CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

MODULE 3: CLUSTERING

DAT405, 2019-2020, READING PERIOD 1

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

Some clustering methods

- K-means clustering
- Density-based clustering (DBSCAN)
- Hierarchical clustering

DBSCAN

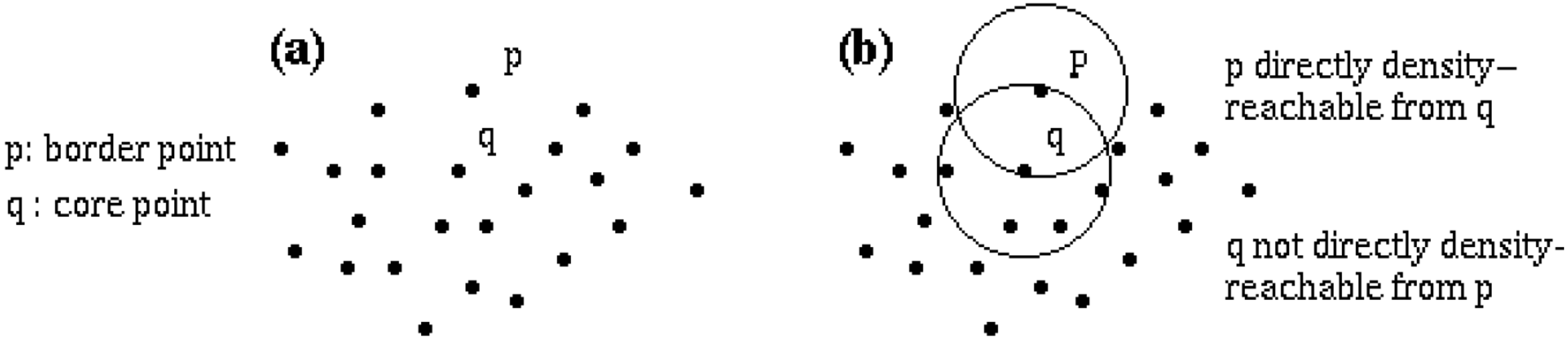
- Density-Based Spatial Clustering of Applications with Noise

- minimum number of neighbours
    - choose the minimum number of samples in the neighbourhood for a point to be considered as a core point

- radius of neighbourhood
    - choose maximum distance between two samples belonging to the same neighbourhood (“eps” or “epsilon”)

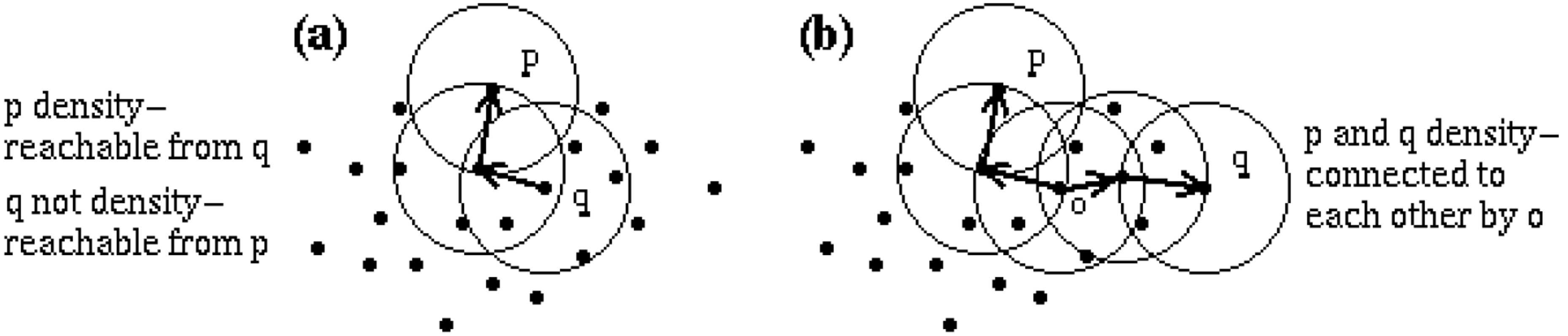
- choose distance metric

Core points and border points



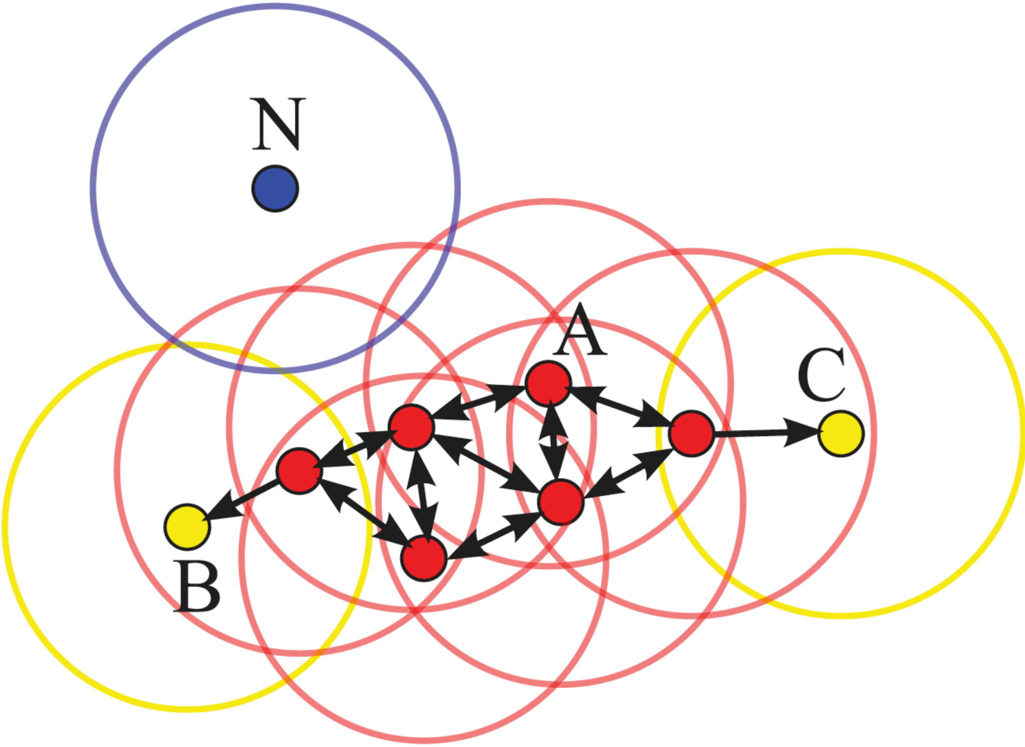
Ester, Kriegel, Sander, Xu (1996), In Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, (KDD), AAAI Press, pp. 226–231

Density-reachability and density-connectivity



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

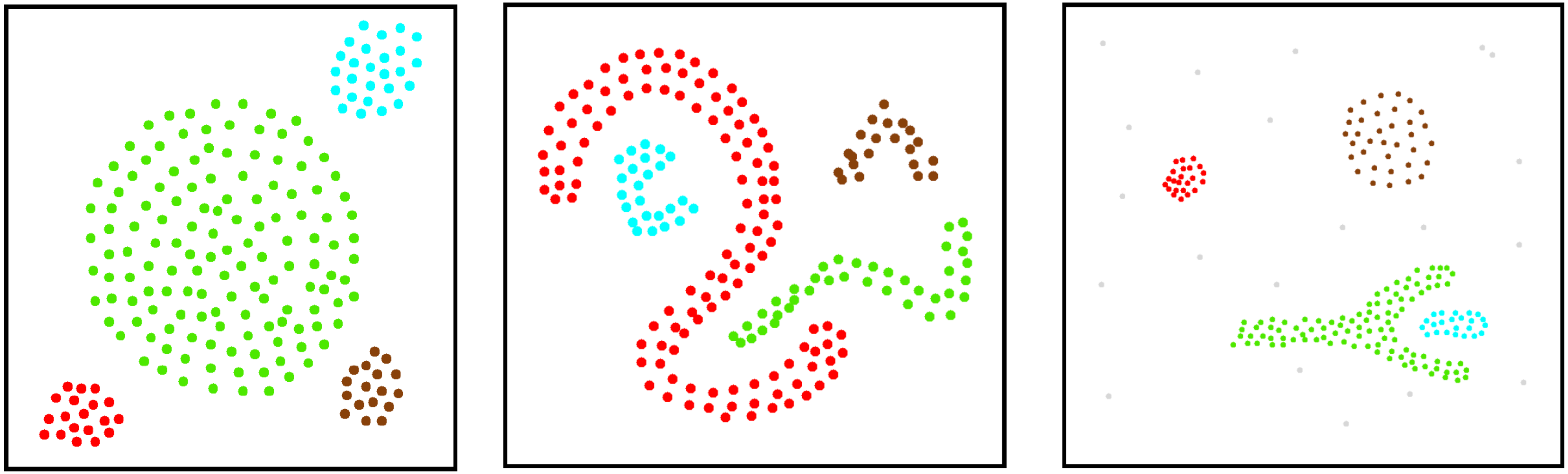
DBSCAN cluster model



minPts = 4

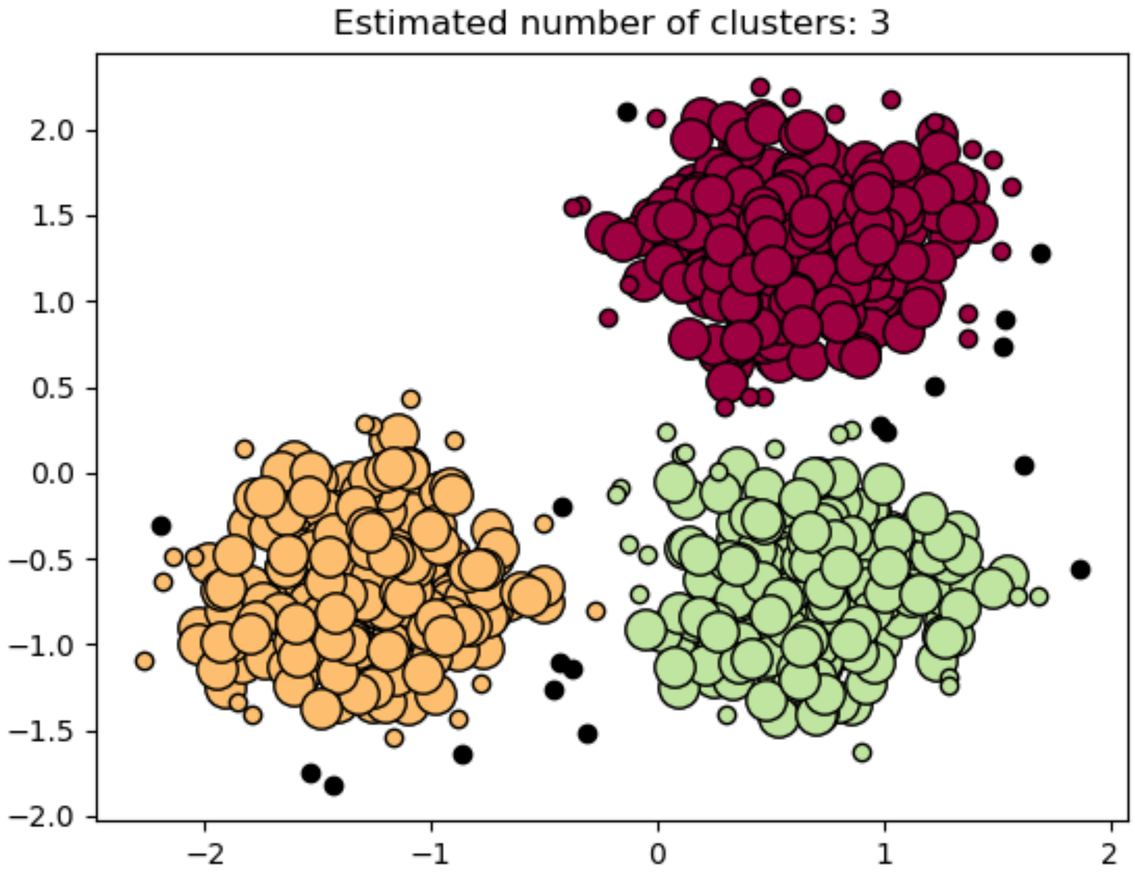
Schubert, Sander, Ester, Kriegel, Xu (2017), ACM Trans. Database Syst. 42, 3, Article 19

Clusterings discovered 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

- Sometimes called agglomerative clustering, when done bottom-up.
- Start with each observation in a cluster of its own, then successively merge the clusters together.
- Various algorithms to choose from, e.g.
  - Ward
  - complete linkage
  - average linkage
  - single linkage
  - neighbour-joining
  - Unweighted Pair Group Method with Arithmetic Mean (UPGMA)
  - Weighted Pair Group Method with Arithmetic Mean (WPGMA)

Neighbour-joining

- Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, *4*(4), 406-425 (over 55,000 citations)
- If input distance matrix is correct, output tree will be correct.
- Doesn’t assume same rate of evolution in all lineages.
- OTUs = “Operational taxonomic units”

Neighbour-joining

- Provide a distance matrix as input
- Start with a completely unresolved star network
- Repeat
    1. Calculate Q matrix
    2. Find entry in Q matrix with lowest value – pair of neighbours (OTUs) to join
    3. Calculate distance from each of joined pair to the new node
    4. Calculate distance from each node outside this pair to the new node
- Until tree is completely resolved and all branch lengths are known

Neighbour-joining

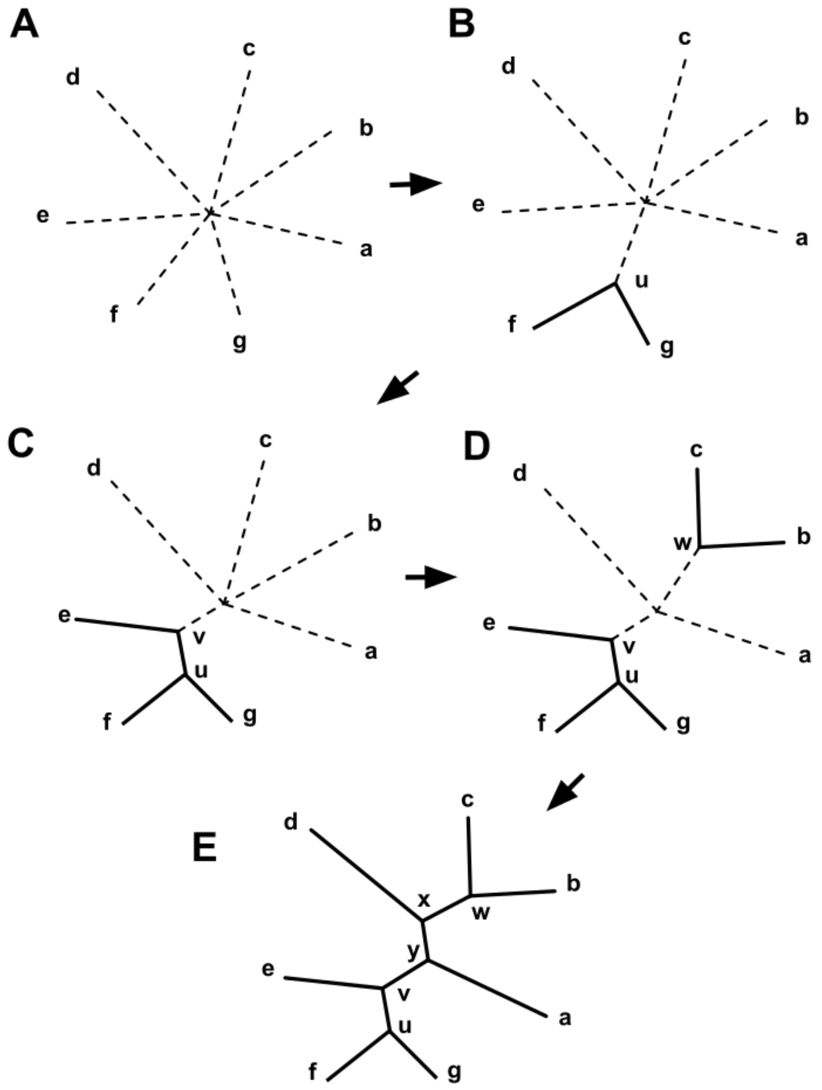
A. Start with a star tree

B. Join f and g, new node is u

C. Join u and e, new node is v

D. Join b and c, new node is w

E. Eventually the tree is fully resolved



https://en.wikipedia.org/wiki/Neighbor_joining#/media/File:Neighbor_joining_7_taxa_start_to_finish_diagram.svg

The Q matrix

$$Q(i,j) = (n-2)d(i,j) - \sum_{k=1}^{n} d(i,k) - \sum_{k=1}^{n} d(j,k)$$

where d(i,j) is the distance between i and j in the distance matrix and n nodes remain to be clustered

Distance from the pair members to new node

$$d(f,u) = \frac{1}{2}d(f,g) + \frac{1}{2(n-2)}\left[\sum_{k=1}^{n} d(f,k) - \sum_{k=1}^{n} d(g,k)\right]$$

and

$$d(g,u) = d(f,g) - d(f,u)$$

where f and g are the pair of neighbouring OTUs being joined together, and u is the new node.

Distance of other OTUs to the new node

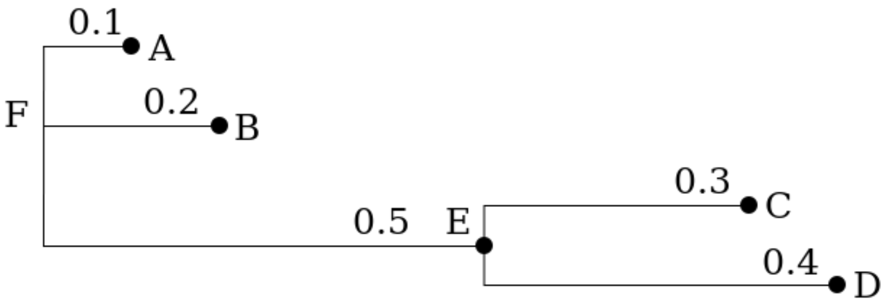
$$d(u,k) = \frac{1}{2}\left[d(f,k) + d(g,k) - d(f,g)\right]$$

where u is the new node, k is the node to which we want to calculate the distance, and f and g are the pair of neighbouring OTUs that have just been joined.

On-line demo of agglomerative clustering

http://rna.informatik.uni-freiburg.de/Teaching/index.jsp?toolName=Agglomerative%20Clustering

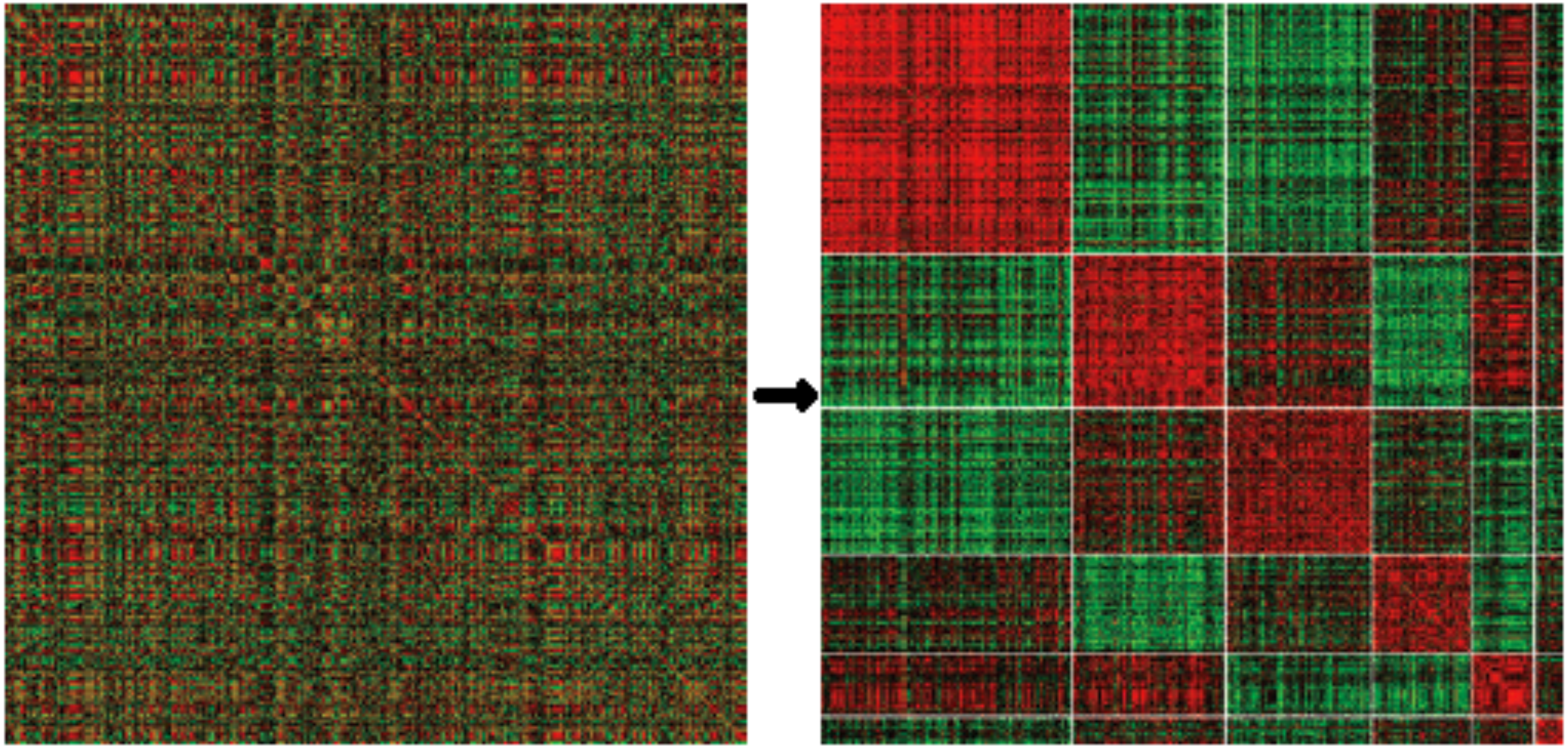
Newick format

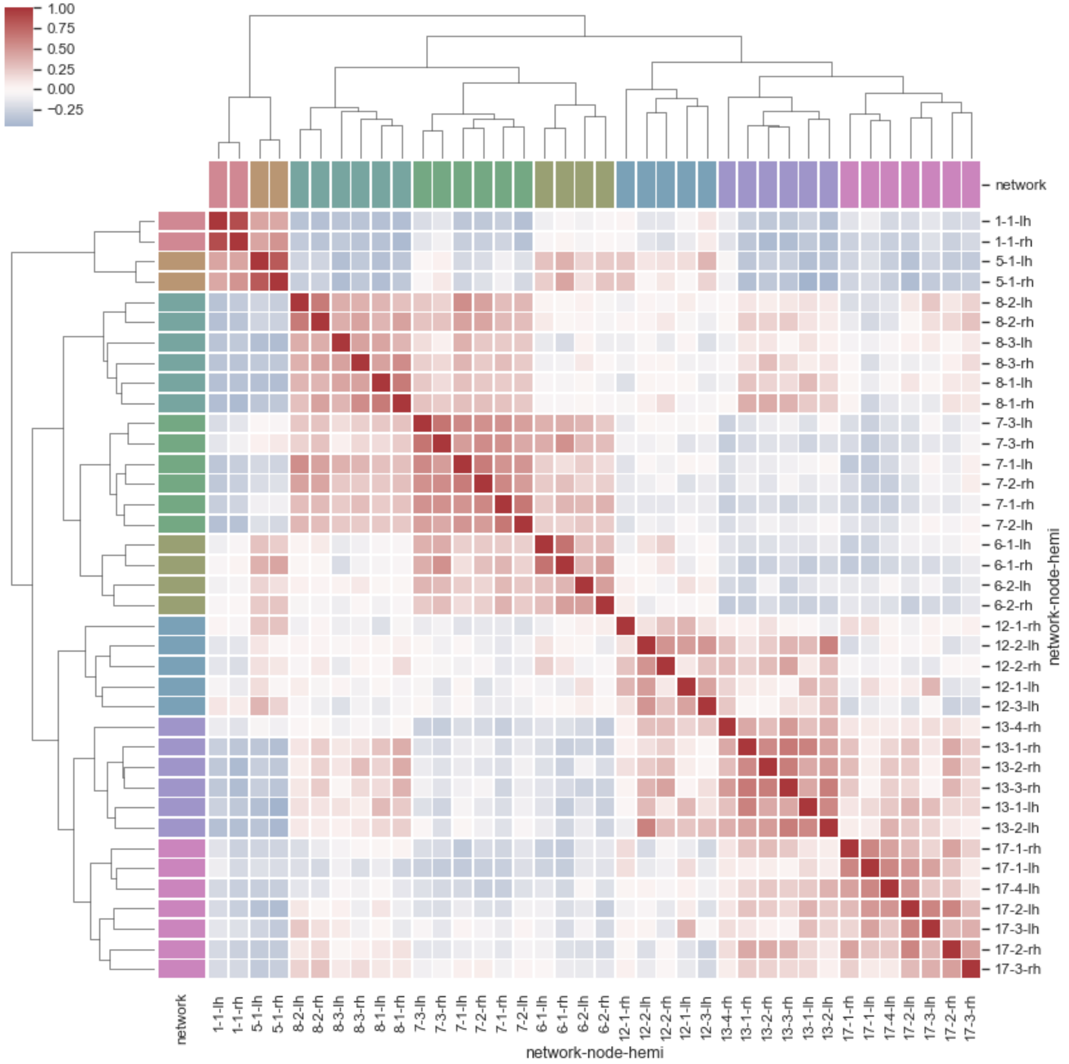


- A standard for representing trees
- Examples:

`(,,(,));`	no nodes are named
`(A,B,(C,D));`	leaf nodes are named
`(A,B,(C,D)E)F;`	all nodes are named
`(:0.1,:0.2,(:0.3,:0.4):0.5);`	all but root node have a distance to parent
`(:0.1,:0.2,(:0.3,:0.4):0.5):0.0;`	all have a distance to parent
`(A:0.1,B:0.2,(C:0.3,D:0.4):0.5);`	distances and leaf names (popular)
`(A:0.1,B:0.2,(C:0.3,D:0.4)E:0.5)F;`	distances and all names
`((B:0.2,(C:0.3,D:0.4)E:0.5)A:0.1)F;`	a tree rooted on a leaf node (rare)

Biclustering





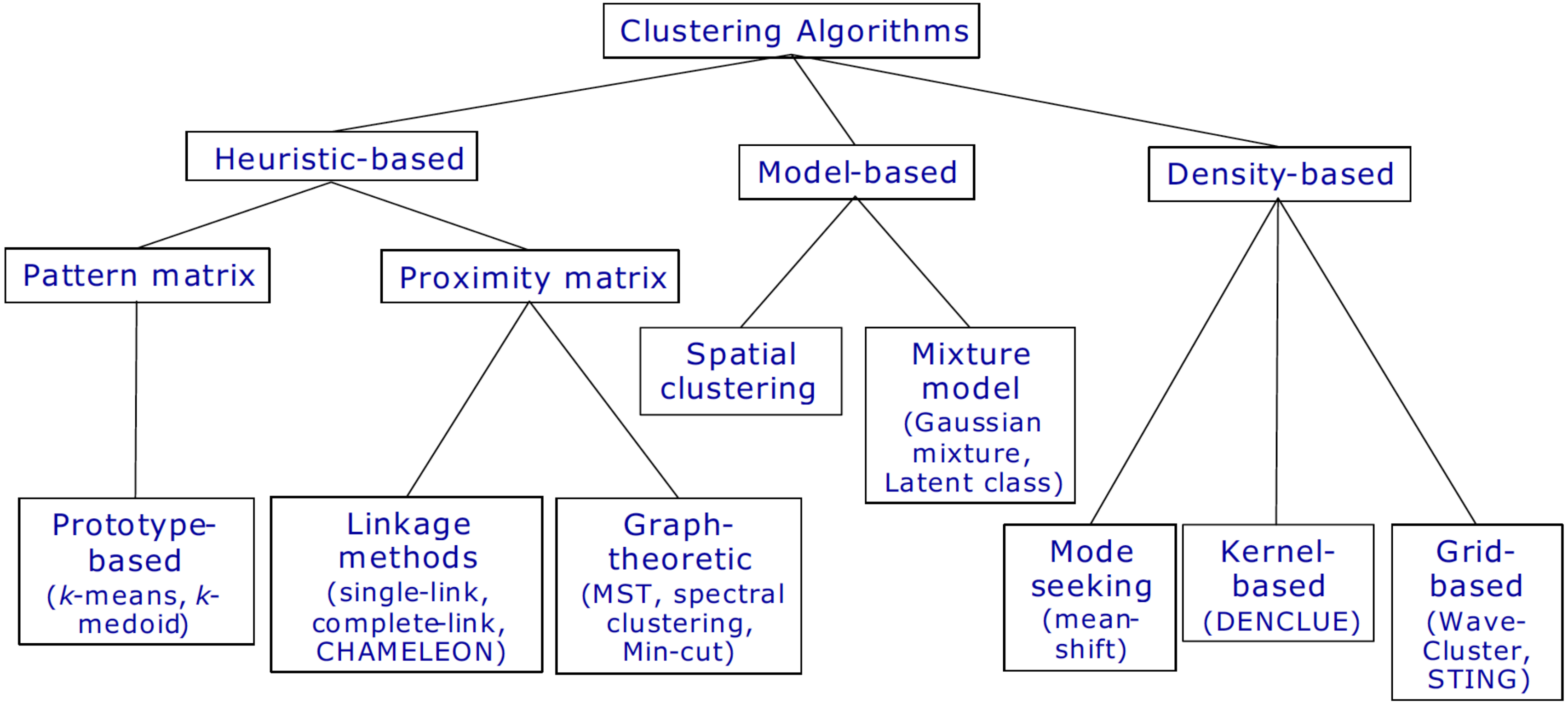
https://seaborn.pydata.org/examples/structured_heatmap.html

Biclustering

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

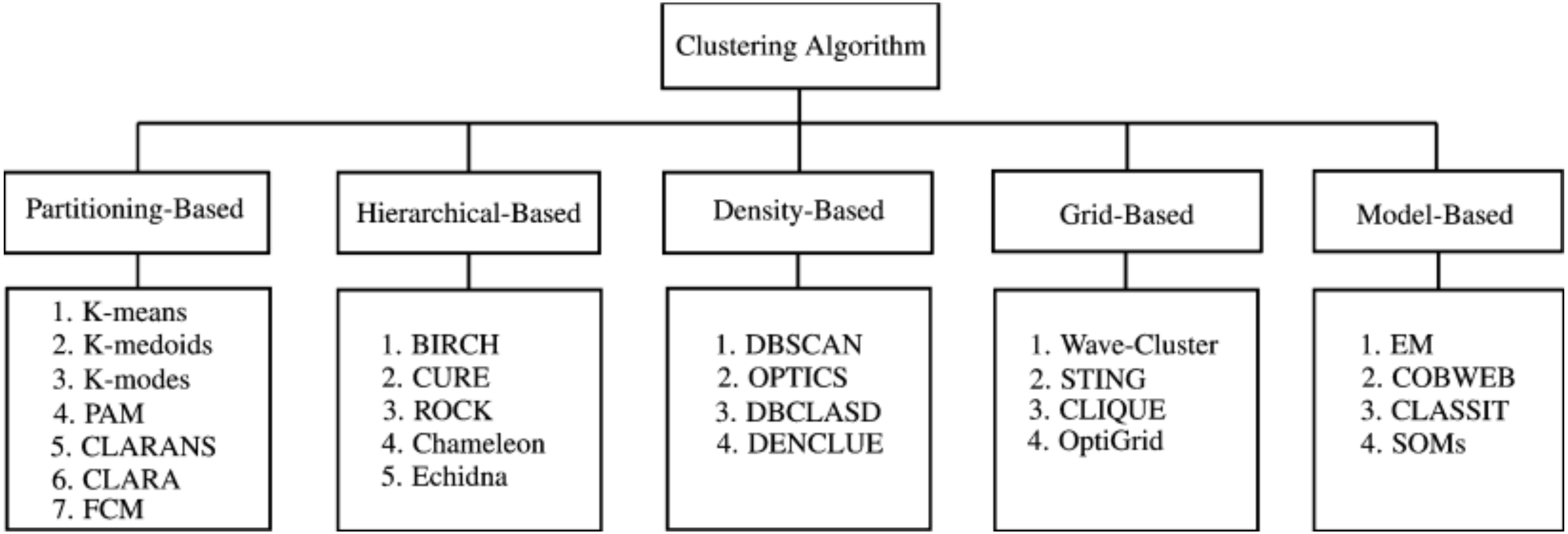
- https://scikit-learn.org/stable/modules/biclustering.html

Clustering clustering algorithms


Clustering Algorithms
Heuristic-based
Model-based
Density-based
Pattern matrix
Proximity matrix
Spatial clustering
Mixture model (Gaussian mixture, Latent class)
Prototype-based (*k*-means, *k*-medoid)
Linkage methods (single-link, complete-link, CHAMELEON)
Graph-theoretic (MST, spectral clustering, Min-cut)
Mode seeking (mean-shift)
Kernel-based (DENCLUE)
Grid-based (Wave-Cluster, STING)

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

Clustering clustering algorithms



Clustering Algorithm

Partitioning-Based

1. K-means
2. K-medoids
3. K-modes
4. PAM
5. CLARANS
6. CLARA
7. FCM

Hierarchical-Based

1. BIRCH
2. CURE
3. ROCK
4. Chameleon
5. Echidna

Density-Based

1. DBSCAN
2. OPTICS
3. DBCLASD
4. DENCLUE

Grid-Based

1. Wave-Cluster
2. STING
3. CLIQUE
4. OptiGrid

Model-Based

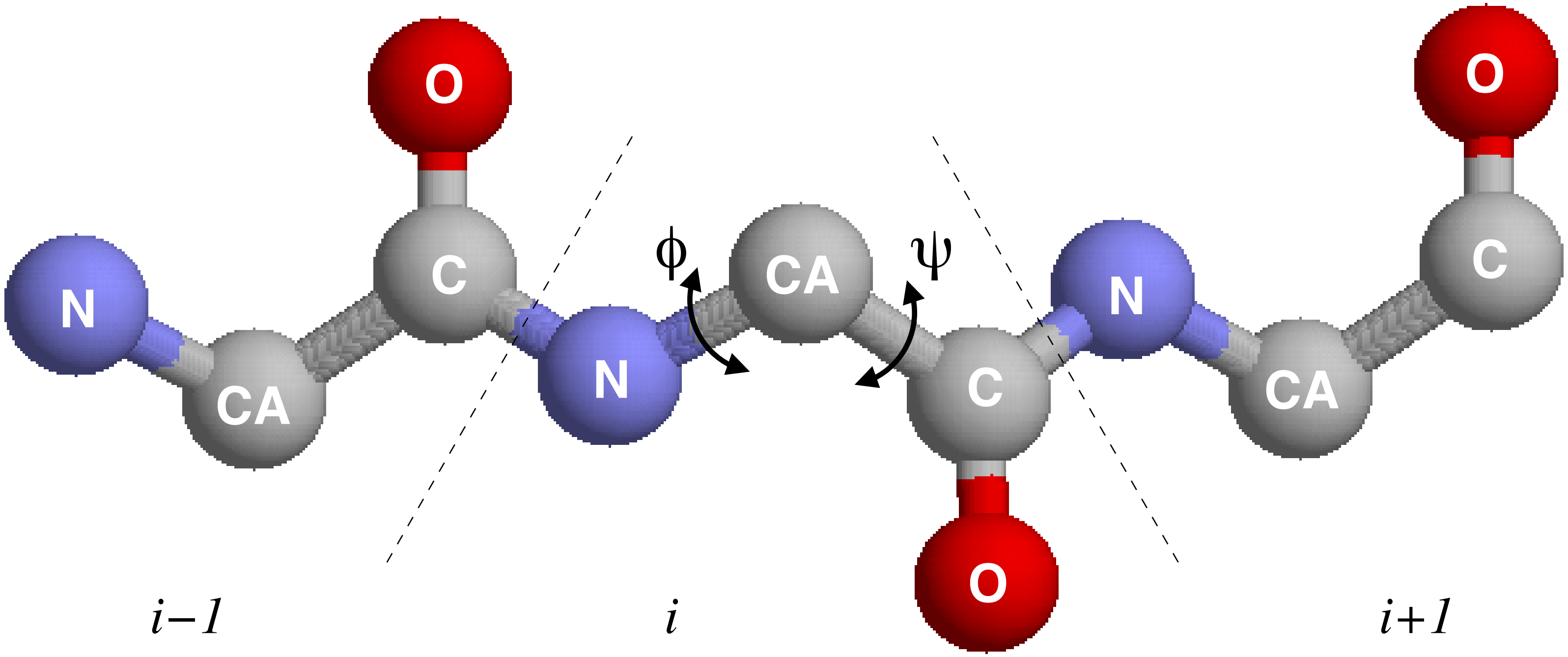
1. EM
2. COBWEB
3. CLASSIT
4. SOMs

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

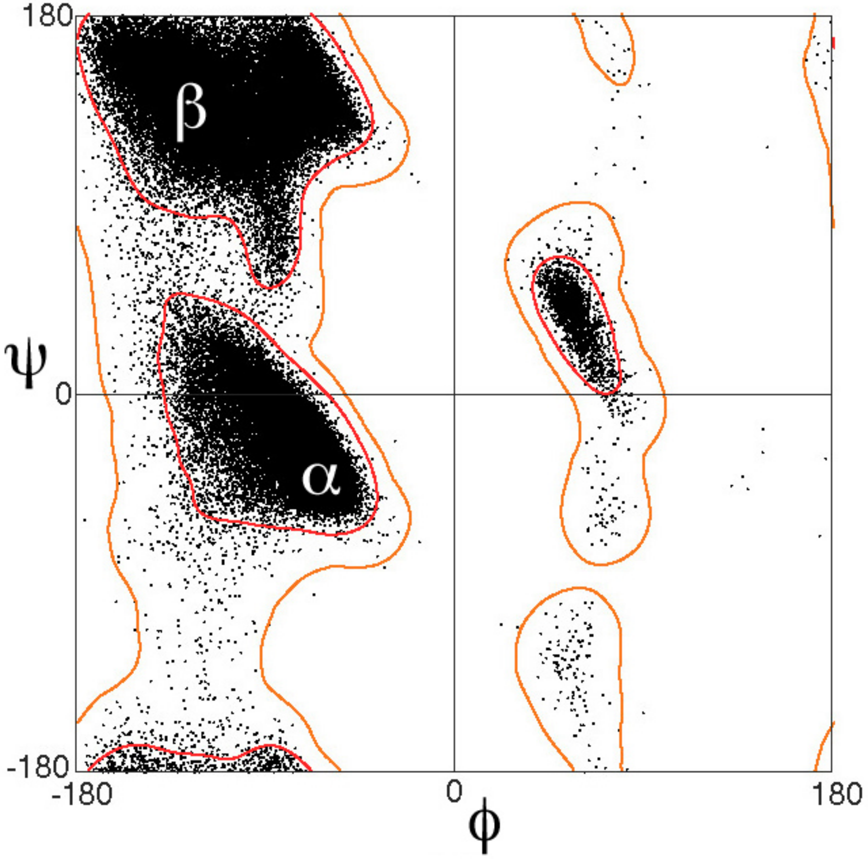
Assignment 3

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

Protein main chain

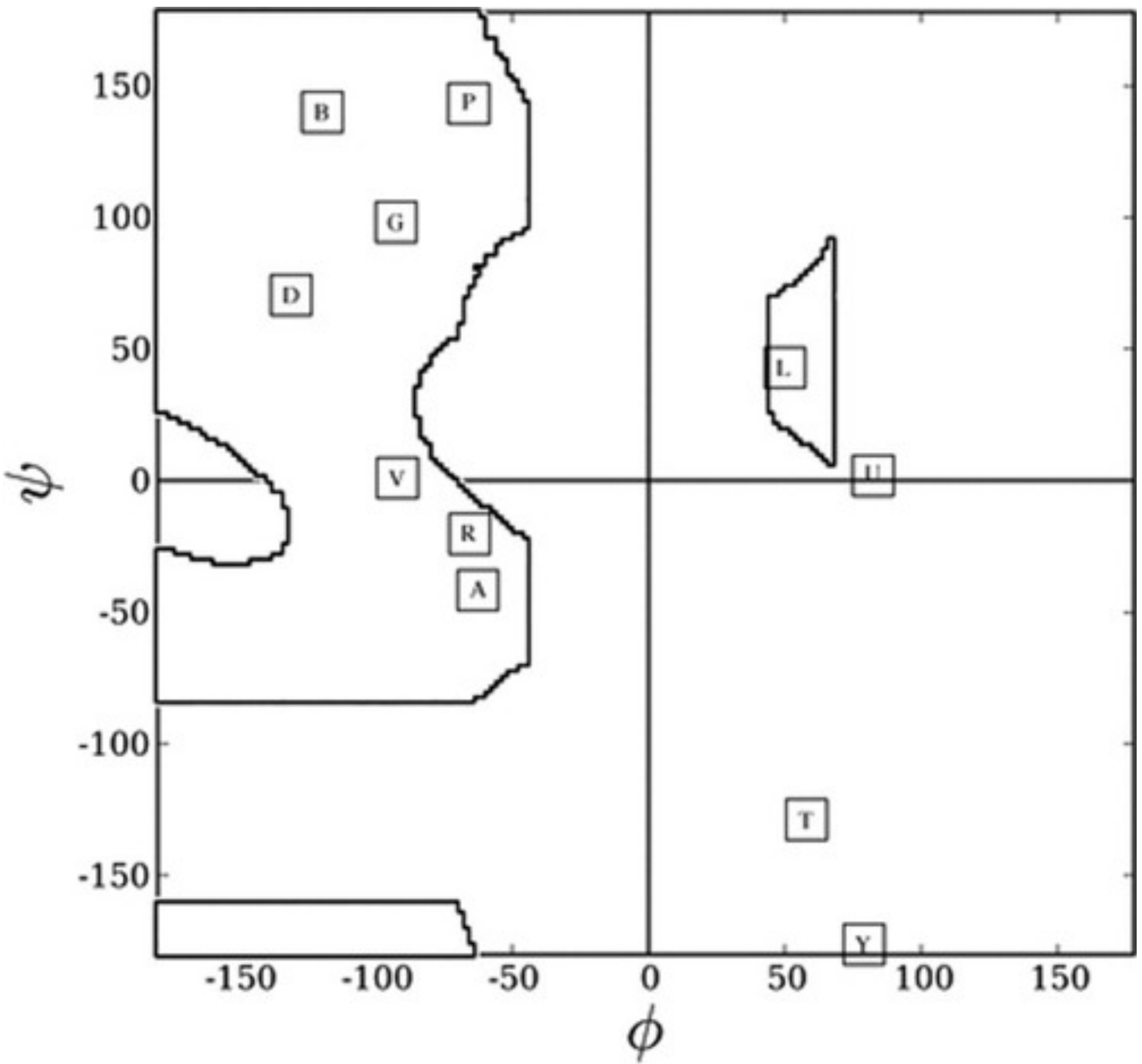


Ramachandran plot



Around 100000 data points shown here

http://bioinformatics.org/molvis/phipsi/



Chellapa and Rose (2012). *Protein Science*, **21**, 1231-1240

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

Assignment 2

- A company is considering using a system that will allow management to track staff behaviour in real-time, gather data on when and who they email, who accesses and edits files, and who meets whom when. The HR (human resources, or personnel) department at the company is in favour of introducing this new system and believes it will benefit the staff at the company. The company's management is also in favour.
- Discuss whether introducing this system raises potential ethical issues.