COMP4331: Introduction to Data Mining

James Kwok

Department of Computer Science and Engineering
Hong Kong University of Science and Technology

Why Data Mining?

We are experiencing an explosive growth of data

- 5 exabytes (10^{18} bytes) of data were created by human until 2003 (kilo, mega, giga, tera; peta; exa; zetta); today this amount of information is created in two days
- in 2012, digital world of data was expanded to 2.72 zettabytes
- it is predicted to double every two years
- IBM indicates that every day 2.5 exabytes of data is created

Major Sources of Abundant Data

- society: news, social networking (Facebook, YouTube)
- business: web, e-commerce, transactions, stocks
- science: sensors, bioinformatics, scientific simulations

Example

- Facebook has 955 million monthly active accounts using 70 languages, 140 billion photos uploaded, 125 billion friend connections, every day 30 billion pieces of content and 2.7 billion likes and comments have been posted
- Every minute, 48 hours of video are uploaded and every day, 4 billion views performed on YouTube
- 1 billion Tweets every 72 hours from more than 140 million active users on Twitter
- 571 new websites are created every minute of the day
- every day 10 billion text messages are sent
- by the year 2020, 50 billion devices will be connected to networks and the internet

Why Data Mining?

We are drowning in data, but starving for knowledge

- while data size and complexity rapidly increase, the number of data analysts remains relatively small

Example

Within the next decade, number of information will increase by 50 times; however number of information technology specialists who keep up with all that data will increase by 1.5 time

- traditional techniques are simply inapplicable
- we need to find efficient ways to analyze the vast quantities of raw data to extract knowledge

Some Motivating Examples

- Business: A book store wishes to make recommendations to customers based on other customers' previous purchases
- Science: A bioinformatics lab wishes to find DNA similarities among different organisms
- Society: Either a company (for marketing purposes) or a lab (for research purposes) wishes to identify the most influential users in a social network

What is Data Mining?

- extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
- exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns

What is Data Mining?...

Is everything "Data Mining"?

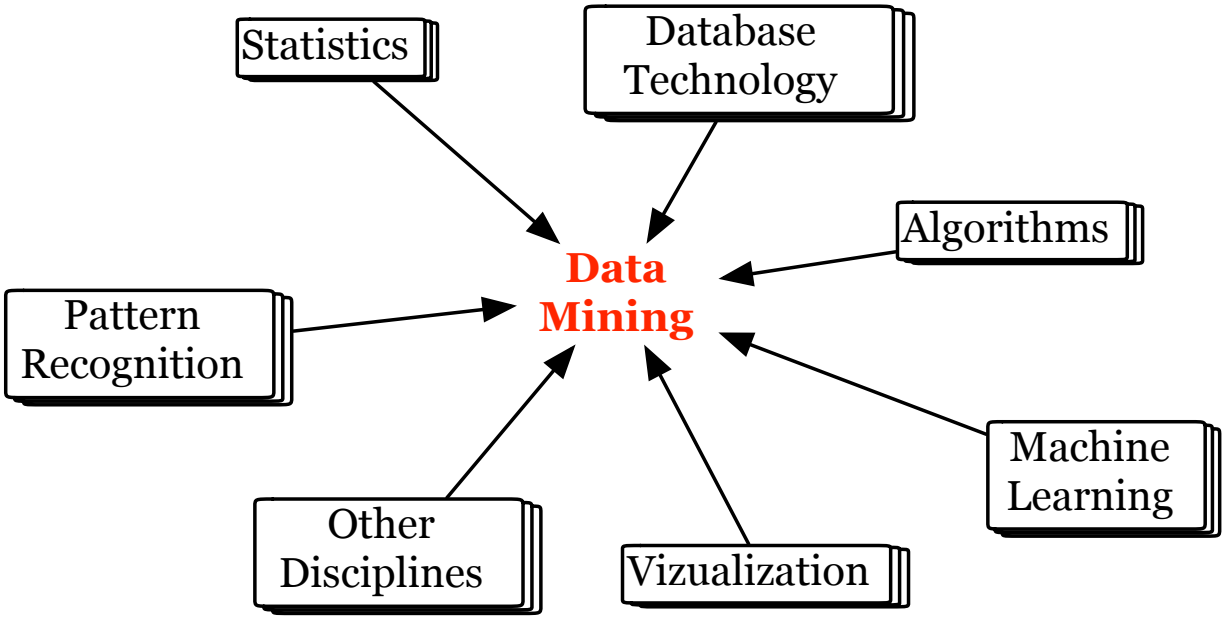
- simple search or query processing should not be confused with data mining

What is Not Data Mining?

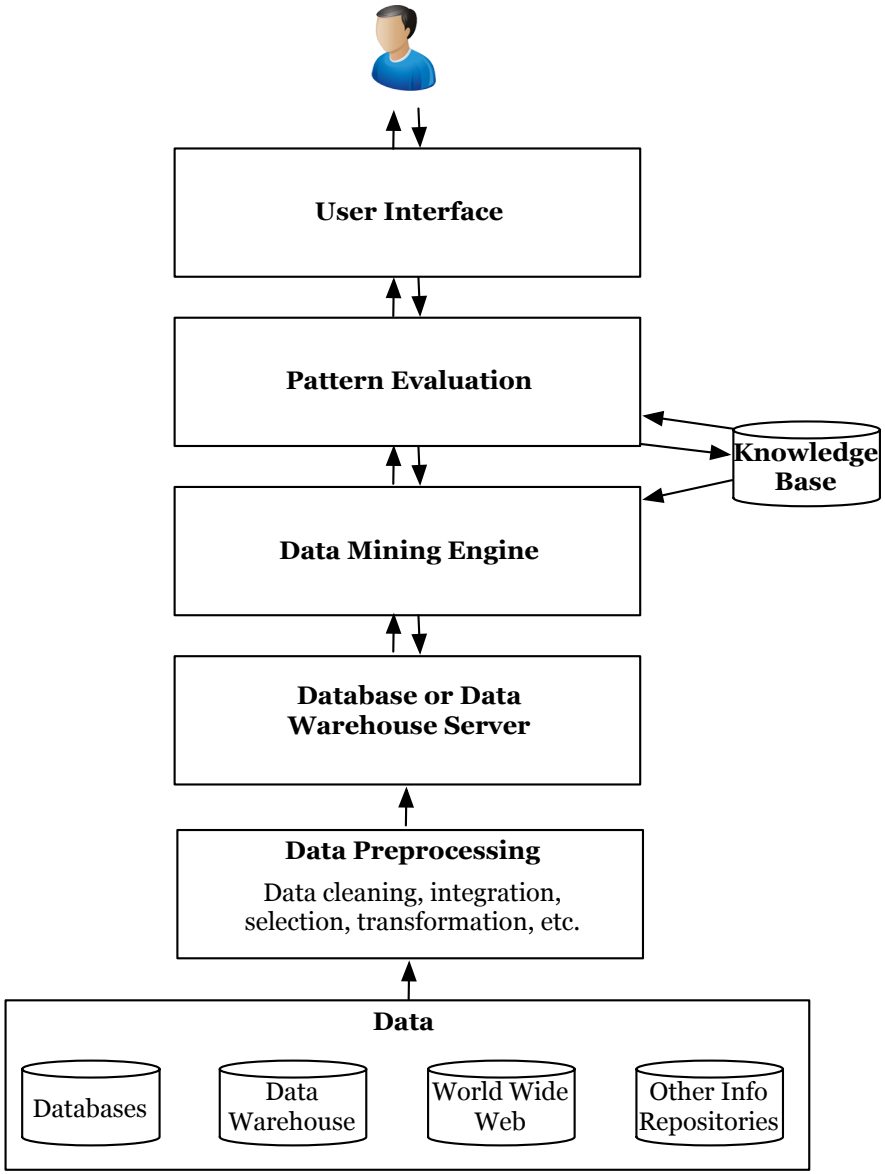
- look up phone number in a phone directory
- query a web search engine for information about "Amazon"

What is Data Mining?...

Data mining is a confluence of several disciplines



Data Mining Architecture



Architecture of a typical data mining system

On What Kind of Data?

Various data repositories

- relational data
- data warehouses
- transactional data
- graph data
- sequence data
- time series
- spatial data
- text & multimedia data

Relational Data

- a relational database consists of a set of tables, each of which consisting of a set of attributes (or columns or fields), and containing a large set of records (or tuples or rows)

Example

Employee

EID	***Name***	***Address***	***Position***	***Salary***
0023	A. Smith	122 Lake Ave., Chicago, IL	Manager	200,000$
...	...	...	...	...

Branch

BID	***Name***	***Address***
005	City Square	356 Michigan Ave., Chicago, IL
...	...	...

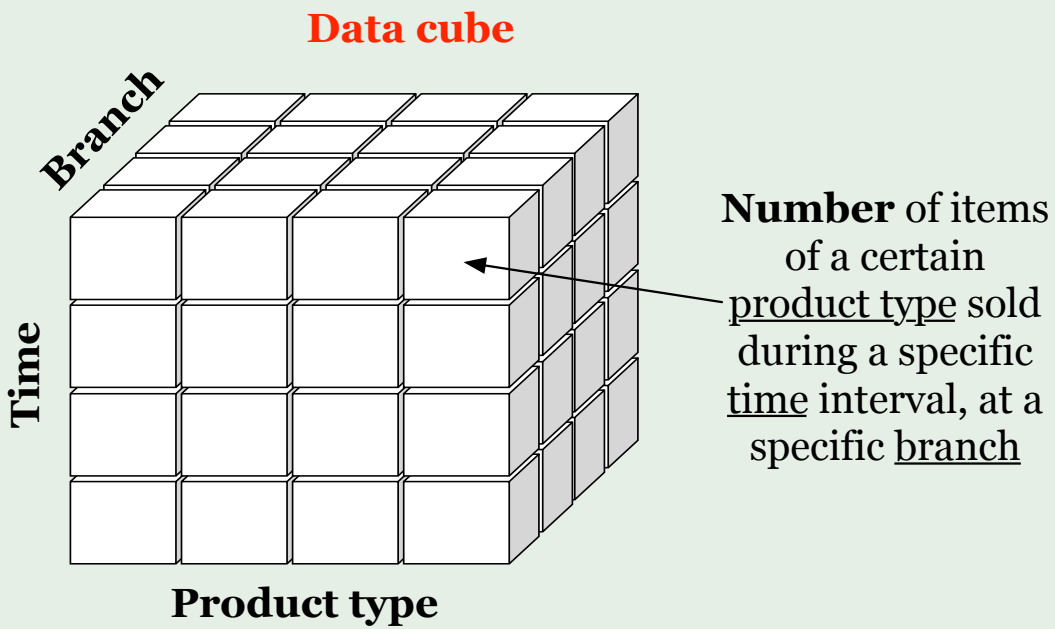
Works_At

EID	***BID***
0023	005
...	...

Data Warehouse

- a data repository of information collected from different sources stored under a unified scheme, and it usually resides at a single site
- the stored data provide information from a historical perspective and they are usually summarized
- the physical structure is typically a multidimensional data cube

Example



Transactional Data

- a special type of relational data, where every record is a transaction and involves a set of items

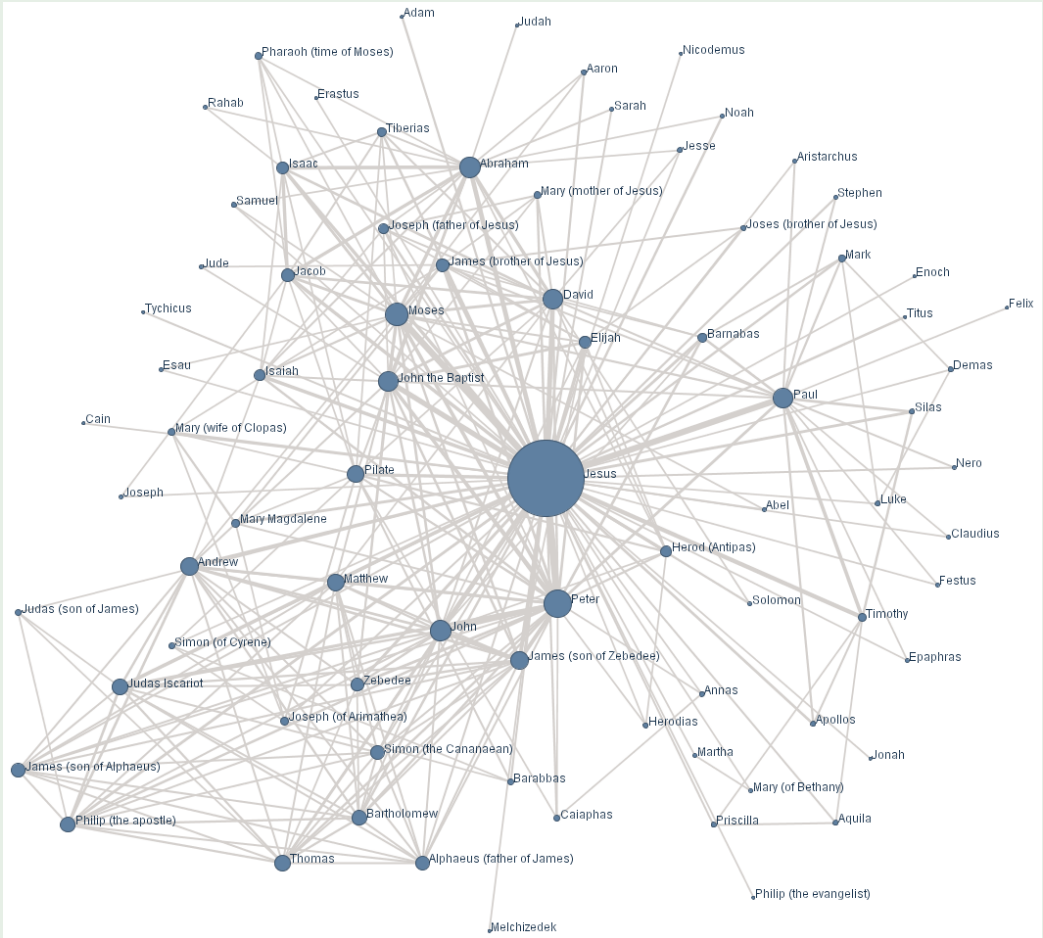
Example (a set of transactions)

TID	*Items*
1	Bread, Butter, Milk, Cereal
2	Beer, Coke
3	Bread, Diaper, Milk, Cereal
4	Beer, Diaper
5	Coke, Bisquits, Milk

Graph Data

- capture relationships among objects

Example (social network)



Graph Data...

Social network analysis

- mainly motivated by the rapid proliferation of social networking (Facebook, YouTube, etc.)

Example (possible tasks)

- discover social communities using similarity metrics
- model the strength of a social link based on the interaction between the users
- identify the most influential users in the social network (for viral marketing purposes)

Sequence Data

- ordered sequences of events with or without a concrete notion of time

Example (social network)

Genomic Sequence Data

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

Time Series Data

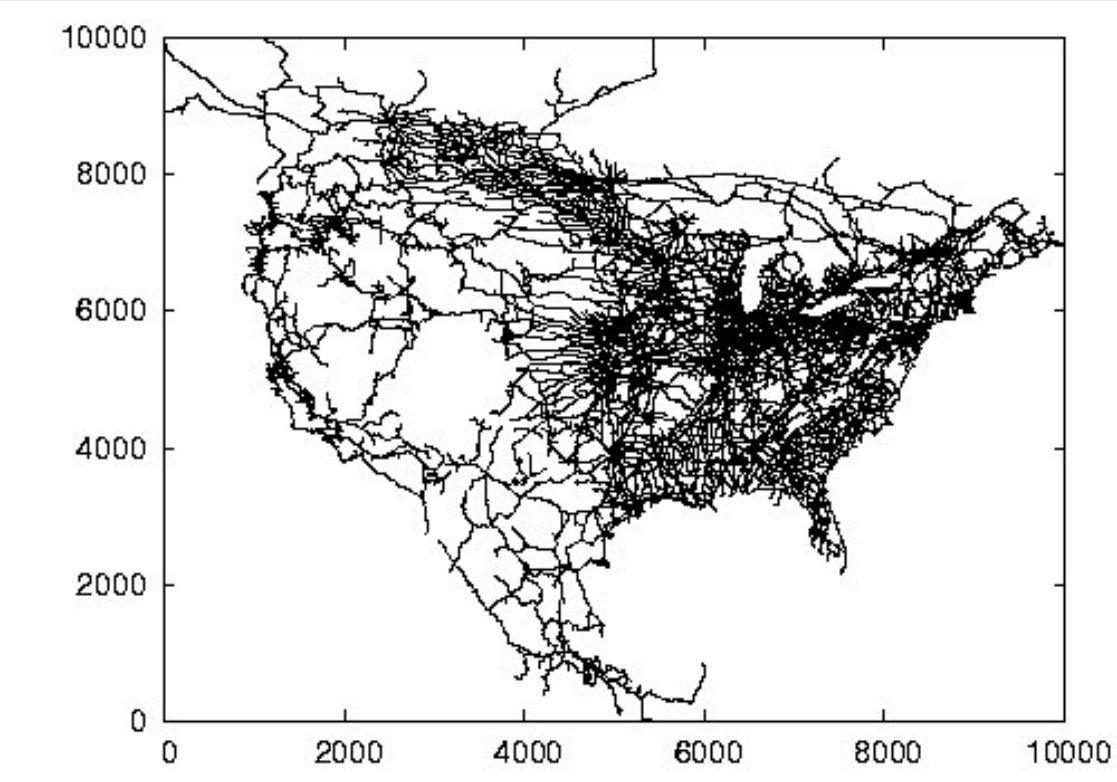
- a special type of sequence data, where the values or events are obtained over repeated measurements of time (e.g., hourly, daily, weekly)

Example (Apple vs. Google Stock)

Spatial Data

- contain geographical attributes (such as spatial coordinates or areas)

Example (Road Network of North America (NA))



Text & Multimedia Data

- text databases contain word descriptions for objects
- multimedia databases store image, audio, and video data

Example

Document Term Vectors

	'Team'	'Coach'	'Timeout'
Doc#1	3	0	0
Doc#2	0	7	0

Image

Data Mining Functionalities

Major data mining tasks

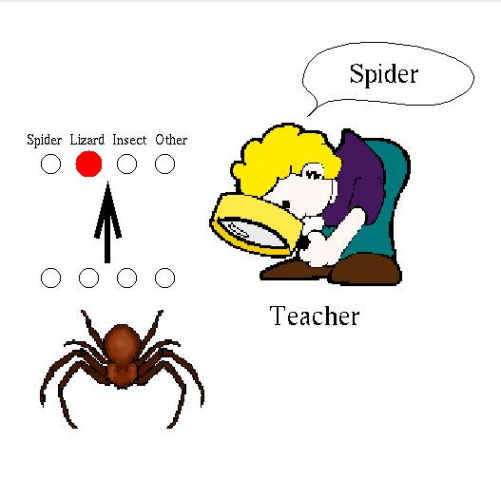
- Classification and regression
- Cluster analysis
- Association analysis

Classification and Regression

Classification

- we have a set of records called training set
- each record contains various attributes, among which there is a categorical (i.e., discrete) attribute referred to as class

Example



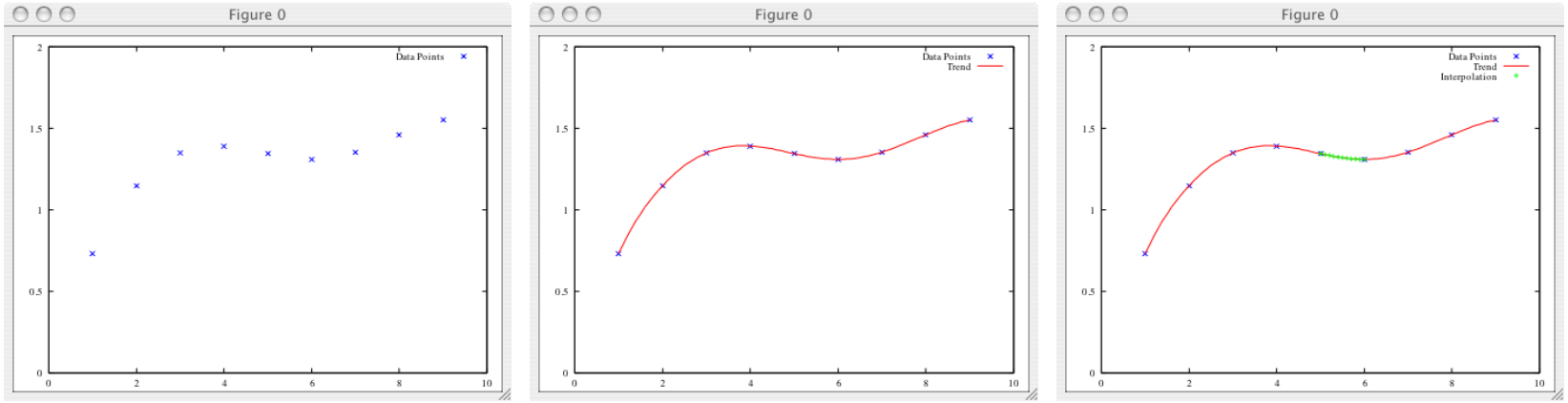
Regression

- classification predicts categorical attribute values; regression predicts numerical attribute values

Classification and Regression...

How to predict the value of a new (i.e., previously unseen) record?

- what about an old record?



- we explore the training set and devise a function called model
- the model takes as input a set of attributes values, and returns a value for the class attribute
- we then predict the class of the new record based on the model

Classification and Regression

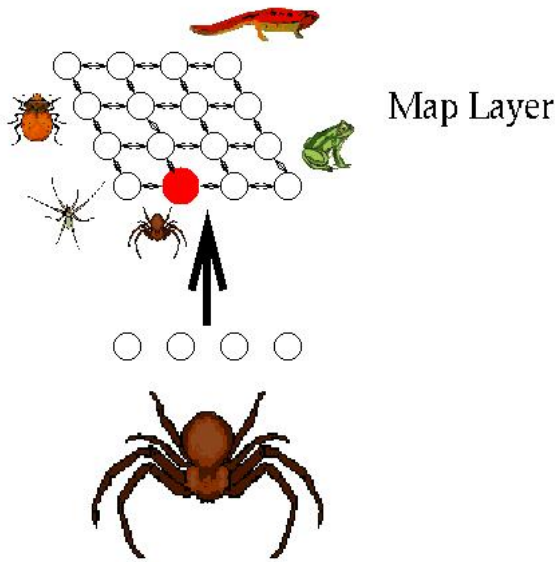
Example (Direct Marketing)

- suppose that we run an electronics consumer store
- we wish to reduce the cost of mailing by targeting only the set of consumers that are likely to buy a new product
- we collect a dataset of consumers that bought a similar product introduced before, as well as various demographic, lifestyle, and other information about them (training set)
- this {buy, don't buy} decision forms the class attribute
- we use the above information to devise a classifier model
- when reviewing the potential mail recipients, we predict if they are likely to buy the new product based on the model

Clustering

Given a set of objects, each having a set of attributes, and a similarity measure among them, find clusters (i.e., groups) such that

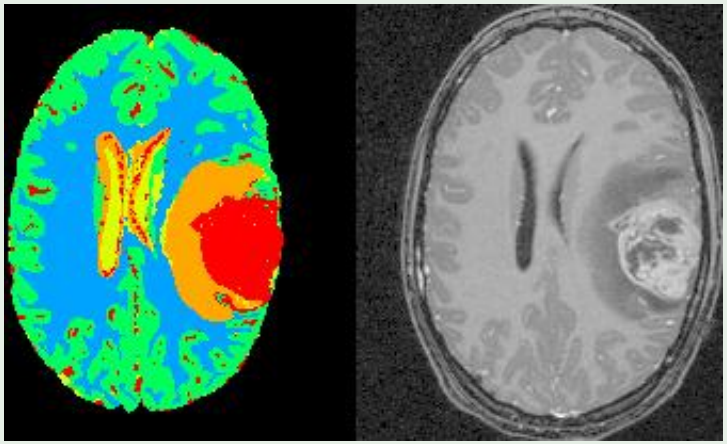
- objects in one cluster are more similar to one another
- objects in separate clusters are less similar to one another

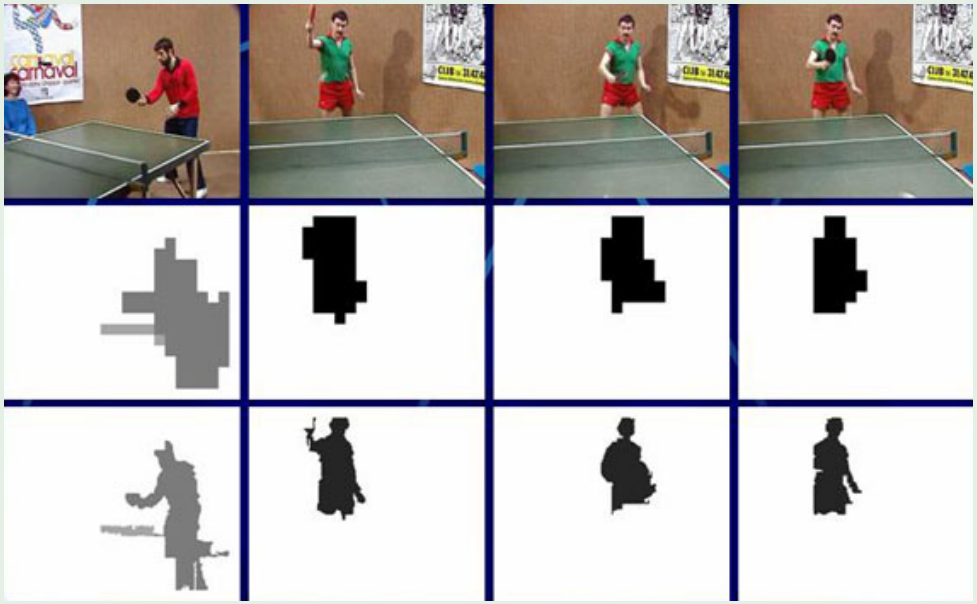


- unlike classification, clustering analyzes objects without consulting a known class label

Clustering...

Example (image segmentation)





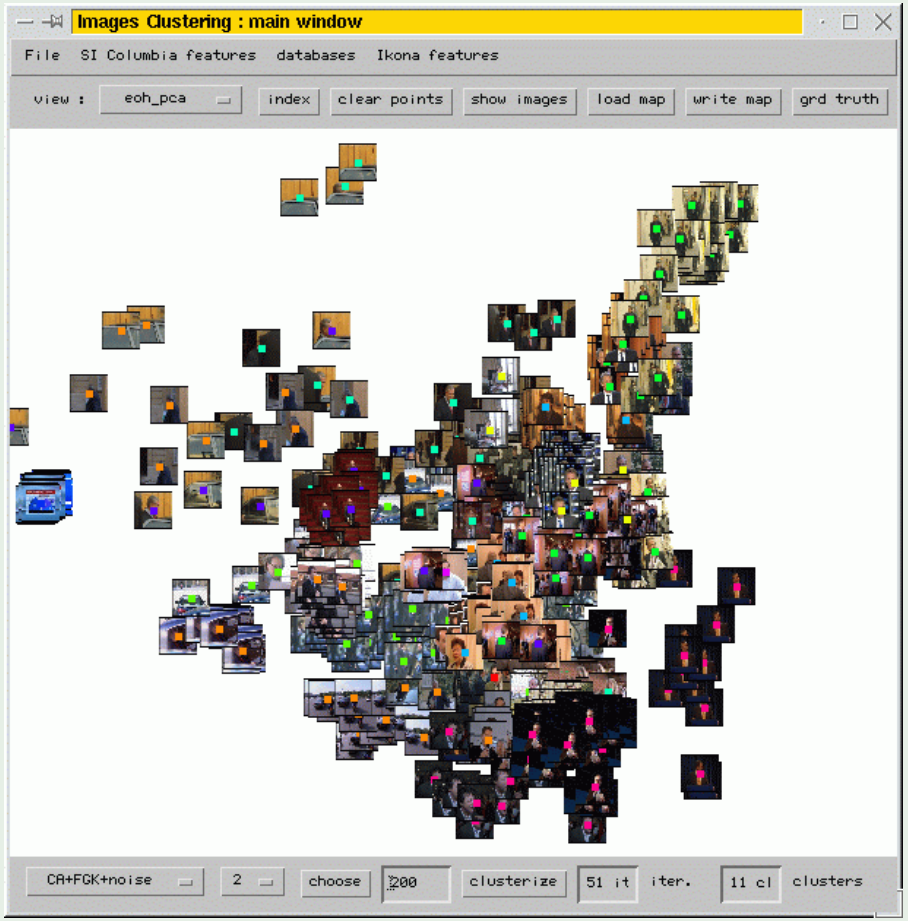
Clustering...

Example

Star Trek IV 0.024 Star Trek II 0.023 Star Trek VI 0.023 Star Trek III 0.021 The Fifth Element 0.018	Dr. Strangelove 0.029 A Clockwork Orange 0.020 Delicatessen 0.018 Cinema Paradiso 0.018 Brazil 0.017717
The Rock 0.553 Eraser 0.232 Independence Day (ID4) 0.089 Mission: Impossible 0.077 Trainspotting 0.021	The Piano 0.288 The Remains of the Day 0.077 In the Name of the Father 0.067 Forrest Gump 0.052 Shadowlands 0.047

Clustering...

Example (content-based image retrieval)



Outlier Detection

- outlier is an object that is "far away" from any cluster
- clustering can be used

Example (Fraud Detection)

- collect old transactions of a credit card holder
- cluster the transactions based on the location and/or the amount of money spent
- detect whether an incoming transaction is considerably dissimilar to all clusters

Association Analysis

Example

TID	Items
1	Bread, Butter, Milk, Cereal
2	Beer, Coke
3	Bread, Diaper, Milk, Cereal
4	Beer, Diaper
5	Coke, Bisquits, Milk

A supermarket wishes to find the products that most frequently co-occur in the customer transactions, in order to strategize effective promotions

Goal

Given a transactional database, find the sets of objects that frequently appear within the same transactions

- also called frequent pattern mining