What is Data Mining?

- Data explosion problem
  - We are drowning in data, but starving for knowledge!

- Solution: Data warehousing and data mining
  - Data warehousing and on-line analytical processing (OLAP)
  - Mining interesting knowledge (rules, regularities, patterns, constraints) from data in large databases
Evolution of Database Technology

- **1960s:**
  - Data collection, database creation, IMS and network DBMS

- **1970s:**
  - **Relational data model,** relational DBMS implementation
  - Tedd Codd (1923-2003)
  - Structured English Query Language (SEQUEL), **SQL**

- **1980s:**
  - Advanced data models (extended-relational, OO, deductive, etc.)
  - Application-oriented DBMS (spatial, scientific, engineering, etc.)

- **1990s:**
  - Data mining, data warehousing, multimedia databases
  - **Web databases** (...Amazon)

- **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 \(\text{non-trivial, implicit, previously unknown}\)
    and potentially useful) patterns or knowledge from huge amount
    of data (\textit{interesting patterns}?)
  - Data mining: a misnomer? (\textit{erro de nome})

- 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”?
  - (Deductive) query processing.
  - Expert systems or small ML/statistical programs

---

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
  - Medical data mining
  - Bioinformatics and bio-data analysis
Example 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

Market Analysis and Management

- Cross-market analysis—Find associations/co-reations 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
Example 2: Corporate Analysis & Risk Management

- Finance planning and asset evaluation
  - cash flow analysis and prediction (feature development)
  - contingent claim analysis to evaluate assets *(componente do ativo)*
  - cross-sectional and time series analysis (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

Example 3: Fraud Detection & Mining Unusual Patterns

- Approaches:
  - Unsupervised Learning: Clustering
  - Supervised Learning: Neuronal Networks

- 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 (vender a varejo)
  - Analysts estimate that 38% of retail shrink is due to dishonest employees
- Anti-terrorism

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
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

Data Mining and Knowledge Discovery (KDD) Process

- Data mining—core of knowledge discovery process
- Data Cleaning
- Data Integration
- Databases
- Data Warehouses
- Task-relevant Data
- Selection
- Data Mart
- Pattern Evaluation

Knowledge
Steps of a KDD Process (1)

- 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!)
- Understand data (statistics)
- Data reduction and transformation
  - Find useful features, dimensionality/variable reduction, invariant representation

Steps of a KDD Process (2)

- 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
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
Data Mining Functionalities (1)

- Multidimensional concept description: Characterization and discrimination
  - Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet regions
- Frequent patterns, association, correlation and causality
  - Smoking → Cancer (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?
- Trend and evolution analysis
  - Trend and deviation: e.g., regression analysis
  - Sequential pattern mining, periodicity analysis
  - Similarity-based analysis
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.

Can We Find All and Only Interesting Patterns?

- **Find all the interesting patterns**: Completeness
  - Can a data mining system find all 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
Data Mining: Confluence of Multiple Disciplines

- Database Technology
- Statistics
- Machine Learning
- Algorithm
- Visualization
- Other Disciplines

Data Mining: Classification Schemes

- General functionality
  - Descriptive data mining
  - Predictive data mining
- Different views lead to different classifications
  - Kinds of data to be mined
  - Kinds of knowledge to be discovered
  - Kinds of techniques utilized
  - Kinds of applications adapted
Data Mining from different perspectives

- **Data to be mined**
  - Object-oriented/relational, 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, machine learning, statistics, visualization, etc.

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

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 (Categories)
- Discrimination
- Association
- Classification/prediction
- Clustering
- Outlier analysis
- Other data mining tasks
Primitive 3:
Background Knowledge

- Schema hierarchy (taxonomy)
  - 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:
Measurements of Pattern Interestingness

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

Architecture: Typical Data Mining System
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 Online Analytical Processing (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 implement 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, etc.
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
  - (constraints, taxonomy)
- Handling noise and incomplete data (preprocessing)
- 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