Introducción a la Minería de Datos

Han & Kamber (2007). Revisado J. M. Matías 2008

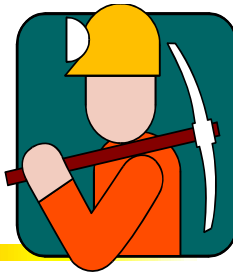
Indice

- Qué es la Minería de Datos?
- Motivación: por qué la Minería de Datos?
- Aplicaciones
- Las Fases del Proceso de Construcción de Conocimiento
- Arquitectura de un Sistema de Minería de Datos
- Minería de Datos: Formas de Conocimiento y Técnicas de Minería de Datos
- Classification of data mining systems
- Major issues in data mining

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

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

Why Data Mining?

- The Explosive Growth of Data: from terabytes to petabytes
  - Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized society
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, ...
    - Science: Remote sensing, bioinformatics, scientific simulation, ...
    - Society and everyone: news, digital cameras, YouTube
- We are drowning in data, but starving for knowledge!
- "Necessity is the mother of invention"—Data mining—Automated analysis of massive data sets

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, 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

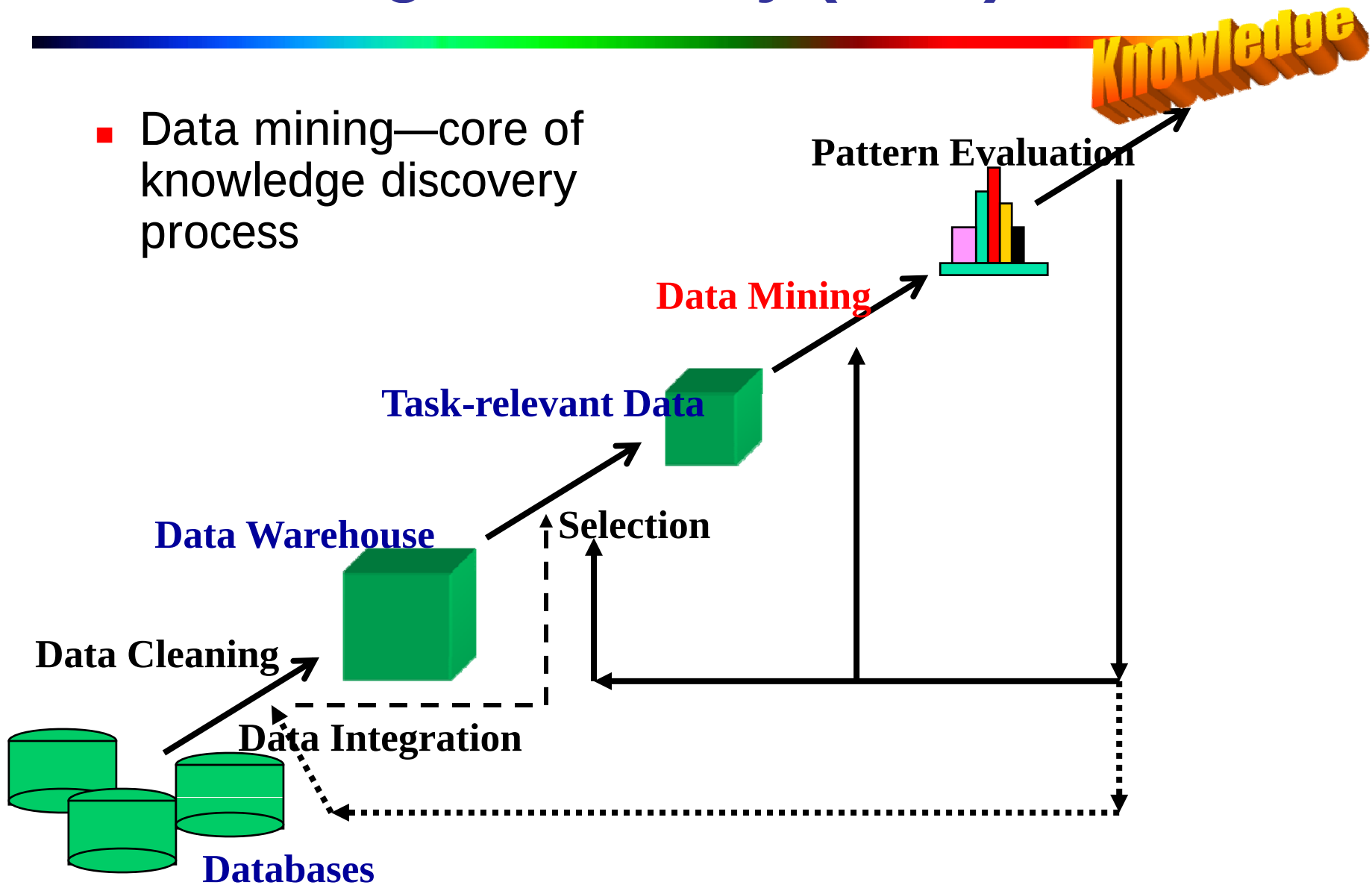
- Where does the data come from?—Credit card transactions, loyalty cards, discount coupons, customer complaint calls, plus (public) lifestyle studies
- Target marketing
  - Find clusters of “model” customers who share the same characteristics: interest, income level, spending habits, etc.
  - Determine 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 groups of customers
  - Predict what factors will attract new customers
- Provision of summary information
  - Multidimensional summary reports
  - Statistical summary information (data central tendency and variation)

KDD Process: Several Key Steps

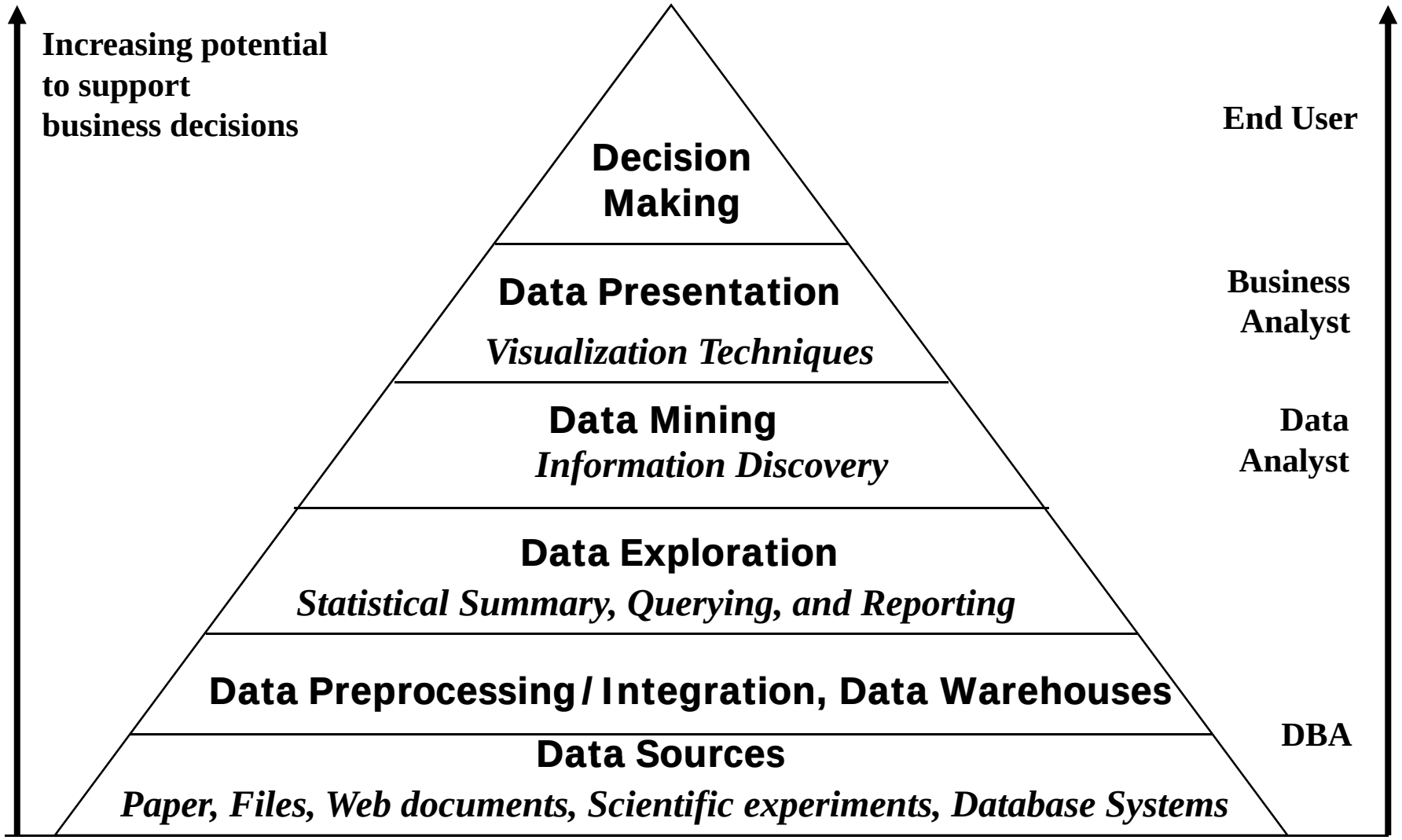
- Learning the application domain
  - relevant prior knowledge and goals of application
- Creating a target data set: data selection
- Data cleaning and preprocessing: (may take 60% of effort!)
- Data reduction and transformation
  - Find useful features, dimensionality/variable reduction, invariant 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

Knowledge Discovery (KDD) Process

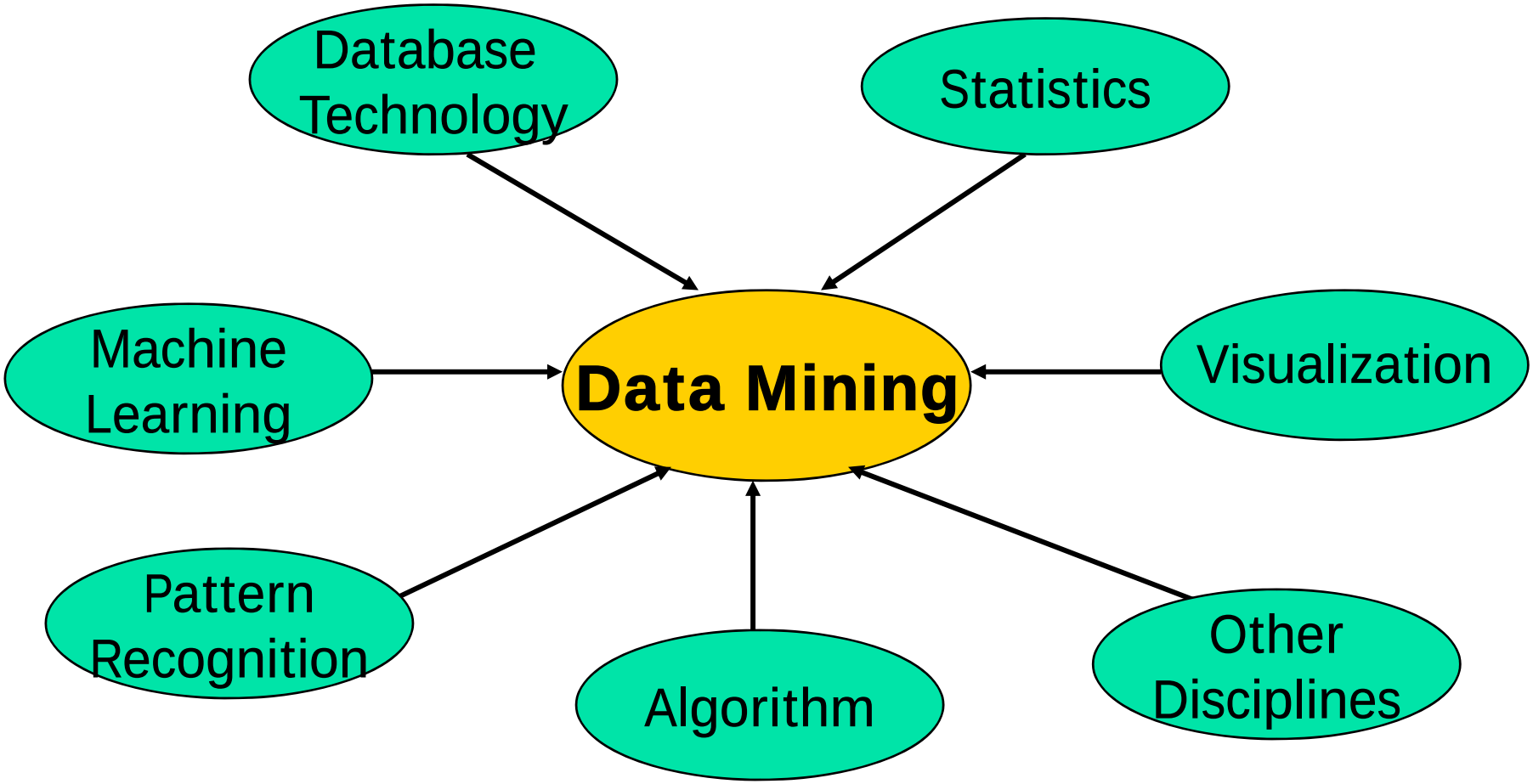
- Data mining—core of knowledge discovery process



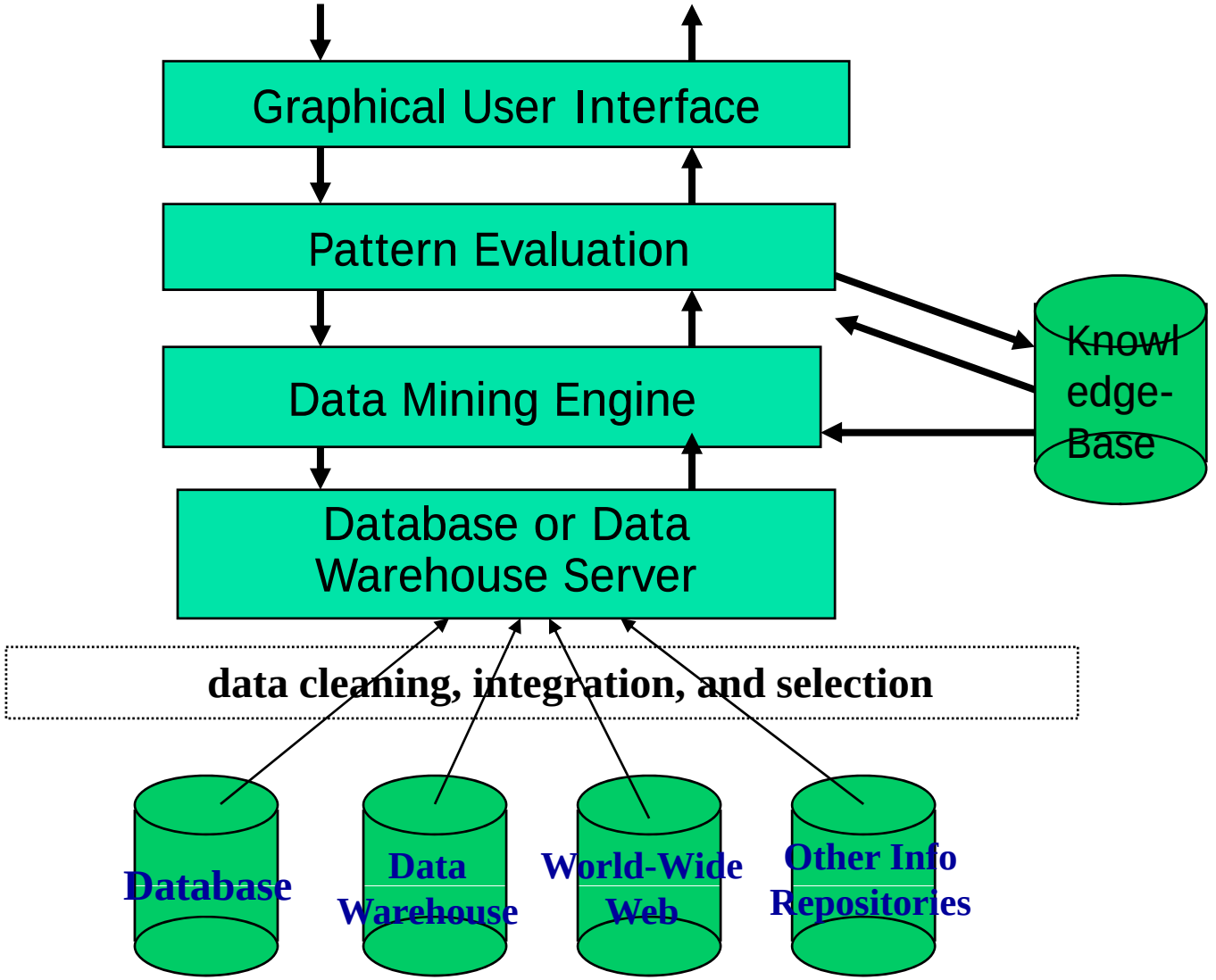
Data Mining and Business Intelligence



Data Mining: Confluence of Multiple Disciplines



Architecture: Typical Data Mining System



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

Formas de Conocimiento y Técnicas de Minería de Datos

Formas de Conocimiento

- Toda la distribución de probabilidad (valores y probabilidades) o función de densidad.
- Partes de la distribución de probabilidad:
  - Una distribución condicionada completa.
  - Trozos de alguna distribución condicionada, p. ej. una proposición o regla del tipo:

    si origen=nacional, entonces tipo = sobremesa con soporte 30% y confianza 80%.
- Algunos parámetros de la distribución probabilidad: p. ej. la media condicionada (regresión), la varianza condicionada, la correlación entre dos variables.
- Los grupos relevantes de una población
  - Uno o varios conceptos basados en individuos (adjetivos calificativos): ej. cinéfilo.
- Las variables más relevantes de una población (las que contienen la mayor parte de la información)
  - Uno o varios conceptos quizás no observables basados en atributos: ej. la inteligencia.

Técnicas de Data Mining

- Supervisadas y no supervisadas.
- Estimación de la distribución de probabilidad:
  - Discretas: Redes Bayesianas.
  - Continuas: Técnicas no paramétricas.
- Obtención de reglas: algoritmo Apriori.
- De regresión:
  - Modelos lineales, modelos lineales generalizados, regresión polinómico local, redes neuronales, Support vector machines (regresión generalizada y splines), expansión en básicas (Fourier, wavelets), árboles de regresión.
- De Clasificación:
  - Naive bayes, modelos lineales generalizados (logit, probit), cualquier técnica de regresión que proporcione probabilidades a posteriori, Support vector machines, cluster para clasificación.
- De Reducción-Compresión de información:
  - De datos: clustering, self-organizing networks.
  - De variables: clustering, componentes principales, principal curves.

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