CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

MODULE 3: CLUSTERING

DAT405 / DIT407, 2022-2023, READING PERIOD 4

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

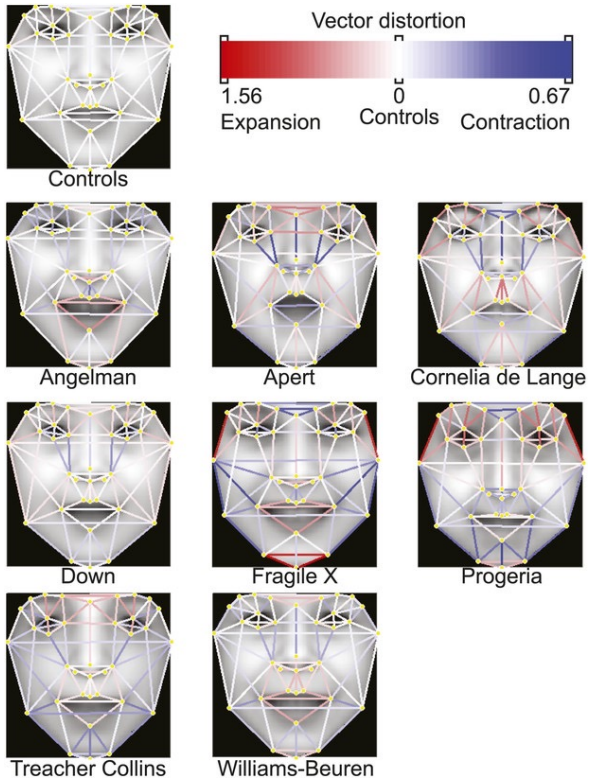
Purpose of data clustering

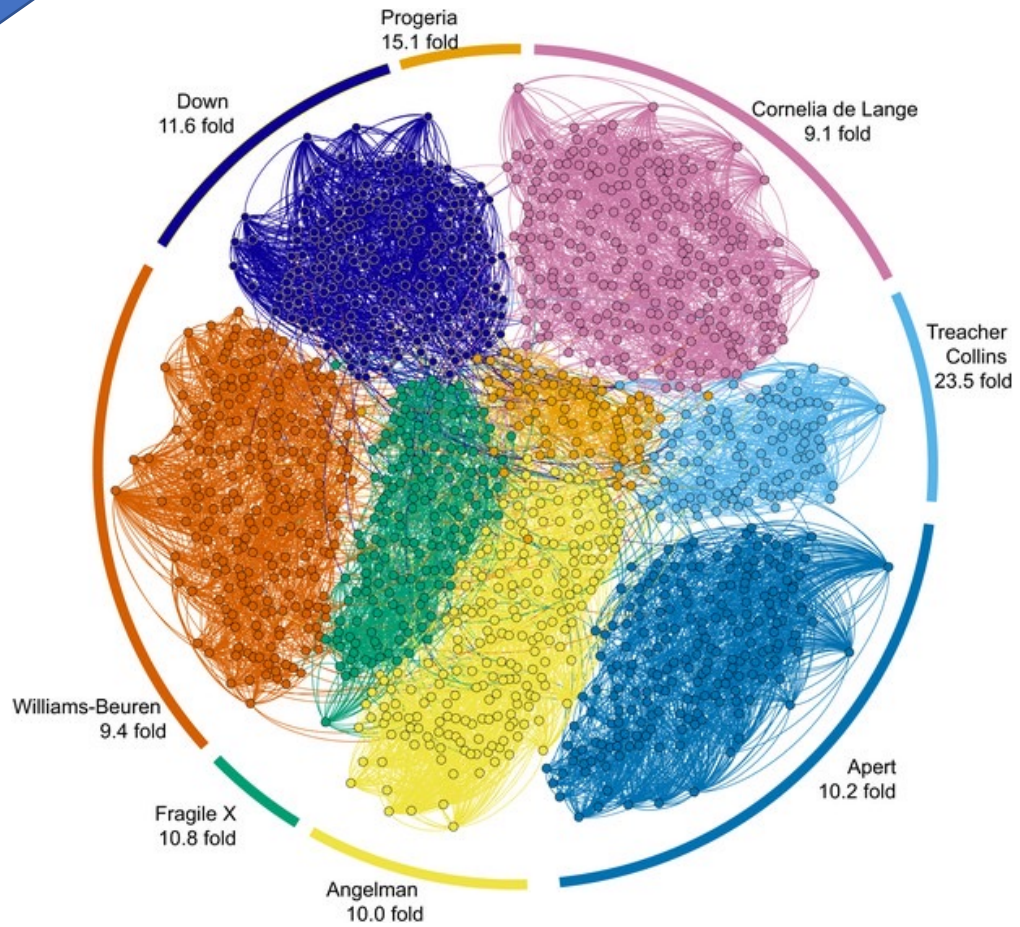
- Underlying structure
  - to gain insights into data, generate hypotheses and identify salient features
- Natural classification
  - to identify the degree of similarity among forms or organisms (phylogenetic relationships)
- Compression
  - to organise data and summarise it through cluster prototypes

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

Clustering faces

Helps discovering and diagnosing genetic disorders

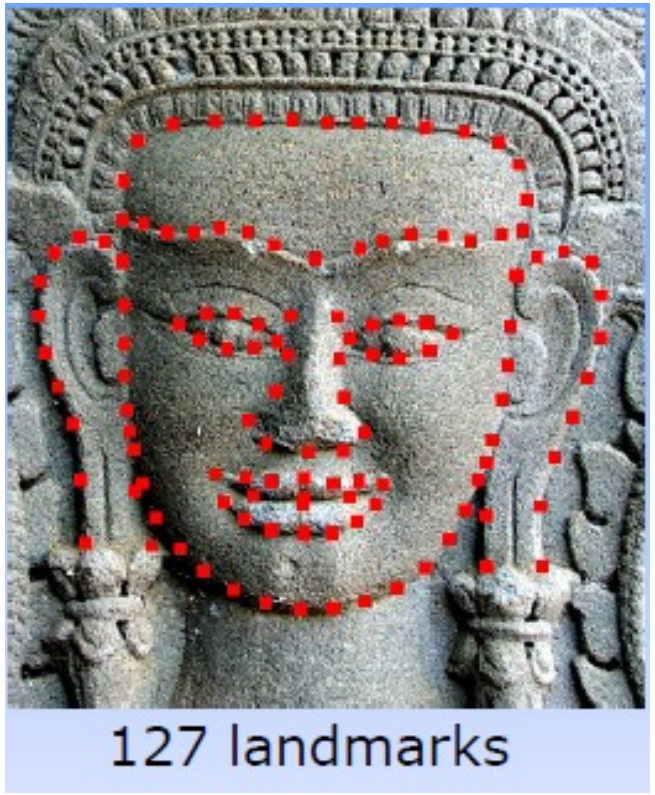




Ferry et al.. eLife (2014)

Clustering statues

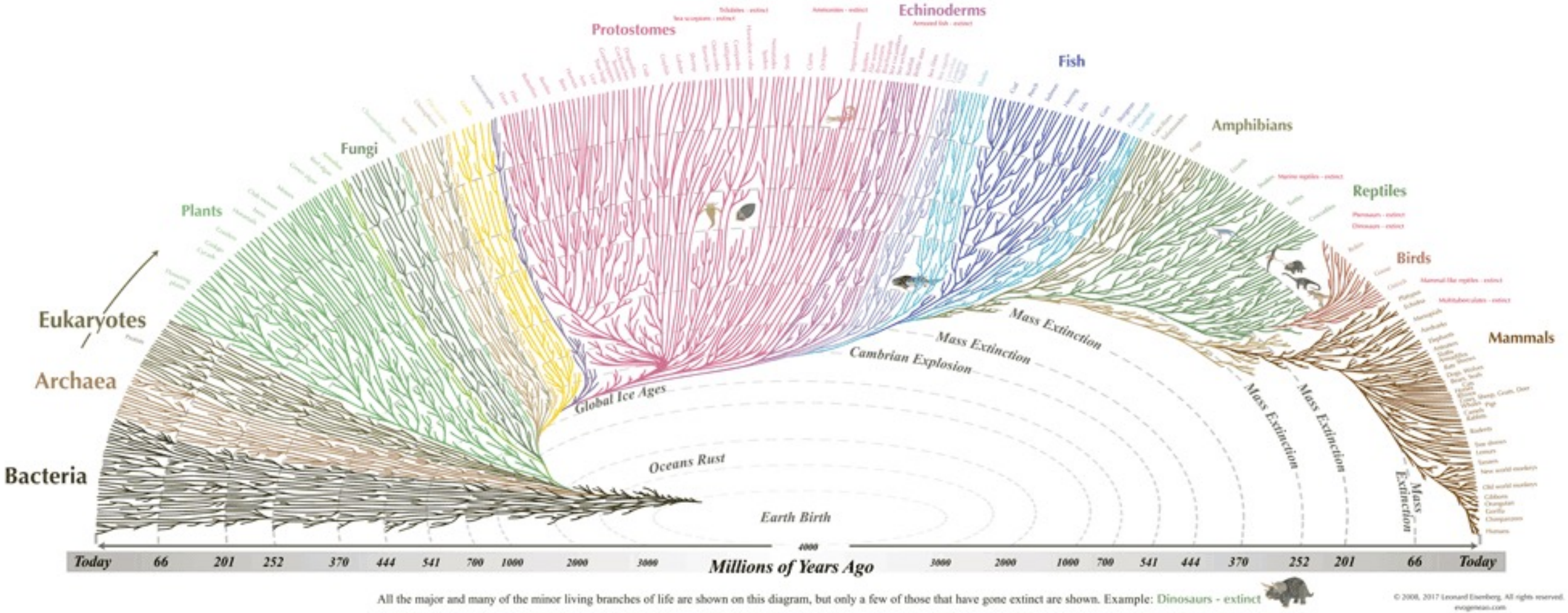
Helps ethnologists understand cultural influences at Angkor Wat via its 2000 statues



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

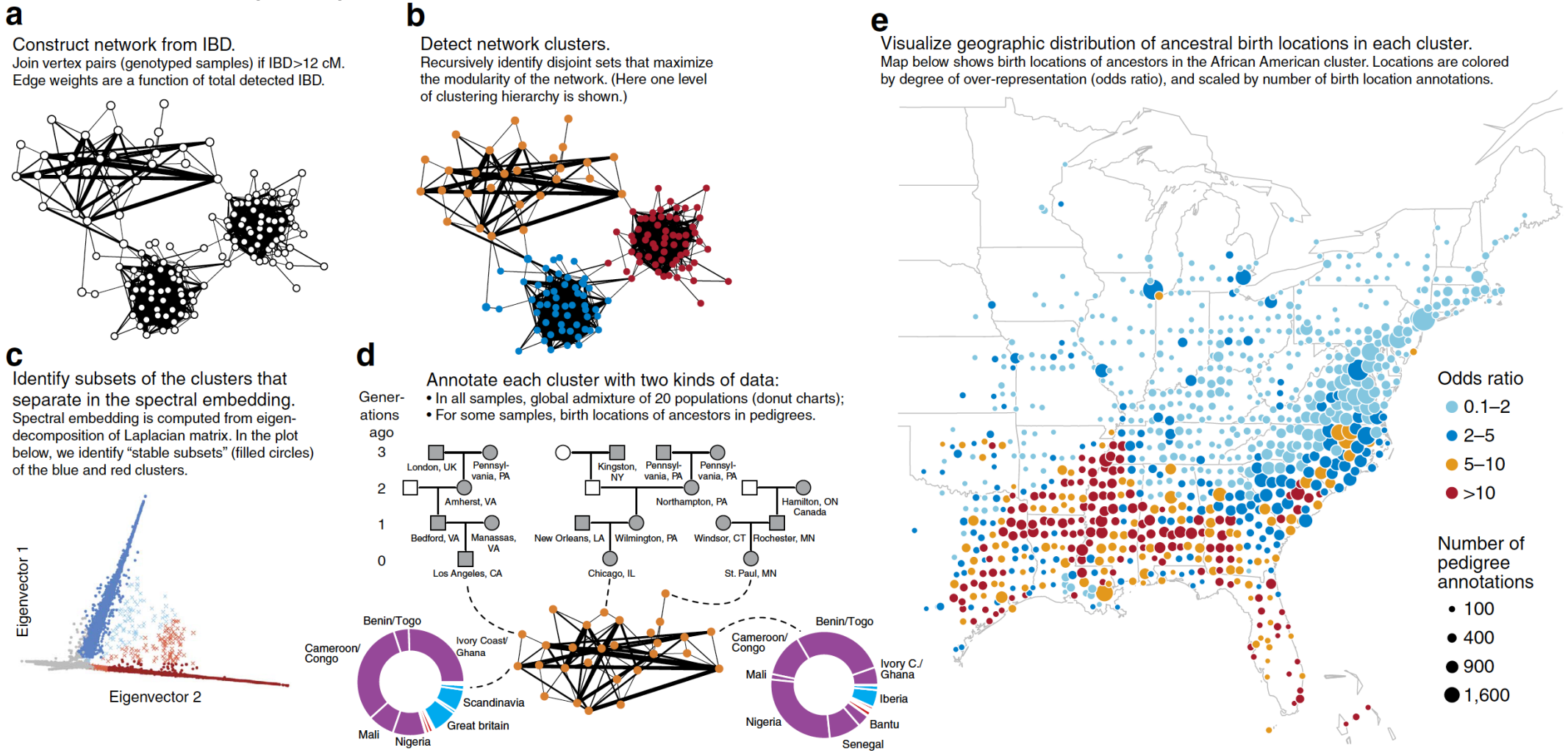
Images from Jain, A.K. (2008) Data clustering: 50 years beyond K-means. King Sun Fu Prize Lecture, ICPR

Tree of life

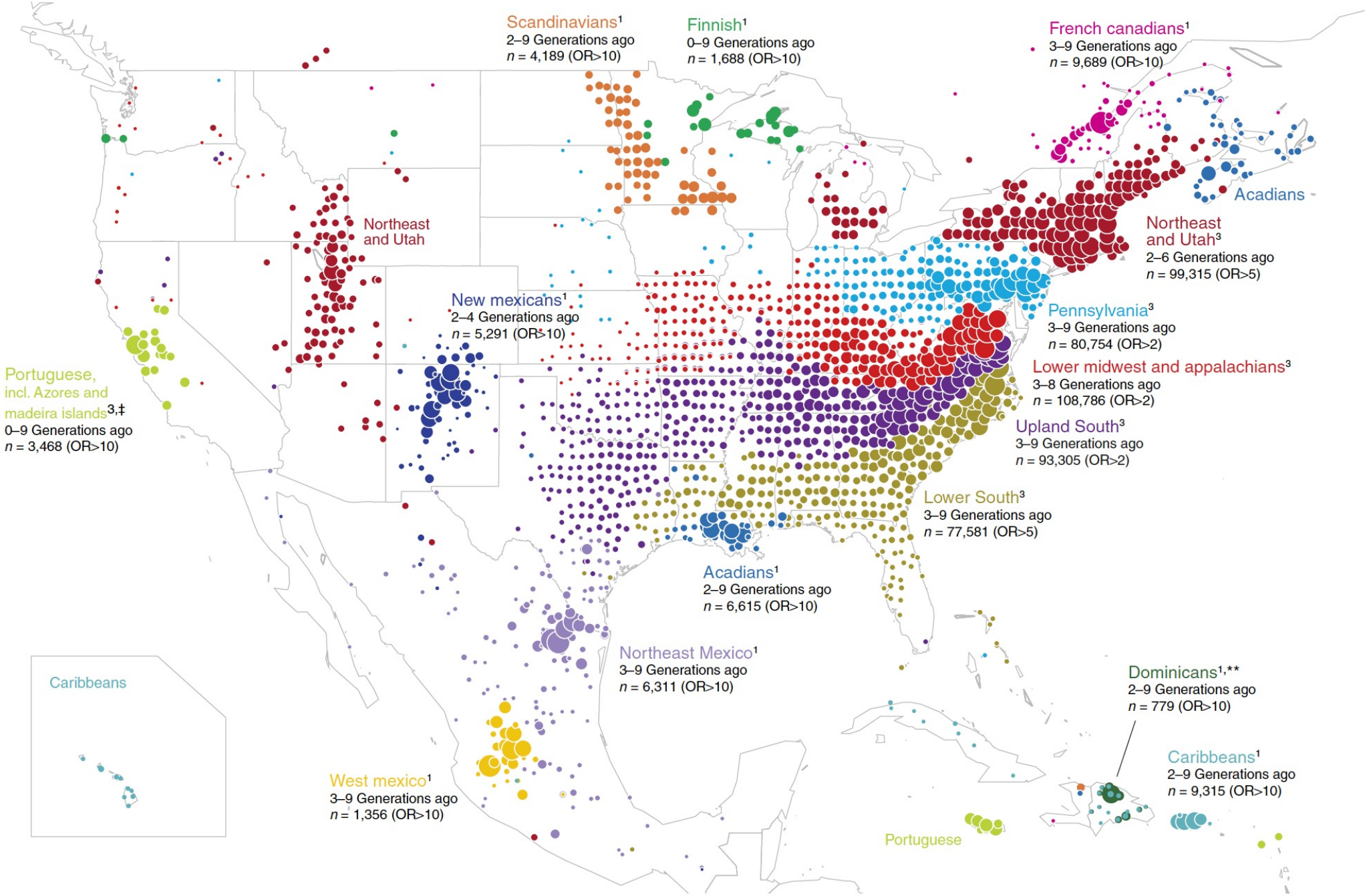


https://www.evogeneao.com/

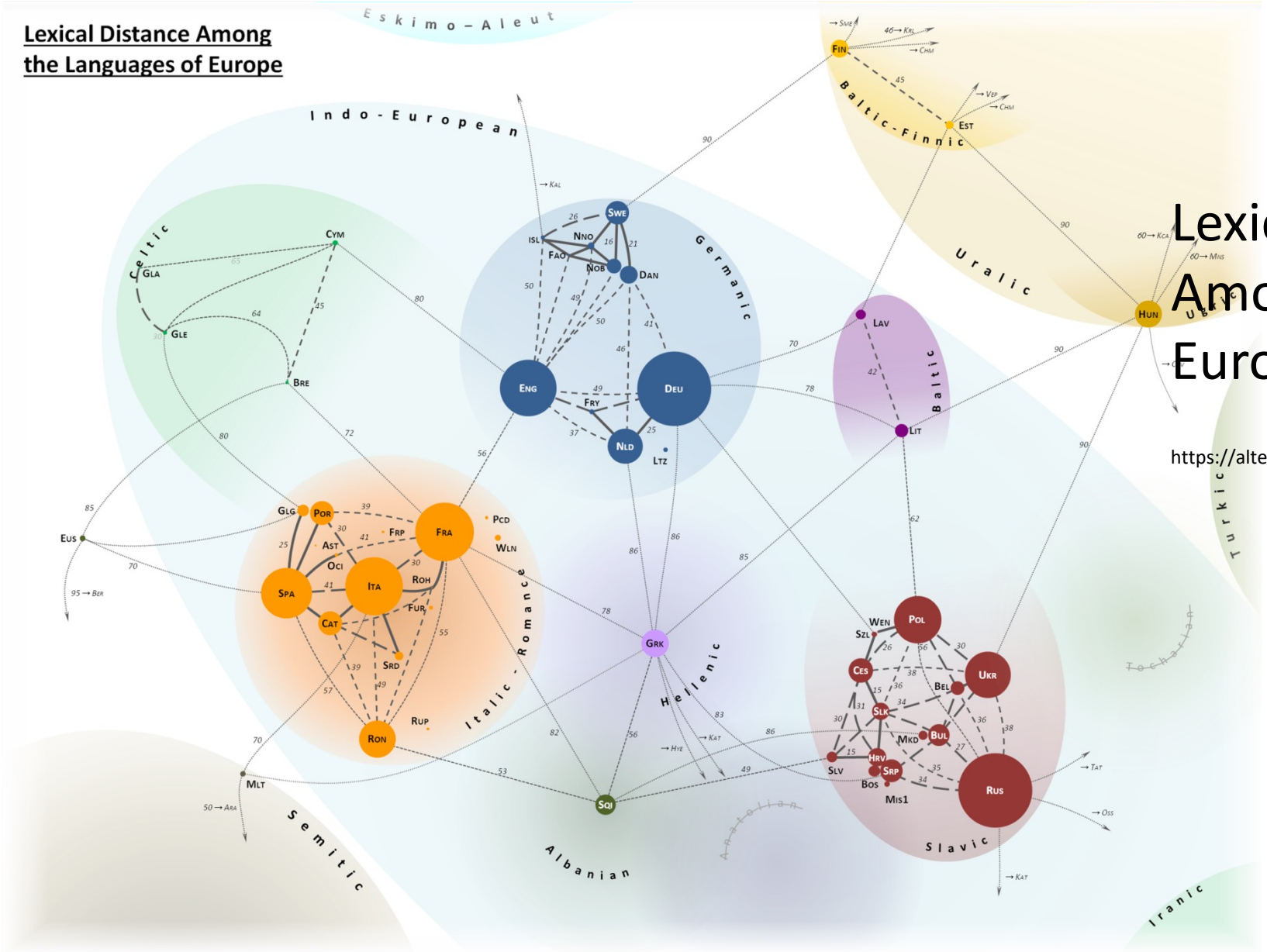
Clustering of 770,000 genomes reveals post-colonial population structure of North America



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



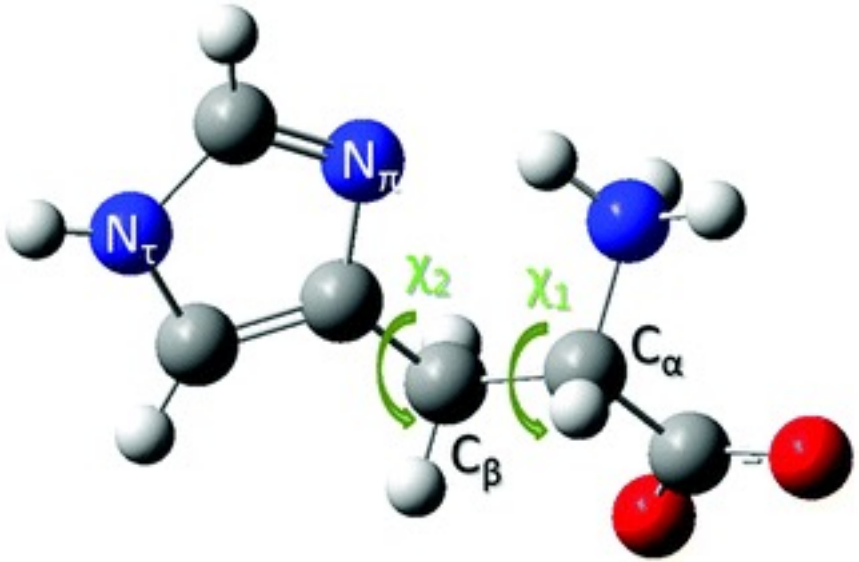
Clustering of 770,000 genomes. Han et al. Nature Comm. 2016



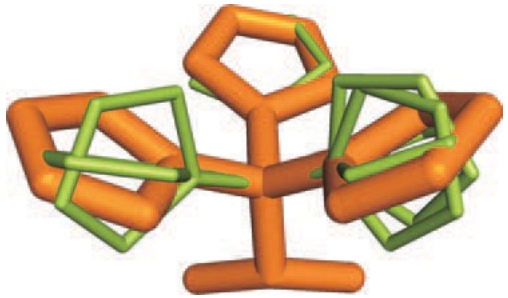
Lexical Distance Among Languages of Europe 2015

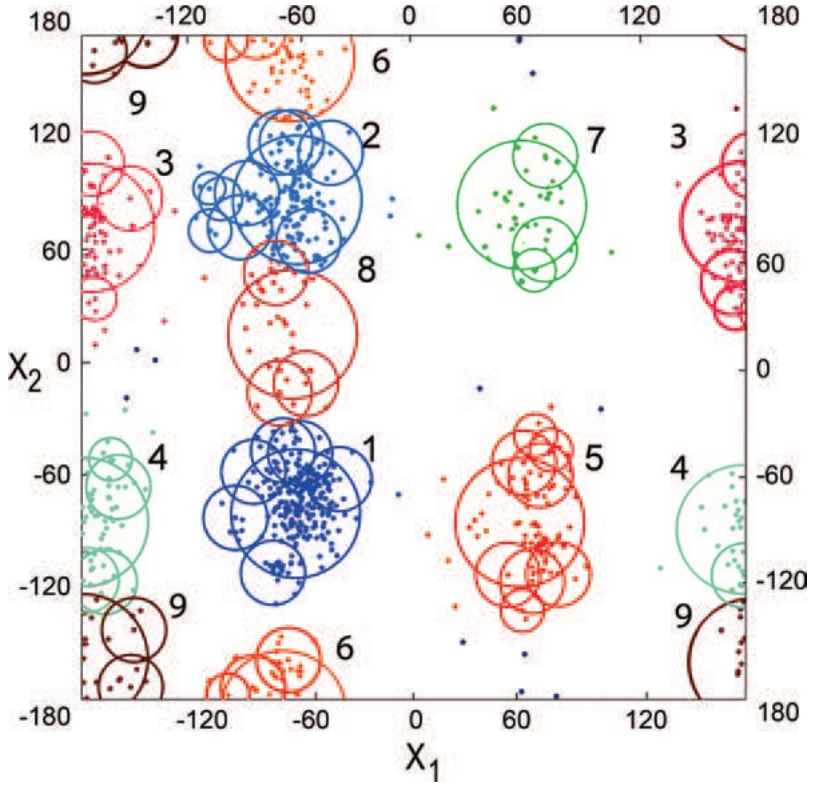
https://alternativetransport.wordpress.com/2015/05/05/34/

Protein side chain rotamers

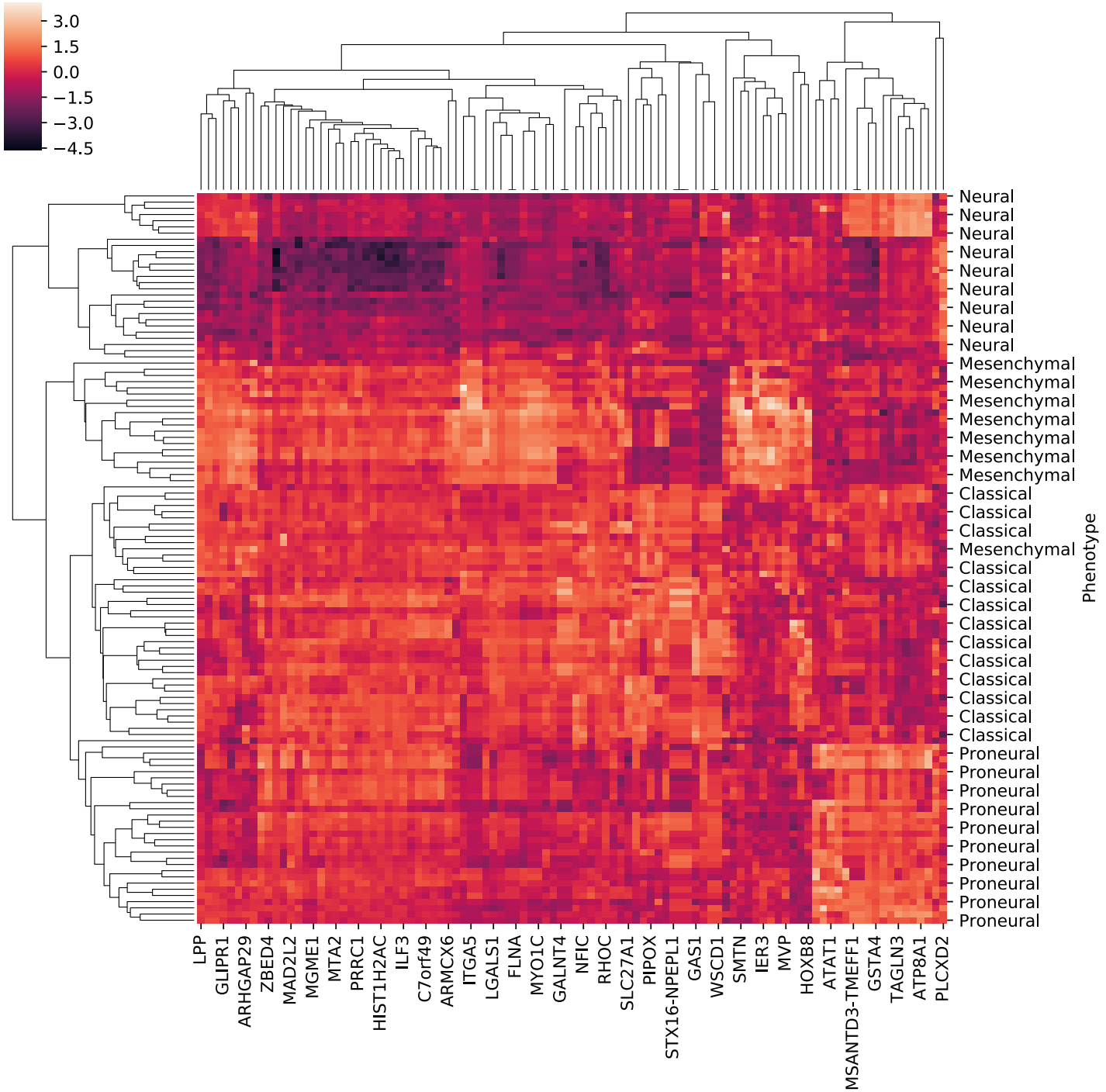


Cardamone et al (2016), Phys. Chem. Chem. Phys, 18, 27377-27389





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).

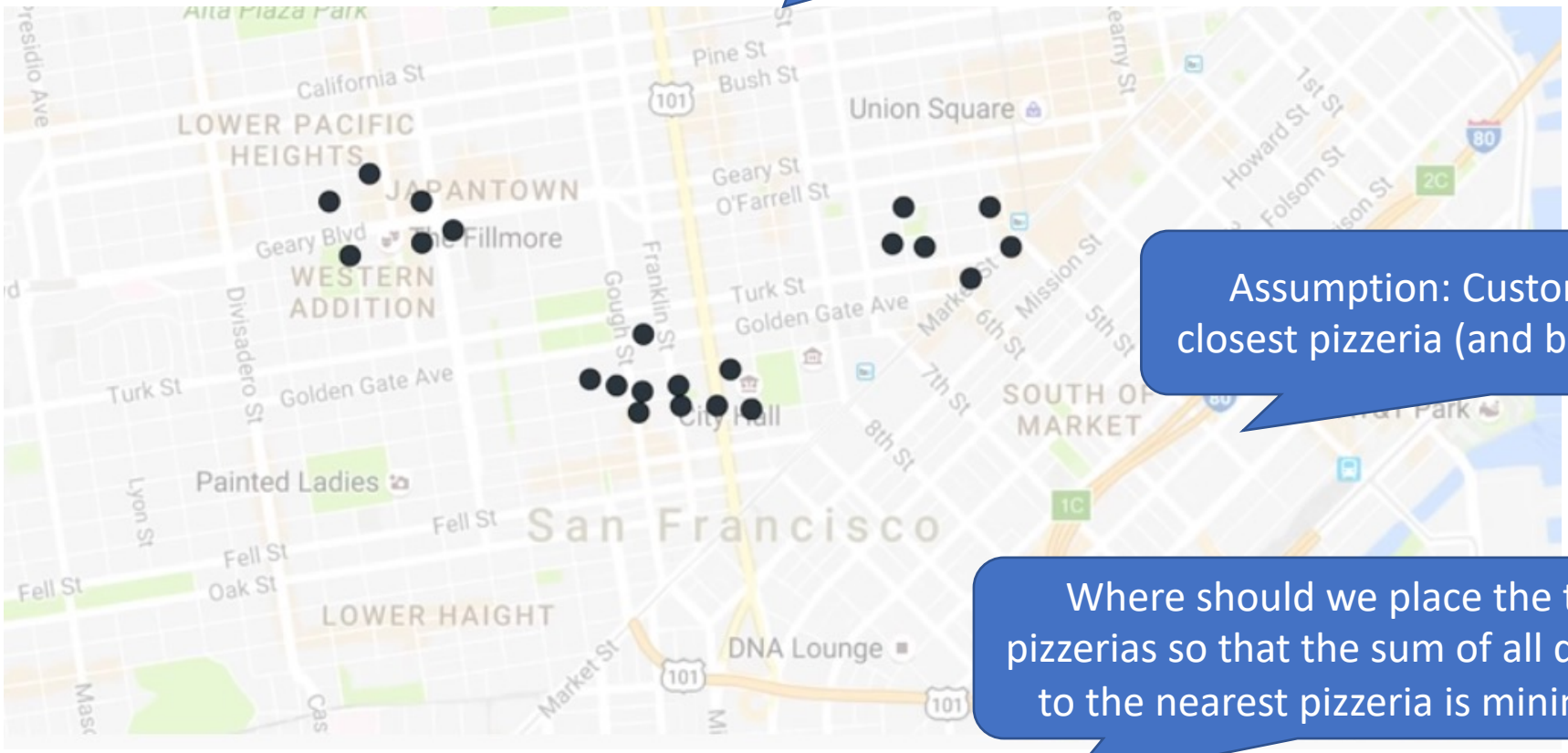
K-means clustering

Pizzerias in San Francisco

"Clustering: K-means and Hierarchical", Luis Serrano, https://www.youtube.com/watch?v=QXOkPvFM6NU

Problem

We want to start three pizzerias that serve these customers

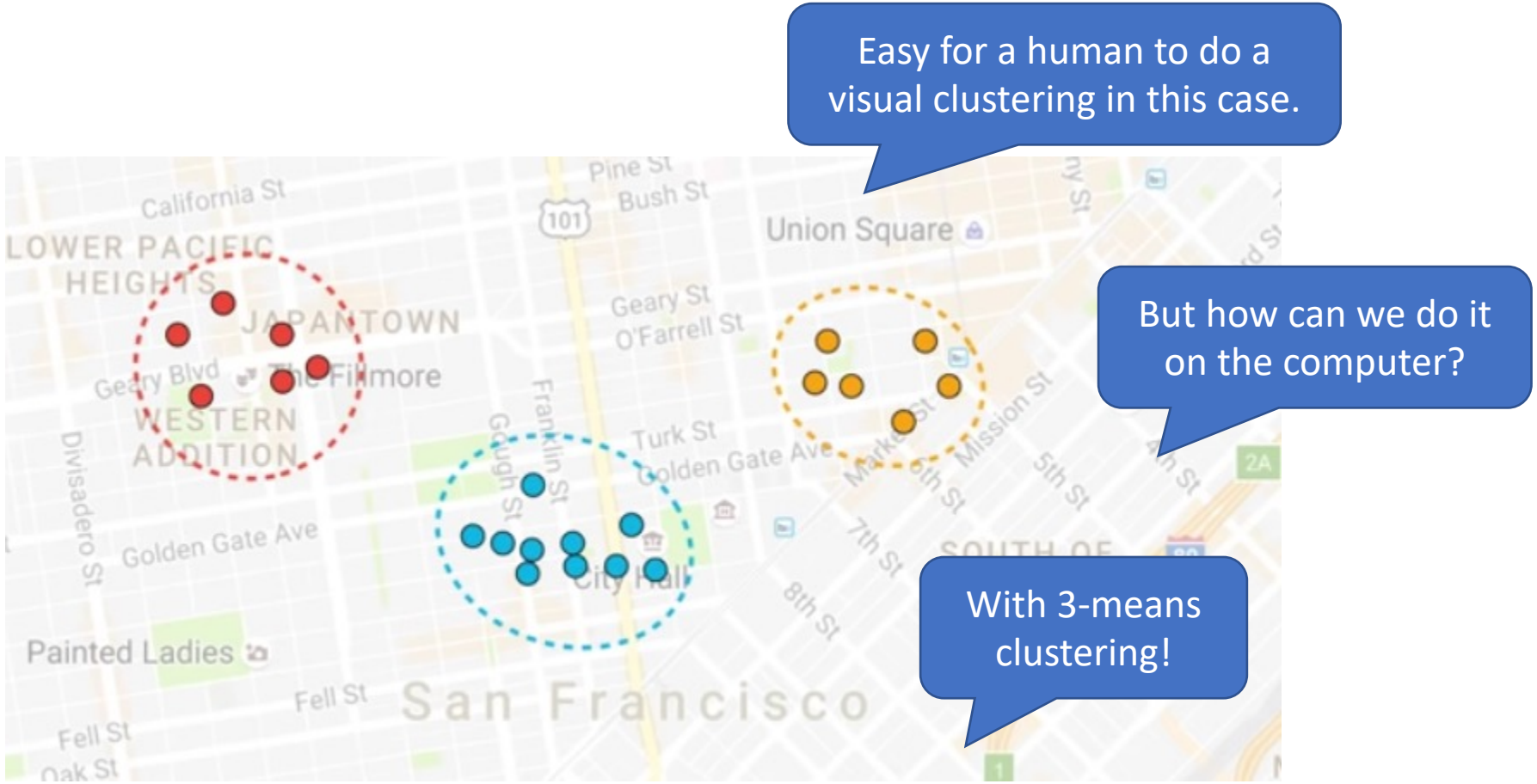


Assumption: Customers will go to the closest pizzeria (and buy more if it is close).

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

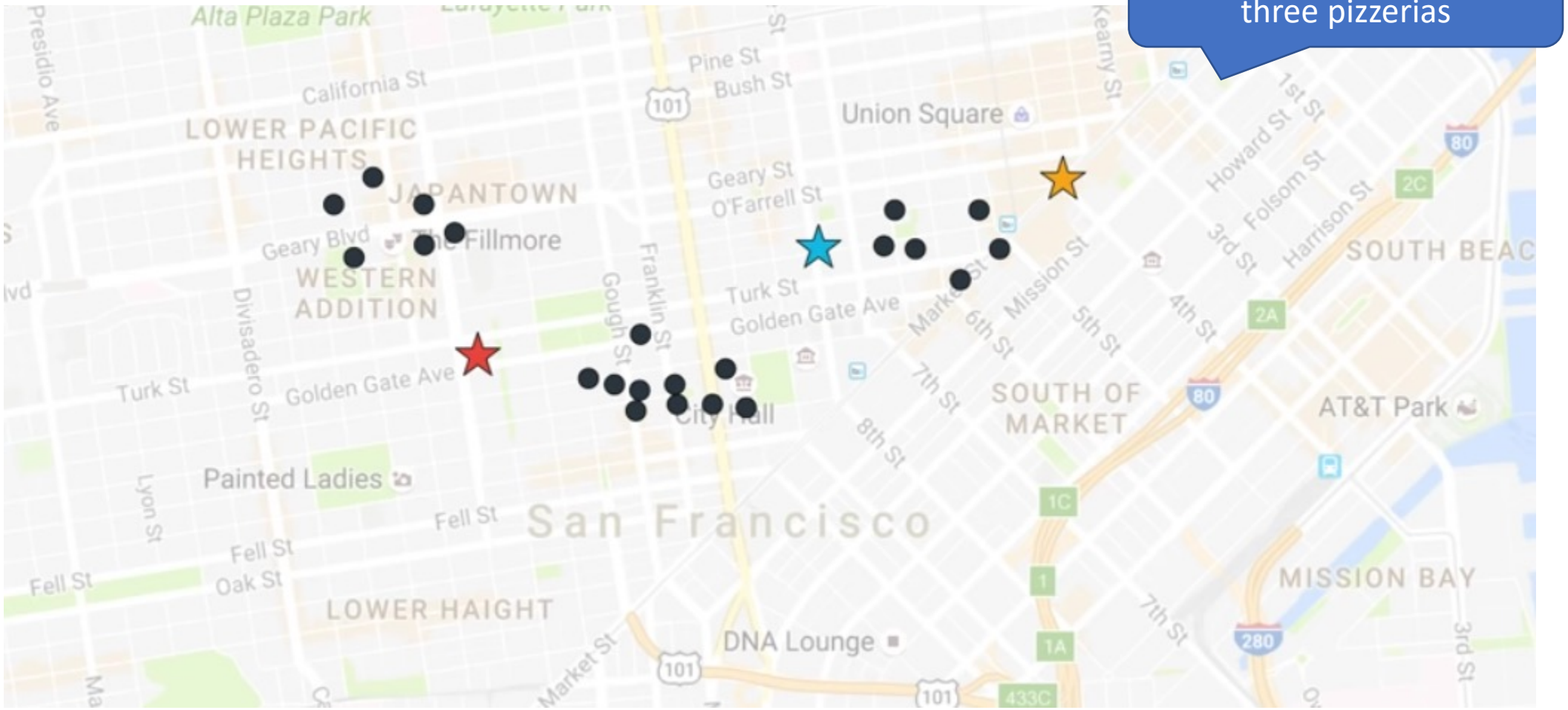
"Clustering: K-means and Hierarchical", Luis Serrano, https://www.youtube.com/watch?v=QXOkPvFM6NU

Visual clustering



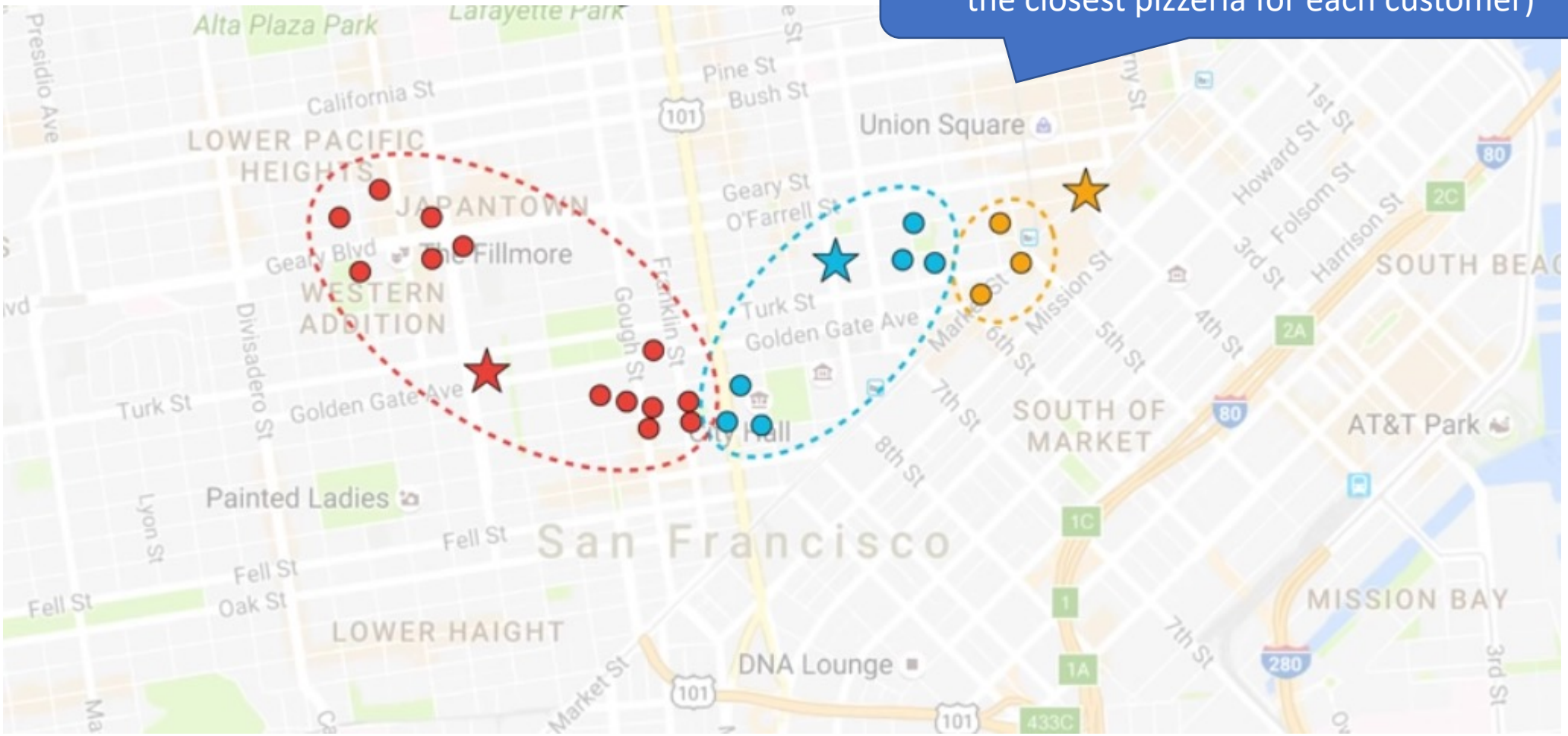
"Clustering: K-means and Hierarchical", Luis Serrano, https://www.youtube.com/watch?v=QXOkPvFM6NU

Initialize the centroids



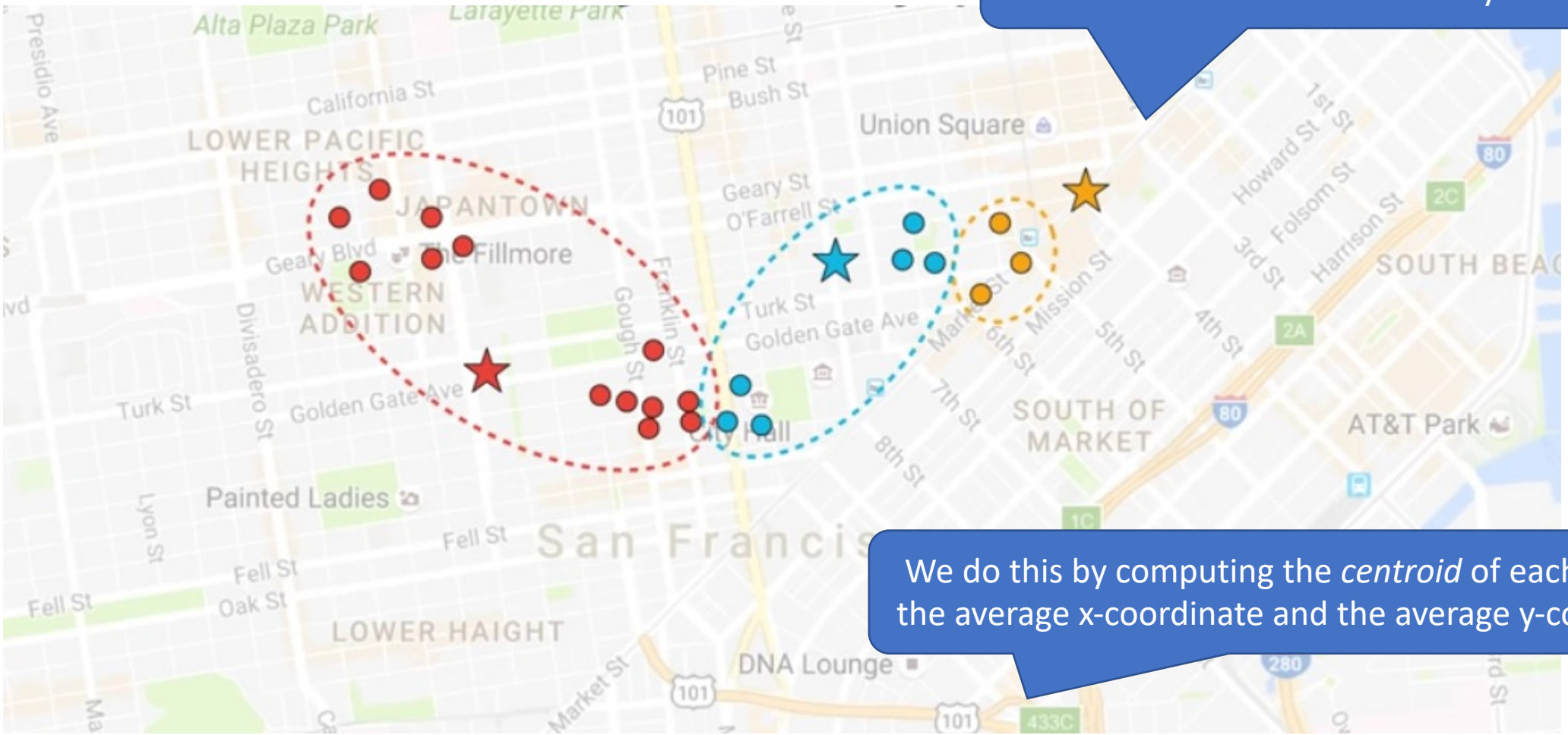
"Clustering: K-means and Hierarchical", Luis Serrano, https://www.youtube.com/watch?v=QXOkPvFM6NU

Form the clusters



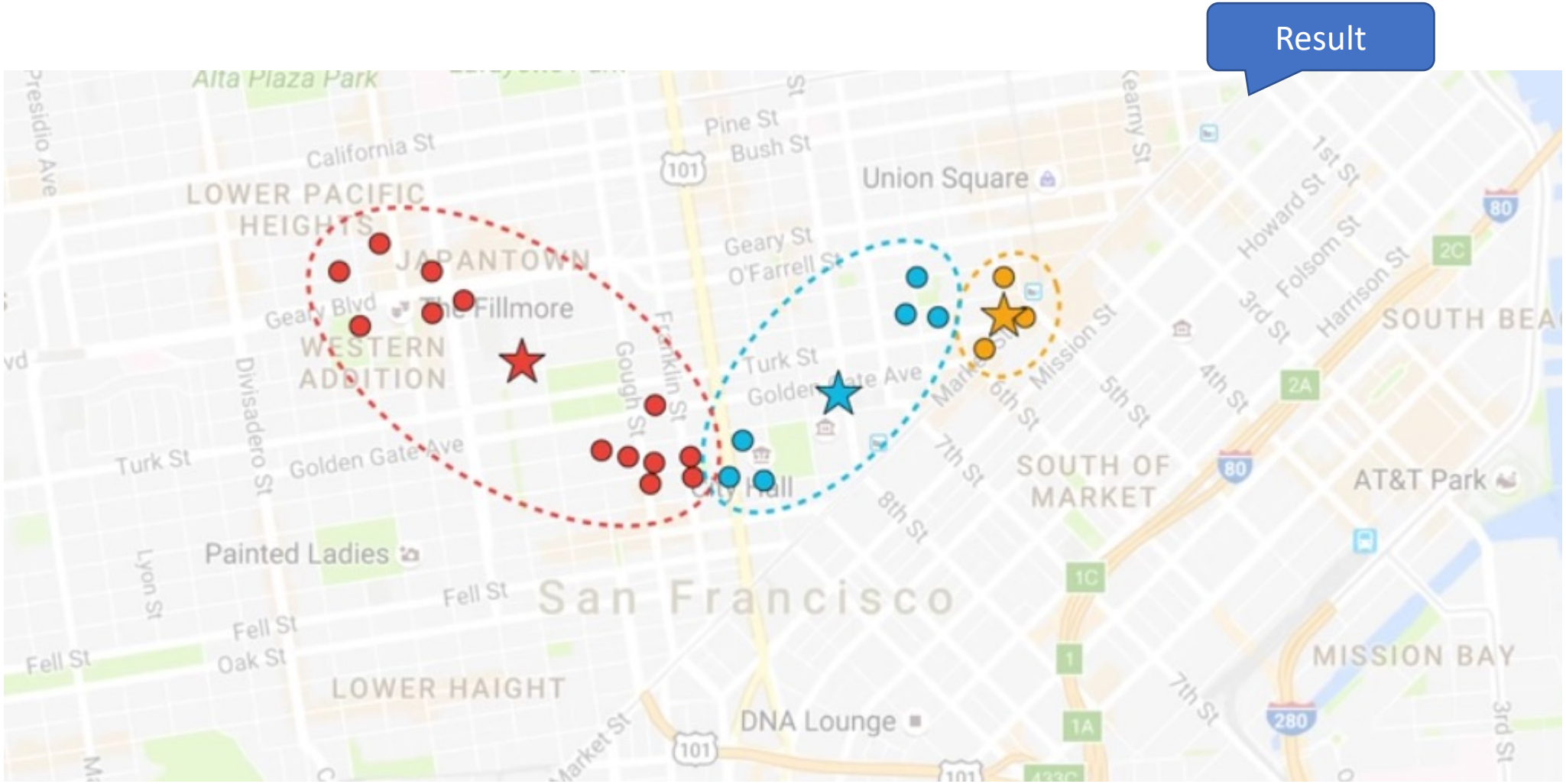
"Clustering: K-means and Hierarchical", Luis Serrano, https://www.youtube.com/watch?v=QXOkPvFM6NU

Move the centroids



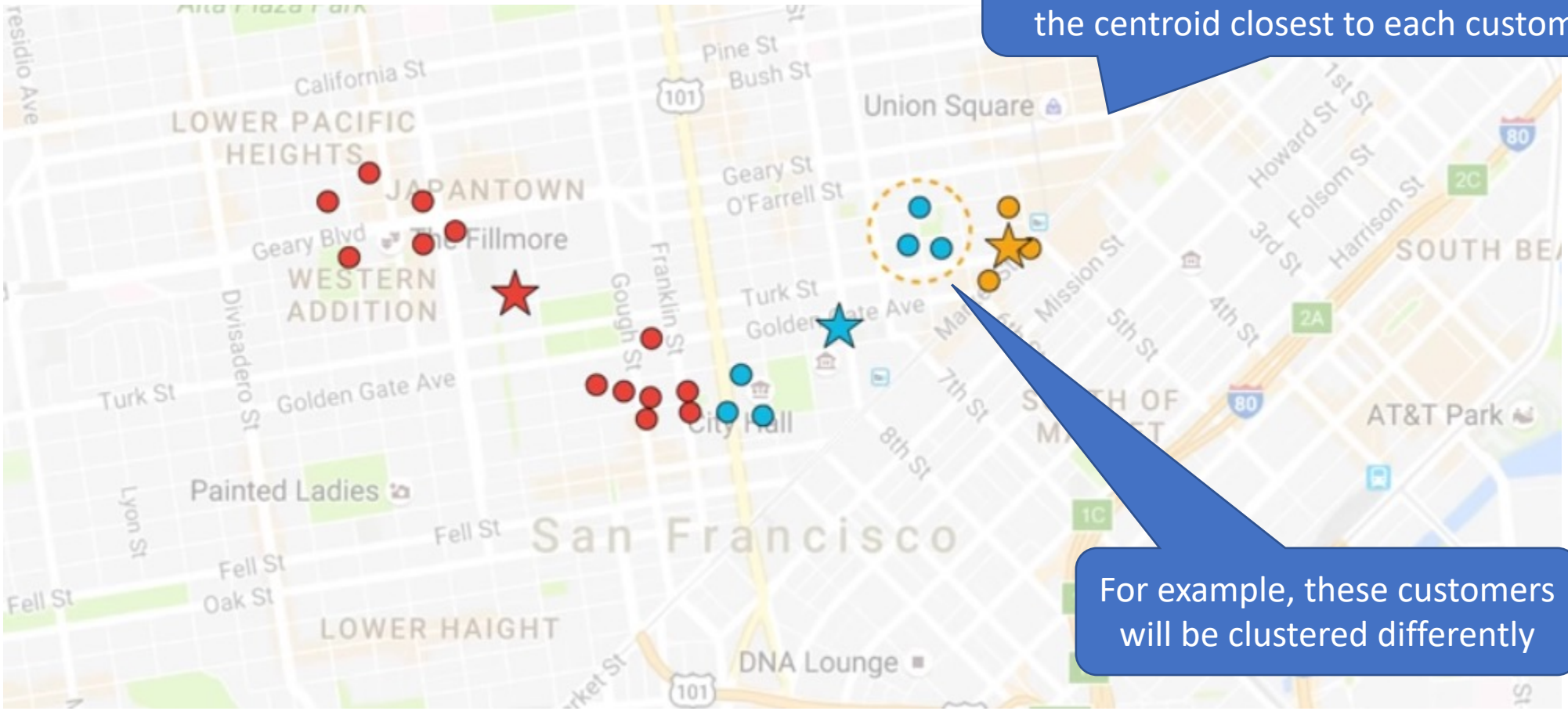
"Clustering: K-means and Hierarchical", Luis Serrano, https://www.youtube.com/watch?v=QXOkPvFM6NU

Move the centroids



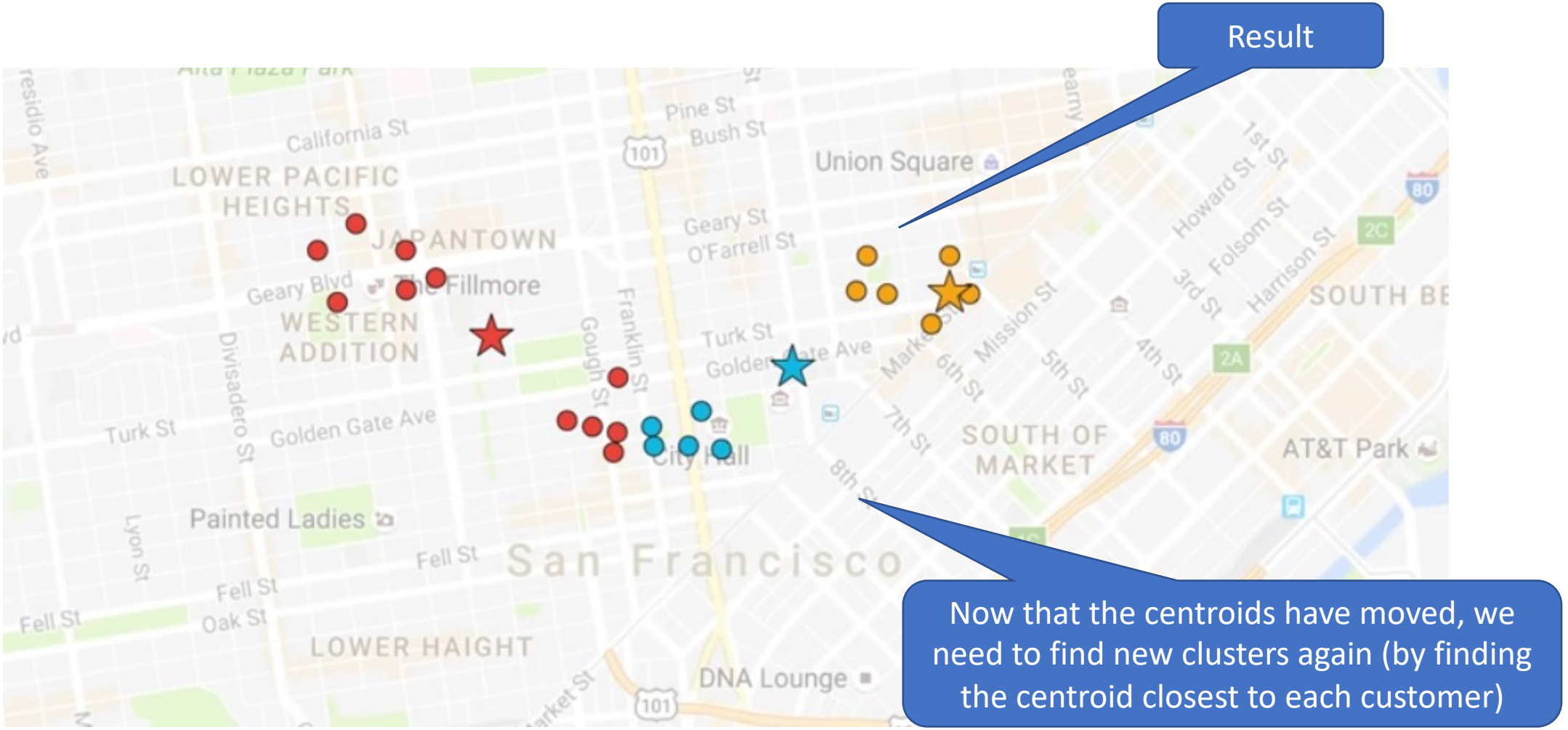
"Clustering: K-means and Hierarchical", Luis Serrano, https://www.youtube.com/watch?v=QXOkPvFM6NU

Find the clusters



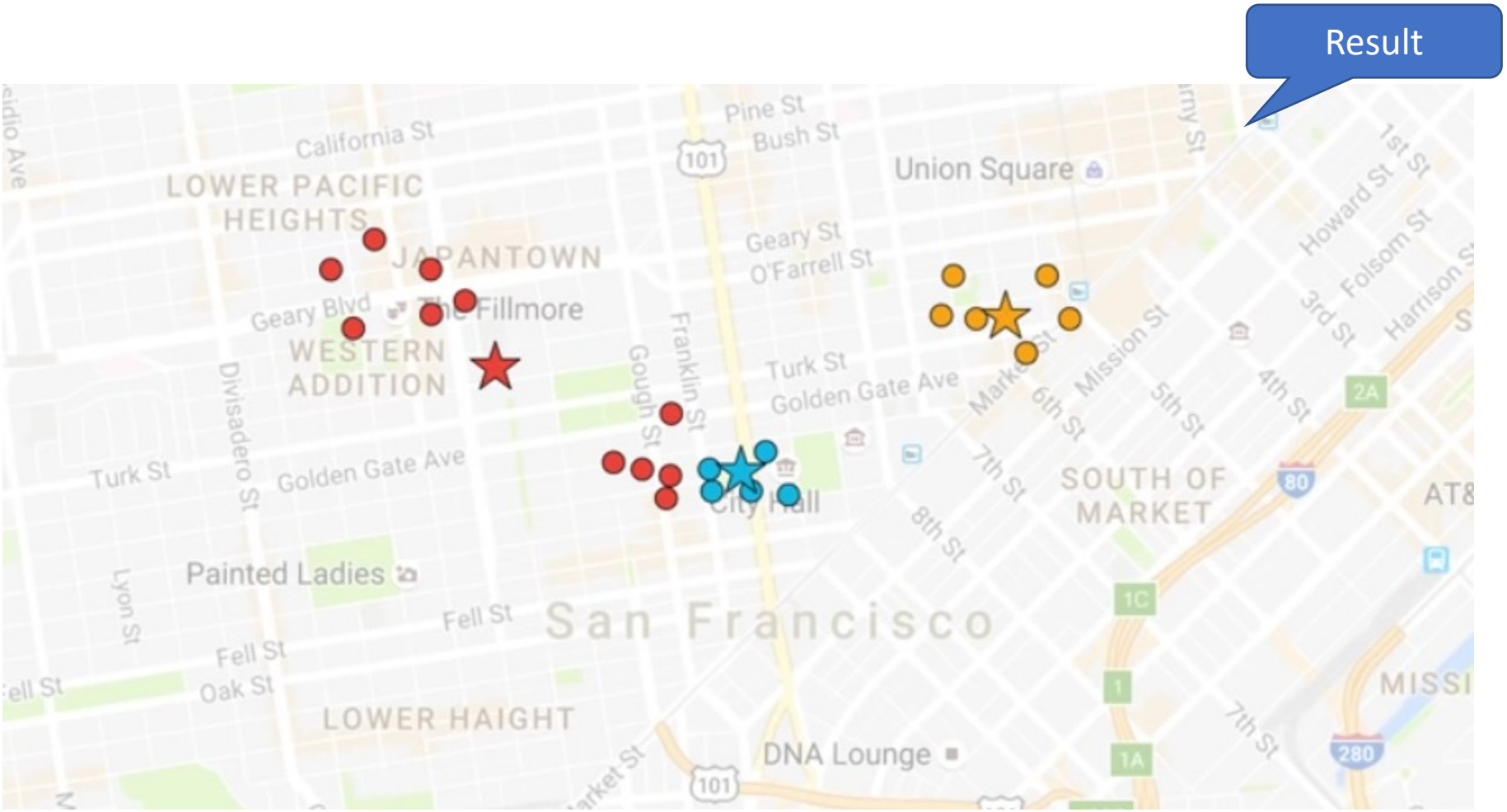
"Clustering: K-means and Hierarchical", Luis Serrano, https://www.youtube.com/watch?v=QXOkPvFM6NU

Find the clusters



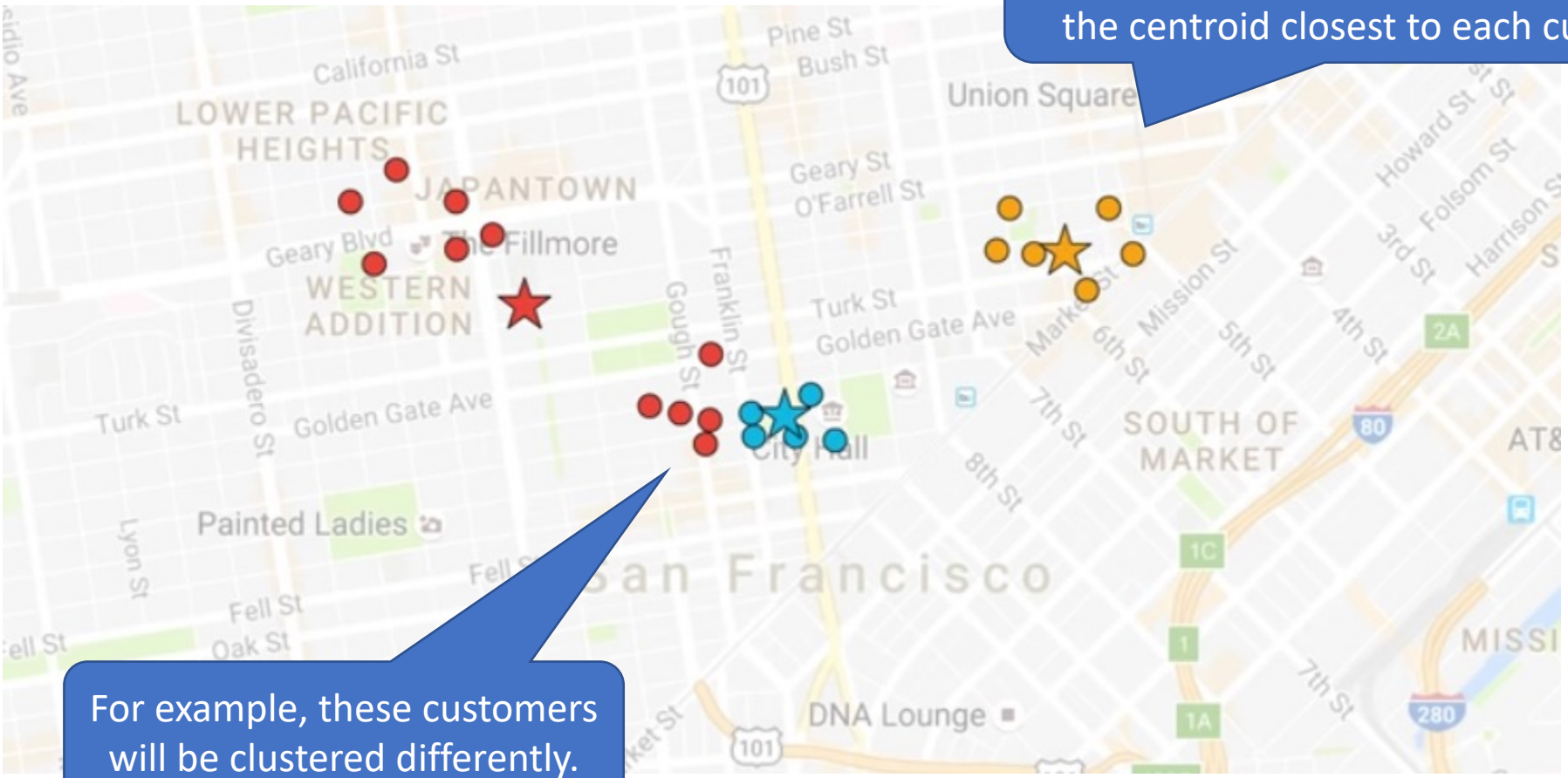
"Clustering: K-means and Hierarchical", Luis Serrano, https://www.youtube.com/watch?v=QXOkPvFM6NU

Move the centroids



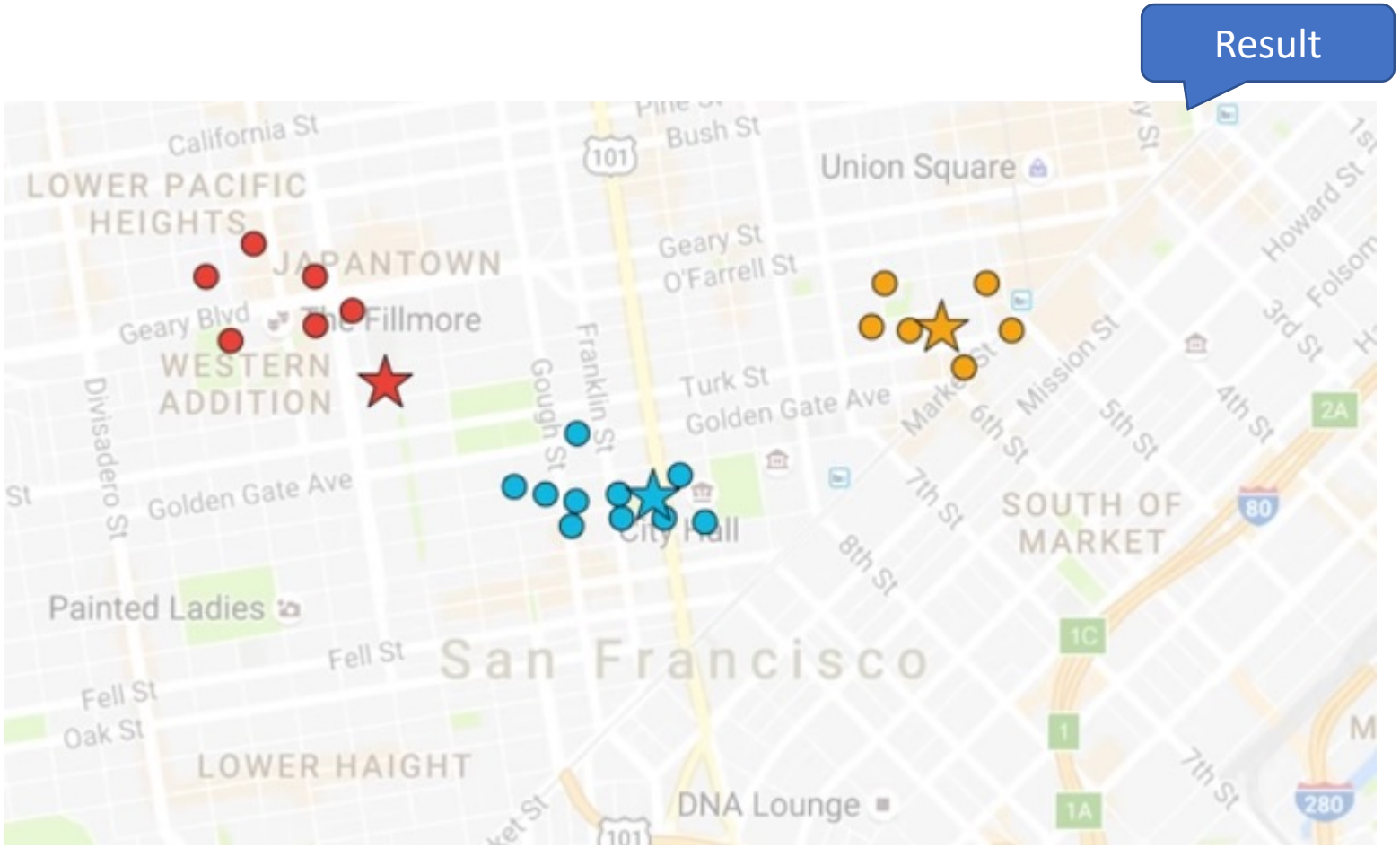
"Clustering: K-means and Hierarchical", Luis Serrano, https://www.youtube.com/watch?v=QXOkPvFM6NU

Find the clusters



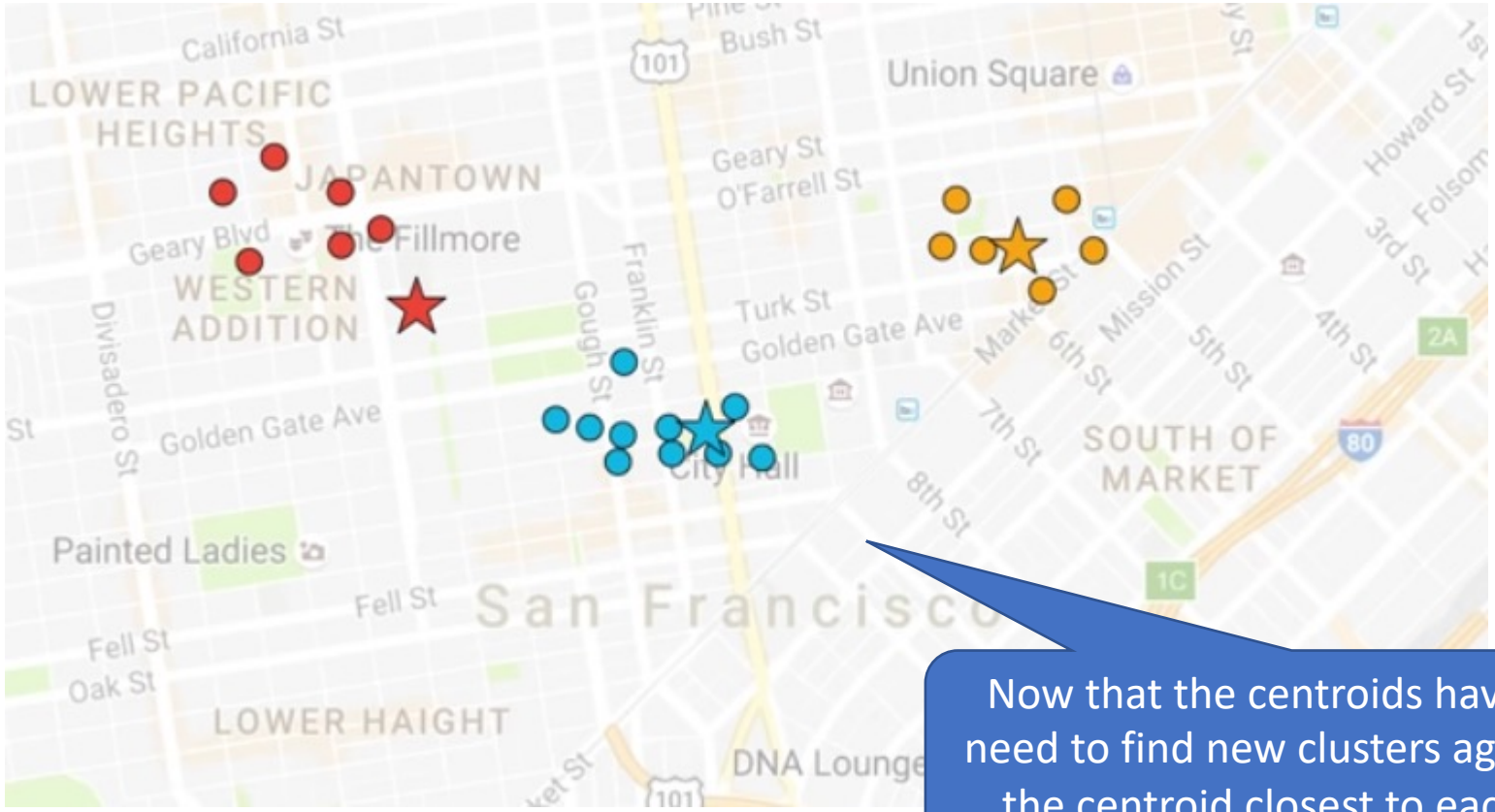
"Clustering: K-means and Hierarchical", Luis Serrano, https://www.youtube.com/watch?v=QXOkPvFM6NU

Find the clusters



"Clustering: K-means and Hierarchical", Luis Serrano, https://www.youtube.com/watch?v=QXOkPvFM6NU

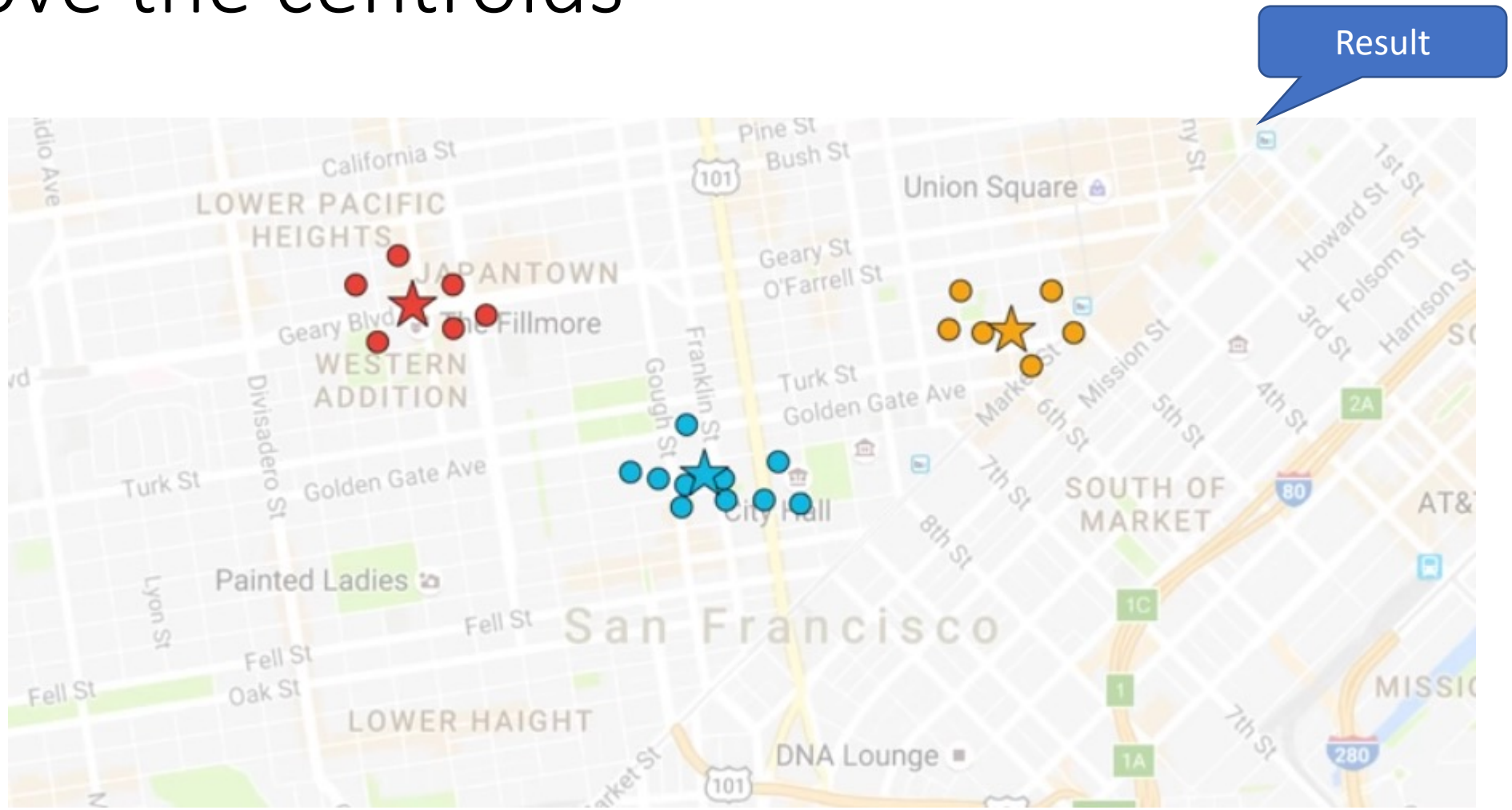
Move the centroids



Now that the centroids have moved, we need to find new clusters again (by finding the centroid closest to each customer)

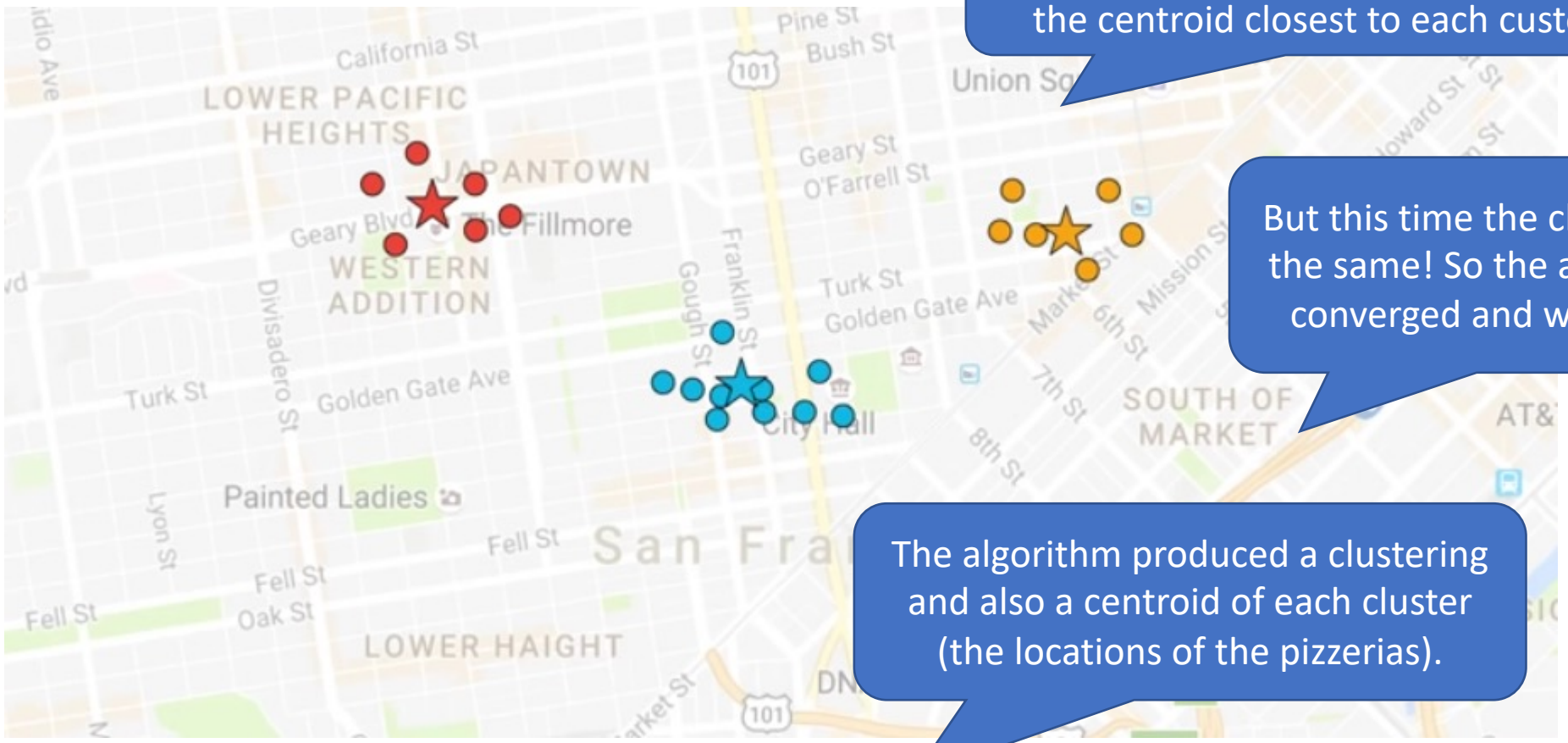
"Clustering: K-means and Hierarchical", Luis Serrano, https://www.youtube.com/watch?v=QXOkPvFM6NU

Move the centroids



"Clustering: K-means and Hierarchical", Luis Serrano, https://www.youtube.com/watch?v=QXOkPvFM6NU

Find the clusters

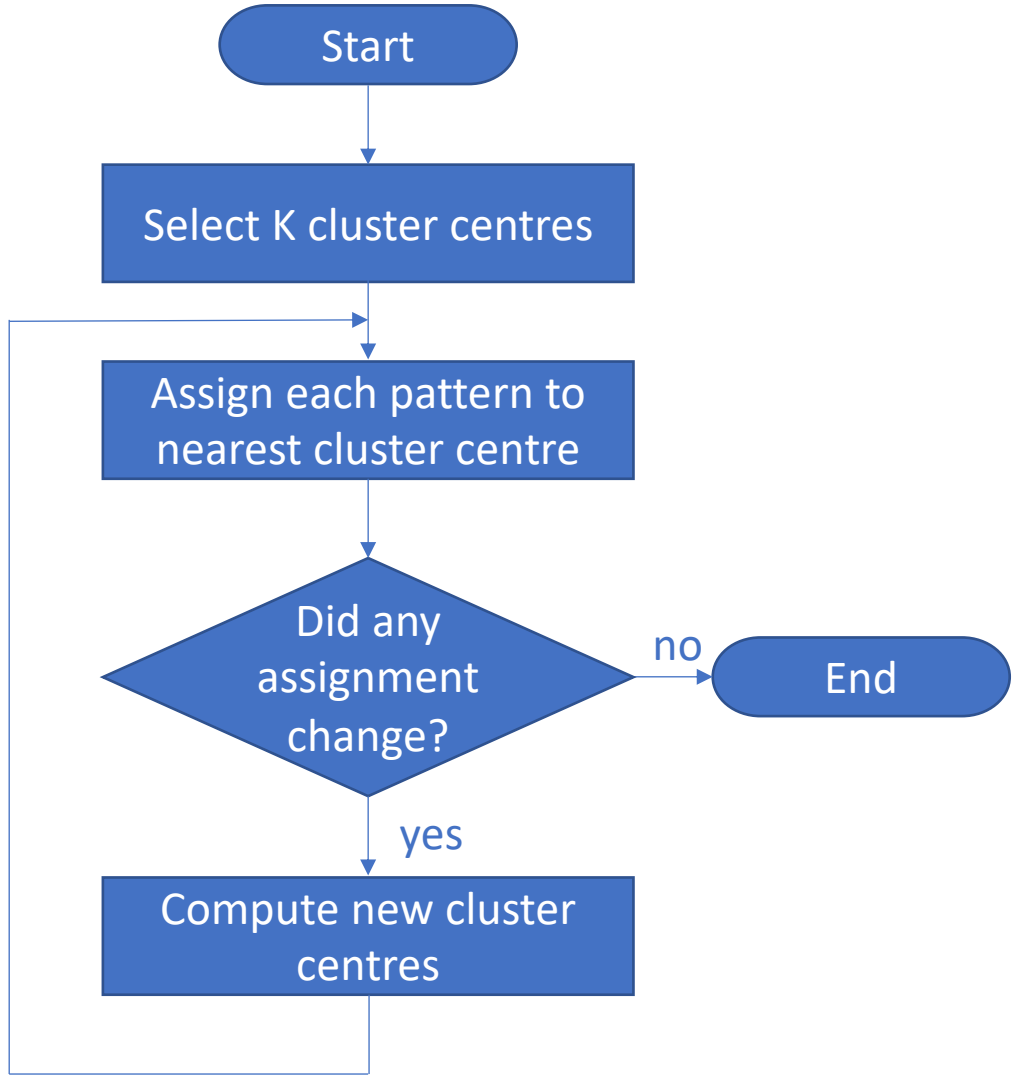


"Clustering: K-means and Hierarchical", Luis Serrano, https://www.youtube.com/watch?v=QXOkPvFM6NU

K-means

- choose value of K
- choose initial positions of the K cluster centres
- choose distance metric

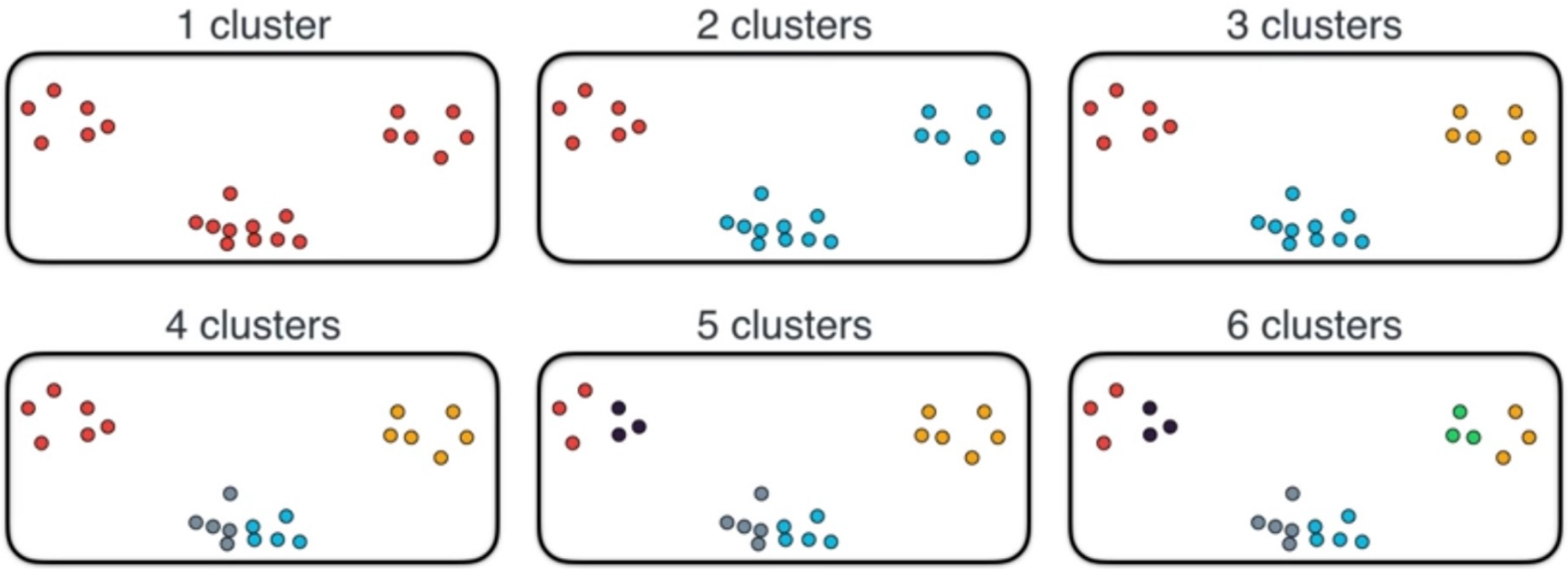
- a greedy algorithm
- converges to a local minimum



The elbow method

The elbow method

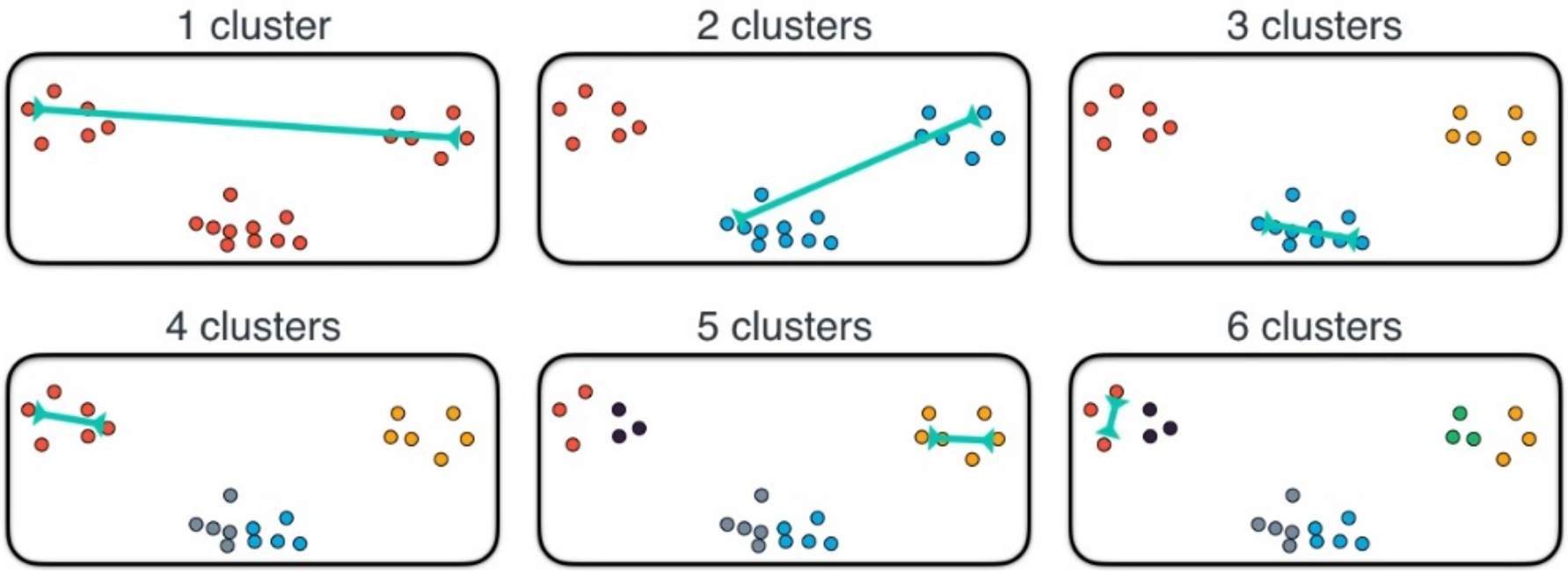
How do we pick K (the number of clusters)?



"Clustering: K-means and Hierarchical", Luis Serrano, https://www.youtube.com/watch?v=QXOkPvFM6NU

The elbow method

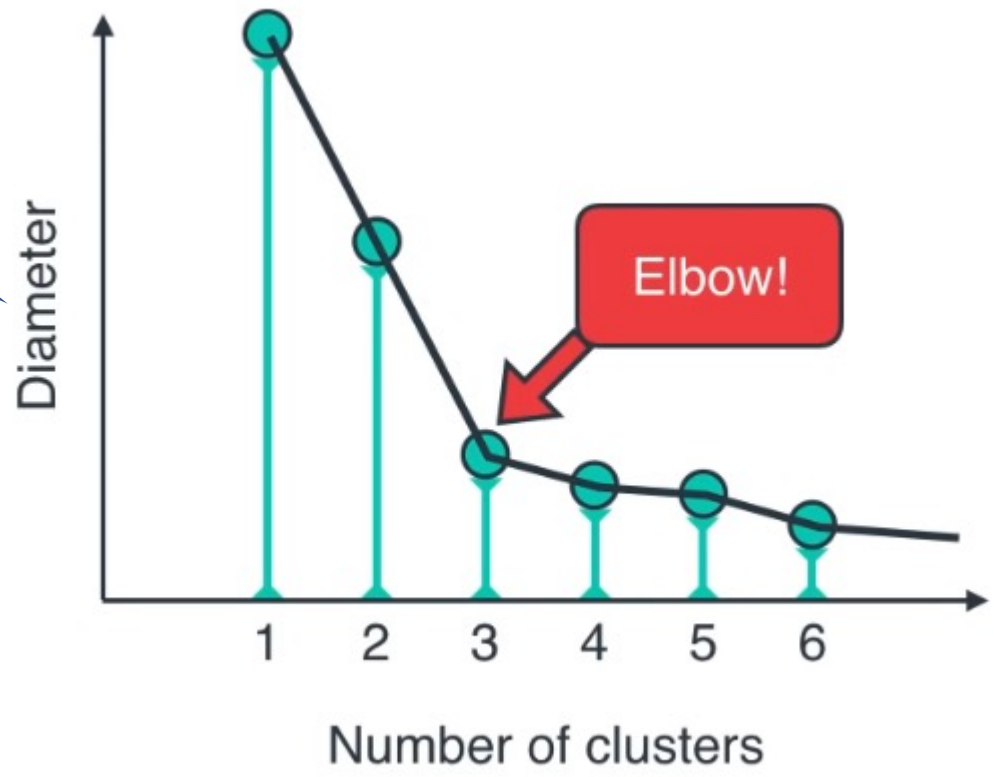
Let the *diameter* of a clustering be the longest intra-cluster distance (longest distance between two points of the same colour)



"Clustering: K-means and Hierarchical", Luis Serrano, https://www.youtube.com/watch?v=QXOkPvFM6NU

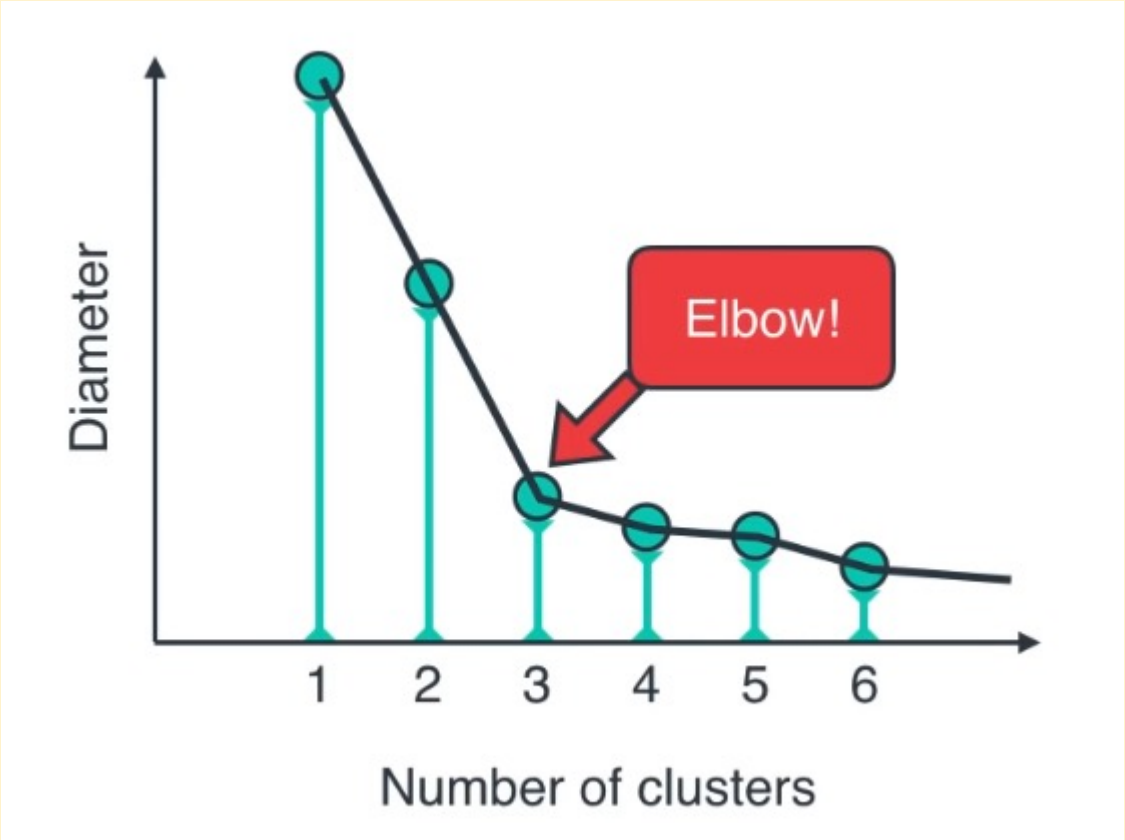
The Elbow method

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



"Clustering: K-means and Hierarchical", Luis Serrano, https://www.youtube.com/watch?v=QXOkPvFM6NU

The elbow method

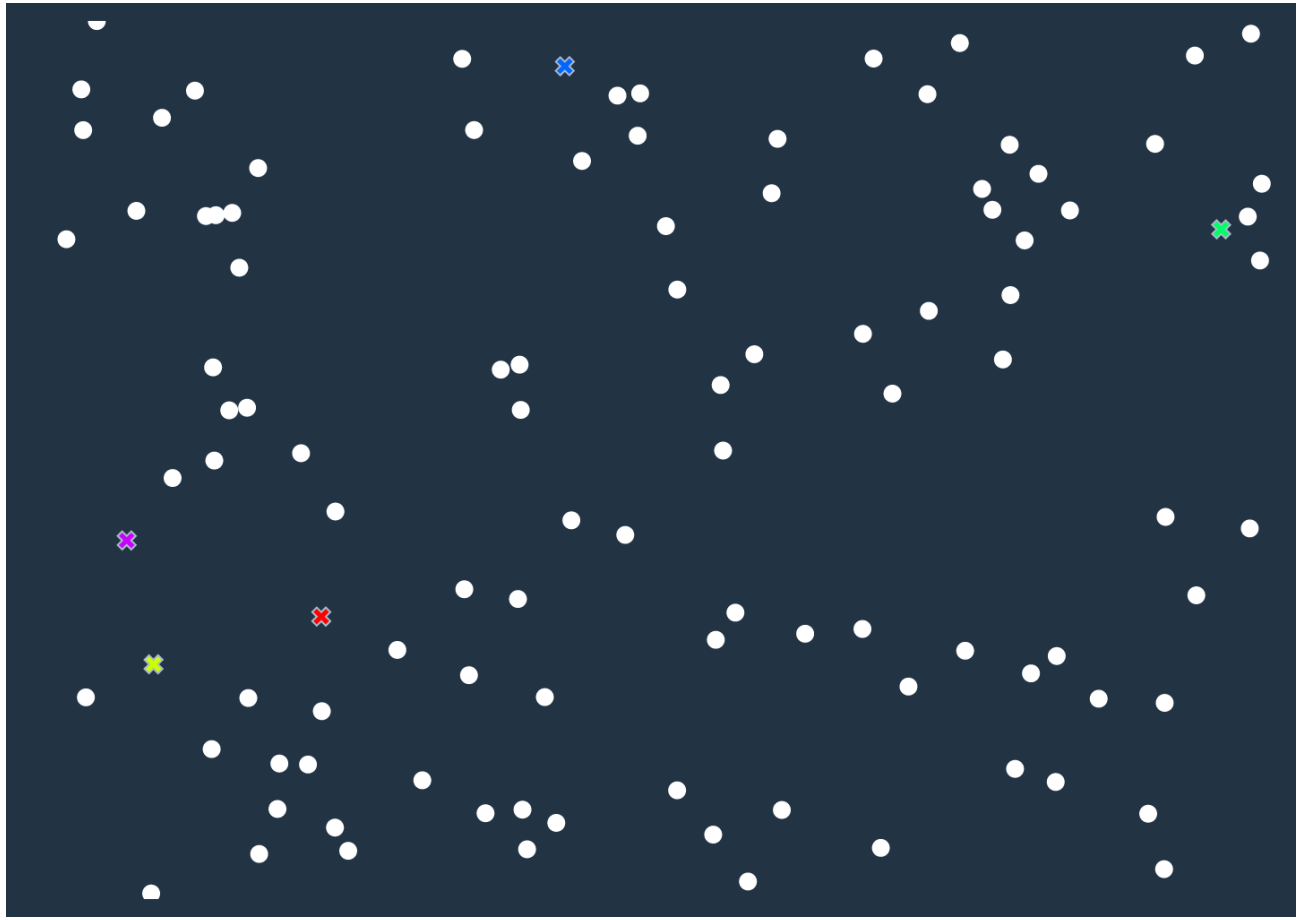


We can see where the elbow is, but how could a program find the elbow?

"Clustering: K-means and Hierarchical", Luis Serrano, https://www.youtube.com/watch?v=QXOkPvFM6NU

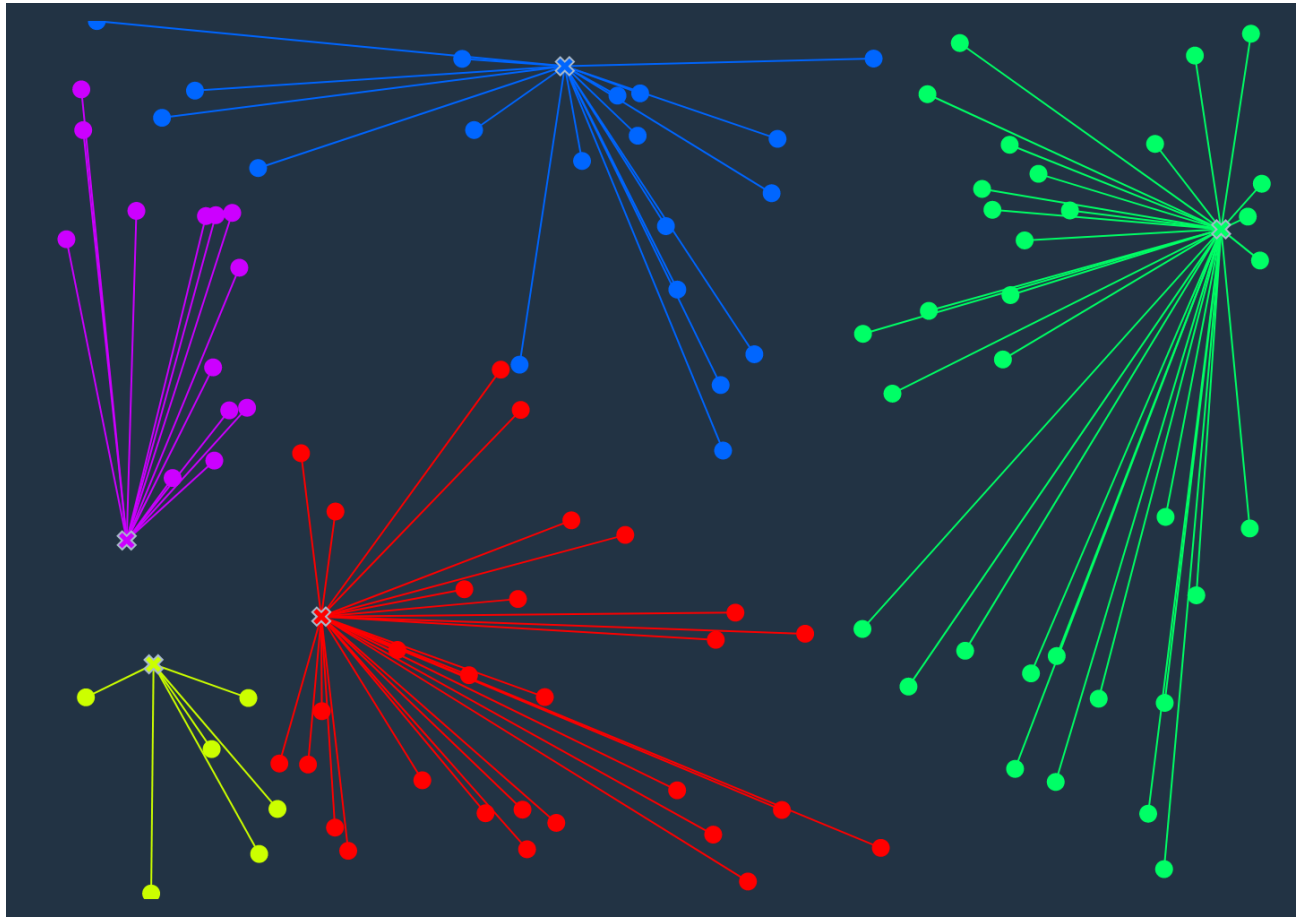
Limitations and problems of k-means clustering

Select initial cluster “centres”



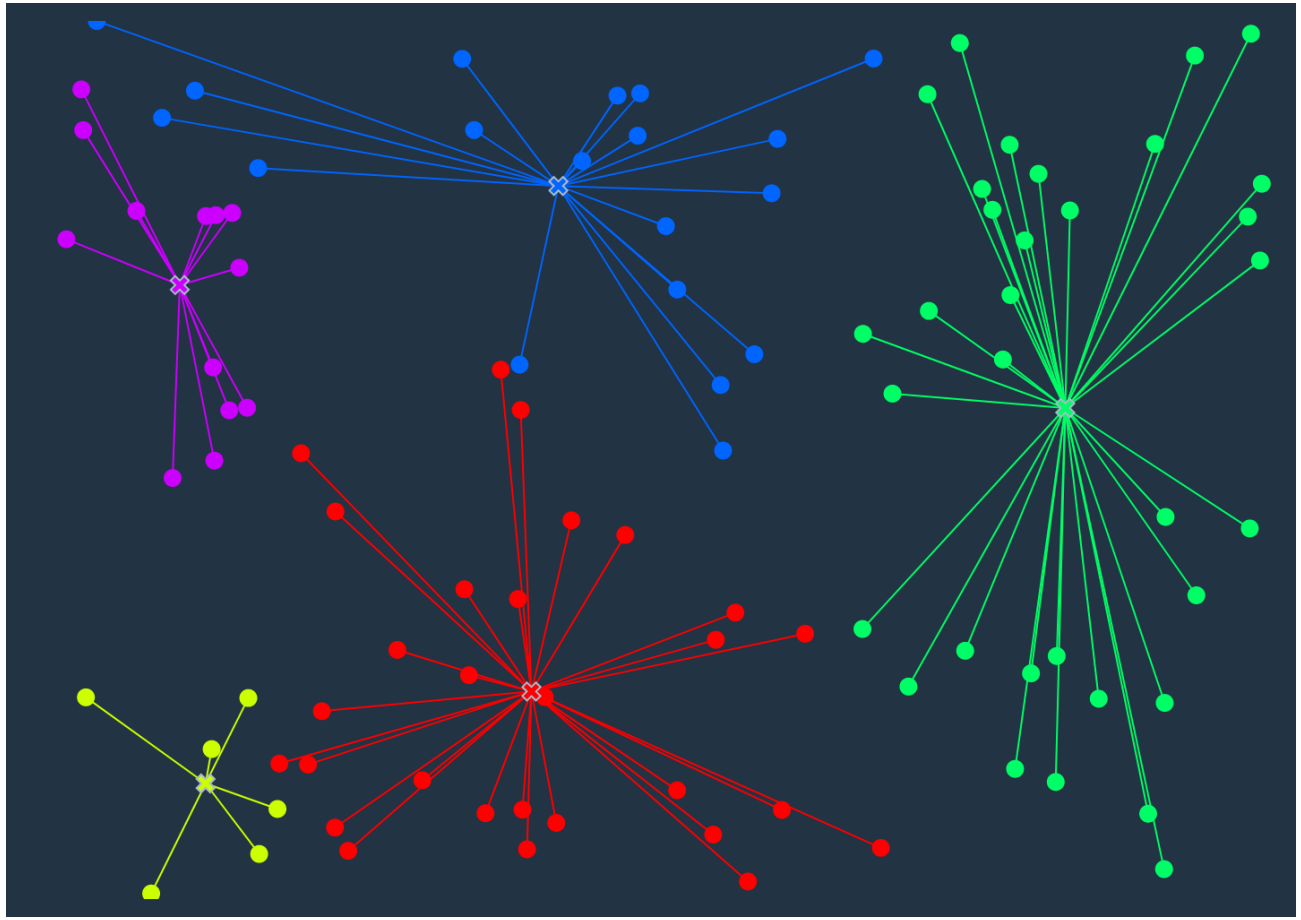
http://tech.nitoyon.com/en/blog/2013/11/07/k-means/

Assign patterns to nearest cluster centre (1)



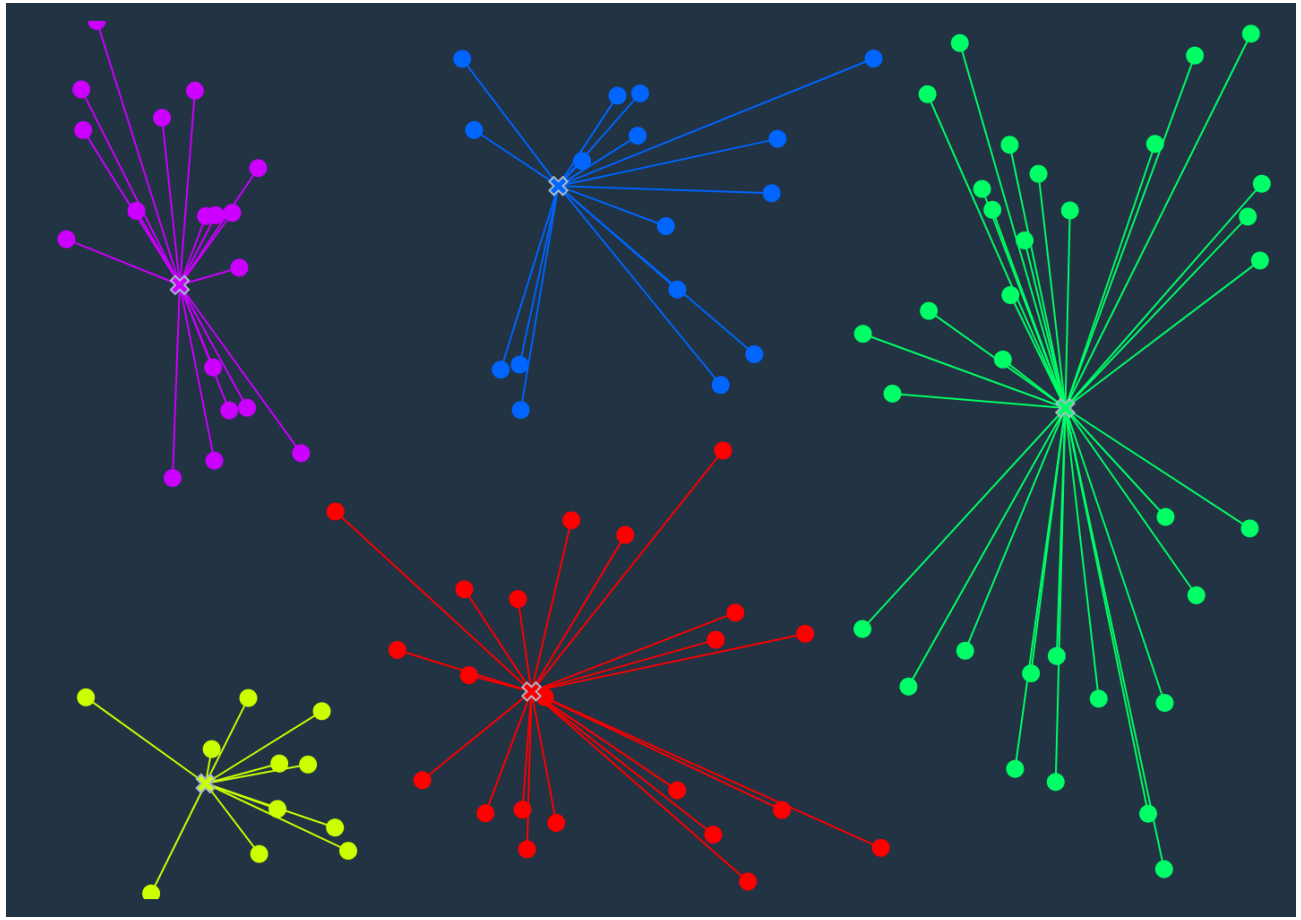
http://tech.nitoyon.com/en/blog/2013/11/07/k-means/

Compute new cluster centres (1)



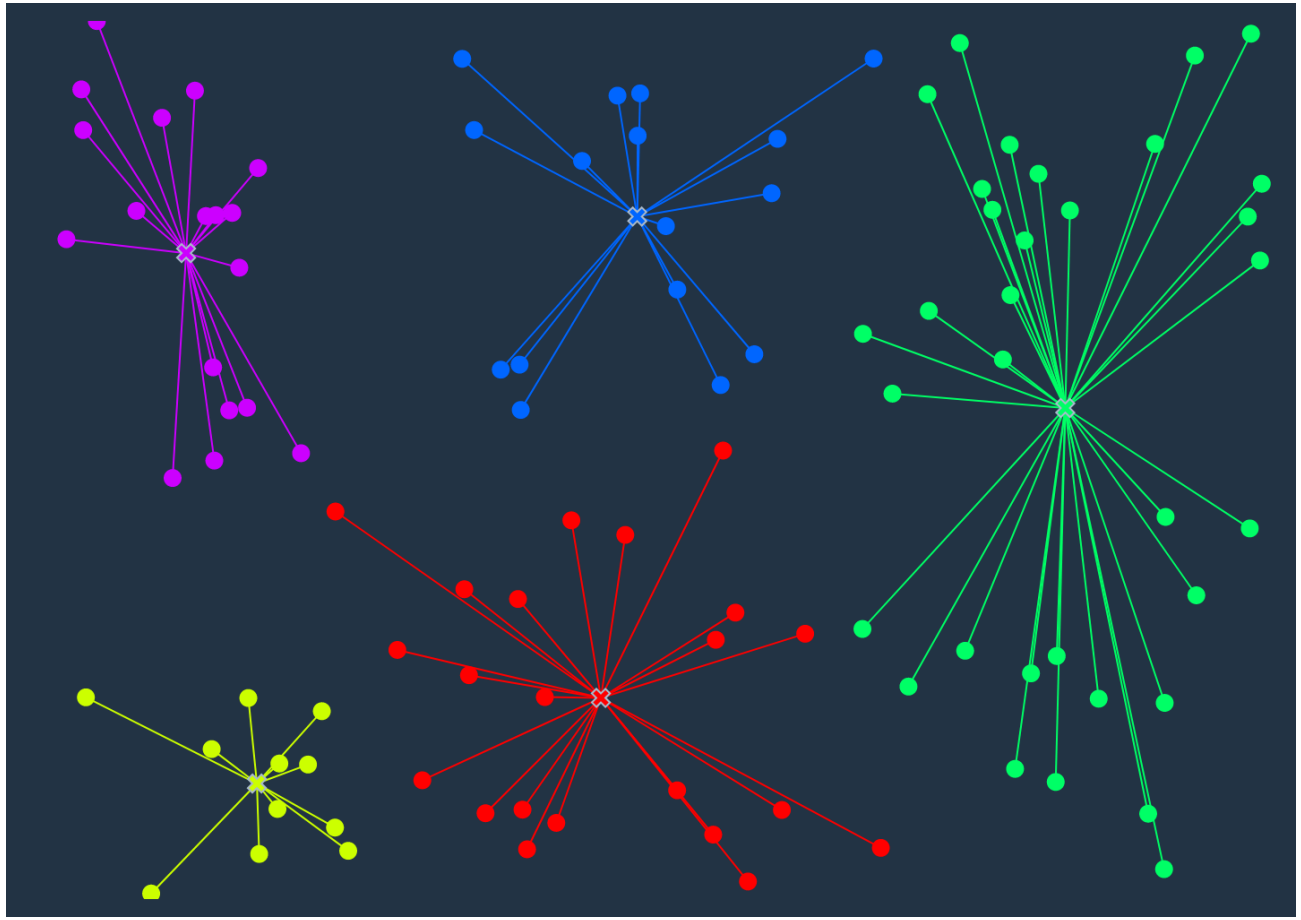
http://tech.nitoyon.com/en/blog/2013/11/07/k-means/

Assign patterns to nearest cluster centre (2)



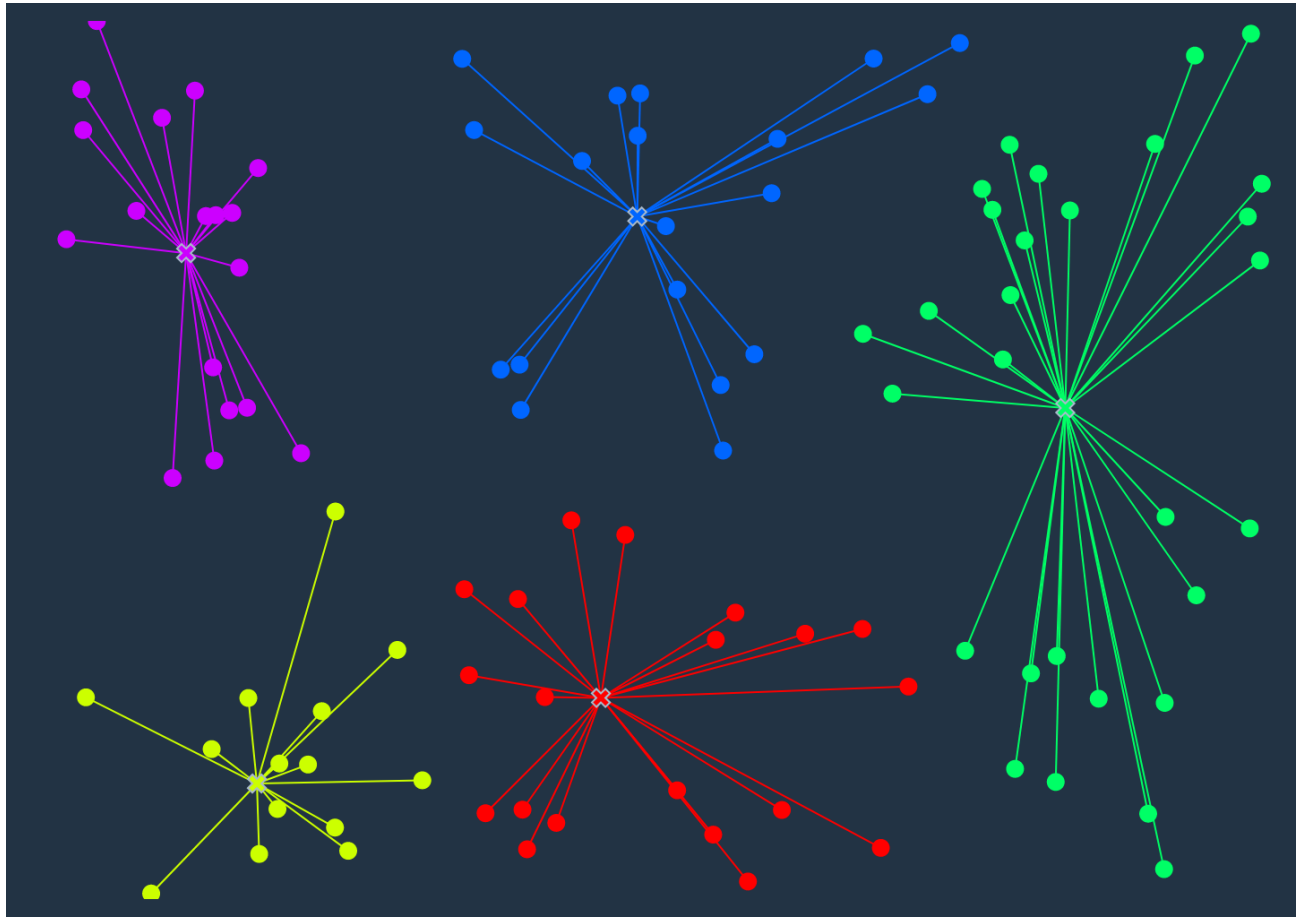
http://tech.nitoyon.com/en/blog/2013/11/07/k-means/

Compute new cluster centres (2)



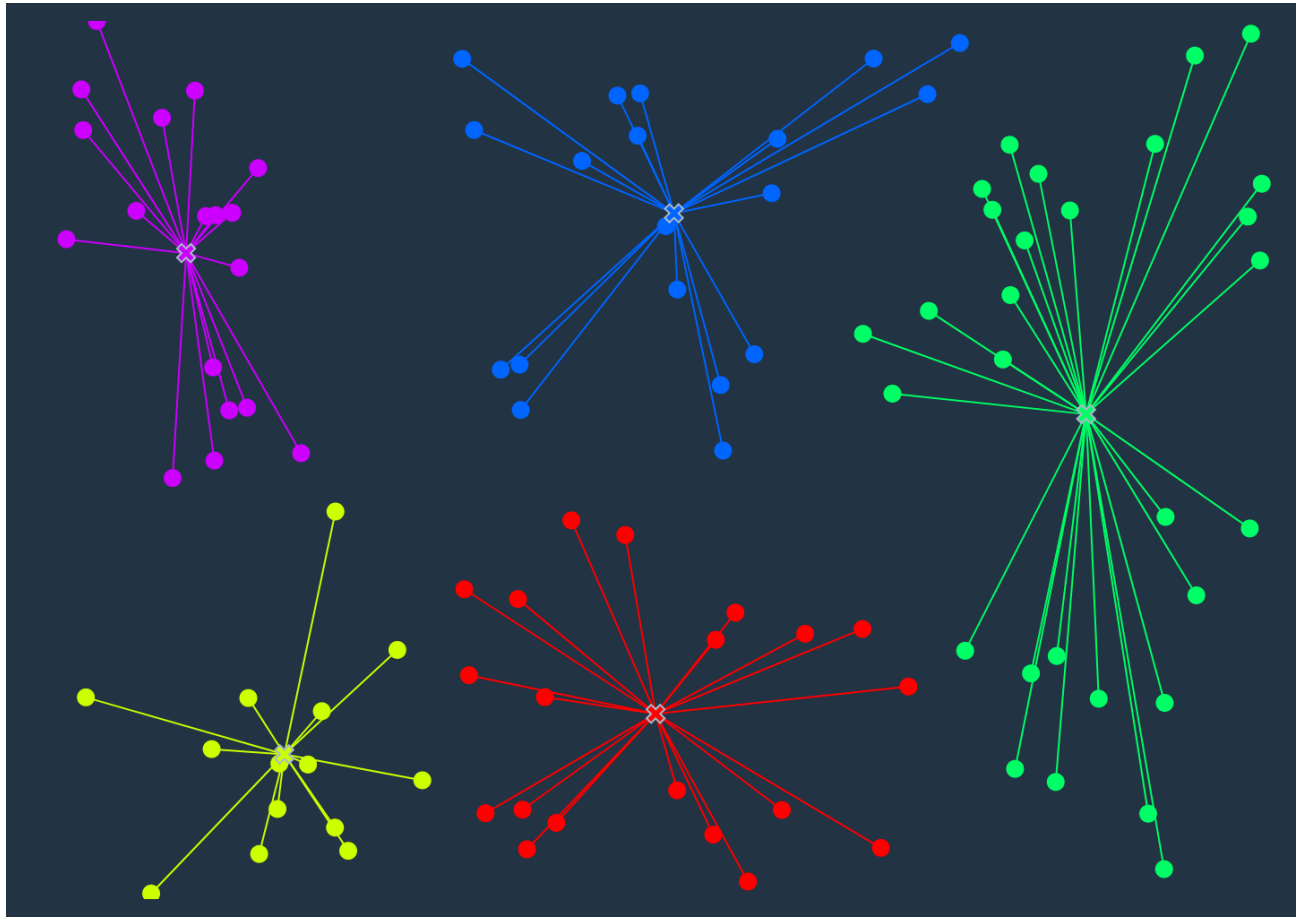
http://tech.nitoyon.com/en/blog/2013/11/07/k-means/

Assign patterns to nearest cluster centre (3)



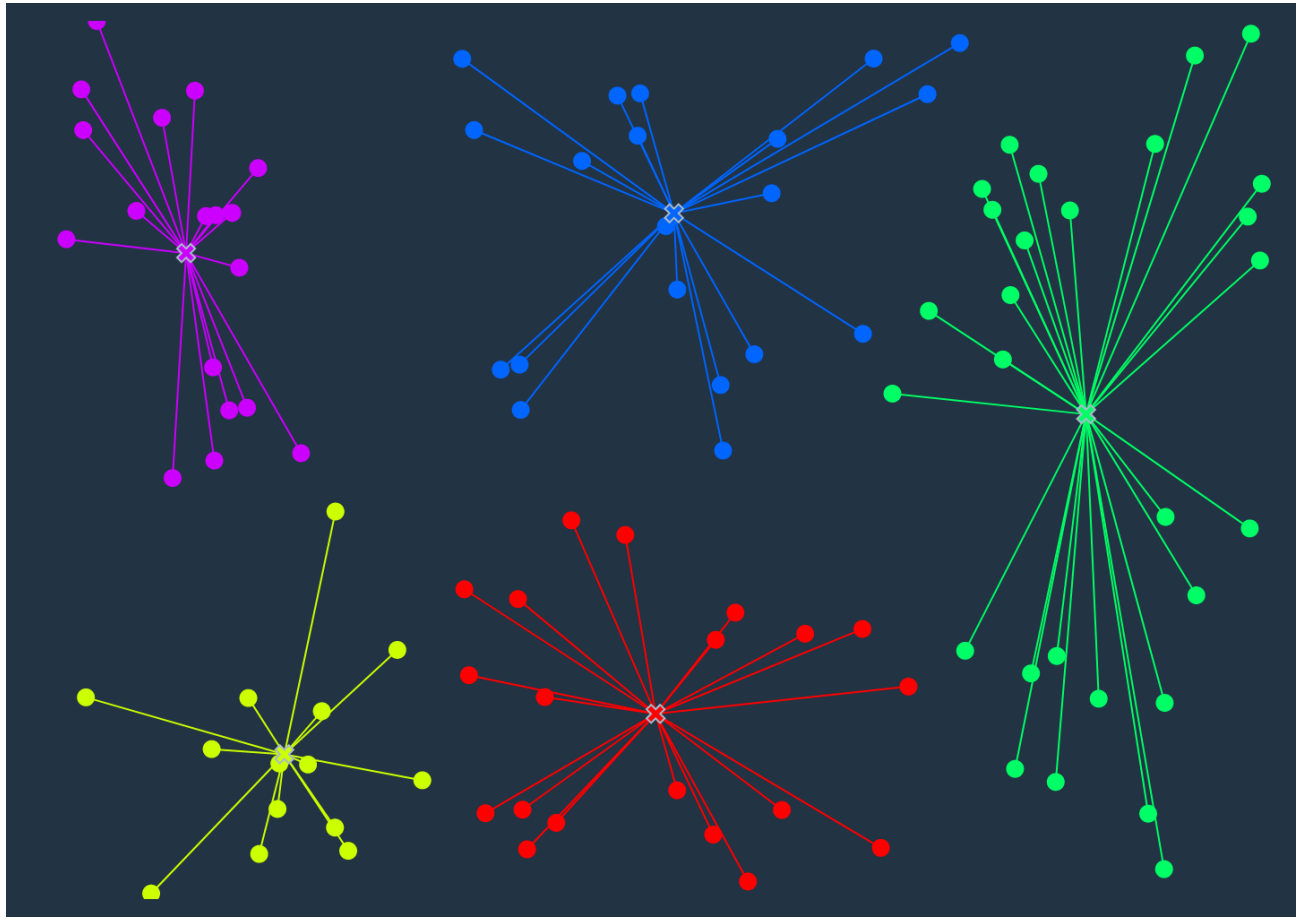
http://tech.nitoyon.com/en/blog/2013/11/07/k-means/

Compute new cluster centres (3)



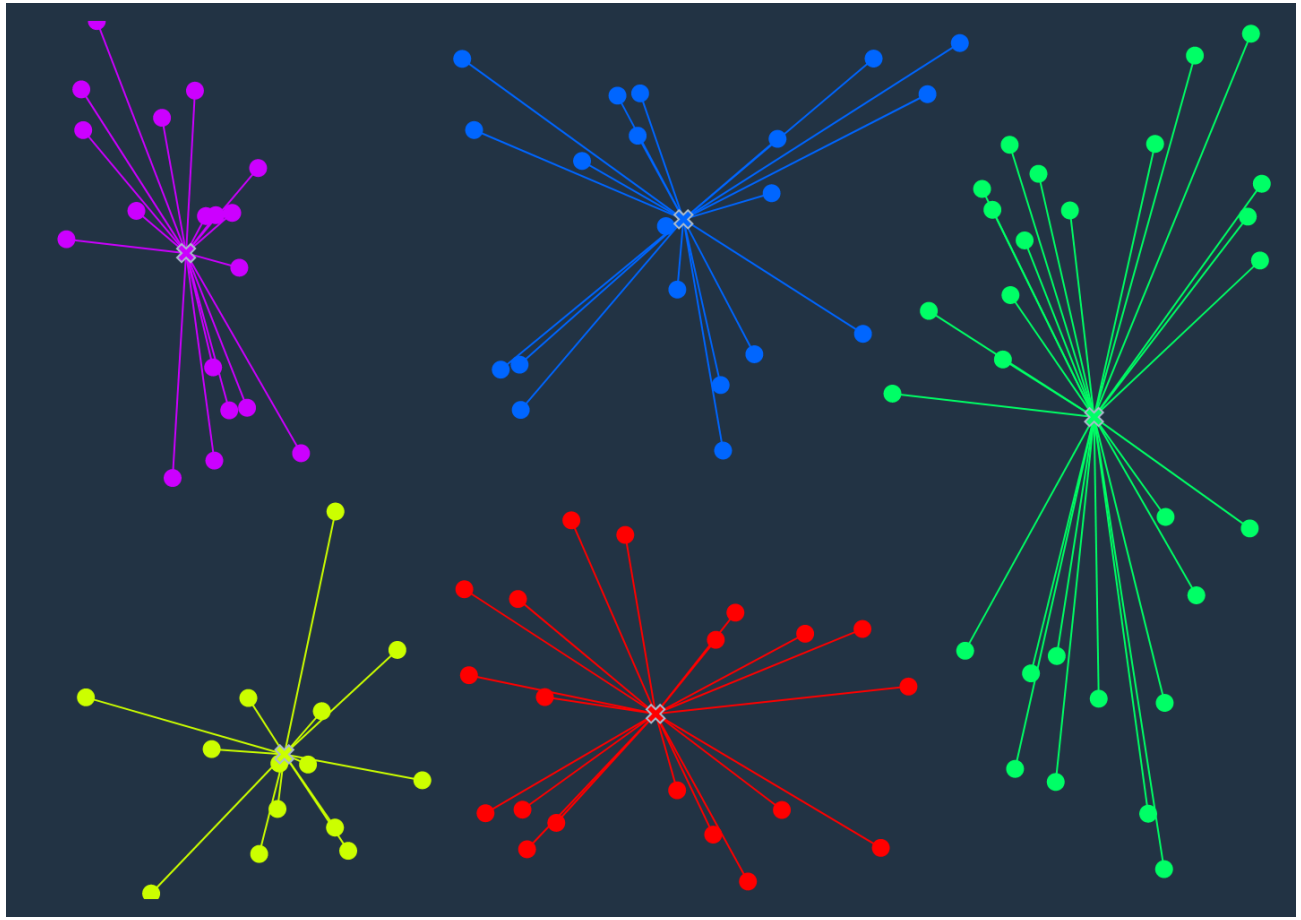
http://tech.nitoyon.com/en/blog/2013/11/07/k-means/

Assign patterns to nearest cluster centre (4)



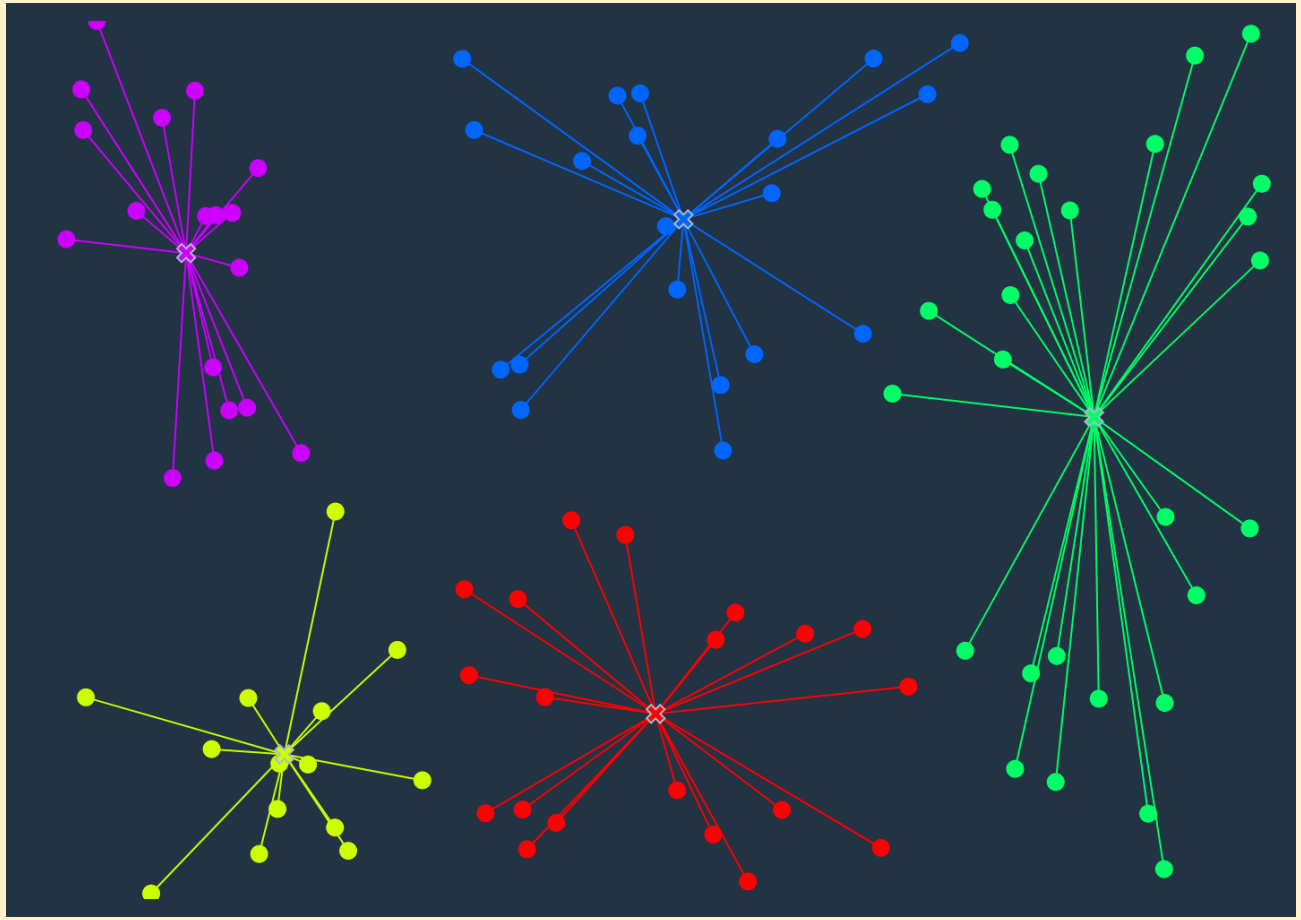
http://tech.nitoyon.com/en/blog/2013/11/07/k-means/

Compute new cluster centres (4)



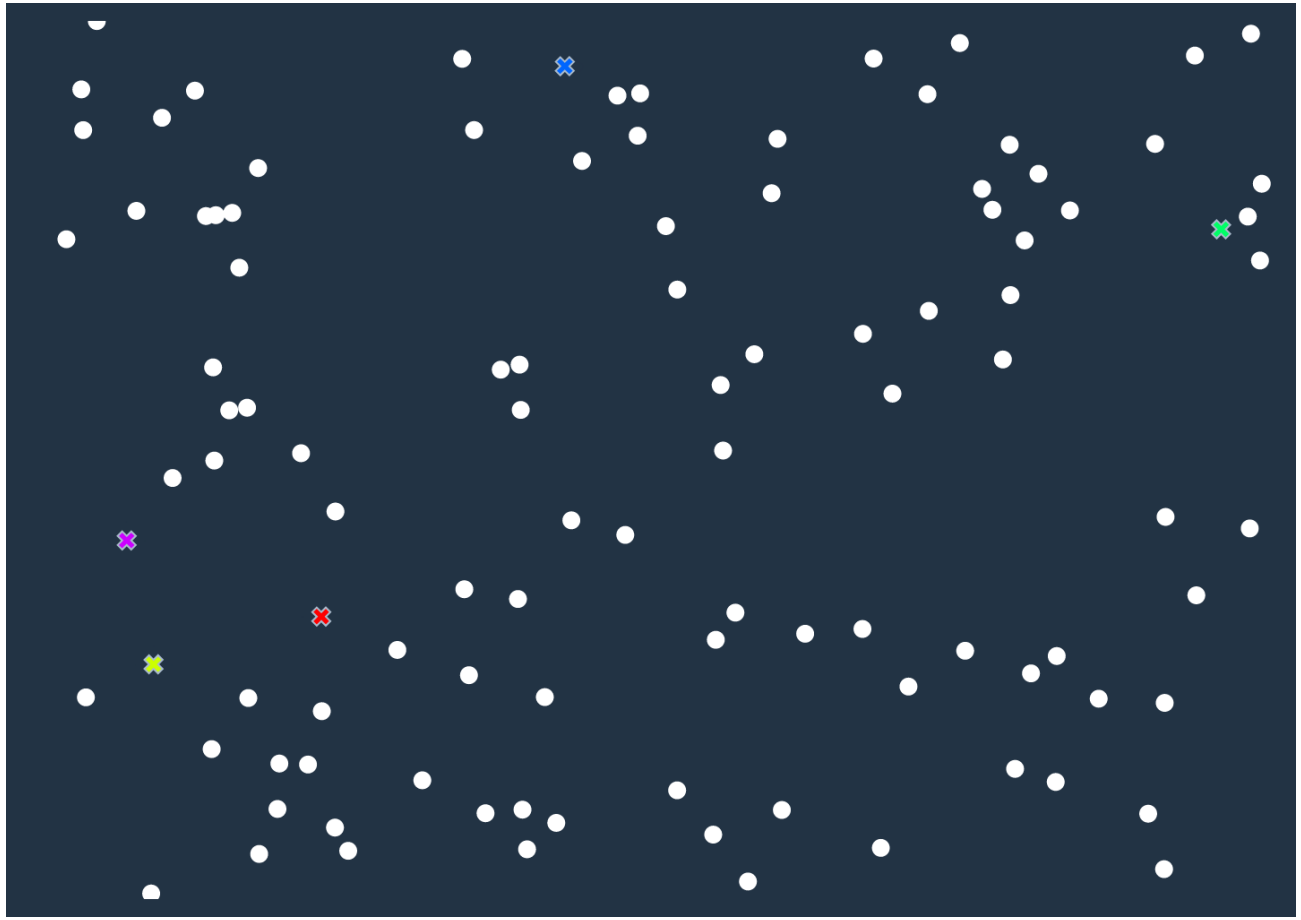
http://tech.nitoyon.com/en/blog/2013/11/07/k-means/

Is anything “wrong” with this?



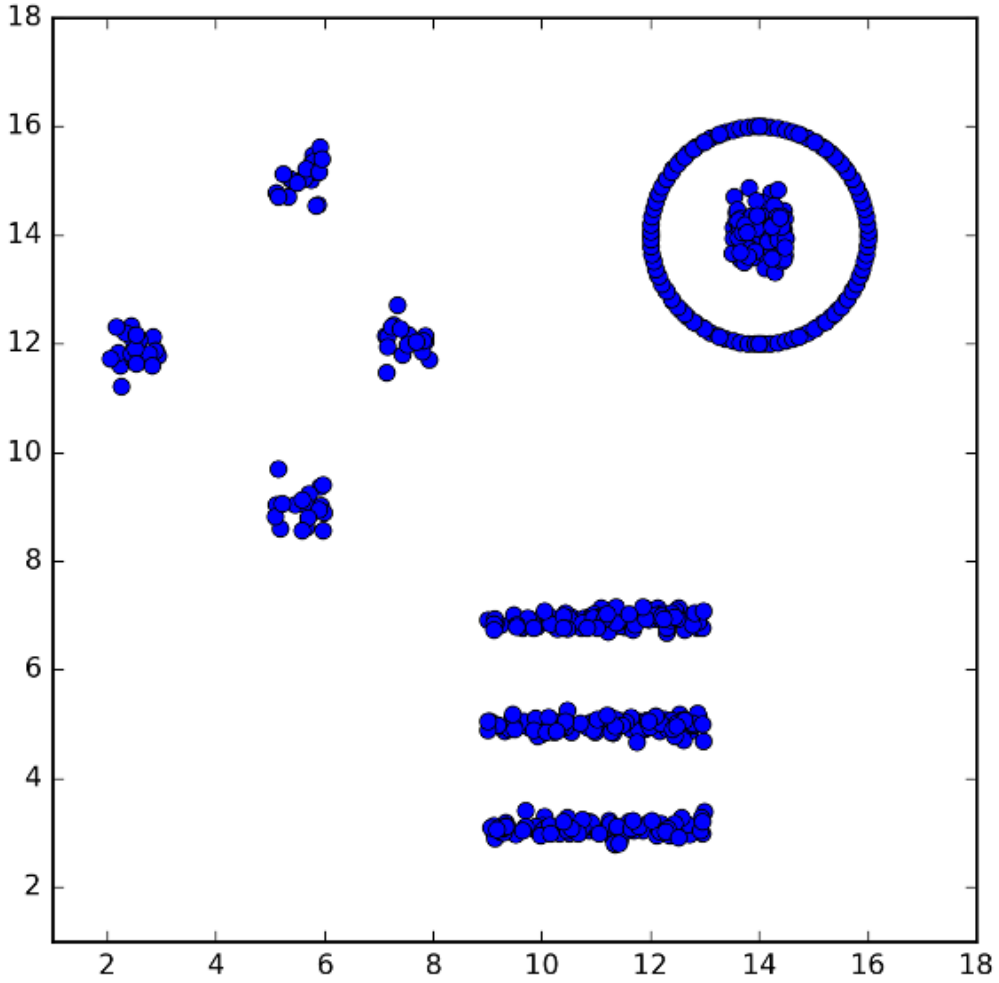
http://tech.nitoyon.com/en/blog/2013/11/07/k-means/

Look at the initial data



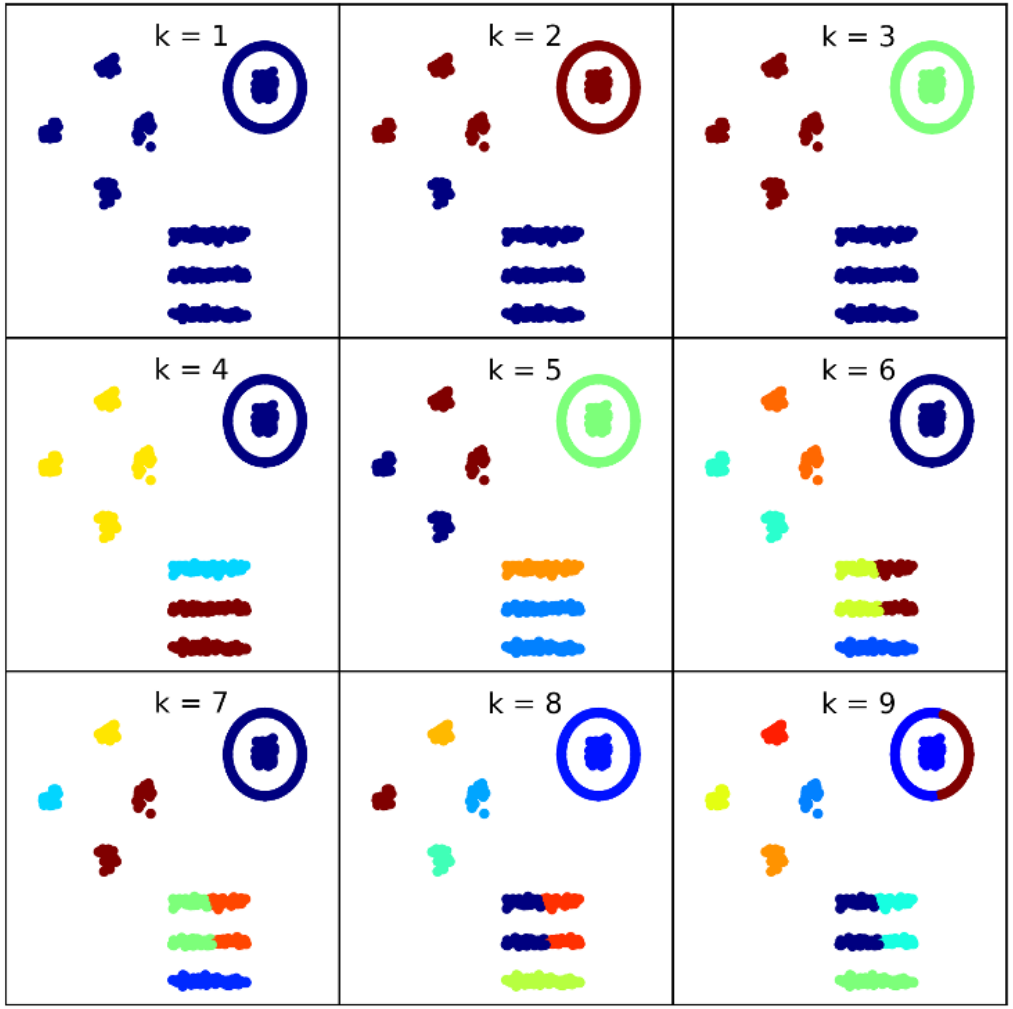
http://tech.nitoyon.com/en/blog/2013/11/07/k-means/

What will K-means do?



Skiena (2017) The Data Science Design Manual, Springer

What will K-means do?



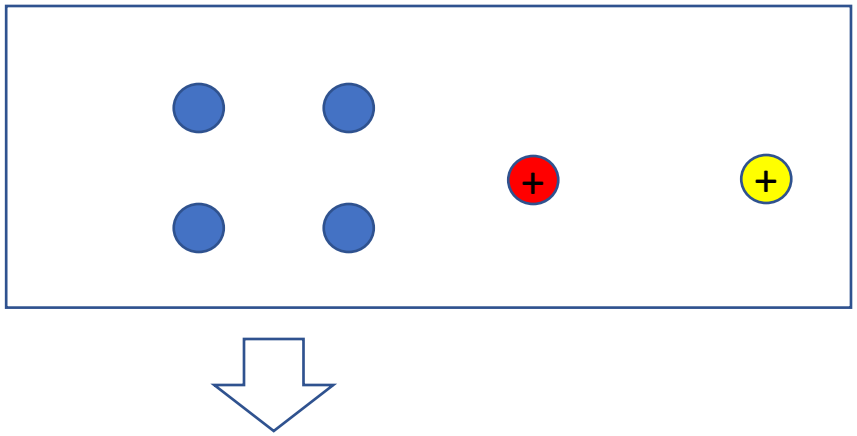
Skiena (2017) The Data Science Design Manual, Springer

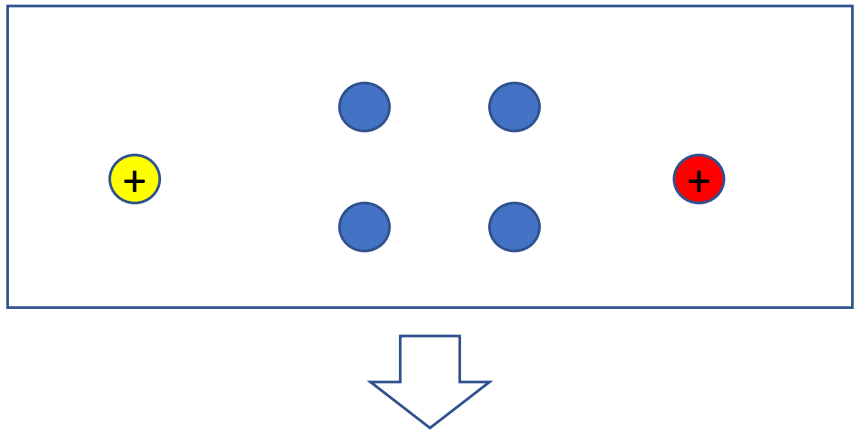
Caveats

K-means works best for

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

The result depends on the initialisation

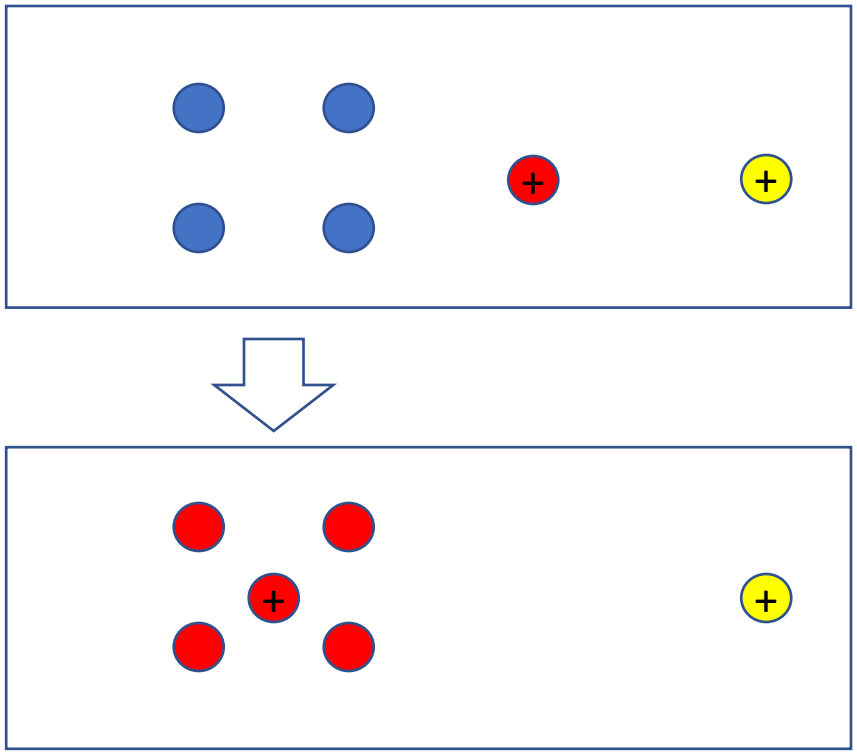


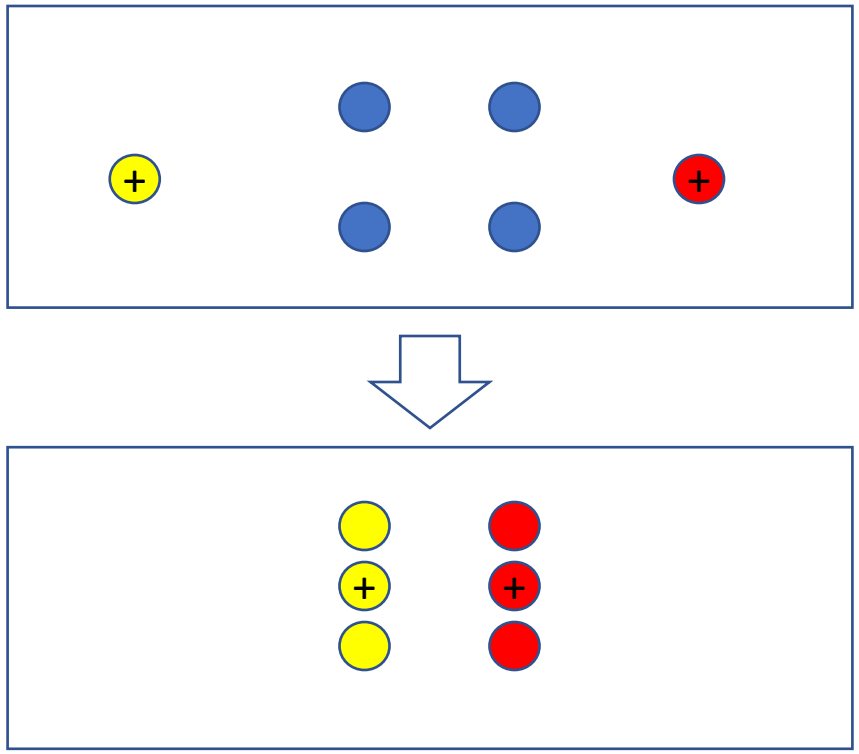


We are going to create two clusters, based on the initial centres that are shown by + and +

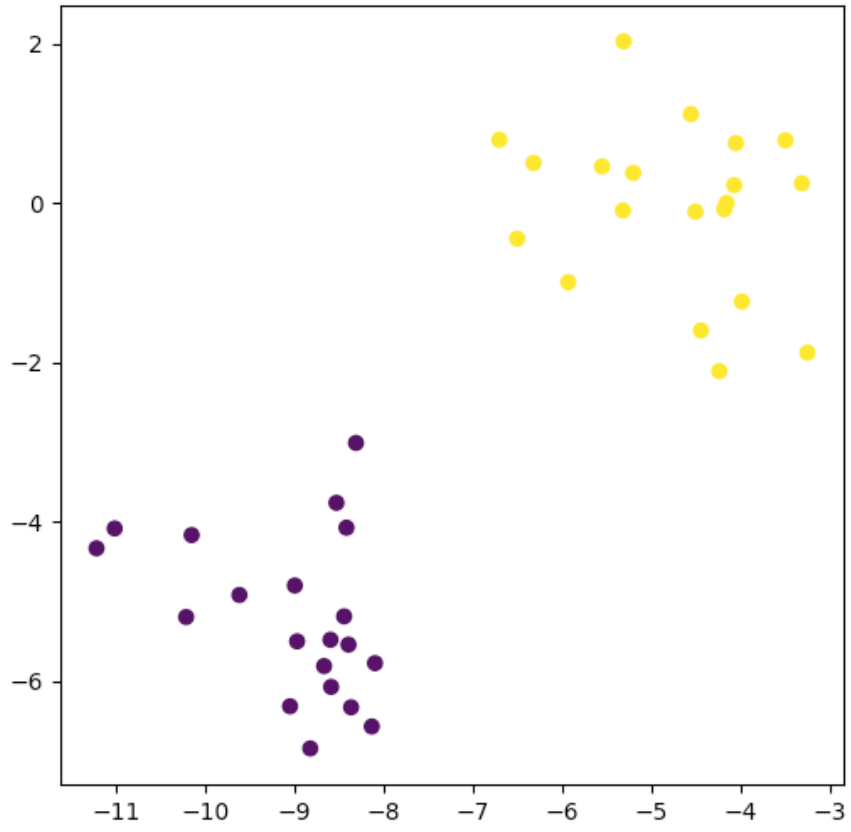
How will the blue points be assigned to the clusters?

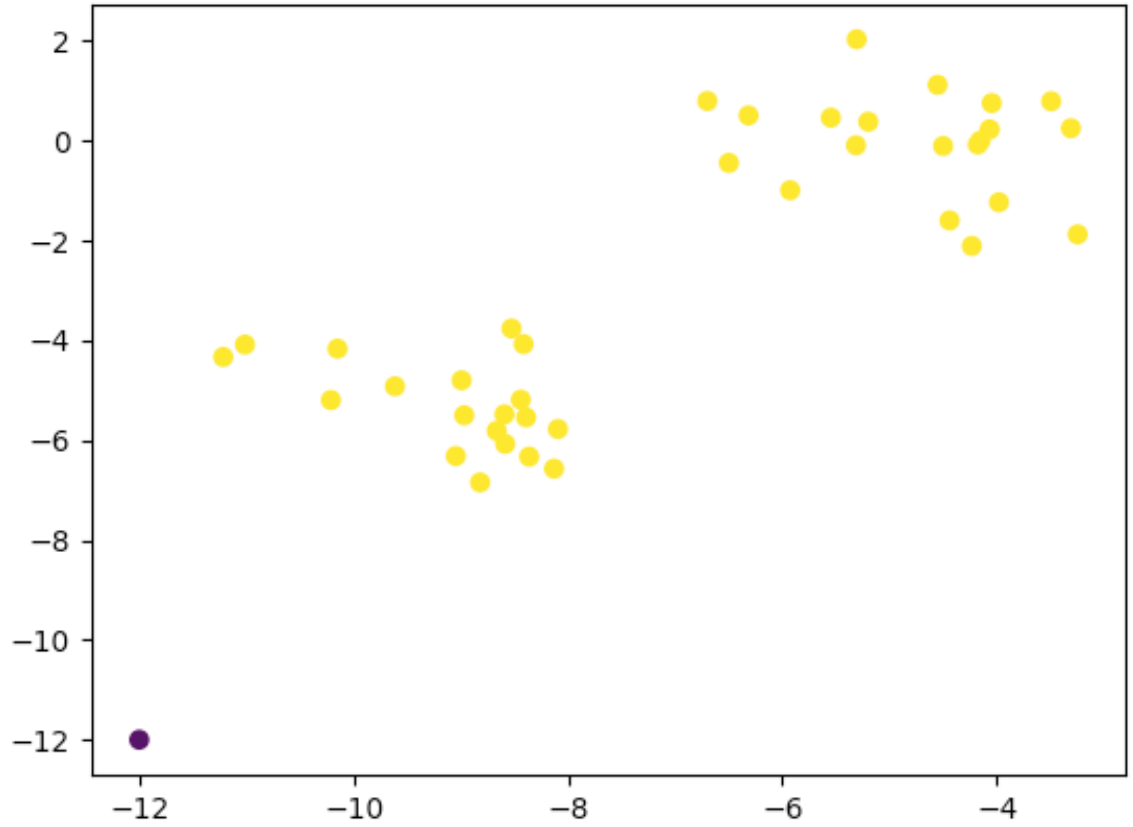
The result depends on the initialisation



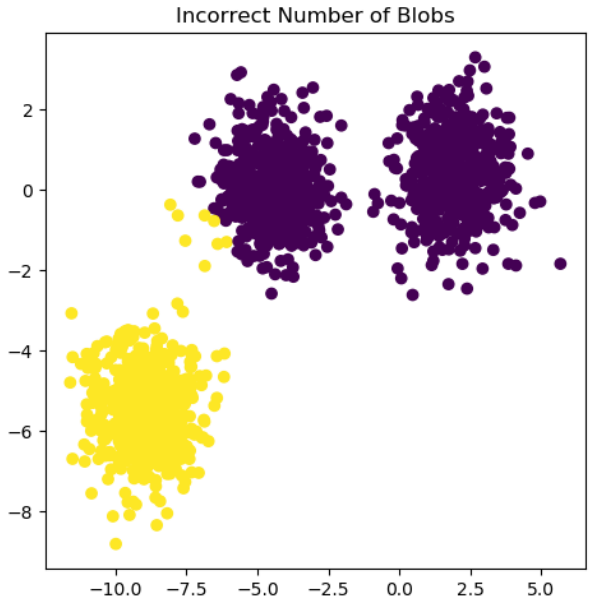


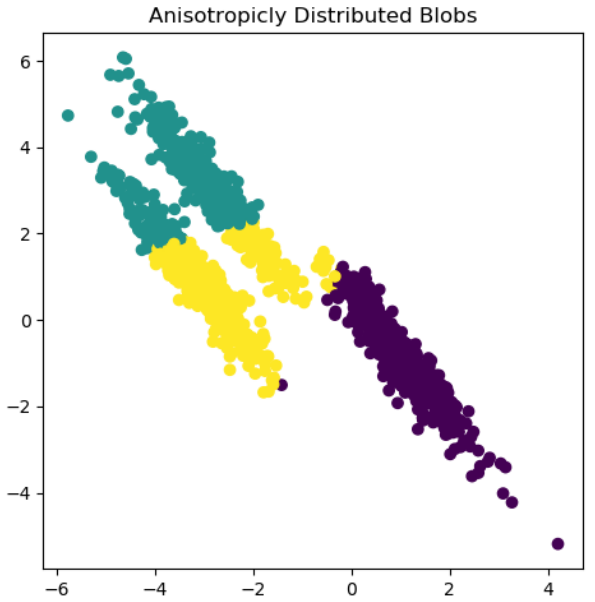
Moving one point, which is then selected as an initial centroid





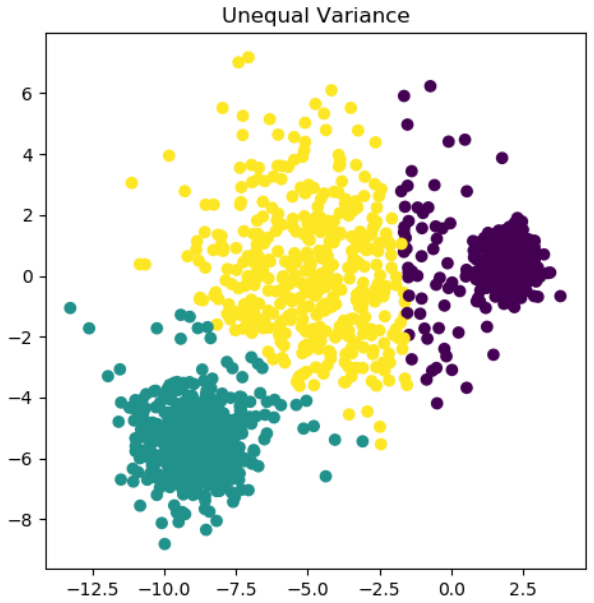
Clustering with k=2, when we “see” that there are 3 natural clusters

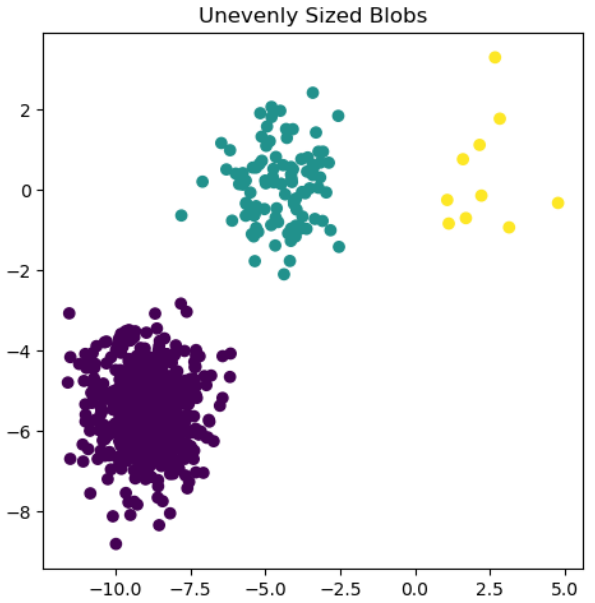




Different properties in different directions (result here is that blobs are not “spherical”)

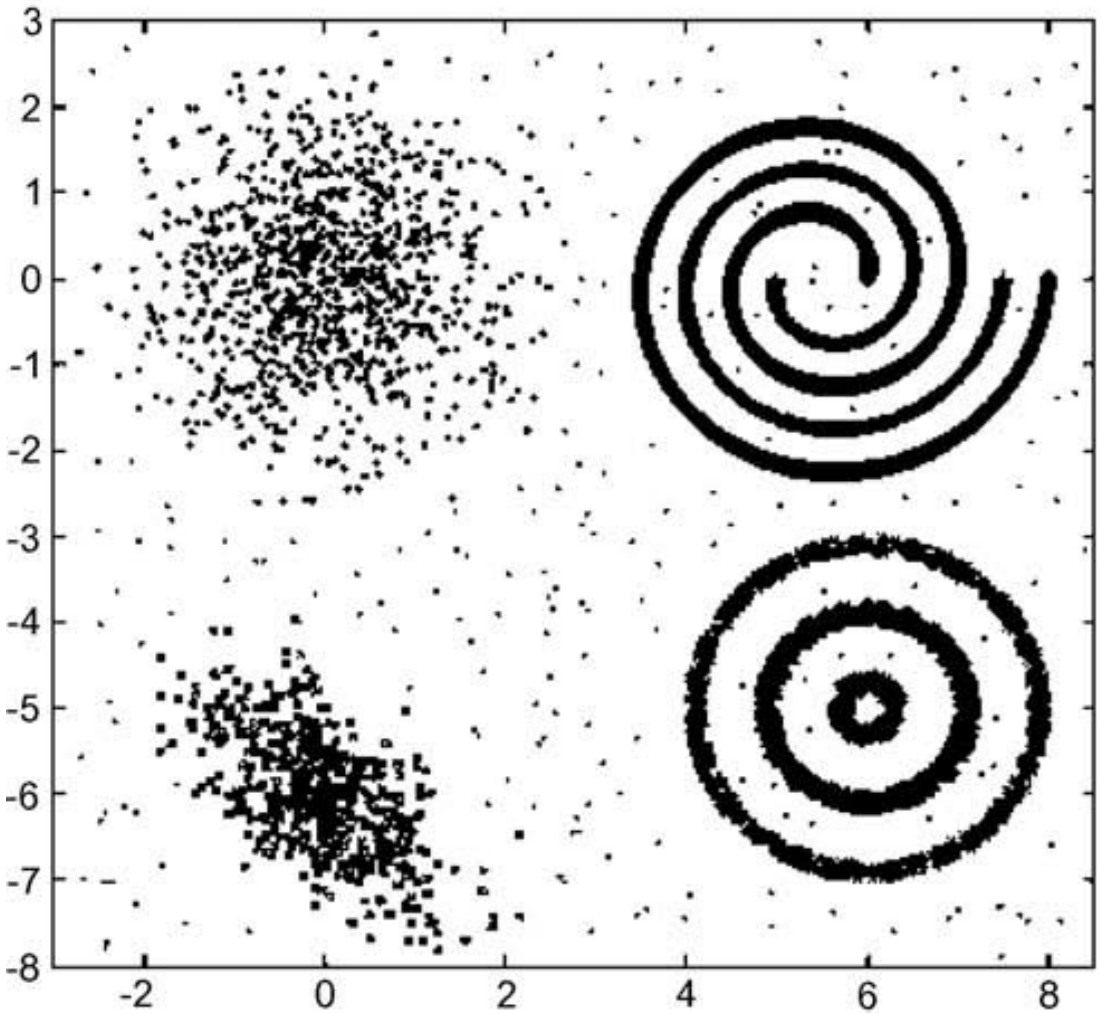
Two of the clusters have very small variance



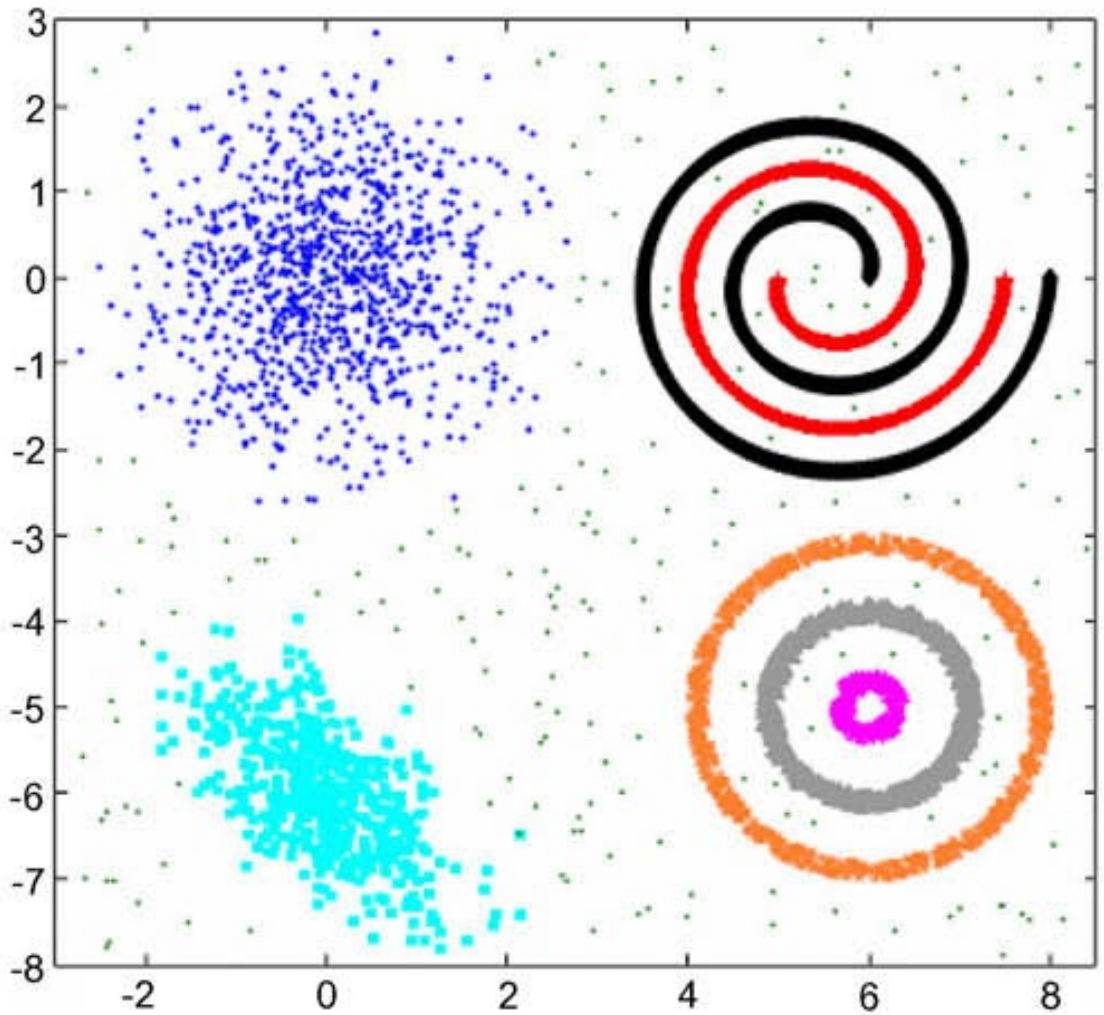


Intuitive result, even though clusters differ in size (number of cluster members)

https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_assumptions.html#sphx-glr-auto-examples-cluster-plot-kmeans-assumptions-py



(a) Input data



(b) Desired clustering

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

Next time

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