Lecture 6

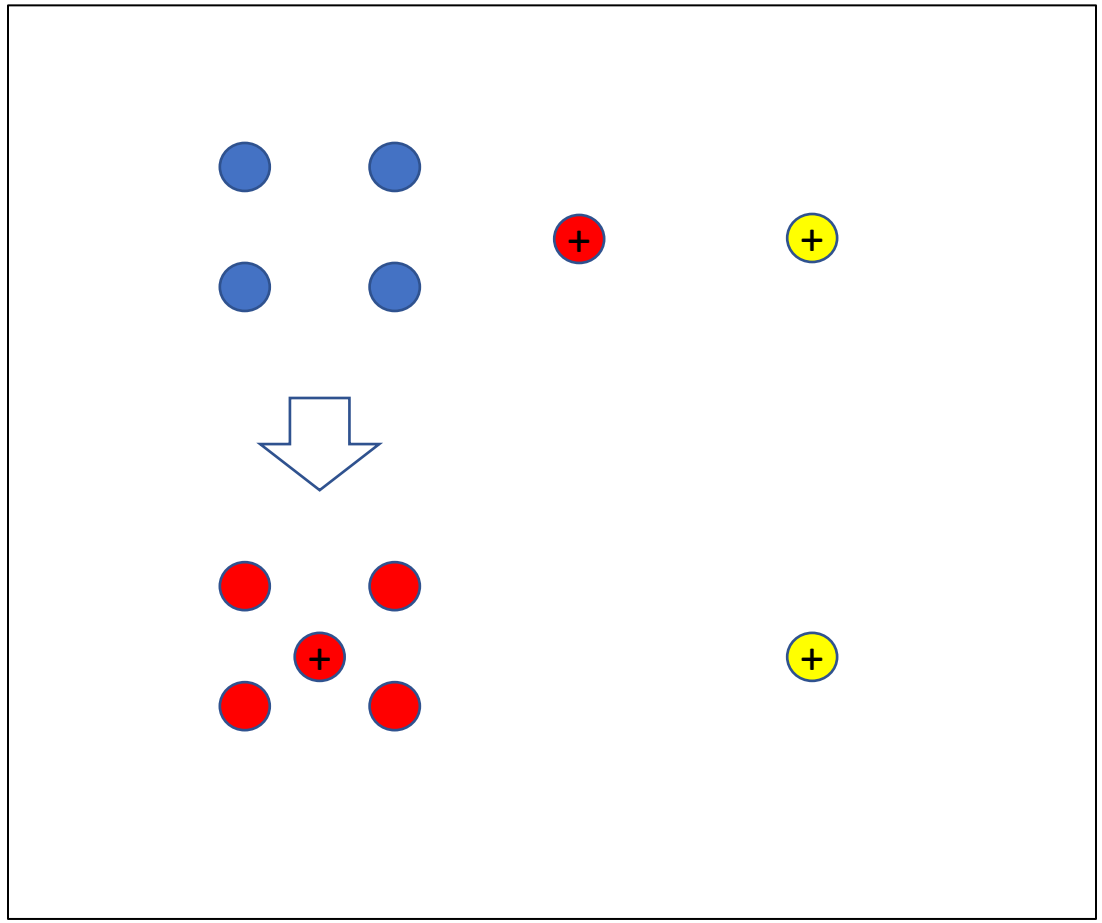
Clustering 2

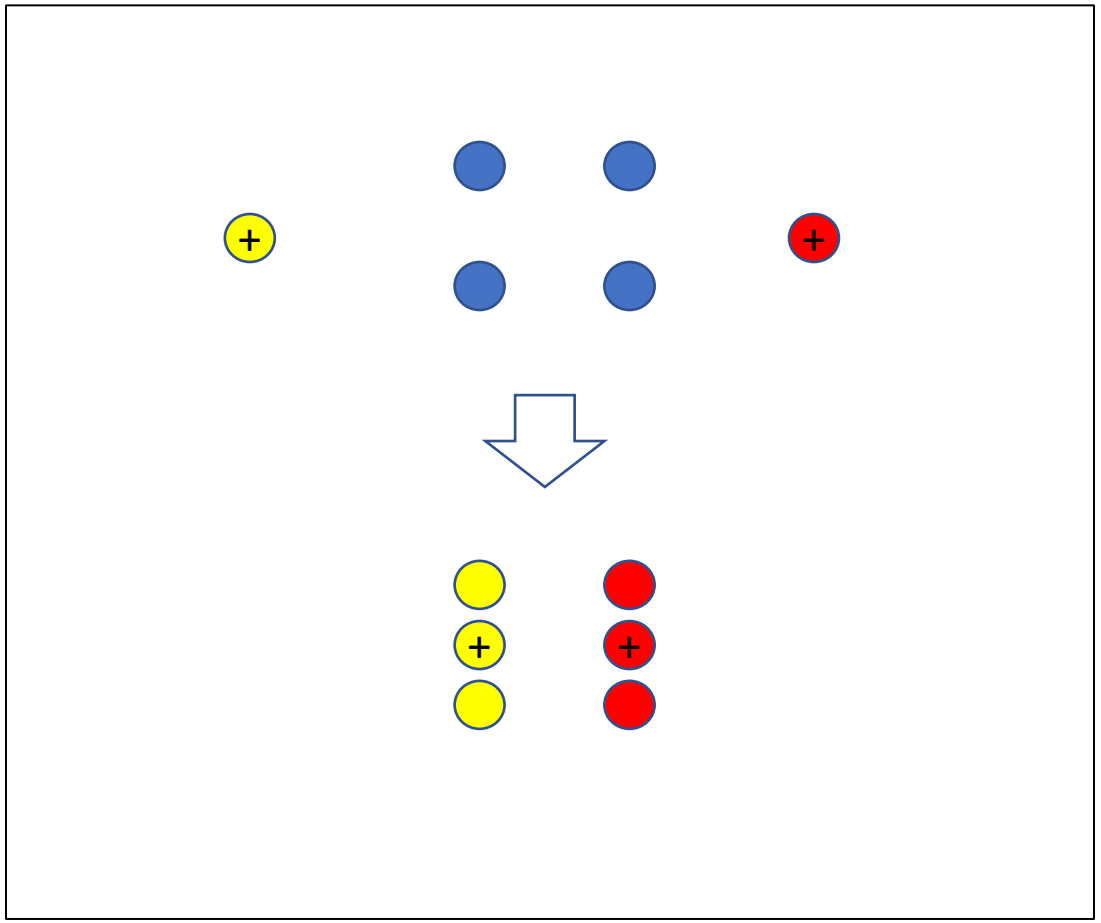
Topics

- DBSCAN clustering
- Hierarchical clustering
- Validating clusterings

Limitations of K-means clustering

K-means: result depends on initialization





DBSCAN clustering

Steven Bierwagen

DBSCAN

- Density-Based Spatial Clustering of Applications with Noise

- From 1996

Ingredients for DBSCAN

- A *distance* measure (or metric or similarity measure)
  - often Euclidean distance

scanning radius

- A number defining the meaning of *neighbor*
  - epsilon: the max distance between two points considered neighbors.

- A number defining the meaning of *cluster* (vs outlier or noise)
  - minpts: the minimum number of points in a cluster.

min points inside radius

Two hyperparameters

Labeling step

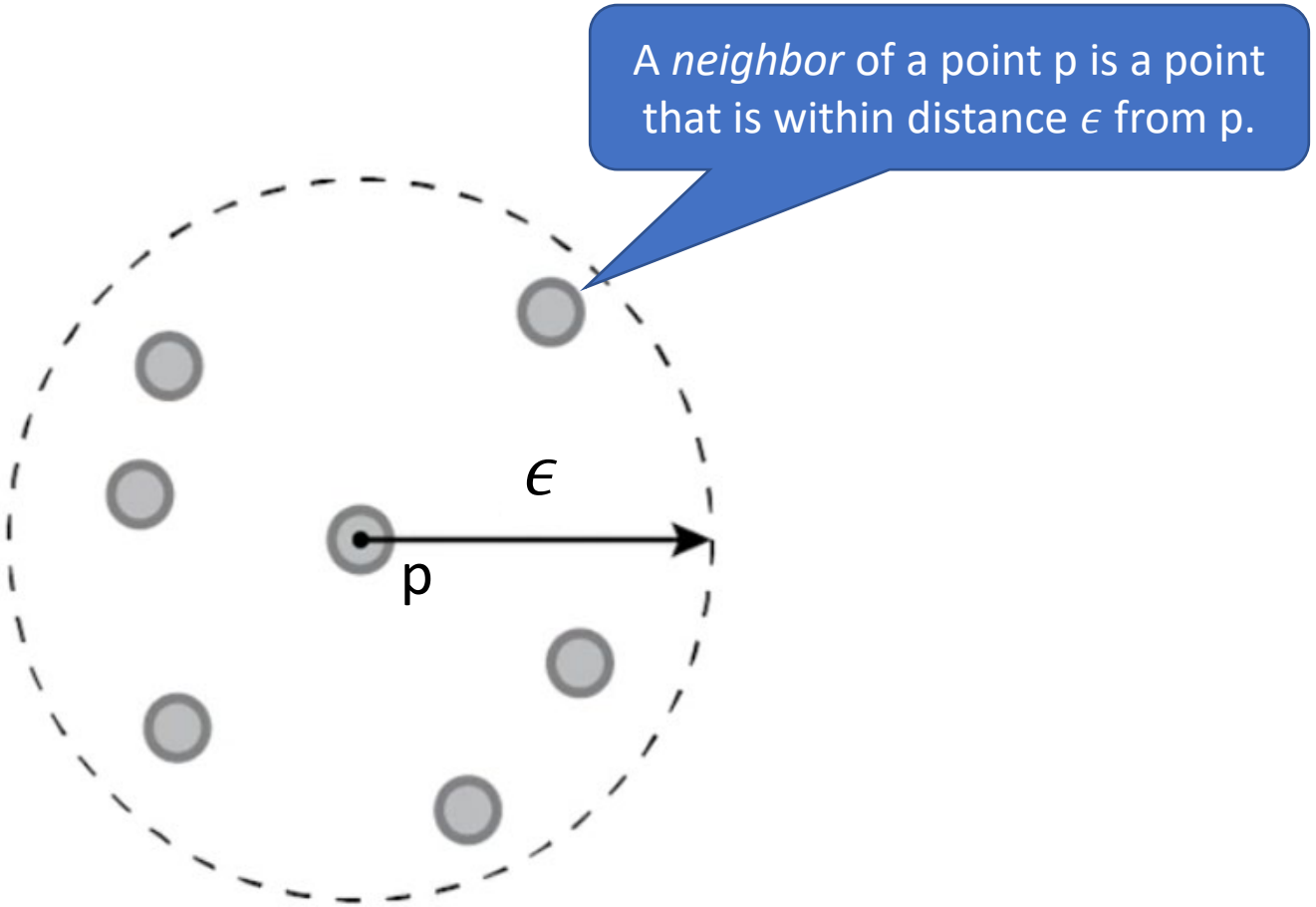
CORE

BORDER

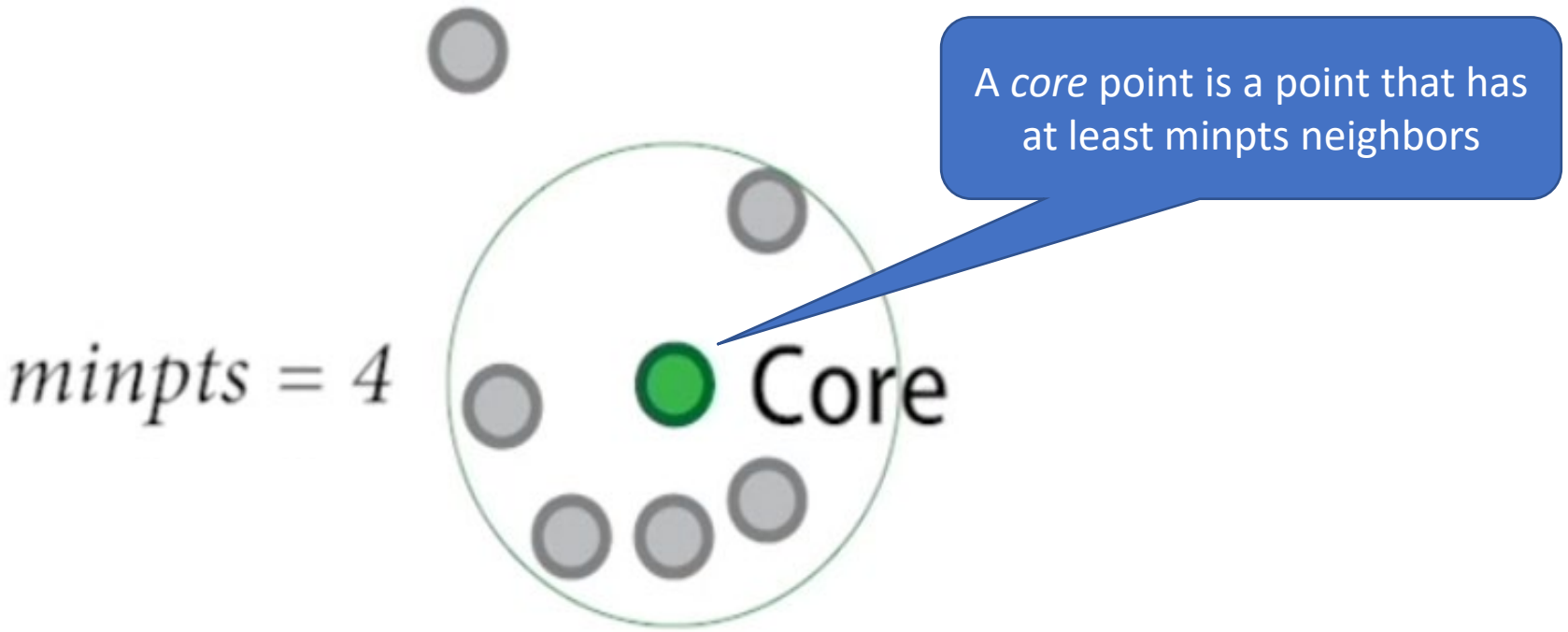
NOISE

All points are labeled with one of these labels

Neighbors

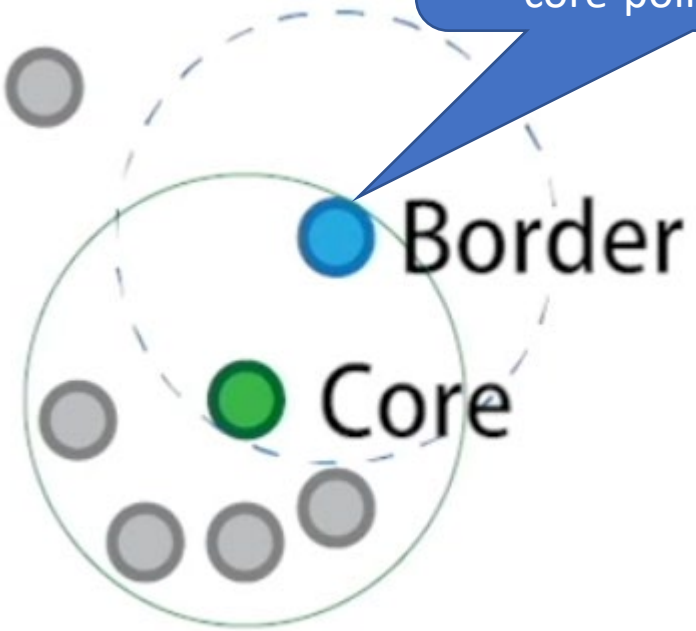

A *neighbor* of a point p is a point that is within distance ϵ from p.
ϵ
p

Core points


minpts = 4
Core
A *core* point is a point that has at least minpts neighbors

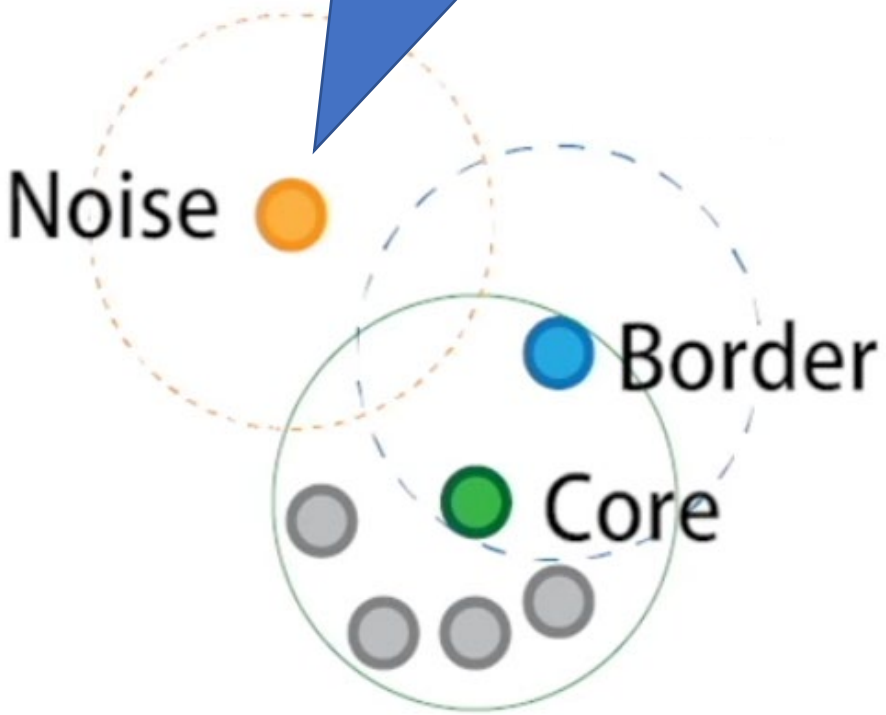
Border points

A *border* point is a non-core point that has at least one core-point as neighbor.



Noise points

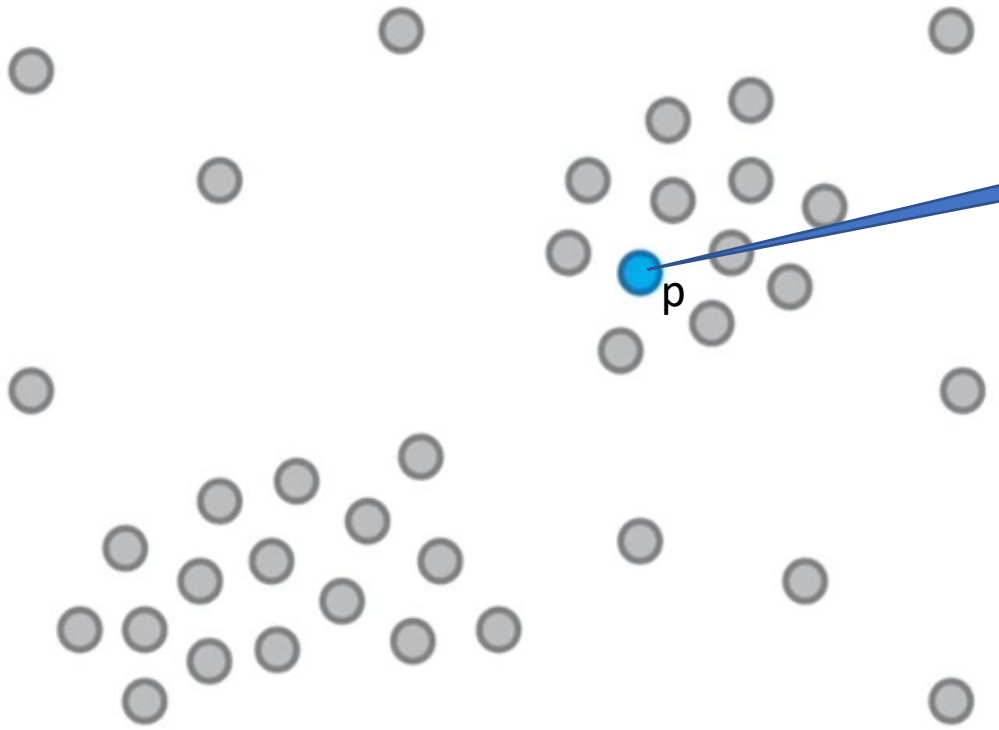
A *noise* point or *outlier* is a non-core and non-border point



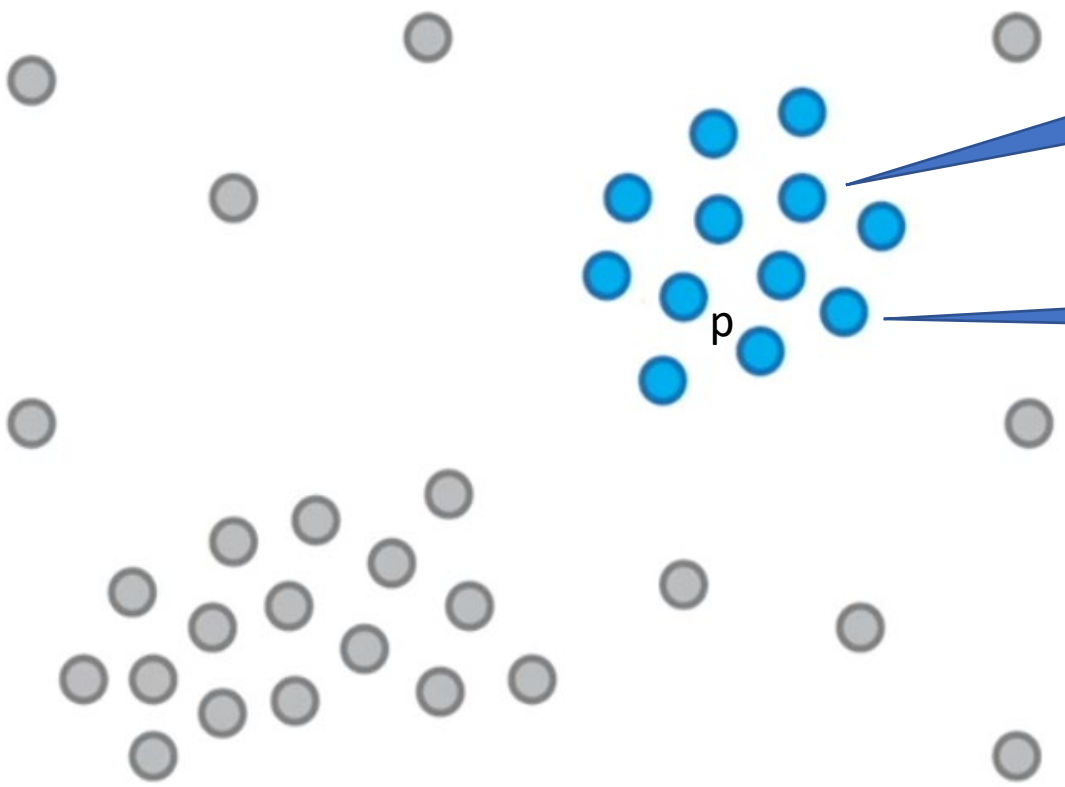
Clustering step

Clusters all core points and border points. Outliers will not be clustered!

Start by picking a new color c and an uncoloured core point p.



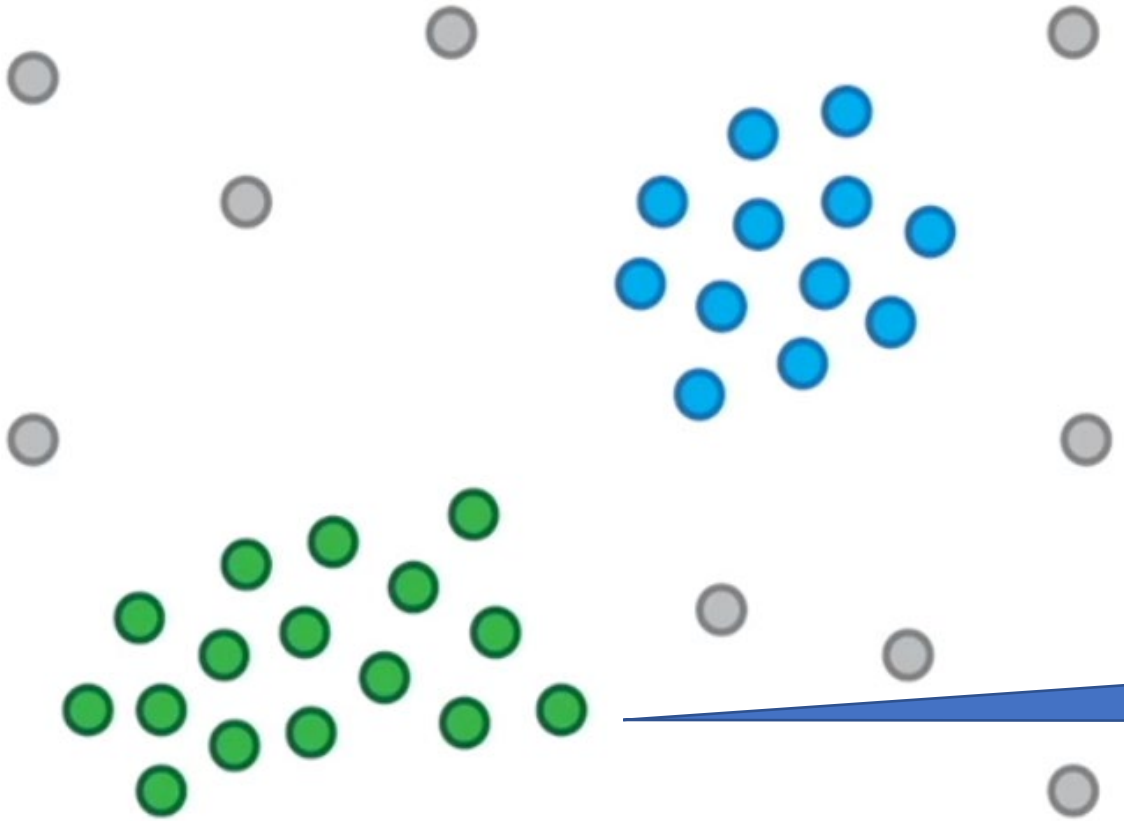
Clustering step



Put an edge between core points that are neighbors. Color those connected components with c

Also color the border points of those nodes with c

Clustering step



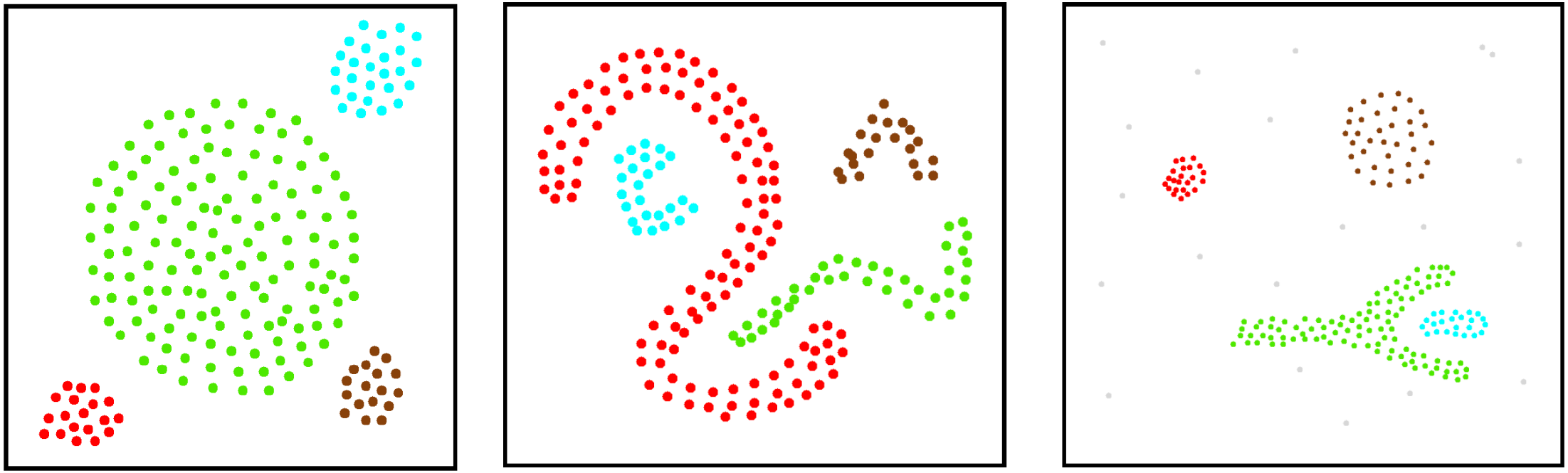
Repeat until all core points and border points have been colored!

Algorithm

Algorithm 8.4 DBSCAN algorithm.

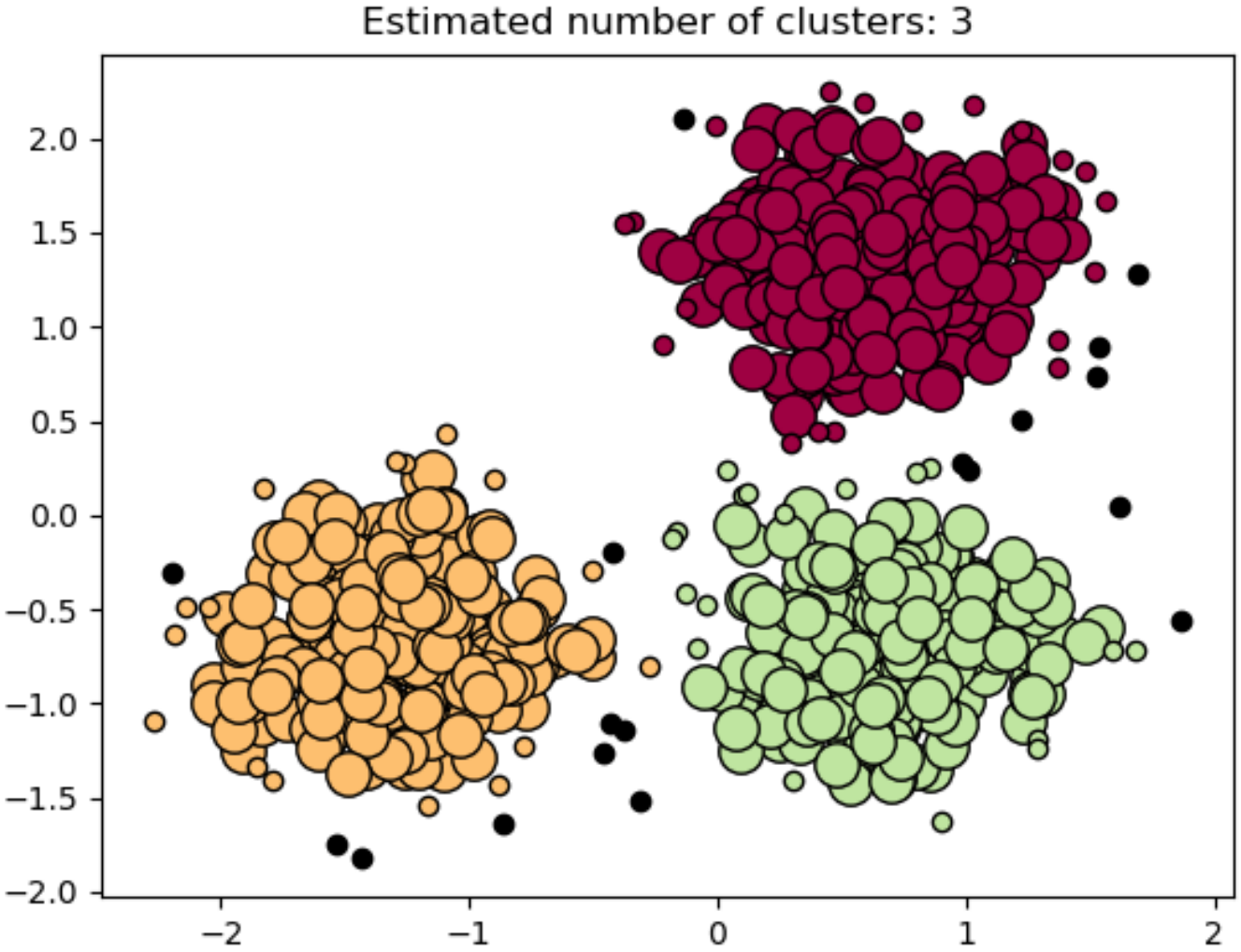
```
Label all points as core, border, or noise points.
Eliminate noise points.
Put an edge between all core points that are within Eps of each other.
Make each group of connected core points into a separate cluster.
Assign each border point to one of the clusters of its associated core points.
```

Clusterings created by DBSCAN



Ester, Kriegel, Sander, Xu (1996), In Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, (KDD), AAAI Press, pp. 226–231

Demo of DBSCAN



large = core
small = border

black = outlier

https://scikit-learn.org/stable/auto_examples/cluster/plot_dbscan.html

K-means vs. DBSCAN

- K-means assigns all points to a cluster, whereas DBSCAN doesn’t necessarily do this. DBSCAN treats outliers as outliers.

- K-means works best when clusters are basically spherical. DBSCAN can find arbitrarily-shaped clusters.

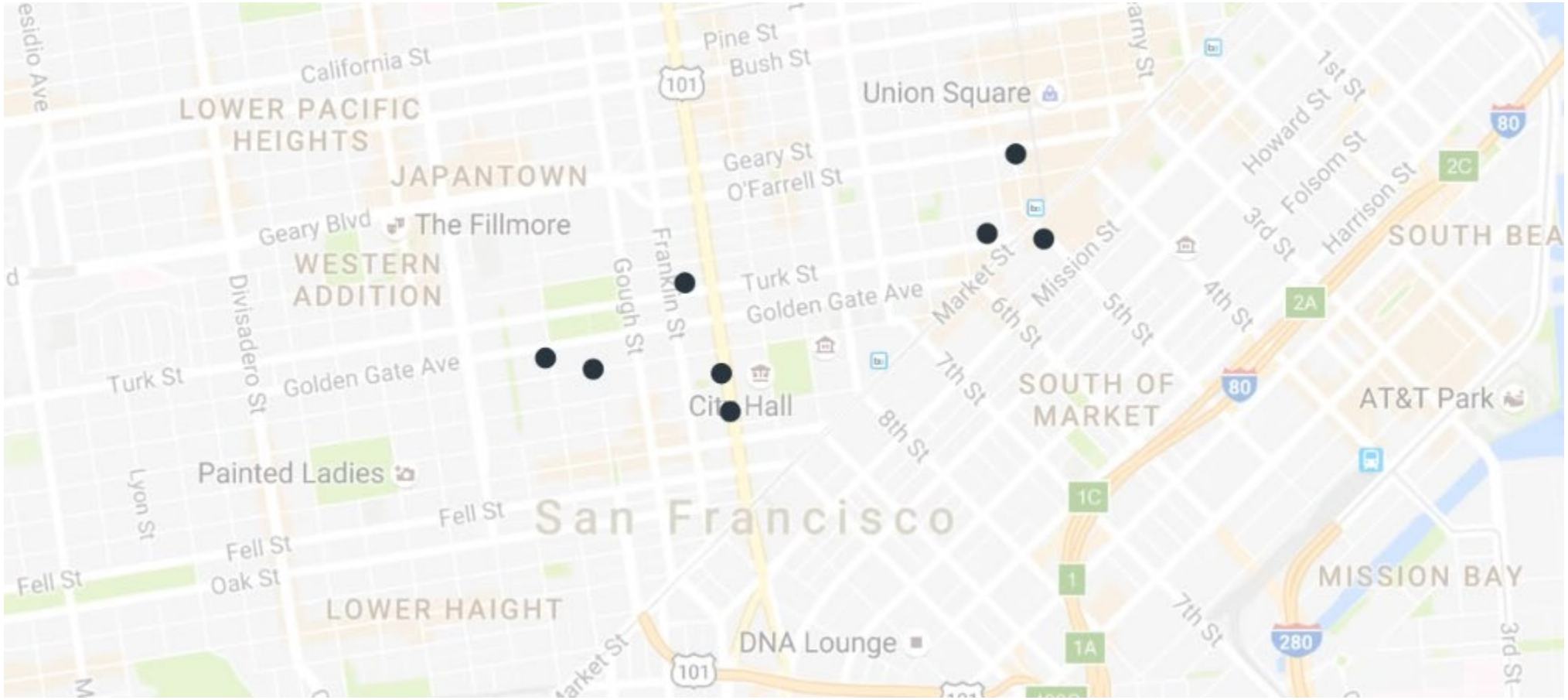
- DBSCAN doesn’t require the number of clusters to be specified by the user.

Hierarchical clustering

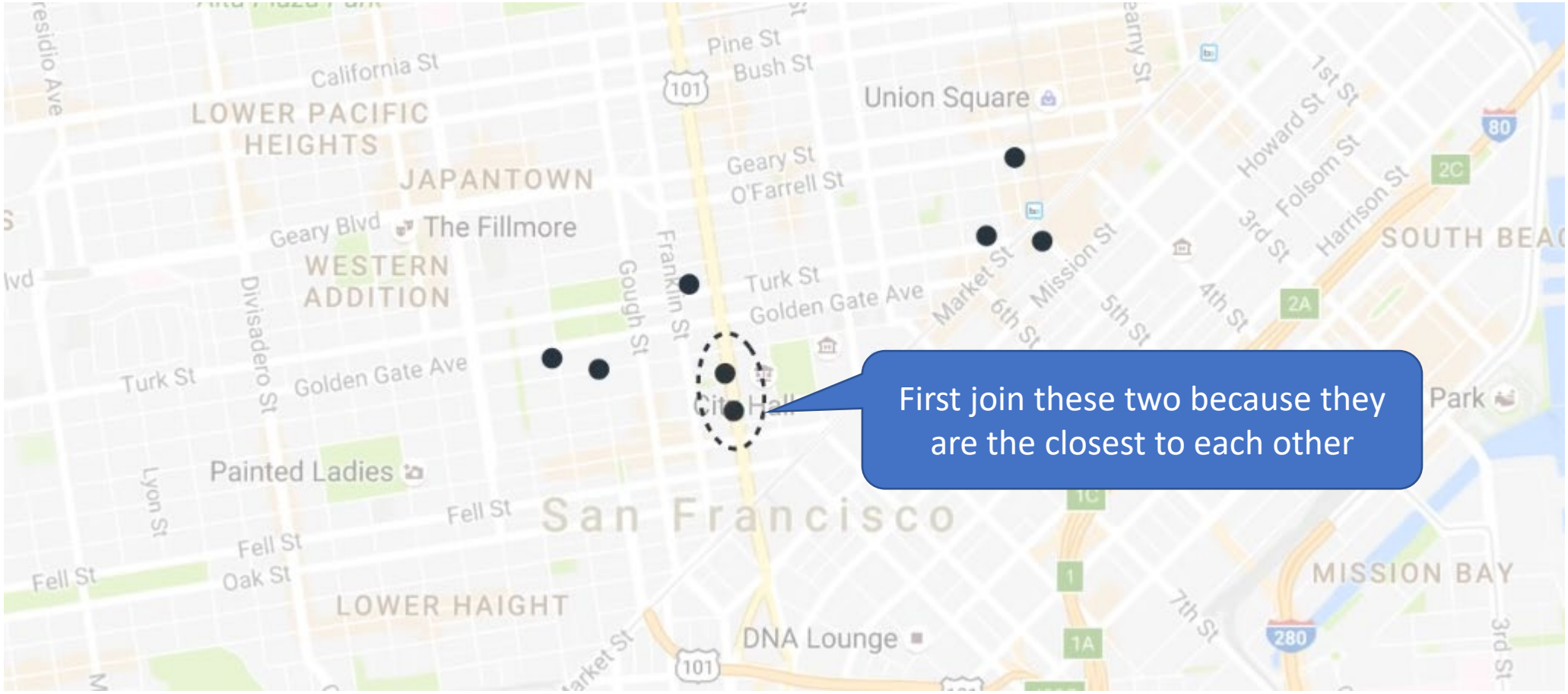
Luis Serrano

A new set of points

Suppose we want to cluster these addresses by proximity. No pizza parlors involved this time!

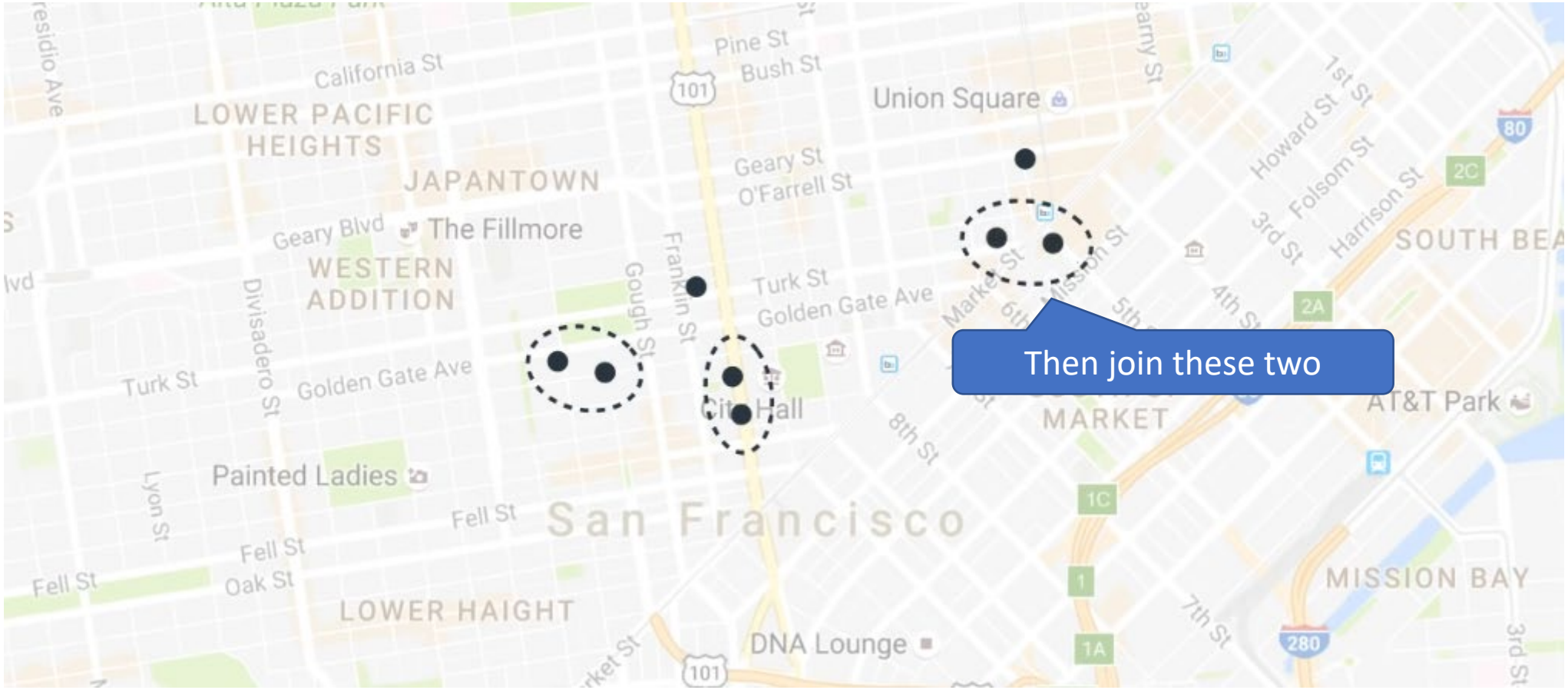


Join the close ones

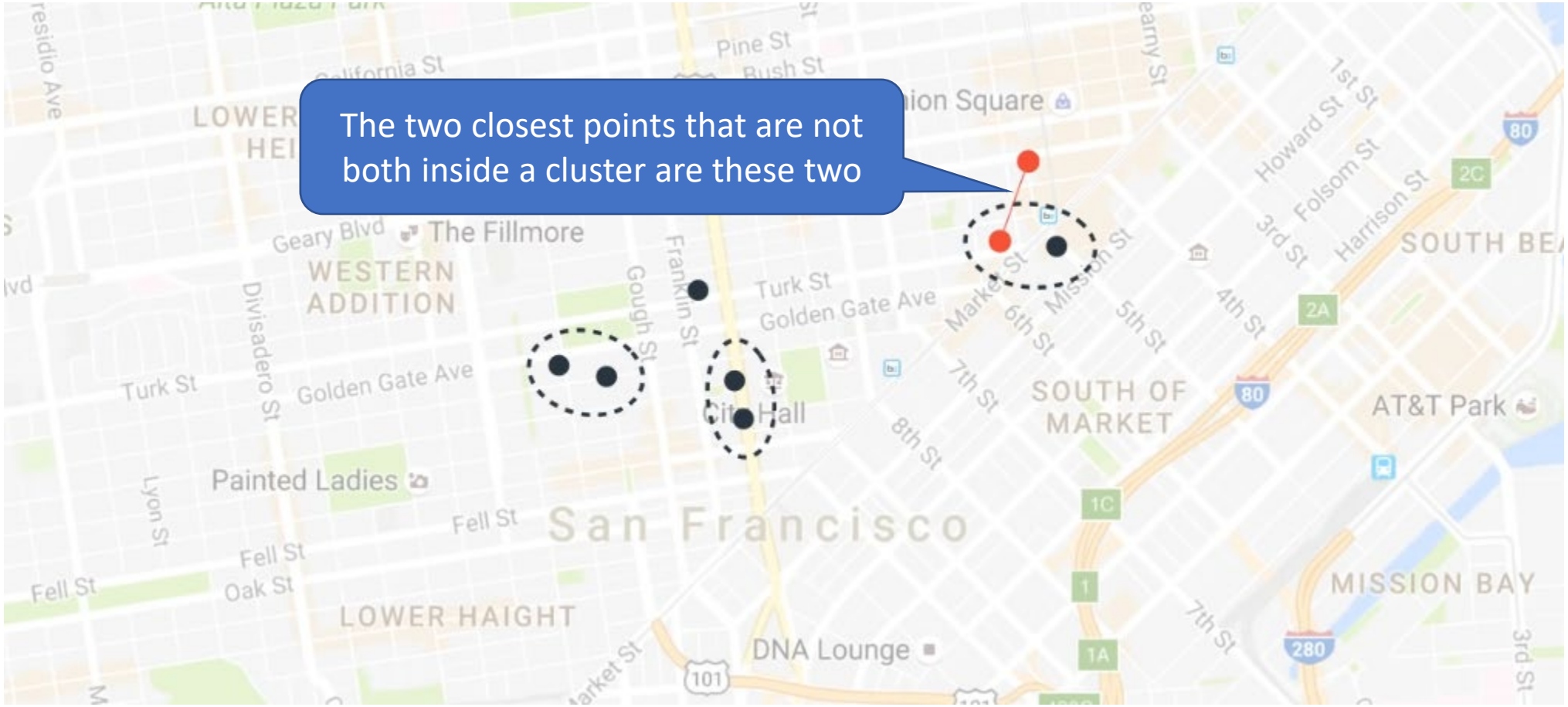


Join the close ones

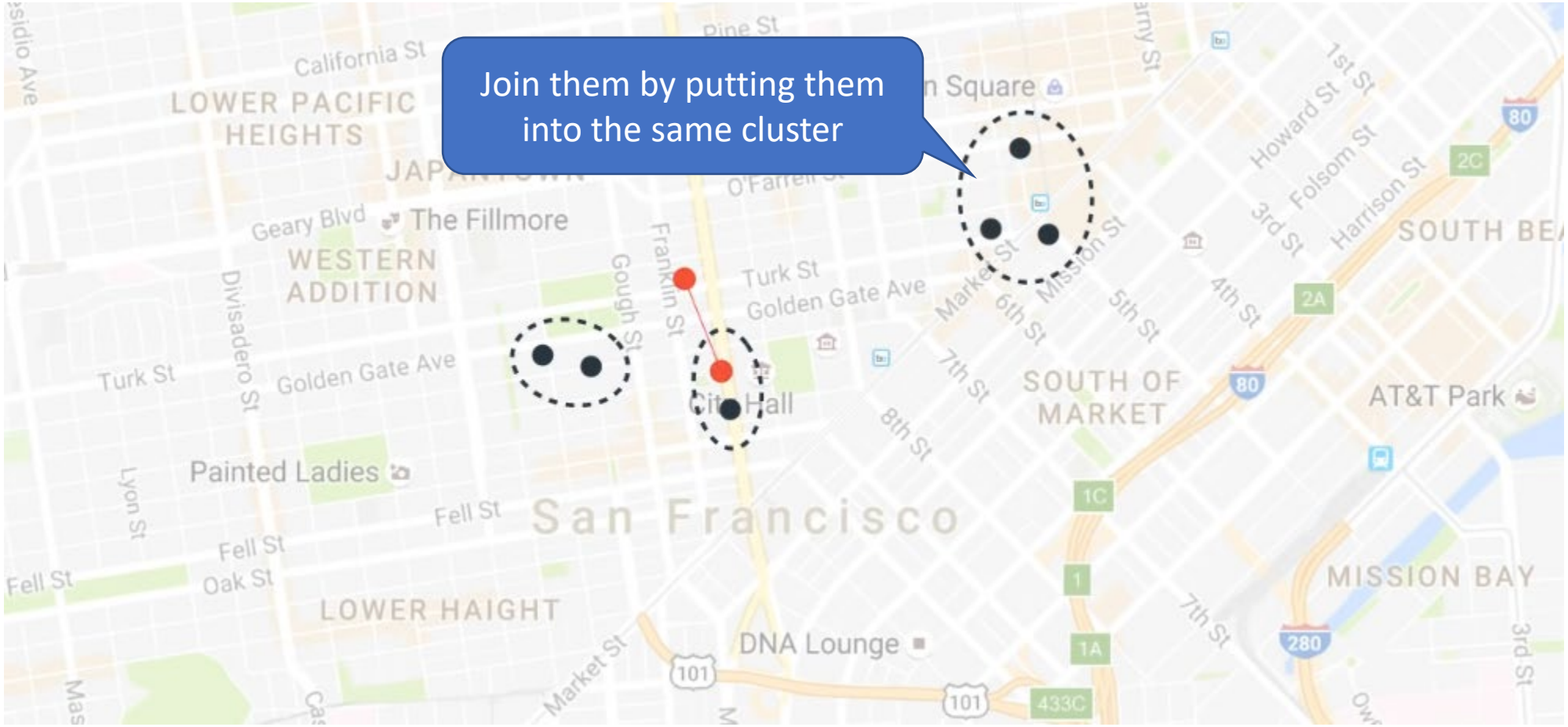
Join the close ones



Join the close ones



Join the close ones

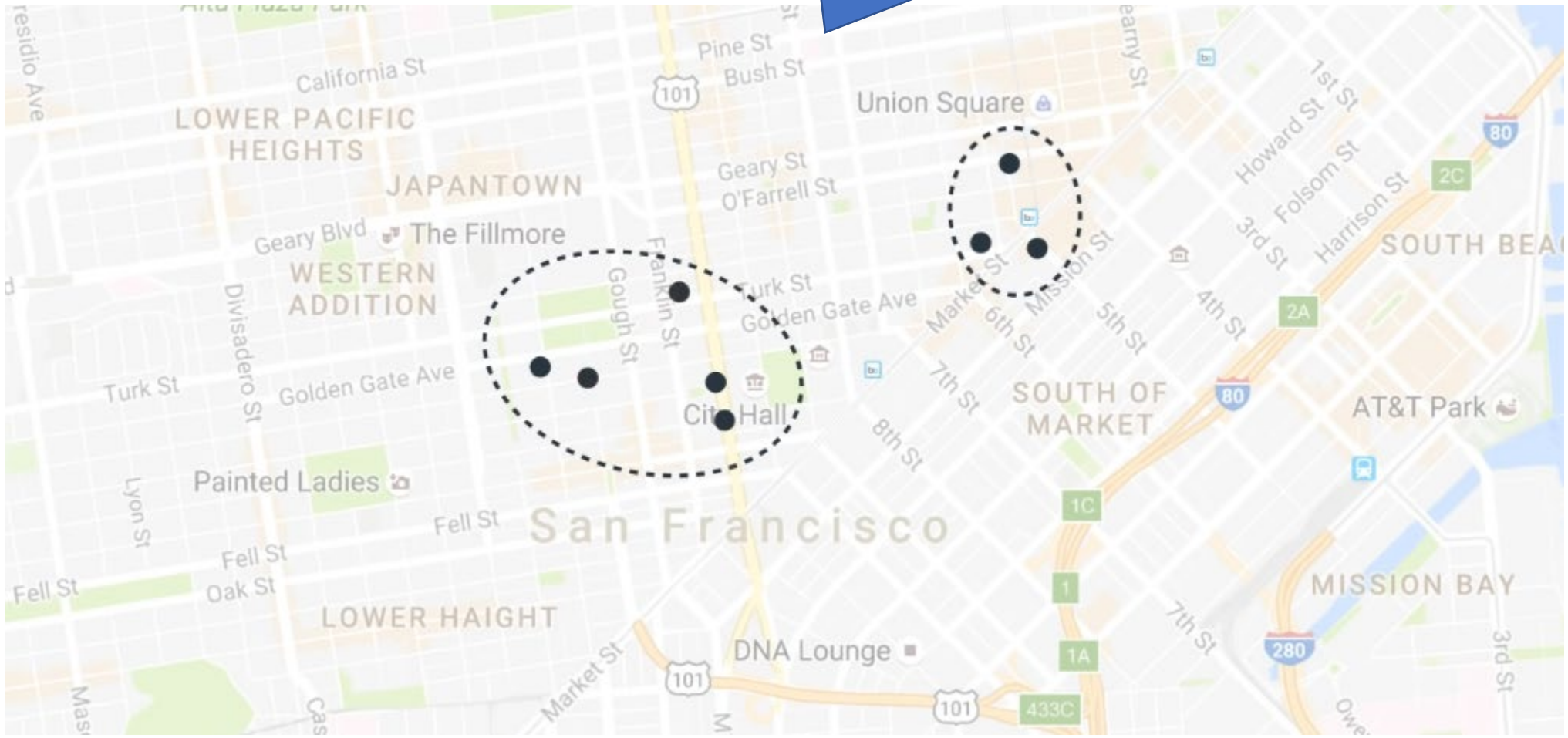


Join the close ones

Continue in the same way…

Join the close ones

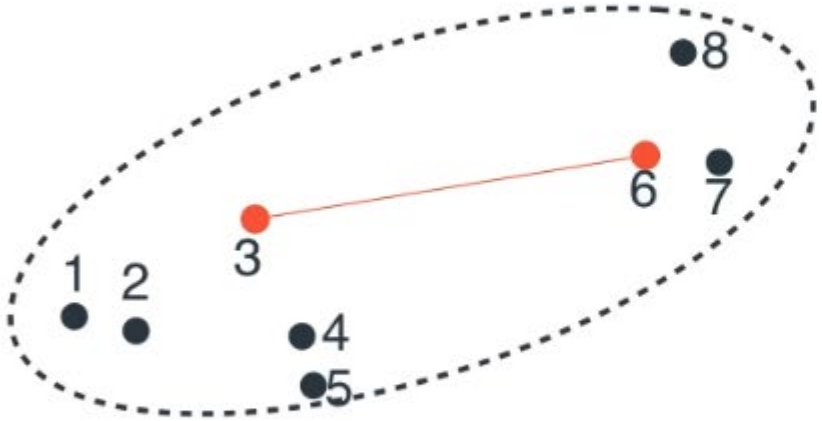
Continue in the same way…

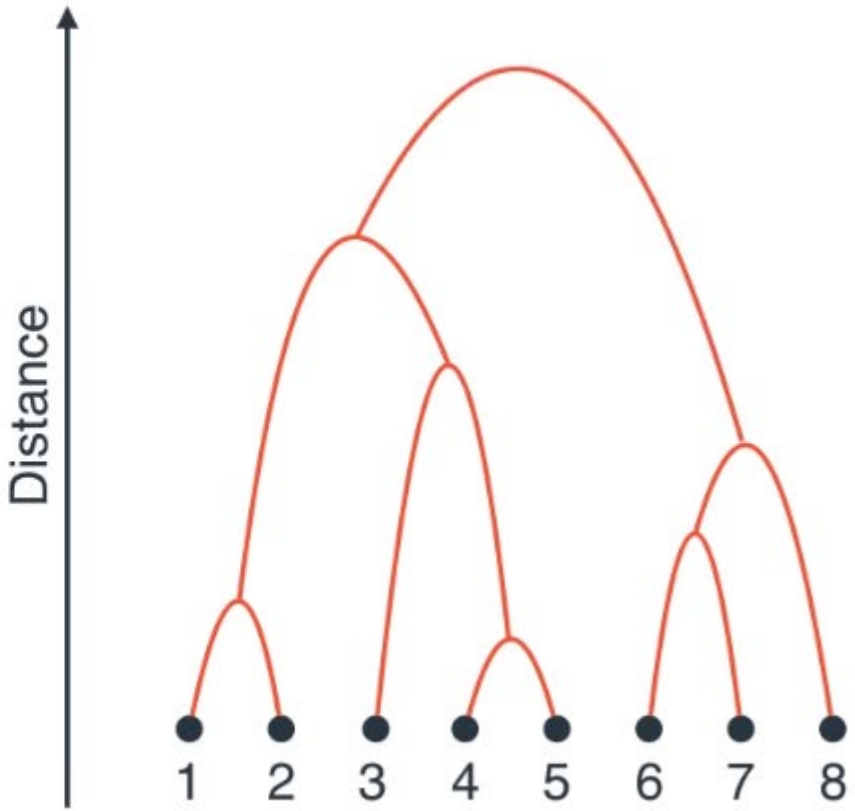


Join the close ones

Dendrogram

A *dendrogram* shows the entire hierarchichal clustering process (without STOP)

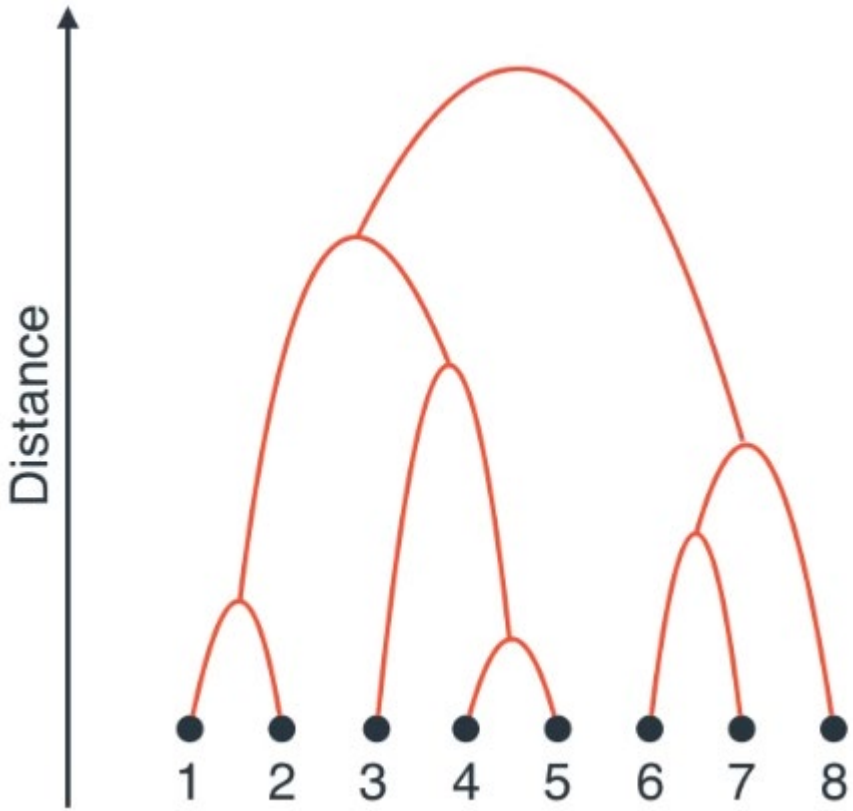




Dendrogram

1
2
3
4
5
6
7
8
Distance
If we want 4 clusters, we can cut here.
If we want 8 clusters, we can cut here.

Dendrogram

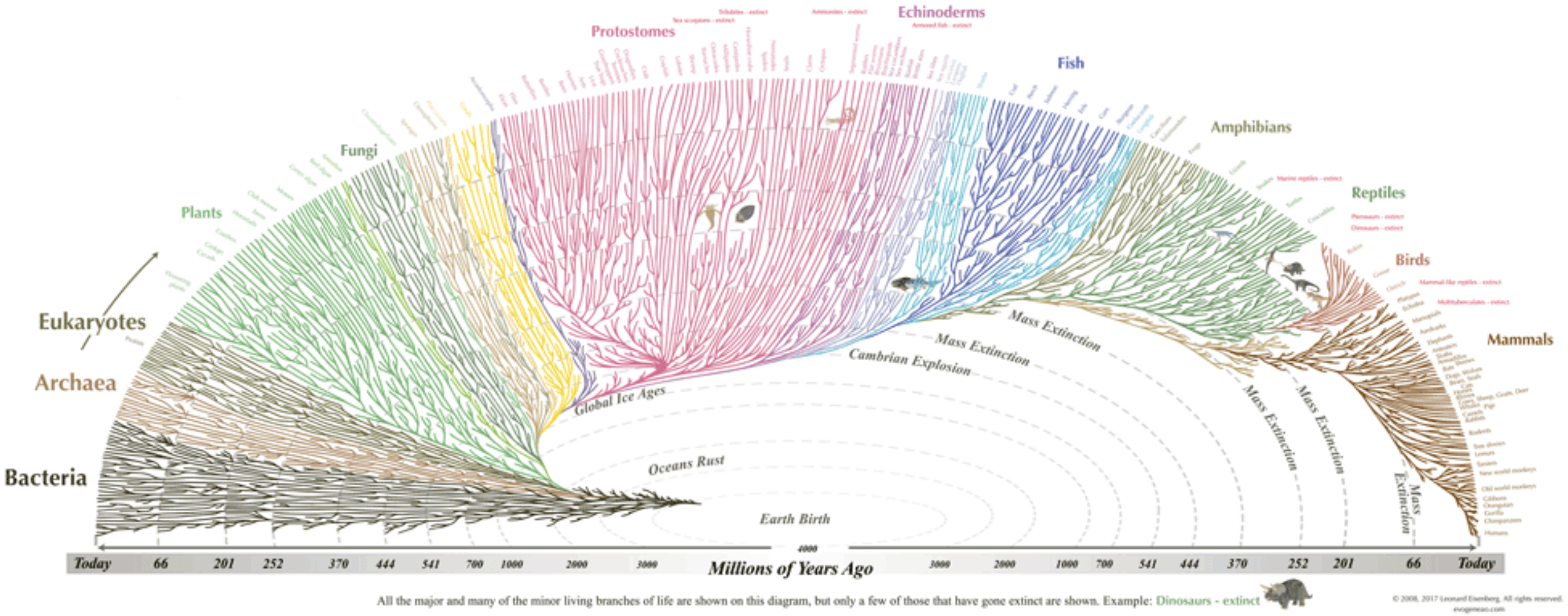


If we have a space with billions of points in thousands of dimensions, the dendrogram is still a 2D graph!

For example the tree of life!

Hierarchichal clustering gives more than a clustering: a hierarchy (or taxonomy)

The tree of life

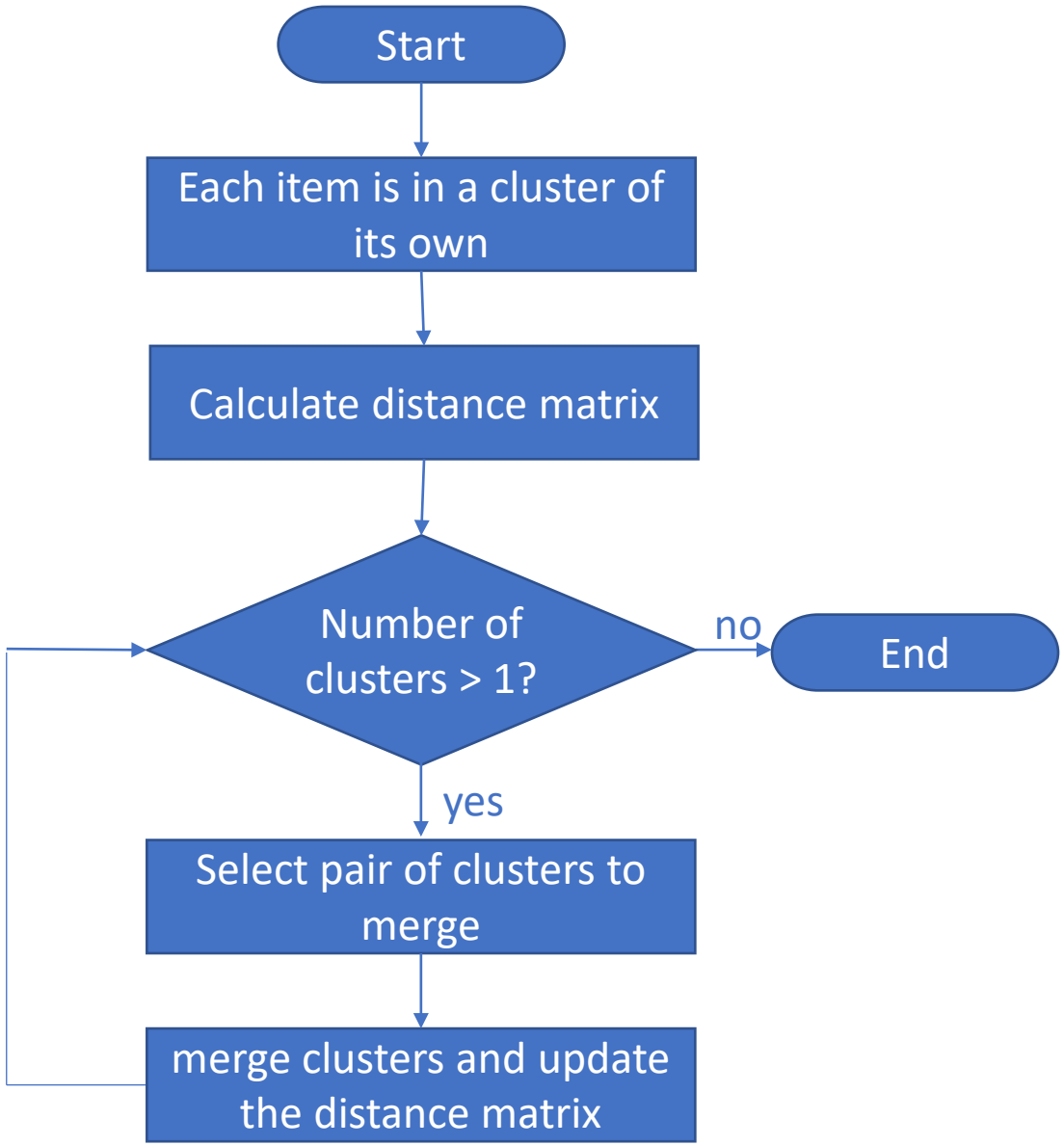


https://www.evogeneao.com/

Hierarchical clustering

- Sometimes called agglomerative clustering, when done bottom-up

- From one extreme case (many clusters, each containing one item) to another (one cluster that contains all items)



Distance matrix

Edit distances between protein sequences (strings)

a. Human haemoglobin alpha chain
b. Human haemoglobin beta chain
c. Horse haemoglobin alpha chain
d. Horse haemoglobin beta chain
e. Marine bloodworm haemoglobin
f. Yellow lupine leghaemoglobin

D	a	b	c	d	e	f
a	0	84	18	86	112	121
b	84	0	85	26	117	119
c	18	85	0	84	112	125
d	86	26	84	0	113	121
e	112	117	112	113	0	119
f	121	119	125	121	119	0

Six proteins with a common evolutionary ancestor

Amino acid sequences of six proteins

```
> human_alpha
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR
> human_beta
VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
> horse_alpha
VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHGKKVADGLTLAVGHLDDLPGALSDLSNLHAHKLRVDPVNFKLLSHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLTSKYR
> horse_beta
VQLSGEEKAAVLALWDKVNEEEVGGEALGRLLVVYPWTQRFFDSFGDLSNPGAVMGNPKVKAHGKKVLHSFGEGVHHLDNLKGTFAALSELHCDKLHVDPENFRLLGNVLALVVARHFGKDFTPELQASYQKVVAGVANALAHKYH
> marine_bloodworm
GLSAAQRQVIAATWKDIAGADNGAGVGKKCLIKFLSAHPQMAAVFGFSGASDPGVAALGAKVLAQIGVAVSHLGDEGKMVAQMKAVGVRHKGYGNKHIKAQYFEPLGASLLSAMEHRIGGKMNAAAKDAWAAAYADISGALISGLQS
> yellow_lupine
GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSEVPQNNPELQAHAGKVFKLVYEAAIQLEVTGVVVTDATLKNLGSVHVSKGVADAHFPVVKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMDDAA
```

Edit distance is the number of single character operations that are required to change one string into another.

Merging clusters

- When clusters are merged, how do we calculate the distance between the merged cluster and each of the other clusters?

- Various algorithms to choose from, e.g.
  - complete linkage (furthest inter-cluster distance)
  - single linkage (closest inter-cluster distance)
  - average linkage
    - Unweighted Pair Group Method with Arithmetic Mean (UPGMA)
    - Weighted Pair Group Method with Arithmetic Mean (WPGMA)
  - neighbour-joining

Neighbour-joining

This has the most complicated method for selecting pairs to merge and updating the distance matrix, but it has a nice property:

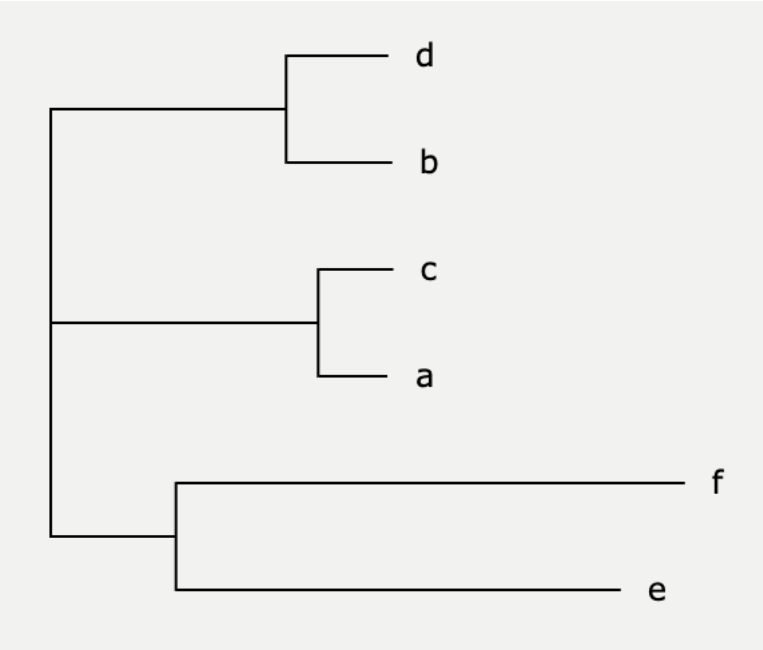
- **If input distance matrix is correct, output tree will be correct** (distance between each pair in the tree matches the distance between that pair in the initial distance matrix).

Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, *4*(4), 406-425

A highly cited paper – over 60,000 citations

Neighbour-joining result

a. Human haemoglobin alpha chain
b. Human haemoglobin beta chain
c. Horse haemoglobin alpha chain
d. Horse haemoglobin beta chain
e. Marine bloodworm haemoglobin
f. Yellow lupine leghaemoglobin

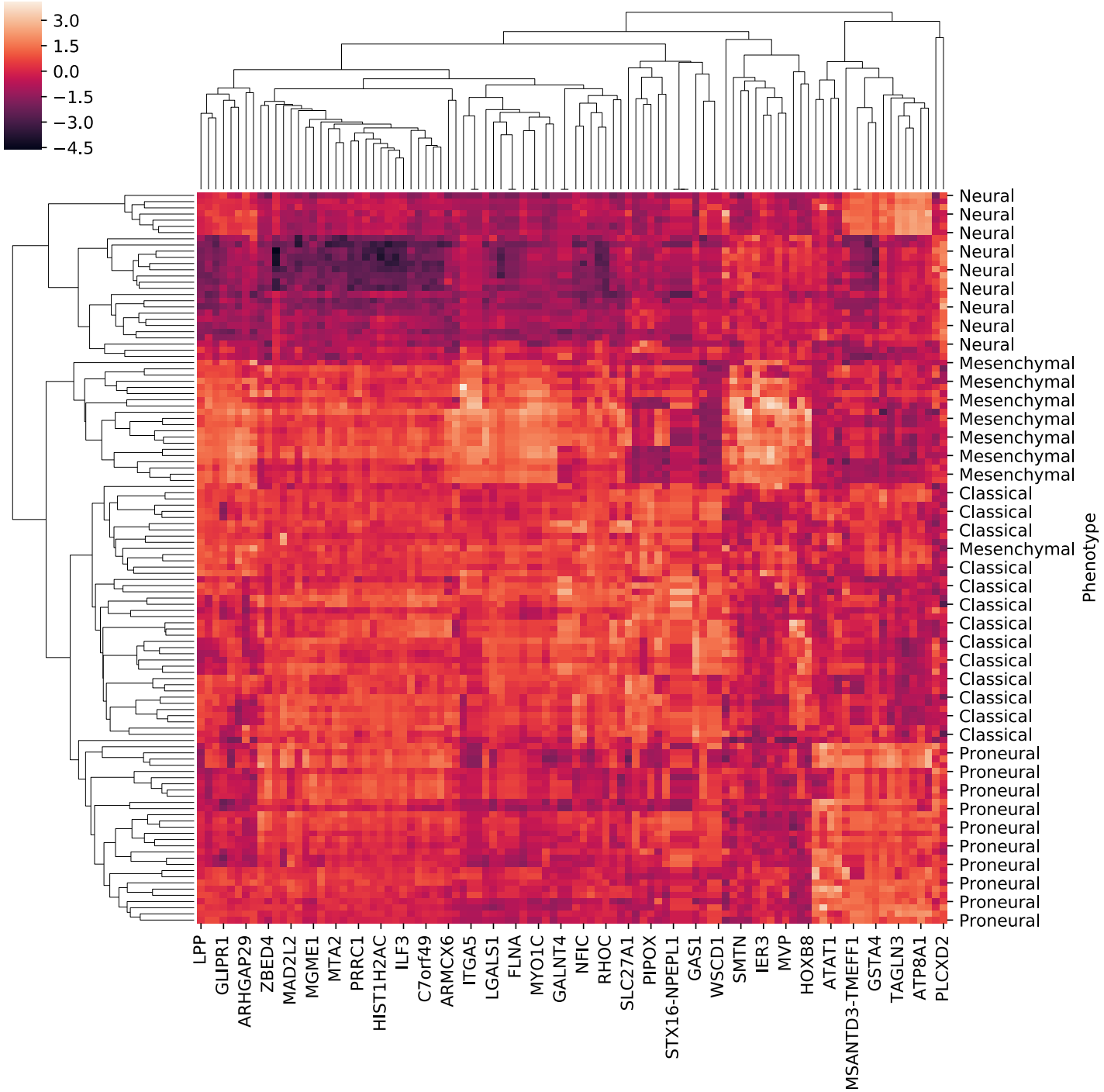


(((d:12.75,b:13.25):29.375,(c:9.375,a:8.625):33.375):0,(f:63.5,e:55.5):15.625);

Biclustering

- Biclustering algorithms simultaneously cluster rows and columns of a data matrix.

- https://scikit-learn.org/stable/modules/biclustering.html



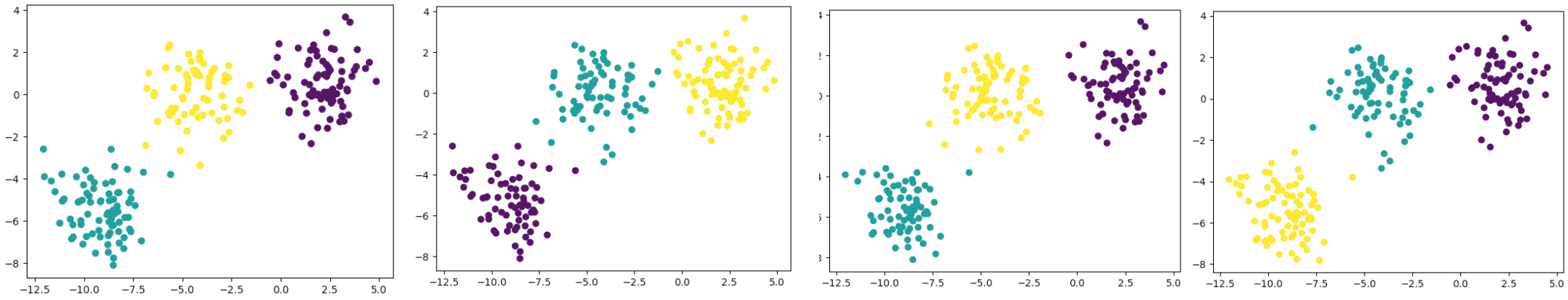
Biclustered heatmap of 118 samples using hierarchical clustering of 100 gene features

Stackhouse et al. (2019) A Novel Assay for Profiling GBM Cancer Model Heterogeneity and Drug Screening. Cells 2019, 8, 702; doi:10.3390/cells8070702

Validating clustering

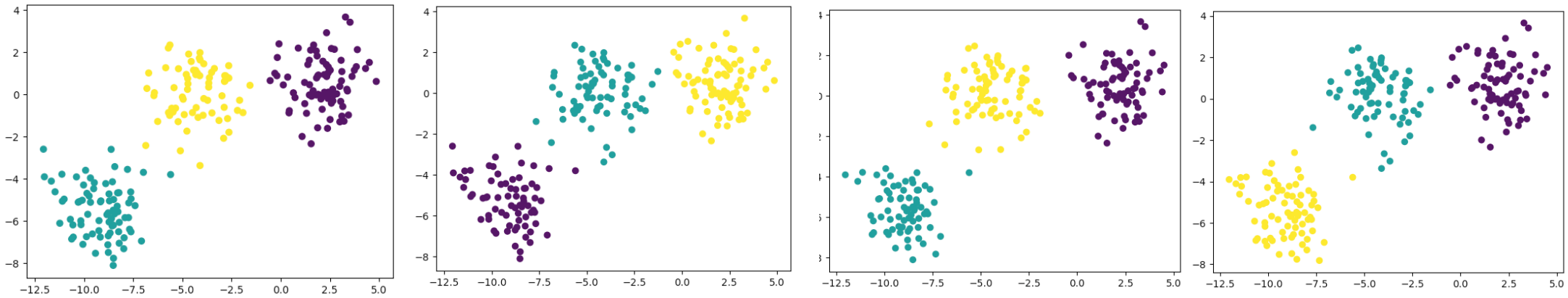
Stability on subsets

Clustering stable if removing a proportion of random points does not change the clustering fundamentally



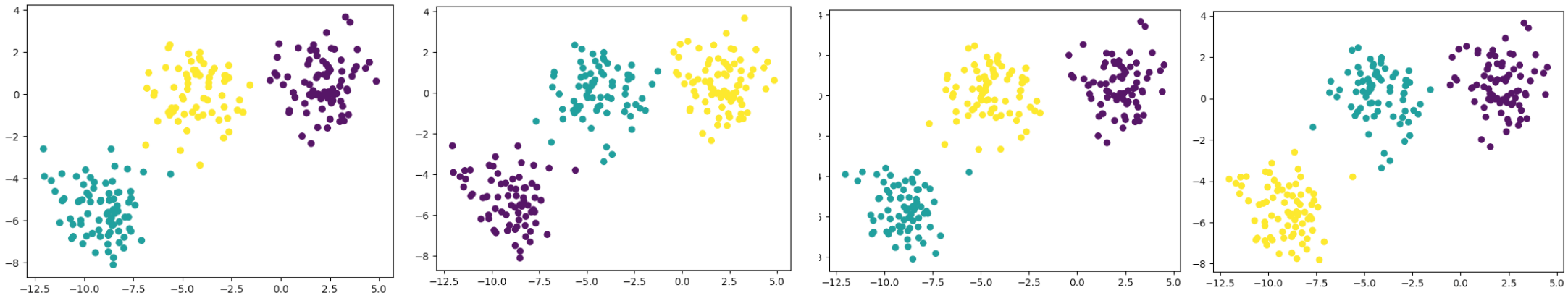
Stability on subsets

Note colors change as labeling clusters into first, second, third … changes!

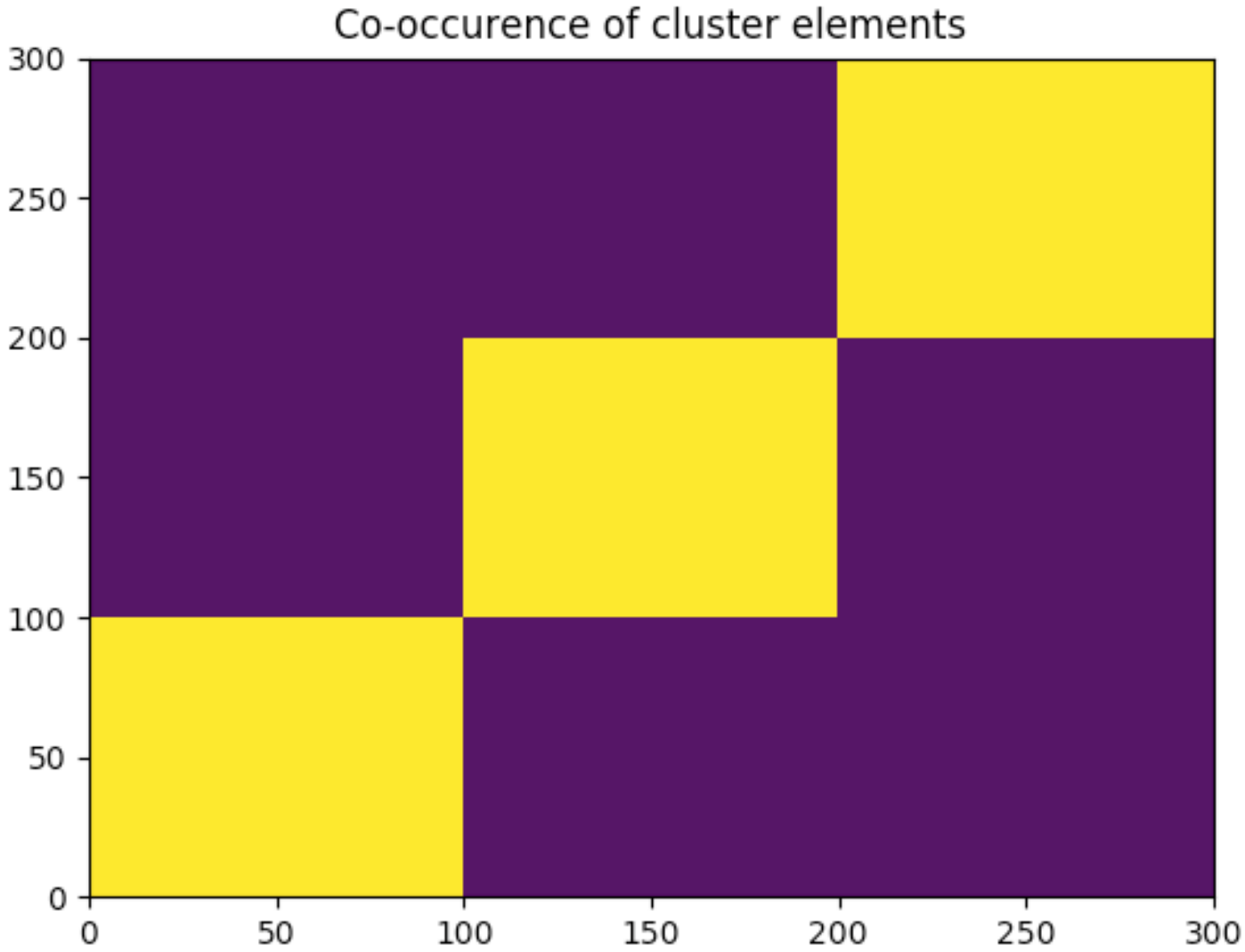


Co-occurrence

For all pairs (i,j) count how frequently i and j are in the same cluster.

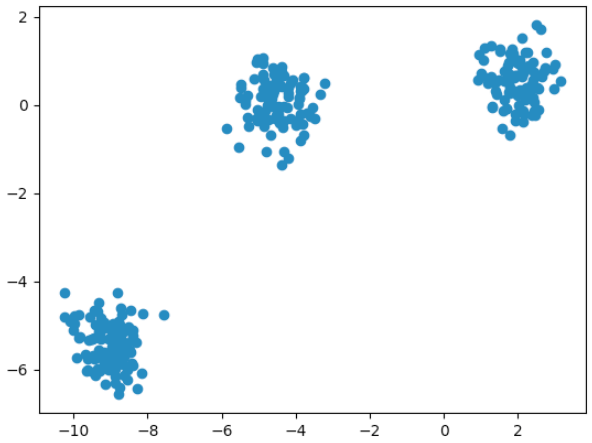


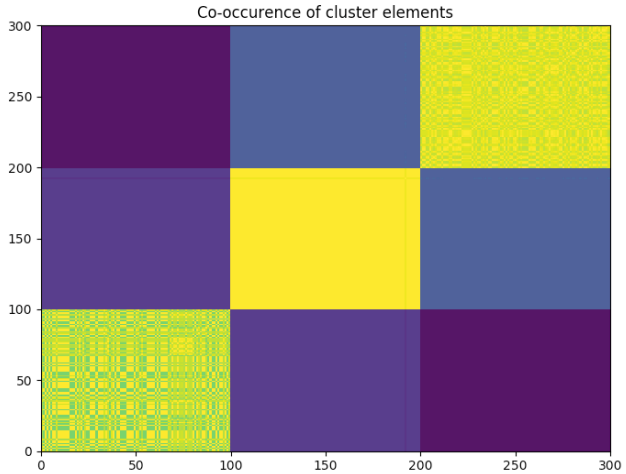
Co-occurrence



Stability over repetitions

- Clustering stable if (almost) always same points end up in the same clusters togethers (co-occurrence frequencies) from random initializations





Silhouette coefficient

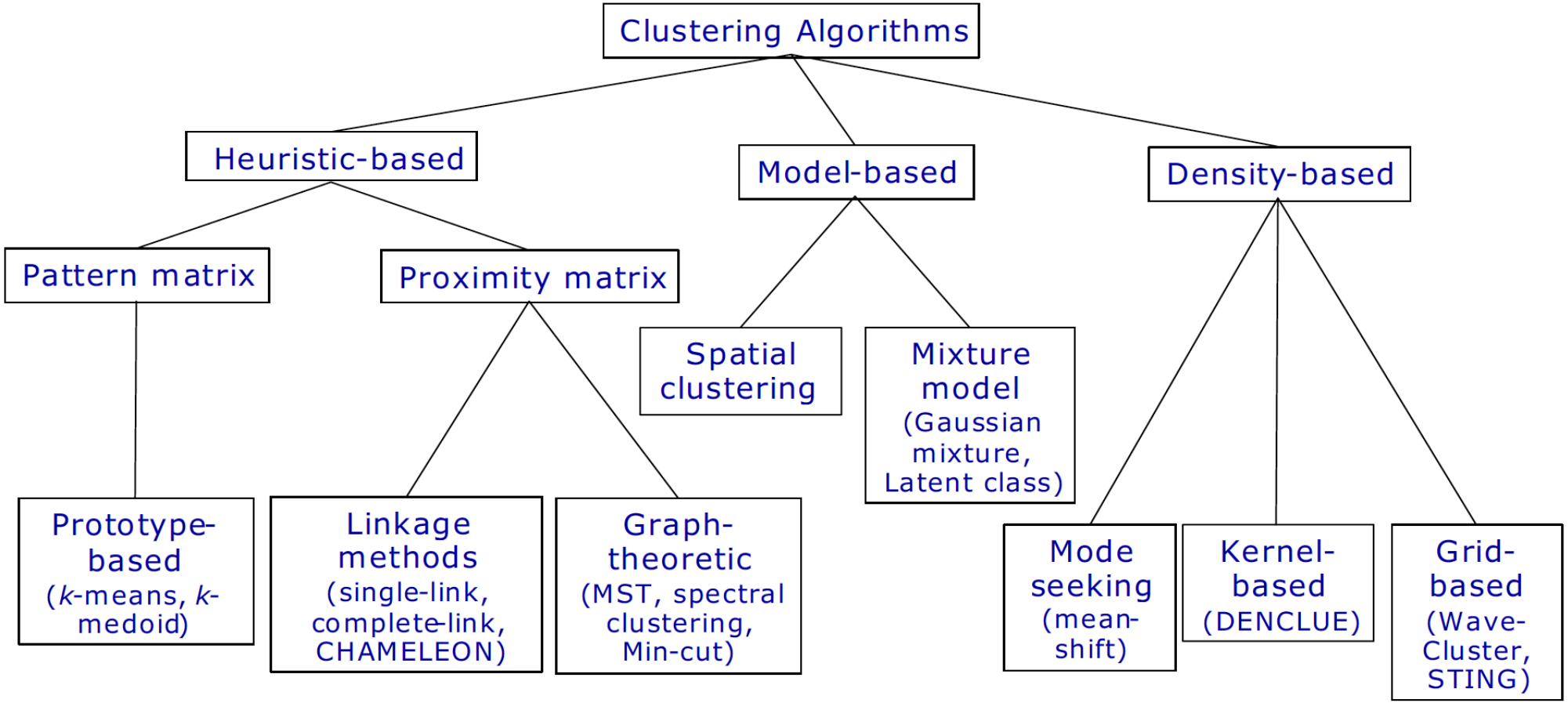
a: The mean distance between a sample and all other points in the same class.

b: The mean distance between a sample and all other points in the *next nearest cluster*.

$$s = \frac{b - a}{\max(a, b)}$$

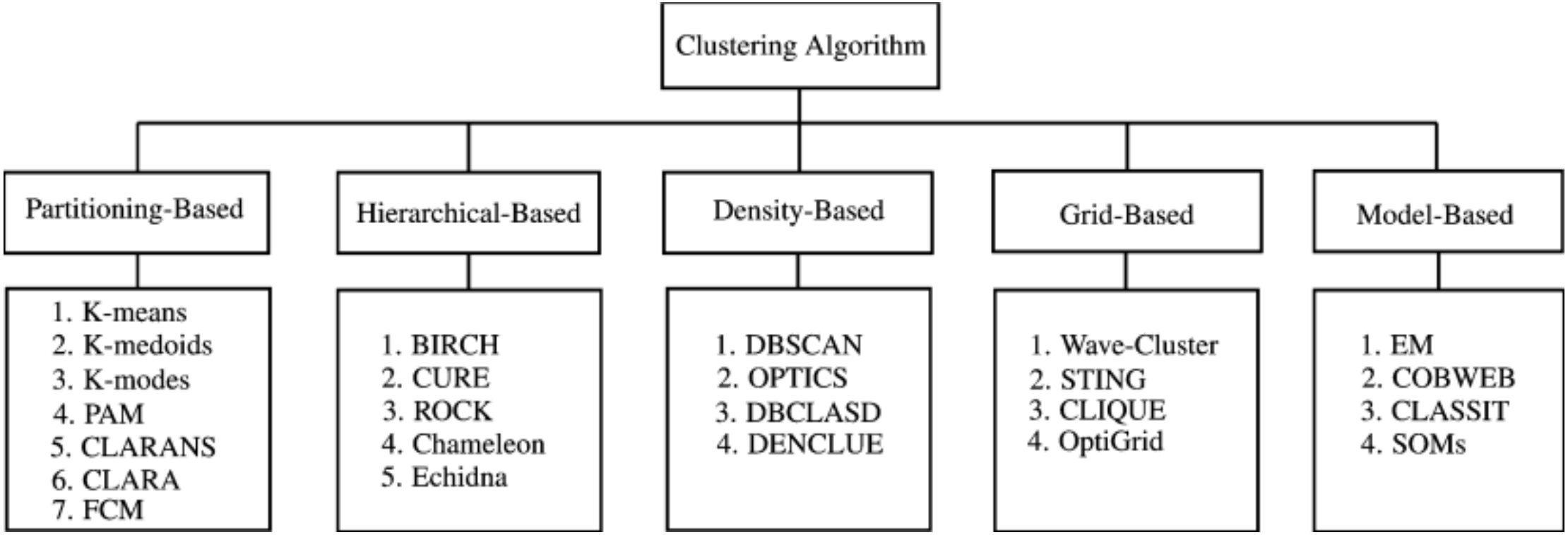
https://scikit-learn.org/stable/modules/clustering.html#silhouette-coefficient

Clustering clustering algorithms



Jian et al.. (2004) Landscape of Clustering Algorithms, Proc. 17th Int. Conf. on Pattern Recognition (ICPR'04)

Clustering clustering algorithms



Fahad et al. (2014) IIEEE Trans. Emerging Topics in Computing, volume 2, 267-279

Combining clustering and classification

Useful idea when labeling is expensive

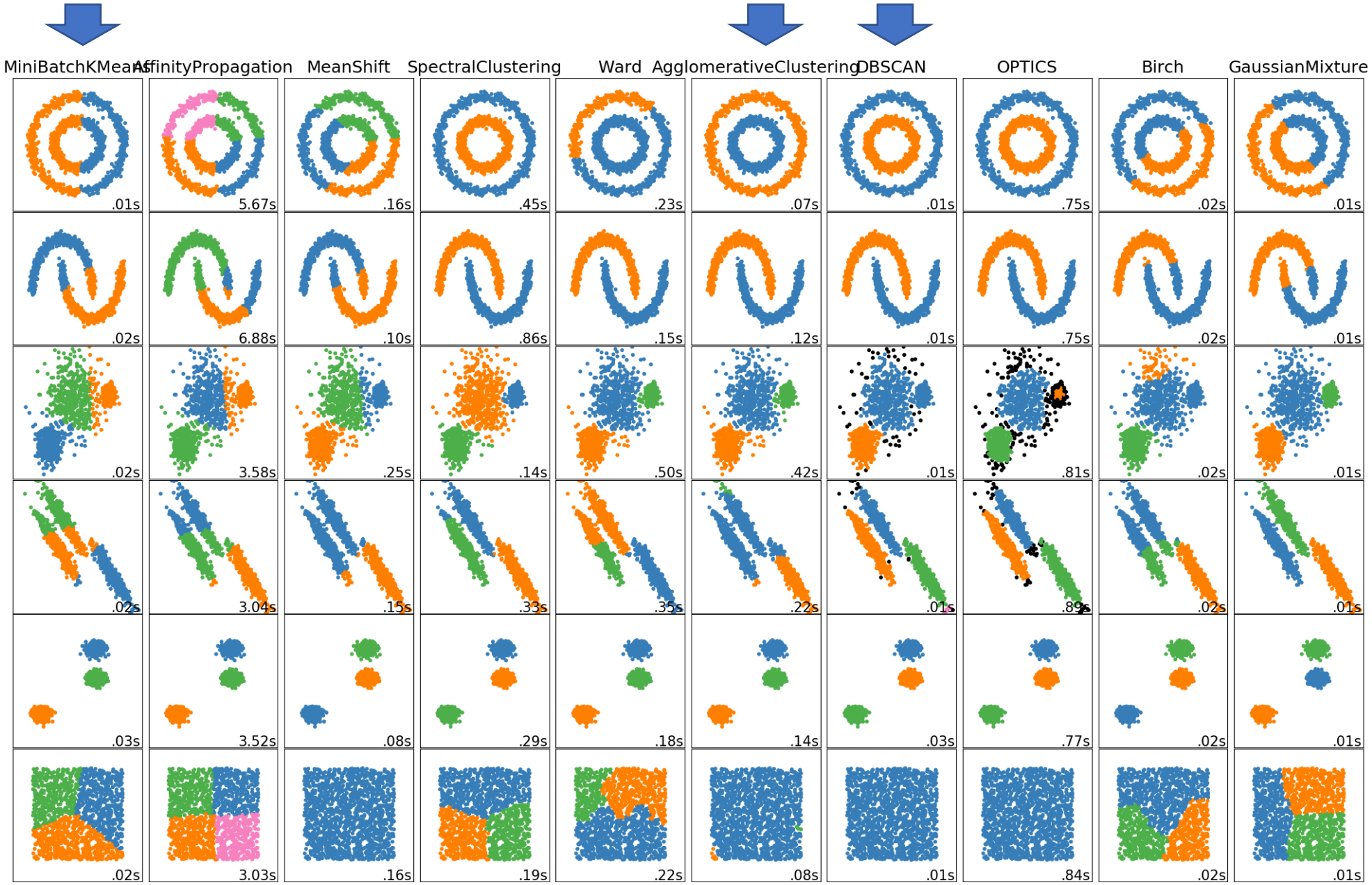
- Take a dataset with handwritten digits
- Provide only one label per digit (10 labels for the whole dataset)
- Use 10-means with the ten labeled images as starting points for clustering the whole dataset.
- Then use 1nn for classifying new handwritten digits.

Some reflections on clustering

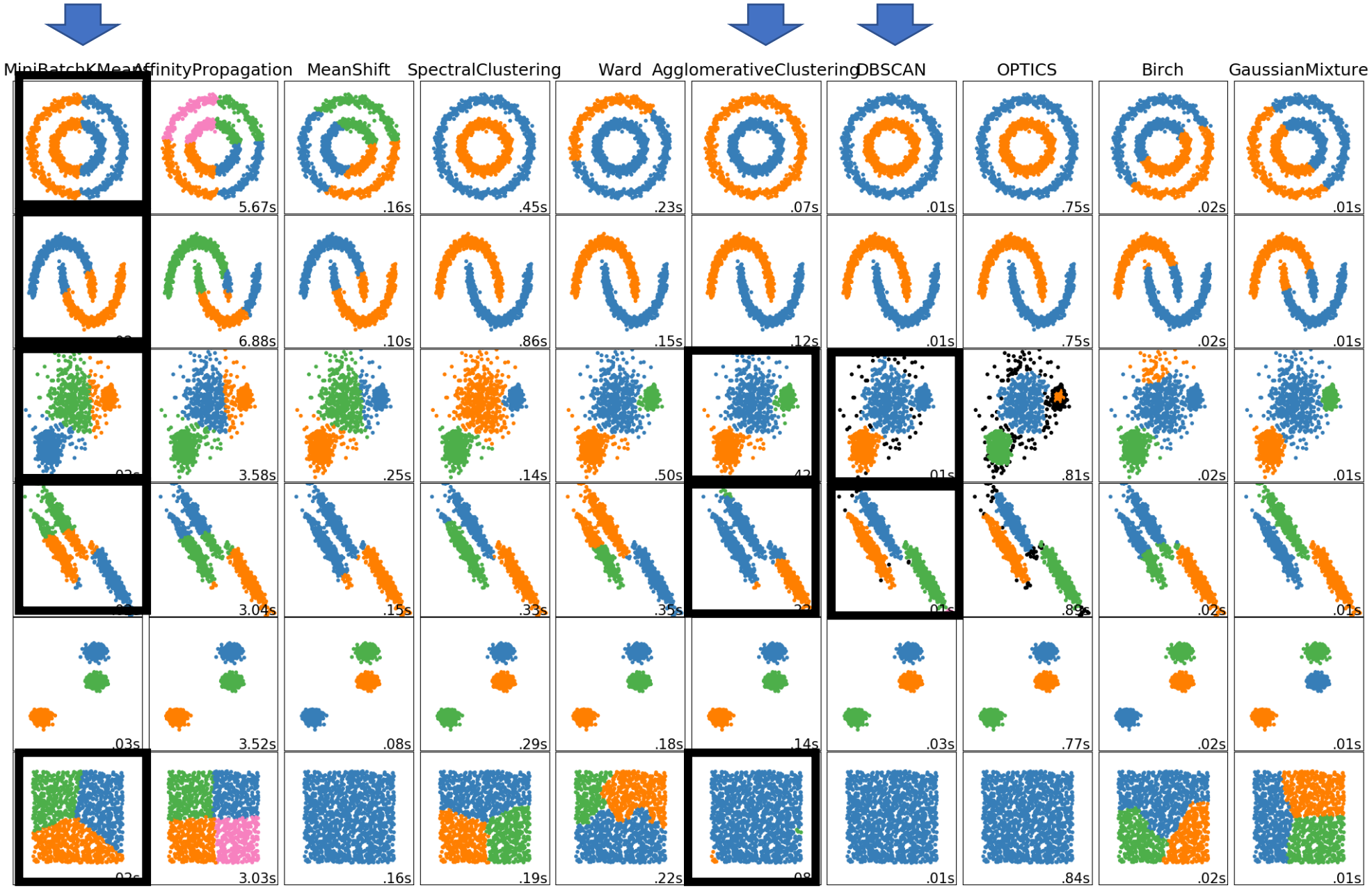
Clustering is successful, but difficult

- Inherent vagueness in the definition of a cluster
- Can be difficult to define an appropriate similarity measure

Jain, A.K. (2010) Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, **31**, 651-666



https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html



Are the framed cases as desired?

Questions about clustering

a) What is a cluster?
b) What features should be used?
c) Should the data be normalized?
d) Does the data contain any outliers?
e) How do we define the pair-wise similarity?
f) How many clusters are present in the data?
g) Which clustering method should be used?
h) Does the data have any clustering tendency?
i) Are the discovered clusters and partition valid?

Jain, A.K. (2010) Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, **31**, 651-666