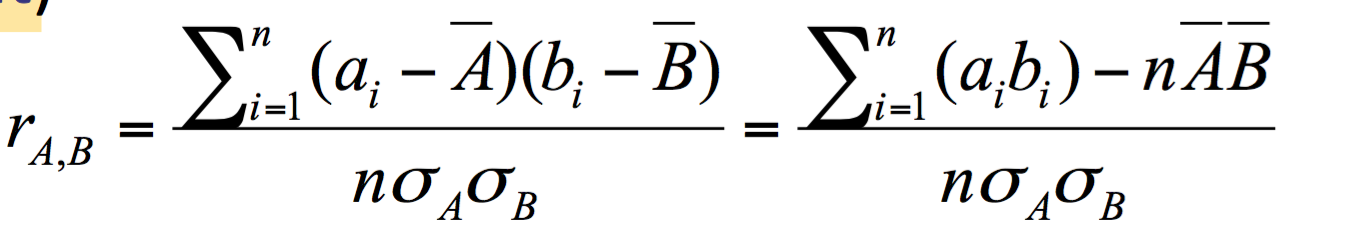
**1. Data Preprocessing**

**Correlation Analysis(Nominal)**

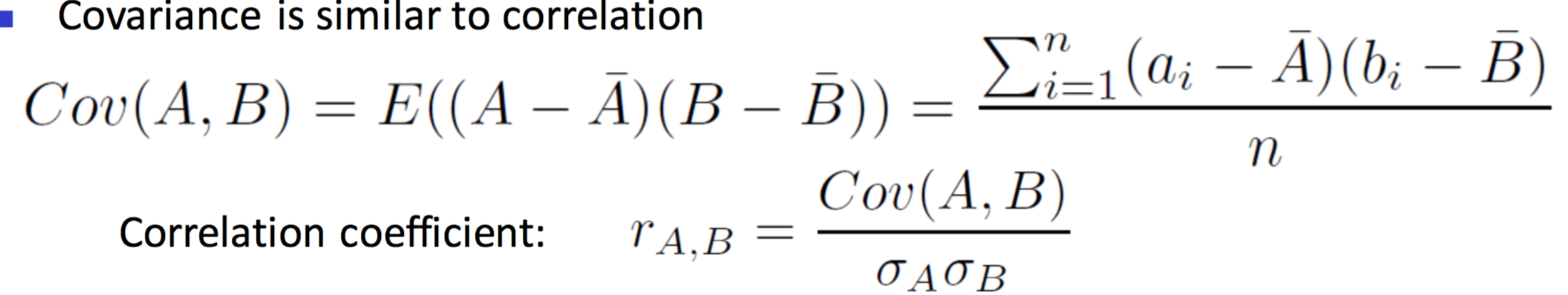
**X2(chi-square):** , Larger, more likely the variables are realted

**Correlation Analysis(Numeric)(Pearson’s product )**

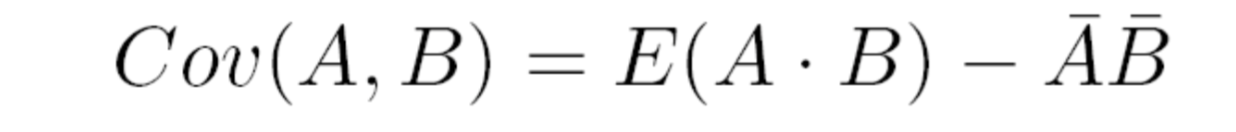
****

>0, positively correlated. A increase as B

**Covariance**

****

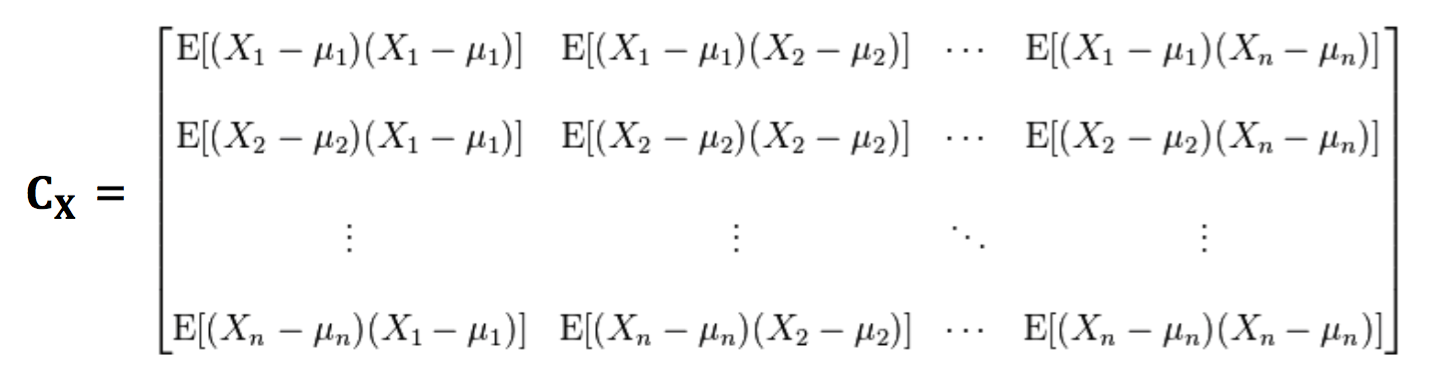
>0, positive covariance, A,B both tend to be larger than expected values.

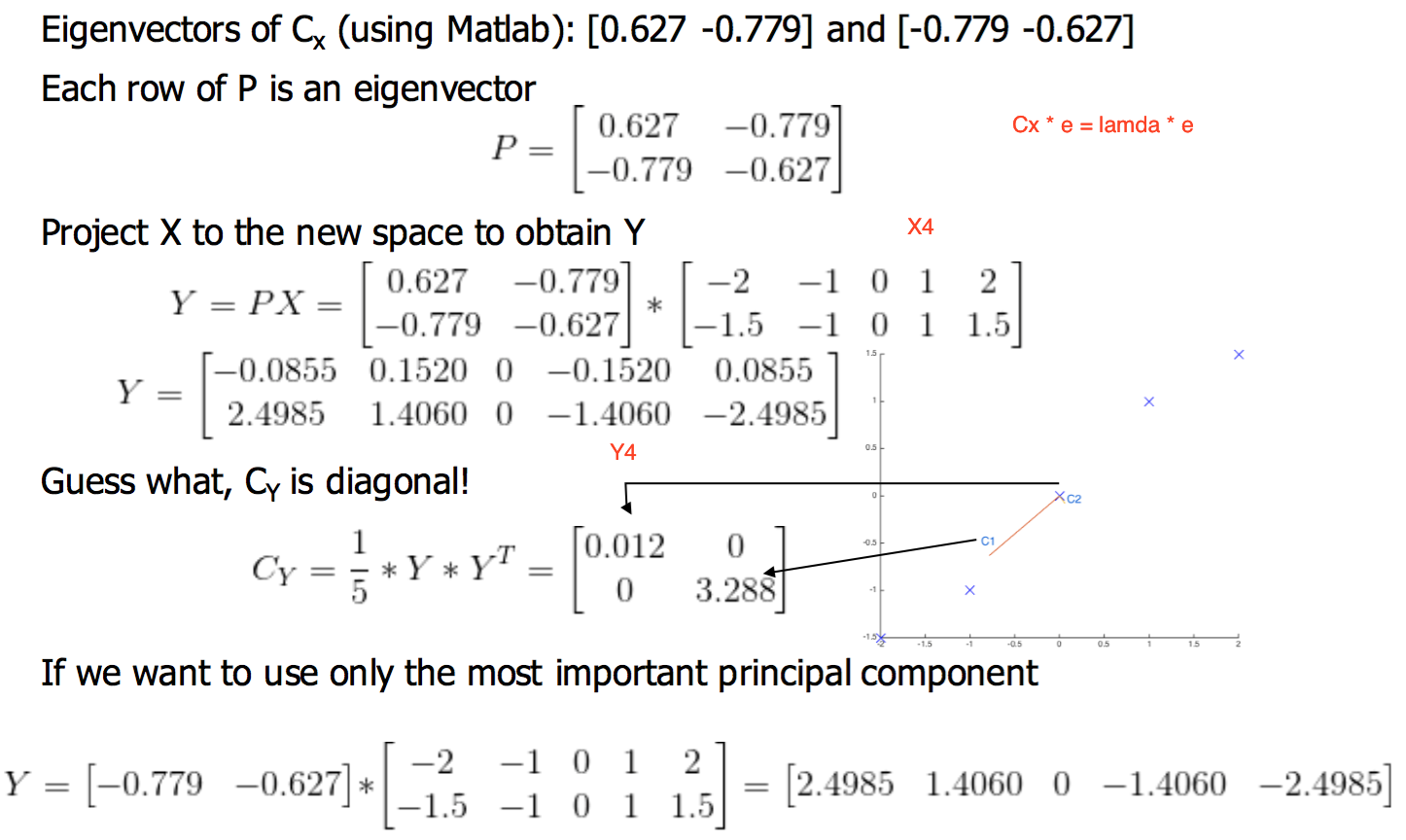


Min\_Max normal, v’ = (v-min)/(max-min)

Z-score normal,v’ = (v-u)/

**PCA:max variances of new dims, min co-variance between new dims, cov matrix of Cy be diagonal**

****



**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

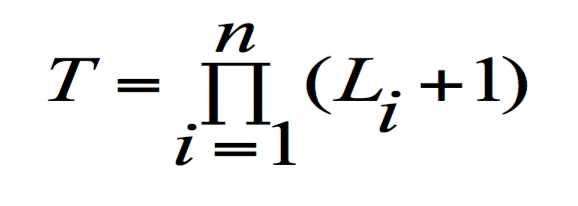
**2. Data Warehousing**

subject-oriented, integrated, time-variant and nonvolatile collection of data in support of management’s decision-making process

**distributive**: sum,count,min,max

**algebraic**: avg,min\_N,max\_N,stddev

**Holistic**: median,mode,rank



Roll-up: summarize data, Drill down, **Slice**: reduce dim, **dice:** reduce data elements**, select**, pivot: change perspectives

Support = #T(XUY)/#T(Total) ,Confidence = #T(XUY)/#T(X) . **Closed**: X is freq and no supper pattern X’ has the same support as X. support(super(x))<support(x) **Max**:X is freq, no super X’ that is also freq, support(super(x))<minsup

**Lift =** P(AUB)/P(A)P(B)

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**3. Mining Frequent Patterns**

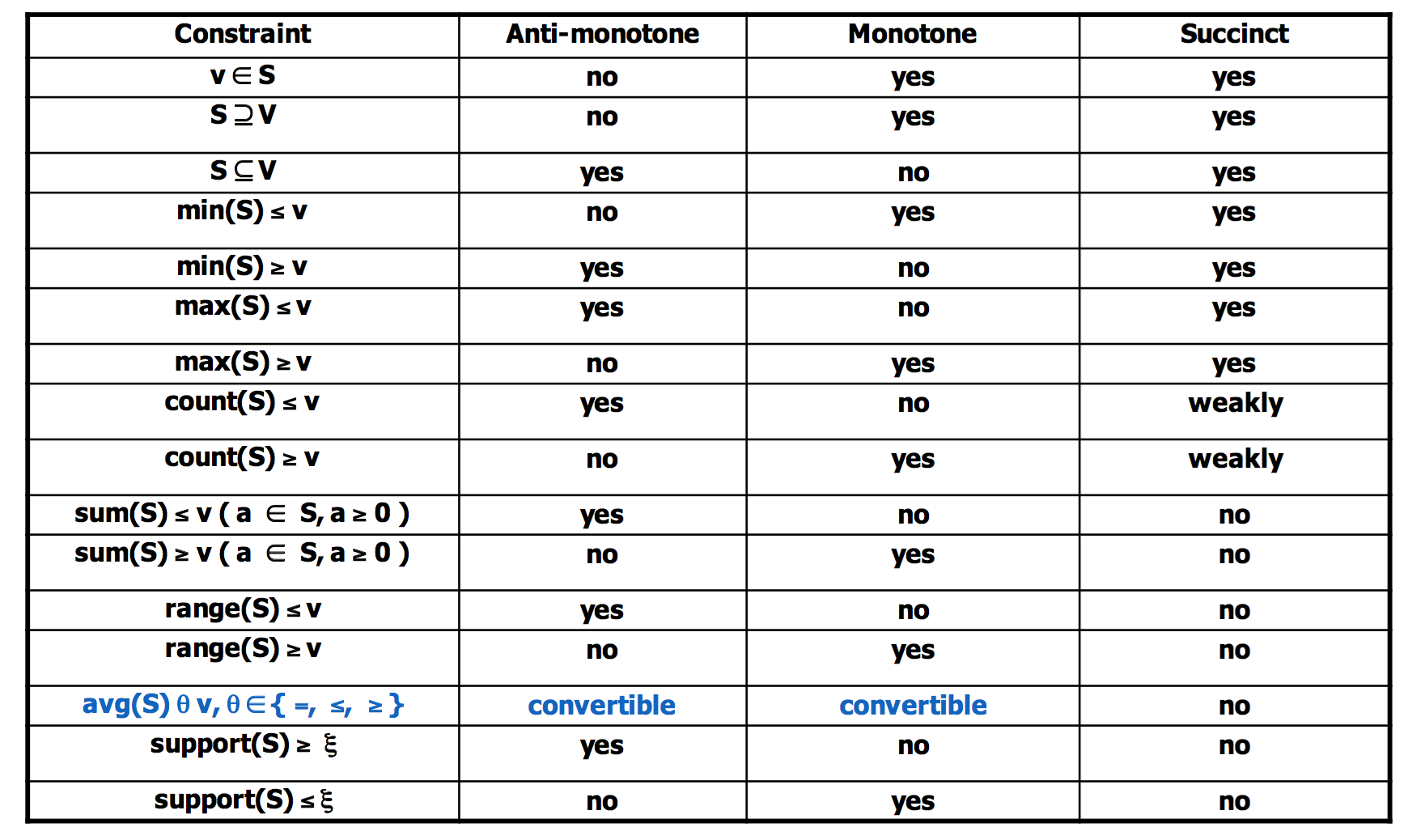
**Constraints-based Frequent Pattern Mining**

**Anti-monotonic:** Satisfying pattern implies satisfying sub-patterns, max(S) ≤v, apriori

**Monotonic:** Satisfying pattern implies satisfying super-patterns, **min(S) ≤V,** FP-growth

**Succinct:** Satisfying patterns all based on one set A, max(S) ≤V (A = {x|x≤v},then S⊆A)

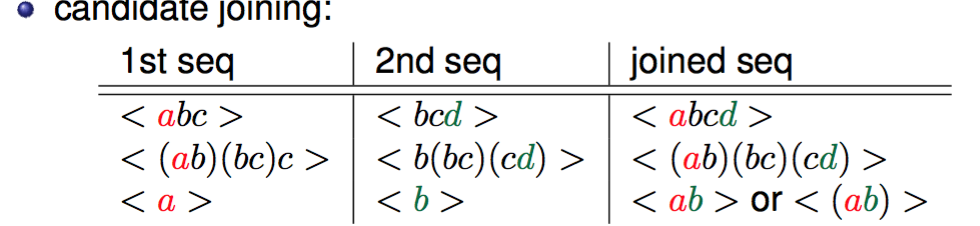
**Convertible:** constrains become anti-monotonic or/and monotonic if properly ordering items, ex, avg(S) ≥V(descending order, anti-mo)



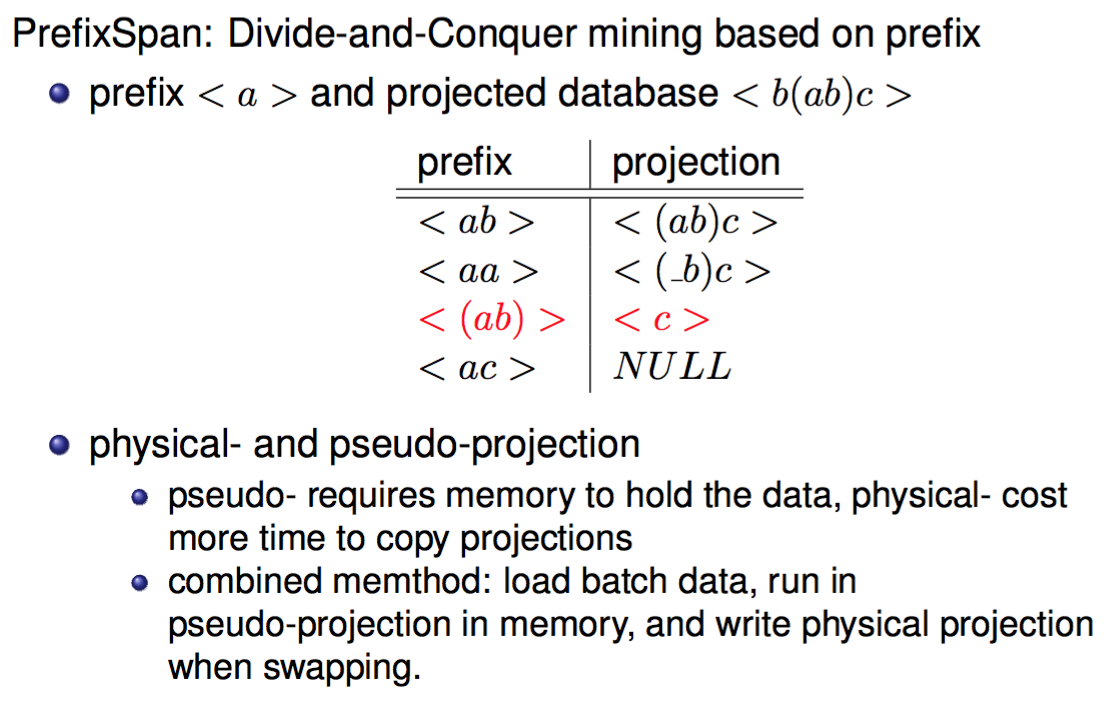
Sequential Pattern Mining

GSP(Apriori): n\*n + n\*(n-1)/2, length-2 candidates

Length-100 sequential pattern



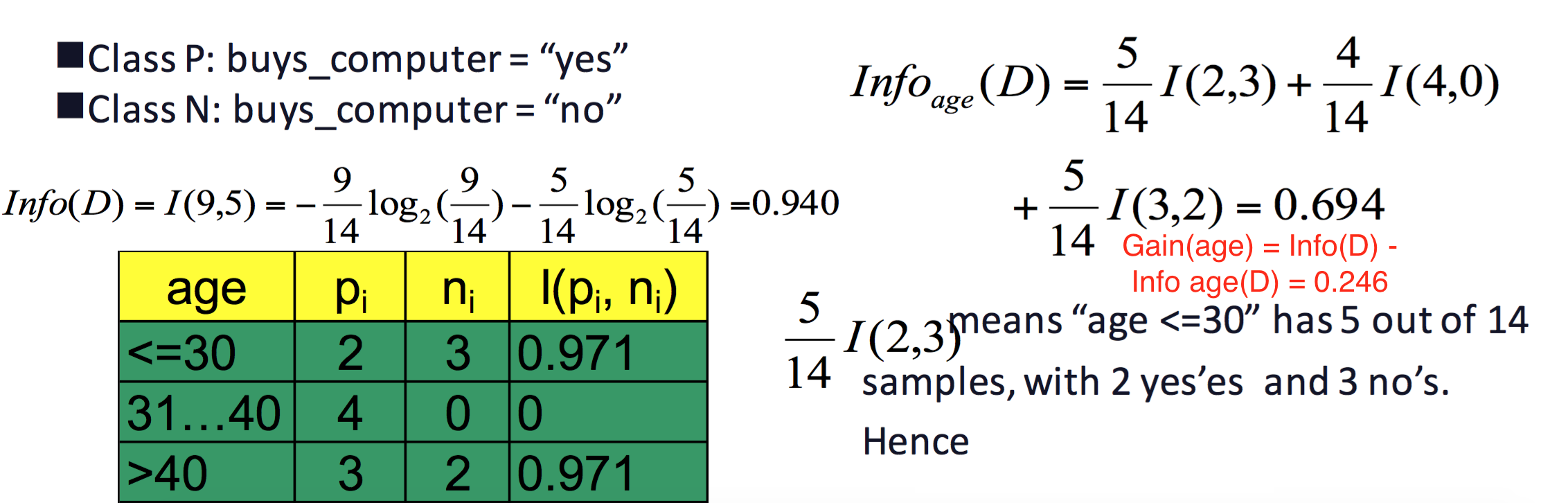
PrefixSpan:

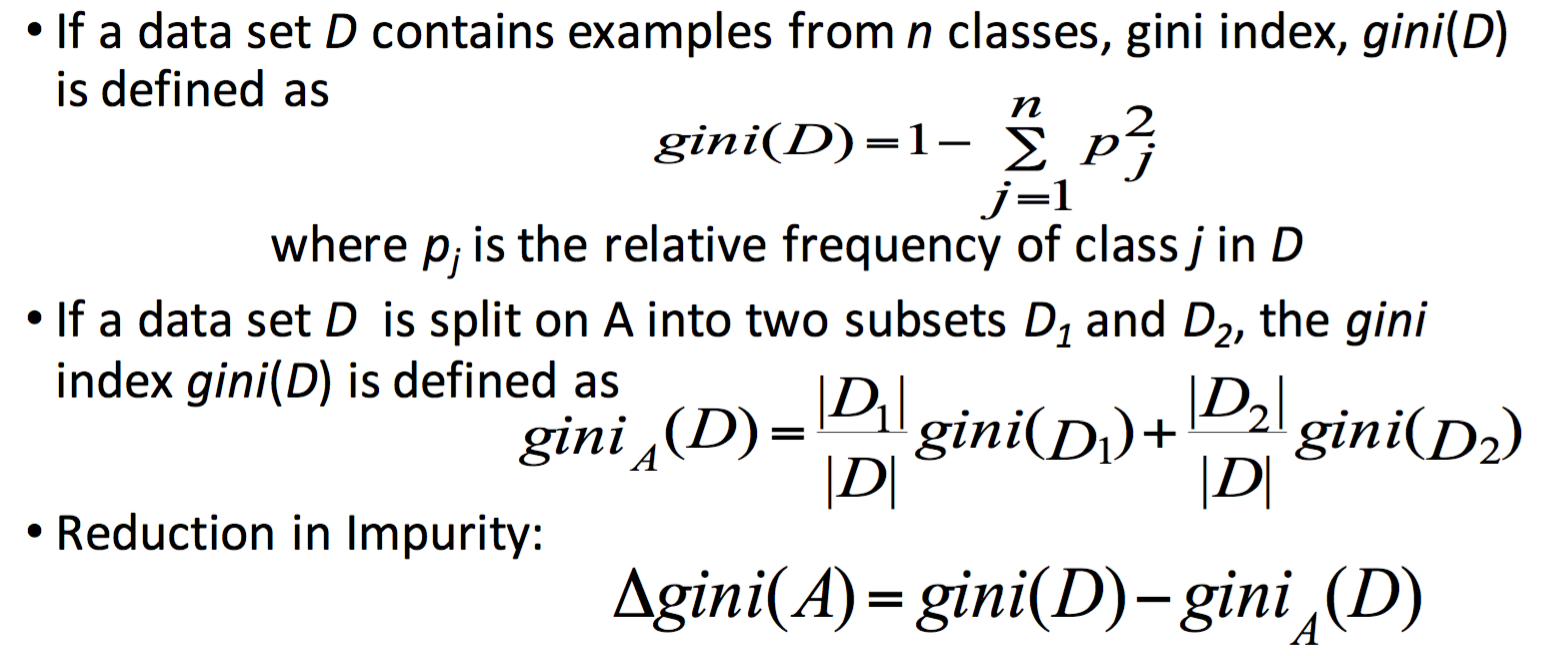


**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**4. Classification**

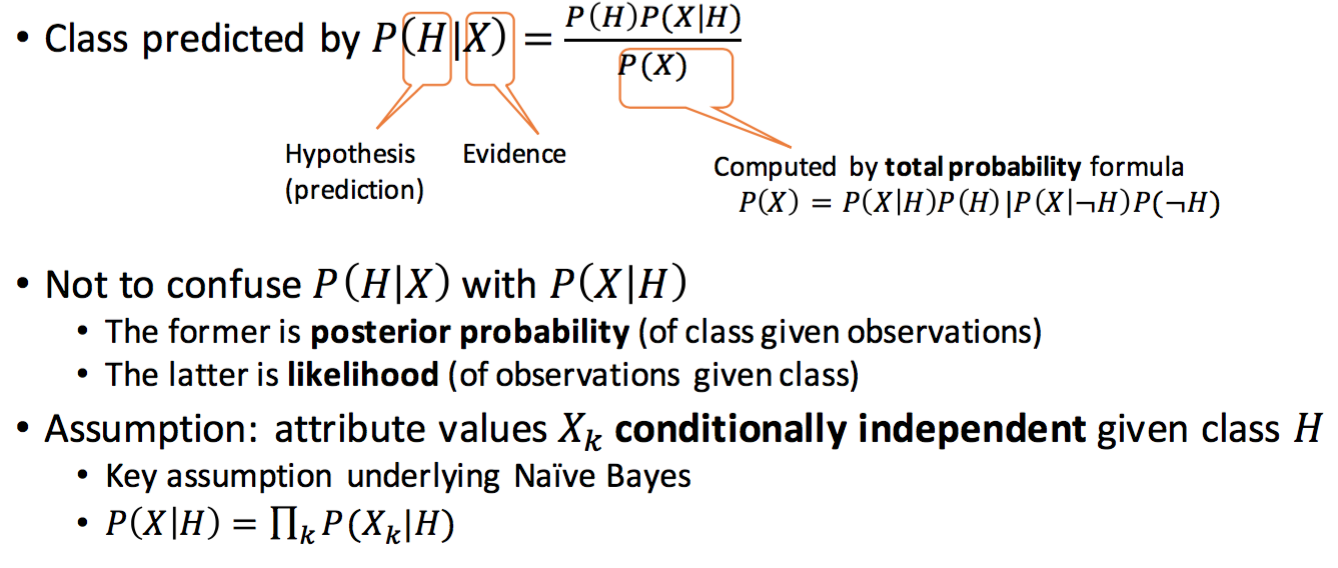
Decision Tree





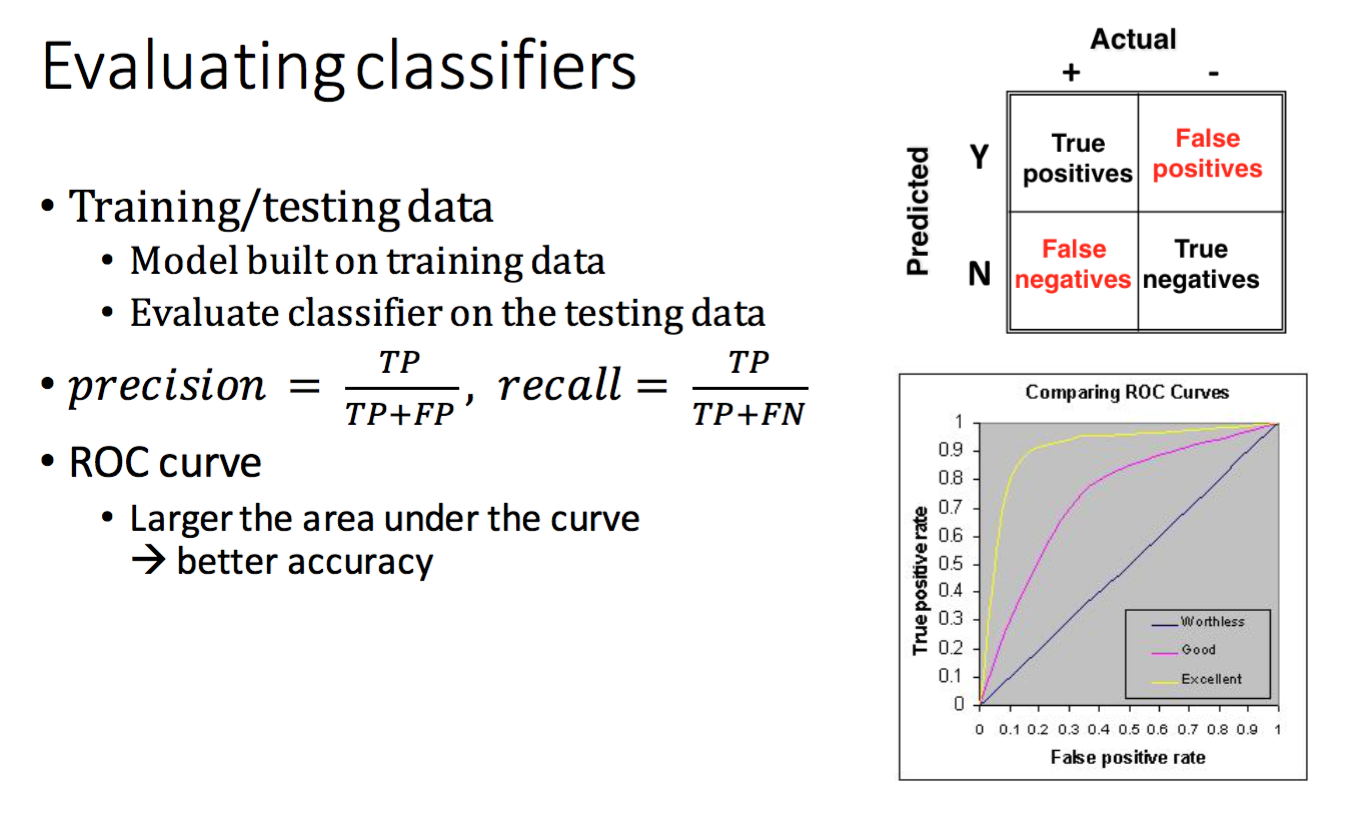
Bayes

Naïve Bayes



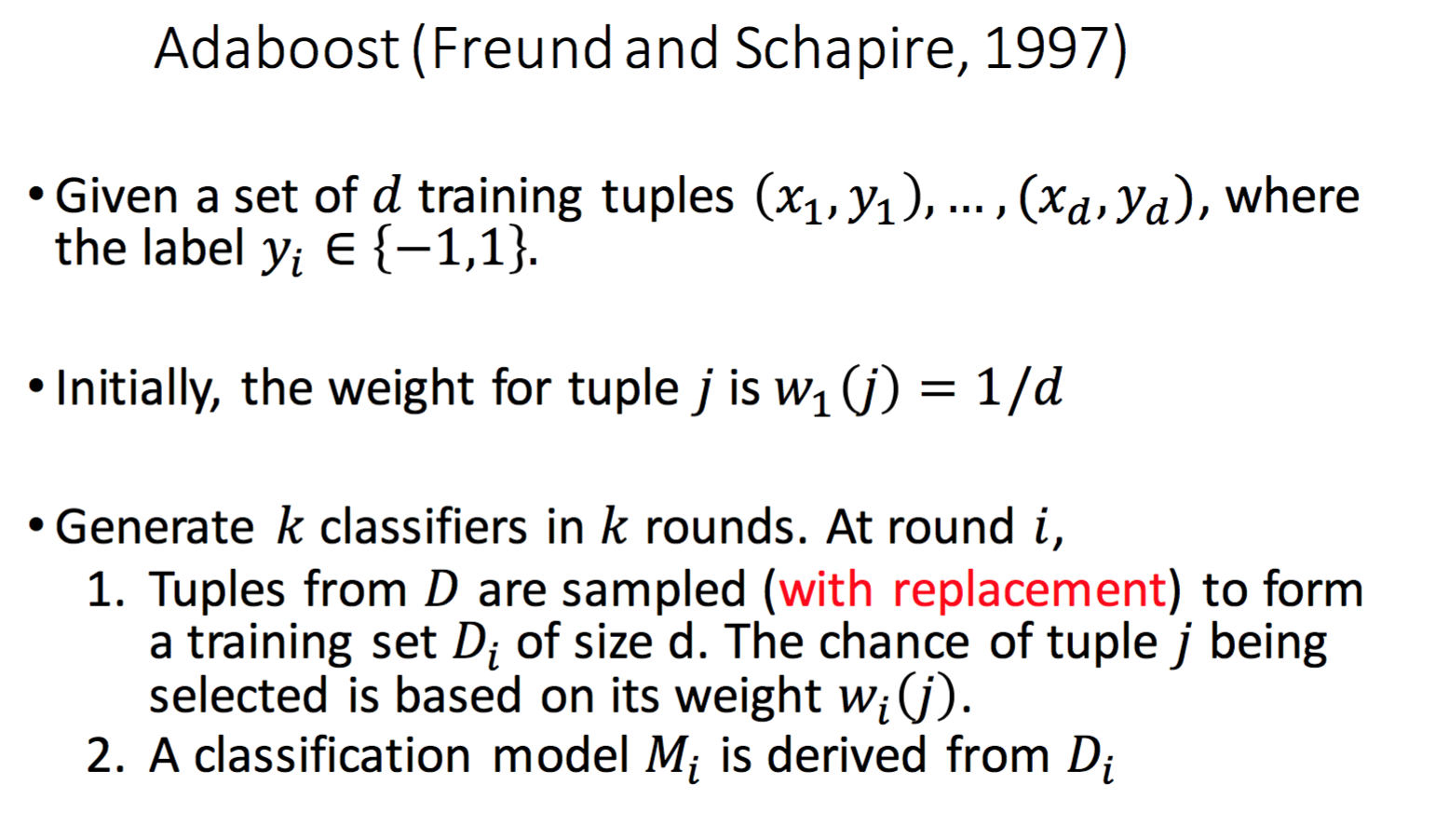
Smoothing: add 1 to each case

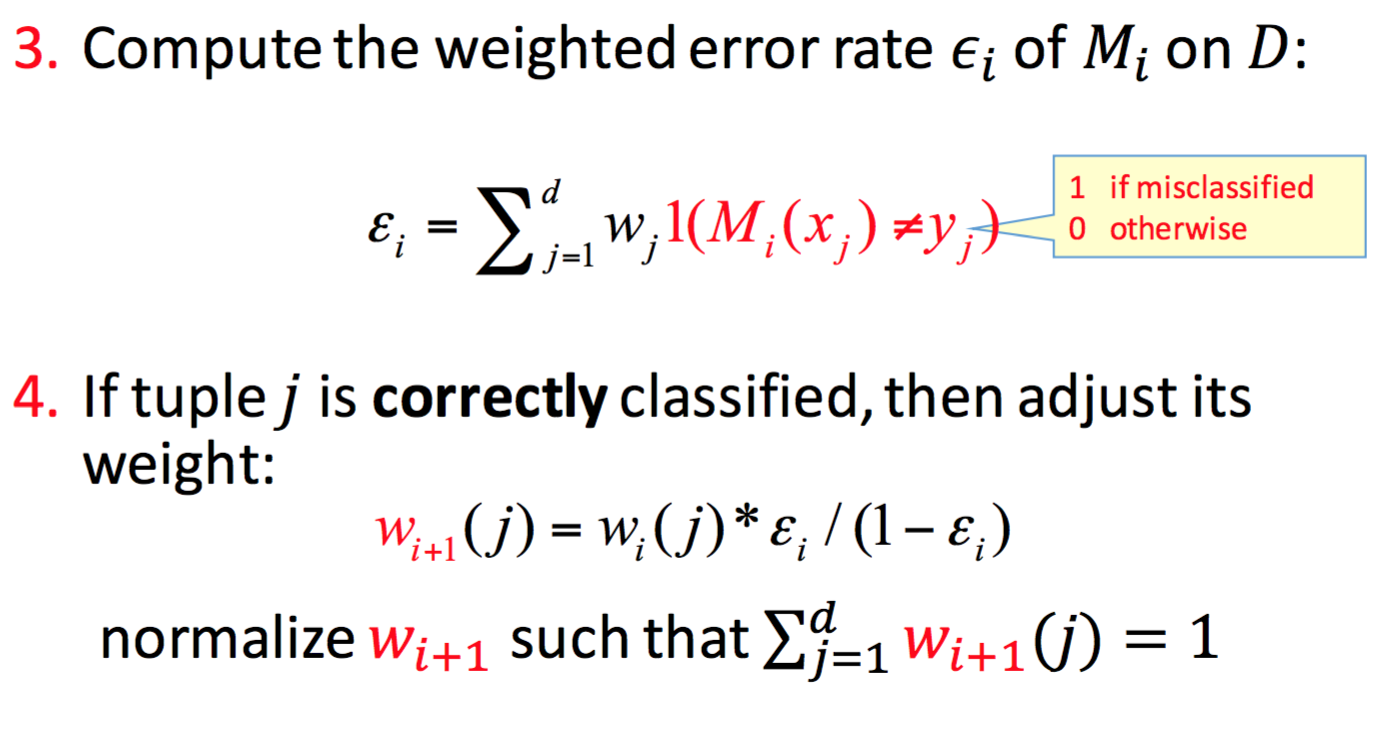
Model Evaluation and Selection

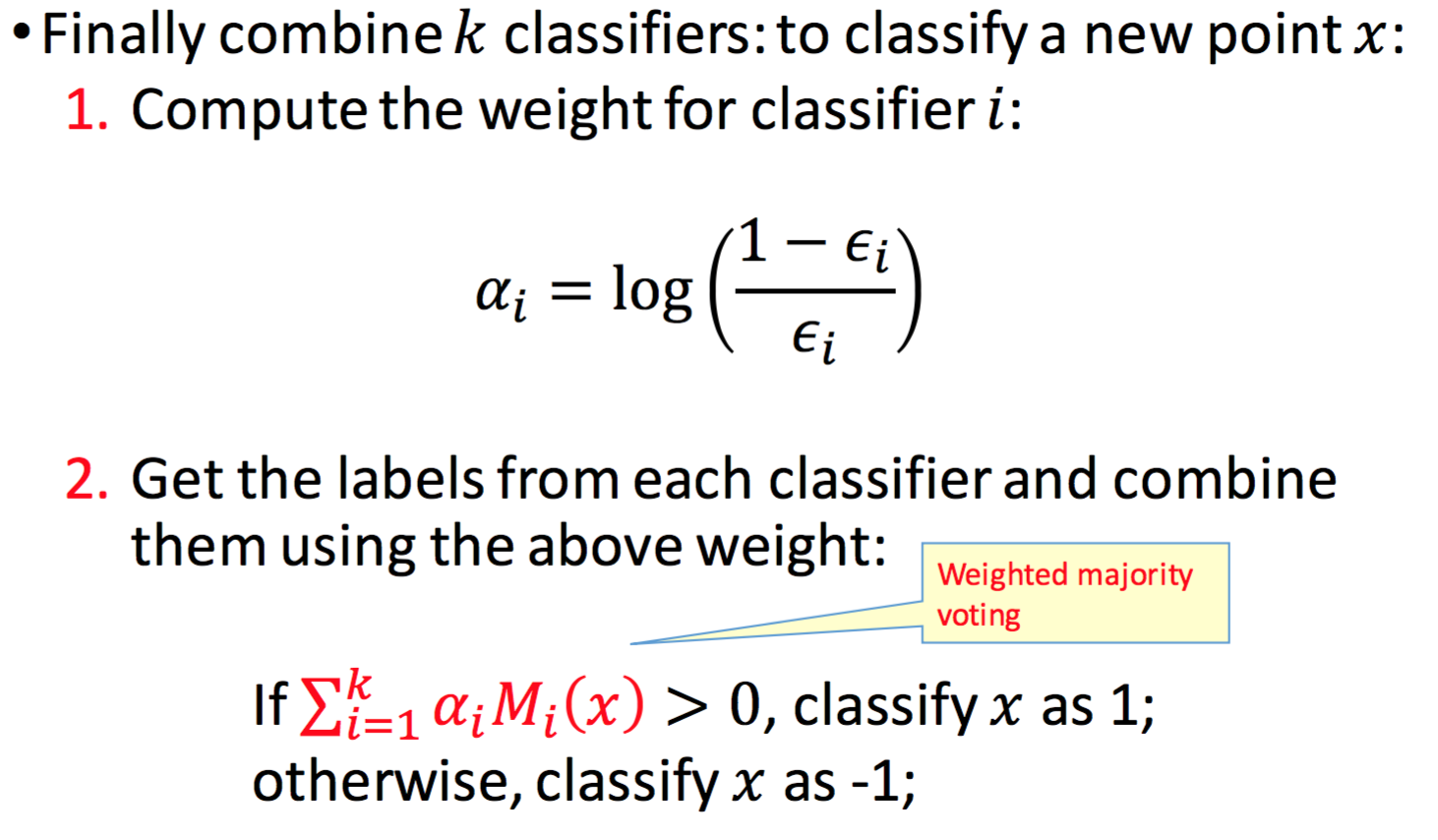


Improve Accuracy: Ensemble Methods

**AdaBoost**

****

****

****

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**5. [Clustering]**

Partitioning Method

K-means: O(tkn). T is iterations. K is clusters. N is objects. t,k<<n. Issues: 1. Terminates at a local optimal: Initialization can be important to find high-quality clusters 2. Sensitive to noise data and outliers

K-Medoids (PAM): Robust but expensive. does not scale well for large dataset. O(k(n-k)2).

Hierarchical Methods：

weak:1. Can never undo what was done previously2. Do not scale well O(n2)

AGNES(agglomerative nesting): O(n2). single-link method and dissimilarity matrix. Merge nodes that have the least dissimilarity. Go on in non-descending fashion. Eventually all nodes belong to same cluster.

**Single link**: smallest distance between an element in one cluster and an element in the other

**Complete link:** largest distance ~

**Average:** avg distance ~

**Centroid:** distance between the centroids of two cls

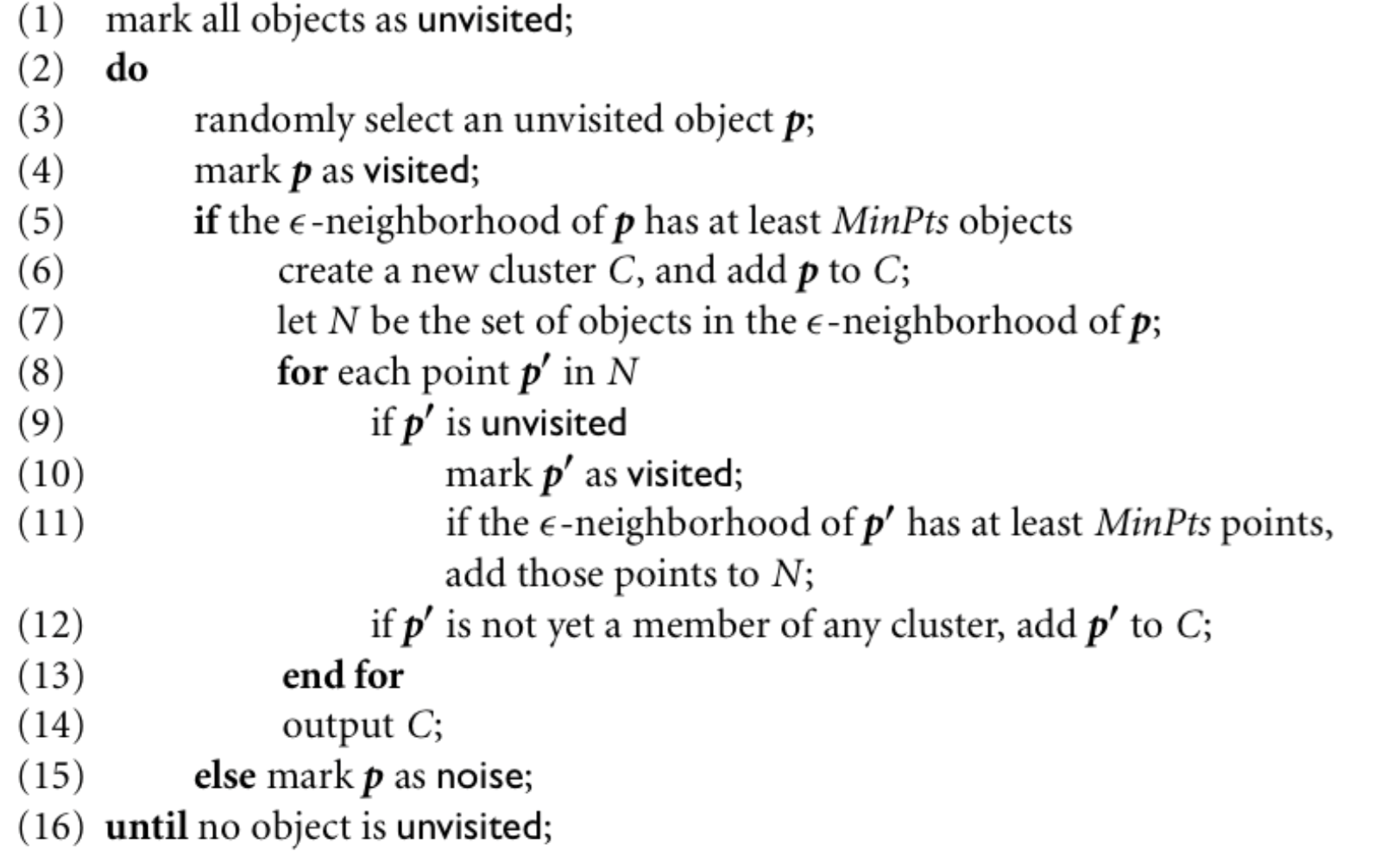
**Medoid:** distance between the medoids of two cls

Density-based Method: Clustering based on density(local cluster criterion), such as density-connected points

**Major features:** 1. Discover clusters of arbitrary shape:dense ara 2. Discover noise: if points do not belong to a dense area 3. Need density parameters as termination condition 4. Require one scan

**Two params: ε:** maximum radius of the neighborhood **MinPts:** Minimum # of pts in an eps-neighborhood of that point

DBSCAN: pros: support arbitrary shape w/ noise. Cons: sensitive to params



B-cubed Precision & Recall:

Pi = # correct in outputchain/# all in outputchain

Ri = # correct in outputchian/# in truth chain

Final P =