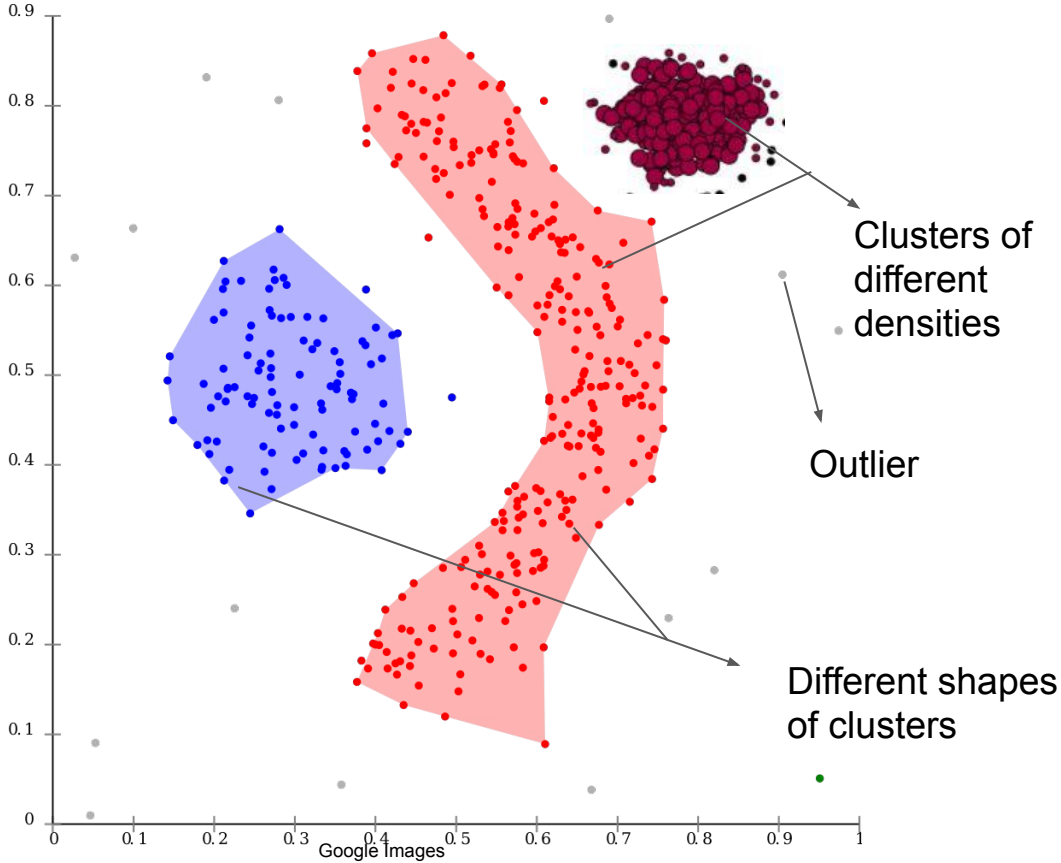
Clustering

OPTICS- Ordering Points to Identify the Clustering Structure



- Density-based Clustering
- Discover Clusters of arbitrary shapes
- Discover Clusters of varying densities
- Dense regions= Clusters
- Outlier points = Noise

OPTICS- Definitions and notations

Ɛ = max distance to consider

MinPts = Number of points required to form the cluster

$N_\varepsilon(p)$ = Neighbourhood of Ɛ

Core point = if atleast MinPts points are within a distance of **Ɛ** to the point

p directly reachable from q if:

1) $p \in N_{Eps}(q)$
2) $|N_{Eps}(q)| \geq MinPts$

Eps = 6mm

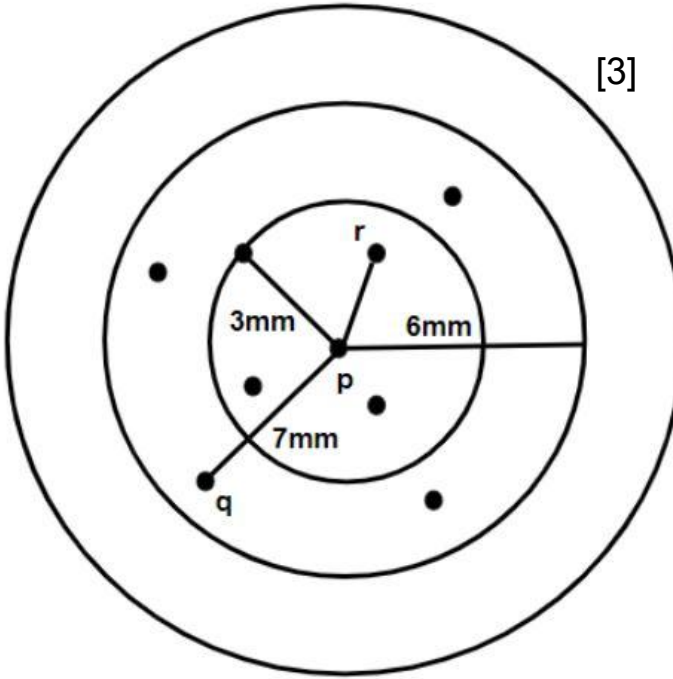
MinPts = 5

Core_Distance(p) = 3mm

Core points = p,r

Reachability_Distance(q,p) = 7mm

Reachability_Distance(r,p) = 3mm



[3]

$$\text{core-dist}_{\varepsilon, MinPts}(p) = \begin{cases} \text{UNDEFINED} & \text{if } |N_\varepsilon(p)| < MinPts \\ MinPts\text{-th smallest distance in } N_\varepsilon(p) & \text{otherwise} \end{cases}$$

$$\text{reachability-dist}_{\varepsilon, MinPts}(o,p) = \begin{cases} \text{UNDEFINED} & \text{if } |N_\varepsilon(p)| < MinPts \\ \max(\text{core-dist}_{\varepsilon, MinPts}(p), \text{dist}(p,o)) & \text{otherwise} \end{cases}$$

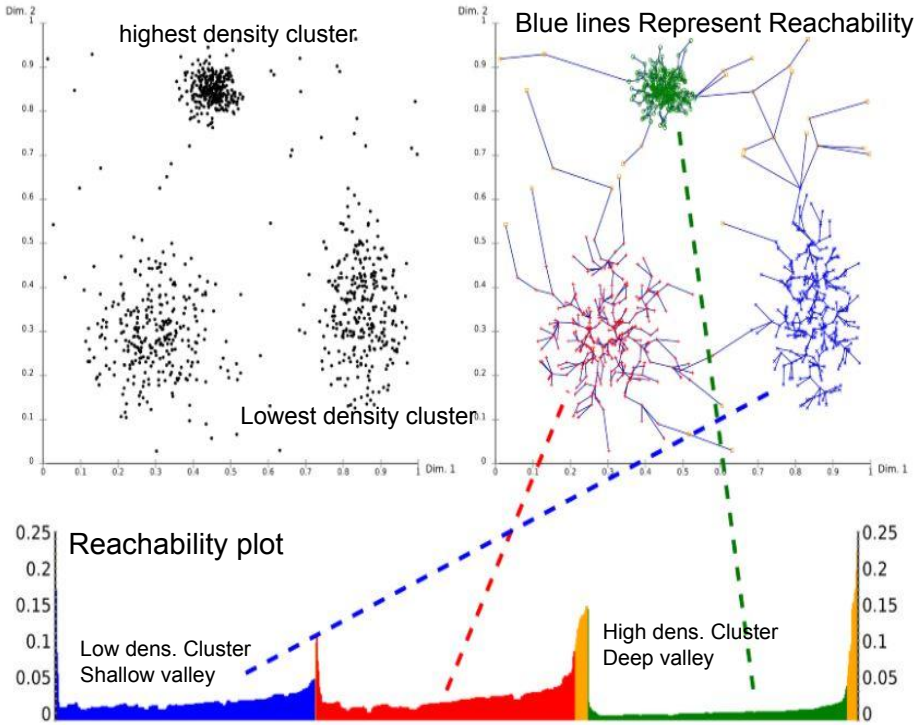
OPTICS - Implementation

OPTICS creates an ordering of data points along with storing the core-distance & reachability distance for objects

Select an arbitrary p from data: if it is not processed then-

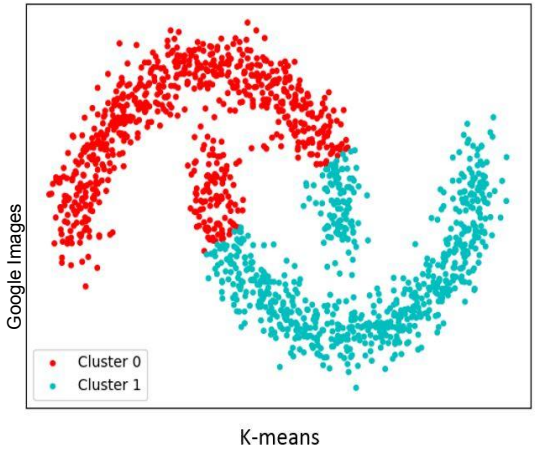
- Retrieve **Nε(p)** & find **core-dist**, **reachability-dist**
  - If p is core point then- iteratively collect its density reachable points wrt **Ɛ** & **MinPts**
    - If new points are directly density reachable, sort them by **reachability-dist**
    - find **Nε** and **core-dist**, add to “order” list along with **reachability-dist**
  - If p is not core point then- add p to “order” list and process next data point

After all the points are processed the core-dist and reachability-dist are plotted in a “Reachability Plot”

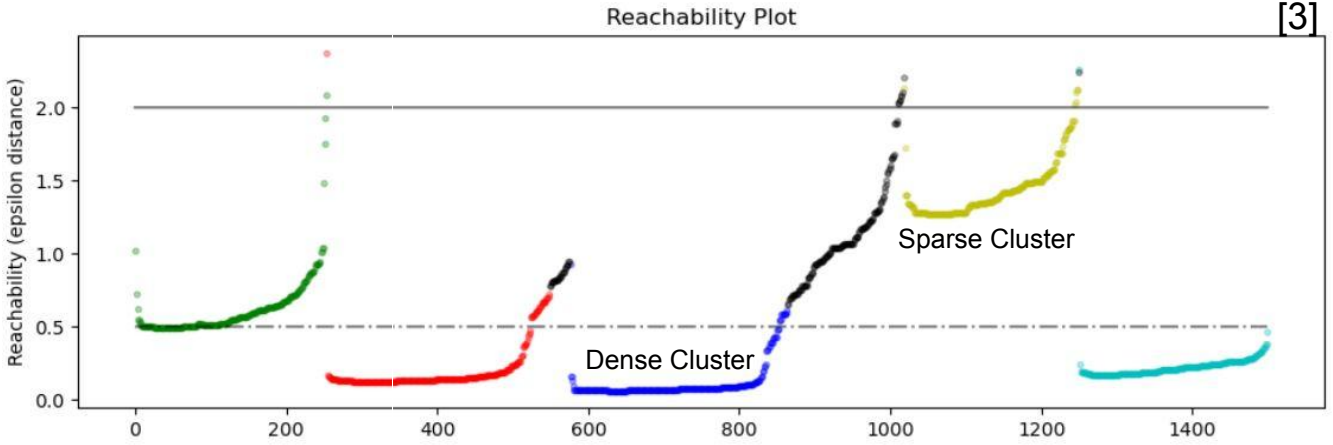


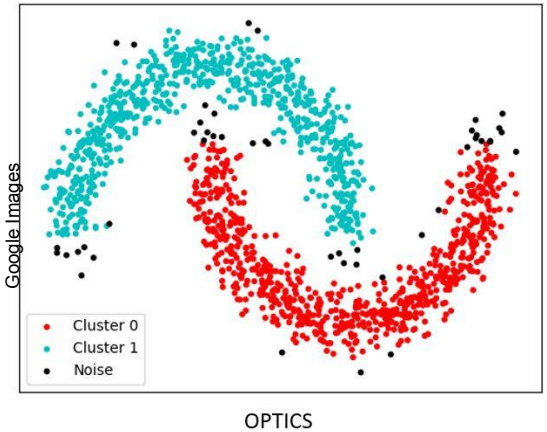
[2]

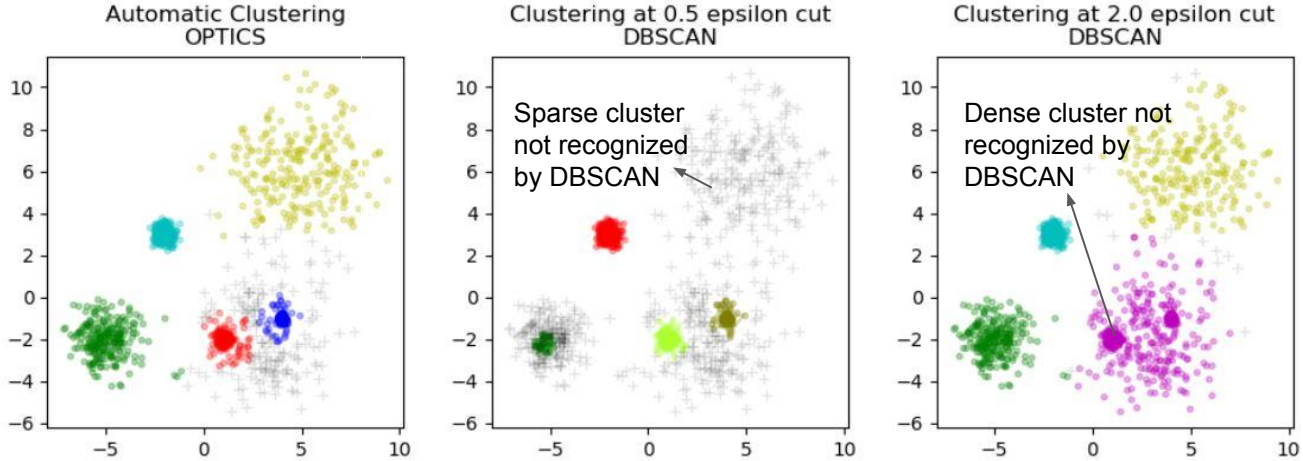
OPTICS vs Other Clustering Algorithms



[3]







OPTICS- Summary

OPTICS is a density based algorithm that uses “core-distance” and “reachability distance” concepts to order data points into clusters

- ✚ Resistant to noise
- ✚ Resistant to arbitrary shapes and varying densities
- ✚ Relatively insensitive to parameter settings
- ▬ Higher Space Complexity than DBSCAN
- ▬ Requires more computational power than DBSCAN
- ▬ Interpretation of results is subjective

Reference materials

[1] Mihael Ankerst; Markus M. Breunig; Hans-Peter Kriegel; Jörg Sander (1999). *OPTICS: Ordering Points To Identify the Clustering Structure*. ACM SIGMOD international conference on Management of data. ACM Press. pp. 49–60.

[2] Martin Ester; Hans-Peter Kriegel; Jörg Sander; Xiaowei Xu (1996). Evangelos Simoudis; Jiawei Han; Usama M. Fayyad (eds.). *A density-based algorithm for discovering clusters in large spatial databases with noise*. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp. 226–231.

[3] Wikipedia contributors. "OPTICS algorithm." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 27 Jul. 2020.

[4] https://medium.com/@xzz201920/optics-d80b41fd042a