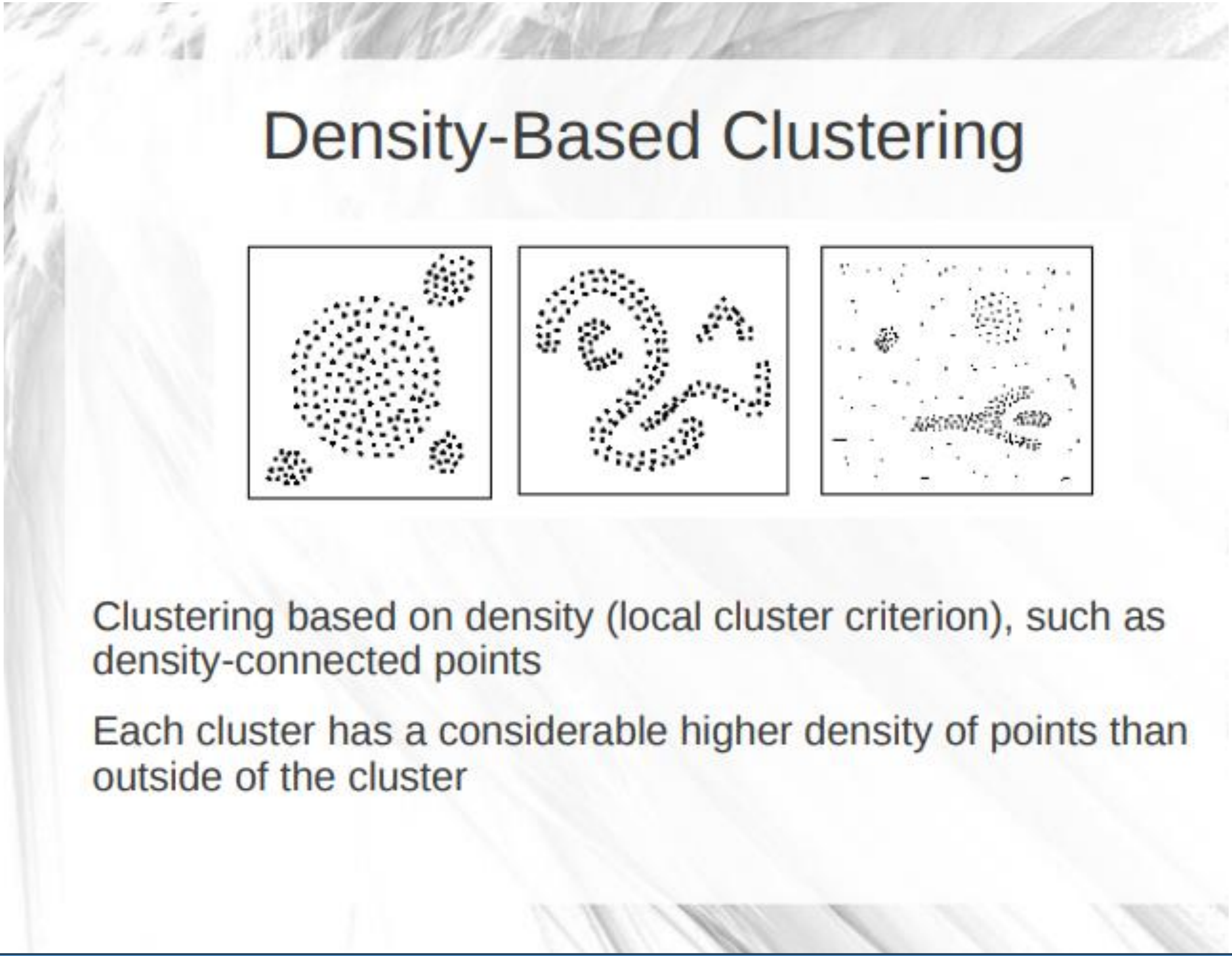
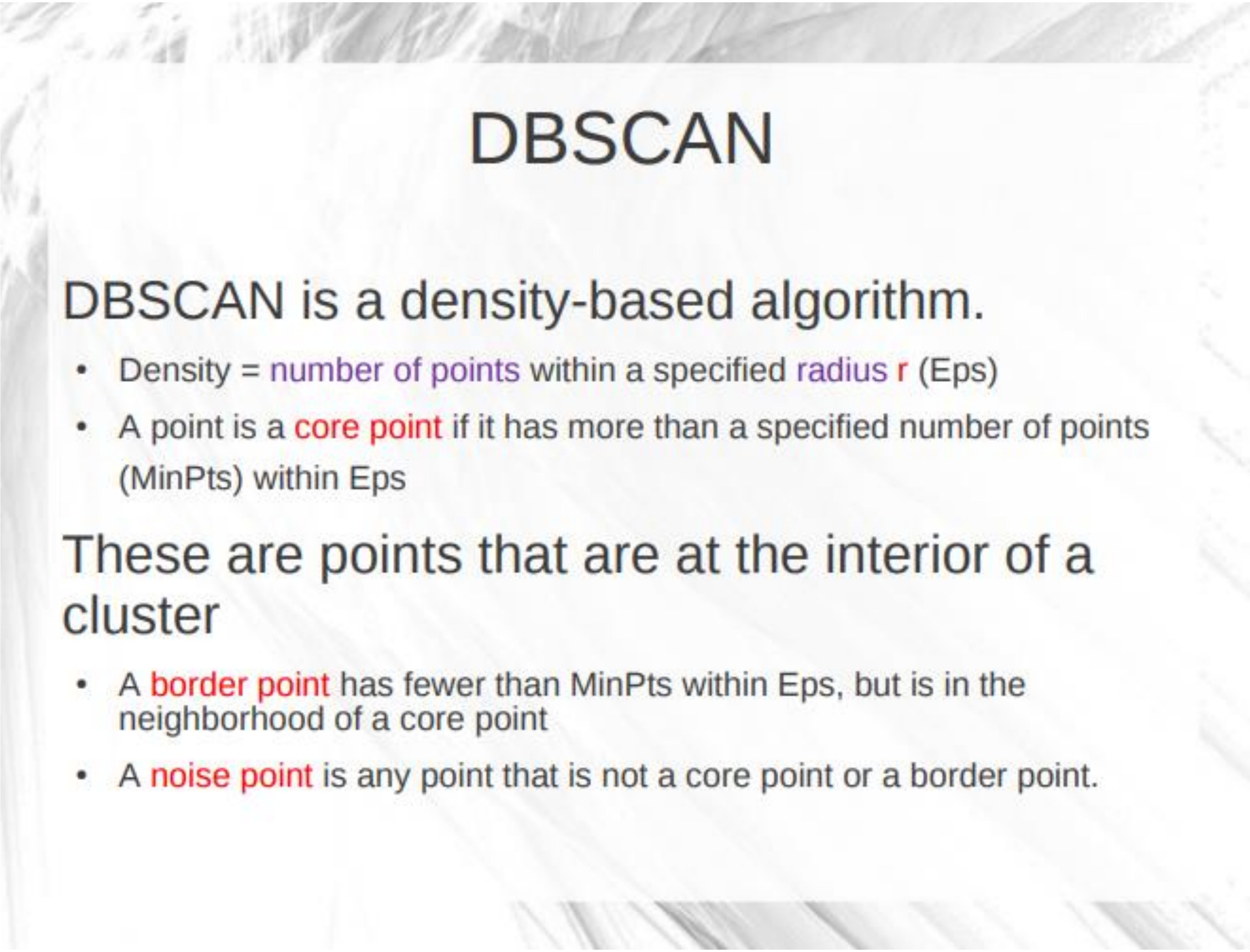
DB-SCAN



Density-Based Clustering

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

Each cluster has a considerable higher density of points than outside of the cluster



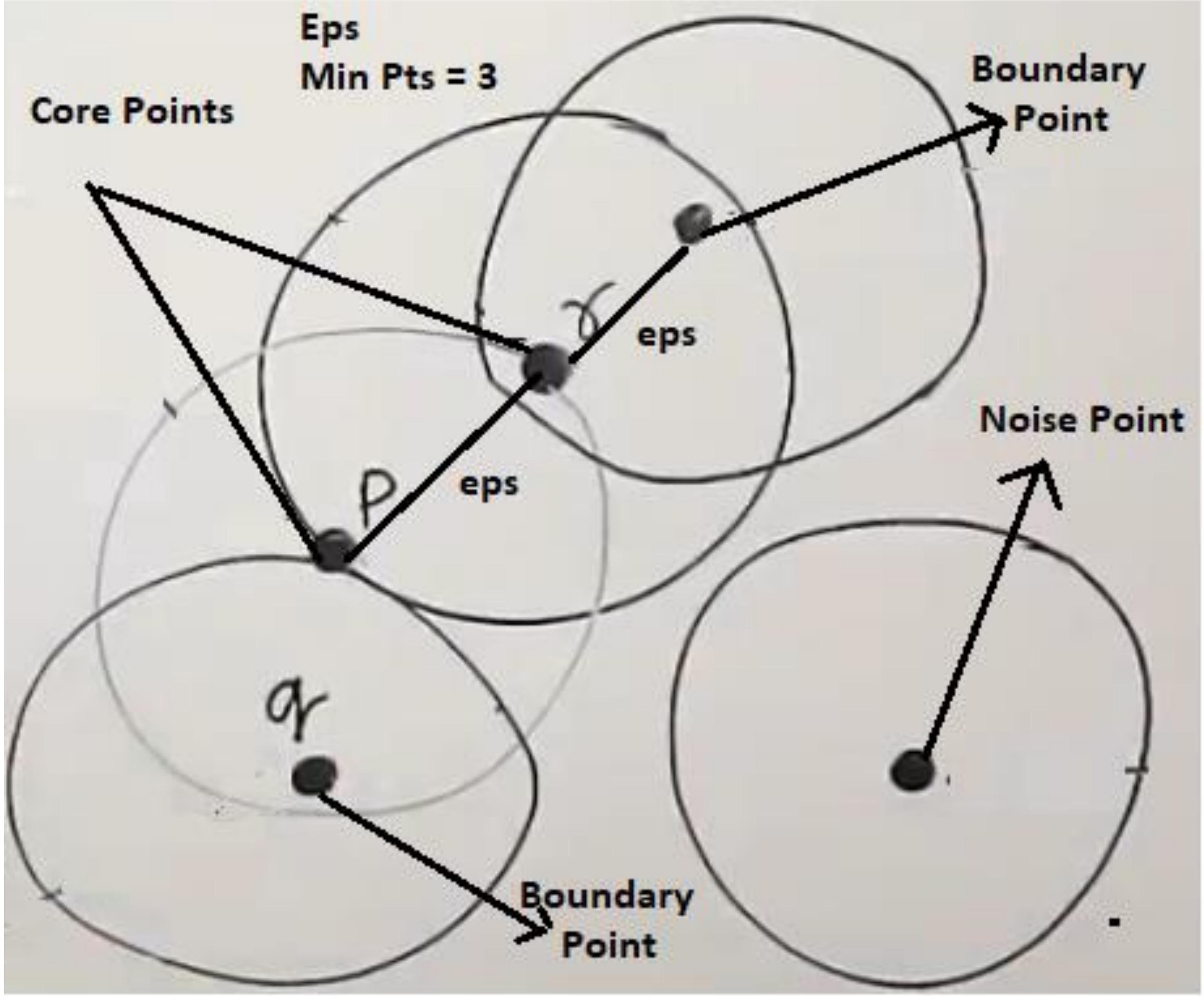
DBSCAN

DBSCAN is a density-based algorithm.

- Density = number of points within a specified radius r (Eps)
- A point is a core point if it has more than a specified number of points (MinPts) within Eps

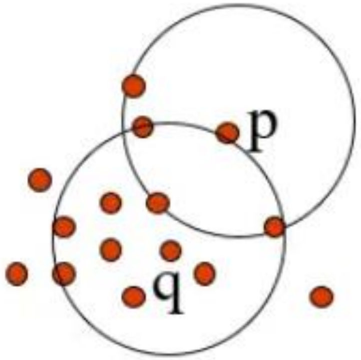
These are points that are at the interior of a cluster

- A border point has fewer than MinPts within Eps, but is in the neighborhood of a core point
- A noise point is any point that is not a core point or a border point.


Eps
Min Pts = 3
Core Points
Boundary Point
r
eps
P
eps
Noise Point
q
Boundary Point

Density-Based Clustering: Basic Concepts

- Two parameters:
  - Eps: Maximum radius of the neighbourhood
  - MinPts: Minimum number of points in an Eps-neighbourhood of that point
- NEps(p): {q belongs to D | dist(p,q) ≤ Eps}
- Directly density-reachable: A point p is directly density-reachable from a point q w.r.t. Eps, MinPts if
  - p belongs to NEps(q)
  - core point condition:
  - |NEps (q)| ≥ MinPts

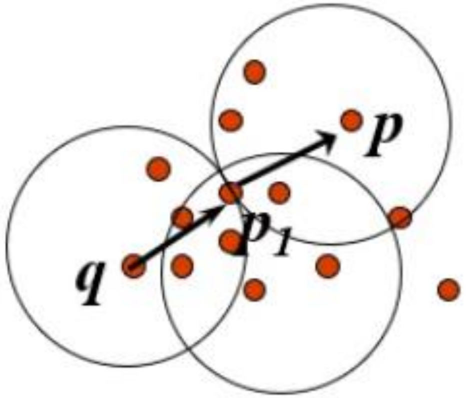


MinPts = 5

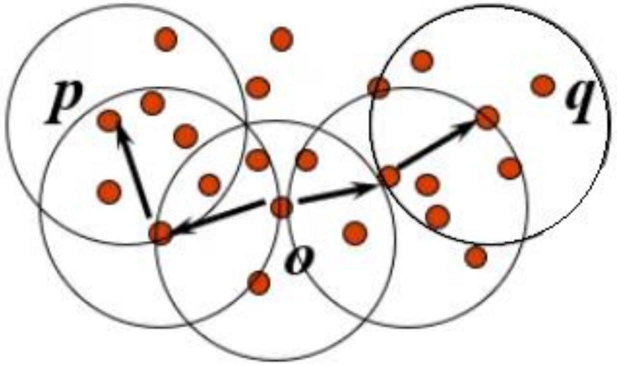
Eps = 1

Density-Reachable and Density-Connected

- Density-reachable:
  - A point p is density-reachable from a point q w.r.t. Eps, MinPts if there is a chain of points p1, …, pn, p1 = q, pn = p such that pi+1 is directly density-reachable from pi

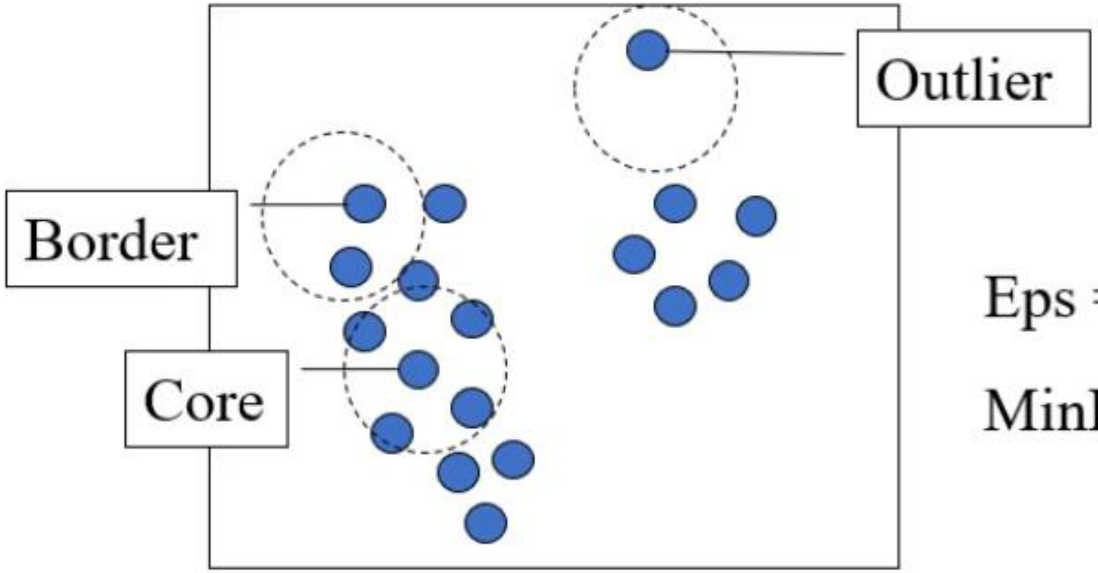


- Density-connected
  - A point p is density-connected to a point q w.r.t. Eps, MinPts if there is a point o such that both, p and q are density-reachable from o w.r.t. Eps and MinPts



DBSCAN: Density-Based Spatial Clustering of Applications with Noise

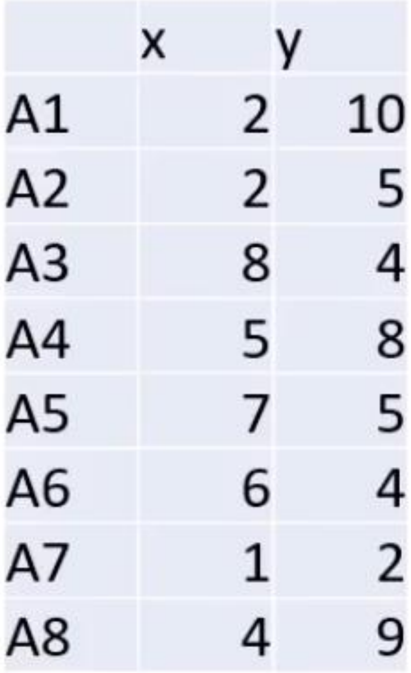
- Relies on a density-based notion of cluster: A cluster is defined as a maximal set of density-connected points
- Discovers clusters of arbitrary shape in spatial databases with noise



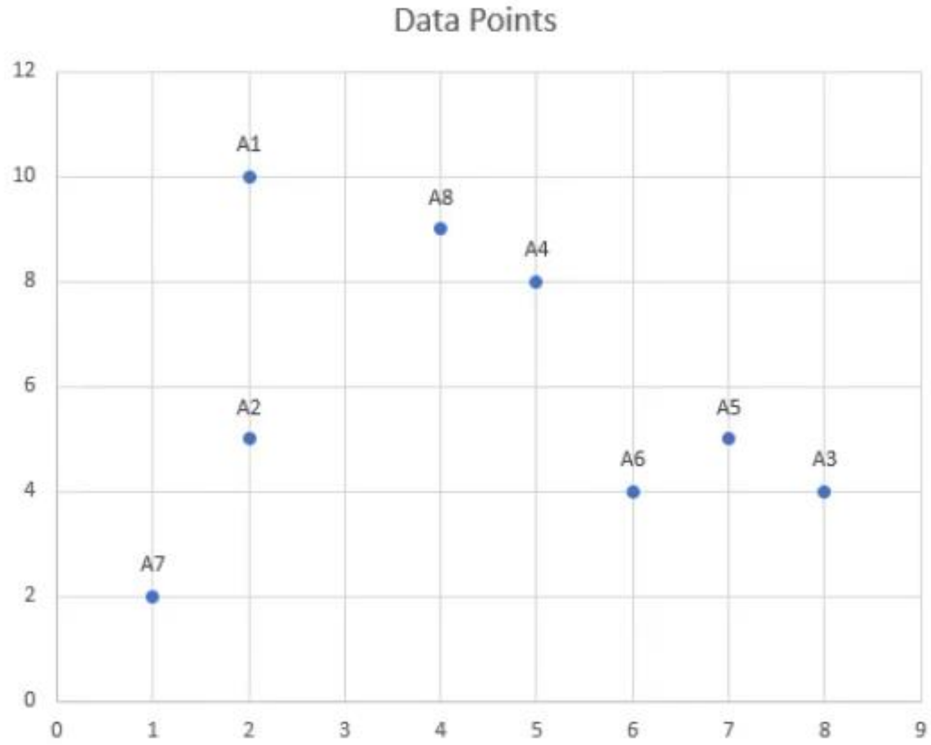
Eps = 1cm

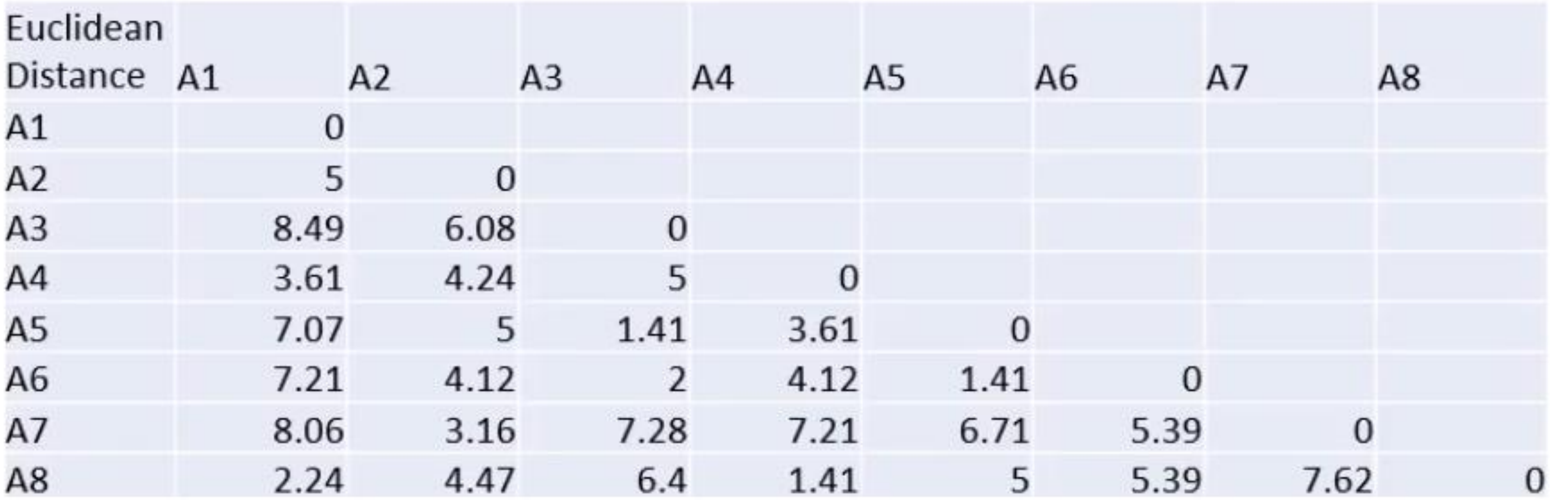
MinPts = 5

Min-pts = 3

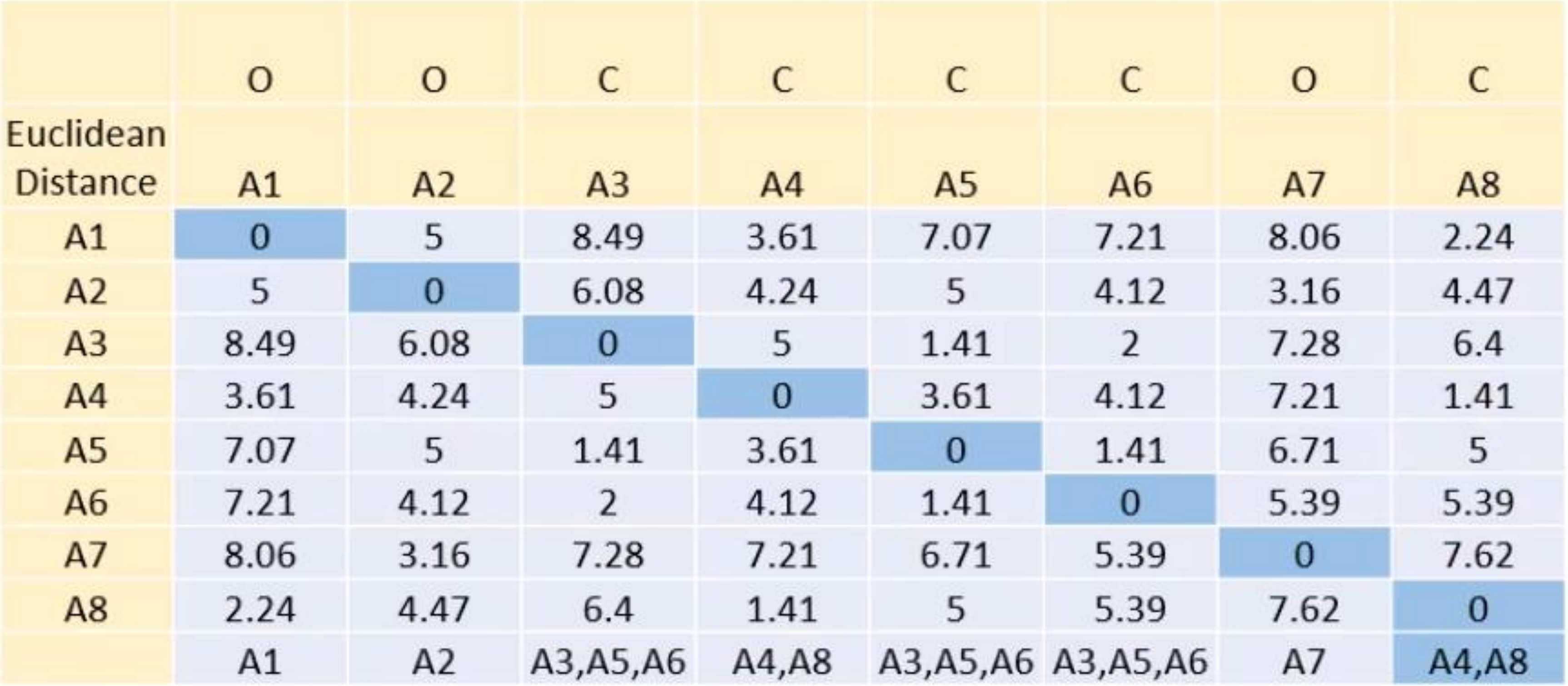


	x	y
A1	2	10
A2	2	5
A3	8	4
A4	5	8
A5	7	5
A6	6	4
A7	1	2
A8	4	9





Euclidean Distance	A1	A2	A3	A4	A5	A6	A7	A8
A1	0							
A2	5	0						
A3	8.49	6.08	0					
A4	3.61	4.24	5	0				
A5	7.07	5	1.41	3.61	0			
A6	7.21	4.12	2	4.12	1.41	0		
A7	8.06	3.16	7.28	7.21	6.71	5.39	0	
A8	2.24	4.47	6.4	1.41	5	5.39	7.62	0



	O	O	C	C	C	C	O	C
Euclidean Distance	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	5	8.49	3.61	7.07	7.21	8.06	2.24
A2	5	0	6.08	4.24	5	4.12	3.16	4.47
A3	8.49	6.08	0	5	1.41	2	7.28	6.4
A4	3.61	4.24	5	0	3.61	4.12	7.21	1.41
A5	7.07	5	1.41	3.61	0	1.41	6.71	5
A6	7.21	4.12	2	4.12	1.41	0	5.39	5.39
A7	8.06	3.16	7.28	7.21	6.71	5.39	0	7.62
A8	2.24	4.47	6.4	1.41	5	5.39	7.62	0
	A1	A2	A3,A5,A6	A4,A8	A3,A5,A6	A3,A5,A6	A7	A4,A8

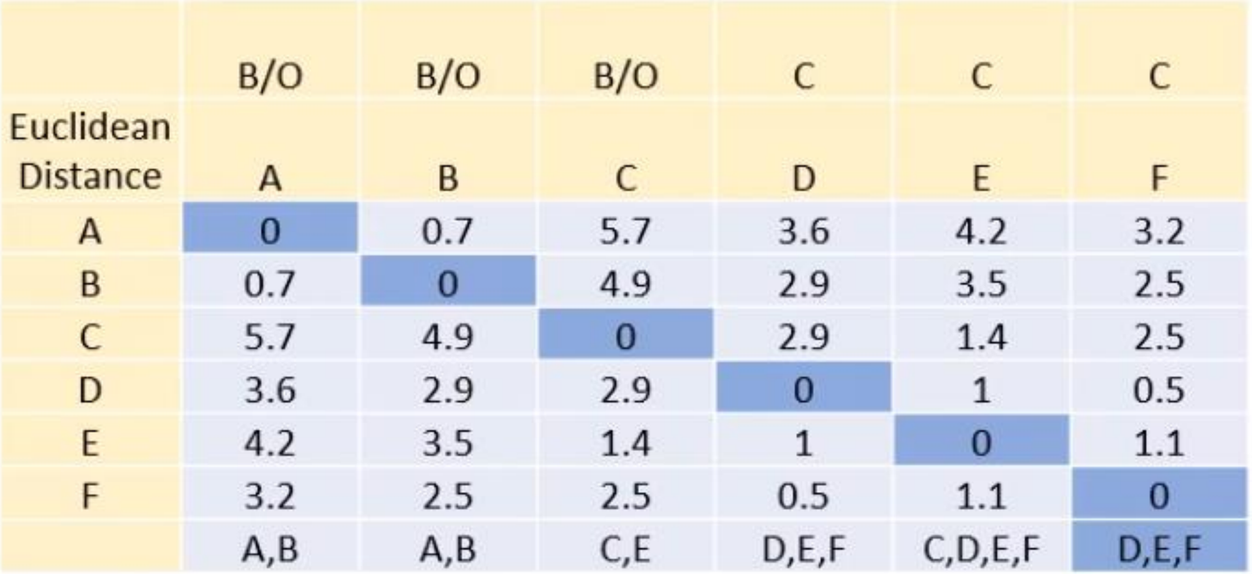
ε = 2 and min pt = 2

A1, A2 and A7 are outliers

There are 2 clusters formed:

Cluster1: A3,A5,A6

Cluster2: A4,A8



	B/O	B/O	B/O	C	C	C
Euclidean Distance	A	B	C	D	E	F
A	0	0.7	5.7	3.6	4.2	3.2
B	0.7	0	4.9	2.9	3.5	2.5
C	5.7	4.9	0	2.9	1.4	2.5
D	3.6	2.9	2.9	0	1	0.5
E	4.2	3.5	1.4	1	0	1.1
F	3.2	2.5	2.5	0.5	1.1	0
	A,B	A,B	C,E	D,E,F	C,D,E,F	D,E,F

ε = 2 and min pt = 3

Visiting the Border/Outlier Points:

C: Neighborhood points = E which is a core point

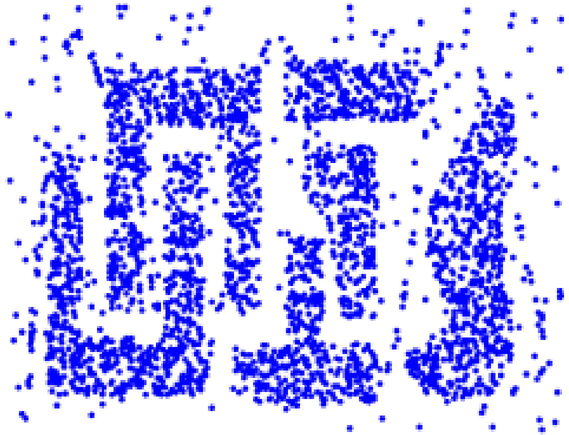
Thus C is a Border point

A: Neighborhood points = B which is a B/O

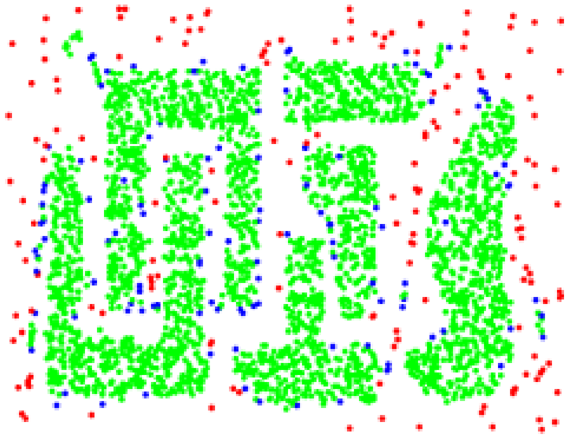
B: Neighborhood points = A which is a B/O

Thus A and B are outlier

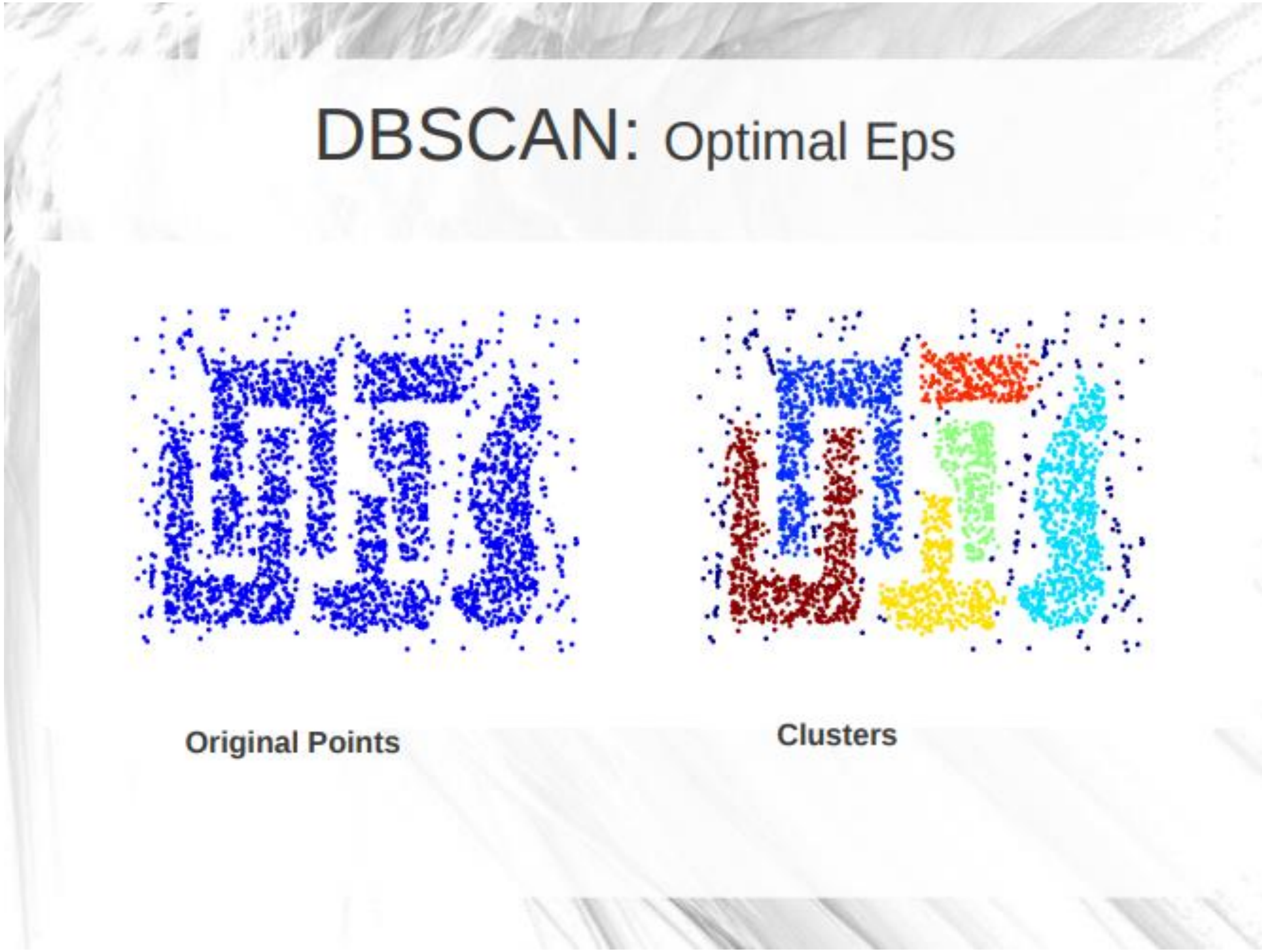
DBSCAN: Large Eps



Original Points



Point types: core, border and noise



DBSCAN: Optimal Eps

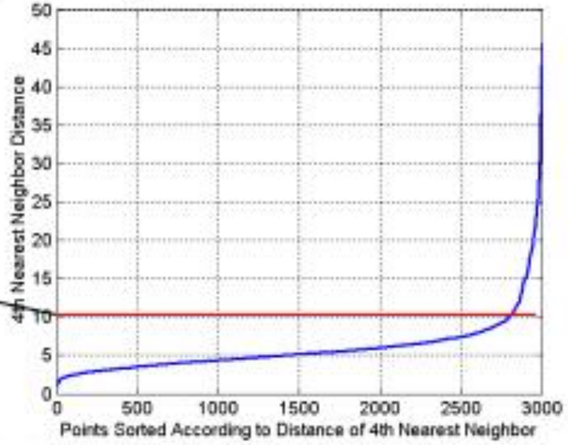
Original Points

Clusters

Determining Eps and MinPts

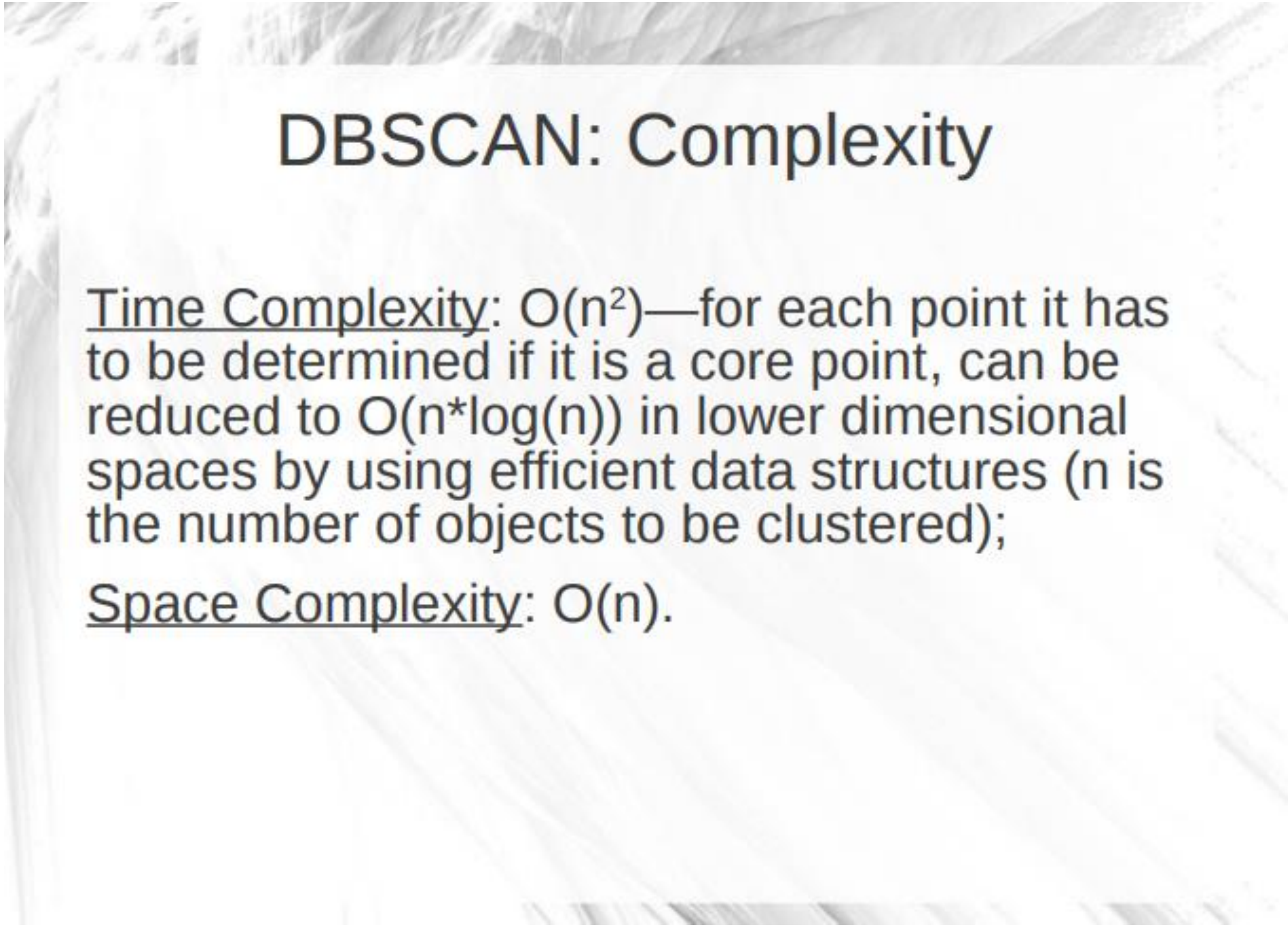
- Idea is that for points in a cluster, their kth nearest neighbors are at roughly the same distance
- Noise points have the kth nearest neighbor at farther distance
- So, plot sorted distance of every point to its kth nearest neighbor (e.g., k=4)

Thus, eps=10



DB Scan algorithm

1. $S=\{x_1,x_2.....x_n\} \in R^m$
2. Chose value for $r>0$ and $\in > 0$ where $r=$ no of points and $\in=$ distance of two points
3. $A_i =\{ x \in S : d(x_i,x) < \in \}$; $i=1,2,3.....n$
4. if $|Ai| >= r$, calculate all those Ai
5. Take union of A_i & A_j if $A_i \cap A_j \neq \emptyset$
6. Repeat 5 till no union take place.



DBSCAN: Complexity

Time Complexity: O(n^2)—for each point it has to be determined if it is a core point, can be reduced to O(n*log(n)) in lower dimensional spaces by using efficient data structures (n is the number of objects to be clustered);

Space Complexity: O(n).

Questions?