



UNIVERSITY OF
EASTERN FINLAND

Centroid index

Cluster level quality measure

Pasi Fränti

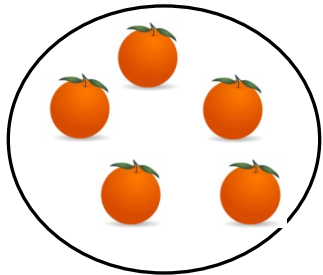
3.9.2018

Clustering accuracy

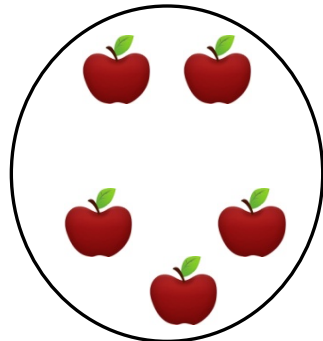
Classification accuracy

Known class labels

Solution A:

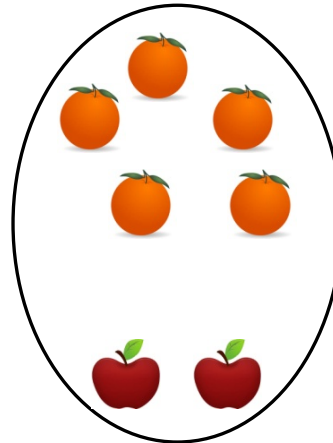


Oranges
100 %

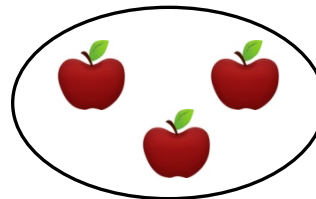


Apples
100 %

Solution B:



Oranges
Precision = $5/7 = 71\%$
Recall = $5/5 = 100\%$

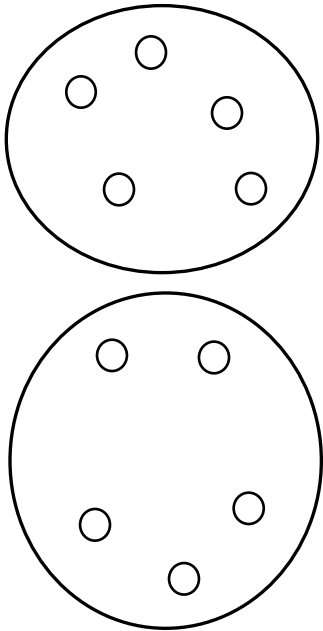


Apples
Precision = $3/3 = 100\%$
Recall = $3/5 = 60\%$

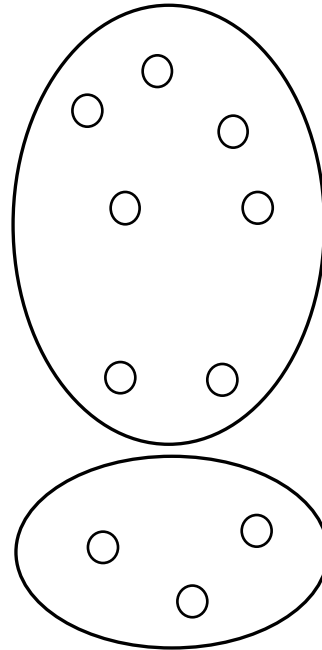
Clustering accuracy

No class labels!

Solution A:



Solution B:



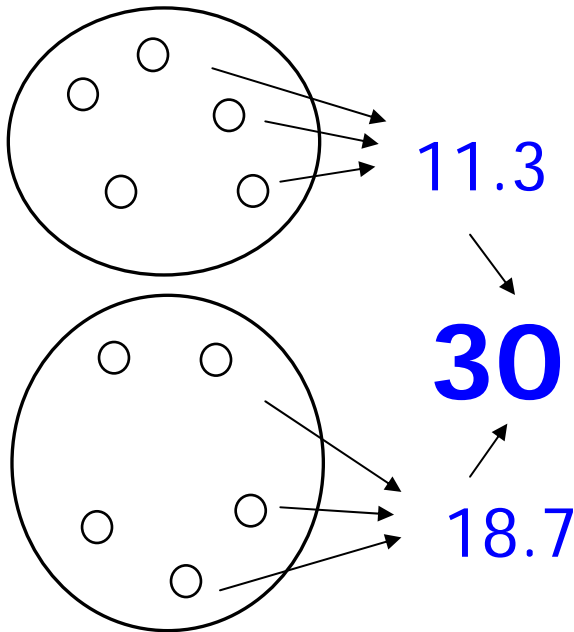
???

Internal index

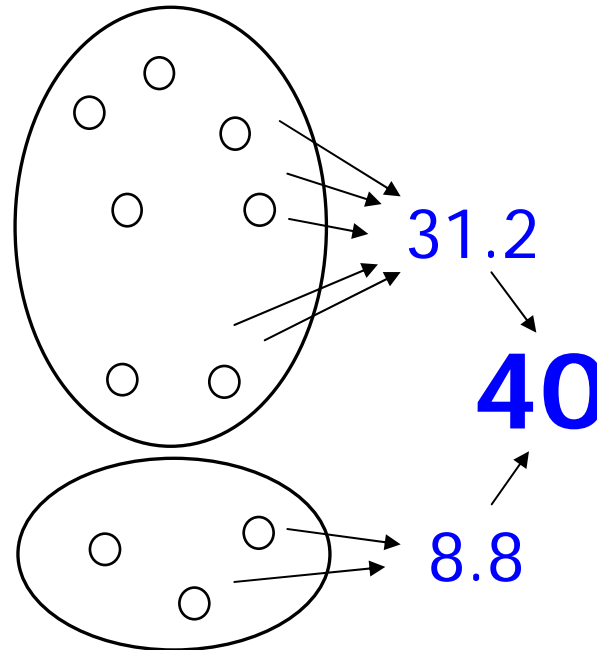
Sum-of-squared error (MSE)

$$f = \frac{1}{N} \sum_{i=1}^N \sum_{x_i \in c_k} \|x_i - c_k\|^2$$

Solution A:



Solution B:

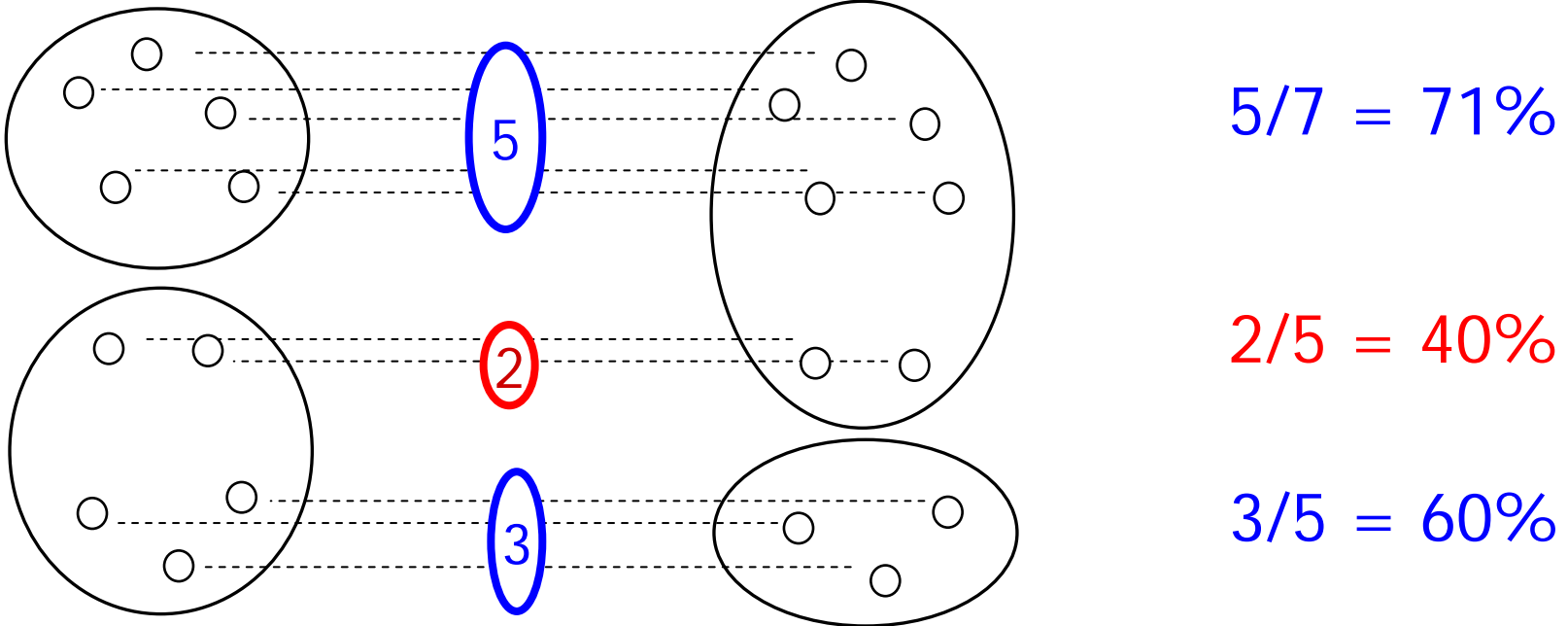


External index

Compare two solutions

Solution A:

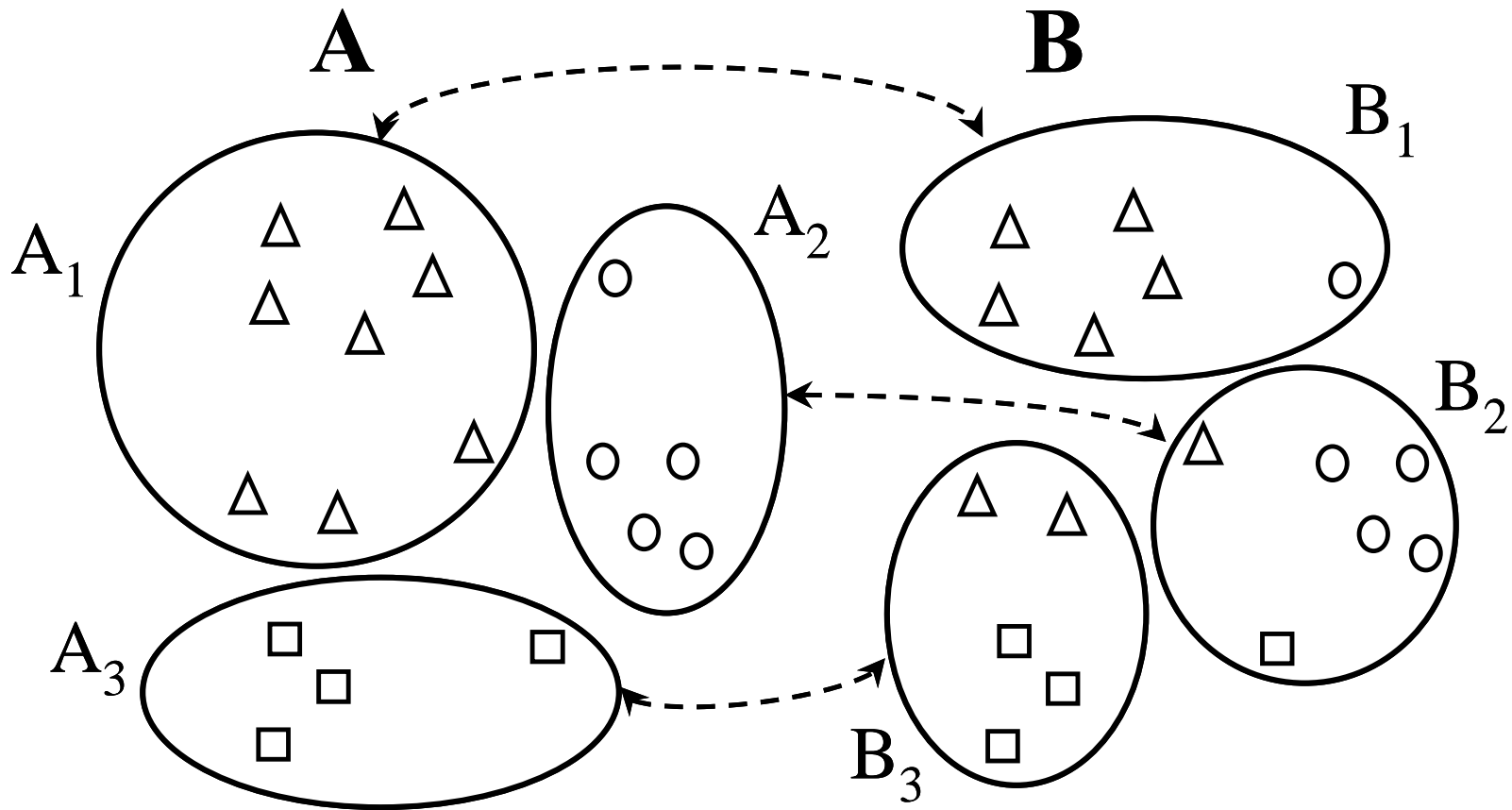
Solution B:



- Two clustering (A and B)
- Clustering against ground truth

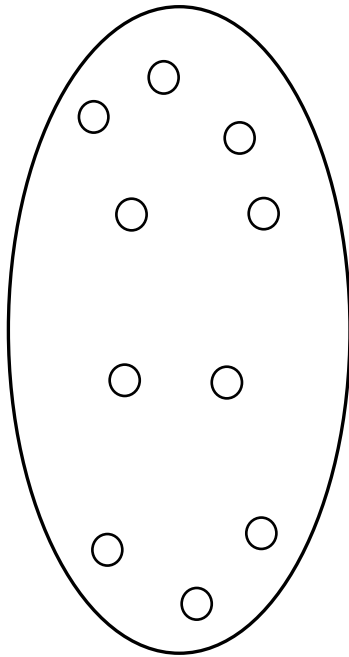
Set-matching based methods

M. Rezaei and P. Fränti, "Set-matching measures for external cluster validity", *IEEE Trans. on Knowledge and Data Engineering*, 28 (8), 2173-2186, August 2016.



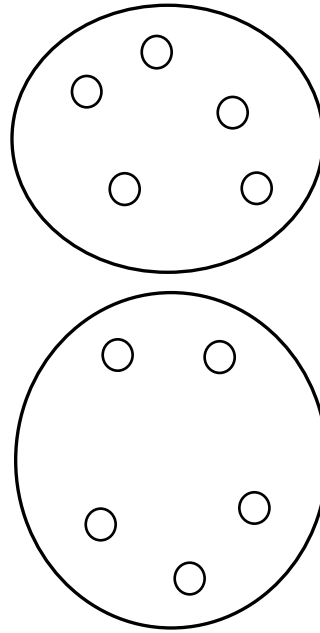
What about this...?

Solution 1:



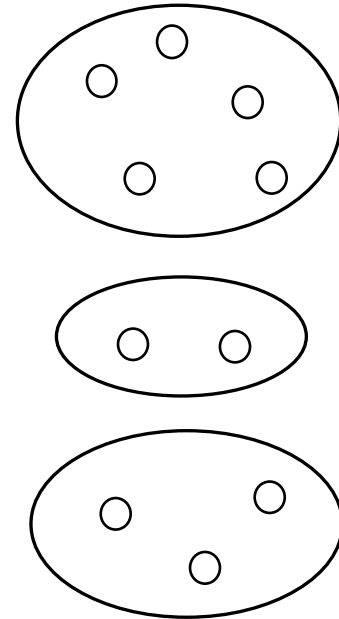
?

Solution 2:



?

Solution 3:



External index

Selection of existed methods

Pair-counting measures

- Rand index (RI) [Rand, 1971]
- Adjusted Rand index (ARI) [Hubert & Arabie, 1985]

Information-theoretic measures

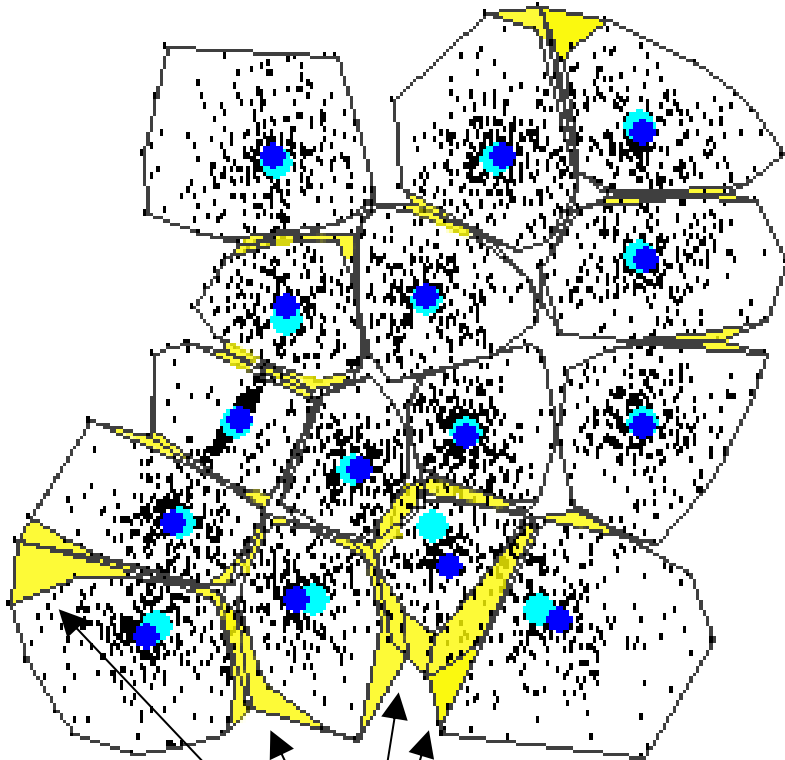
- Mutual information (MI) [Vinh, Epps, Bailey, 2010]
- Normalized Mutual information (NMI) [Kvalseth, 1987]

Set-matching based measures

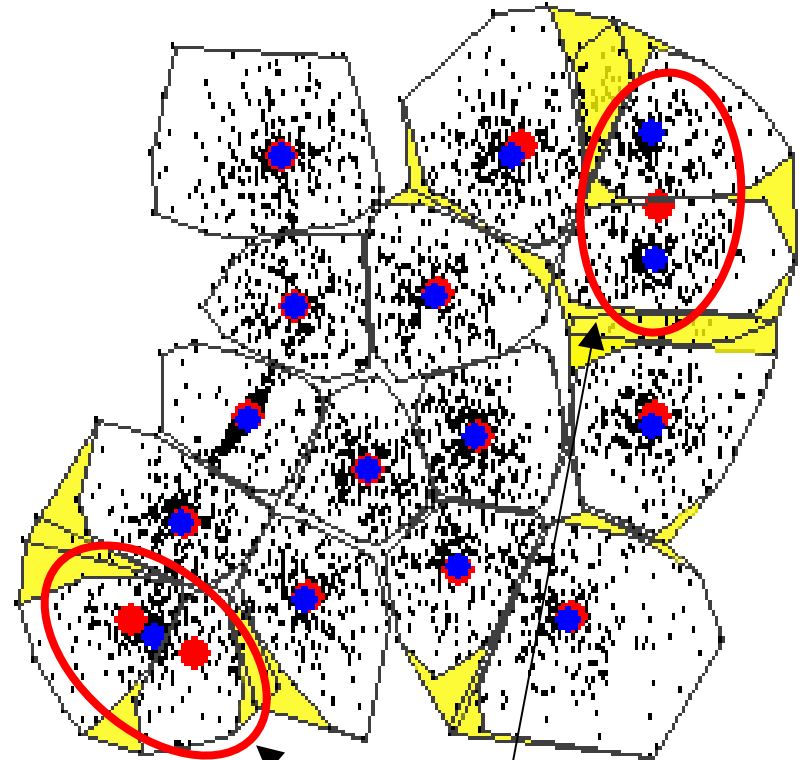
- Normalized van Dongen (NVD) [Kvalseth, 1987]
- Criterion H (CH) [Meila & Heckerman, 2001]
- Purity [Rendon et al, 2011]
- Centroid index (CI) [Fränti, Rezaei & Zhao, 2014]

Cluster level measure

Point level vs. cluster level



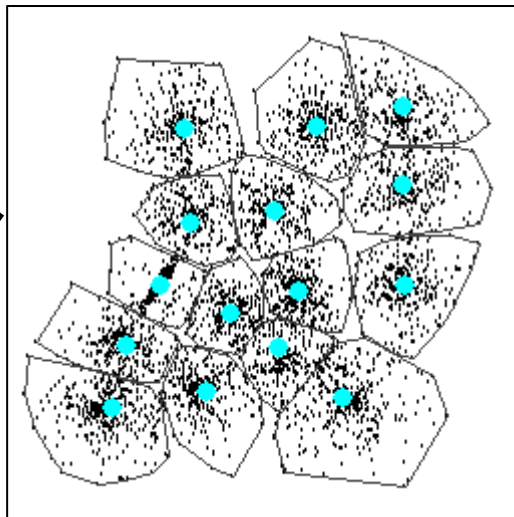
**Point-level
differences**



**Cluster-level
mismatches**

Point level vs. cluster level

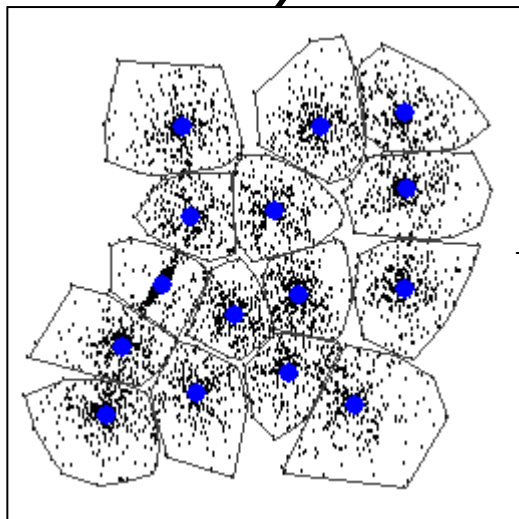
Agglomerative (AC)



ARI=0.91

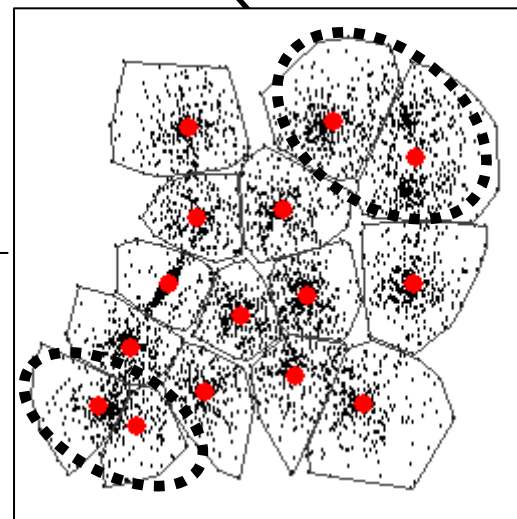
ARI=0.82

Random Swap (RS)

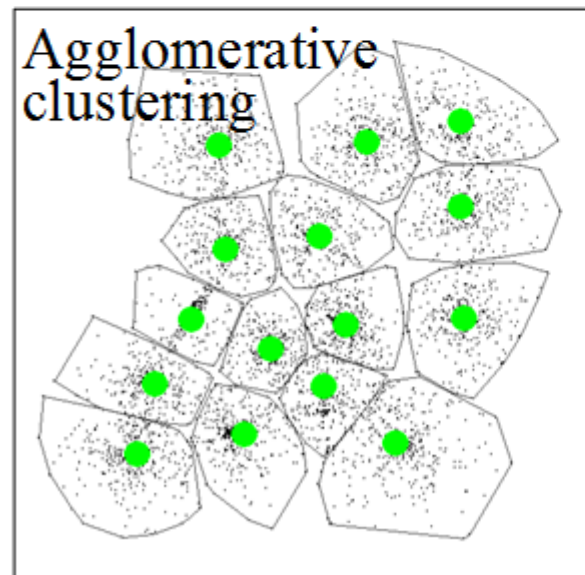
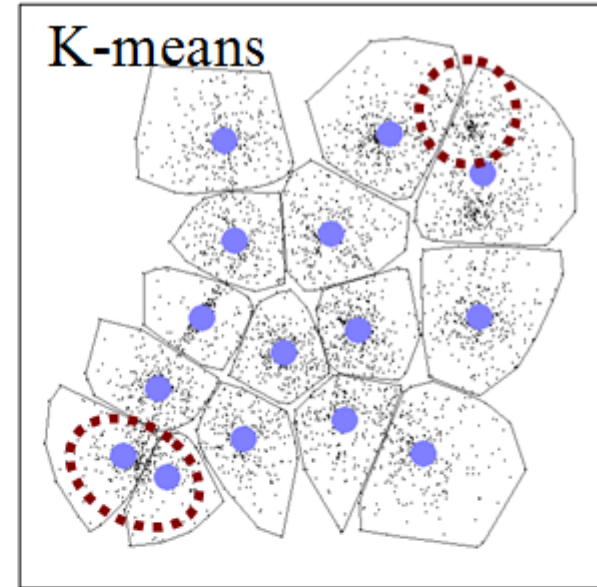
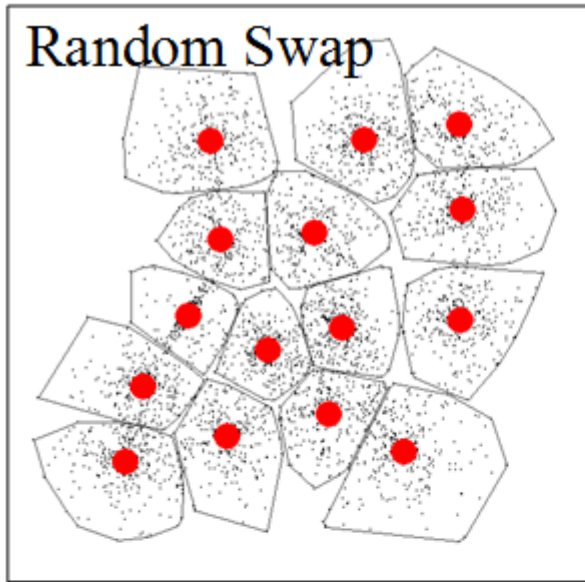


ARI=0.88

K-means



Point level vs. cluster level

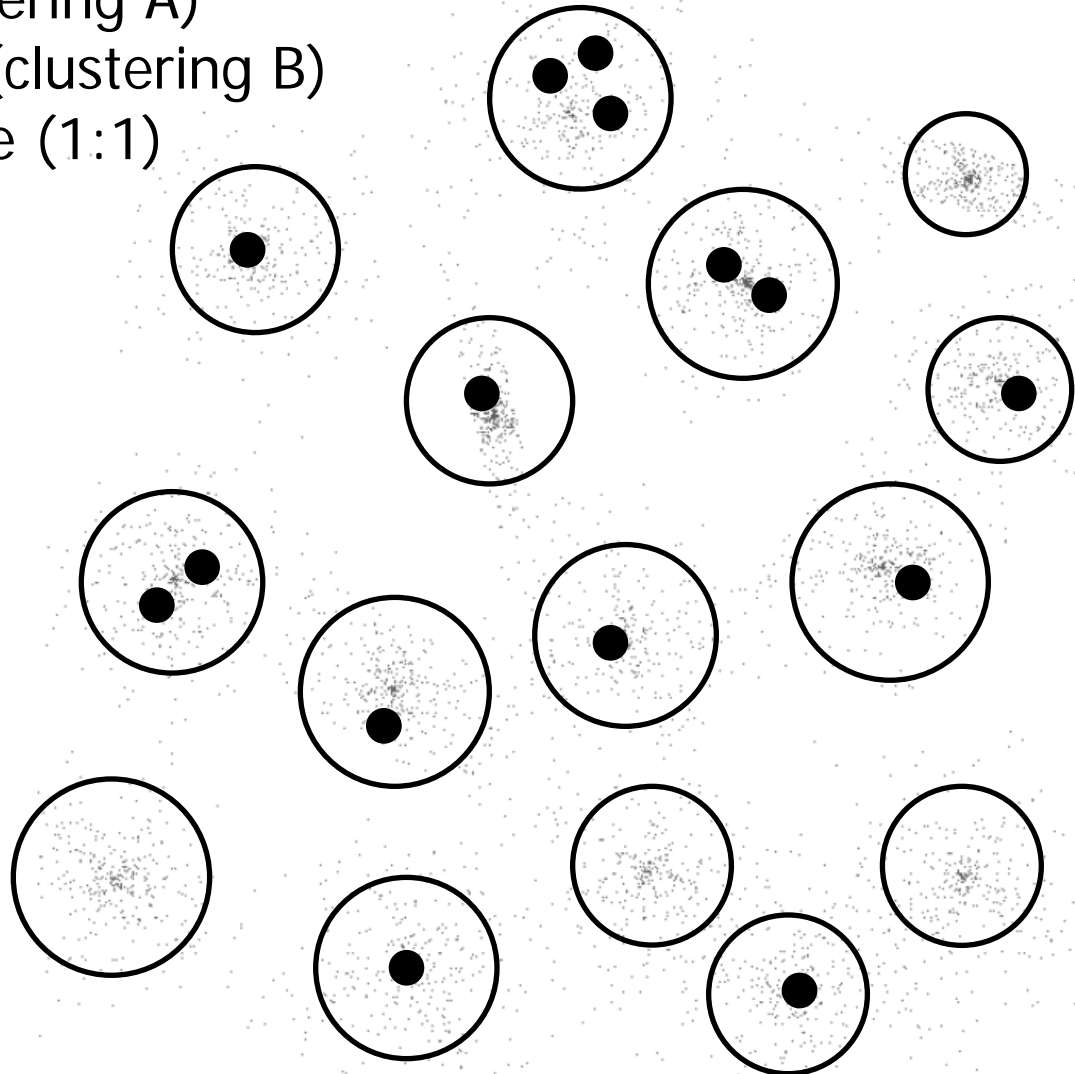


	RS AC	RS KM	AC KM
RI	0.99	0.99	0.98
ARI	0.91	0.88	0.82
MI	3.64	3.64	3.48
NMI	0.93	0.94	0.90
NVD	0.05	0.07	0.10
CH	0.05	0.10	0.14
CI	0	1	1

Centroid index

Pigeon hole principle

- 15 pigeons (clustering A)
- 15 pigeon holes (clustering B)
- Only one bijective (1:1) mapping exists



Definitions

Find nearest centroids ($A \rightarrow B$):

$$NN(A_i) = \arg \min_{1 \leq j \leq k} \|c[A_i] - c[B_j]\|$$

Detect prototypes with no mapping:

$$Orphan(B) = \begin{cases} 1 & InDegree(A) = 0 \\ 0 & InDegree(A) > 0 \end{cases}$$

Number of orphans:

$$CI_1(A \rightarrow B) = \sum_{j=1}^k Orphan(B_j)$$

Number of zero mappings!



Centroid index:

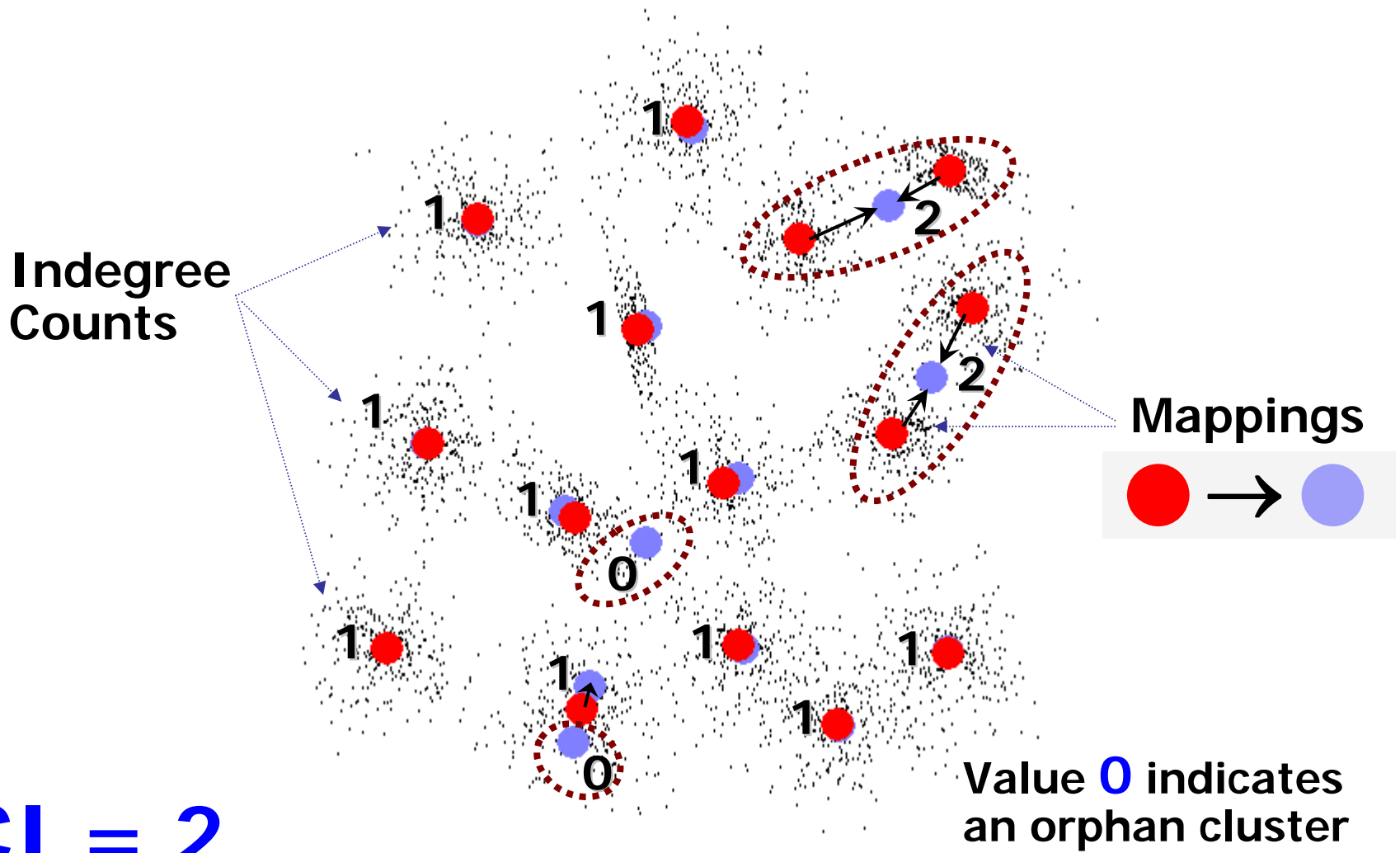
$$CI(A, B) = \max\{CI_1(A \rightarrow B), CI_1(B \rightarrow A)\}$$

Mapping both ways



Example

S_2

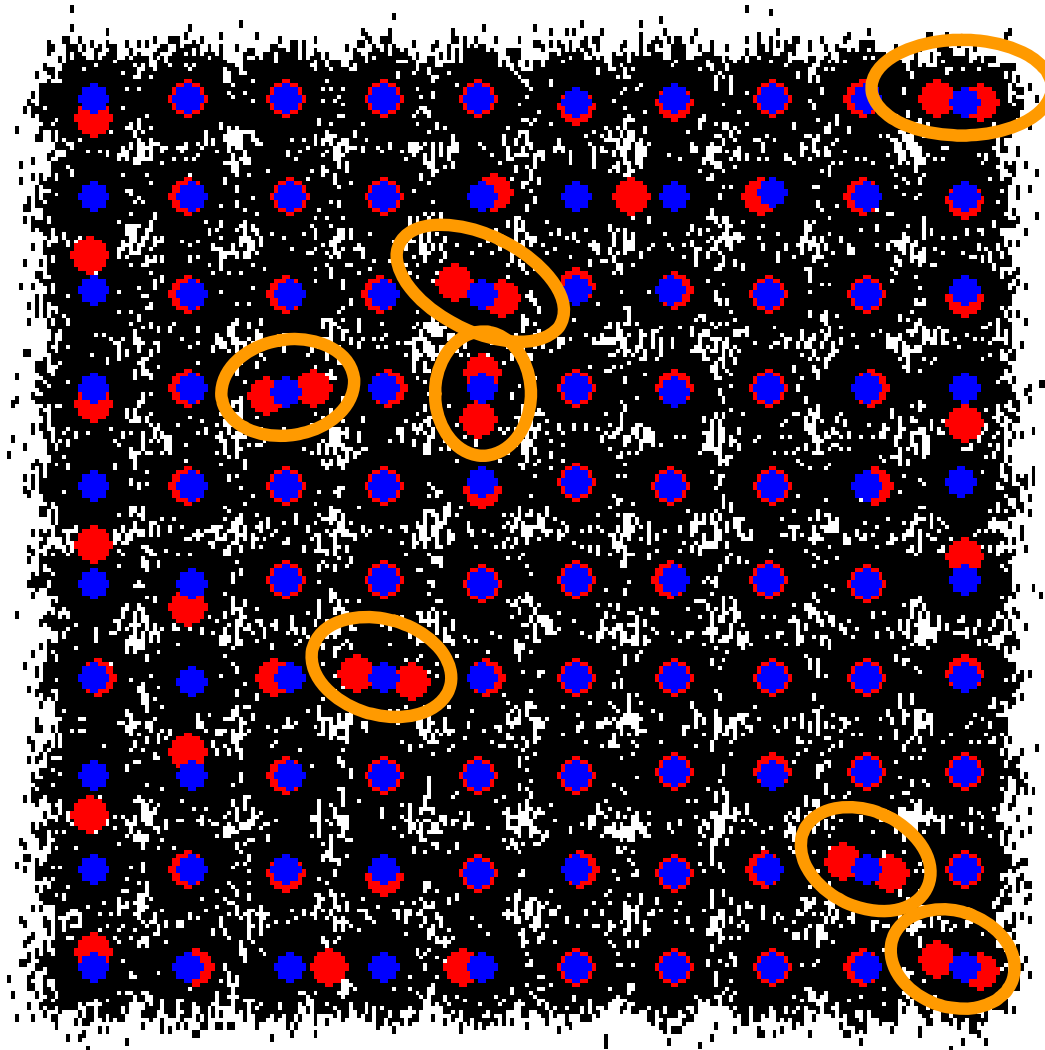
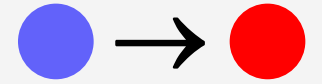


CI = 2

Example

Birch1

Mappings

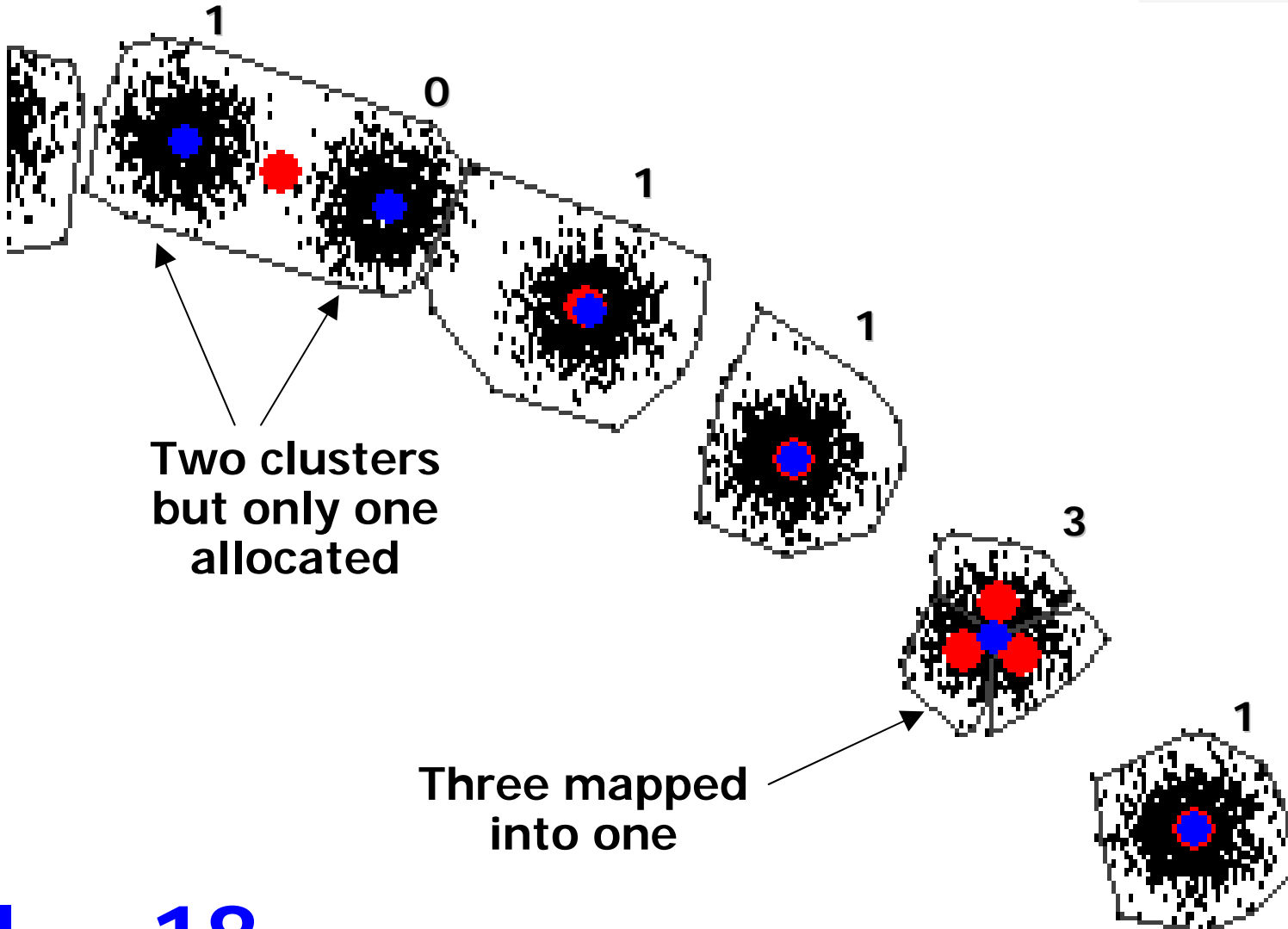
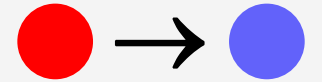


CI = 7

Example

Birch2

Mappings

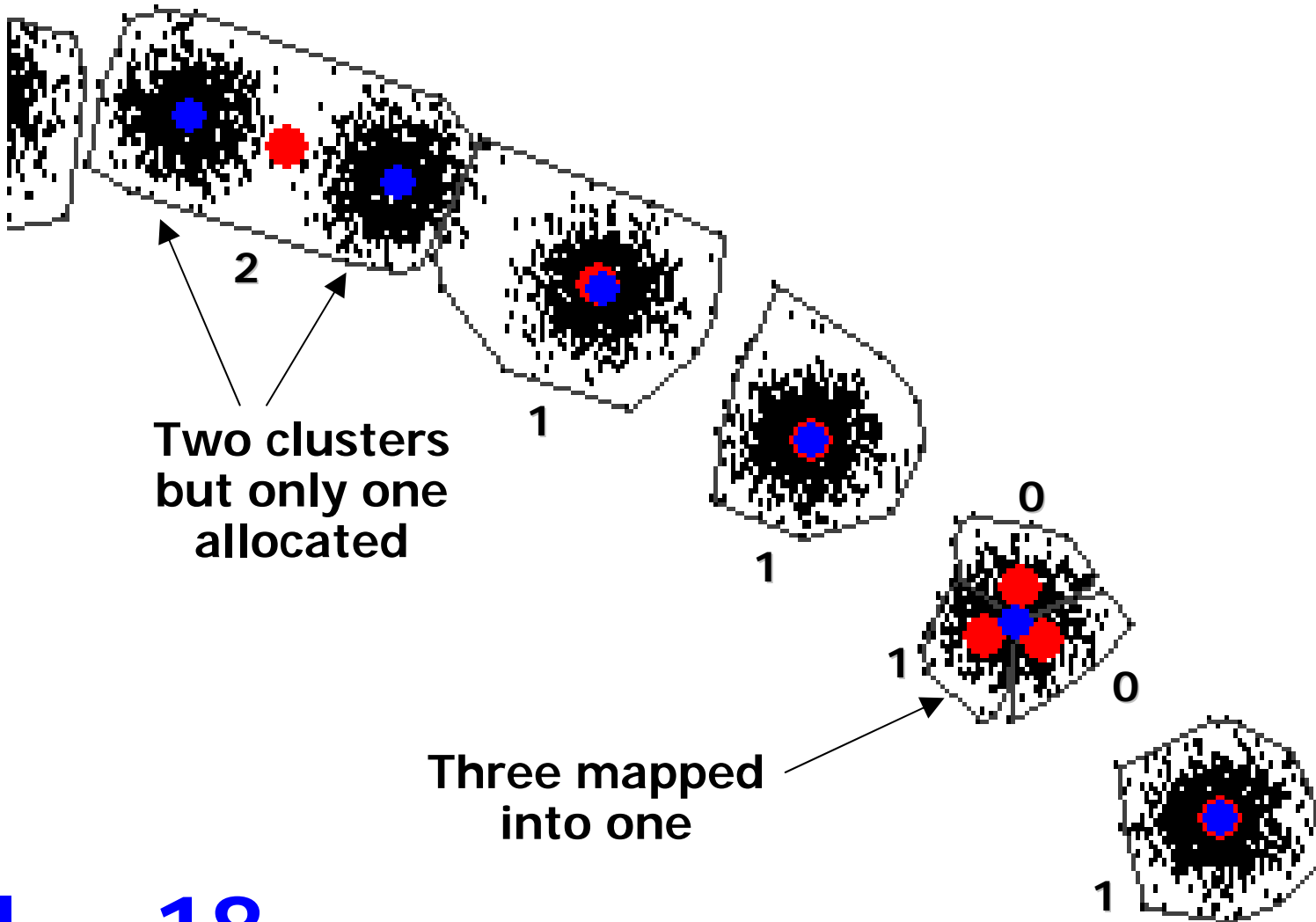
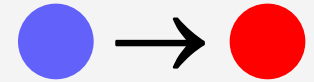


CI = 18

Example

Birch2

Mappings



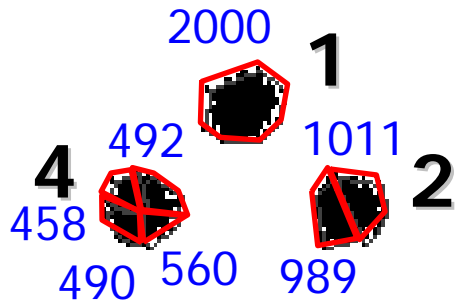
CI = 18

Unbalanced example

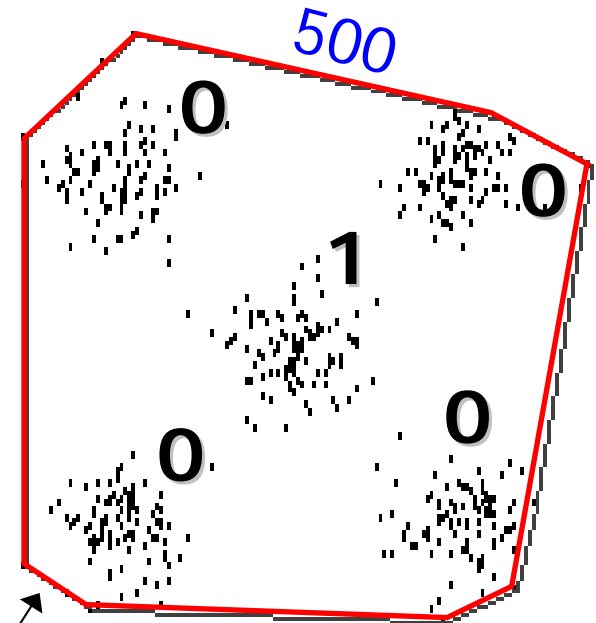
K-means result

KM → GT

CI = 4



K-means tend to put
too many clusters here ...



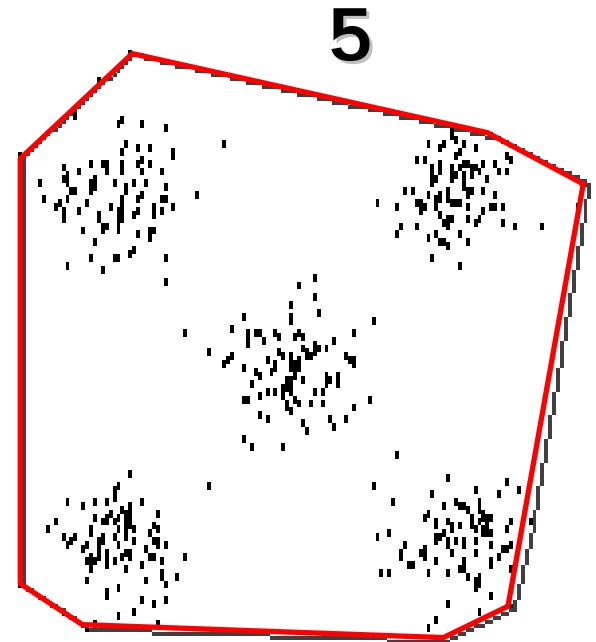
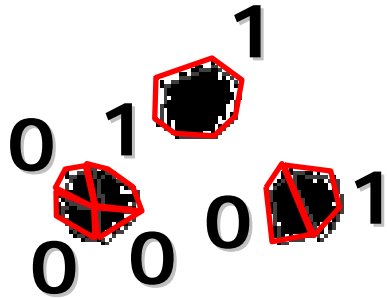
... and too few here

Unbalanced example

K-means result

GT → KM

CI = 4



Experiments

Mean Squared Errors

Green = Correct clustering structure

Data set	Clustering quality (MSE)							
	KM	RKM	KM++	XM	AC	RS	GKM	GA
<i>Bridge</i>	179.76	176.92	173.64	179.73	168.92	164.64	164.78	161.47
<i>House</i>	6.67	6.43	6.28	6.20	6.27	5.96	5.91	5.87
<i>Miss America</i>	5.95	5.83	5.52	5.92	5.36	5.28	5.21	5.10
<i>House</i>	3.61	3.28	2.50	3.57	2.62	2.83	-	2.44
<i>Birch</i> ₁	5.47	5.01	4.88	5.12	4.73	4.64	-	4.64
<i>Birch</i> ₂	7.47	5.65	3.07	6.29	2.28	2.28	-	2.28
<i>Birch</i> ₃	2.51	2.07	1.92	2.07	1.96	1.86	-	1.86
<i>S</i> ₁	19.71	8.92	8.92	8.92	8.92	8.92	8.92	8.92
<i>S</i> ₂	20.58	13.28	13.28	15.87	13.44	13.28	13.28	13.28
<i>S</i> ₃	19.57	16.89	16.89	16.89	17.70	16.89	16.89	16.89
<i>S</i> ₄	17.73	15.70	15.70	15.71	17.52	15.70	15.71	15.70

Raw numbers
don't tell much

Adjusted Rand Index

[Hubert & Arabie, 1985]

Data set	Adjusted Rand Index (ARI)							
	KM	RKM	KM++	XM	AC	RS	GKM	GA
<i>Bridge</i>	0.38	0.40	0.39	0.37	0.43	0.52	0.50	1
<i>House</i>	0.40	0.40	0.44	0.47	0.43	0.53	0.53	1
<i>Miss America</i>	0.19	0.19	0.18	0.20	0.20	0.20	0.23	1
<i>House</i>	0.46	0.49	0.52	0.46	0.49	0.49	-	1
<i>Birch</i> ₁	0.85	0.93	0.98	0.91	0.96	1.00	-	1
<i>Birch</i> ₂	0.81	0.86	0.95	0.86	1	1	-	1
<i>Birch</i> ₃	0.74	0.82	0.87	0.82	0.86	1.00	-	1
<i>S</i> ₁	0.83	1.00	1.00	1.00	1.00	1.00	1.00	1.00
<i>S</i> ₂	0.80	0.99	0.99	0.89	0.98	0.99	0.99	0.99
<i>S</i> ₃	0.86	0.96	0.96	0.96	0.92	0.96	0.96	0.96
<i>S</i> ₄	0.82	0.93	0.93	0.94	0.77	0.93	0.93	0.93

How high is good?

Normalized Mutual information

[Kvalseth, 1987]

Data set	Normalized Mutual Information (NMI)							
	KM	RKM	KM++	XM	AC	RS	GKM	GA
<i>Bridge</i>	0.77	0.78	0.78	0.77	0.80	0.83	0.82	1.00
<i>House</i>	0.80	0.80	0.81	0.82	0.81	0.83	0.84	1.00
<i>Miss America</i>	0.64	0.64	0.63	0.64	0.64	0.66	0.66	1.00
<i>House</i>	0.81	0.81	0.82	0.81	0.81	0.82	-	1.00
<i>Birch</i> ₁	0.95	0.97	0.99	0.96	0.98	1.00	-	1.00
<i>Birch</i> ₂	0.96	0.97	0.99	0.97	1.00	1.00	-	1.00
<i>Birch</i> ₃	0.90	0.94	0.94	0.93	0.93	0.96	-	1.00
<i>S</i> ₁	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00
<i>S</i> ₂	0.90	0.99	0.99	0.95	0.99	0.93	0.99	0.99
<i>S</i> ₃	0.92	0.97	0.97	0.97	0.94	0.97	0.97	0.97
<i>S</i> ₄	0.88	0.94	0.94	0.95	0.85	0.94	0.94	0.94

Normalized Van Dongen

[Kvalseth, 1987]

Data set	Normalized Van Dongen (NVD)							
	KM	RKM	KM++	XM	AC	RS	GKM	GA
<i>Bridge</i>	0.45	0.42	0.43	0.46	0.38	0.32	0.33	0.00
<i>House</i>	0.44	0.43	0.40	0.37	0.40	0.33	0.31	0.00
<i>Miss America</i>	0.60	0.60	0.61	0.59	0.57	0.55	0.53	0.00
<i>House</i>	0.40	0.37	0.34	0.39	0.39	0.34	-	0.00
<i>Birch</i> ₁	0.09	0.04	0.01	0.06	0.02	0.00	-	0.00
<i>Birch</i> ₂	0.12	0.08	0.03	0.09	0.00	0.00	-	0.00
<i>Birch</i> ₃	0.19	0.12	0.10	0.13	0.13	0.13	0.13	0.00
<i>S</i> ₁	0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00
<i>S</i> ₂	0.11	0.00	0.00	0.06	0.01	0.04	0.00	0.00
<i>S</i> ₃	0.08	0.02	0.02	0.02	0.05	0.00	0.00	0.02
<i>S</i> ₄	0.11	0.04	0.04	0.03	0.13	0.04	0.04	0.04

Lower is better

Centroid Similarity Index

[Fränti, Rezaei, Zhao, 2014]

Data set	Centroid Similarity Index (CSI)							
	KM	RKM	KM++	XM	AC	RS	GKM	GA
<i>Bridge</i>	0.47	0.51	0.49	0.45	0.57	0.62	0.63	1.00
<i>House</i>	0.49	0.50	0.54	0.57	0.55	0.63	0.66	1.00
<i>Miss America</i>	0.32	0.32	0.32	0.33	0.38	0.40	0.42	1.00
<i>House</i>	0.54	0.57	0.63	0.54	0.57	0.62	---	1.00
<i>Birch</i> ₁	0.87	0.94	0.98	0.93	0.99	1.00	---	1.00
<i>Birch</i> ₂	0.76	0.84	0.94	0.83	1.00	0.99	0.99	1.00
<i>Birch</i> ₃	0.71	0.82	0.87	0.81	0.86	0.99	0.99	1.00
<i>S</i> ₁	0.83	1.00	1.00	1.00	1.00	1.00	1.00	1.00
<i>S</i> ₂	0.82	1.00	1.00	0.91	1.00	1.00	1.00	1.00
<i>S</i> ₃	0.89	0.99	0.99	0.99	0.98	0.99	0.99	0.99
<i>S</i> ₄	0.87	0.98	0.98	0.99	0.85	0.98	0.98	0.98

Ok but lacks
threshold

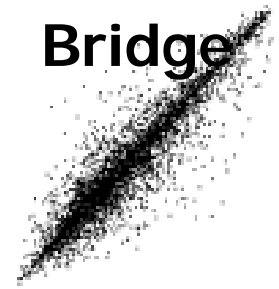
Centroid Index

[Fränti, Rezaei, Zhao, 2014]

Data set	C-Index (CI_2)							
	KM	RKM	KM+ +	XM	AC	RS	GKM	GA
<i>Bridge</i>	74	63	58	81	33	33	35	0
<i>House</i>	56	45	40	37	31	22	20	0
<i>Miss America</i>	88	91	67	88	38	43	36	0
<i>House</i>	43	39	22	47	26	23	---	0
<i>Birch</i> ₁	7	3	1	4	0	0	---	0
<i>Birch</i> ₂	18	11	4	12	0	0	---	0
<i>Birch</i> ₃	23	11	7	10	7	2	---	0
<i>S</i> ₁	2	0	0	0	0	0	0	0
<i>S</i> ₂	2	0	0	1	0	0	0	0
<i>S</i> ₃	1	0	0	0	0	0	0	0
<i>S</i> ₄	1	0	0	0	1	0	0	0



Going deeper...



Accurate clustering

GAIS-2002 similar to GAIS-2012 ?

	Method	MSE
GKM	Global K-means	164.78
RS	Random swap (5k)	164.64
GA	Genetic algorithm	161.47
<hr/>		
RS _{8M}	Random swap (8M)	161.02
GAIS-2002	GAIS	160.72
+ RS _{1M}	GAIS + RS (1M)	160.49
+ RS _{8M}	GAIS + RS (8M)	160.43
<hr/>		
GAIS-2012	GAIS	160.68
+ RS _{1M}	GAIS + RS (1M)	160.45
+ RS _{8M}	GAIS + RS (8M)	160.39
+ PRS	GAIS + PRS	160.33
+ RS _{8M} +	GAIS + RS (8M) + PRS	160.28

-0.04

-0.29

-0.29

GAIS-2002

GAIS-2012

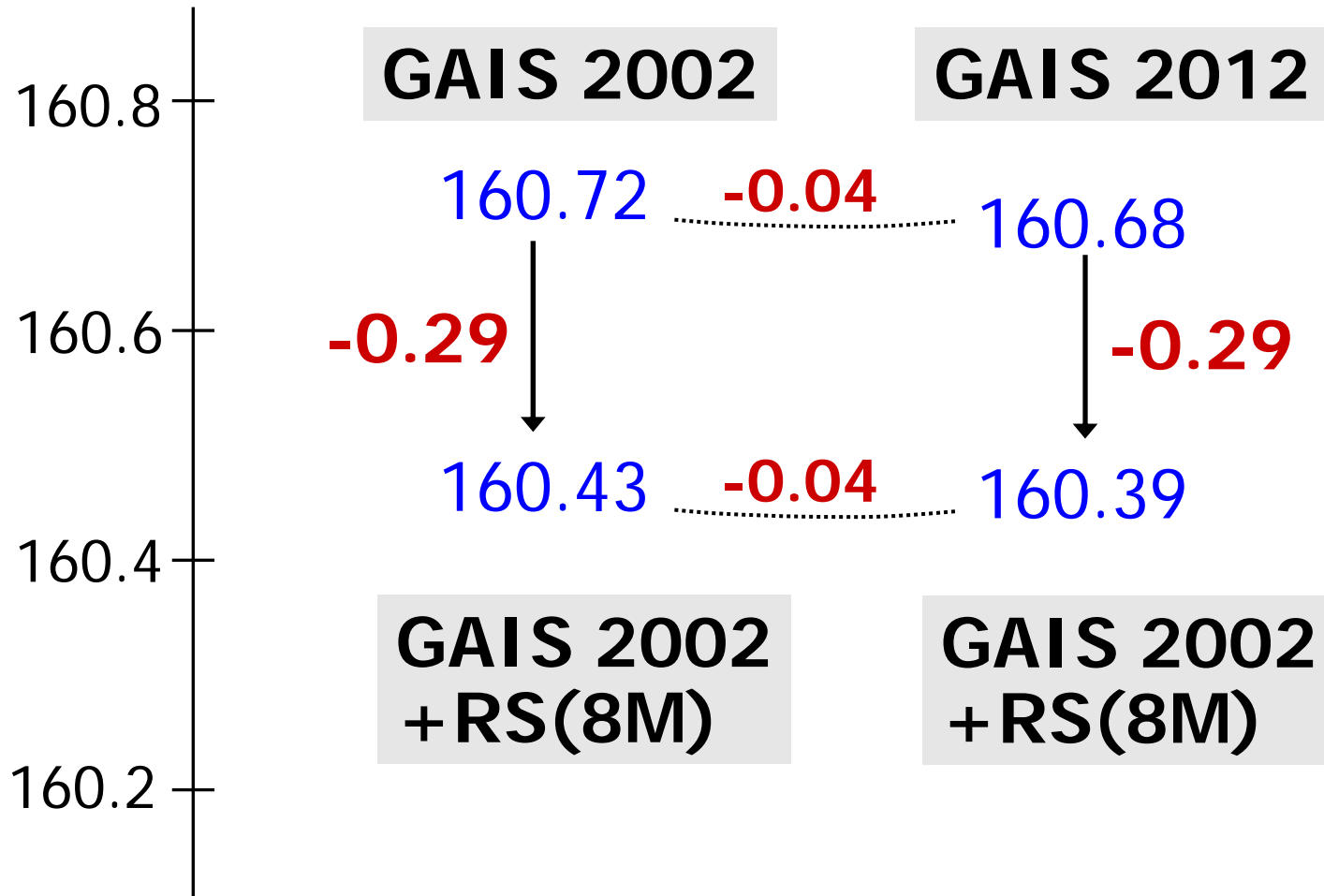
160.72

160.68



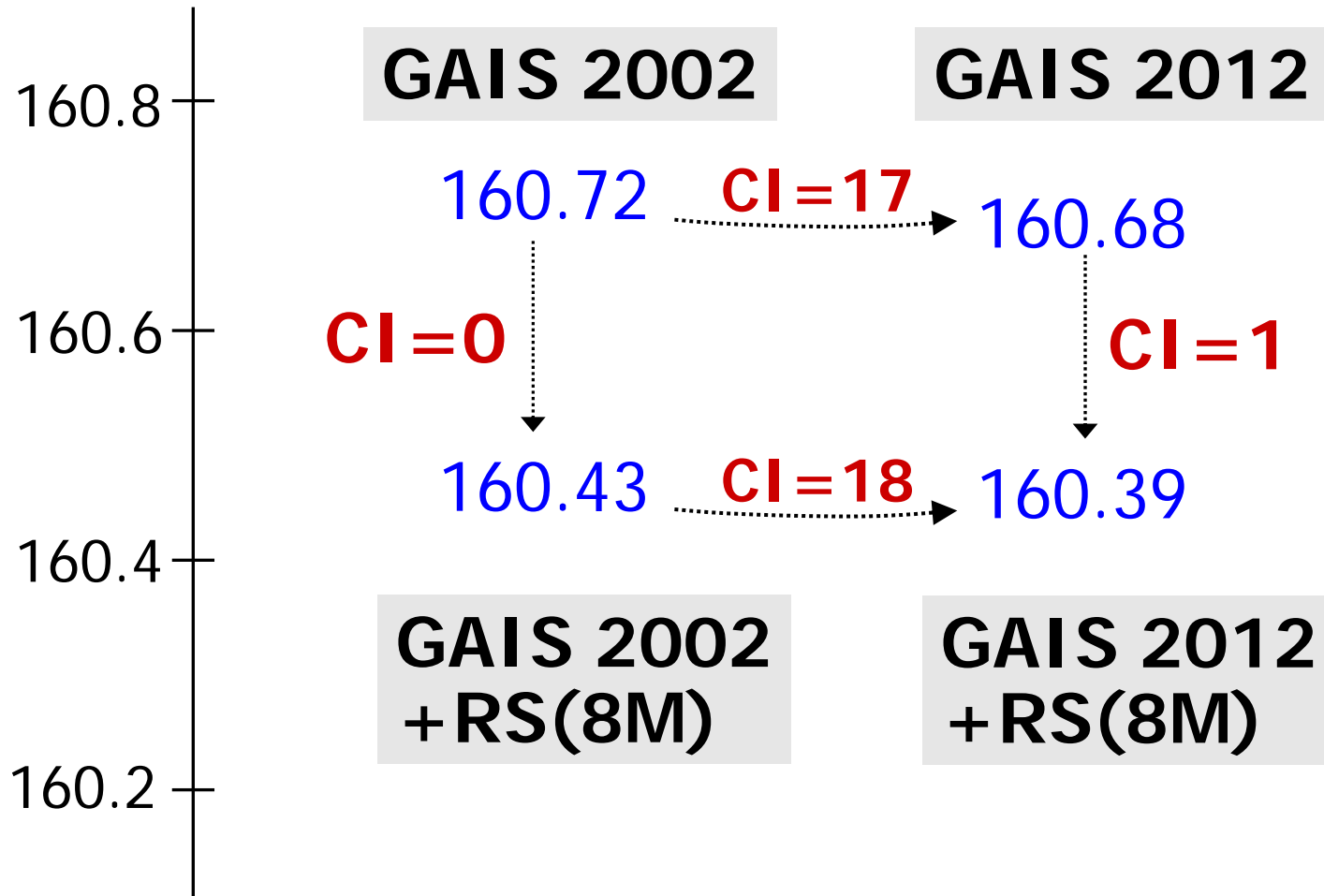
GAIS'02 and GAIS'12 the same?

Virtually the same MSE-values



GAIS'02 and GAIS'12 the same?

But different structure!



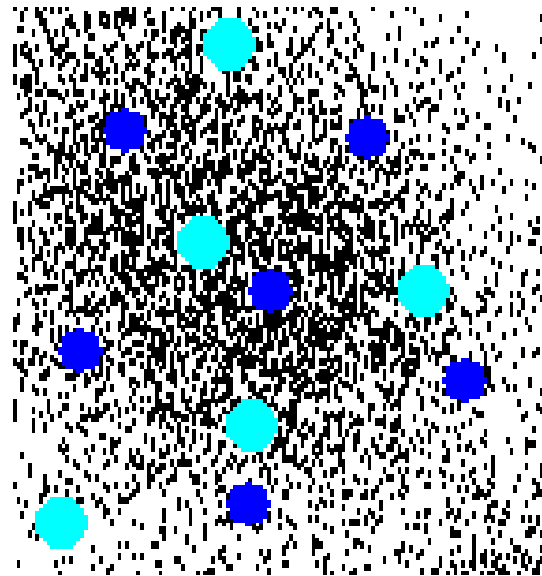
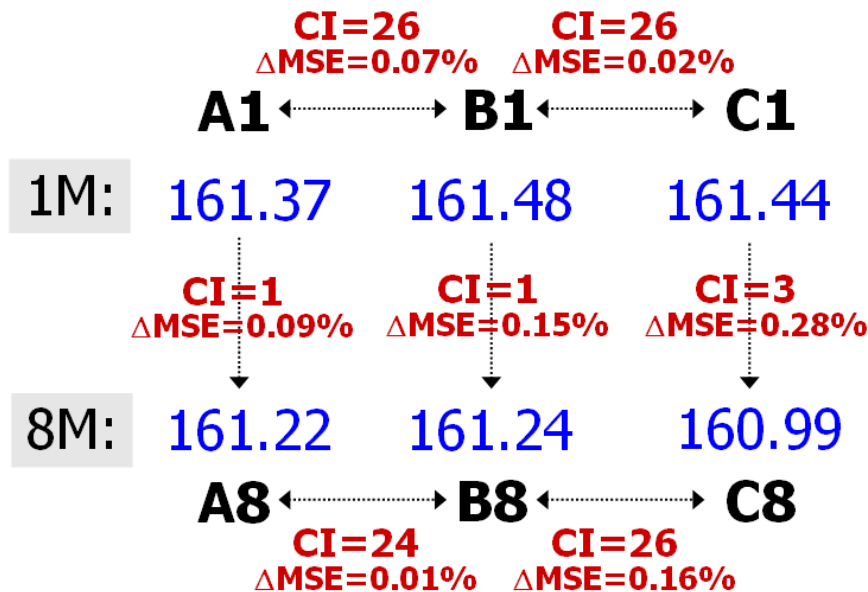
Seemingly the same solutions

Same structure
"same family"

Main algorithm: + Tuning 1 + Tuning 2	RS _{8M}	GAIS 2002			GAIS 2012				
		×	RS _{1M}	RS _{8M}	×	RS _{1M}	RS _{8M}	×	RS _{8M}
RS _{8M}	---	19	19	19	23	24	24	23	22
GAIS (2002)	23	---	0	0	14	15	15	14	16
+ RS _{1M}	23	0	---	0	14	15	15	14	13
+ RS _{8M}	23	0	0	---	14	15	15	14	13
GAIS (2012)	25	17	18	18	---	1	1	1	1
+ RS _{1M}	25	17	18	18	1	---	0	0	1
+ RS _{8M}	25	17	18	18	1	0	---	0	1
+ PRS	25	17	18	18	1	0	0	---	1
+ RS _{8M} + PRS	24	17	18	18	1	1	1	1	---

But why?

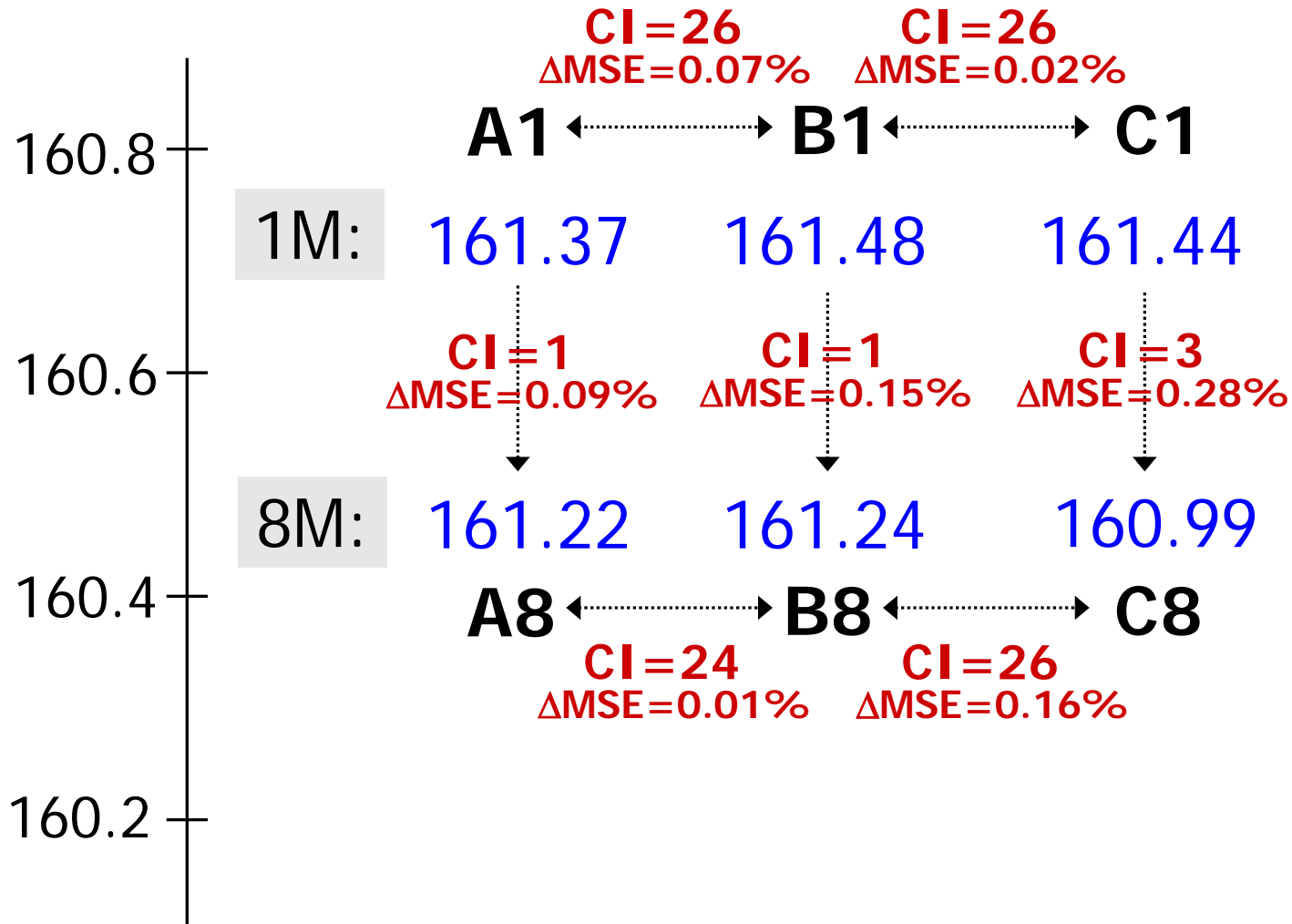
- Real cluster structure missing
- Clusters allocated like well optimized "grid"
- Several grids results different allocation
- Overall clustering quality can still be the same



RS runs

	A1	A8	B1	B8	C1	C8
A1	-	1	26	24	25	25
A8	1	-	26	24	25	25
B1	26	26	-	1	26	24
B8	24	24	1	-	25	24
C1	25	25	26	25	-	3
C8	25	25	24	24	3	-

RS runs 1M vs 8M

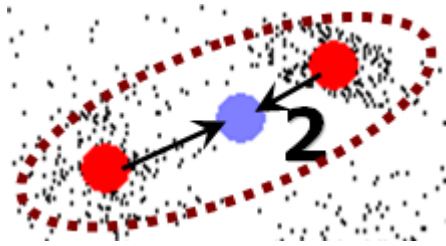


Generalization

Three alternatives

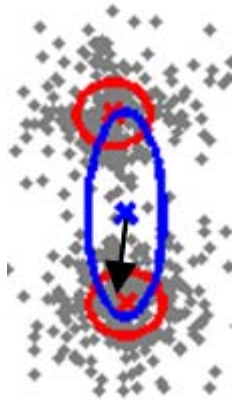
1. Prototype similarity

Prototype must exist

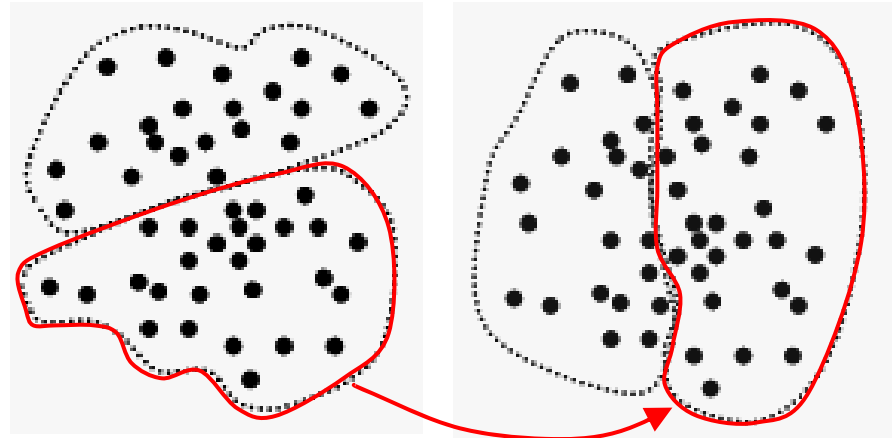


2. Model similarity

Derived from model



3. Partition similarity

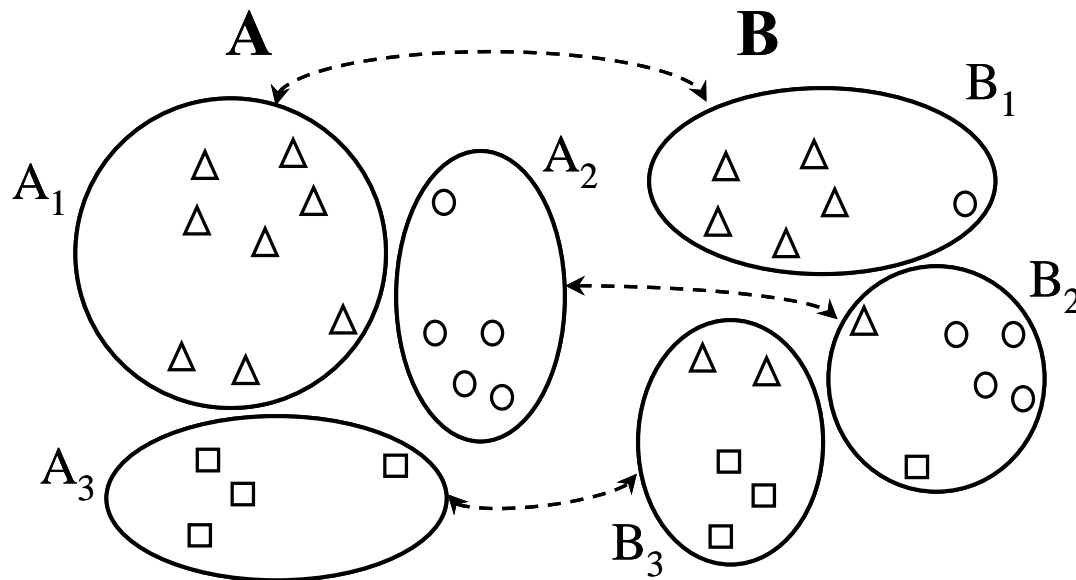


$$J = \frac{|P_i \cap G_j|}{|P_i \cup G_j|}$$

Partition similarity

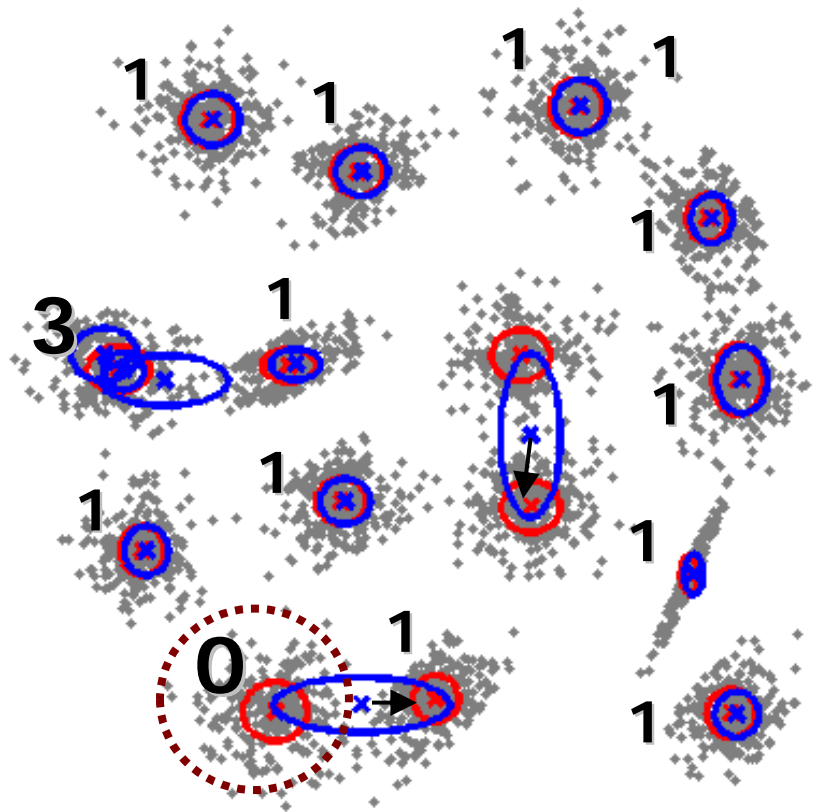
- Cluster similarity using Jaccard
- Calculated from contingency table

	B_1	B_2	B_3	Σ
A_1	5	1	2	8
A_2	1	4	0	5
A_3	0	1	3	4
Σ	6	6	5	17



Gaussian mixture model

$$S_{BC} = \frac{1}{8} (c[A_i] - c[B_j])^T \Sigma^{-1} (c[A_i] - c[B_j]) + \frac{1}{2} \ln \left(\frac{|\Sigma|}{\sqrt{|\Sigma_1| |\Sigma_2|}} \right)$$



RI	0.98
ARI	0.84
<hr/>	
MI	3.60
NMI	0.94
<hr/>	
NVD	0.08
CH	0.16

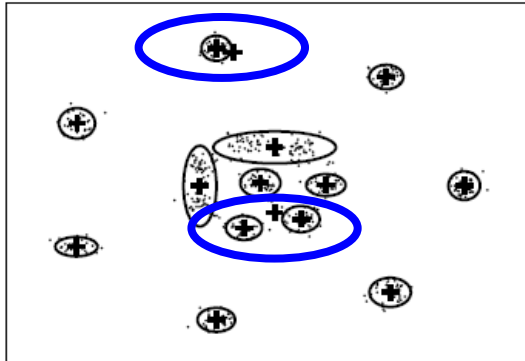
CI=2

- Split-and-Merge EM
- Random Swap EM

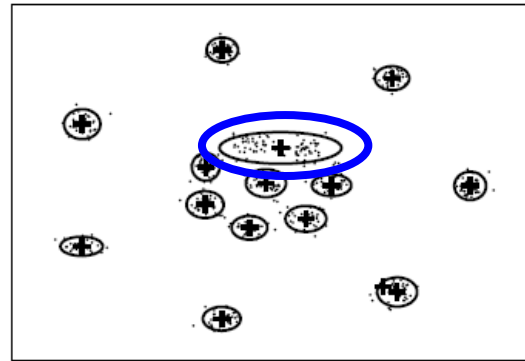
Gaussian mixture model

CI=2

REM:-6.42



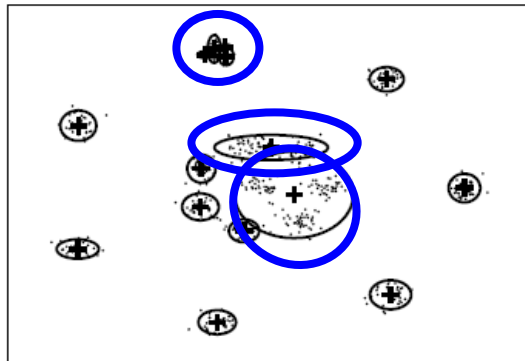
REM:-6.37



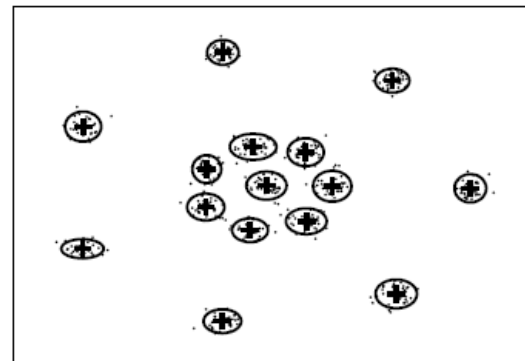
CI=1

CI=3

SMEM:-6.53



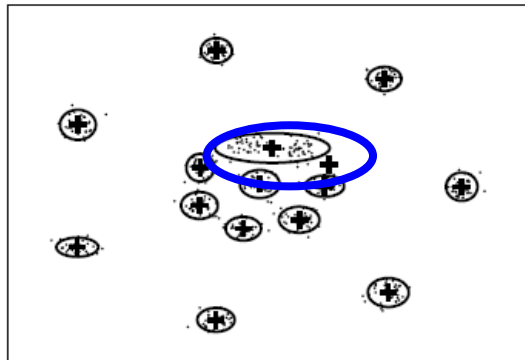
SMEM:-6.33



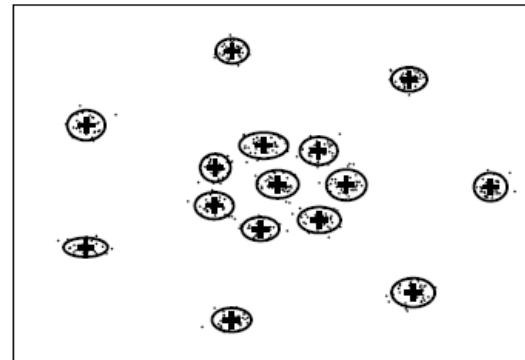
CI=0

CI=1

RSEM:-6.37

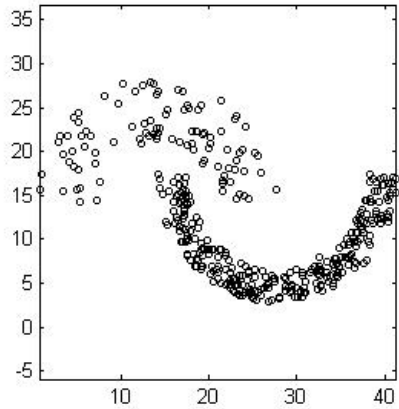


RSEM:-6.33

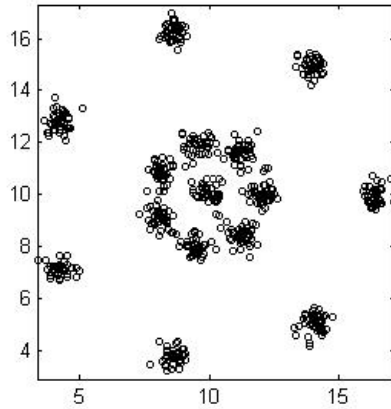


CI=0

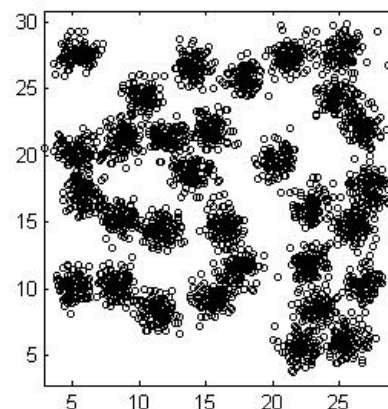
Arbitrary-shape data



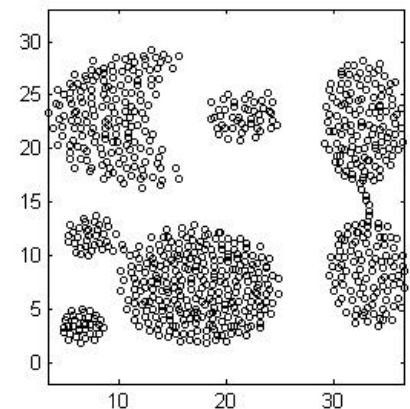
A.K. Jain's Toy problem



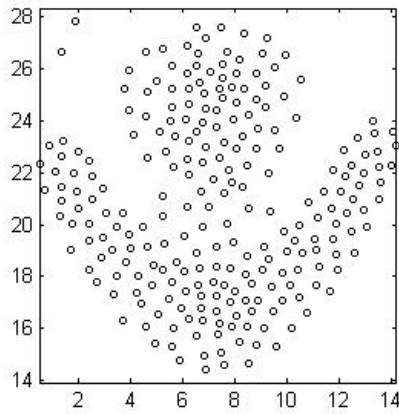
R15



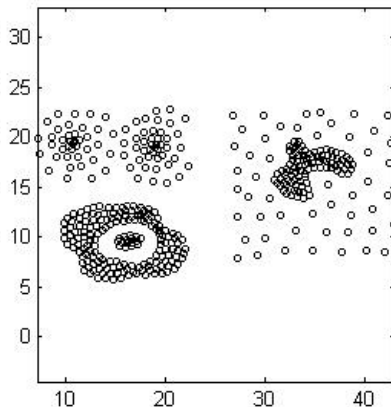
D31



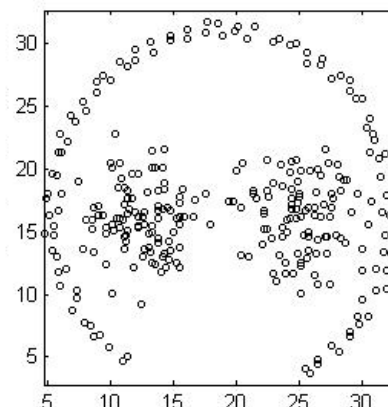
Aggregation



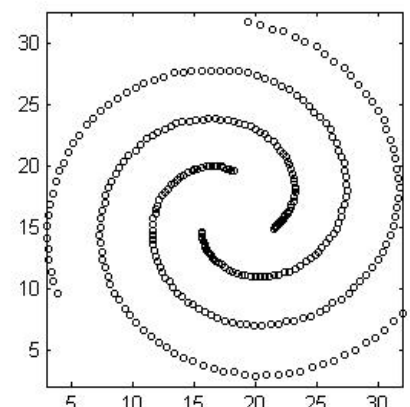
Flame



Zahn's Compound

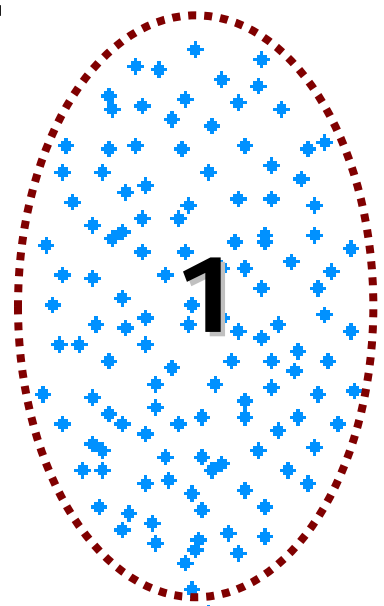
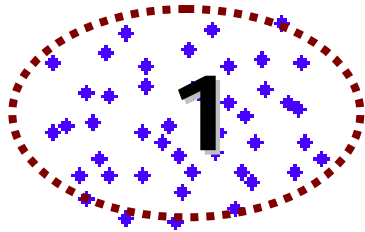
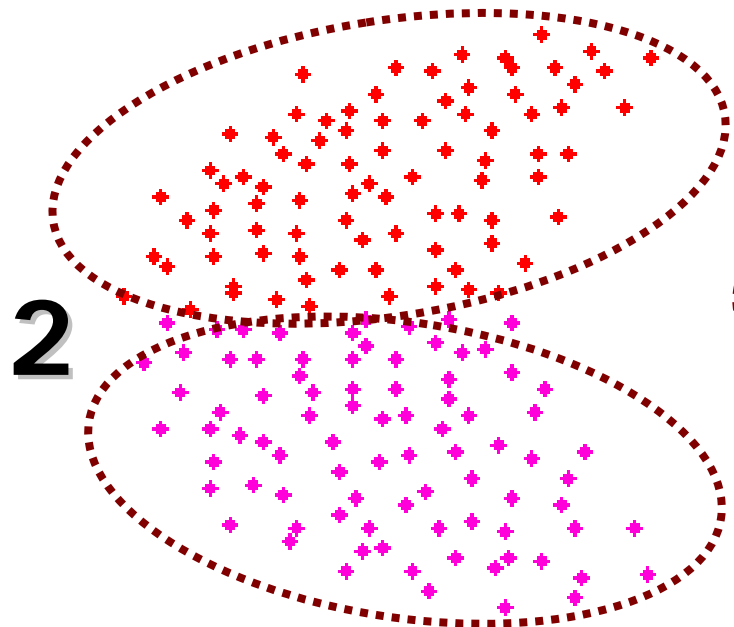


Path-based1

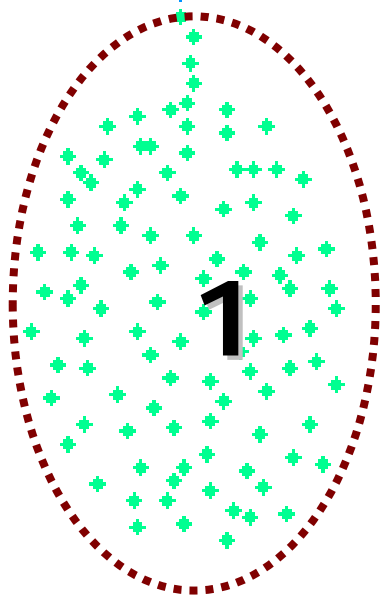
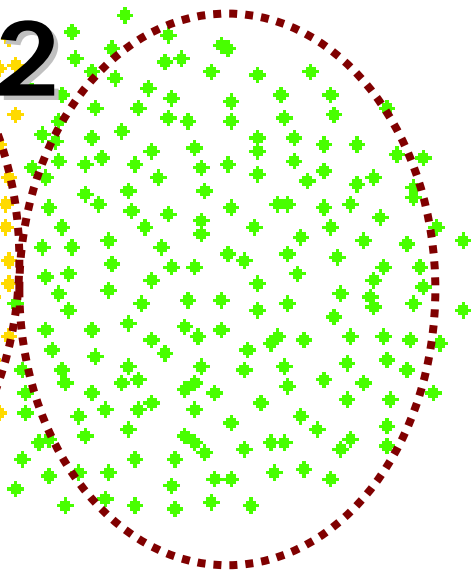
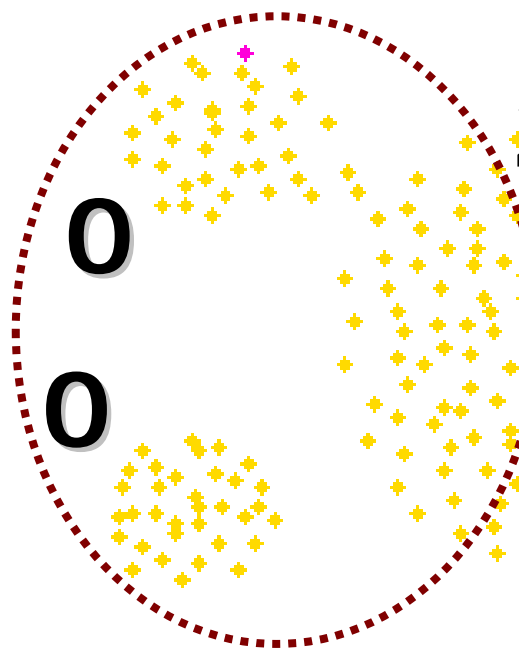


path-based2: spiral

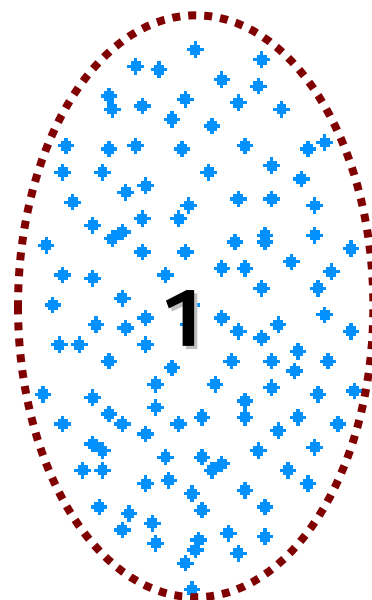
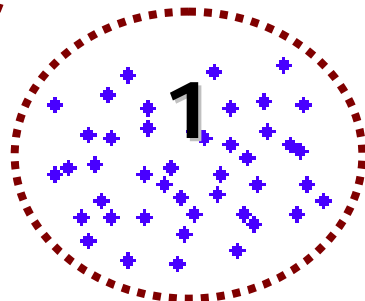
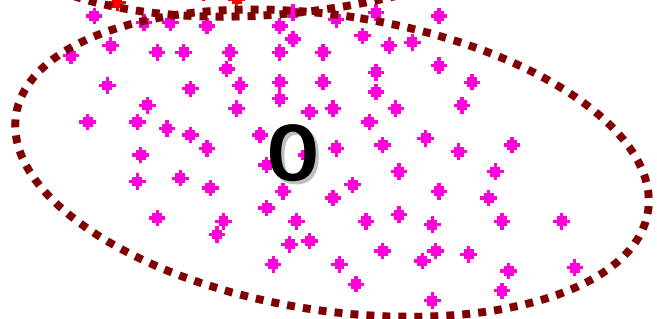
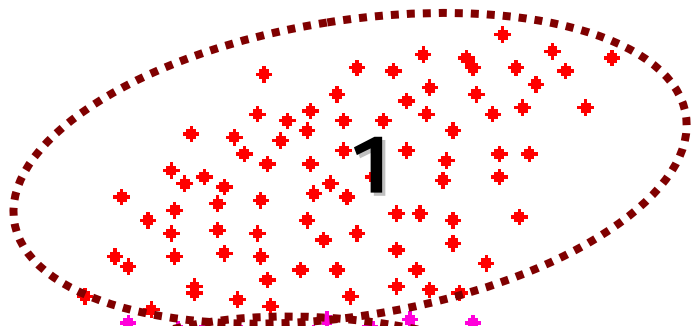
KM→GT



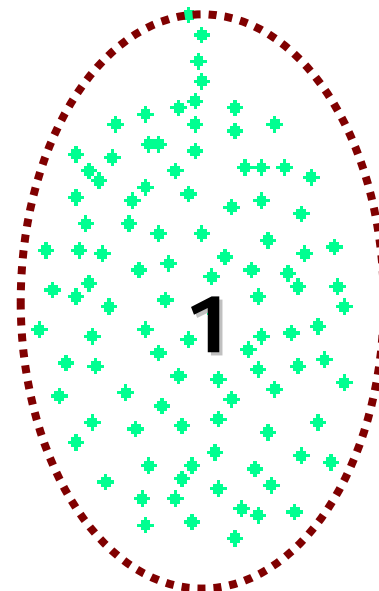
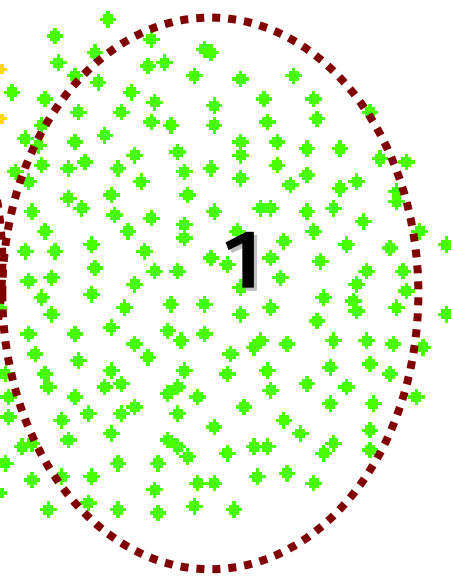
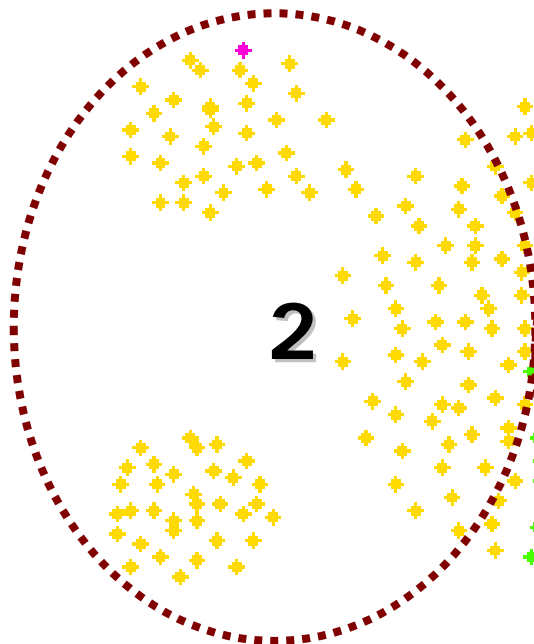
CI = 2



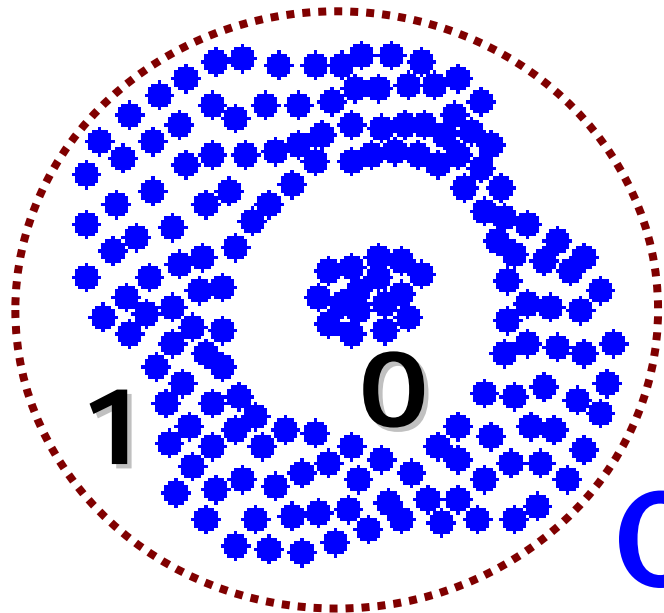
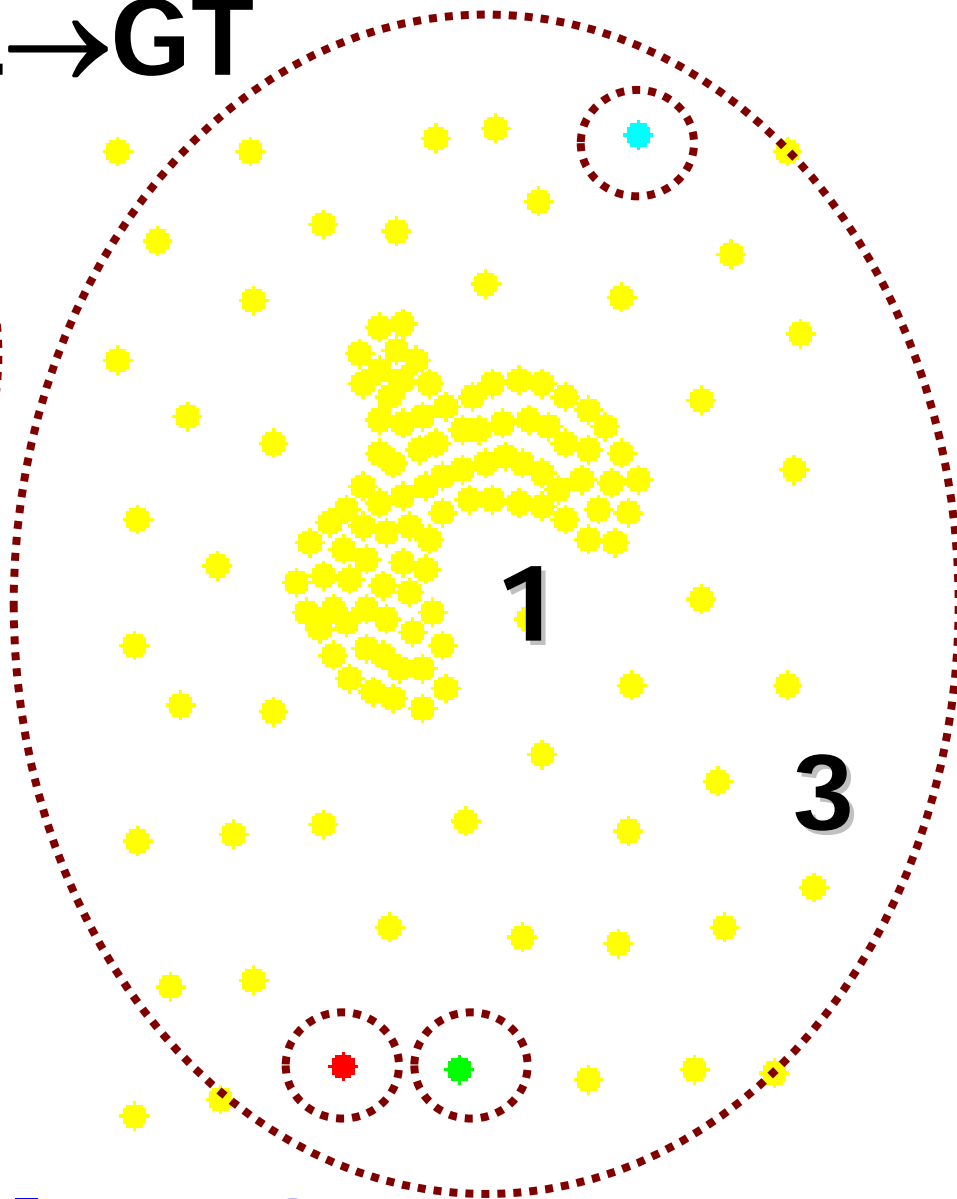
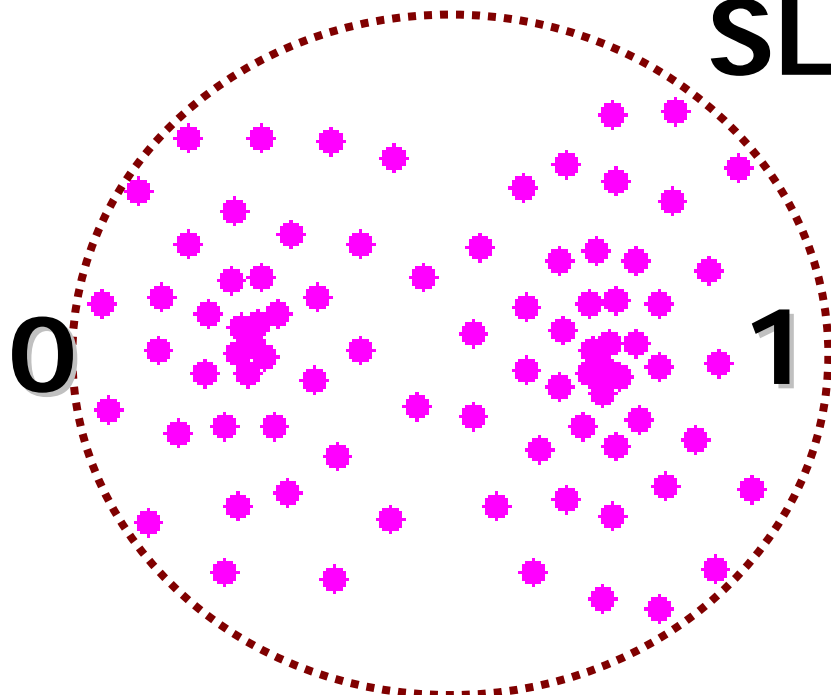
GT → KM



CI = 1

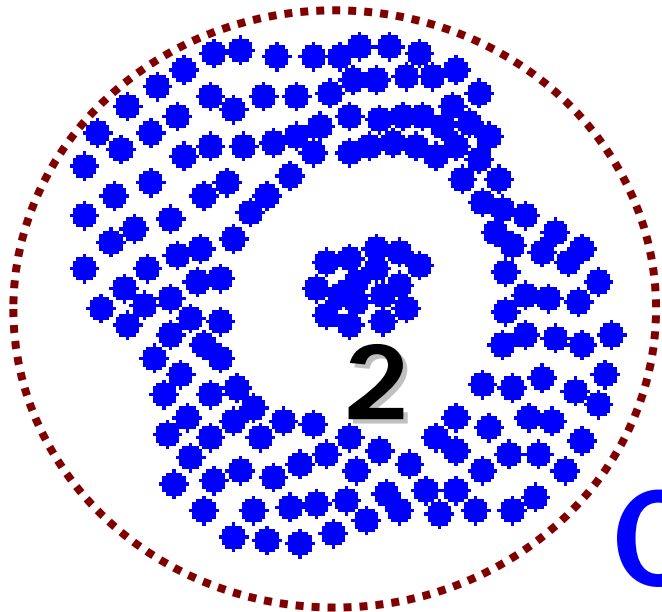
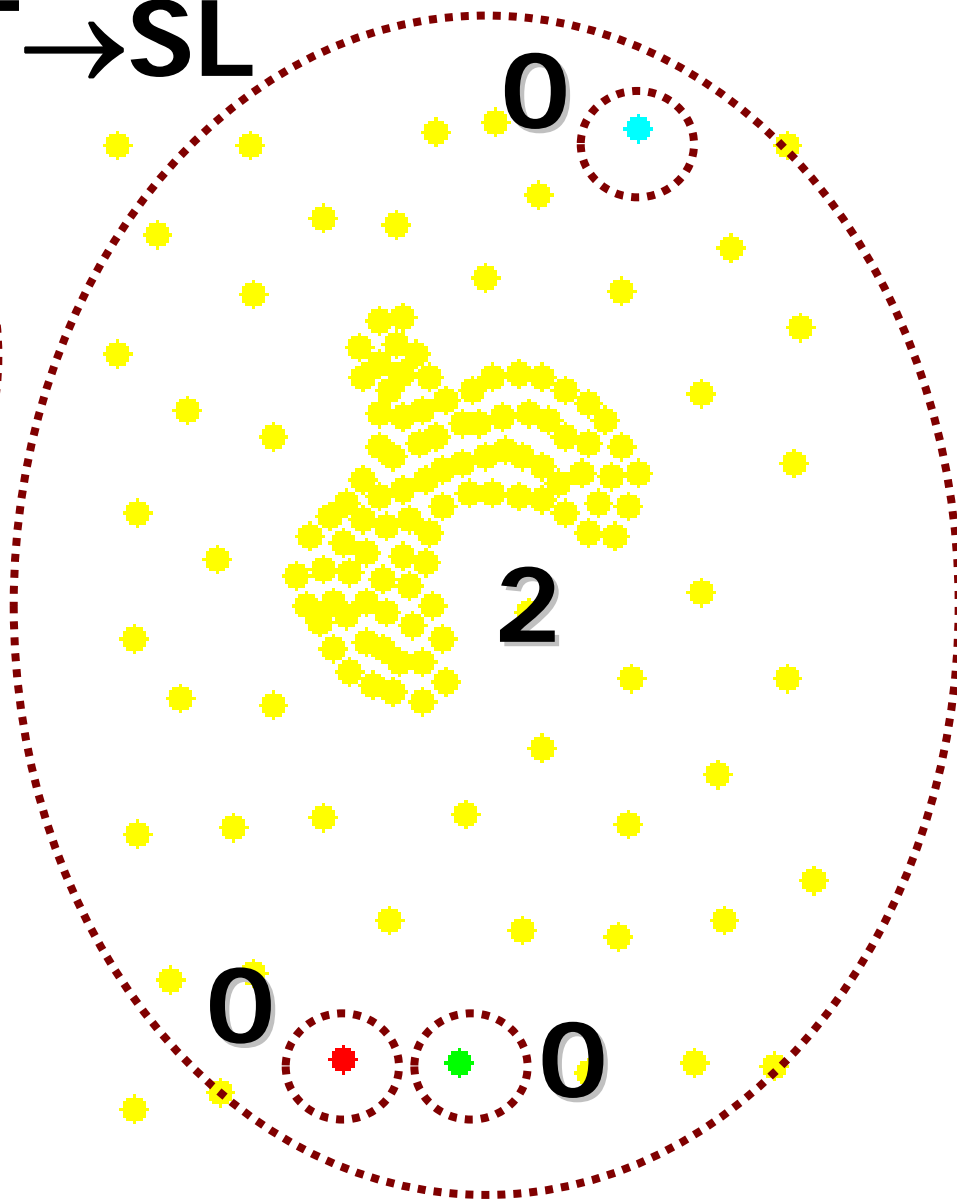
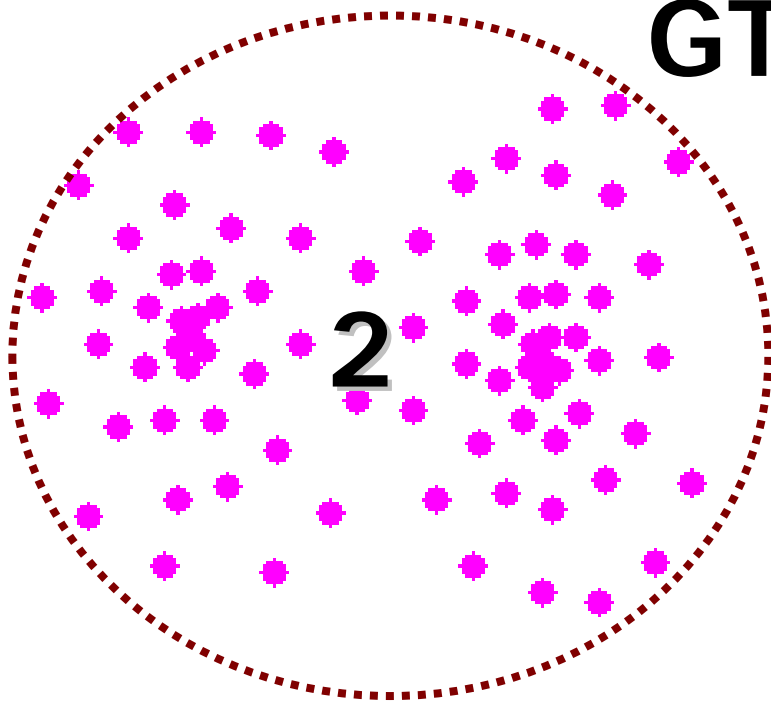


SL → GT



CI = 2

GT → SL



CI = 3

Summary of experiments

prototype similarity

	RI	ARI	MI	NMI	NVD	CH	CSI	CI
Birch2								
KM	1.00	0.81	6.26	0.96	0.12	0.24	0.88	18
KM++	1.00	0.95	6.54	0.99	0.03	0.06	0.97	4
RS	1.00	1.00	6.64	1.00	0.00	0.00	1.00	0
GA	1.00	1.00	6.64	1.00	0.00	0.00	1.00	0
S1								
KM	0.98	0.82	3.57	0.93	0.09	0.17	0.83	2
KM++	1.00	1.00	3.90	0.98	0.00	0.00	1.00	0
RS	1.00	1.00	3.90	0.98	0.00	0.00	1.00	0
GA	1.00	1.00	3.90	0.98	0.00	0.00	1.00	0
S2								
KM	0.97	0.80	3.46	0.90	0.11	0.18	0.82	2
KM++	1.00	0.99	3.87	0.99	0.00	0.00	1.00	0
RS	1.00	0.99	3.87	0.99	0.00	0.00	1.00	0
GA	1.00	0.99	3.87	0.99	0.00	0.00	1.00	0

Summary of experiments

partition similarity

	RI	ARI	MI	NMI	NVD	CH	CSI	CI
Unbalanced								
KM	0.92	0.79	1.85	0.81	0.14	0.29	0.86	4
KM++	1.00	1.00	2.03	1.00	0.00	0.00	1.00	0
RS	1.00	1.00	2.03	1.00	0.00	0.00	1.00	0
GA	1.00	1.00	2.03	1.00	0.00	0.00	1.00	0
SL	1.00	0.99	1.91	0.97	0.02	0.05	0.98	3
DBSCAN	1.00	1.00	2.02	0.99	0.00	0.00	1.00	0
SAM	0.93	0.81	1.85	0.82	0.12	0.25	0.88	4
Aggregate								
KM	0.91	0.71	2.16	0.84	0.14	0.24	0.86	2
SL	0.93	0.80	1.96	0.88	0.09	0.18	0.91	2
DBSCAN	0.99	0.98	2.41	0.98	0.01	0.01	0.99	0
SAM	1.00	1.00	2.45	1.00	0.00	0.00	1.00	0
Compound								
KM	0.84	0.54	1.71	0.72	0.25	0.34	0.75	2
SL	0.89	0.74	1.54	0.80	0.13	0.26	0.87	3
DBSCAN	0.95	0.88	1.90	0.87	0.10	0.12	0.90	2
SAM	0.83	0.53	1.78	0.76	0.19	0.34	0.81	2

Conclusions

- Simple cluster level measure
- Generalized to GMM and arbitrary-shaped data
- Value has clear interpretation:

