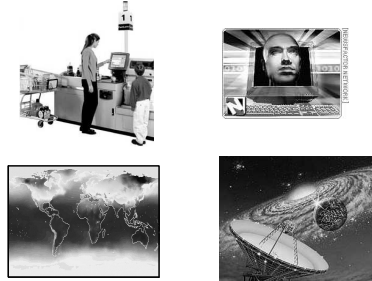


## CS522 Advanced Database Systems Course Overview

Chengyu Sun  
California State University, Los Angeles

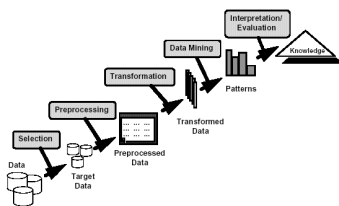
## Why Data Mining?



©Tan, Steinbach, Kumar Introduction to Data Mining 2004

## Data Mining

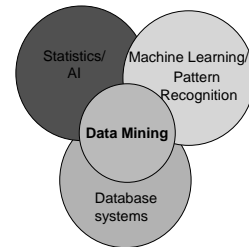
- ◆ Extracting knowledge from large amounts of data



©Tan, Steinbach, Kumar Introduction to Data Mining 2004

## 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



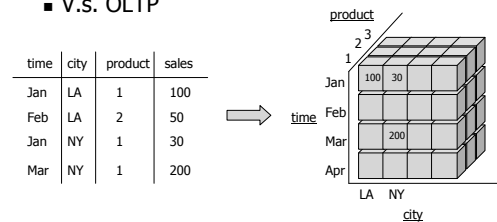
©Tan, Steinbach, Kumar Introduction to Data Mining 2004

## Topics Covered

- ◆ Data warehouse and OLAP
- ◆ Mining frequent patterns
- ◆ Classification and regression
- ◆ Clustering

## OLAP

- ◆ Online Analytic Processing
  - V.s. OLTP



## Mining Frequent Itemsets

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

Frequent Itemsets:  
**{Coke, Milk}**  
**{Beer, Diaper, Milk}**

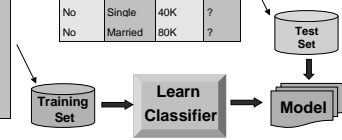
Association Rules:  
**{Milk} --> {Coke}**  
**{Diaper, Milk} --> {Beer}**

©Tan, Steinbach, Kumar Introduction to Data Mining 2004

## 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	?



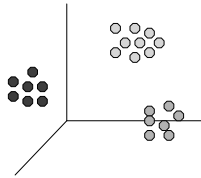
©Tan, Steinbach, Kumar Introduction to Data Mining 2004

## Clustering

☒ Euclidean Distance Based Clustering in 3-D space.

Intracluster distances are minimized

Intercluster distances are maximized



©Tan, Steinbach, Kumar Introduction to Data Mining 2004