

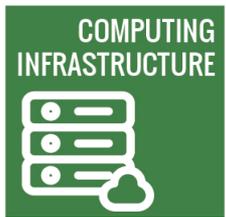
Data Mining: Introduction

Lecture Notes for Chapter 1

Introduction to Data Mining, 2nd Edition

by

Tan, Steinbach, Karpatne, Kumar



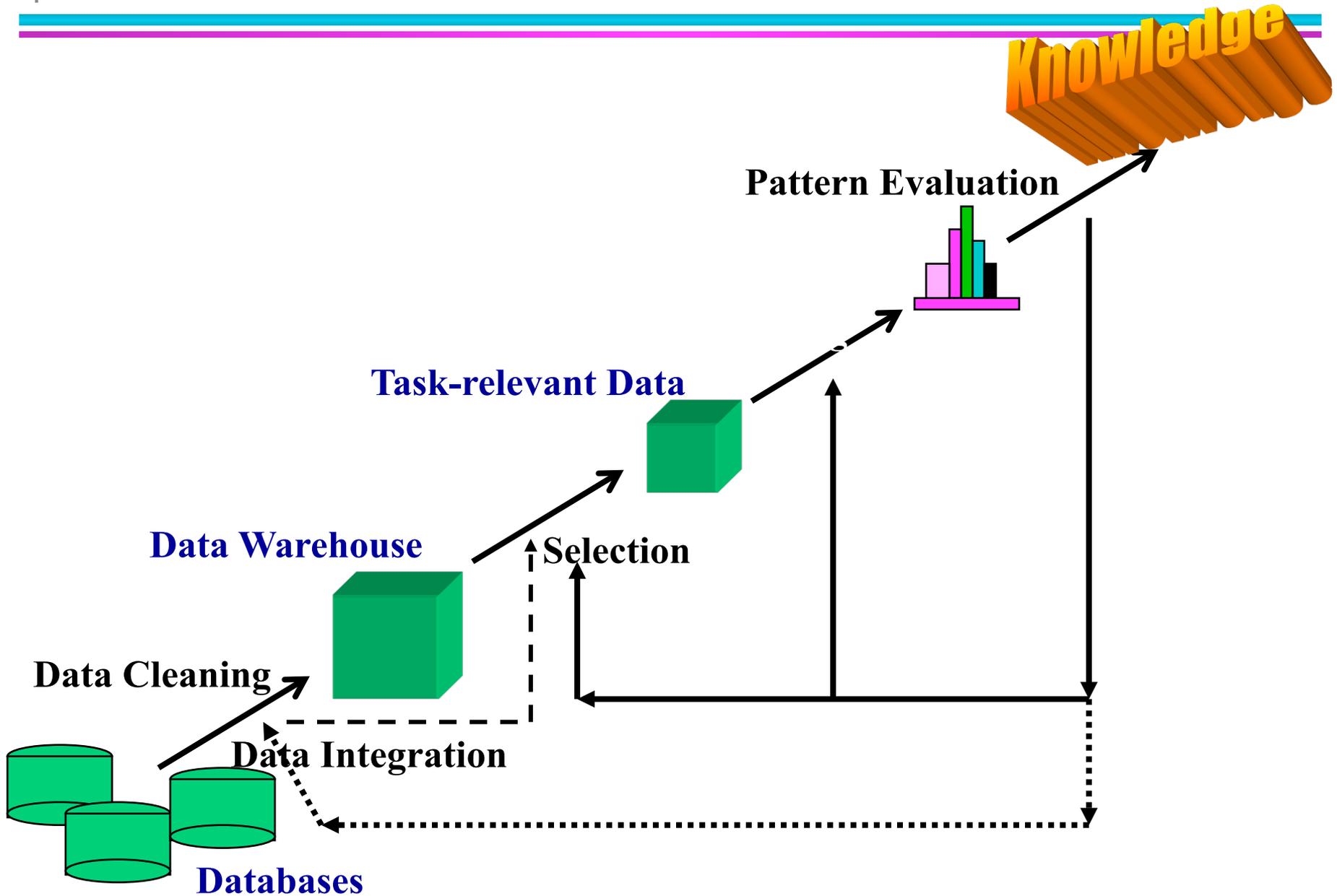
What Is Data Mining?

Data mining (knowledge discovery from data)

Data mining is the use of **efficient** techniques for the analysis of **very large collections of data** and the **extraction** of useful and possibly unexpected patterns in data (**hidden knowledge**).

The KDD Process

4



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.



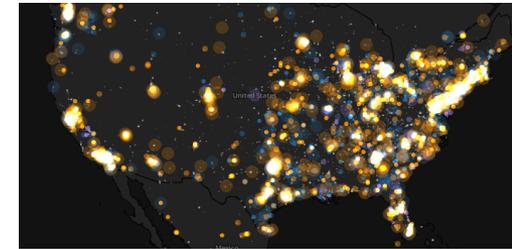
Cyber Security



E-Commerce



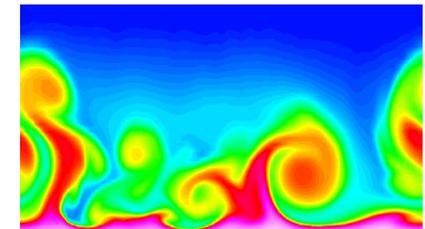
Traffic Patterns



Social Networking: Twitter



Sensor Networks



Computational Simulations

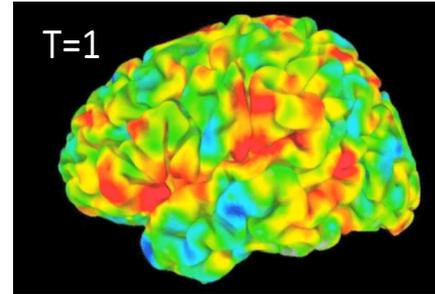
Why Data Mining? Commercial Viewpoint

- 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)

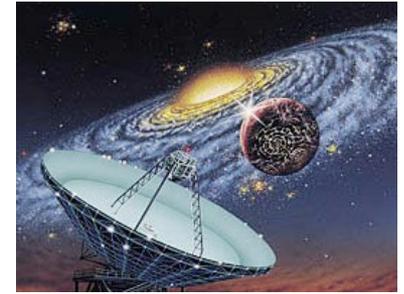


Why Data Mining? Scientific Viewpoint

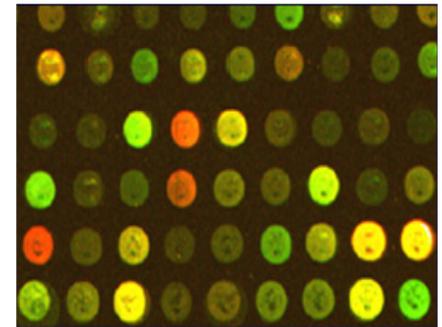
- Data collected and stored at enormous speeds
 - remote sensors on a satellite
 - ◆ NASA EOSDIS archives over petabytes of earth science data / year
 - telescopes scanning the skies
 - ◆ Sky survey data
 - High-throughput biological data
 - scientific simulations
 - ◆ terabytes of data generated in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



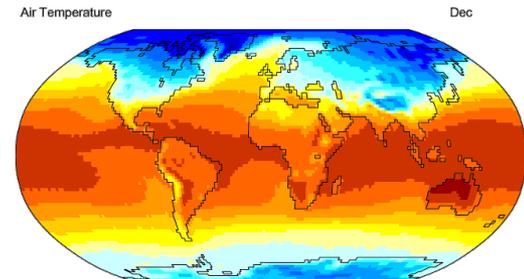
fMRI Data from Brain



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

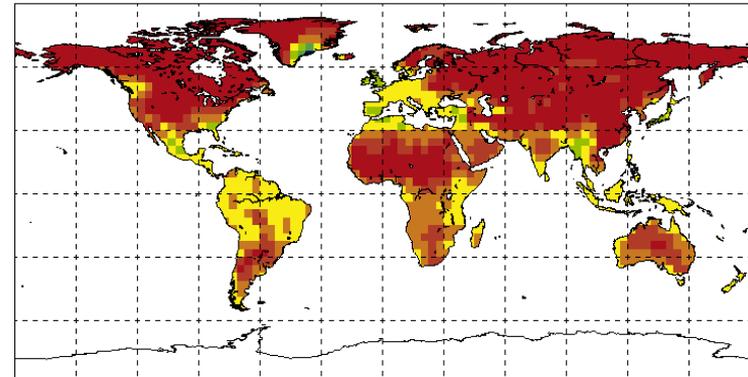
- \$300 billion** potential annual value to US health care—more than double the total annual health care spending in Spain
- €250 billion** potential annual value to Europe's public sector administration—more than GDP of Greece
- \$600 billion** potential annual consumer surplus from using personal location 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



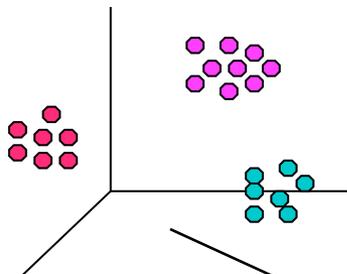
Reducing hunger and poverty by increasing agriculture production

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 ...



Clustering

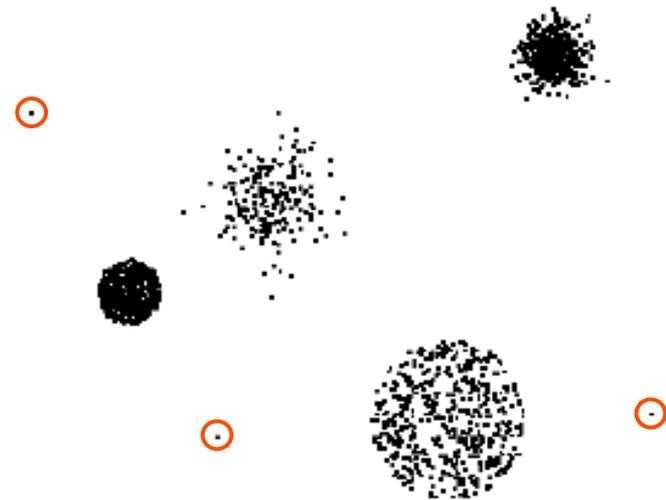
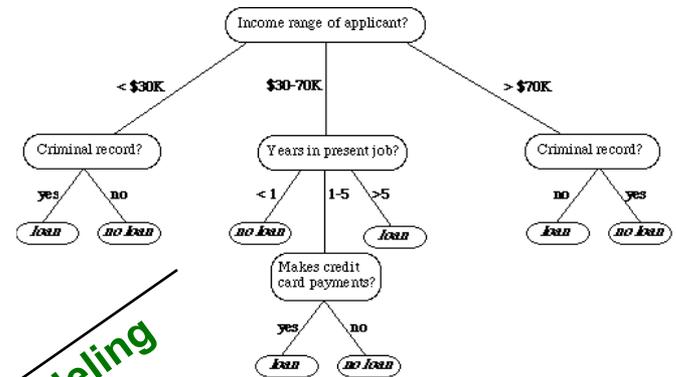
Data

| 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



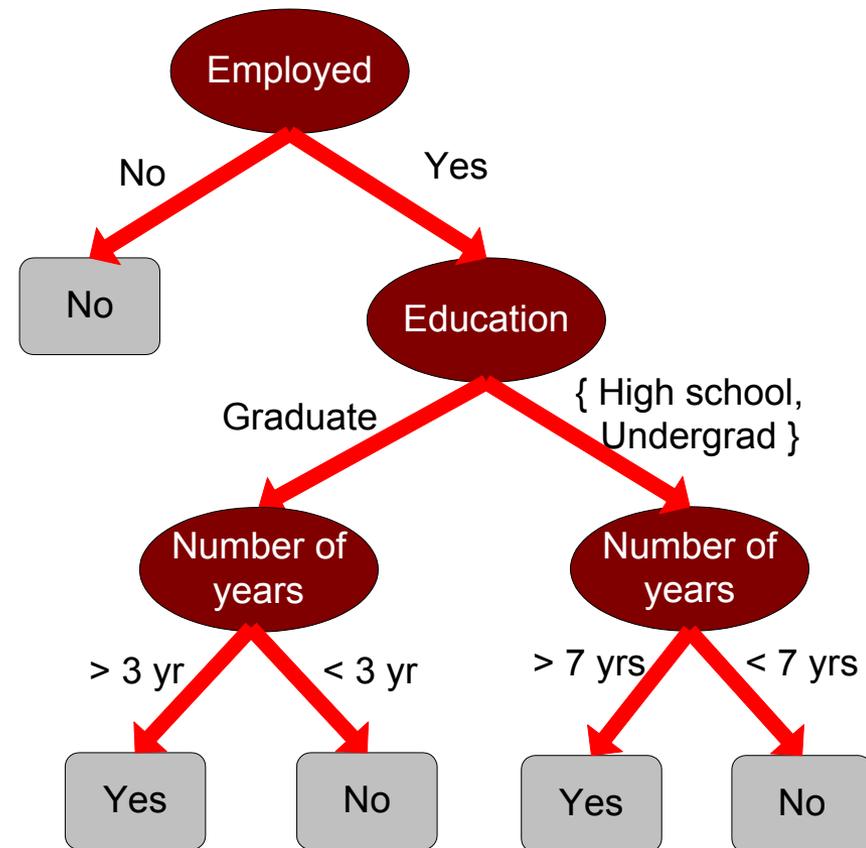
Predictive Modeling: Classification

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

Class

| <i>Tid</i> | 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

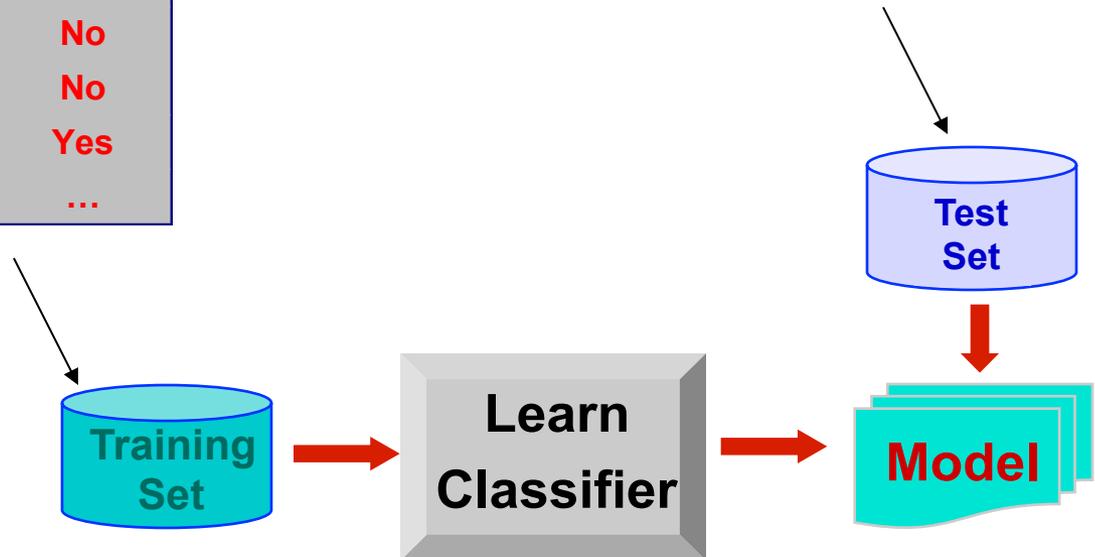


Classification Example

categorical categorical quantitative class

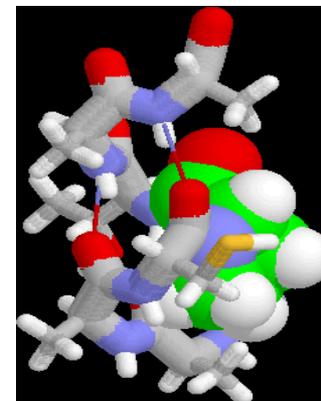
| <i>Tid</i> | Employed | Level of Education | # years at present address | Credit Worthy |
|------------|----------|--------------------|----------------------------|---------------|
| 1 | Yes | Graduate | 5 | Yes |
| 2 | Yes | High School | 2 | No |
| 3 | No | Undergrad | 1 | No |
| 4 | Yes | High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

| <i>Tid</i> | Employed | Level of Education | # years at present address | Credit Worthy |
|------------|----------|--------------------|----------------------------|---------------|
| 1 | Yes | Undergrad | 7 | ? |
| 2 | No | Graduate | 3 | ? |
| 3 | Yes | High School | 2 | ? |
| ... | ... | ... | ... | ... |

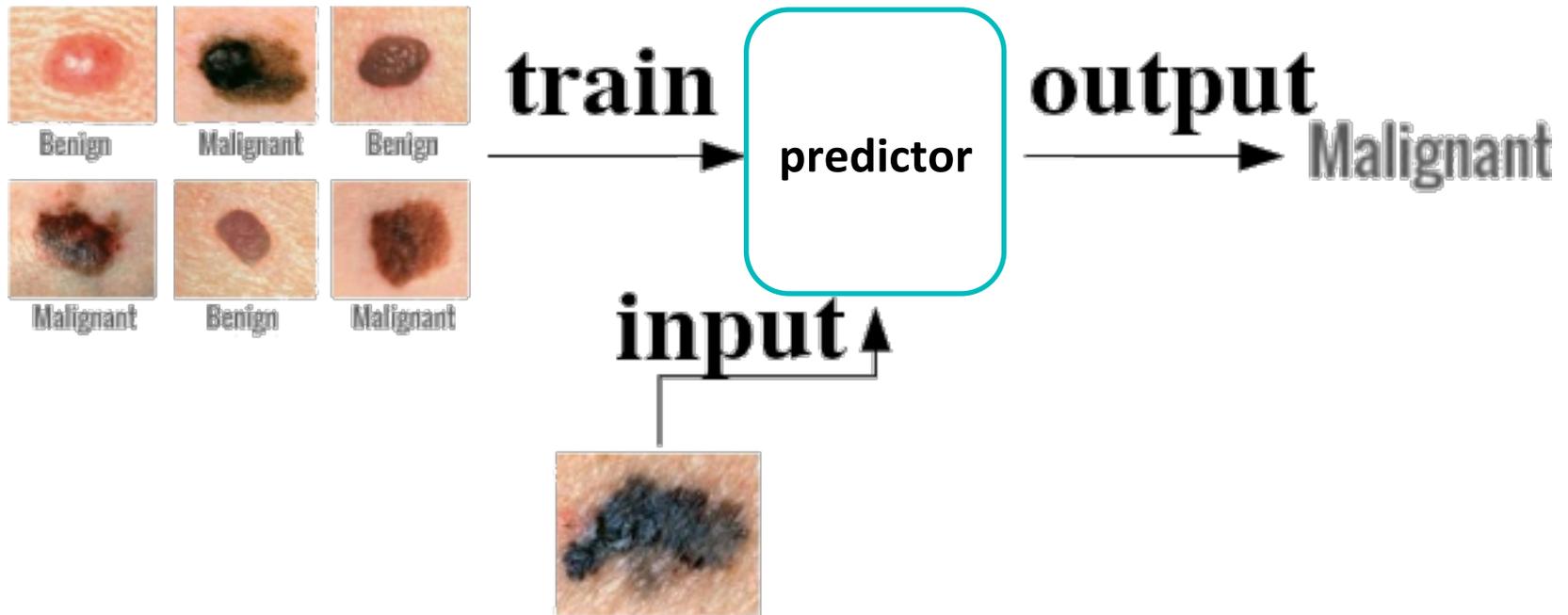


Examples of Classification Task

- 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



AI = Machine Learning + Big Data



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.

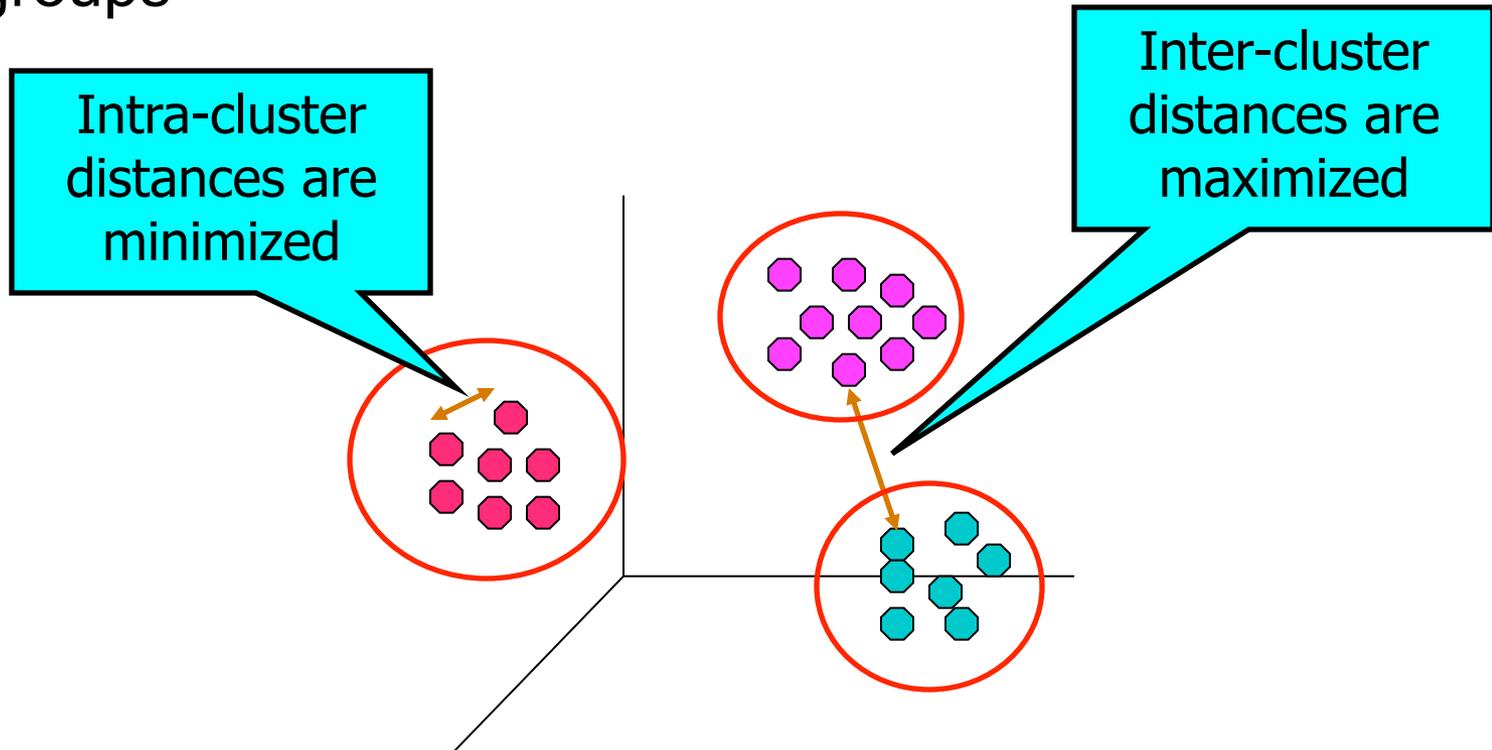
Classification: Application 2

- 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

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



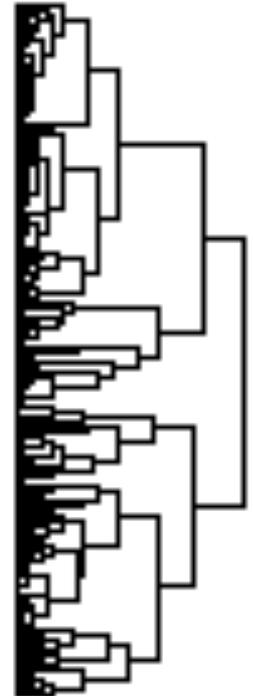
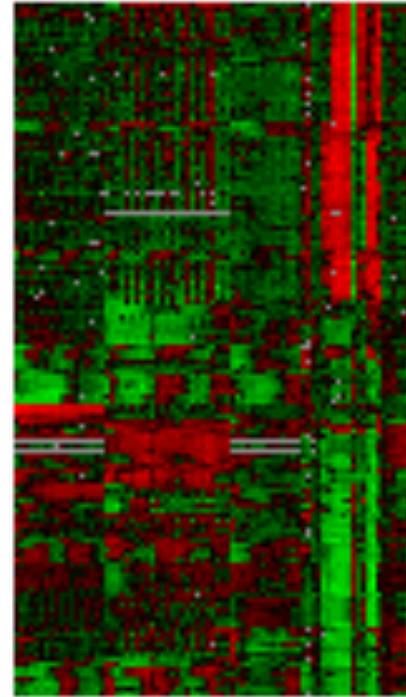
Applications of Cluster Analysis

● 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

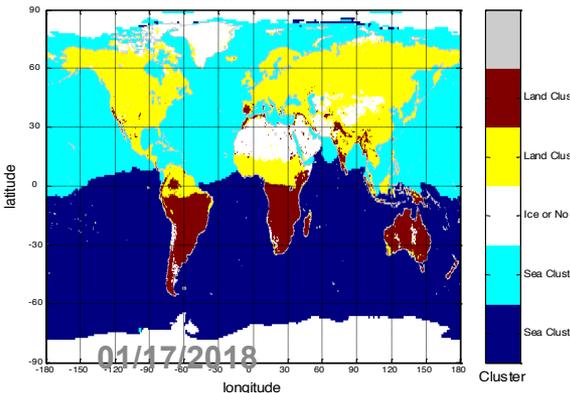
● Summarization

- Reduce the size of large data sets



Courtesy: Michael Eisen

Clusters for Raw SST and Raw NPP



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

A screenshot of a Google News search results page. The page shows the Google logo, search bar, and navigation links. The main content area displays a list of top stories, including "Four more SARS deaths in Hong Kong" and "Bechtel Gets Huge Contract for Iraq Work". The page is dated 01/17/2018.

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.

A Behavior Based Segmentation

Using unsupervised clustering segmentation for a grocery chain which would like better product assortment for its high profitable customers

Potential Inputs

Value

- Basket Size
- Visit Frequency

Basket

- Spend by category
- Type of category
- Brand spend (i.e. private label)

Promotions

- % bought on targeted promotion
- % bought from flyer

Time

- Time of day
- Day of week

Location

- Store format
- Area population density

Clustering approach



Deal Seeking Mom

Key Differentiators



- Full store shop
- High avg. basket size / # trips



- High % purchased on promotion
- Rewards seeker

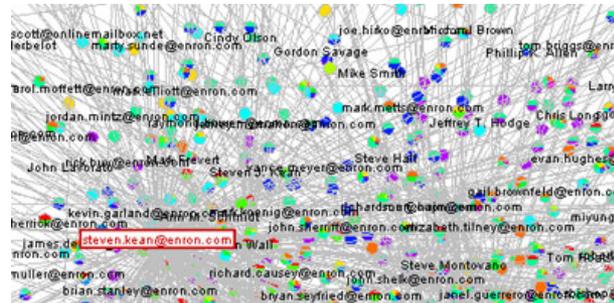


- High spend categories
 - Fresh produce
 - Organic food
 - Multipack juice, snack

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:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

Association Analysis: Applications

- **Market-basket analysis**

- Rules are used for sales promotion, shelf management, and inventory management

- **Telecommunication alarm diagnosis**

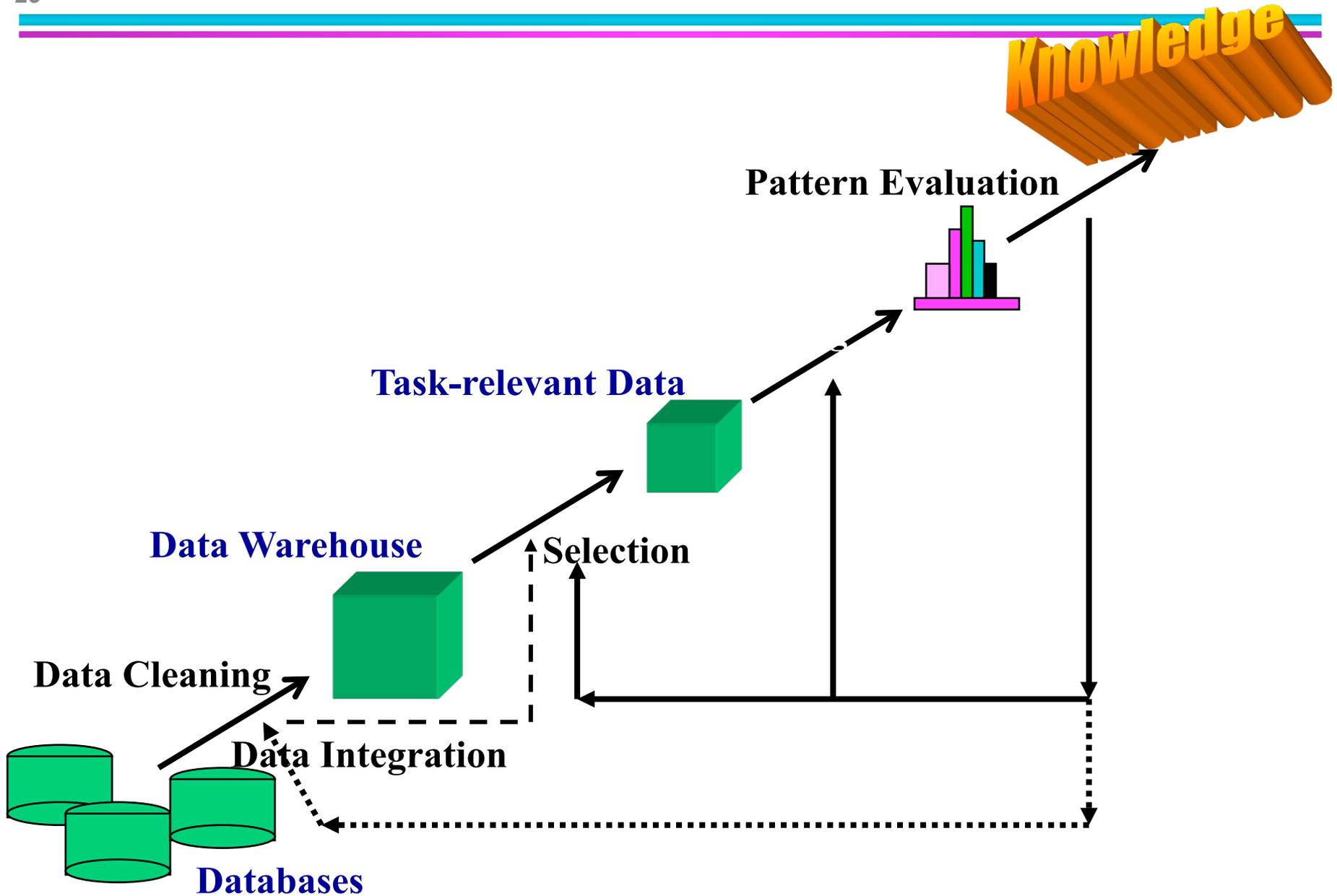
- Rules are used to find combination of alarms that occur together frequently in the same time period

- **Medical Informatics**

- Rules are used to find combination of patient symptoms and test results associated with certain diseases

The KDD Process

25





DATA

What is Data?

- Collection of **data objects** and their **attributes**
- An **attribute** is a property or characteristic of an object
 - Examples: eye color of a person, temperature, etc.
 - Attribute is also known as variable, field, characteristic, dimension, or feature
- A collection of attributes describe an **object**
 - Object is also known as record, point, case, sample, entity, or instance

Attributes

| <i>Tid</i> | 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 |

Objects

Types of data sets

- Record
 - Data Matrix
 - Document Data
 - Transaction Data
- Graph
 - World Wide Web
 - Molecular Structures
- Ordered
 - Spatial Data
 - Temporal Data
 - Sequential Data
 - Genetic Sequence Data

Data Matrix

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute
- Such data set can be represented by an m by n matrix, where there are m rows, one for each object, and n columns, one for each attribute

| Projection of x Load | Projection of y load | Distance | Load | Thickness |
|-----------------------------|-----------------------------|-----------------|-------------|------------------|
| 10.23 | 5.27 | 15.22 | 2.7 | 1.2 |
| 12.65 | 6.25 | 16.22 | 2.2 | 1.1 |

Document Data

- Each document becomes a 'term' vector
 - Each term is a component (attribute) of the vector
 - The value of each component is the number of times the corresponding term occurs in the document.

| | team | coach | play | ball | score | game | win | lost | timeout | season |
|------------|------|-------|------|------|-------|------|-----|------|---------|--------|
| Document 1 | 3 | 0 | 5 | 0 | 2 | 6 | 0 | 2 | 0 | 2 |
| Document 2 | 0 | 7 | 0 | 2 | 1 | 0 | 0 | 3 | 0 | 0 |
| Document 3 | 0 | 1 | 0 | 0 | 1 | 2 | 2 | 0 | 3 | 0 |

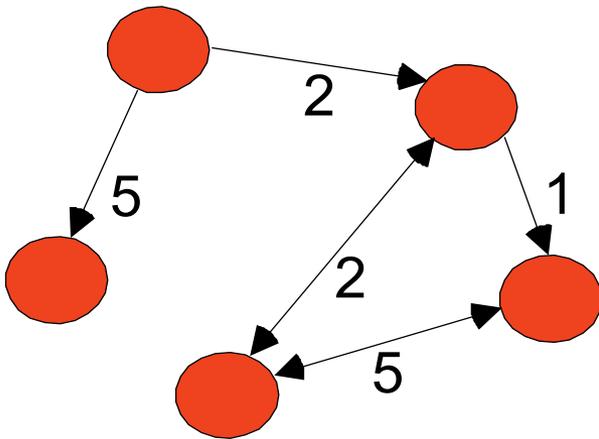
Transaction Data

- A special type of record data, where
 - Each record (transaction) involves a set of items.
 - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the 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 |

Graph Data

- Examples: Generic graph, a molecule, and webpages



Useful Links:

- [Bibliography](#)
- Other Useful Web sites
 - [ACM SIGKDD](#)
 - [KDnuggets](#)
 - [The Data Mine](#)

Knowledge Discovery and Data Mining Bibliography

(Gets updated frequently, so visit often!)

- [Books](#)
- [General Data Mining](#)

Book References in Data Mining and Knowledge Discovery

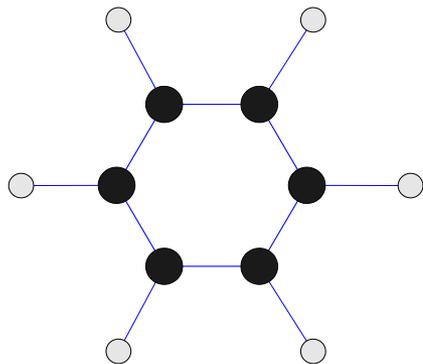
Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy uthurasamy, "Advances in Knowledge Discovery and Data Mining", AAAI Press/the MIT Press, 1996.

J. Ross Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993.
Michael Berry and Gordon Linoff, "Data Mining Techniques (For Marketing, Sales, and Customer Support)", John Wiley & Sons, 1997.

General Data Mining

Usama Fayyad, "Mining Databases: Towards Algorithms for Knowledge Discovery", Bulletin of the IEEE Computer Society Technical Committee on data Engineering, vol. 21, no. 1, March 1998.

Christopher Matheus, Philip Chan, and Gregory Piatetsky-Shapiro, "Systems for knowledge Discovery in databases", IEEE Transactions on Knowledge and Data Engineering, 5(6):903-913, December 1993.

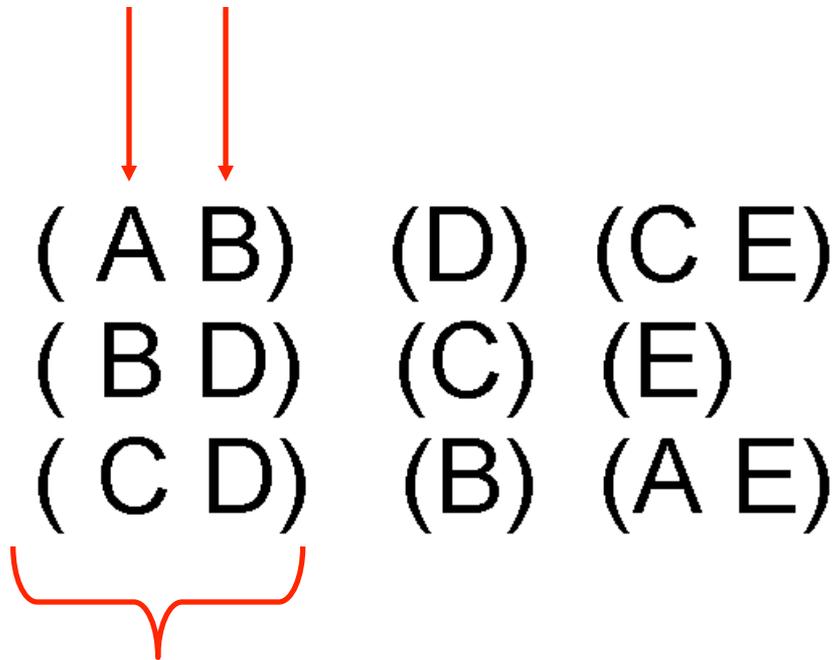


Benzene Molecule: C6H6

Ordered Data

- Sequences of transactions

Items/Events



**An element of
the sequence**

Ordered Data

- Genomic sequence data

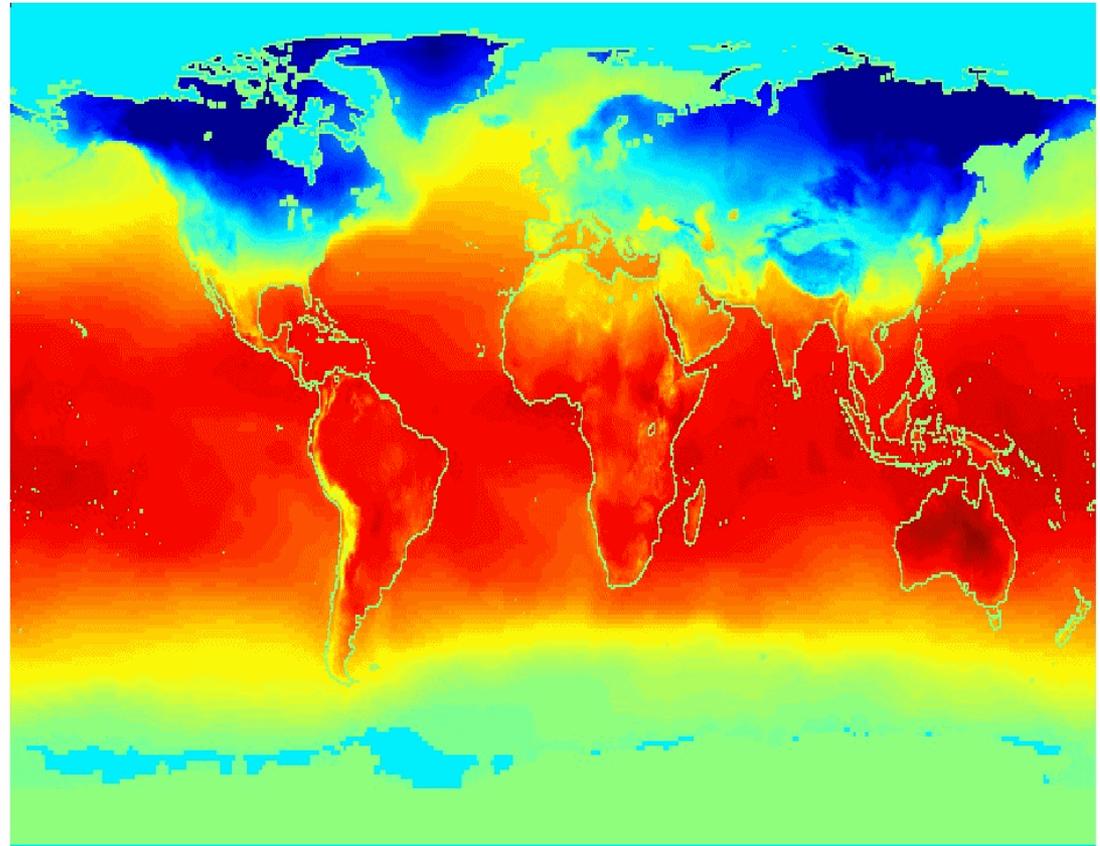
**GGTTC CGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG**

Ordered Data

- Spatio-Temporal Data

**Average Monthly
Temperature of
land and ocean**

Jan



Data Quality

- Poor data quality negatively affects many data processing efforts

“The most important point is that poor data quality is an unfolding disaster.

- Poor data quality costs the typical company at least ten percent (10%) of revenue; twenty percent (20%) is probably a better estimate.”

Thomas C. Redman, DM Review, August 2004

- Data mining example: a classification model for detecting people who are loan risks is built using poor data
 - Some credit-worthy candidates are denied loans
 - More loans are given to individuals that default

Data Quality ...

- What kinds of data quality problems?
- How can we detect problems with the data?
- What can we do about these problems?

- Examples of data quality problems:
 - Noise and outliers
 - Missing values
 - Duplicate data
 - Wrong data