CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

# MODULE 3: CLUSTERING

DAT405, 2019-2020, READING PERIOD 1

# Confounding variables

- “extra” variables that you didn’t account for
- usually an error in data collection or measurement

- effect on dependent variable might not be entirely due to values of given independent variables – confounding variables can cause the effect of the given independent variable to be overestimated or underestimated
- take care when presenting relationship between independent and dependent variables since there might be confounding variables

https://www.statisticshowto.datasciencecentral.com/experimental-design/confounding-variable/

# Core data science tasks

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

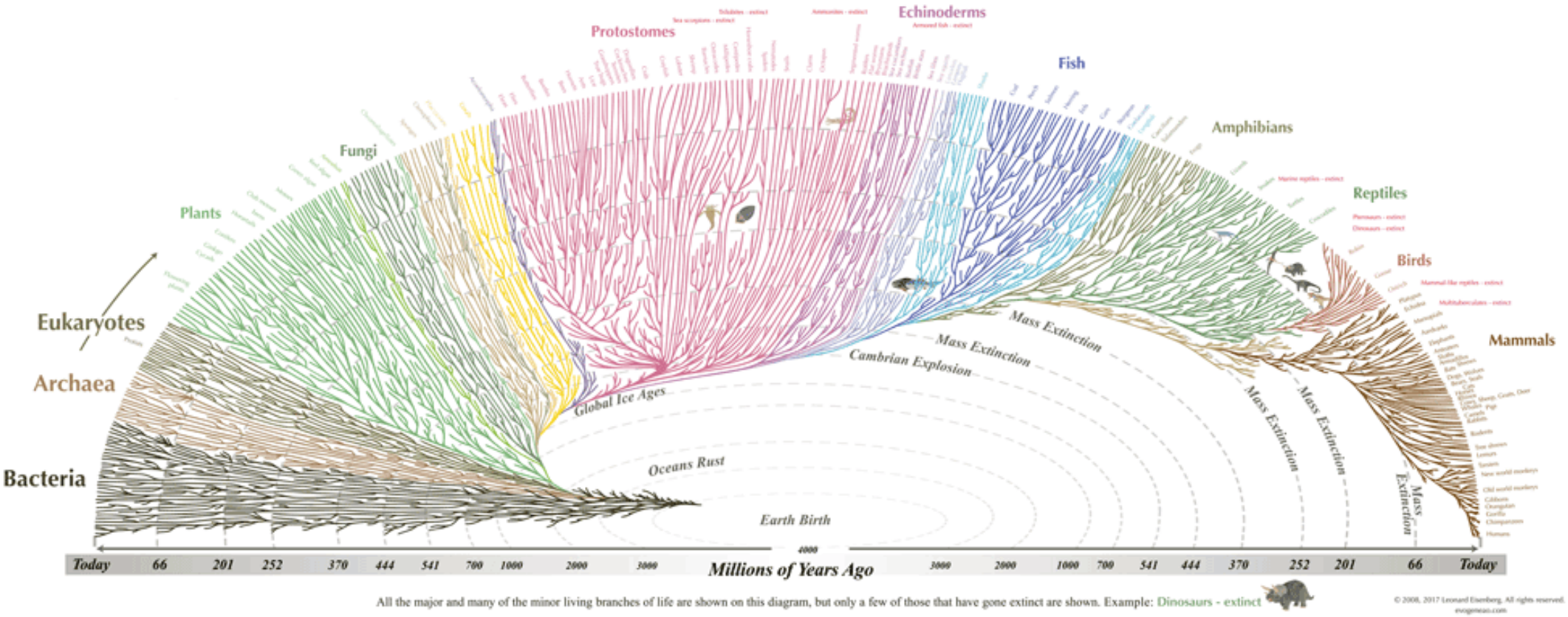
# CLUSTERING

- Grouping items by similarity

# Clustering: Use Cases

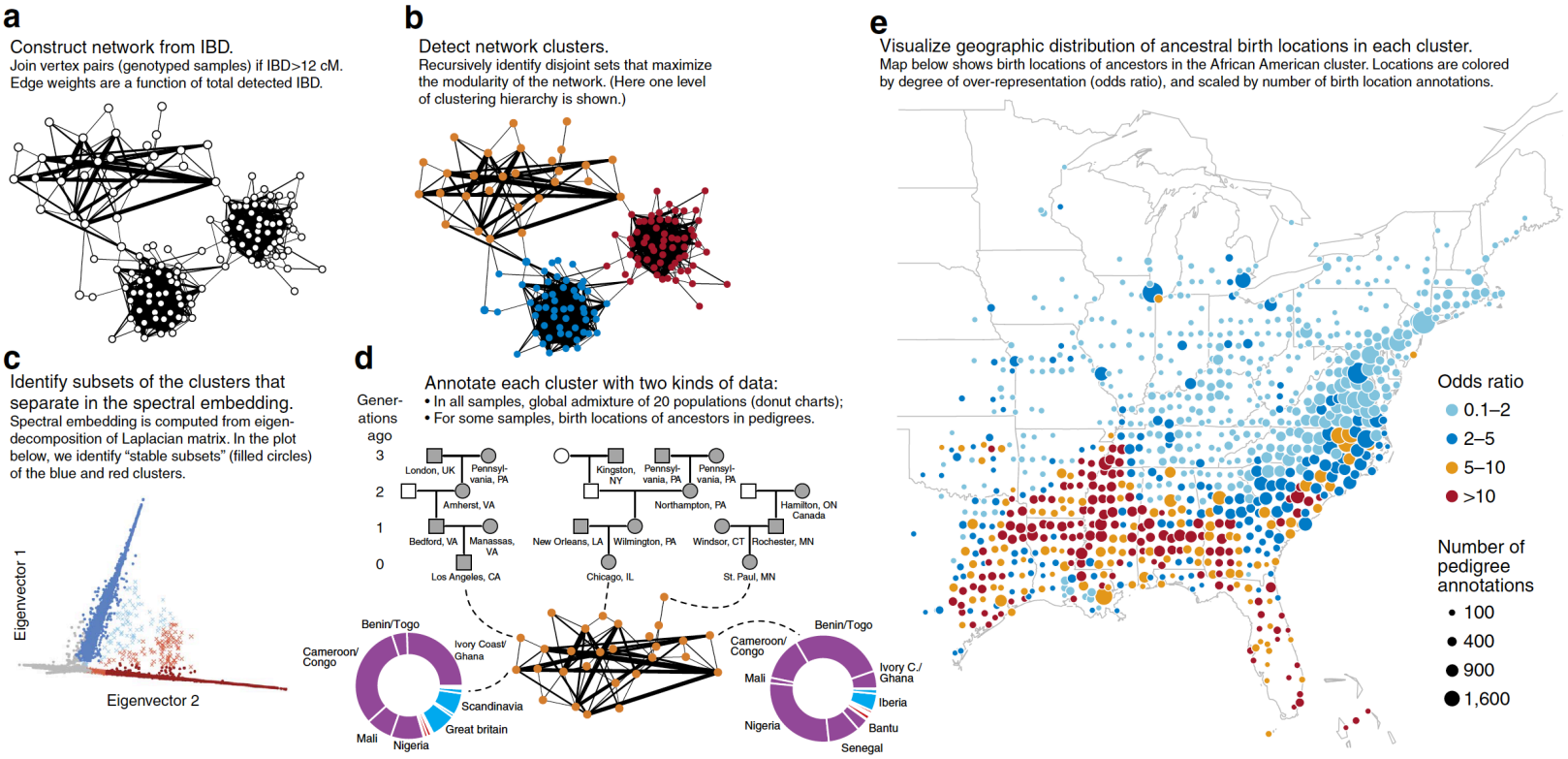
- Organize data for human analysis
- Infer prototypes for distinct groups
- Detect novel groups in the data
- Test for biases in data sets
- Find smaller, representative subsets of data
- …

# Tree of life




https://www.evogeneao.com/

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



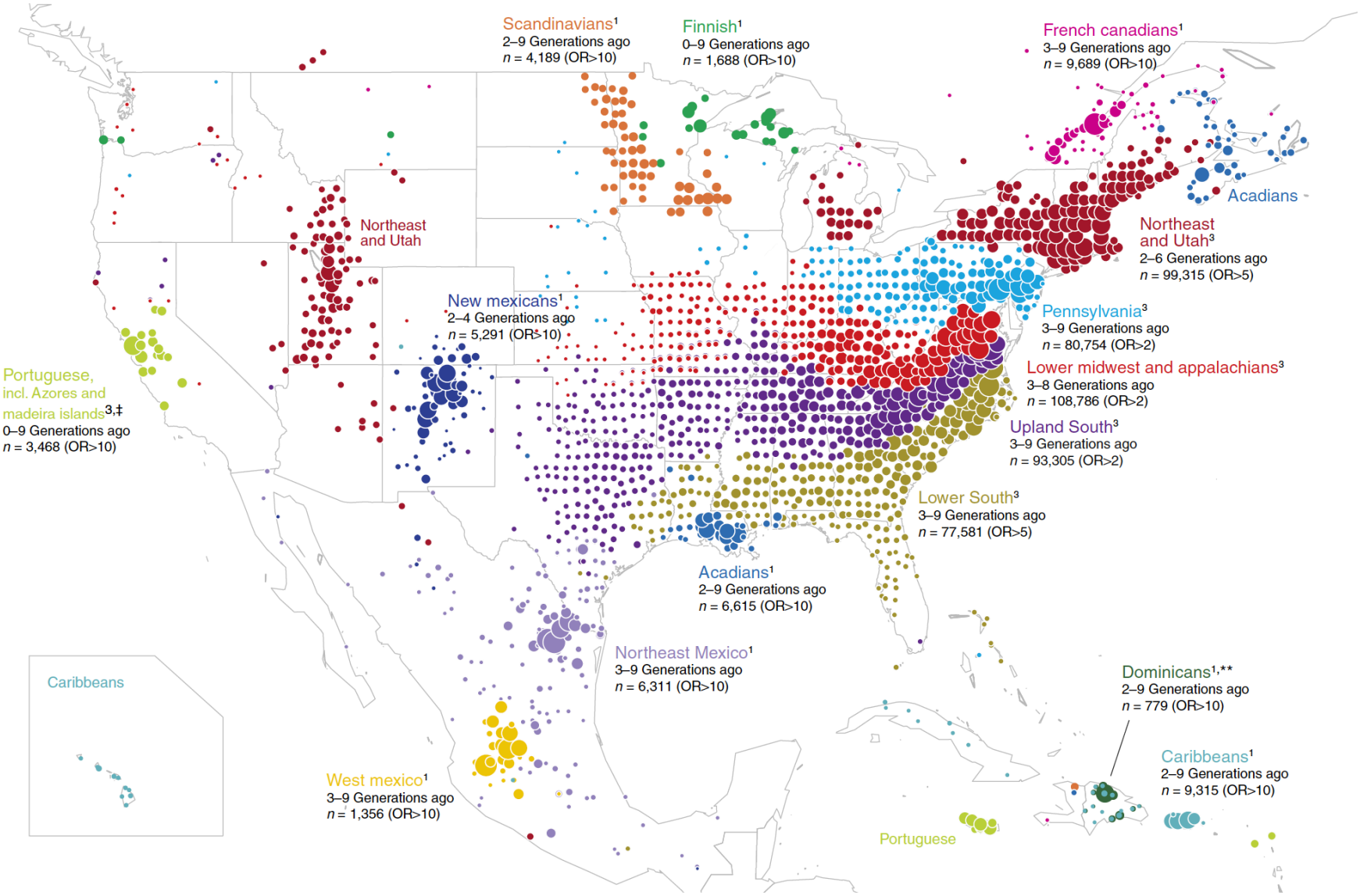

Clustering of 770,000 genomes. Han et al. Nature Comm. 2016

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

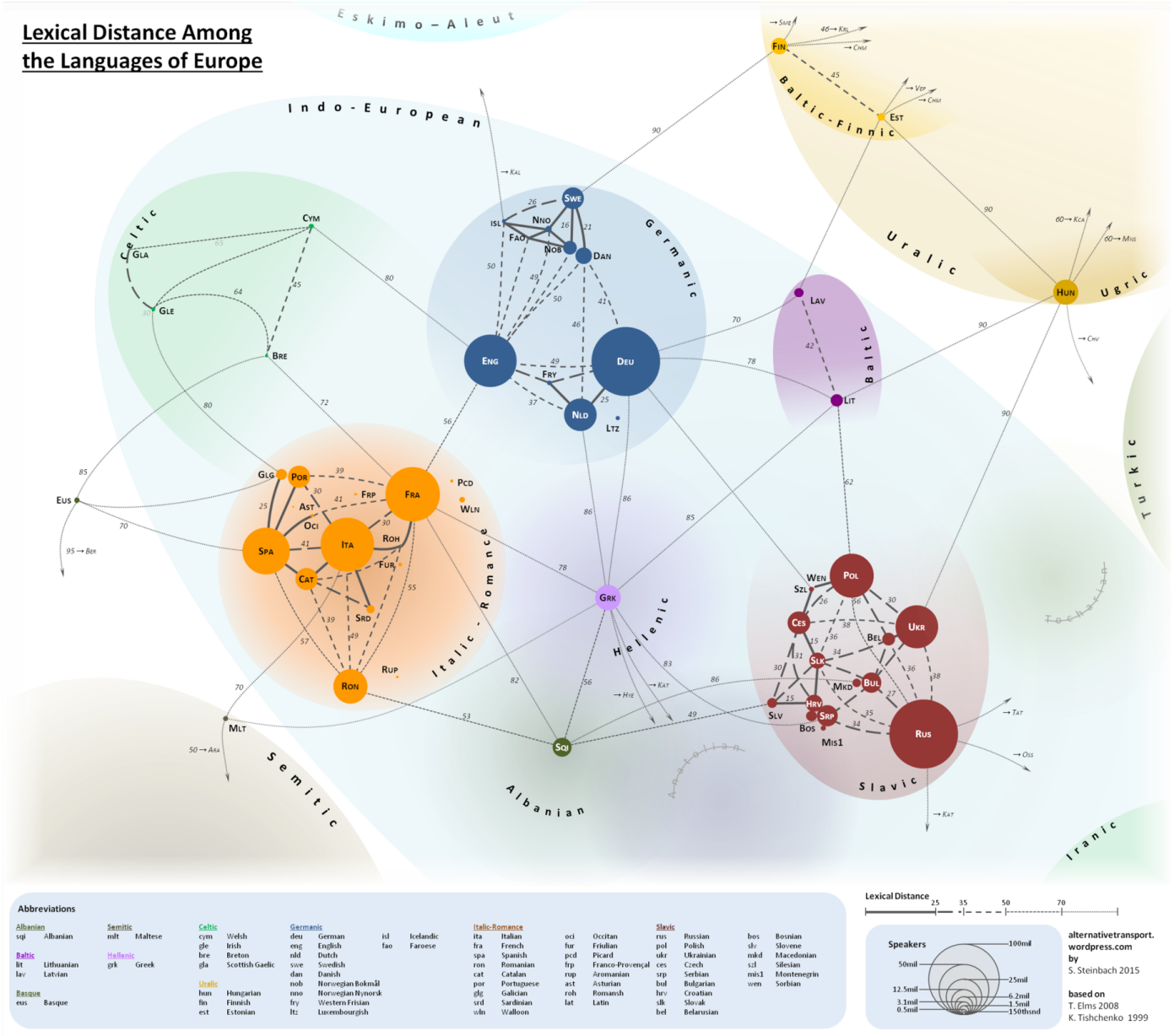
- ”computational methods reveal densely connected clusters, in which the members of each cluster are subtly more related to each other.”

- ”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



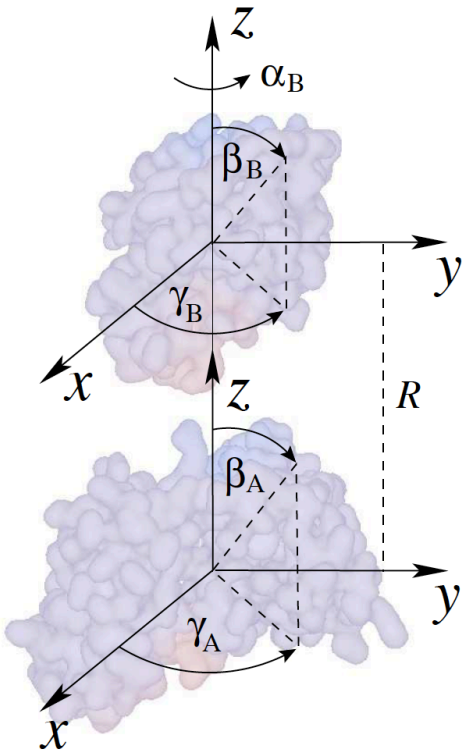

Clustering of 770,000 genomes. Han et al. Nature Comm. 2016




# Lexical Distance Among Languages of Europe 2015

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

# Protein-protein docking



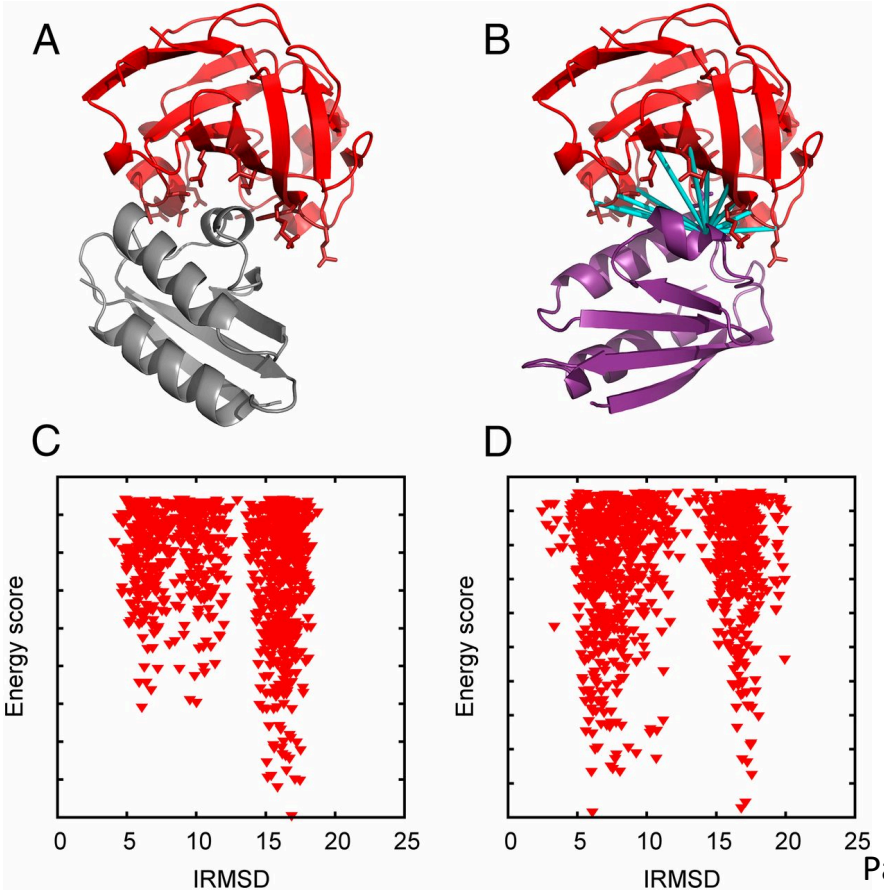

Dave Ritchie

When predicting how proteins A and B dock together, we might sample millions of possible conformations.

We can sort these and prefer the conformations with low energy.

Unlikely (but possible) that correct conformation is top of our list, but it might be in (say) the top 1000.

# Protein-protein docking



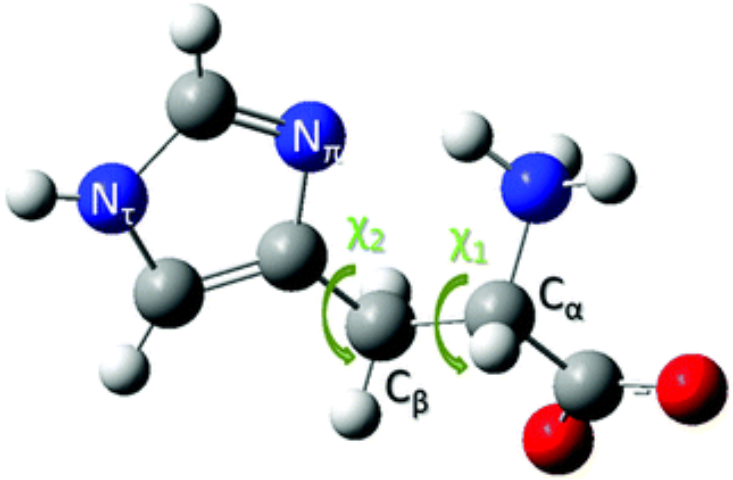

In docking challenges, can typically report up to 100 models.

How to select, and how to maximise chance that a prediction close to the correct conformation is high on the list?

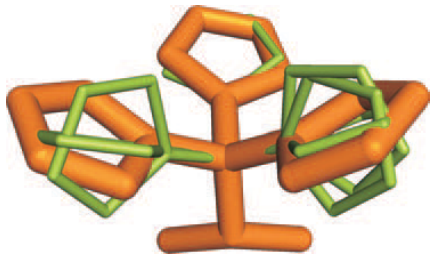
Interface root mean squared distance (IRMSD) vs. energy.

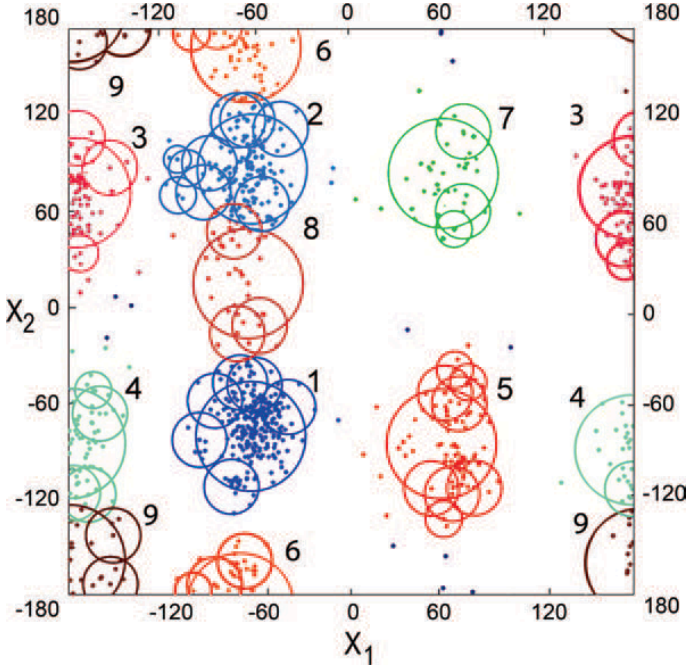
Padhorny et al. (2016) PNAS July 26, 2016 113 (30) E4286-E4293

# Protein side chain rotamers

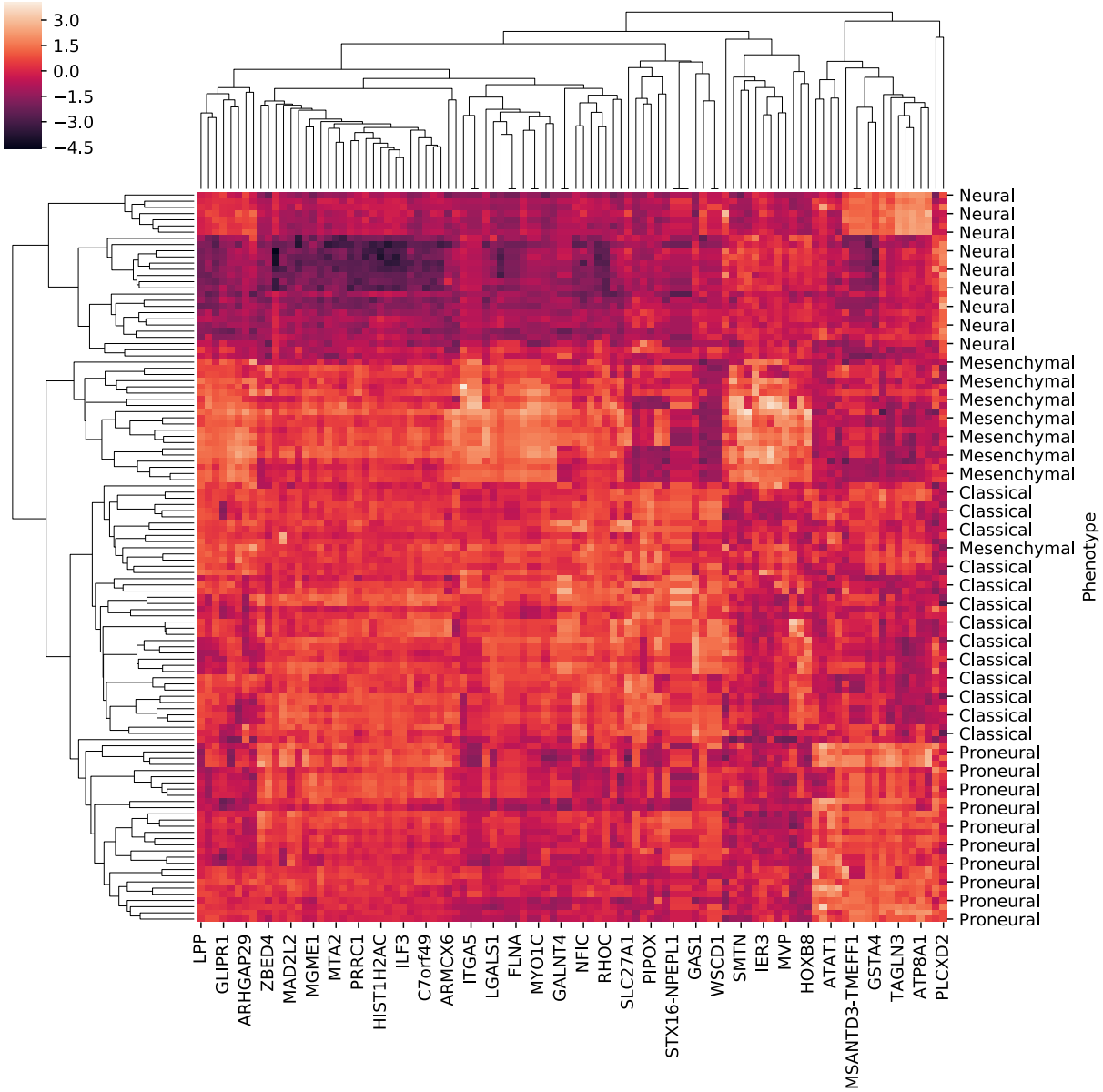



Cardamone et al (2016), Phys. Chem. Chem. Phys, 18, 27377-27389






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

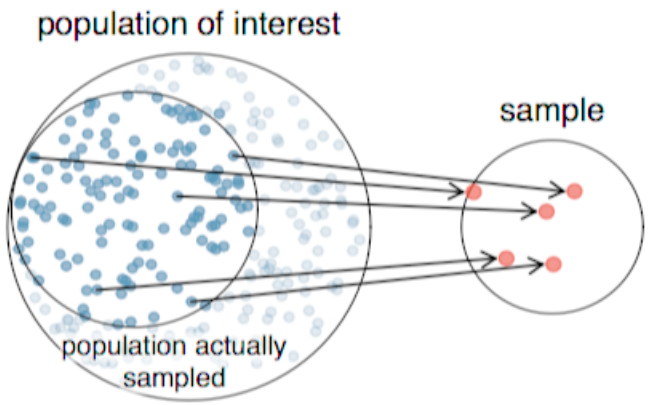


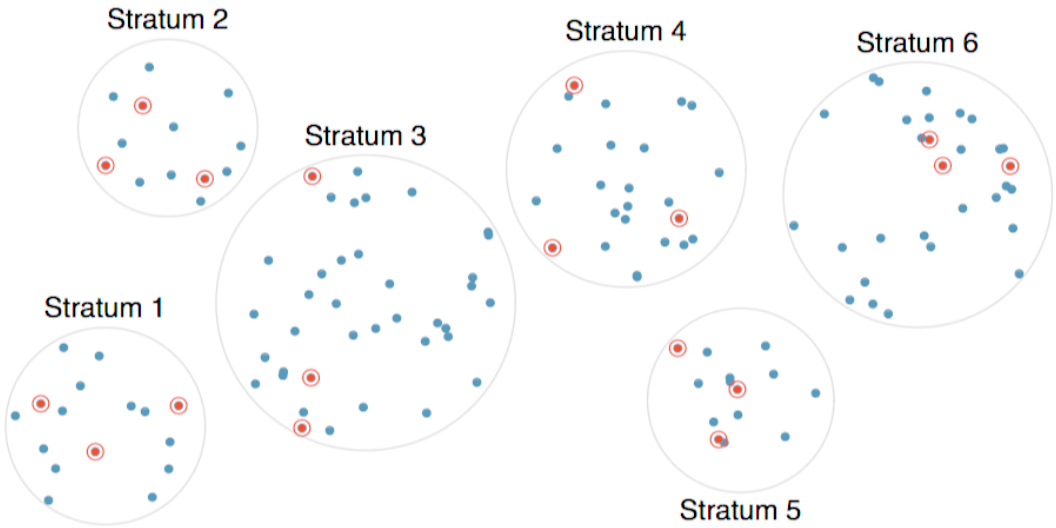

# Bi-clustered heatmap of 118 samples using hierarchical clustering of 100 gene features

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

# Case study:
# Creating Unbiased Data Sets

# Correct for different survey response rates



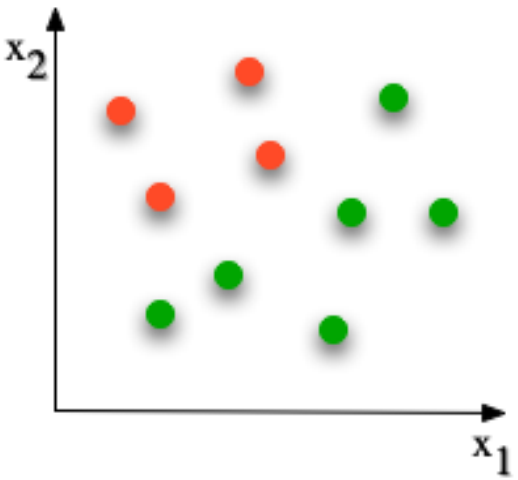

Figures from OpenIntro statistics

# Purpose of data clustering

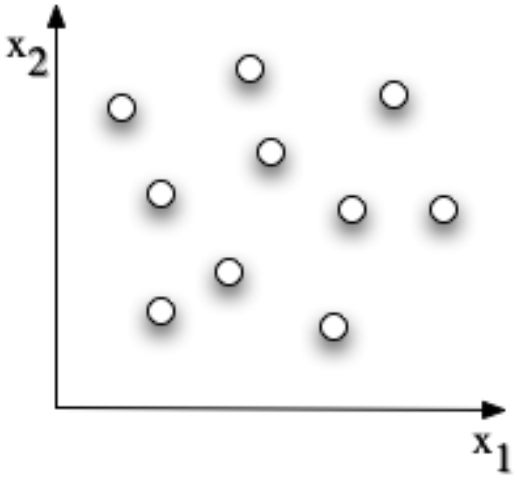
- Underlying structure
  - to gain insights into data, generate hypotheses and identify salient features
- Natural classification
  - to identify the degree of similarity among forms or organisms (phylogenetic relationships)
- Compression
  - to organise data and summarise it through cluster prototypes

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

# (Un)Supervised Learning

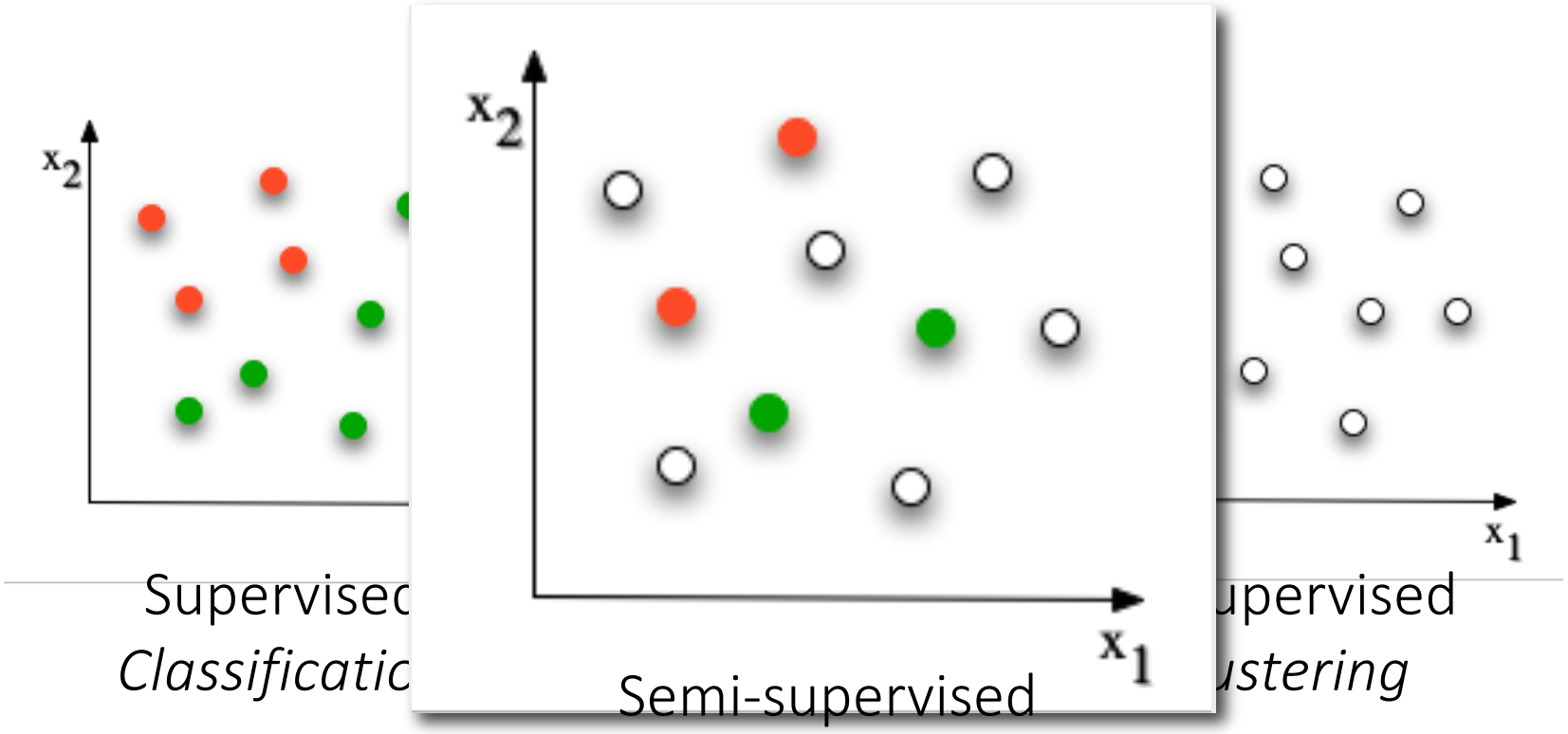



Supervised
*Classification*

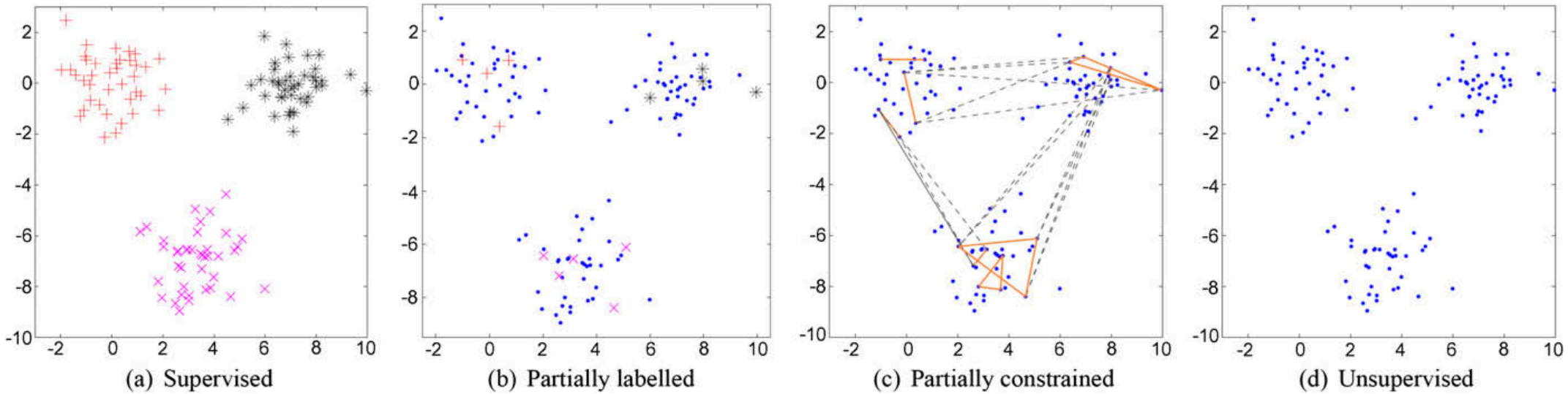



Unsupervised
*Clustering*

# (Un)Supervised Learning



# Spectrum of learning problems

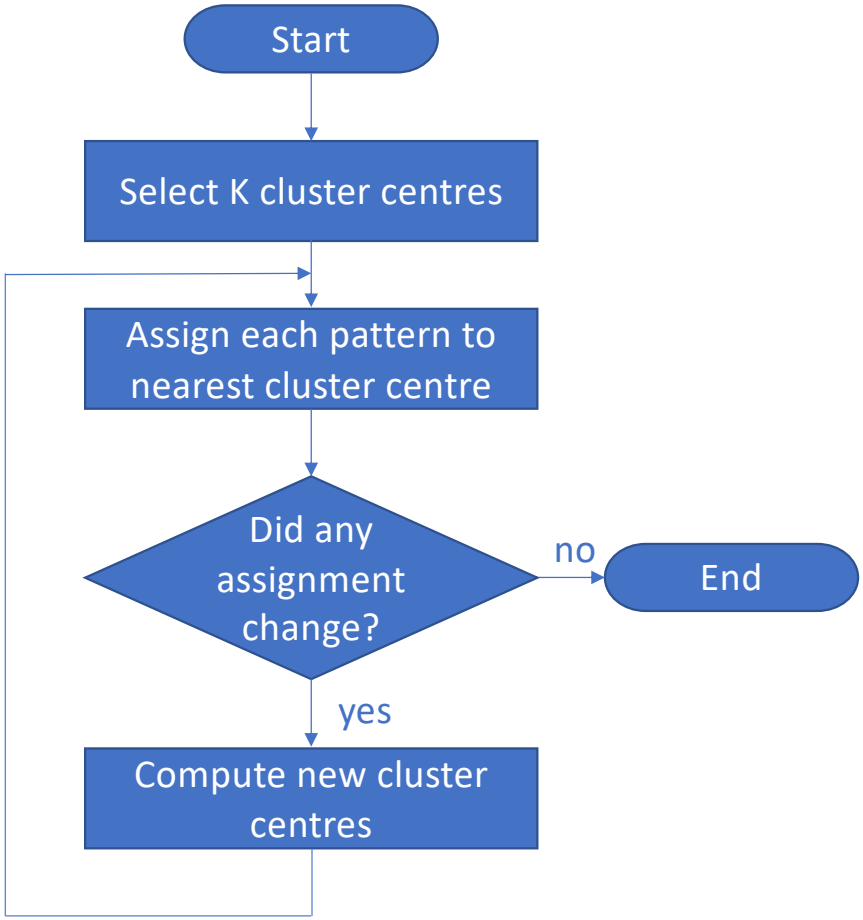



Jain, A.K. (2010) Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, **31**, 651-666

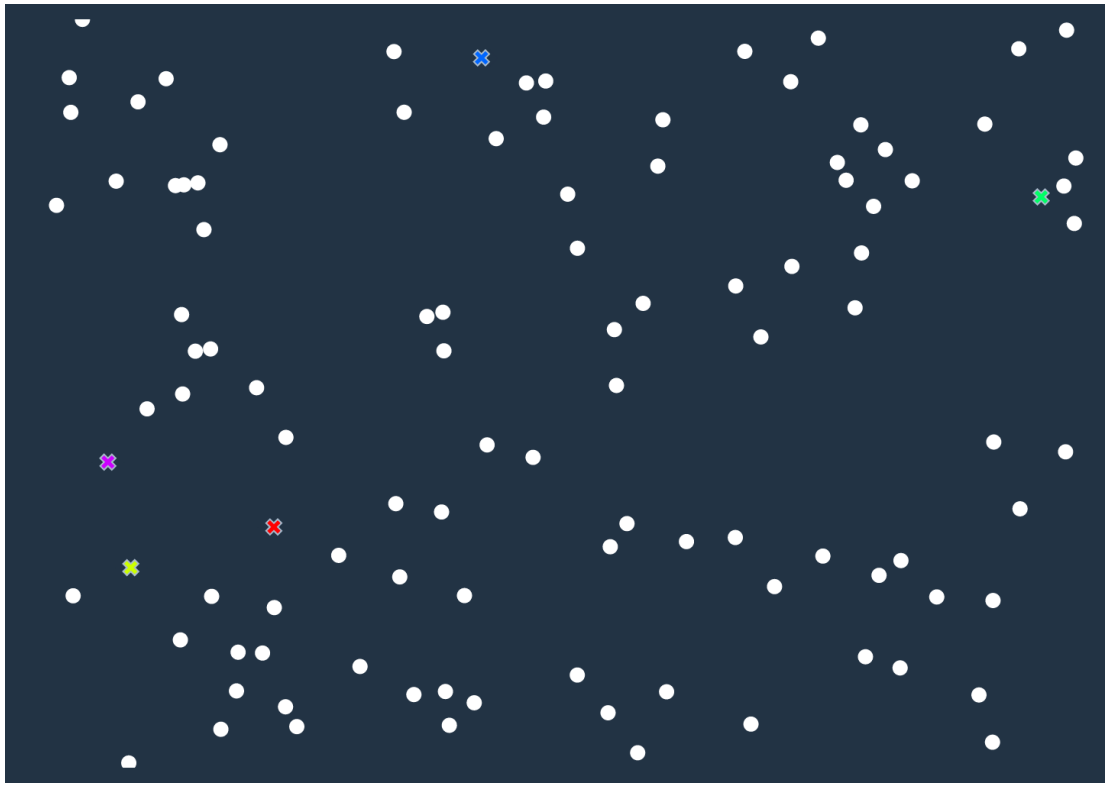
# K-means

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

- a greedy algorithm
- converges to a local minimum

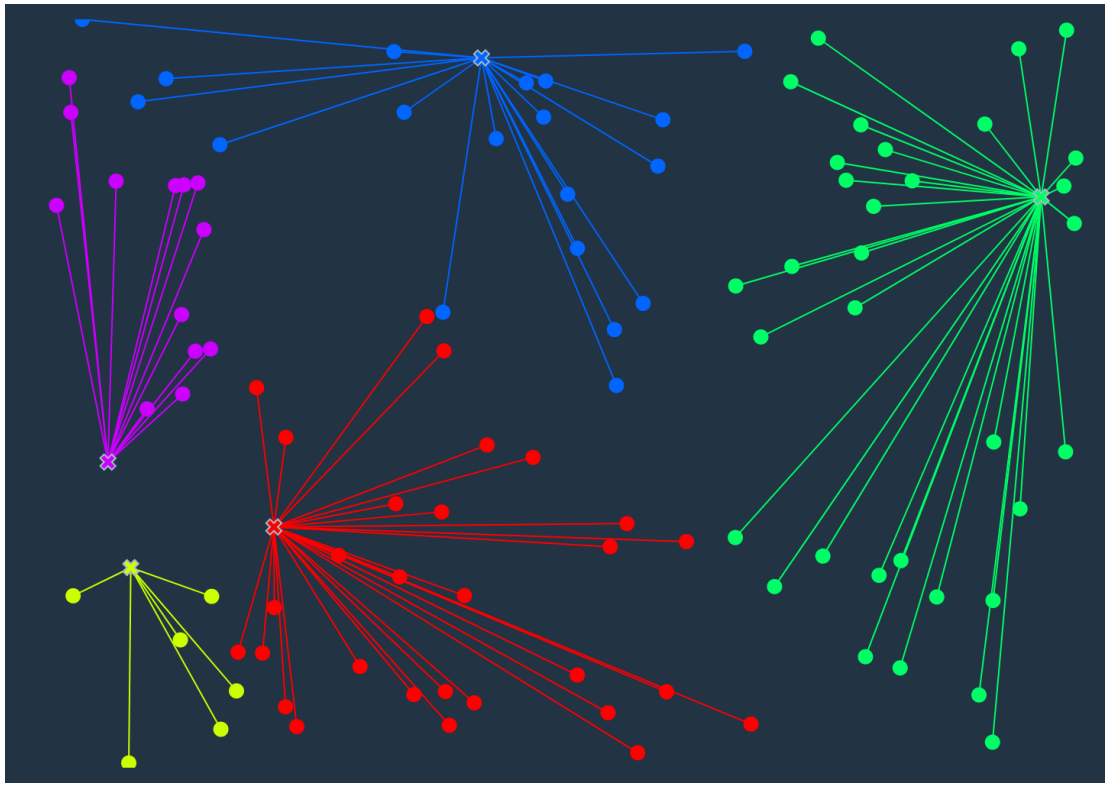


# Select initial cluster “centres”



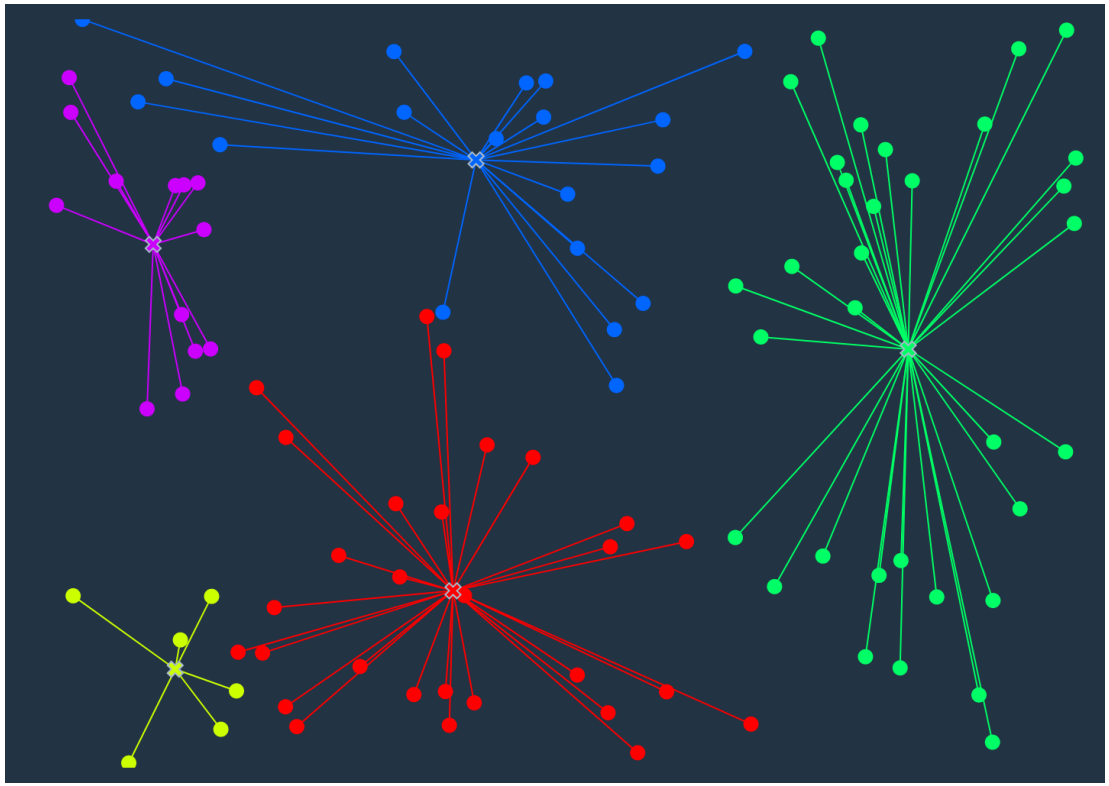
http://tech.nitoyon.com/en/blog/2013/11/07/k-means/

# Assign patterns to nearest cluster centre (1)



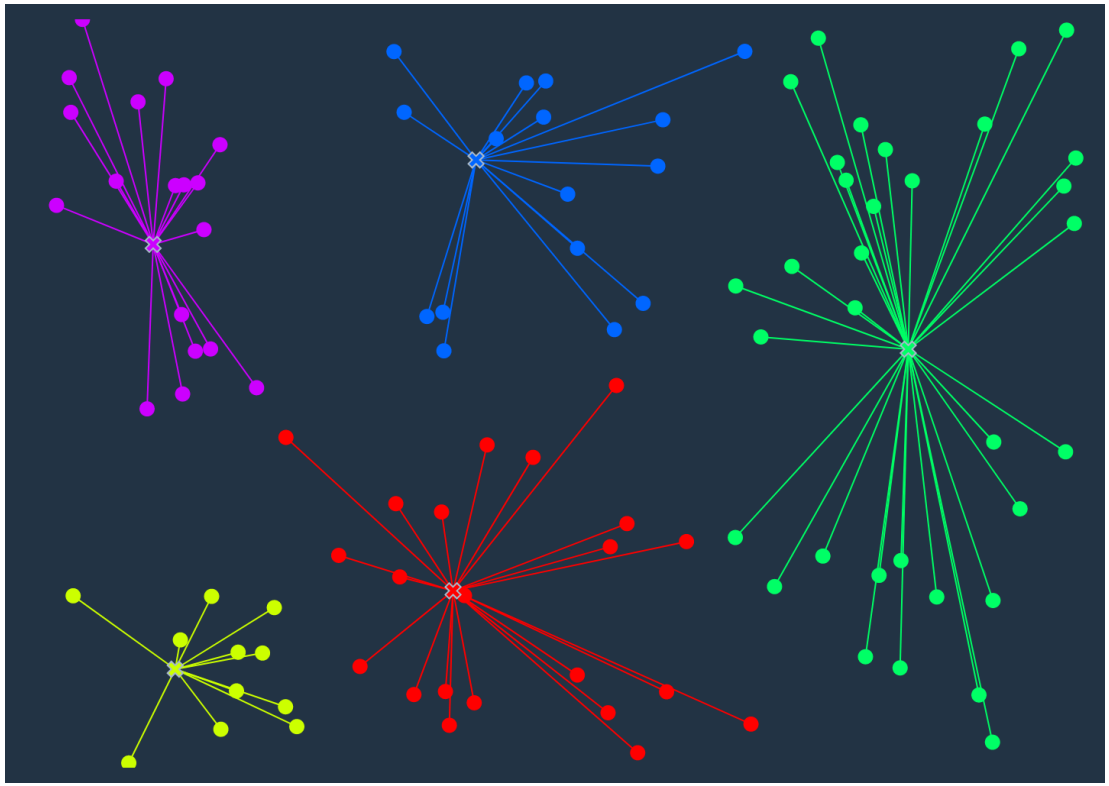
http://tech.nitoyon.com/en/blog/2013/11/07/k-means/

# Compute new cluster centres (1)



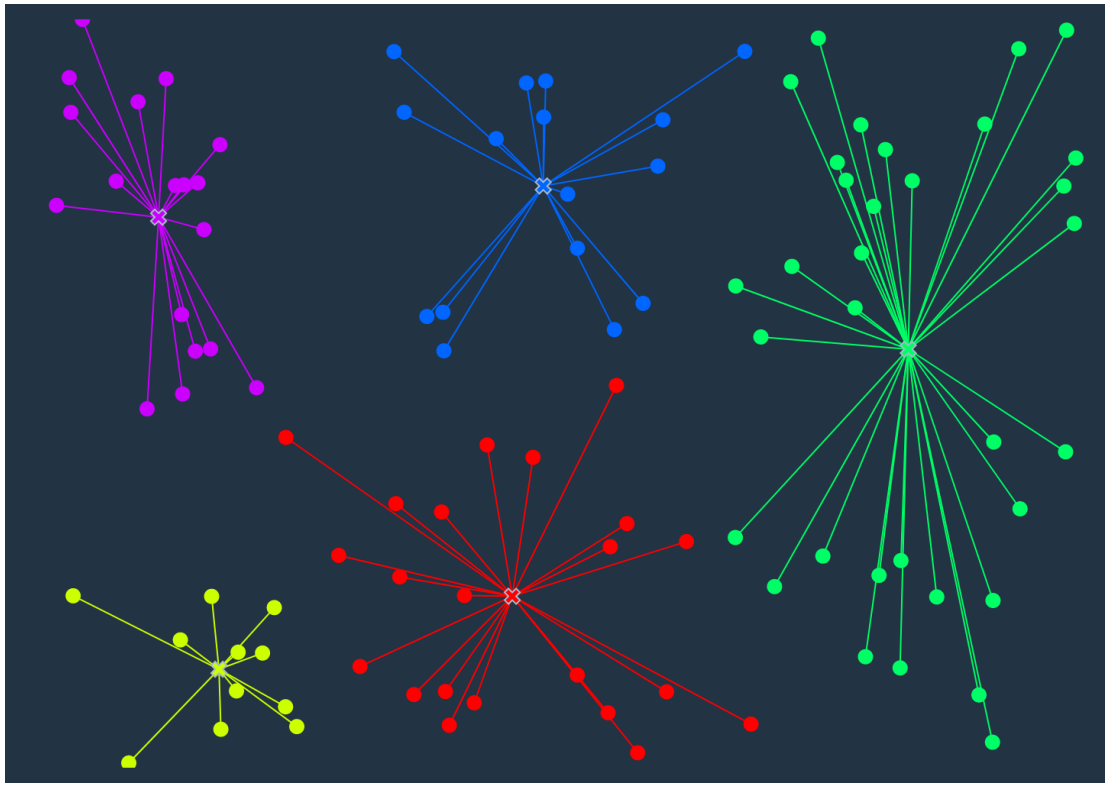
http://tech.nitoyon.com/en/blog/2013/11/07/k-means/

# Assign patterns to nearest cluster centre (2)



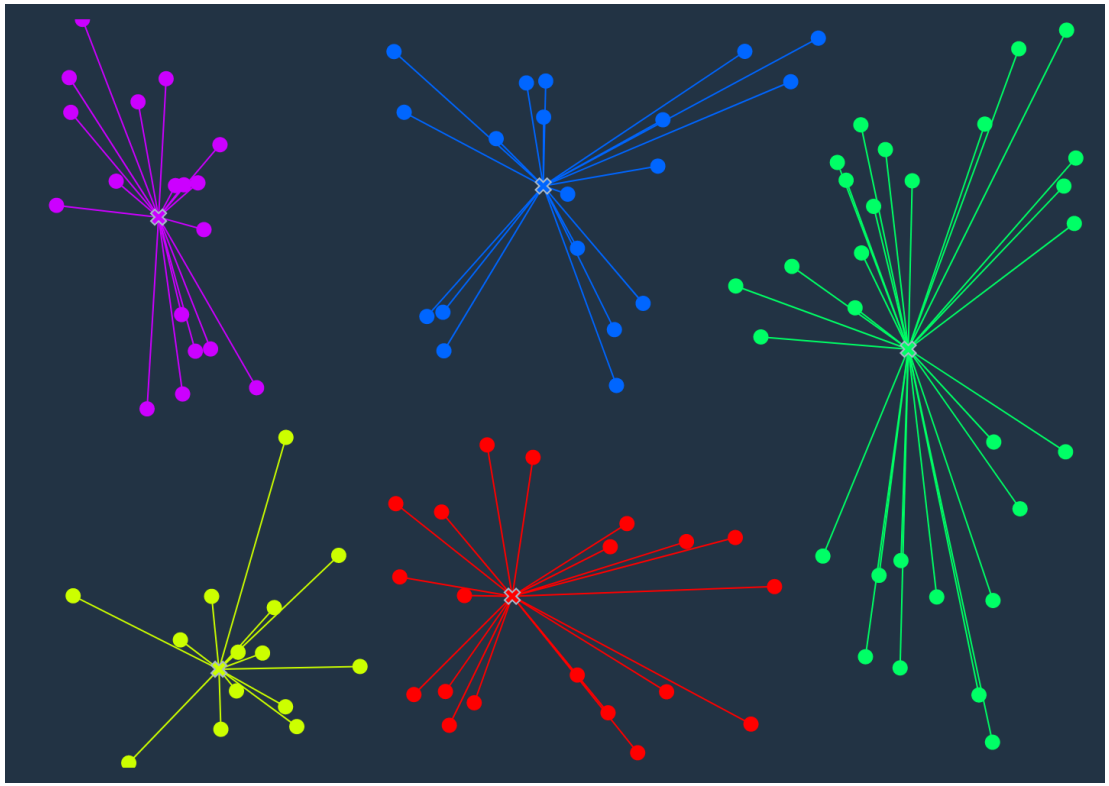
http://tech.nitoyon.com/en/blog/2013/11/07/k-means/

# Compute new cluster centres (2)



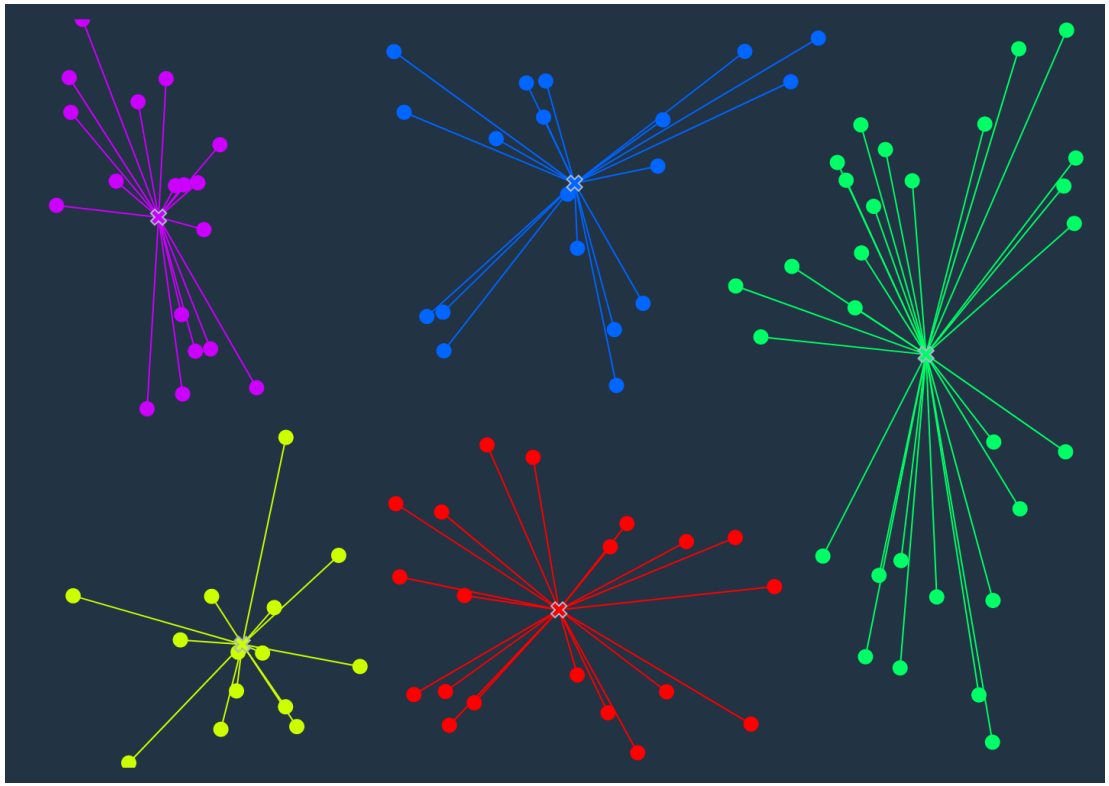
http://tech.nitoyon.com/en/blog/2013/11/07/k-means/

# Assign patterns to nearest cluster centre (3)



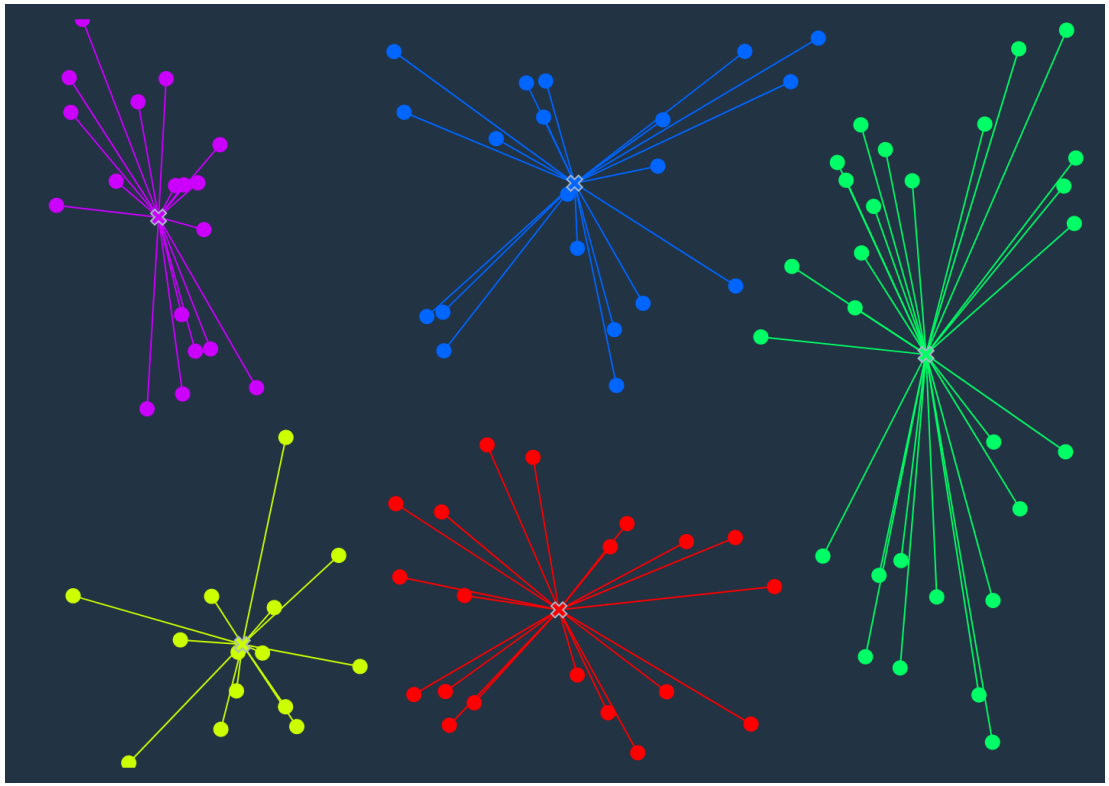
http://tech.nitoyon.com/en/blog/2013/11/07/k-means/

# Compute new cluster centres (3)



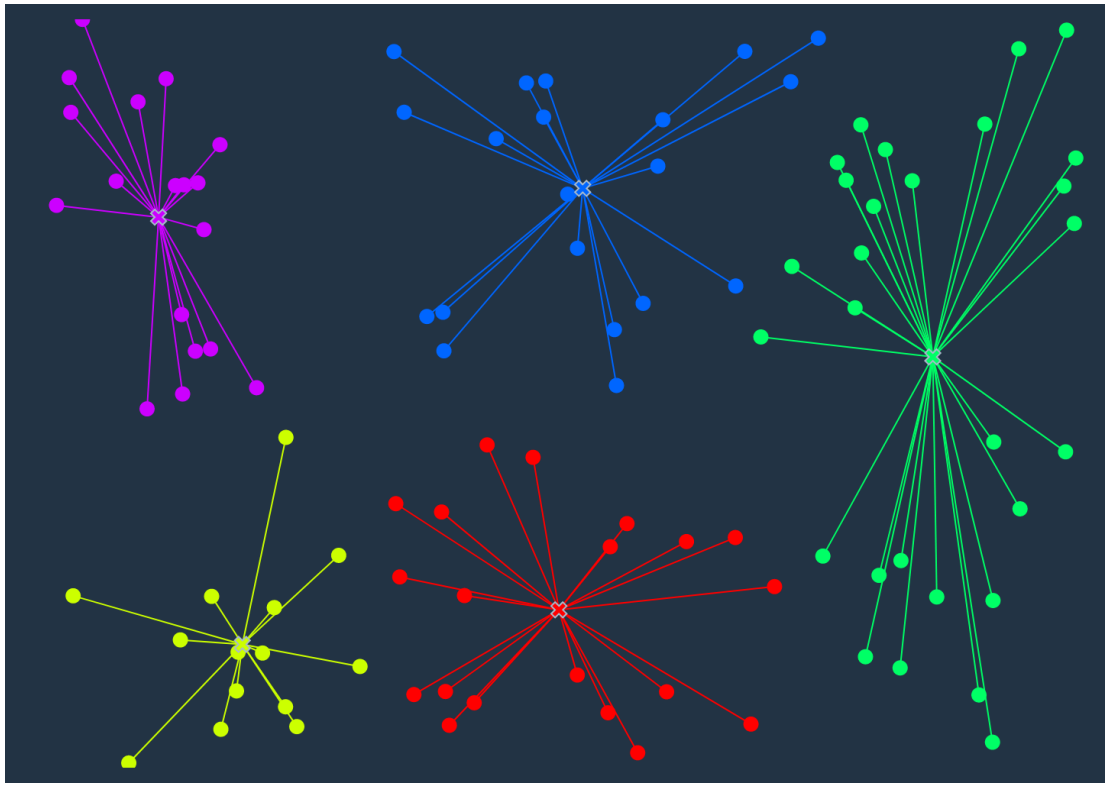
http://tech.nitoyon.com/en/blog/2013/11/07/k-means/

# Assign patterns to nearest cluster centre (4)



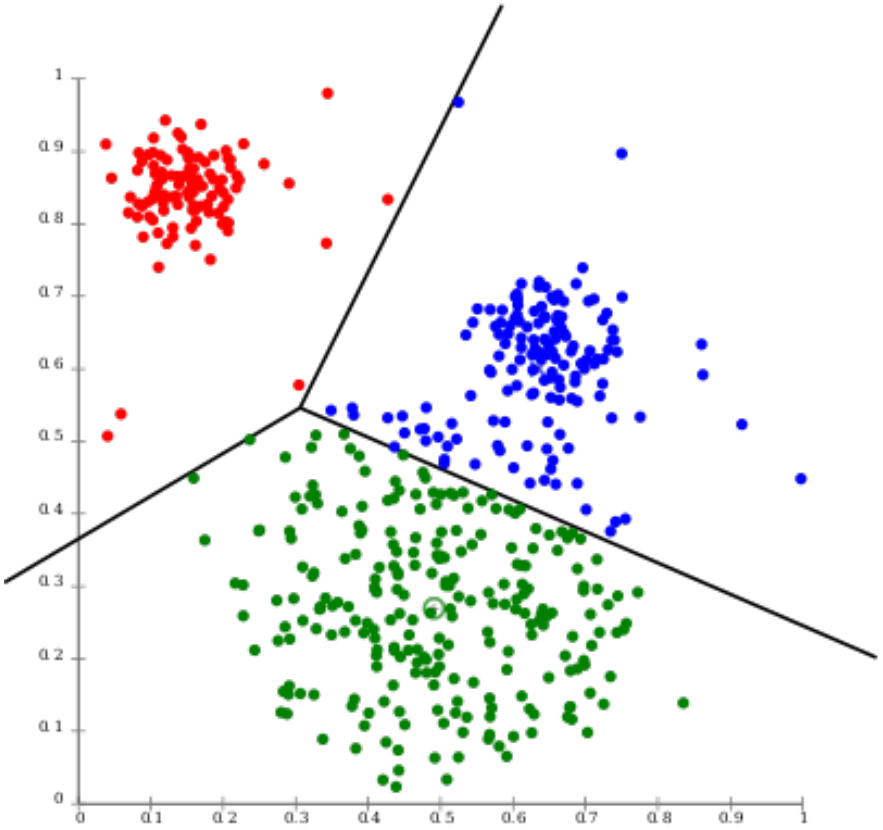
http://tech.nitoyon.com/en/blog/2013/11/07/k-means/

# Compute new cluster centres (4)



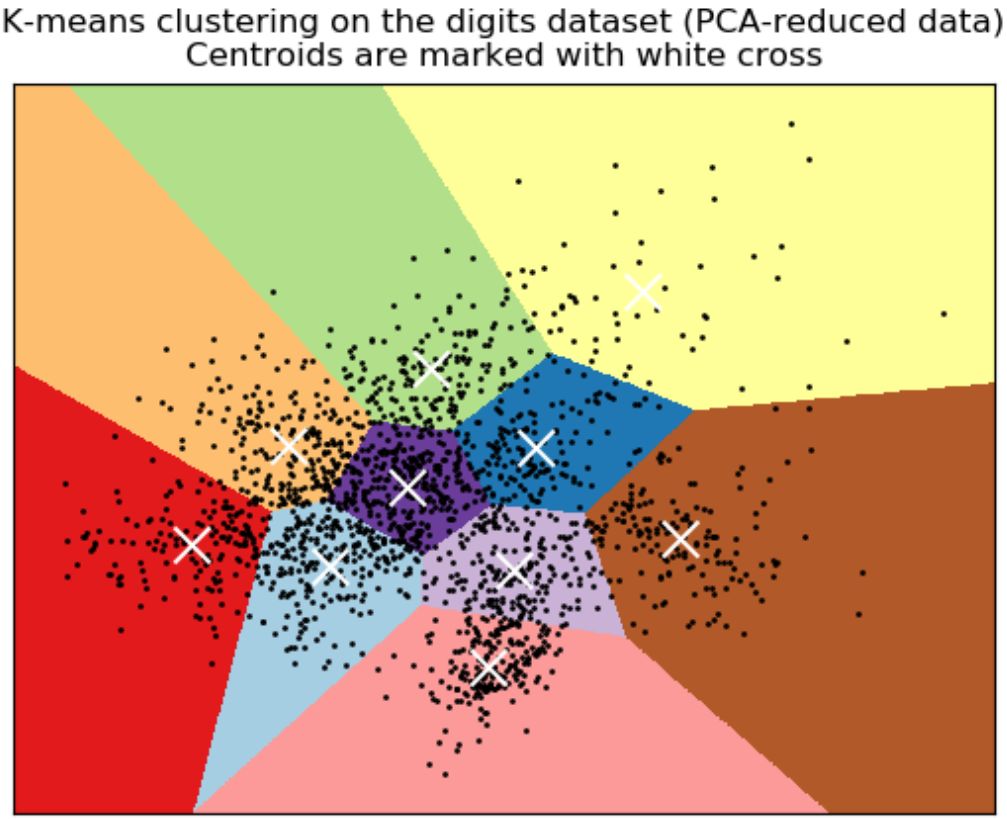
http://tech.nitoyon.com/en/blog/2013/11/07/k-means/

# Geometric Interpretation



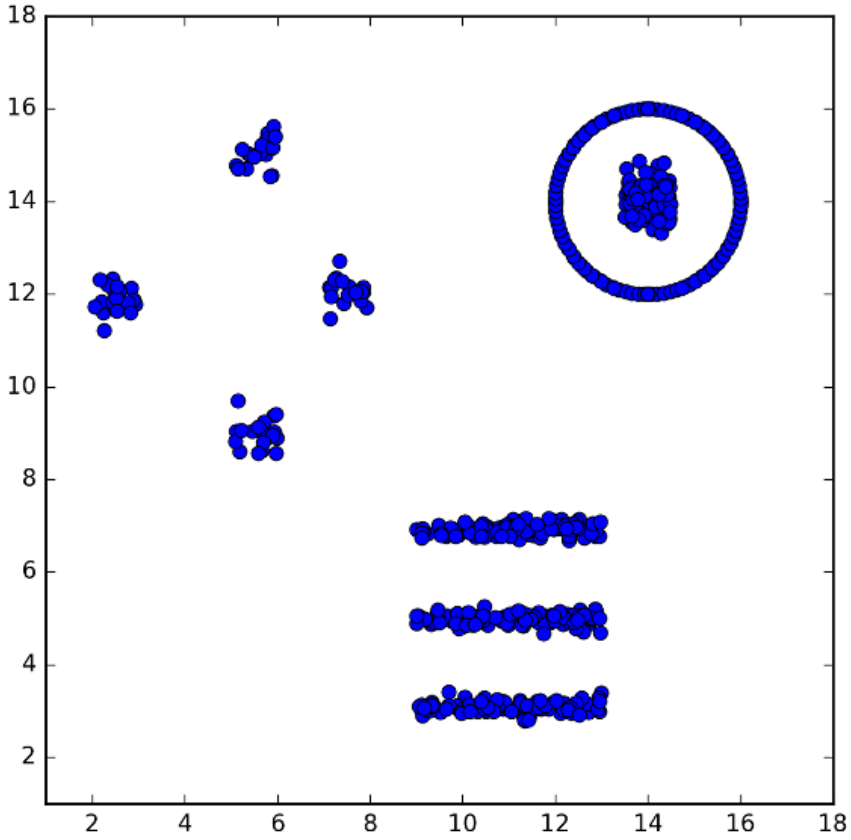

*Voronoi diagram* for centroids as seeds. Points in area between lines closer to their centroid than to any other

# Clustering hand-written digits



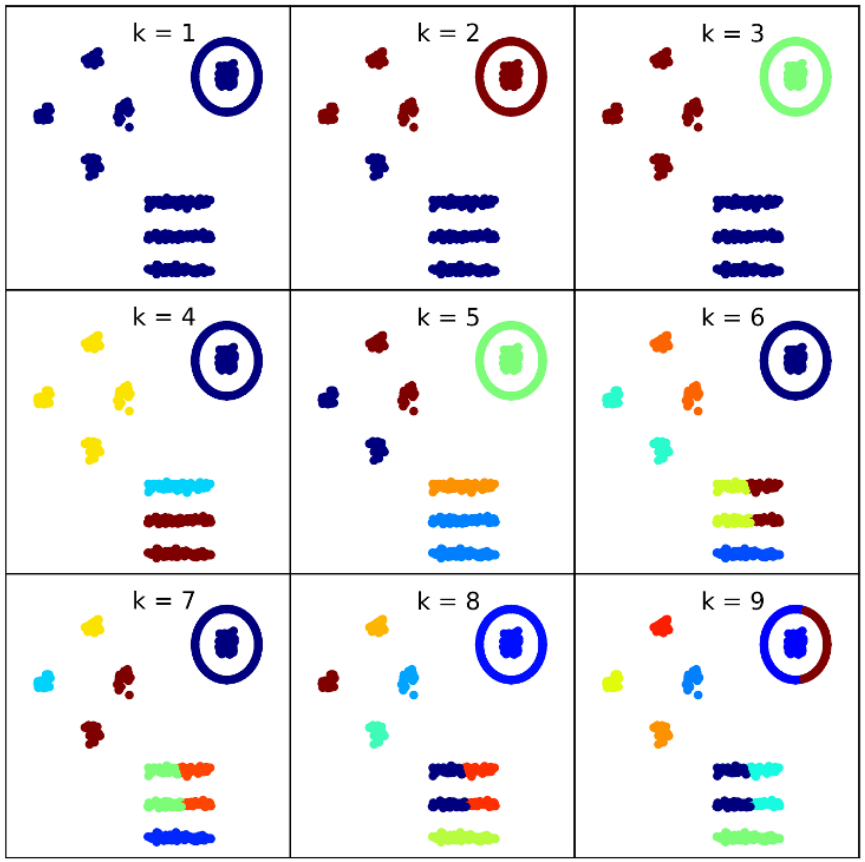

https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_digits.html#sphx-glr-auto-examples-cluster-plot-kmeans-digits-py

# What will K-means do?



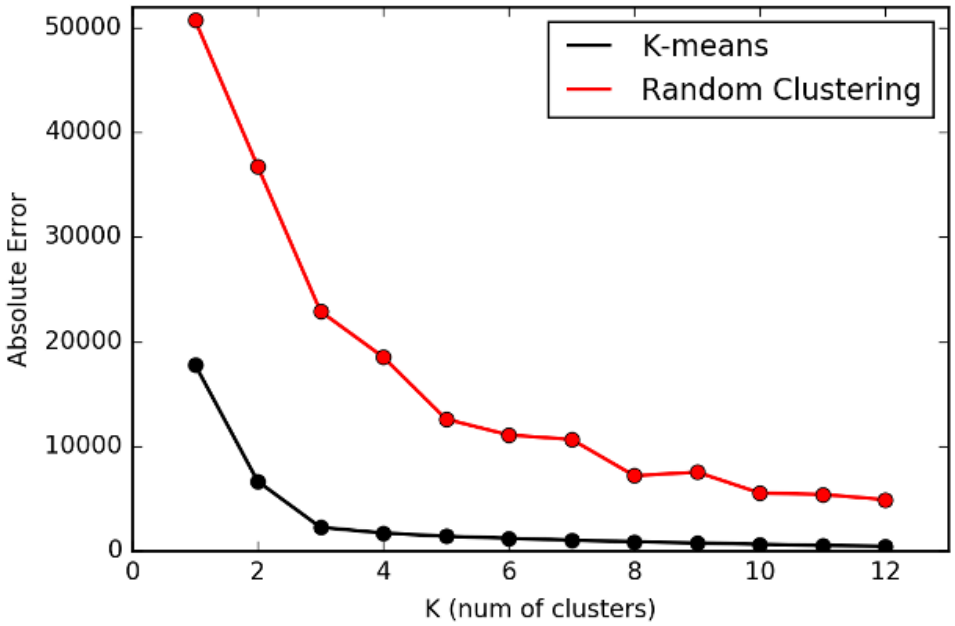

Skiena (2017) The Data Science Design Manual, Springer

# What will K-means do?




Skiena (2017) The Data Science Design Manual, Springer

# “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

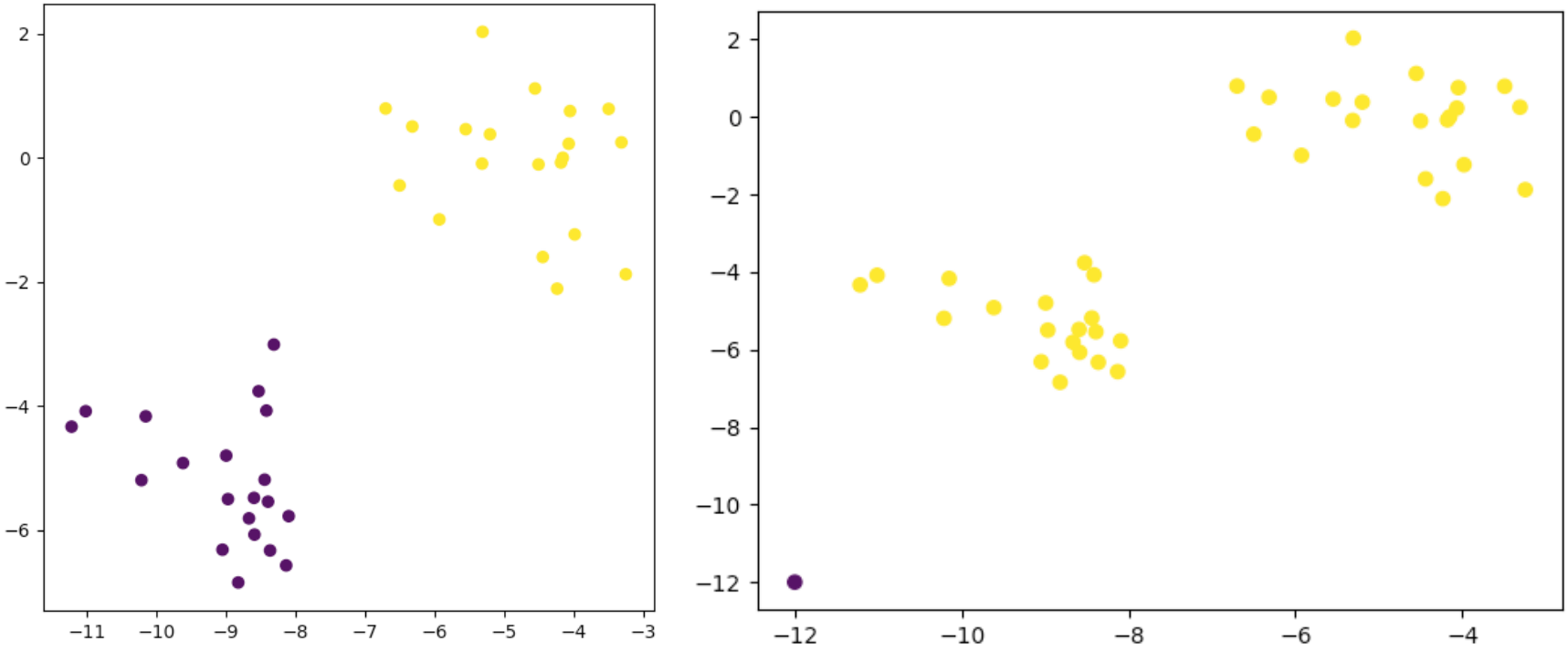
Skiena (2017) The Data Science Design Manual, Springer

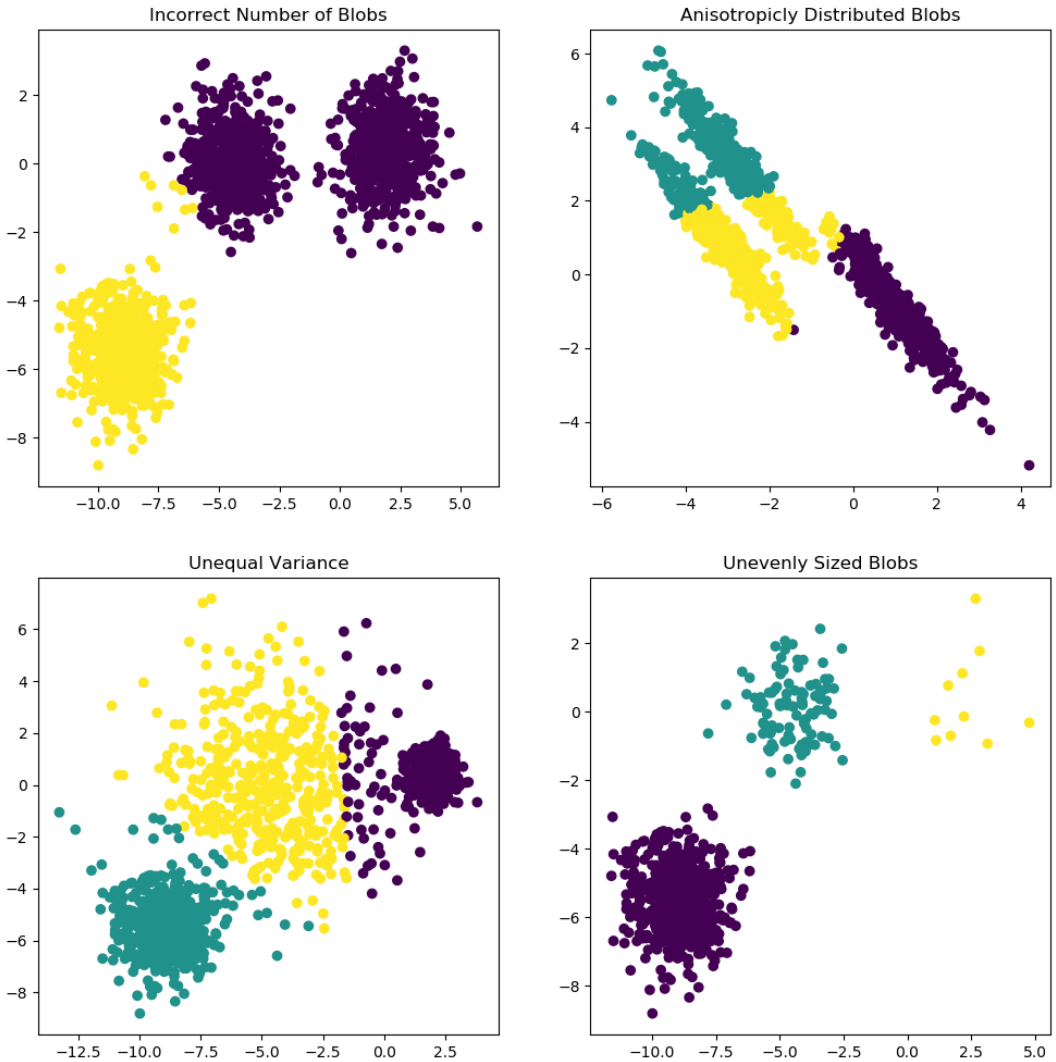
# Caveats

K-means works best for

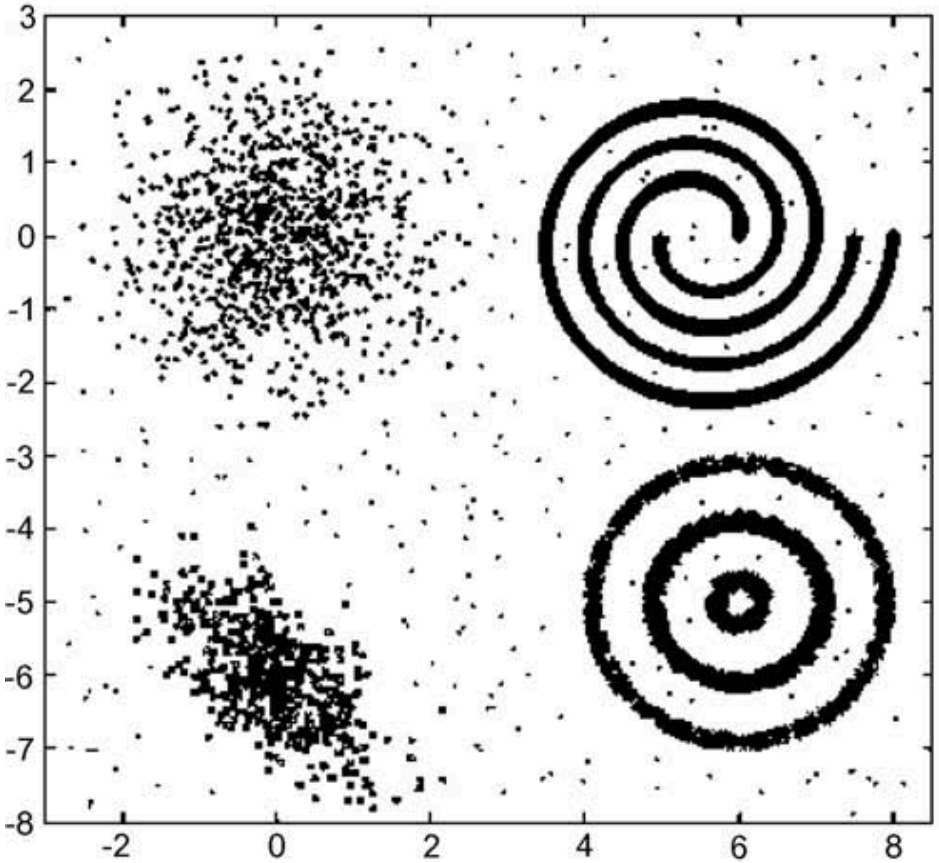
- spherical clusters
- equal diameter (equal variance)
- equal cluster size

# Caveats



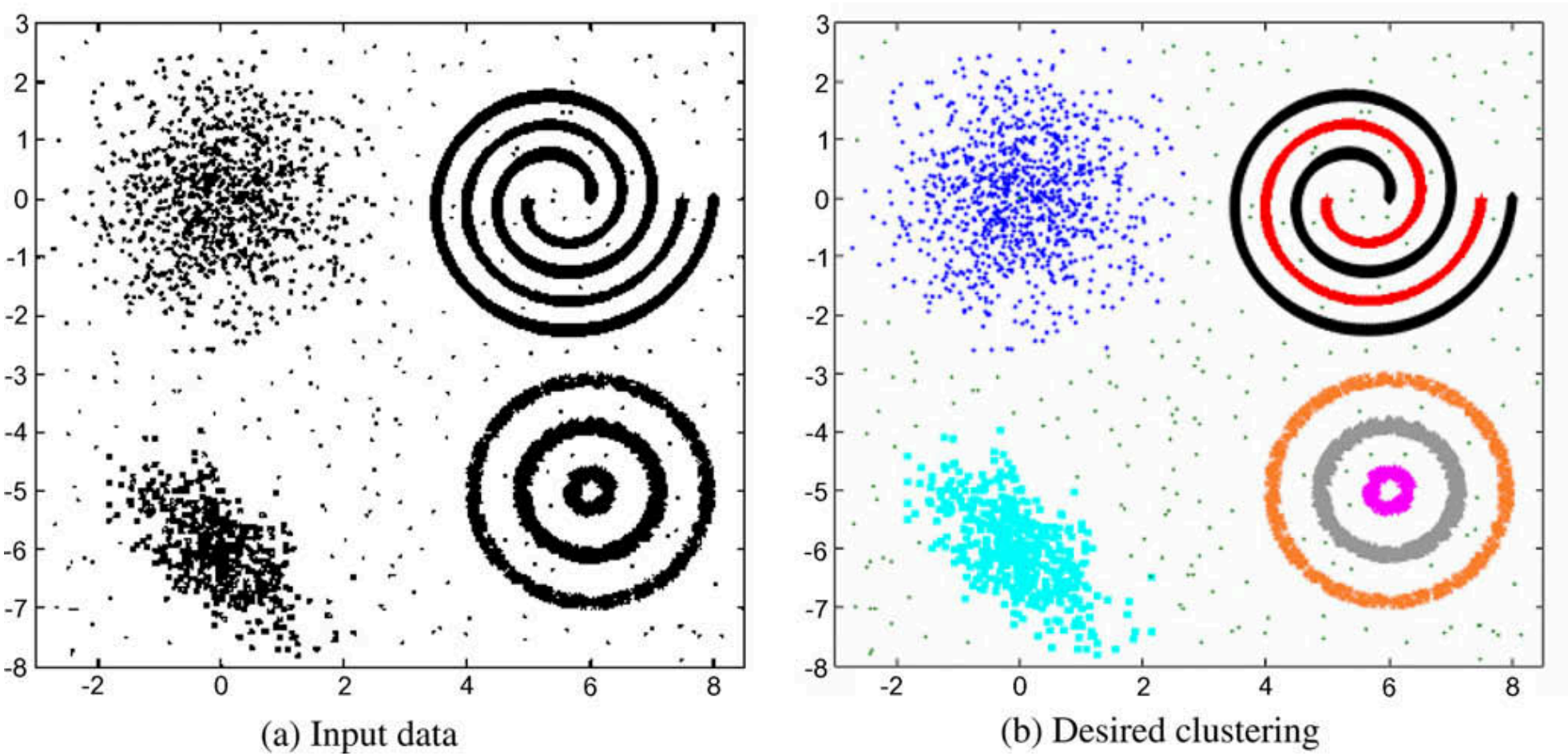



https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_assumptions.html#sphx-glr-auto-examples-cluster-plot-kmeans-assumptions-py

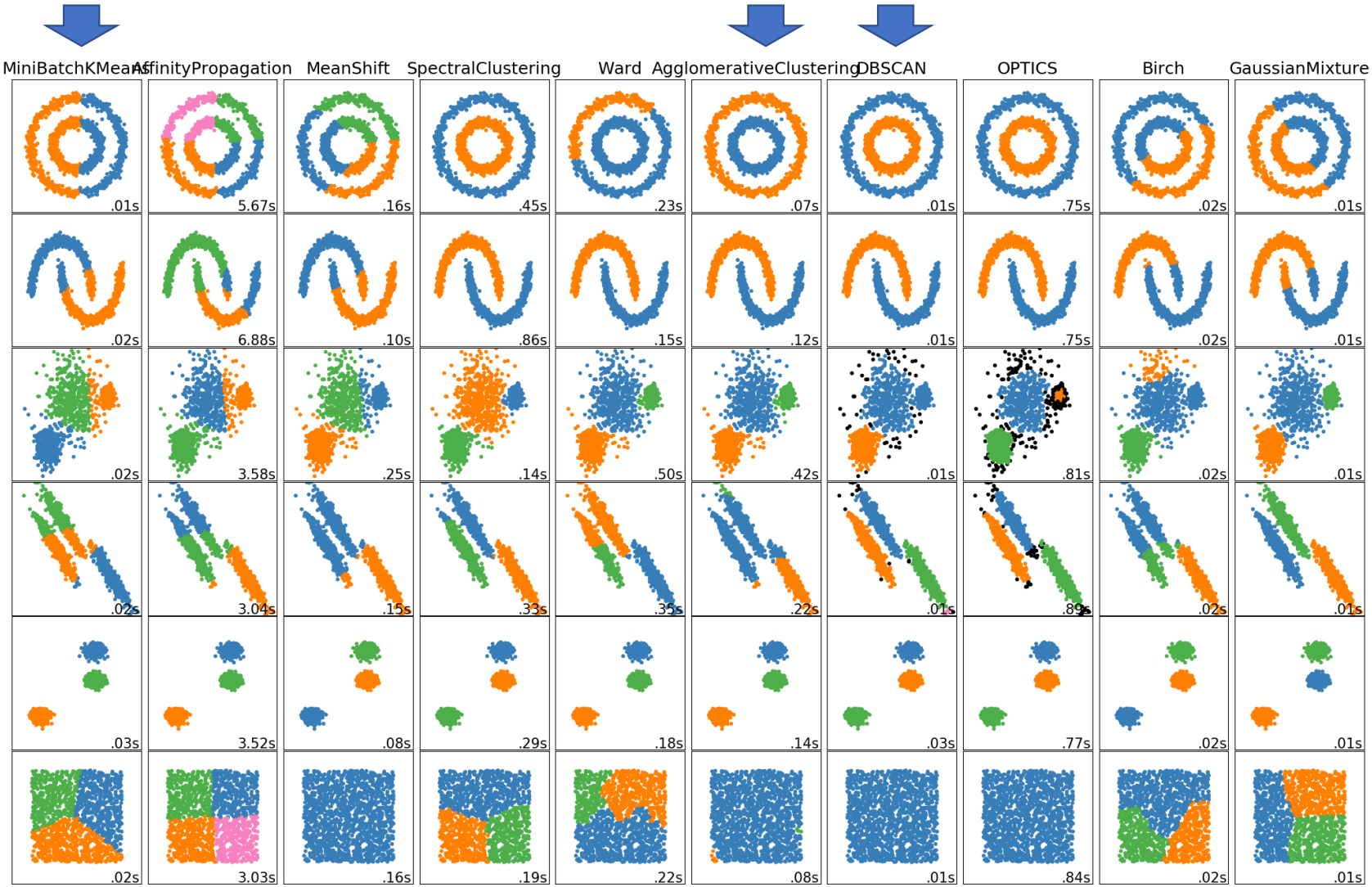



(a) Input data

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




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

# Clustering is successful, but difficult

- inherent vagueness in the definition of a cluster
- can be difficult to defining an appropriate similarity measure and objective function.

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

# Questions about clustering

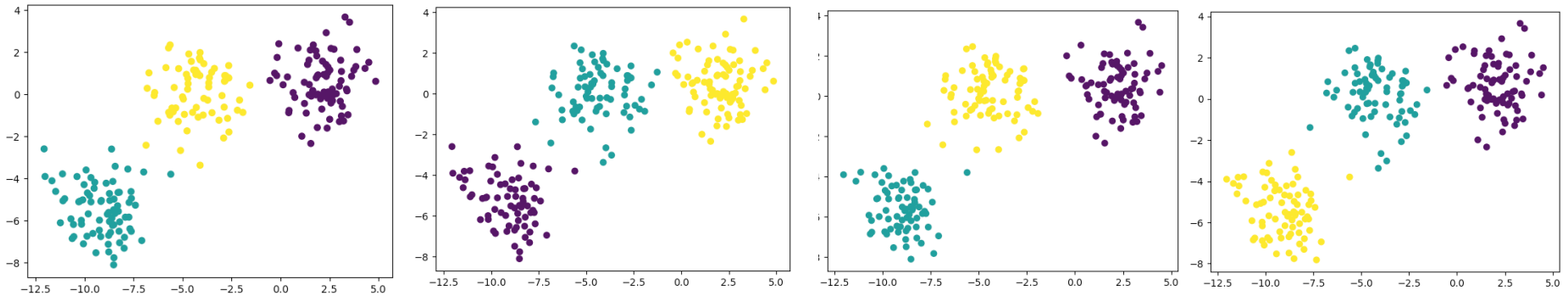
a) What is a cluster?
b) What features should be used?
c) Should the data be normalized?
d) Does the data contain any outliers?
e) How do we define the pair-wise similarity?
f) How many clusters are present in the data?
g) Which clustering method should be used?
h) Does the data have any clustering tendency?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

# Validating clusterings

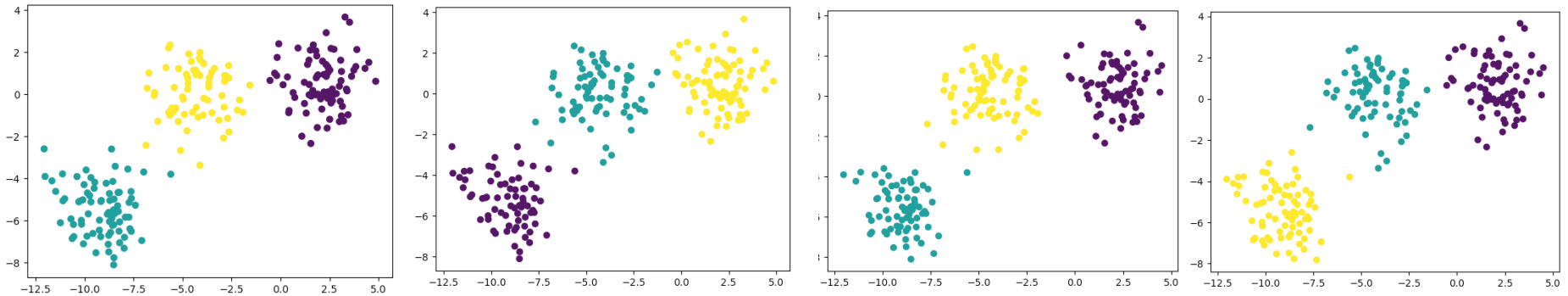
# Stability on subsets

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



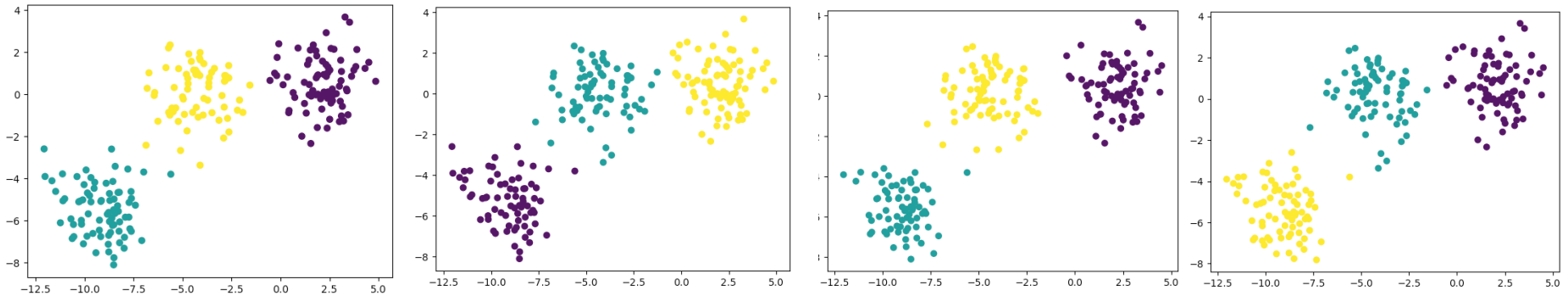
# Stability on subsets

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

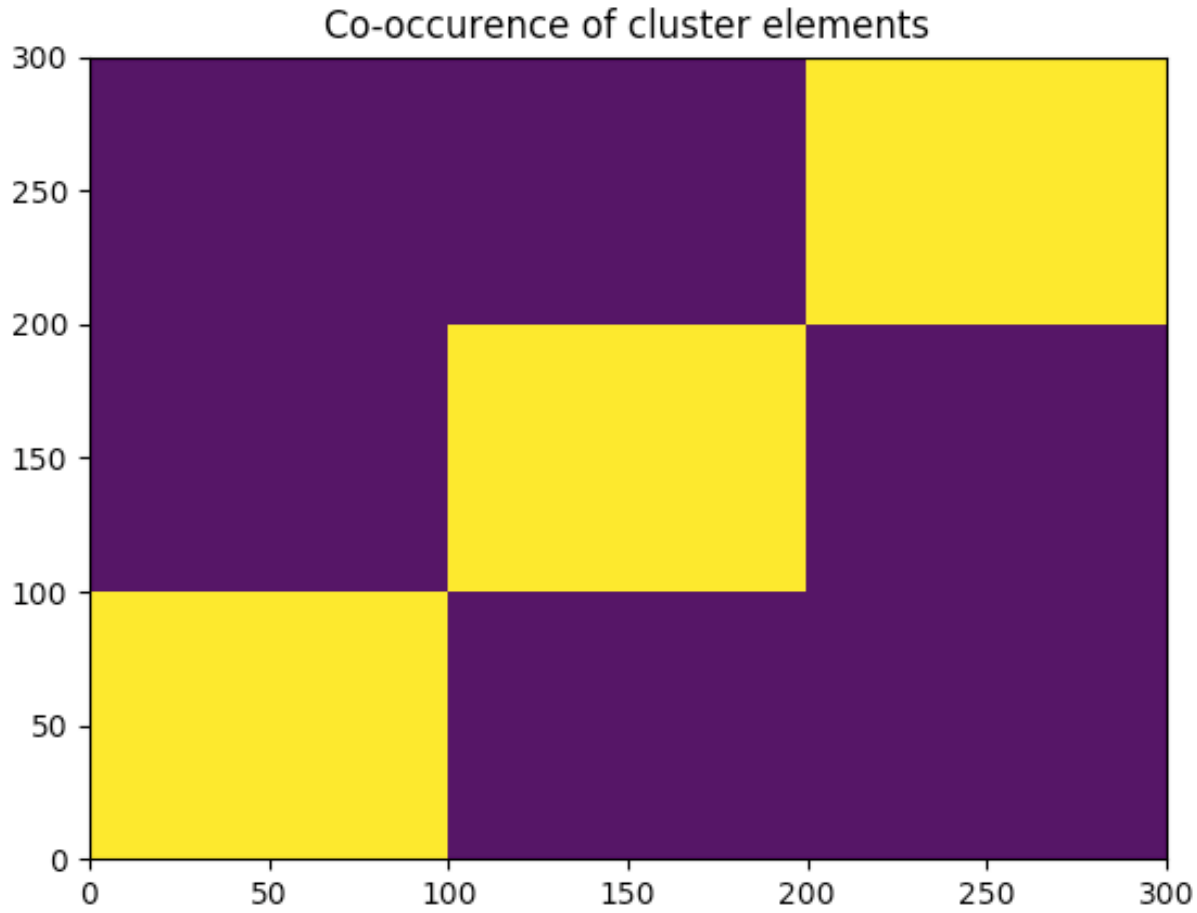


# Co-occurrence

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

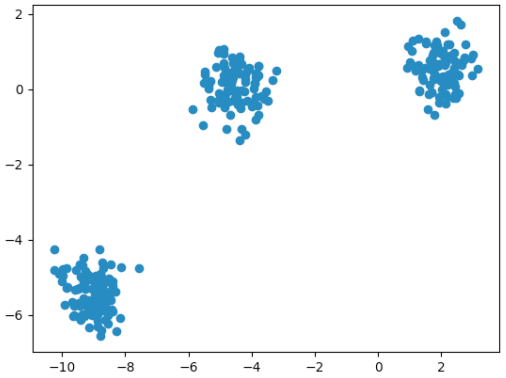


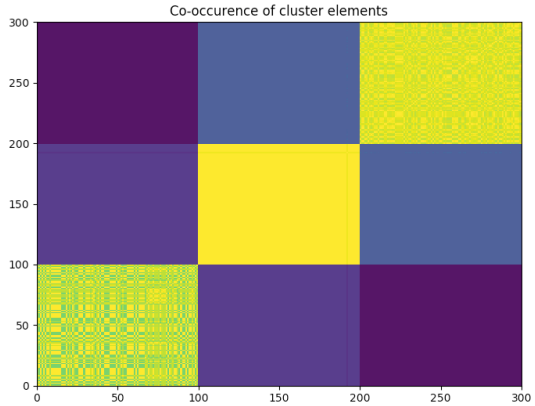
# Co-occurrence



# 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