DATA MINING INTRODUCTION

PRESENTED BY

DR. NEHA SHARMA, INDIA

OBJECTIVE

To understand-

- Motivation: Why Data Mining
- What is Data Mining?
- What Kind of Data can be Mined?
- What kind of patterns can be mined?
- Which technologies are used?
- Which kinds of Applications are targeted?

BASICS…

- Data : Any facts, figures, numbers, or text that can be processed by a computer.
- Information: The patterns, associations or relationships among all this *data* can provide information.
- Knowledge: Information can be converted into *knowledge* about historical patterns and future trends.

BASICS- EVOLUTION OF SCIENCE…

- 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 information 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

BASICS- 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

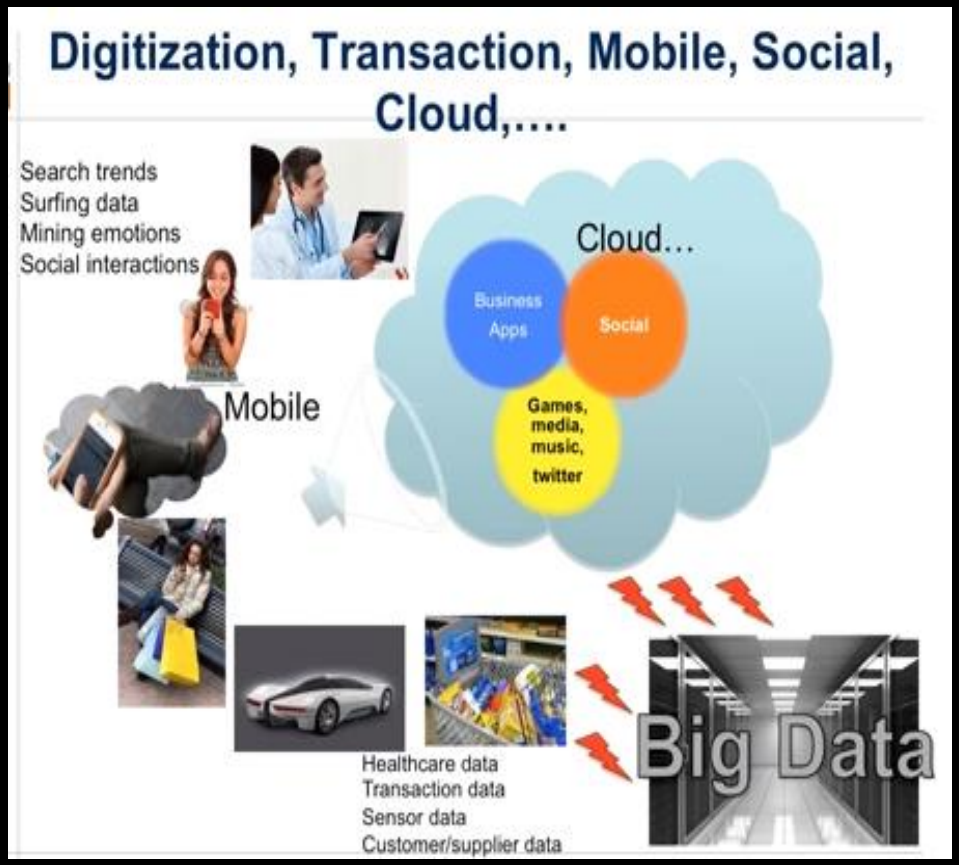
WHY DATA MINING ??

Trends leading to Data Flood…

WHY DATA MINING ??

Trends leading to Data Flood…

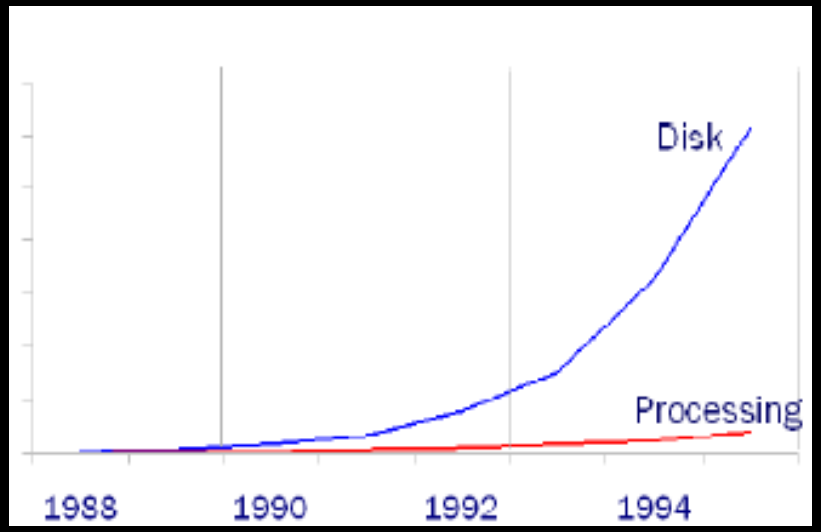
- The Explosive Growth of Data: from petabytes to zettabytes
  - More data is generated (Major sources of abundant data)
    - Business: Web, e-commerce, transactions, stocks, bank, telecom …
    - Science: Remote sensing, bioinformatics, scientific simulation, astronomy, biology….
    - Society and everyone: news, digital cameras, YouTube, social media…..
  - More data collection and data availability
    - Faster and cheaper storage technology
    - Automated data collection tools, database systems, Web, computerized society



WHY DATA MINING ??

Growth Trends…

- Moore's law
  - Computer Speed doubles every 18 months
- Storage law
  - Total storage doubles every 9 months
- Consequence -Human analysis skills are inadequate:
  - Volume and dimensionality of the data
  - High data growth rate
  - Very little data will ever be looked at by a human



Knowledge Discovery is **NEEDED** to make sense and use of data.

MOTIVATION FOR DATA MINING ??

- Availability of:
  - Data
  - Storage
  - Computational power
  - Off-the-shelf software
  - Expertise
- Evolving Data Science as search, mobile, social interactions, sensors, transactions, scientific computing etc., which contributed hugely to generate big data.
- We are drowning in data, but starving for knowledge!
- "Necessity is the mother of invention"—Data Mining—Automated analysis of massive data sets
- Hence, there is a significant demand for people who can manage, process, analyze and discover insights from massive data using quantitative and technical expertise to solve business, social and economic problems.

WHY SO MUCH OF INTEREST ???

- Cheap computing power
- Statistics
- Machine learning approaches
- Database technologies
- Search data
- Information retrieval methods
- Virtual communities
- Recommendation systems & online reviews
- Visualization
- Surfing data
- Social networks
- Twitter text data

Confluence

Scientific applications
Weather forecast
Genomics database

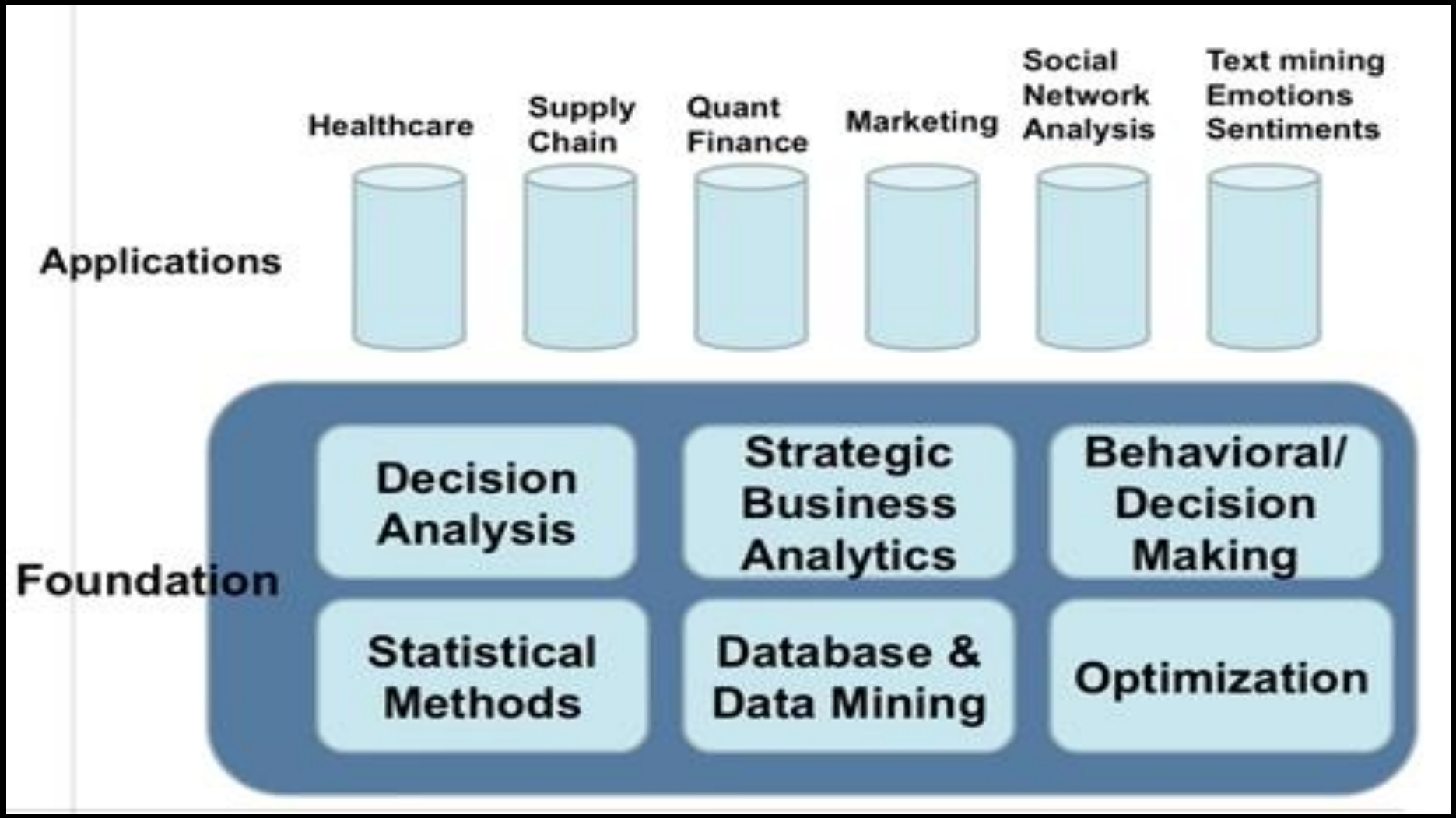
Macroeconomic analyses

Business applications

Social program effectiveness

Homeland security

IF YOU WANT TO BE AN EXPERT



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

WHAT IS DATA MINING ???

DATA MINING

- Data mining is process of extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
- Overall goal of data mining is to extract information from a data set and transform it into an understandable structure for further use.
- The term is a misnomer, because the goal is to extract the pattern and knowledge from large database, not the extraction of data itself.
- Data mining is automated or semi-automated process, in which the data is stored electronically and the analysis is either automated or semi-automated (augmented by computer), to extract previously unknown, interesting patterns such as group of data records (cluster analysis), unusual records (anomaly detection), and dependencies (association rule mining)

DATA MINING: Alternative Names

- 1960 – Data Fishing or Data Dredging – Statisticians
- 1989- Knowledge discovery in databases (KDD) – coined by Gregory Piatetsky-Shapiro and popular amongst AI and Machine Learning Community.
- 1990 – Data Mining – Database and Research community. Other terms used include Data Archaeology, Information Harvesting, Information Discovery, Knowledge Extraction, etc
- Term Data Mining became popular in business community.
- Google Search - 2000000 pages for Data Mining / 300000 pages for KDD
- Currently, Data Mining and Knowledge Discovery are used interchangeably

DATA MINING

- Data Mining != Data Analysis
- Data analysis is inline with standard statistical software, which usually present information about subsets and relations within the recorded data set. Example – browser/search engine usage, average visit time
- Data mining software uses some intelligence implies over simple grouping and partitioning to infer new information

Not Data Mining	Data Mining
Simple Search and Query Processing	Online retailer is interested in knowing consumers' Market Basket to better advertise, display items for cross-selling and upselling, discount items and target customers.
Deductive – Expert System	
Search for phone number in directory	A bank wants to know which customer to target for credit cards and loans
Web search for information about "Amazon"	An insurance company is looking for fraud pattern in health care

DATA MINING & KDD PROCESS

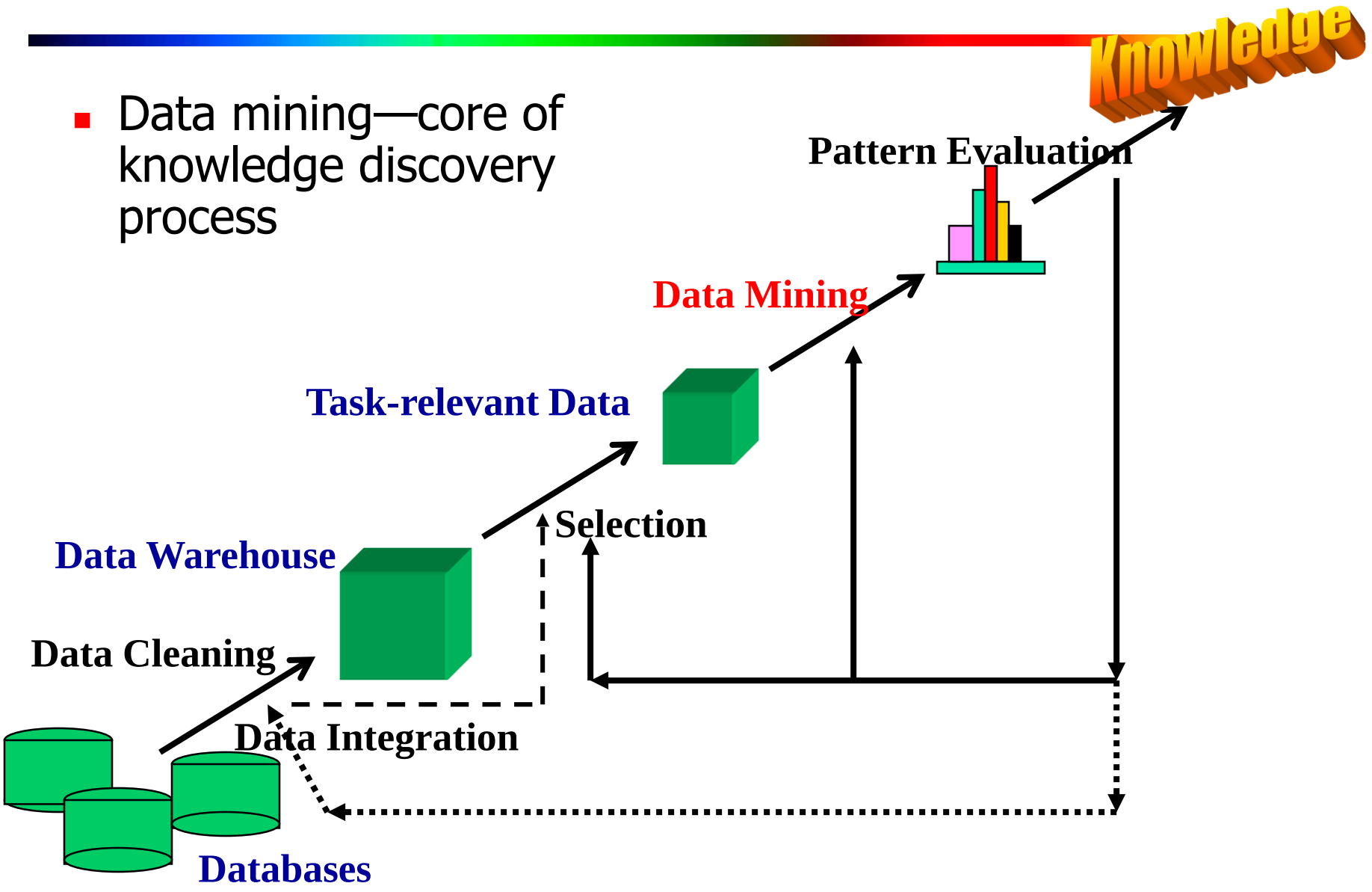
- Data mining is also referred as KDD
- KDD is Knowledge Discovery in Databases process
- It is a non-trivial process of identifying patterns that are:
  - Valid: The patterns hold in general.
  - Novel: We did not know the pattern beforehand.
  - Useful: We can devise actions from the patterns.
  - Understandable: We can interpret and comprehend the patterns.

KDD PROCESS

- Data mining emphasize on the analysis step of the "Knowledge Discovery in Databases" process
- KDD is not a single step solution of applying machine learning method / statistical technique / data mining task to a dataset, but is a continuous process with many loops and feedbacks.
- This process has been formalized by an industry group called CRISP-DM, which stands for CRoss Industry Standard Process for Data Mining

KNOWLEDGE DISCOVERY IN DATA BASES

- Data mining—core of knowledge discovery process



STEPS OF A KDD PROCESS

- Learning the application domain:
  - Relevant prior knowledge and goals of application
- Data Integration:
  - Relevant data from multiple source is combined
- Data preprocessing: (may take 60% of effort!)
  - Data selection - Creating a target data set
  - Data cleaning - To remove Noise, Missing and Inconsistent data
  - Data reduction and transformation:
    - Find useful features, dimensionality/variable reduction, generate new fields, invariant representation (create common units)

STEPS OF A KDD PROCESS

- 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 is an iterative process

DATA MINING: On What Kind 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: Patterns that can be Mined

- 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: Patterns that can be Mined

- 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 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

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, e.g., Illinois vs. Champaign rule implication support ratio)

TYPES OF VARIABLES

- Nominal or categorical:

  Domain is a finite set without any natural ordering (e.g., occupation, marital status, race)

  Examples: Gender, eye color, zip codes

- Ordinal:

  Domain is ordered, but absolute differences between values is unknown (e.g., preference scale, severity of an injury)

  Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall-01, medium-02, short-03}

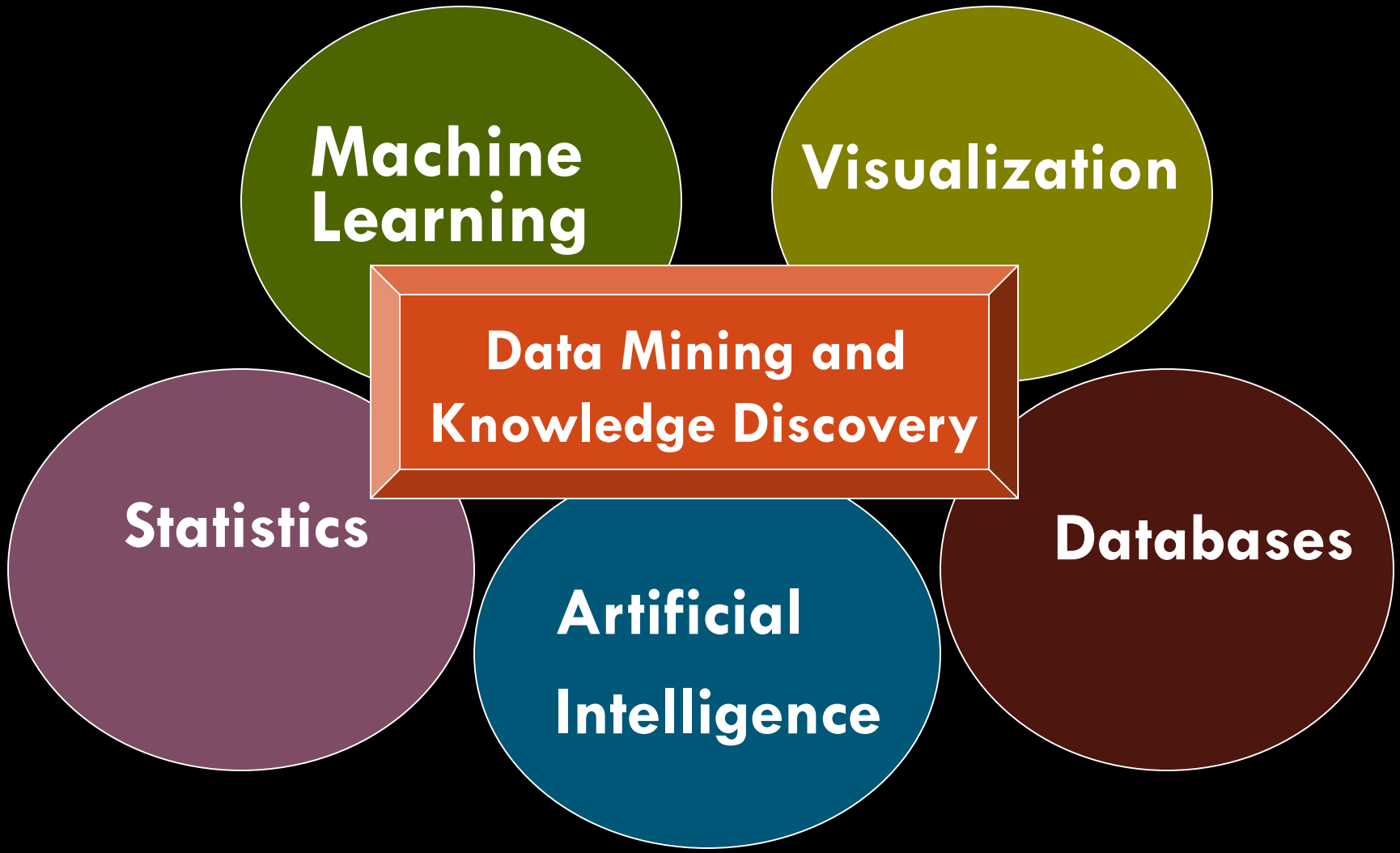
- Numerical:

  Domain is ordered and can be represented on the real line (e.g., age, income)

  - Interval Scaled Variable - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
  - Ratio Scaled Variable - Examples: temperature in Kelvin, length, time, counts

WHICH TECHNOLOGY ARE USED?

Confluence of Multiple Discipline



DATA MINING & KDD:

Confluence of Multiple Discipline

- Data Mining and Knowledge Discovery is an interdisciplinary subfield of computer science, which builds upon a foundation provided by **databases and statistics** and applies methods from **machine learning and visualization** in order to find the useful patterns.
- Other related fields include also information retrieval, artificial intelligence, OLAP, etc.
- Data mining, statistics and machine learning has many things in common. However, there are differences.

DATA MINING: STATISTICS

- Statistics provides a solid theory for dealing with randomness and tools for testing hypotheses.
- It has a set of mathematical functions that describes the behavior of the objects in a target class in terms of random variable and their associated probability distribution.
- Statistical models are the outcome of data mining tasks like characterization and classification.
- Statistical methods are used to verify data mining results using statistical hypothesis test or confirmatory data analysis.
- However, Statistics does not study topics such as data preprocessing or results visualization, which are part of data mining.

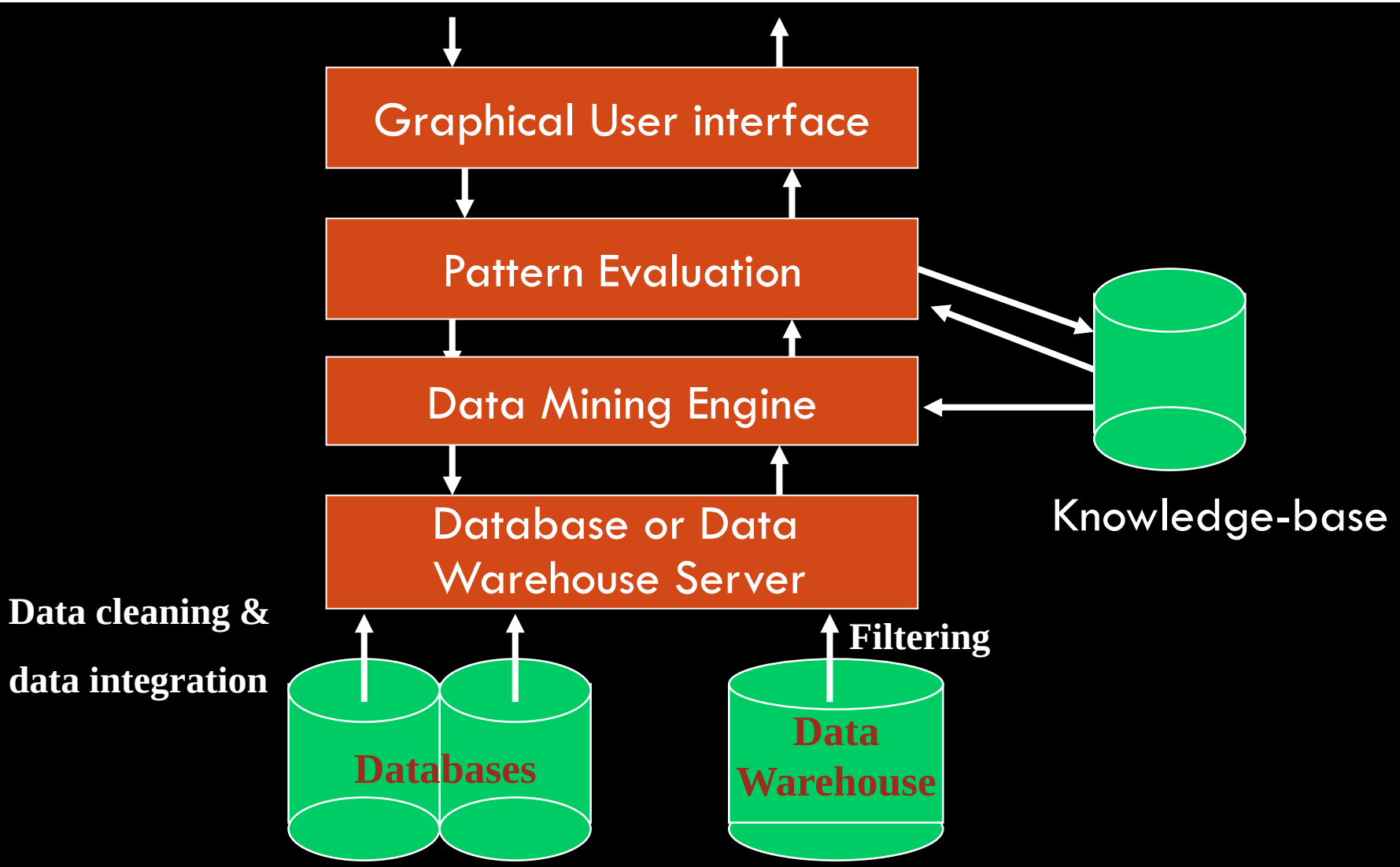
DATA MINING: Machine Learning

- Machine learning has a more heuristic approach and is focused on improving performance of a learning agent.
  - Teaching the machine to teach itself
  - Writing programs that can learn
- Technical basis for data mining: algorithms for acquiring structural descriptions from examples (training data) which can be used to:
  - predict outcome in new situation
  - understand and explain how prediction is derived
- Methods originate from artificial intelligence, statistics and research on databases
- It also has other subfields such as real-time learning and robotics - which are not part of data mining.

DATA MINING: Machine Learning

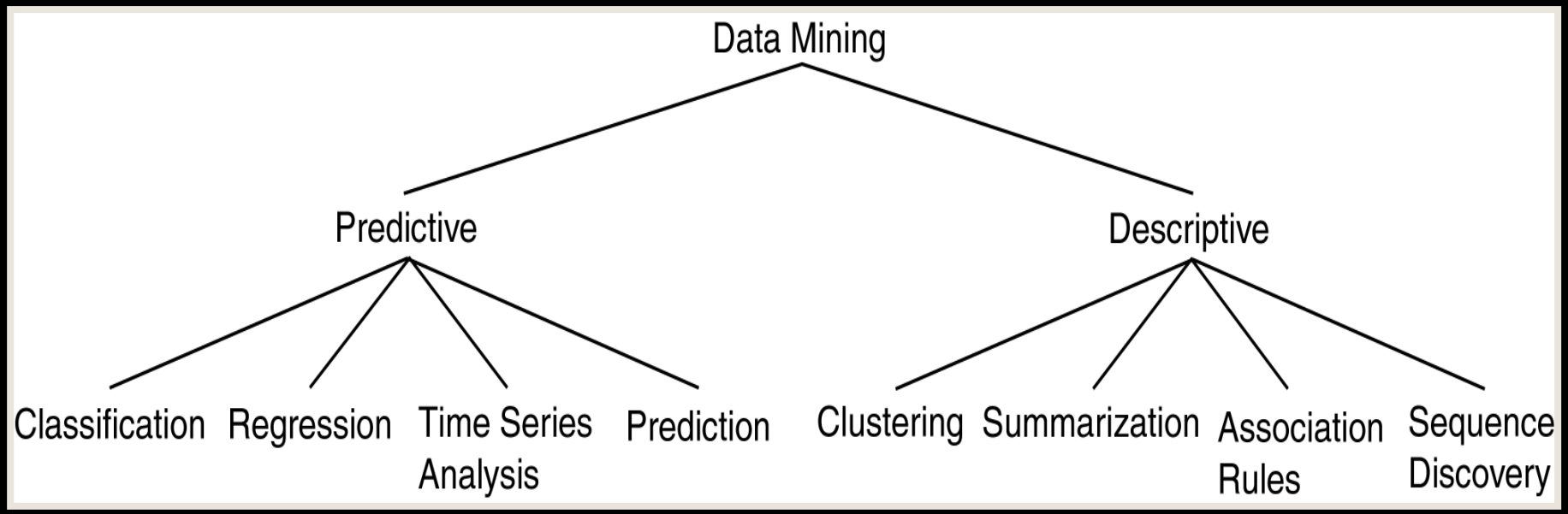
- Area of AI that examines how to write programs that can learn.
- Often used in classification and prediction
- Types
  - Supervised Learning: learns by example.
  - Unsupervised Learning: learns without knowledge of correct answers.
- Machine learning often deals with small static datasets.
- Data Mining: Uses many machine learning techniques.
- Data Mining and Knowledge Discovery field integrates theory and heuristics. It focuses on the entire process of knowledge discovery, including data cleaning, learning, and integration and visualization of results.

DATA MINING ARCHITECTURE



BASIC DATA MINING OPERATIONS

- Predictive: Supervised Learning
  - Use some variables to predict unknown or future values of other variables.
- Descriptive: Unsupervised learning
  - Find human-interpretable patterns that describe the data.



DATA MINING TECHNIQUES

- Supervised Learning
  - Classification and Regression – Predicting an Item's Class
- Unsupervised learning
  - Clustering – Grouping similar data
- Dependency modeling
  - Associations, summarization, causality – Frequency of Co-occurrences
- Outlier and deviation detection – Finding Changes
- Trend analysis and change detection
- Estimation – prediction of continuous value
- Link Analysis – Finding relationship

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 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)

More Examples…..

- Example 2: **US Presidential Election**: Polling data can be analyzed using data mining techniques to accurately predict the outcome in each state
- Example 3: **The gene pattern**: Vast amount of data is collected in molecular biology research scientists use large amounts of genomic data to better understand the structure and function of genes.
- Example 4: **Google's Flu trend**: The search data available at the Google search engine can be mined to predict flu in a particular area at much faster pace than that of traditional system.

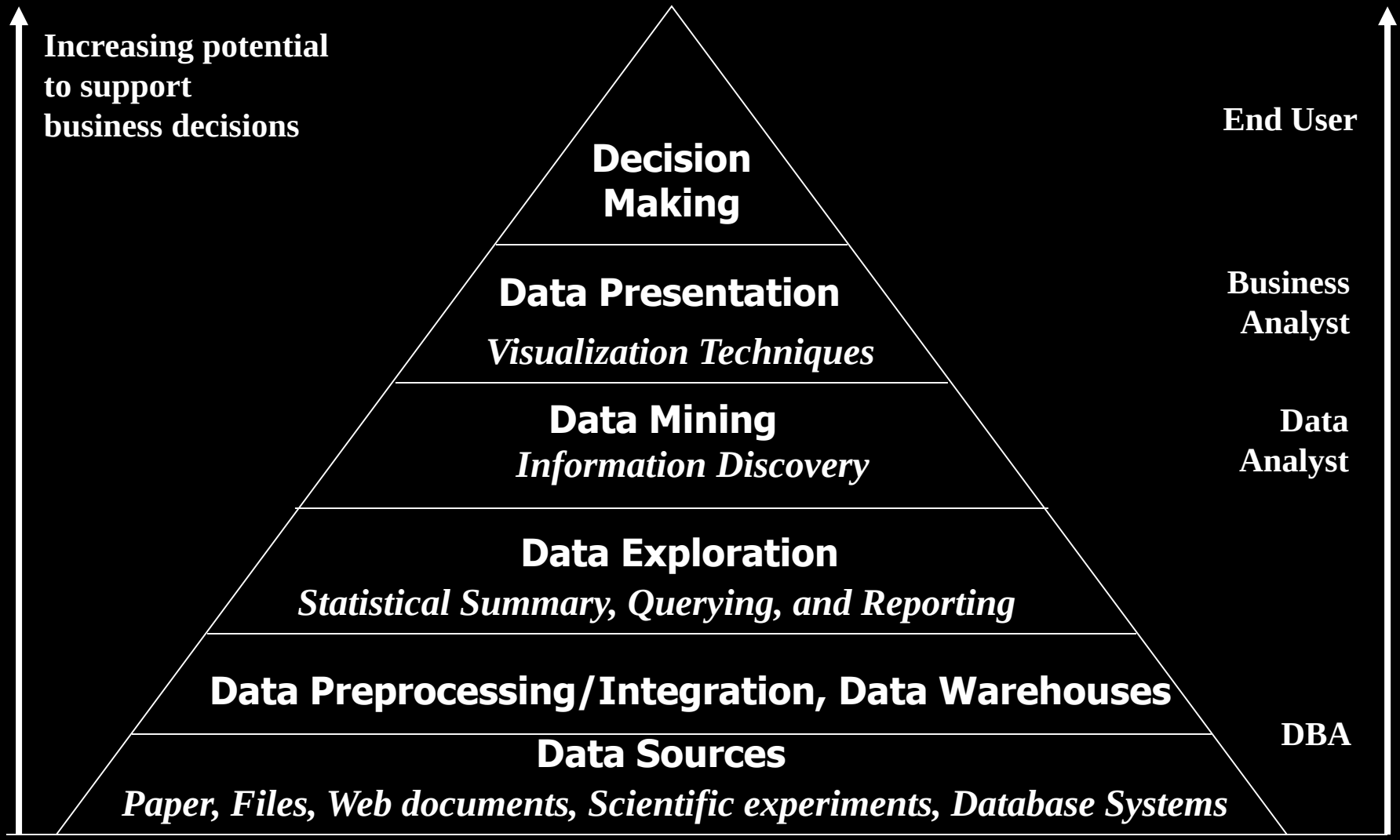
More Examples…

- Example 5: **Earth's Climate System**: Earth scientists track and analyze weather data over time to help predict, prepare for and perhaps even prevent catastrophic changes in climate.
- Example 6: **Healthcare System**: Information gleaned from big data can even identify disease outbreaks in time to impose quarantines that will prevent an epidemic.

Ex. 8: 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

SUMMARY: DATA MINING & BUSINESS INTELLIGENCE



DATA MINING – Multidimensional View

- **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.

Thank you

QUESTIONS????