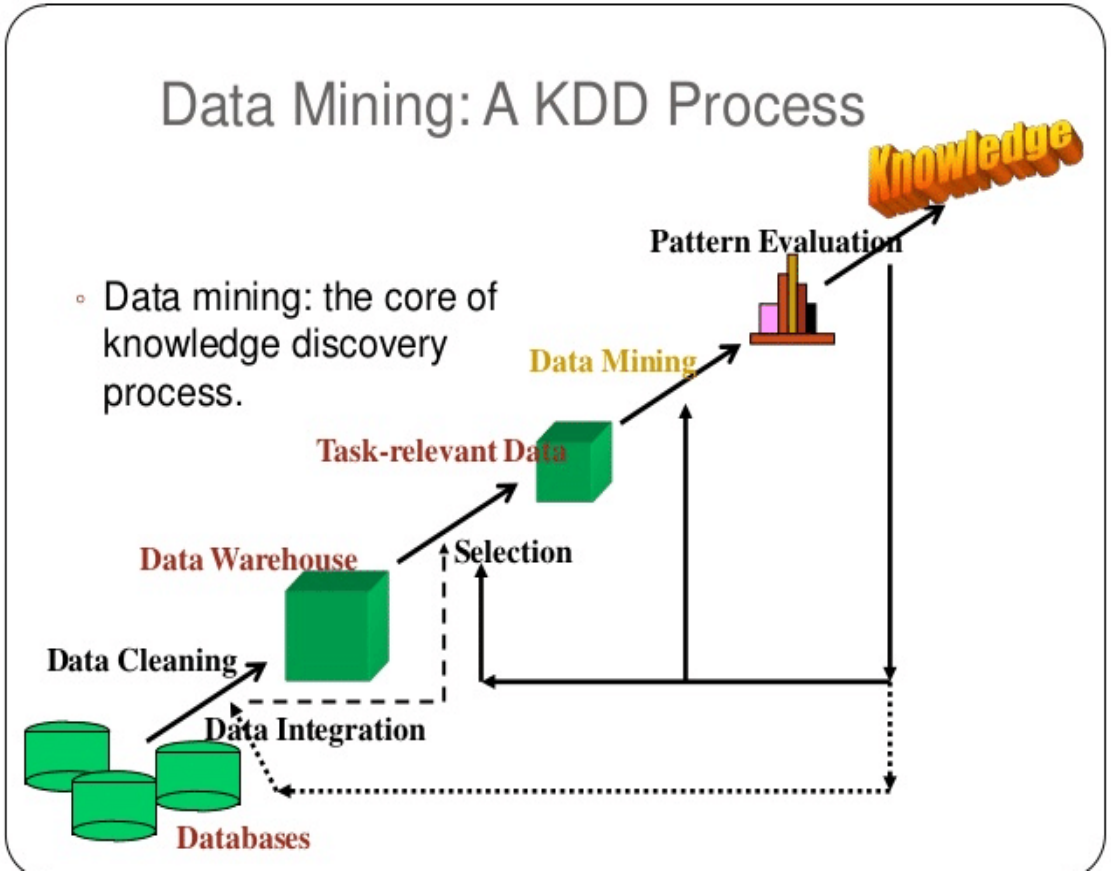
Watch out: Is everything "data mining"?

- 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

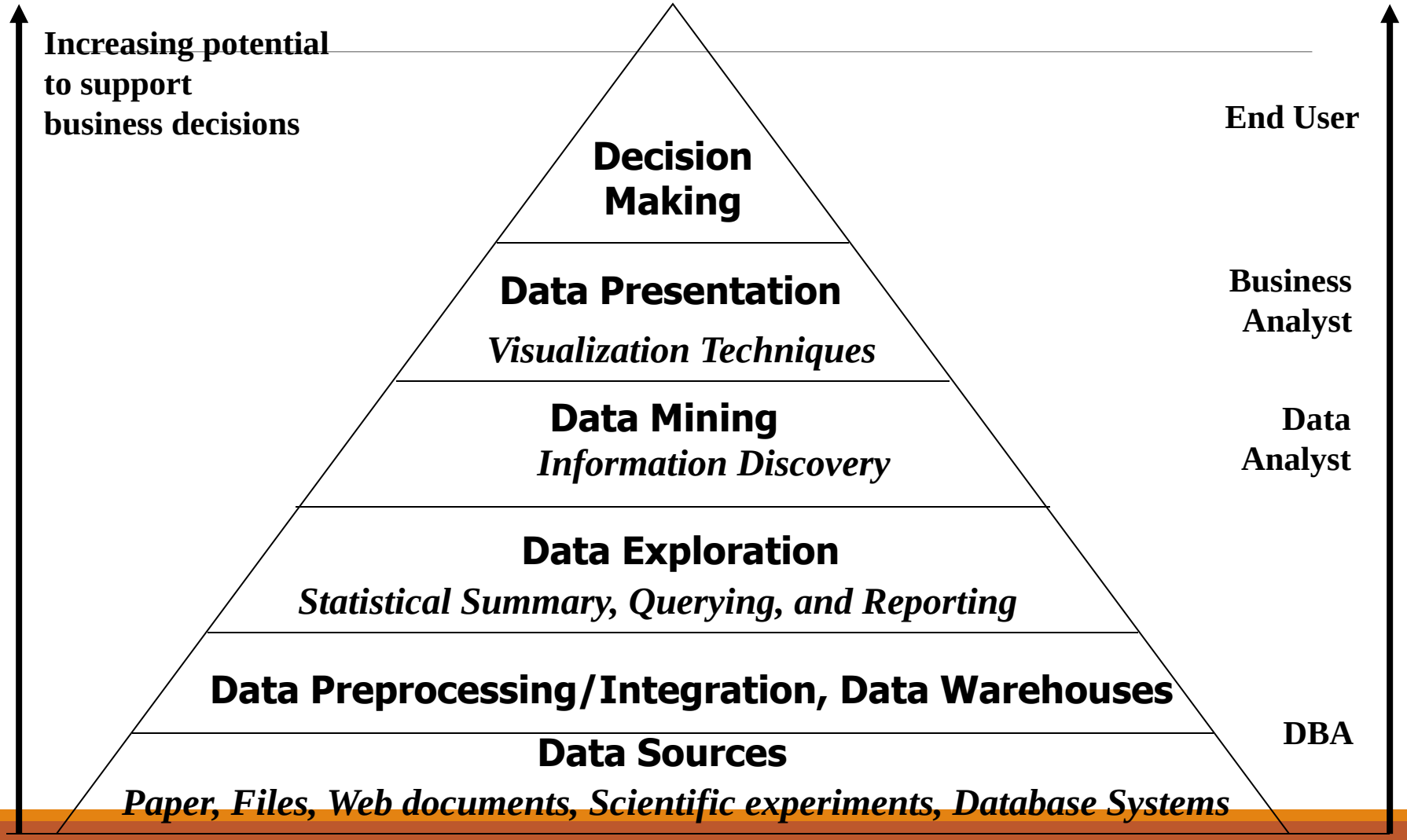


Example: A Web Mining Framework

Web mining usually involves

- Data cleaning
- Data integration from multiple sources
- Warehousing the data
- Data cube construction
- Data selection for data mining
- Data mining
- Presentation of the mining results
- Patterns and knowledge to be used or stored into knowledge-base

Data Mining in Business Intelligence



Example: Mining vs. Data Exploration

Business intelligence view

- Warehouse, data cube, reporting but not much mining

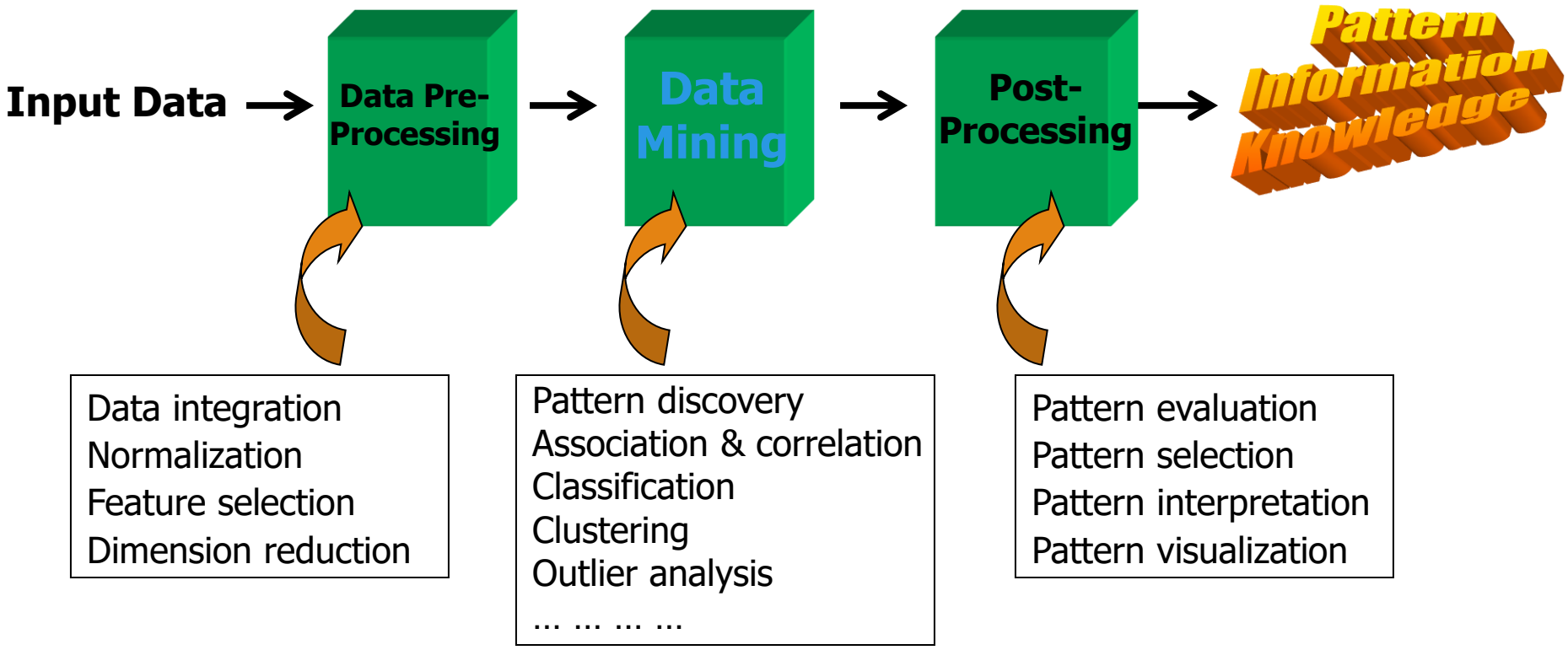
Business objects vs. data mining tools

Supply chain example: tools

Data presentation

Exploration

KDD Process: A Typical View from ML and Statistics



This is a view from typical machine learning and statistics communities

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

Multi-Dimensional View of Data Mining

Data to be mined

- Database data (extended-relational, object-oriented, heterogeneous, legacy), data warehouse, transactional data, stream, spatiotemporal, time-series, sequence, text and web, multi-media, graphs & social and information networks

Knowledge to be mined (or: Data mining functions)

- Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
- Descriptive vs. predictive data mining
- Multiple/integrated functions and mining at multiple levels

Techniques utilized

- Data-intensive, data warehouse (OLAP), machine learning, statistics, pattern recognition, visualization, high-performance, etc.

Applications adapted

- Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

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, 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 Mining Function: Generalization

Information integration and data warehouse construction

- Data cleaning, transformation, integration, and multidimensional data model

Data cube technology

- Scalable methods for computing (i.e., materializing) multidimensional aggregates
- OLAP (online analytical processing)

Multidimensional concept description: Characterization and discrimination

- Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet region

Data Mining Function: Association and Correlation Analysis

Frequent patterns (or frequent itemsets)

- What items are frequently purchased together in your Walmart?

Association, correlation vs. causality

- A typical association rule
  - Diaper → Beer [0.5%, 75%] (support, confidence)
- Are strongly associated items also strongly correlated?

How to mine such patterns and rules efficiently in large datasets?

How to use such patterns for classification, clustering, and other applications?

Data Mining Function: 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 some unknown class labels

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, direct marketing, classifying stars, diseases, web-pages, …

Data Mining Function: Cluster Analysis

Unsupervised learning (i.e., Class label is unknown)

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

Data Mining Function: Outlier Analysis

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: by product of clustering or regression analysis, …
- Useful in fraud detection, rare events analysis

Time and Ordering: Sequential Pattern, Trend and Evolution Analysis

Sequence, 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
- Motifs and biological sequence analysis
  - Approximate and consecutive motifs
- Similarity-based analysis

Mining data streams

- Ordered, time-varying, potentially infinite, data streams

Structure and Network Analysis

Graph mining
- Finding frequent subgraphs (e.g., chemical compounds), trees (XML), substructures (web fragments)

Information network analysis
- Social networks: actors (objects, nodes) and relationships (edges)
  - e.g., author networks in CS, terrorist networks
- Multiple heterogeneous networks
  - A person could be multiple information networks: friends, family, classmates, …
- Links carry a lot of semantic information: Link mining

Web mining
- Web is a big information network: from PageRank to Google
- Analysis of Web information networks
  - Web community discovery, opinion mining, usage mining, …

Evaluation of Knowledge

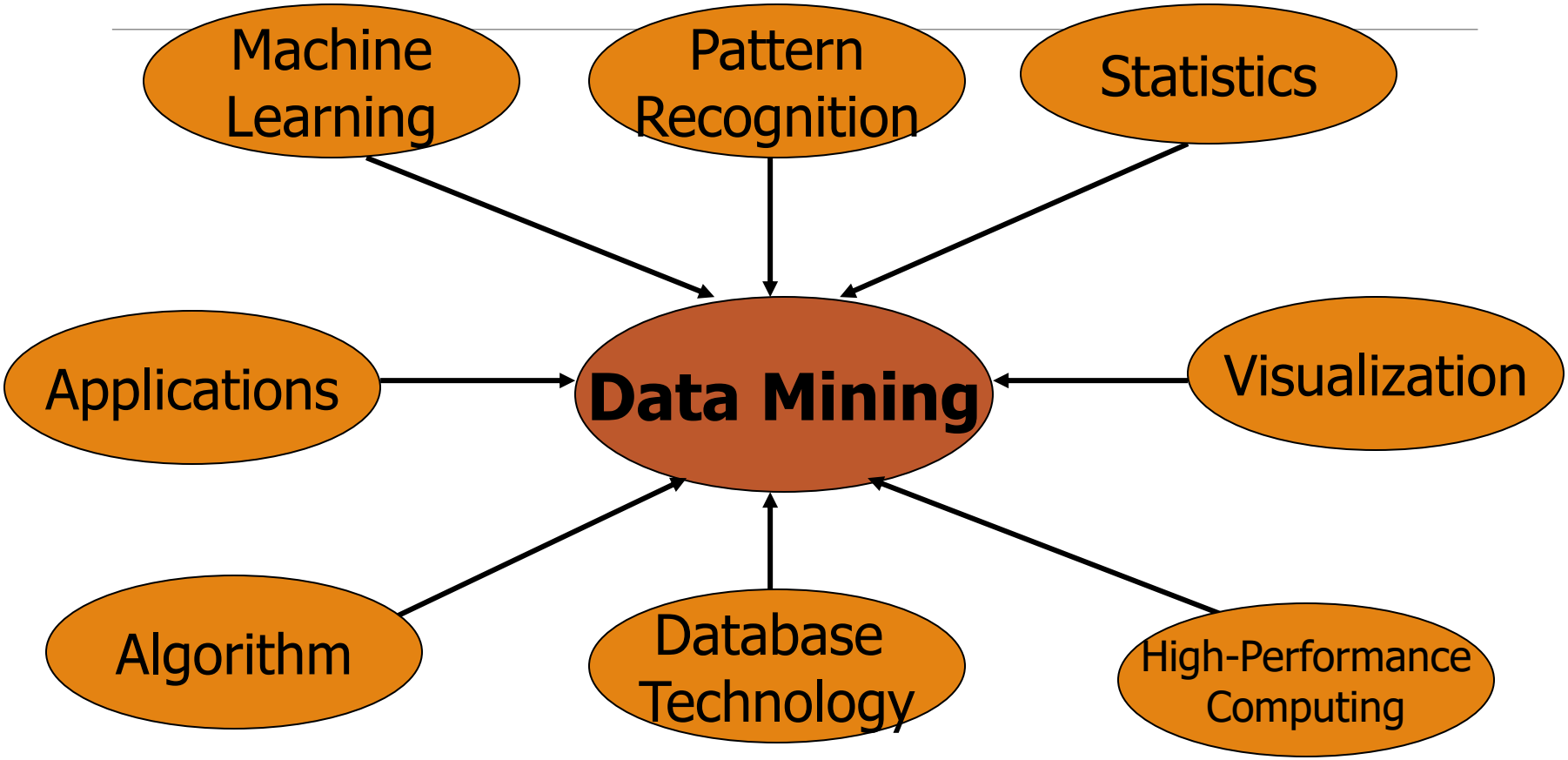
Are all mined knowledge interesting?

- 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, …

Evaluation of mined knowledge → directly mine only interesting knowledge?

- Descriptive vs. predictive
- Coverage
- Typicality vs. novelty
- Accuracy
- Timeliness
- …

Data Mining: Confluence of Multiple Disciplines



Why Confluence of Multiple Disciplines?

Tremendous amount of data

- Algorithms must be highly scalable to handle such as tera-bytes of data

High-dimensionality of data

- Micro-array may have tens of thousands of dimensions

High complexity of data

- Data streams and sensor data
- Time-series data, temporal data, sequence data
- Structure data, graphs, social networks and multi-linked data
- Heterogeneous databases and legacy databases
- Spatial, spatiotemporal, multimedia, text and Web data
- Software programs, scientific simulations

New and sophisticated applications

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 (e.g., IEEE Computer, Aug. 2009 issue)
- 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

- Mining various and new kinds of knowledge
- Mining knowledge in multi-dimensional space
- Data mining: An interdisciplinary effort
- Boosting the power of discovery in a networked environment
- Handling noise, uncertainty, and incompleteness of data
- Pattern evaluation and pattern- or constraint-guided mining

User Interaction

- Interactive mining
- Incorporation of background knowledge
- Presentation and visualization of data mining results

Major Issues in Data Mining

Efficiency and Scalability

- Efficiency and scalability of data mining algorithms
- Parallel, distributed, stream, and incremental mining methods

Diversity of data types

- Handling complex types of data
- Mining dynamic, networked, and global data repositories

Data mining and society

- Social impacts of data mining
- Privacy-preserving data mining
- Invisible data mining

Activity

- Discuss any 4 real world applications of data mining.