Data Science

Data Mining Introduction

Themis Palpanas
University of Paris

Thanks for slides to:

- Jiawei Han
- Jeff Ullman

Roadmap

- Motivation: Why data mining?
- What is data mining?
- Data Mining: On what kind of data?
- Data mining functionality
- Are all the patterns interesting?
- Classification of data mining systems
- Data Mining Task Primitives
- Integration of data mining system with a DB and DW System
- Major issues in data mining

Why Massive Data Analytics?

- 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,
- We are drowning in data, but starving for knowledge!
- Data mining: Automated analysis of massive data sets

Why Massive Data Analytics?

- examples of data sizes
  - telecommunications industry (AT&T)
    - 7GB/day call detail data
    - 15GB/day IP network monitoring data
  - web sites
    - 10TB/day click data for Yahoo!
  - retailers
    - 20 million sales transactions/day for WalMart
  - scientific projects
    - 1.2TB/day for Earth Observing System (NASA)
    - 100PB/year for European Organization for Nuclear Research (CERN)

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

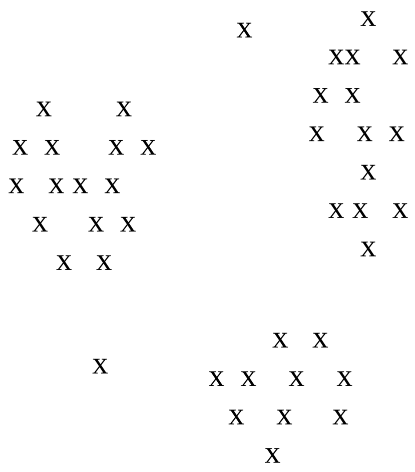
What Is Data Mining?

- 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

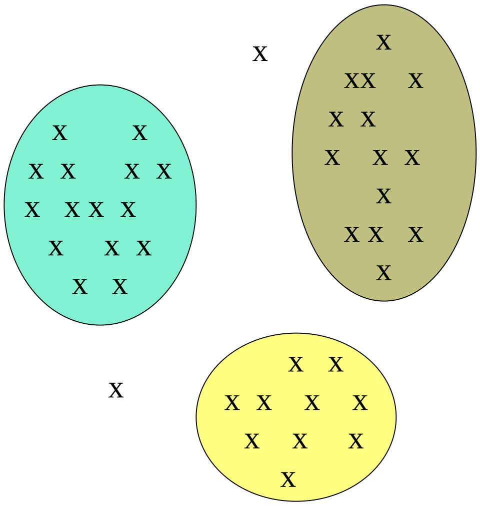
Typical Kinds of Patterns

1. Decision trees: succinct ways to classify by testing properties.
2. Clusters: another succinct classification by similarity of properties.
3. Bayes models, hidden-Markov models, frequent-itemsets: expose important associations within data.

Example: Clusters



Example: Clusters



Example: Frequent Itemsets

- A common marketing problem: examine what people buy together to discover patterns.
  1. What pairs of items are unusually often found together at Safeway checkout?
     - Answer: diapers and beer.
  2. What books are likely to be bought by the same Amazon customer?

Why Data Mining?—Potential Applications

- Data analysis and decision support
  - Market analysis and management
    - Target marketing, customer relationship management (CRM), market basket analysis, cross selling, market segmentation
  - Risk analysis and management
    - Forecasting, customer retention, improved underwriting, quality control, competitive analysis
  - Fraud detection and detection of unusual patterns (outliers)
- Other Applications
  - Text mining (news group, email, documents) and Web mining
  - Stream data mining
  - Bioinformatics and bio-data analysis

Ex. 1: Market Analysis and Management

- Where does the data come from?—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
- Cross-market analysis—Find associations/co-relations between product sales, & predict based on such association
- Customer profiling—What types of customers buy what products (clustering or classification)
- Customer requirement analysis
  - Identify the best products for different customers
  - Predict what factors will attract new customers
- Provision of summary information
  - Multidimensional summary reports
  - Statistical summary information (data central tendency and variation)

Ex. 2: Corporate Analysis & Risk Management

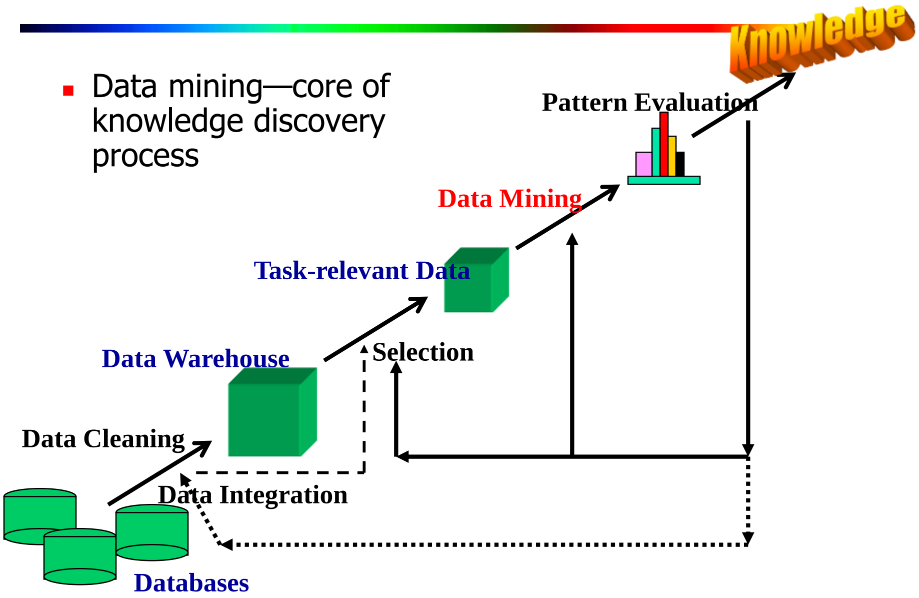
- Finance planning and asset evaluation
  - cash flow analysis and prediction
  - contingent claim analysis to evaluate assets
  - cross-sectional and time series analysis (financial-ratio, trend analysis, etc.)
- Resource planning
  - summarize and compare the resources and spending
- Competition
  - monitor competitors and market directions
  - group customers into classes and a class-based pricing procedure
  - set pricing strategy in a highly competitive market

Ex. 3: Fraud Detection & Mining Unusual Patterns

- Approaches: Clustering & model construction for frauds, outlier analysis
- Applications: Health care, retail, credit card service, telecomm.
  - Auto insurance: ring of collisions
  - Money laundering: suspicious monetary transactions
  - Medical insurance
    - Professional patients, ring of doctors, and ring of references
    - Unnecessary or correlated screening tests
  - Telecommunications: phone-call fraud
    - Phone call model: destination of the call, duration, time of day or week. Analyze patterns that deviate from an expected norm
  - Retail industry
    - Analysts estimate that 38% of retail shrink is due to dishonest employees
  - Anti-terrorism

Knowledge Discovery (KDD) Process

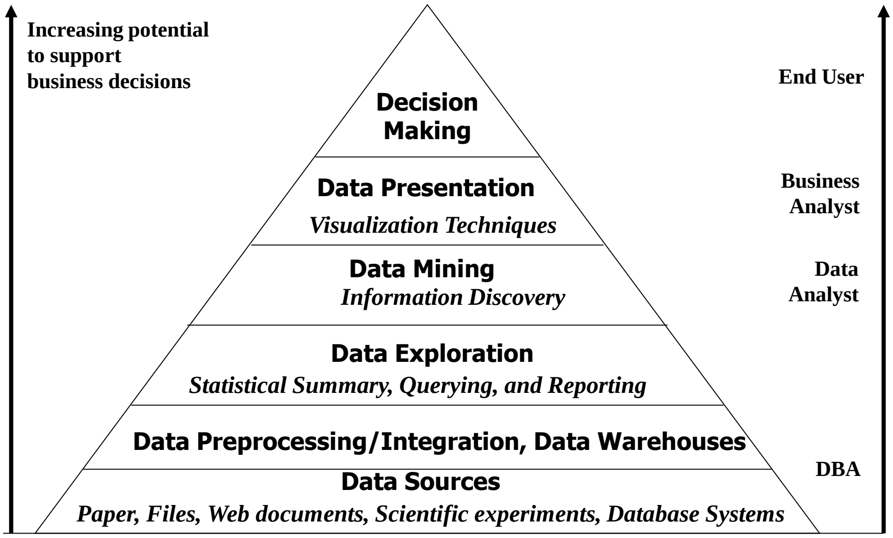
- Data mining—core of knowledge discovery process



KDD Process: Several Key Steps

- 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 and 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

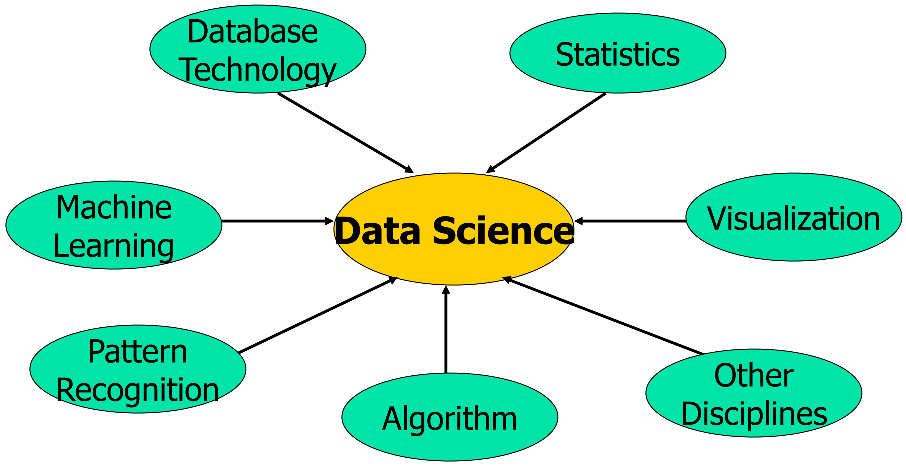
DBA

Data Science

- **Data science** is an inter-disciplinary field that
- uses scientific methods, processes, algorithms and systems to
- extract knowledge and insights from
- structured and unstructured data

source: Wikipedia

Data Science: Confluence of Multiple Disciplines



Database Technology

Statistics

Machine Learning

Data Science

Visualization

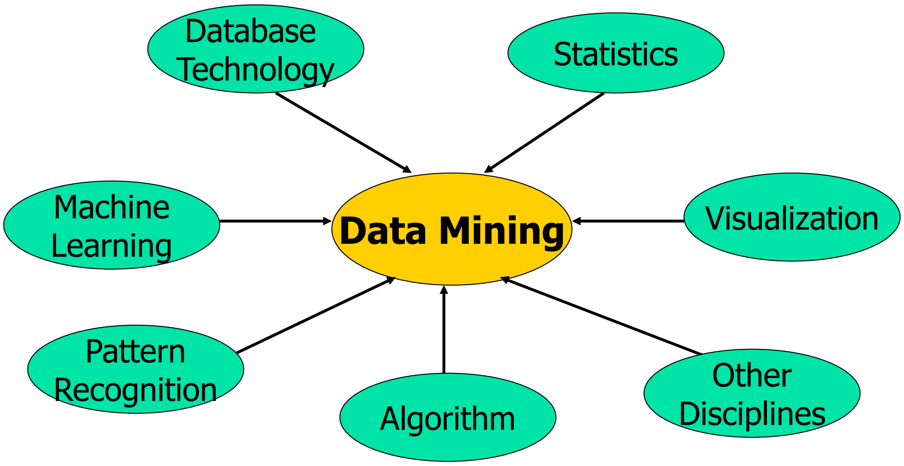
Pattern Recognition

Algorithm

Other Disciplines

year: 2020

Data Mining: Confluence of Multiple Disciplines



year: 2000

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

Cultures

- Databases: concentrate on large-scale (non-main-memory) data.
- AI (machine-learning): concentrate on complex methods, small data.
- Statistics: concentrate on models.

Models vs. Analytic Processing

- To a database person, data-mining is an extreme form of analytic processing --- queries that examine large amounts of data.
  - Result is the data that answers the query.
- To a statistician, data-mining is the inference of models.
  - Result is the parameters of the model.

Why Not Traditional Data Analysis?

- 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

Multi-Dimensional View of Data Mining

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

Data Mining: Classification Schemes

- General functionality
  - Descriptive data mining
  - Predictive data mining
- Different views lead to different classifications
  - Data view: Kinds of data to be mined
  - Knowledge view: Kinds of knowledge to be discovered
  - Method view: Kinds of techniques utilized
  - Application view: Kinds of applications adapted

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 Functionalities

- Multidimensional concept description: Characterization and discrimination
  - Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet regions
- Frequent patterns, association, correlation vs. causality
  - Diaper → Beer [0.5%, 75%] (Correlation or causality?)
- Classification and prediction
  - Construct models (functions) that 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 or missing numerical values

Data Mining Functionalities (2)

- Cluster analysis
  - Class label is unknown: Group data to form new classes, e.g., cluster houses to find distribution patterns
  - Maximizing intra-class similarity & minimizing interclass similarity
- Outlier analysis
  - Outlier: Data object that does not comply with the general behavior of the data
  - Noise or exception? Useful in fraud detection, rare events analysis
- Trend and evolution analysis
  - Trend and deviation: e.g., regression analysis
  - Sequential pattern mining: e.g., digital camera → large SD memory
  - Periodicity analysis
  - Similarity-based analysis
- Other pattern-directed or statistical analyses

Are All the "Discovered" Patterns Interesting?

- Data mining may generate thousands of patterns: Not all of them are interesting
  - Suggested approach: Human-centered, query-based, focused mining
- **Interestingness measures**
  - A pattern is interesting if it is easily understood by humans, valid on new or test data with some degree of certainty, potentially useful, novel, or validates some hypothesis that a user seeks to confirm
- **Objective vs. subjective interestingness measures**
  - Objective: based on statistics and structures of patterns, e.g., support, confidence, etc.
  - Subjective: based on user's belief in the data, e.g., unexpectedness, novelty, actionability, etc.

Find All and Only Interesting Patterns?

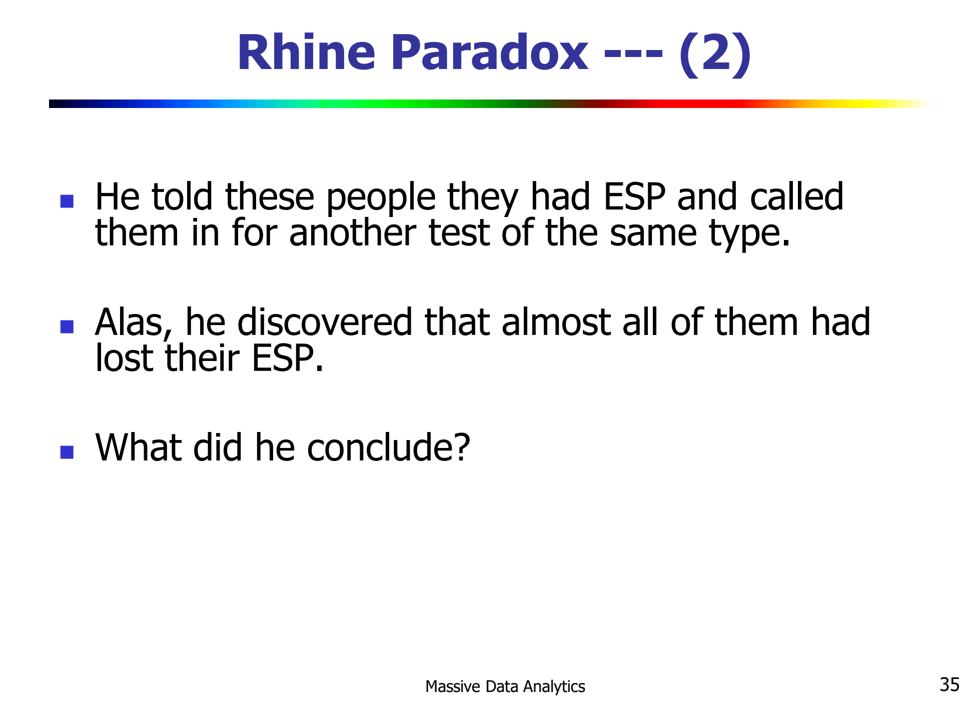
- Find all the interesting patterns: Completeness
  - Can a data mining system find all the interesting patterns? Do we need to find all of the interesting patterns?
  - Heuristic vs. exhaustive search
  - Association vs. classification vs. clustering
- Search for only interesting patterns: An optimization problem
  - Can a data mining system find only the interesting patterns?
  - Approaches
    - First general all the patterns and then filter out the uninteresting ones
    - Generate only the interesting patterns—mining query optimization

Meaningfulness of Answers

- A big risk when data mining is that you will "discover" patterns that are meaningless.
- Statisticians call it Bonferroni's principle: (roughly) if you look in more places for interesting patterns than your amount of data will support, you are bound to find crap.

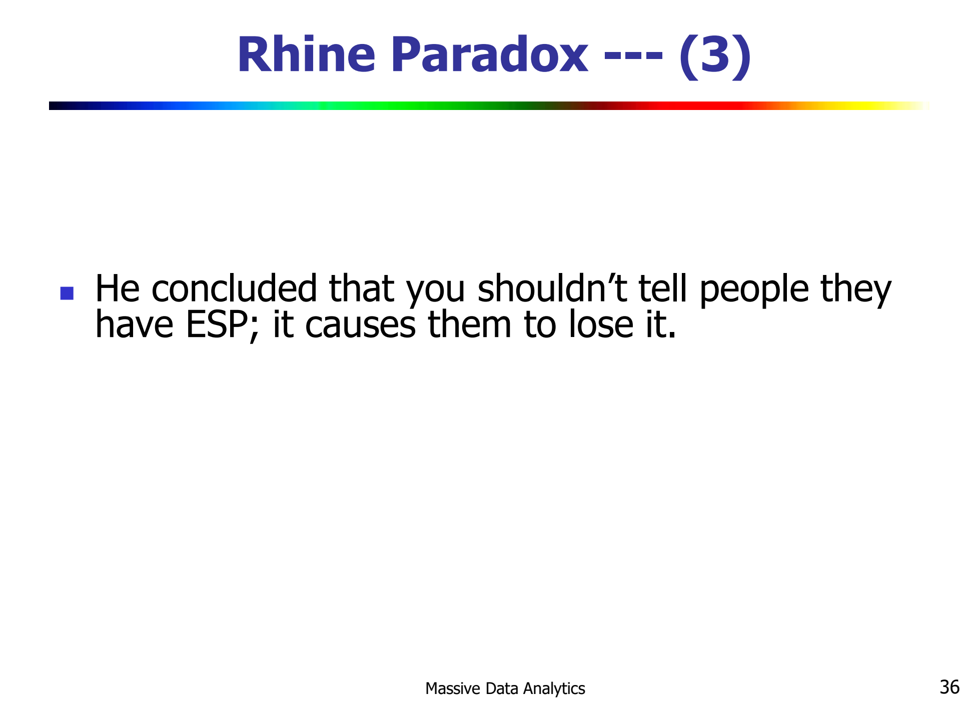
Rhine Paradox --- (1)

- Joseph Rhine was a parapsychologist in the 1950's who hypothesized that some people had Extra-Sensory Perception.
- He devised an experiment where subjects were asked to guess 10 hidden cards --- red or blue.
- He discovered that almost 1 in 1000 had ESP --- they were able to get all 10 right!



Rhine Paradox --- (2)

- He told these people they had ESP and called them in for another test of the same type.
- Alas, he discovered that almost all of them had lost their ESP.
- What did he conclude?



Rhine Paradox --- (3)

- He concluded that you shouldn't tell people they have ESP; it causes them to lose it.

Rhine Paradox --- (4)

- The Rhine Paradox: a great example of how not to conduct scientific research.
- When looking for a property, make sure that there are not so many possibilities that random data will produce facts “of interest.”

Other Pattern Mining Issues

- Precise patterns vs. approximate patterns
  - Association and correlation mining: possible find sets of precise patterns
    - But approximate patterns can be more compact and sufficient
    - How to find high quality approximate patterns??
  - Gene sequence mining: approximate patterns are inherent
    - How to derive efficient approximate pattern mining algorithms??
- Constrained vs. non-constrained patterns
  - Why constraint-based mining?
  - What are the possible kinds of constraints? How to push constraints into the mining process?

Why Data Mining Query Language?

- Automated vs. query-driven?
  - Finding all the patterns autonomously in a database?—unrealistic because the patterns could be too many but uninteresting
- Data mining should be an interactive process
  - User directs what to be mined
- Users must be provided with a set of primitives to be used to communicate with the data mining system
- Incorporating these primitives in a data mining query language
  - More flexible user interaction
  - Foundation for design of graphical user interface
  - Standardization of data mining industry and practice

Primitives that Define a Data Mining Task

- Task-relevant data
- Type of knowledge to be mined
- Background knowledge
- Pattern interestingness measurements
- Visualization/presentation of discovered patterns

Primitive 1: Task-Relevant Data

- Database or data warehouse name
- Database tables or data warehouse cubes
- Condition for data selection
- Relevant attributes or dimensions
- Data grouping criteria

Primitive 2: Types of Knowledge to Be Mined

- Characterization
- Discrimination
- Association
- Classification/prediction
- Clustering
- Outlier analysis
- Other data mining tasks

Primitive 3: Background Knowledge

- A typical kind of background knowledge: Concept hierarchies
- Schema hierarchy
  - E.g., street < city < province_or_state < country
- Set-grouping hierarchy
  - E.g., {20-39} = young, {40-59} = middle_aged
- Operation-derived hierarchy
  - email address: hagonzal@cs.uiuc.edu

    login-name < department < university < country
- Rule-based hierarchy
  - low_profit_margin (X) <= price(X, P_1) and cost (X, P_2) and (P_1 - P_2) < $50

Primitive 4: Pattern Interestingness Measure

- Simplicity

  e.g., (association) rule length, (decision) tree size
- Certainty

  e.g., confidence, P(A|B) = #(A and B)/ #(B), classification reliability or accuracy, certainty factor, rule strength, rule quality, discriminating weight, etc.
- Utility

  potential usefulness, e.g., support (association), noise threshold (description)
- Novelty

  not previously known, surprising (used to remove redundant rules)

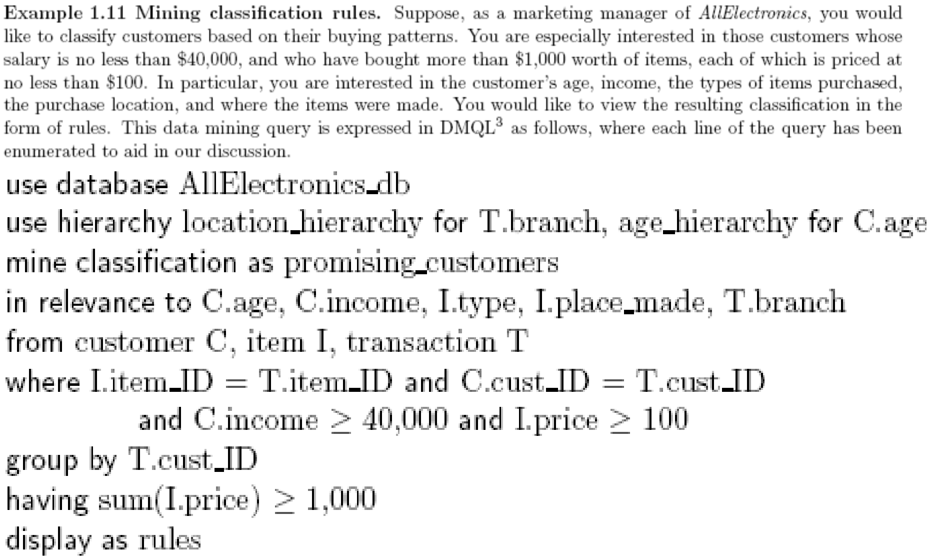
Primitive 5: Presentation of Discovered Patterns

- Different backgrounds/usages may require different forms of representation
  - E.g., rules, tables, crosstabs, pie/bar chart, etc.
- Concept hierarchy is also important
  - Discovered knowledge might be more understandable when represented at high level of abstraction
  - Interactive drill up/down, pivoting, slicing and dicing provide different perspectives to data
- Different kinds of knowledge require different representation: association, classification, clustering, etc.

DMQL—A Data Mining Query Language

- Motivation
  - A DMQL can provide the ability to support ad-hoc and interactive data mining
  - By providing a standardized language like SQL
    - Hope to achieve a similar effect like that SQL has on relational database
    - Foundation for system development and evolution
    - Facilitate information exchange, technology transfer, commercialization and wide acceptance
- Design
  - DMQL is designed with the primitives described earlier

An Example Query in DMQL



Example 1.11 Mining classification rules. Suppose, as a marketing manager of *AllElectronics*, you would like to classify customers based on their buying patterns. You are especially interested in those customers whose salary is no less than \$40,000, and who have bought more than \$1,000 worth of items, each of which is priced at no less than \$100. In particular, you are interested in the customer's age, income, the types of items purchased, the purchase location, and where the items were made. You would like to view the resulting classification in the form of rules. This data mining query is expressed in DMQL^3 as follows, where each line of the query has been enumerated to aid in our discussion.

```
use database AllElectronics_db
use hierarchy location_hierarchy for T.branch, age_hierarchy for C.age
mine classification as promising_customers
in relevance to C.age, C.income, I.type, I.place_made, T.branch
from customer C, item I, transaction T
where I.item_ID = T.item_ID and C.cust_ID = T.cust_ID
          and C.income ≥ 40,000 and I.price ≥ 100
group by T.cust_ID
having sum(I.price) ≥ 1,000
display as rules
```

Other Data Mining Languages & Standardization Efforts

- Association rule language specifications
  - MSQL (Imielinski & Virmani'99)
  - MineRule (Meo Psaila and Ceri'96)
  - Query flocks based on Datalog syntax (Tsur et al'98)
- OLEDB for DM (Microsoft'2000) and DMX (Microsoft SQLServer 2005)
  - Based on OLE, OLE DB, OLE DB for OLAP, C#
  - Integrating DBMS, data warehouse and data mining
- DMML (Data Mining Mark-up Language) by DMG (www.dmg.org)
  - Providing a platform and process structure for effective data mining
  - Emphasizing on deploying data mining technology to solve business problems

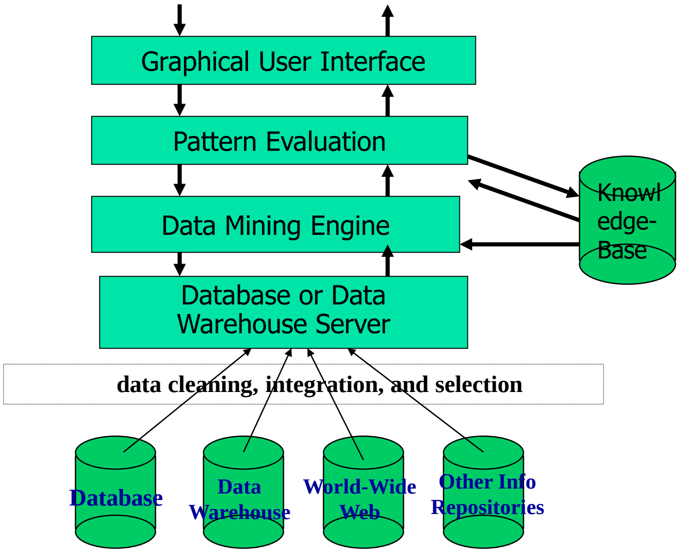
Integration of Data Mining and Data Warehousing

- **Data mining systems, DBMS, Data warehouse systems coupling**
  - No coupling, loose-coupling, semi-tight-coupling, tight-coupling
- **On-line analytical mining data**
  - integration of mining and OLAP technologies
- **Interactive mining multi-level knowledge**
  - Necessity of mining knowledge and patterns at different levels of abstraction by drilling/rolling, pivoting, slicing/dicing, etc.
- **Integration of multiple mining functions**
  - Characterized classification, first clustering and then association

Coupling Data Mining with DB/DW Systems

- No coupling—flat file processing, not recommended
- Loose coupling
  - Fetching data from DB/DW
- Semi-tight coupling—enhanced DM performance
  - Provide efficient implementations of a few data mining primitives in a DB/DW system, e.g., sorting, indexing, aggregation, histogram analysis, multiway join, precomputation of some stat functions
- Tight coupling—A uniform information processing environment
  - DM is smoothly integrated into a DB/DW system, mining query is optimized based on mining query, indexing, query processing methods, etc.

Architecture: Typical Data Mining System



Graphical User Interface

Pattern Evaluation

Data Mining Engine

Database or Data Warehouse Server

Knowl edge-Base

data cleaning, integration, and selection

Database

Data Warehouse

World-Wide Web

Other Info Repositories

Major Issues in Data Mining

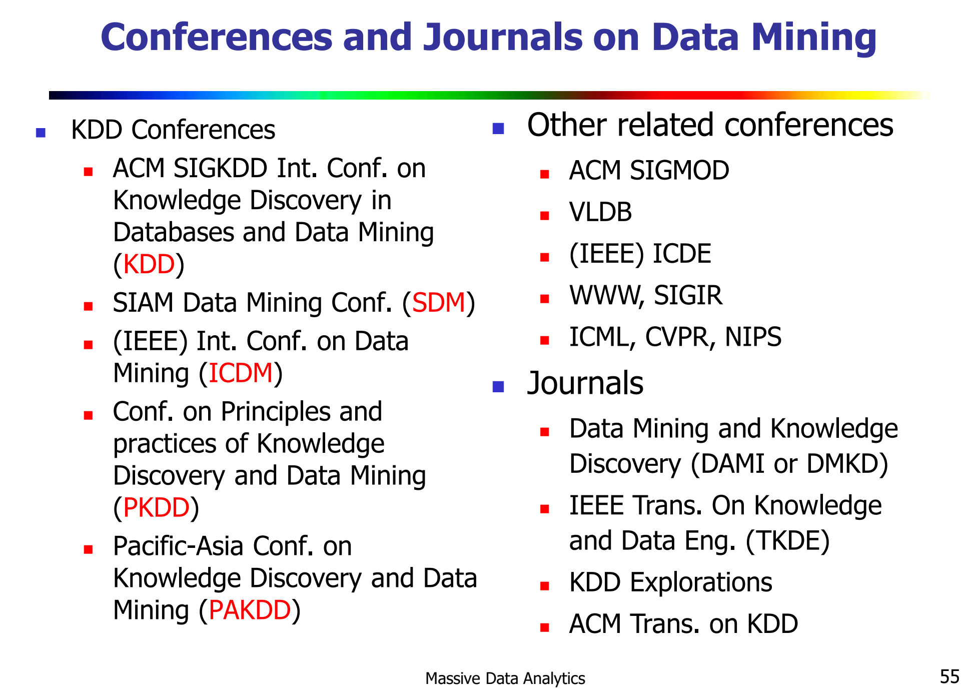
- Mining methodology
  - Mining different kinds of knowledge from diverse data types, e.g., bio, stream, Web
  - Performance: efficiency, effectiveness, and scalability
  - Pattern evaluation: the interestingness problem
  - Incorporation of background knowledge
  - Handling noise and incomplete data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

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 functionalities: characterization, discrimination, association, classification, clustering, outlier and trend analysis, etc.
- Data mining systems and architectures
- Major issues in data mining

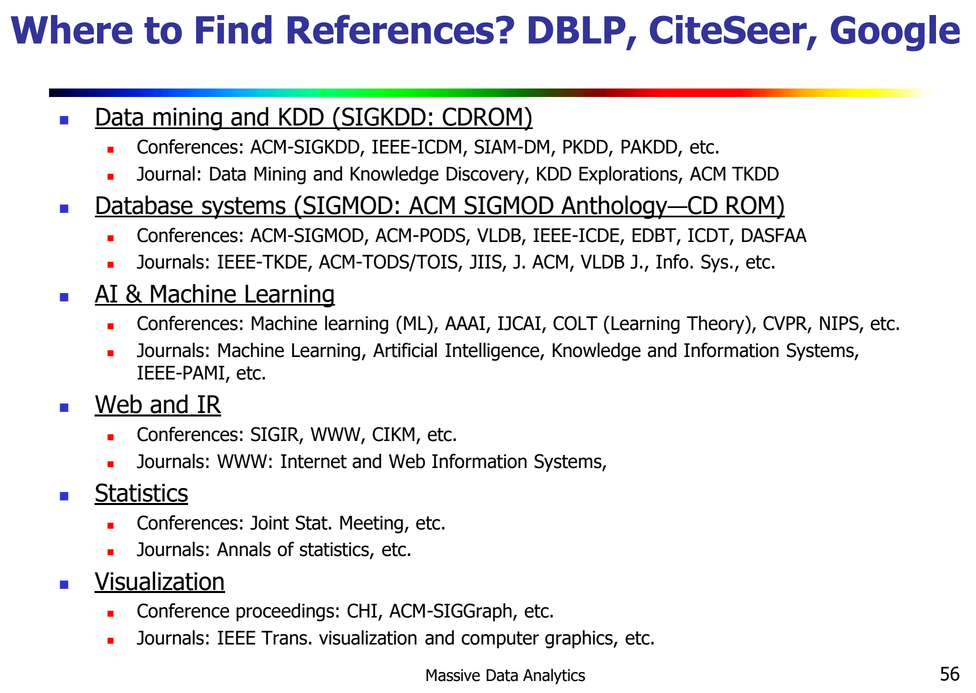
A Brief History of Data Mining Society

- 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)
  - Conf. on Principles and practices of Knowledge Discovery and Data Mining (PKDD)
  - Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD)
- Other related conferences
  - ACM SIGMOD
  - VLDB
  - (IEEE) ICDE
  - WWW, SIGIR
  - ICML, CVPR, NIPS
- 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.