Introduction to Artificial Intelligence G51IAI

An Introduction to Data Mining

Learning Objectives

- Introduce a range of data mining techniques used in AI systems including :
  - Neural networks
  - Decision trees
  - …
- Present some real life data mining applications.

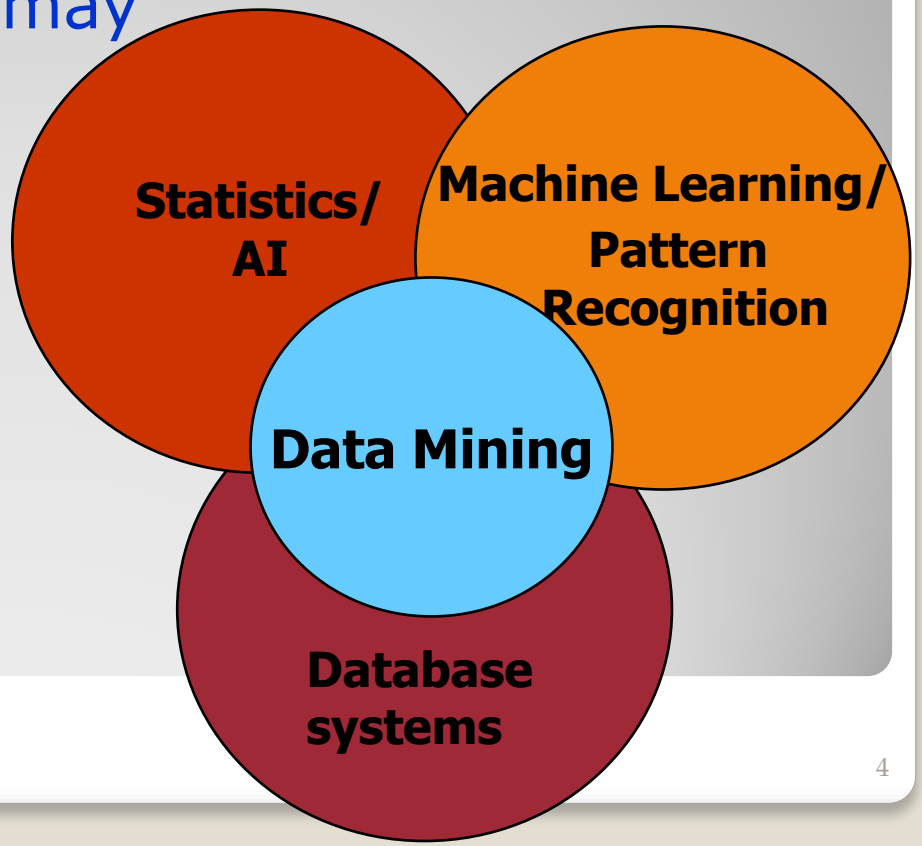
Road Map

- **Data mining overview**
- Data mining tasks
  - Classification (supervised learning)
  - Clustering (unsupervised learning)
  - Association rule discovery
- Summary

Note: These lecture materials are based on invited lectures from Dr Li in the School of Computer Science, University of Nottingham, 2008.

Origins of Data Mining?

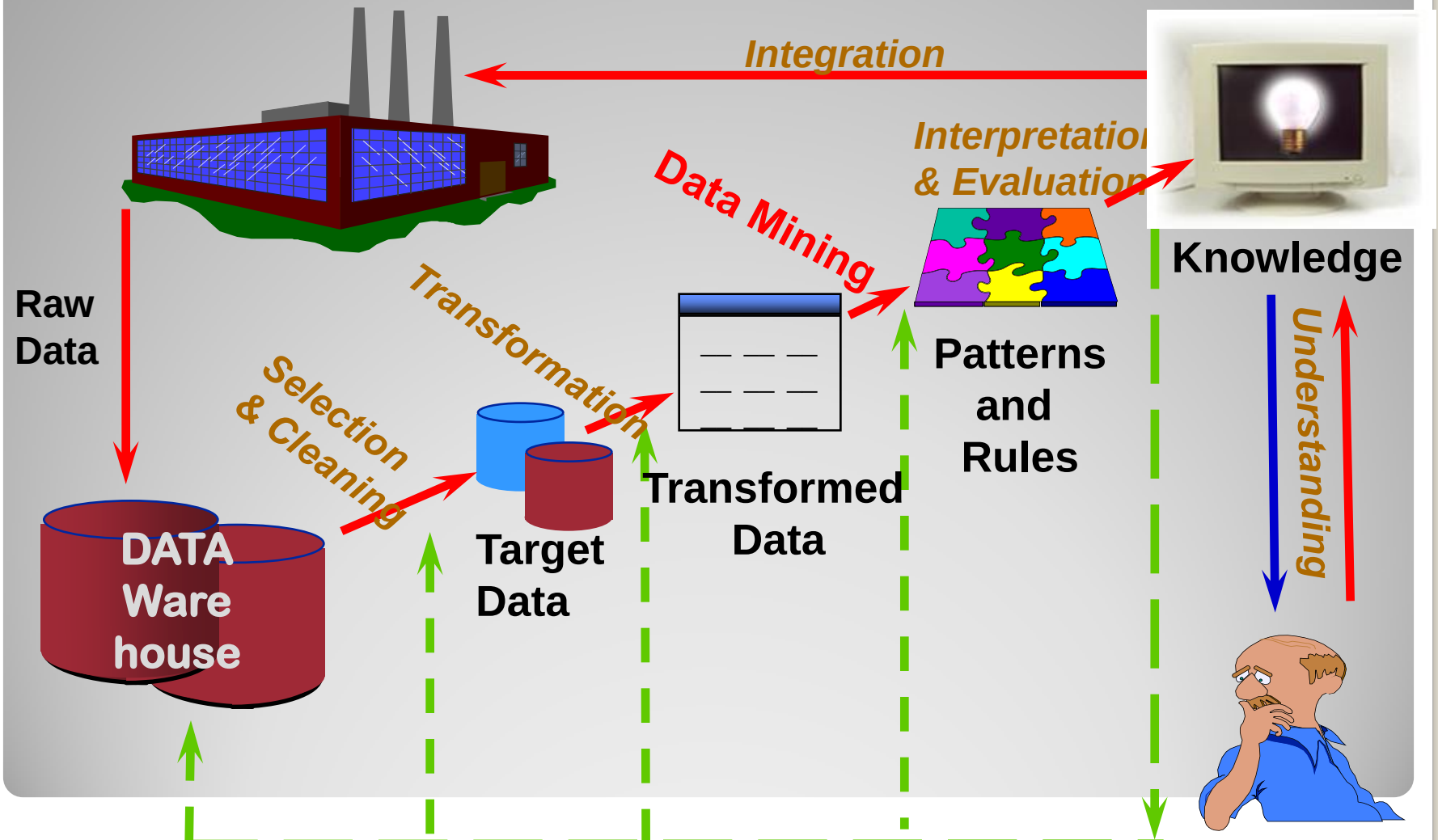
- Draws ideas from machine learning / AI, pattern recognition, statistics, and database systems
- Traditional Techniques may be unsuitable due to
  - High dimensionality of data
  - Heterogeneous, distributed nature of data



What is Data Mining?

- Data mining is the exploration and analysis of large quantities of data in order to discover *valid*, *novel*, potentially *useful*, and ultimately *understandable* patterns in data.
  - **Valid**: hold on new data with some certainty
  - **Novel**: non-obvious to the system
  - **Useful**: should be possible to act on the item
  - **Understandable**: humans should be able to interpret the pattern
  - Also known as Knowledge Discovery in Databases (**KDD**)

Data Mining: the core of KDD



What is (Not) Data Mining? – Examples

- **What is not Data Mining?**
  - Look up phone number in phone directory
  - Query a Web search engine for information about "Amazon"

- **What is Data Mining?**
  - Certain names are more prevalent in certain US locations (O'Brien, O'Rurke, O'Reilly… in Boston area)
  - Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com)

Why Data Mining? - Commercial Viewpoint

- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - Purchases at department/ grocery stores
  - Bank/Credit Card transactions

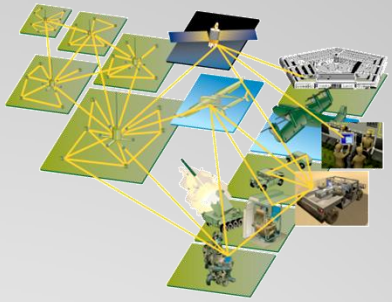
- Computers have become cheaper and more powerful
- Competition pressure is getting stronger
  - Provide better, customized services, e.g. in Customer Relationship Management (CRM)

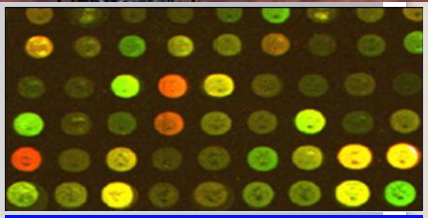
Why Data Mining? - Commercial Viewpoint

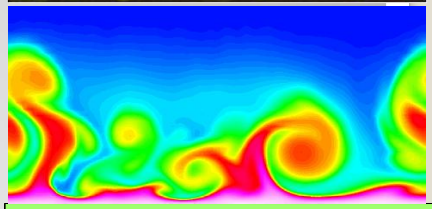
- Banking: loan/credit card approval
  - predict good customers based on old customers
- Fraud detection: network security, financial transactions
  - use historical data to build models of fraudulent behavior and use data mining to help identify similar instances
- Customer relationship management (CRM)
  - Which of my customers are likely the most loyal
  - Which are most likely to leave for a competitor?
  - Identify likely responders to sales promotions
- ...

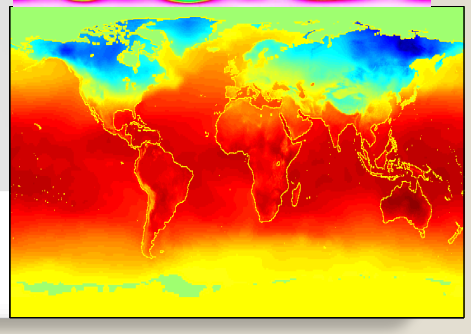
Why Data Mining? – Scientific Viewpoint

- Data collected and stored at enormous speeds
  - Remote sensors on a satellite
  - Telescopes scanning the sky
  - Microarrays generating gene expression data
  - Scientific simulations generating terabytes of data









- Traditional techniques infeasible for raw data
- Data mining may help scientists in
  - Classifying and segmenting data
  - Hypothesis Formation

Why Data Mining? – Scientific Viewpoint

- Medicine: disease outcome, effectiveness of treatments
  - analyze patient disease history & find relationship between diseases
- Astronomy: scientific data analysis
  - identify new galaxies by the stages of formation
- Web site design and promotion:
  - find affinity of visitor to pages and modify layout

Road Map

- Data mining overview
- **Data mining tasks**
  - Classification (supervised learning)
  - Clustering (unsupervised learning)
  - Association rule discovery
- Summary

Data Mining tasks

Predictive: use some variables to predict unknown or future values of other variables

Descriptive: Find human-interpretable patterns that describe the data

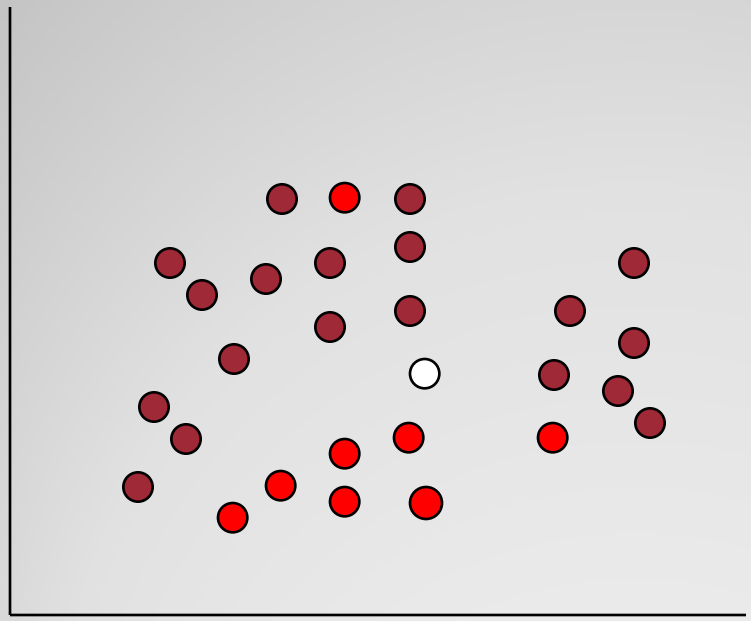
- Classification -- **Predictive**
- Clustering -- **Descriptive**
- Association rule discovery -- **Descriptive**

Road Map

- Data mining overview
- Data mining tasks
  - **Classification (supervised learning)**
  - Clustering (unsupervised learning)
  - Association rule discovery
- Summary

Classification: Illumination

- Learn a method to predict the instance class from pre-labeled (classified) instances



Many approaches:
Regression,
Decision Trees,
Neural Networks,
...

Classification (Supervised Learning)

- Given a collection of records
  - Each record contains a set of **attributes**
  - One of the attributes is the **class attribute**
- Find a model for class attribute as a function of the values of other attributes
- Goal: assign a class to unseen records correctly
- Approach
  - Divide the given data set into training & test sets
  - Use training set to build the model
  - Use test set to validate the model

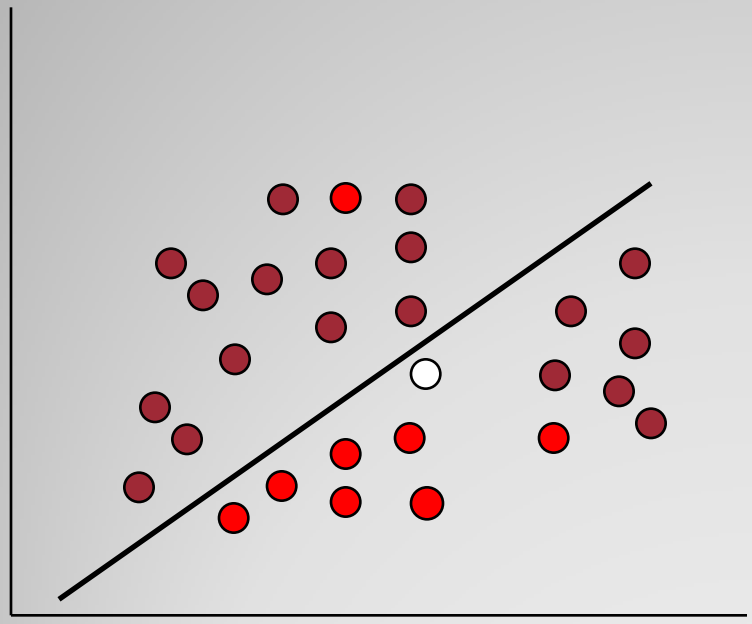
Classification: Methods

- **Goal**: Predict class yi = f(x1, x2, .. Xn)
- **Regression**: (linear or any other polynomial)
  a*x1 + b*x2 + c = yi
- **Decision trees**: divide decision space into piecewise constant regions.
- **Neural networks**: partition by non-linear boundaries
- **…**

Classification: Regression

- Predict the value of a given variable Y based on the values of other variables X
  - assuming a linear or nonlinear model of dependency.
- Examples:
  - Predicting sales amounts of new product based on advertising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.

Linear Regression



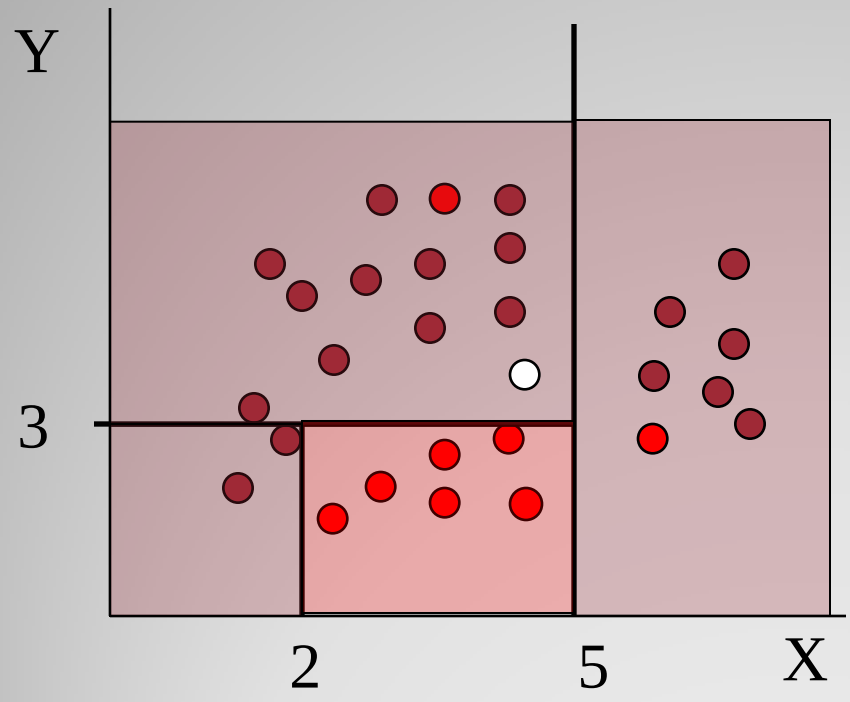
- Linear Regression

  $w_0 + w_1 x + w_2 y >= 0$

- Regression computes **wi** from data to minimize squared error to 'fit' the data

- Not flexible enough

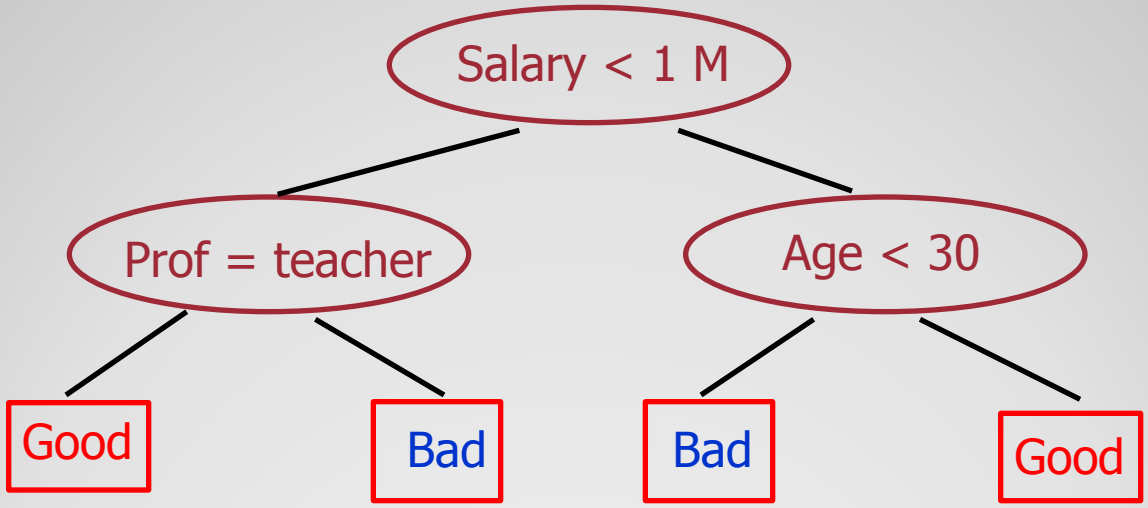
Classification: Decision Trees



if X > 5 then dark red
else if Y > 3 then dark red
else if X > 2 then red
else dark red

Decision Trees

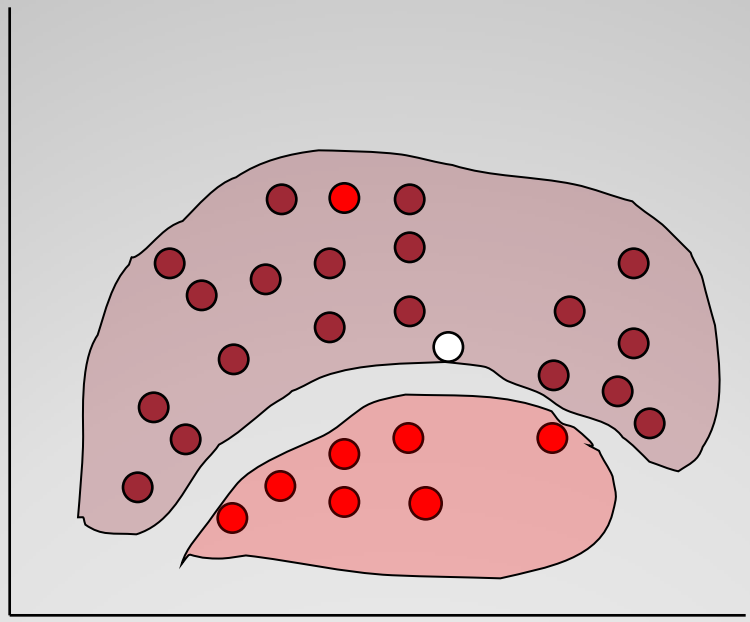
- **Internal node**: a simple decision rule on one or more attributes
- **Leaf node**: a predicted class label



Decision Tree - Cont.

- **Pros**
  - \+ Reasonable training time
  - \+ Easy to interpret
  - \+ Easy to implement
  - \+ Can handle large number of attributes
- **Cons**
  - \- Simple decision boundaries
  - \- Cannot handle complicated relationship between attributes
  - \- Problems with lots of missing data

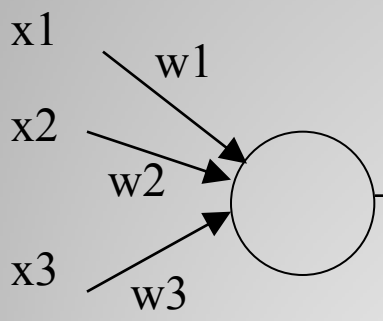
Classification: Neural Networks



Neural Networks

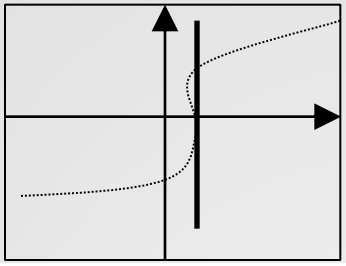
- Set of neurons connected by directed weighted edges

Basic NN unit

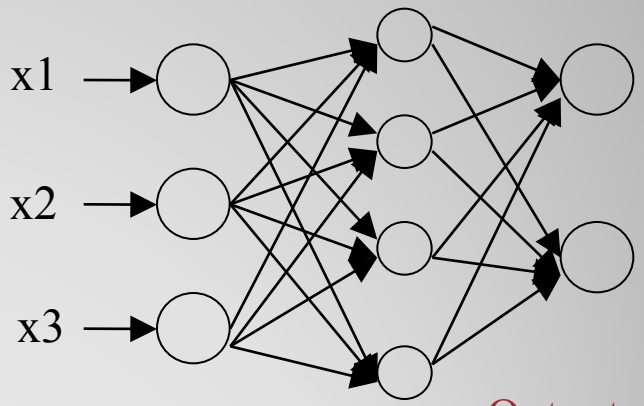


$$o = \sigma\left(\sum_{i=1}^{n} w_i x_i\right)$$

$$\sigma(y) = \frac{1}{1+e^{-y}}$$



A more typical NN



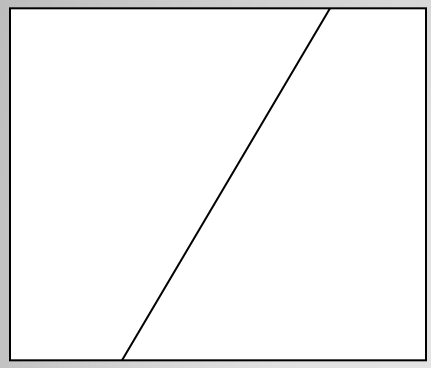
Output nodes

Hidden nodes

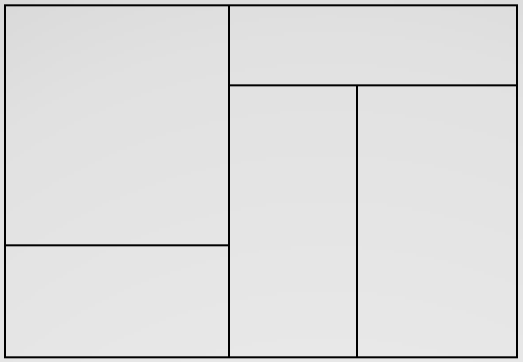
Neural Networks

- Useful for learning complex data like speech, image / handwriting recognition

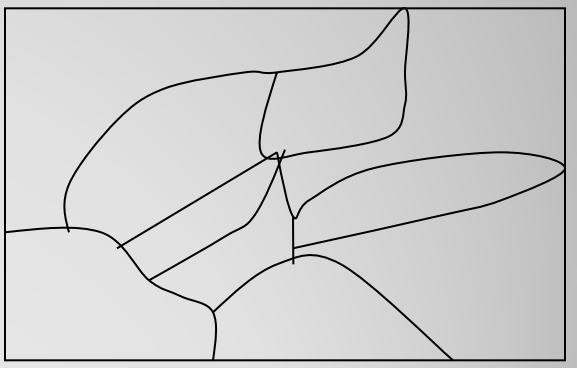
Decision boundaries:



Linear regression



Decision tree



Neural network

Neural Networks

➢ Pros

+ Can learn more complicated class boundaries
+ Can be more accurate
+ Can handle large number of features

➢ Cons

- Slow training time
- Hard to interpret
- Hard to implement: trial and error for choosing parameters and network structure
- Can overfit the data: find patterns in random noise

Classification: Application 1

- Target marketing
  - **Goal:** Reduce cost of mailing by targeting consumers who are likely to buy a new cell-phone product.
  - **Approach:**
    - Find the old data for a similar product.
    - Collect information of all customers.
      - Business type, where they stay, how much they earn, …
    - We know previous customers decision. This *{buy, don't buy}* decision forms the *class attribute*.
    - Use this information to learn a classifier model.

Classification: Application 2 - Banking

- 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 a customer buys, what he buys, how often he pays on time, etc
    - Label past transactions as fraud or fair. This forms the class attribute.
    - Learn a model for fraudulence.
    - Use this model to detect fraud by observing credit card transactions on an account.

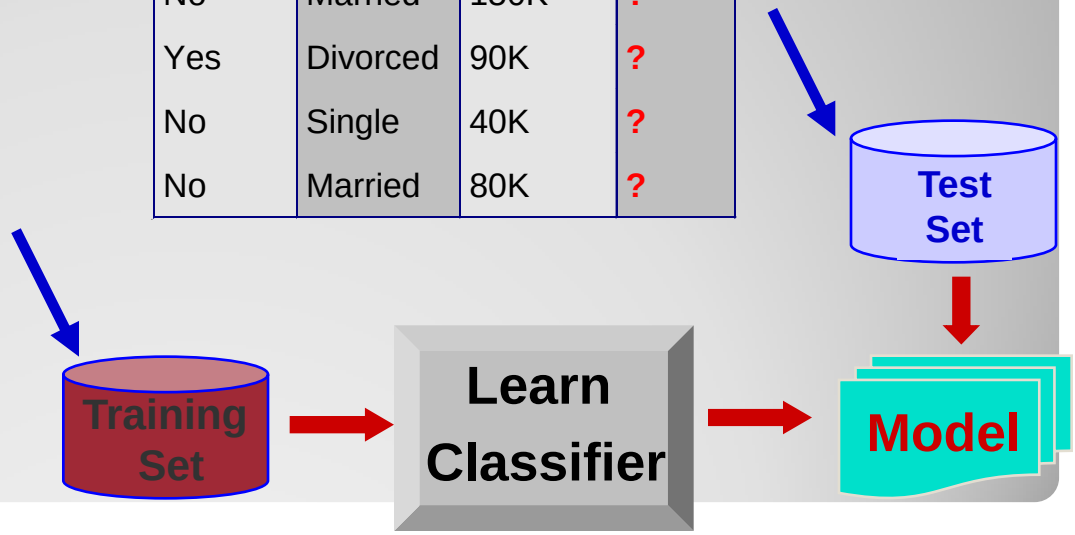
Classification: Application 2 - Banking

categorical | categorical | continuous | class

Tid	Refund	Marital Status	Taxable Income	Good
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

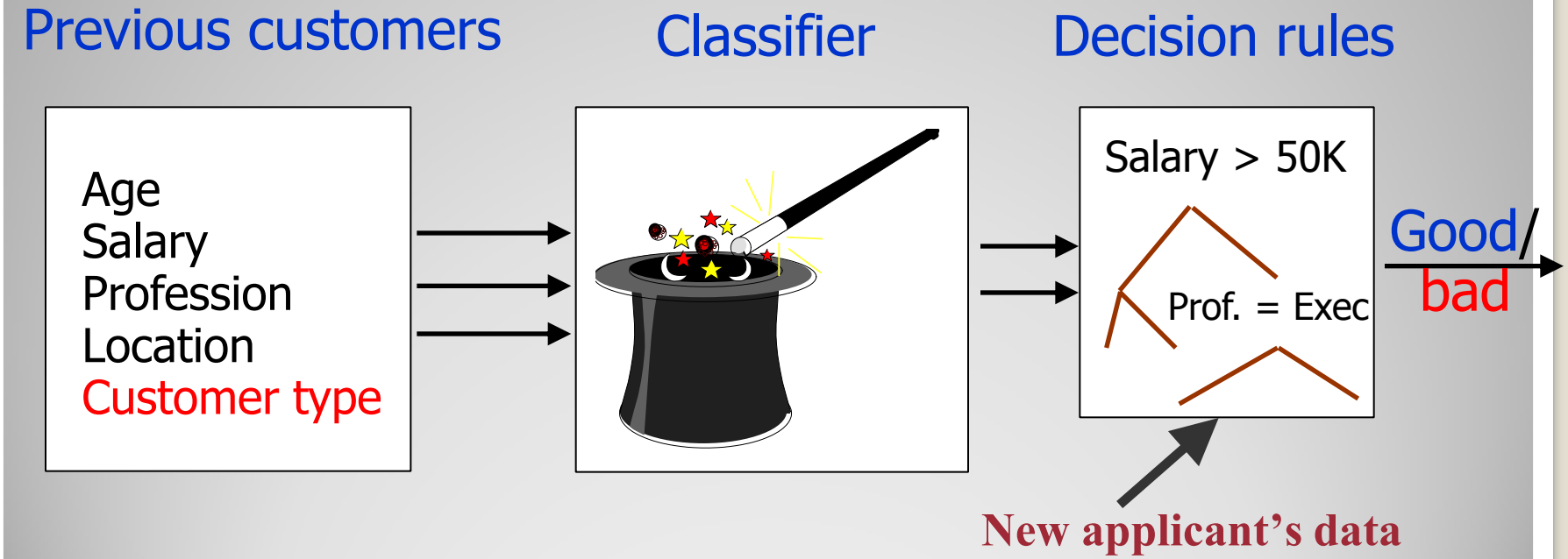
record

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



Classification: Application 2 - Banking

- Loan approval: given old data about customers and payments, predict new applicant's loan eligibility.



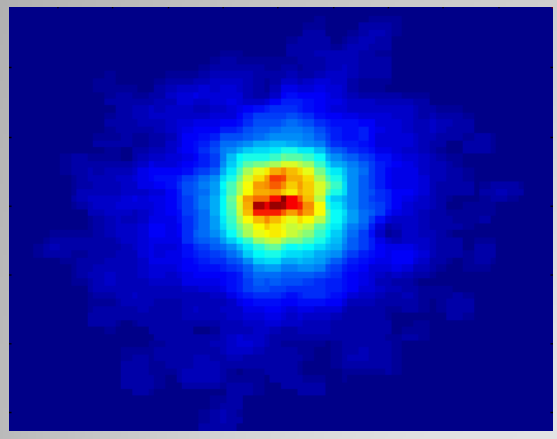
Classification: Application 3

- Customer relationship management
  - **Goal:** To predict whether a customer is likely to be lost to a competitor.
  - **Approach:**
    - Use detailed record of transactions of each 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.

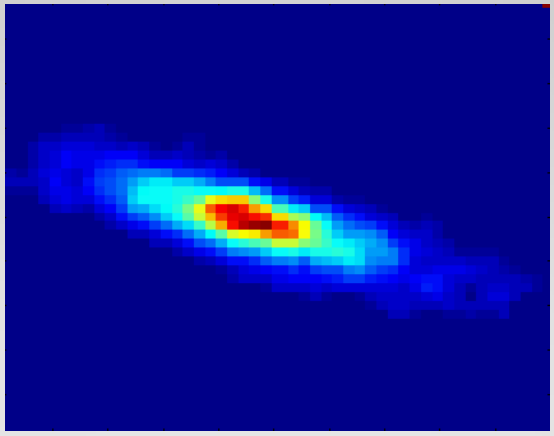
Classification: Application 4

- Classifying galaxies

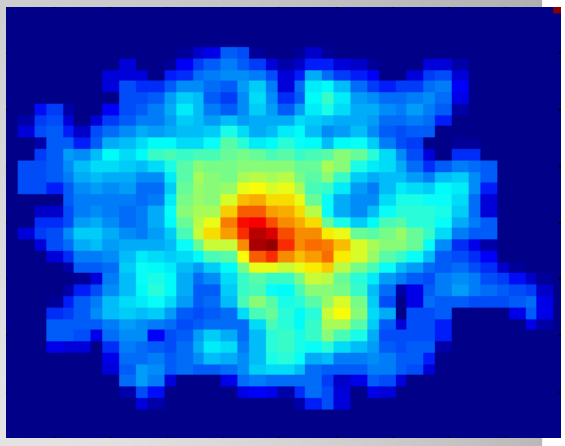
Early



Intermediate



Late



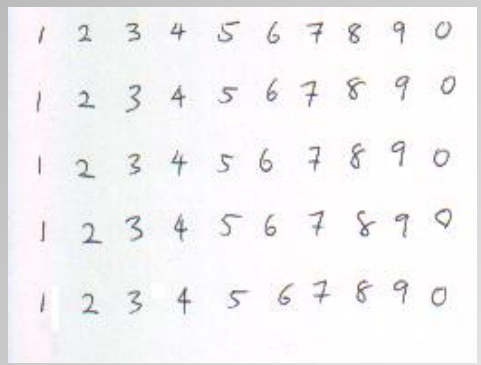
Data Size:
- 20M galaxies
- Over 10K images
- 600M pixels per image

Class:
- Stages of Formation {Early, Intermediate, Late}

Attributes (over 40):
- Image features
- Features such as light waves received, ...

Classification: Application 5

- Handwriting / pattern recognition



Build a ANN in Matlab:
- Overfitting

ASIMO robot:
- James May

Data Mining

- Data mining overview
- Data mining tasks
  - Classification (supervised learning)
    - Regression
    - Decision tree
    - Neural network
  - **Clustering (unsupervised learning)**
  - Association rule discovery
- Summary

Supervised Learning

- F(x): true function (usually not known)
- D: training sample (x, F(x))

57,M,195,0,125,95,39,25,0,1,0,0,0,1,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0	0
78,M,160,1,130,100,37,40,1,0,0,0,1,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0	1
69,F,180,0,115,85,40,22,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0	0
18,M,165,0,110,80,41,30,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	0
54,F,135,0,115,95,39,35,1,1,0,0,0,1,0,0,0,1,0,0,0,0,1,0,0,0,1,0,0,0,0	1

- G(x): model learned from D

71,M,160,1,130,105,38,20,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0	?

- Goal: $E[(F(x)-G(x))^2]$ is small (near zero) for future samples

Un-Supervised Learning

Training dataset:

57,M,195,0,125,95,39,25,0,1,0,0,0,1,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0	0
78,M,160,1,130,100,37,40,1,0,0,0,1,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0	1
69,F,180,0,115,85,40,22,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0	0
18,M,165,0,110,80,41,30,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	1
54,F,135,0,115,95,39,35,1,1,0,0,0,1,0,0,0,1,0,0,0,0,1,0,0,0,1,0,0,0	
84,F,210,1,135,105,39,24,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0	
89,F,135,0,120,95,36,28,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,1,0,0	
49,M,195,0,115,85,39,32,0,0,0,1,1,0,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0	
40,M,205,0,115,90,37,18,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	
74,M,250,1,130,100,38,26,1,1,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	1
77,F,140,0,125,100,40,30,1,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,1,1	0

Test dataset:

71,M,160,1,130,105,38,20,1,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0 **?**

Un-Supervised Learning

Data set:

```
57,M,195,0,125,95,39,25,0,1,0,0,0,1,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0
78,M,160,1,130,100,37,40,1,0,0,0,1,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0
69,F,180,0,115,85,40,22,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0
18,M,165,0,110,80,41,30,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
54,F,135,0,115,95,39,35,1,1,0,0,0,1,0,0,0,1,0,0,0,0,1,0,0,0,1,0,0,0,0
84,F,210,1,135,105,39,24,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0
89,F,135,0,120,95,36,28,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,1,0,0
49,M,195,0,115,85,39,32,0,0,0,1,1,0,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0
40,M,205,0,115,90,37,18,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
74,M,250,1,130,100,38,26,1,1,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
77,F,140,0,125,100,40,30,1,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,1,1
```

Supervise Vs. Un-Supervised Learning

Supervised

- **y=F(x)**: true function
- **D**: labeled training set
- **D**: $\{x_i, F(x_i)\}$
- **Learn**:
  G(x): model trained to predict labels of new cases
- **Goal**:
  $E[(F(x)-G(x))^2] \approx 0$
- **Well defined criteria**:
  Mean square error

Un-supervised

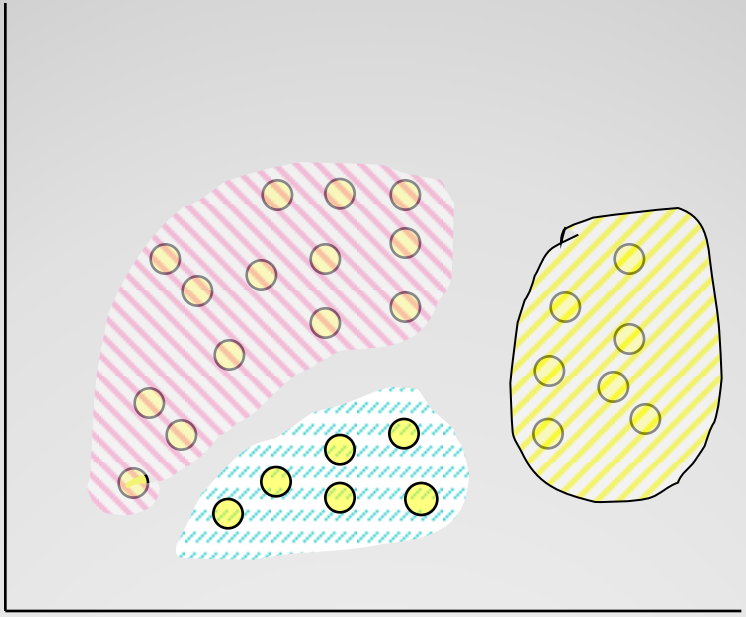
- **y=?**: no true function
- **D**: unlabeled data set
- **D**: $\{x_i\}$
- **Learn**
  ??????????
- **Goal**:
  ??????????
- **Well defined criteria**:
  ??????????

Clustering (Unsupervised Learning)

- What we have:
  - Data Set D
  - Similarity/distance metric
- What we need to do:
  - Find Partitioning of data, or groups of similar/close items

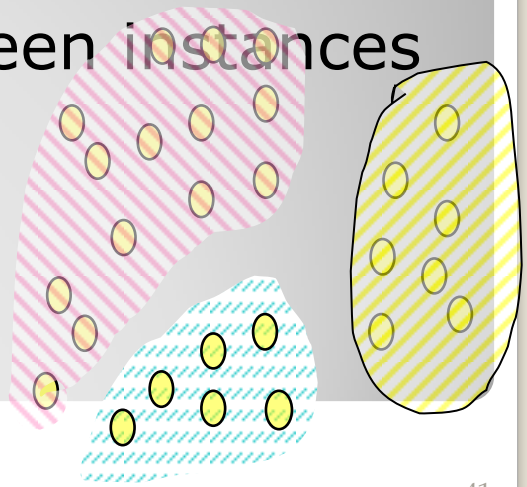
Clustering: Illumination

- Find "natural" grouping of instances given un-labeled data



Clustering

- Given a set of data points, each having a set of attributes and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Key requirement:
  - Need a measure of similarity between instances
    - Manhattan & Euclidean distances
    - Hamming distance
    - Other problem-specific measures



Clustering: Similarity?

- Groups of similar customers
  - Similar demographics
  - Similar buying behavior
  - Similar health
- Similar products
  - Similar cost
  - Similar function
  - Similar store
  - …
- Similarity usually is domain/problem specific

Clustering: Distance Functions

- Numeric data:
  - Euclidean distance
  - Manhattan distance
- Categorical data (0 / 1 indicating presence / absence):
  - Hamming distance (# dissimilarity)
- Combined numeric and categorical data:
  - weighted normalized distance

Manhattan & Euclidean Distance

Consider two records $x=(x_1,\ldots,x_d)$, $y=(y_1,\ldots,y_d)$:

$$d(x,y)=\sqrt[p]{|x_1-y_1|^p+|x_2-y_2|^p+\ldots+|x_d-y_d|^p}$$

Special cases:

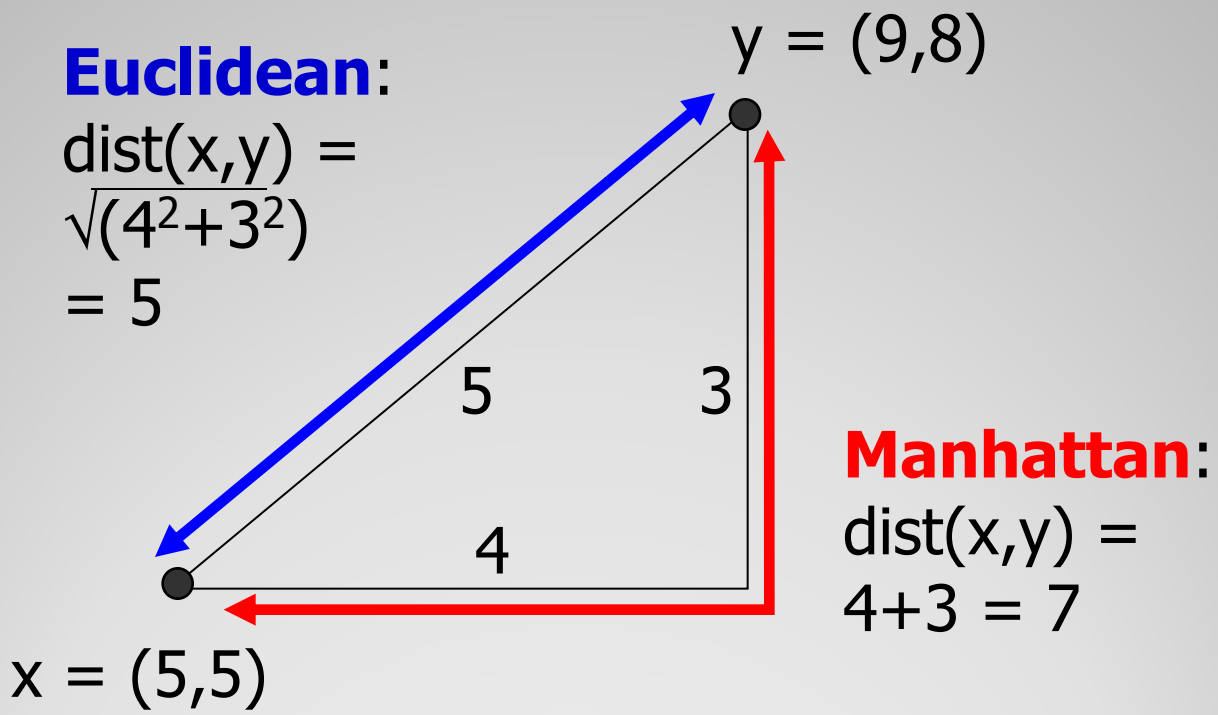
- p=1: Manhattan distance

$$d(x,y)=|x_1-y_1|+|x_2-y_2|+\ldots+|x_p-y_p|$$

- p=2: Euclidean distance

$$d(x,y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\ldots+(x_d-y_d)^2}$$

Manhattan & Euclidean Distances: Example

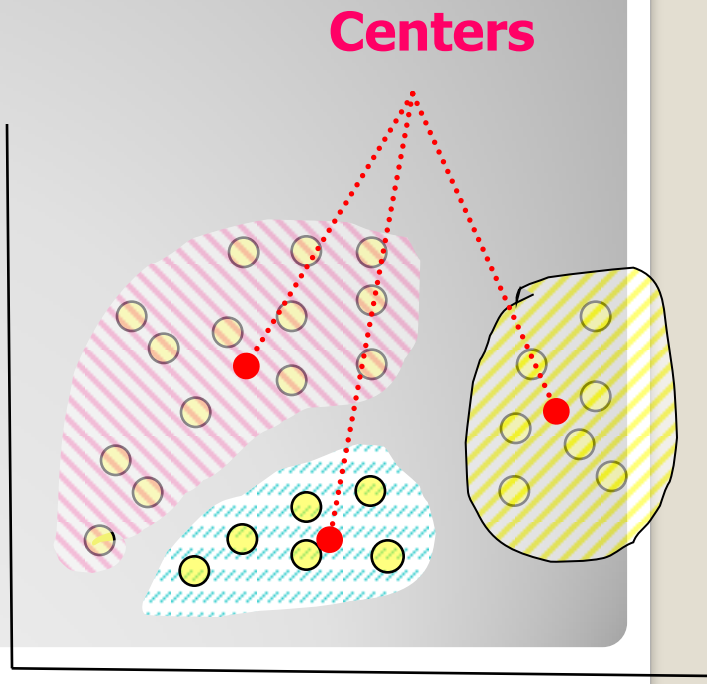


Clustering: Methods

- Partitioning-based clustering
  - **K-means clustering**
  - K-medoids clustering
  - EM (expectation maximization) clustering
- Density-based clustering
  - Separate regions of dense points by sparser regions of relatively low density

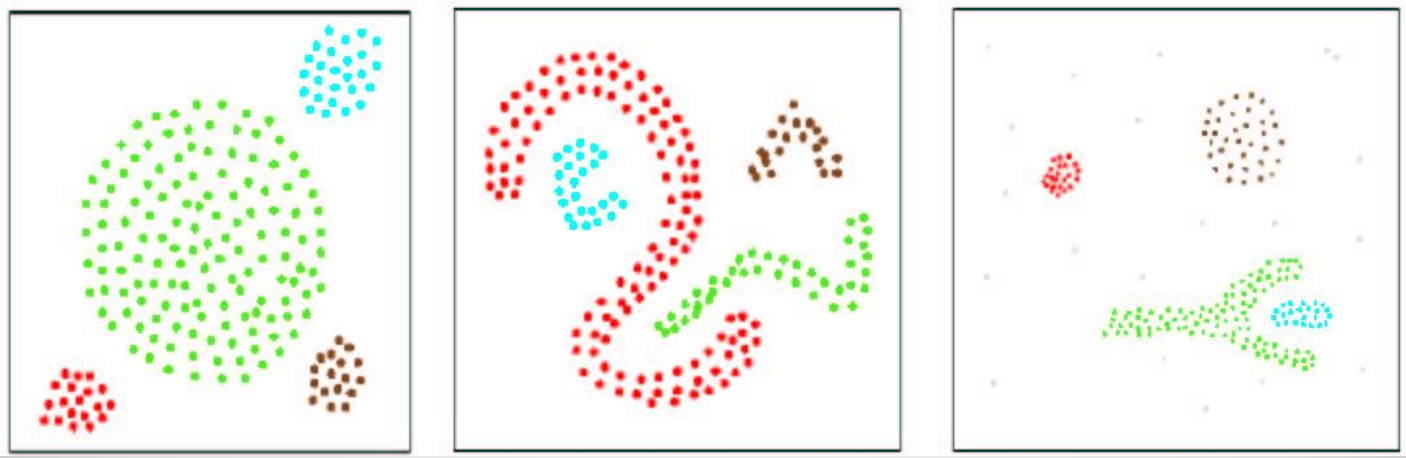
Partitioning-based Clustering: K-Means

- Goal: minimize sum of square of distance
  - Between each point and centers of the cluster.
  - Between each pair of points in the cluster
- Algorithm:
  - Initialize K cluster centers
    - random, first K, K separated points
  - Repeat until stabilization:
    - Assign each point to closest cluster center
    - Generate new cluster centers
    - Adjust clusters by merging or splitting



Density-Based Clustering

- **A cluster**: a connected dense component
- **Density**: the number of neighbors of a point
- Can find clusters of arbitrary shape



Clustering: Application 1

- Market Segmentation:
  - **Goal:** divide a market into distinct subsets of customers, where any subset may conceivably be selected as a market target.
  - **Approach:**
    - Collect different attributes of customers, based on their demographical and lifestyle related information.
    - Find clusters of similar customers.
    - Evaluate the clustering quality by observing buying patterns of customers in the same cluster vs. those from different clusters.

Clustering: Application 2

- **Document Clustering**
  - Clustering Points: 3204 Articles of Evening Post.
  - Similarity Measure: How many words are common in these documents.

Category	*Total Articles*	*Correctly Placed*
Financial	555	**364**
Foreign	341	**260**
National	273	**36**
Metro	943	**746**
Sports	738	**573**
Entertainment	354	**278**

Road Map

- Data mining overview
- Data mining tasks
  - Classification (supervised learning)
  - Clustering (unsupervised learning)
  - **Association rule discovery**
- Summary

Market Basket Analysis

- Consider shopping cart filled with several items
- **Market basket analysis** tries to answer the following questions:
  - Who makes purchases?
  - What do customers buy together?
  - In what order do customers purchase items?

Market Basket Analysis

- Co-occurrences
  - 80% of all customers purchase items X, Y and Z **together**.
- **Association rules**
  - 60% of all customers who purchase X and Y **also buy** Z.
- Sequential patterns
  - 40% of customers who first buy X also purchase Y **within** three weeks.

Association Rule Discovery: Definition

- Giving a set of records, each of which contain some number of items
  - Produce dependency rules, which predict occurrence of an item based on occurrences of other items.

Market-Basket transactions

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

Example Association Rules

{Diaper} → {Beer},
{Beer, Bread} → {Milk},
{Milk, Bread} → {Eggs, Coke}

Association Rule Discovery: Applications

- Supermarket shelf management.
  - **Goal:** To identify items that are bought together by sufficiently many customers.
  - **Approach:**
    - Process the point-of-sale data collected with barcode scanners.
    - find dependencies among items.
  - **A classic rule:**
    - If a customer buys diaper and milk, then he is very likely to buy beer.
    - So, **don't be surprised** if you find six-packs stacked next to diapers!

Summary

- Data mining
  - Discovering interesting patterns from large database
  - A natural evolution of database technology, in great demand, with wide applications
- Data mining functionalities
  - **Classification**
  - **Clustering**
  - **Association rule discovery**