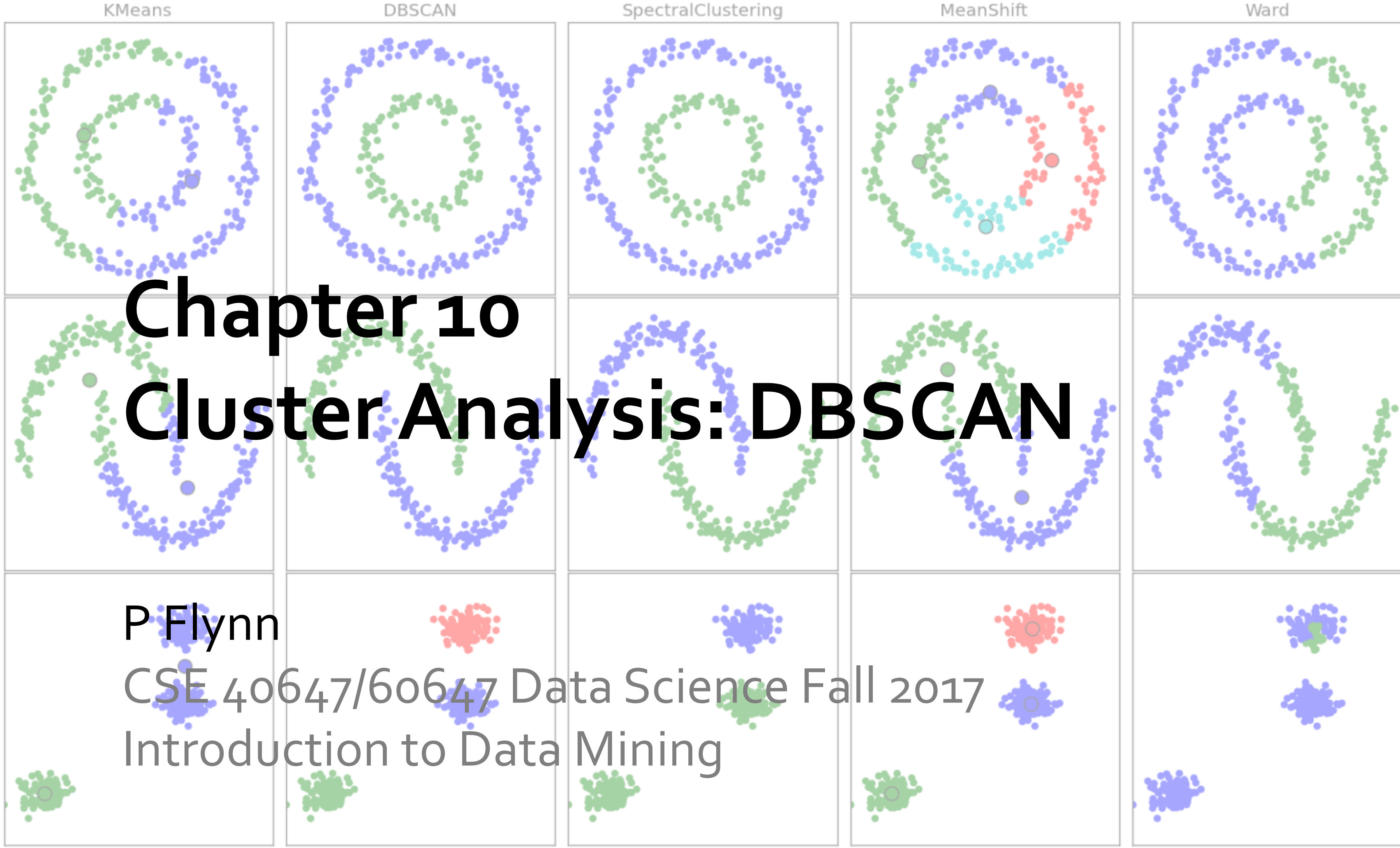


Chapter 10
Cluster Analysis: DBSCAN

P Flynn

CSE 40647/60647 Data Science Fall 2017

Introduction to Data Mining

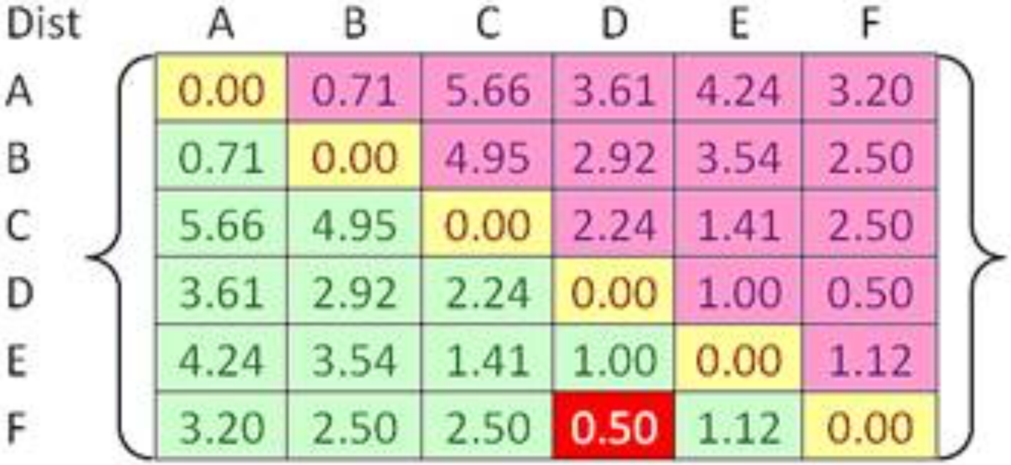
Some material from Margareta Ackerman, Santa Clara University

Book-keeping

- Link
- HW5 due 11/28 – you have everything you need to do this assignment
- Project due 11/30

A dangling thought from last time

- We assume we have “items” (“points”, “patterns”) with locations in some feature space.
- Really?
- How about JUST a distance matrix?



Dist	A	B	C	D	E	F
A	0.00	0.71	5.66	3.61	4.24	3.20
B	0.71	0.00	4.95	2.92	3.54	2.50
C	5.66	4.95	0.00	2.24	1.41	2.50
D	3.61	2.92	2.24	0.00	1.00	0.50
E	4.24	3.54	1.41	1.00	0.00	1.12
F	3.20	2.50	2.50	0.50	1.12	0.00

http://people.revoledu.com/kardi/tutorial/Clustering/Numerical%20Example.htm

Cluster Analysis

- Cluster Analysis: An Introduction
- Partitioning Methods
- **Density-based Methods**
- Evaluation of Clustering

Density-Based and Grid-Based Clustering Methods

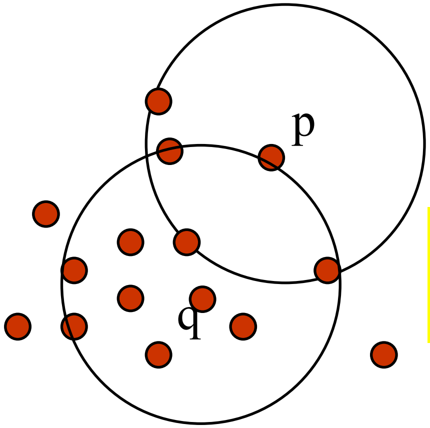
- Density-Based Clustering
  – Basic Concepts
  – **DBSCAN: A Density-Based Clustering Algorithm**
  – OPTICS: Ordering Points To Identify Clustering Structure
- Grid-Based Clustering Methods
  – Basic Concepts
  – STING: A Statistical Information Grid Approach
  – CLIQUE: Grid-Based Subspace Clustering

Density-Based Clustering Methods

- Clustering based on density (a local cluster criterion), such as density-connected points
- Major features:
  - Discover clusters of arbitrary shape
  - Handle noise
  - One scan (only examine the local region to justify density)
  - Need density parameters as termination condition
- Several interesting studies:
  - DBSCAN: Ester, et al. (KDD'96)
  - OPTICS: Ankerst, et al (SIGMOD'99)
  - DENCLUE: Hinneburg & D. Keim (KDD'98)
  - CLIQUE: Agrawal, et al. (SIGMOD'98) (also, grid-based)

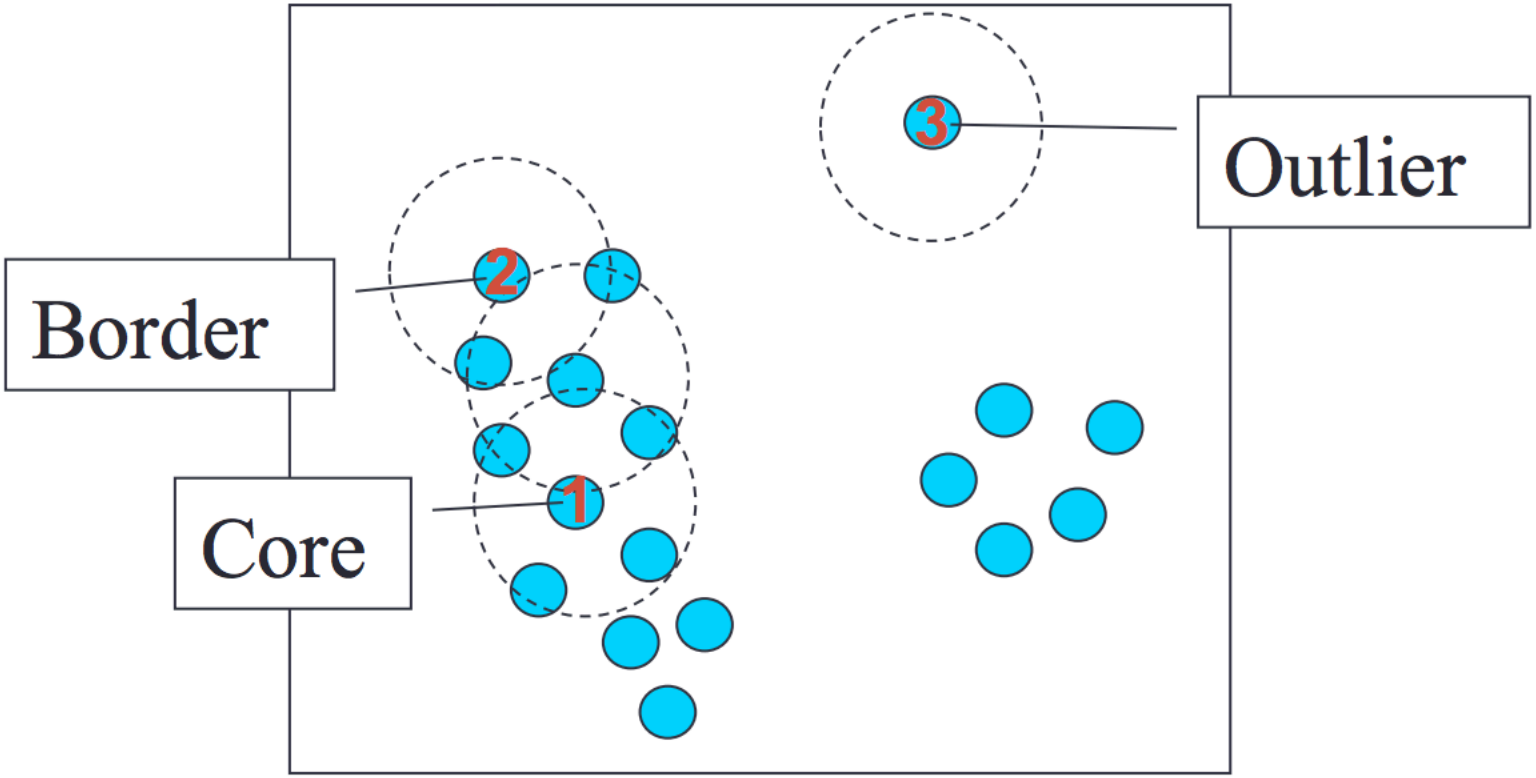
DBSCAN: A Density-Based Spatial Clustering Algorithm

- DBSCAN (M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, KDD'96)
  - Discovers clusters of arbitrary shape: Density-Based Spatial Clustering of Applications with Noise
- 2014 KDD test-of-time award
- A ***cluster*** is defined as a **maximal** set of **density-connected** points
- "Finds core samples of high density and expands clusters from them"
- Definitions
  - Core point
  - Border point
  - outlier



MinPts = 5

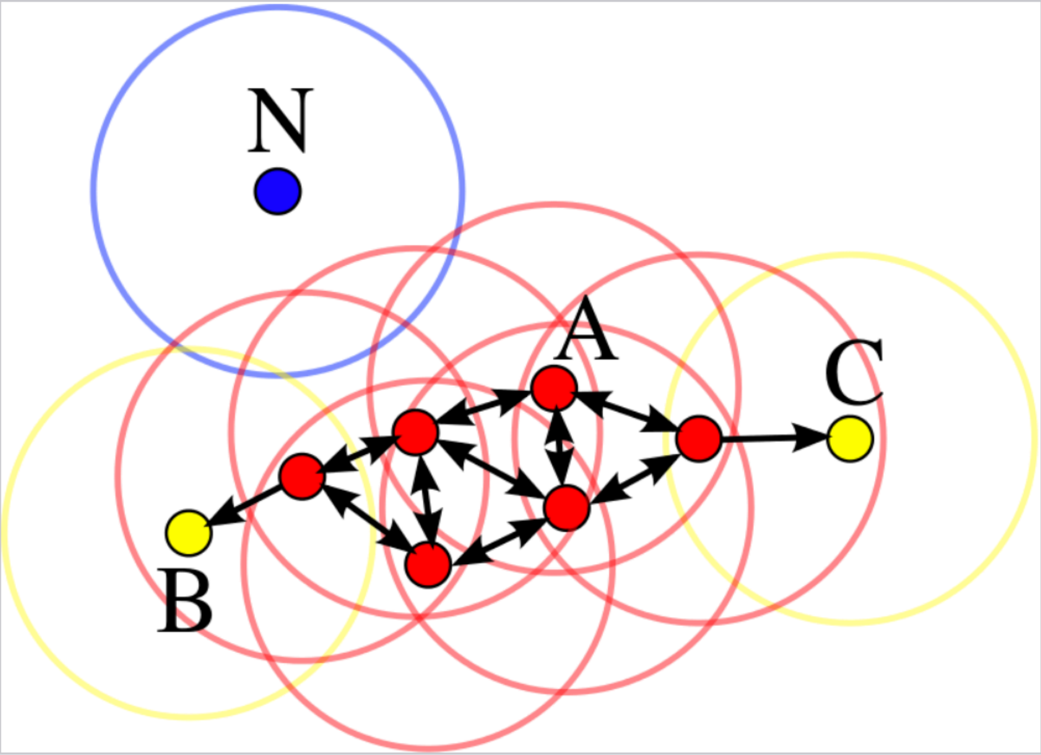
ε = 1 cm



MinPts = 5

Source: Margareta Ackerman, Santa Clara University

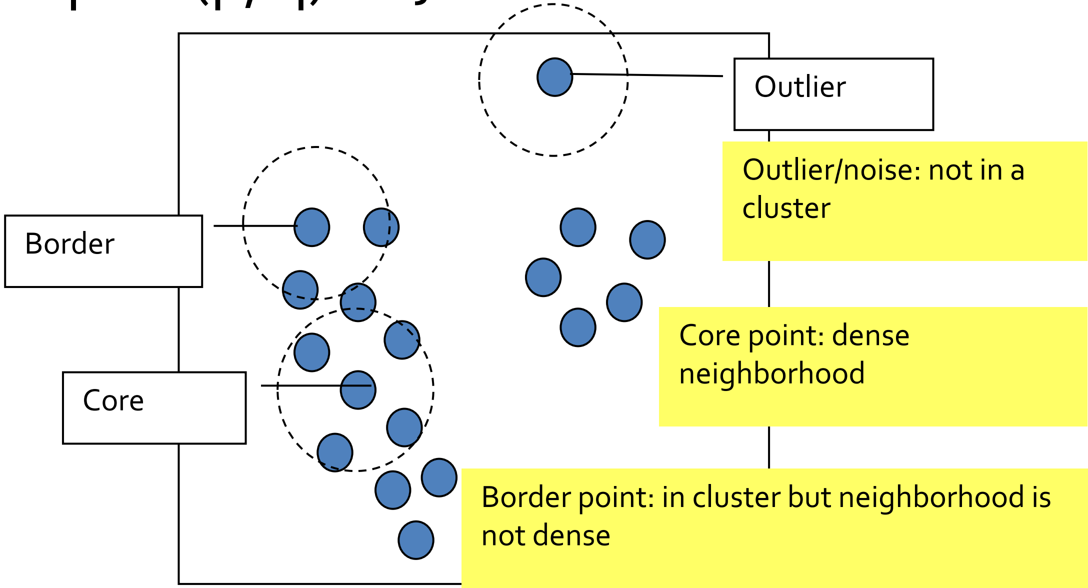
One cluster, one outlier (Wikipedia)



In this diagram, $\mathrm{minPts} = 4$. Point A and the other red points are core points, because the area surrounding these points in an ε radius contain at least 4 points (including the point itself). Because they are all reachable from one another, they form a single cluster. Points B and C are not core points, but are reachable from A (via other core points) and thus belong to the cluster as well. Point N is a noise point that is neither a core point nor directly-reachable.

Definitions

- Two parameters*:*
  - *Eps* (ε): Maximum radius of the neighborhood
  - *MinPts*: Minimum number of points in the Eps-neighborhood of a point
- The Eps(ε)-neighborhood of a point *q*:
  - $N_{Eps}(q)$: {p belongs to D | dist(p, q) ≤ ε}



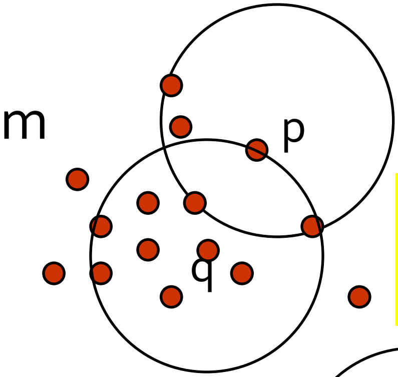
Outlier/noise: not in a cluster

Core point: dense neighborhood

Border point: in cluster but neighborhood is not dense

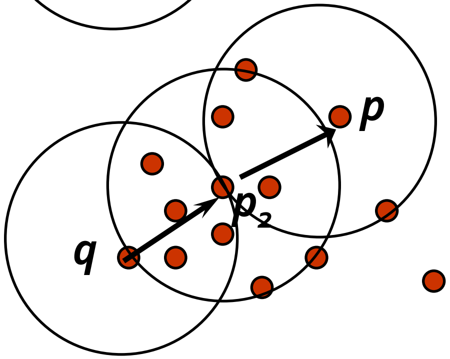
DBSCAN: Density-Reachable and Density-Connected

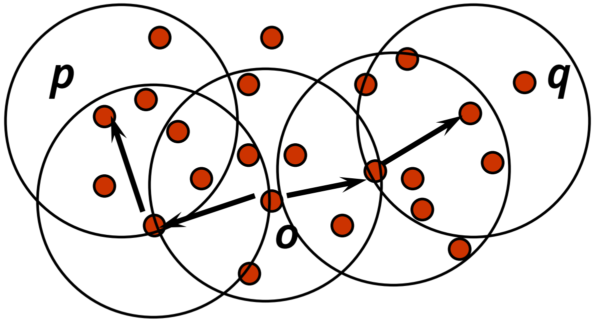
- **Directly density-reachable**:
  - A point *p* is directly density-reachable from a point *q* w.r.t., *MinPts* if
    - *p* belongs to $N_{\varepsilon}(q)$
    - **core point** condition: $|N_{\varepsilon}(q)| \geq MinPts$
- **Density-reachable**:
  - A point *p* is density-reachable from a point *q* w.r.t. *ε*, *MinPts* if there is a chain of points $p_1, \ldots, p_n$, $p_1 = q$, $p_n = p$ such that p_{i+1} is directly density-reachable from p_i
- **Density-connected:**
  - A point *p* is density-connected to a point *q* w.r.t. *ε*, *MinPts* if there is a point *o* such that *p* and *q* are **both** density-reachable from *o* w.r.t. *ε* and *MinPts*



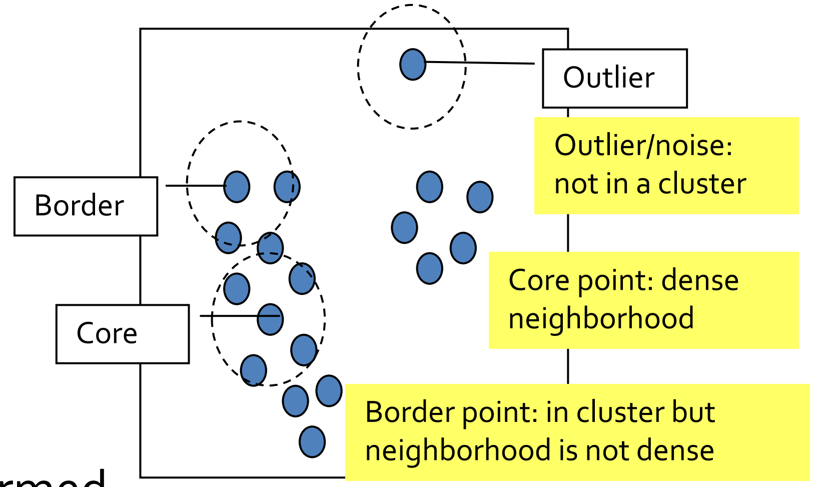
MinPts = 5

ε = 1 cm





DBSCAN: The Algorithm



Outlier/noise: not in a cluster

Core point: dense neighborhood

Border point: in cluster but neighborhood is not dense

- **Algorithm**
  - Specify *ε , MinPts*
  - Arbitrarily select a point *p*
  - Retrieve all points density-reachable from *p*
    - If *p* is a core point, a cluster is formed
    - If *p* is a border point, no points are density-reachable from *p*, and DBSCAN visits the next point of the database
  - Continue until all of the points have been processed

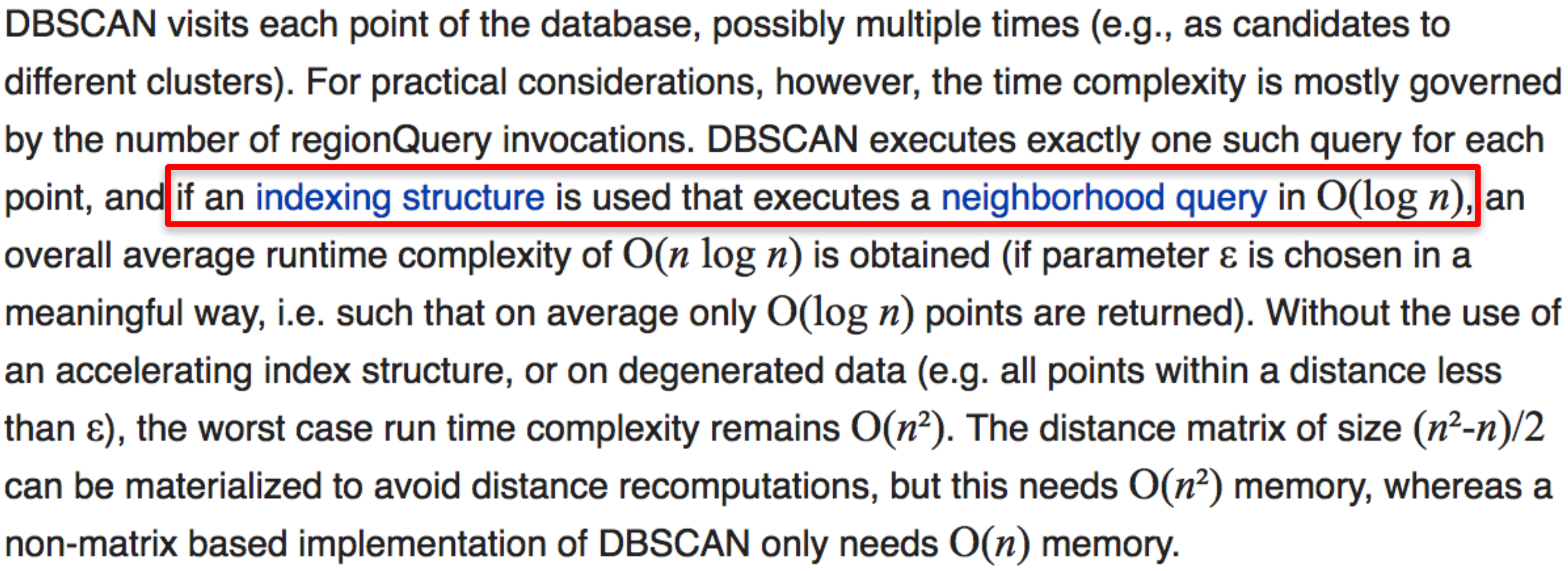
Maximality: if p in a cluster, and q is density-reachable from p, then q is in the same cluster

Connectivity: any pair of points p,q in a cluster are density-connected

DBSCAN: The Algorithm

- **Computational complexity**
  - If a spatial index is used, the computational complexity of DBSCAN is O(nlogn), where n is the number of database objects
  - Otherwise, the complexity is $O(n^2)$

https://en.wikipedia.org/wiki/DBSCAN



DBSCAN visits each point of the database, possibly multiple times (e.g., as candidates to different clusters). For practical considerations, however, the time complexity is mostly governed by the number of regionQuery invocations. DBSCAN executes exactly one such query for each point, and if an indexing structure is used that executes a neighborhood query in $O(\log n)$, an overall average runtime complexity of $O(n \log n)$ is obtained (if parameter ε is chosen in a meaningful way, i.e. such that on average only $O(\log n)$ points are returned). Without the use of an accelerating index structure, or on degenerated data (e.g. all points within a distance less than ε), the worst case run time complexity remains $O(n^2)$. The distance matrix of size $(n^2\text{-}n)/2$ can be materialized to avoid distance recomputations, but this needs $O(n^2)$ memory, whereas a non-matrix based implementation of DBSCAN only needs $O(n)$ memory.

Advantages

- Resistant to noise (outlier class)
- Arbitrary cluster shape is OK

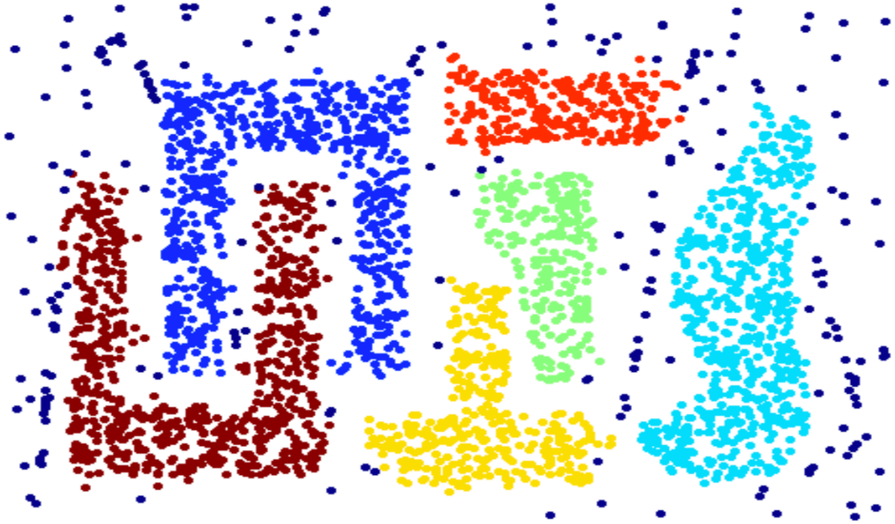


Figure from Ackerman’s site

DBSCAN Is Sensitive to the Setting of Parameters

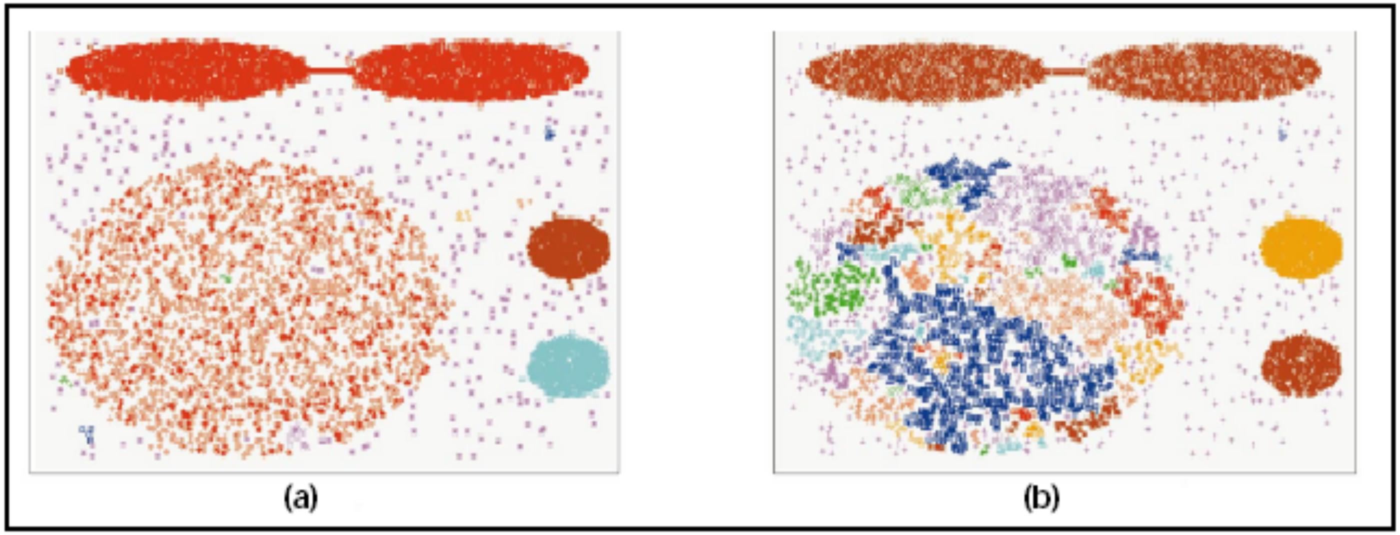


Figure 8. DBScan results for DS1 with MinPts at 4 and Eps at (a) 0.5 and (b) 0.4.

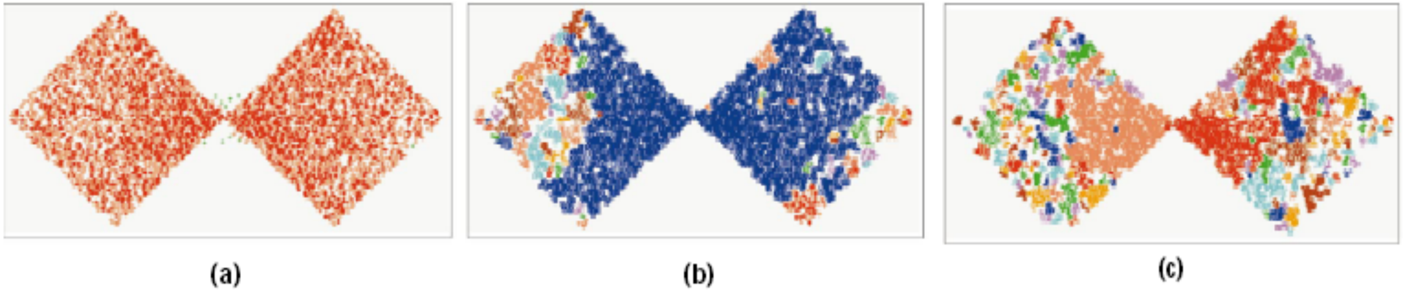
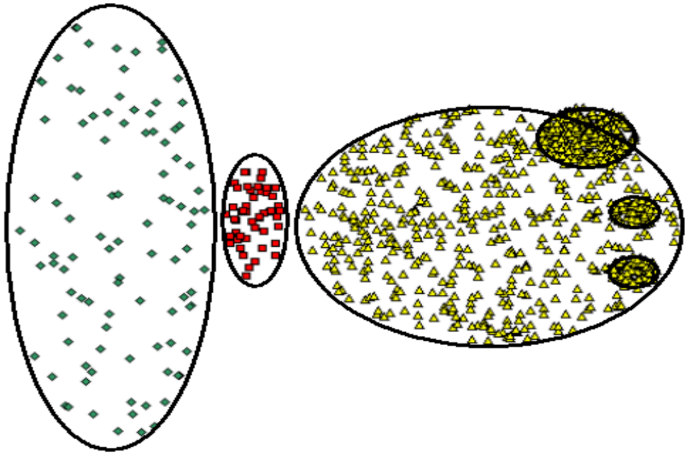
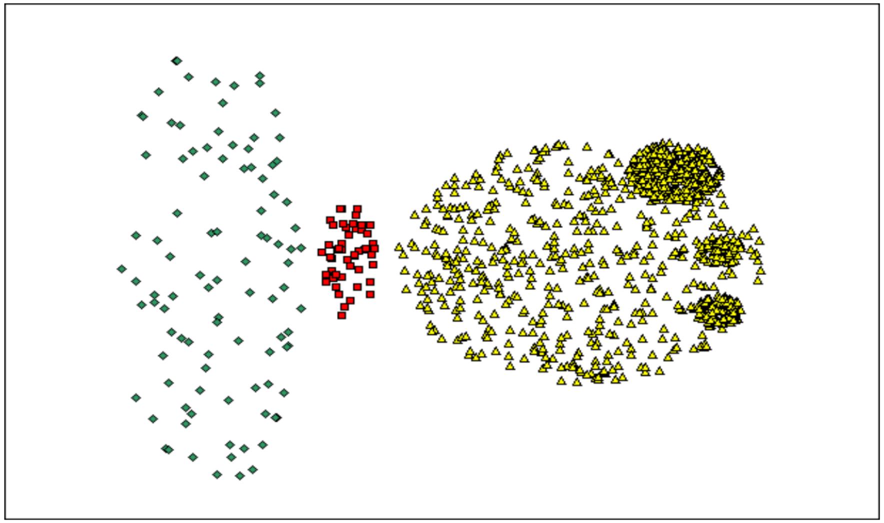


Figure 9. DBScan results for DS2 with MinPts at 4 and Eps at (a) 5.0, (b) 3.5, and (c) 3.0.

Ack. Figures from G. Karypis, E.-H. Han, and V. Kumar, *COMPUTER*, 32(8), 1999



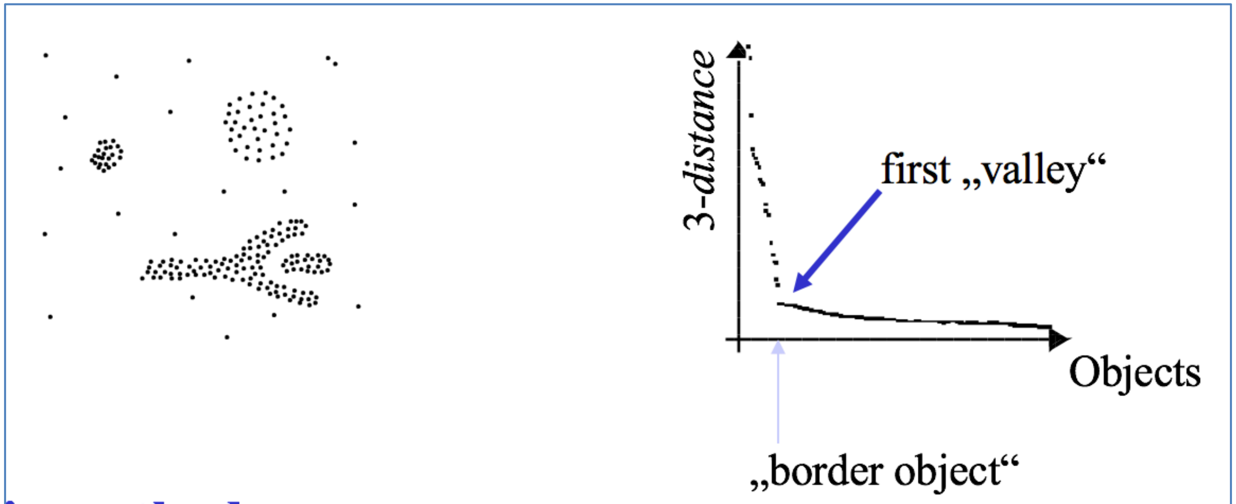
Original Points



(ε =9.92, MinPts=4)

Parameter estimation

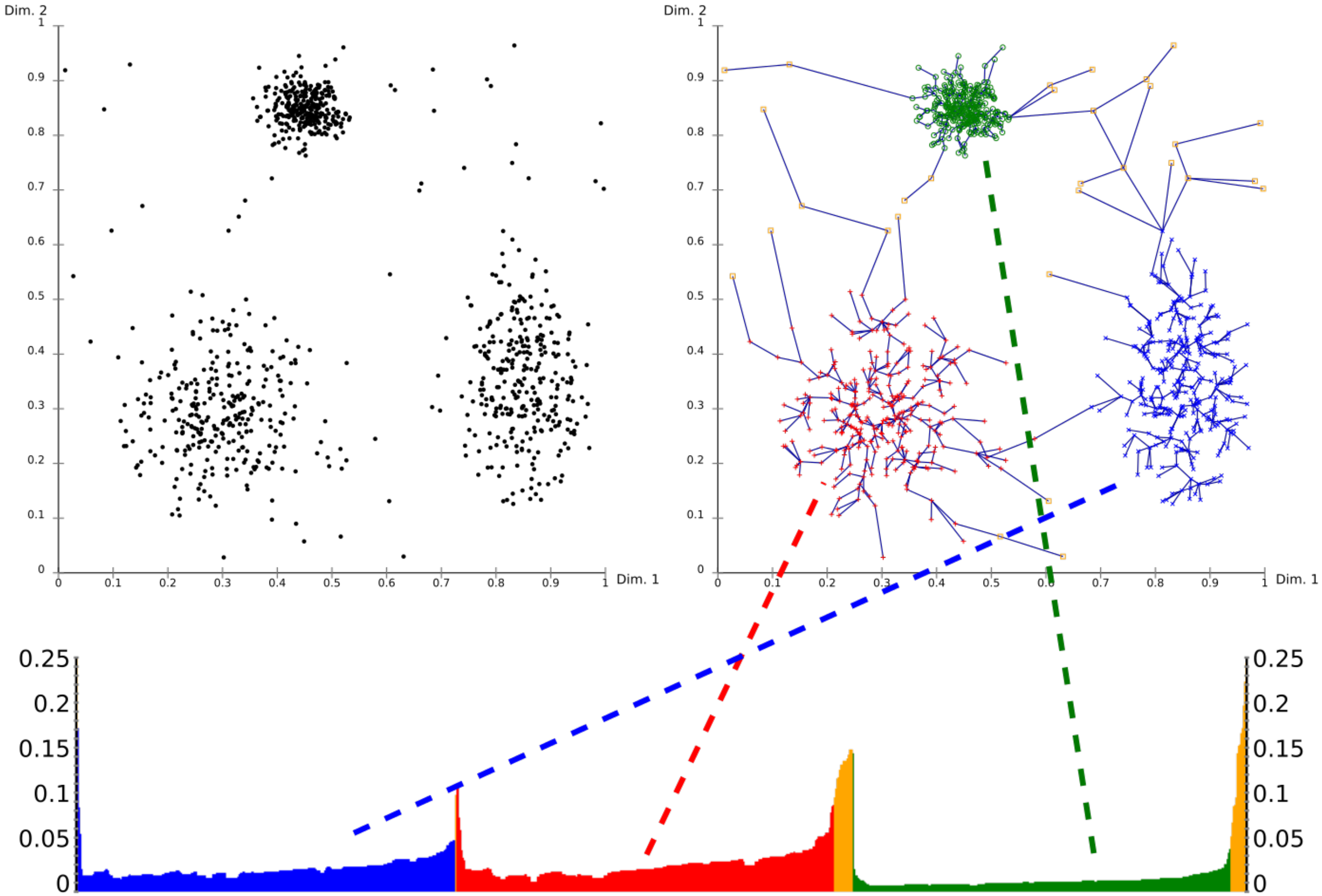
- How to choose MinPts, Eps?
- For MinPts, guess
- Heuristic: use the radius of the *least dense cluster* that you will tolerate
- Plot distance to k'th nearest neighbor and look for a "knee" in the curve



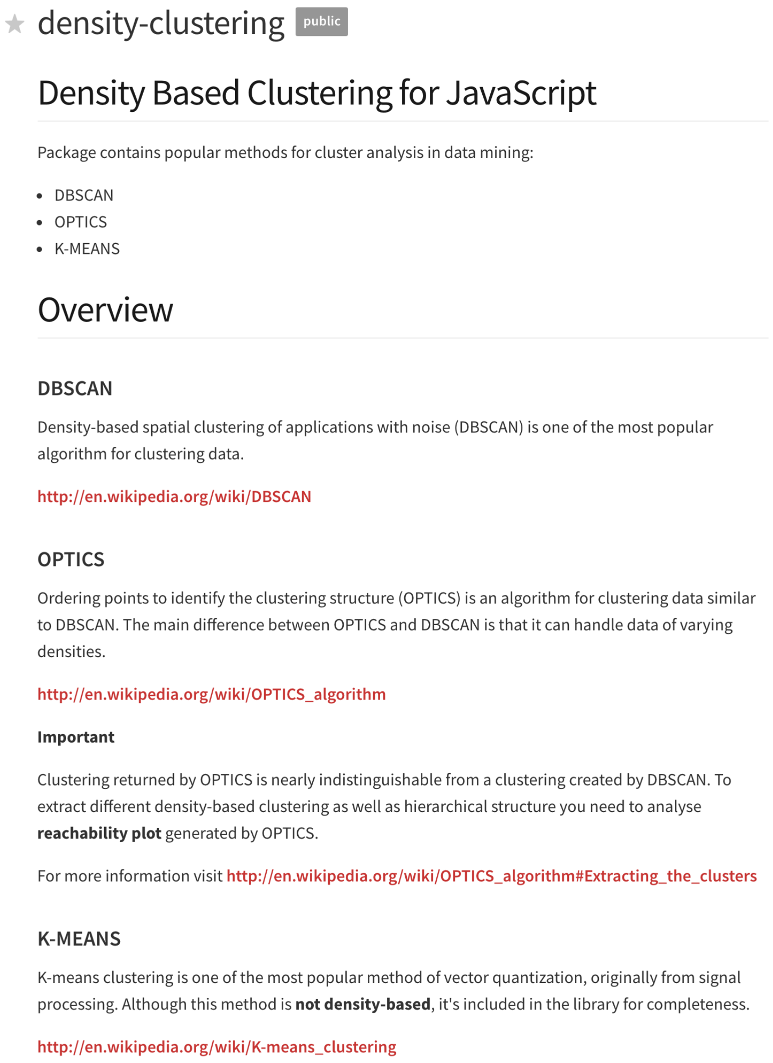
OPTICS: A Cluster-Ordering Method (1999)

- OPTICS: Ordering Points To Identify the Clustering Structure (Ankerst et al., SIGMOD 1999)
- According to Wikipedia:
  - Overcome a major limitation of DBSCAN: handling clusters of varying density (implicitly, variable ε)
  - When building a new cluster, start with a seed and then find and output points in order by reachability
- Good for both automatic and interactive cluster analysis, including finding intrinsic clustering structure
- Can be represented graphically or using visualization techniques

Data, spanning tree, and reachability plot (Wikipedia)



For fun



★ density-clustering public

Density Based Clustering for JavaScript

Package contains popular methods for cluster analysis in data mining:

- DBSCAN
- OPTICS
- K-MEANS

Overview

DBSCAN

Density-based spatial clustering of applications with noise (DBSCAN) is one of the most popular algorithm for clustering data.

http://en.wikipedia.org/wiki/DBSCAN

OPTICS

Ordering points to identify the clustering structure (OPTICS) is an algorithm for clustering data similar to DBSCAN. The main difference between OPTICS and DBSCAN is that it can handle data of varying densities.

http://en.wikipedia.org/wiki/OPTICS_algorithm

Important

Clustering returned by OPTICS is nearly indistinguishable from a clustering created by DBSCAN. To extract different density-based clustering as well as hierarchical structure you need to analyse **reachability plot** generated by OPTICS.

For more information visit http://en.wikipedia.org/wiki/OPTICS_algorithm#Extracting_the_clusters

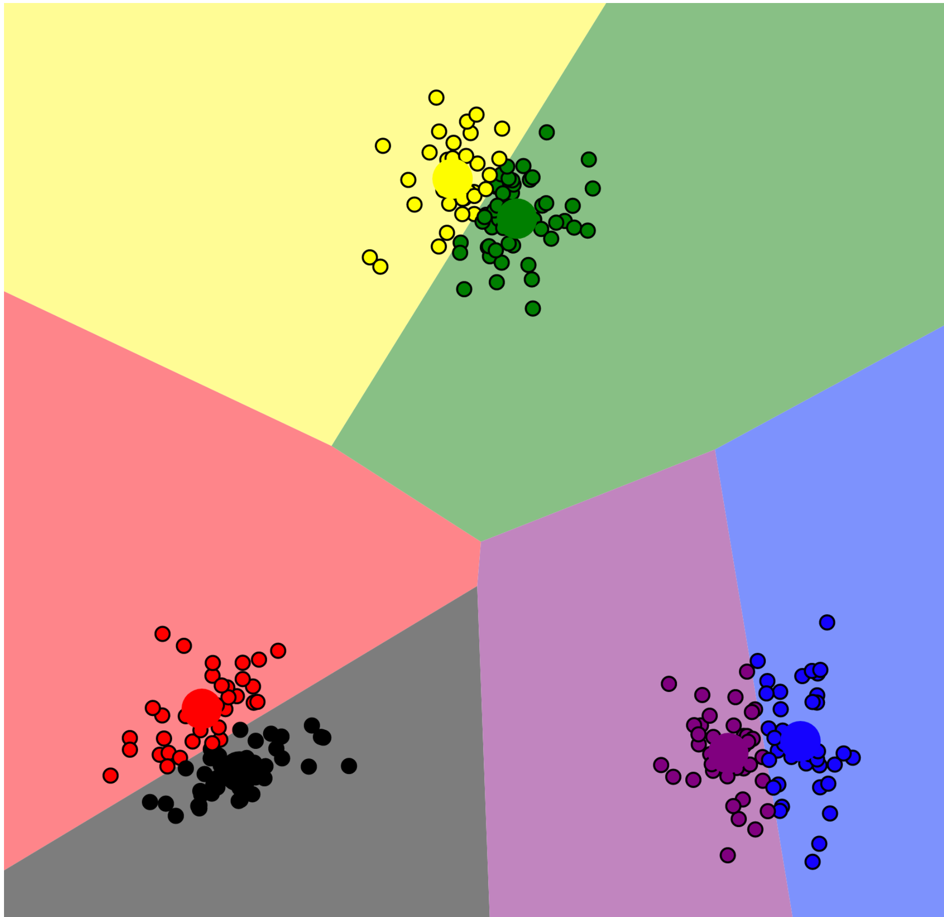
K-MEANS

K-means clustering is one of the most popular method of vector quantization, originally from signal processing. Although this method is **not density-based**, it's included in the library for completeness.

http://en.wikipedia.org/wiki/K-means_clustering

For more fun

- https://www.naftaliharris.com/blog/visualizing-k-means-clustering/
- DBSCAN too



References: (III) Density-based Methods

- M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases. KDD'96
- W. Wang, J. Yang, R. Muntz, STING: A Statistical Information Grid Approach to Spatial Data Mining, VLDB’97
- R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. SIGMOD’98
- A. Hinneburg and D. A. Keim. An Efficient Approach to Clustering in Large Multimedia Databases with Noise. KDD’98
- M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander. Optics: Ordering Points to Identify the Clustering Structure. SIGMOD’99
- M. Ester. Density-Based Clustering. In (Chapter 5) Aggarwal and Reddy (eds.), Data Clustering: Algorithms and Applications . CRC Press. 2014
- W. Cheng, W. Wang, and S. Batista. Grid-based Clustering. In (Chapter 6) Aggarwal and Reddy (eds.), Data Clustering: Algorithms and Applications. CRC Press. 2014