

MODULE 3: CLUSTERING

DAT405 / DIT407, 2022-2023, READING PERIOD 2

Core data science tasks

- Regression
 - Predicting a numerical quantity
- Classification
 - Assigning a label from a discrete set of possibilities
- Clustering
 - Grouping items by similarity



CLUSTERING

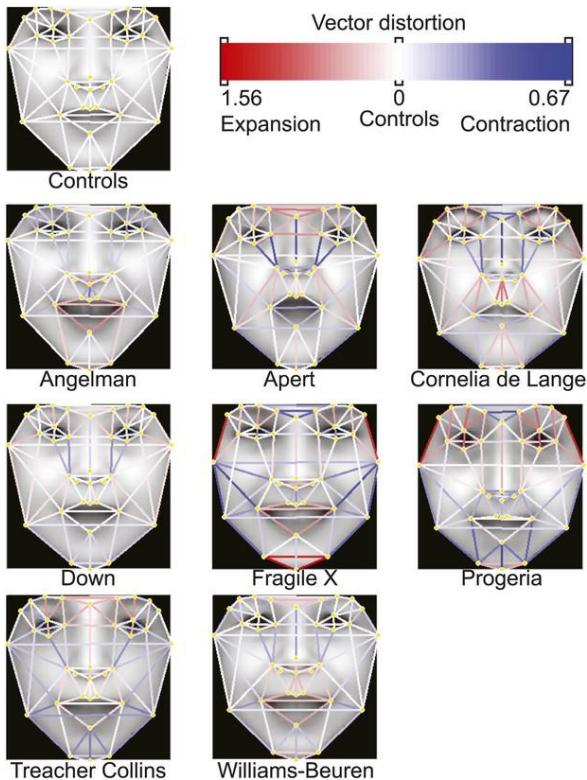
- Grouping items by similarity

Clustering

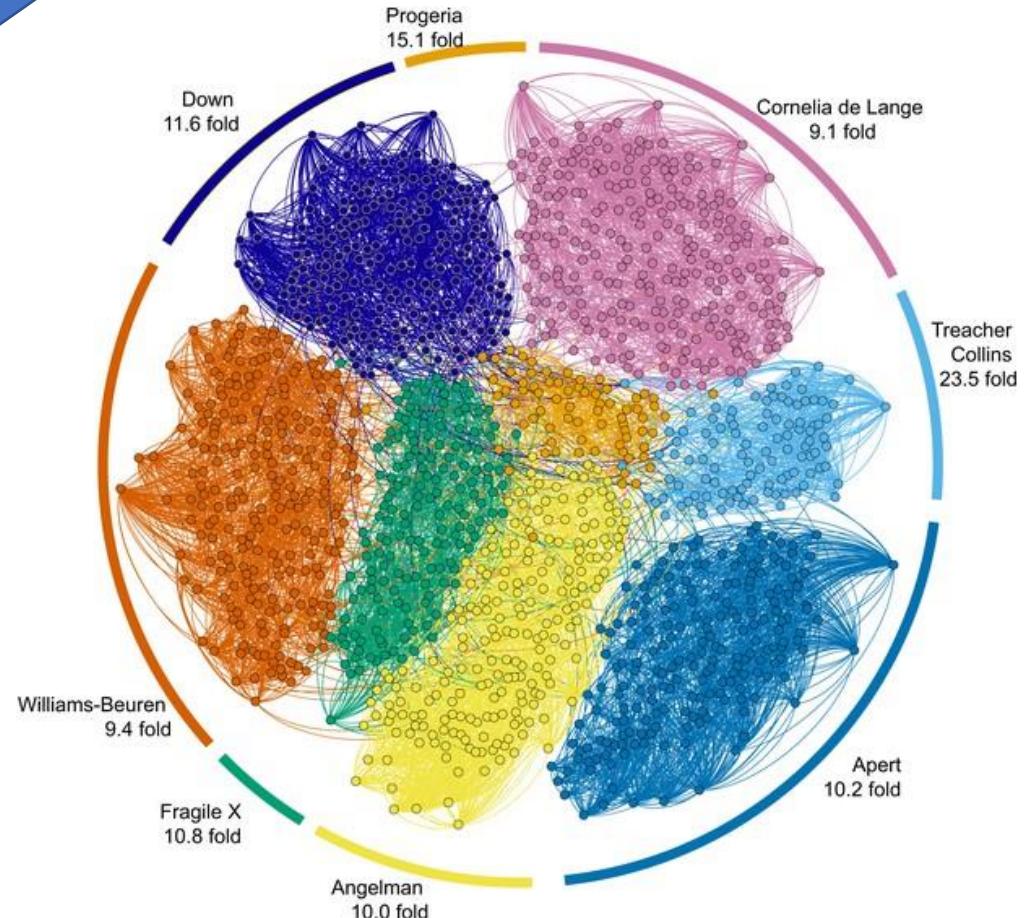
Purpose of data clustering

- Underlying structure
 - to gain insights into data, generate hypotheses and identify salient features
- Natural classification
 - to identify the degree of similarity among forms or organisms (phylogenetic relationships)
- Compression
 - to organise data and summarise it through cluster prototypes

Clustering faces

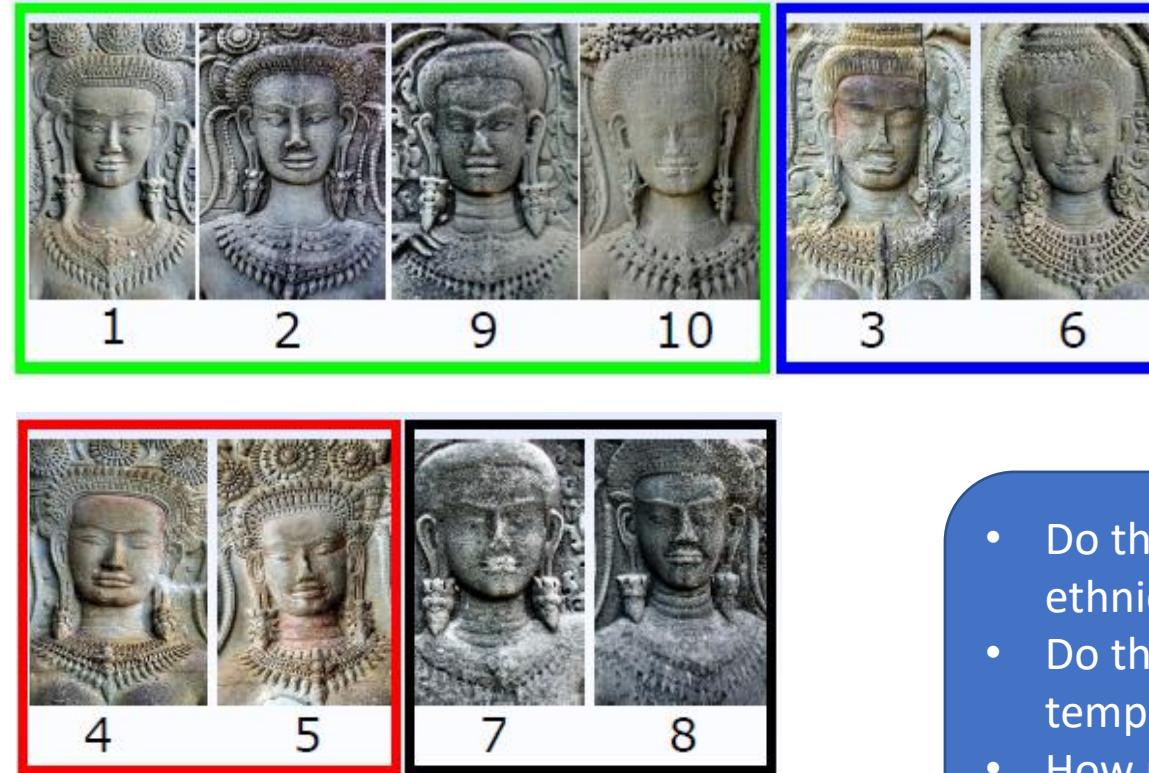
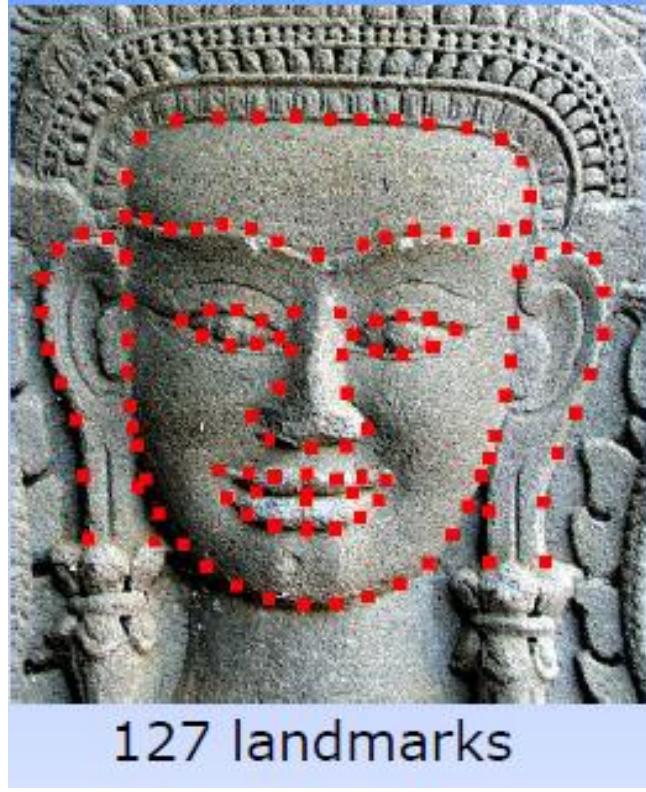


Helps discovering and diagnosing genetic disorders



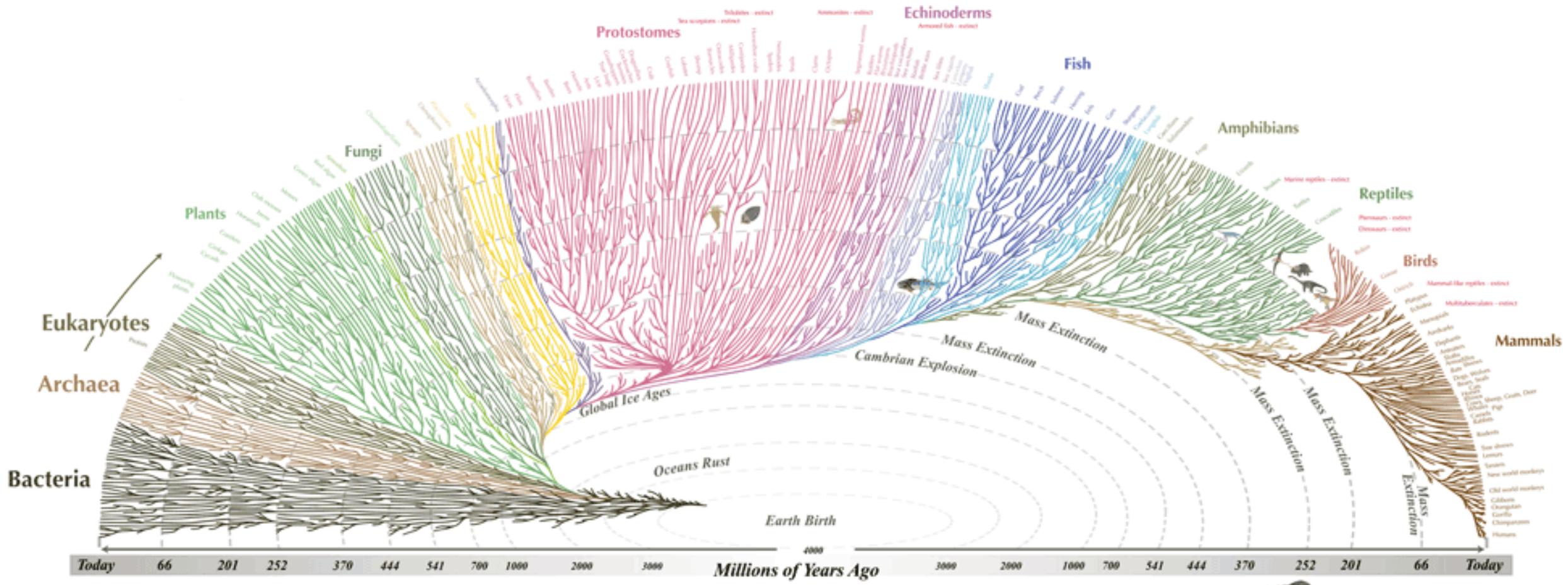
Clustering statues

Helps ethnologists understand cultural influences at Angkor Wat via its 2000 statues



- Do they represent different ethnic groups?
- Do their location in the temple have meaning?
- How many sculptors created the carvings?

Tree of life



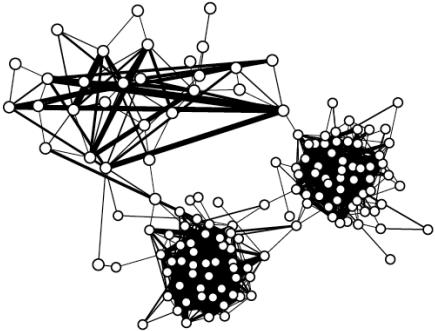
All the major and many of the minor living branches of life are shown on this diagram, but only a few of those that have gone extinct are shown. Example: **Dinosaurs - extinct**

© 2008, 2017 Leonard Eisenberg. All rights reserved.
evogeneao.com

Clustering of 770,000 genomes reveals post-colonial population structure of North America

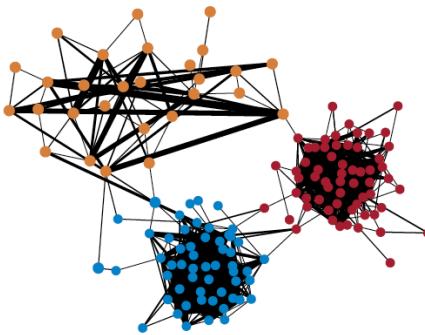
a

Construct network from IBD.
Join vertex pairs (genotyped samples) if $\text{IBD} > 12 \text{ cM}$.
Edge weights are a function of total detected IBD.



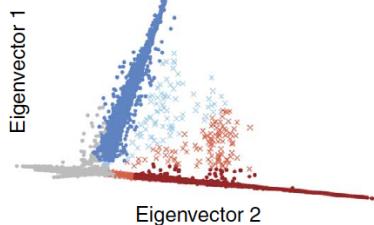
b

Detect network clusters.
Recursively identify disjoint sets that maximize the modularity of the network. (Here one level of clustering hierarchy is shown.)



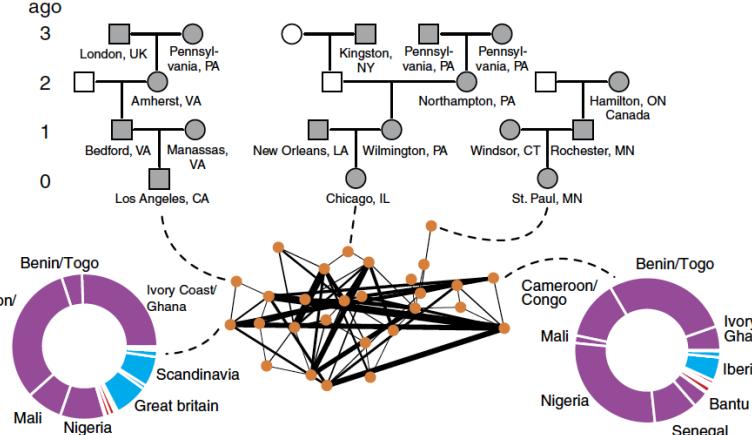
c

Identify subsets of the clusters that separate in the spectral embedding. Spectral embedding is computed from eigen-decomposition of Laplacian matrix. In the plot below, we identify "stable subsets" (filled circles) of the blue and red clusters.



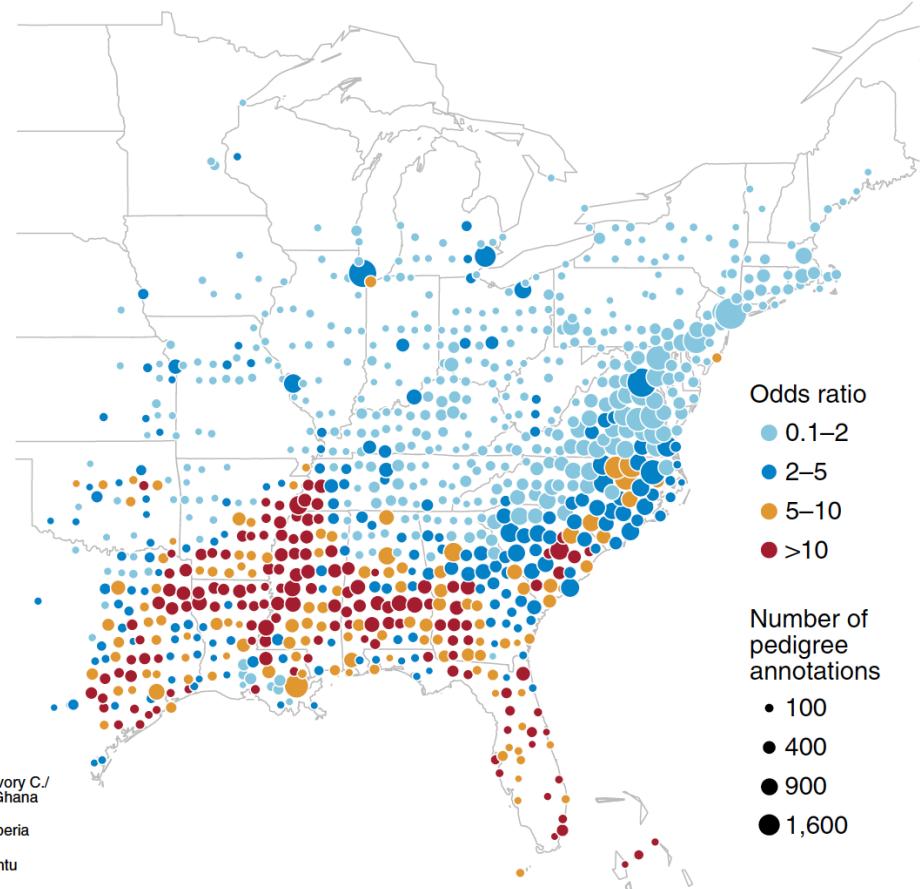
d

Annotate each cluster with two kinds of data:
• In all samples, global admixture of 20 populations (donut charts);
• For some samples, birth locations of ancestors in pedigrees.



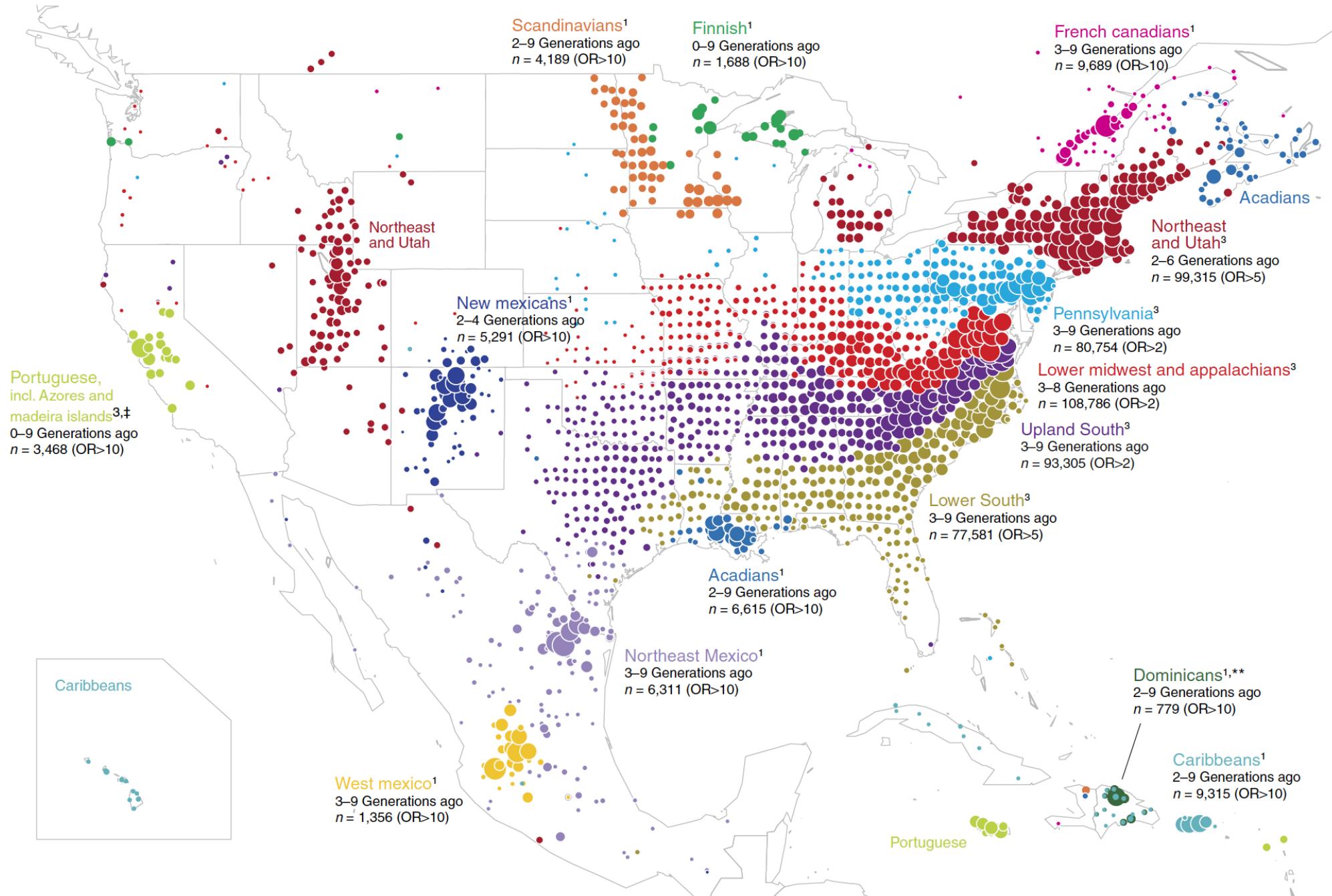
e

Visualize geographic distribution of ancestral birth locations in each cluster.
Map below shows birth locations of ancestors in the African American cluster. Locations are colored by degree of over-representation (odds ratio), and scaled by number of birth location annotations.



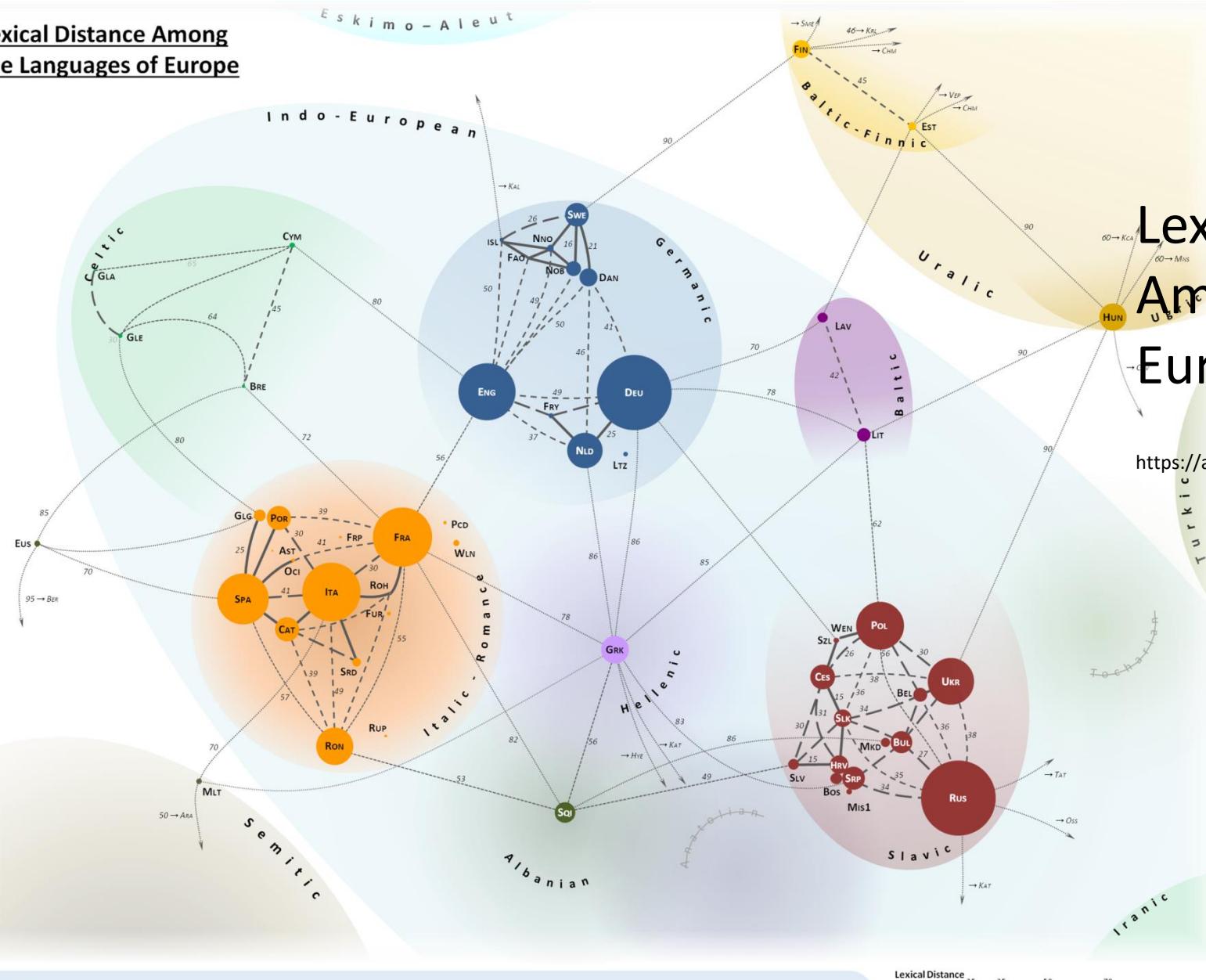
Clustering of 770,000 genomes reveals post-colonial population structure of North America

- “computational methods reveal densely connected clusters, in which the members of each cluster are subtly more related to each other.”
- “we annotate these densely connected clusters to identify the putative historical origins of such population substructure, and to infer temporal and geographic patterns of migration and settlement”



Clustering of 770,000 genomes. Han et al. Nature Comm. 2016

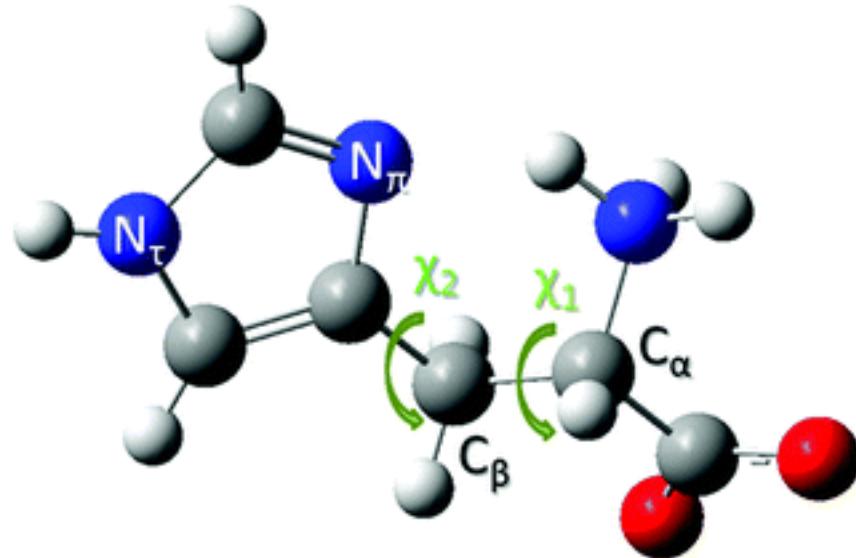
Lexical Distance Among the Languages of Europe



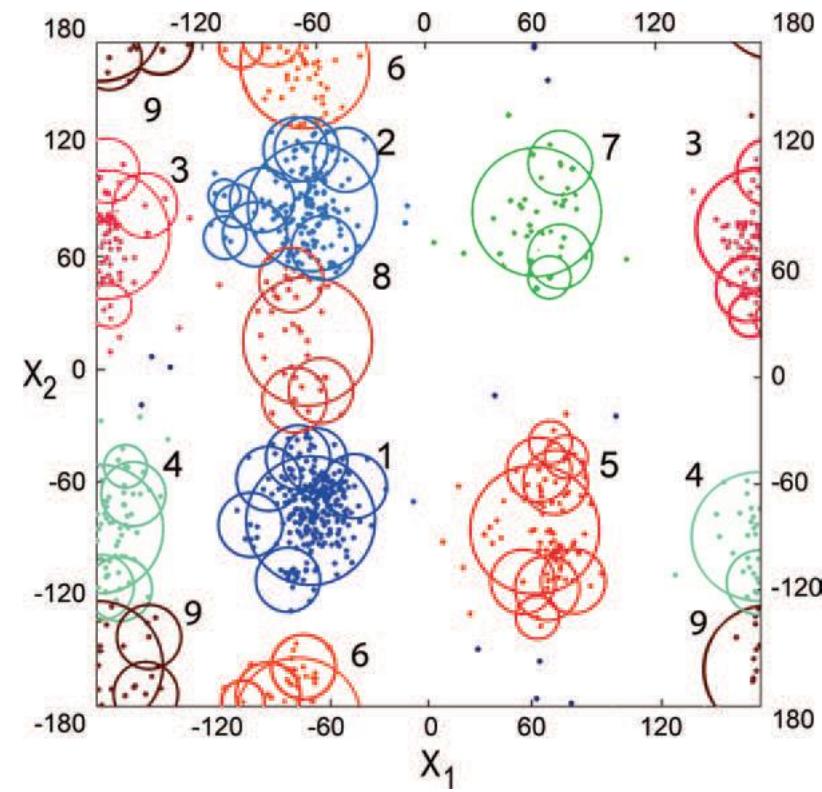
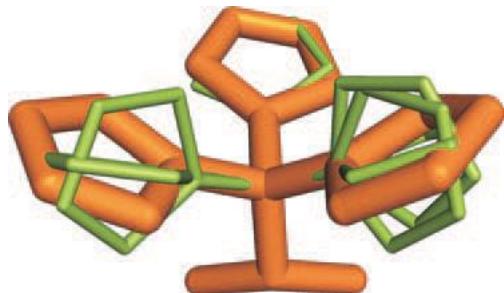
Lexical Distance Among Languages of Europe 2015

<https://alternativetransport.wordpress.com/2015/05/05/34/>

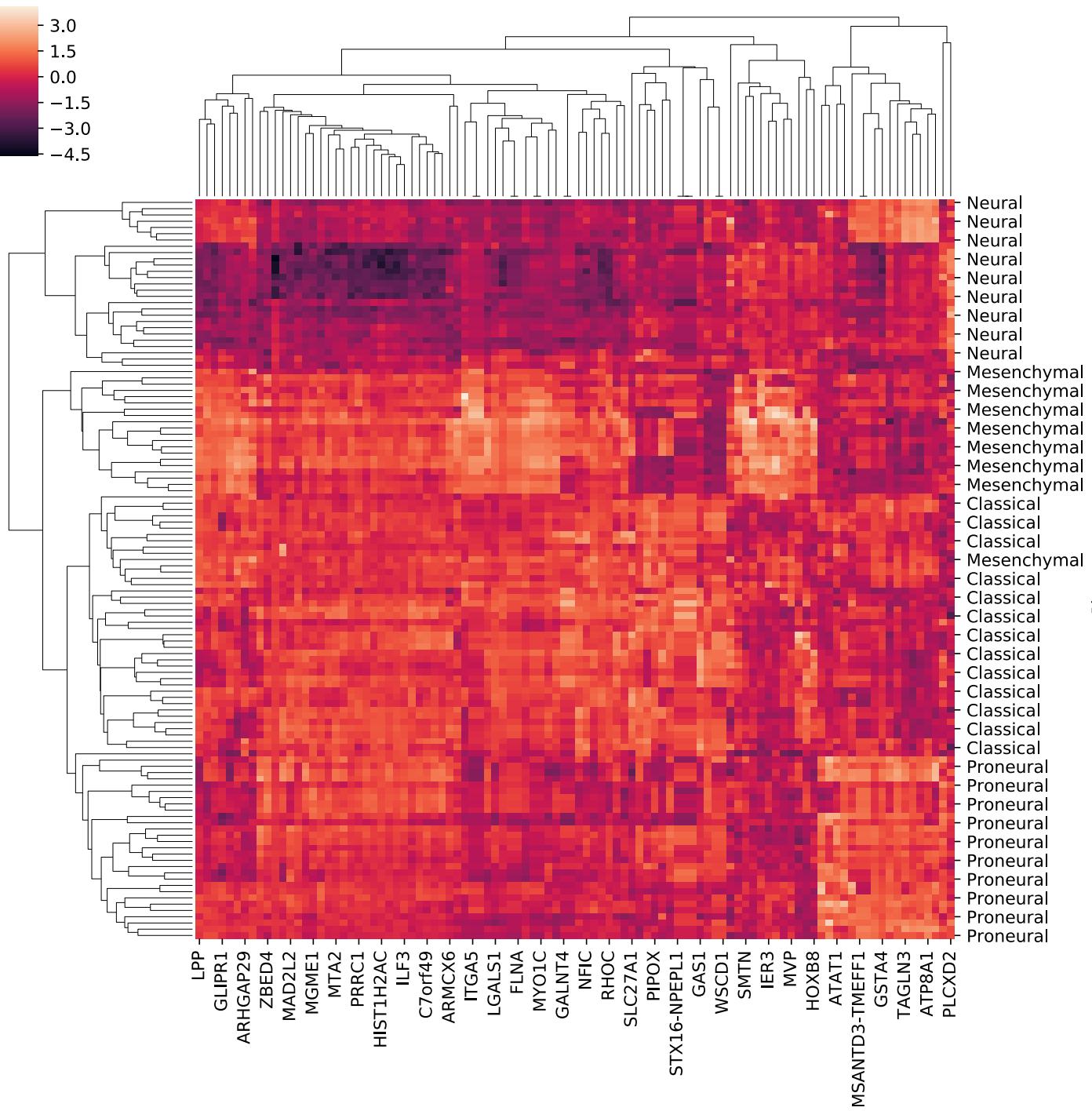
Protein side chain rotamers



Cardamone et al (2016),
Phys. Chem. Chem. Phys.,
18, 27377-27389



Kirys et al. (2012), Proteins., 80(8):2089-98. doi: 10.1002/prot.24103



Bi-clustered 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

Customer/market segmentation

- Demographic
 - e.g. age, gender, income, education, family size/situation
- Geographic
 - e.g. postcode, region, city, country
- Psychographic
 - e.g. lifestyle, personality characteristics
- Behaviourial
 - e.g. purchase frequency, customer loyalty, engagement with website

Clustering methods

We shall look at three clustering approaches that build on quite different ideas.

- K-means clustering
 - Partition into k clusters (global method)
- DBSCAN clustering
 - Start at a point and build a cluster by joining its neighbors (local method)
- Hierarchical clustering
 - Build bigger clusters by joining smaller clusters that are close together (global method).

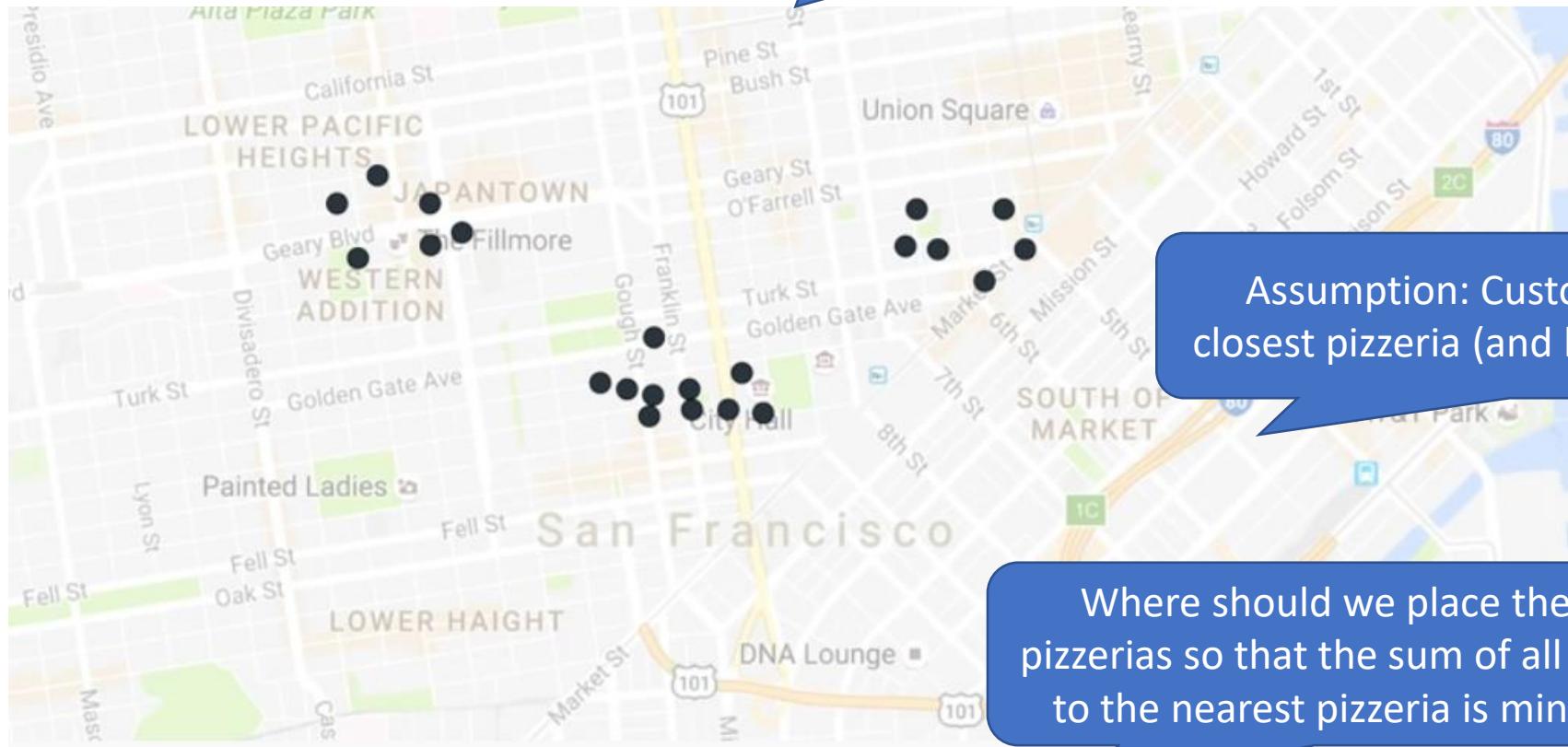
K-means clustering

Pizzerias in San Francisco



Problem

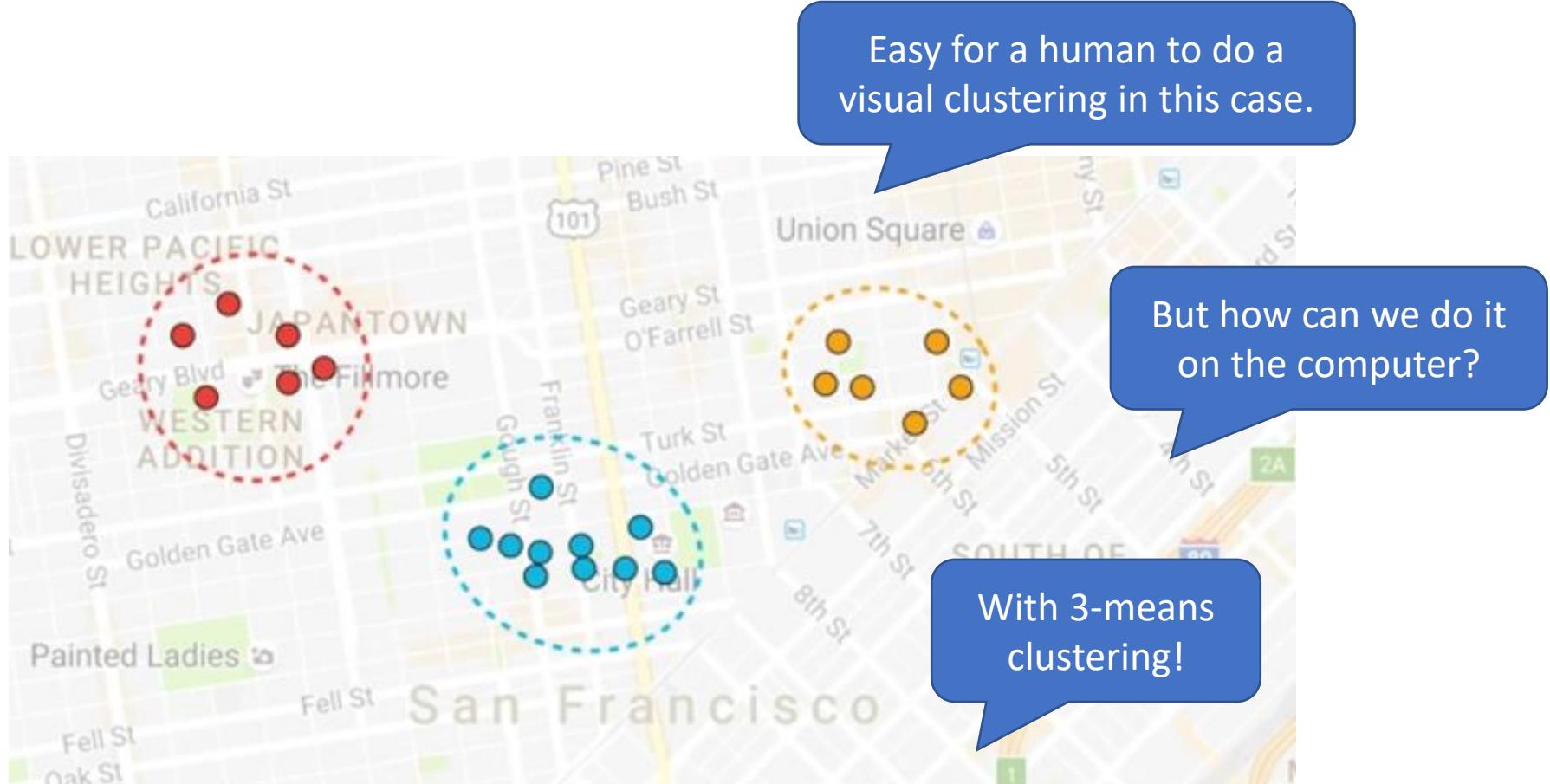
We want to start three pizzerias
that serve these customers



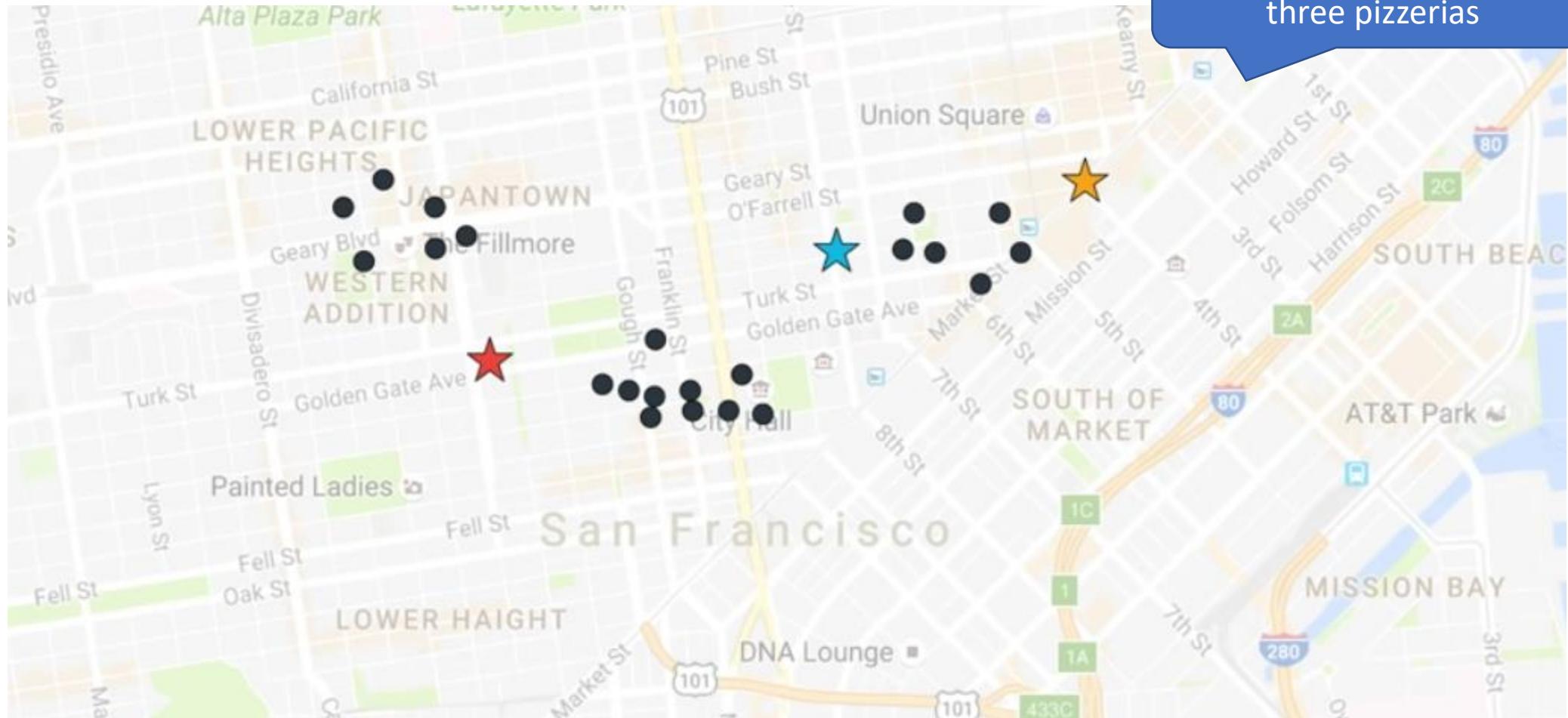
Assumption: Customers will go to the closest pizzeria (and buy more if it is close).

Where should we place the three pizzerias so that the sum of all distances to the nearest pizzeria is minimised?

Visual clustering

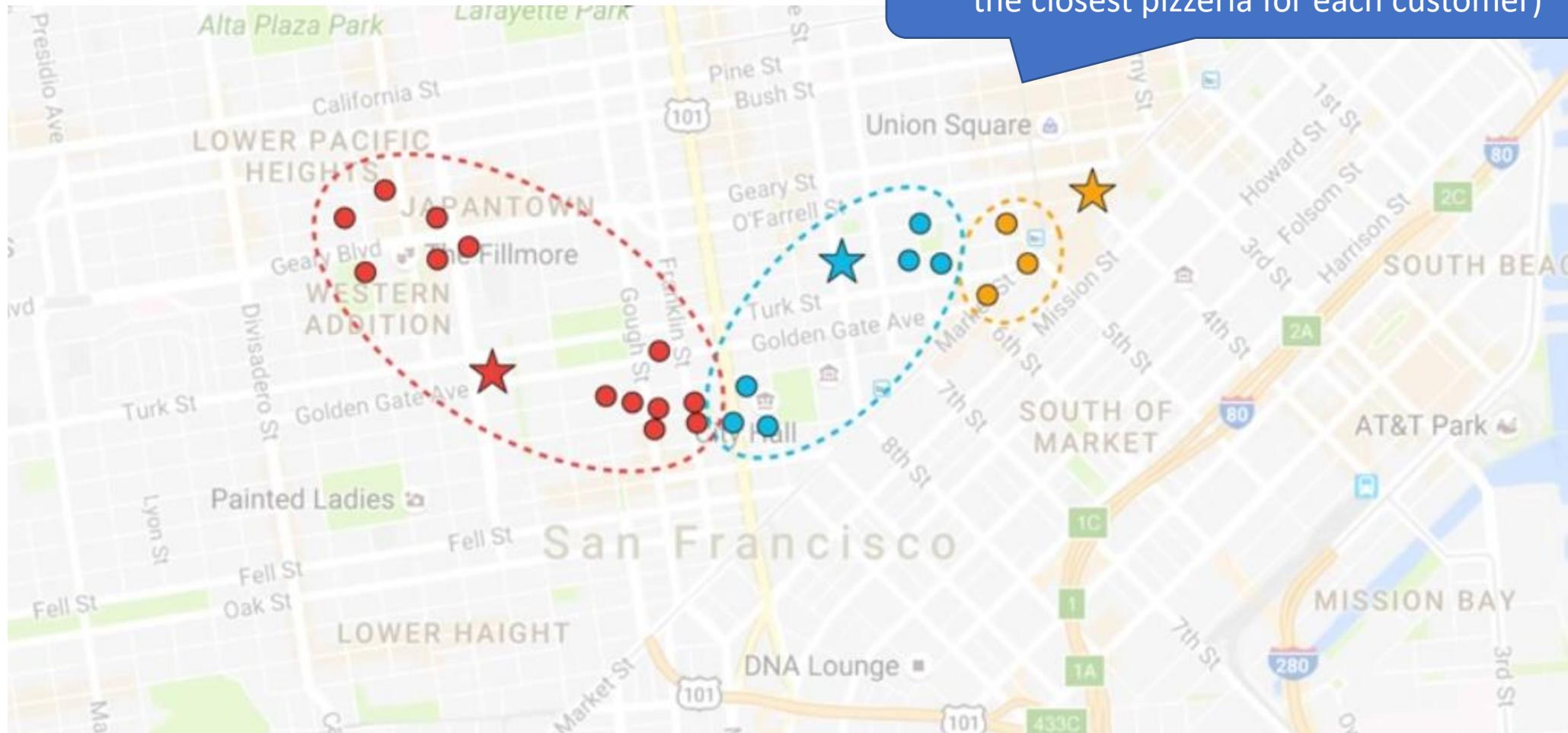


Initialize the centroids



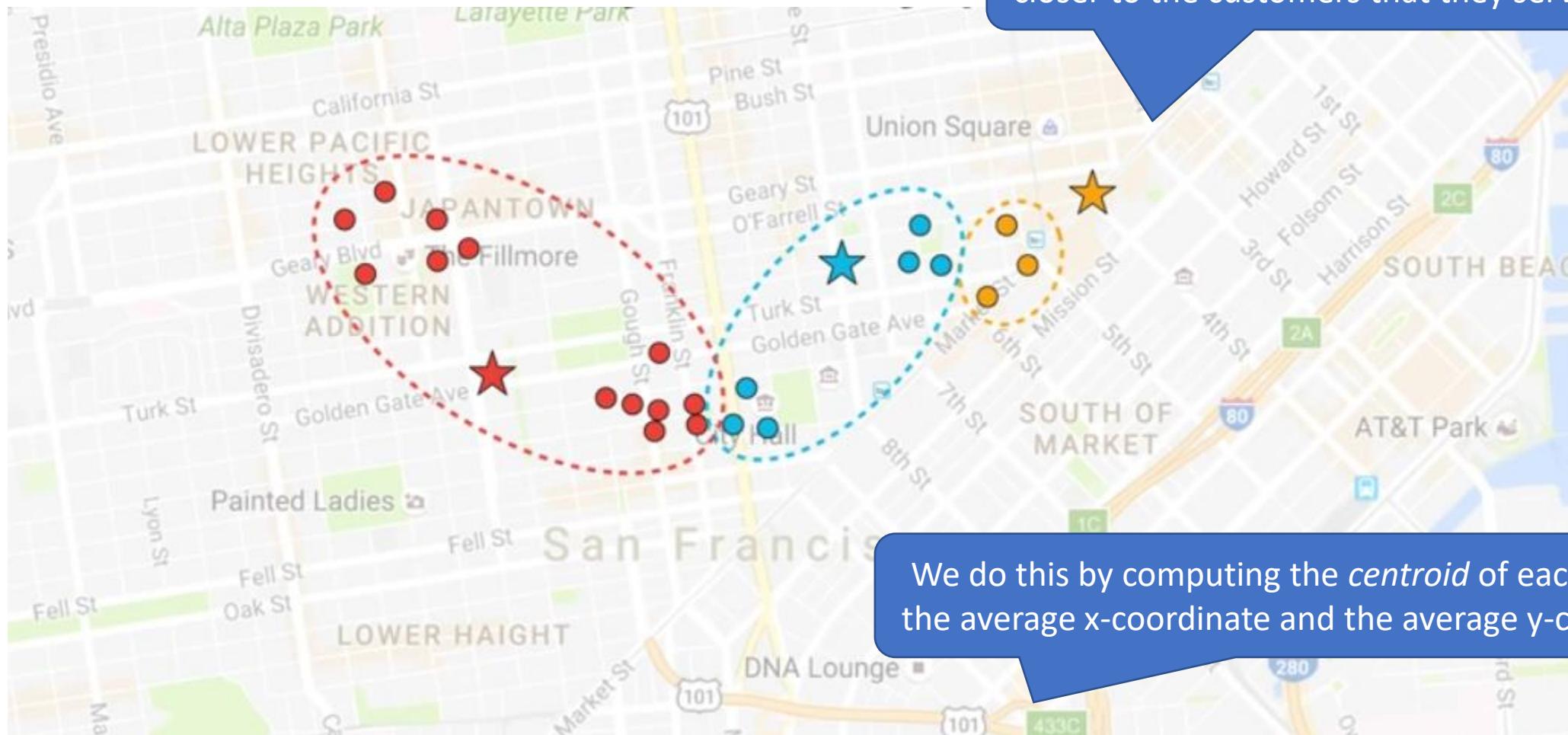
Form the clusters

Then these clusters are generated (by finding the closest pizzeria for each customer)



Move the centroids

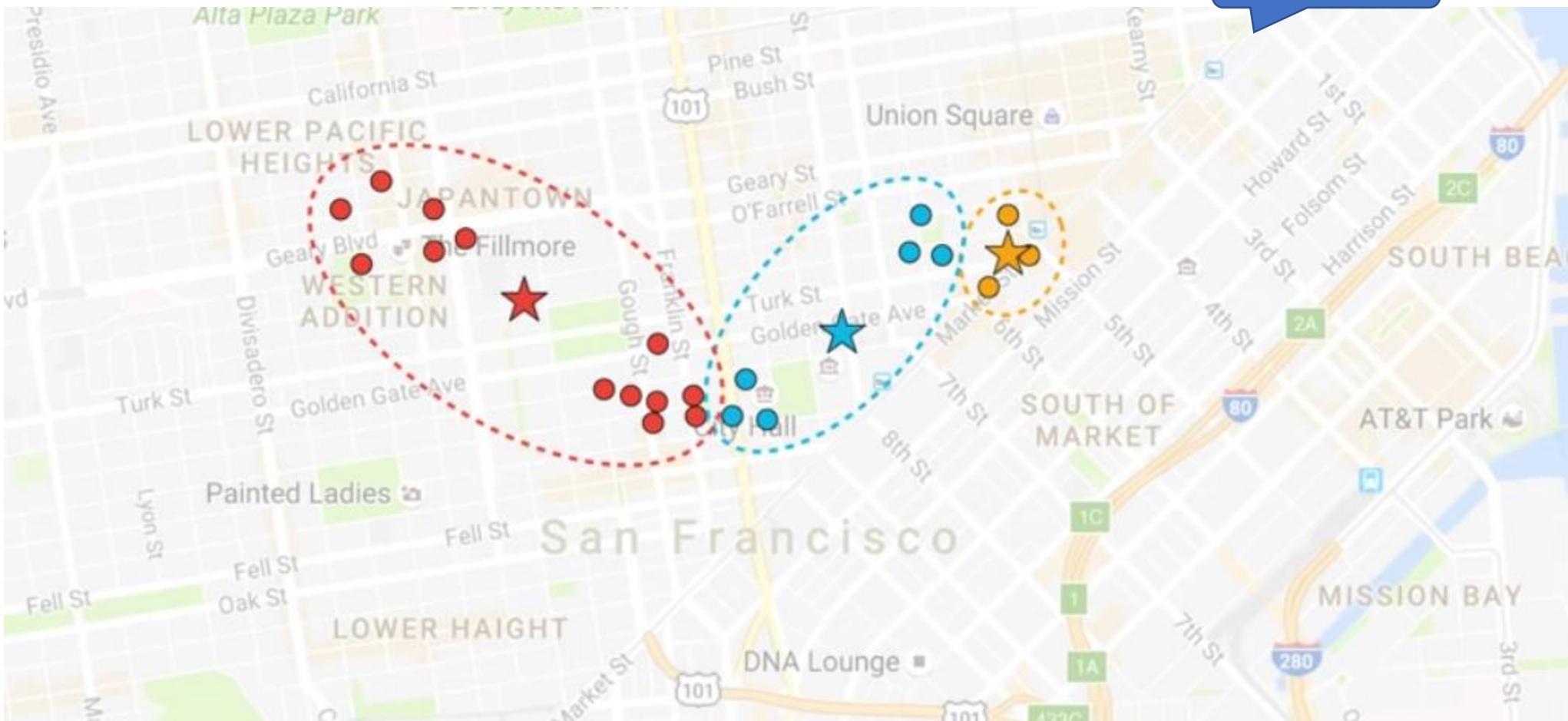
Move the pizzerias so that they are closer to the customers that they serve.



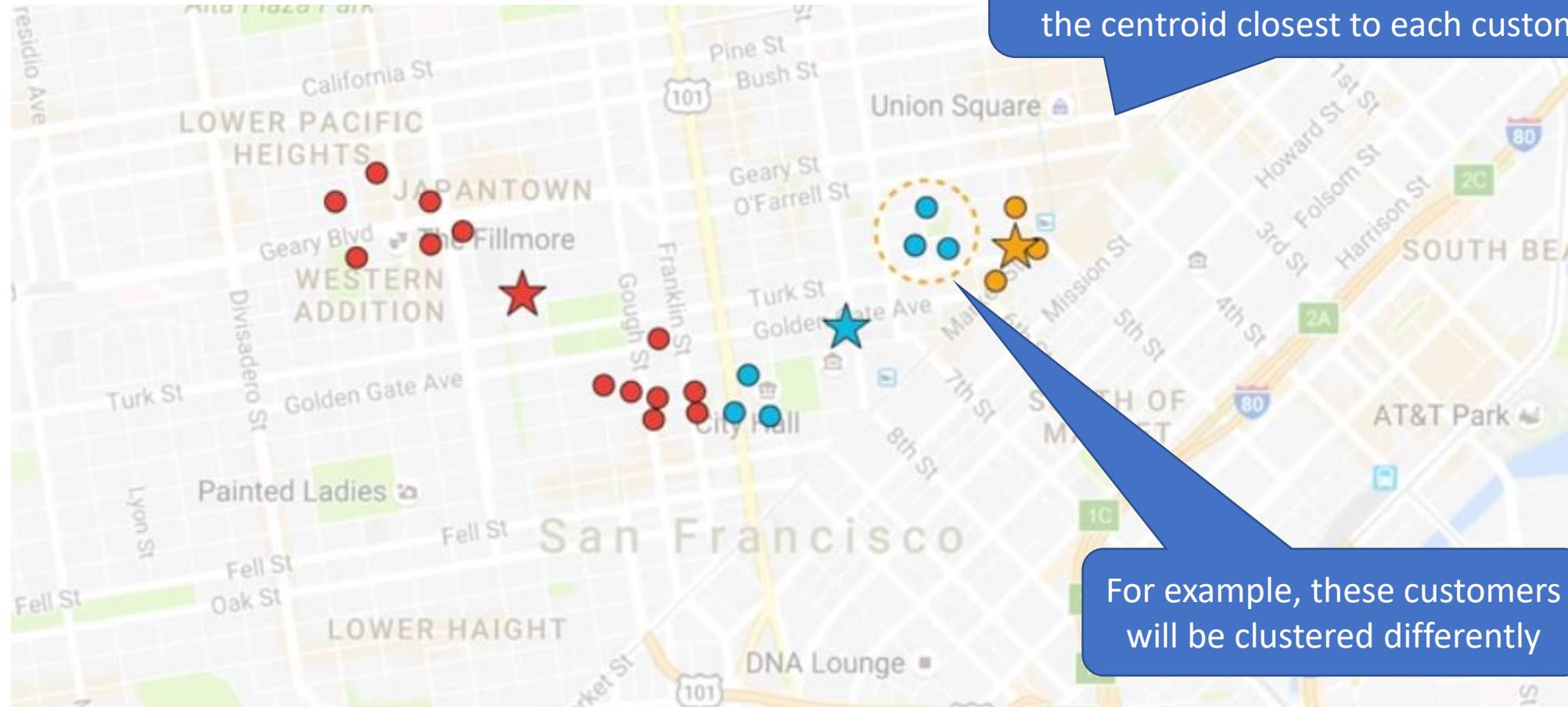
We do this by computing the *centroid* of each cluster:
the average x-coordinate and the average y-coordinate

Move the centroids

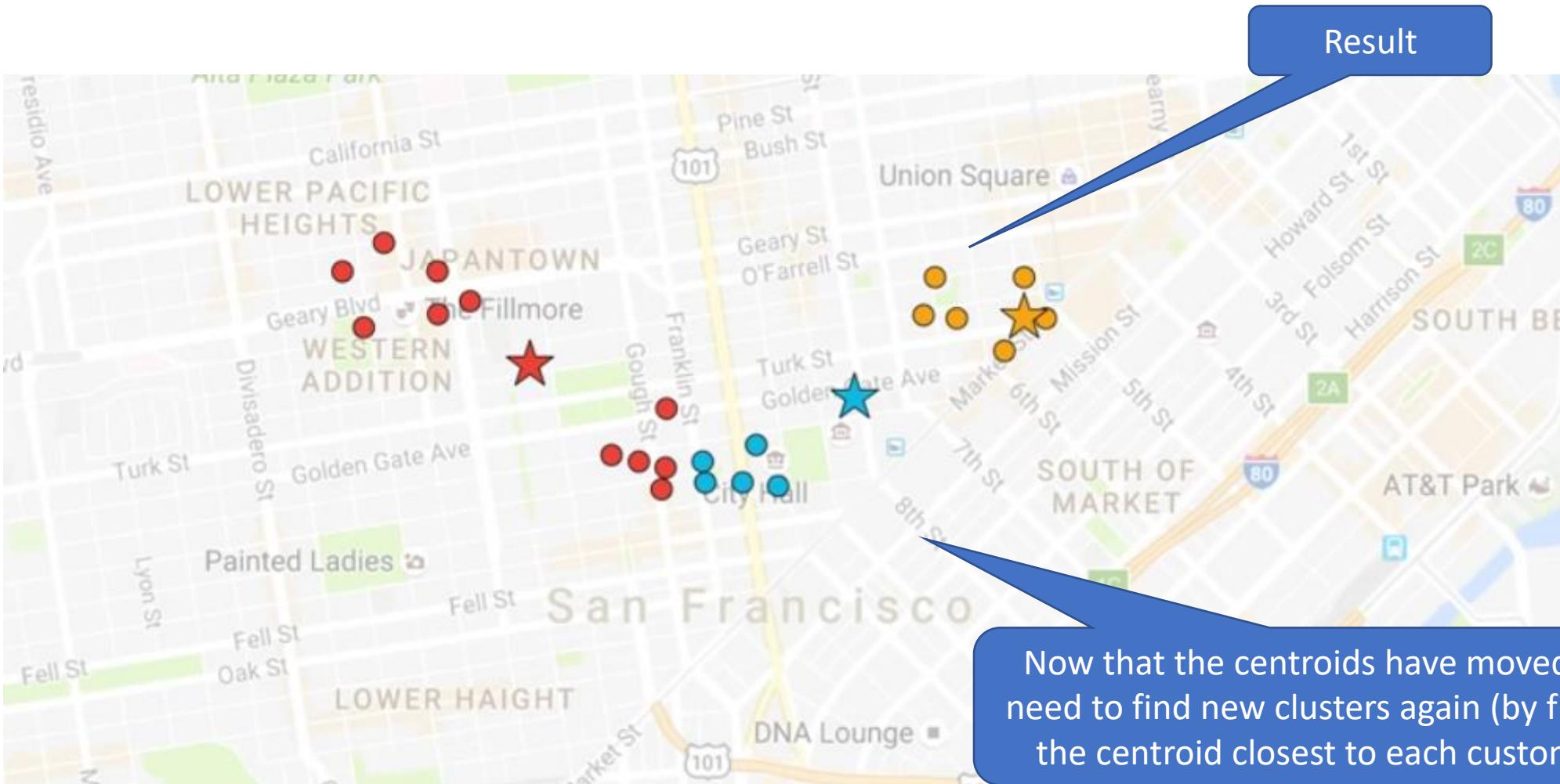
Result



Find the clusters

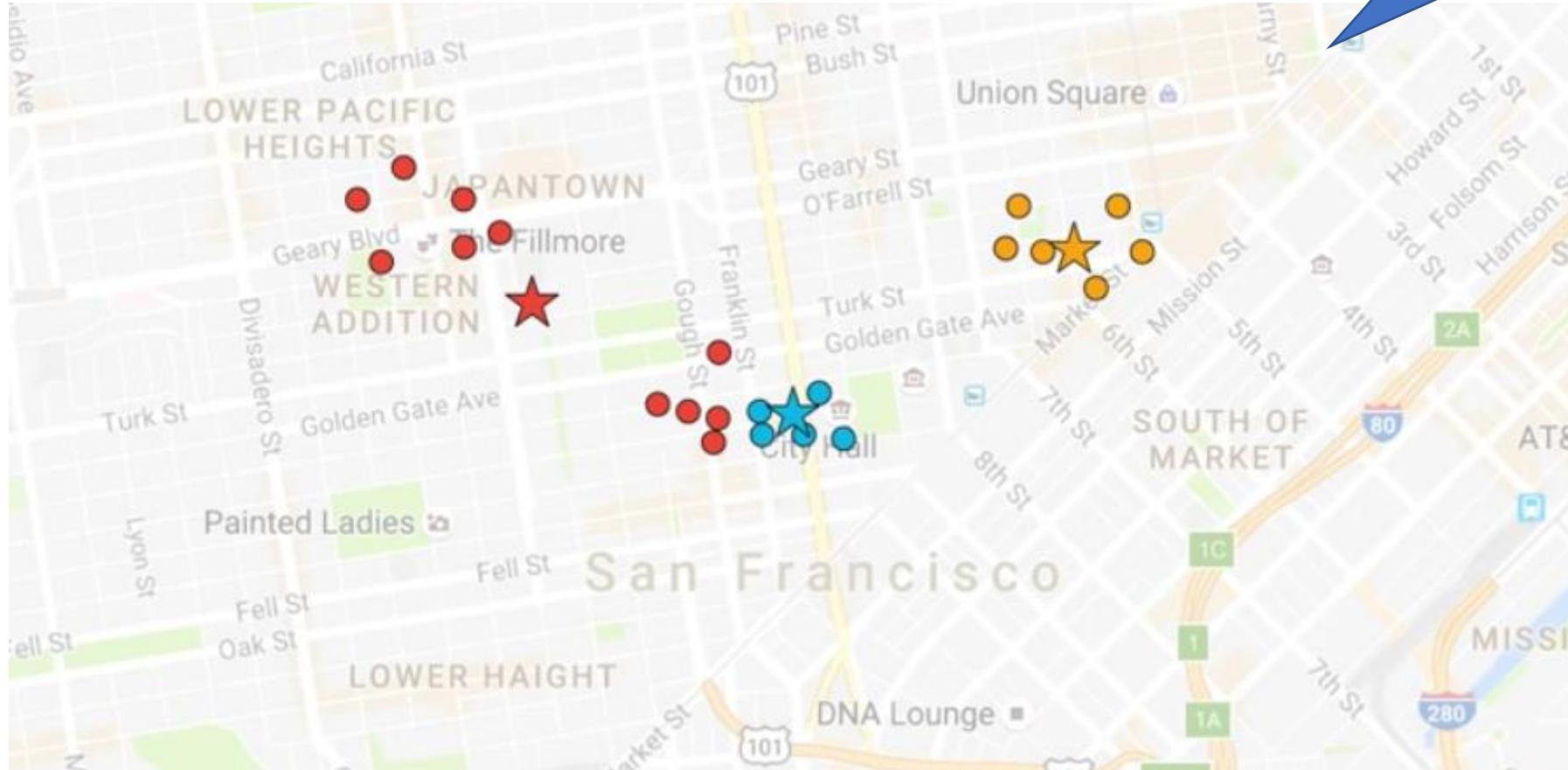


Find the clusters

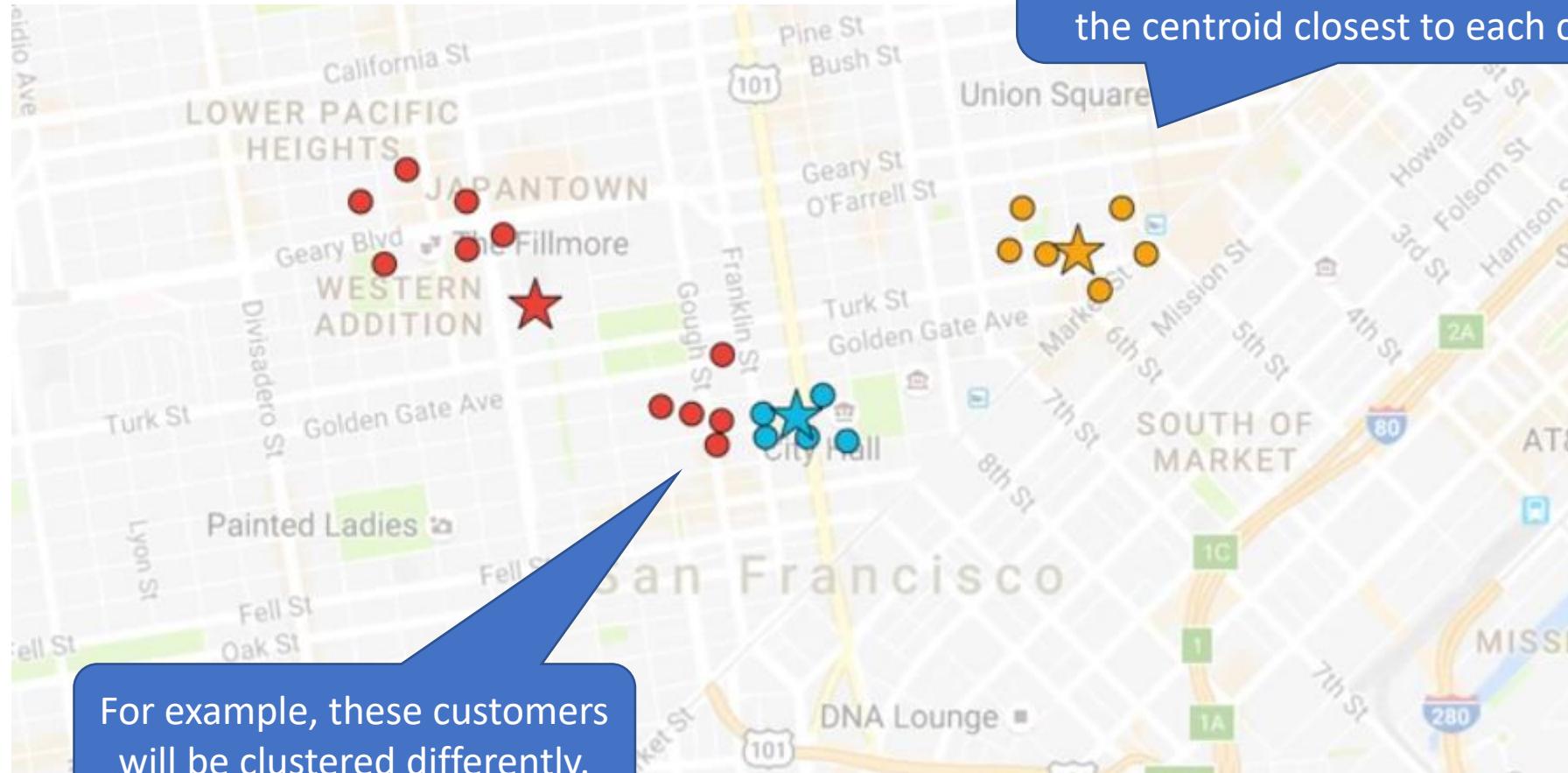


Move the centroids

Result

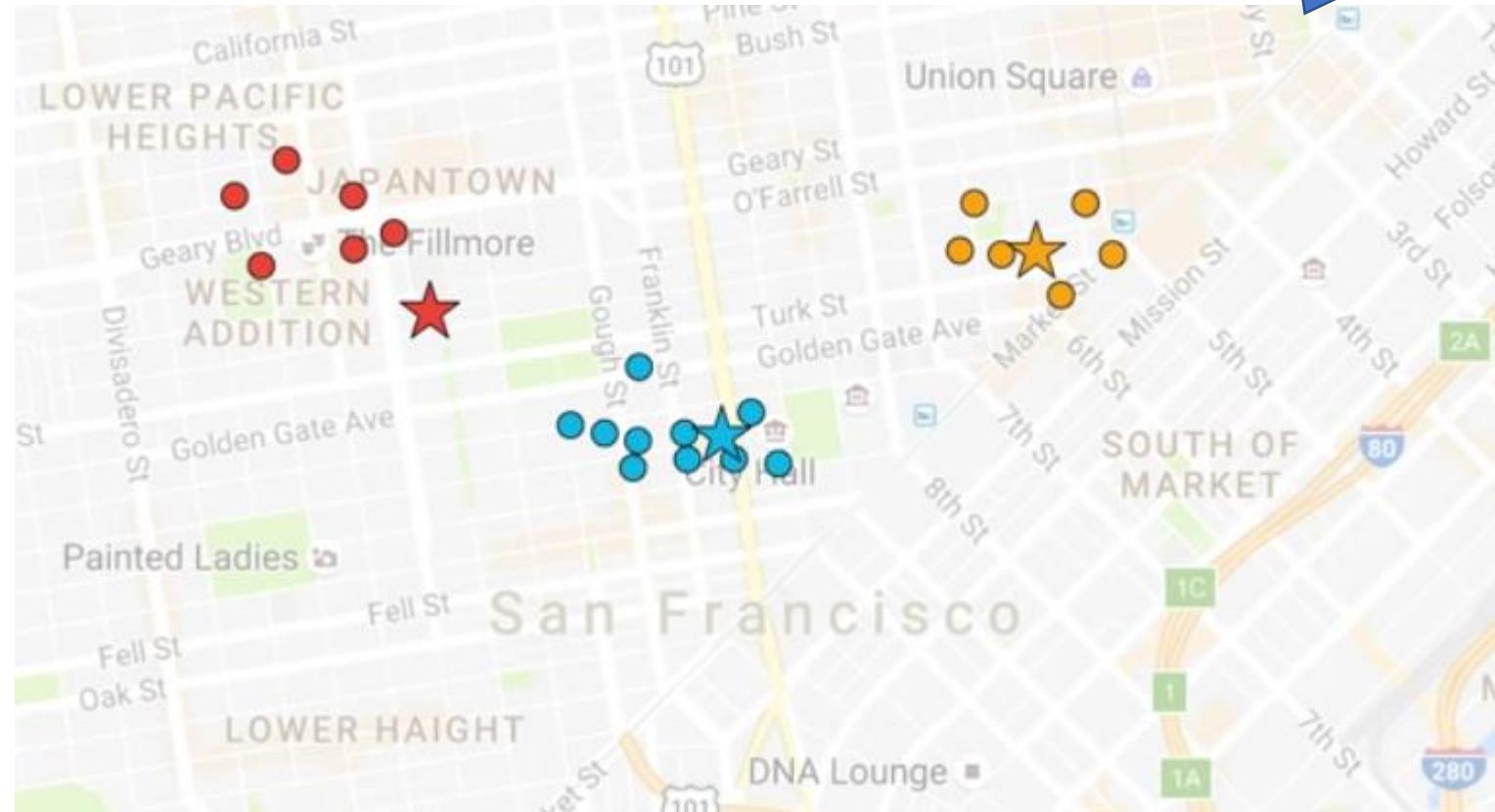


Find the clusters

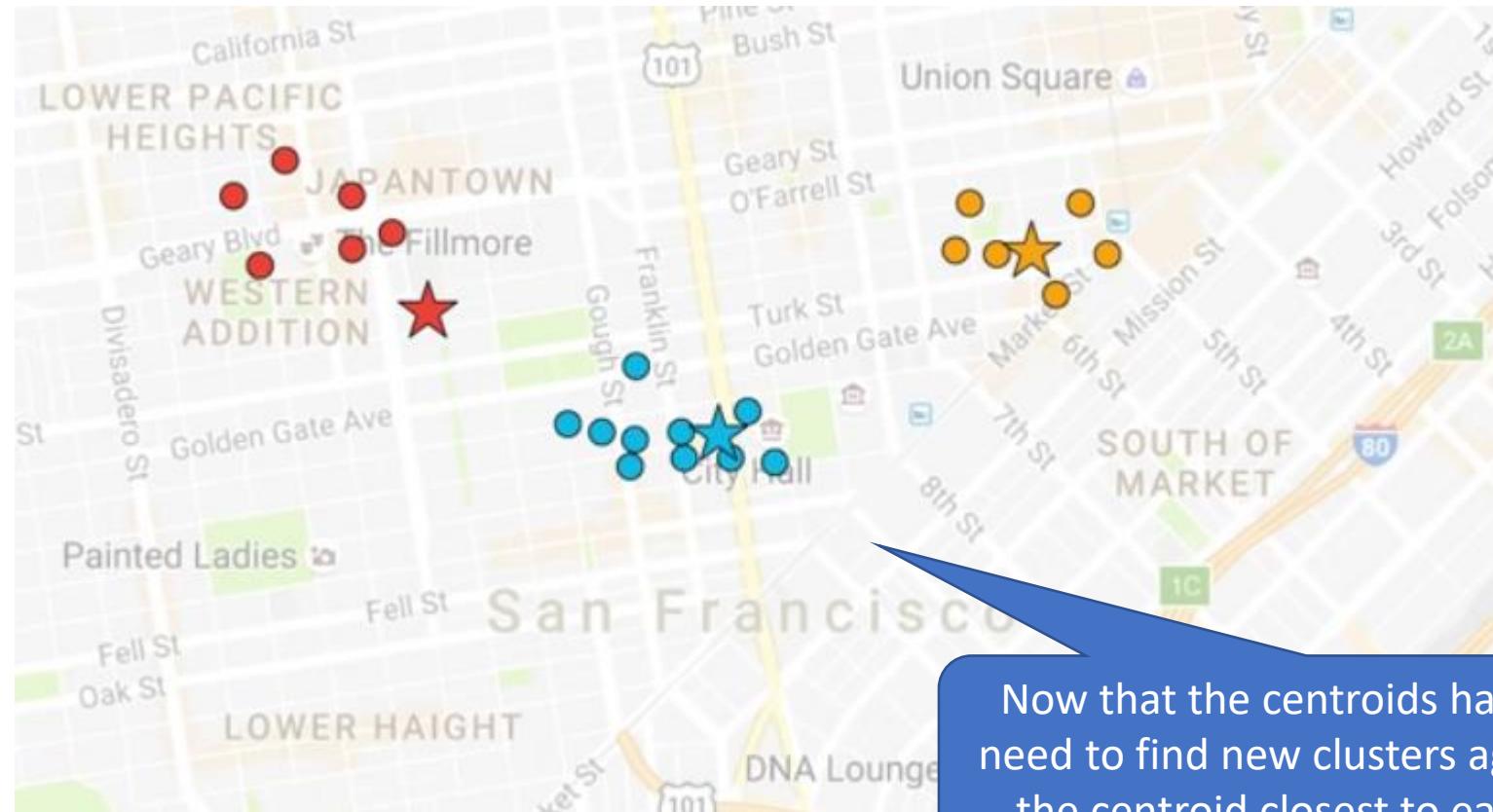


Find the clusters

Result

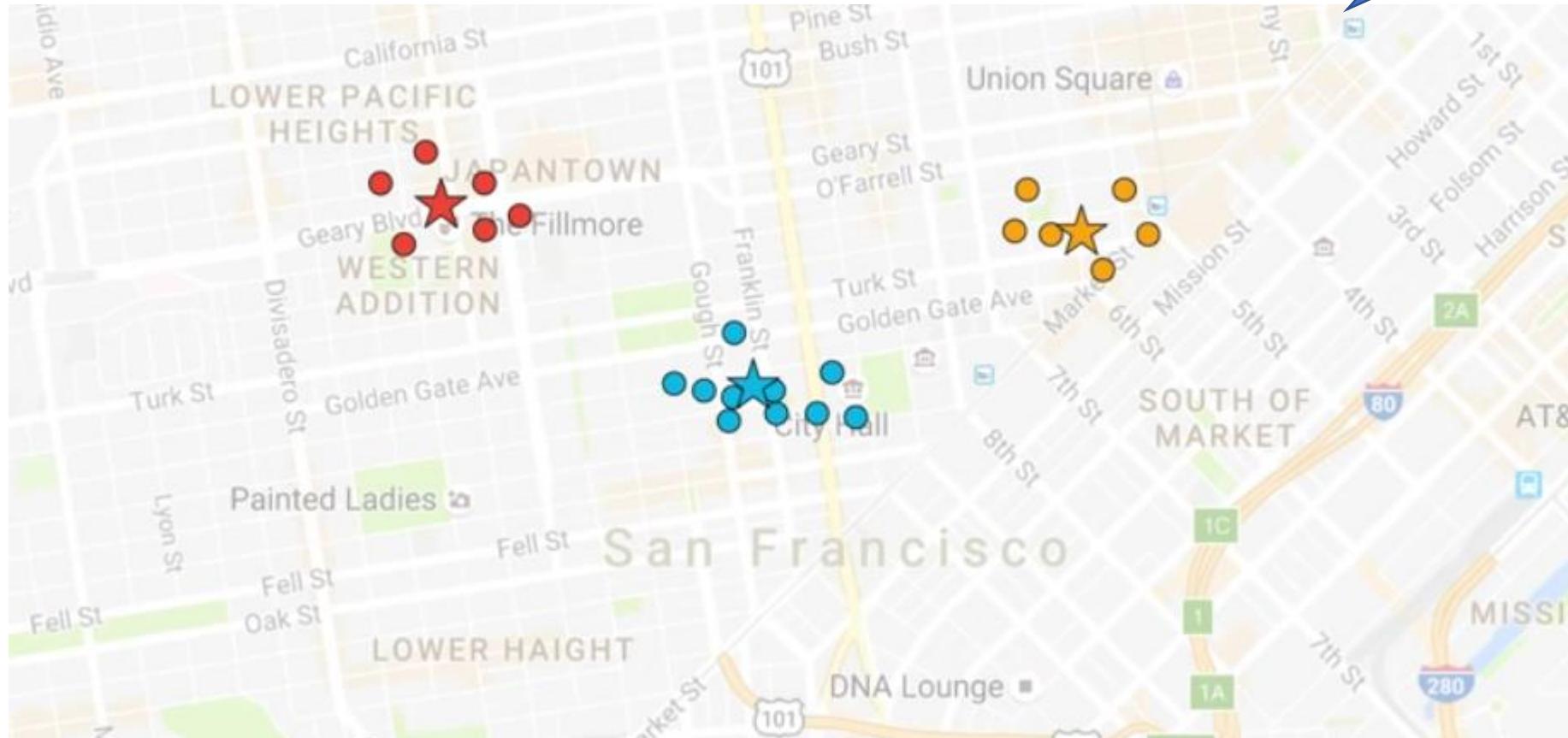


Move the centroids

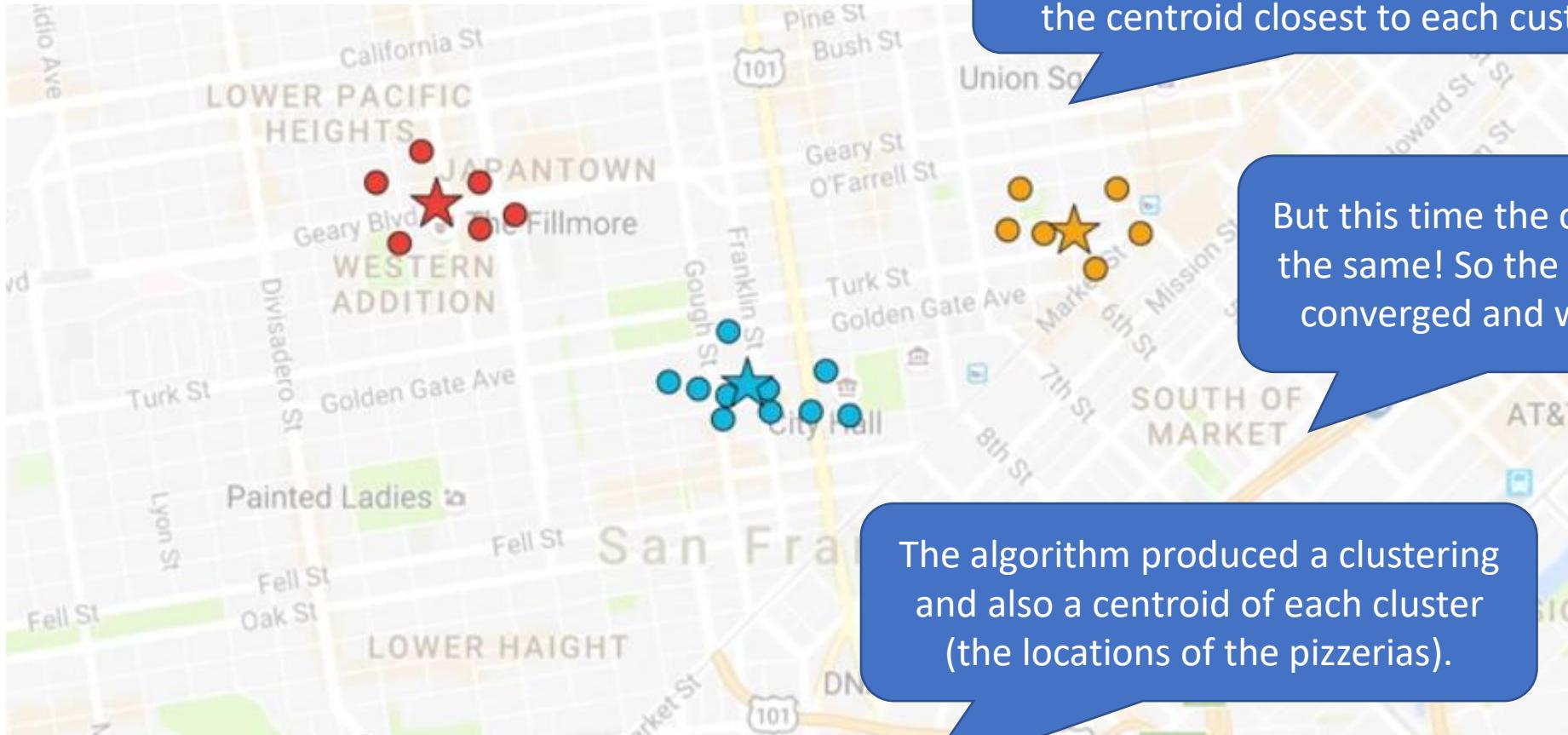


Move the centroids

Result

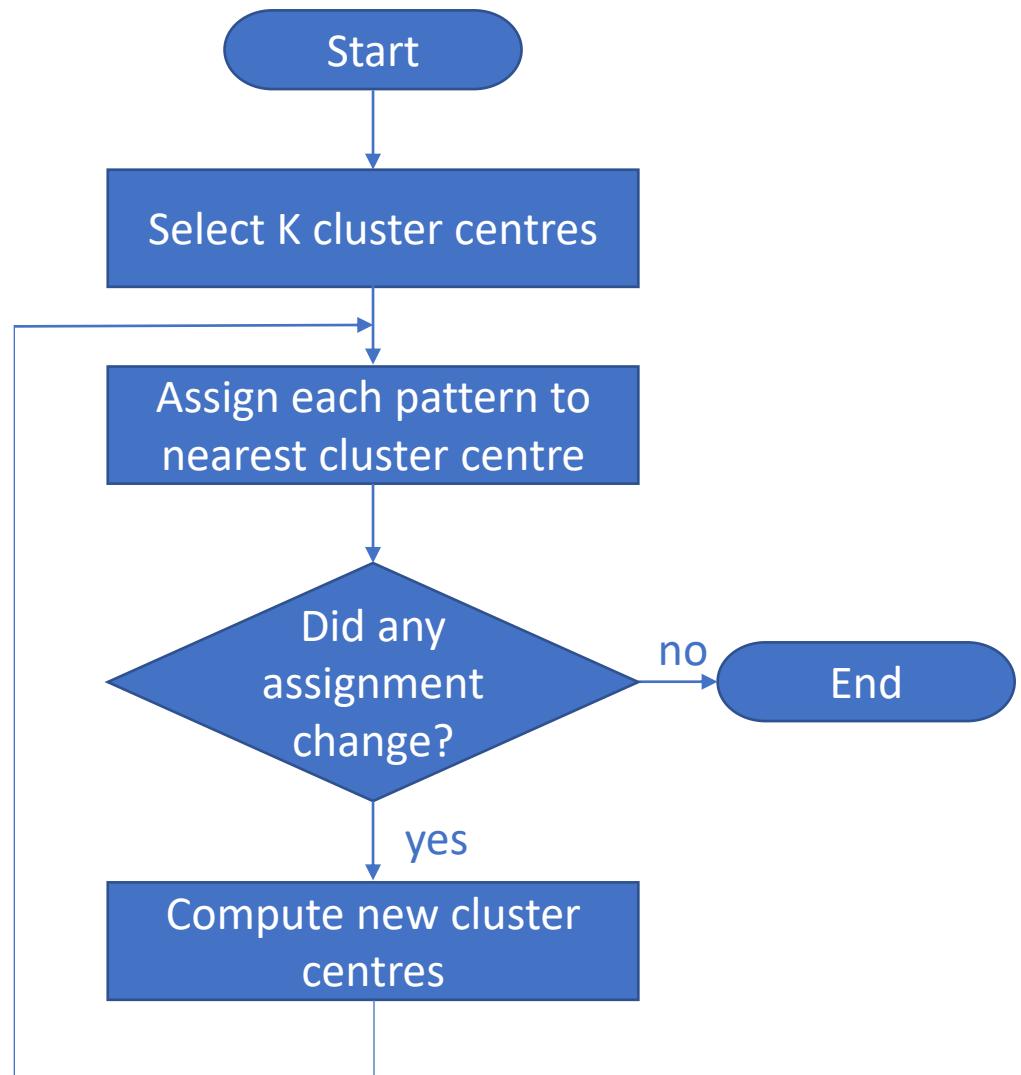


Find the clusters

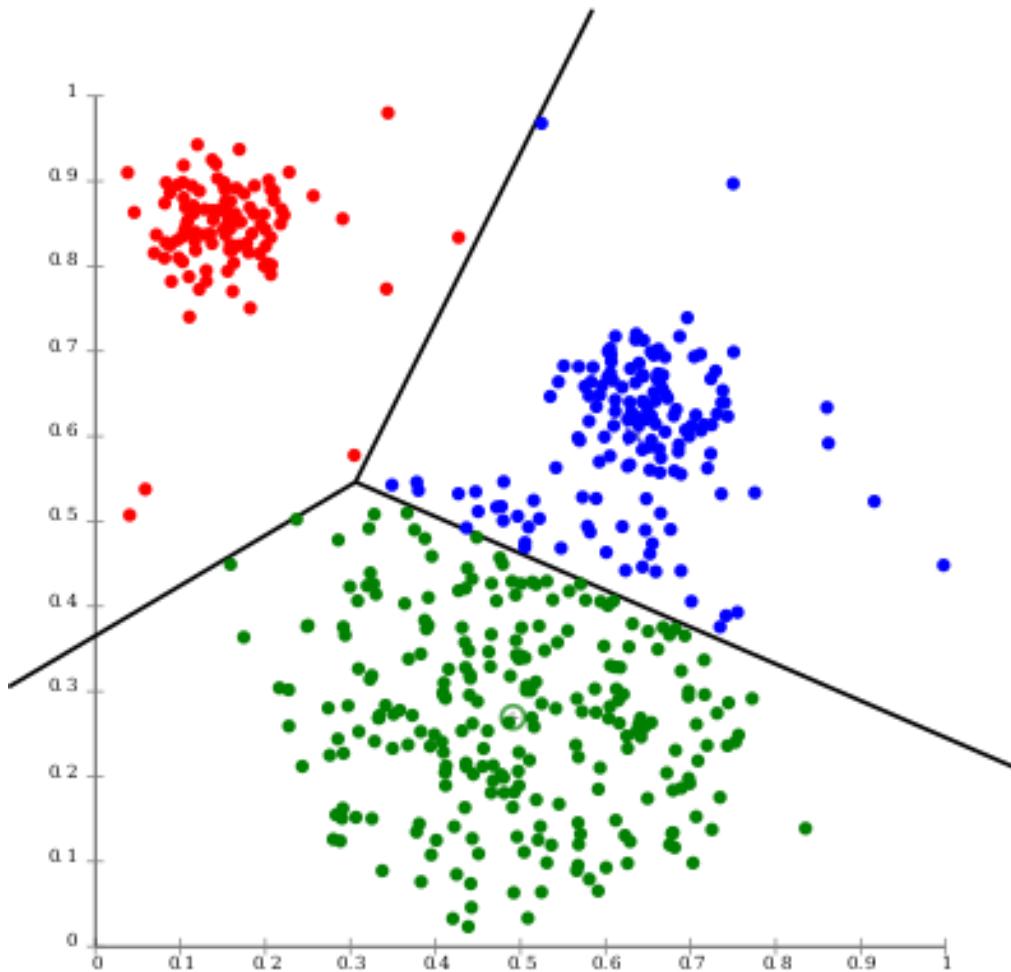


K-means

- choose value of K
- choose initial positions of the K cluster centres
- choose distance metric
- a greedy algorithm
- converges to a local minimum



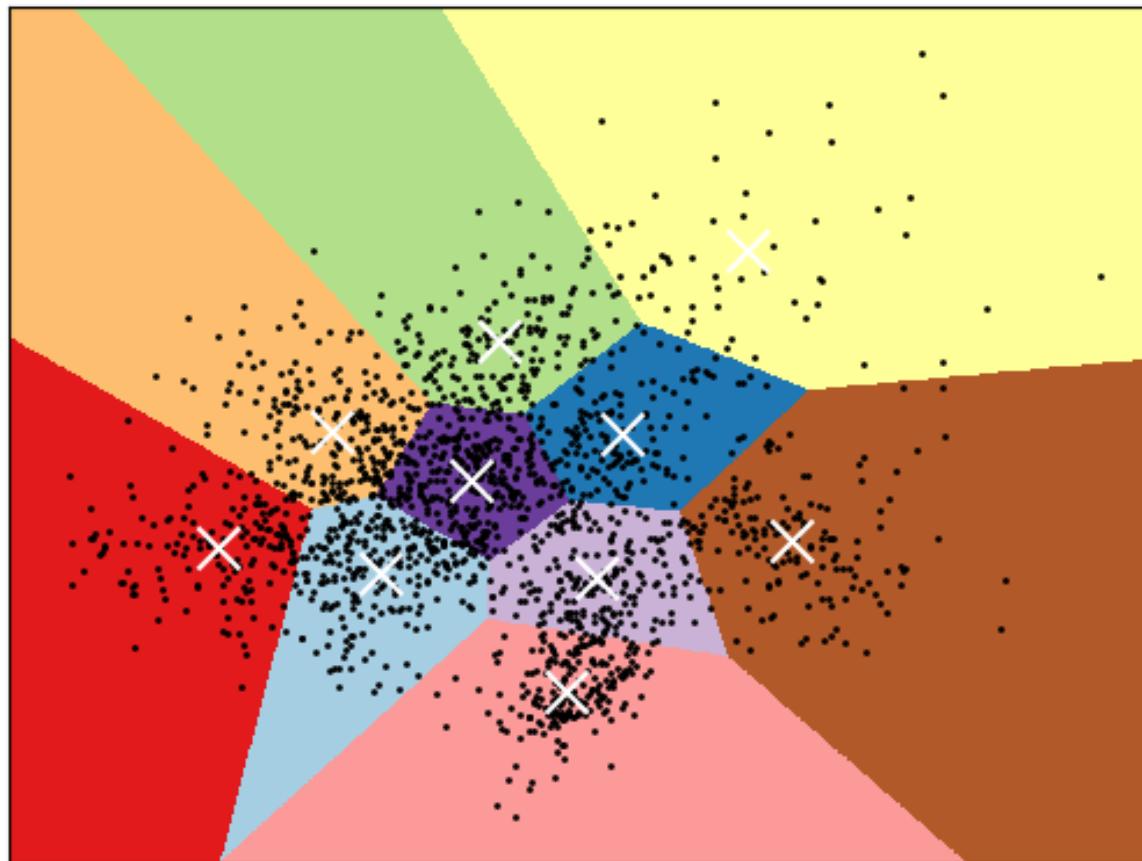
Geometric Interpretation



Voronoi diagram for centroids as seeds.
Points in area between lines closer to their centroid than to any other

Clustering hand-written digits

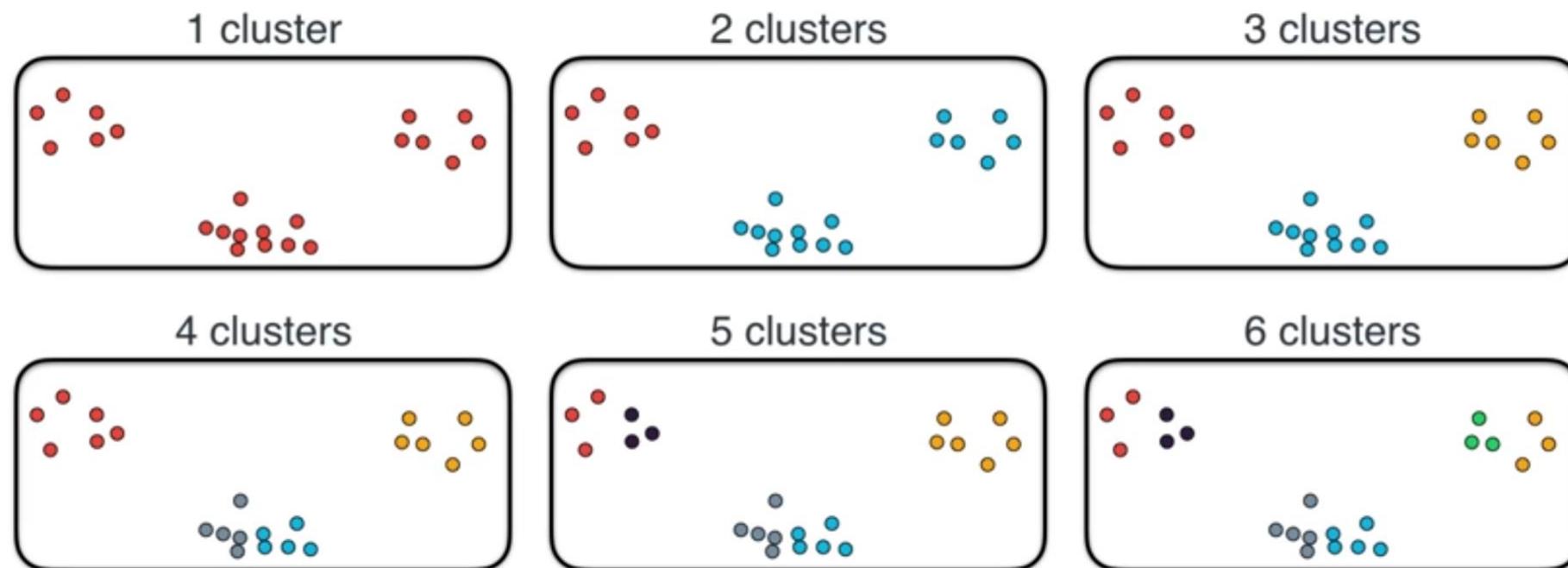
K-means clustering on the digits dataset (PCA-reduced data)
Centroids are marked with white cross



The elbow method

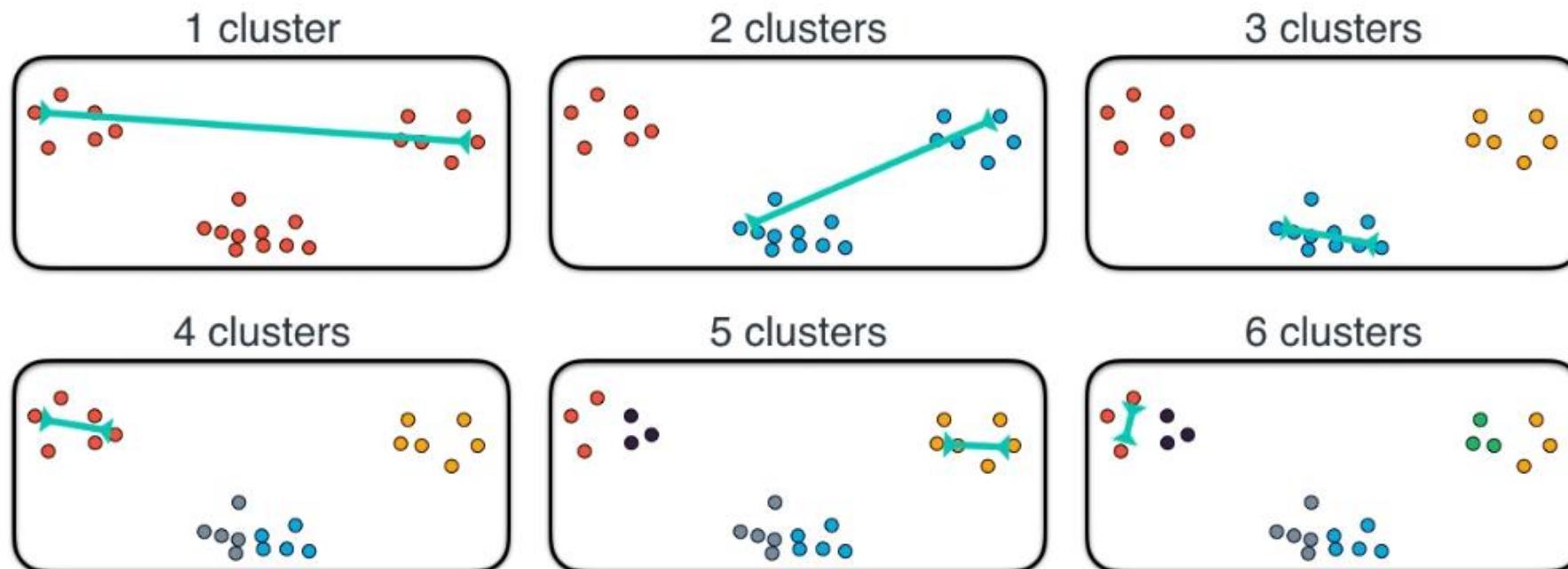
The elbow method

How do we pick K (the number of clusters)?



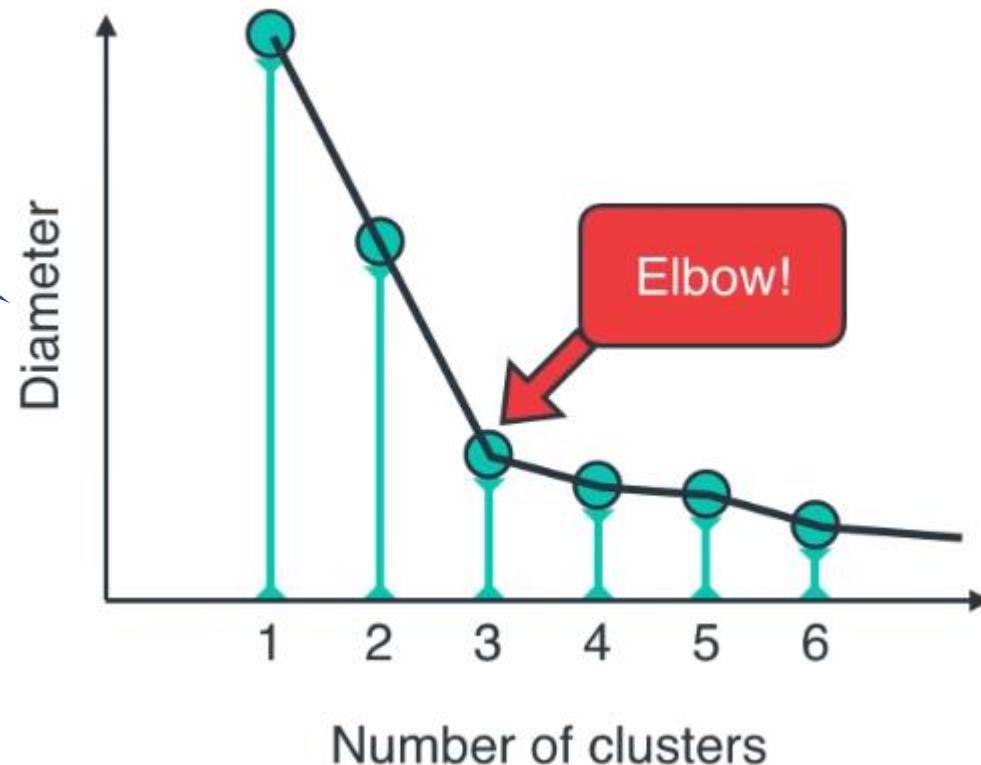
The elbow method

Let the *diameter* of a clustering be the longest intra-cluster distance
(longest distance between two points of the same colour)



The Elbow method

The diameter is a scalar
also when the data are
multi-dimensional

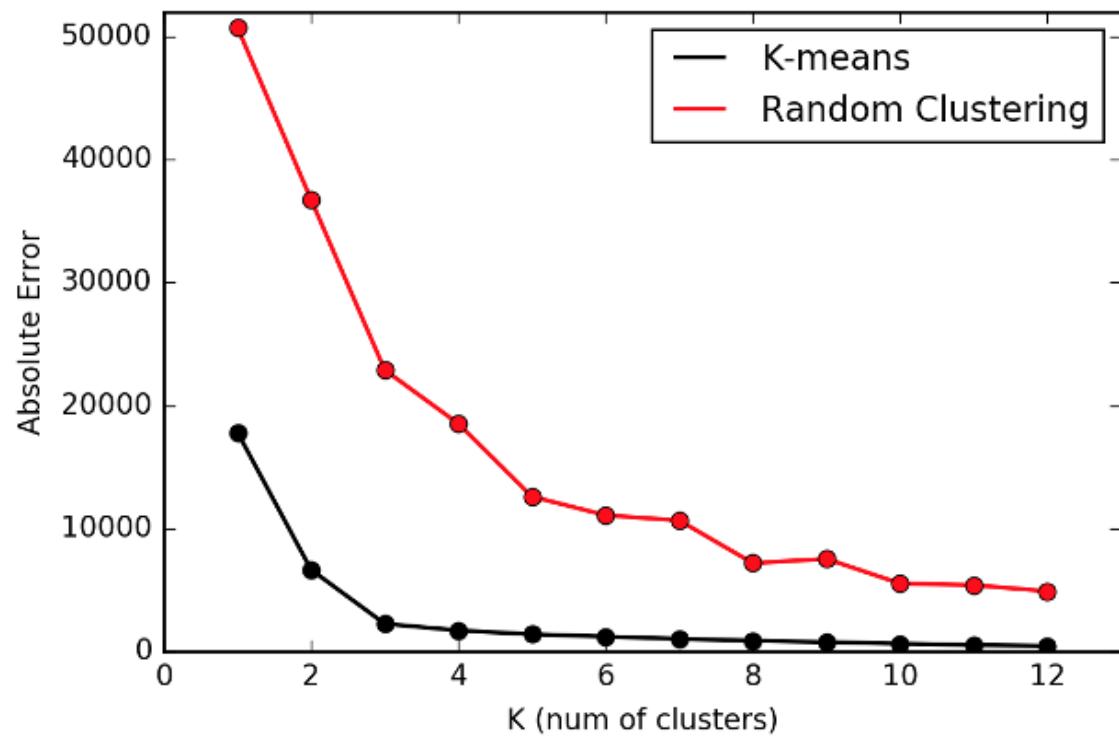


The elbow method



We can see where the elbow is, but how could a program find the elbow?

“Elbow method” for selecting K

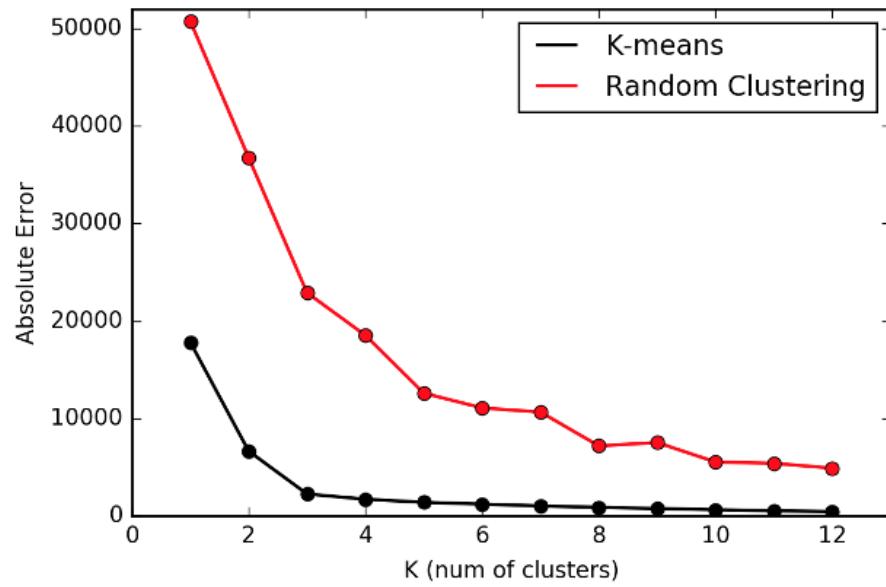


Balance between wanting

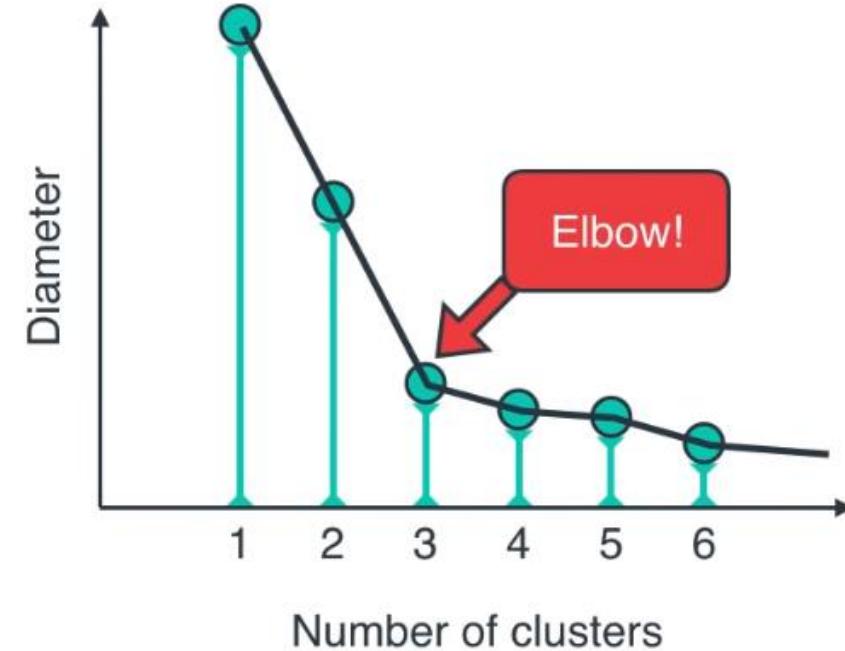
- as few clusters as possible and
- that each cluster is as “tight” as possible

Usual to use sum of squared errors

“Elbow method”



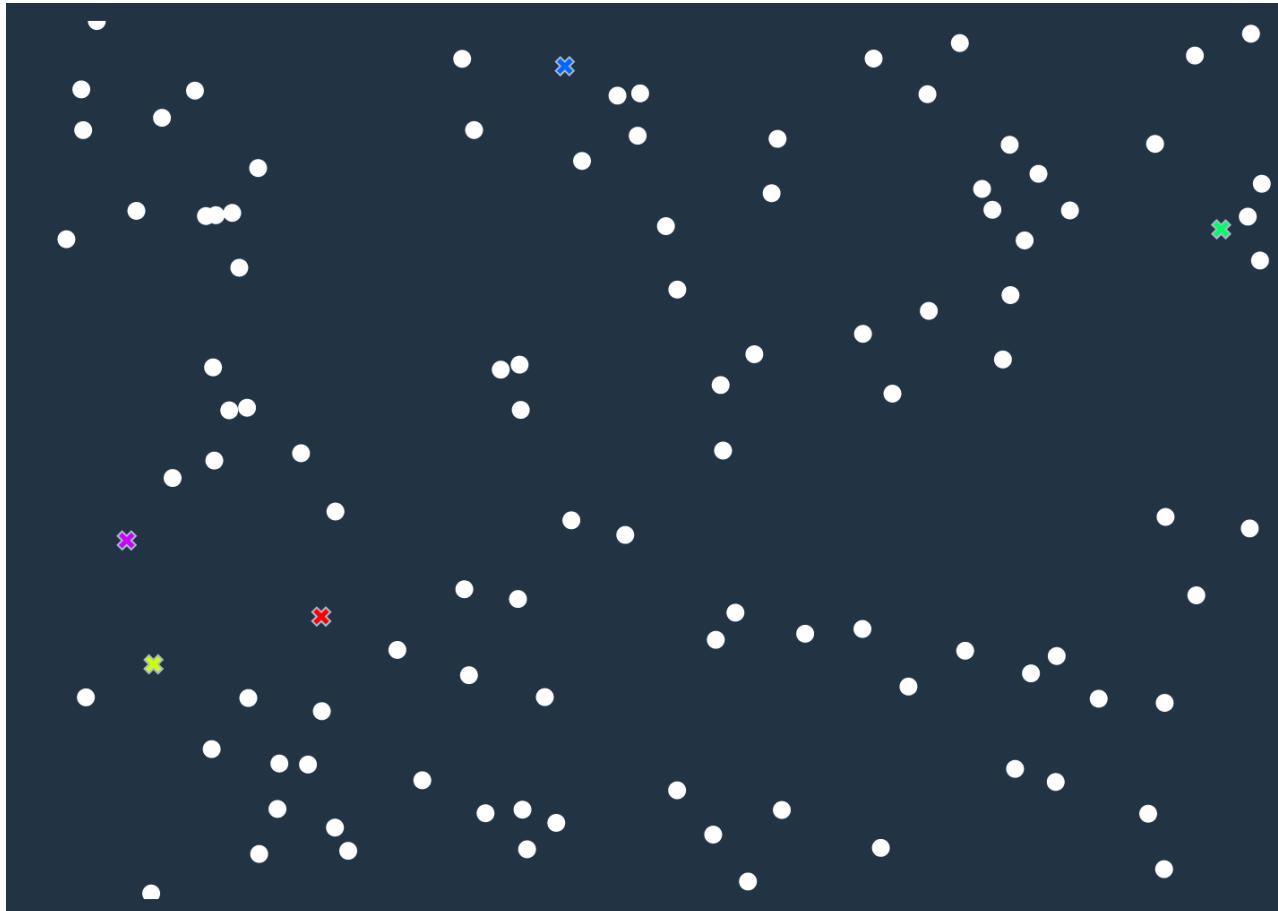
In this example, the **sum of squared errors** was used.



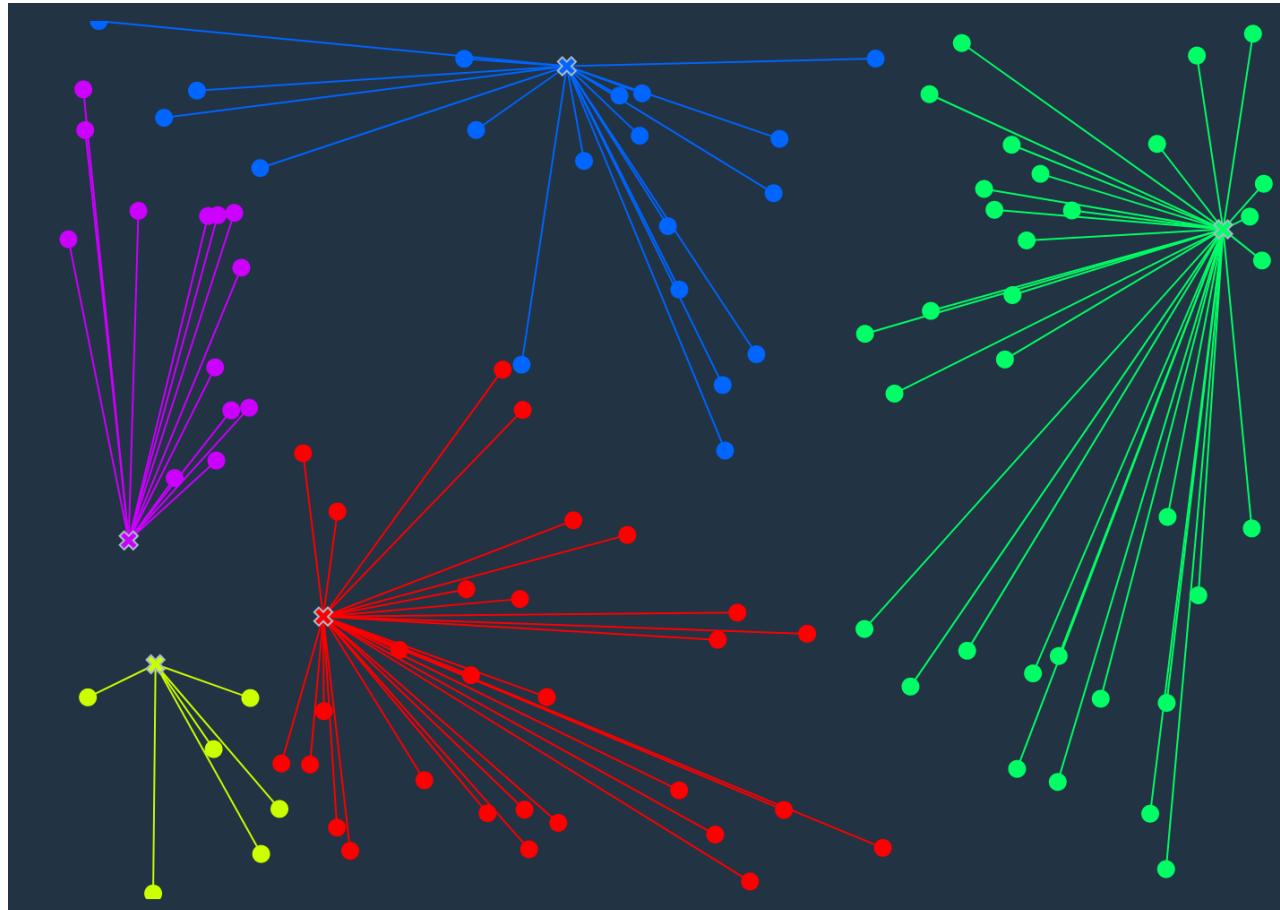
Here, the largest cluster **diameter** was used, i.e. the largest distance between two points in the same cluster.

Limitations and problems of k-means clustering

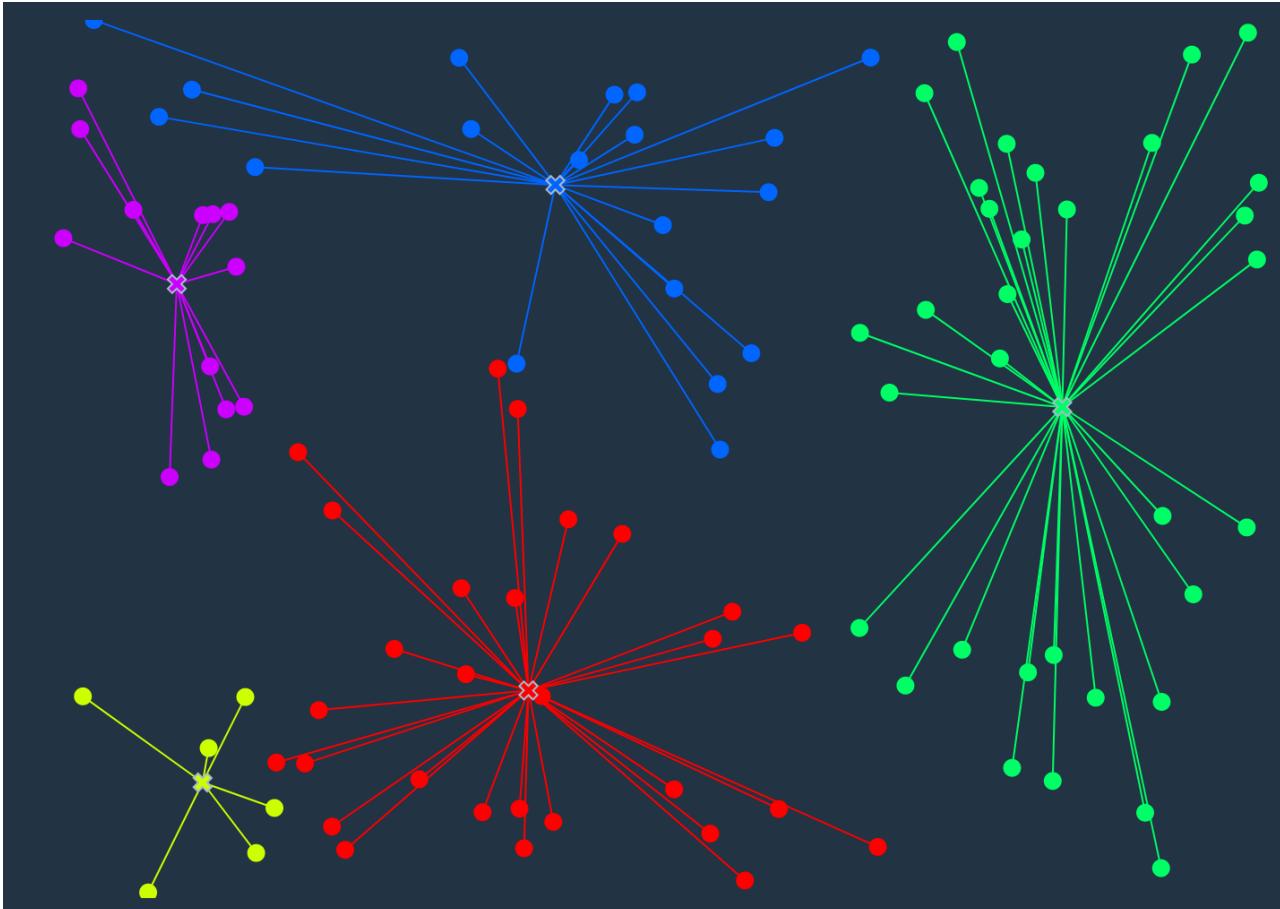
Select initial cluster “centres”



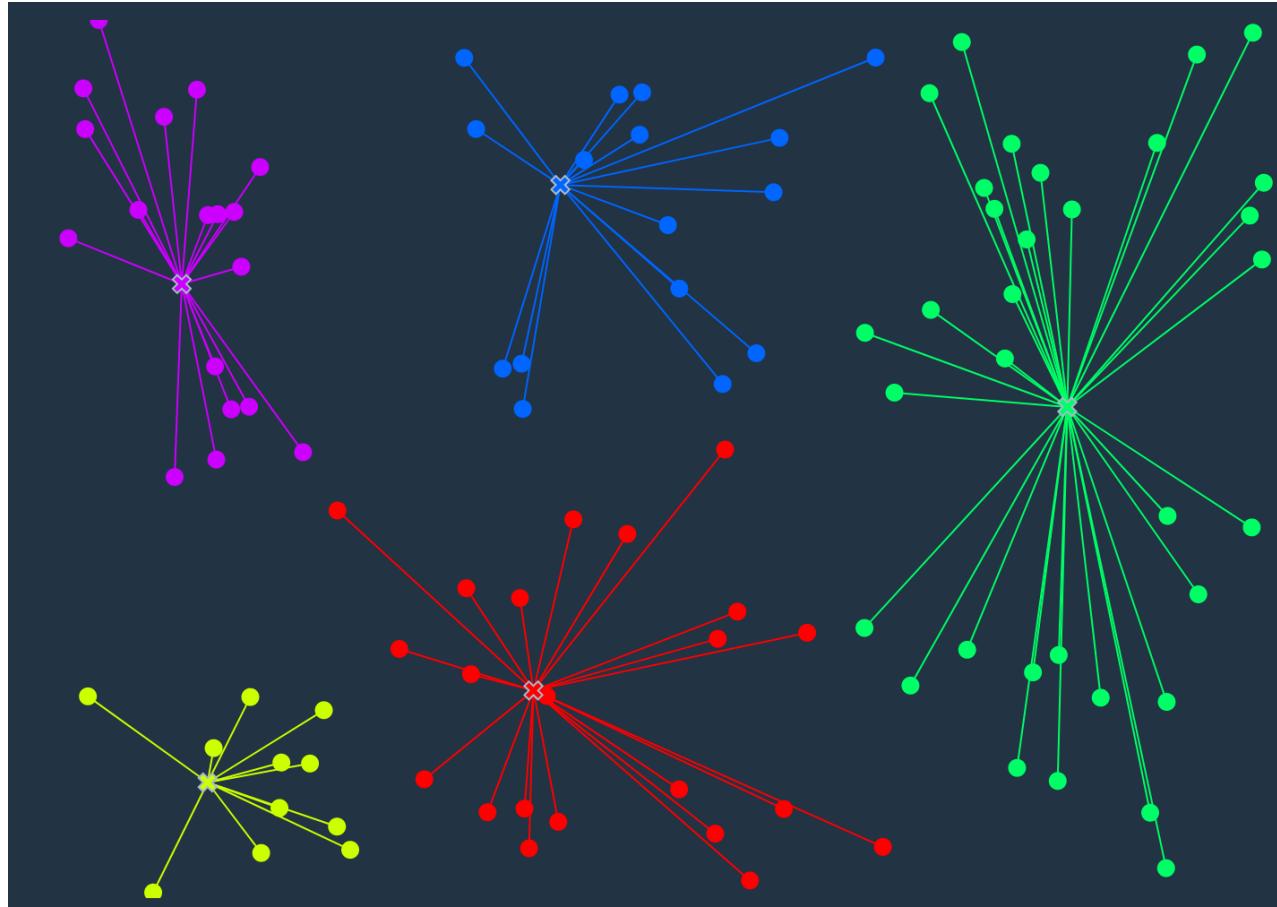
Assign patterns to nearest cluster centre (1)



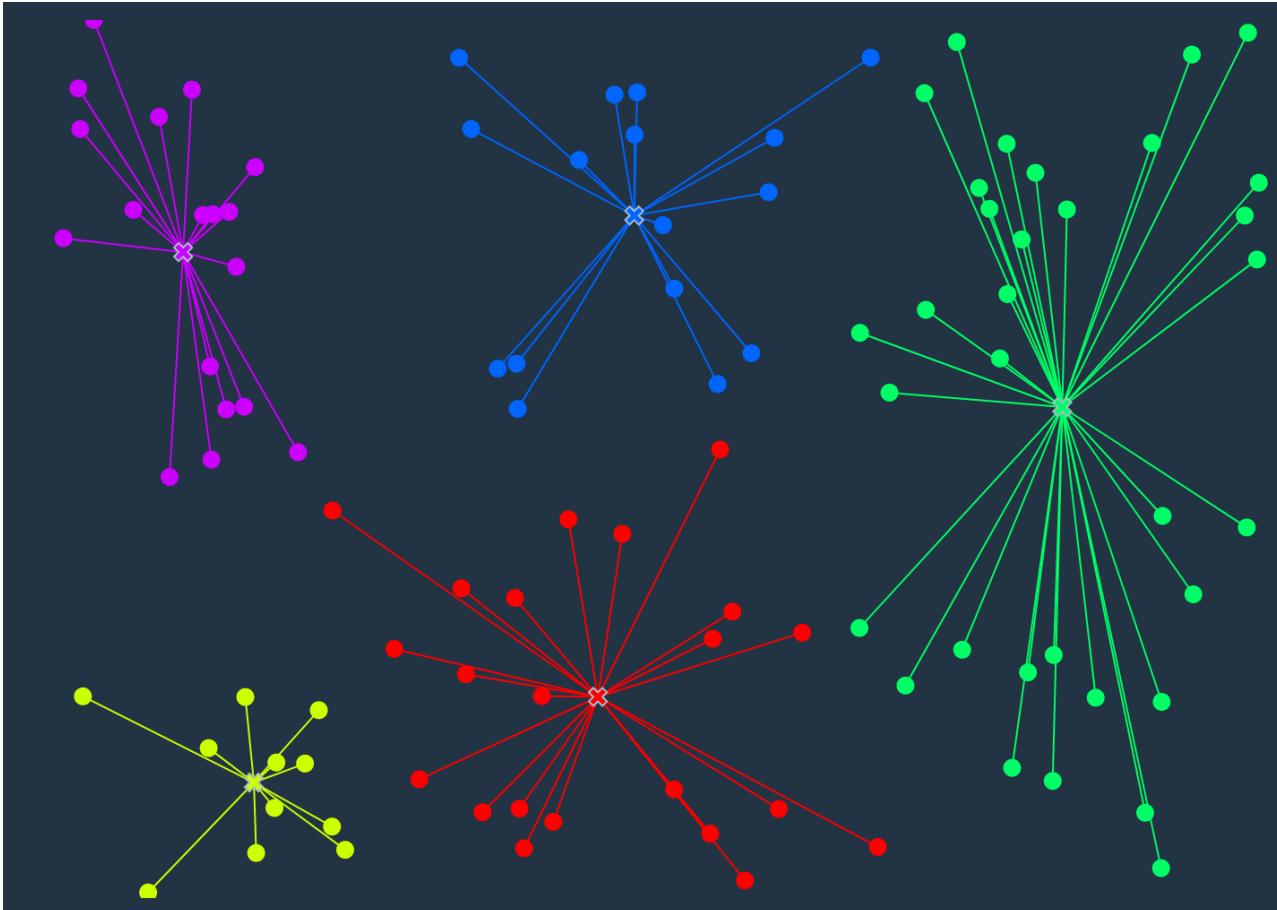
Compute new cluster centres (1)



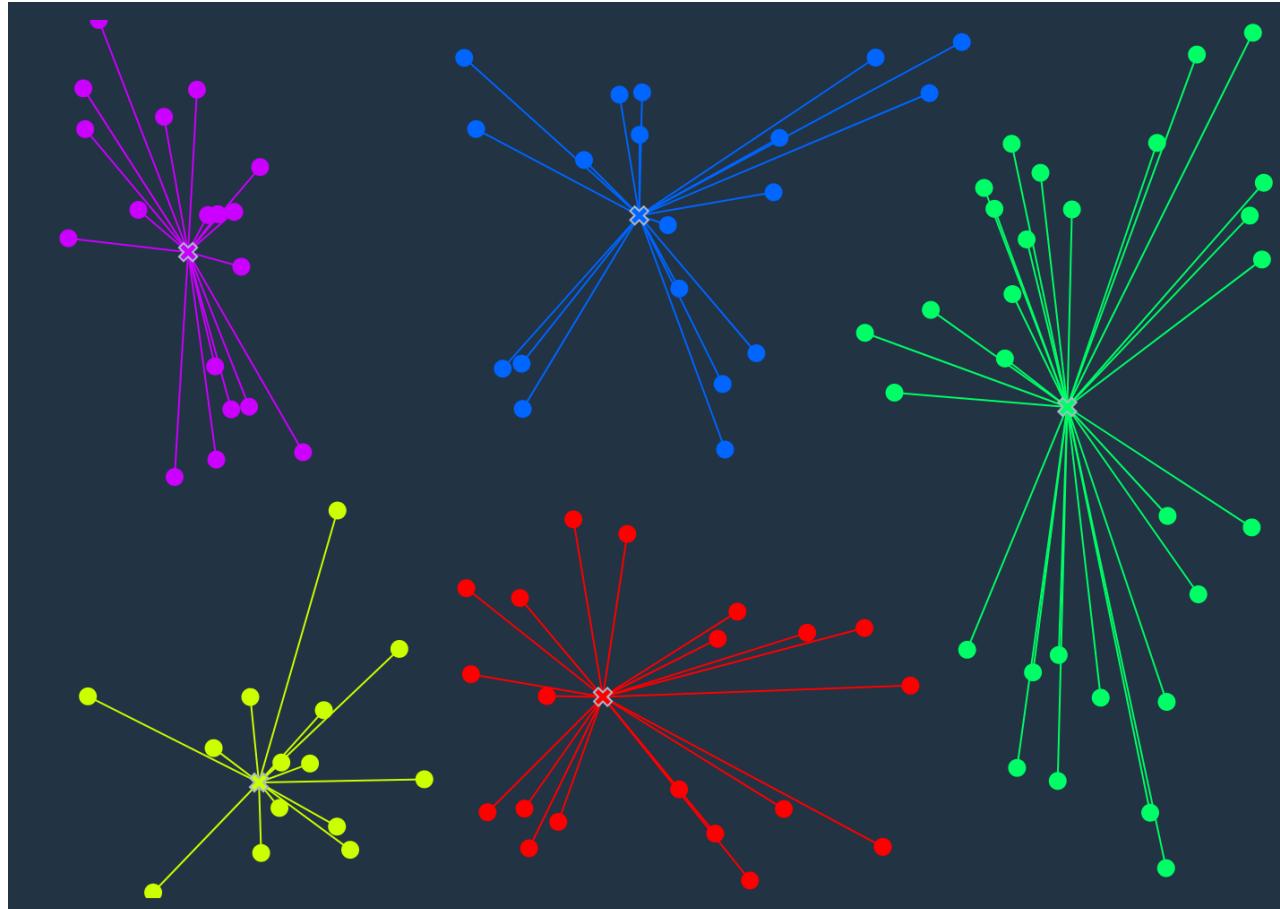
Assign patterns to nearest cluster centre (2)



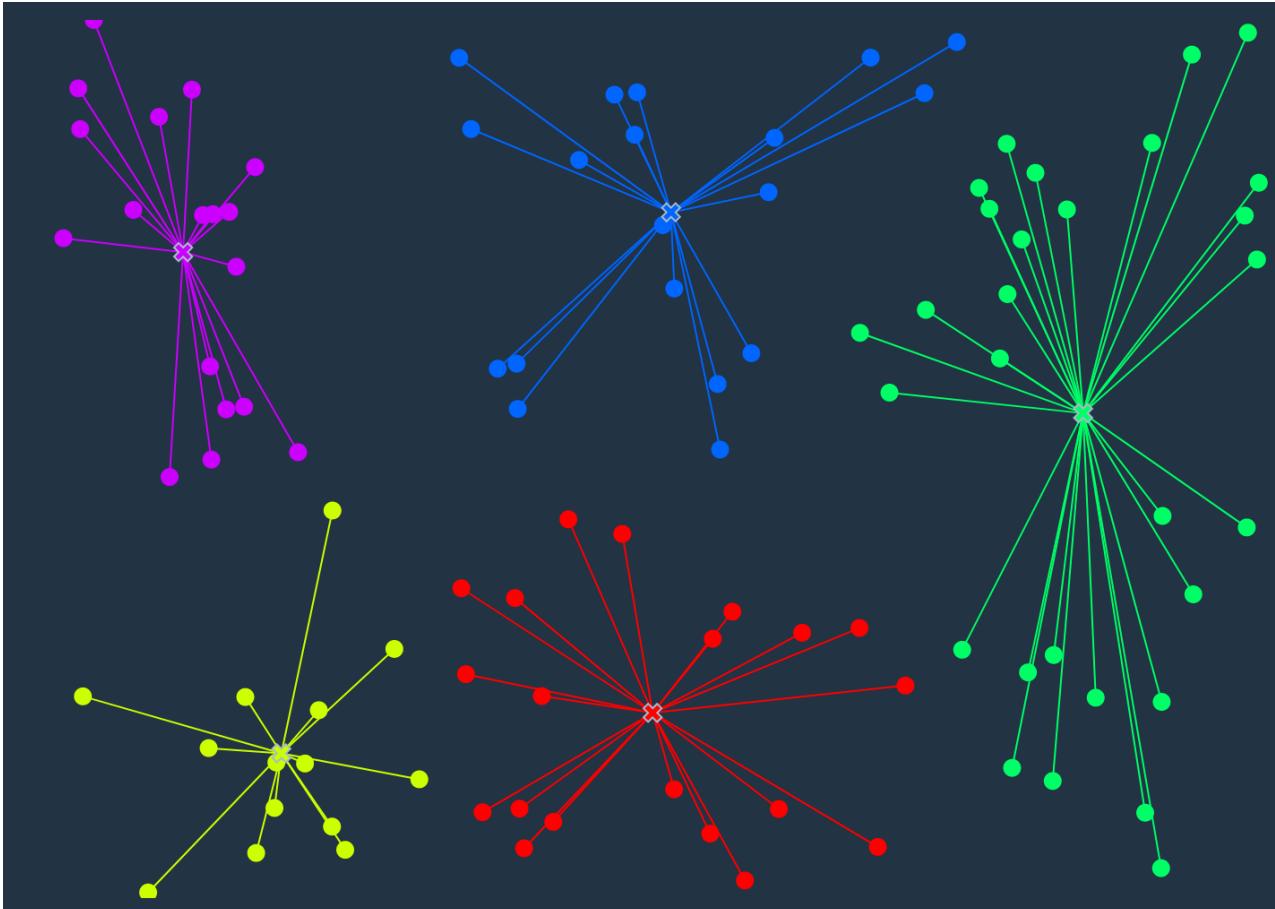
Compute new cluster centres (2)



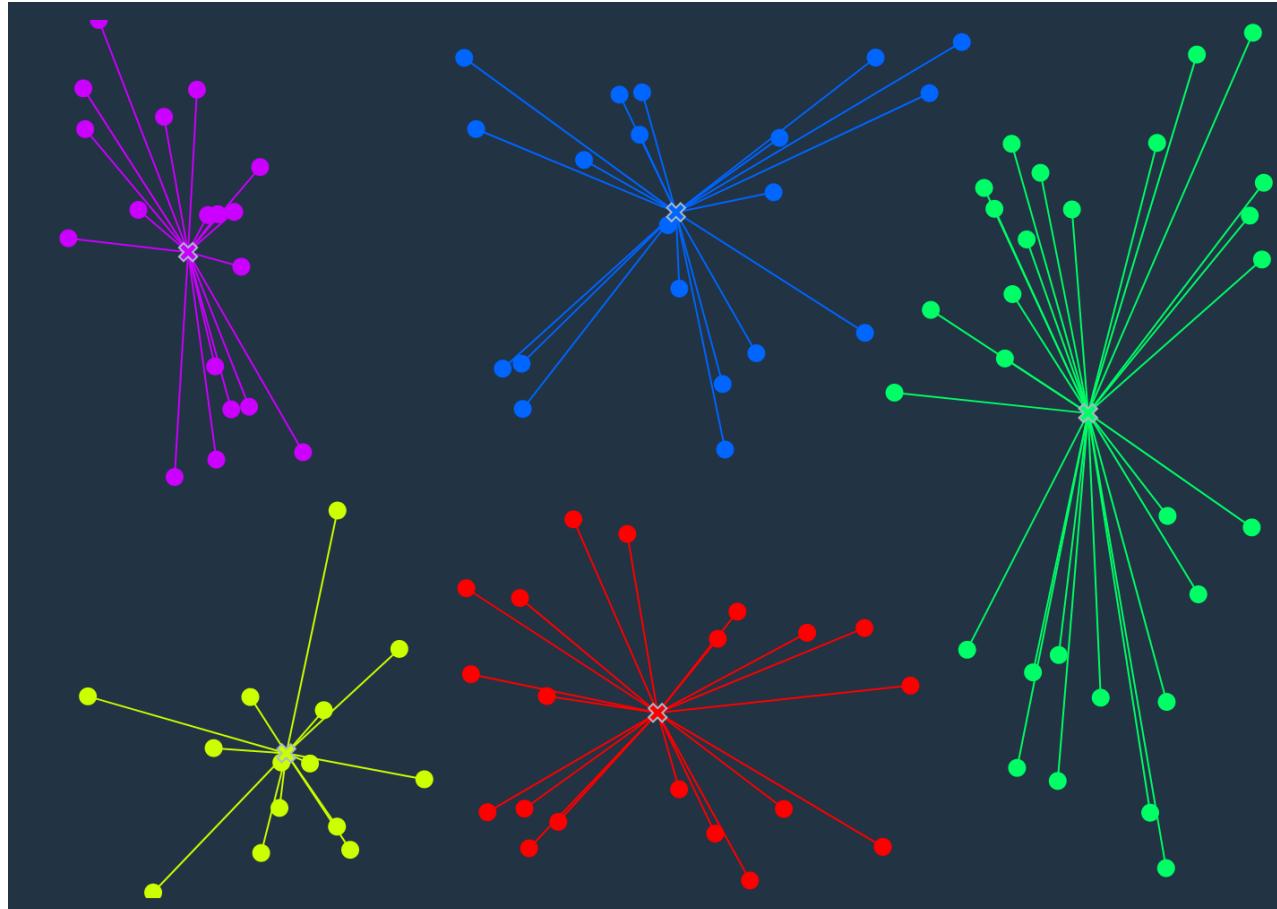
Assign patterns to nearest cluster centre (3)



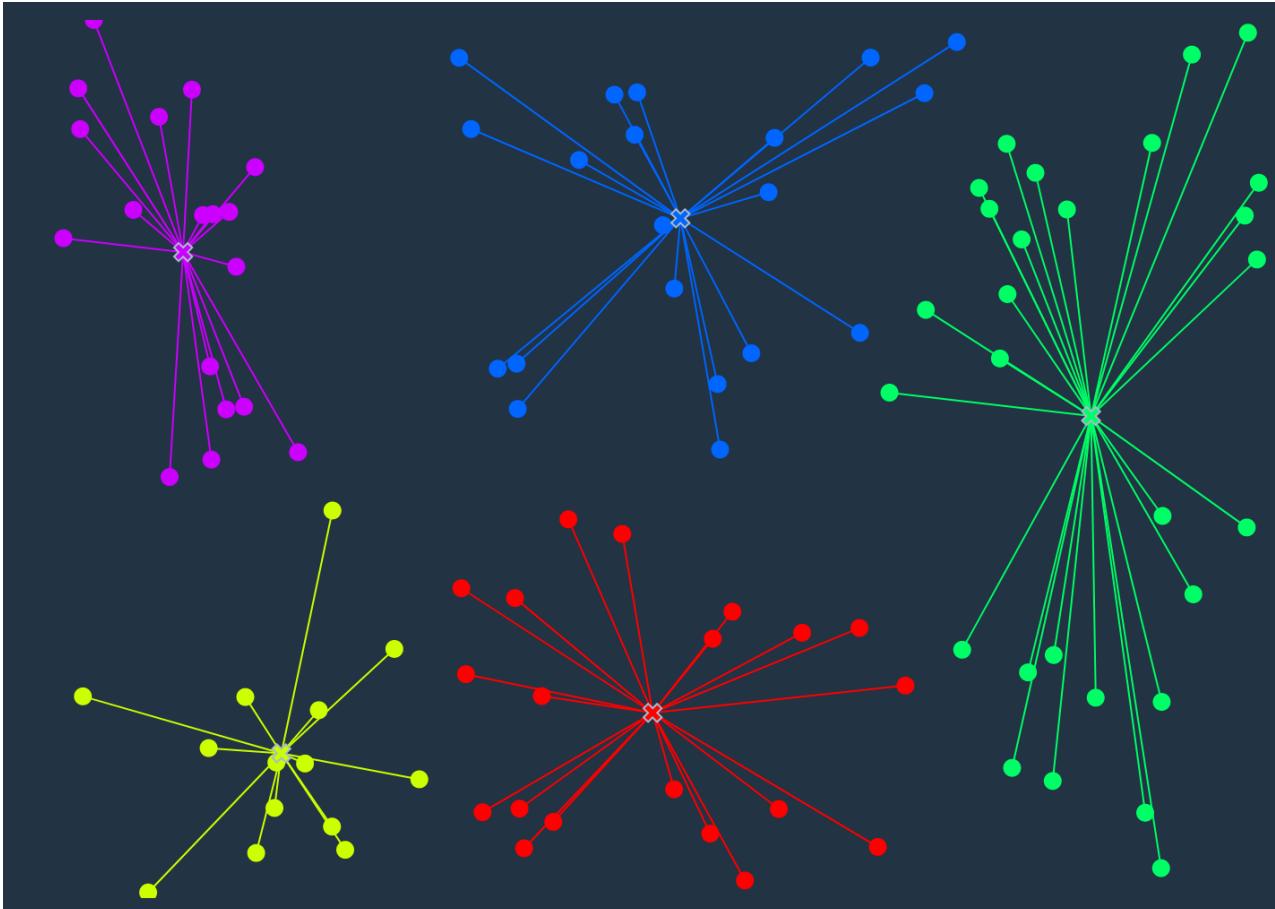
Compute new cluster centres (3)



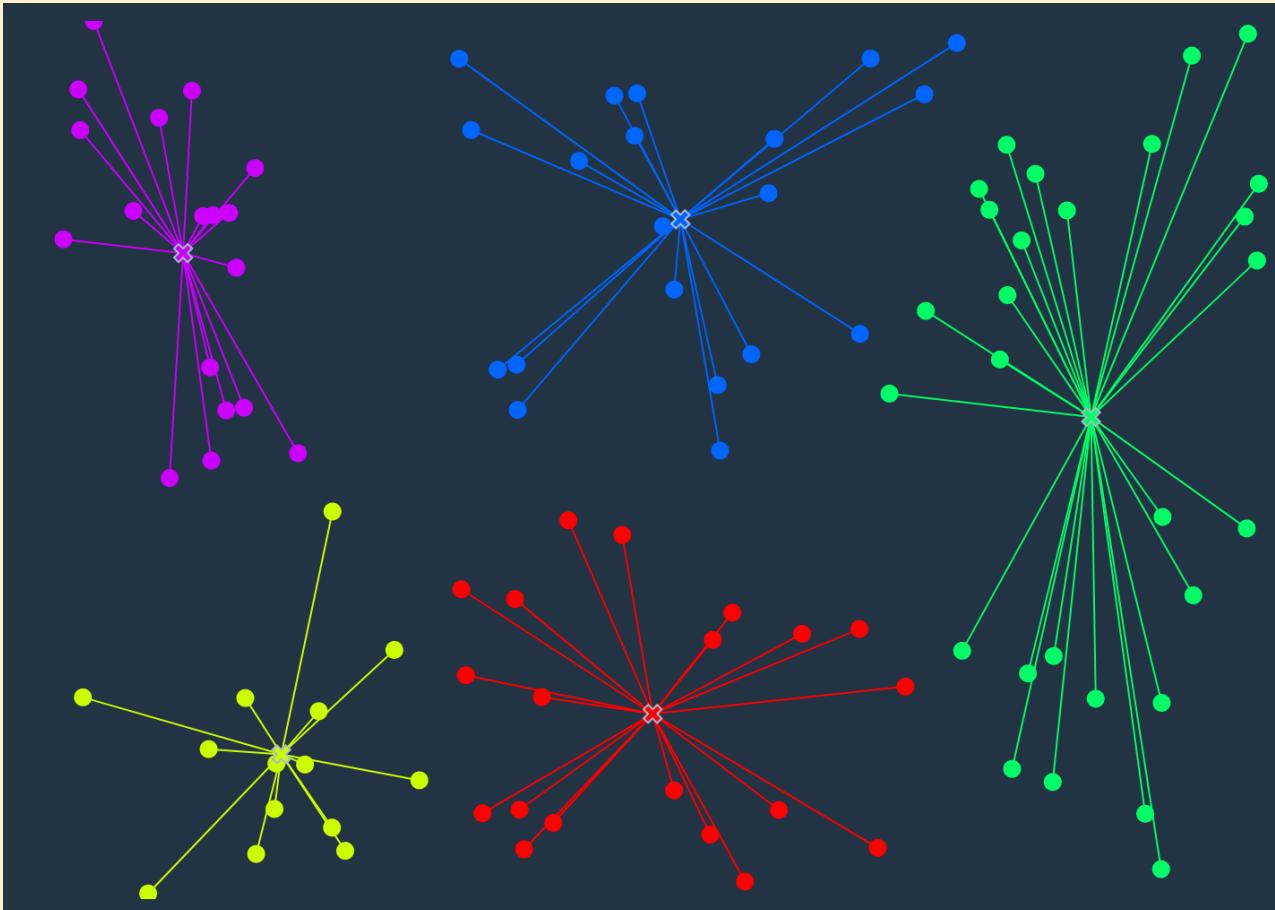
Assign patterns to nearest cluster centre (4)



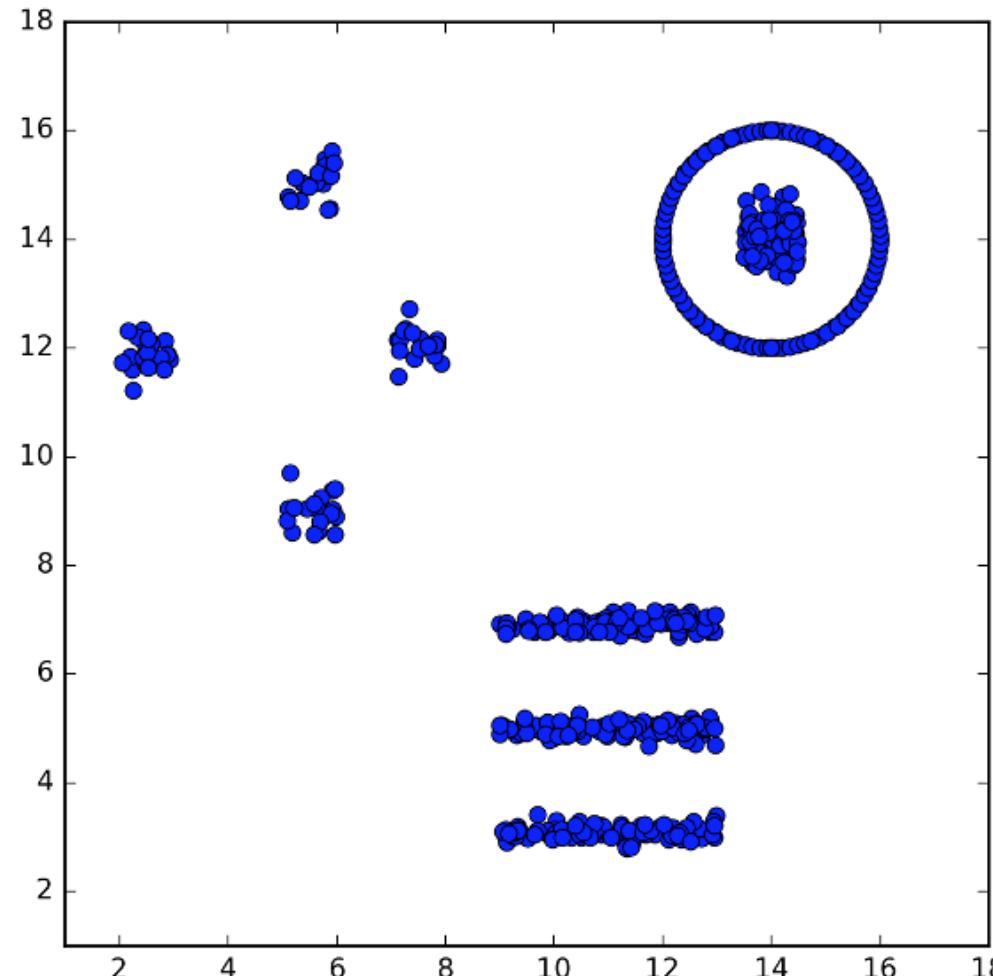
Compute new cluster centres (4)



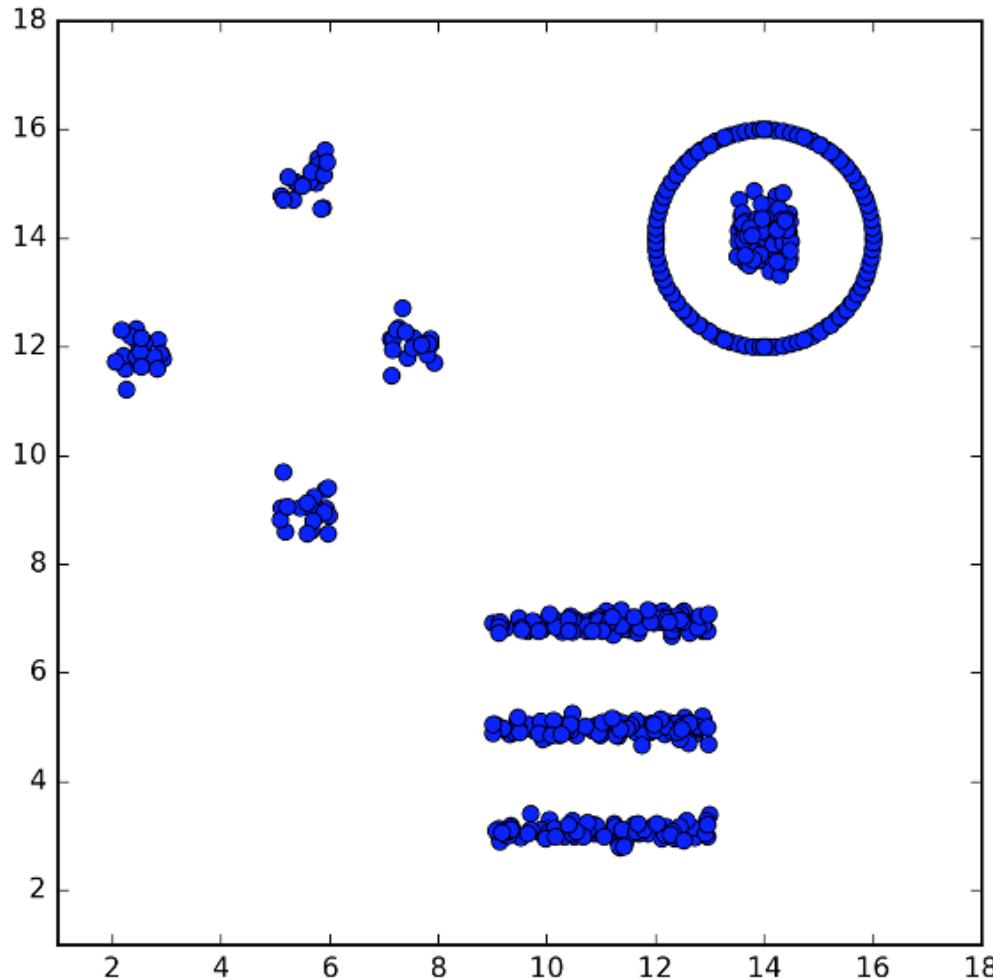
Is anything “wrong” with this?



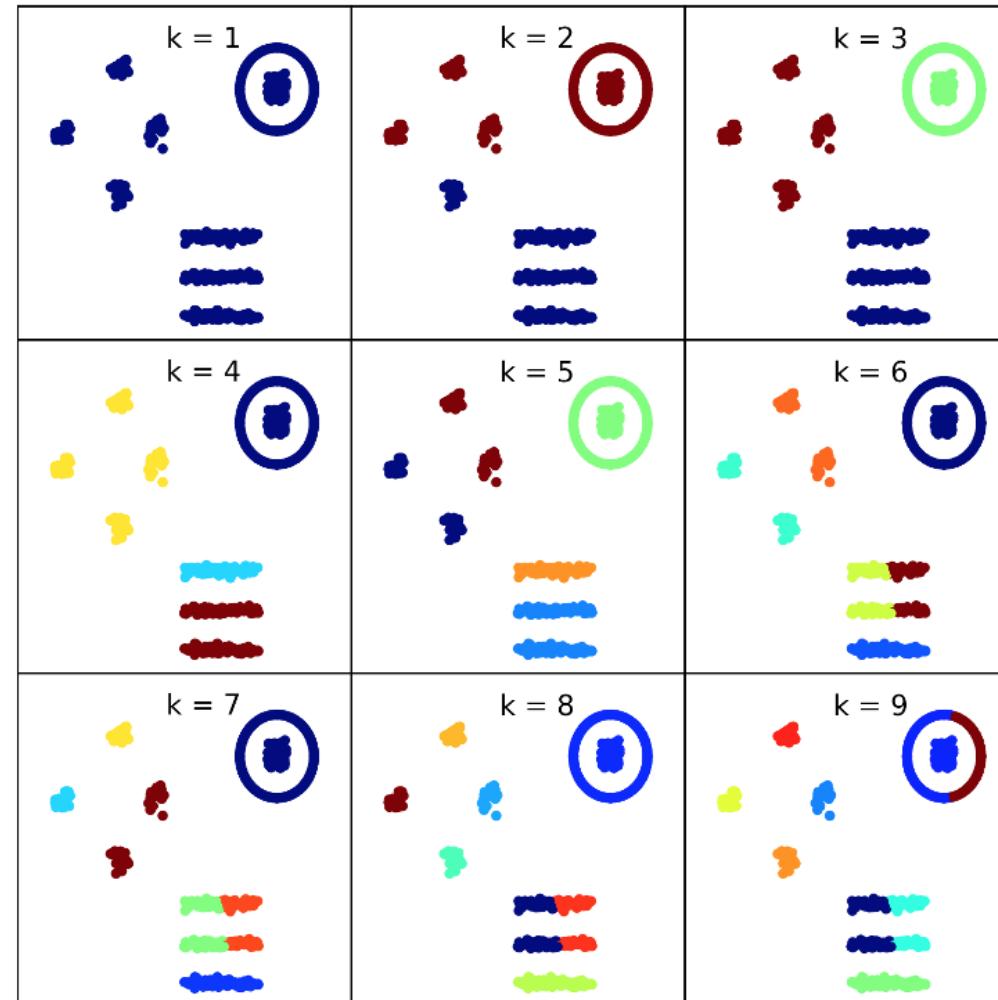
What will K-means do?



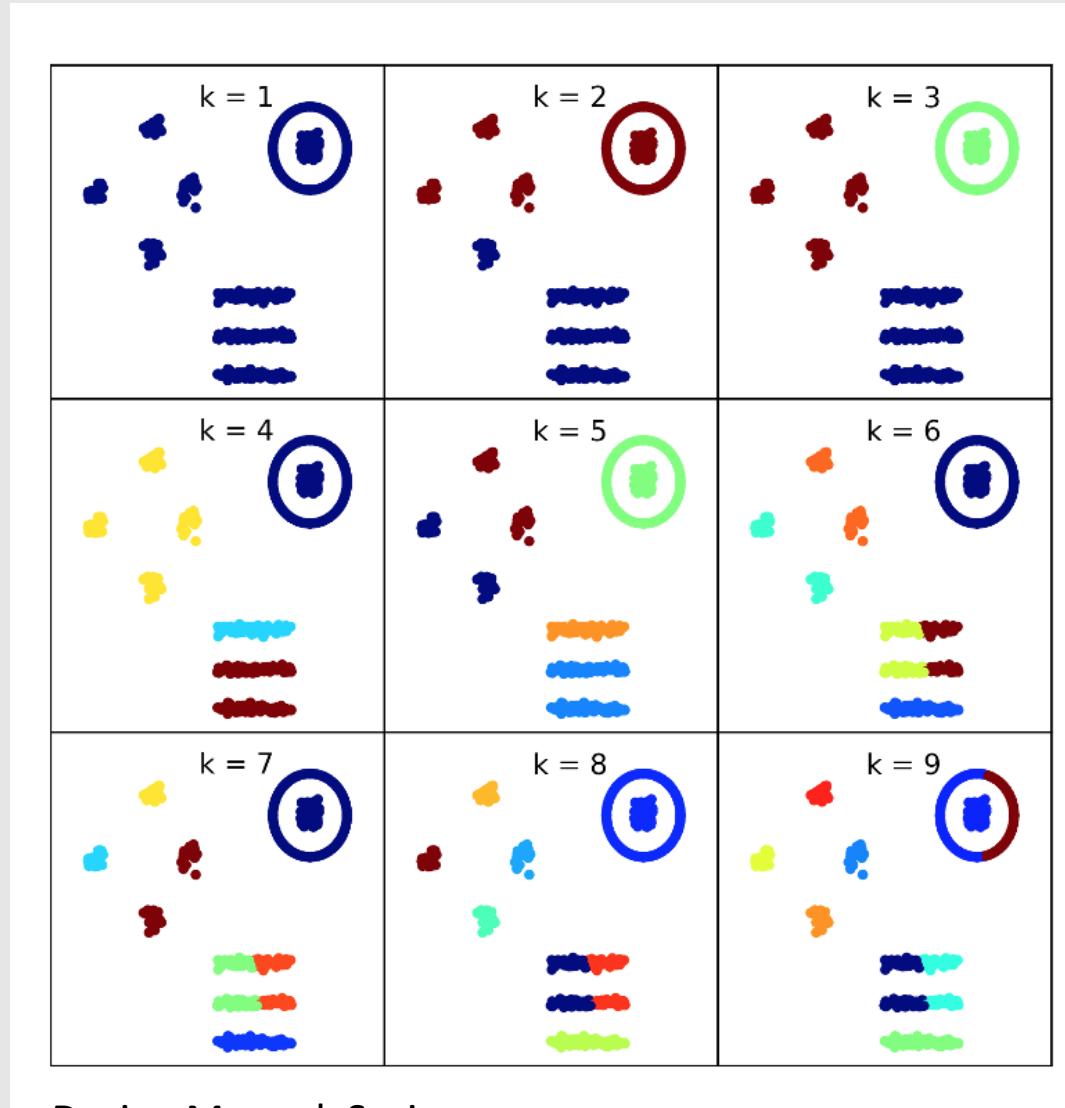
What will K-means do?



What will K-means do?



What will K-means do?



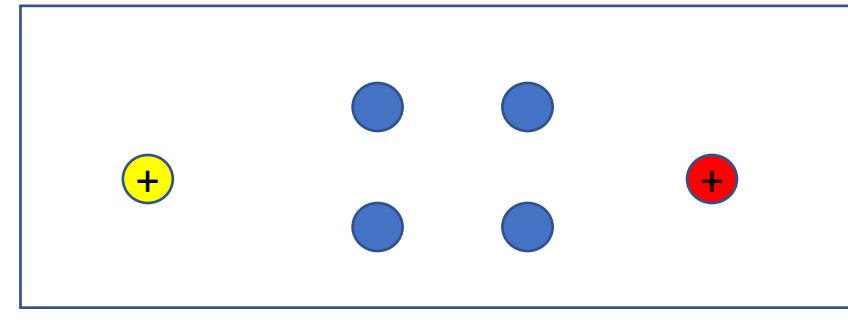
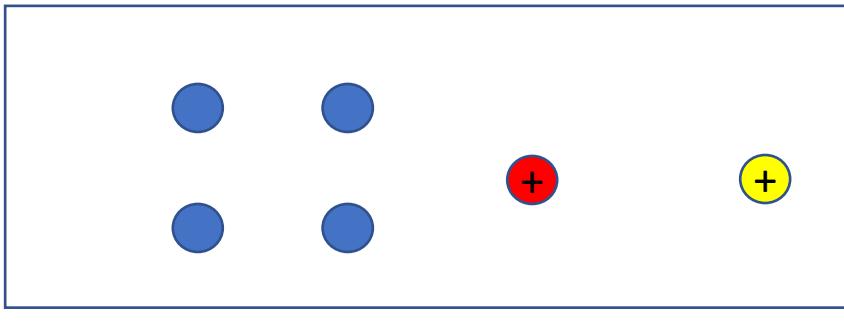
Which clustering
is best?

Caveats

K-means works best for

- spherical clusters
- equal diameter (equal variance)
- equal cluster size

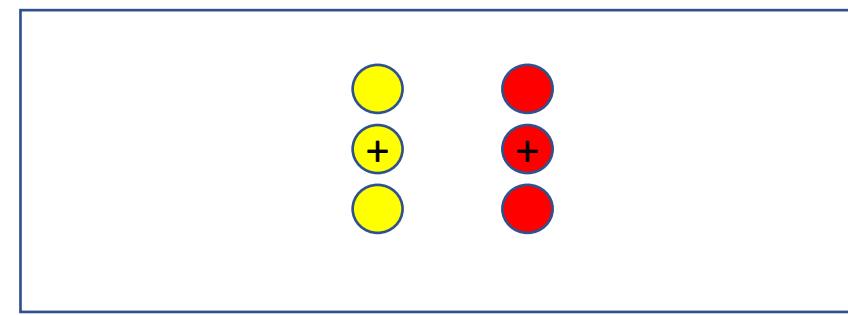
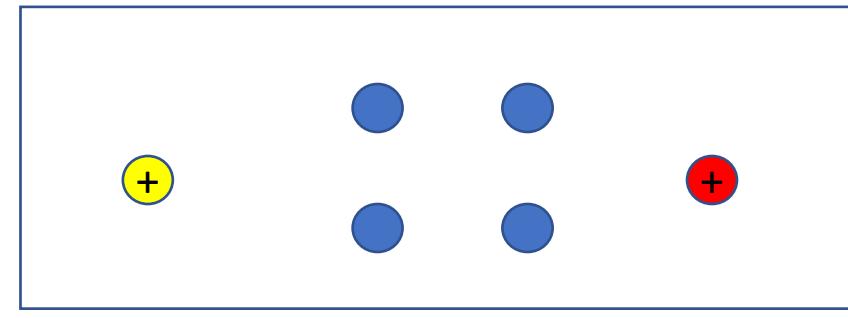
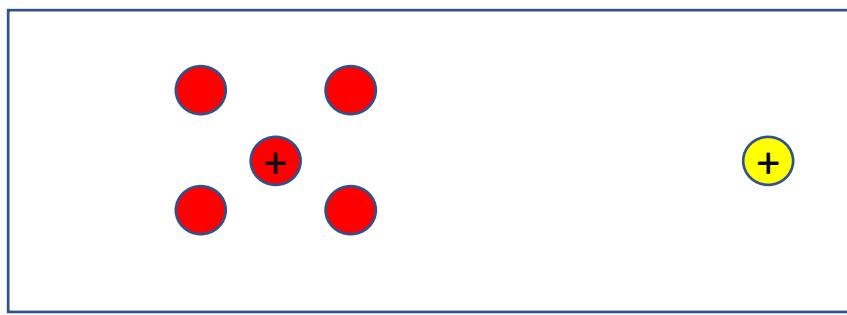
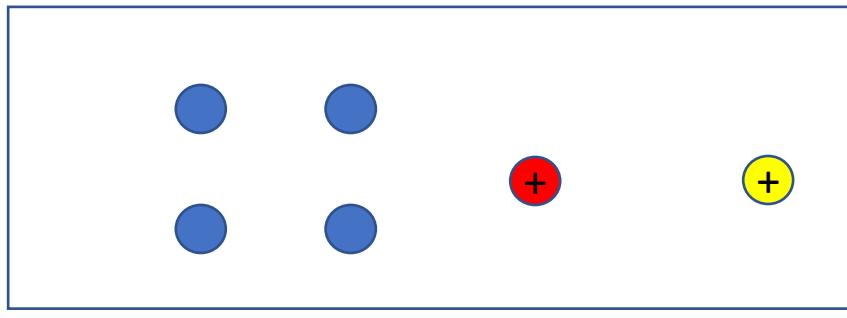
The result depends on the initialisation



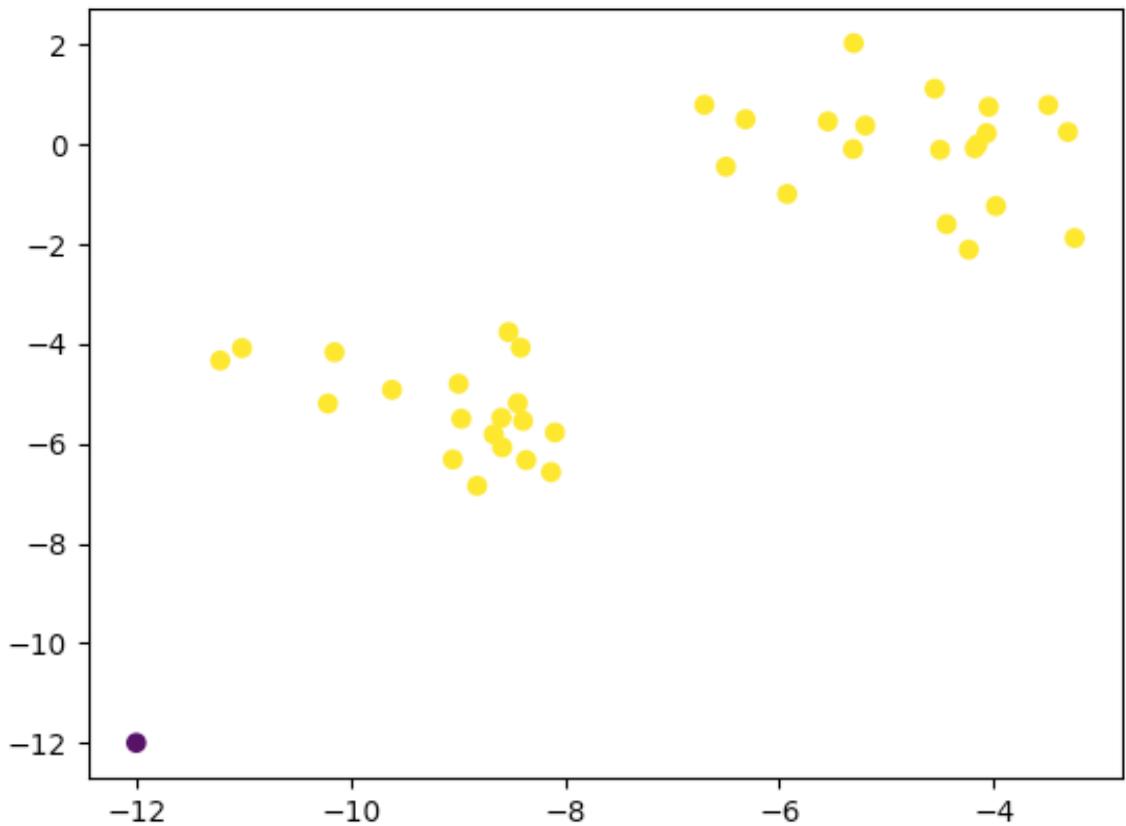
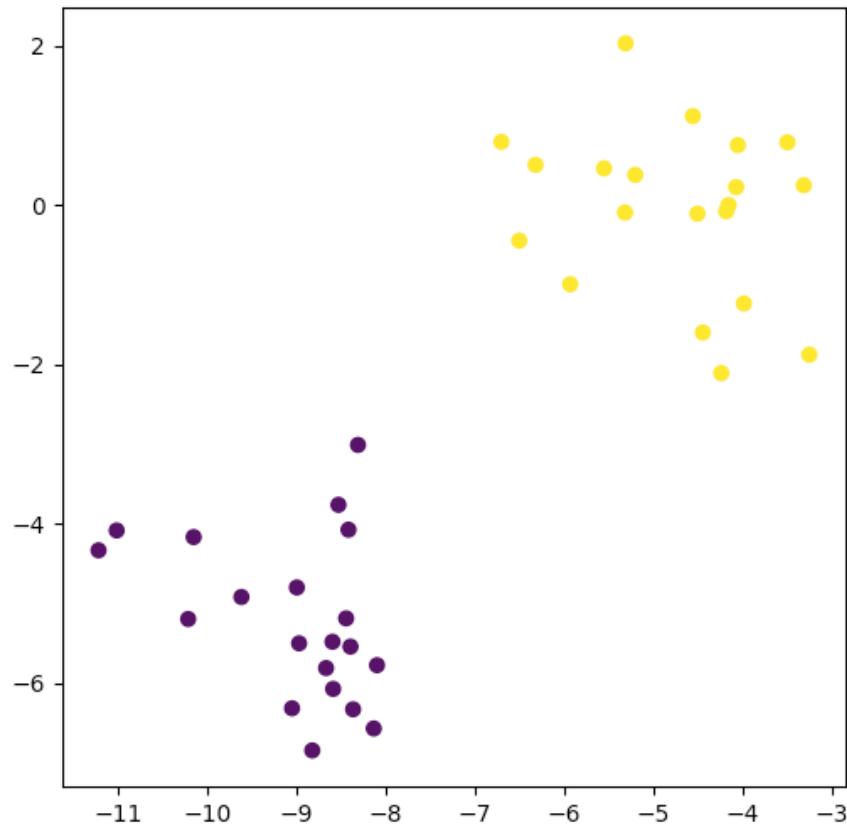
We are going to create two clusters, based on the initial centres that are shown by and .

How will the blue points be assigned to the clusters?

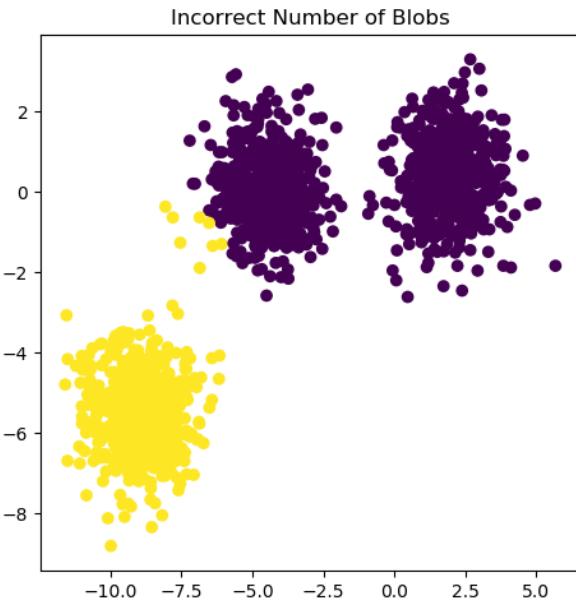
The result depends on the initialisation



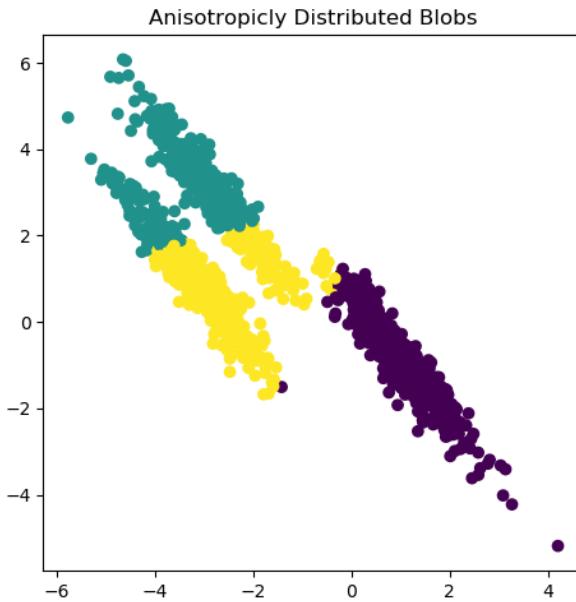
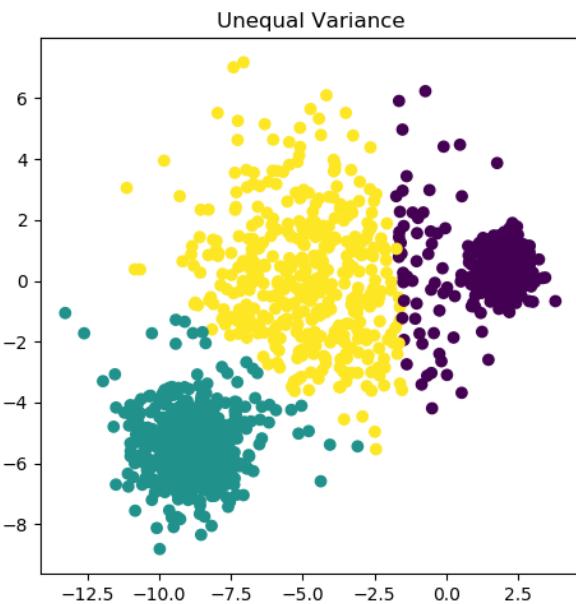
Moving one point, which is then selected as an initial centroid



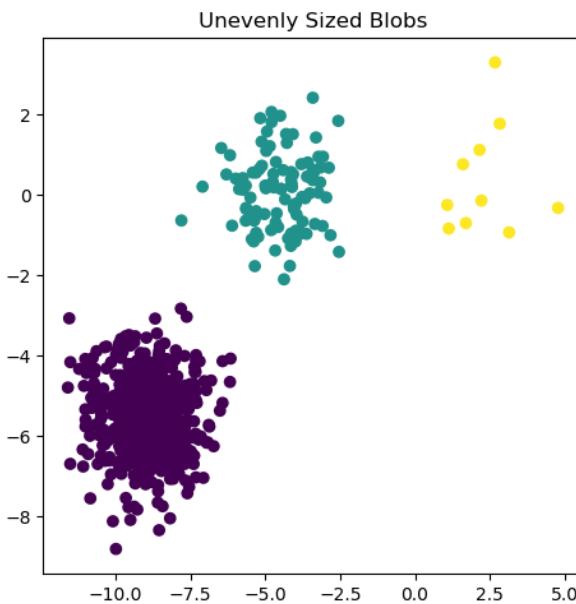
Clustering with k=2,
when we “see” that
there are 3 natural
clusters



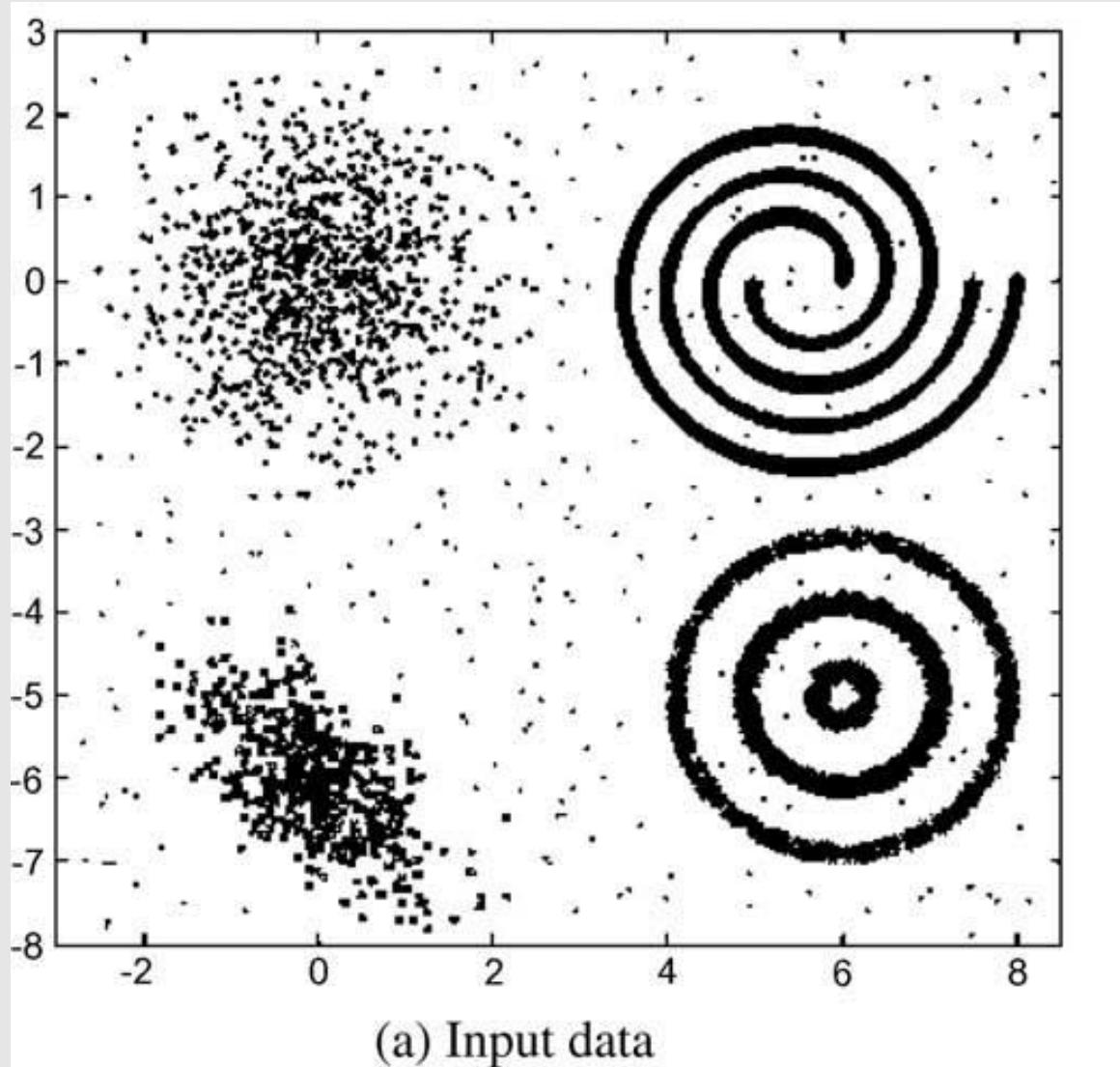
Two of the clusters
have very small
variance



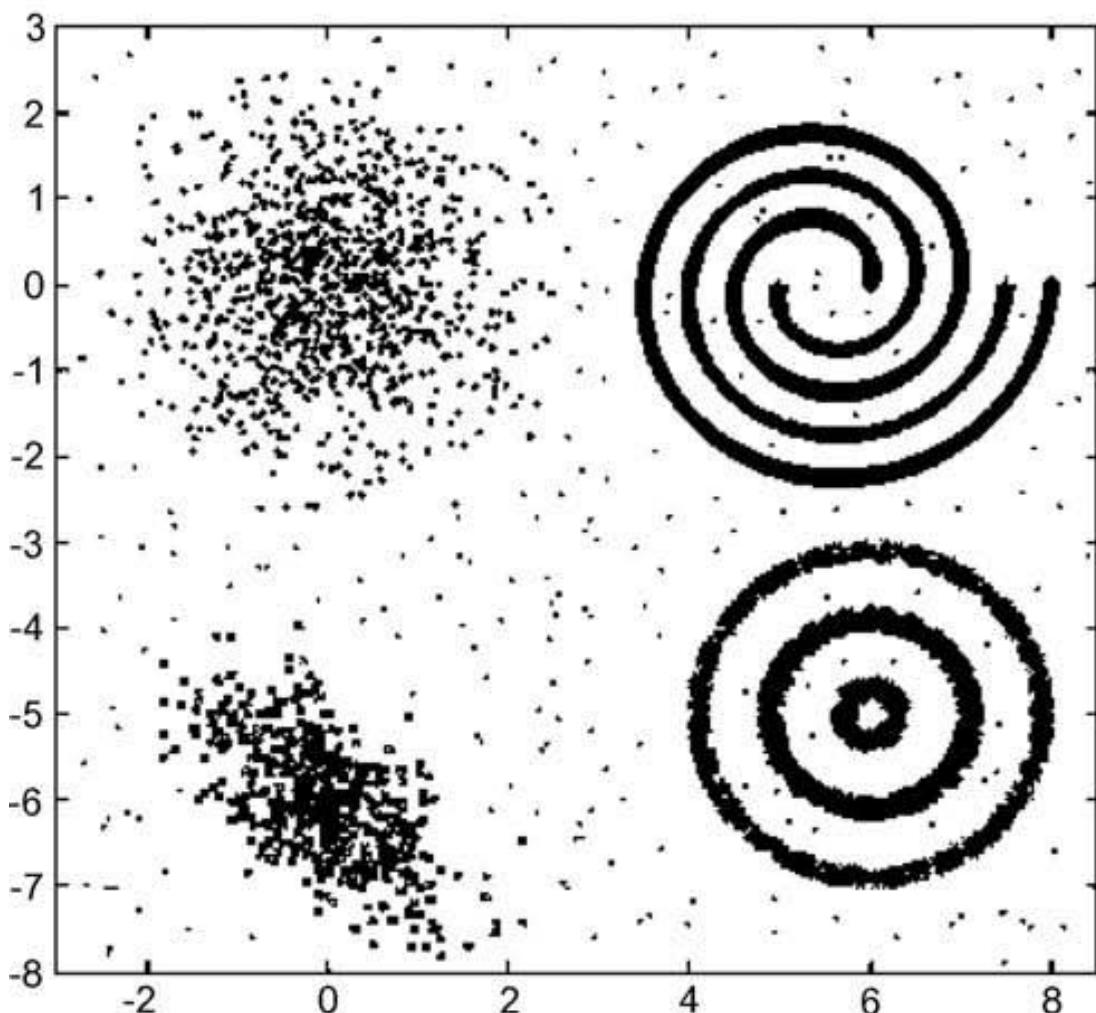
Different properties
in different
directions (result
here is that blobs
are not “spherical”)



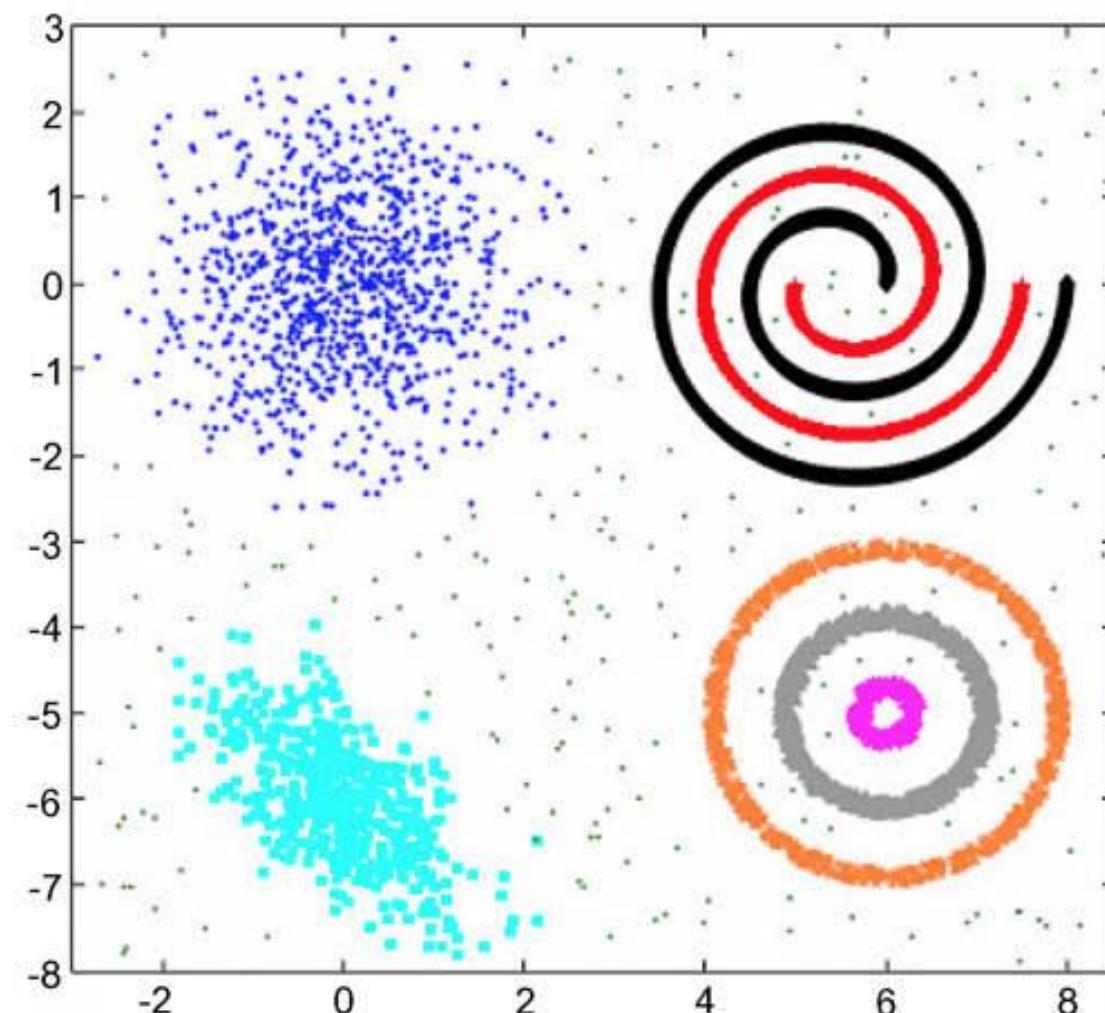
Intuitive result, even
though clusters
differ in size
(number of cluster
members)



- How many clusters should be found?



(a) Input data



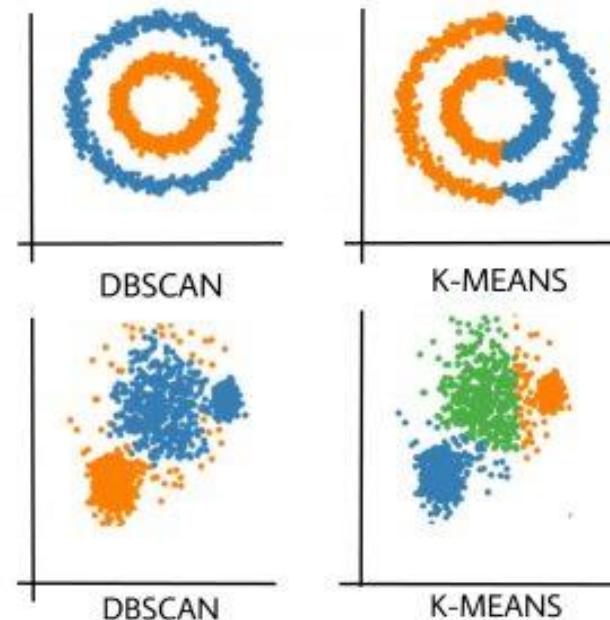
(b) Desired clustering

Limitations of K-means



Next time

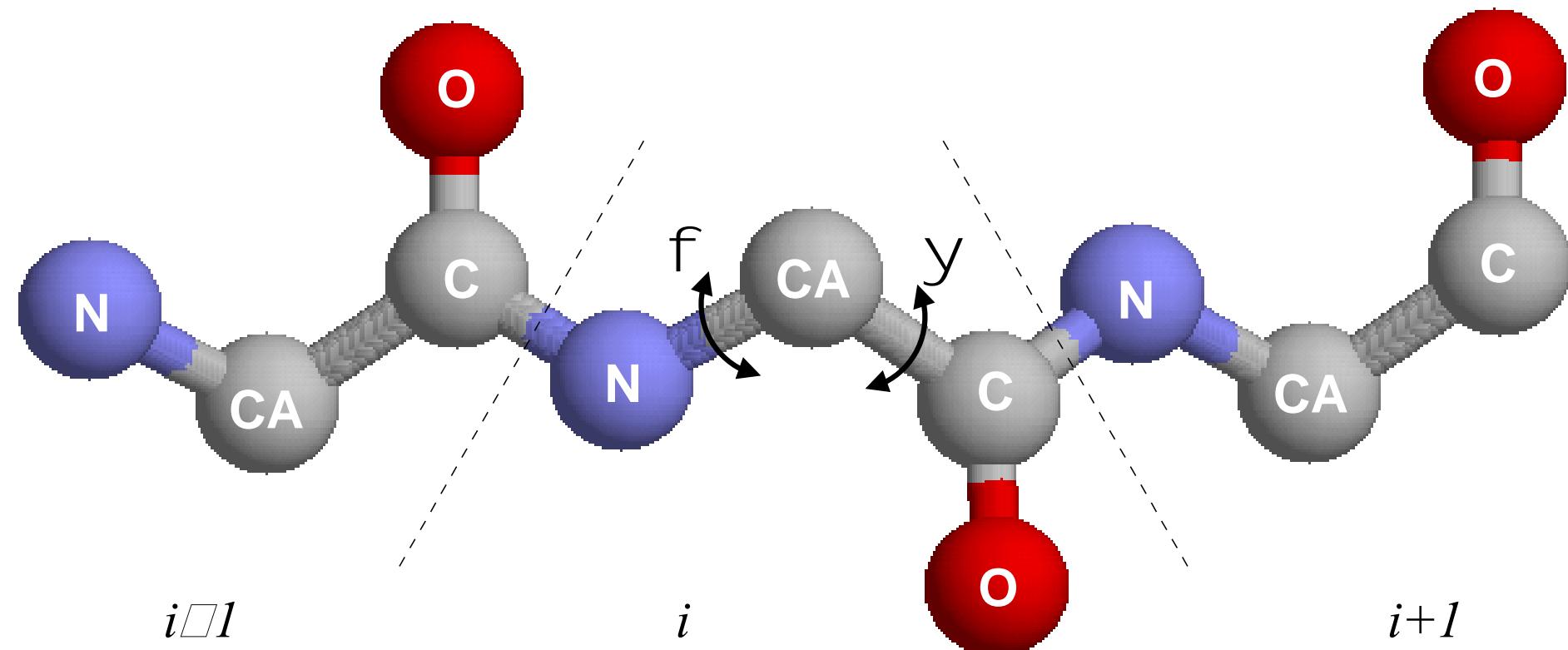
- We shall look at density-based clustering (e.g. DBSCAN) and hierarchical clustering.



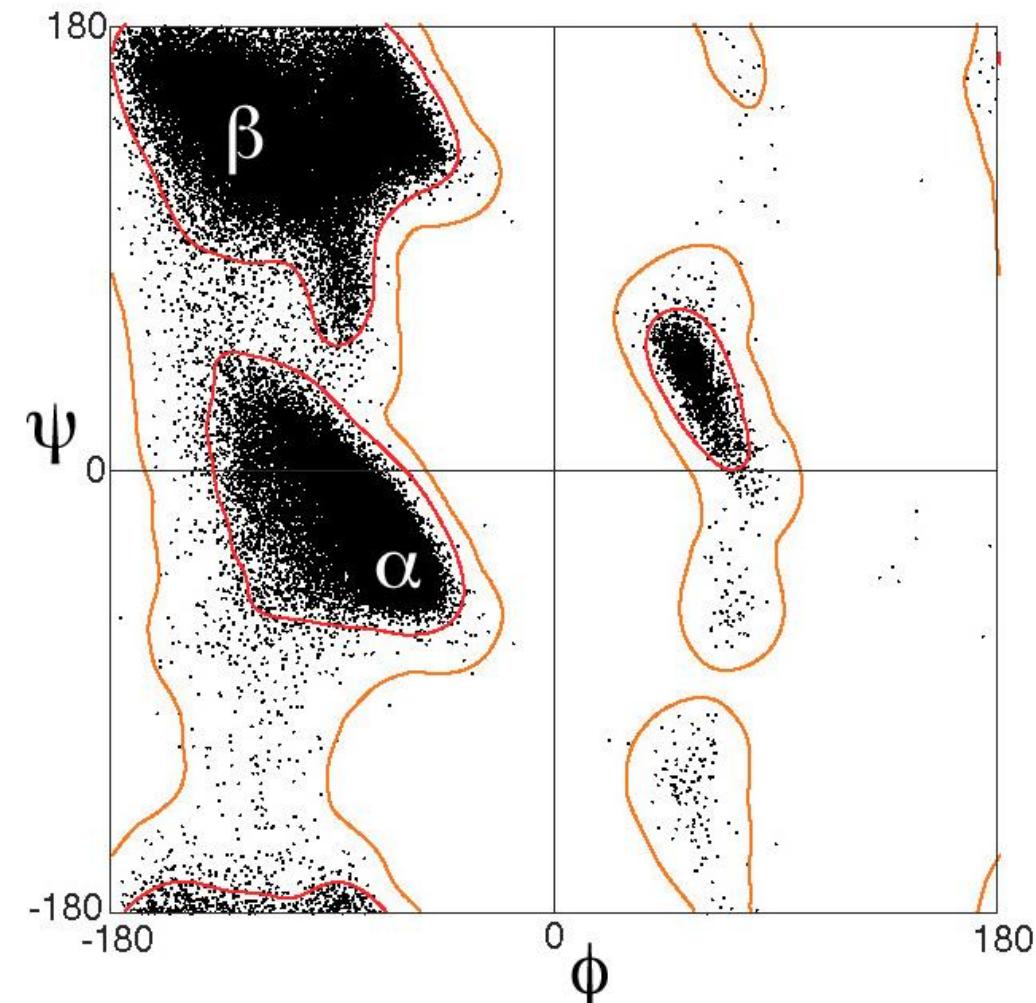
Assignment 3

Using K-means and density-based clustering to cluster the main chain conformations of amino acid residues in proteins.

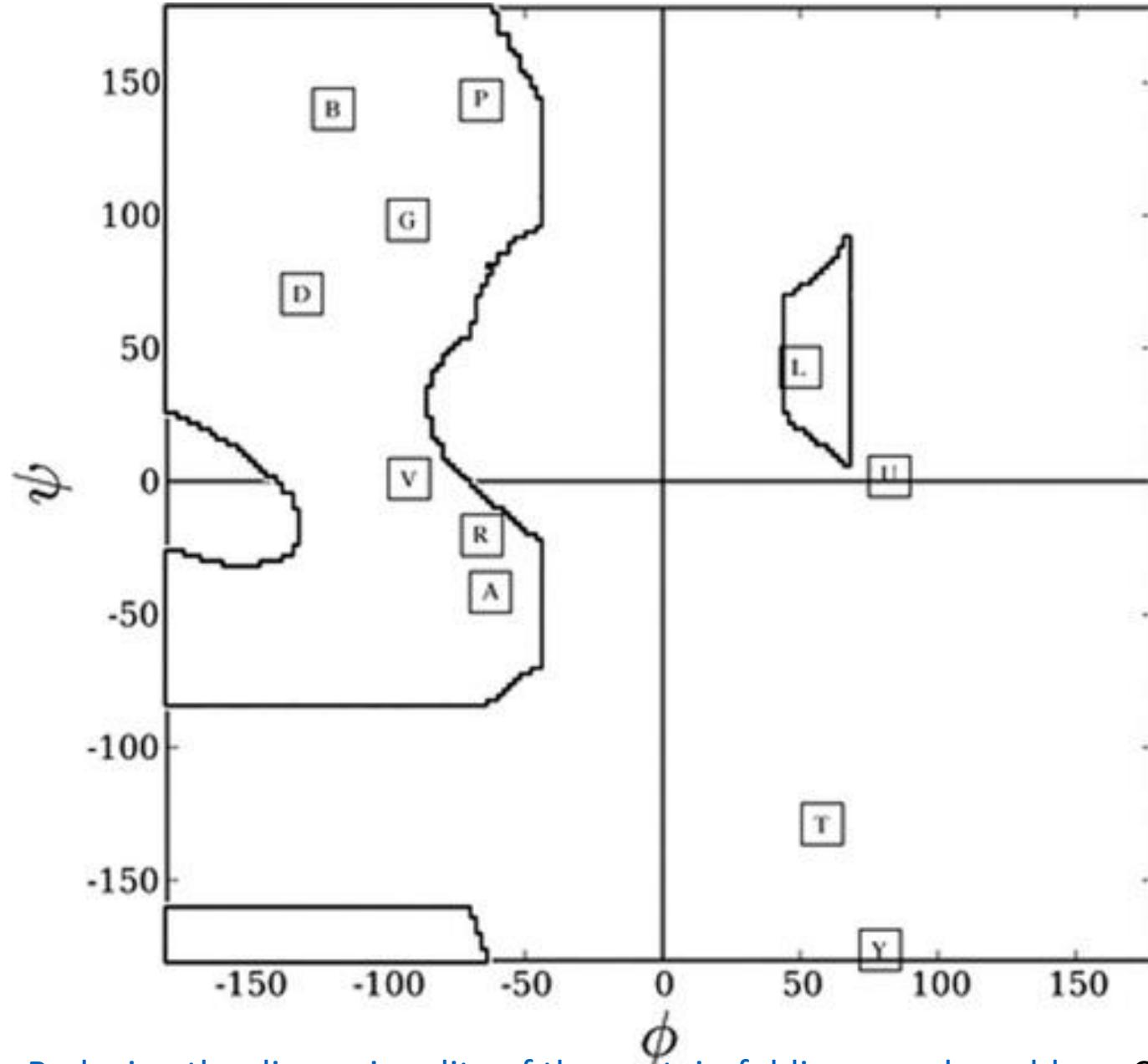
Protein main chain



Ramachandran plot



Around 100000
data points
shown here



“the 11 most populated residue basins in a database of high-resolution protein structures”