BRACE YOURSELF
CLUSTERING IS COMING

IT'S A SECRET ART PASSED DOWN 1,000 GENERATIONS
K-MEANS
memegenerator.net

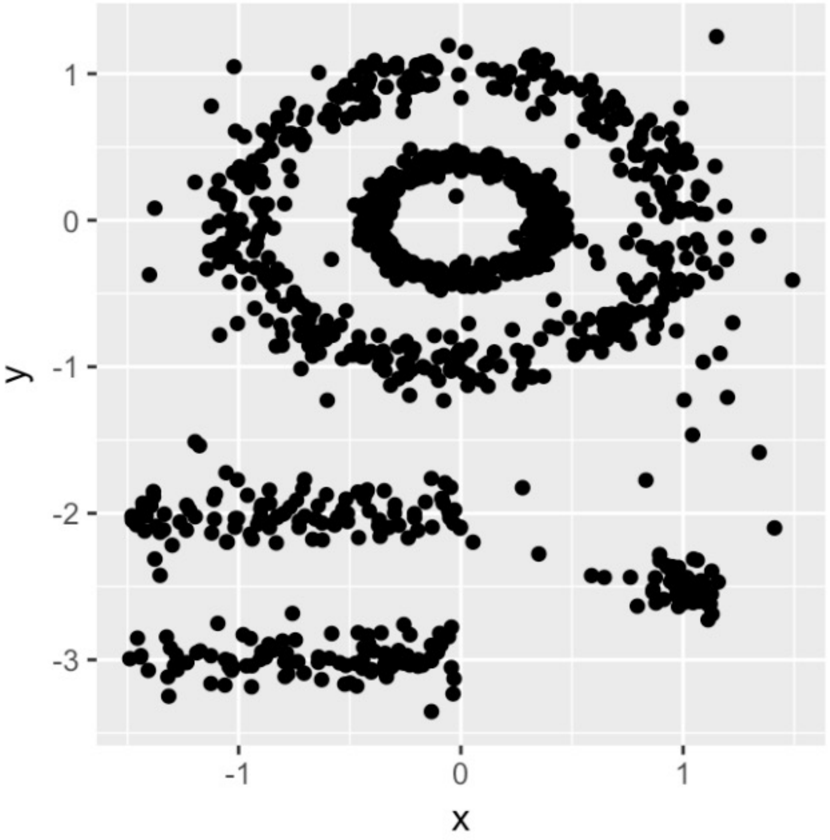
“We were able to form a model to predict the personality of every single adult in the United States of America – 220 million people.” – Alexander Nix, CEO Cambridge Analytica

“For a highly neurotic and conscientious audience, the threat of a burglary.”

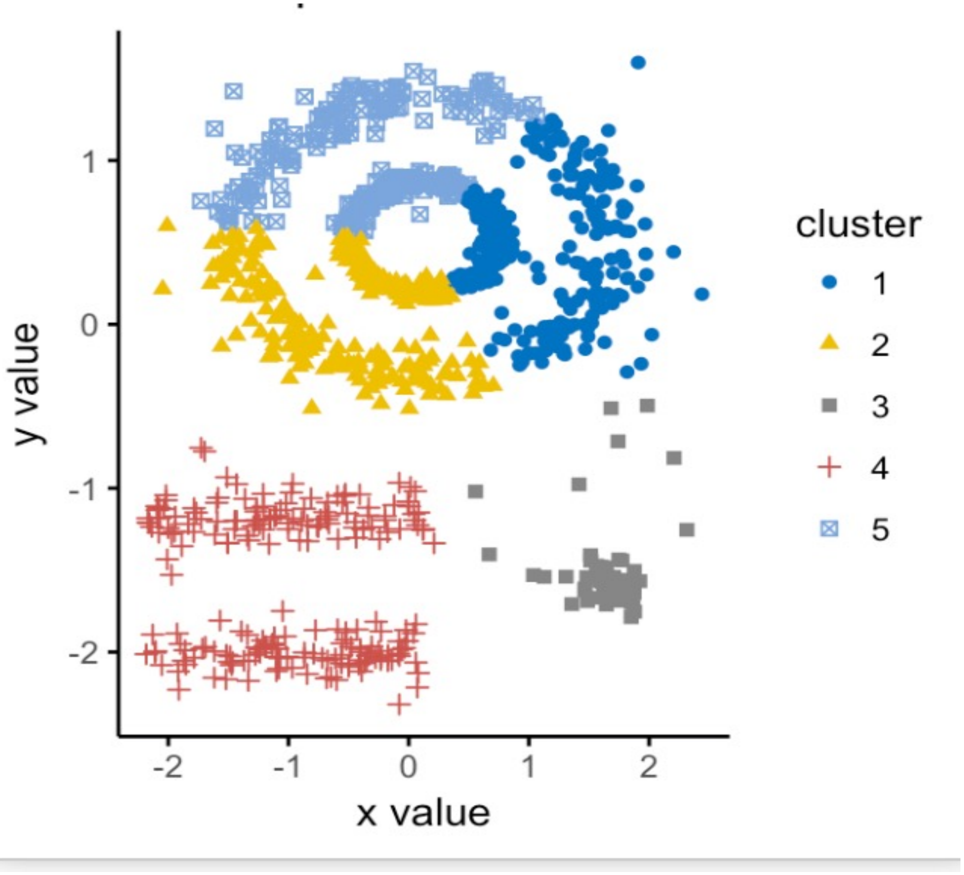
“Conversely, for a closed and agreeable audience. People who care about tradition, and habits, and family.”

“The model of the voter as a bundle of psychological vulnerabilities to be carefully exploited reduces people to mathematical inputs.” –Adrian Chen, *The New Yorker*

Density Based Clustering



Original Points



K-means, K=5

Density-Based Clustering

- Locates regions of high density that are separated by regions of low density

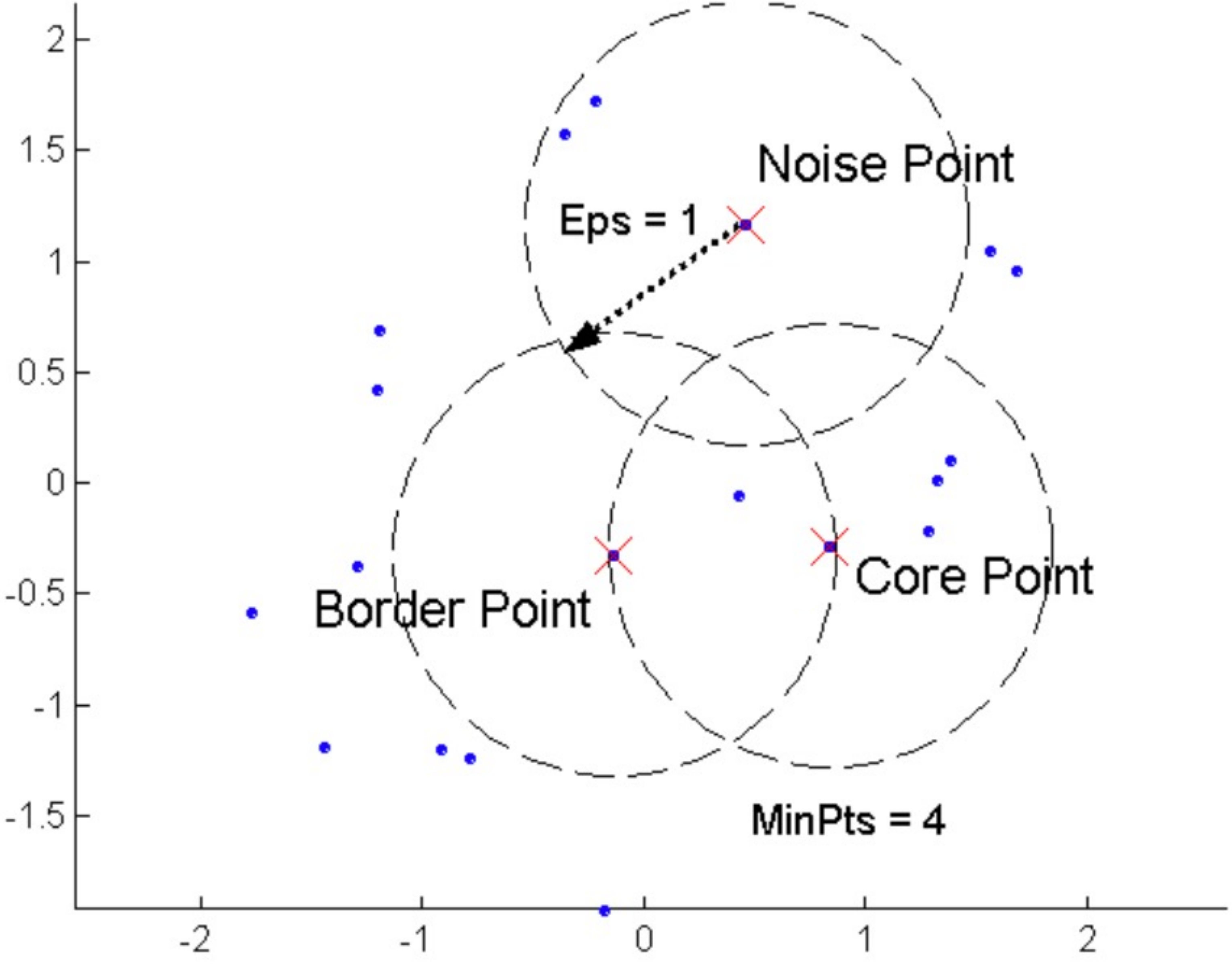
- **Center-based density**: density is calculated for a particular point in the data set by counting the number of points within a specified radius, Eps, of that point (the count includes the point itself).

Classifying Points

- **Core points**: Points in the interior of a dense region. If the number of points within Eps of this point meets a certain threshold, $MinPts$, this point is a core point.

- **Border points**: Points on the edge of a dense region. A point that is not a core point, but falls within the neighborhood (within Eps) of a core point.

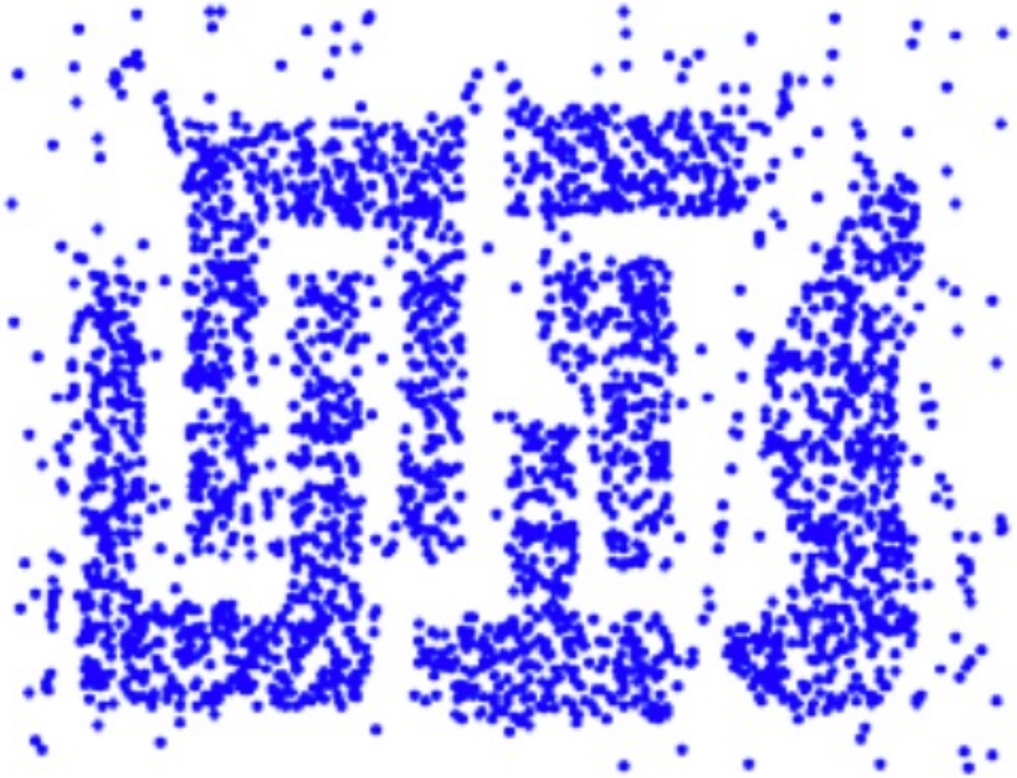
- **Noise points**: Points in sparse regions. Any point that is neither a core point, nor a border point.


Noise Point
Eps = 1
Border Point
Core Point
MinPts = 4
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-1
0
1
2

DBSCAN Algorithm

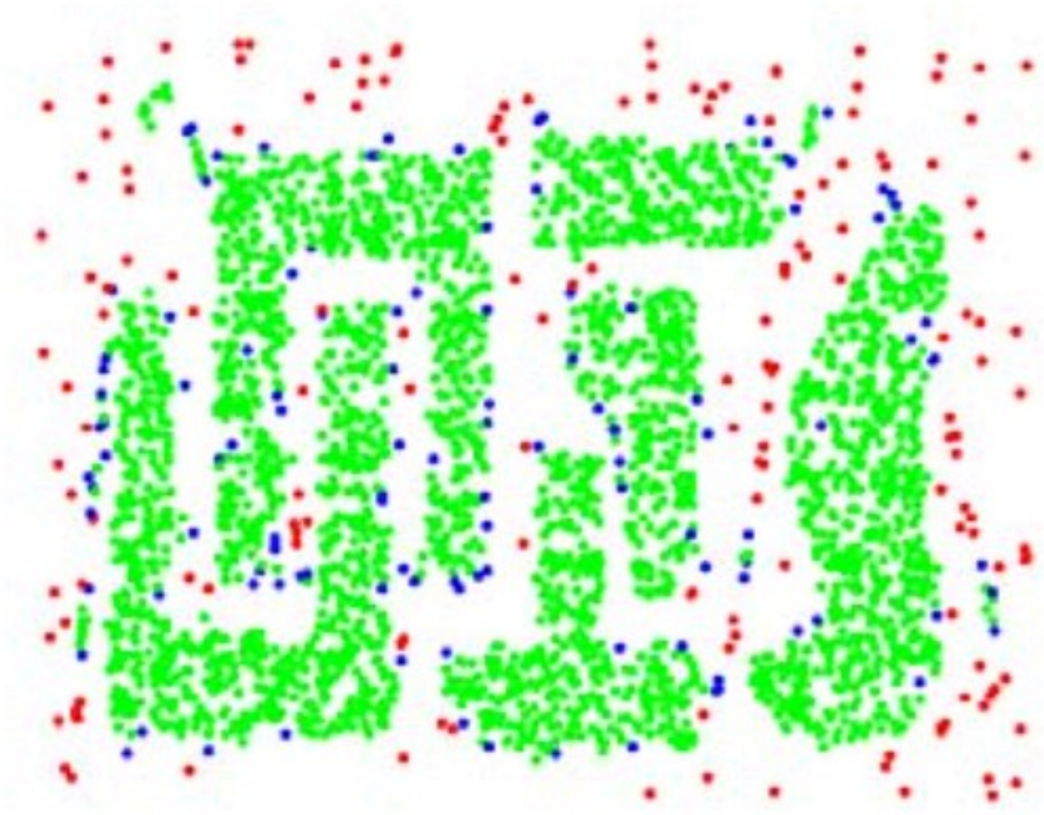
- Classify all points as core, border, or noise, using Eps and $MinPts$
- Eliminate noise points
- Any two core points that are within Eps of each other are put in the same cluster
- Any border point that is within Eps of a core point is put into the same cluster as the core point. (Ties may need to be resolved.)

Example



Eps = 10, MinPts = 4

Example



Eps = 10, MinPts = 4

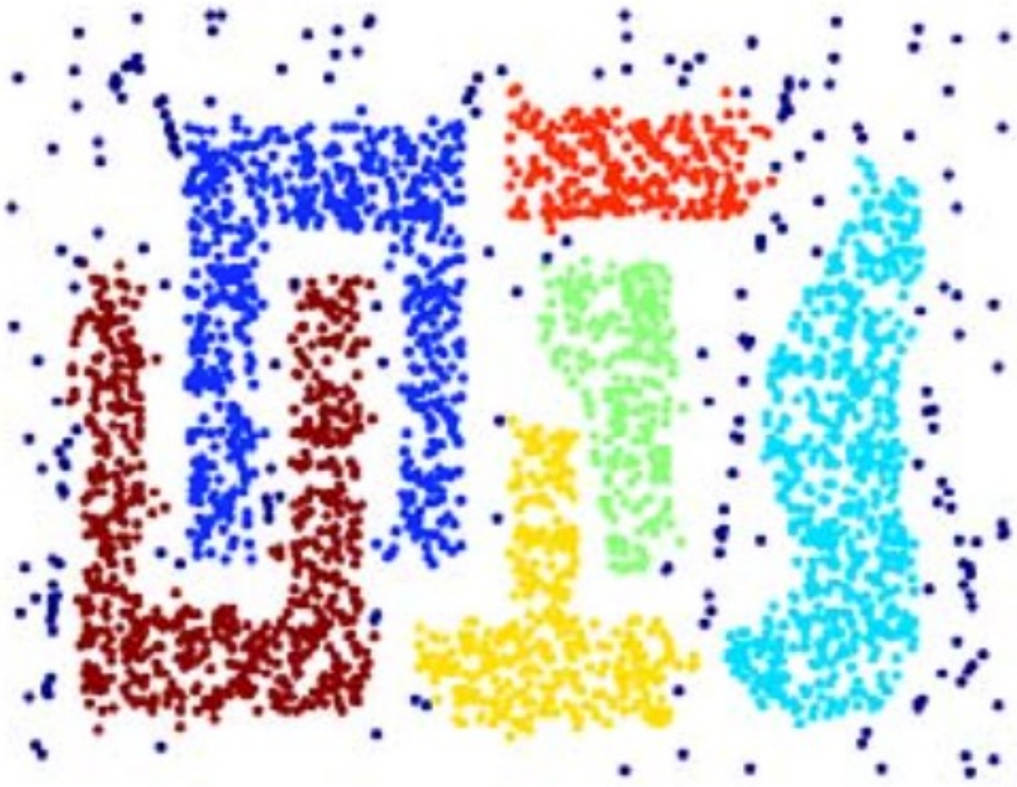
Point types:

Core

Border

Noise

Example

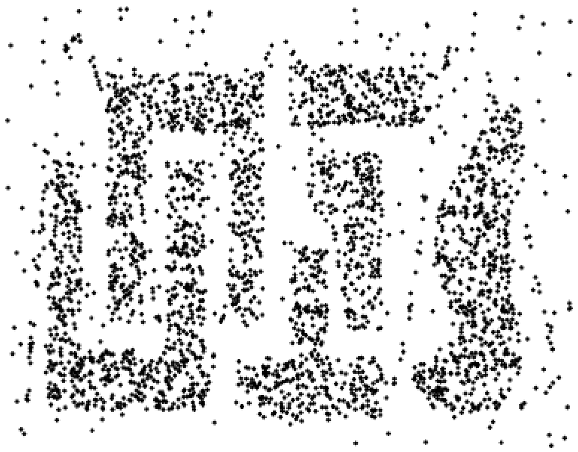


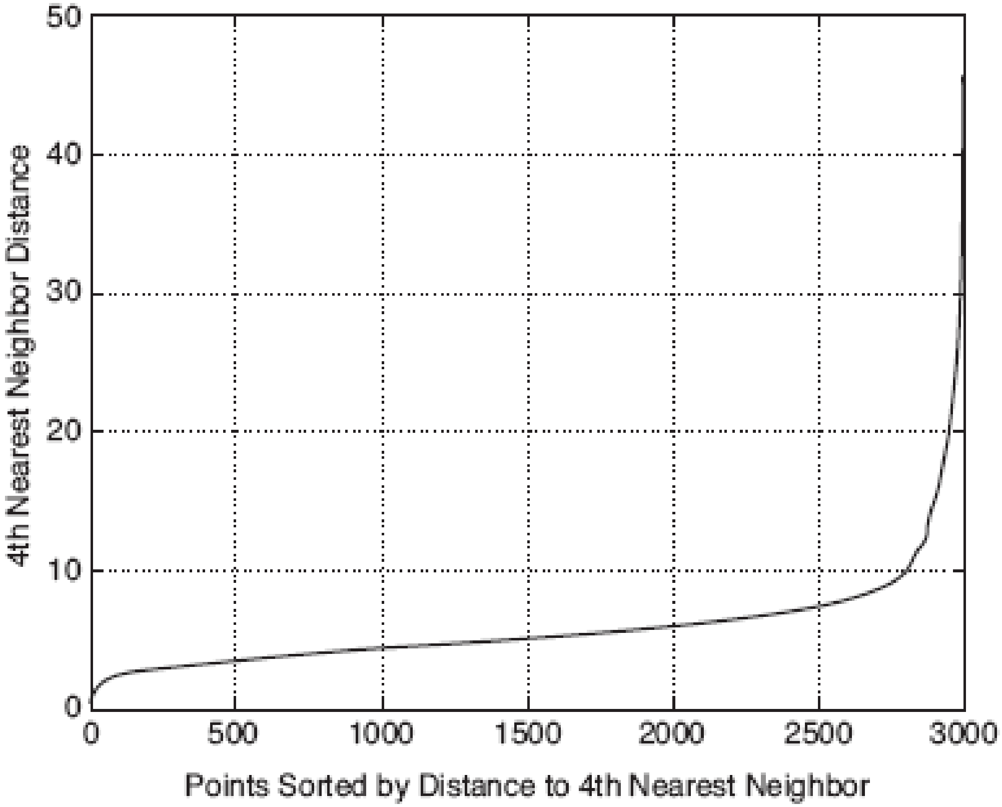
Eps = 10, MinPts = 4

Final Clustering

Determining Parameters

- *k-dist*: The distance from each point to it’s k^{th} nearest neighbor
- Select some k (typically based on domain knowledge, or often k=4 is used)
- Compute the *k-dist* for all data points, sort them in increasing order.
- There will be a sharp change at the value of *k-dist* that corresponds to a suitable value of *Eps*. Select this distance to be *Eps* and k to be *MinPts*.





k-dist plot

Characteristics of DBSCAN

- Can handle clusters of arbitrary shapes and sizes
- Resistant to noise & outliers
- Curse of dimensionality: distance between points, and thus density, becomes less meaningful as dimensionality increases
- DBSCAN can struggle with clusters of different densities