

DATA SCI 8410: Data Mining & Information Retrieval

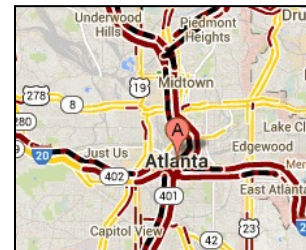
Tozammel Hossain



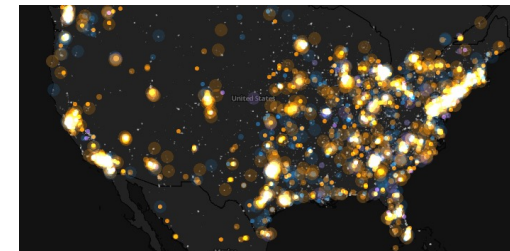
Data Science & Analytics
University of Missouri

Large-scale Data is Everywhere!

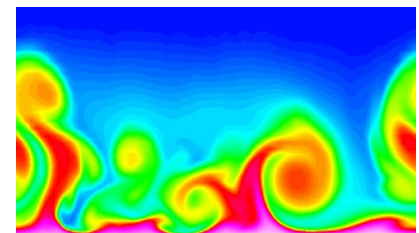
- There has been enormous data growth in both commercial and scientific databases due to advances in data generation and collection technologies
- **New mantra**
 - Gather whatever data you can whenever and wherever possible.
- **Expectations**
 - Gathered data will have value either for the purpose collected or for a purpose not envisioned.



Traffic Patterns



Social Networking: Twitter



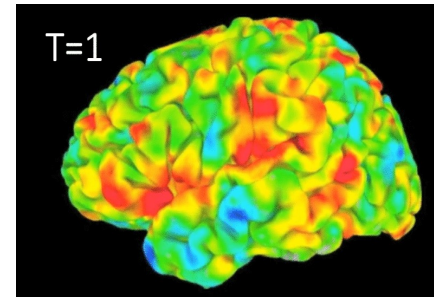
Why Data Mining? Commercial Viewpoint

- **Lots of data is being collected and warehoused**
 - Web data
 - Google has Peta Bytes of web data
 - Facebook has billions of active users
 - purchases at department/grocery stores, e-commerce
 - Amazon handles millions of visits/day
 - Bank/Credit Card transactions
- **Computers have become cheaper and more powerful**
- **Competitive Pressure is Strong**
 - Provide better, customized services for an edge (e.g. in Customer Relationship Management)



Why Data Mining? Scientific Viewpoint

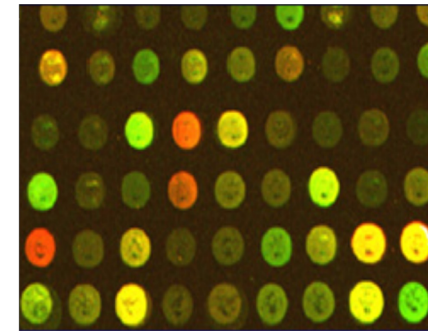
- **Data collected and stored at enormous speeds**
 - remote sensors on a satellite
 - NASA EOSDIS archives over petabytes of earth science data / year
 - telescopes scanning the skies
 - Sky survey data
 - High-throughput biological data
 - scientific simulations
 - terabytes of data generated in a few hours
- **Data mining helps scientists**
 - in automated analysis of massive datasets
 - In hypothesis formation



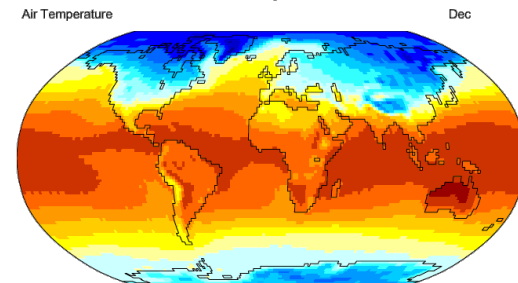
fMRI Data from Brain



Sky Survey Data



Gene Expression Data



Surface Temperature of Earth

Great opportunities to improve productivity in all walks of life

McKinsey Global Institute

Big data: The next frontier for innovation, competition, and productivity

Big data—a growing torrent

\$600 to buy a disk drive that can store all of the world's music

5 billion mobile phones in use in 2010

30 billion pieces of content shared on Facebook every month

40% projected growth in global data generated per year vs. **5%** growth in global IT spending

235 terabytes data collected by the US Library of Congress in April 2011

15 out of 17 sectors in the United States have more data stored per company than the US Library of Congress

Big data—capturing its value

\$300 billion potential annual value to US health care—more than double the total annual health care spending in Spain

€250 billion potential annual value to Europe's public sector administration—more than GDP of Greece

\$600 billion potential annual consumer surplus from using personal location data globally

60% potential increase in retailers' operating margins possible with big data

140,000–190,000 more deep analytical talent positions, and

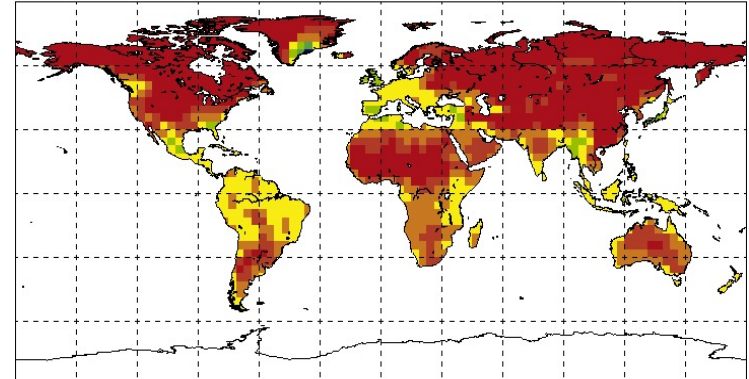
1.5 million more data-savvy managers needed to take full advantage of big data in the United States

Great Opportunities to Solve Society's Major Problems



Improving health care and reducing costs

CCCma/A2a January to January Mean Temperature (degrees C) 2080s relative to 1961-90



Predicting the impact of climate change



Finding alternative/ green energy sources



Reducing hunger and poverty by increasing agriculture production

What is Data Mining?

- **Many Definitions**

- Non-trivial extraction of implicit, previously unknown and potentially useful information from data
- Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns

- **Alternative names**

- Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

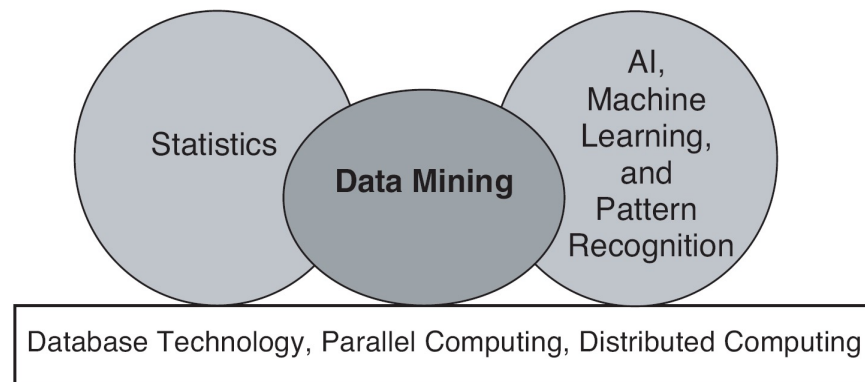
- **Watch out: Is everything “data mining”?**

- Simple search and query processing

Origins of Data Mining

- **Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems**
- **Traditional techniques may be unsuitable due to data that is**

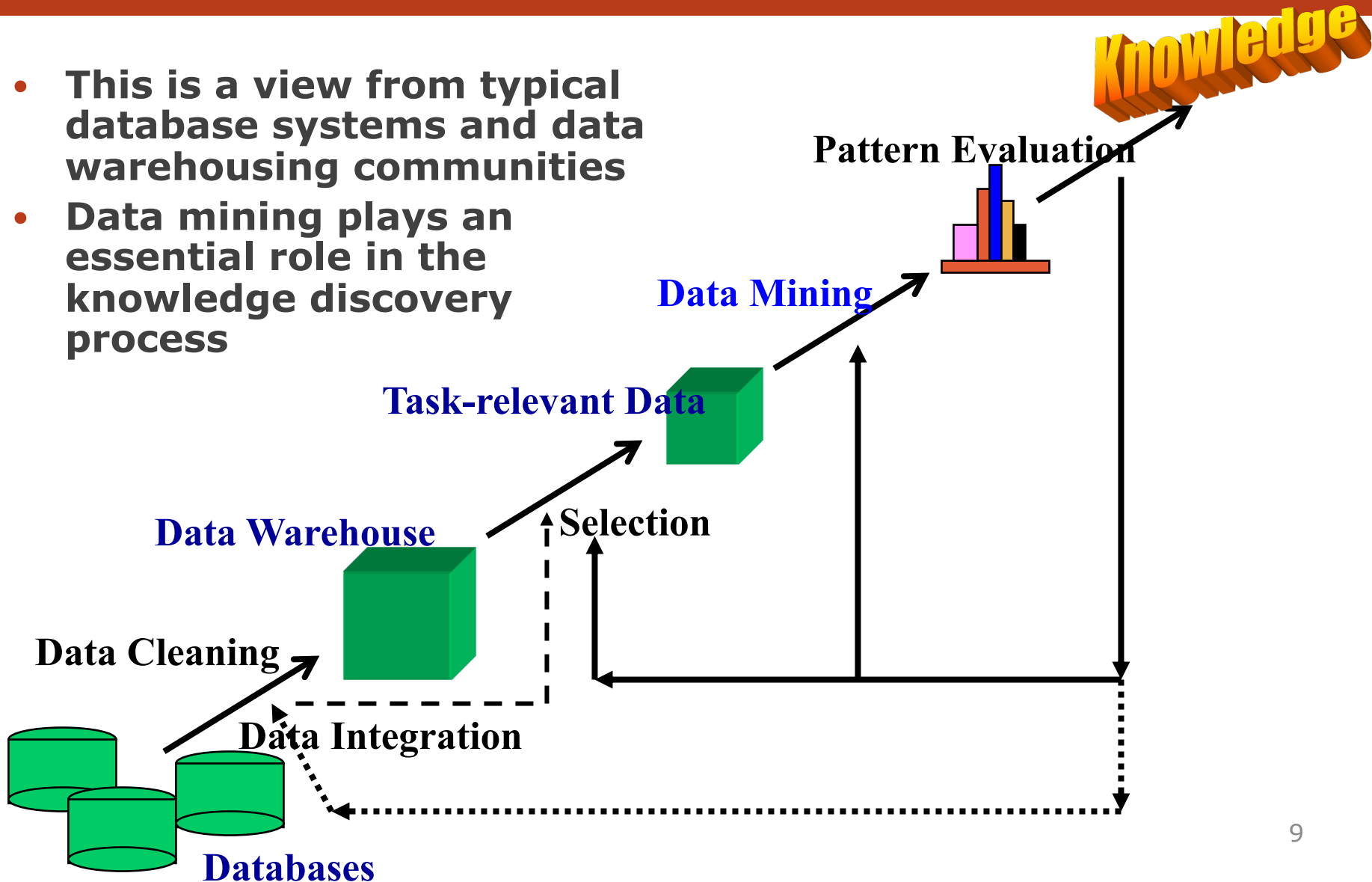
- Large-scale
- High dimensional
- Heterogeneous
- Complex
- Distributed



- **A key component of the emerging field of data science and data-driven discovery**

Knowledge Discovery (KDD) Process

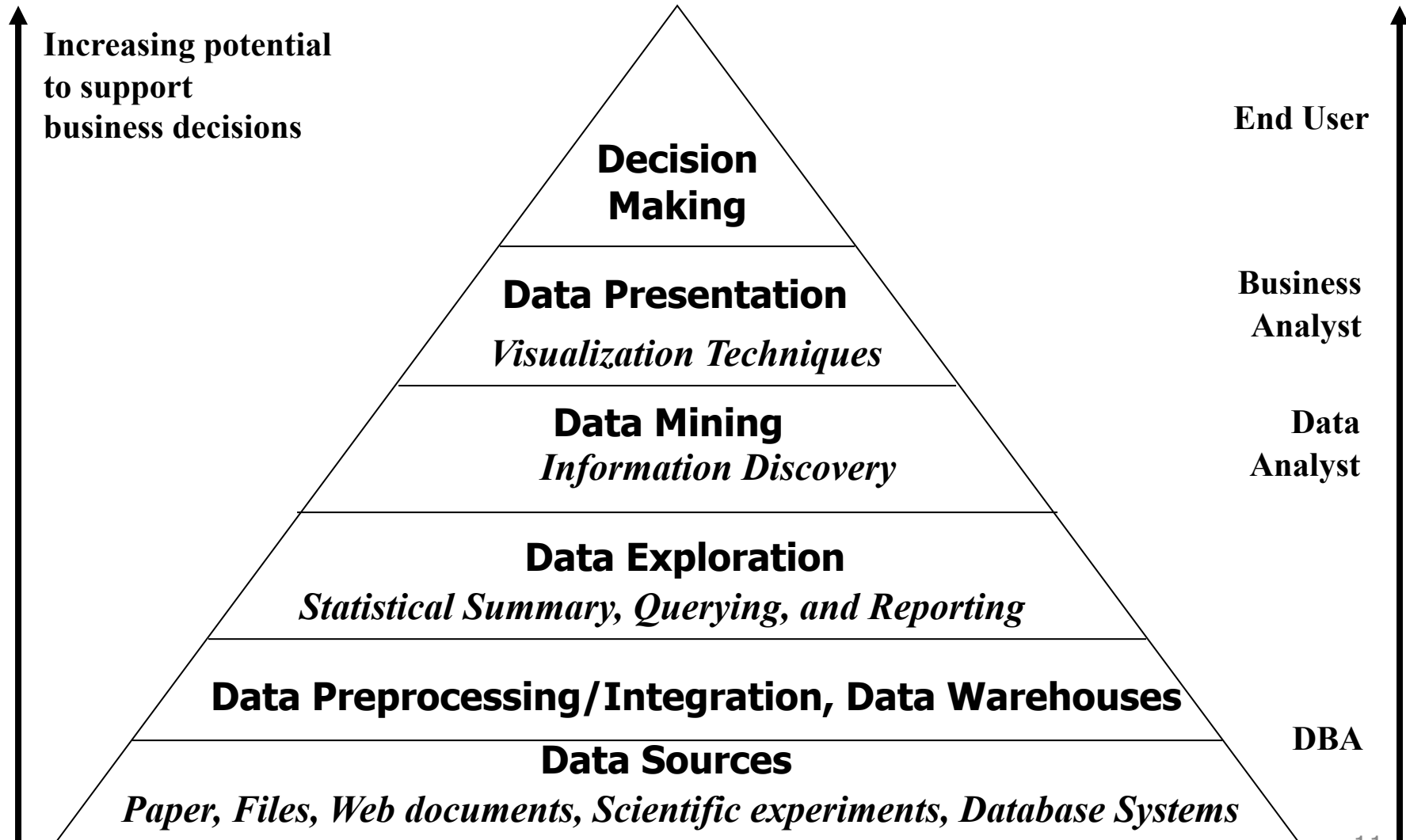
- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process



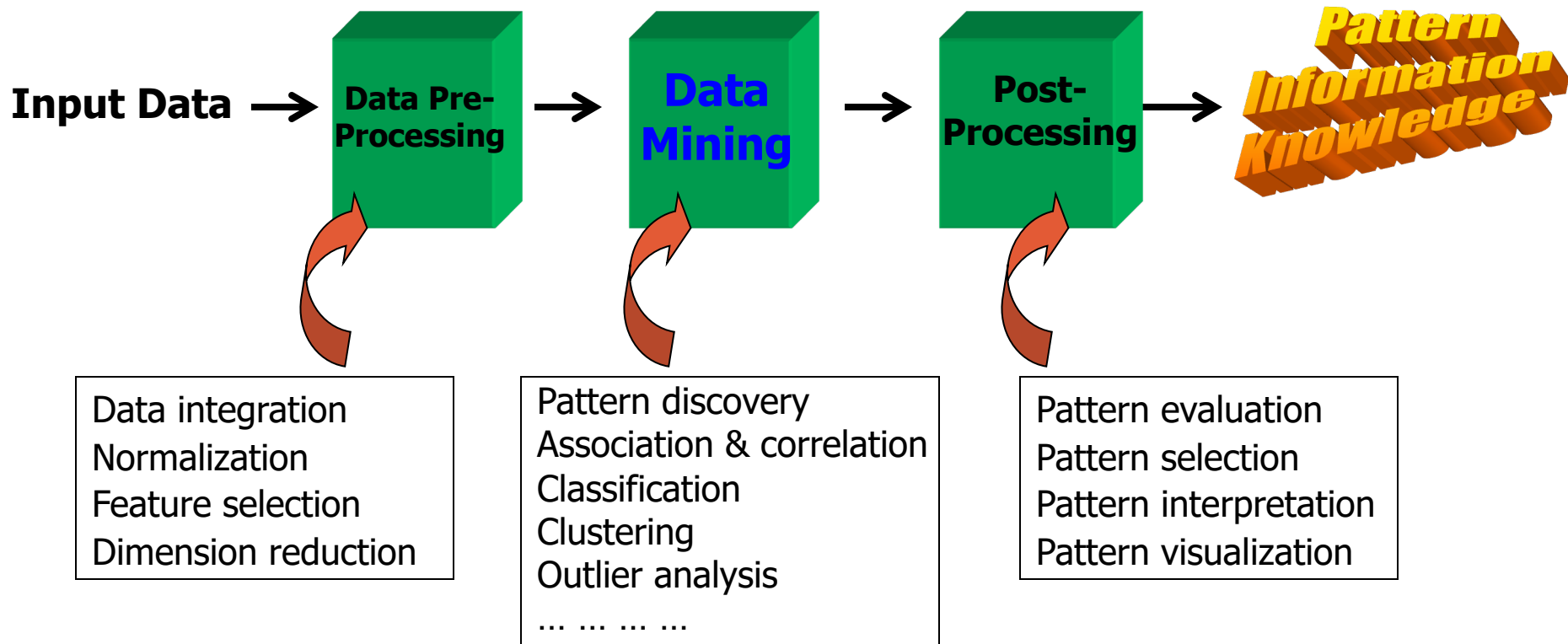
Example: A Web Mining Framework

- **Web mining usually involves**
 - Data cleaning
 - Data integration from multiple sources
 - Warehousing the data
 - Data cube construction
 - Data selection for data mining
 - Data mining
 - Presentation of the mining results
 - Patterns and knowledge to be used or stored into knowledge-base

Data Mining in Business Intelligence



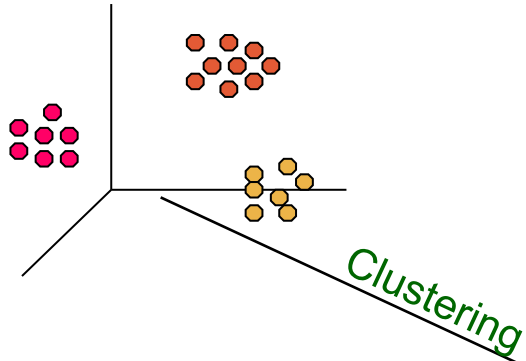
KDD Process: A typical view



Data Mining Tasks

- **Prediction Methods**
 - Use some variables to predict unknown or future values of other variables.
- **Description Methods**
 - Find human-interpretable patterns that describe the data.

Data Mining Tasks ...



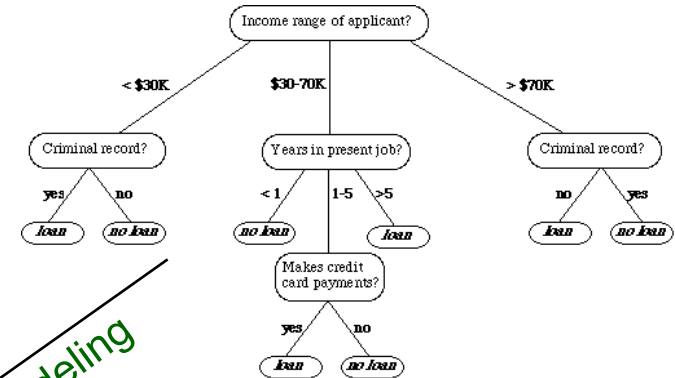
Data

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

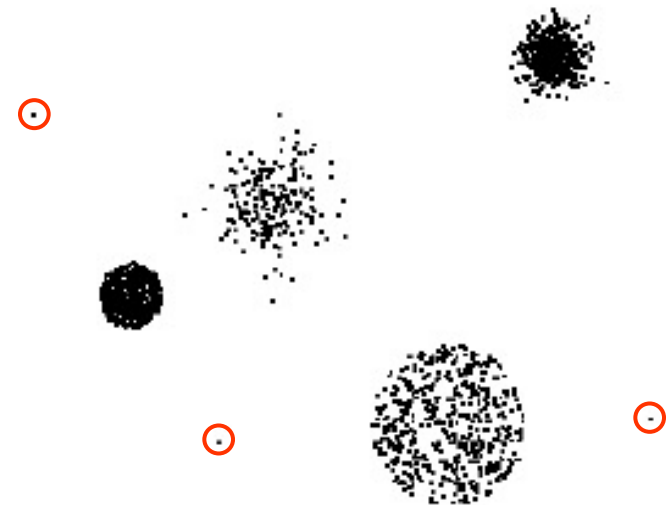
Association Rules



Predictive Modeling



Anomaly Detection



Multi-Dimensional View of Data Mining

- **Data to be mined**
 - Database data (extended-relational, object-oriented, heterogeneous, legacy), data warehouse, transactional data, stream, spatiotemporal, time-series, sequence, text and web, multi-media, graphs & social and information networks
- **Knowledge to be mined (or: Data mining functions)**
 - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
 - Descriptive vs. predictive data mining
 - Multiple/integrated functions and mining at multiple levels
- **Techniques utilized**
 - Data-intensive, data warehouse (OLAP), machine learning, statistics, pattern recognition, visualization, high-performance, etc.
- **Applications adapted**
 - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

Information Retrieval

- **Information Retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).**
 - These days we frequently think first of **web search**, but there are many other cases:
 - E-mail search
 - Searching your laptop
 - Corporate knowledge bases
 - Legal information retrieval

Basic assumptions of Information Retrieval

- **Collection: A set of documents**
 - Assume it is a static collection for the moment
- **Goal: Retrieve documents with information that is relevant to the user's information need and helps the user complete a task**

How good are the retrieved docs?

- ***Precision*** : Fraction of retrieved docs that are relevant to the user's **information need**
- ***Recall*** : Fraction of relevant docs in collection that are retrieved
 - More precise definitions and measurements to follow later

Thanks