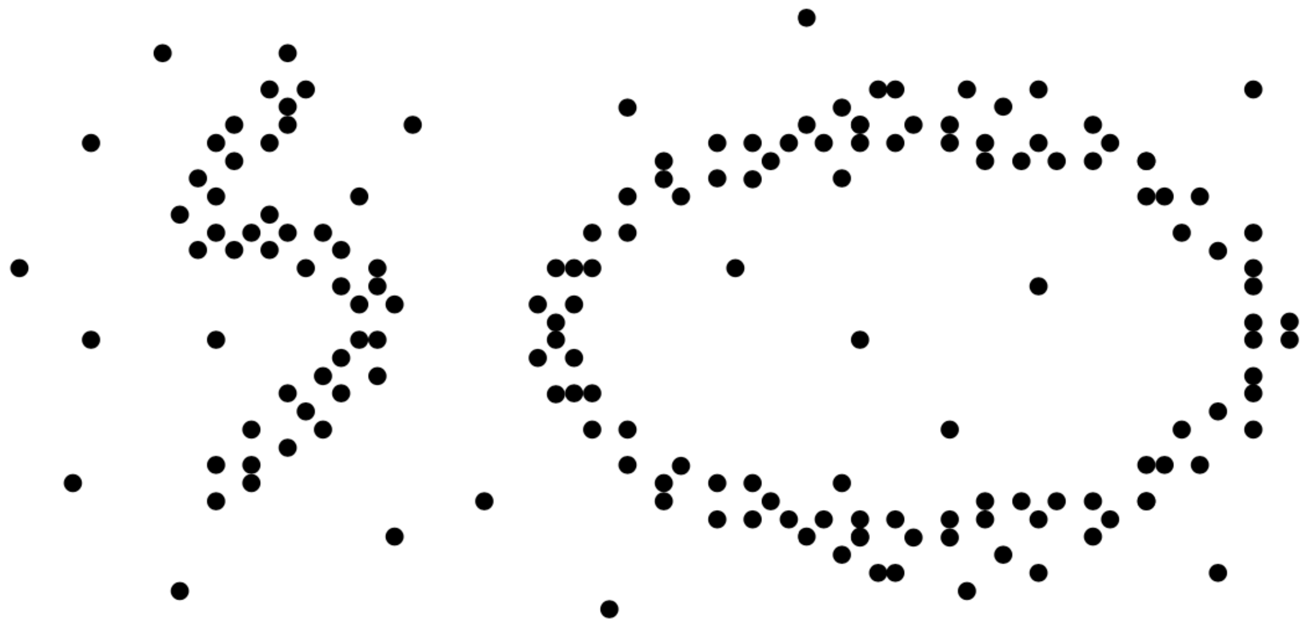
Density-based clustering methods

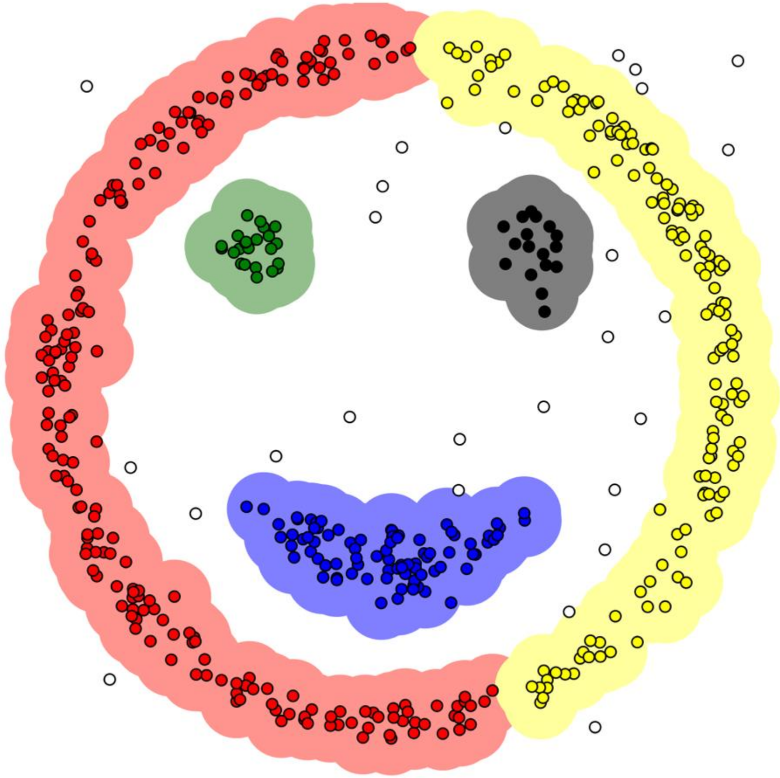
Introduction to Density-based and Grid-based Clustering

- Traditional clustering techniques struggle with non-spherical, arbitrary shapes.
- Example: Clusters with shapes like "S" and ovals
- Partitioning and hierarchical methods are suited for spherical clusters
- Difficulty in identifying clusters with arbitrary shapes
- Noise and outliers are often included in clusters inaccurately

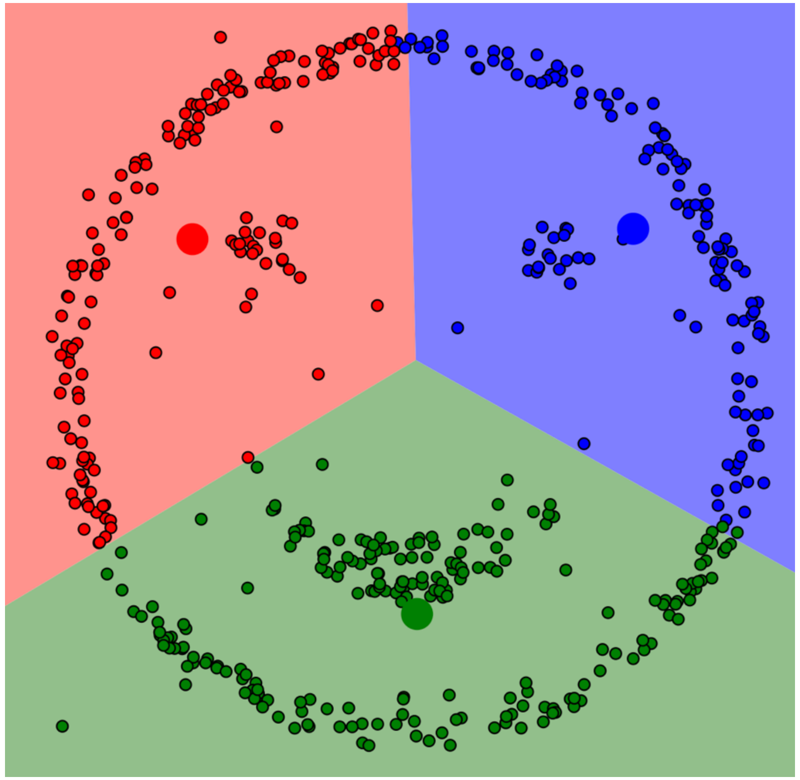


Density Vs. Partitioning based clustering

DBSCAN



***k*Means Clustering**



Links: 1 and 2

Density-Based Clustering Concepts

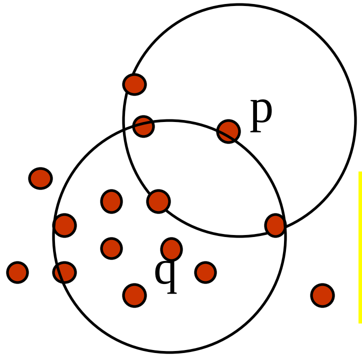
- Key idea: **Model clusters as dense regions separated by sparse regions**
- Capable of discovering non-spherical clusters
- Accurate identification of complex-shaped clusters
- Example: DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

Introduction to DBSCAN

- 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
- DBSCAN: Density-Based Spatial Clustering of Applications with Noise
- Main Ideas:
  - Identify dense regions to form clusters
  - Identifies core objects (dense regions) and forms clusters by connecting core objects and their neighborhoods

DBSCAN: A Density-Based Spatial Clustering Algorithm

- A density-based notion of cluster
  - A cluster is defined as a maximal set of density-connected points
  - 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:
  - NEps(q): {p belongs to D | dist(p, q) ≤ Eps}
- If MinPts = 5, then p will be considered a core point if there are at least 4 other points within its ϵ-radius (making a total of 5 points, including p itself).

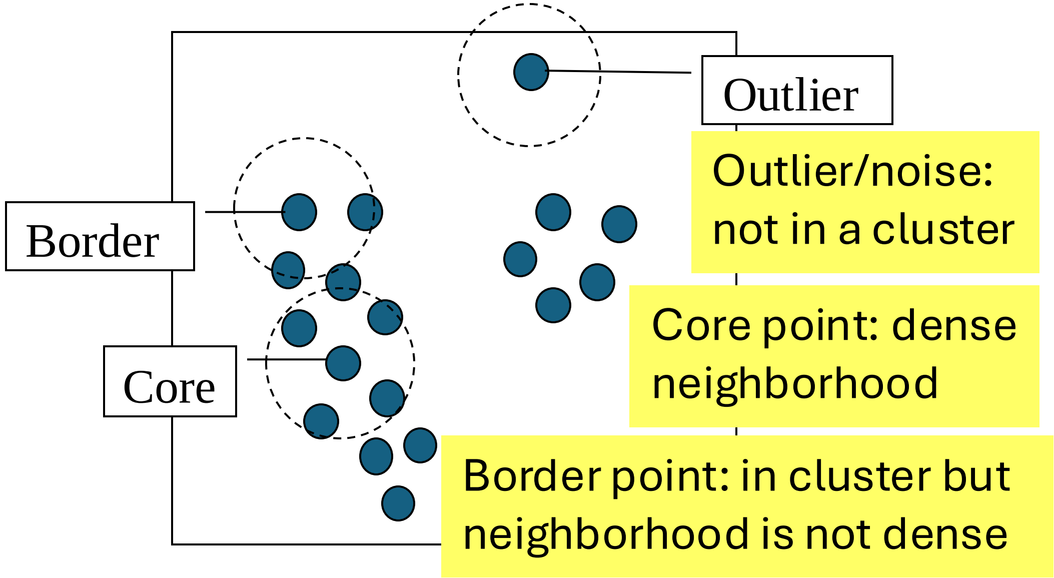


MinPts = 5

Eps = 1 cm

DBSCAN: A Density-Based Spatial Clustering Algorithm

- **Core Object:** An object with at least **MinPts** points in its ε-neighborhood
- Border points are non-core objects that lie within the ε-neighborhood of a core object.
- Outliers, or noise points, are objects that do not belong to the ε-neighborhood of any core object.



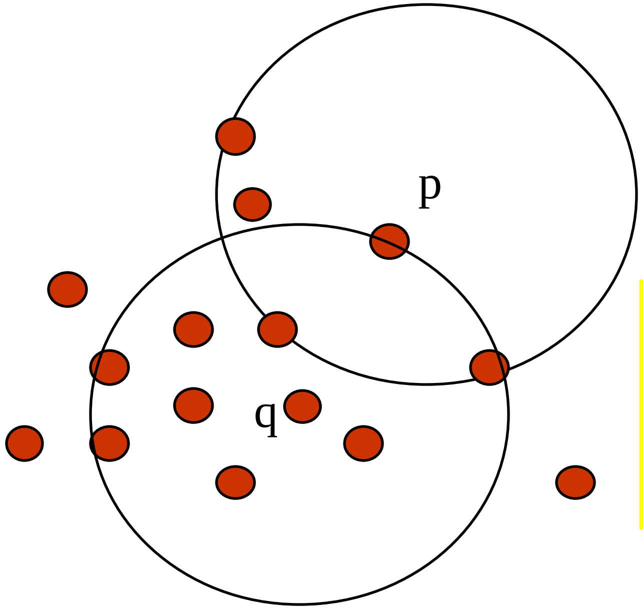
Outlier/noise: not in a cluster

Core point: dense neighborhood

Border point: in cluster but neighborhood is not dense

DBSCAN: Directly Density-Reachable

- A point p is directly density-reachable from a point q w.r.t. Eps (ε), MinPts if
  - p belongs to NEps(q)
  - core point condition: |NEps (q)| ≥ MinPts

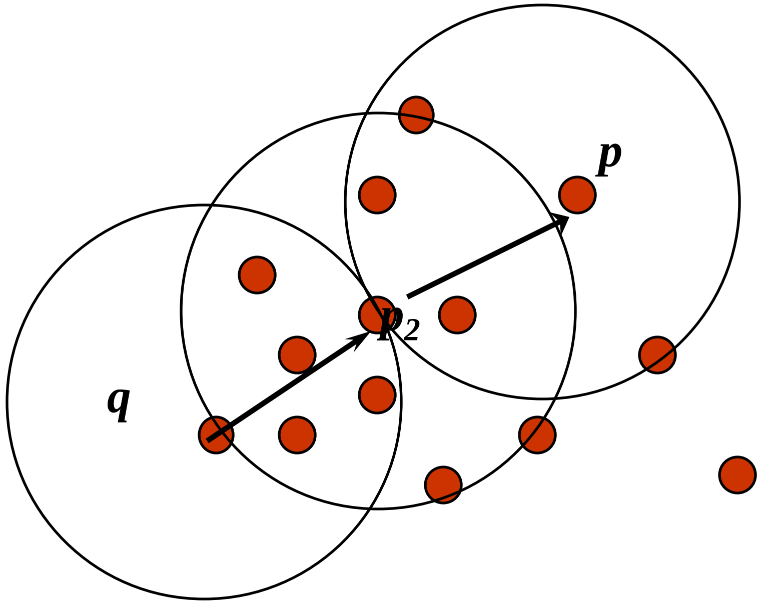


MinPts = 5

Eps = 1 cm

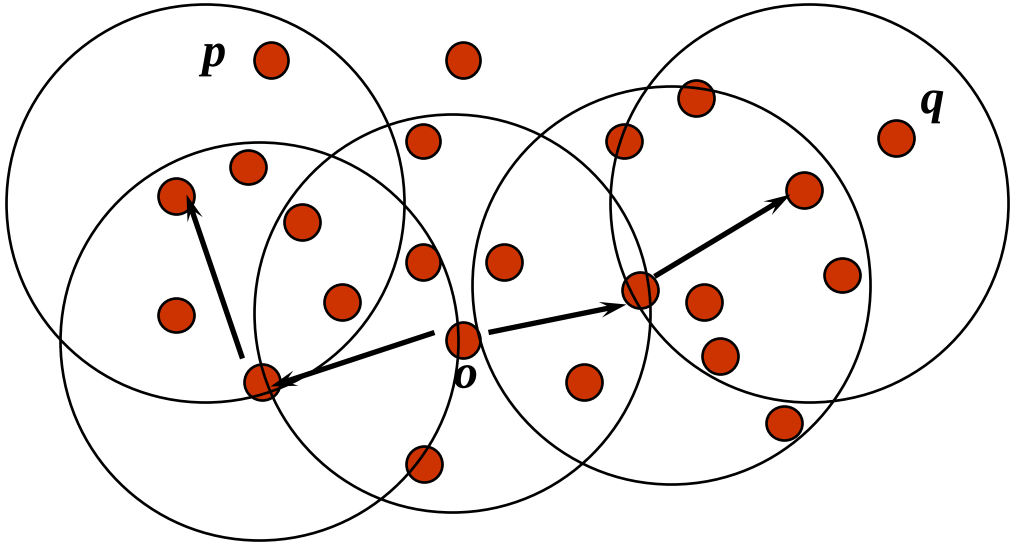
DBSCAN: Density-Reachable

A point p is density-reachable from a point q w.r.t. Eps, MinPts if there is a chain of points $p_1, \dots, p_n, p_1 = q, p_n = \; p$ such that $p_i + 1$ is directly density-reachable from p_i

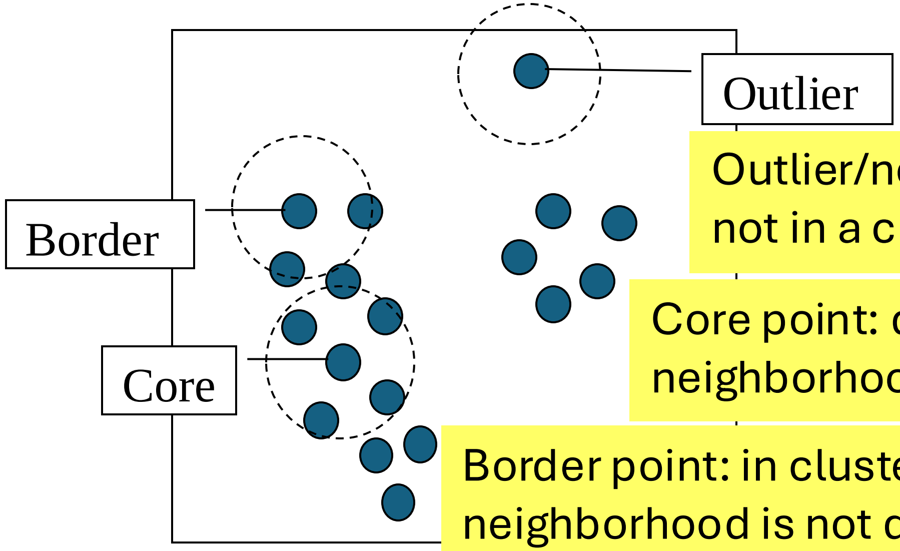


DBSCAN: Density-Connected

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: The Algorithm



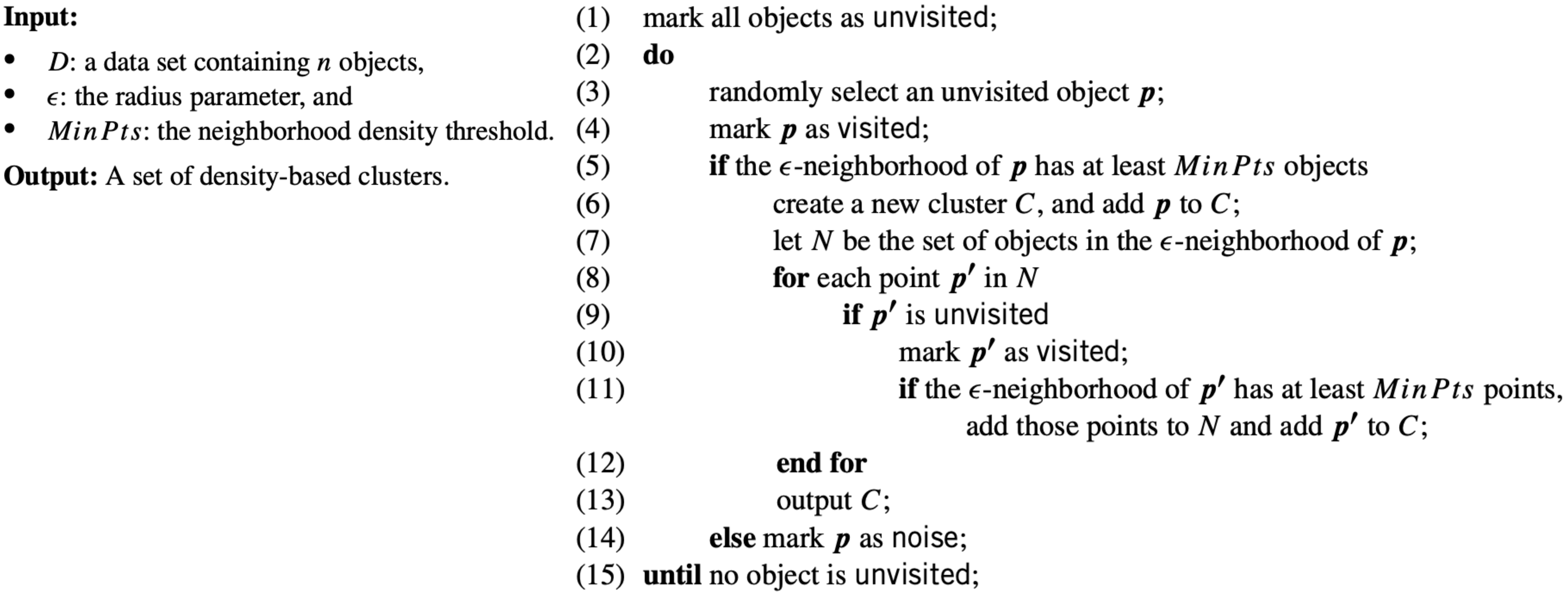
Outlier/noise: not in a cluster

Core point: dense neighborhood

Border point: in cluster but neighborhood is not dense

- Algorithm
  - Arbitrarily select a point p
  - Retrieve all points density-reachable
      - from p w.r.t. Eps and MinPts
    - If p is a core point, a cluster is formed
    - If p is a border point, no points are directly density-reachable from p, and DBSCAN visits the next point of the database
  - Continue the process until all of the points have been processed
- Computational complexity
  - If a spatial index is used, the computational complexity of DBSCAN is $O(n \log n)$, where n is the number of database objects
  - Otherwise, the complexity is O(n^2)

DBSCAN Algorithm: A closer Look



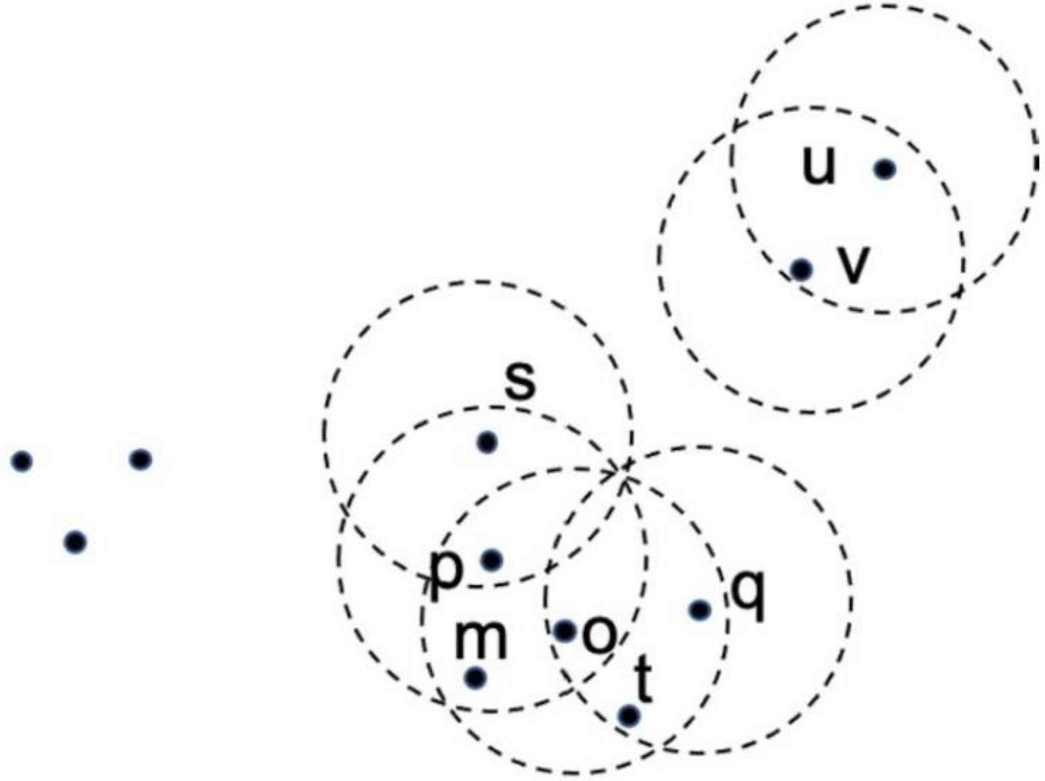
Input:

- D: a data set containing n objects,
- ϵ: the radius parameter, and
- $MinPts$: the neighborhood density threshold.

Output: A set of density-based clusters.

```
mark all objects as unvisited;
do
    randomly select an unvisited object $p$;
    mark $p$ as visited;
    if the $\epsilon$-neighborhood of $p$ has at least $MinPts$ objects
        create a new cluster $C$, and add $p$ to $C$;
        let $N$ be the set of objects in the $\epsilon$-neighborhood of $p$;
        for each point $p'$ in $N$
            if $p'$ is unvisited
                mark $p'$ as visited;
                if the $\epsilon$-neighborhood of $p'$ has at least $MinPts$ points,
                    add those points to $N$ and add $p'$ to $C$;
        end for
        output $C$;
    else mark $p$ as noise;
until no object is unvisited;
```

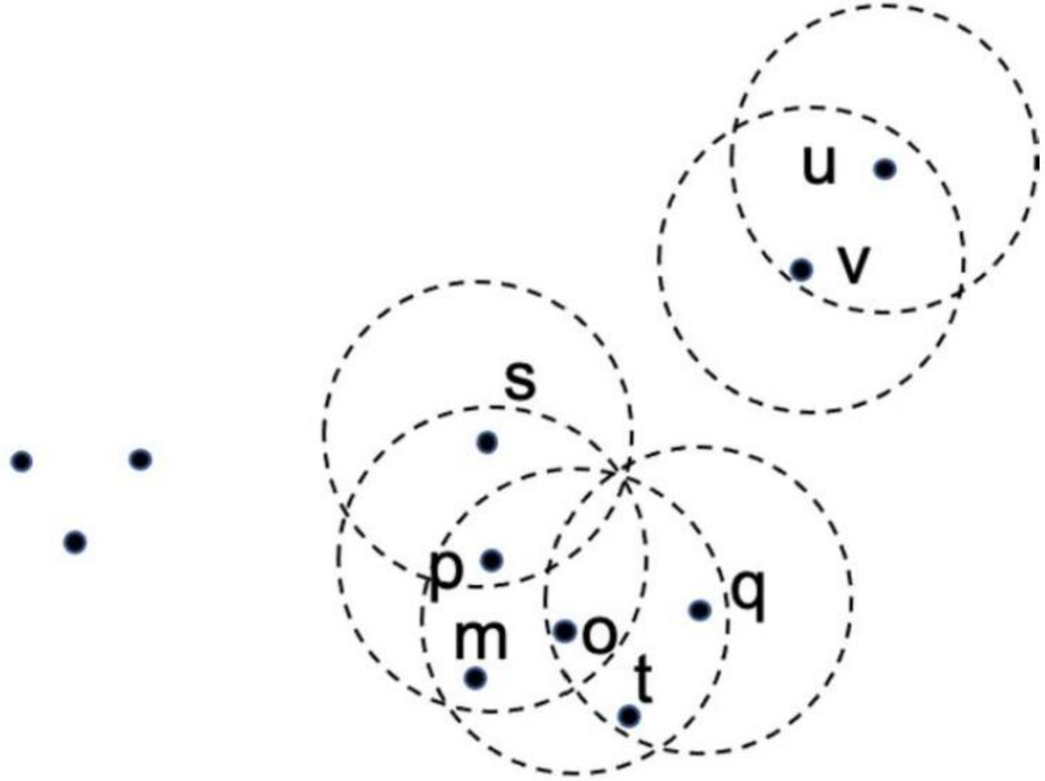
DBSCAN Example: Identifying Core Objects



MinPts= 3

- Of the labeled objects, p, m, o, q, and t are core objects, since each of the ϵ-neighborhoods (dashed circles in the figure) of them contains at least three objects.
- Objects p and o are ϵ-reachable, so are o and q.
- Thus p and q are density-connected.
- **Note**: When calculating if MinPts condition is met, we include the object itself.
  - **Example**: N(q) = {t, o} so it meets the requirement because we count 3 (q, t and o).

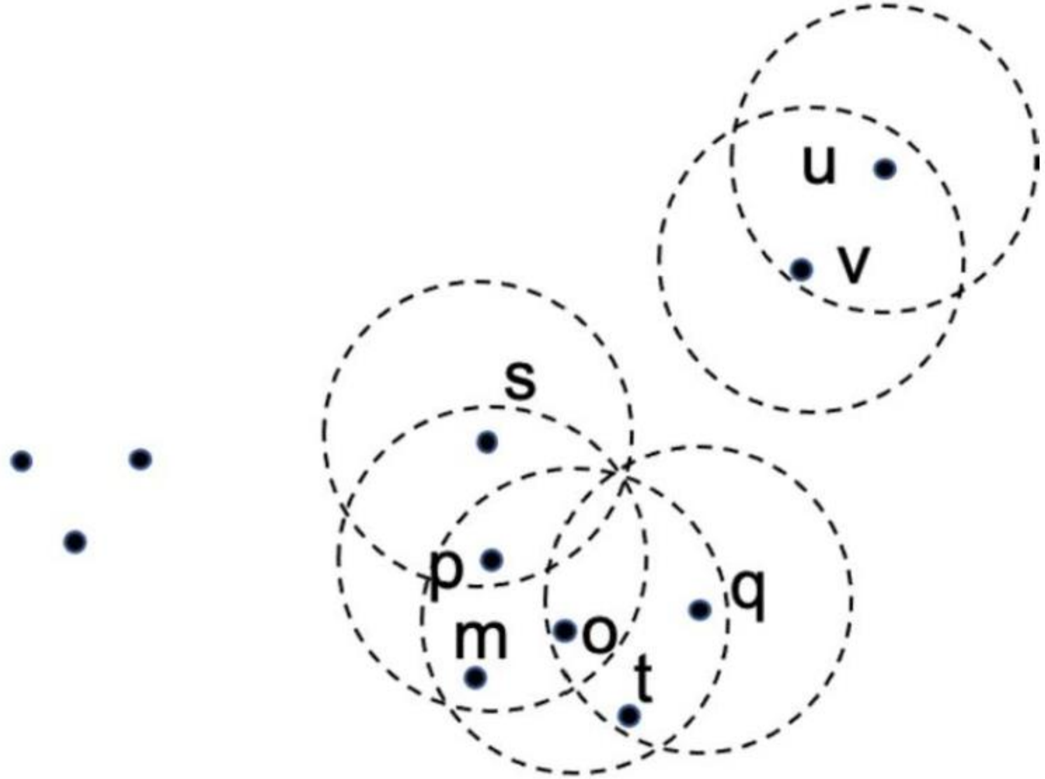
DBSCAN Example



MinPts= 3

Core objects p, m, o, q, and t form a cluster, since each two among them are density-connected and no other core objects can be added into this group so that the pairwise density-connectivity is maintained.

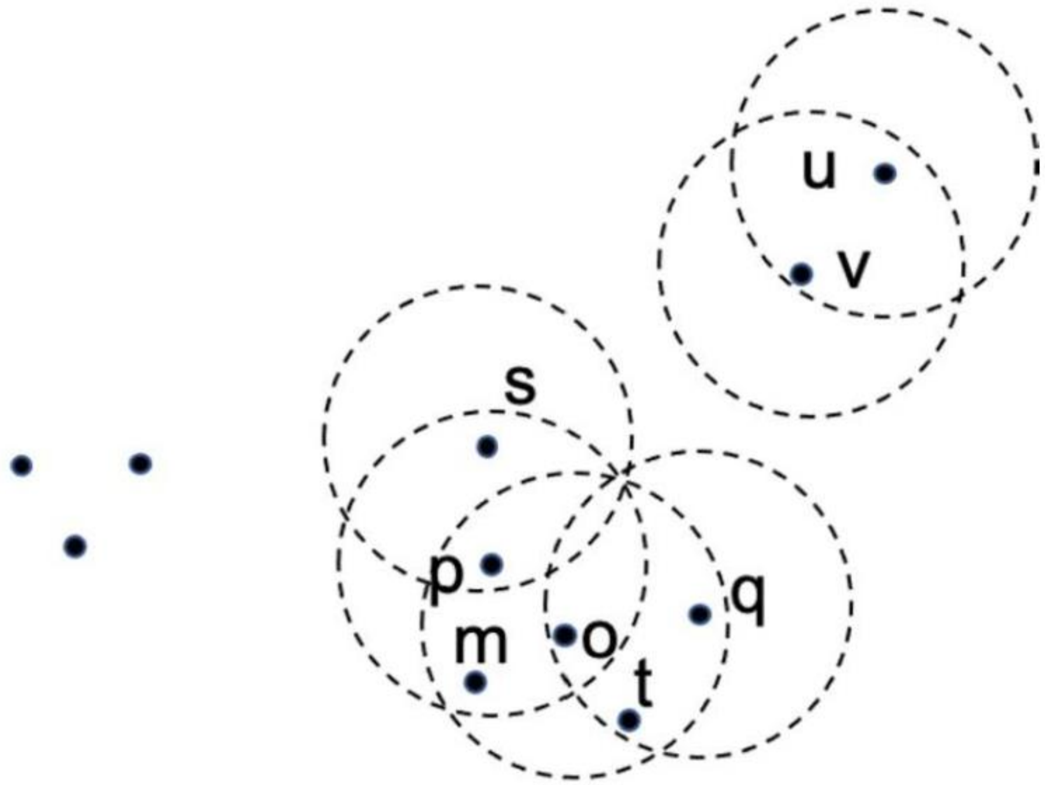
DBSCAN Example



MinPts= 3

Object s is not a core object, since the ϵ-neighborhood of s contains only two objects. However, s is in the ϵ-neighborhood of core object p, thus s is a border object.

DBSCAN Example



MinPts= 3

Objects u and v are not core objects, and they do not belong to the ϵ-neighborhood of any core objects. Thus they are outliers.

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.

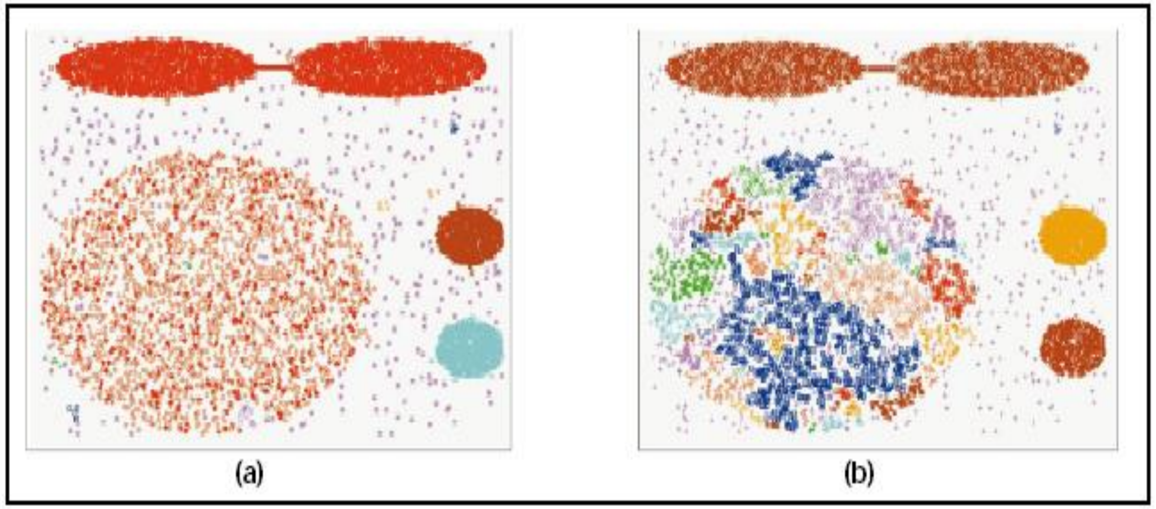
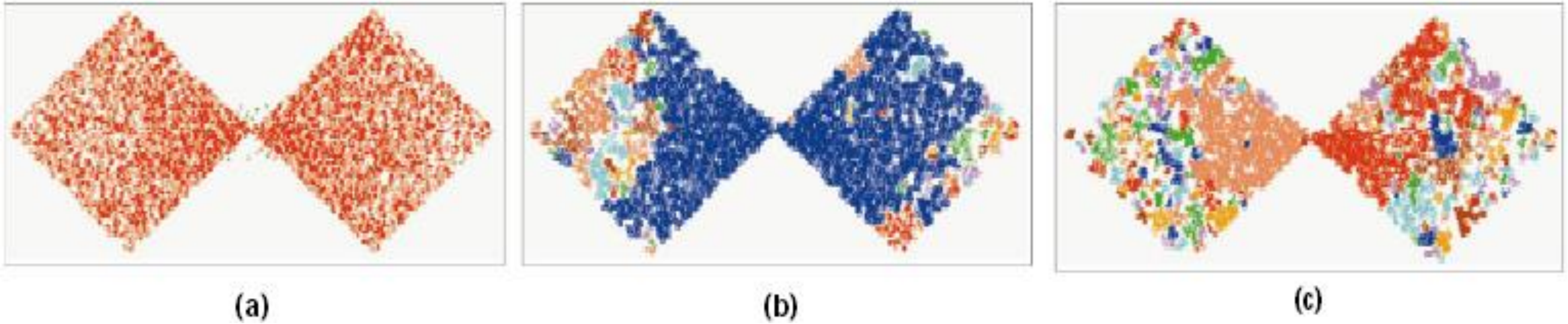


Figure 9. DBScan results for DS2 with MinPts at 4 and Eps at (a) 5.0, (b) 3.5, and (c) 3.0.



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