

Artificial Intelligence

Lecture 1. Introduction

Spring 2022

Prof. Jonghoon Chun, Ph.D.

E-mail : jchun@mju.ac.kr
Lecture Note : <http://lms.mju.ac.kr>

데이터가 없으면 인공지능도 없다!

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.



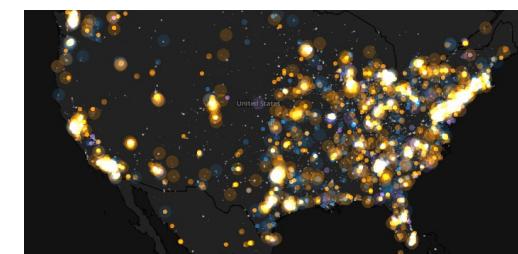
Cyber Security



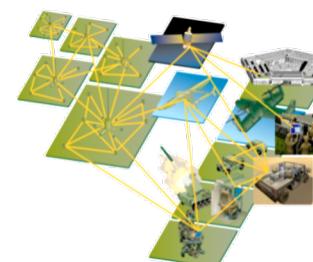
E-Commerce



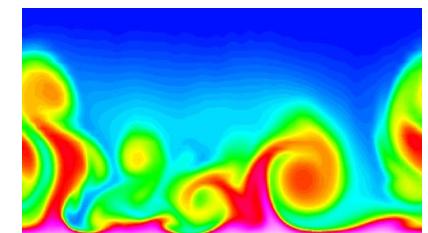
Traffic Patterns



Social Networking: Twitter



Sensor Networks



Computational Simulations

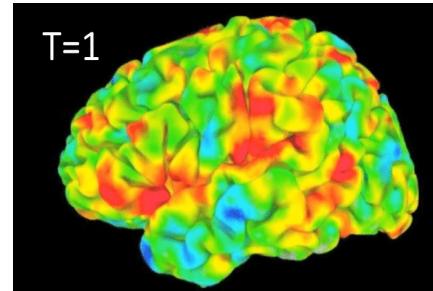
Why? Commercial Viewpoint

- Lots of data is being collected and warehoused
 - Web data
 - Yahoo 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? Scientific Viewpoint

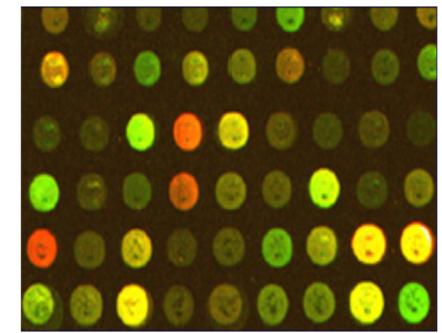
- Data collected and stored at enormous speeds
 - remote sensors on a satellite
 - NASA EOSDIS archives over petabytes of earth science data / year
 - telescopes scanning the skies
 - Sky survey data
 - High-throughput biological data
 - scientific simulations
 - terabytes of data generated in a few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



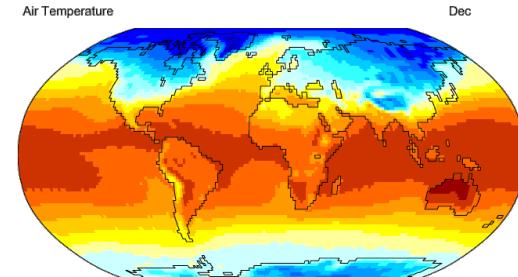
MRI Data from Brain



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

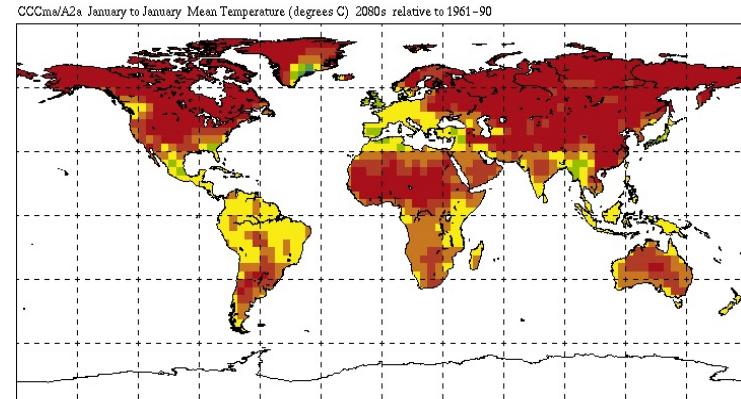
Great Opportunities to Solve Society's Major Problems



Improving health care and reducing costs



Finding alternative/ green energy sources



Predicting the impact of climate change



Reducing hunger and poverty by increasing agriculture production



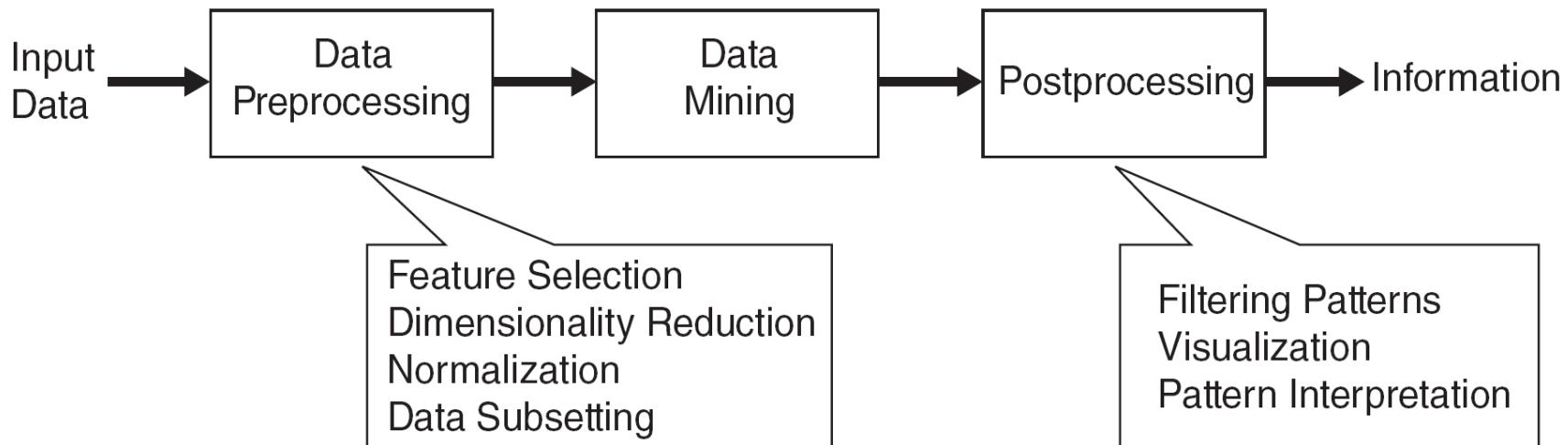
What is Data Mining?

- Data mining (knowledge discovery in databases):
 - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases
- Alternative names and their “inside stories”:
 - Data mining: a misnomer?
 - Knowledge discovery(mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, etc.
 - advanced OLAP from DW perspectives



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



This is a view from typical machine learning and statistics communities!

What is (not) Data Mining?

What is not Data Mining?

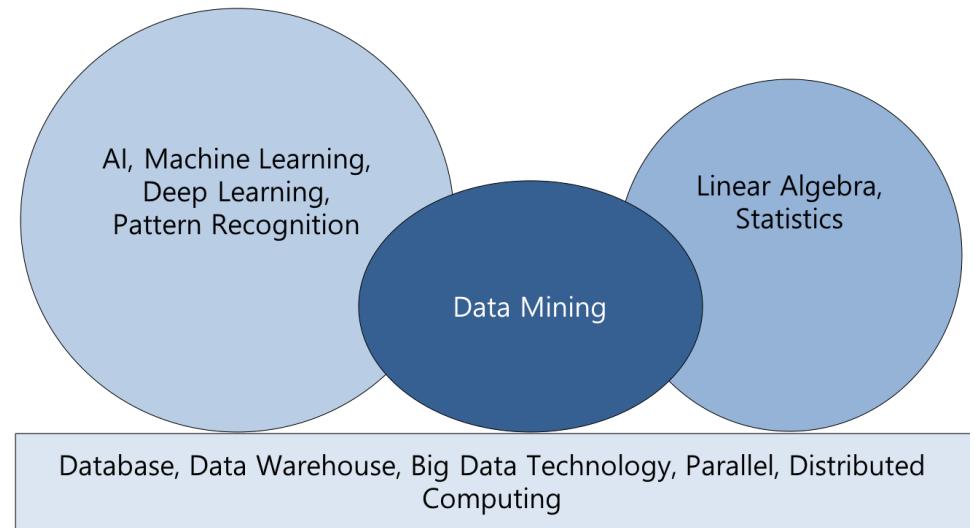
- Look up phone number in phone directory
- Query a Web search engine for information about "Amazon"
- (Deductive) query processing
- Expert systems
- Small machine learning / statistical programs

What is Data Mining?

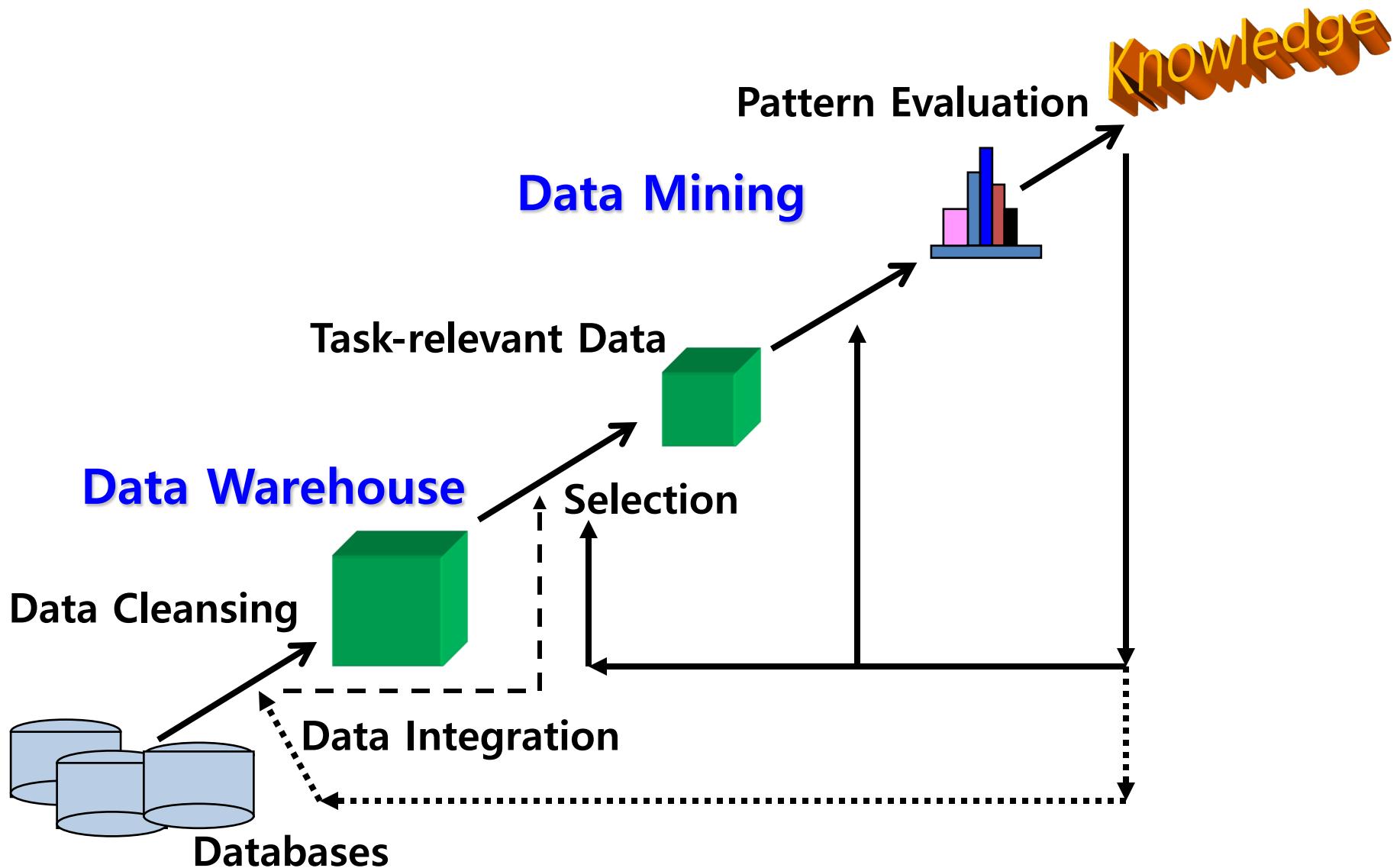
- Certain names are more prevalent in certain US locations (O'Brien, O'Rourke, O'Reilly... in Boston area)
- Group together similar documents returned by search engine according to their context (e.g., Amazon rainforest, Amazon.com)

Origins of Data Mining

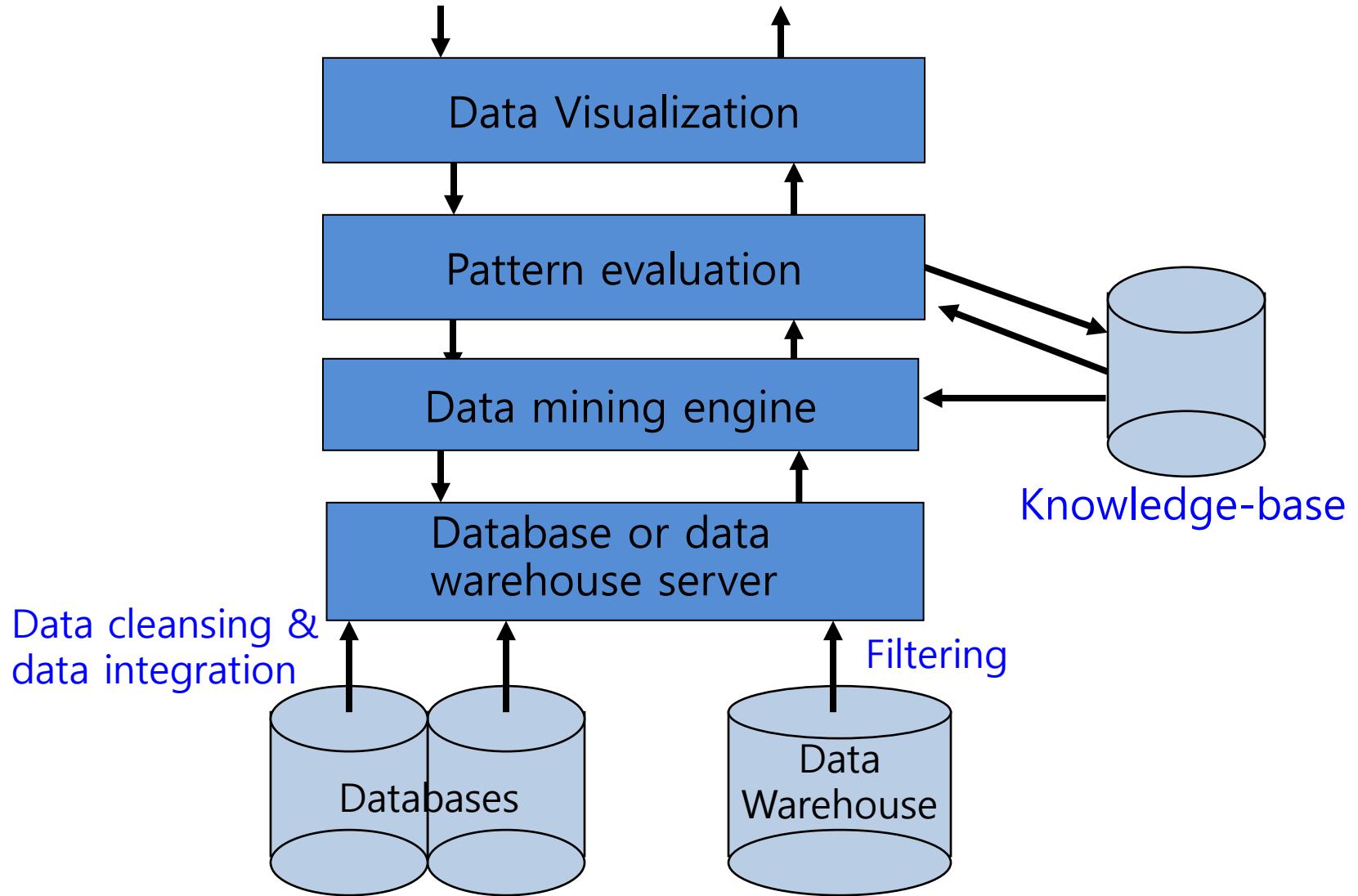
- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems
- Traditional techniques may be unsuitable due to data that is
 - Large-scale
 - High dimensional
 - Heterogeneous
 - Complex
 - Distributed
- A key component of the emerging field of data science and data-driven discovery (vs. query-driven dataset discovery)



Data Mining: A KDD Process



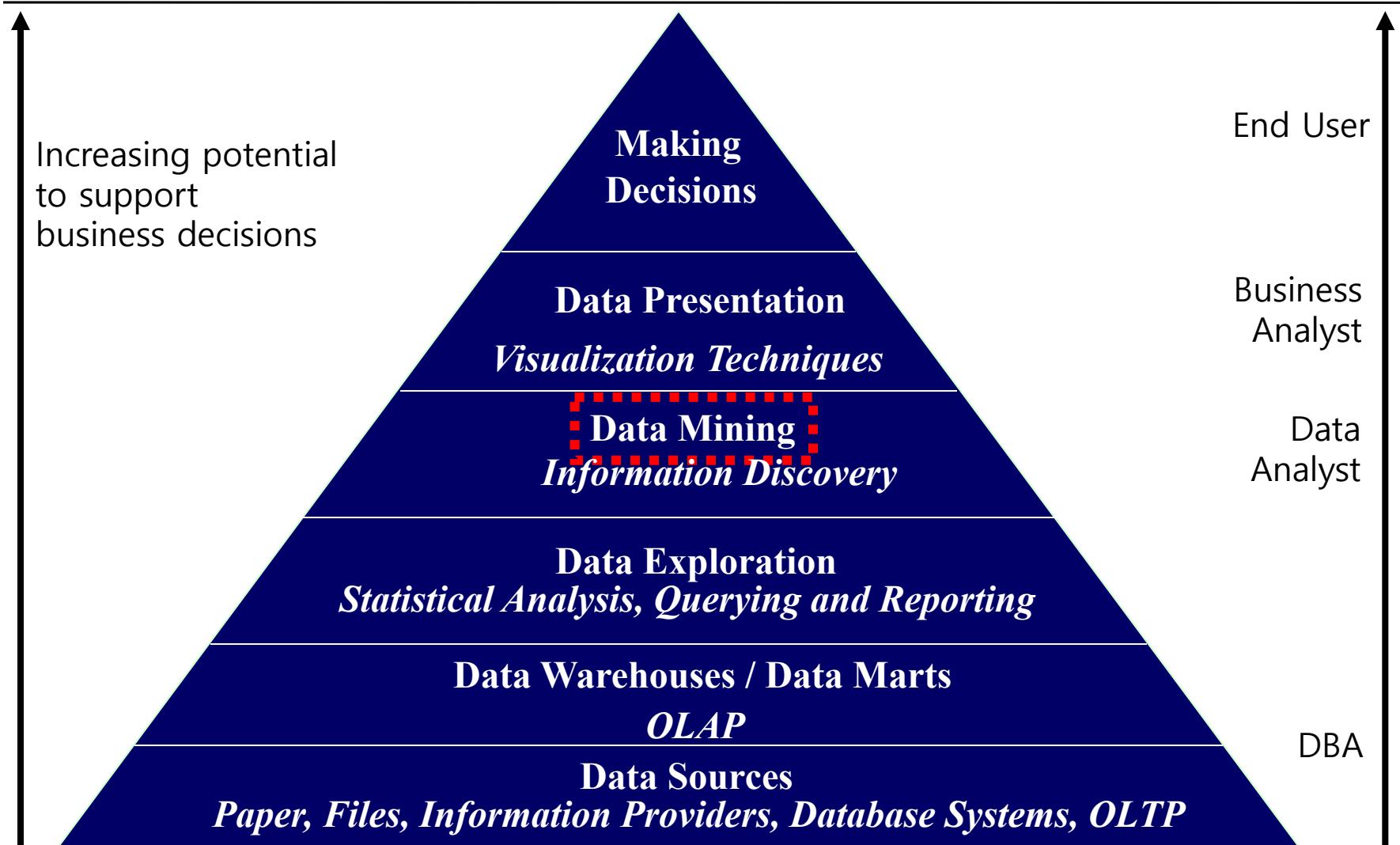
Architecture of a Typical Data Mining System



Components of a typical Data Mining System

- Database, data warehouse, other information repository
 - Data cleansing / integration needed
- DB/DW Server
- Knowledge base
 - Domain knowledge used to guide the search & evaluate the resulting patterns
- Data Mining Engine
 - Characterization, association, classification, cluster analysis, evolution & deviation analysis
- Pattern Evaluation Module
 - Uses interestingness measure to evaluate patterns
 - Maybe integrated with the mining modules
- Graphical User Interface
 - Data visualization

Data Mining and Business Intelligence



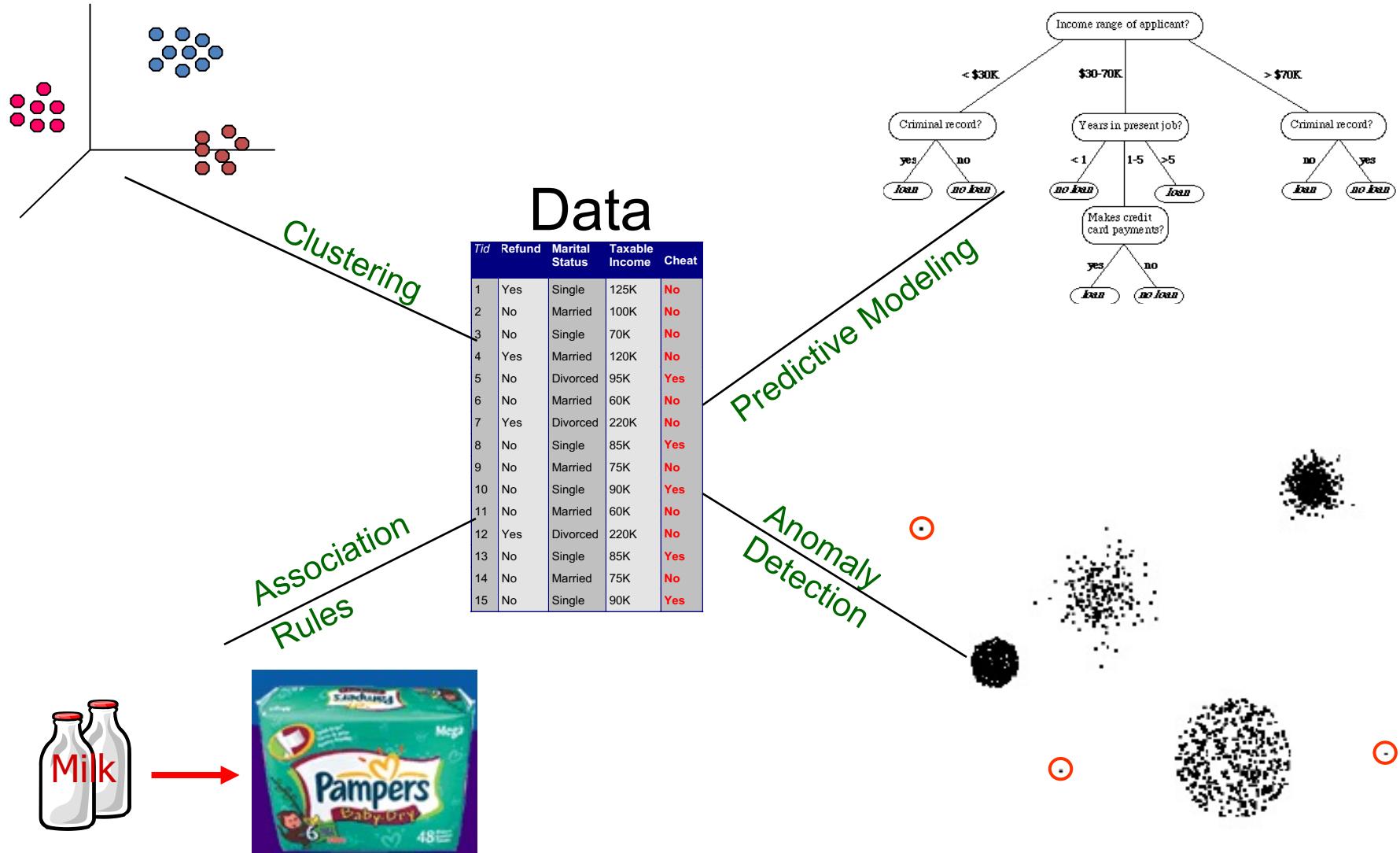
Data Mining Tasks

- Prediction Methods
 - Use some variables to predict unknown or future values of other variables.

- Description Methods
 - Find human-interpretable patterns that describe the data.

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Data Mining Tasks...

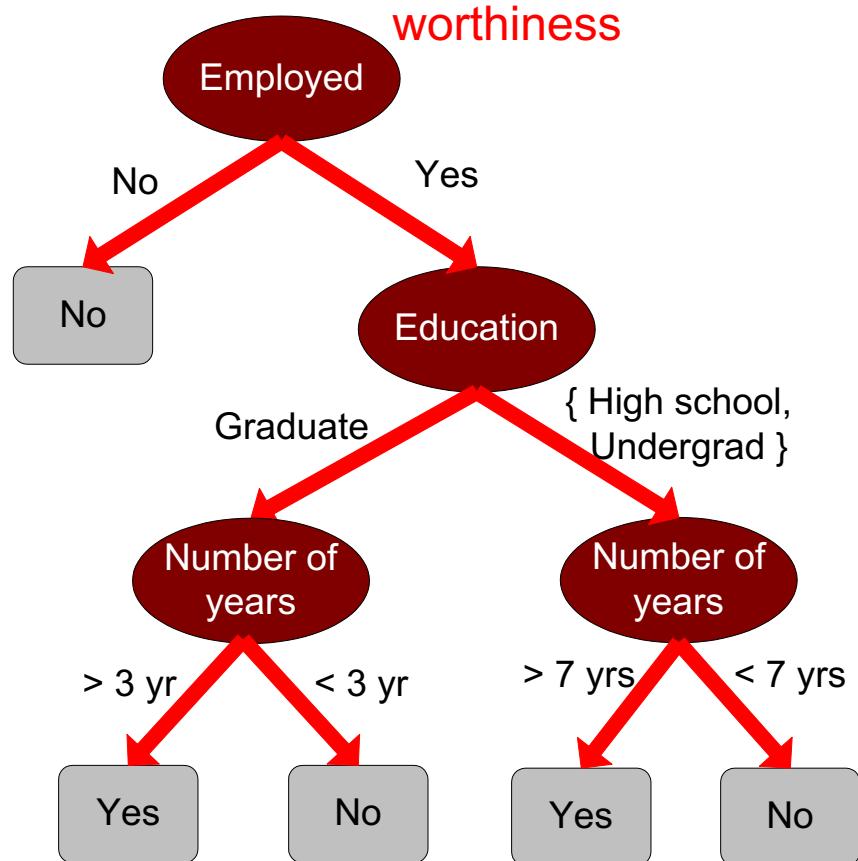


Predictive Modeling: Classification

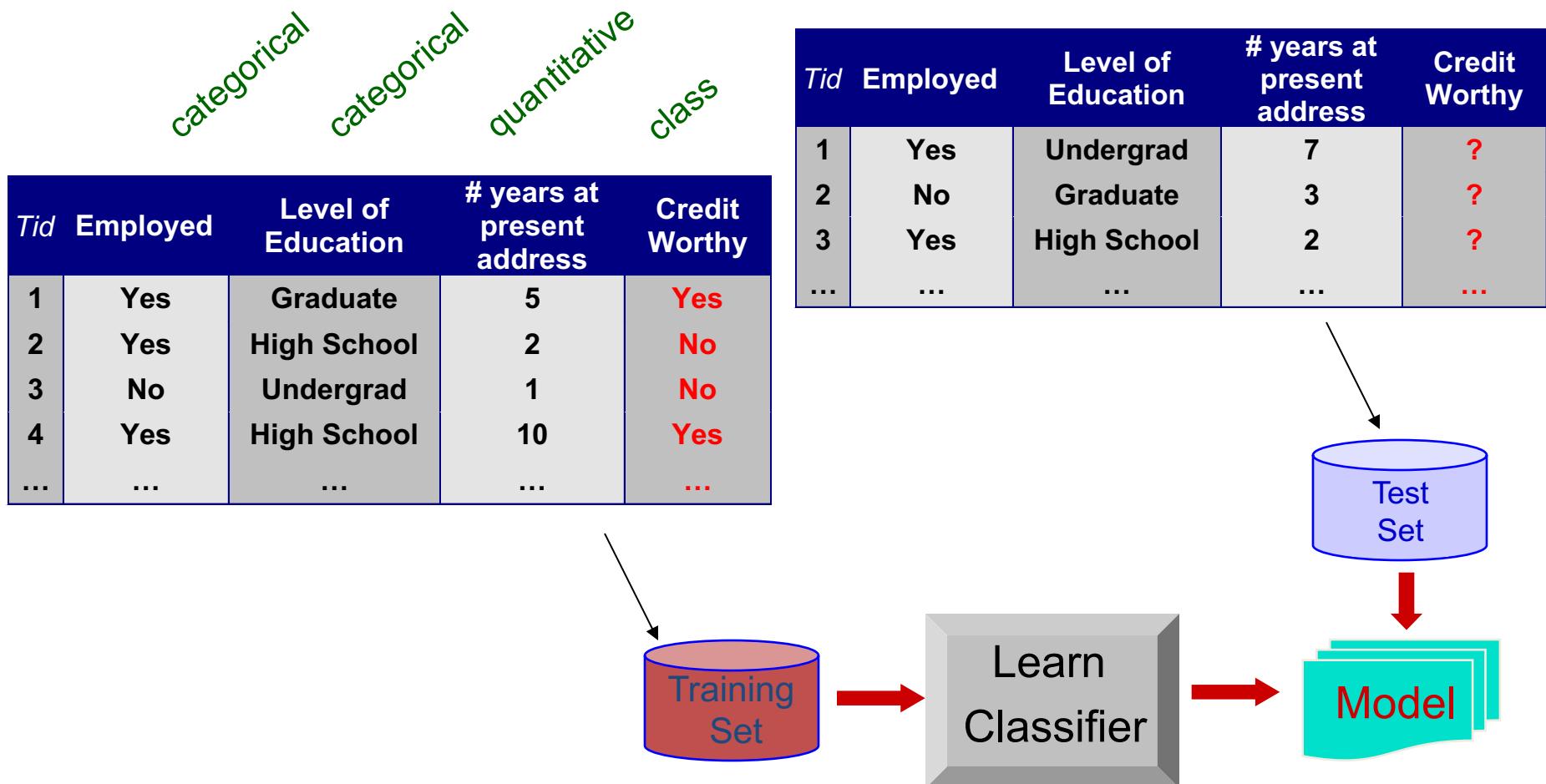
- Find a model for class attribute as a function of the values of other attributes

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
...

Model for predicting credit worthiness

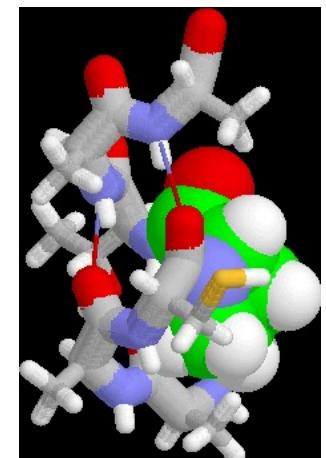


Classification Example



Examples of Classification Task

- Classifying credit card transactions as legitimate or fraudulent
- Classifying land covers (water bodies, urban areas, forests, etc.) using satellite data
- Categorizing news stories as finance, weather, entertainment, sports, etc
- Identifying intruders in the cyberspace
- Predicting tumor cells as benign or malignant
- Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil



Classification: Application 1

- Fraud Detection
 - **Goal:** Predict fraudulent cases in credit card transactions.
 - **Approach:**
 - Use credit card transactions and the information on its account-holder as attributes.
 - ✓ When does a customer buy, what does he buy, how often he pays on time, etc
 - Label past transactions as fraud or fair transactions. This forms the class attribute.
 - Learn a model for the class of the transactions.
 - Use this model to detect fraud by observing credit card transactions on an account.

Classification: Application 2

- Churn prediction for telephone customers
 - **Goal:** To predict whether a customer is likely to be lost to a competitor.
 - **Approach:**
 - Use detailed record of transactions with each of the past and present customers, to find attributes.
 - ✓ How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
 - Label the customers as loyal or disloyal.
 - Find a model for loyalty.

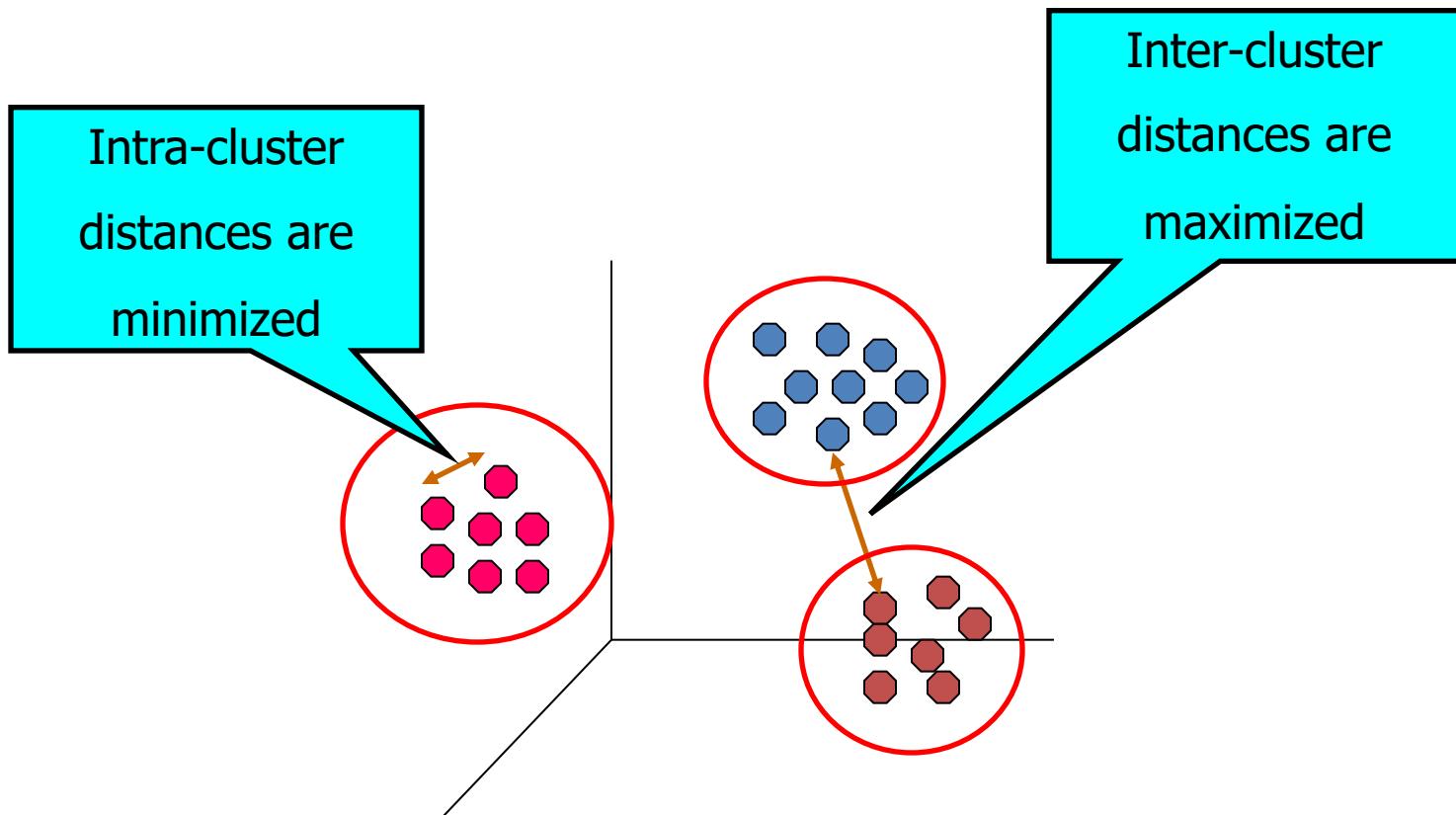
From [Berry & Linoff] Data Mining Techniques, 1997

Regression

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Extensively studied in statistics, neural network fields.
- Examples:
 - Predicting sales amounts of new product based on advertising expenditure.
 - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
 - Time series prediction of stock market indices.

Clustering

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



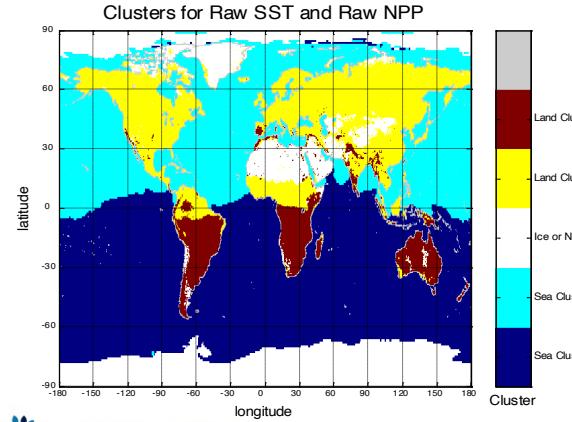
Applications of Cluster Analysis

▪ Understanding

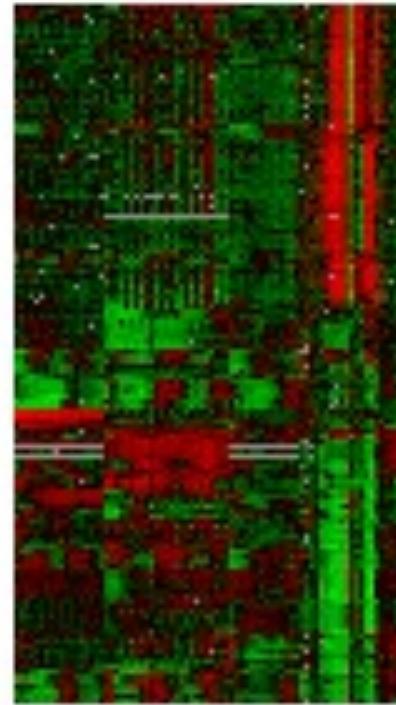
- Custom profiling for targeted marketing
- Group related documents for browsing
- Group genes and proteins that have similar functionality
- Group stocks with similar price fluctuations

▪ Summarization

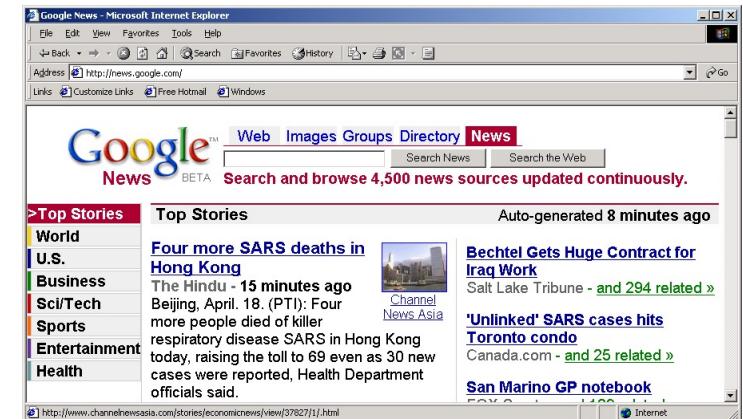
- Reduce the size of large data sets



Use of K-means to partition Sea Surface Temperature (SST) and Net Primary Production (NPP) into clusters that reflect the Northern and Southern Hemispheres.



Courtesy: Michael Eisen



Clustering: Application 1

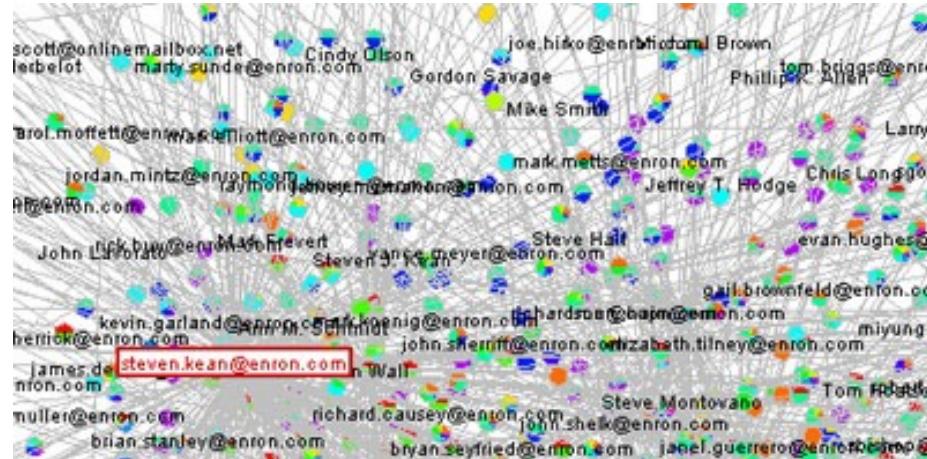
- Market Segmentation:
 - **Goal:** subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
 - **Approach:**
 - Collect different attributes of customers based on their geographical and lifestyle related information.
 - Find clusters of similar customers.
 - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Clustering: Application 2

- Document Clustering:

- **Goal:** To find groups of documents that are similar to each other based on the important terms appearing in them.
- **Approach:** To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.

Enron email dataset



Association Rule Discovery: Definition

- Given a set of records each of which contain some number of items from a given collection
 - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:

$\{\text{Milk}\} \rightarrow \{\text{Coke}\}$

$\{\text{Diaper}, \text{Milk}\} \rightarrow \{\text{Beer}\}$

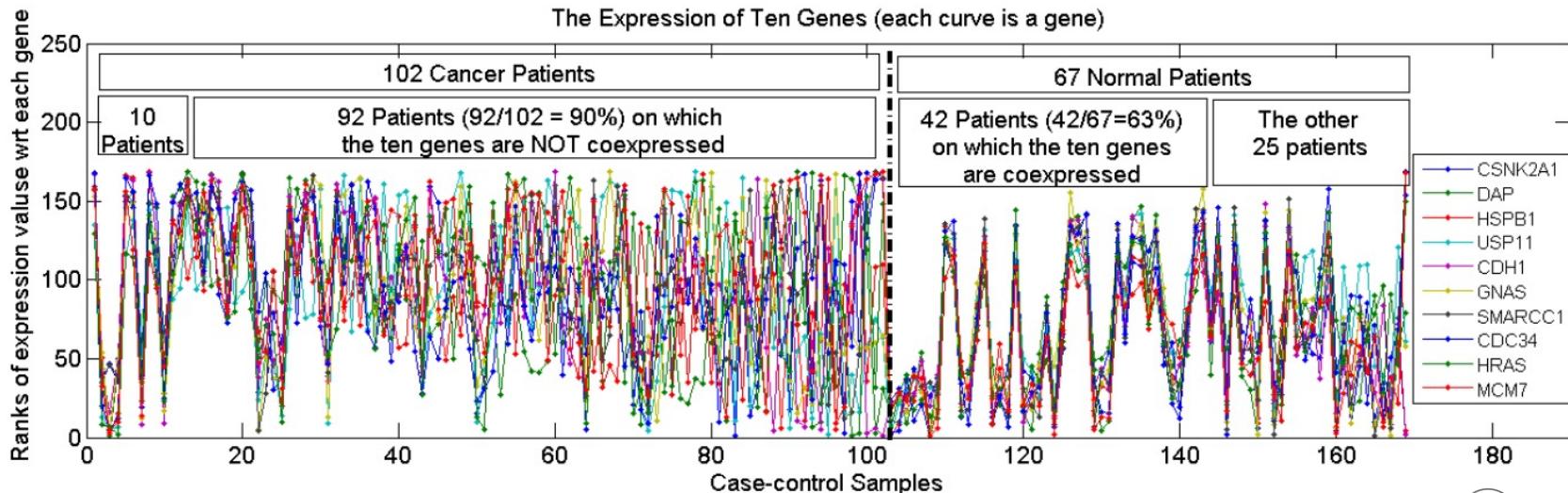
Association Analysis: Applications

- Market-basket analysis
 - Rules are used for sales promotion, shelf management, and inventory management
- Telecommunication alarm diagnosis
 - Rules are used to find combination of alarms that occur together frequently in the same time period
- Medical Informatics
 - Rules are used to find combination of patient symptoms and test results associated with certain diseases

Association Analysis: Applications

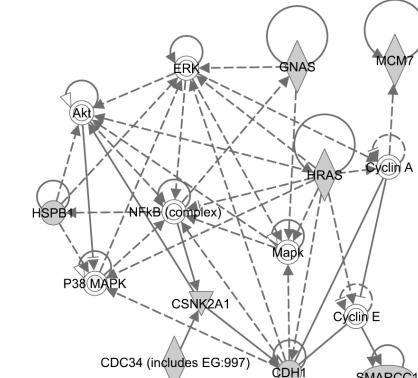
- An Example Subspace Differential Coexpression Pattern from lung cancer dataset

Three lung cancer datasets [Bhattacharjee et al. 2001], [Stearman et al. 2005], [Su et al. 2007]



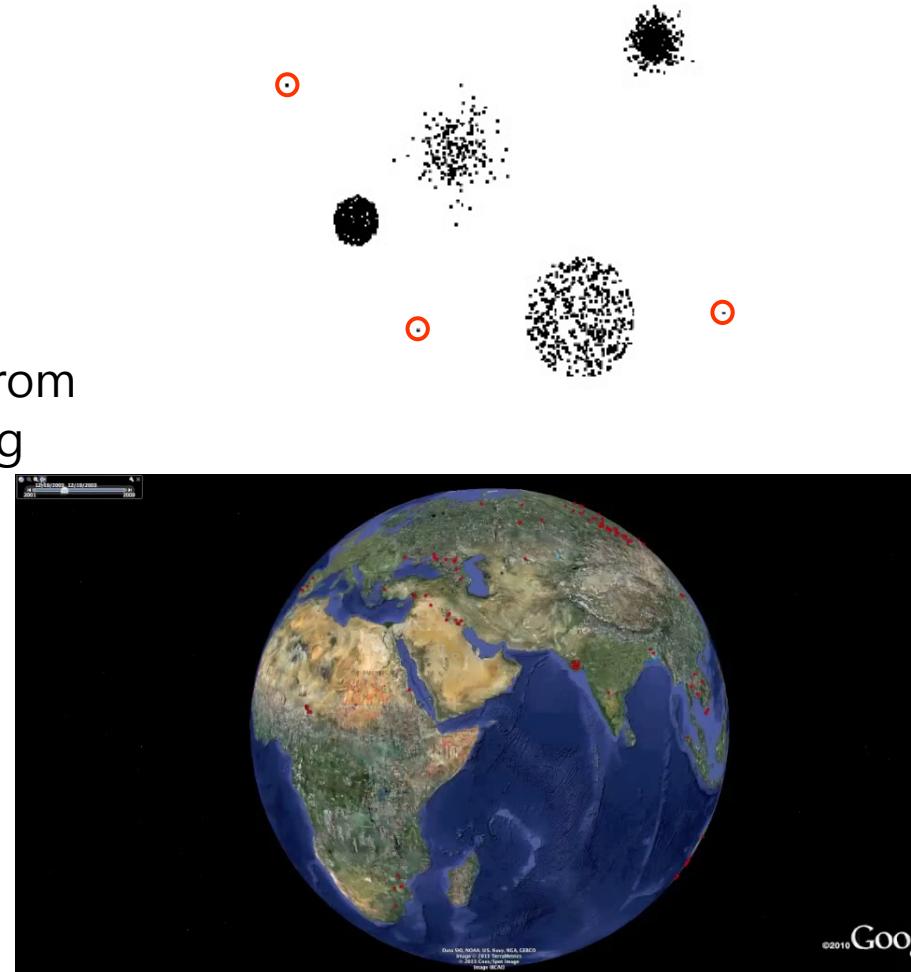
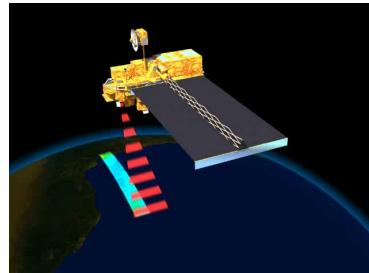
Enriched with the TNF/NFB signaling pathway
which is well-known to be related to lung cancer
P-value: 1.4×10^{-5} (6/10 overlap with the pathway)

[Fang et al PSB 2010]



Deviation/Anomaly/Change Detection

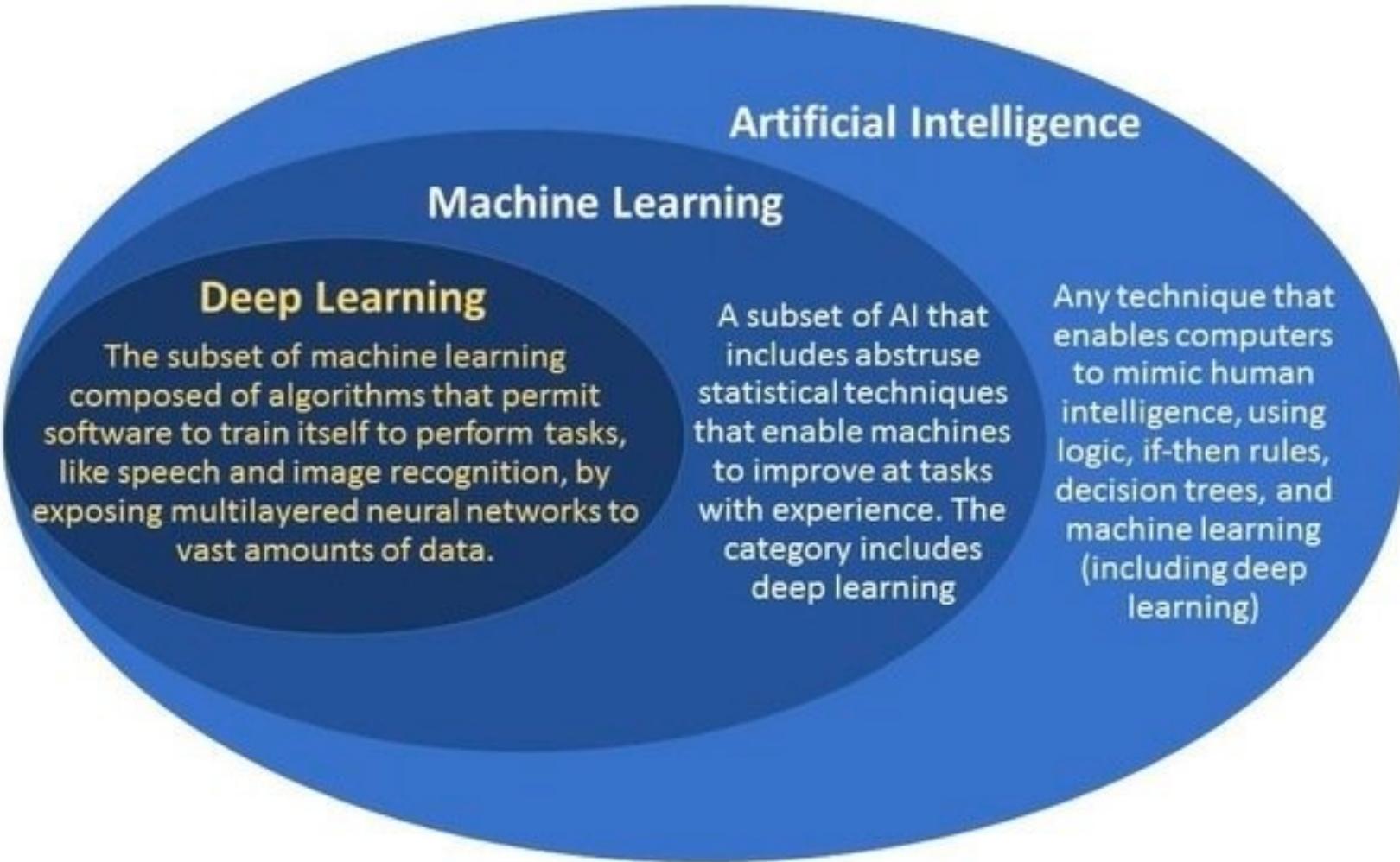
- Detect significant deviations from normal behavior
- Applications:
 - Credit Card Fraud Detection
 - Network Intrusion Detection
 - Identify anomalous behavior from sensor networks for monitoring and surveillance.
 - Detecting changes in the global forest cover.



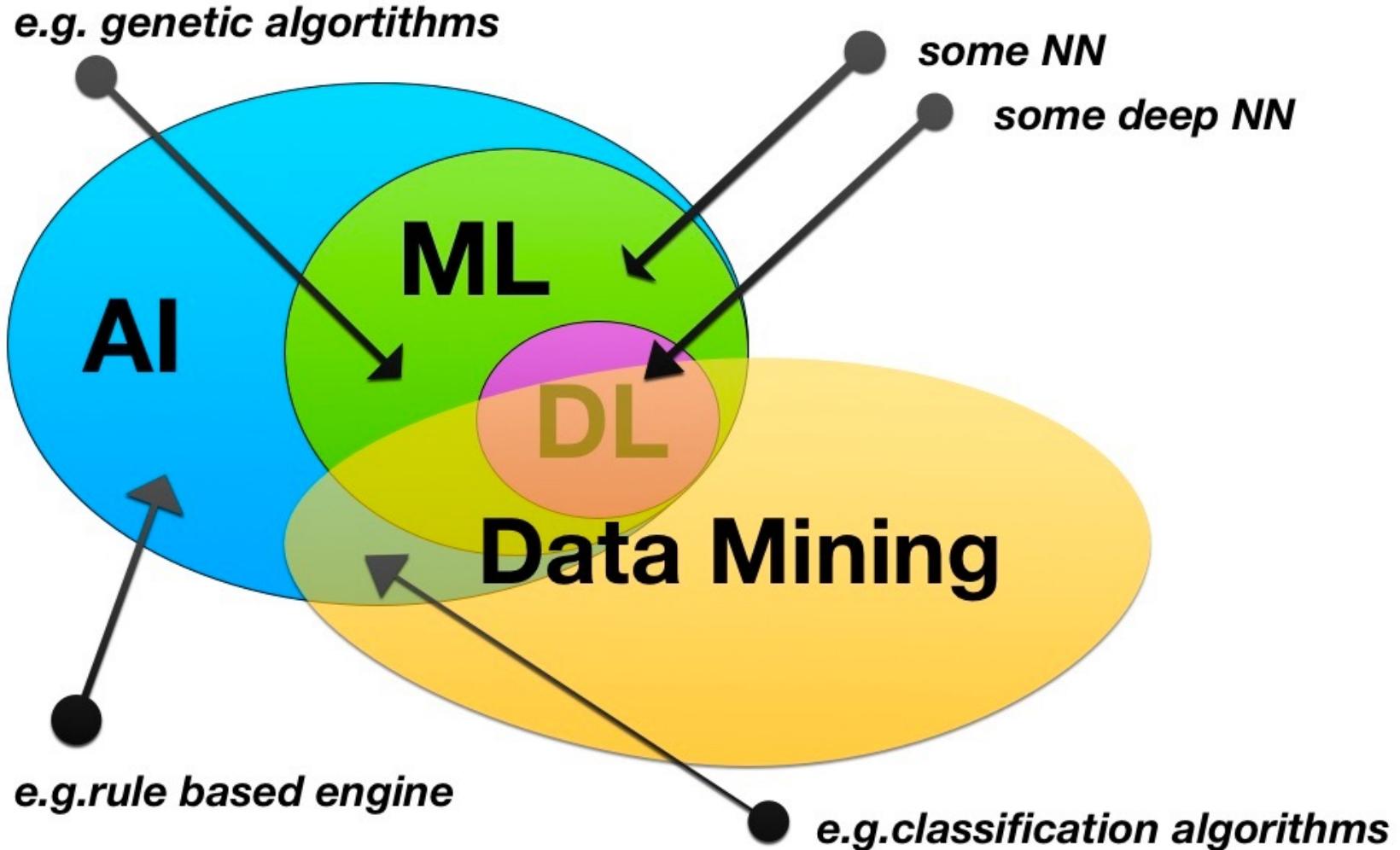
Motivating Challenges

- Scalability
- High Dimensionality
- Heterogeneous and Complex Data
- Data Ownership and Distribution
- Non-traditional Analysis

AI, Machine Learning, Deep Learning



... and Data Mining ...



ACKNOWLEDGMENTS

- Some of the slides and examples are provided by the authors of the books "**Data Mining: Concepts and Techniques**" by J. Han et al. and "**Introduction to Data Mining,**" by P. Tan et al. which were modified into their current form.

END