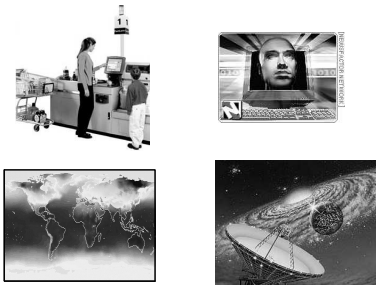
CS522 Advanced Database Systems

Course Overview

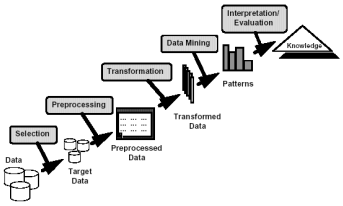
Chengyu Sun
California State University, Los Angeles

Why Data Mining?



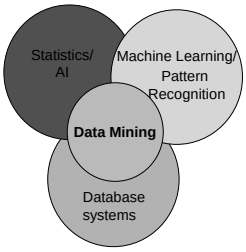
Data Mining

- Extracting knowledge from large amounts of data



Origins of Data Mining

- Traditional techniques may not be suitable due to
  - Enormity of data
  - High dimensionality of data
  - Heterogeneous, distributed nature of the data



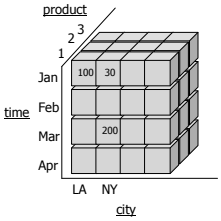
Topics Covered

- Data warehouse and OLAP
- Association rule mining
- Classification and regression
- Clustering

OLAP

- Online Analytic Processing
  - V.s. OLTP

time	city	product	sales
Jan	LA	1	100
Feb	LA	2	50
Jan	NY	1	30
Mar	NY	1	200



Association Rule Mining

- 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.

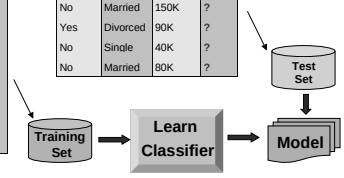
TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:
{Milk} --> {Coke}
{Diaper, Milk} --> {Beer}

Classification

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

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Clustering

- Euclidean Distance Based Clustering in 3-D space.

Intracluster distances are minimized

Intercluster distances are maximized