UNIVERSITY OF
EASTERN FINLAND

K-means properties

Pasi Fränti

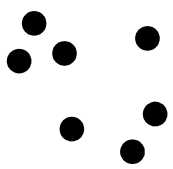
17.4.2017

K-means properties on six clustering benchmark datasets
Pasi Fränti and Sami Sieranoja
Algorithms, 2017.

Goal of k-means

Input N points:

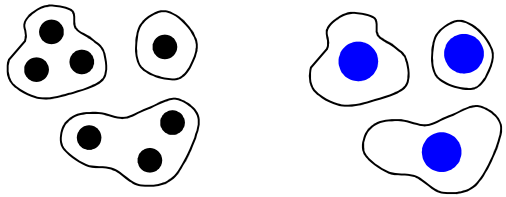
$X=\{x_1, x_2, \ldots, x_N\}$



Output partition and k centroids:

$P=\{p_1, p_2, \ldots, p_k\}$

$C=\{c_1, c_2, \ldots, c_k\}$



Objective function:

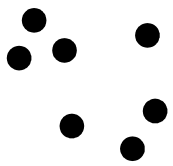
$$SSE = \sum_{i=1}^{N} \left\| x_i - c_j \right\|^2$$

SSE = **s**um-of-**s**quared **e**rrors

Goal of k-means

Input *N* points:

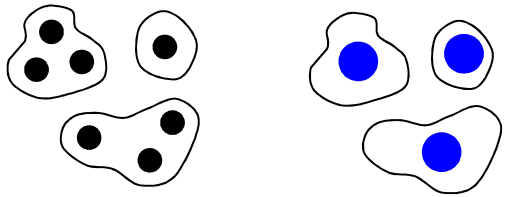
$X=\{x_1, x_2, \ldots, x_N\}$



Output partition and *k* centroids:

$P=\{p_1, p_2, \ldots, p_k\}$

$C=\{c_1, c_2, \ldots, c_k\}$



Objective function:

$$SSE = \sum_{i=1}^{N} \left\| x_i - c_j \right\|^2$$

Assumptions:

- SSE is suitable
- k is known

K-means algorithm

X = Data set
C = Cluster centroids
P = Partition

K-Means(X, C) $\rightarrow$ (C, P)

REPEAT

$C_{prev} \leftarrow C$;

FOR i=1 TO N DO
$p_i \leftarrow$ FindNearest(x_i, C);

Assignment step

FOR j=1 TO k DO
$c_j \leftarrow$ Average of $x_i \ \forall \ p_i = j$;

Centroid step

UNTIL C = C_{prev}

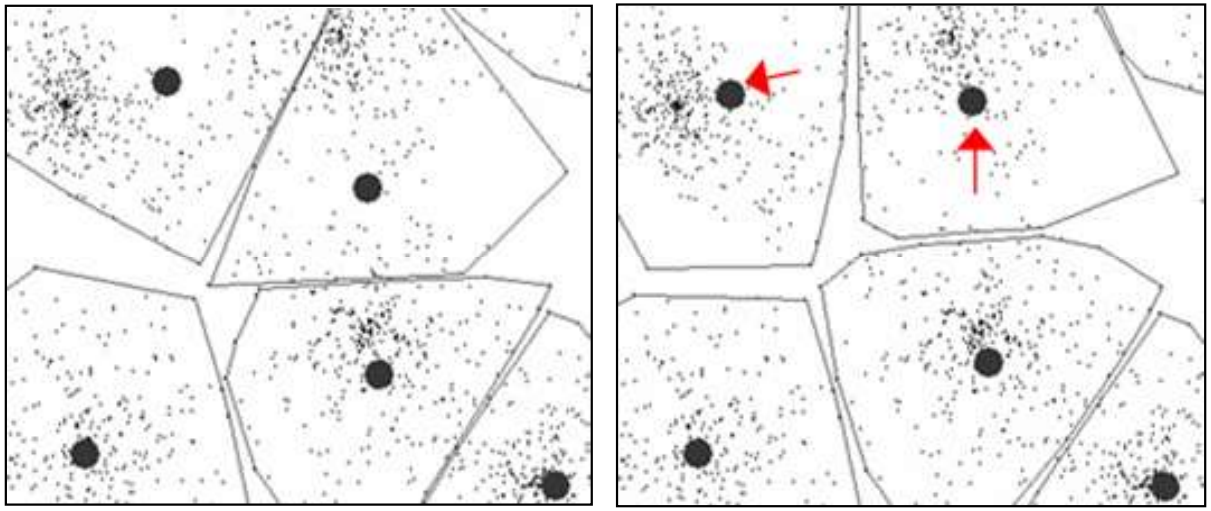
K-means optimization steps

Assignment step:

$$P_i = \arg\min_{1 \le j \le k} \left\| x_i - c_j \right\|^2 \qquad \forall i \in [1, N]$$

Centroid step:

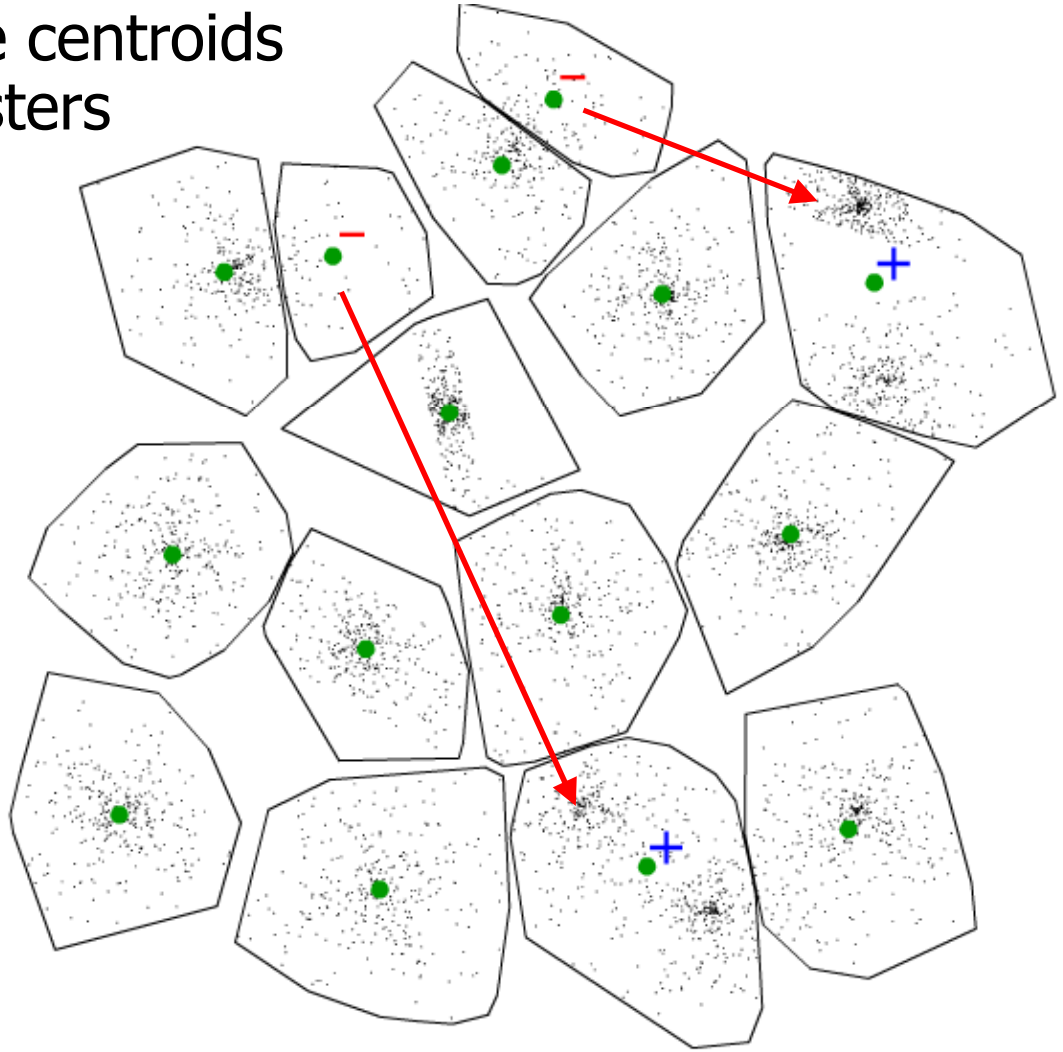
$$c_j = \sum_{P_i = j} x_i \Big/ \sum_{P_i = j} 1 \qquad \forall j \in [1, k]$$



Problems of k-means

Distance of clusters

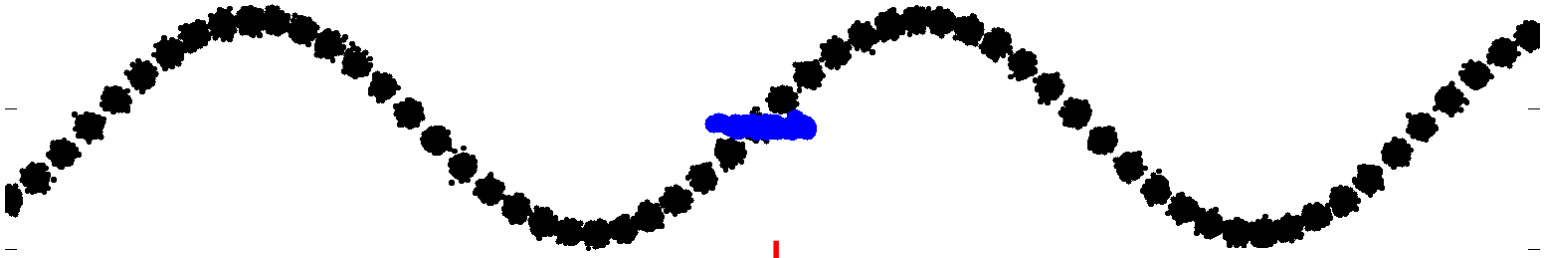
Cannot move centroids between clusters far away



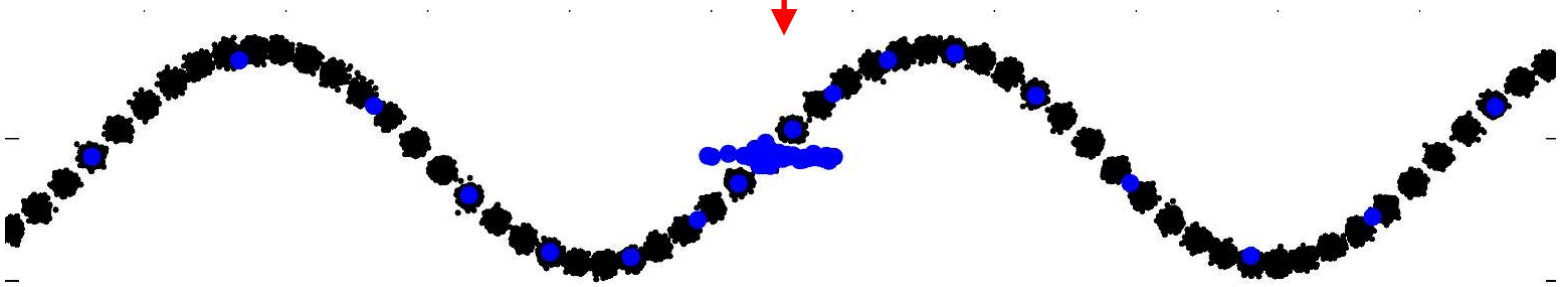
Problems of k-means

Dependency of initial solution

Initial solution:



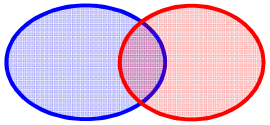
After k-means:



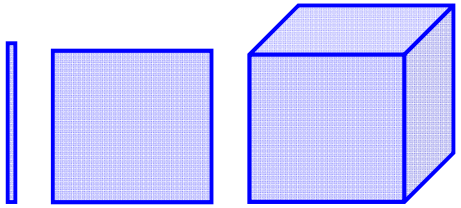
K-means performance

How affected by?

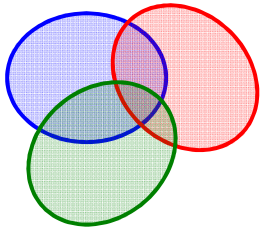
1. Overlap



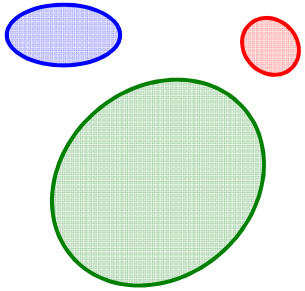
3. Dimensionality



2. Number of clusters



4. Unbalance of cluster sizes

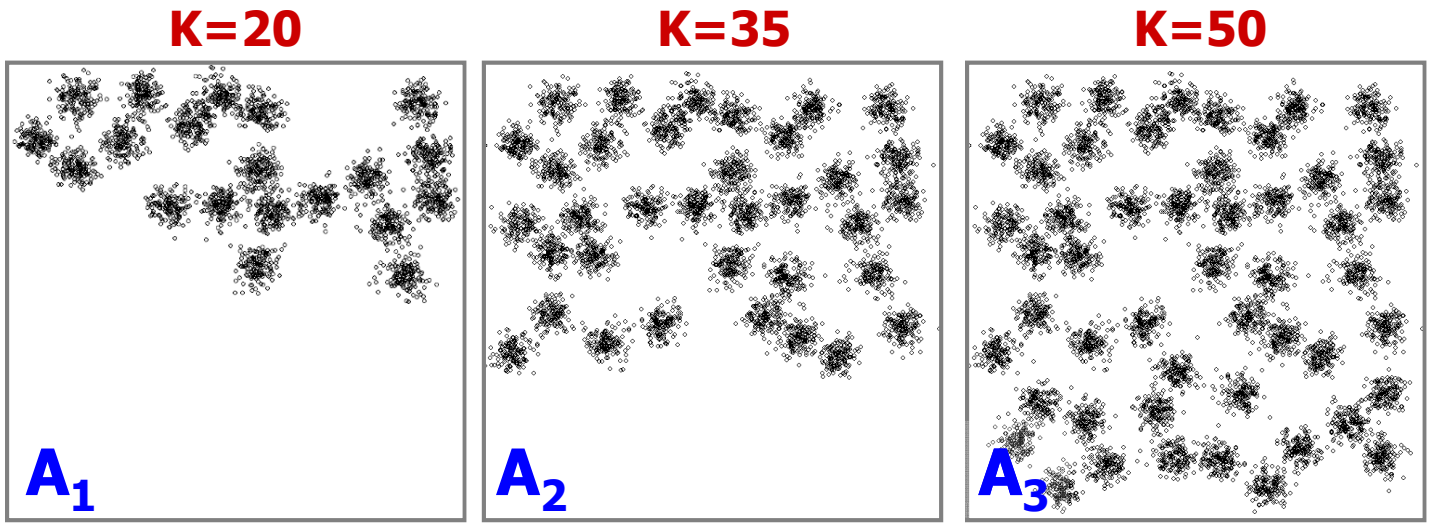


Basic Benchmark

Data sets statistics

Dataset	Varying	Size	Clusters	Per cluster
A	Number of clusters	3000 – 7500	20 - 50	150
S	Overlap	5000	15	333
Dim	Dimensions	1024	16	64
G2	Dimensions + overlap	2048	2	1024
Birch	Structure	100,000	100	1000
Unbalance	Balance	6500	8	100-2000

A sets



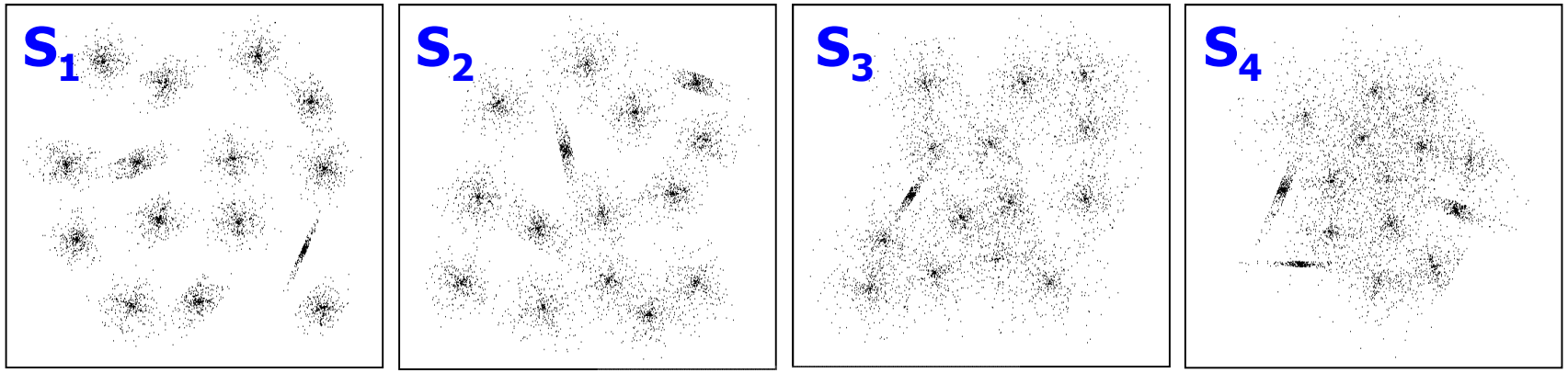
- Spherical clusters
- Number of clusters changing from k=20 to 50
- Subsets of each other: A1 $\subset$ A2 $\subset$ A3.
- Other parameters fixed:

- Cluster size = 150
- Deviation = 1402
- Overlap = 0.30
- Dimensionality = 2

S sets

K=15

Gaussian clusters (few truncated)



Least overlap

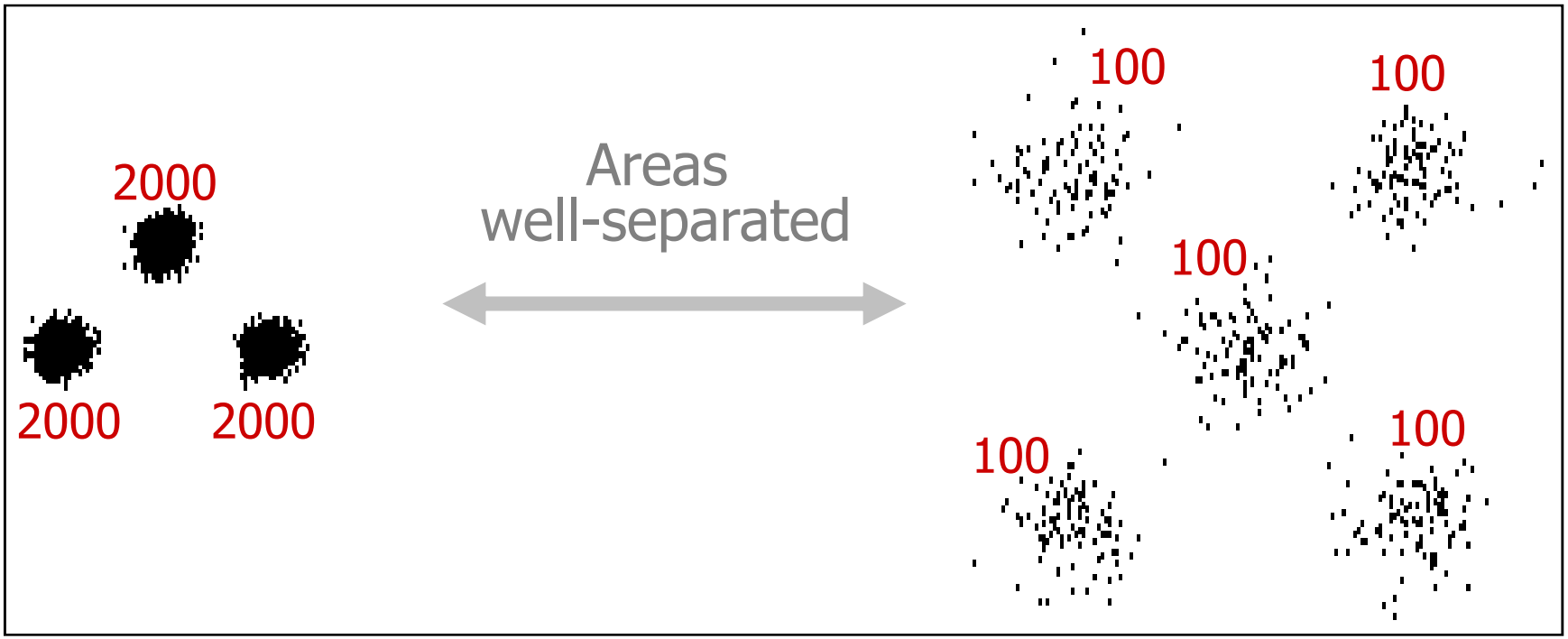
Strong overlap but the clusters still recognizable

Unbalance

K=8

Dense clusters
st.dev=2043

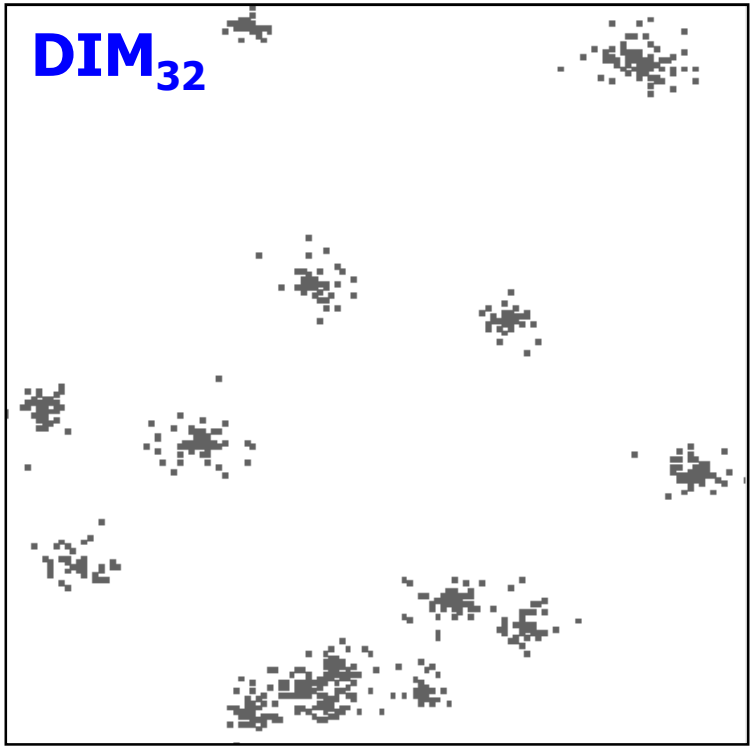
Sparse clusters
st.dev=6637



*Correct clustering can be obtained by minimizing SSE

DIM sets

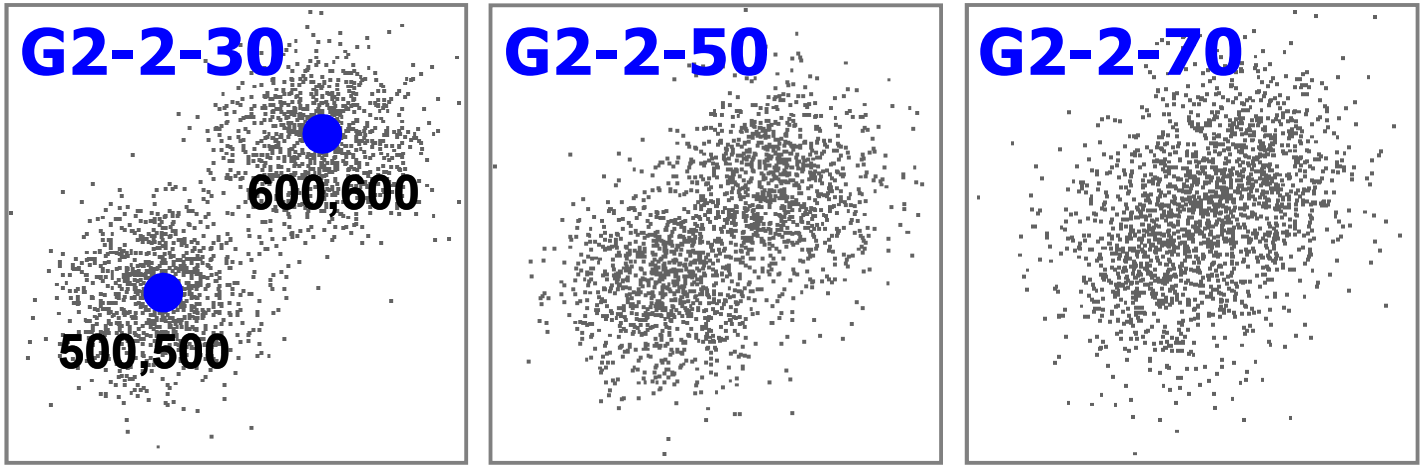
K=16



- Well-separated clusters in high-dimensional spaces
- Dimensions vary: 32, 64, 128, 256, 512, 1024

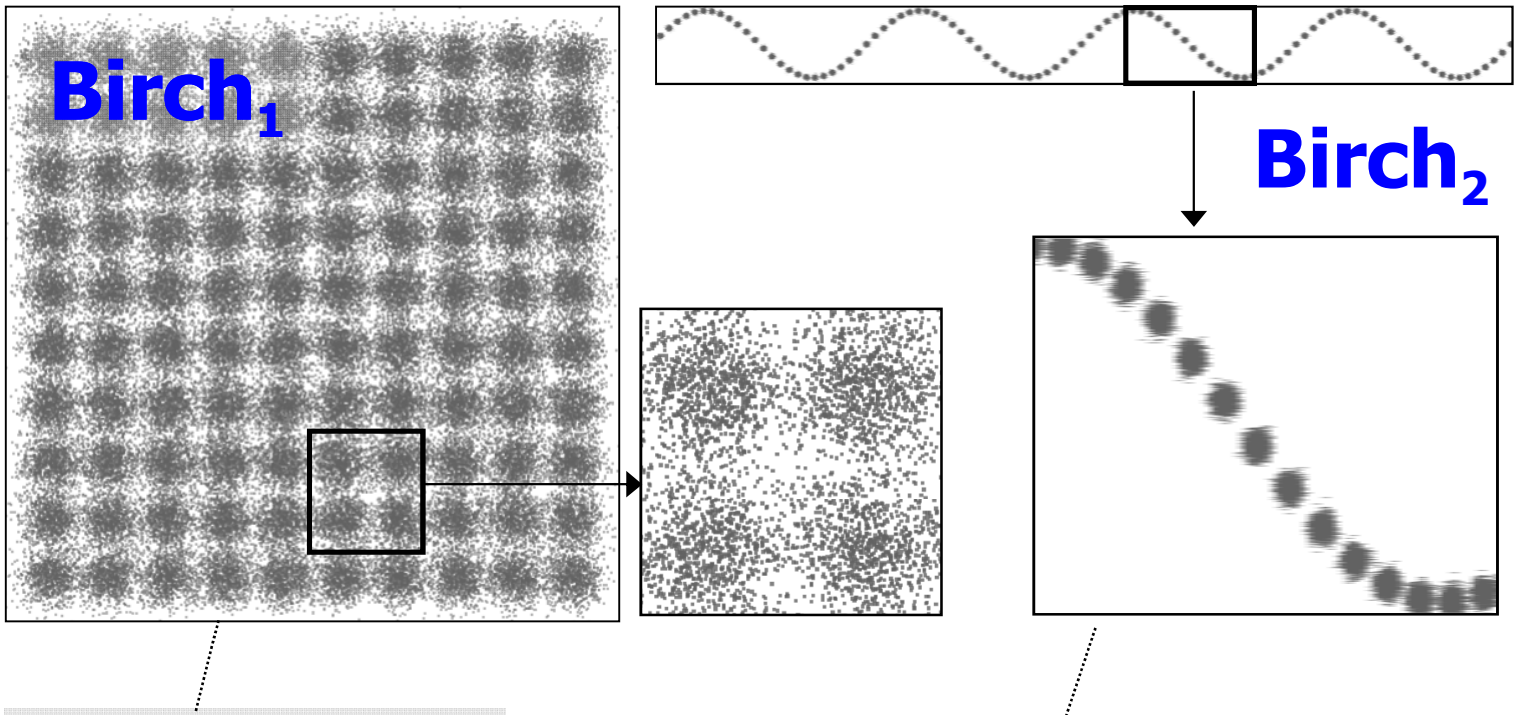
G2 Datasets

K=2



Dataset name:	G2-dim-sd
Centroid 1:	[500,500, ...]
Centroid 2:	[600,600, ...]
Dimensions:	1,2,4,8,16, ... 1024
St.dev.	10,20,30,40 ... 100

Birch

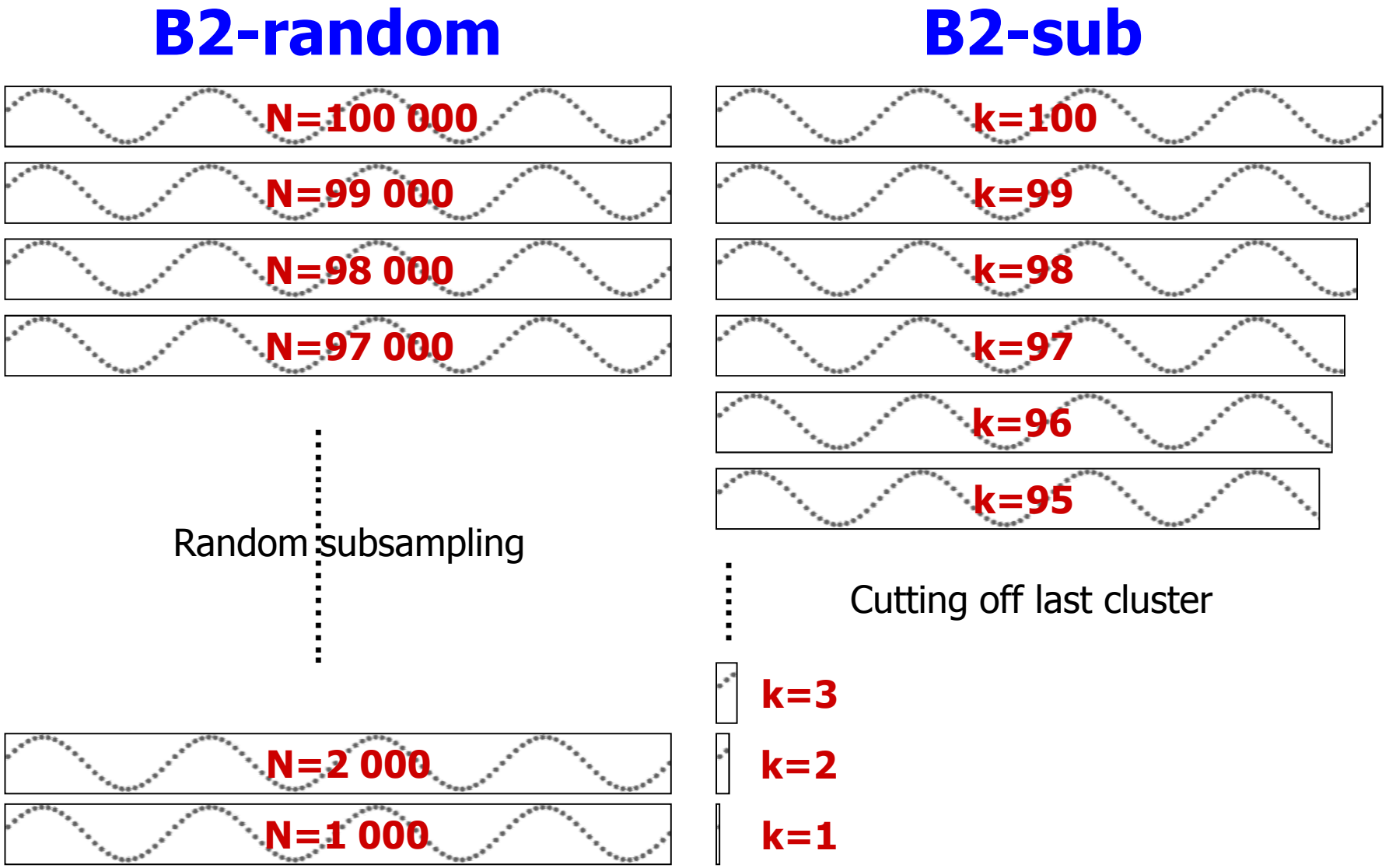


Regular 10x10 grid
Constant variance

offset	=	43659
amplitude	=	-37819
phaseshift	=	20.8388
frequency	=	0.000004205

$$y(x) = \text{amplitude} * \sin(2*\pi*\text{frequency}*x + \text{phaseshift}) + \text{offset}$$

Birch$_2$ subsets



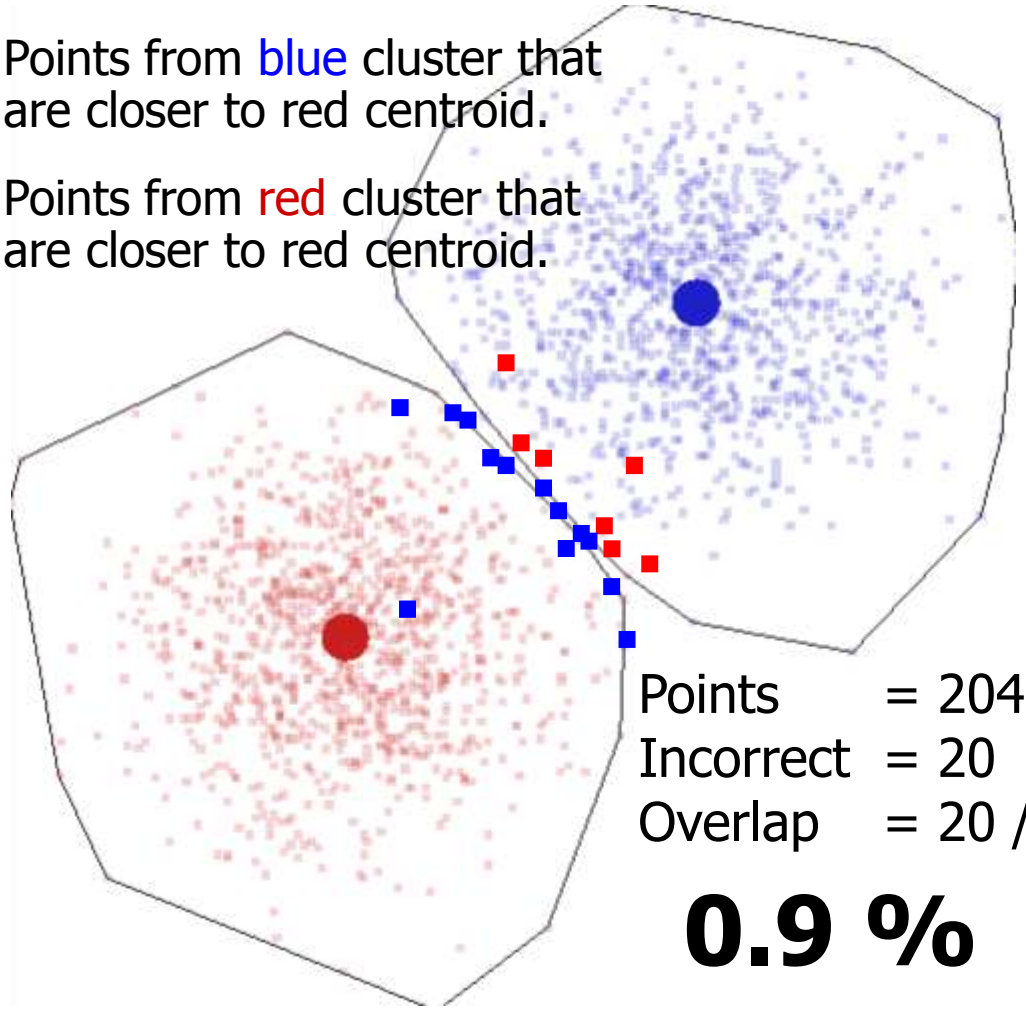
Properties

Measured properties

- Overlap
- Contrast
- Intrinsic dimensionality
- H-index
- Distance profiles

Overlap by misclassification probability

- Points from blue cluster that are closer to red centroid.
- Points from red cluster that are closer to red centroid.



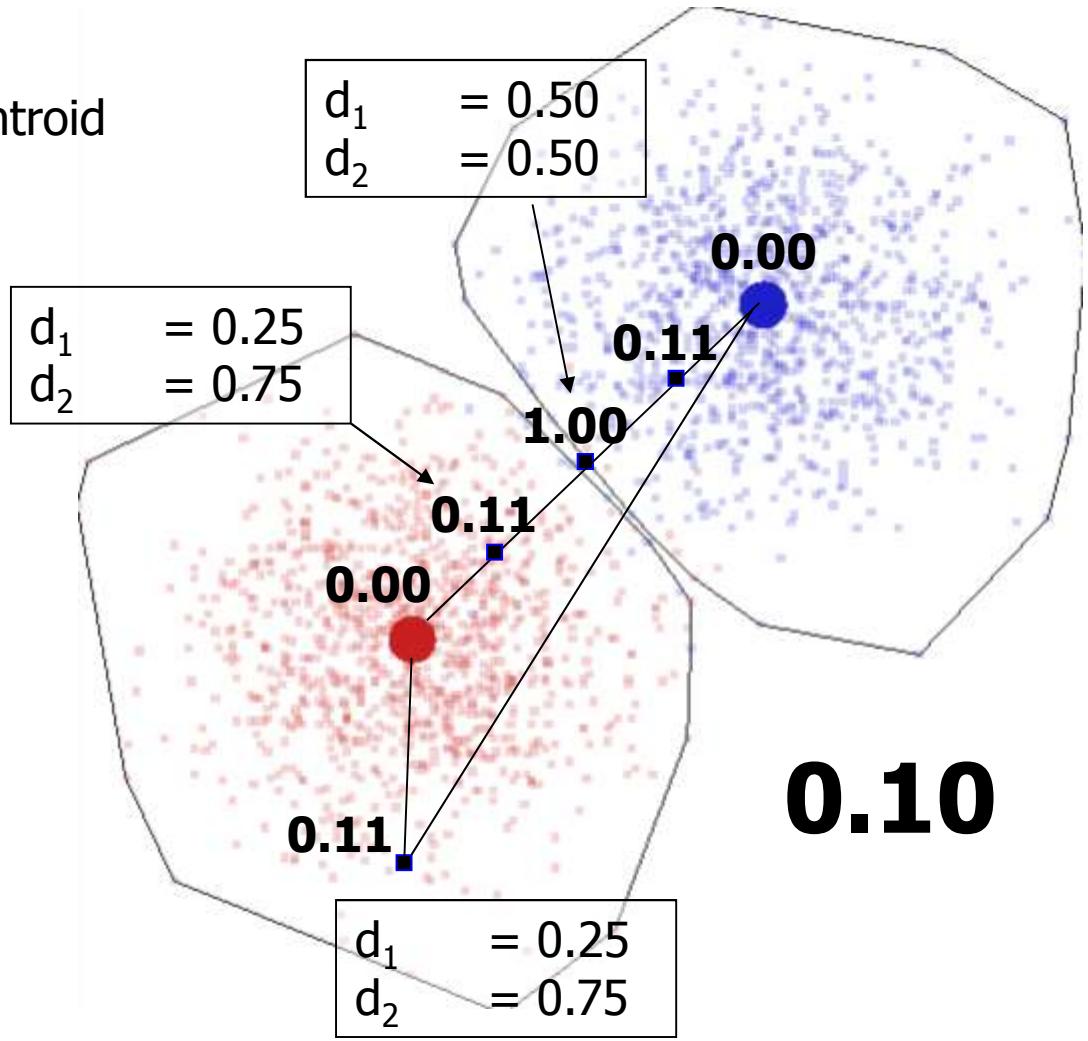
Points = 2048
Incorrect = 20
Overlap = 20 / 2048

0.9 %

Overlap by separation

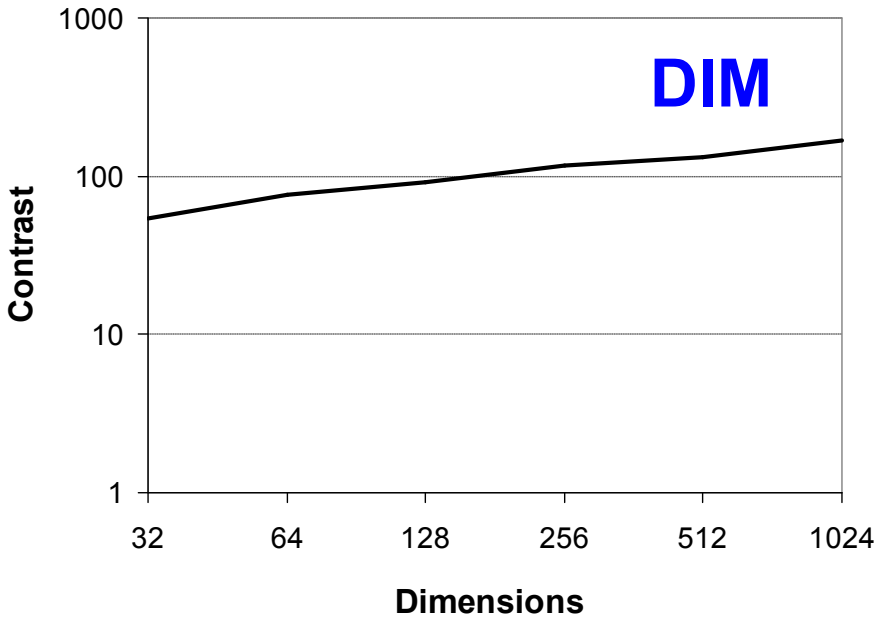
$$overlap = \frac{1}{N} \cdot \sum \left(\frac{d_1}{d_2} \right)^2$$

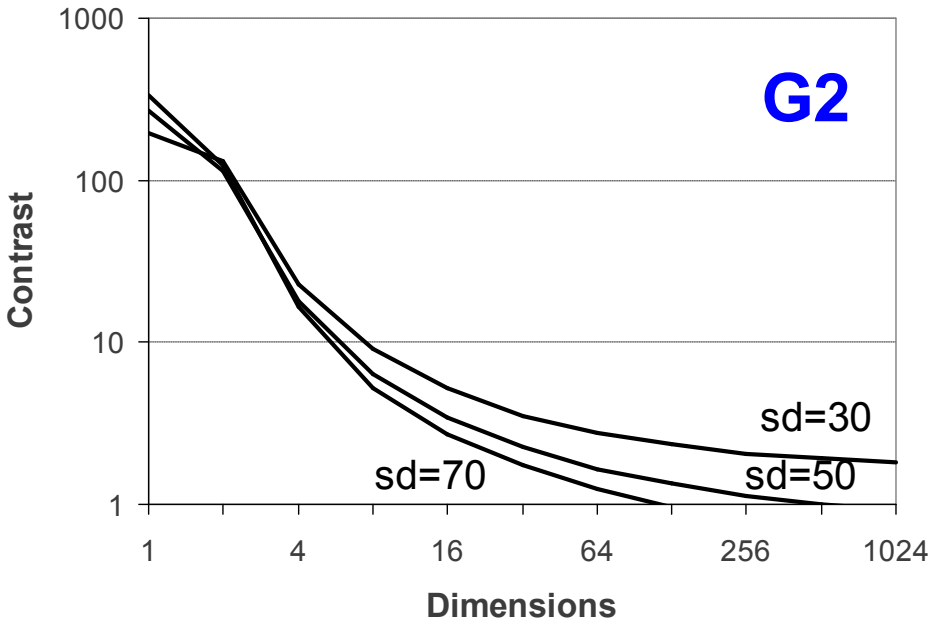
d_1 = distance to nearest centroid
d_2 = distance to 2nd nearest



Contrast

$$contrast = median\left(\frac{(d_{\max} - d_{\min})}{d_{\min}}\right)$$



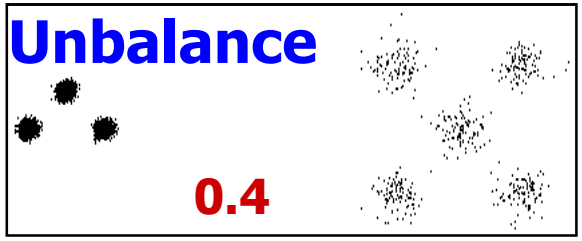


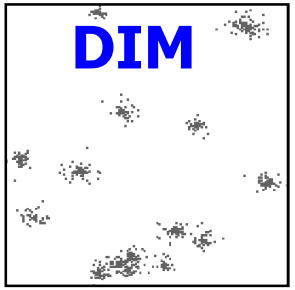
Intrinsic dimensionality

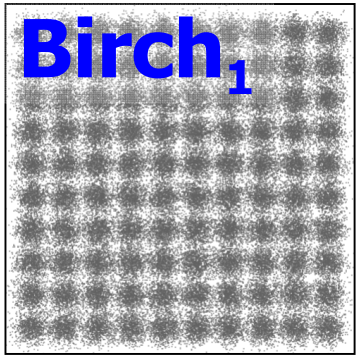
$$ID = \frac{\hat{d}^2}{2\sigma^2}$$

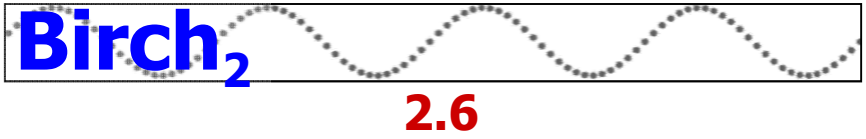
$\hat{d}$ Average of distances

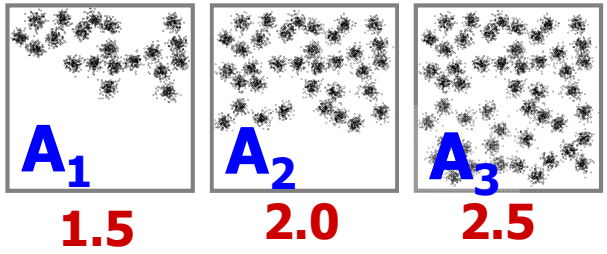
σ^2 Variance of distances

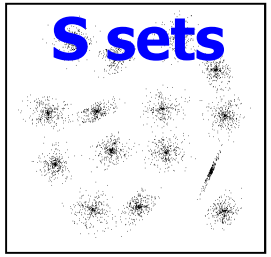






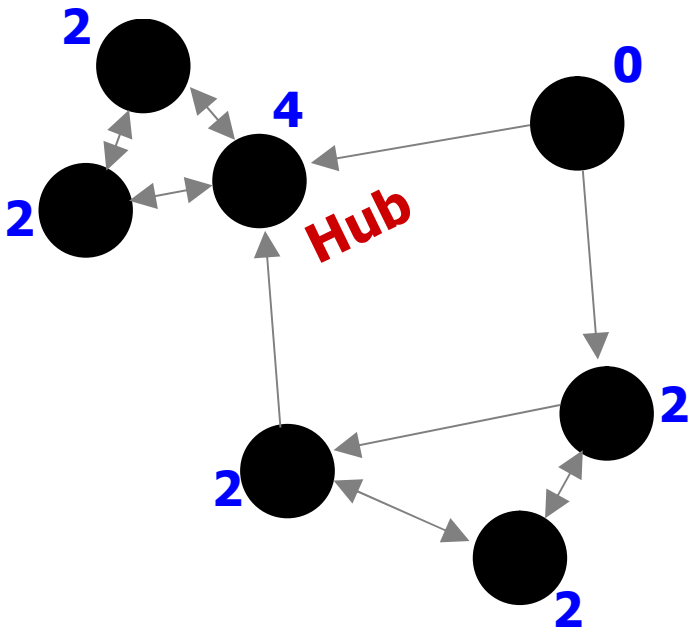






H-index

Hubness values:

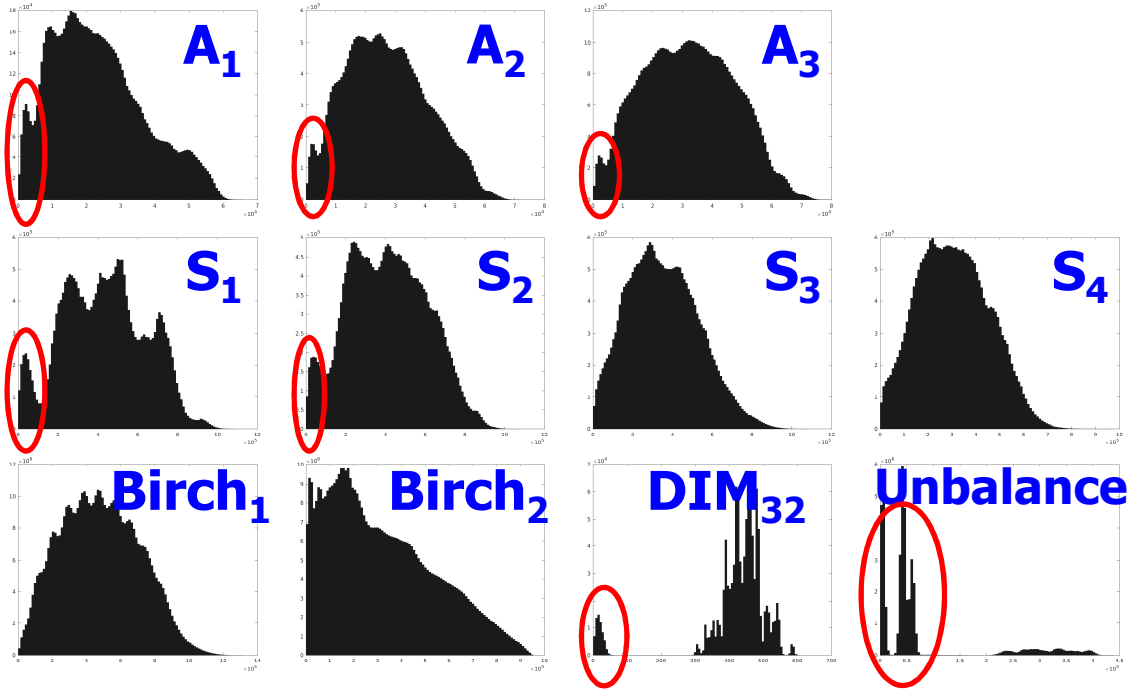


Rank:	1	**2**	3	4	5	6	7
Hub:	4	2	2	2	2	2	0

Distance profiles

Data that contains clusters tends to have two peaks:

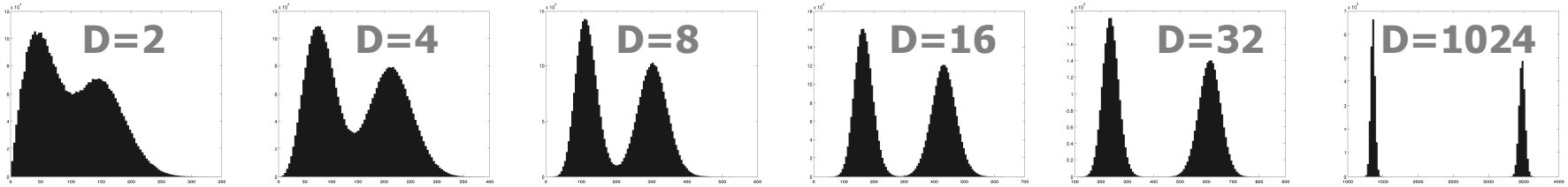
- **Local distances**: distances inside the clusters
- **Global distances**: distances across different clusters



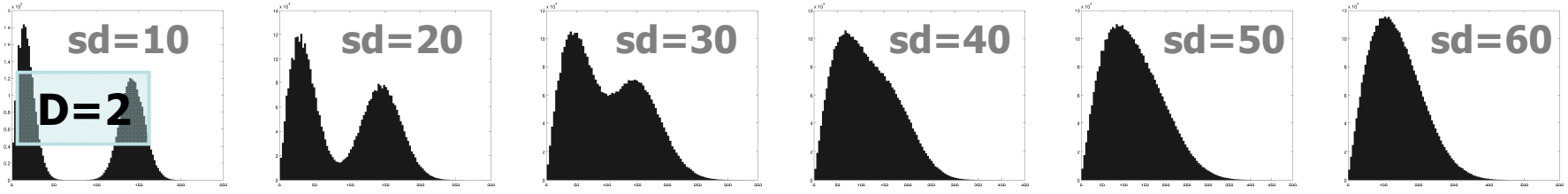
Distance profiles

G2 datasets

G2: dimension increases



G2: overlap increases



Summary of the properties

Dataset	Overlap	Contrast	Intrinsic dim.	H-index
A1	0.30	227	1.5	2
A2	0.30	261	2.0	3
A3	0.30	294	2.5	3
S1	0.20	320	2.2	2
S2	0.27	257	2.2	3
S3	0.37	210	2.0	3
S4	0.39	205	2.2	2
Dim32-1024	0.01 – 0.04	54 – 167	6.6 – 7.5	7-11
G2	0.00 – 0.15*	0.37 – 494	0.7 – 43.4	2-17
Birch1	0.42	799	8.3	3
Birch2	0.19	8308	2.6	3
Unbalance	0.10	2983	0.4	3

G2 overlap

Overlap decreases →

Overlap increases ↓

σ \ dim	2	4	8	16	32	64	128	256	512	1024
10	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
20	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
30	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%
40	4%	1%	0%	0%	0%	0%	0%	0%	0%	0%
50	8%	2%	0%	0%	0%	0%	0%	0%	0%	0%
60	12%	4%	1%	0%	0%	0%	0%	0%	0%	0%
70	15%	8%	2%	0%	0%	0%	0%	0%	0%	0%
80	19%	9%	4%	1%	0%	0%	0%	0%	0%	0%
90	22%	12%	6%	2%	0%	0%	0%	0%	0%	0%
100	25%	15%	7%	2%	0.1%	0%	0%	0%	0%	0%

G2 contrast

Contrast decreases →

Contrast decreases ↓

σ \ dim	2	4	8	16	32	64	128	256	512	1024
10	172	53.8	22.3	14.0	10.7	8.9	7.9	7.3	6.9	6.7
20	161	31.6	12.4	7.4	5.2	4.3	3.7	3.4	3.1	3.0
30	131	22.8	9.1	5.2	3.5	2.7	2.3	2.1	1.9	1.8
40	123	20.0	7.5	4.0	2.7	2.1	1.7	1.5	1.3	1.2
50	115	18.0	6.3	3.4	2.2	1.7	1.3	1.1	1.0	0.9
60	108	16.2	5.5	3.0	1.9	1.4	1.1	0.9	0.8	0.7
70	123	16.4	5.2	2.7	1.8	1.2	1.0	0.8	0.7	0.6
80	122	16.0	5.2	2.6	1.6	1.1	0.9	0.7	0.6	0.5
90	116	15.0	4.9	2.5	1.5	1.0	0.8	0.6	0.5	0.4
100	110	15.3	4.8	2.4	1.4	1.0	0.7	0.6	0.4	0.4

G2 Intrinsic dimensionality

ID increases (if overlap)

ID increases

σ \ dim	2	4	8	16	32	64	128	256	512	1024
10	0.8	0.8	0.8	0.9	0.9	0.9	0.9	0.9	0.9	0.9
20	1.1	1.3	1.4	1.5	1.5	1.5	1.5	1.5	1.5	1.5
30	1.4	1.9	2.2	2.4	2.5	2.5	2.6	2.6	2.6	2.6
40	1.6	2.4	3.1	3.6	3.9	4.1	4.2	4.2	4.2	4.3
50	1.7	2.8	4.1	5.0	5.9	6.3	6.6	6.7	6.9	6.8
60	1.8	3.1	5.0	6.6	8.1	9.2	10.0	10.3	10.5	10.5
70	1.8	3.3	5.6	8.5	11.0	13.1	14.4	15.1	15.7	15.9
80	1.8	3.5	6.1	9.9	13.9	17.3	19.6	21.4	22.1	22.3
90	1.8	3.6	6.4	11.2	16.6	22.3	26.9	29.1	30.5	31.5
100	1.8	3.6	6.7	12.2	19.1	26.9	34.0	39.1	41.2	43.4

Most significant

G2 H-index

H-index increases

σ\dim	2	4	8	16	32	64	128	256	512	1024
10	4	3	4	5	9	11	14	16	15	14
20	3	3	4	6	9	12	14	15	15	16
30	2	3	3	6	10	11	14	15	14	15
40	3	3	4	6	10	11	15	13	17	16
50	2	3	4	7	9	12	13	14	15	17
60	2	3	4	6	9	13	14	14	16	16
70	2	3	4	6	10	11	12	13	14	16
80	2	3	4	6	9	12	15	14	16	17
90	2	3	4	6	9	12	15	14	18	14
100	2	3	4	5	9	11	15	15	15	15

No change

Most significant

Evaluation

Internal measures

Sum of squared distances (SSE)

$$SSE = \sum_{i=1}^{N} \left\| x_i - c_j \right\|^2$$

Normalized mean square error (nMSE)

$$nMSE = \frac{SSE}{N \cdot D}$$

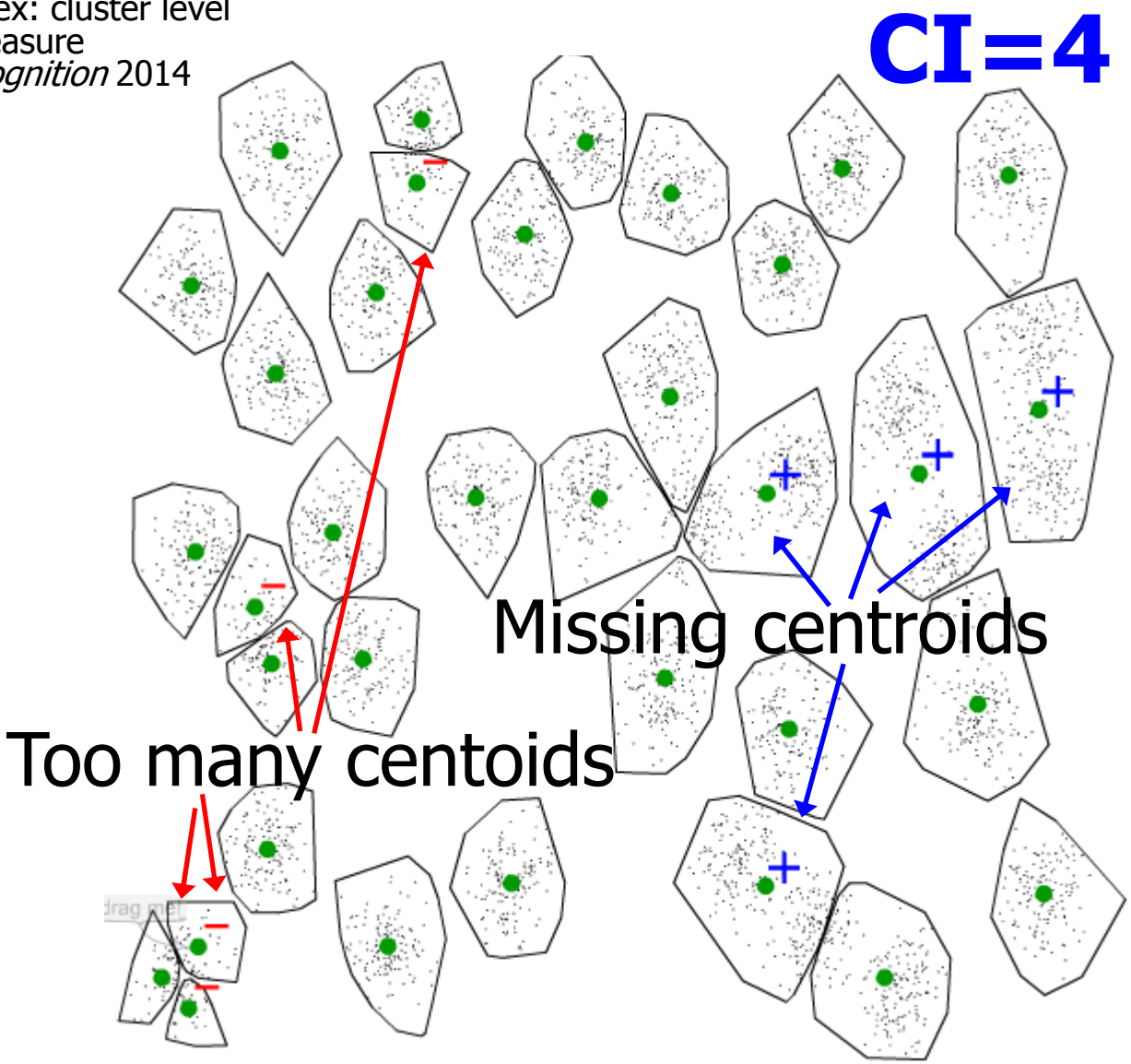
Approximation ratio (ε)

$$\varepsilon = \frac{\left(SSE - SSE_{opt} \right)}{SSE_{opt}}$$

External measures

Centroid index

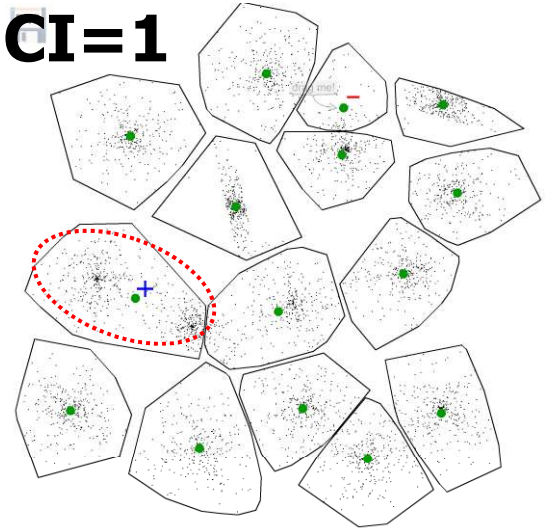
P. Fränti, M. Rezaei, Q. Zhao
Centroid index: cluster level similarity measure
Pattern Recognition 2014

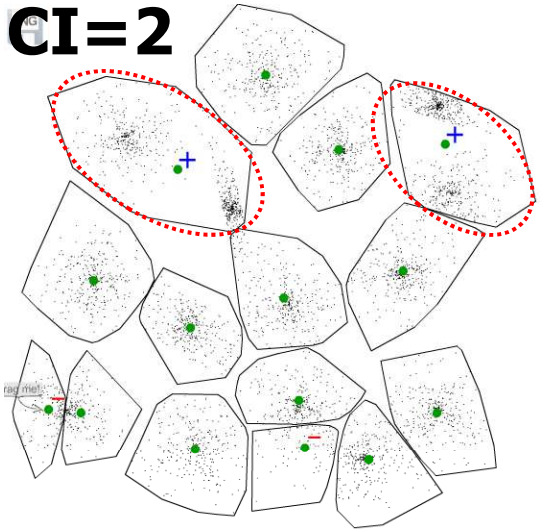


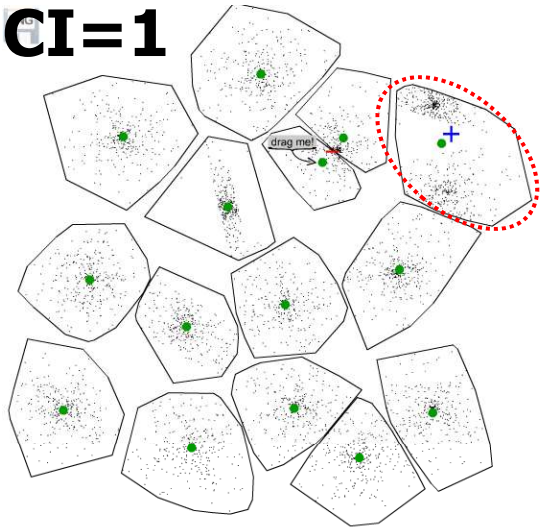
External measures

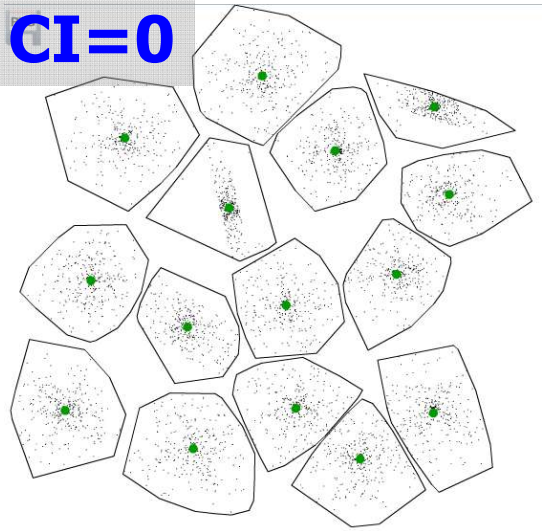
Success rate

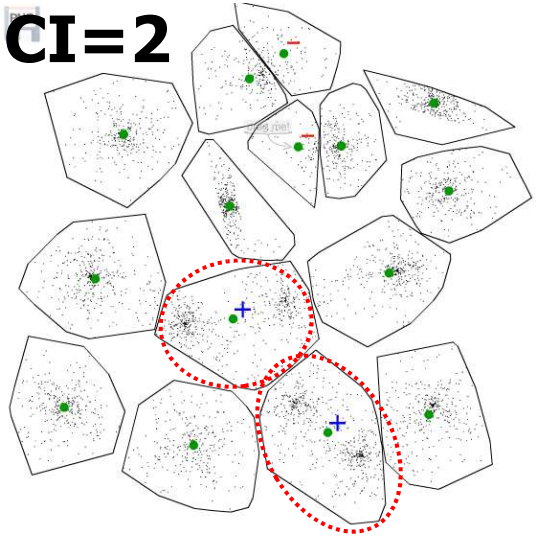
17%

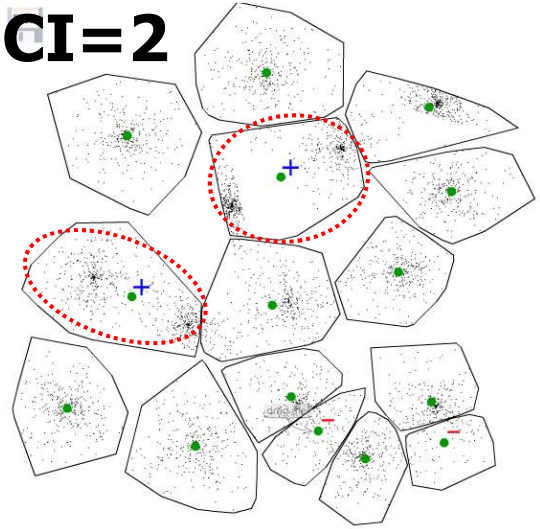












Results

Summary of results

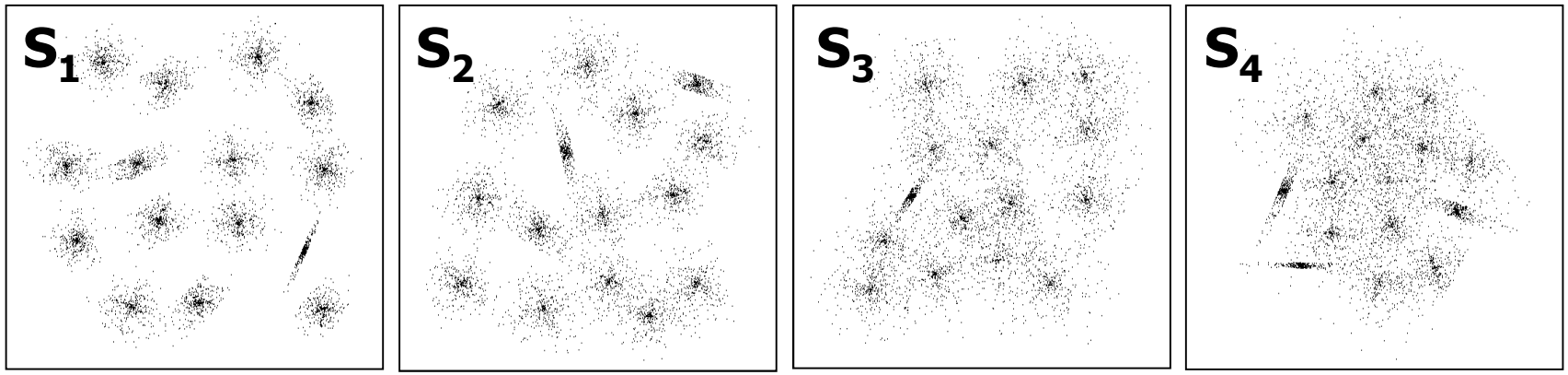
Dataset	Success	Clustering quality			Objective function		
		CI	rel-CI	ARI	SSE	nMSE	ε
A1	1%	2.5	13%	0.82	1.98	3.31	0.64
A2	0%	4.5	13%	0.82	3.39	3.23	0.67
A3	0%	6.6	13%	0.82	4.90	3.27	0.69
S1	3%	1.8	12%	0.85	18.84	18.84	1.11
S2	11%	1.4	9%	0.86	19.79	19.79	0.48
S3	12%	1.3	9%	0.84	19.51	19.51	0.14
S4	26%	0.9	6%	0.84	17.00	17.00	0.07
Unbalance	0%	3.9	49%	0.64	2.10	1.61	8.81
Birch1	0%	6.6	7%	0.85	10.95	5.47	0.18
Birch2	0%	16.6	17%	0.81	15.75	7.87	2.45
Dim32	0%	3.6	23%	0.76	16.5	504	68.34
Average:	**5%**	4.5	16%	0.81	---	---	7.60

Dependency on overlap

S datasets

Success rates and CI-values:

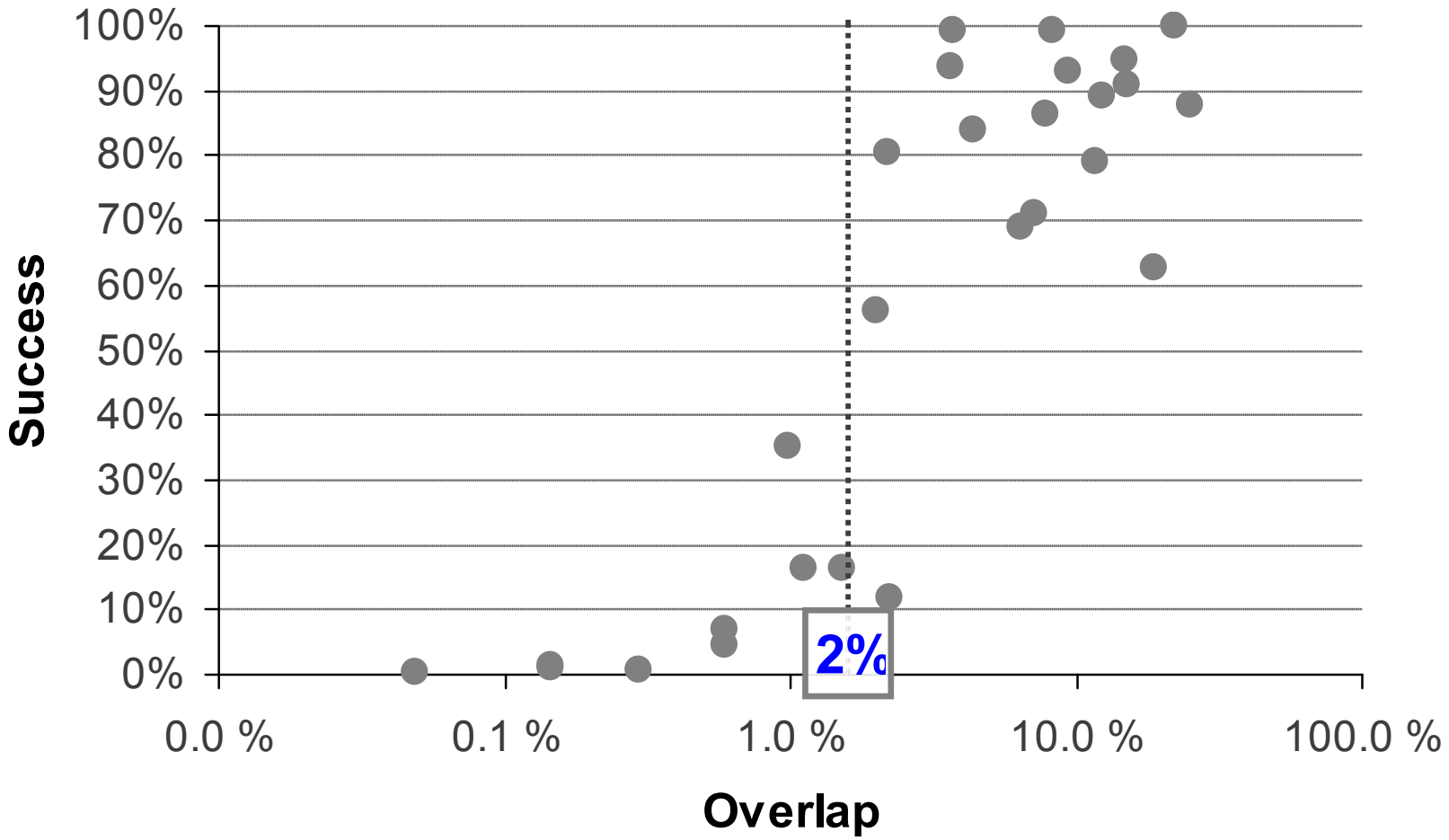
overlap increases



3%	**11%**	**12%**	**26%**
1.8	**1.4**	**1.3**	**0.9**

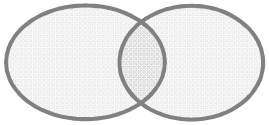
Dependency on overlap

G2 datasets



Main observation

1. Overlap

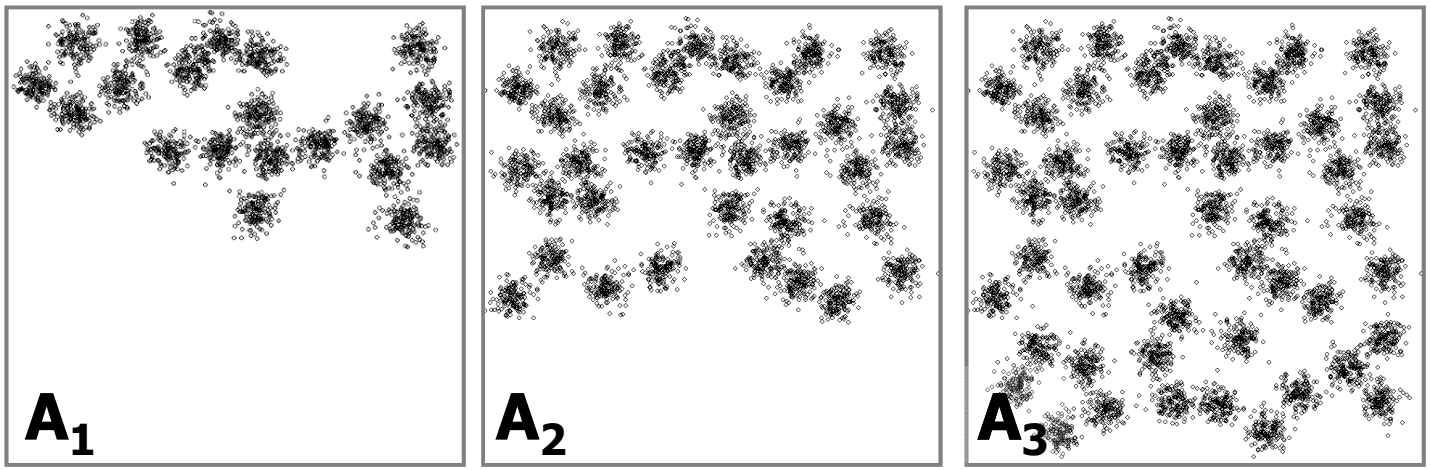


Overlap is good!

Dependency on clusters (k)

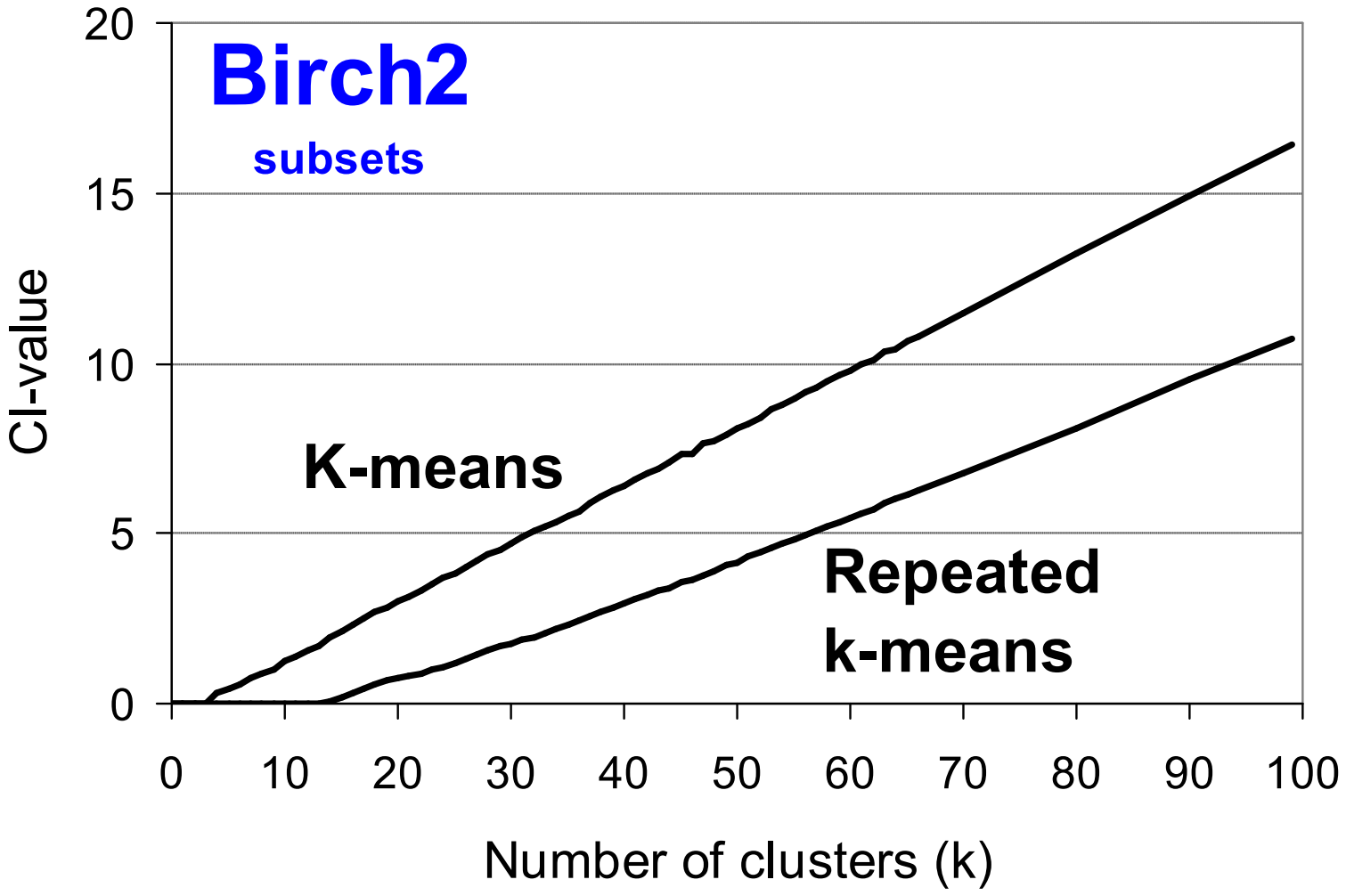
A datasets

Clusters increases

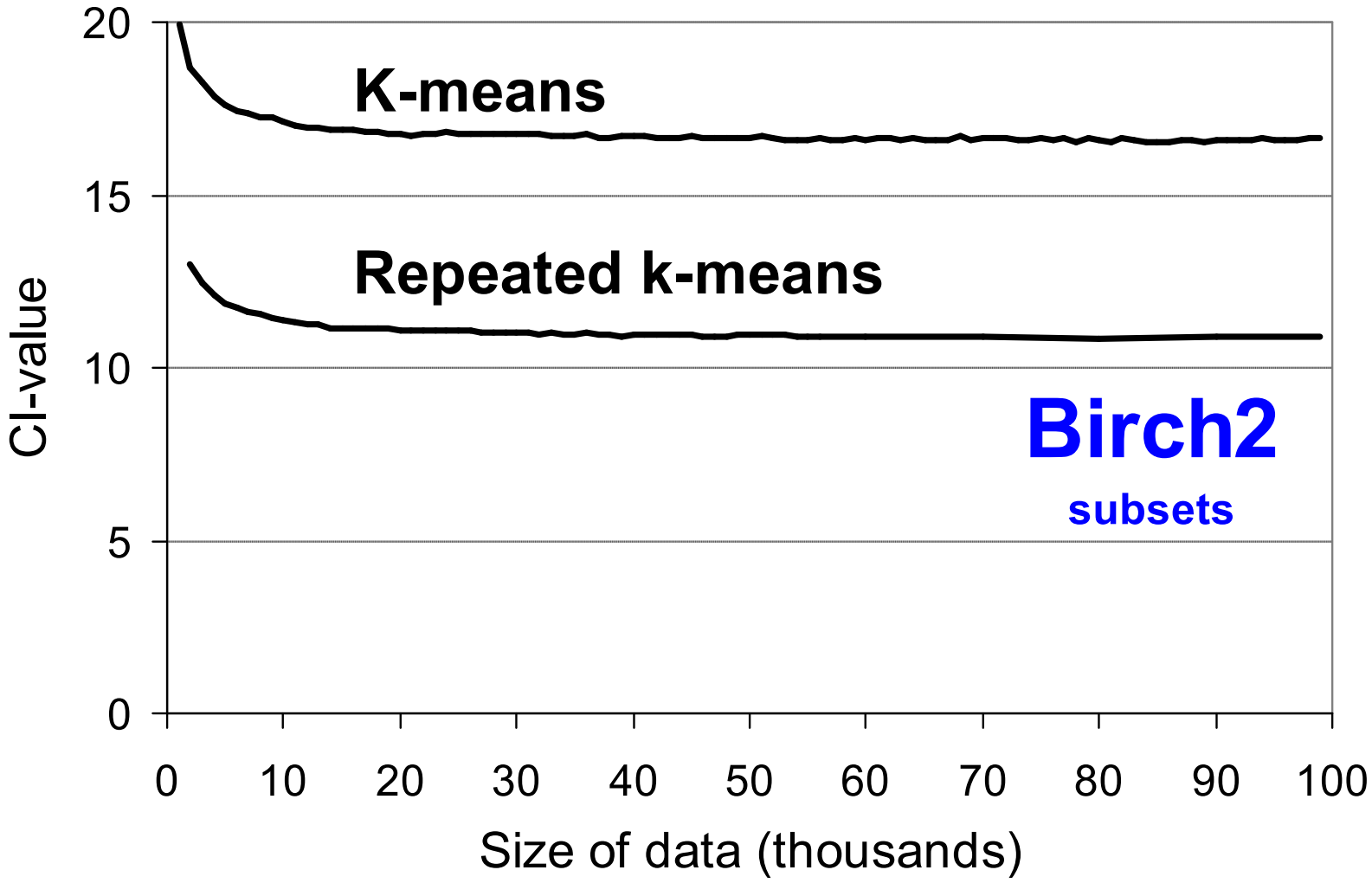


Success:	1%	0%	0%
CI:	**2.5**	**4.5**	**6.6**
Relative CI:	13%	13%	13%

Dependency on clusters (k)

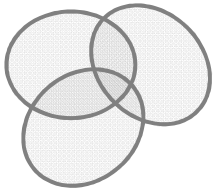


Dependency on data size (N)



Main observation

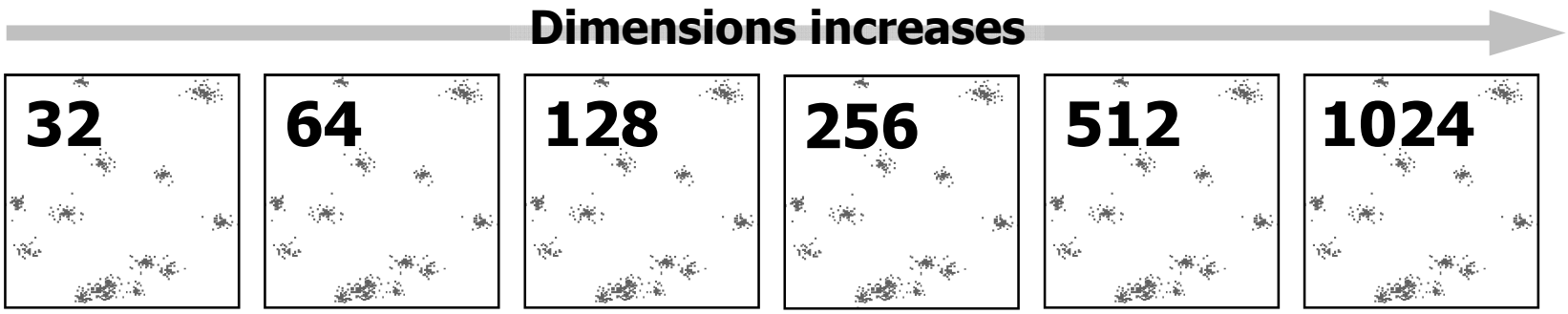
2. Number of clusters



Linear increase with k!

Dependency on dimensions

DIM datasets



CI:	3.6	3.5	3.8	3.8	3.9	3.7

Success rate: **0%**

Dependency on dimensions

G2 datasets

Success degrades →

Success improves ↓

σ \ dim	2	4	8	16	32	64	128	256	512	1024
10	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
20	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
30	35%	1%	0%	0%	0%	0%	0%	0%	0%	0%
40	99%	7%	0%	0%	0%	0%	0%	0%	0%	0%
50	99%	80%	1%	0%	0%	0%	0%	0%	0%	0%
60	79%	84%	16%	1%	0%	0%	0%	0%	0%	0%
70	95%	86%	56%	1%	0%	0%	0%	0%	0%	0%
80	63%	93%	94%	5%	0%	0%	0%	0%	0%	0%
90	100%	89%	69%	16%	0%	0%	0%	0%	0%	0%
100	88%	91%	71%	12%	1%	0%	0%	0%	0%	0%

Lack of overlap is the cause!

σ\dim	2	4	8	16	32	64	128	256	512	1024
10										
20										
30	1%									
40	4%	1%								
50	8%	2%								
60	12%	4%	1%							
70	15%	8%	2%							
80	19%	9%	4%	1%						
90	22%	12%	6%	2%						
100	25%	15%	7%	2%	0.1%					

Overlap

σ\dim	2	4	8	16	32	64	128	256	512	1024
10										
20										
30	35%	1%								
40	99%	7%								
50	99%	80%	1%							
60	79%	84%	16%	1%						
70	95%	86%	56%	1%						
80	63%	93%	94%	5%						
90	100%	89%	69%	16%						
100	88%	91%	71%	12%	1%					

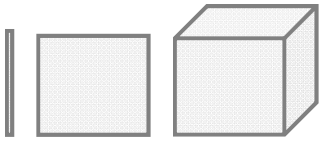
Success

Correlation:

0.88

Main observation

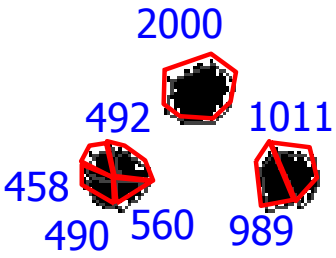
3. Dimensionality



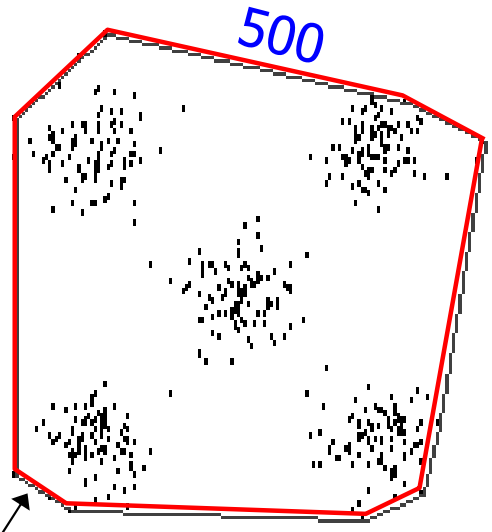
No direct effect!

Effect of unbalance

DIM datasets



K-means tend to put
too many clusters here …



… and too few here

Success:

0%

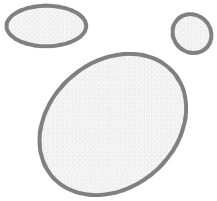
Average CI:

3.9

→ Problem originates from the random initialization.

Main observation

4. Unbalance of cluster sizes



Unbalance is bad!

Improving k-means

Better initialization technique

Simple initializations:

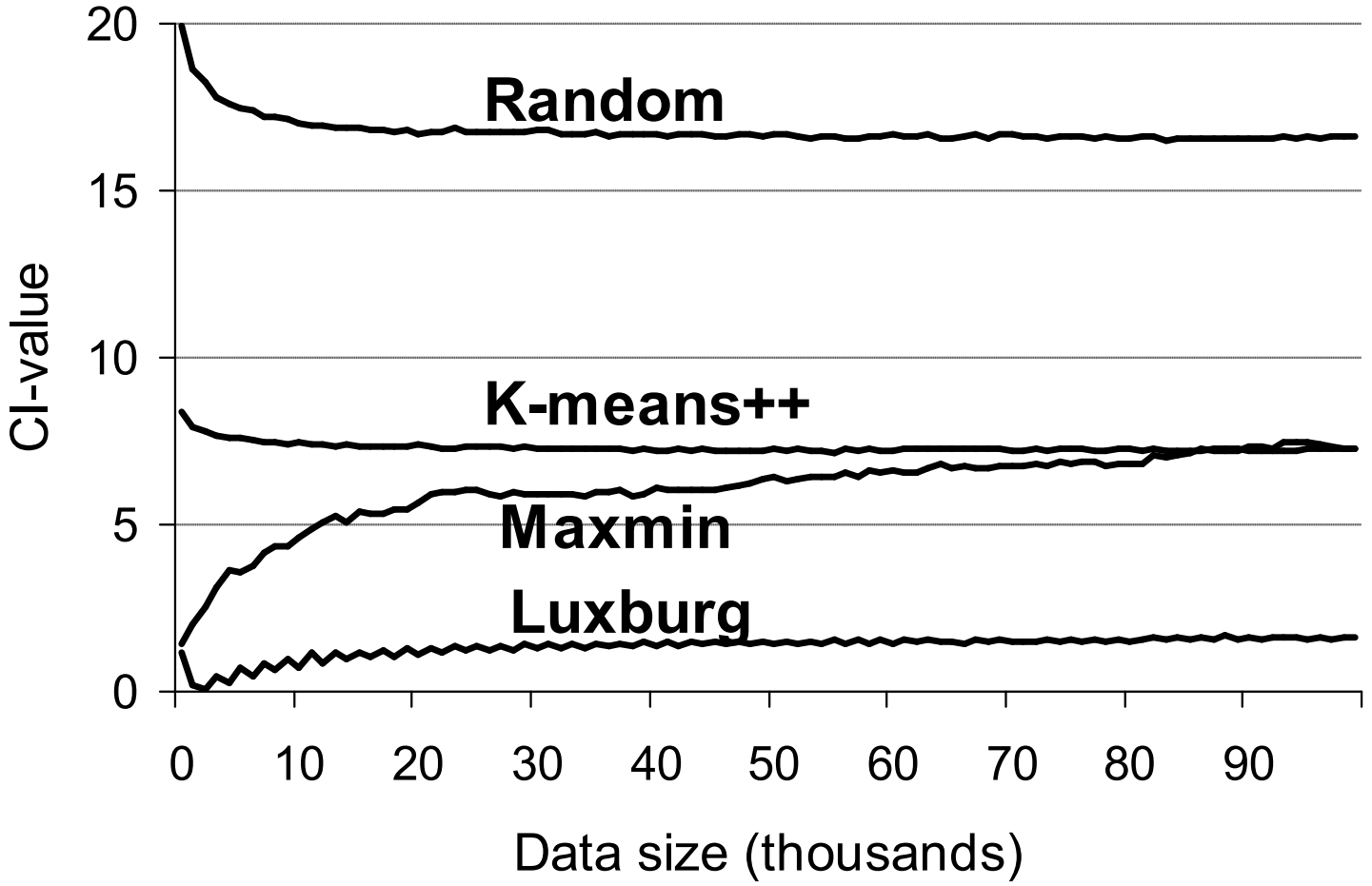
- Random centroids (Random) [Forgy][MacQueen]
- Further point heuristic (max) [Gonzalez]

More complex:

- K-means++ [Vasilievski]
- Luxburg [Luxburg]

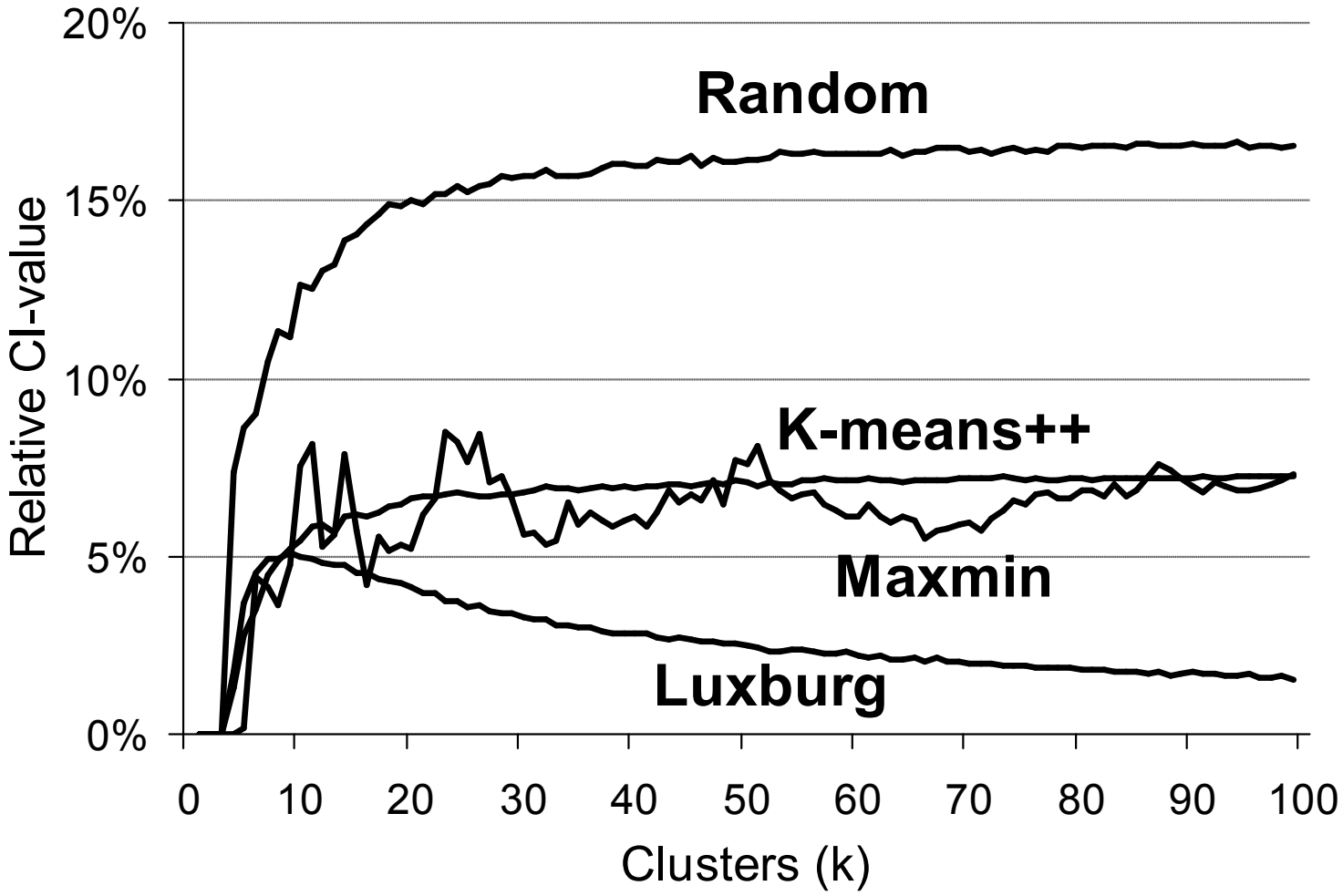
Initialization techniques

Varying **N**

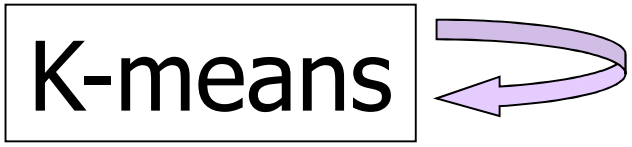


Initialization techniques

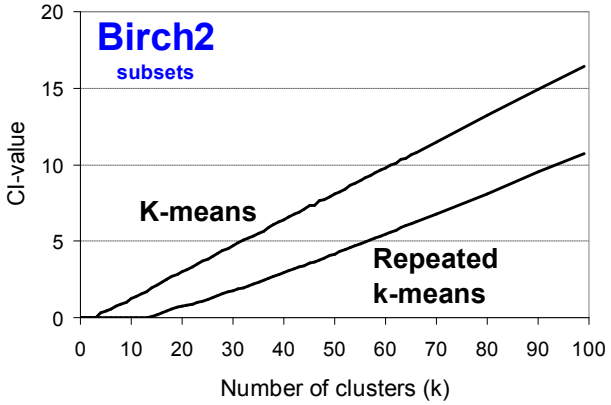
Varying **k**

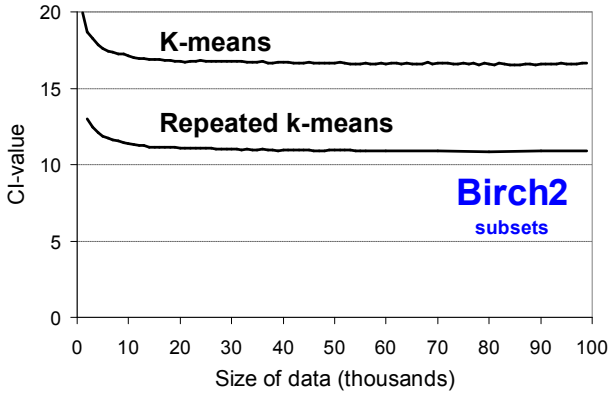


Repeated k-means (RKM)



- Repeat 100 times
- Can increase changes to success significantly
- In principle, running forever would solve
- Limitations if *k* is large





A better algorithm

Random Swap (RS)

Random Swap(X) $\rightarrow$ C, P

$C \leftarrow$ Select random representatives(X);
$P \leftarrow$ Optimal partition(X, C);
REPEAT T times
  (C^{new},j) $\leftarrow$ Random swap(X, C);
  $P^{new} \leftarrow$ Local repartition(X, C^{new}, P, j);
  C^{new}, $P^{new} \leftarrow$ *Kmeans*(X, C^{new}, P^{new});
  IF $f(C^{new}, P^{new}) < f(C, P)$ THEN
    (C, P) $\leftarrow$ C^{new}, P^{new};
RETURN (C, P);

http://cs.uef.fi/pages/franti/research/rs.txt

Genetic Algorithm (GA)

GeneticAlgorithm(X) $\rightarrow$ (C, P)
FOR $i \leftarrow 1$ TO Z DO
  $C^i \leftarrow$ RandomCodebook(X);
  $P^i \leftarrow$ OptimalPartition(X, C^i);
SortSolutions(C,P);
REPEAT
  {C,P} $\leftarrow$ CreateNewSolutions({C,P});
  SortSolutions(C,P);
UNTIL no improvement;

CreateNewSolutions({C, P}) $\rightarrow$ {C^{new}, P^{new} }
$C^{new\text{-}1}$, $P^{new\text{-}1} \leftarrow C^1$, P^1;
FOR $i \leftarrow 2$ TO Z DO
  (a,b) $\leftarrow$ SelectNextPair;
  $C^{new\text{-}i}$, $P^{new\text{-}i} \leftarrow$ Cross(C^a, P^a, C^b, P^b);
  IterateK-Means($C^{new\text{-}i}$, $P^{new\text{-}i}$);

Cross(C^1, P^1, C^2, P^2) $\rightarrow$ (C^{new}, P^{new})
$C^{new} \leftarrow$ CombineCentroids(C^1, C^2);
$P^{new} \leftarrow$ CombinePartitions(P^1, P^2);
$C^{new} \leftarrow$ UpdateCentroids(C^{new}, P^{new});
RemoveEmptyClusters(C^{new}, P^{new});
IS(C^{new}, P^{new});

CombineCentroids(C^1, C^2) $\rightarrow$ C^{new}
$C^{new} \leftarrow C^1 \cup C^2$

CombinePartitions(C^{new}, P^1, P^2) $\rightarrow$ P^{new}
FOR $i \leftarrow 1$ TO N DO
  IF $\left\| x_i - c_{p_i^1} \right\|^2 \leq \left\| x_i - c_{p_i^2} \right\|^2$ THEN
    $p_i^{new} \leftarrow p_i^1$
  ELSE
    $p_i^{new} \leftarrow p_i^2$
END-FOR

UpdateCentroids(C^1, C^2) $\rightarrow$ C^{new}
FOR $j \leftarrow 1$ TO $|C^{new}|$ DO
  $c_j^{new} \leftarrow$ CalculateCentroid(P^{new}, j);

cs.uef.fi/pages/franti/research/ga.txt

Overall comparison

CI-values

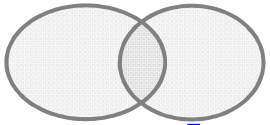
Dataset	K-means					RKM	RS	GA
	Ste	Ran	Max	K++	Lux			
A1	6.0	2.5	1.0	1.5	0.6	0.3	0.0	0.0
A2	10.7	4.5	2.6	2.9	0.9	1.8	0.0	0.0
A3	17.9	6.6	2.9	4.2	1.0	2.9	0.0	0.0
S1	3.2	1.8	0.7	1.0	0.5	0.1	0.0	0.0
S2	0.6	1.4	1.0	0.9	0.4	0.0	0.0	0.0
S3	1.2	1.3	0.7	1.0	0.6	0.0	0.0	0.0
S4	0.4	0.9	1.0	0.8	0.4	0.0	0.0	0.0
Unbalance	4.0	3.9	0.9	0.5	4.0	2.9	0.0	0.0
Birch1	11.3	6.6	5.5	4.9	2.7	2.8	0.0	0.0
Birch2	75.5	16.6	7.3	7.2	1.6	10.9	0.0	0.0
Dim32	5.5	3.6	0.0	0.1	0.04	1.1	0.0	0.0
Average:	**12.4**	**4.5**	**2.2**	**2.3**	**1.2**	**2.1**	**0.0**	**0.0**

Conclusions

Conclusions

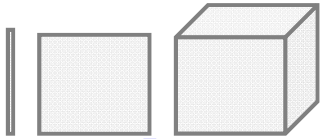
How did K-means perform?

1. Overlap



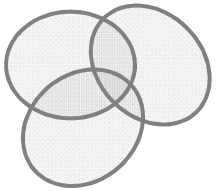
Good!

3. Dimensionality



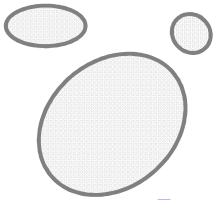
No change

2. Number of clusters



Bad!

4. Unbalance of cluster sizes



Bad!

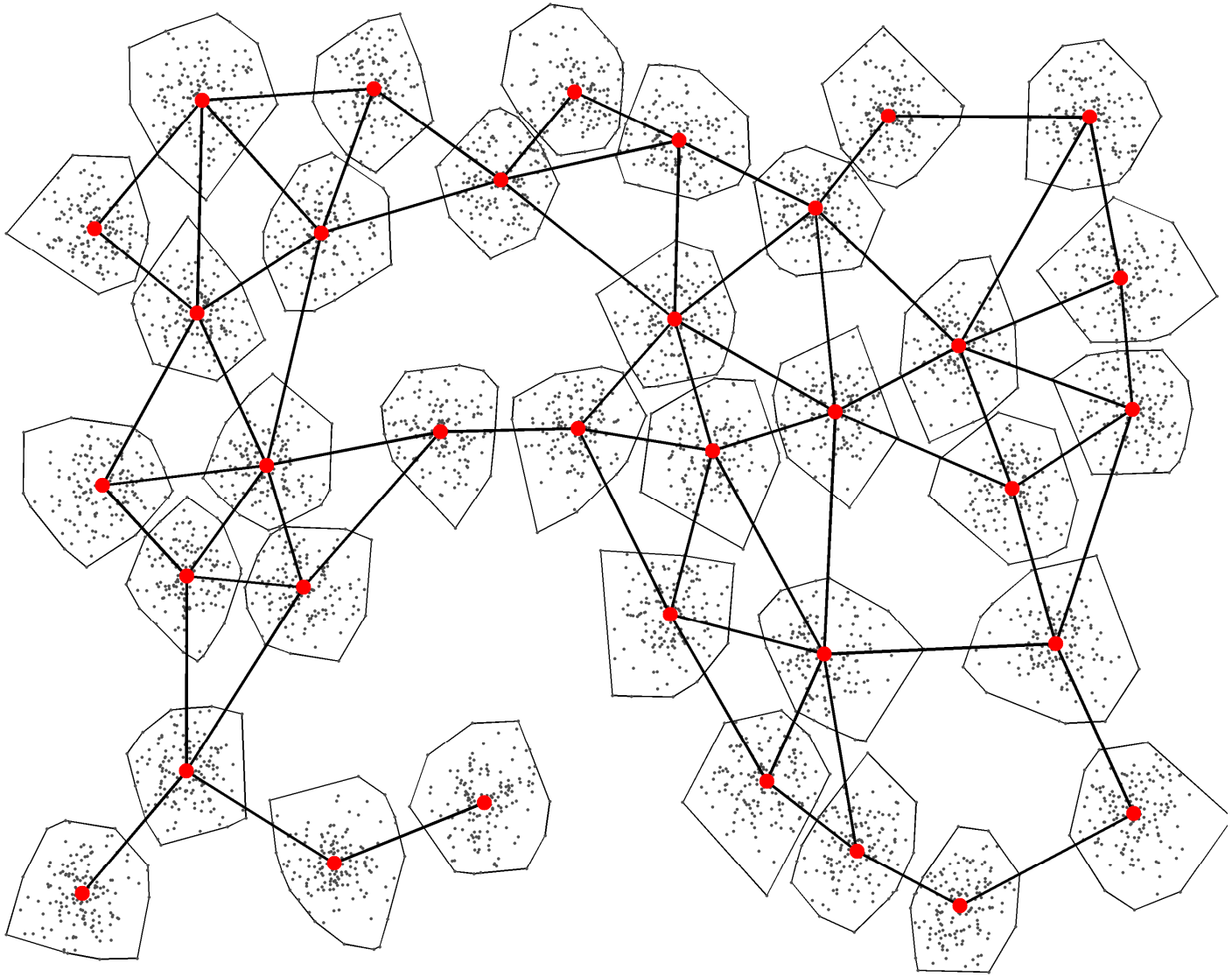
References

- J. MacQueen, Some methods for classification and analysis of multivariate observations, *Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1: Statistics, pp. 281-297, University of California Press, Berkeley, Calif., 1967.
- S.P. Lloyd, Least squares quantization in PCM, *IEEE Trans. on Information Theory*, 28 (2), 129–137, 1982.
- Forgy, E. (1965). Cluster analysis of multivariate data: Efficiency vs. interpretability of classification. *Biometrics*, 21, 768.
- M. Steinbach, L. Ertöz, V. Kumar, The challenges of clustering high dimensional data, *New Vistas in Statistical Physics -- Applications in Econophysics, Bioinformatics, and Pattern Recognition*, Springer-Verlag, 2003.
- U. Luxburg, R.C. Williamson, I. Guyon, "Clustering: Science or Art?", *J. Machine Learning Research*, 27: 65–79, 2012.
- P. Fränti, "Genetic algorithm with deterministic crossover for vector quantization", *Pattern Recognition Letters*, 21 (1), 61-68, 2000
- P. Fränti and J. Kivijärvi, "Randomized local search algorithm for the clustering problem", *Pattern Analysis and Applications*, 3 (4), 358-369, 2000.
- P. Fränti, O. Virmajoki and V. Hautamäki, Fast agglomerative clustering using a k-nearest neighbor graph, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28 (11), 1875-1881, November 2006.
- Zhang R. Ramakrishnan and M. Livny, BIRCH: A new data clustering algorithm and its applications, *Data Mining and Knowledge Discovery*, 1 (2), 141-182, 1997.
- I. Kärkkäinen and P. Fränti, Dynamic local search algorithm for the clustering problem, *Research Report A-2002-6*.
- P. Fränti and O. Virmajoki, Iterative shrinking method for clustering problems, *Pattern Recognition*, 39 (5), 761-765, May 2006.
- P. Fränti R. Mariescu-Istodor and C. Zhong, XNN graph *IAPR Joint Int. Workshop on Structural, Syntactic, and Statistical Pattern Recognition Merida, Mexico, LNCS 10029,* 207-217, November 2016.
- M. Rezaei and P. Fränti, "Set-matching methods for external cluster validity", *IEEE Trans. on Knowledge and Data Engineering*, 28 (8), 2173-2186, August 2016.
- E. Chavez and G. Navarro, A probabilistic spell for the curse of dimensionality. *Workshop on Algorithm Engineering and Experimentation*, LNCS 2153, 147-160, 2001.
- N. Tomasev, M. Radovanovi, D. Mladeni and M. Ivanovi, "The role of hubness in clustering high-dimensional data", *IEEE Trans. on Knowledge and Data Engineering*, 26 (3), 739-751, March 2014.
- D. Steinley, Local optima in k-means clustering: what you don't know may hurt you", *Psychological Methods*, 8, 294–304, 2003.
- P. Fränti, M. Rezaei and Q. Zhao, "Centroid index: Cluster level similarity measure", *Pattern Recognition*, 47 (9), 3034-3045, 2014.
- T. Gonzalez, Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38 (2–3), 293–306, 1985.

Back-up material

K-means neighbors

Estimation

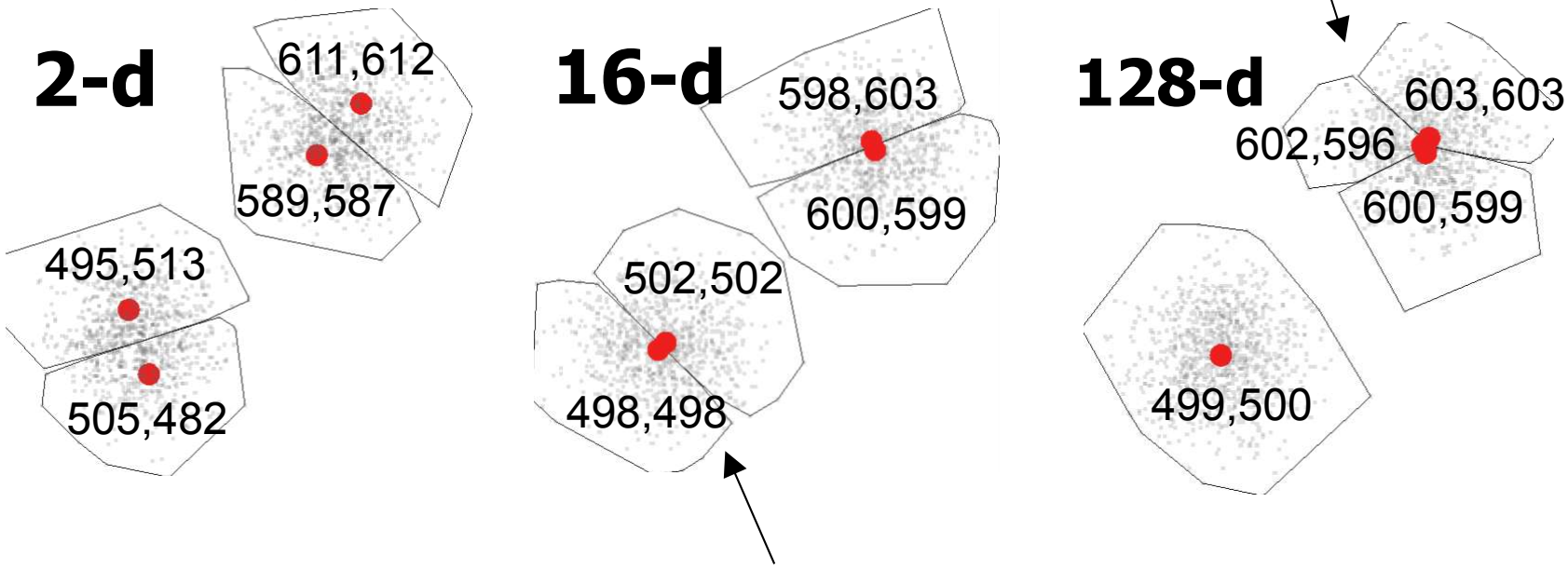


Shrinking of the space

Clusters divided to two halves

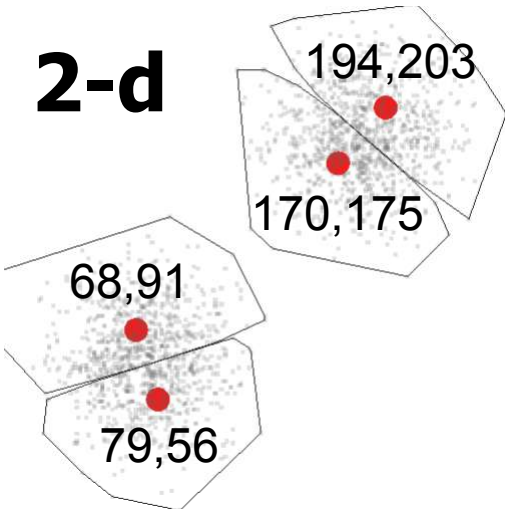
Centroids clearly separated

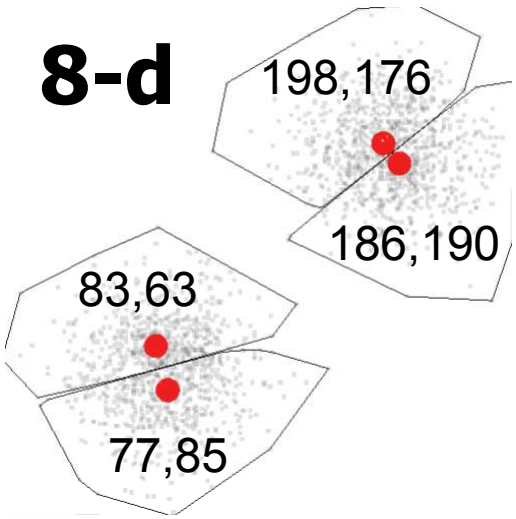
No separation anymore.
Can even cause 3:1 division.

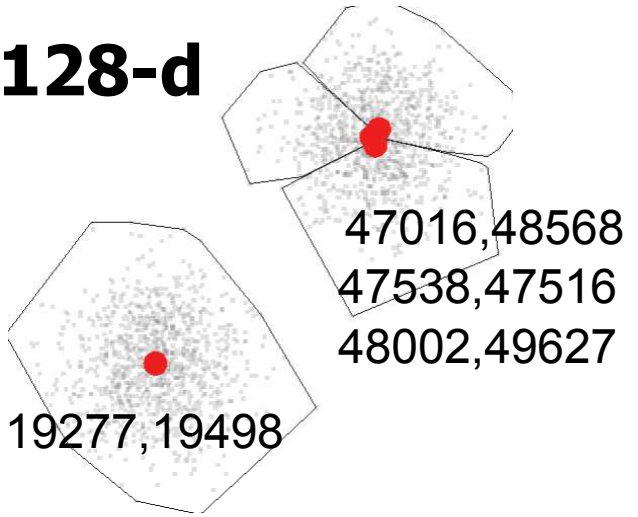


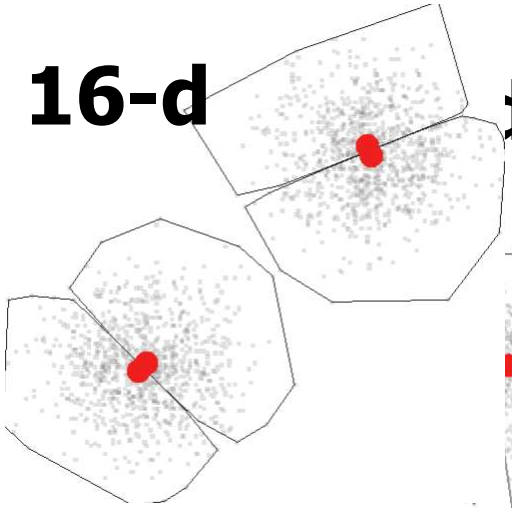
Centroids closing to the group center

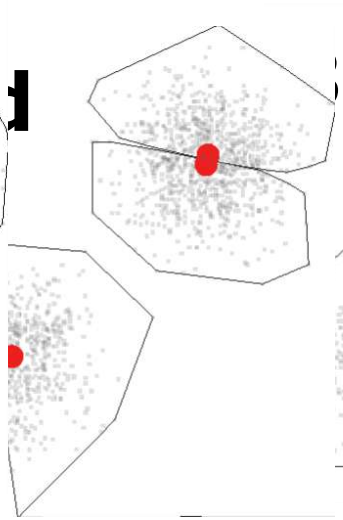
Backup material

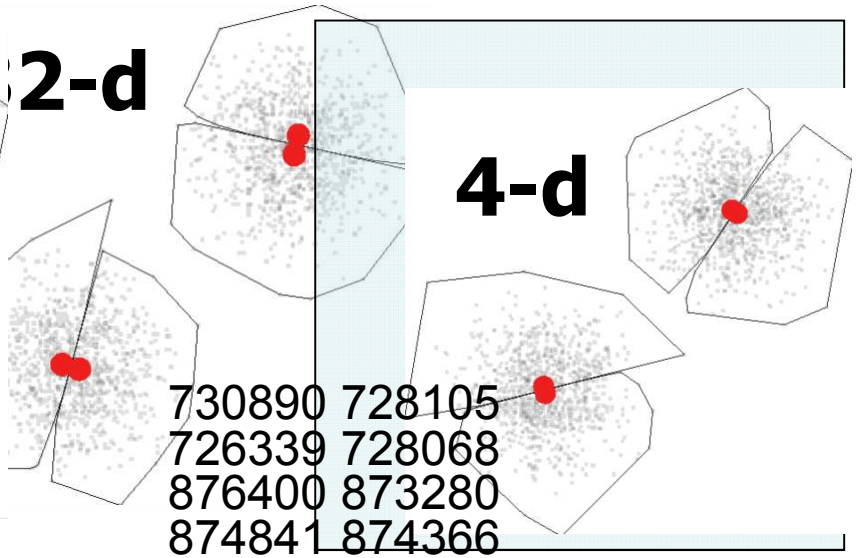




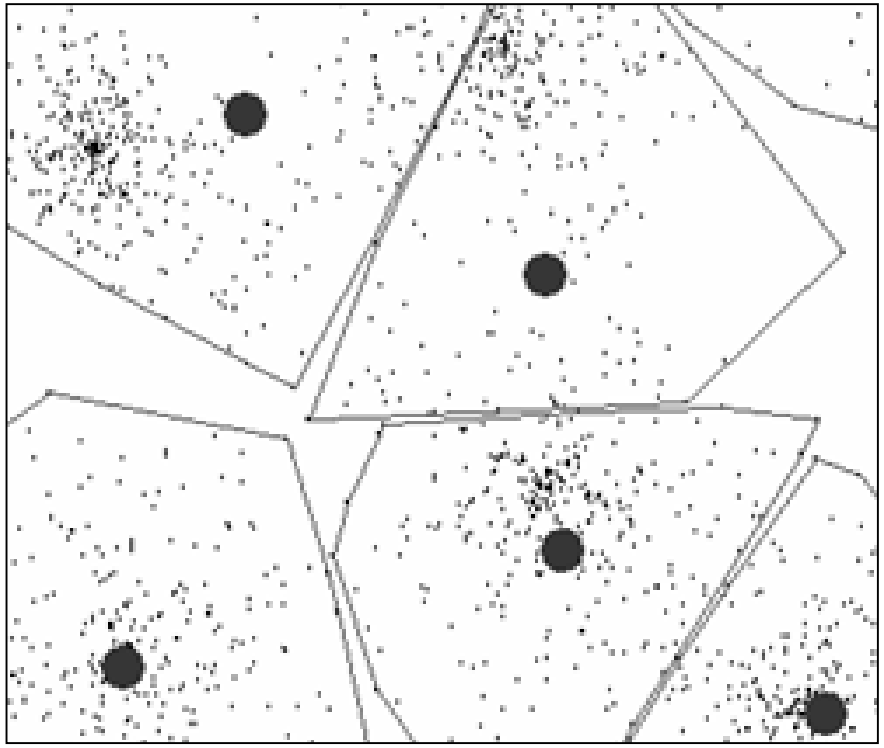


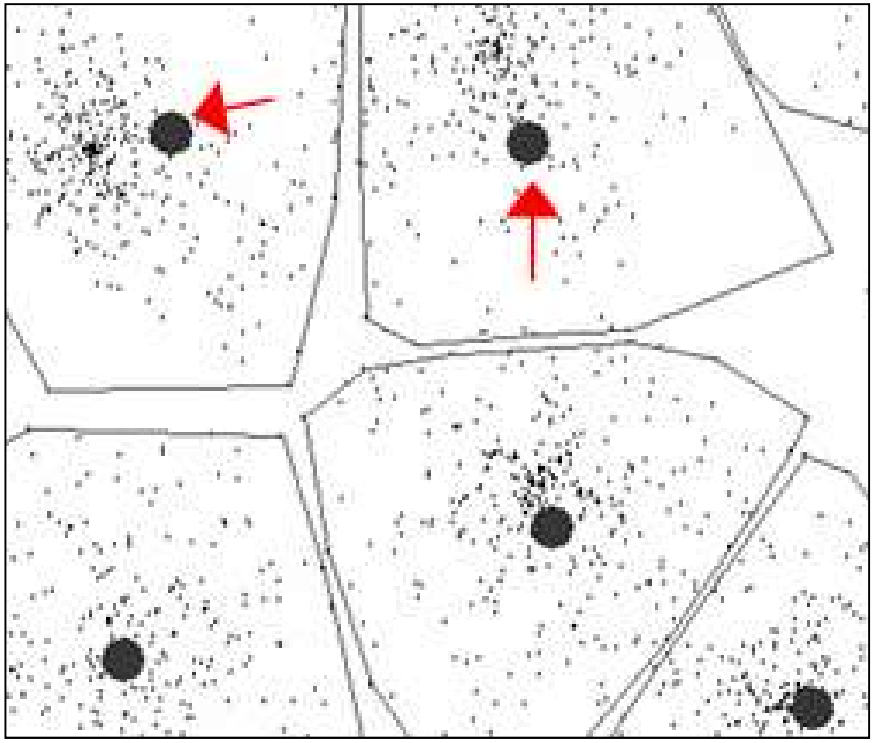






How K-means process goes





Time complexity

Efficiency of k-means

Total time to find correct clustering:

- Time per iteration $\times$ Number of iterations

Time complexity of single iteration:

- Swap: $O(1)$
- Remove cluster: $2k \cdot N/k = O(N)$
- Add cluster: $2N = O(N)$
- Centroids: $2N/k + 2N/k + 2\alpha = O(N/k)$
- K-means: $I \cdot k \cdot N = O(IkN)$

Efficiency of fast k-means

Total time to find correct clustering:

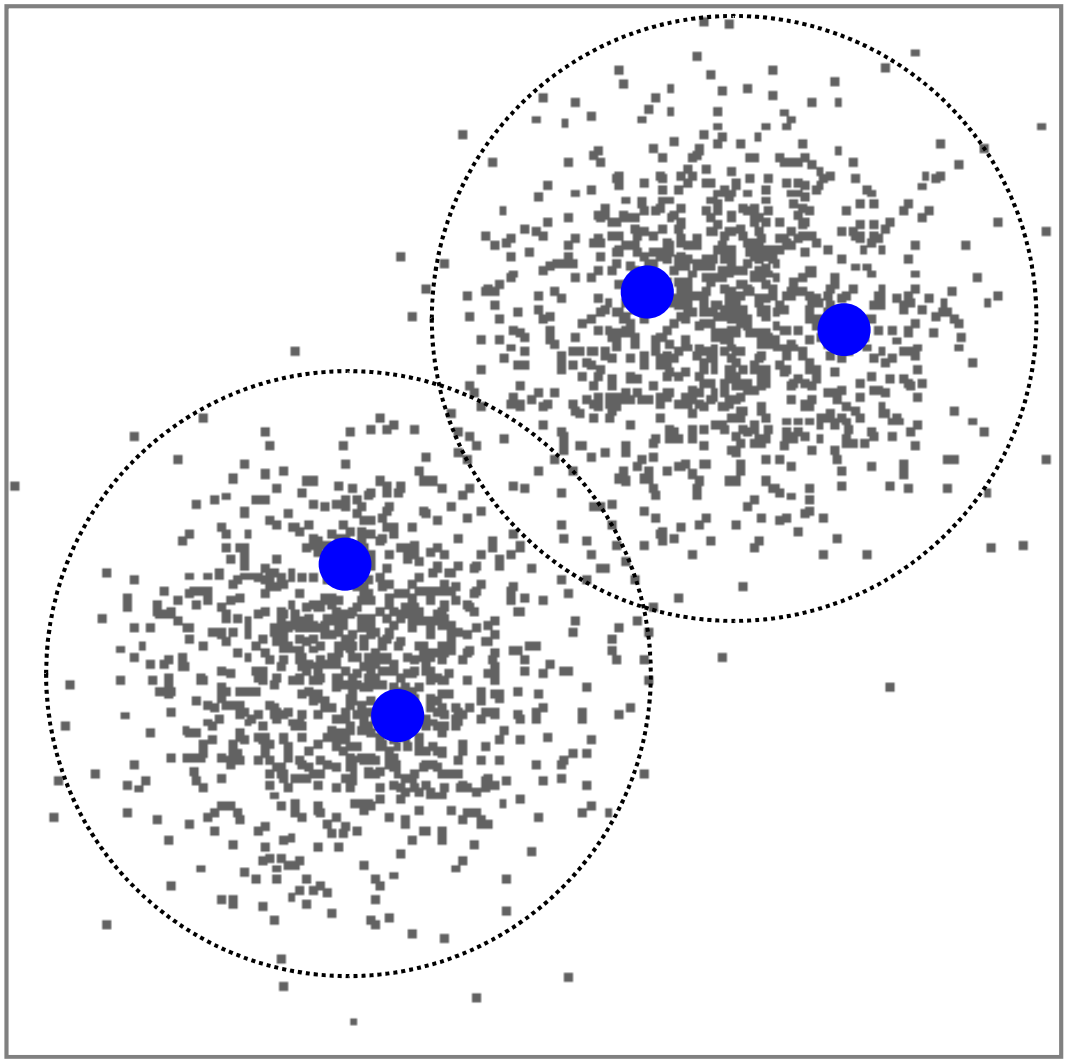
- Time per iteration × Number of iterations

Time complexity of single iteration:

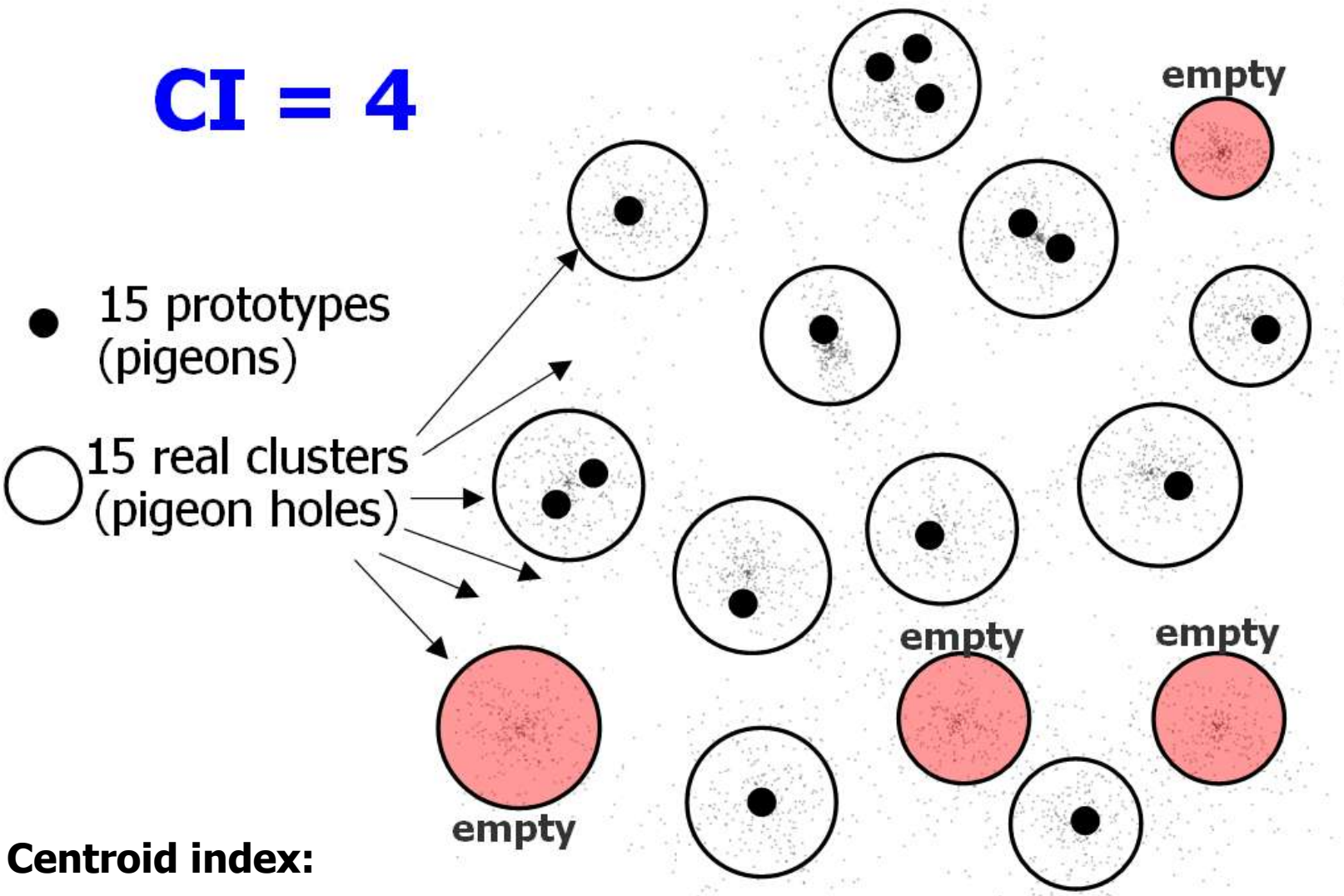
- Swap: $O(1)$
- Remove cluster: $2k \cdot N/k = O(N)$
- Add cluster: $2N = O(N)$
- Centroids: $2N/k + 2N/k + 2\alpha = O(N/k)$
- (Fast) K-means: $4\alpha \cdot N = O(\alpha N)$

2 iterations only!

T. Kaukoranta, P. Fränti and O. Nevalainen
"A fast exact GLA based on code vector activity detection"
IEEE Trans. on Image Processing, 9 (8), 1337-1342, August 2000.



Centroid index



CI = Centroid index:

P. Fränti, M. Rezaei and Q. Zhao
"Centroid index: cluster level similarity measure"
Pattern Recognition, 47 (9), 3034-3045, September 2014, 2014.

Data sets visualized

Artificial