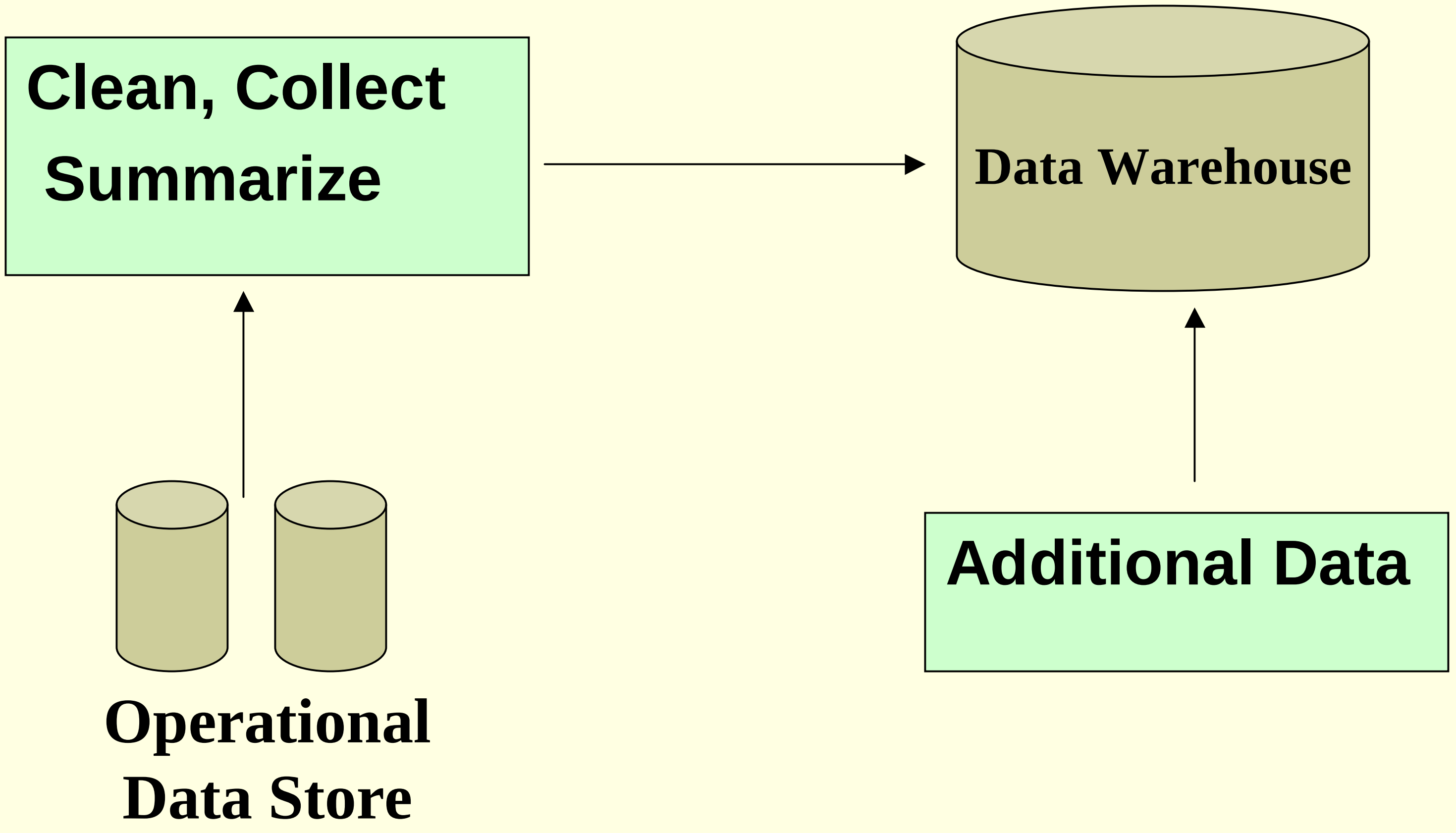
What is Data Mining

- Many Definitions
  - Search for valuable information in large amounts of data
  - Automated or Semi Automated Exploration and Analysis of large volumes of data in order to discover meaningful patterns
  - A step in KDD process
  - …

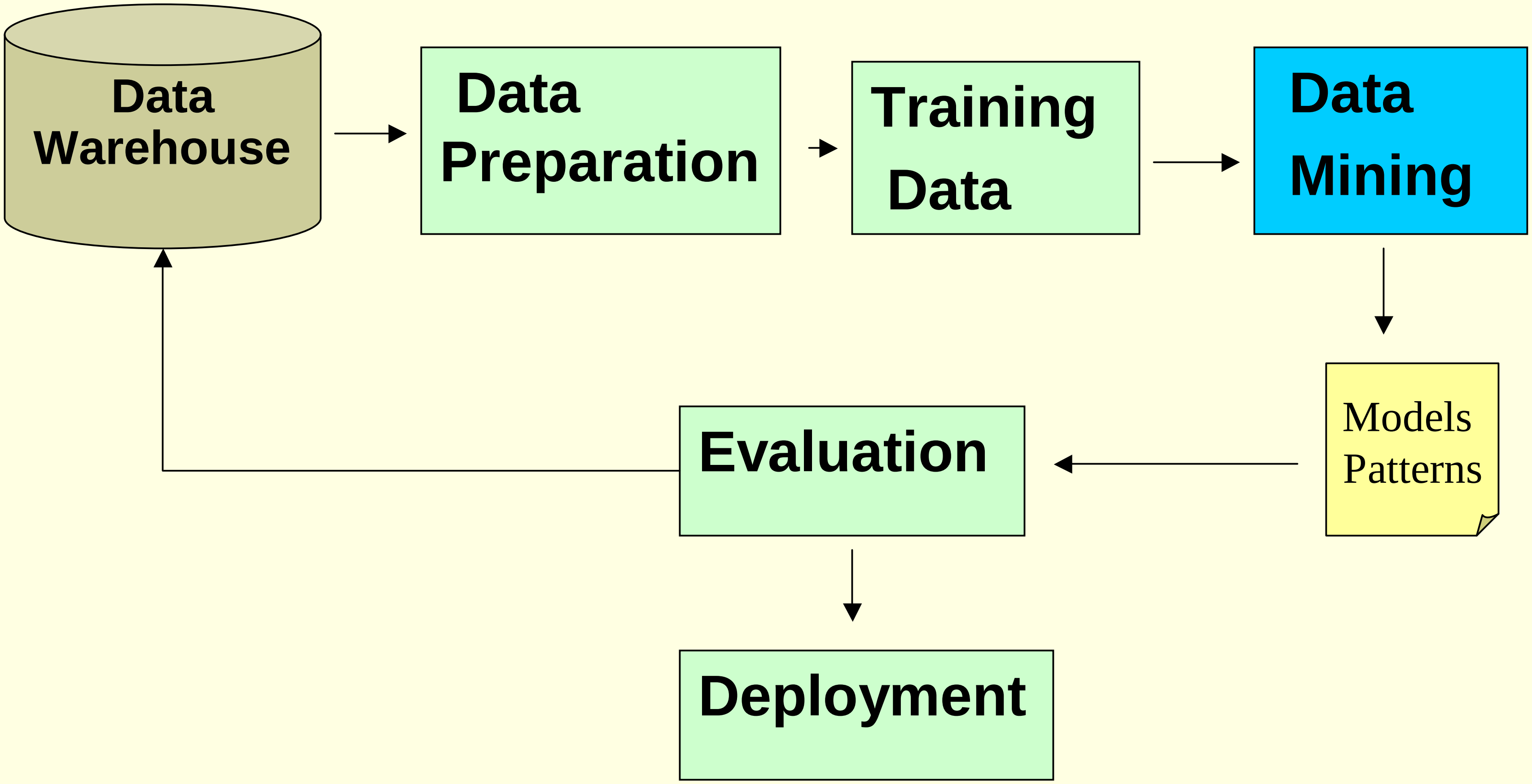
KDD Process

- KDD is a non trivial process of identifying novel valid and potentially useful patterns in data
- Divided into
  - Data Collection into a Data Warehouse
  - Data Mining

KDD Process -1 Data Warehousing



KDD Process-2 Data Mining



Data Mining

- Salient features
  - Large volumes of data
  - Process for discovery information or patterns
  - Automated or semi automated process
  - Useful
  - Understandable

Why Data Mining

- From a scientific viewpoint
  - Data is collected at enormous speeds
    - Microarray experiments producing gene expression data
    - Clinical data
    - Images
  - Data is heterogenous
  - Data is stored in Relational Databases
  - Data mining can be used for summarizing
    - Conversion into understandable form
    - Hypothesis formation

Origins

- Data mining is an interdisciplinary field
- Draws on
  - Computer Science
    - Databases
    - Algorithm theory
    - Machine learning/ AI
  - Statistics
  - Visualization

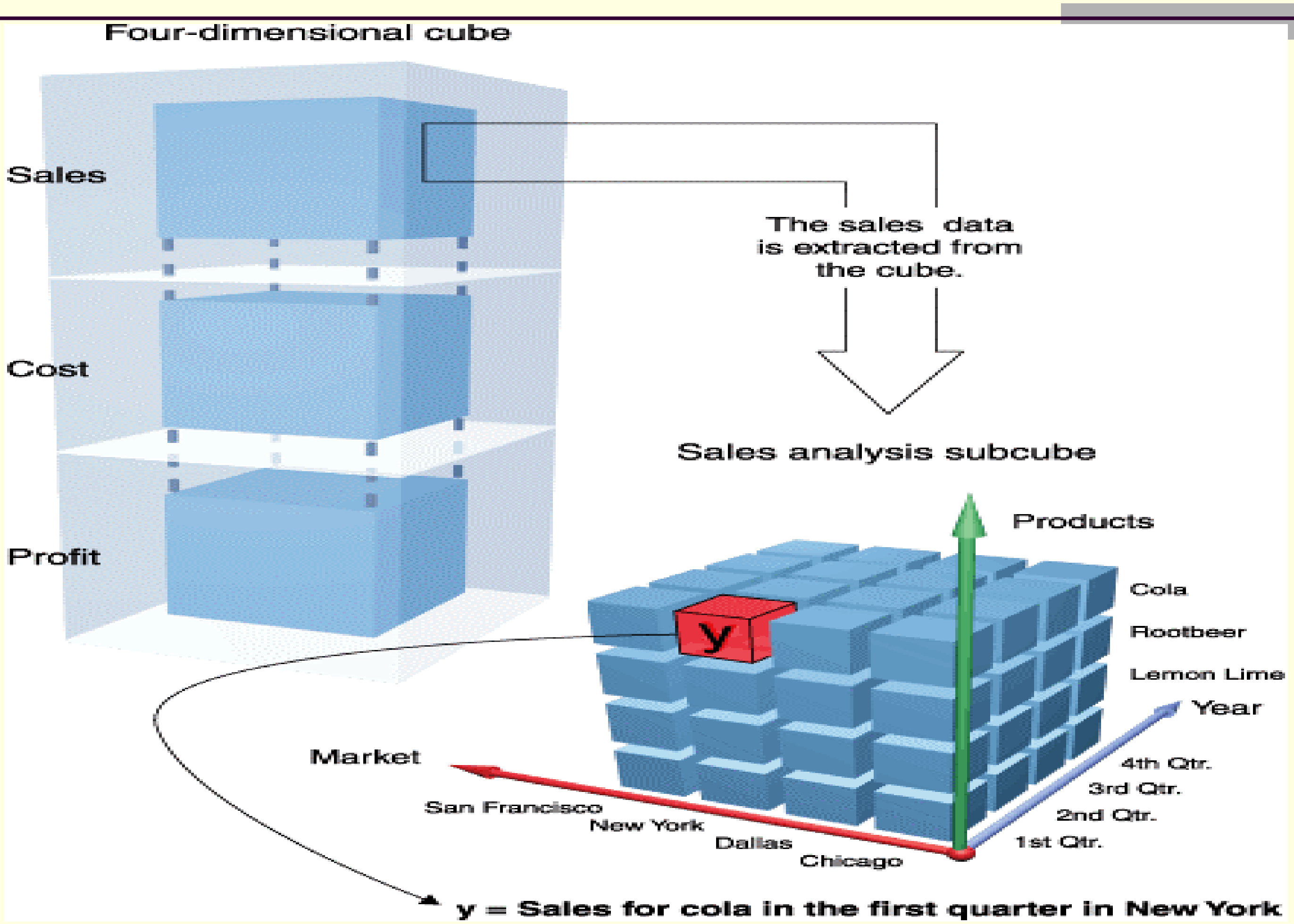
Data Mining Tasks

- Model building
  - Create a model that does a task in an automated manner
    - Unsupervised – dependent variable is absent
    - Supervised   - dependent variable is present
- Descriptive
  - Aid a human in getting information that he desires
    - Adhoc Reports
    - OLAP - FASMI
    - Visualization

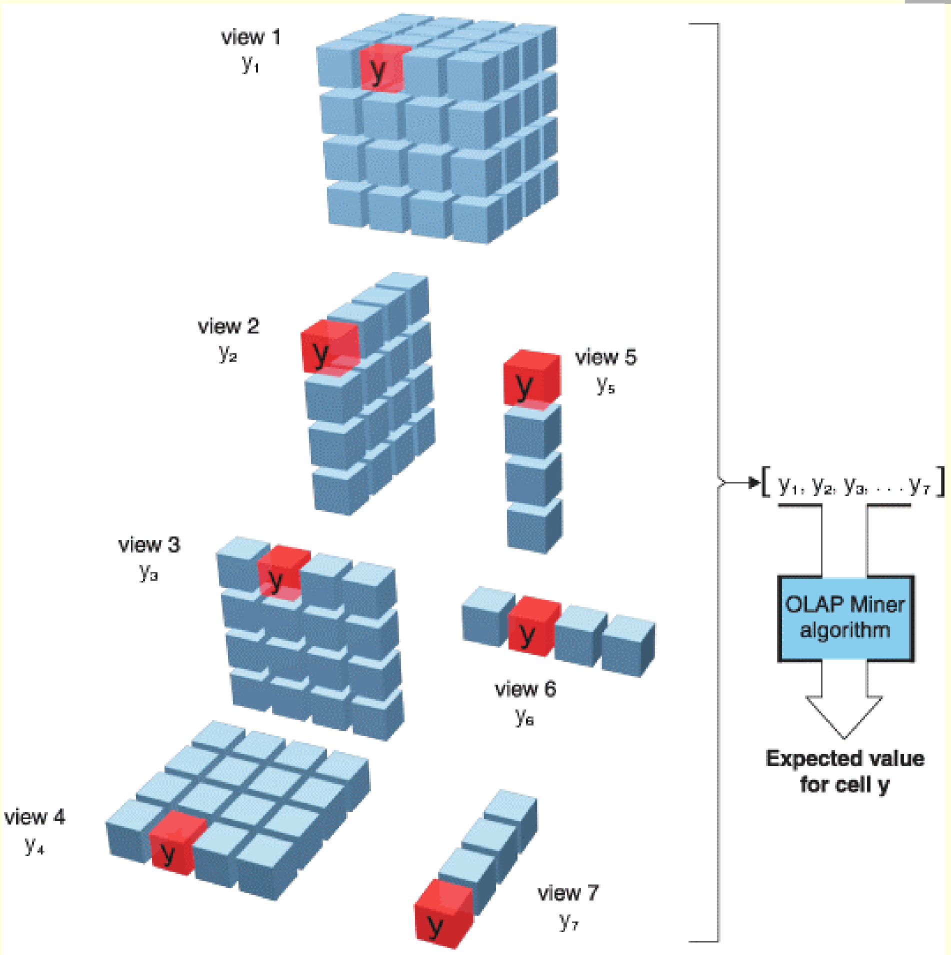
OLAP

- ROLAP
- MOLAP
- Hybrid
- Facts or measurements about the business ----Sale invoices
- Dimensions
    - Products
    - Markets
    - Time

Cubes from OLAP-miner (IBM)

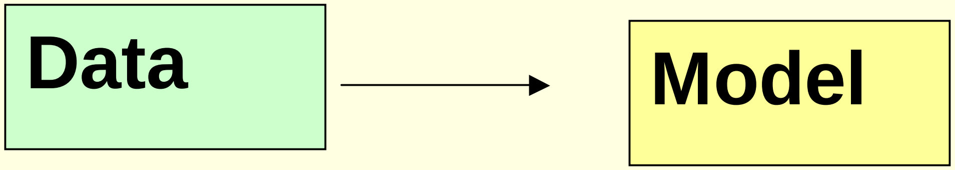


Cubes …


view 1
y_1
y
view 2
y_2
y
view 5
y_5
y
view 3
y_3
y
y
view 6
y_6
view 4
y_4
y
y
view 7
y_7
[y_1, y_2, y_3, . . . y_7]
OLAP Miner algorithm
Expected value for cell y

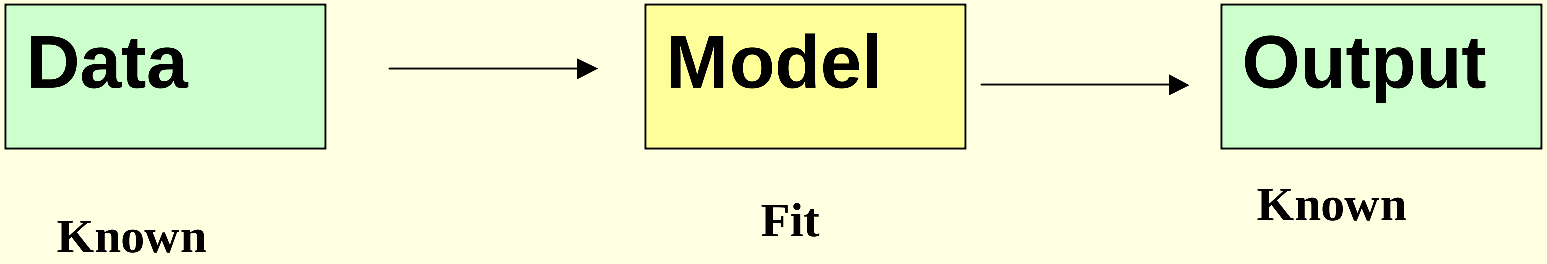
Inductive Models

Unsupervised



Data → Model

Supervised



Data → Model → Output

Known

Fit

Known

Unsupervised Models

- Examples
  - Clustering
  - Association rules
  - Outlier detection
- No apriori dependent variables
  - More flexible
  - Difficult to evaluate accuracy
  - Only criterion is usefulness

Clustering Definition

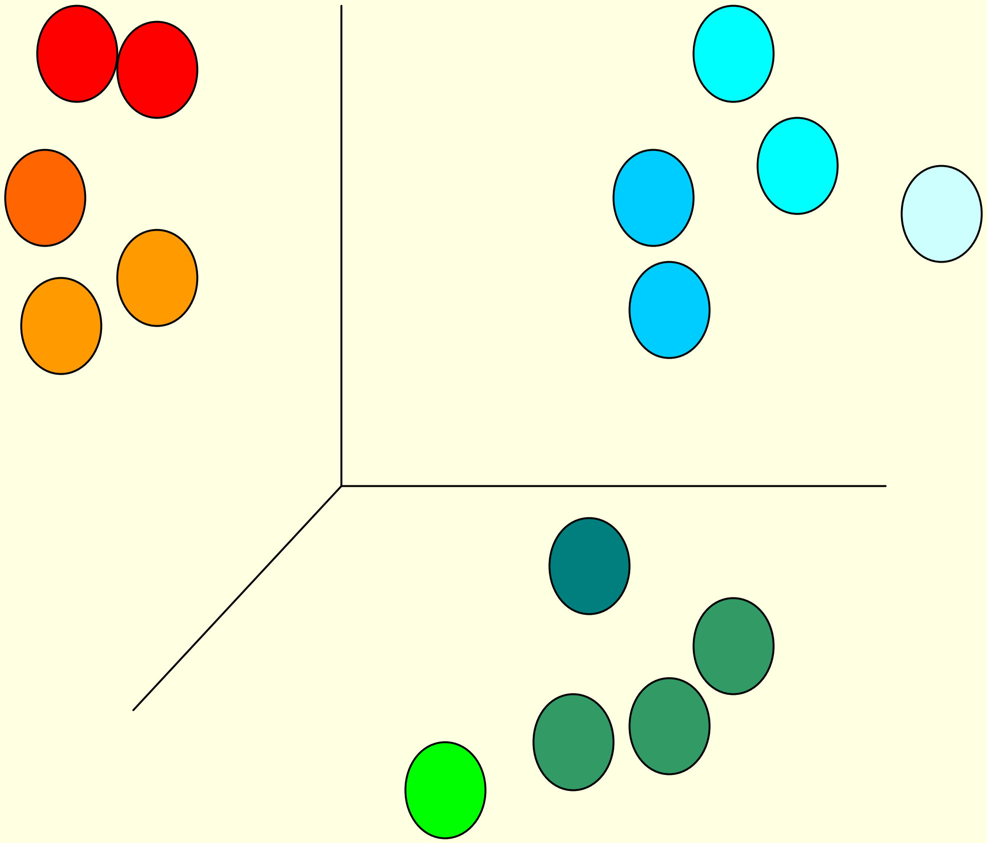
- Given a set of data points, each having a set of attributes and a similarity measure defined find clusters such that
  - Data points in a cluster are similar to each other
  - Data points in different clusters are not similar to each other
- Similarity Measures

  Euclidean distance

  Pearson correlation coefficient

  Jaccard coefficient

Clustering Illustration

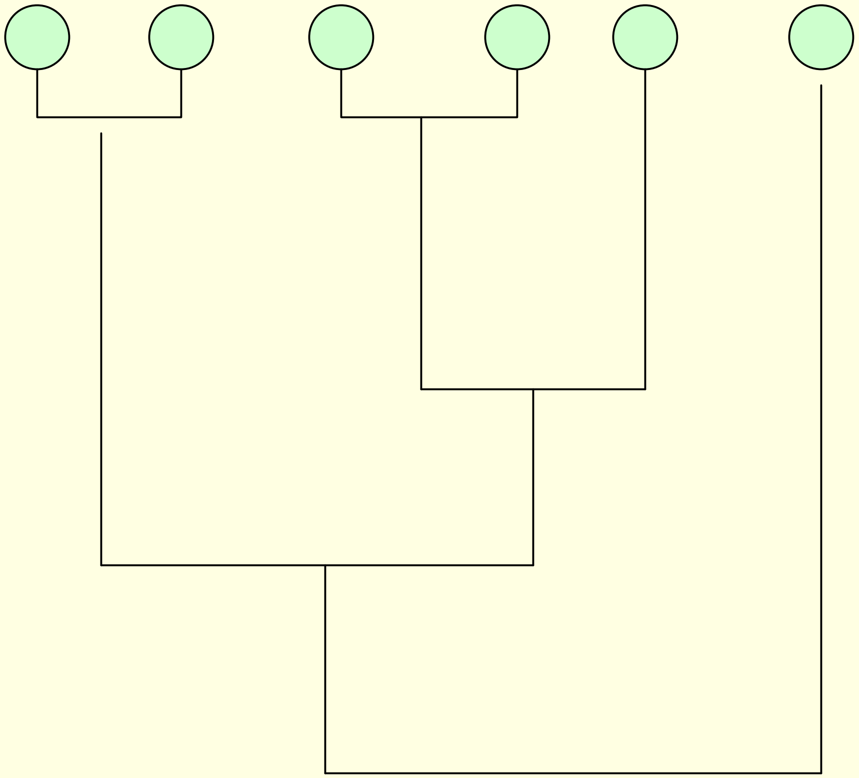


Clustering Algorithms

- Hierarchical: A sequence of nested partitions
  - Agglomerative : Iterative combination of multiple partitions to form a single partition
  - Divisive : Iterative breaking up from one partition to form multiple partitions
- Partitional: a single set of partitions

Hierarchical Agglomerative Clustering

- Dendogram representation



Agglomerative Clustering

- A graphical representation
- Nodes are merged based on a similarity measure defined on groups
    - Single link join based on closest in the groups
    - Complete link based on farthest points in the groups

Partitional Clustering

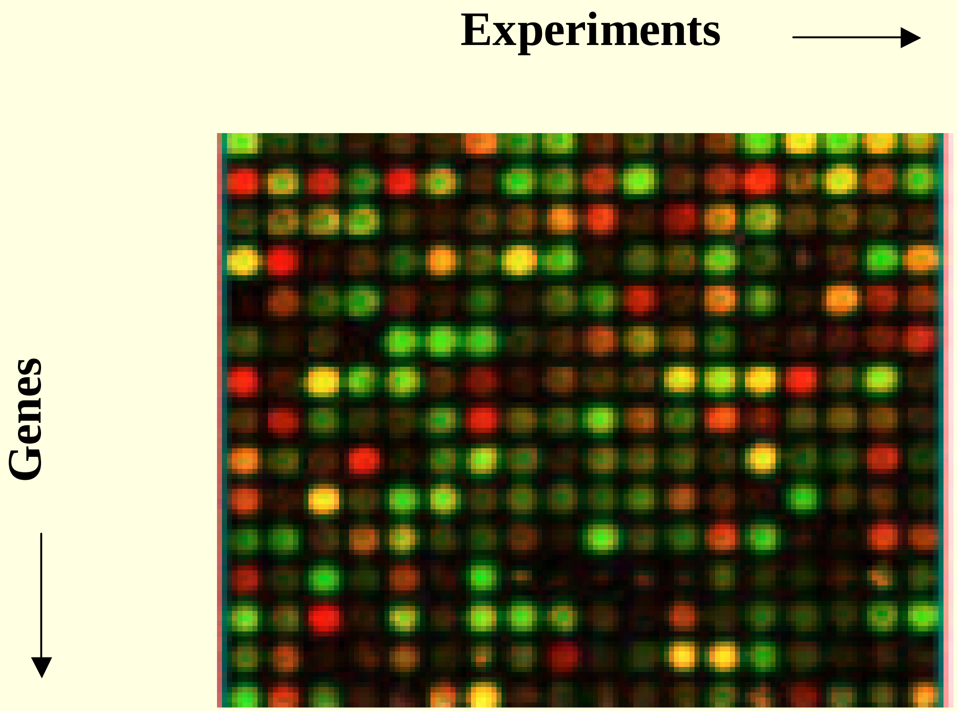
- All data points divided into a fixed number of partitions
  - Divide the data based on prototypes
    - Kmeans Clustering
    - Kohonen Clustering
  - Graph based approaches such as CAST

Nearest Neighbor Clustering

- Input
  - A threshold **t** on the nearest neighbor distance
  - A set of data points $\{x_1,x_2,\ldots,x_n\}$
- Algorithm
  - Initialize assign set i=1, k=1 x_i to C_k
  - Set i=i+1 Find nearest neighbor of x_i among points already assigned to clusters
  - Let the nearest neighbor be in cluster m
  - If distance to the nearest neighbor is < t
    - Assign x_i to m
    - Else increment k and assign x_i to C_k
    - If all points are assigned then stop

Clustering Applications

- **Microarray Data**

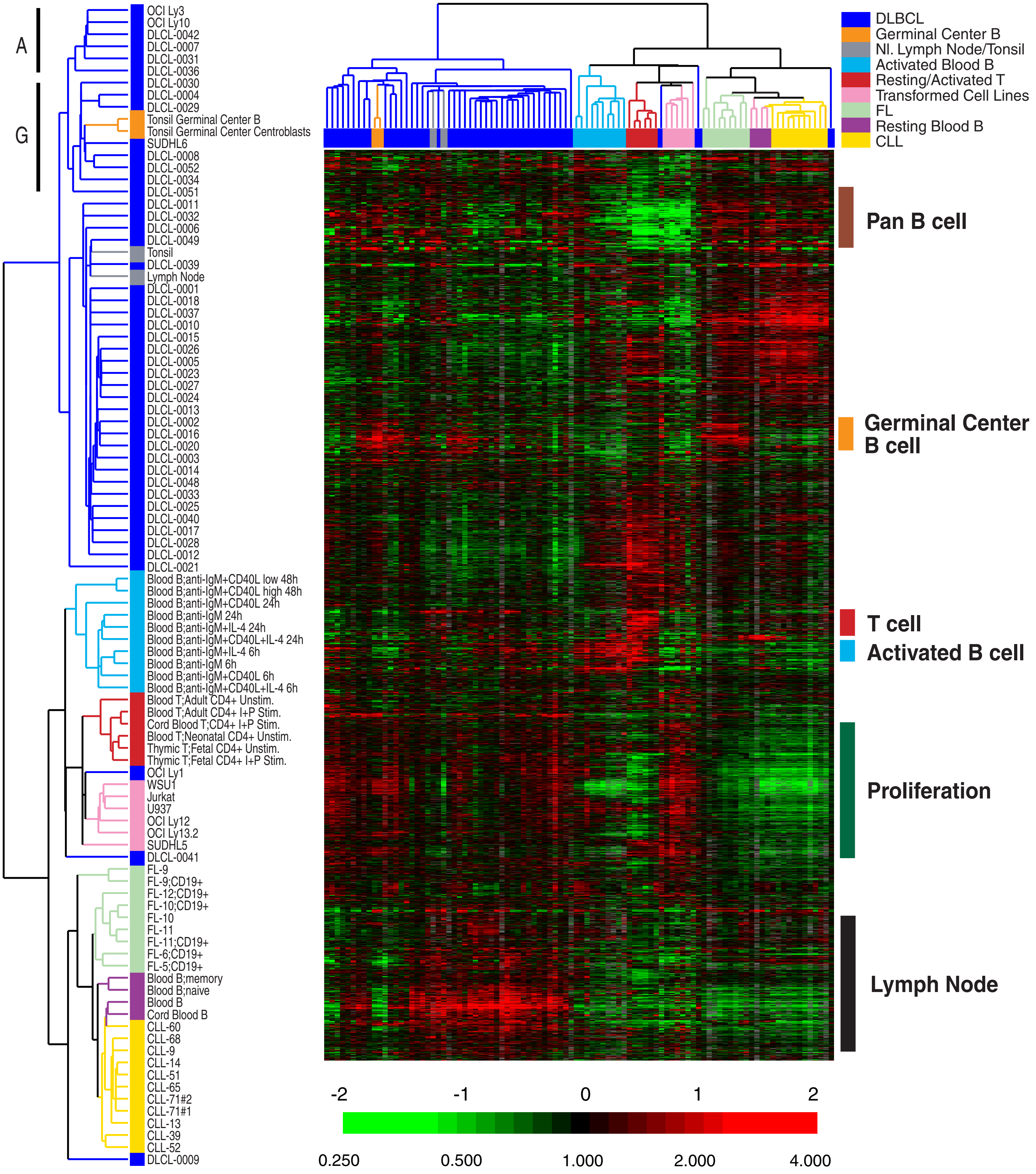


Experiments

Genes

Example of hierarchical clustering

- Use acrobat reader


A
G
OCI Ly3
OCI Ly10
DLCL-0042
DLCL-0007
DLCL-0031
DLCL-0036
DLCL-0030
DLCL-0004
DLCL-0029
Tonsil Germinal Center B
Tonsil Germinal Center Centroblasts
SUDHL6
DLCL-0008
DLCL-0052
DLCL-0034
DLCL-0051
DLCL-0011
DLCL-0032
DLCL-0006
DLCL-0049
Tonsil
DLCL-0039
Lymph Node
DLCL-0001
DLCL-0018
DLCL-0037
DLCL-0010
DLCL-0015
DLCL-0026
DLCL-0005
DLCL-0023
DLCL-0027
DLCL-0024
DLCL-0013
DLCL-0002
DLCL-0016
DLCL-0020
DLCL-0003
DLCL-0014
DLCL-0048
DLCL-0033
DLCL-0025
DLCL-0040
DLCL-0017
DLCL-0028
DLCL-0012
DLCL-0021
Blood B;anti-IgM+CD40L low 48h
Blood B;anti-IgM+CD40L high 48h
Blood B;anti-IgM+CD40L 24h
Blood B;anti-IgM 24h
Blood B;anti-IgM+IL-4 24h
Blood B;anti-IgM+CD40L+IL-4 24h
Blood B;anti-IgM+IL-4 6h
Blood B;anti-IgM 6h
Blood B;anti-IgM+CD40L 6h
Blood B;anti-IgM+CD40L+IL-4 6h
Blood T;Adult CD4+ Unstim.
Blood T;Adult CD4+ I+P Stim.
Cord Blood T;CD4+ I+P Stim.
Blood T;Neonatal CD4+ Unstim.
Thymic T;Fetal CD4+ Unstim.
Thymic T;Fetal CD4+ I+P Stim.
OCI Ly1
WSU1
Jurkat
U937
OCI Ly12
OCI Ly13.2
SUDHL5
DLCL-0041
FL-9
FL-9;CD19+
FL-12;CD19+
FL-10;CD19+
FL-10
FL-11
FL-11;CD19+
FL-6;CD19+
FL-5;CD19+
Blood B;memory
Blood B;naive
Blood B
Cord Blood B
CLL-60
CLL-68
CLL-9
CLL-14
CLL-51
CLL-65
CLL-71#2
CLL-71#1
CLL-13
CLL-39
CLL-52
DLCL-0009
DLBCL
Germinal Center B
Nl. Lymph Node/Tonsil
Activated Blood B
Resting/Activated T
Transformed Cell Lines
FL
Resting Blood B
CLL
Pan B cell
Germinal Center B cell
T cell
Activated B cell
Proliferation
Lymph Node
-2
-1
0
1
2
0.250
0.500
1.000
2.000
4.000

Clustering applications -documents

- To find groups of documents that are similar to each other
  - Use frequencies of words occurring within documents and a similarity measure to group documents together
    - Can be used for automatic categorization of documents
      - Assigning emails automatically for complaint handling

Association rules

- Given a set of records each of which contains some items from a given collection
- Produce dependency rules that will predict occurrence of an item based on occurrence of other items
- Rules discovered
- {Milk} → {Bread}
- {Bread} → {Milk}

1	Bread, Milk
2	Eggs, Bread, Milk
3	Bagels, cream cheese, orange juice
4	Coke, Potato chips
5	Bread, milk, orange juice

Association rules

- Usefulness
- Super market shelf arrangement
- Product pricing and promotion
- Predict normal behavior for Fraud detection

Outlier Detection

- An interesting problem – reamins to be solved for many practical applications
  - Requires a model for “normal”
  - Lots of applications
    - Telecom fraud detection
    - Intrusion detection
    - Medicare fraud detection

Supervised methods

- An output label is available for the data
  - Classification : the output variable is categorical
    - Classification of tissues into cancer types
  - Prediction : The output variable is continuous
    - Prediction of S&P 500 Index

Classification

- Given a collection of records
  - Each record containing a set of attributes or features and a class
- Derive a model that can assign a record to a class as accurately as possible

Set of records :

training set

test set

k-fold Cross validation

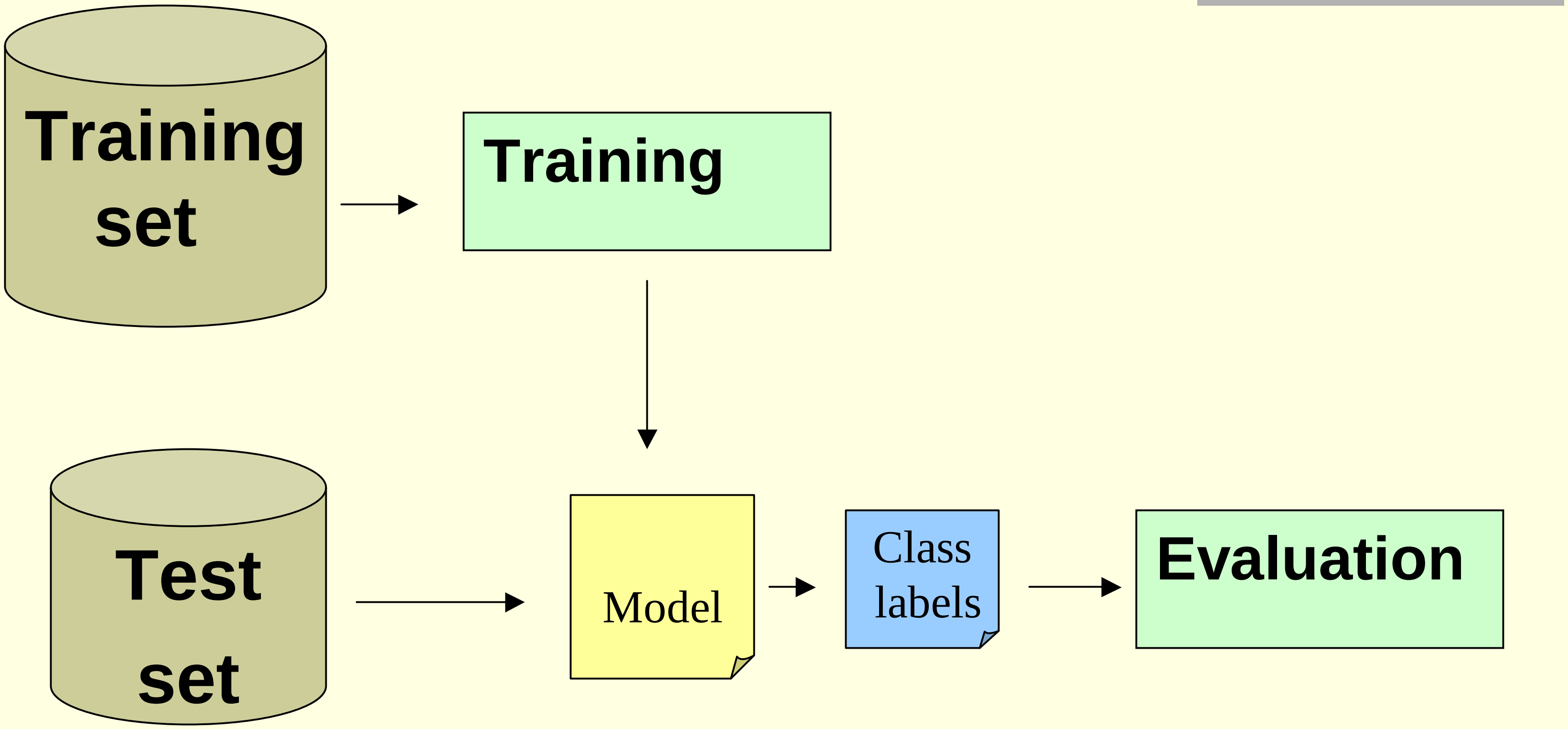
Classification example IRS

Row	Tax. Income	EIC	Marital Status	Child	Refund	Fraud
1	125K	Yes	Single	1	yes	No
2	100k	No	Married	2	no	No
3	40K	Yes	Divorced	0	no	Yes
4	180K	No	Single	0	yes	No
5	100K	Yes	Married	2	no	No
6	50K	Yes	Single	1	Yes	Yes
7	100K	No	Married	1	no	No

Classification example IRS

Row	Tax. Income	EIC	Marital Status	Child	Refund	Fraud
1	100K	No	Single	1	yes	?
2	115k	yes	Married	2	no	?
3	50K	Yes	Divorced	0	no	?
4	140K	No	Single	0	yes	?
5	85K	Yes	Married	2	no	?
6	70K	No	Single	1	Yes	?
7	100K	Yes	Married	1	no	?

Classification Model



Training set

Training

Test set

Model

Class labels

Evaluation

Classification Example 1

- Marketing response
  - Goal : To find a set of customers that will buy vacation property
  - Approach:
    - Collect customer attributes
      - Credit score
      - Income
      - Other purchases
    - Create a classification model {promising, not promising}
    - Send mail and evaluate results

Classification Example 2

Mortgage Loan

- Goal : To grant or reject loan application
- Approach:
  - Collect customer attributes
    - Credit score
    - Income
    - Expenses
    - Credit history
  - Create a classification model {acceptable, not acceptable }
  - Evaluate results

Classification algorithms

- Nearest Neighbor
- Discriminant analysis
- Logistic Regression
- Rule based systems
- Decision trees
- Support vector machines
- Bayesian networks

Nearest Neighbor Algorithm

- Define a distance measure
  - Euclidean distance
  - Manhattan distance
  - Pearson correlation coefficient
  - Find k nearest neighbors
- Classify to the class of the majority

Decision Trees

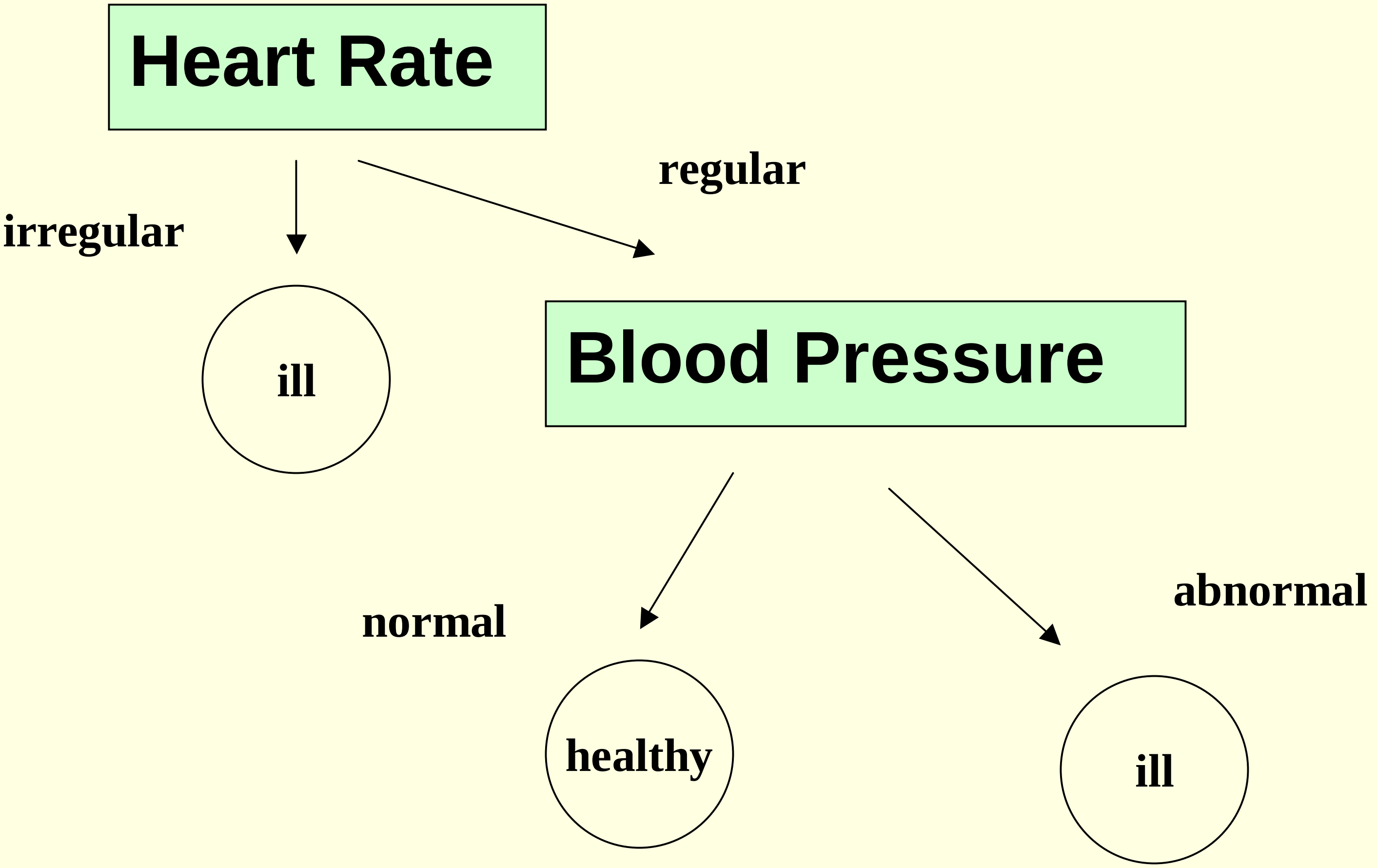
- Repeatedly partition the feature space
  - IDE3
  - CART
  - C4.5

  - Evaluate All variables/combinations
  - Splits on single variables /combinations
    - Mutual Information
    - GINI criterion

Decision Trees

Patient No.	Heart Rate	Blood Pressure	Class
1	irregular	normal	Severely ill
2	regular	normal	healthy
3	irregular	abnormal	severely ill
4	irregular	normal	severely ill
5	regular	normal	Healthy
6	regular	abnormal	ill
7	regular	normal	healthy
8	regular	normal	healthy

Decision Tree induced



Rules Induced

- Can give a better mental fit
- **If** Heart rate is irregular **then** Patient is severely **ill**
- **If** Heart rate is normal and Blood Pressure is abnormal **then** Patient is **ill**
- **If** heart rate is normal **and** blood pressure is normal **then** patient is **healthy**

Prediction

- Given a collection of records
  - Each record containing a set of attributes or features including a dependent variable
- Derive a model that can predict the dependent variable as accurately as possible from the rest of the attributes

Set of records :

training set

test set

k-fold Cross validation

Prediction Example 1

- Credit score
  - Goal: To assign a score to each individual that is an indicator of loan default
  - Approach:
    - Collect training set
      - Credit history
      - Outstanding balances
      - Rent or own
      - Loan defaults
    - Create a prediction model

Prediction Example 2

Weather forecasting

Goal: Predict probability of rain one day in advance

Approach:

- Collect past data
  - humidity
  - pressure
  - temperature
  - rainfall
- Create a prediction model

Prediction Algorithms

- Linear Regression
- Polynomial Nets
- Neural Networks
- Multiple Adaptive Regression Splines

Products- Adhoc queries/reports

Business Objects

Impromtu from Cognos

GQL from Anadyne

Browser from Oracle

Brio Query from Brio technology

Discoverer from Oracle

Products OLAP

- Microsoft
- Hyperion
- Cognos
- Business Objects
- Microstrategy
- SAP
- Oracle

Products - Modeling

General
- Clementine from SPSS
- Enterprise Miner from SAS
- Oracle Data Mining Suite
- Oracle 9i
- IBM Intelligent miner for data
- IBM intelligent miner for text

Specific:
- CART
- Neuroshell

Public domain:
- MLC++
- WEKA
- R

Text Mining

- Text data is unstructured
    - A collection of documents
        - Each document is a collection of words
        - Few cases class label
    - NLP based approaches
        - Natural language understanding
    - Statistics based approaches
    - Mixed approaches

Text Mining – NLP based approaches

- Based on understanding of a language information can be extracted through patterns
  - Can be used directly
  - Convert into structured data

Statistics based approaches

- Need to handle sparse data
  - Lots of possible words
  - Each document contains only a few words
  - 10100000001010100010000010100000000
- TFIDF
  - Term Frequency
  - Inverse document frequency
- Text clustering
  - TFIDF approaches
- Text classification

Text Clustering

- Goal: Divide a set of documents into groups where the number of groups is not known
- Approach:
  - Define a distance measure suitable for binary sparse vectors
    - Commonly used is the cosine distance

      x .y/(|x||y|)

    Use modifications of algorithms that can handle large data size

CNN and Reuters news stories Jan-Feb 95

Size	**Top ranking words per cluster**
330	clinton congress house amend
217	Simpson trial jury prosecute
98	Israel palestine gaza peace arafat
97	Japan kobe earthquake
93	Russian grozn yeltsin chechnya

Document Classification

- Goal : Classify email into spam and non spam
- Approach:
  - Create a corpus of spam and non spam email
  - Train a text classifier (naïve bayes)
  - Evaluate on a test set
  - Accuracy obtained was of the order 99.85%

Questions