CS4.405 : Data Analytics-I

P. Krishna Reddy

IIIT Hyderabad
E-mail: pkreddy@iiit.ac.in
http://www.iiit.net/~pkreddy

Introduction: Presentation Outline

- **Why Data Analytics (or data mining)?**
  - **Courses you have Completed**
  - Content of human mind
  - Sample data mining problems
- Data mining definition and KDD process
- Multidimensional view of data mining
- Overview of data mining functionalities
- Issues in data mining
- Course outline
- Summary and History of data mining society

Courses You have Completed

- Digital Logic Design
  - Able to carry out arithmetic and logic operations
- Operating systems
  - Able to store the file and retrieve the file
- Programming language
  - Able to write a program, given the requirement
- Algorithms, data structures
  - Efficient computation given the requirement: sorting and searching
- Database systems
  - Able to store data and retrieve data based on SQL

- Questions
  - What will you do with 10,000 or 100,000 rows result?
  - Can we get some new insights? If yes, how?

Introduction: Presentation Outline

- **Why Data Analytics (or data mining)?**
  - Courses you have Completed
  - **Content of human mind**
  - Sample data mining problems
- Data mining definition and KDD process
- Multidimensional view of data mining
- Overview of data mining functionalities
- Issues in data mining
- Course outline
- Summary and History of data mining society

Data, Information, Knowledge, and Wisdom
by Gene Bellinger, Durval Castro, Anthony Mills

- According to Russell Ackoff, content of human mind can be classified into five categories: Data, Information, Knowledge, Understanding and wisdom

- Data: Symbols
  - **Data represents a fact or a statement of event without relation to other things.**
  - Data is raw. It simply exists and has no significance beyond its existence (in and of itself). It can exist in any form, usable or not. It does not have meaning of itself. In computer parlance, a spreadsheet generally starts out by holding data.
  - Ex: It is raining.

Content of Human Mind

- **Information:** Data that are processed to be useful; provides answer to “who”, “what”, “where”, and “when” questions.
  - Information is data that has been given meaning by way of relational connection. This "meaning" can be useful, but does not have to be.
  - Information embodies **the understanding** of a relationship of some sort, possibly cause and effect.
  - Example The temperature dropped 15 degrees and then it started raining.
  - In computer parlance, a relational database makes information from the data stored within it.

Content of Human Mind

- Knowledge: application of data and information; answers "how" questions.
  - Knowledge represents a pattern that connects and generally providing a high level of predictability as what is described or what will happen next.
  - Knowledge is the appropriate collection of information, such that it's intent is to be useful. Knowledge is a deterministic process.
  - When someone "memorizes" information (as less-aspiring test-bound students often do), then they have amassed knowledge. This knowledge has useful meaning to them, but it does not provide for, in and of itself, an integration such as would infer further knowledge.

Content of Human Mind

- Knowledge: application of data and information; answers “how” questions.
  - …
  - For example, elementary school children memorize, or amass knowledge of, the "times table". They can tell you that "2 x 2 = 4" because they have amassed that knowledge (it being included in the times table). But when asked what is "1267 x 300", they can not respond correctly because that entry is not in their times table.
  - To correctly answer such a question requires a true cognitive and analytical ability that is only encompassed in the next level... understanding.
  - Ex: If the humidity is very high and the temperature drops suddenly the atmosphere is often unlikely to be able to hold the moisture so it rains.
  - In computer parlance, most of the applications we use (modeling, simulation, etc.) exercise some type of stored knowledge.

Content of Human Mind

- Understanding: appreciation of why
  - Understanding is an interpolative and probabilistic process. It is cognitive and analytical.
  - It is the process by which I can take knowledge and synthesize new knowledge from the previously held knowledge.
  - The difference between understanding and knowledge is the difference between "learning" and "memorizing".
  - People who have understanding can undertake useful actions because they can synthesize new knowledge, or in some cases, at least new information, from what is previously known (and understood).
  - **That is, understanding can build upon currently held information, knowledge and understanding itself.**
  - In computer parlance, AI systems possess understanding in the sense that they are able to synthesize new knowledge from previously stored information and knowledge.

Content of human mind

- **Wisdom: evaluated understanding**
- Wisdom embodies more of an understanding of fundamental principles embodied within the knowledge that are essentially the basis for the knowledge being what it is. Wisdom is essentially systemic.
  - Ex: It rains because it rains. And this encompasses an understanding of all the interactions that happen between raining, evaporation, air currents, temperature gradients, changes, raining.
  - Wisdom is an extrapolative and non-deterministic, non-probabilistic process.
  - It calls upon all the previous levels of consciousness, and specifically upon special types of human programming (moral, ethical codes, etc.).
  - **It beckons to give us understanding about which there has previously been no understanding, and in doing so, goes far beyond understanding itself.**

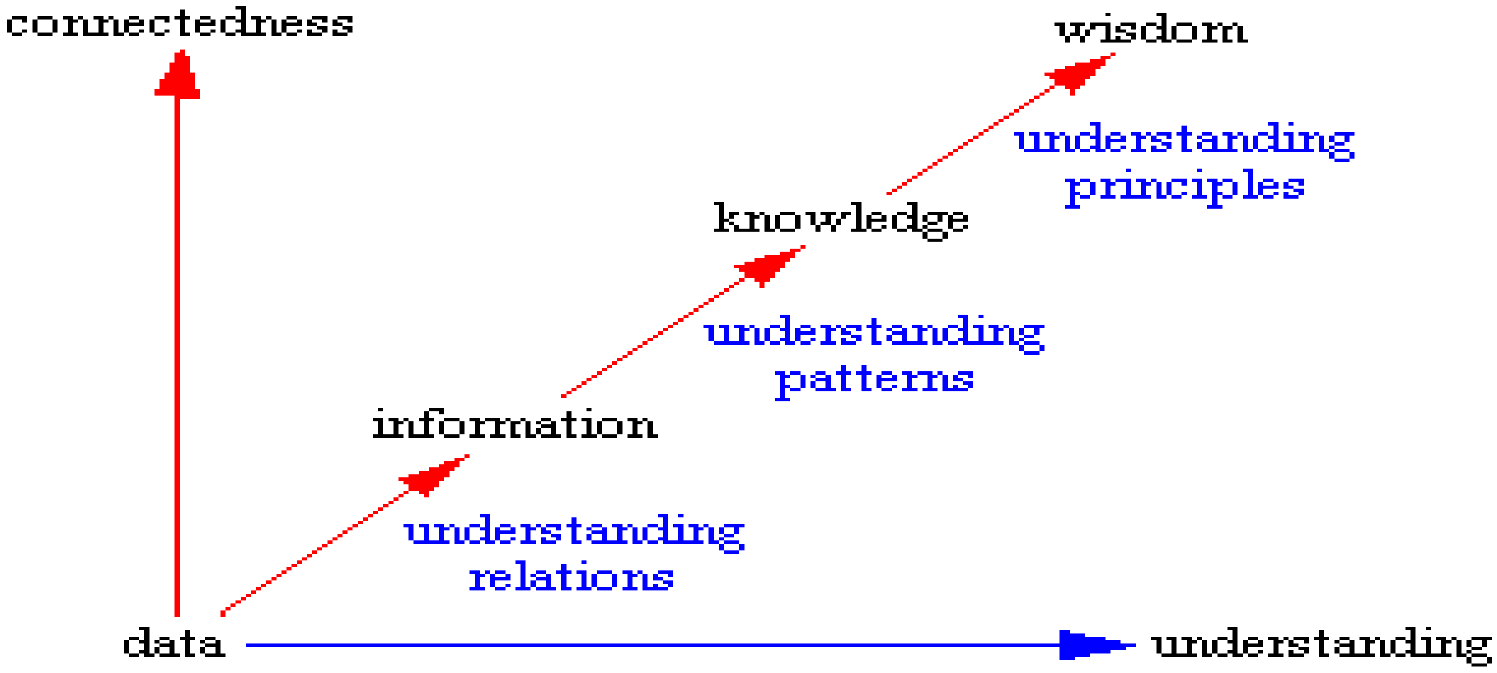
Content of human mind

- **Wisdom: evaluated understanding**
  - It is the essence of philosophical probing.
  - Unlike the previous four levels, it asks questions to which there is no (easily-achievable) answer, and in some cases, to which there can be no humanly-known answer period.
  - Wisdom is therefore, the process by which we also discern, or judge, between right and wrong, good and bad.

  - I personally believe that computers do not have, and will never have the ability to posses wisdom.
  - Wisdom is a uniquely human state, or as I see it, wisdom requires one to have a soul, for it resides as much in the heart as in the mind. And a soul is something machines will never possess (or perhaps I should reword that to ay, a soul is something that, in general, will never possess a machine).

Content of human mind

- Understanding that supports the transition from each stage to the next.
- Understanding is not a separate level of its own.
- Important: Ability to connect



Examples

- Example 1
  - Abugt dbesbt regtc uatn s uitrzt.
  - ubtxte pstye ysote anet sser extess
  - bxtedstes bet3 ibtes otesb tapbesct ehracts
- There is no foundation for you to connect with the pattern. If you know the translation, these are Newton's 3 laws of motion

- Example 2
  - I have a box.
  - The box is 3' wide, 3' deep, and 6' high.
  - The box is very heavy.
  - The box has a door on the front of it.
  - When I open the box it has food in it.
  - It is colder inside the box than it is outside.
  - You usually find the box in the kitchen.
  - There is a smaller compartment inside the box with ice in it.
  - When you open the door the light comes on.
  - When you move this box you usually find lots of dirt underneath it.
  - Junk has a real habit of collecting on top of this box.
- It is a refrigerator
- But, if you lived in a society that had never seen a refrigerator you might still be scratching your head as to what the sequence of statements referred to.

Introduction: Presentation Outline

- **Why Data Analytics (or data mining)?**
  - Courses you have Completed
  - Content of human mind
  - **Sample data mining problems**
- Data mining definition and KDD process
- Multidimensional view of data mining
- Overview of data mining functionalities
- Issues in data mining
- Course outline
- Summary and History of data mining society

Sample data mining problem # 1

I manage a supermarket (restaurant, video store, book store) and my cash register (or web site) pumps transactions into my DB.

- Can you help me visualize my sales ?
- Can you profile my customers ?
- Tell me something interesting …
- I do not know statistics, and I do not want to hire statisticians.

Sample data mining problem #2

- I am an astronomer and I have sky survey 3 tera bytes of data, 2 billion objects.
  - Can you help to recognize the objects ?
  - Most of my data is beyond my reach.
  - Can you find new/unusual items in my data ?
  - Can you help me with basic manipulation, so I can focus on basic science ?

  - I know my data and statistics, but that is not enough …

Issue of Knowledge

- 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

Introduction: Presentation Outline

- Why Data Analytics (or data mining)?
  - Courses you have Completed
  - Content of human mind
  - Sample data mining problems
- **Data mining definition and KDD process**
- Multidimensional view of data mining
- Overview of data mining functionalities
- Issues in data mining
- Course outline
- Summary and History of data mining society

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

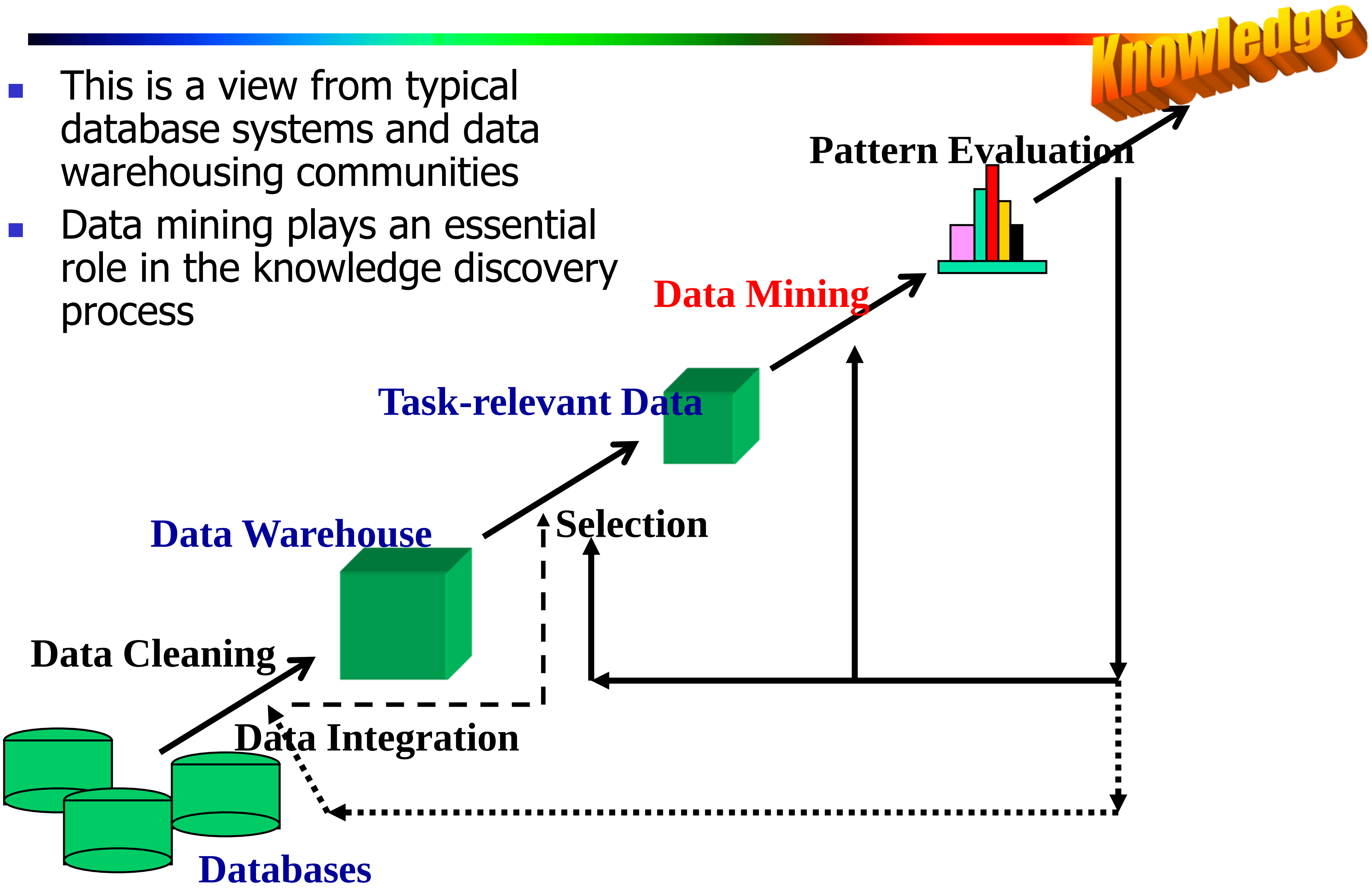
What is (not) Data Mining?

- **What is not Data Mining?**
  - Look up phone number in phone directory
  - Query a Web search engine for information about “Amazon”

- **What is Data Mining?**
  - Certain names are more prevalent in certain US locations (O’Brien, O’Rurke, O’Reilly… in Boston area)
  - Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com,)

Knowledge Discovery (KDD) Process

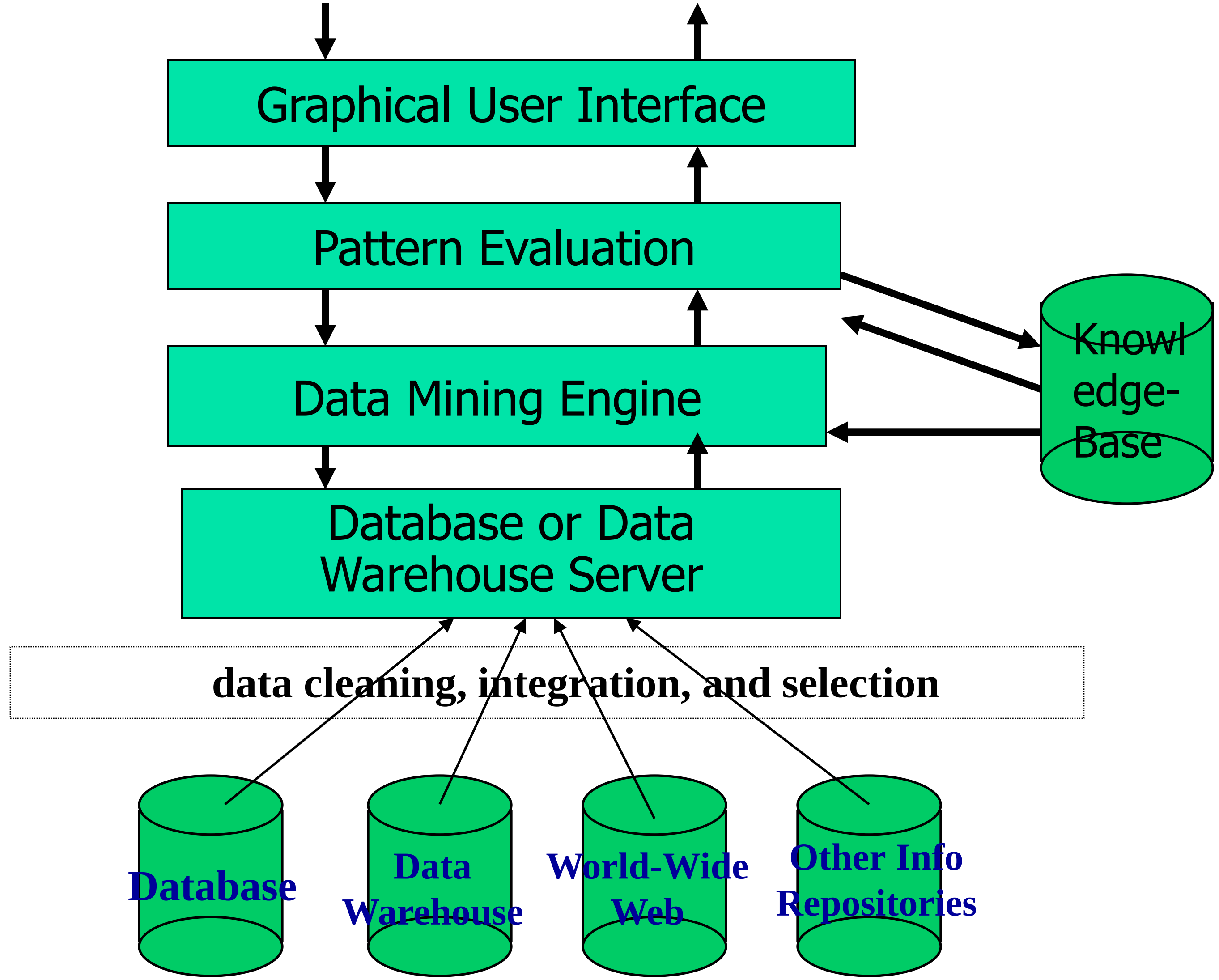
- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process



Steps of a KDD Process

- Learning the application domain:
  - relevant prior knowledge and goals of application
- Data cleaning: to remove noise and inconsistent data
- Data integration: Multiple data sources can be combined
- Creating a target data set: data selection
- Data cleaning and preprocessing: (may take 60% of effort!)
- Data reduction and transformation:
  - Find useful features, dimensionality/variable reduction, invariant representation.
- Choosing functions of data mining
  - summarization, association, classification, clustering.
- Choosing the mining algorithm(s)
- Data mining: search for patterns of interest
- Pattern evaluation and knowledge presentation
  - visualization, transformation, removing redundant patterns, etc.
- Use of discovered knowledge

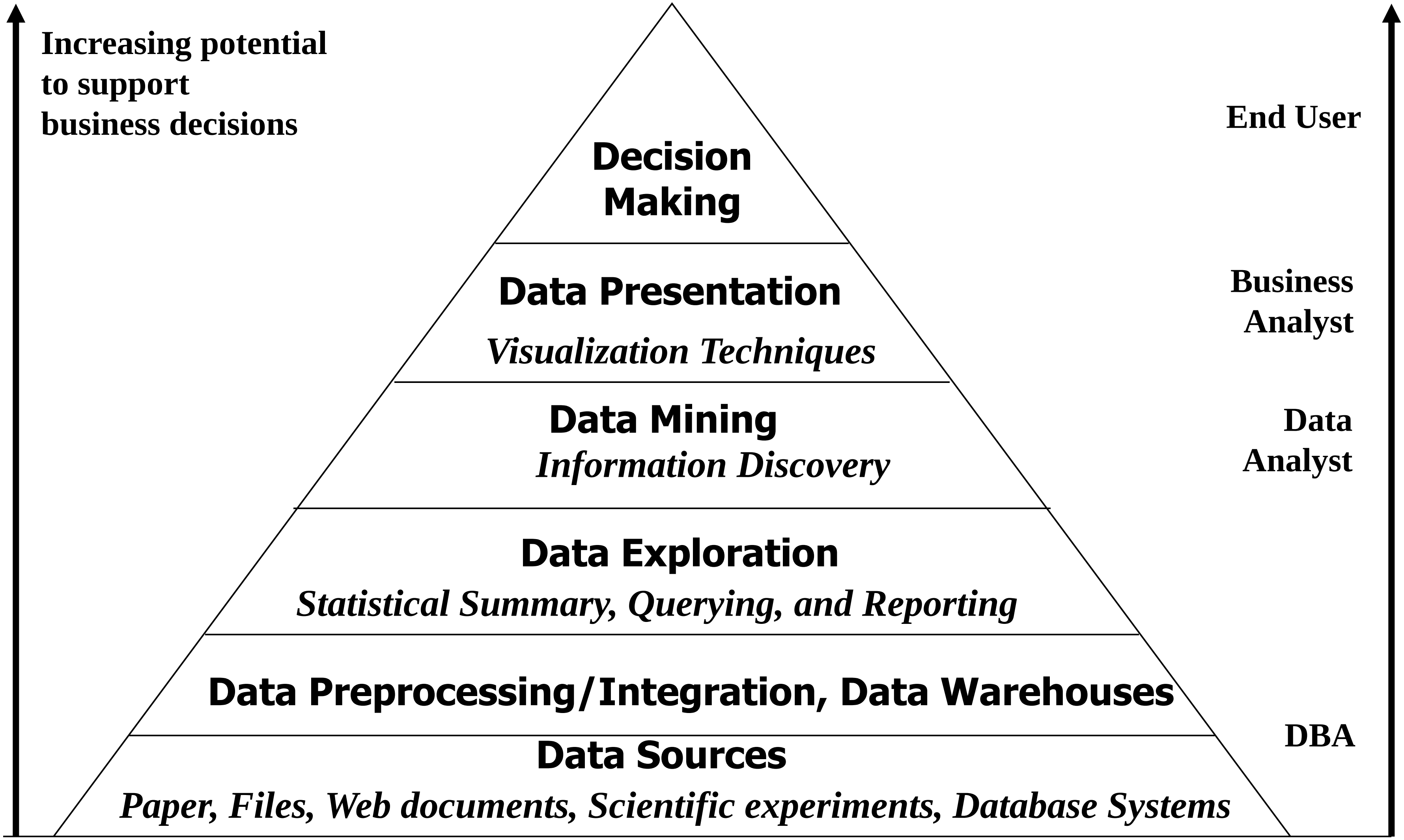
Architecture: Typical Data Mining System



Components of data mining system

- **Database, Data warehouse, World Wide Web or other information Repository**
  - Data cleaning and data integration techniques are performed on this data
- **Database and data warehouse server:** Responsible for fetching the relevant data, based on the user's data mining request.
- **Knowledge-base:** Domain knowledge which is used to guide the data mining process.
  - Attribute levels, semantics, user beliefs, pattern interestingness, thrsholds, meta data
- **Data mining engine:** Set of functional modules for tasks such as characterization, summarization, association, classification, clustering, outlier extraction
- **Pattern evaluation:** Employees interestingness measures
  - Put the evaluation pattern as much deep as you can so that one can optimize.
- **User interface:** communication between users and the data mining system.

Data Mining in Business Intelligence



Increasing potential
to support
business decisions

Decision Making

Data Presentation

Visualization Techniques

Data Mining

Information Discovery

Data Exploration

Statistical Summary, Querying, and Reporting

Data Preprocessing/Integration, Data Warehouses

Data Sources

Paper, Files, Web documents, Scientific experiments, Database Systems

End User

Business Analyst

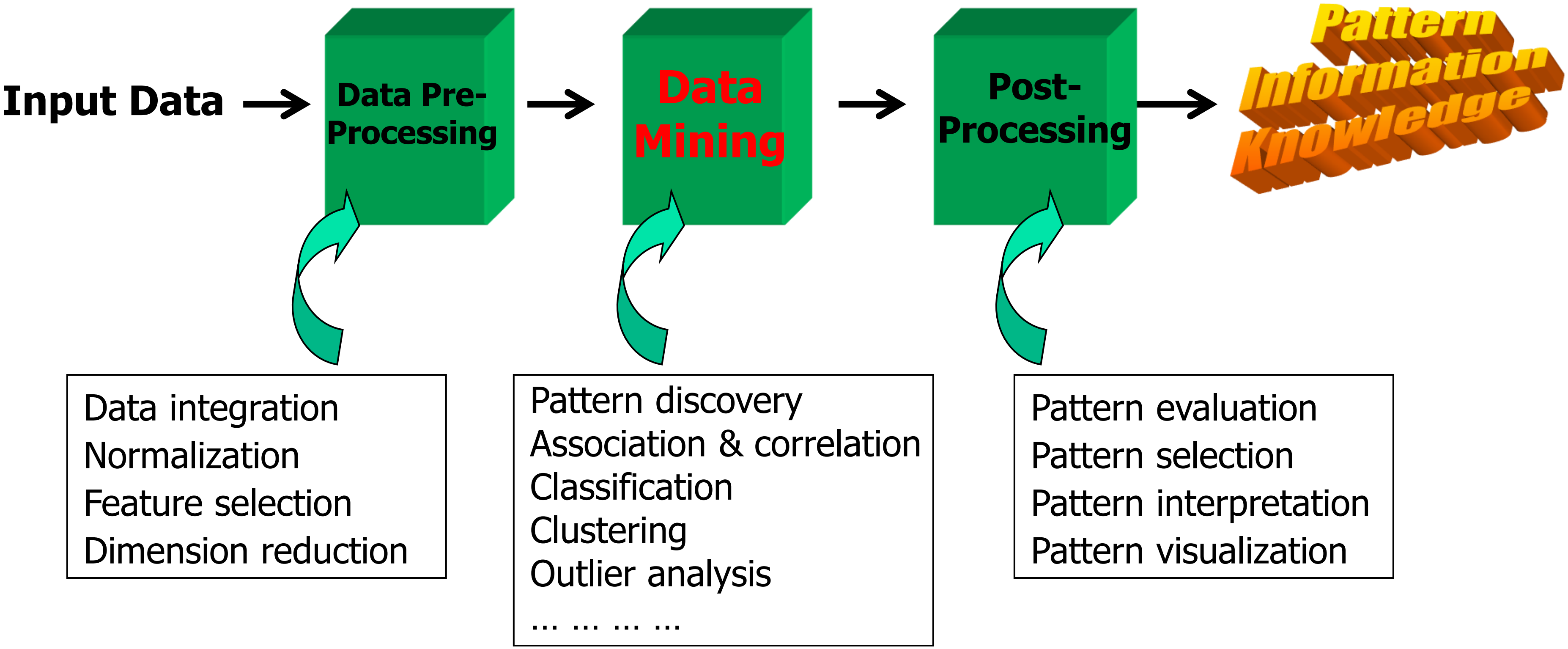
Data Analyst

DBA

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



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

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

Introduction: Presentation Outline

- Why Data Analytics (or data mining)?
  - Courses you have Completed
  - Content of human mind
  - Sample data mining problems
- Data mining definition and KDD process
- **Multidimensional view of data mining**
- Overview of data mining functionalities
- Issues in data mining
- Course outline
- Summary and History of data mining society

Introduction: Presentation Outline

- Why Data Analytics (or data mining)?
  - Courses you have Completed
  - Content of human mind
  - Sample data mining problems
- Data mining definition and KDD process
- **Multidimensional view of data mining**
- Overview of data mining functionalities
- Issues in data mining
- Course outline
- Summary and History of data mining society

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

Introduction: Presentation Outline

- Why Data Analytics (or data mining)?
  - Courses you have Completed
  - Content of human mind
  - Sample data mining problems
- Data mining definition and KDD process
- Multidimensional view of data mining
- **Overview of data mining functionalities**
- Issues in data mining
- Course outline
- Summary and History of data mining society

Data Mining Functionalities

- Generalization/Summarization: characterization and discrimination
- Pattern mining, Association mining, correlation
- Classification
- Clustering
- Outlier analysis
- Sequential, trend and evolution analysis
- Structure and network analysis

Data Mining Function: (1) 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: (2) 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: (3) 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: (4) 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: (5) 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

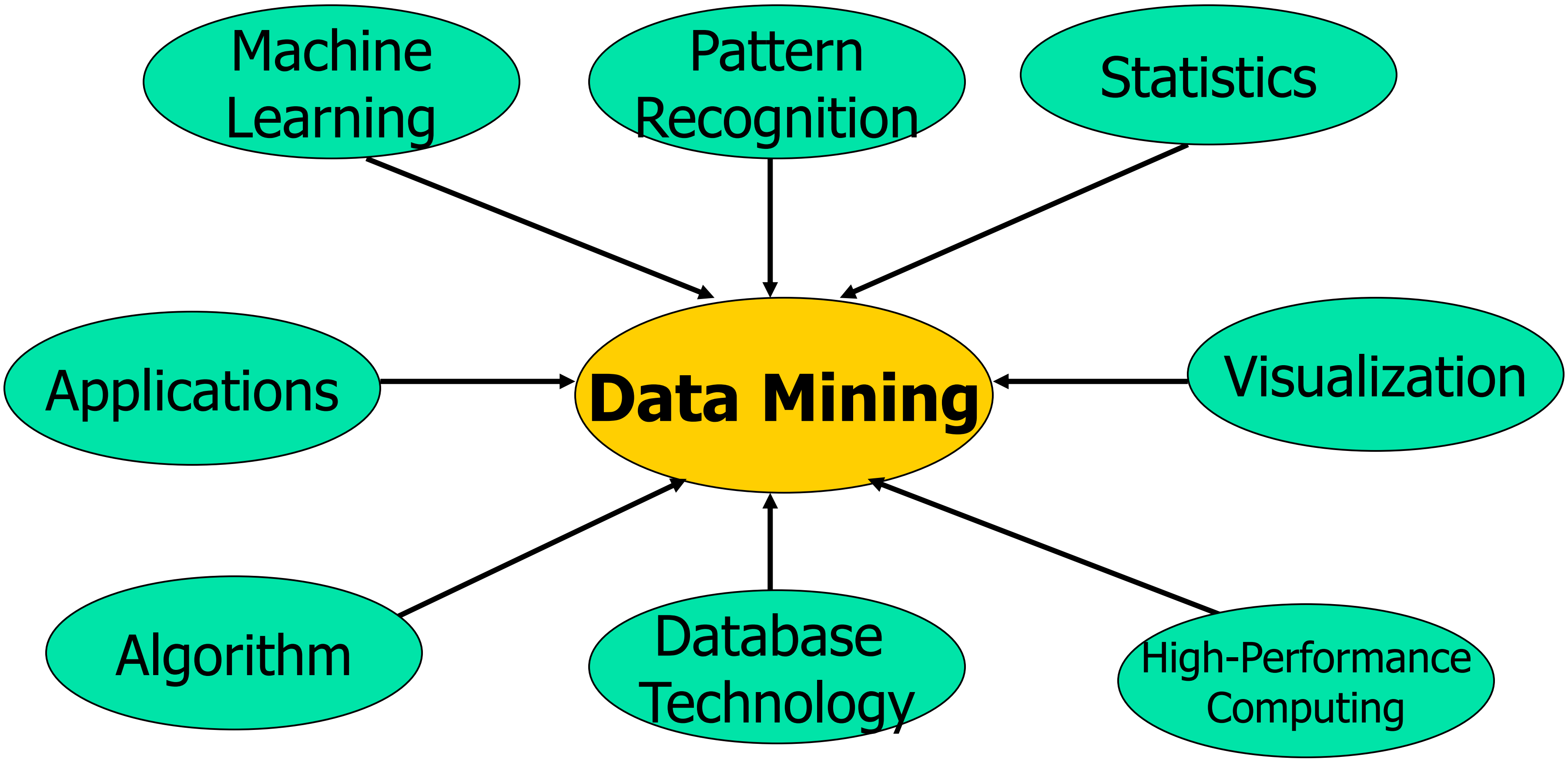
Structure and Network Analysis

- 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

Introduction: Presentation Outline

- Why Data Analytics (or data mining)?
  - Courses you have Completed
  - Content of human mind
  - Sample data mining problems
- Data mining definition and KDD process
- Multidimensional view of data mining
- Overview of data mining functionalities
- **Issues in data mining**
- Course outline
- Summary and History of data mining society

Major Issues in Data Mining (1)

- 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 (2)

- 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

Introduction: Presentation Outline

- Why Data Analytics (or data mining)?
  - Courses you have Completed
  - Content of human mind
  - Sample data mining problems
- Data mining definition and KDD process
- Multidimensional view of data mining
- Overview of data mining functionalities
- Issues in data mining
- **Course outline**
- Summary and History of data mining society

Course Outline
(CS4.405: Data Analytics-I)

Prerequisite Course / Knowledge:

(i) Data and Applications, *or equivalent courses that cover Data modelling, normalization, SQL*

(ii) First courses on programming, data-structures and algorithms

(iii) Basics of Python language, to be able to use relevant libraries and toolkits for data analytics

Objective:

- In a computerized and networked society, vast amount of data is being collected every day in multiple domains.
- We are drowning in data, but starving for knowledge or actionable insights.
- Data mining or data analytics constitute a collection of concepts and algorithms, which are being developed to answer “how” questions by extracting interesting and useful knowledge of from large data.
- Data analytics based platforms are being operated in multiple domains to extract valuable and actionable insights from the data to improve the business performance.
- **The objective** of this first level course is to learn the important concepts and algorithms related to data analytics and data mining functionalities such as summarization, pattern mining, classification, clustering and outlier analysis.

Course Outcomes (COs)

- After completing the course successfully, the students are able to
  - CO-1. describe the concepts of data summarization, data warehousing, pattern mining, classification and clustering approaches
  - CO-2. perform the task of data summarization, pattern mining, classification and clustering based on the requirement.
  - CO-3. prescribe a single or a combination of data summarization, pattern mining, classification and clustering approaches for the problem scenario of a business/ organization.
  - CO-4. construct the improved data analytics methods for the existing services.
  - CO-5. formulate new data mining problems for creating new services and design the corresponding solutions

Detailed Syllabus

- **Introduction (1.5 hour): Definition, KDD framework, Issues in data mining.**
- Data summarization (7.5 hrs): Characterization, Discrimination, data warehousing techniques (Multidimensional data model, Data warehousing architecture, Data cube computation and OLAP technology)
- Concepts and algorithms for mining patterns and associations (9 hours) (Frequent item-set generation, A priori and FP-growth algorithm, Evaluation of Association patterns) and preprocessing
- Concepts and algorithms related to classification and regression (9hrs) (Overview, Decision tree induction, Over-fitting and under-fitting, Scalable decision tree algorithms, Bayesian Classification, Regression-based Prediction methods (9 hours)
- Concepts and algorithms for clustering the data (9 hours) (Overview, Types of Data, K-means, Aglomerative clustering, Clustering algorithms (DBSCAN, BIRCH, CURE, ROCK, CHAMELEON)).
- Outlier analysis and future trends (graph mining, spatio-temporal mining). (3 hours)

Lab

- Five mini projects related to the above syllabus will be done by students in the laboratory

Reference Books and materials:

1. **Book: Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, Third edition, 2012, Elseiver Inc.**

2. Book: Pang-Nong Tan, Michael Steinbach and Vipin Kumar, Introduction to Data Mining, 2006, Pearson Education.

1. Research Papers: About 25 research papers from the proceeding of the conferences and journals related to data summarization, data warehousing, pattern mining, classification, clustering, outlier detection.

Assessment methods and weightages

- Two Class Room tests: 10 marks (5+5)
- Mid Semester Examination in theory: 20 marks,
- End Semester Examination in Theory: 40 marks,
- Assessment of five mini projects in Laboratory: 30 marks

Lab Assignments

- You will be given the lab assignments in advance
  - The instructor will provide you with adequate background information to do the assignment
  - Some installations will be necessary. Install well in advance and if you encounter any problems with any installation, you can contact your TA
- Policies for lab
  - You should try to solve any programming issues by yourself first, remember that troubleshooting is how you would actually be learning a lot
  - Every time you encounter any minor programming problem, if you ask the TA or instructor, you will lose out on an excellent learning experience
  - Only if you are unable to solve the problem after putting in a reasonable amount of effort, you should contact your TA or your instructor
  - You should also look online for solutions to your problems, but do not copy code from anywhere
- Grading criteria
  - The overall quality of your lab assignment submission in terms of correctness, quality of code, system design etc.

Plagiarism

- This course has a zero-tolerance policy w.r.t. plagiarism
- Any instance of plagiarism will result in serious penalties (e.g., an F grade for the entire course, among other penalties)
- Forget about doing any kind of plagiarism

Deadlines

- Strict deadlines: You will not be able to submit after the deadline has already passed.

Accessing the Course Materials

- The presentations, the materials, lab assignments are available on the course portal
  - Pls. access the course portal regularly
- All students should procure tపెం book, as soon as possible

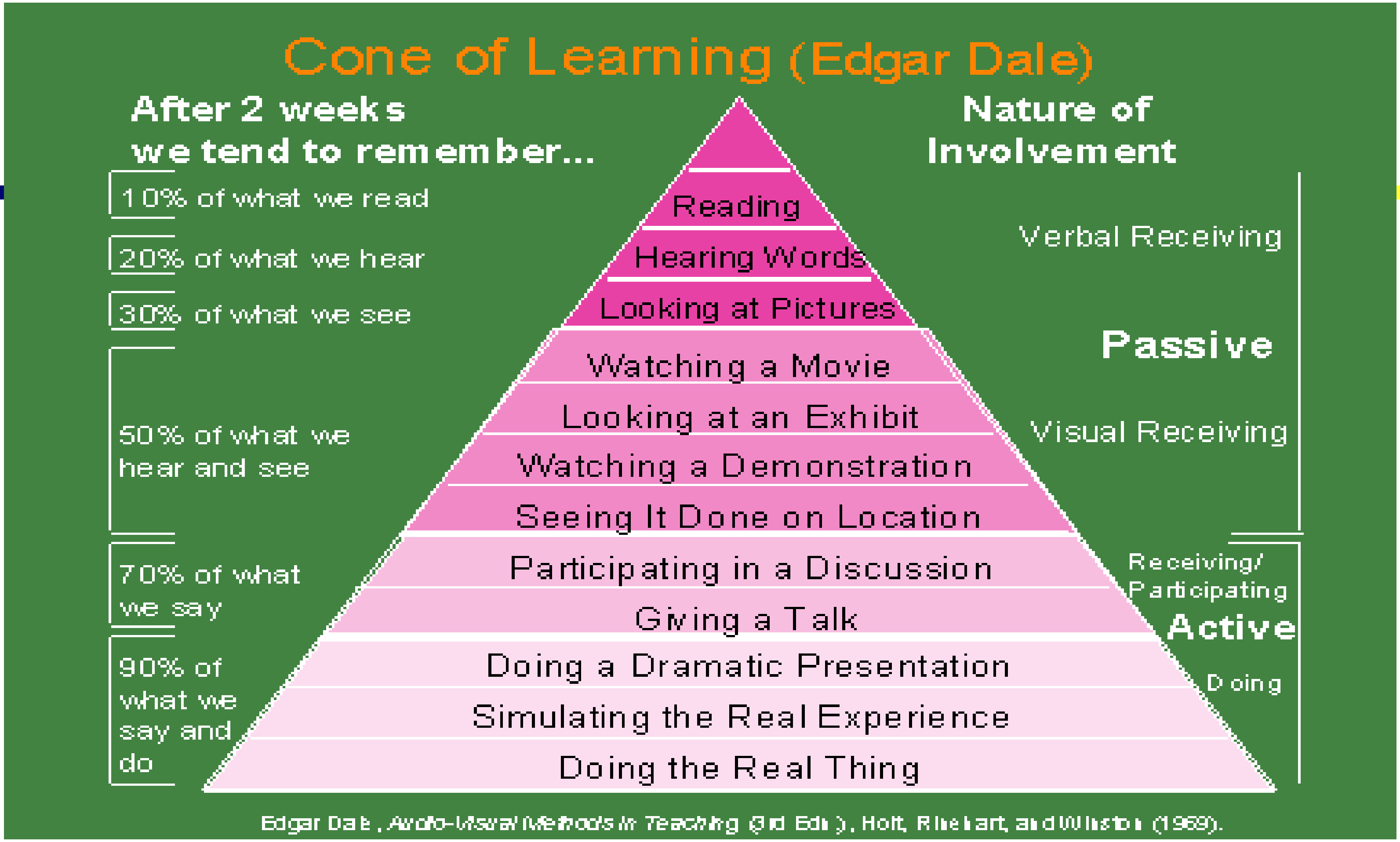
- **Book: Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, Third edition, 2012, Elseiver Inc.**

Asking Questions

- Theory and lab related questions can be asked through the course portal
  - It is better, because all other students can gain the related knowledge
  - Try to ask as many subject/lab questions as possible.
  - Do not hesitate to ask silly/simple questions regarding lab or theory
    - And try to get the problem resolved as soon as possible

Maximize the benefit from the course

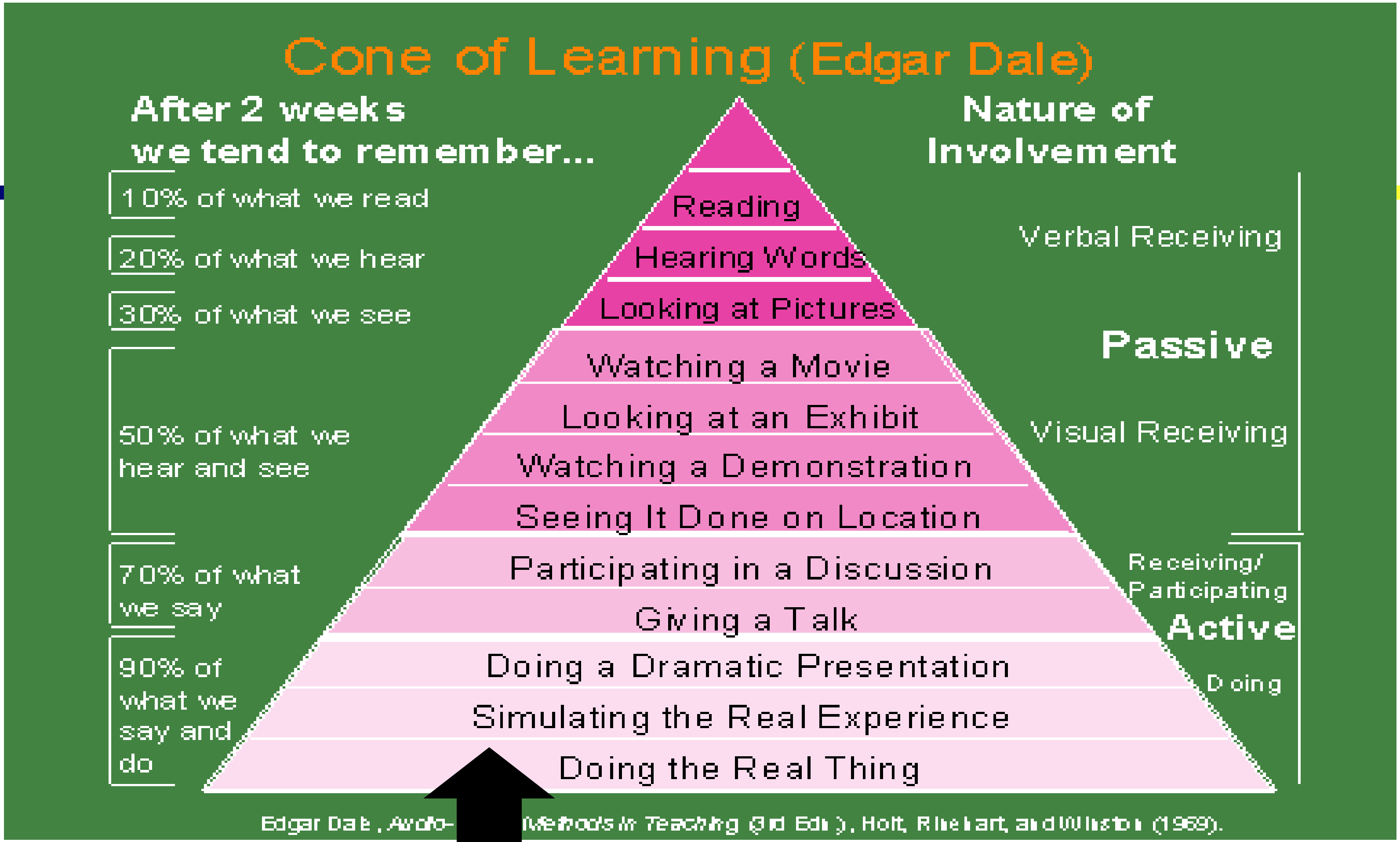
- In the course, we are going study about an important technology
  - Data scientist or data analyst
- Focus on a thorough understanding of the concepts, not on memorizing
  - Understand every sentence of the book



Cone of Learning (Edgar Dale)

Edgar Dale, *Audio-Visual Methods in Teaching* (3rd Edn.), Holt, Rinehart, and Winston (1969).

Source: http://www.cals.ncsu.edu/agexed/sae/ppt1/sld012.htm



Edgar Dale, *Audio-* *Methods in Teaching* (3rd Edn.), Holt, Rinehart, and Winston (1969).

Bottomline: Do the assignments sincerely because it will facilitate you in **INTERNALIZING** the ideas/techniques you learnt in this course.

Source: http://www.cals.ncsu.edu/agexed/sae/ppt1/sld012.htm

Presentation Outline

- Why Data Analytics (or data mining)?
  - Courses you have Completed
  - Content of human mind
  - Sample data mining problems
- Data mining definition and KDD process
- Multidimensional view of data mining
- Overview of data mining functionalities
- Issues in data mining
- Course outline
- **Summary and History of data mining society**

Summary

- Data mining: Discovering interesting patterns and knowledge from massive amount of data
- A natural evolution of database technology, in great demand, with wide applications
- A KDD process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation
- Mining can be performed in a variety of data
- **Data mining functionalities:**
  - **characterization,**
  - **discrimination,**
  - **association,**
  - **classification,**
  - **clustering,**
  - **outlier and trend analysis, etc.**

A Brief History of Data Mining Society

- 1989 IJCAI Workshop on Knowledge Discovery in Databases
  - Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994 Workshops on Knowledge Discovery in Databases
  - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
  - Journal of Data Mining and Knowledge Discovery (1997)
- ACM SIGKDD conferences since 1998 and SIGKDD Explorations
- More conferences on data mining
  - PAKDD (1997), PKDD (1997), SIAM-Data Mining (2001), (IEEE) ICDM (2001), etc.
- ACM Transactions on KDD starting in 2007

Conferences and Journals on Data Mining

- KDD Conferences
  - ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases and Data Mining (KDD)
  - SIAM Data Mining Conf. (SDM)
  - (IEEE) Int. Conf. on Data Mining (ICDM)
  - European Conf. on Machine Learning and Principles and practices of Knowledge Discovery and Data Mining (ECML-PKDD)
  - Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD)
  - Int. Conf. on Web Search and Data Mining (WSDM)

- Other related conferences
  - DB conferences: ACM SIGMOD, VLDB, ICDE, EDBT, ICDT, …
  - Web and IR conferences: WWW, SIGIR, WSDM
  - ML conferences: ICML, NIPS
  - PR conferences: CVPR,
- Journals
  - Data Mining and Knowledge Discovery (DAMI or DMKD)
  - IEEE Trans. On Knowledge and Data Eng. (TKDE)
  - KDD Explorations
  - ACM Trans. on KDD

Where to Find References? DBLP, CiteSeer, Google

- Data mining and KDD (SIGKDD: CDROM)
  - Conferences: ACM-SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, PAKDD, etc.
  - Journal: Data Mining and Knowledge Discovery, KDD Explorations, ACM TKDD
- Database systems (SIGMOD: ACM SIGMOD Anthology—CD ROM)
  - Conferences: ACM-SIGMOD, ACM-PODS, VLDB, IEEE-ICDE, EDBT, ICDT, DASFAA
  - Journals: IEEE-TKDE, ACM-TODS/TOIS, JIIS, J. ACM, VLDB J., Info. Sys., etc.
- AI & Machine Learning
  - Conferences: Machine learning (ML), AAAI, IJCAI, COLT (Learning Theory), CVPR, NIPS, etc.
  - Journals: Machine Learning, Artificial Intelligence, Knowledge and Information Systems, IEEE-PAMI, etc.
- Web and IR
  - Conferences: SIGIR, WWW, CIKM, etc.
  - Journals: WWW: Internet and Web Information Systems,
- Statistics
  - Conferences: Joint Stat. Meeting, etc.
  - Journals: Annals of statistics, etc.
- Visualization
  - Conference proceedings: CHI, ACM-SIGGraph, etc.
  - Journals: IEEE Trans. visualization and computer graphics, etc.

More References

- S. Chakrabarti. Mining the Web: Statistical Analysis of Hypertex and Semi-Structured Data. Morgan Kaufmann, 2002
- R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, 2ed., Wiley-Interscience, 2000
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley & Sons, 2003
- U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996
- U. Fayyad, G. Grinstein, and A. Wierse, Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed., 2011
- D. J. Hand, H. Mannila, and P. Smyth, Principles of Data Mining, MIT Press, 2001
- T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., Springer-Verlag, 2009
- B. Liu, Web Data Mining, Springer 2006.
- T. M. Mitchell, Machine Learning, McGraw Hill, 1997
- G. Piatetsky-Shapiro and W. J. Frawley. Knowledge Discovery in Databases. AAAI/MIT Press, 1991
- P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Wiley, 2005
- S. M. Weiss and N. Indurkhya, Predictive Data Mining, Morgan Kaufmann, 1998
- I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2nd ed. 2005