Clustering
Lecture 7: Clustering Ensemble

Jing Gao
SUNY Buffalo

Outline

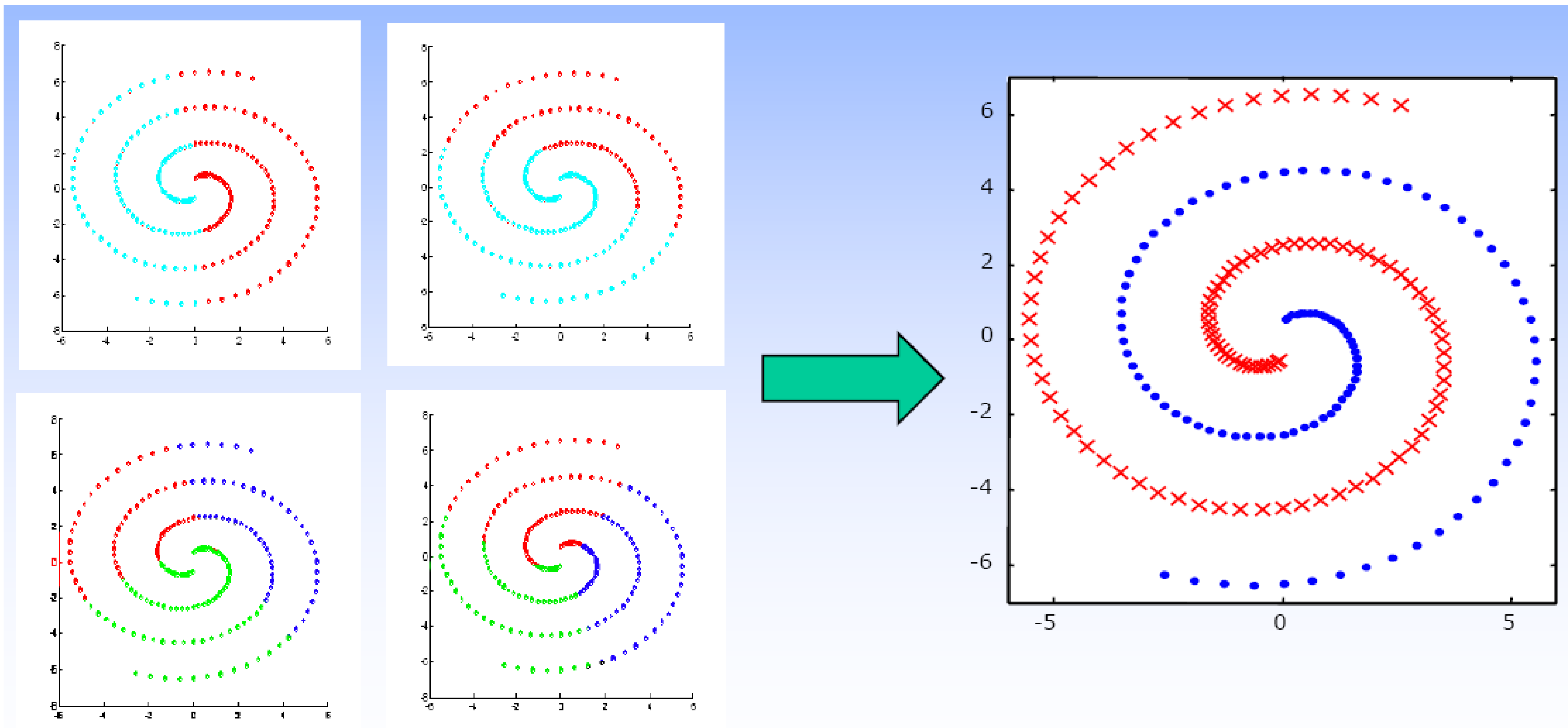
- **Basics**
  - Motivation, definition, evaluation
- **Methods**
  - Partitional
  - Hierarchical
  - Density-based
  - Mixture model
  - Spectral methods
- **Advanced topics**
  - Clustering ensemble
  - Clustering in MapReduce
  - Semi-supervised clustering, subspace clustering, co-clustering, etc.

Clustering Ensemble

- Problem
  - Given an unlabeled data set $D=\{x_1,x_2,\dots,x_n\}$
  - An ensemble approach computes:
    - A set of clustering solutions $\{C_1,C_2,\dots,C_k\}$, each of which maps data to a cluster: $f_j(x)=m$
    - A unified clustering solutions f^* which combines base clustering solutions by their consensus
- Challenges
  - The correspondence between the clusters in different clustering solutions is unknown

Motivations

- Goal
  – Combine “weak” clusterings to a better one



[PTJ05]

An Example

base clustering models

	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_3$	$\mathcal{C}$
v_1	1	1	1	1
v_2	1	2	2	2
v_3	2	1	1	1
v_4	2	2	2	2
v_5	3	3	3	3
v_6	3	4	3	3

objects

they may not represent the same cluster!

The goal: get the consensus clustering

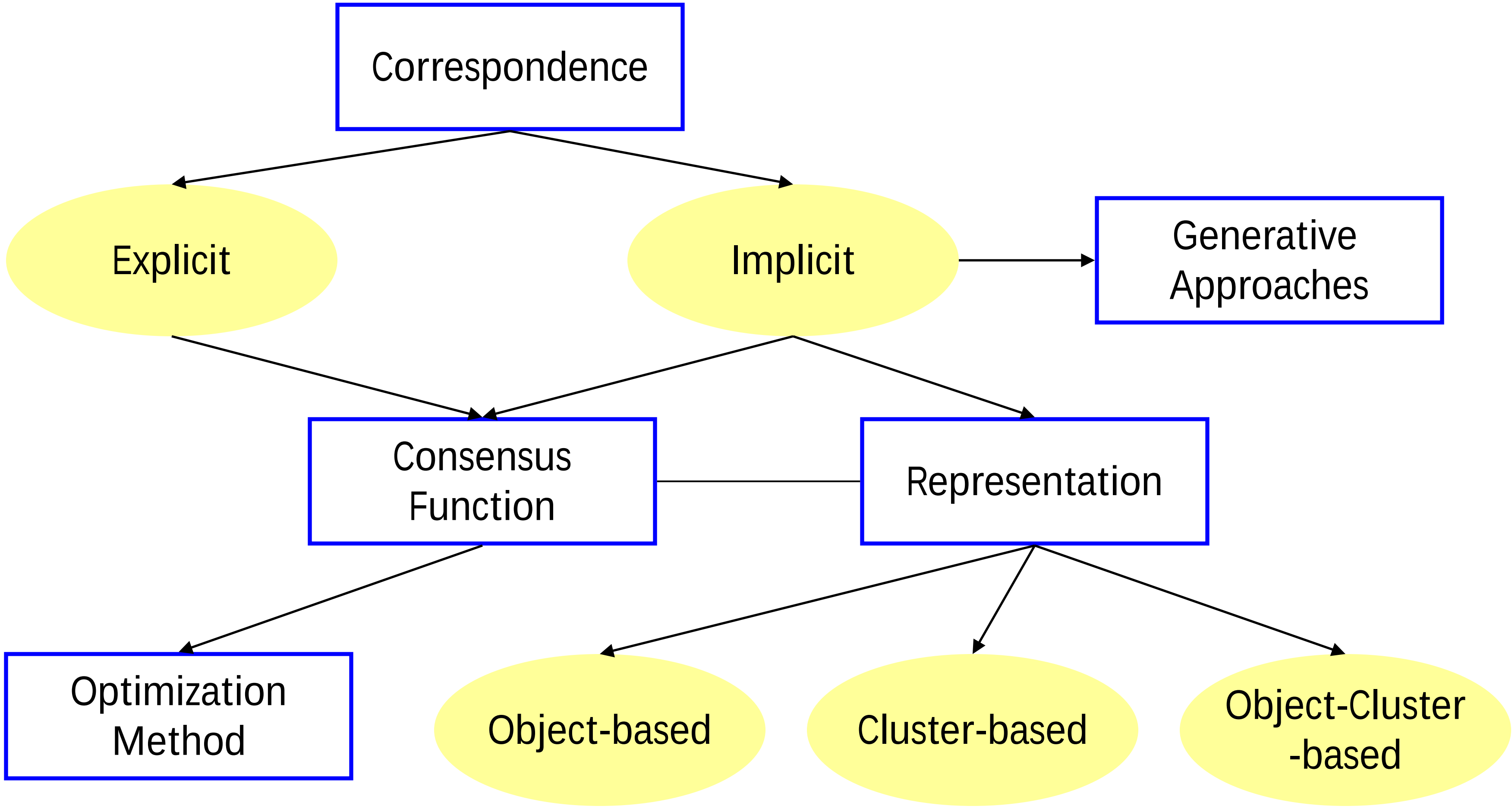
[GMT07]

Methods (1)

- How to get base models?
  - Bootstrap samples
  - Different subsets of features
  - Different clustering algorithms
  - Random number of clusters
  - Random initialization for K-means
  - Incorporating random noises into cluster labels
  - Varying the order of data in on-line methods

Methods (2)

- How to combine the models?



Hard Correspondence (1)

- Re-labeling+voting
  - Find the correspondence between the labels in the partitions and fuse the clusters with the same labels by voting [DuFr03,DWH01]

	C_1	C_2	C_3
v_1	1	3	2
v_2	1	3	2
v_3	2	1	2
v_4	2	1	3
v_5	3	2	1
v_6	3	2	1

Re-labeling

	C_1	C_2	C_3
v_1	1	1	1
v_2	1	1	1
v_3	2	2	1
v_4	2	2	2
v_5	3	3	3
v_6	3	3	3

Voting

C^*
1
1
2
2
3
3

Hard Correspondence (2)

- Details
  - Minimize match costs
  - Match to a reference clustering or match in a pairwise manner
- Problems
  - In most cases, clusters do not have one-to-one correspondence

Soft Correspondence* (1)

- Notations
  - Membership matrix $M_1, M_2, \ldots, M_k$
  - Membership matrix of consensus clustering M
  - Correspondence matrix $S_1, S_2, \ldots, S_k$
  - $M_i S_i = M$

	C_1	C_2	C_3
v_1	1	3	2
v_2	1	3	2
v_3	2	1	2
v_4	2	1	3
v_5	3	2	1
v_6	3	2	1

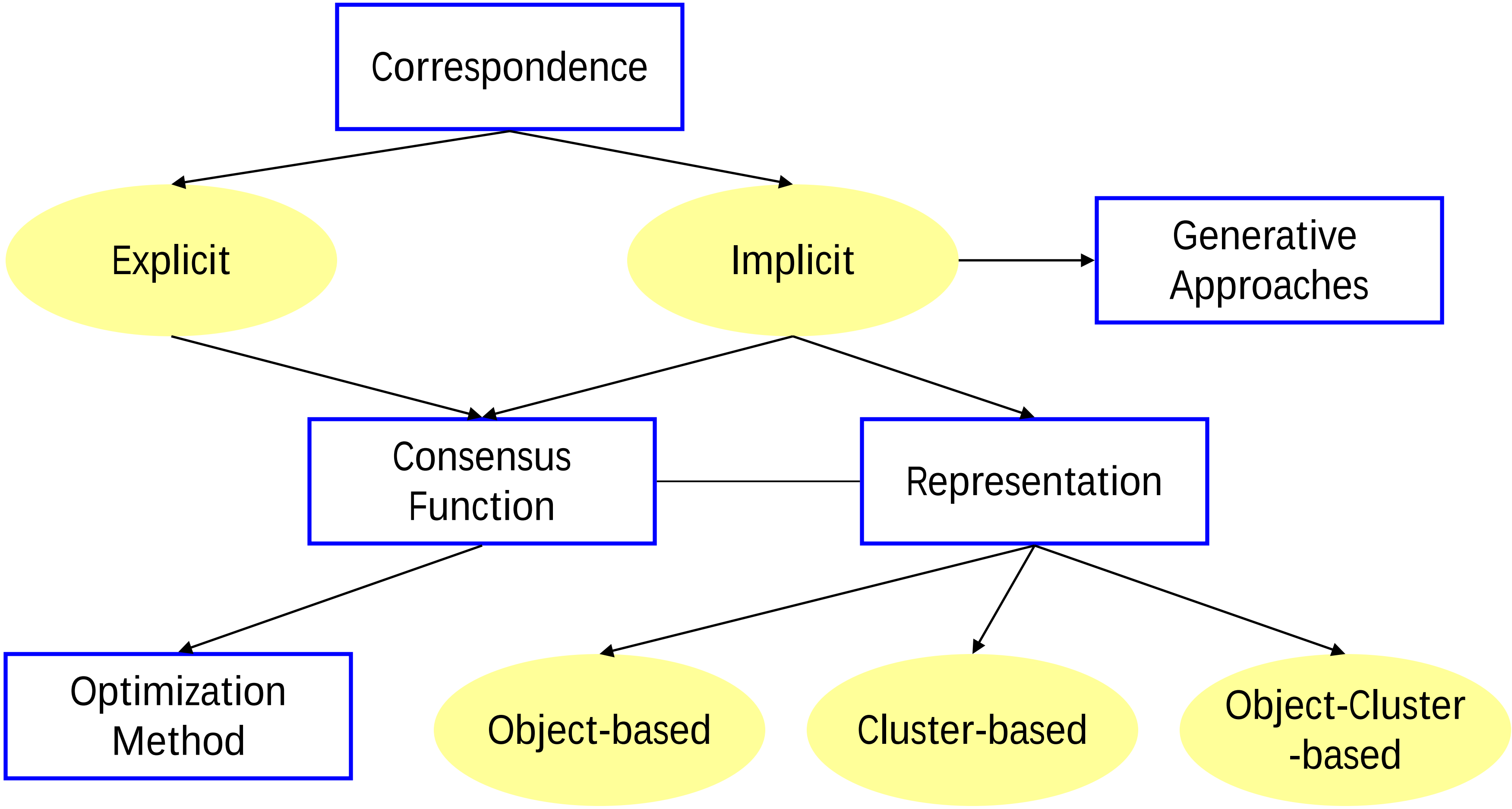
$$\underset{M_2}{\begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix}} \times \underset{S_2}{\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}} = \underset{M}{\begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}}$$

*[LZY05]

Soft Correspondence (2)

- Consensus function
  - Minimize disagreement $\min \sum_{j=1}^{k} \| M - M_j S_j \|^2$
  - Constraint 1: column-sparseness
  - Constraint 2: each row sums up to 1
  - Variables: M, S_1, S_2, ... ,S_k
- Optimization
  - EM-based approach
  - Iterate until convergence
    - Update S using gradient descent
    - Update M as $M = \frac{1}{k} \sum_{j=1}^{k} M_j S_j$

- How to combine the models?

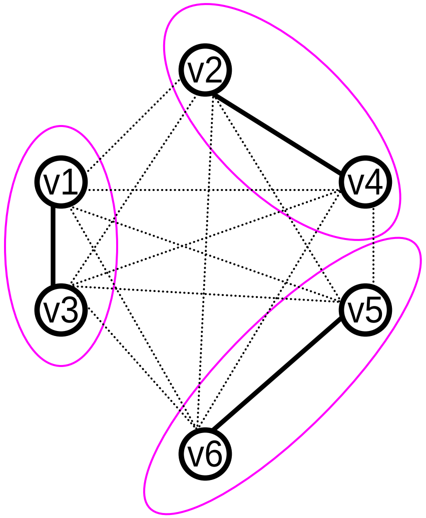


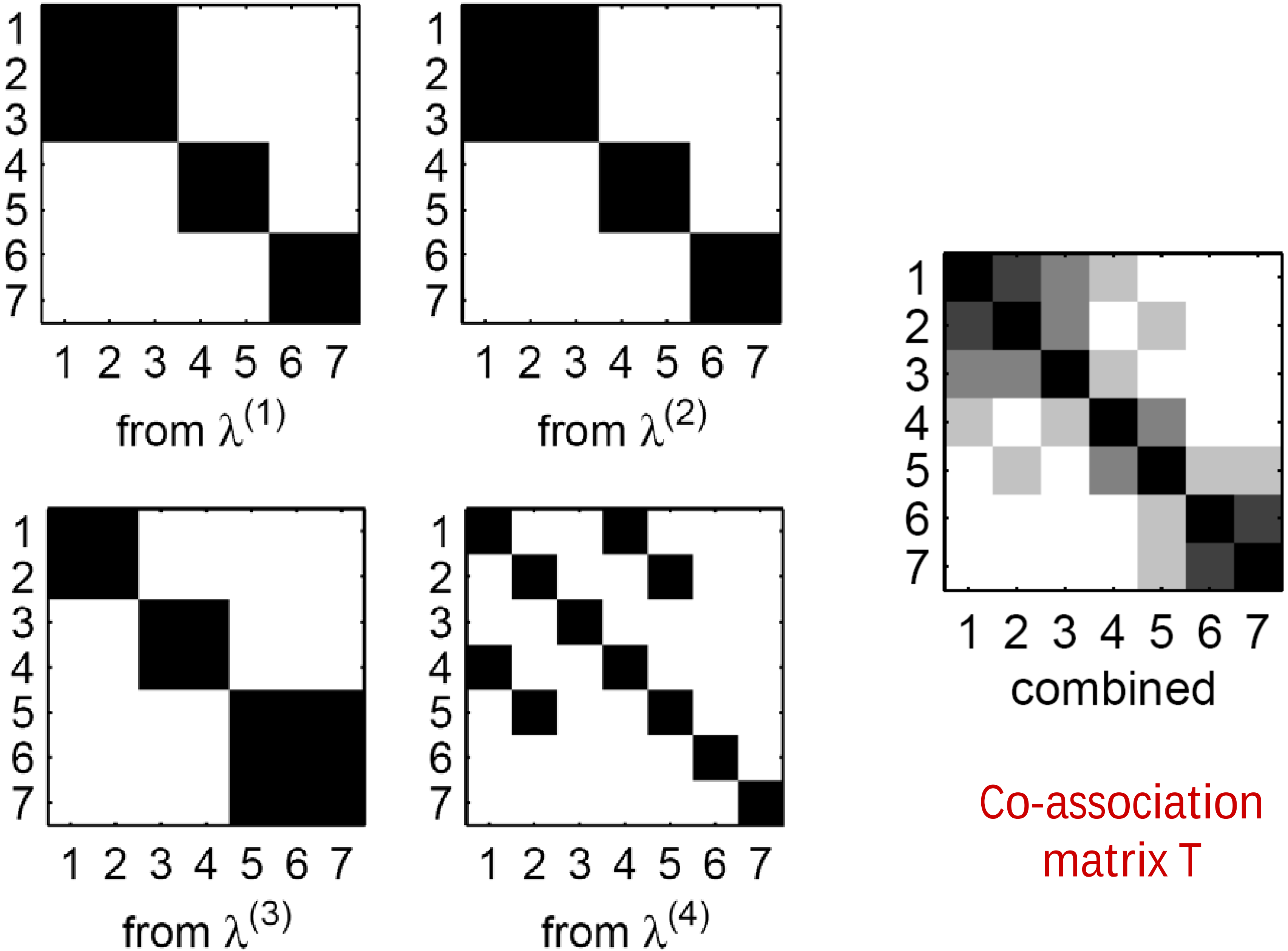
Object-based Methods (1)

- Clustering objects
  - Define a similarity or distance measure:
    - Similarity between two objects can be defined as the percentage of clusterings that assign the two objects into same clusters
    - Distance between two objects can be defined as the percentage of clusterings that assign the two objects into different clusters
  - Conduct clustering on the new similarity (distance) matrix
  - Result clustering represents the consensus
  - Can view this approach as clustering in the new feature space where clustering results are the categorical features

Object-based Methods (2)

	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_3$	$\mathcal{C}$
v_1	1	1	1	1
v_2	1	2	2	2
v_3	2	1	1	1
v_4	2	2	2	2
v_5	3	3	3	3
v_6	3	4	3	3





Co-association matrix T

[StGh03]

Consensus Function

- Minimizing disagreement
  - Information-theoretic [StGh03]

$$\max \frac{1}{k}\sum_{j=1}^{k} NMI(T, T_j) \qquad NMI(T, T_j) = \frac{I(T, T_j)}{\sqrt{H(T)H(T_j)}}$$

  - Median partition [LDJ07]

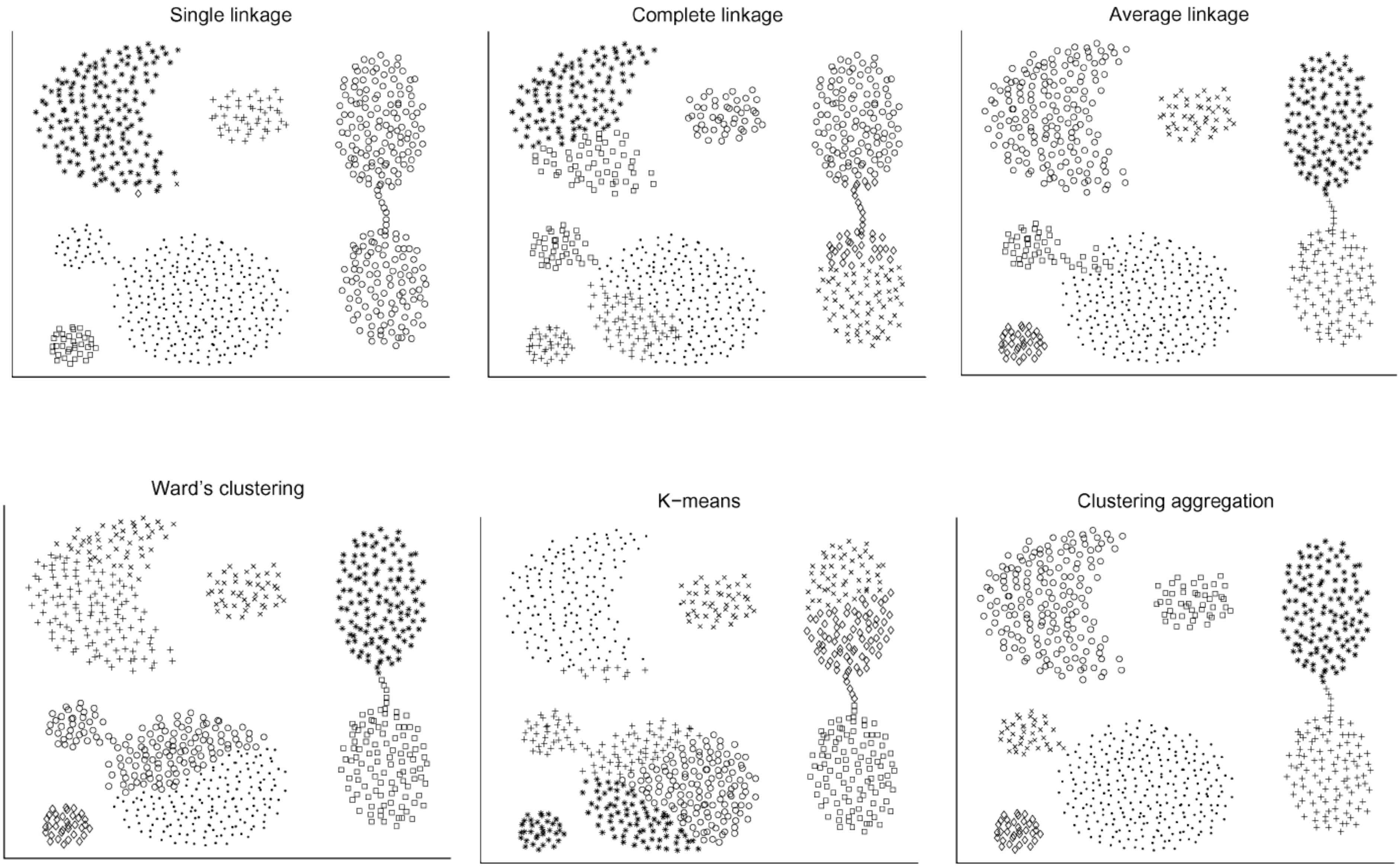
$$\overline{T} = \frac{1}{k}\sum_{j=1}^{k} T_j \qquad \min \left\| \overline{T} - T \right\|^2$$

  - Correlation clustering [GMT07]

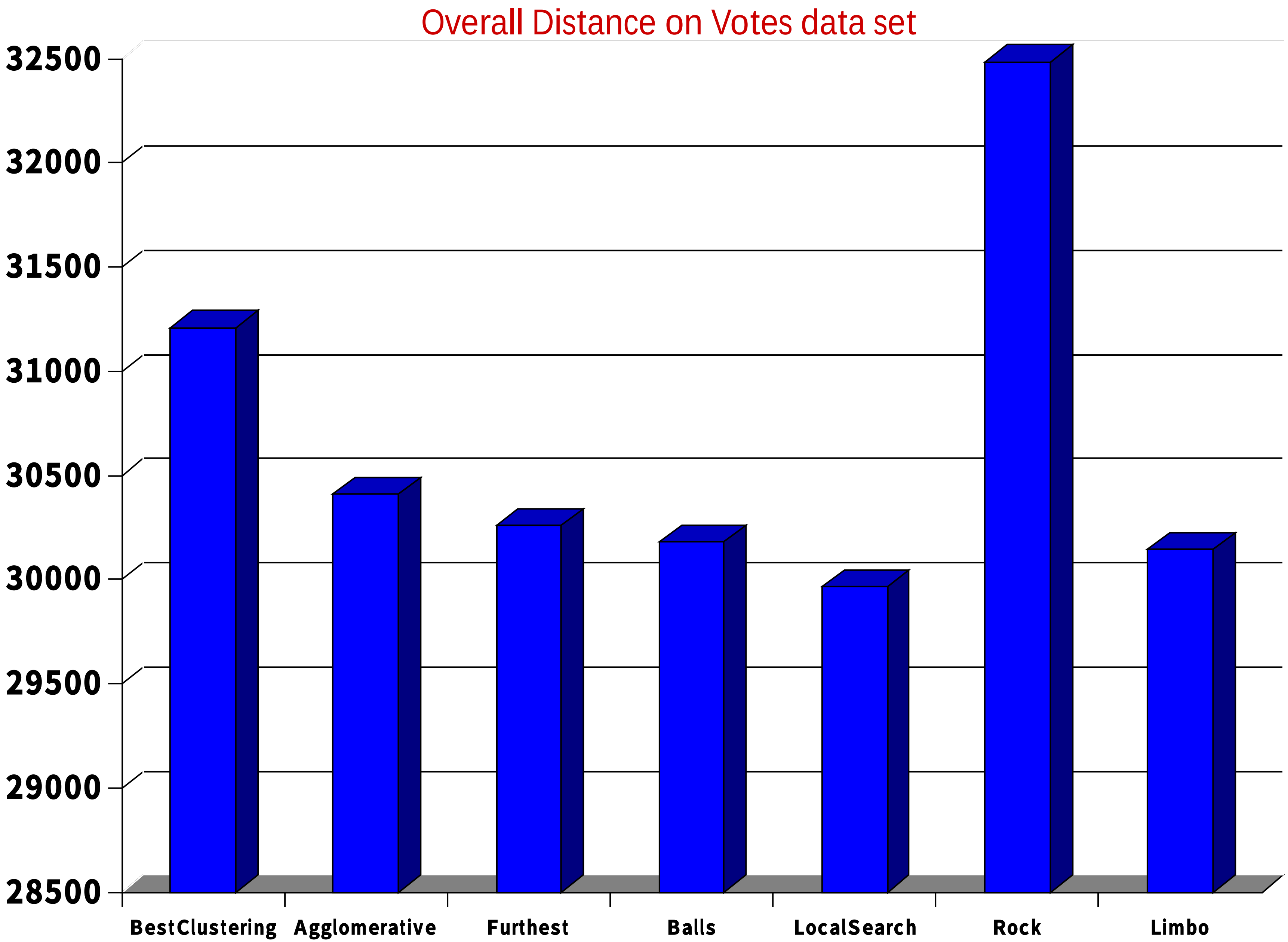
$$\max \sum_{\substack{(u,v)\\ C(u)=C(v)}} \overline{T}_{uv} + \sum_{\substack{(u,v)\\ C(u)\neq C(v)}} (1 - \overline{T}_{uv})$$

Optimization Method

- Approximation
  - Agglomerative clustering (bottom-up) [FrJa02,GMT07]
    - Single link, average link, complete link
  - Divisive clustering (top-down) [GMT07]
    - Furthest
  - LocalSearch [GMT07]
    - Place an object into a different cluster if objective function improved
    - Iterate the above until no improvements can be made
  - BestClustering [GMT07]
    - Select the clustering that maximize (minimize) the objective function
  - Graph partitioning [StGh03]
  - Nonnegative matrix factorization [LDJ07,LiDi08]

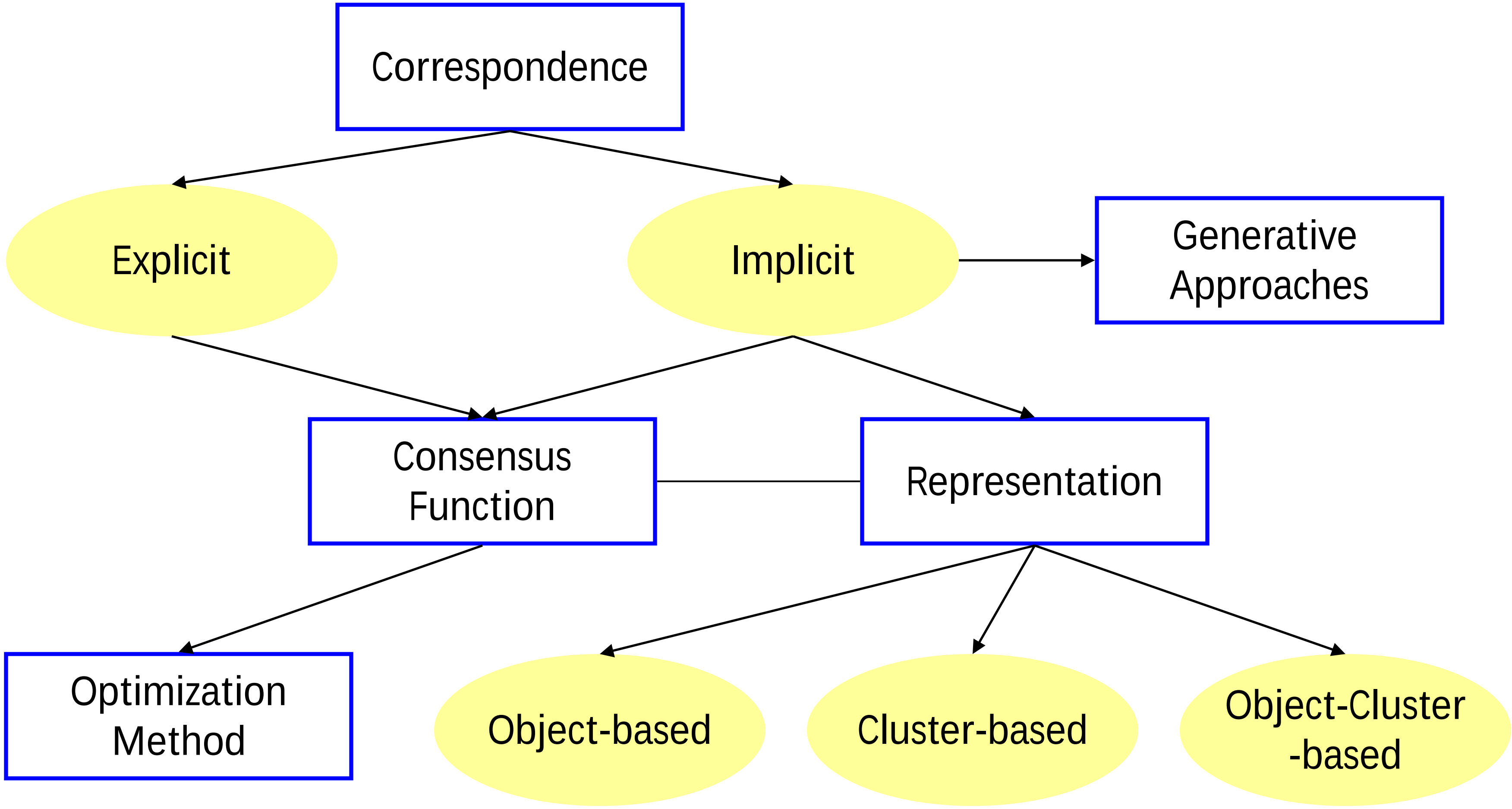

Single linkage
Complete linkage
Average linkage
Ward's clustering
K−means
Clustering aggregation

[GMT07]


Overall Distance on Votes data set
32500
32000
31500
31000
30500
30000
29500
29000
28500
BestClustering
Agglomerative
Furthest
Balls
LocalSearch
Rock
Limbo

[GMT07]

- How to combine the models?



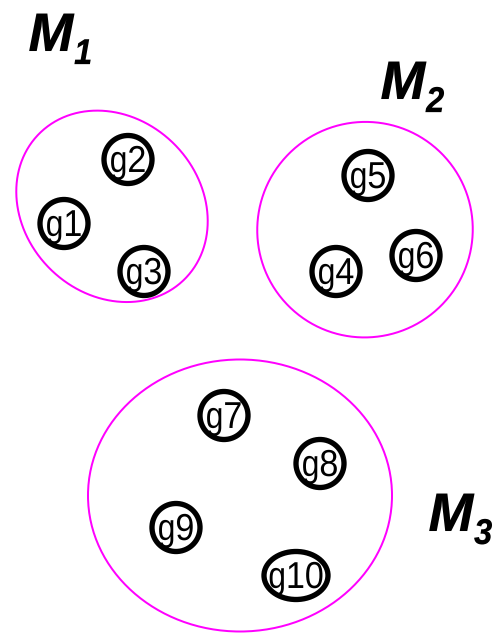
Cluster-based Methods

- Clustering clusters
  - Regard each cluster from a base model as a record
  - Similarity is defined as the percentage of shared common objects
    - eg. Jaccard measure
  - Conduct clustering on these clusters
  - Assign an object to its most associated consensus cluster

Meta-Clustering Algorithm (MCLA)*

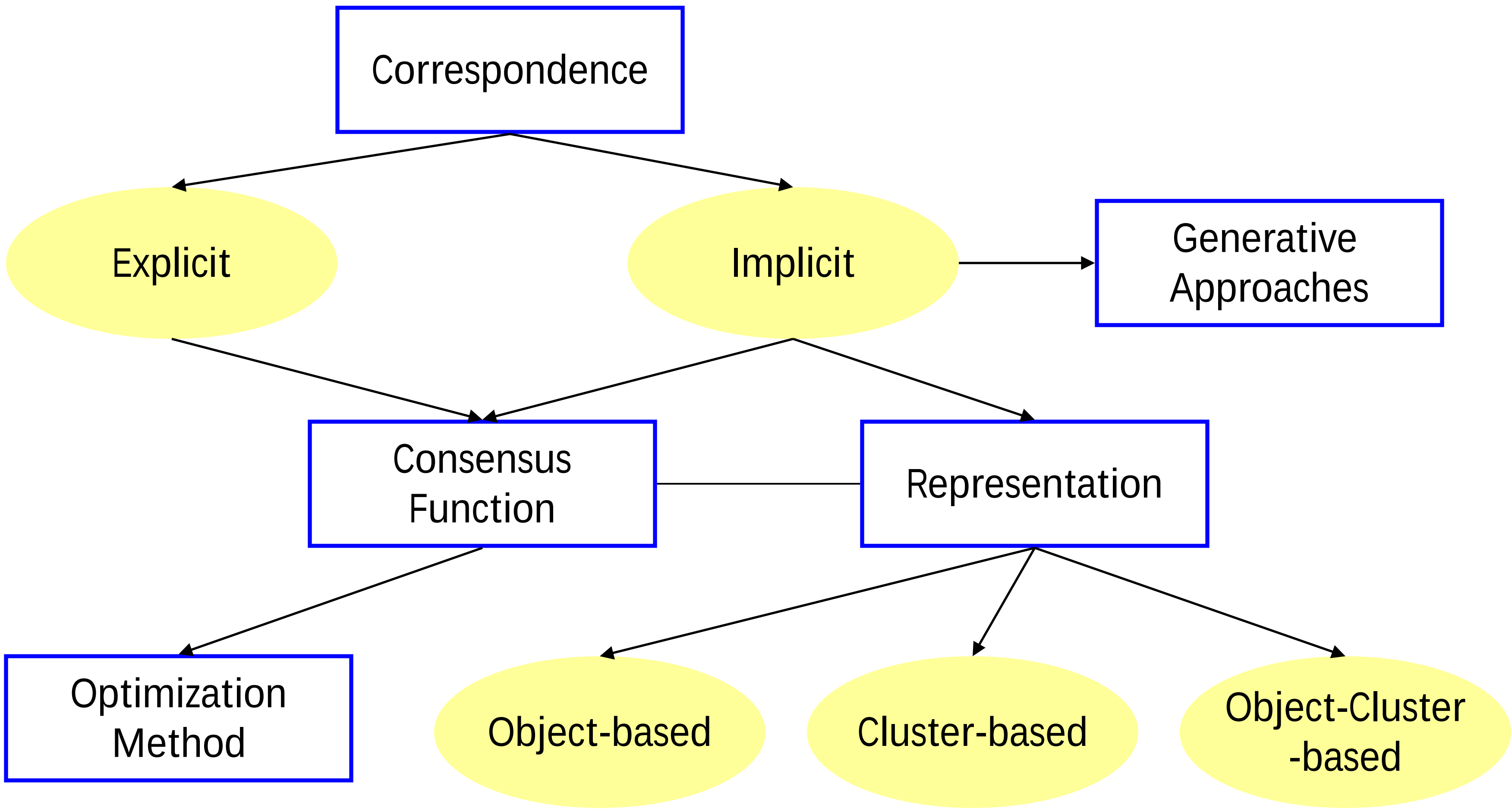
	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_3$	$\mathcal{C}$
v_1	1	1	1	1
v_2	1	2	2	2
v_3	2	1	1	1
v_4	2	2	2	2
v_5	3	3	3	3
v_6	3	4	3	3

M_1	M_2	M_3
3	0	0
1	2	0
2	1	0
0	3	0
0	0	3
0	0	3



*[StGh03]

- How to combine the models?



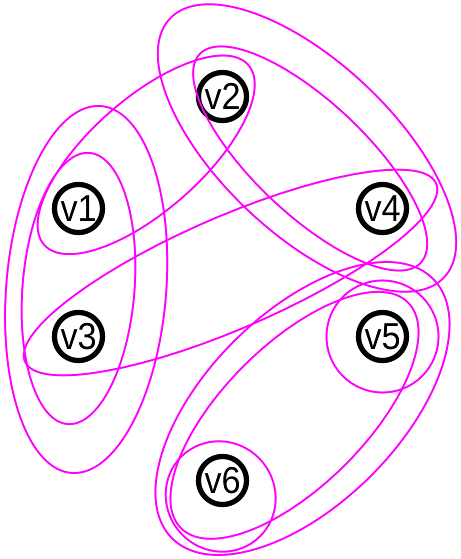
HyperGraph-Partitioning Algorithm (HGPA)*

- Hypergraph representation and clustering
  - Each node denotes an object
  - A hyperedge is a generalization of an edge in that it can connect any number of nodes
  - For objects that are put into the same cluster by a clustering algorithm, draw a hyperedge connecting them
  - Partition the hypergraph by minimizing the number of cut hyperedges
  - Each component forms a consensus cluster

*[StGh03]

HyperGraph-Partitioning Algorithm (HGPA)

	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_3$	$\mathcal{C}$
v_1	1	1	1	1
v_2	1	2	2	2
v_3	2	1	1	1
v_4	2	2	2	2
v_5	3	3	3	3
v_6	3	4	3	3



Hypergraph representation– a circle denotes a hyperedge

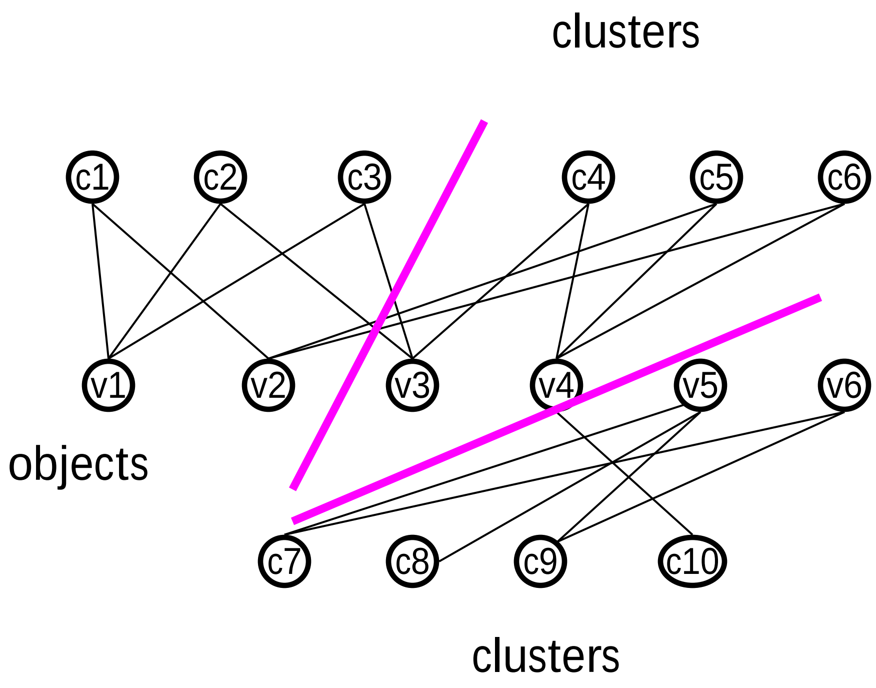
Bipartite Graph Partitioning*

- Hybrid Bipartite Graph Formulation
  - Summarize base model output in a bipartite graph
  - Lossless summarization—base model output can be reconstructed from the bipartite graph
  - Use spectral clustering algorithm to partition the bipartite graph
  - Time complexity $O(nkr)$—due to the special structure of the bipartite graph
  - Each component represents a consensus cluster

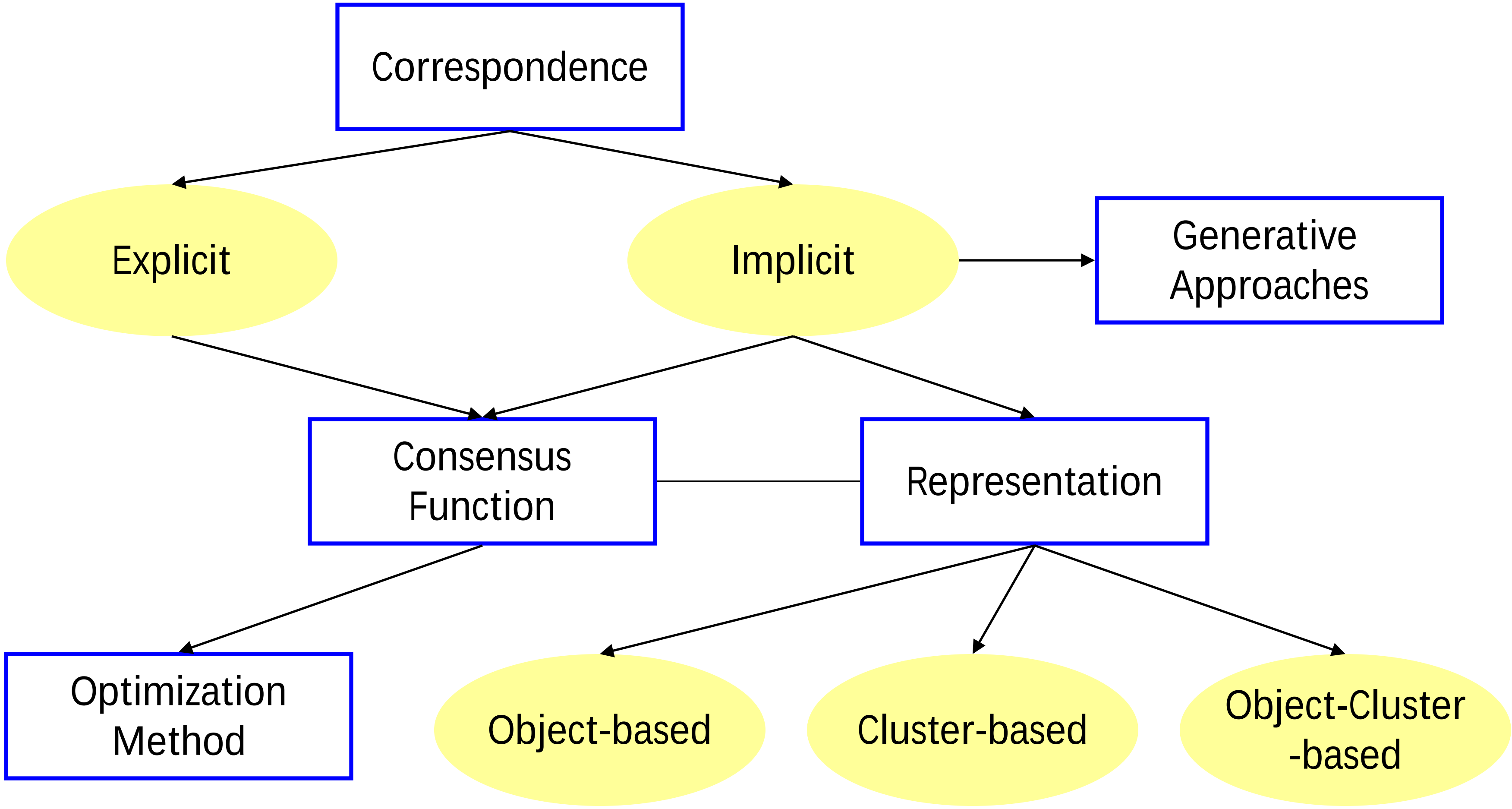
*[FeBr04]

Bipartite Graph Partitioning

	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_3$	$\mathcal{C}$
v_1	1	1	1	1
v_2	1	2	2	2
v_3	2	1	1	1
v_4	2	2	2	2
v_5	3	3	3	3
v_6	3	4	3	3



- How to combine the models?



A Mixture Model of Consensus*

- Probability-based
  - Assume output comes from a mixture of models
  - Use EM algorithm to learn the model
- Generative model
  - The clustering solutions for each object are represented as nominal features--v_i
  - v_i is described by a mixture of k components, each component follows a multinomial distribution
  - Each component is characterized by distribution parameters q_j

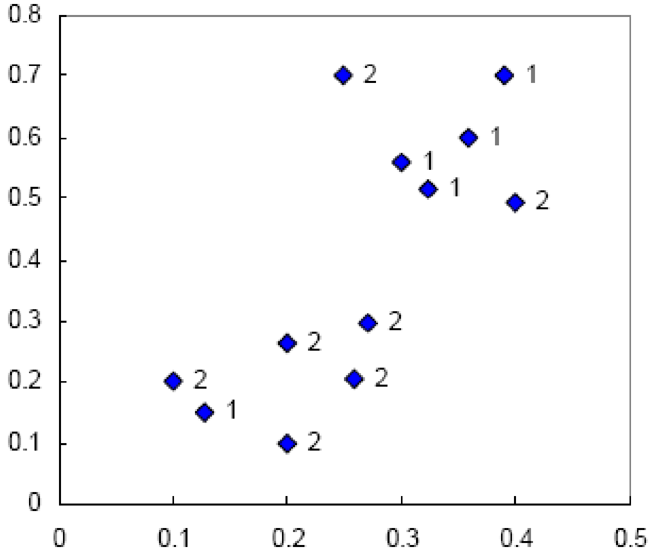
*[PTJ05]

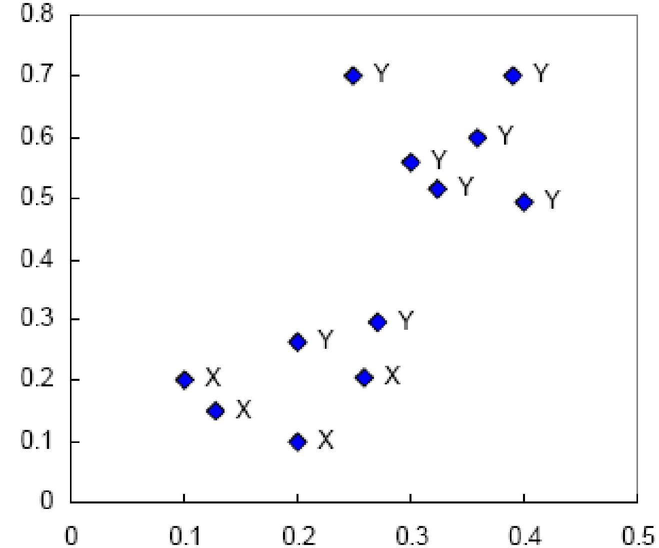
EM Method

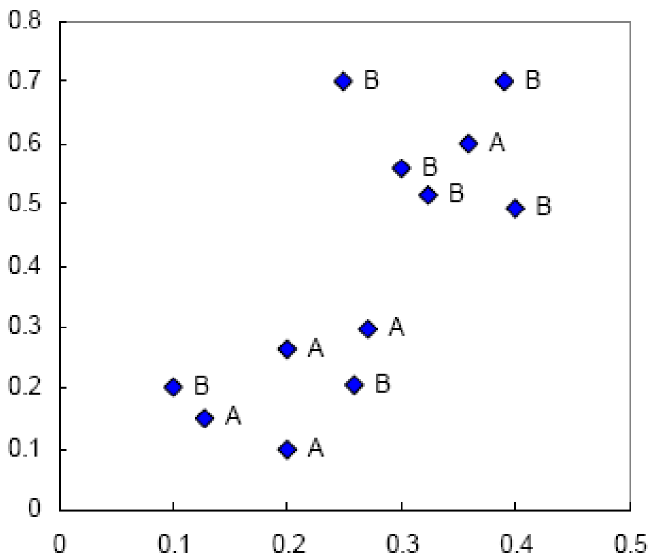
- Maximize log likelihood

$$\sum_{i=1}^{n}\log\left(\sum_{j=1}^{k}\alpha_j P(v_i|\theta_j)\right)$$

- Hidden variables
  - z_i denotes which consensus cluster the object belongs to
- EM procedure
  - E-step: compute expectation of z_i
  - M-step: update model parameters to maximize likelihood


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
2
1
1
1
1
2
2
2
2
2
1
2


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
Y
Y
Y
Y
Y
Y
Y
Y
X
X
X
X


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
B
B
A
B
B
B
A
A
B
B
A
A

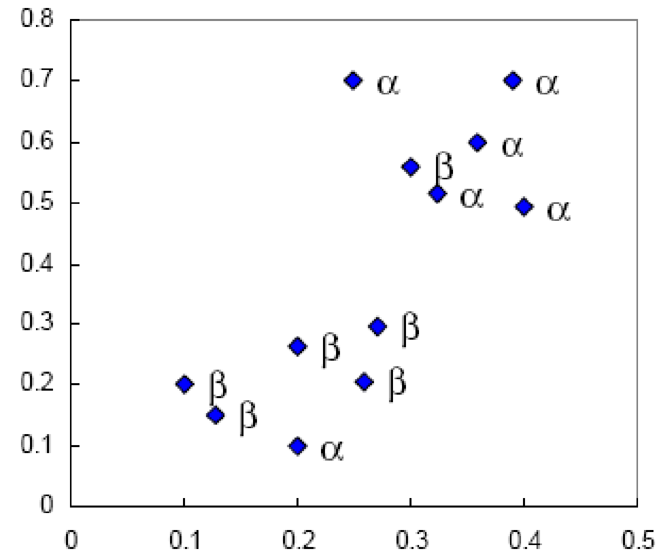
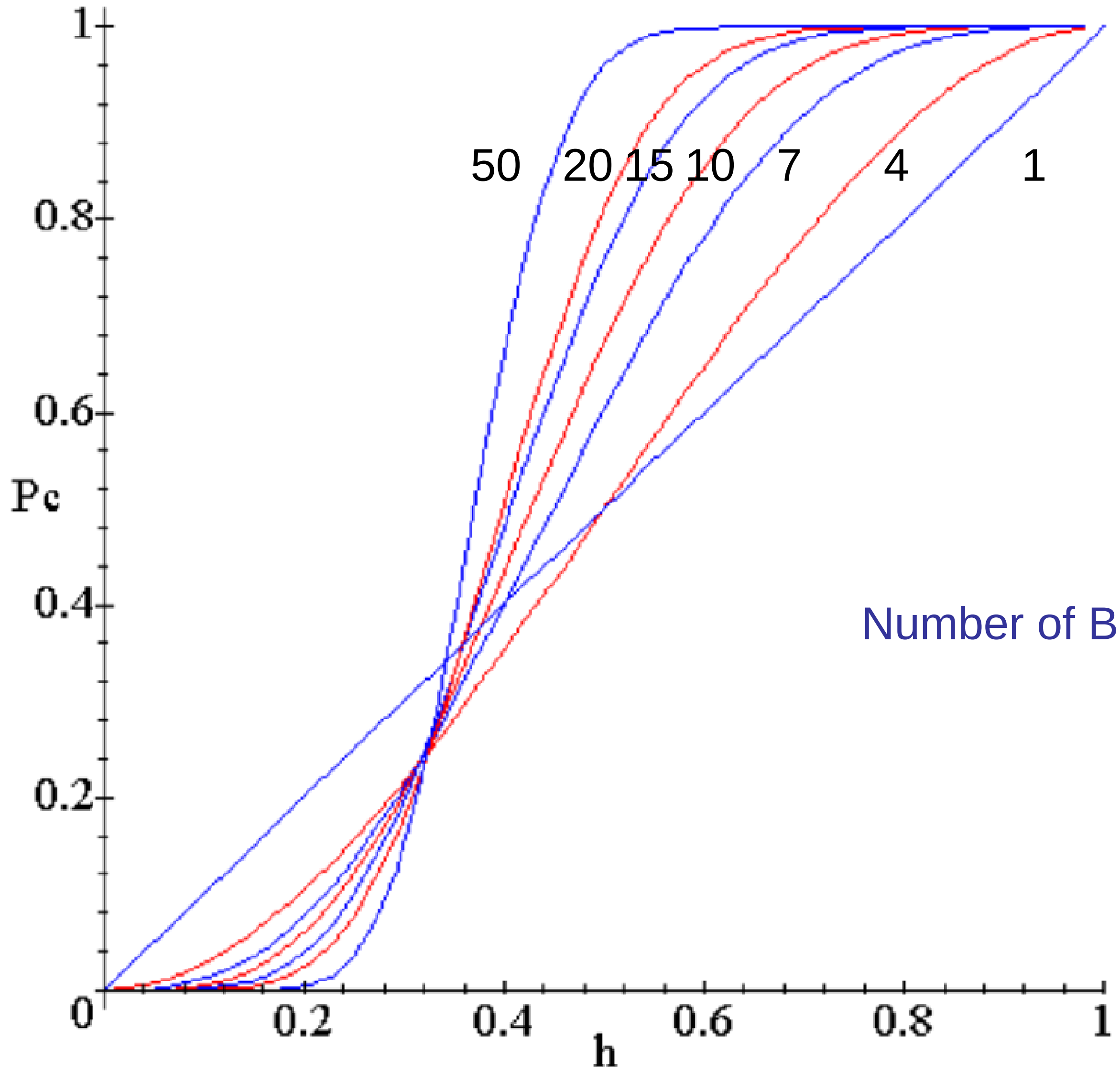

0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
α
α
α
β
α
α
β
β
β
β
β
α

Table 1: Clustering ensemble and consensus solution

	π_1	π_2	π_3	π_4	$E[z_{i1}]$	$E[z_{i2}]$	Consensus
$\mathbf{y}_1$	2	B	X	β	0.999	0.001	**1**
$\mathbf{y}_2$	2	A	X	α	0.997	0.003	**1**
$\mathbf{y}_3$	2	A	Y	β	0.943	0.057	**1**
$\mathbf{y}_4$	2	B	X	β	0.999	0.001	**1**
$\mathbf{y}_5$	1	A	X	β	0.999	0.001	**1**
$\mathbf{y}_6$	2	A	Y	β	0.943	0.057	**1**
$\mathbf{y}_7$	2	B	Y	α	0.124	0.876	**2**
$\mathbf{y}_8$	1	B	Y	α	0.019	0.981	**2**
$\mathbf{y}_9$	1	B	Y	β	0.260	0.740	**2**
$\mathbf{y}_{10}$	1	A	Y	α	0.115	0.885	**2**
$\mathbf{y}_{11}$	2	B	Y	α	0.124	0.876	**2**
$\mathbf{y}_{12}$	1	B	Y	α	0.019	0.981	**2**

Ensemble

Accuracy



Number of Base Models *k*

Base Model Accuracy

[TLJ+04]

Take-away Message

- Clustering ensemble
- Different approaches to combine multiple clustering solutions to one solution