COMP4433
Data Mining & Data Warehousing

Dr. Chung Fu Lai Korris
COMP@PolyU

Acknowledgements

- Part of the slides for this course were prepared based on Han and Kamber’s powerpoint slides for their popular textbook “Data Mining: Concepts and Techniques”
- Some figures and tables were adopted directly from this textbook and others reference materials, including web sites (e.g. Kdnuggets.com) and conference presentation slides

Roadmap

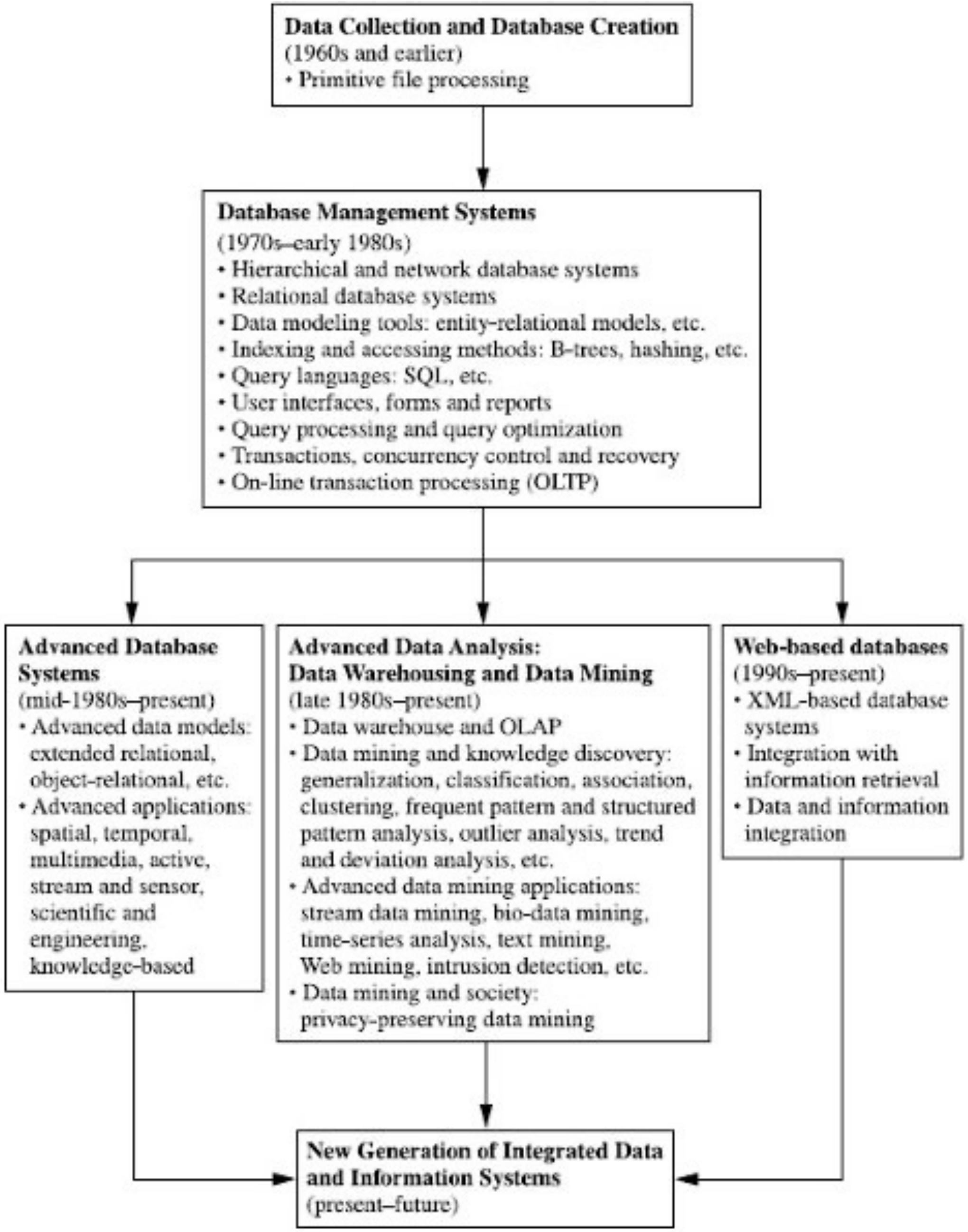
- Why data mining?
- What is data mining? Where is data mining?
- Data Scientist and Machine Learning Engineer
- Data mining tasks
- Potential applications
- KDD vs. DM, DM & BI
- How to mine data?
  - On what kind of data?
  - Classification of data mining systems
  - Major issues/problems in DM
- Data mining tools

Why data mining?

"Necessity is the Mother of Invention"

- *Data explosion problem*
  - Automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, data warehouses and other information repositories
- *We are drowning in data, but starving for knowledge!*
- *Solution: Data warehousing and data mining*
  - Data warehousing and on-line analytical processing (OLAP)
  - Extraction of interesting knowledge (rules, regularities, patterns, constraints) from data in large databases

Evolution of Database Technology



Data Collection and Database Creation
(1960s and earlier)
- Primitive file processing

Database Management Systems
(1970s–early 1980s)
- Hierarchical and network database systems
- Relational database systems
- Data modeling tools: entity-relational models, etc.
- Indexing and accessing methods: B-trees, hashing, etc.
- Query languages: SQL, etc.
- User interfaces, forms and reports
- Query processing and query optimization
- Transactions, concurrency control and recovery
- On-line transaction processing (OLTP)

Advanced Database Systems
(mid-1980s–present)
- Advanced data models: extended relational, object-relational, etc.
- Advanced applications: spatial, temporal, multimedia, active, stream and sensor, scientific and engineering, knowledge-based

Advanced Data Analysis: Data Warehousing and Data Mining
(late 1980s–present)
- Data warehouse and OLAP
- Data mining and knowledge discovery: generalization, classification, association, clustering, frequent pattern and structured pattern analysis, outlier analysis, trend and deviation analysis, etc.
- Advanced data mining applications: stream data mining, bio-data mining, time-series analysis, text mining, Web mining, intrusion detection, etc.
- Data mining and society: privacy-preserving data mining

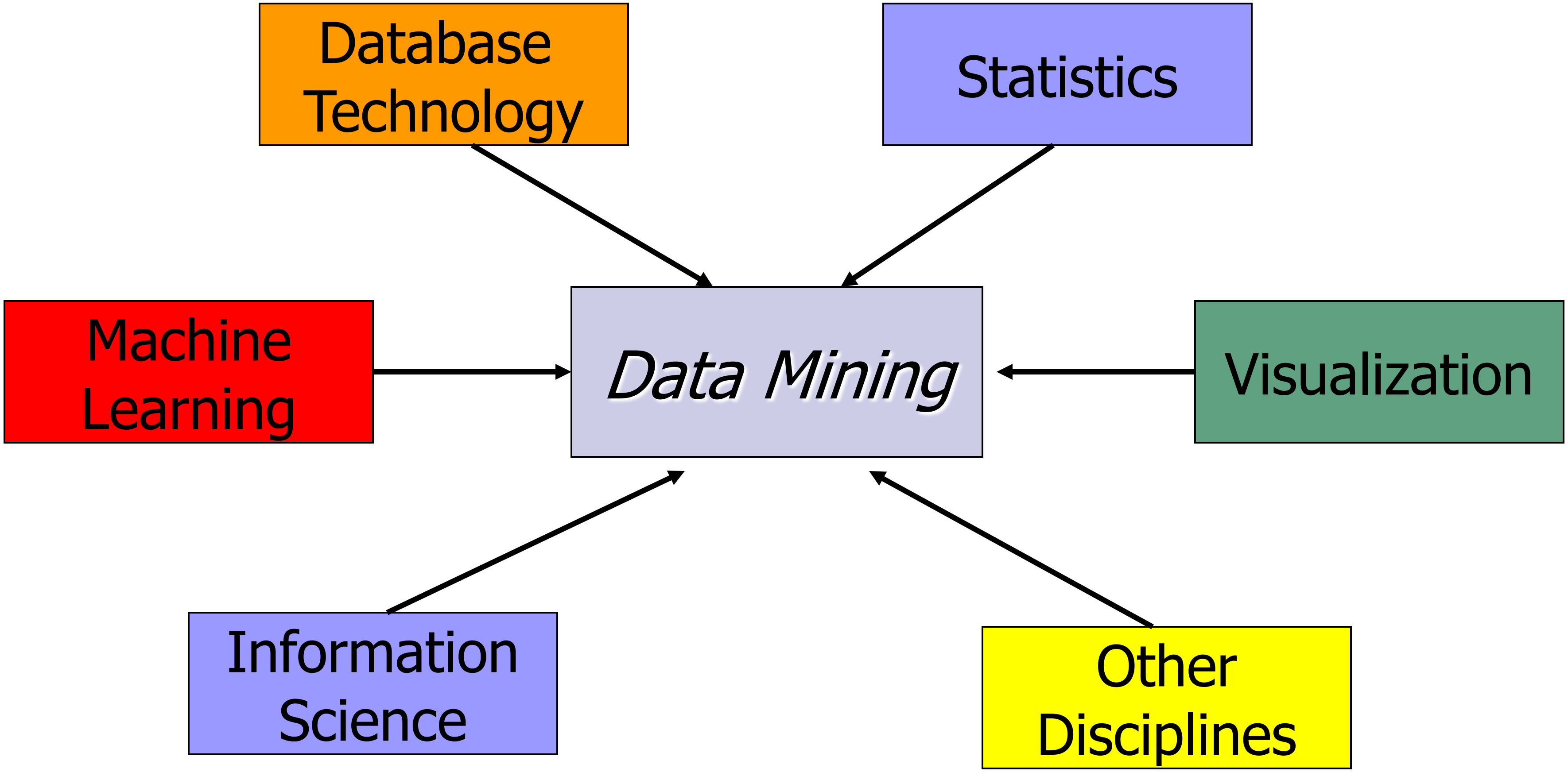
Web-based databases
(1990s–present)
- XML-based database systems
- Integration with information retrieval
- Data and information integration

New Generation of Integrated Data and Information Systems
(present–future)

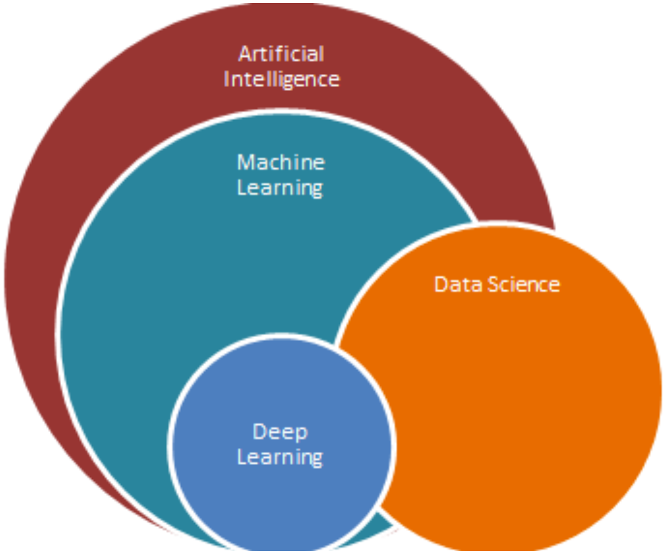
What is Data Mining?

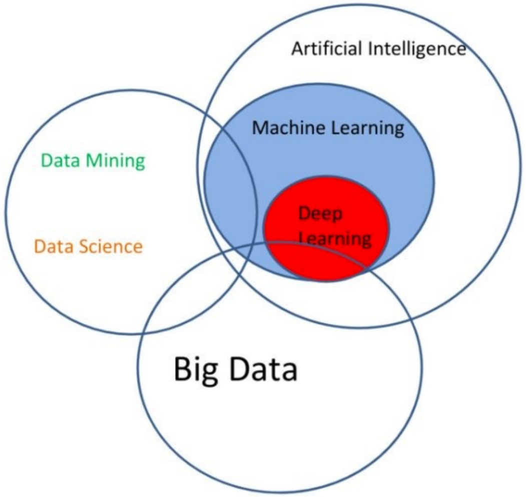
- Data mining (knowledge discovery in databases):
  - □ "the nontrivial extraction of implicit, previously unknown and potentially useful information from data in large databases"
    - 60% of the customers buy diapers also buy beer
- Alternative names:
  - □ Knowledge discovery in databases (KDD), intelligent data/pattern analysis, data archeology, information harvesting, business intelligence, etc.
  - □ Now, big data or data analytics
- What is not data mining?
  - □ (Deductive) query processing.
  - □ Expert systems or small ML/statistical programs

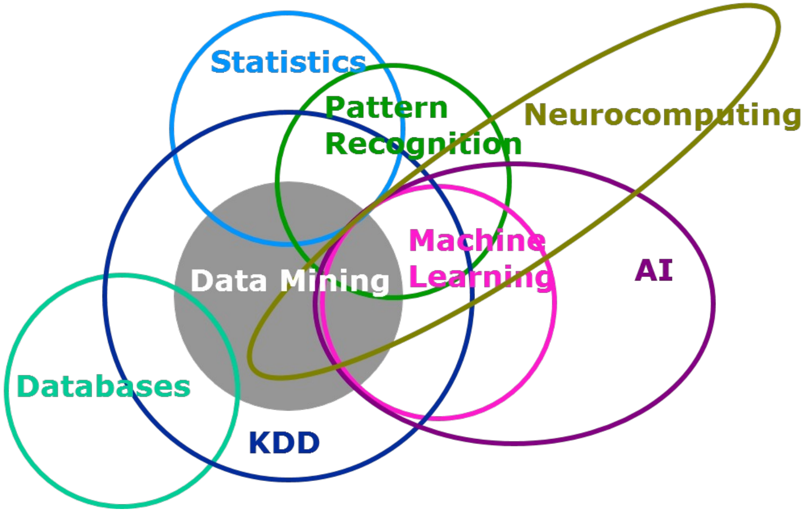
Data Mining:
Confluence of Multiple Disciplines



Where is Data Mining?







Data Scientist vs Machine Learning Engineer

Source: Kdnuggets' Blog on "The Difference Between Data Scientists and ML Engineers"

Responsibilities:

- Data scientists follow data science process which consists of
  - □ **Stage 1:** Understanding the Business Problem
  - □ **Stage 2:** Data Collection
  - □ **Stage 3:** Data Cleaning & Exploration
  - □ **Stage 4:** Model Building
  - □ **Stage 5:** Communicate and Visualize Insights
- Machine Learning Engineers are responsible for creating and maintaining the Machine Learning infrastructure that permits them to deploy the models built by Data Scientists to a production environment.
- Note that we do have ML developers/scientists who work more on Stage 4, i.e. model building!

Salary:

- Generally, ML Engineer is higher (western context)!

Data Scientist vs Machine Learning Engineer

Source: Kdnuggets' Blog on "The Difference Between Data Scientists and ML Engineers"

Expertise:

- Both are expected to have good knowledge of
  - Supervised & Unsupervised Learning
  - Machine Learning & Predictive Modelling
  - Mathematics and Statistics
  - Python (or R)
- Data Scientists are typically extremely good data storytellers. They can just use tools PowerBI and Tableau to share insights to the business.
- Machine Learning engineer is expected to have a strong foundation in computer science and software engineering.
- Yes, their expertise could be overlapped.

Data Mining Tasks

- *Association (correlation and causality)*
  - the most well-known one or the most unique one
  - shows attribute-value conditions that occur frequently together in a given set of data
  - age(X, “20..29”) ^ income(X, “20..29K”)
    ⇒ buys(X, “PC”) [support = 2%, confidence = 60%]
  - contains(transaction, “computer”) ⇒ contains(transaction, “software”) [support = 1%, confidence = 75%]

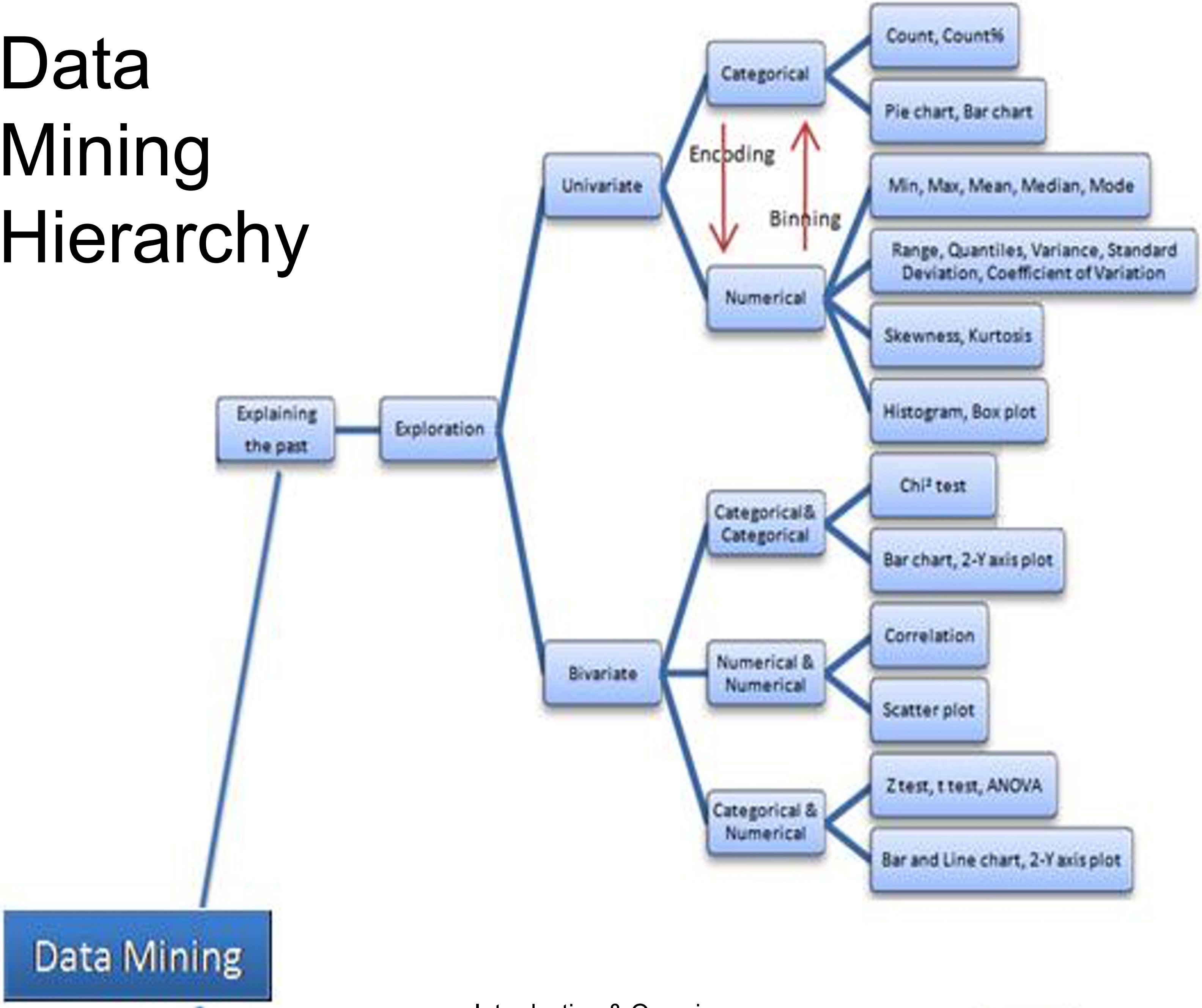
Data Mining Tasks

- *Classification and Prediction*
  - Finding models that describe and distinguish classes or concepts for future prediction
  - E.g., classify countries based on climate, or classify cars based on gas mileage, classify students based on their academic strength
  - Frequently used models: decision-tree, classification rule, neural networks, support vector machine (SVM)
  - Prediction: Predict some unknown or missing numerical values; e.g. Predict the Hang Seng Index (HSI), Stock Price, Power consumption level, Weather
- *Cluster Analysis*
  - Class label is unknown: Group data to form new classes, e.g., cluster houses to find distribution patterns, categorize web pages to define topics
  - Clustering is typically based on the principle of "maximizing the intra-class similarity and minimizing the interclass similarity"

Data Mining Tasks

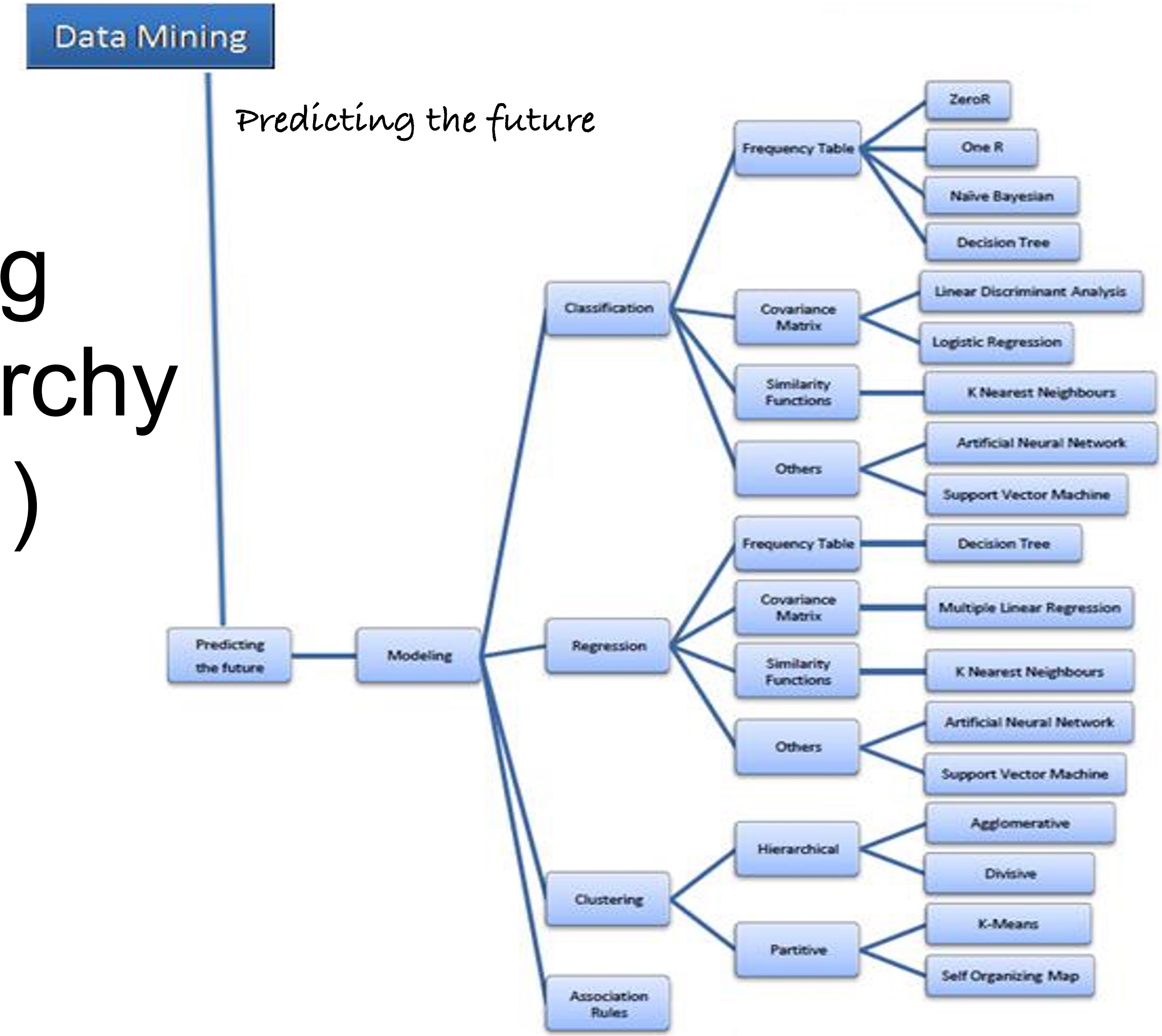
- *Outlier Analysis*
  - Outlier: a data object that does not comply with the general behavior of the data (e.g. a computer hacker vs multiple ordinary users)
  - It can be considered as noise or exception but is quite useful in fraud detection, rare events analysis, network intrusion detection
- *Trend and Sequence Analysis*
  - Trend and deviation: regression analysis
  - Sequential pattern mining, periodicity analysis
  - Similarity-based analysis
- *Other pattern-directed or statistical analysis*

Data Mining Hierarchy



Data Mining Hierarchy (cont.)

Explaining the past



Predicting the future

Potential Applications of DM

Many, many, many…

Whenever you have data, it can be applied!

Prominent one:

- Market analysis and management
  - target marketing, market basket analysis, market segmentation

Application Examples: Market Analysis and Management

- Where are the data sources for analysis?
  - Credit card transactions, loyalty cards, discount coupons, customer complaint calls, plus (public) lifestyle studies
- Target marketing
  - Find clusters of “model” customers who share the same characteristics: interest, income level, spending habits, etc.
- Determine customer purchasing patterns over time
  - Conversion of single to a joint bank account: marriage, etc.
- Cross-market analysis
  - Association/correlation between product sales
  - Prediction based on the association information

Application Examples: Market Analysis and Management

- Customer profiling
  - data mining can tell you what types of customers buy what products (clustering or classification)
- Identifying customer requirements
  - identifying the best products for different customers
  - use prediction to find what factors will attract new customers
- Provides summary information
  - various multidimensional summary reports
  - statistical summary information (data central tendency and variation)
  - mainly through data warehousing

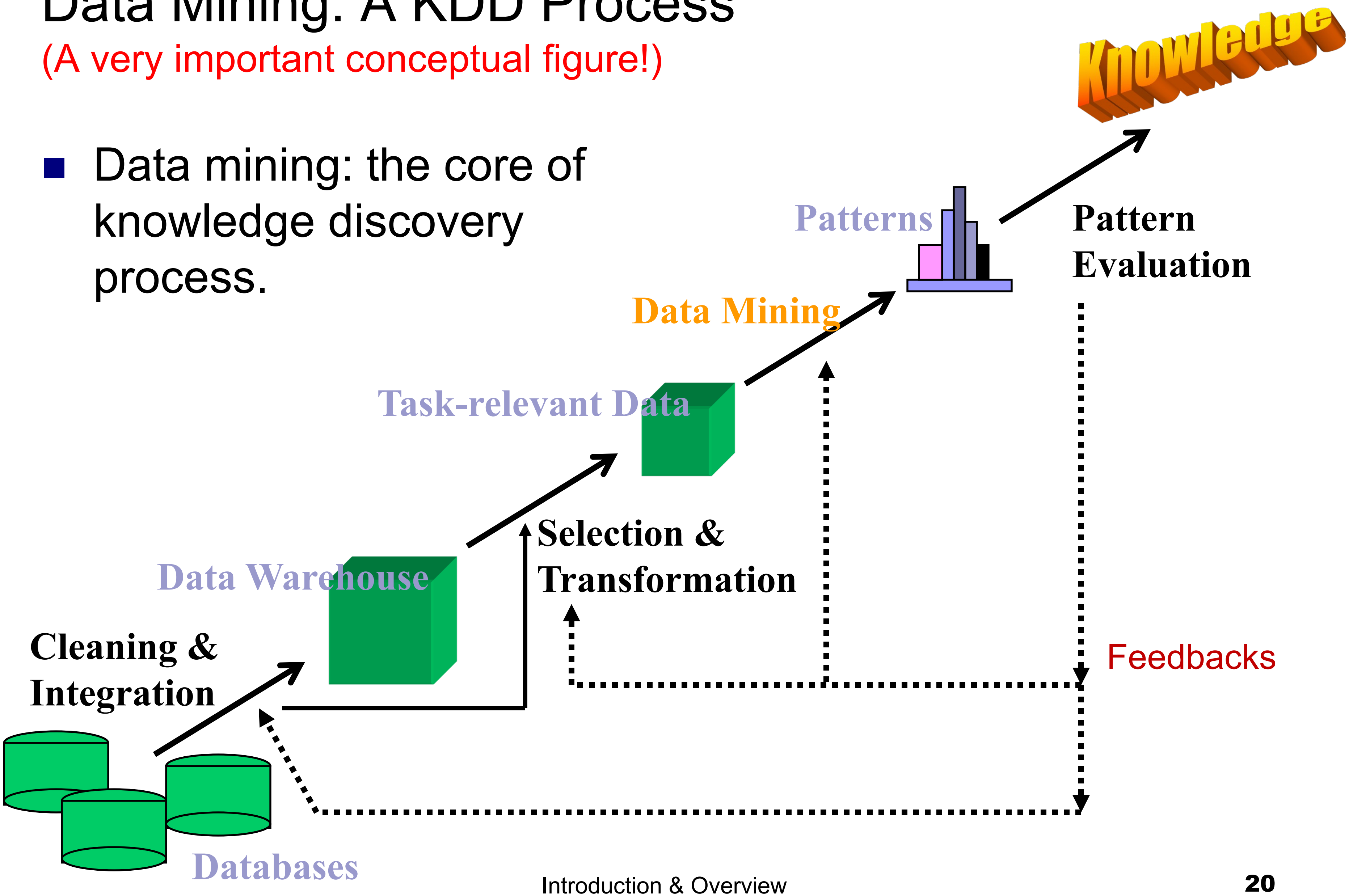
Other Data Type Based Applications

- Web Mining
  - applies mining algorithms to Web access logs for discovering customer preference and behavior, analyzing effectiveness of Web marketing, improving Web site organization, etc.
  - e-CRM
  - Web Analytics (Google Analytics)
- Text mining (email, documents)
  - SPAM Filtering, Opinion mining, (Microsoft) email decluttering
  - Social network analysis
- Spatial-temporal data mining, Time series data mining, Multimedia data mining, Stream data mining

Data Mining: A KDD Process

(A very important conceptual figure!)

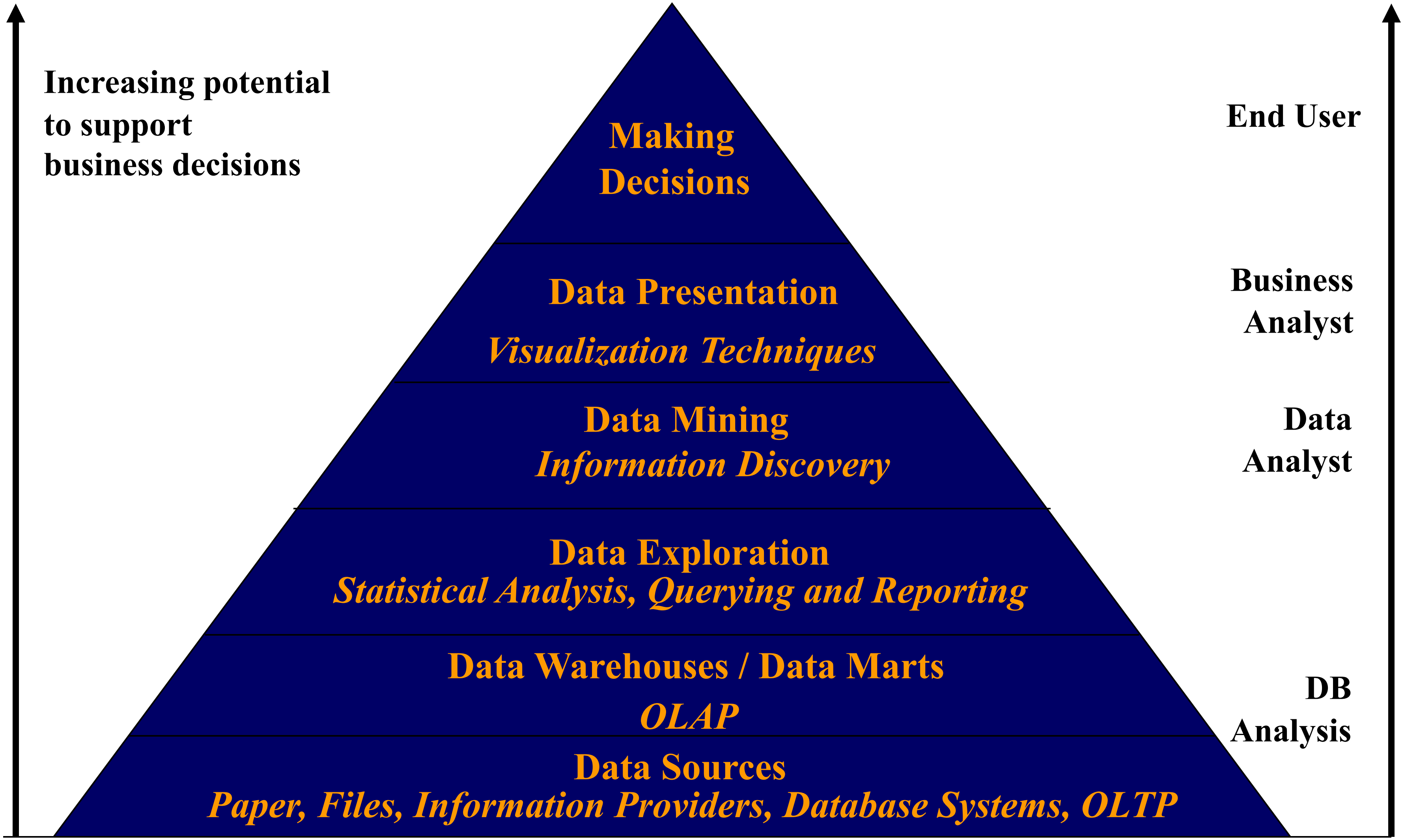
- Data mining: the core of knowledge discovery process.



Steps of a KDD Process

- Learning the application domain!!
  - relevant prior knowledge and goals of application
- 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, classification, regression, association, 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

Data Mining (DM) & Business Intelligence (BI)


Increasing potential to support business decisions
Making Decisions
End User
Data Presentation
Visualization Techniques
Business Analyst
Data Mining
Information Discovery
Data Analyst
Data Exploration
Statistical Analysis, Querying and Reporting
Data Warehouses / Data Marts
OLAP
DB Analysis
Data Sources
Paper, Files, Information Providers, Database Systems, OLTP

How to mine data?

On what kind of data

- Relational databases
- Transactional databases
- Data warehouses
- Advanced DB and information repositories
  - Object-oriented and object-relational databases
  - Spatial databases
  - Time-series data and temporal data
  - Text databases and multimedia databases
  - Heterogeneous and legacy databases
  - Web database
  - Bioinformatics database
  - Stream database

How to mine data?

Classification of Data Mining Systems

- **Databases to be mined**
  - Relational, transactional, object-oriented, object-relational, active, spatial, time-series, text, multi-media, heterogeneous, legacy, WWW, bioinformatics, stream, etc.
- **Knowledge to be mined**
  - Characterization, discrimination, association, classification, clustering, trend, deviation and outlier analysis, etc.
  - Multiple/integrated functions and mining at multiple levels
- **Techniques utilized**
  - Database-oriented, data warehouse (OLAP), machine learning, statistics, visualization, neural network, GA, fuzzy rules, etc.
- **Applications adapted**
  - Retail, telecommunication, banking, fraud analysis, DNA mining, stock market analysis, Web mining, Weblog analysis, etc.

How to mine data?
Major issues in Data Mining

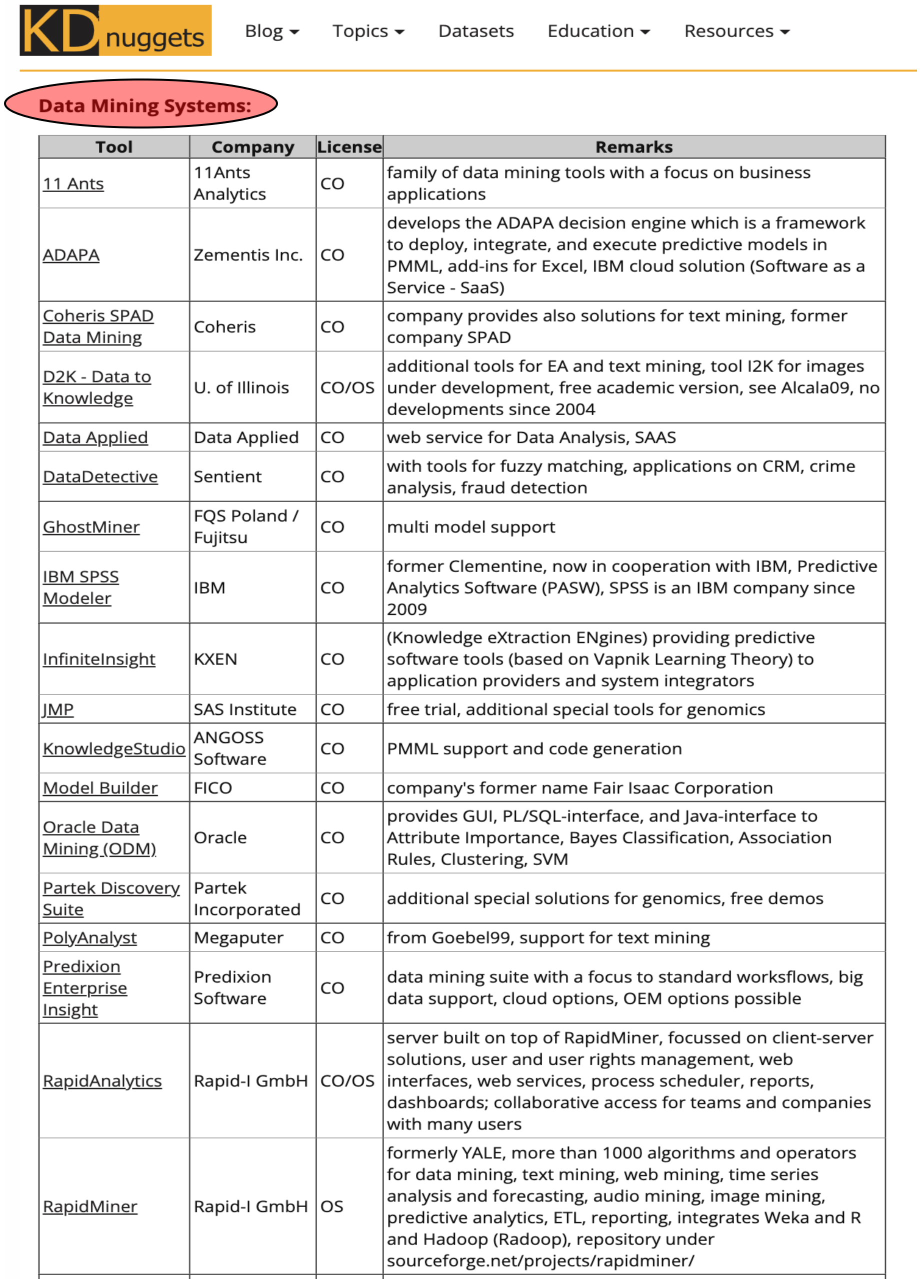
- Mining methodology and user interaction
  - Mining different kinds of knowledge in databases
  - Interactive mining of knowledge at multiple levels of abstraction
  - Incorporation of background knowledge
  - Data mining query languages and ad-hoc data mining
  - Expression and visualization of data mining results
  - Handling noise and incomplete (missing) data
  - Pattern evaluation: the interestingness problem
- Performance and scalability
  - Efficiency and scalability of data mining algorithms
  - Parallel, distributed and incremental mining methods

How to mine data?
Major issues in Data Mining

- Issues relating to the diversity of data types
    - Handling relational and complex types of data
    - Mining information from heterogeneous databases and global information systems (Web)
    - How about social network graphs and GPS data?
- Issues related to applications and social impacts
    - Application of discovered knowledge
        - Domain-specific data mining tools
        - Intelligent query answering
        - Process control and decision making
    - Integration of the discovered knowledge with existing knowledge: A knowledge fusion problem
    - Protection of data security, integrity, and privacy

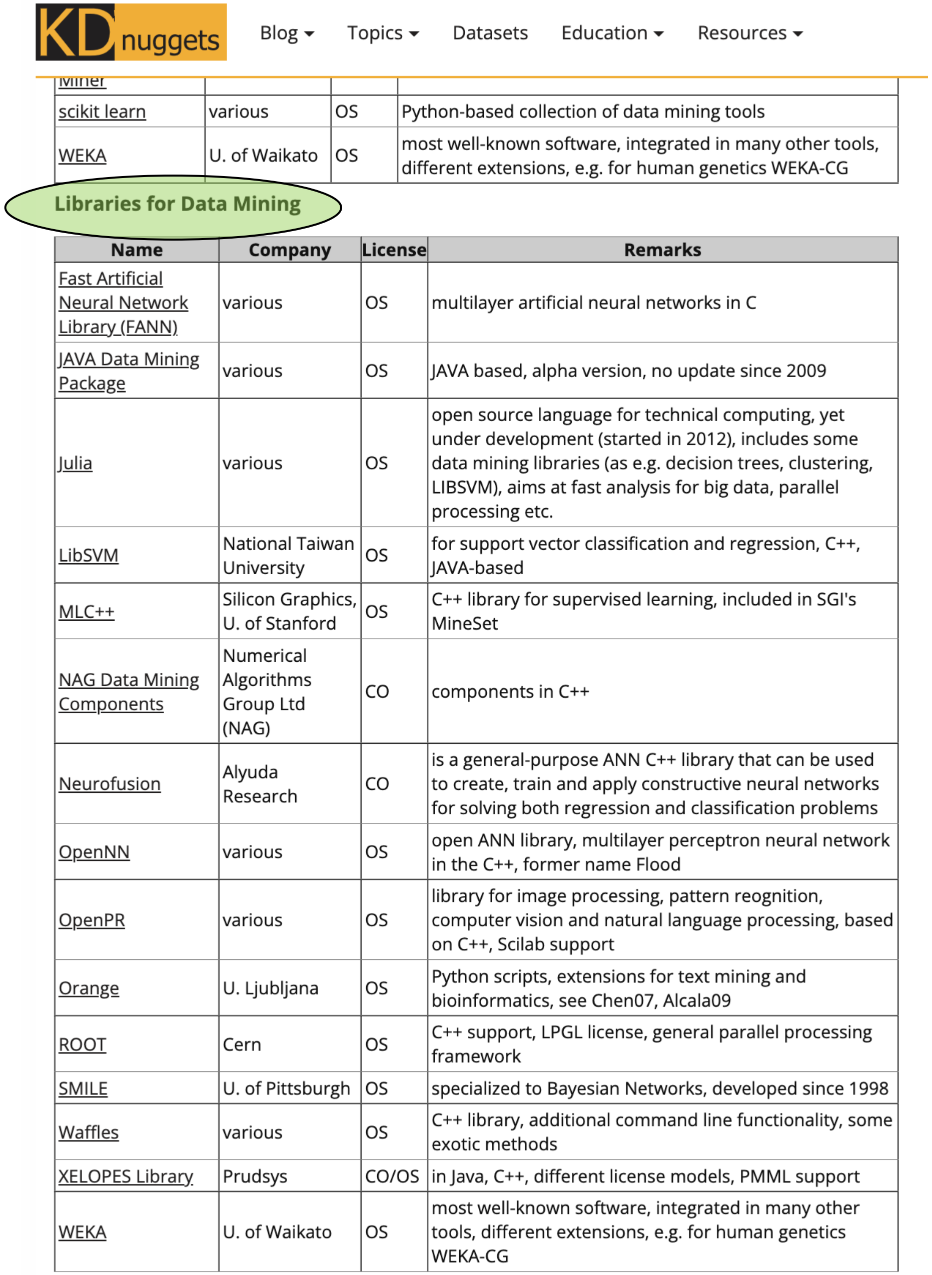
        ⇒ Privacy preserving data mining

Data Mining Tools from KDnuggets



Data Mining Systems:

Tool	Company	License	Remarks
11 Ants	11Ants Analytics	CO	family of data mining tools with a focus on business applications
ADAPA	Zementis Inc.	CO	develops the ADAPA decision engine which is a framework to deploy, integrate, and execute predictive models in PMML, add-ins for Excel, IBM cloud solution (Software as a Service - SaaS)
Coheris SPAD Data Mining	Coheris	CO	company provides also solutions for text mining, former company SPAD
D2K - Data to Knowledge	U. of Illinois	CO/OS	additional tools for EA and text mining, tool I2K for images under development, free academic version, see Alcala09, no developments since 2004
Data Applied	Data Applied	CO	web service for Data Analysis, SAAS
DataDetective	Sentient	CO	with tools for fuzzy matching, applications on CRM, crime analysis, fraud detection
GhostMiner	FQS Poland / Fujitsu	CO	multi model support
IBM SPSS Modeler	IBM	CO	former Clementine, now in cooperation with IBM, Predictive Analytics Software (PASW), SPSS is an IBM company since 2009
InfiniteInsight	KXEN	CO	(Knowledge eXtraction ENgines) providing predictive software tools (based on Vapnik Learning Theory) to application providers and system integrators
JMP	SAS Institute	CO	free trial, additional special tools for genomics
KnowledgeStudio	ANGOSS Software	CO	PMML support and code generation
Model Builder	FICO	CO	company's former name Fair Isaac Corporation
Oracle Data Mining (ODM)	Oracle	CO	provides GUI, PL/SQL-interface, and Java-interface to Attribute Importance, Bayes Classification, Association Rules, Clustering, SVM
Partek Discovery Suite	Partek Incorporated	CO	additional special solutions for genomics, free demos
PolyAnalyst	Megaputer	CO	from Goebel99, support for text mining
Predixion Enterprise Insight	Predixion Software	CO	data mining suite with a focus to standard worksflows, big data support, cloud options, OEM options possible
RapidAnalytics	Rapid-I GmbH	CO/OS	server built on top of RapidMiner, focussed on client-server solutions, user and user rights management, web interfaces, web services, process scheduler, reports, dashboards; collaborative access for teams and companies with many users
RapidMiner	Rapid-I GmbH	OS	formerly YALE, more than 1000 algorithms and operators for data mining, text mining, web mining, time series analysis and forecasting, audio mining, image mining, predictive analytics, ETL, reporting, integrates Weka and R and Hadoop (Radoop), repository under sourceforge.net/projects/rapidminer/



Miner			
scikit learn	various	OS	Python-based collection of data mining tools
WEKA	U. of Waikato	OS	most well-known software, integrated in many other tools, different extensions, e.g. for human genetics WEKA-CG

Libraries for Data Mining

Name	Company	License	Remarks
Fast Artificial Neural Network Library (FANN)	various	OS	multilayer artificial neural networks in C
JAVA Data Mining Package	various	OS	JAVA based, alpha version, no update since 2009
Julia	various	OS	open source language for technical computing, yet under development (started in 2012), includes some data mining libraries (as e.g. decision trees, clustering, LIBSVM), aims at fast analysis for big data, parallel processing etc.
LibSVM	National Taiwan University	OS	for support vector classification and regression, C++, JAVA-based
MLC++	Silicon Graphics, U. of Stanford	OS	C++ library for supervised learning, included in SGI's MineSet
NAG Data Mining Components	Numerical Algorithms Group Ltd (NAG)	CO	components in C++
Neurofusion	Alyuda Research	CO	is a general-purpose ANN C++ library that can be used to create, train and apply constructive neural networks for solving both regression and classification problems
OpenNN	various	OS	open ANN library, multilayer perceptron neural network in the C++, former name Flood
OpenPR	various	OS	library for image processing, pattern reognition, computer vision and natural language processing, based on C++, Scilab support
Orange	U. Ljubljana	OS	Python scripts, extensions for text mining and bioinformatics, see Chen07, Alcala09
ROOT	Cern	OS	C++ support, LPGL license, general parallel processing framework
SMILE	U. of Pittsburgh	OS	specialized to Bayesian Networks, developed since 1998
Waffles	various	OS	C++ library, additional command line functionality, some exotic methods
XELOPES Library	Prudsys	CO/OS	in Java, C++, different license models, PMML support
WEKA	U. of Waikato	OS	most well-known software, integrated in many other tools, different extensions, e.g. for human genetics WEKA-CG

Summary

- Data mining: discovering interesting patterns from large amounts 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 information repositories
- Data mining tasks: characterization, discrimination, association, classification, clustering, outlier and trend analysis, etc.
- Classification of data mining systems
- Major issues in data mining