Information Retrieval
& Data Mining

Information Retrieval & Data Mining
Universität des Saarlandes, Saarbrücken
Winter Semester 2013/14

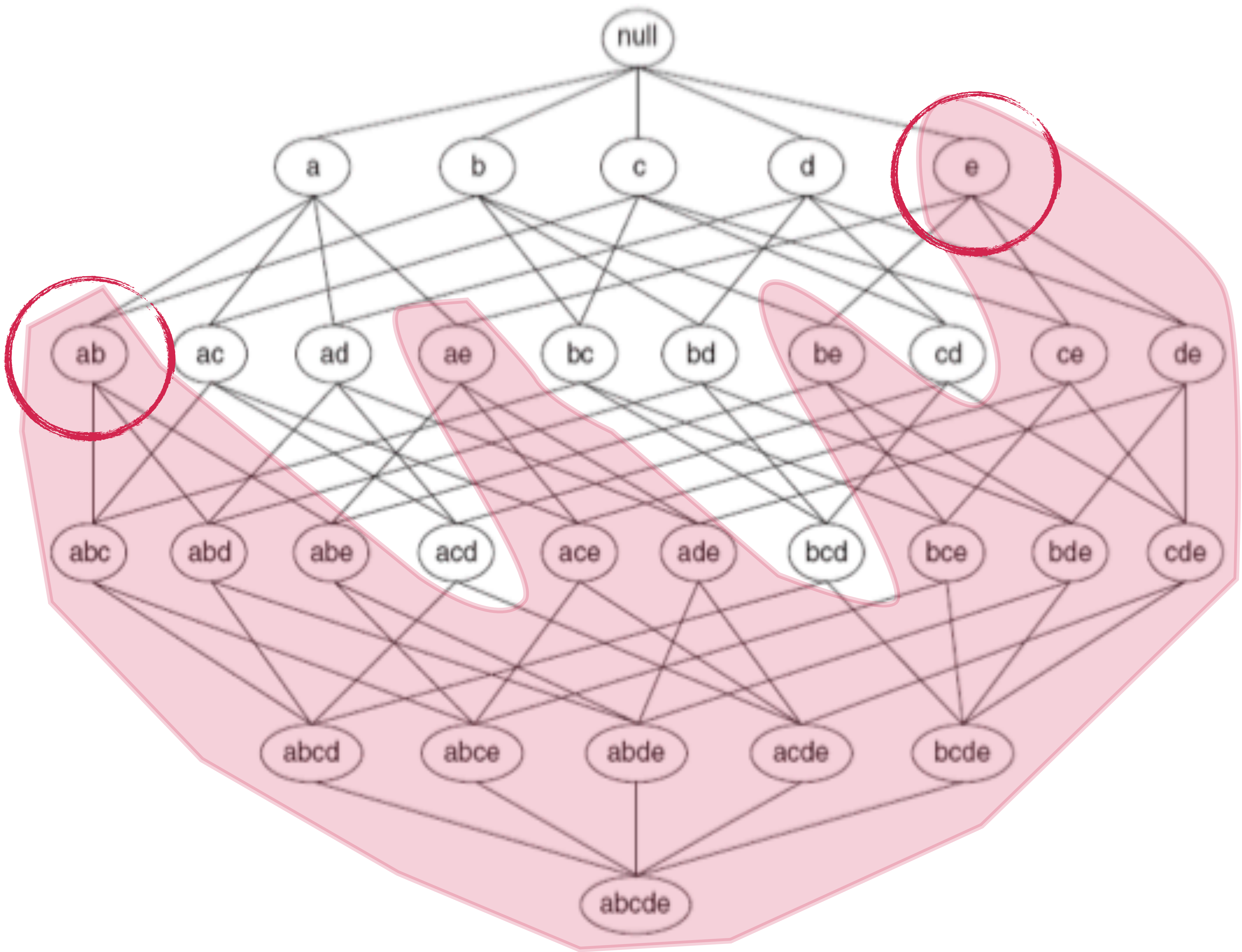
Chapter VII: Frequent Itemsets and Association Rules*

1. **Definitions: Frequent Itemsets and Association Rules**
2. **Algorithms for Frequent Itemset Mining**
   - **Monotonicity and candidate pruning, Apriori, ECLAT, FPGrowth**
3. **Association Rules**
   - **Measures of interestingness**
4. **Summarizing Itemsets**
   - **Closed, maximal, and non-derivable itemsets**

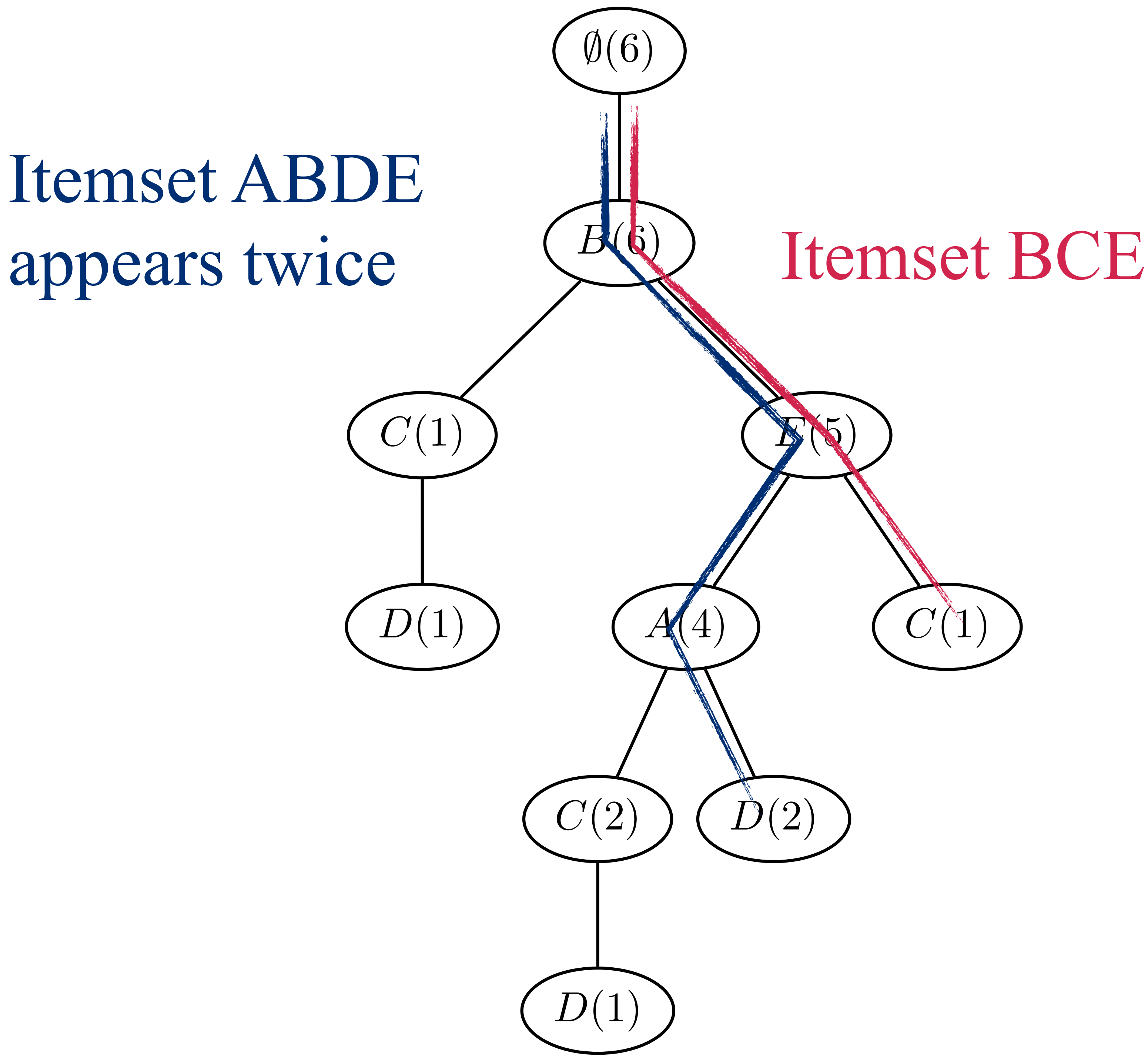
*Zaki & Meira, Chapters 10 and 11; Tan, Steinbach & Kumar, Chapter 6

Example of pruning itemsets

If {e} and {ab} are infrequent



FP-tree example



From Figure 8.9 of Zaki & Meira

Pseudo-code for generating association rules

Algorithm 8.6: Algorithm AssociationRules

AssociationRules ($\mathcal{F}$, *minconf*):

foreach $Z \in \mathcal{F}$, *such that* $|Z| \geq 2$ **do**
    $\mathcal{A} \leftarrow \{X \mid X \subset Z, X \neq \emptyset\}$
    while $\mathcal{A} \neq \emptyset$ **do**
        $X \leftarrow$ maximal element in $\mathcal{A}$
        $\mathcal{A} \leftarrow \mathcal{A} \setminus X$ `// remove` X `from` $\mathcal{A}$
        $c \leftarrow sup(Z)/sup(X)$
        if $c \geq minconf$ **then**
            print $X \longrightarrow Y$, $sup(Z)$, c
        else
            $\mathcal{A} \leftarrow \mathcal{A} \setminus \{W \mid W \subset X\}$ `// remove all subsets of` X `from` $\mathcal{A}$

Algorithm 8.6 of Zaki & Meira

Measures for association rules

Measure (Symbol)	Definition
Goodman-Kruskal (λ)	$\Big(\sum_j \max_k f_{jk} - max_k f_{+k}\Big)\Big/\Big(N - \max_k f_{+k}\Big)$
Mutual Information (M)	$\Big(\sum_i \sum_j \frac{f_{ij}}{N} \log \frac{N f_{ij}}{f_{i+} f_{+j}}\Big)\Big/\Big(-\sum_i \frac{f_{i+}}{N} \log \frac{f_{i+}}{N}\Big)$
J-Measure (J)	$\frac{f_{11}}{N} \log \frac{N f_{11}}{f_{1+} f_{+1}} + \frac{f_{10}}{N} \log \frac{N f_{10}}{f_{1+} f_{+0}}$
Gini index (G)	$\frac{f_{1+}}{N} \times (\frac{f_{11}}{f_{1+}})^2 + (\frac{f_{10}}{f_{1+}})^2] - (\frac{f_{+1}}{N})^2$ $+ \frac{f_{0+}}{N} \times [(\frac{f_{01}}{f_{0+}})^2 + (\frac{f_{00}}{f_{0+}})^2] - (\frac{f_{+0}}{N})^2$
Laplace (L)	$\Big(f_{11} + 1\Big)\Big/\Big(f_{1+} + 2\Big)$
Conviction (V)	$\Big(f_{1+} f_{+0}\Big)\Big/\Big(N f_{10}\Big)$
Certainty factor (F)	$\Big(\frac{f_{11}}{f_{1+}} - \frac{f_{+1}}{N}\Big)\Big/\Big(1 - \frac{f_{+1}}{N}\Big)$
Added Value (AV)	$\frac{f_{11}}{f_{1+}} - \frac{f_{+1}}{N}$

Tan, Steinbach & Kumar Table 6.12

Example of maximal frequent itemsets

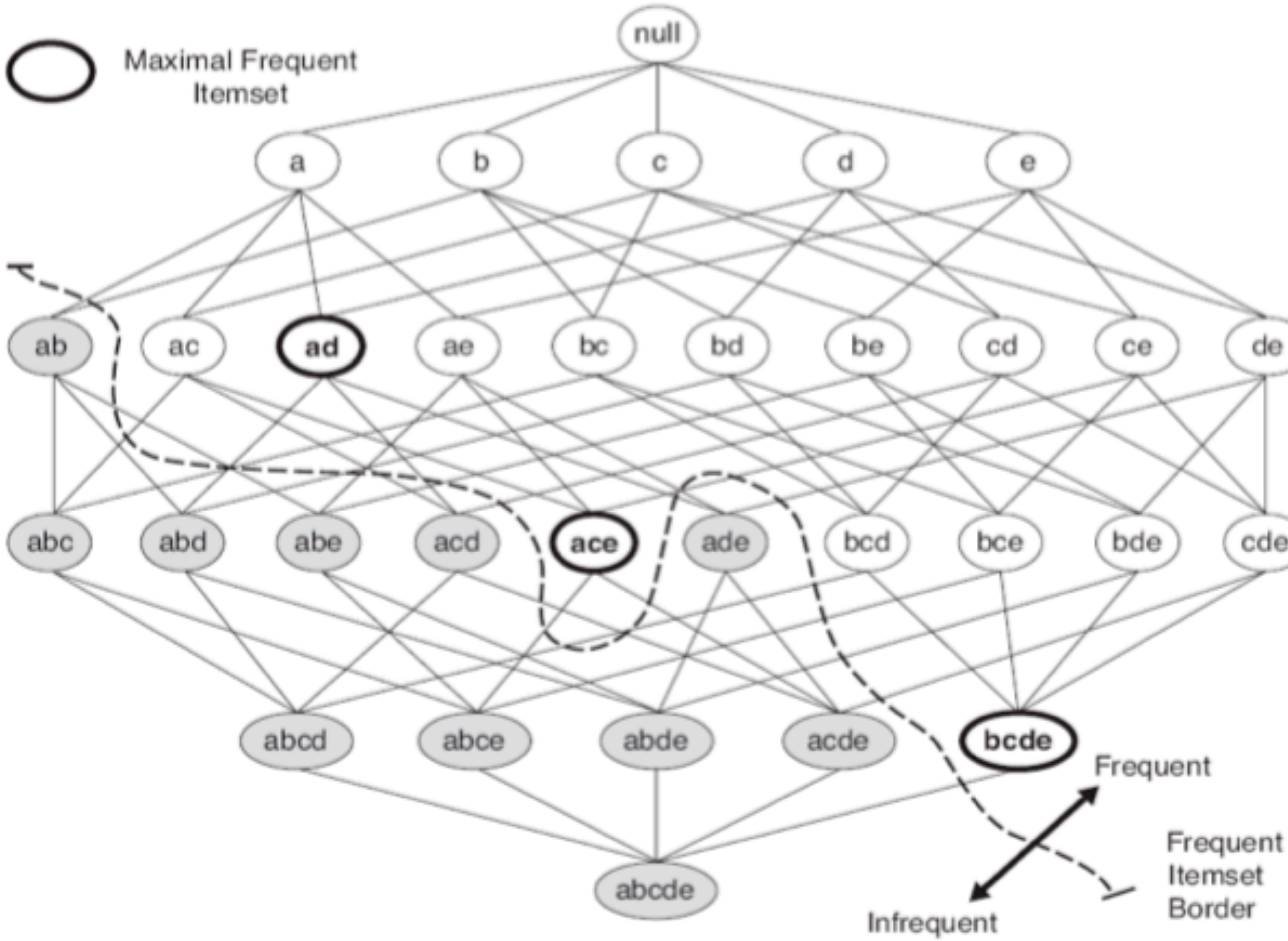
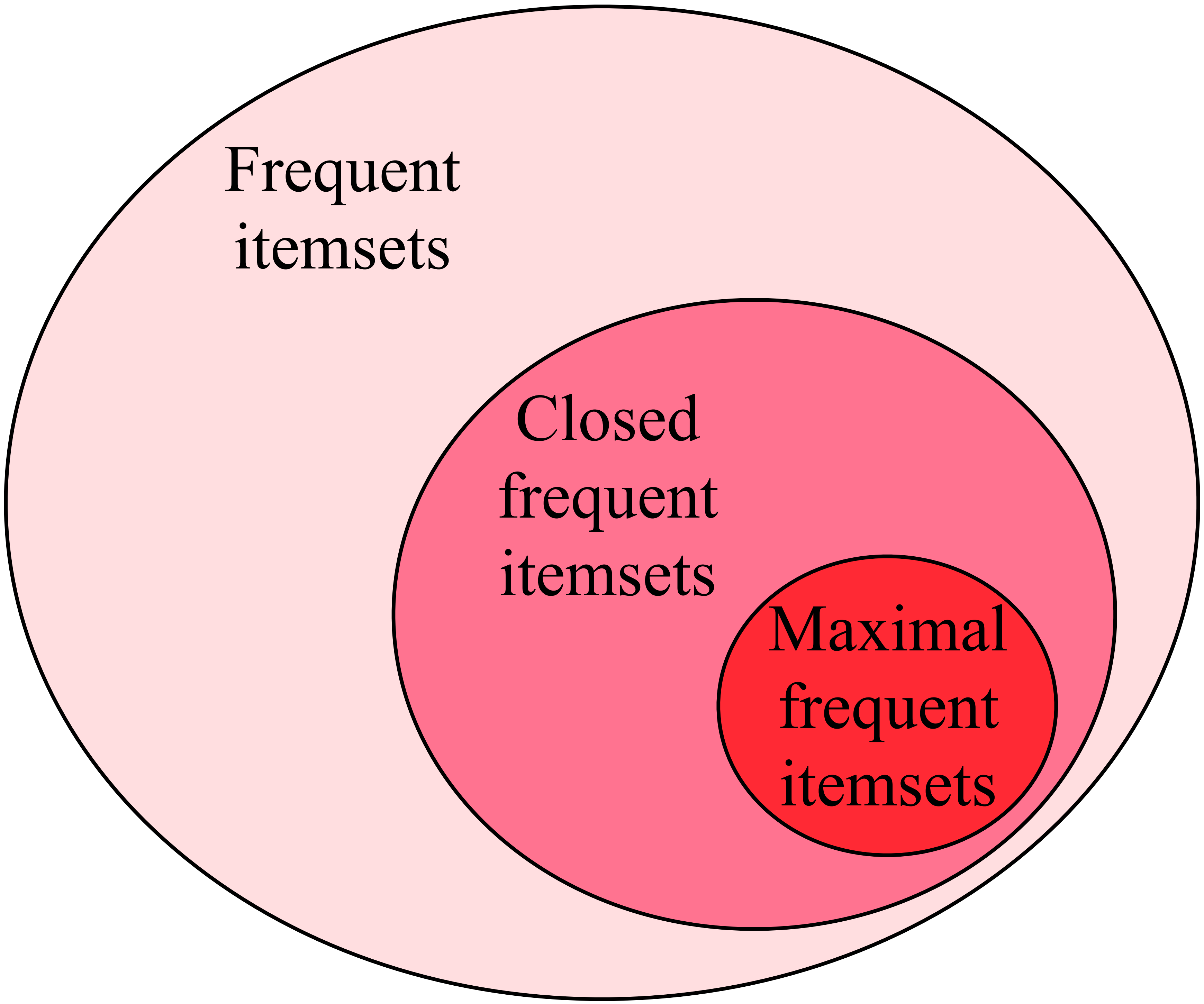


Figure 6.16. Maximal frequent itemset.

Itemset taxonomy



Non-Derivable Itemsets

- Let F be the set of all frequent itemsets. Itemset $X \in F$ is **non-derivable** if we cannot derive its support from its subsets.
  - We can derive the support of X from its subsets if, by knowing the supports of all of the subsets of X we can compute the support of X
- If X is derivable, it doesn't add any new information
  - Knowing just the non-derivable frequent itemsets, we can construct every frequent itemset
  - We only return itemsets that add new information on top of what we already knew

Chapter VIII: Clustering*

1. Basic idea
2. Representative-based clustering
    2.1. *k*-means
    2.2. EM-clustering
3. Hierarchical clustering
    3.1. Basic idea
    3.2. Cluster distances
4. Density-based clustering
5. Co-clustering
6. Discussion and clustering applications

*Zaki & Meira, Chapters 13–15; Tan, Steinbach & Kumar, Chapter 8

An iterative *k*-means algorithm

1. select k random cluster centroids
2. assign each point to its closest centroid and compute the error
3. **do**
   - 3.1. **for each** cluster C_i
     - 3.1.1. compute new centroid as $\boldsymbol{\mu}_i = \frac{1}{|C_i|}\sum_{\mathbf{x}_j \in C_i} \mathbf{x}_j$
   - 3.2. **for each** element $x_j \in U$
     - 3.2.1. assign x_j to its closest cluster centroid
4. **while** error decreases

The general EM clustering algorithm

- Initialization
  - Initialize parameters θ randomly
- Expectation step
  - Compute the posterior probability $P(C_i \mid x_j)$
  - Per Bayes's theorem

$$P(C_i \mid \boldsymbol{x}_j) = \frac{P(\boldsymbol{x}_j \mid C_i)P(C_i)}{\sum_{a=1}^{k} P(\boldsymbol{x}_j \mid C_a)P(C_a)}$$

- Maximization step
  - Re-estimate θ given $P(C_i \mid x_j)$
- Repeat E and M steps until convergence

The general EM algorithm

- A way to find maximum-likelihood parameters when the model depends on latent variables
  - In clustering, the latent variables are the cluster indicators
    - And the parameters are those for the distribution
- We're given data X, we assume there's some latent variables Z and parameters θ together with a log-likelihood function $L(\theta; X, Z)$
- In **E-step** we compute the expectation of L over Z given X and $\theta(t)$, $Q(\theta \mid \theta^{(t)}) = E_{Z|X,\theta^{(t)}}[L(\theta; X, Z)]$
- In **M-step** we maximize Q,

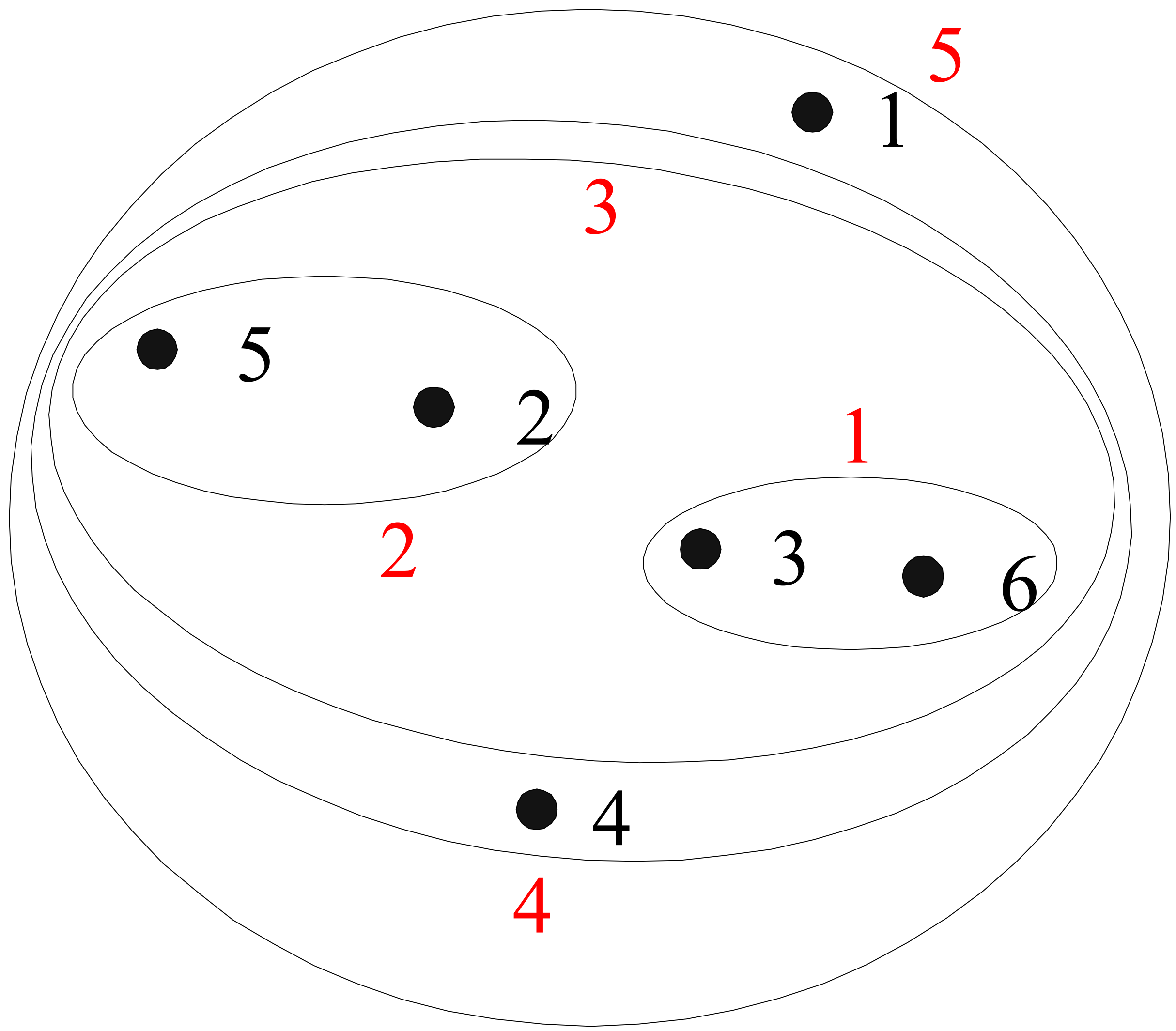
$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta \mid \theta^{(t)})$$

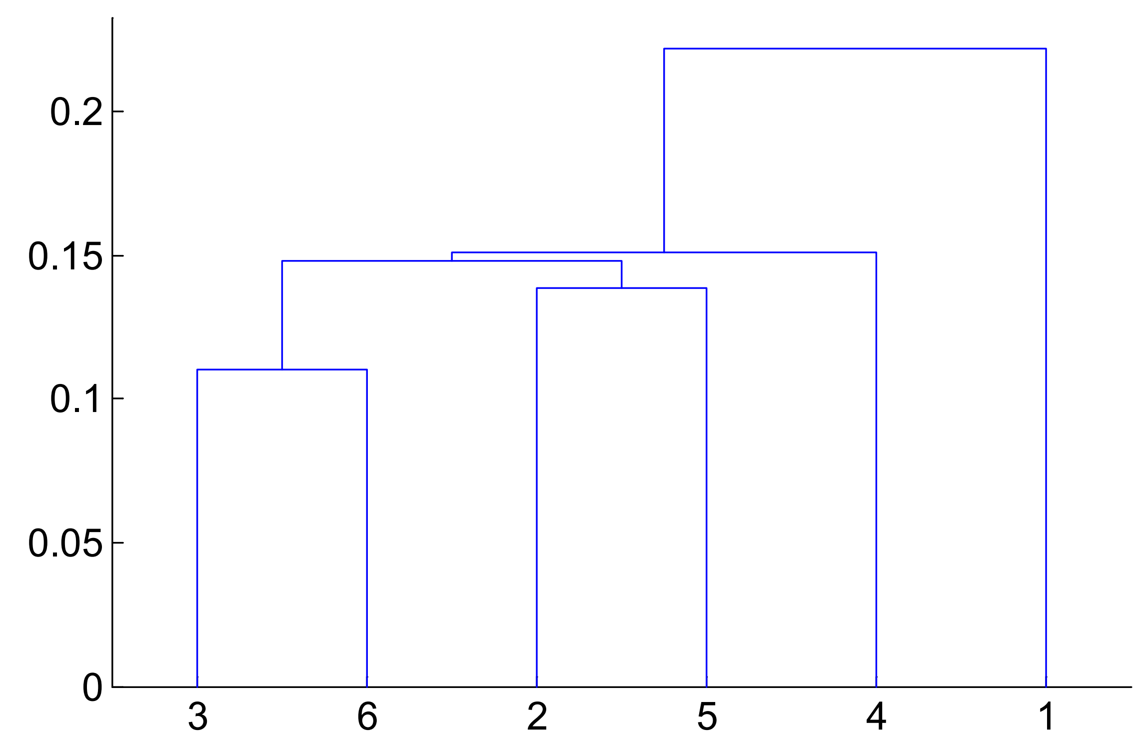
EM in IR&DM

- Latent topic models
  - Parameters for pLSI and LDA
- Hidden Markov models in IE
  - Parameters for the models
- Clustering
  - Parameters for the Gaussian distributions
  - *k*-means

Single link

- The distance between two clusters is the distance between the closest points
  - $d(B,C) = \min\{d(x,y) : x \in B \text{ and } y \in C\}$





Density-based clusters

- A **density-based cluster** is a maximal set of density connected points

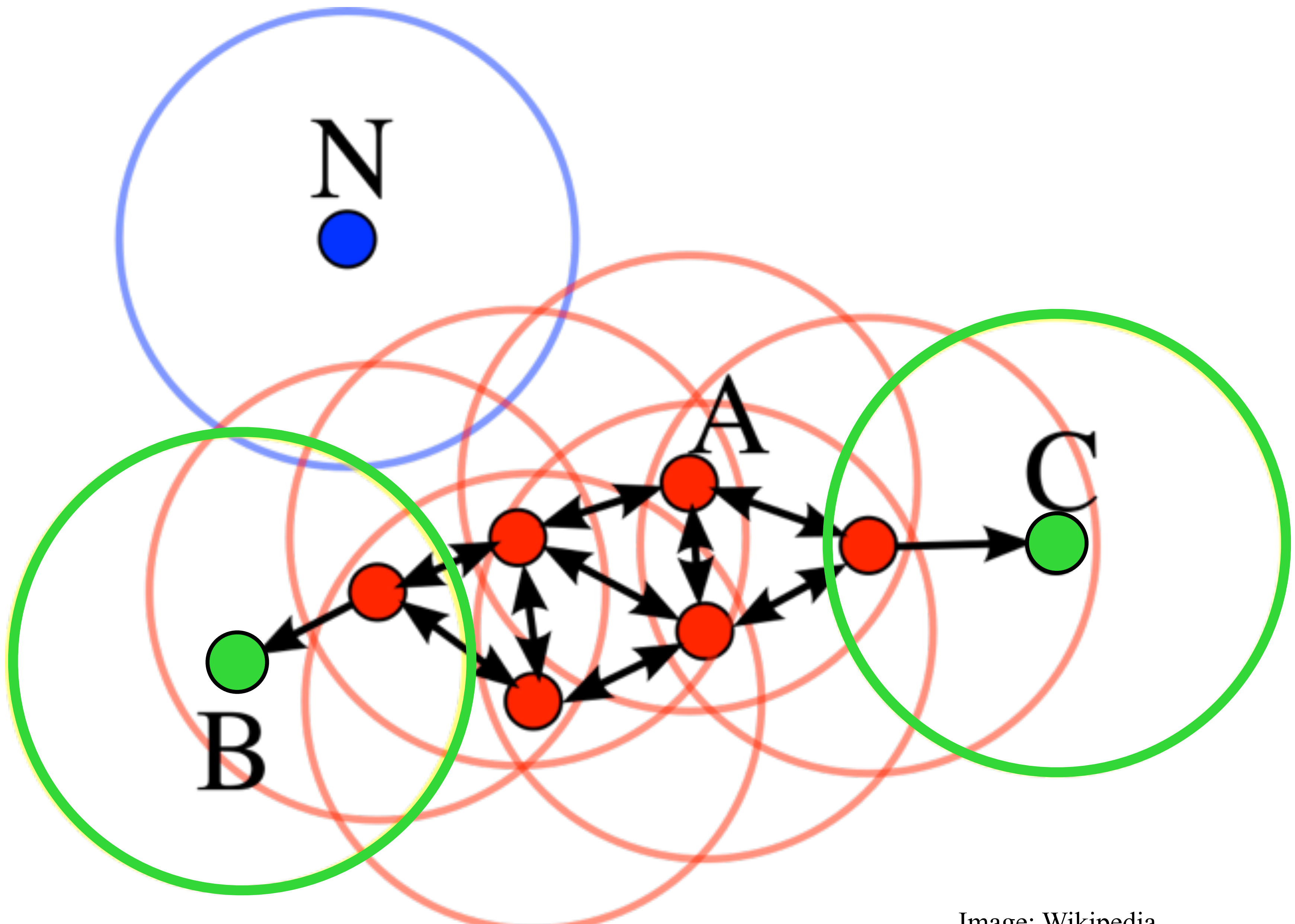


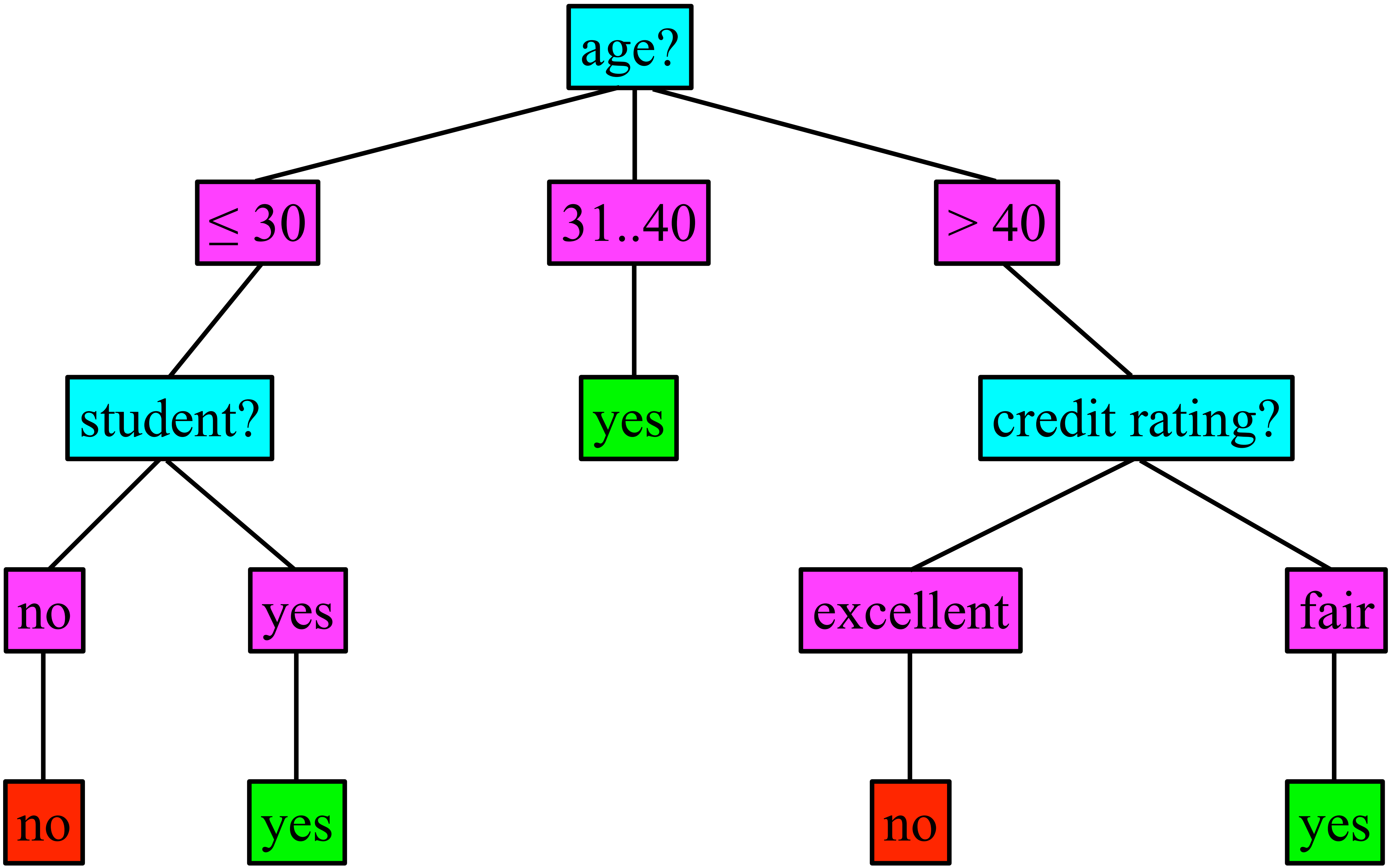
Image: Wikipedia

Chapter IX: Classification*

1. Basic idea
2. Decision trees
3. Naïve Bayes classifier
4. Support vector machines
5. Ensemble methods

* Zaki & Meira: Ch. 18, 19, 21 & 22; Tan, Steinbach & Kumar: Ch. 4, 5.3–5.6

Example: decision tree

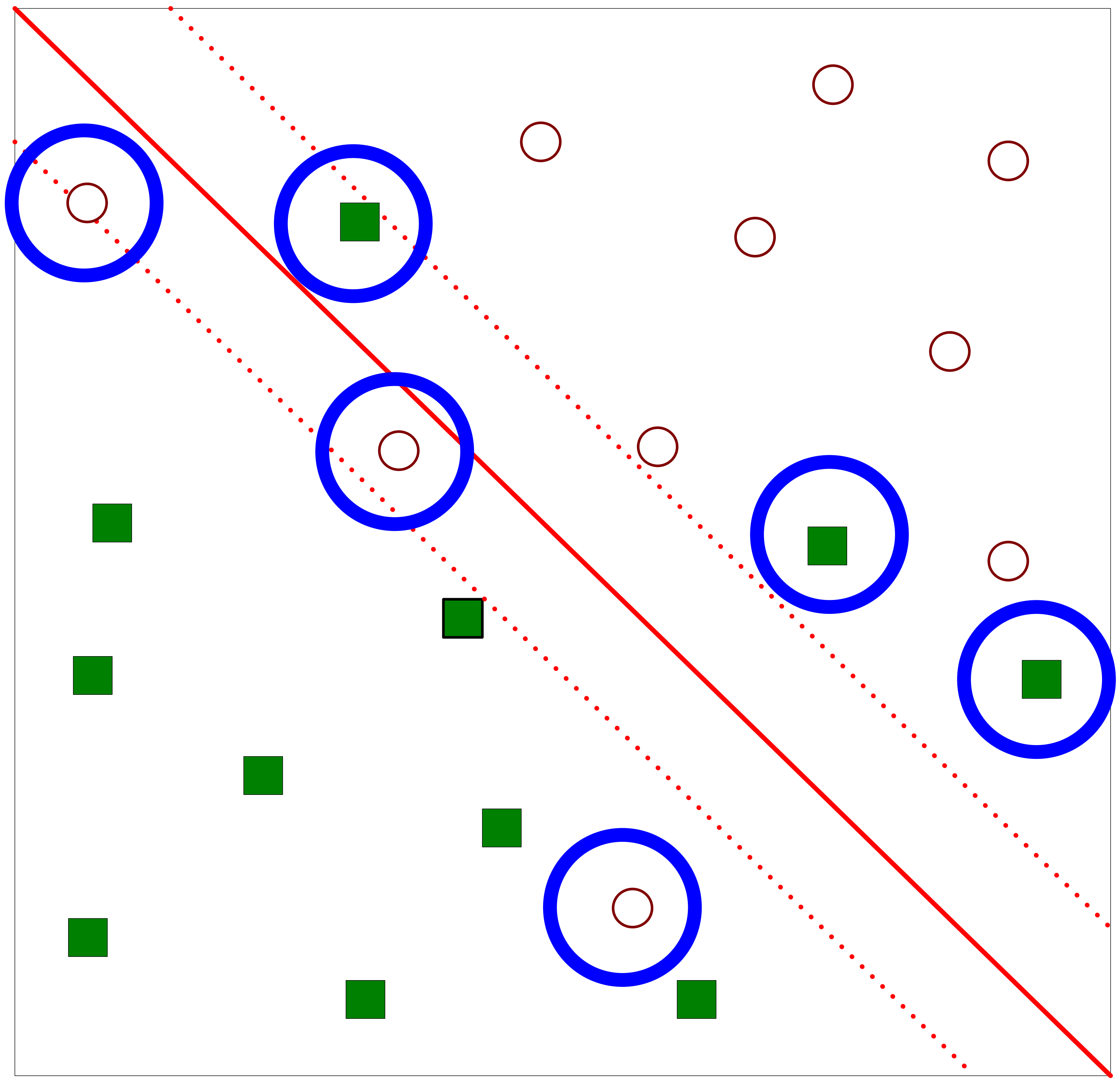


Building the classifier

- **Training phase**
  - Learn the posterior probabilities $\Pr[Y \mid X]$ for every combination of X and Y based on training data
- **Test phase**
  - For test record X', compute the class Y' that *maximizes the posterior probability* $\Pr[Y' \mid X']$
    - $Y' = \arg\max_i \{\Pr[c_i \mid X']\} = \arg\max_i \{\Pr[X' \mid c_i]\Pr[c_i]/\Pr[X']\}$
      $= \arg\max_i \{\Pr[X' \mid c_i]\Pr[c_i]\}$
- So we need $\Pr[X' \mid c_i]$ and $\Pr[c_i]$
  - $\Pr[c_i]$ is the fraction of test records that belong to class c_i
  - $\Pr[X' \mid c_i]$?

Linear, non-separable SVM

- What if the data is not linearly separable?



The dual

$$\frac{\partial L_p}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{N} \lambda_i y_i \boldsymbol{x}_i = 0 \Rightarrow \boldsymbol{w} = \sum_{i=1}^{N} \lambda_i y_i \boldsymbol{x}_i$$

$$\frac{\partial L_p}{\partial b} = -\sum_{i=1}^{N} \lambda_i y_i = 0$$

Partial derivatives

$$\frac{\partial L_p}{\partial \xi_i} = C - \lambda_i - \mu_i = 0 \Rightarrow \lambda_i + \mu_i = C$$

substitute to Lagrangian

Dual Lagrangian

$$L_D = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j$$

The same as before!

Linear, non-separable SVM, dual form.

$$\max_{\lambda} L_d = \sum_i \lambda_i - 1/2 \sum_{i,j} \lambda_i \lambda_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j$$

subject to $0 \leq \lambda_i \leq C$, $i = 1, \ldots, N$

Solving weight and bias with kernel

substitute

n = # of support vectors

$$\boldsymbol{w} = \sum_{i=1}^{N} \lambda_i y_i \Phi(\boldsymbol{x}_i) \qquad b = \frac{1}{n}\left(\sum_{i:\lambda_i>0} y_i - \sum_{i:\lambda_i>0} \boldsymbol{w}^T \Phi(\boldsymbol{x}_i)\right)$$

Has Φ

substitute

$$b = \frac{1}{n}\left(\sum_{i:\lambda_i>0} y_i - \sum_{i:\lambda_i>0}\sum_{j:\lambda_j>0} \lambda_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}_j)\right)$$

Has kernel

Classify new z:

$$\hat{y} = \operatorname{sign}(\boldsymbol{w}^T \Phi(\boldsymbol{z}) + b)$$

$$= \operatorname{sign}\left(\sum_{i:\lambda_i>0} \lambda_i y_i K(\boldsymbol{x}_i, \boldsymbol{z}) + b\right)$$

Chapter X: Graph Mining

1. Introduction to Graph Mining
2. Centrality and Other Graph Properties
3. Frequent Subgraph Mining
    3.1. Graphs and Isomorphism
    3.2. Canonical Codes
    3.3. gSpan
4. Graph Clustering
    4.1. Clustering as Graph Cutting
    4.2. Spectral Clustering
    4.3. Markov Clustering

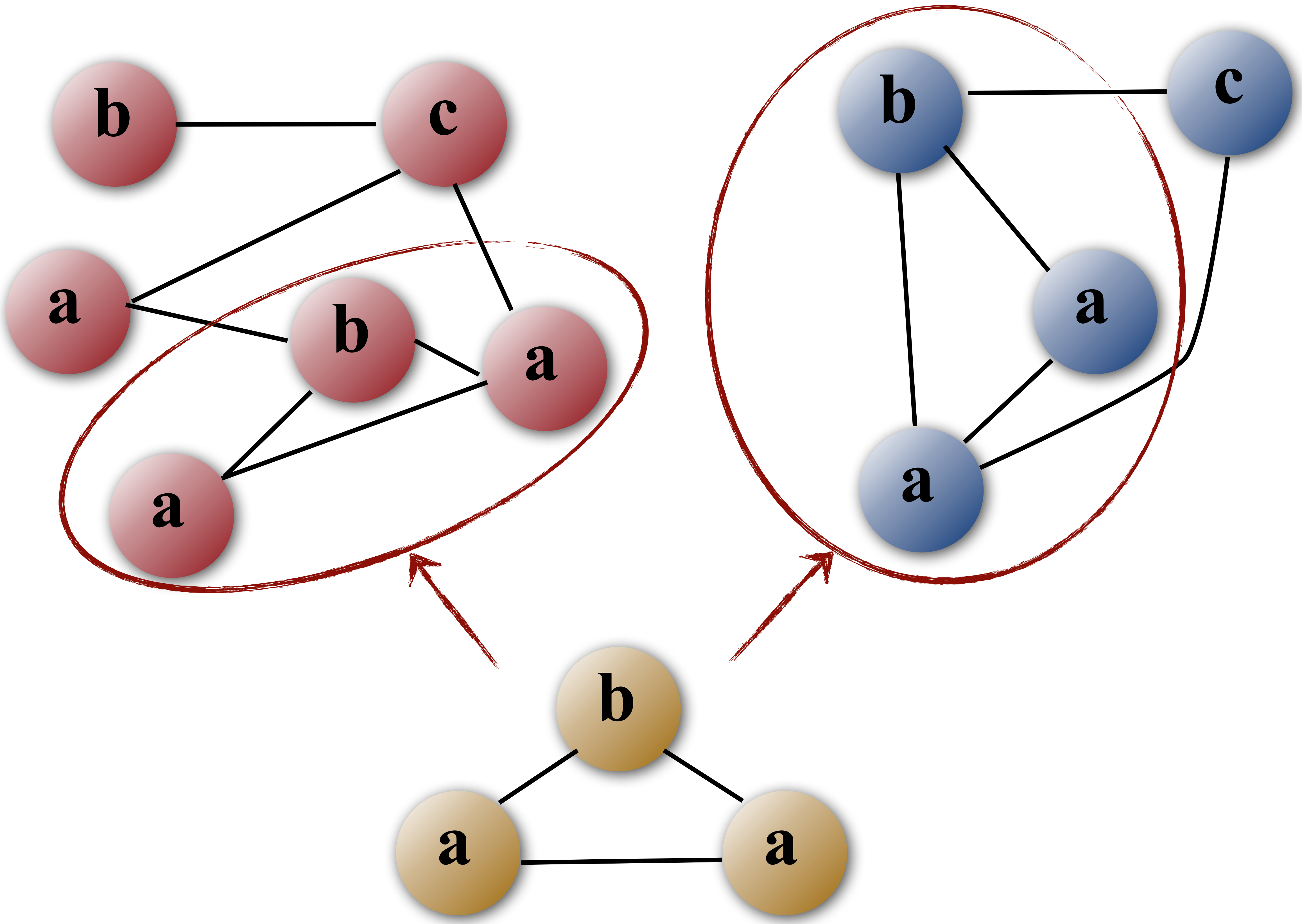
ZM Ch. 4, 11, 16

Centrality

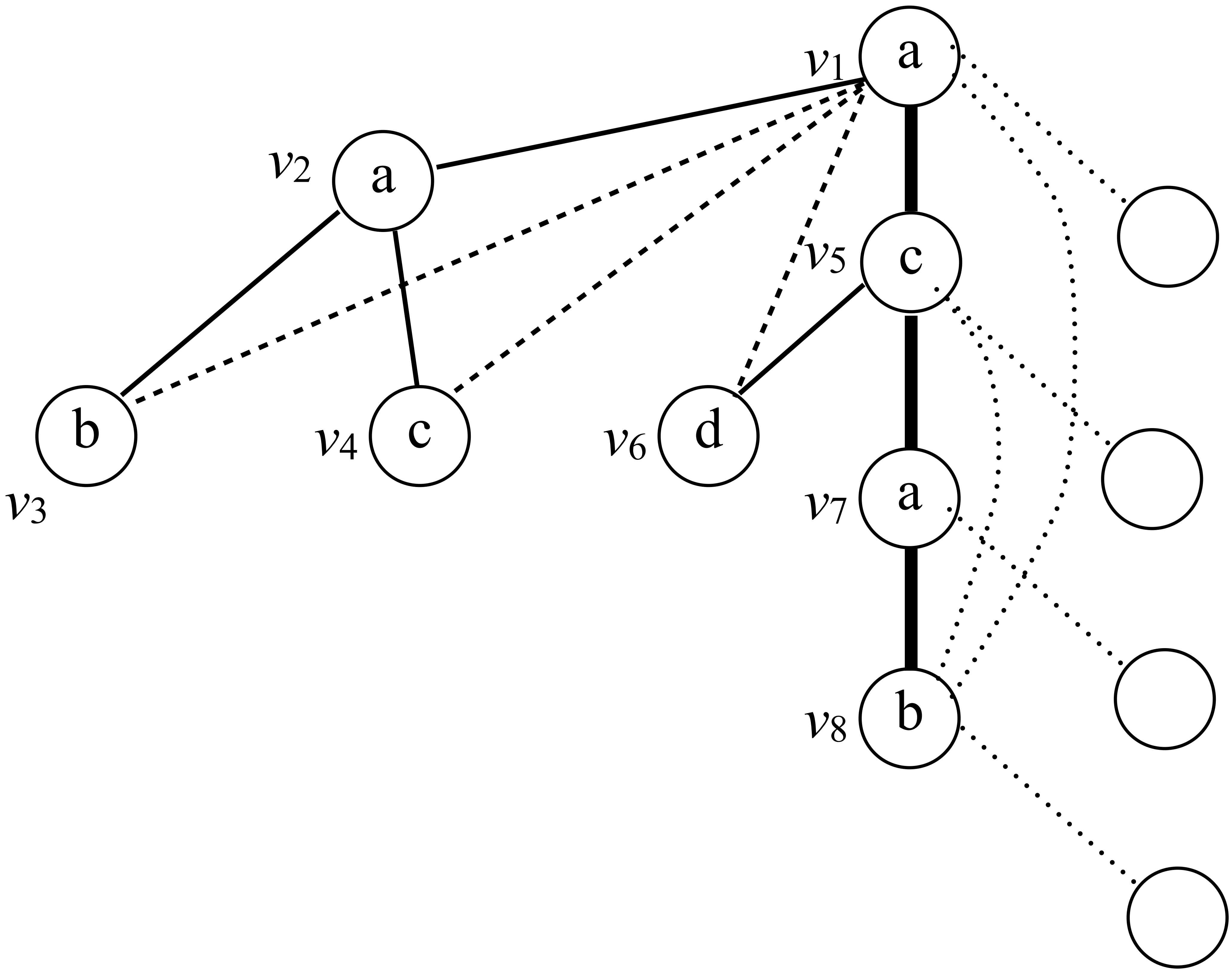
- Six degrees of Kevin Bacon
  - ”Every actor is related to Kevin Bacon by no more than 6 hops”
  - Kevin Bacon has acted with many, that have acted with many others, that have acted with many others…
- That makes Kevin Bacon a *centre* of the co-acting graph
  - Although he’s not the centre: the average distance to him is 2.998 but to Harvey Keitel it is only 2.848

http://oracleofbacon.org

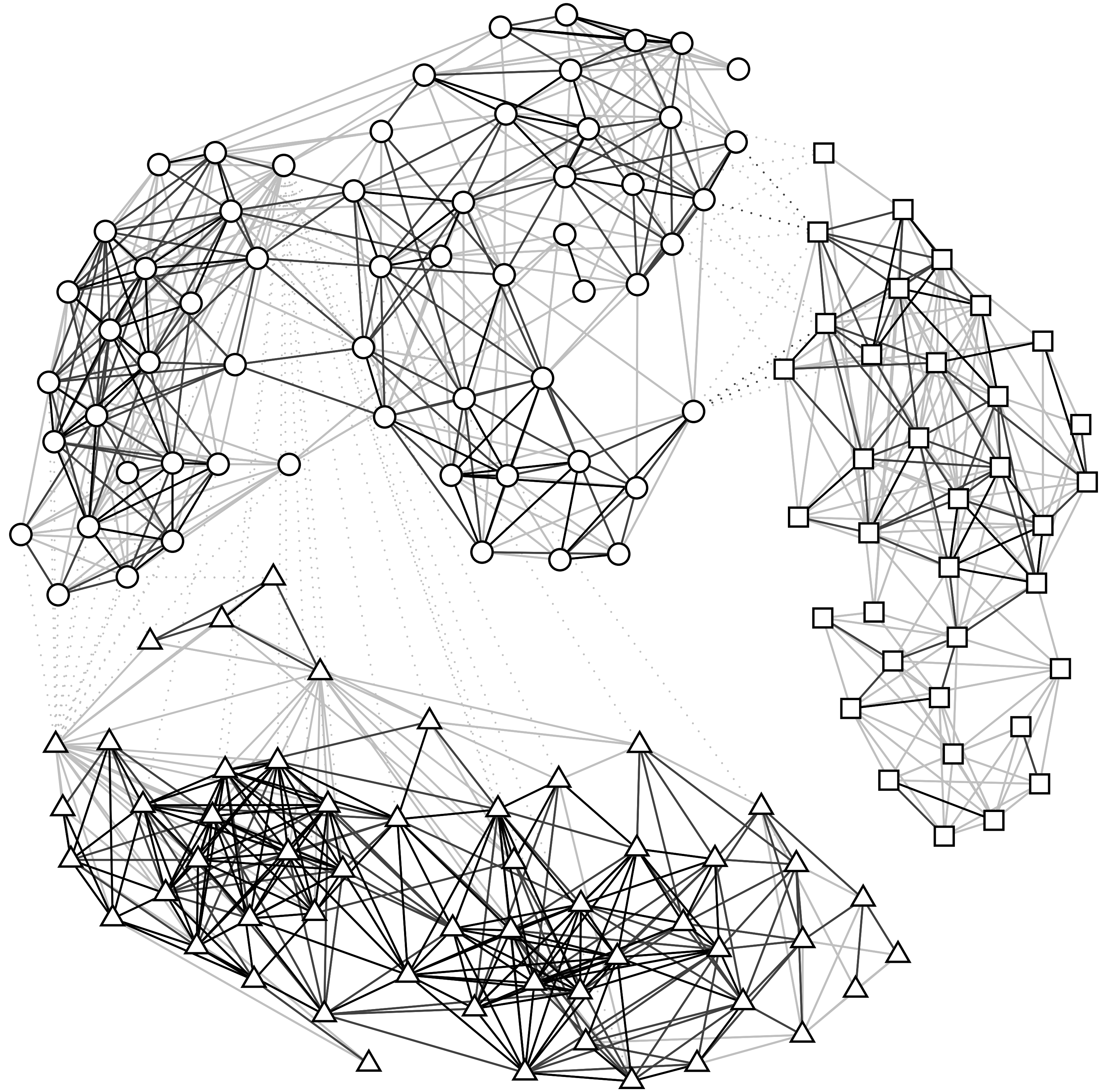
An Example



An Example



Example



ZM Figure 16.4

Cuts using matrices

$$\text{RatioCut} = \sum_{i=1}^{k} \frac{W(C_i, V \setminus C_i)}{|C_i|} = \sum_{i=1}^{k} \frac{\boldsymbol{c}_i^T \boldsymbol{L} \boldsymbol{c}_i}{\|\boldsymbol{c}_i\|^2}$$

$$\text{NormalizedCut} = \sum_{i=1}^{k} \frac{W(C_i, V \setminus C_i)}{vol(C_i)} = \sum_{i=1}^{k} \frac{\boldsymbol{c}_i^T \boldsymbol{L} \boldsymbol{c}_i}{\boldsymbol{c}_i^T \boldsymbol{\Delta} \boldsymbol{c}_i}$$

Chapter XI: Two Matrix Factorizations

1. Non-Negative Matrix Factorization
    1.1. Idea and motivation
    1.2. Algorithms
2. Boolean Matrix Factorization
    2.1. Idea and motivation
    2.2. Algorithms

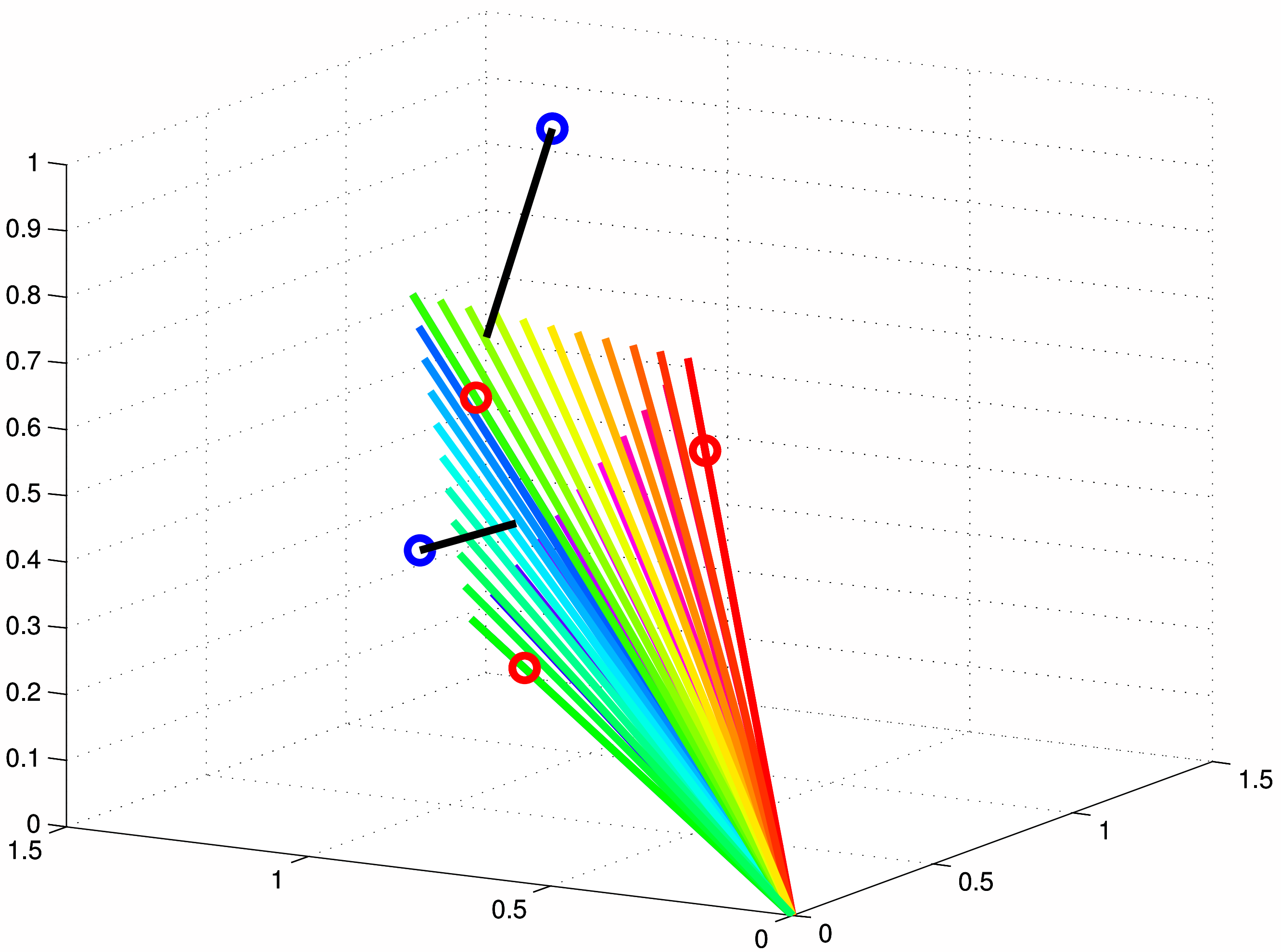
Geometry of NMF

NMF factors

Data points

Convex cone

Projections



Boolean Matrix Factorization

Long-haired	✔	✔	✘
Well-known	✔	✔	✔
Male	✘	✔	✔

BMF example

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix} = \mathbf{A} \circ \mathbf{B}$$

$$\begin{pmatrix} 1 & 1 & 0 \end{pmatrix} = \mathbf{b}_1$$

$$\mathbf{a}_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \mathbf{a}_1 \mathbf{b}_1$$

$$\begin{pmatrix} 0 & 1 & 1 \end{pmatrix} = \mathbf{b}_2$$

$$\mathbf{a}_2 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix} = \mathbf{a}_2 \mathbf{b}_2$$

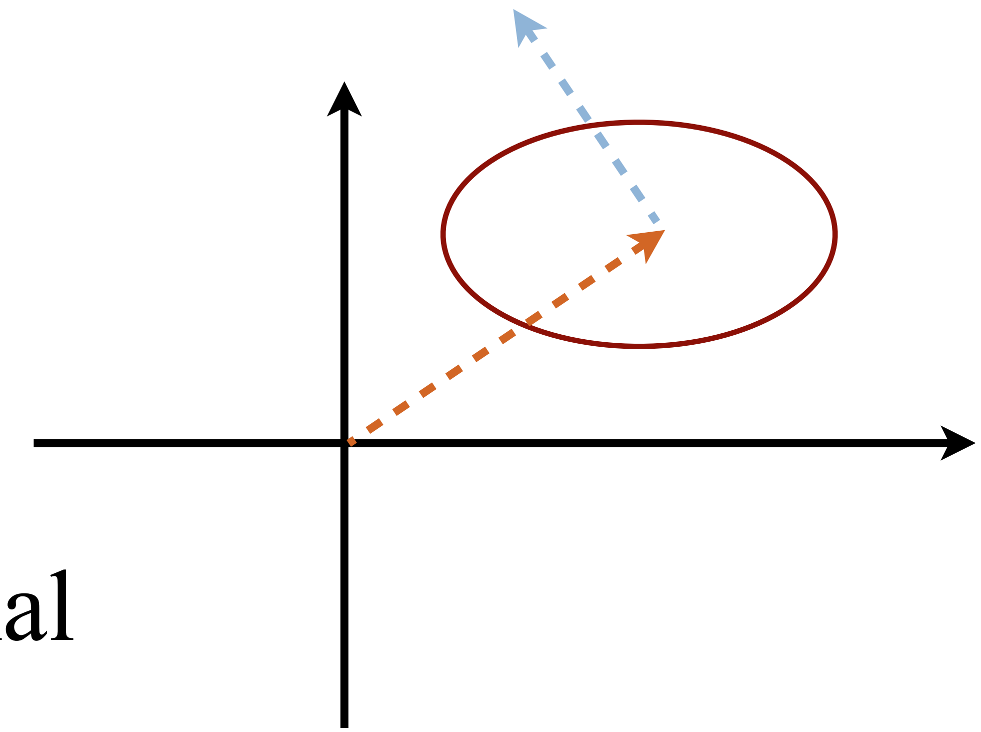
Chapter XII: Data Pre and Post Processing

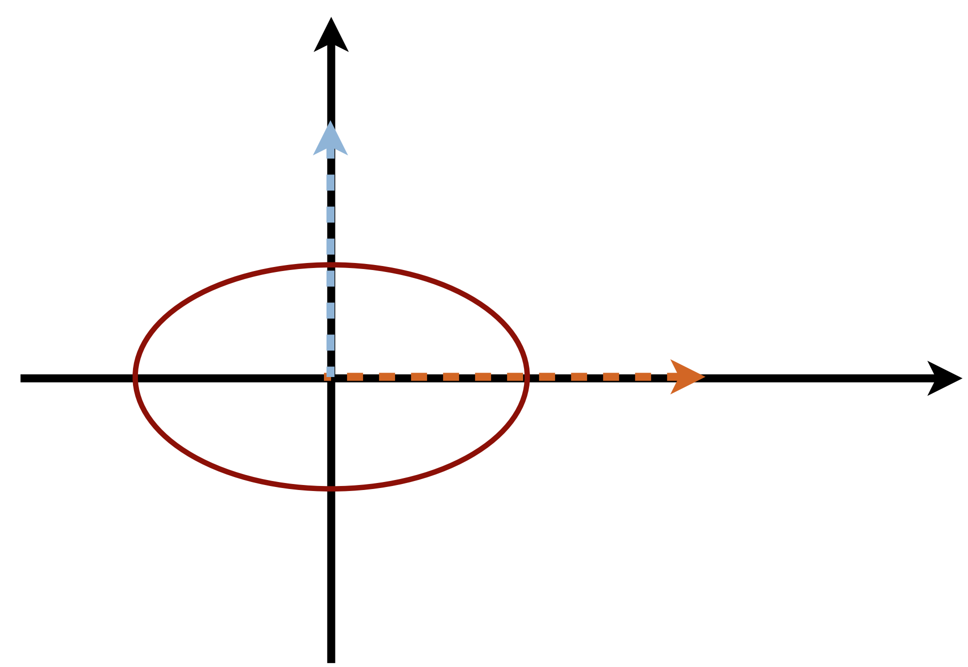
1. Data Normalization
2. Missing Values
3. Curse of Dimensionality
4. Feature Extraction and Selection
    4.1. PCA and SVD
    4.2. Johnson–Lindenstrauss lemma
    4.3. CX and CUR decompositions
5. Visualization and Analysis of the Results
6. Tales from the Wild

Zaki & Meira, Ch. 2.4, 6 & 8

Why centering?

- Consider the red data ellipse
  - The main direction of variance is from the origin to the data
  - The second direction is orthogonal to the first
  - These don't tell the variance of the data!
- If we center the data, the directions are correct





Example of 1-D PCA

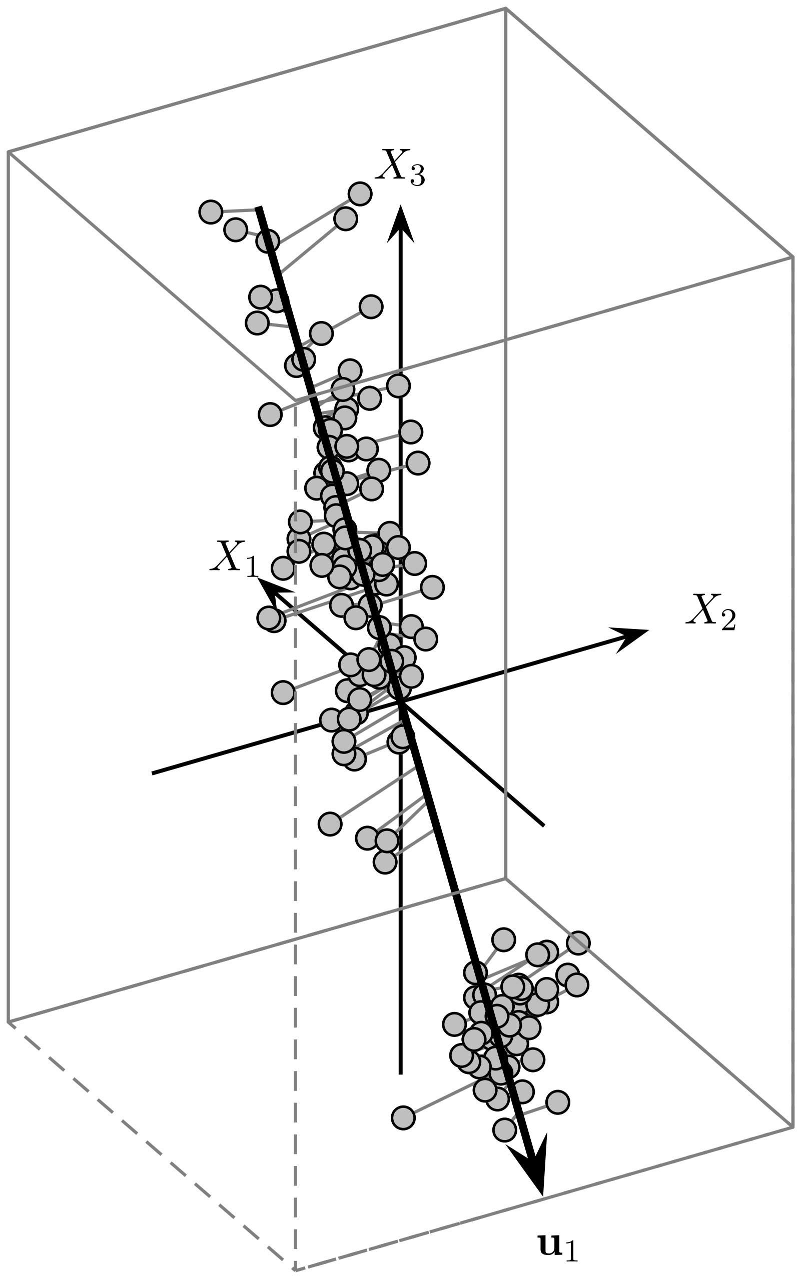


Figure 7.2: Best One-dimensional or Line Approximation

Heat maps with dendrograms

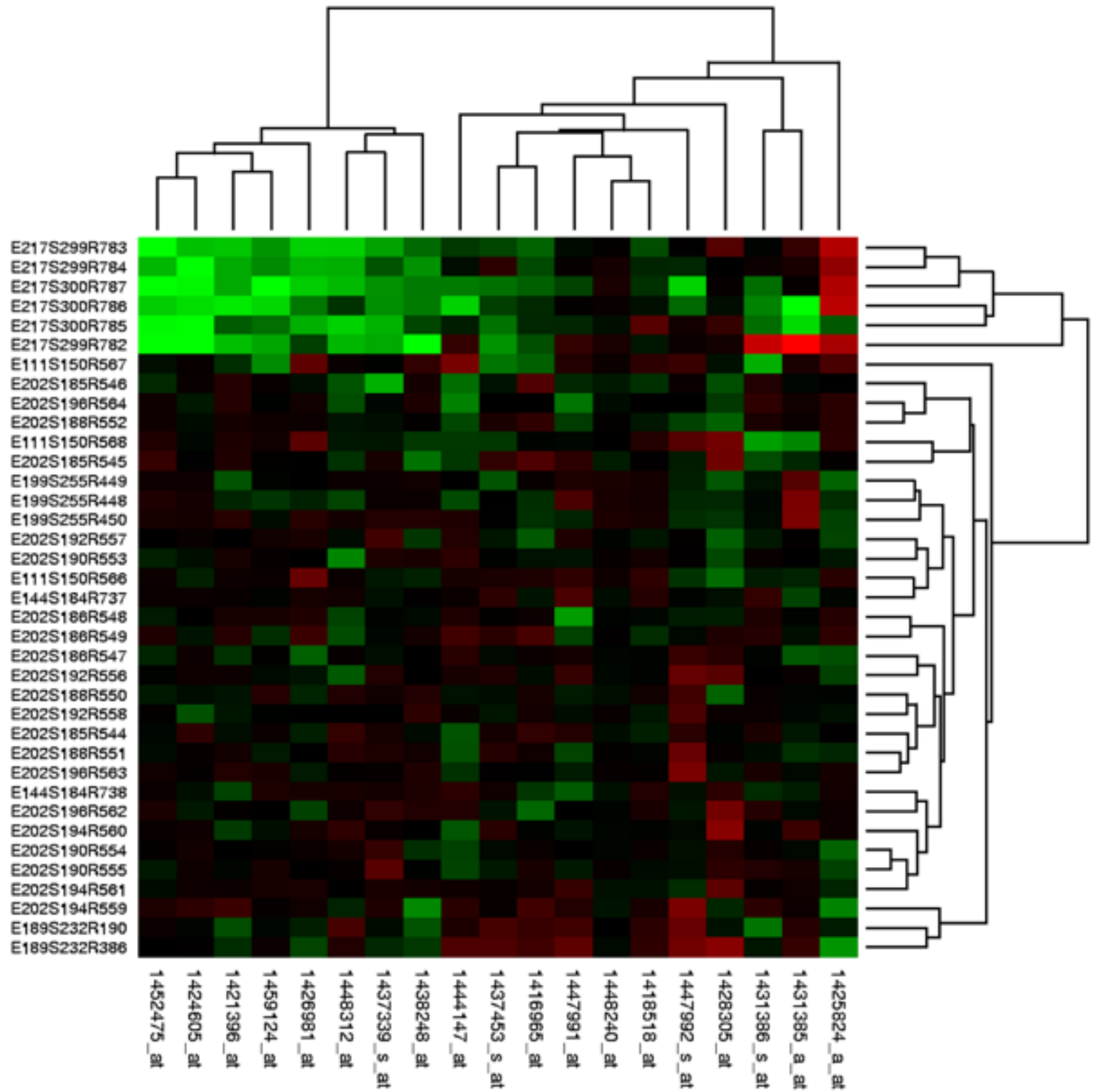


Image: Wikipedia

Data mining = voodoo science