University of Glasgow

Case study 1: Model selection for clustering

Ke Yuan

Ke.Yuan@glasgow.ac.uk

(Slides from Lucas.Farndale@glasgow.ac.uk)

University of Glasgow

Overview

- What is Model selection?
- Challenges in model selection for clustering
- Clustering algorithms
- Colorectal Tissue Biopsy and Clustering

- Your task and dataset details
- Expected results
- Summary

University of Glasgow

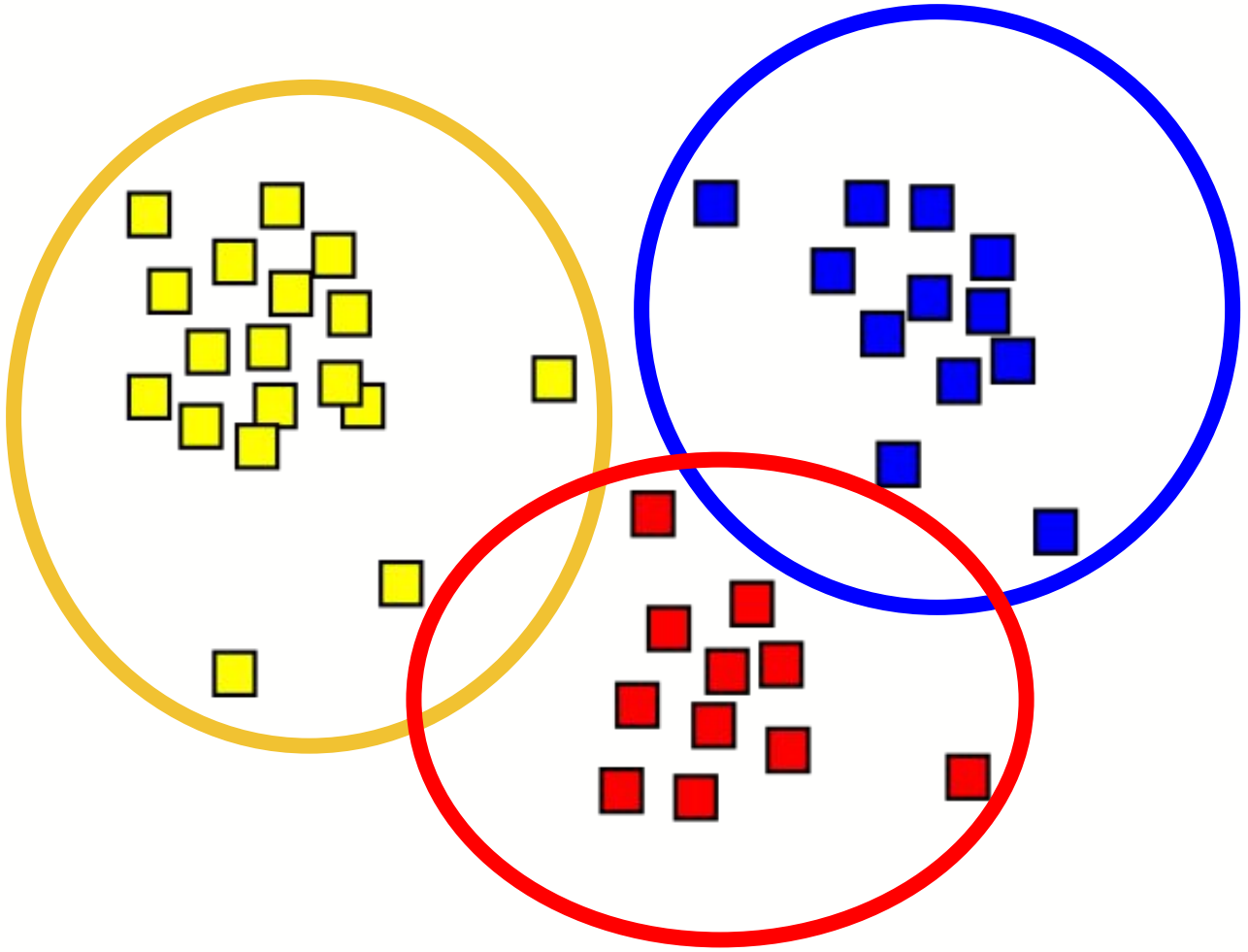
Model selection

- Choosing the best model candidate
  - family of algorithms (e.g. Logistic regression, KNN)
  - different hyperparameters (e.g. regularisation strength, number of neighbours)

- What is ‘the best model’?
  - Define objective (e.g. accuracy, minimise false positives, etc)
  - Complexity
  - Computability
  - Ease of Implementation

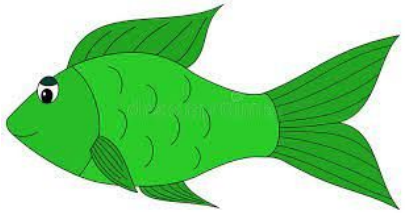
University of Glasgow

Why do we need model selection for clustering?



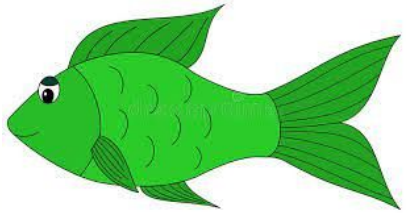
University of Glasgow

Why do we need model selection for clustering?



University of Glasgow

Why do we need model selection for clustering?

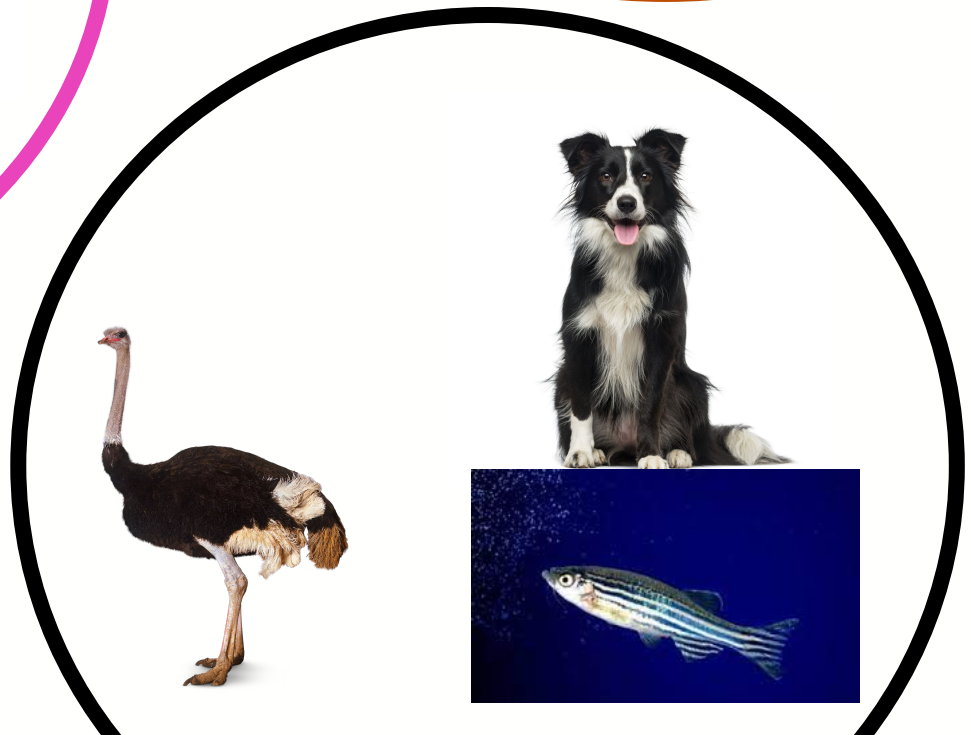


University of Glasgow

Why do we need model selection for clustering?

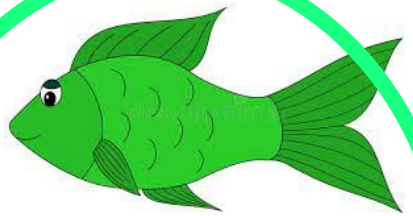
University of Glasgow

Why do we need model selection for clustering?



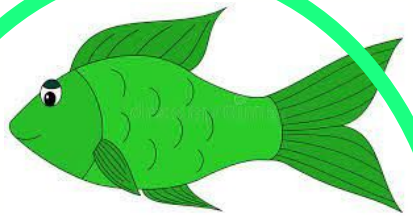
University of Glasgow

Why do we need model selection for clustering?



University of Glasgow

Why do we need model selection for clustering?

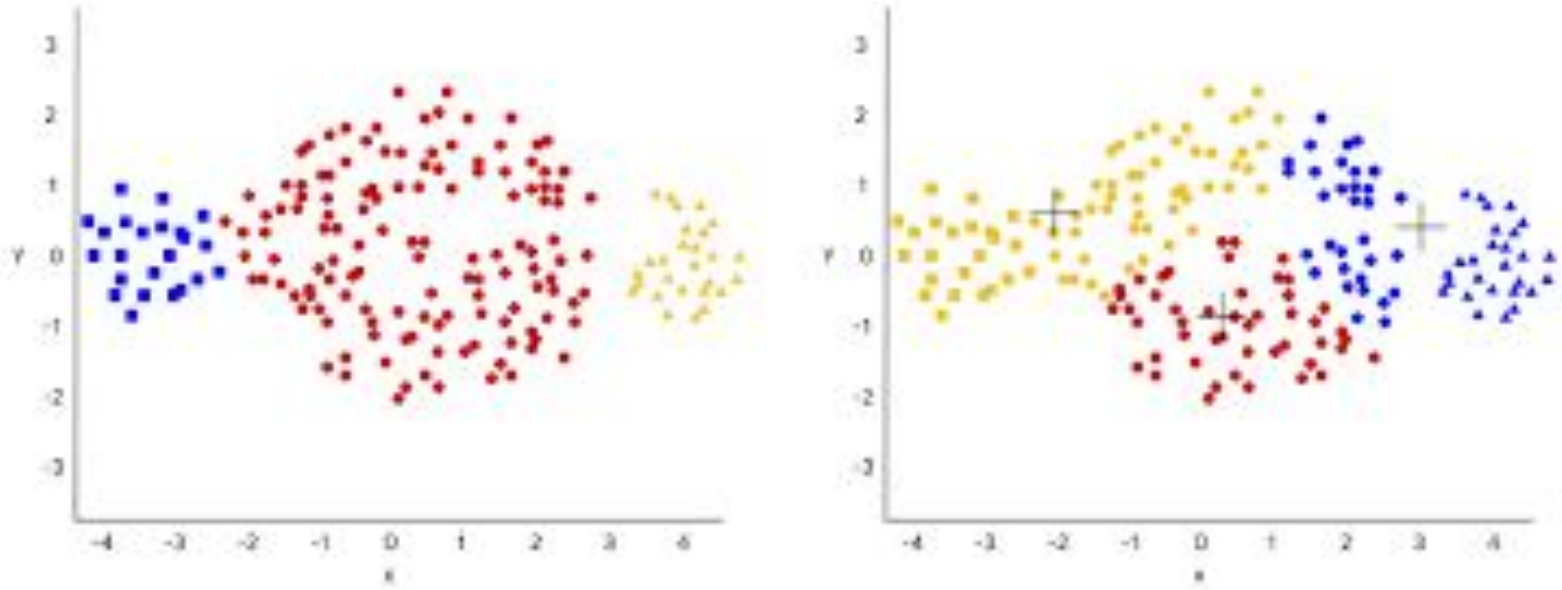


University of Glasgow

Why do we need model selection for clustering?

- Group similar objects together
- Constraints on clusters
- Understand the structure of a dataset

The best clustering model will best describe the structure of the data.



University of Glasgow

Examples of clustering algorithms

Cluster numbers can be explicitly specified:

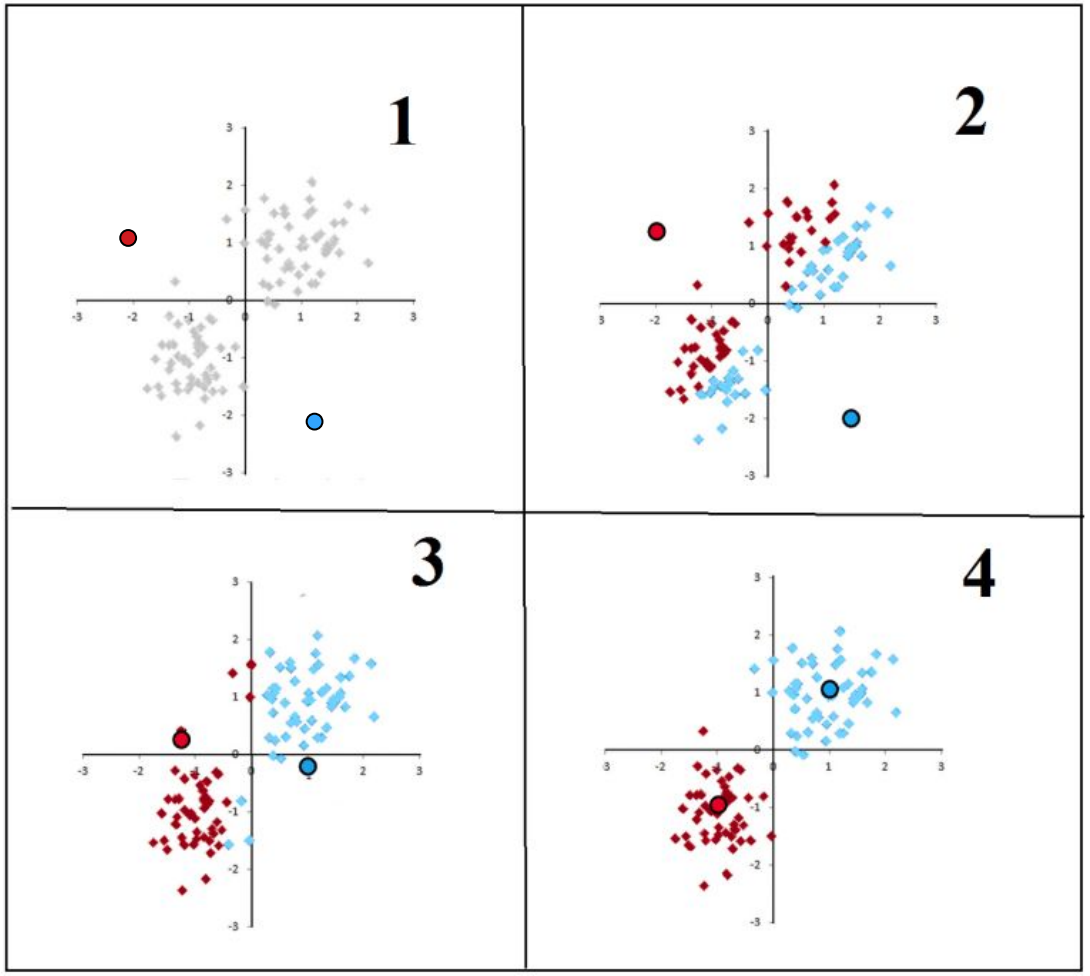
1. K-means
2. Gaussian Mixture Model (GMM)

Or inferred:

1. Hierarchical clustering
2. Louvain Clustering

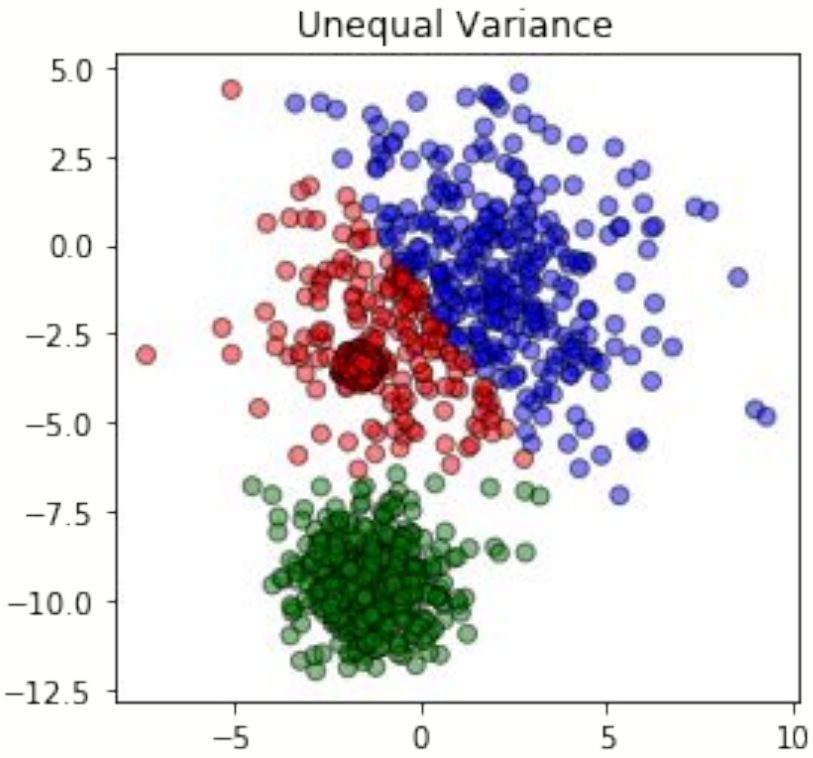
K-means

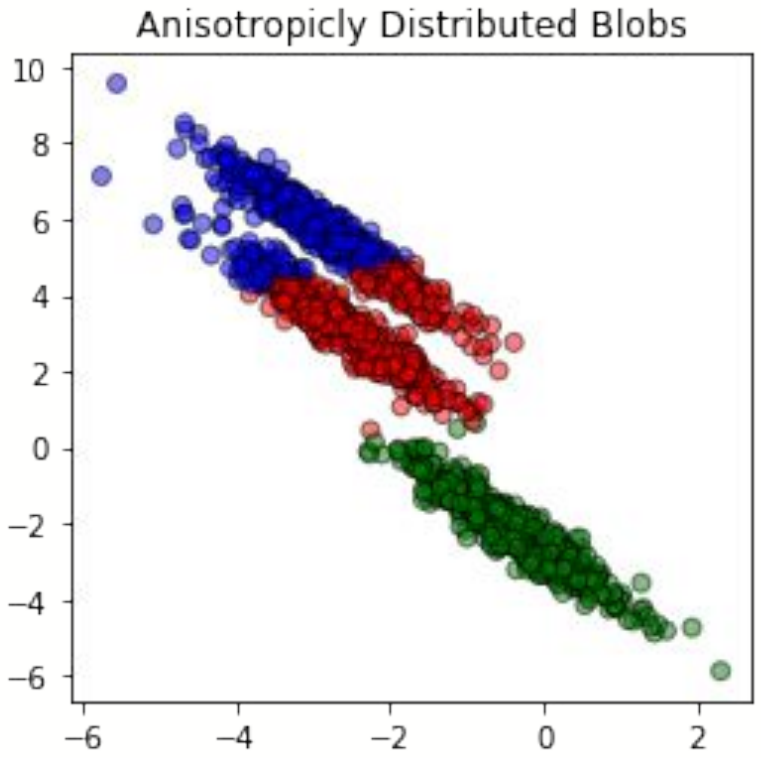
1. Choose *k* random points
2. Partition objects into k subset
3. Compute the new *centroids* (mean points) of the clusters
4. Repeat steps 2+3 until convergence

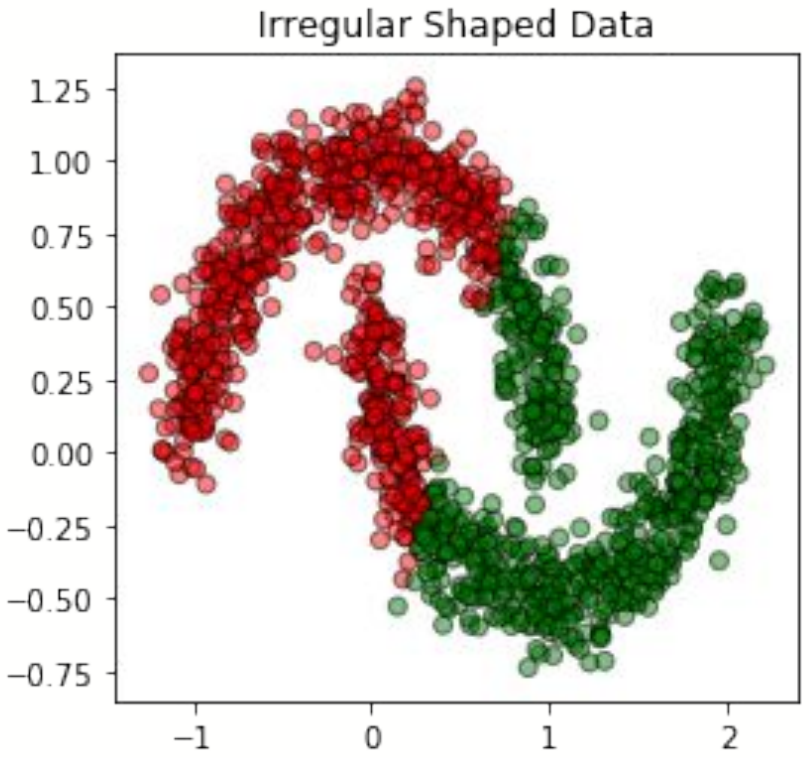


University of Glasgow

K-means







Gaussian Mixture Model (GMM)

- k-means only considers mean points
- GMM considers mean and (co-)variance
- Fit *M* Gaussian components by maximising log-likelihood

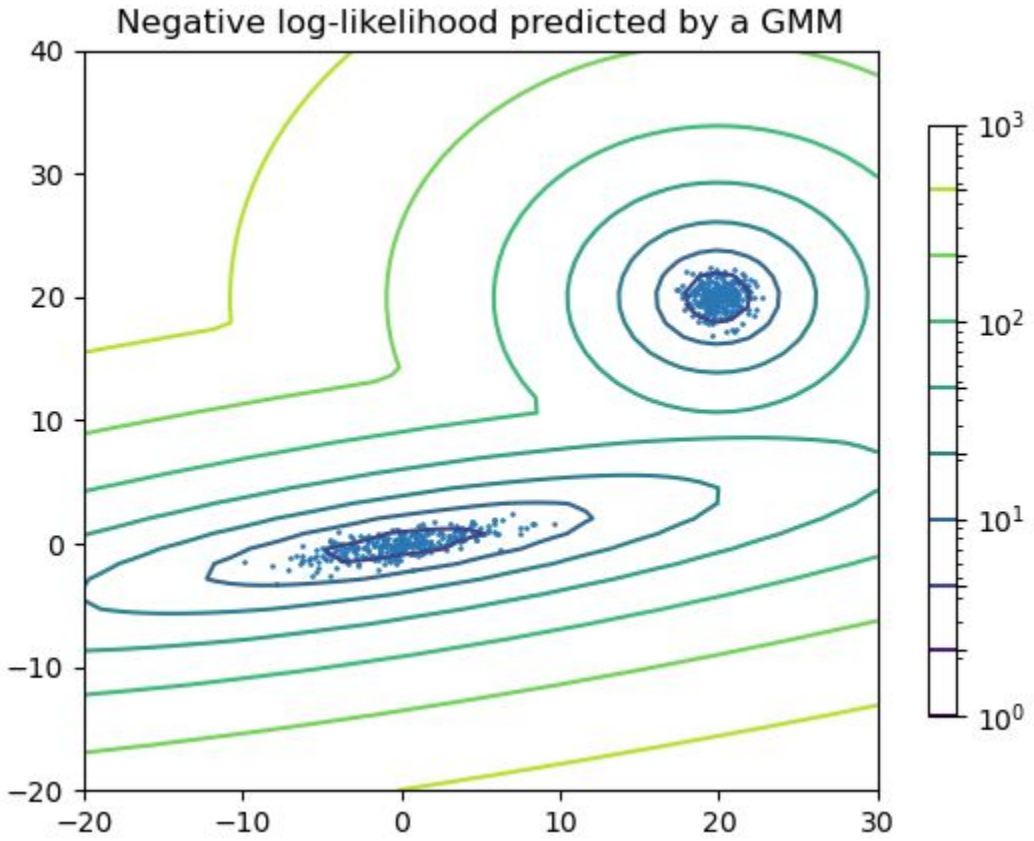
$$\ell(\theta|\mathbf{x}) = \sum_{i=1}^{N} \log\left(\sum_{m=1}^{M} \alpha_m \phi(x_i|\mu_m, \Sigma_m)\right)$$

x – datapoint

θ - gaussian parameters $\{(\mu_m, \Sigma_m): 0 \leq m \leq M\}$

α - mixing coefficient

Φ – probability density function



University of Glasgow

Initial Random Guess

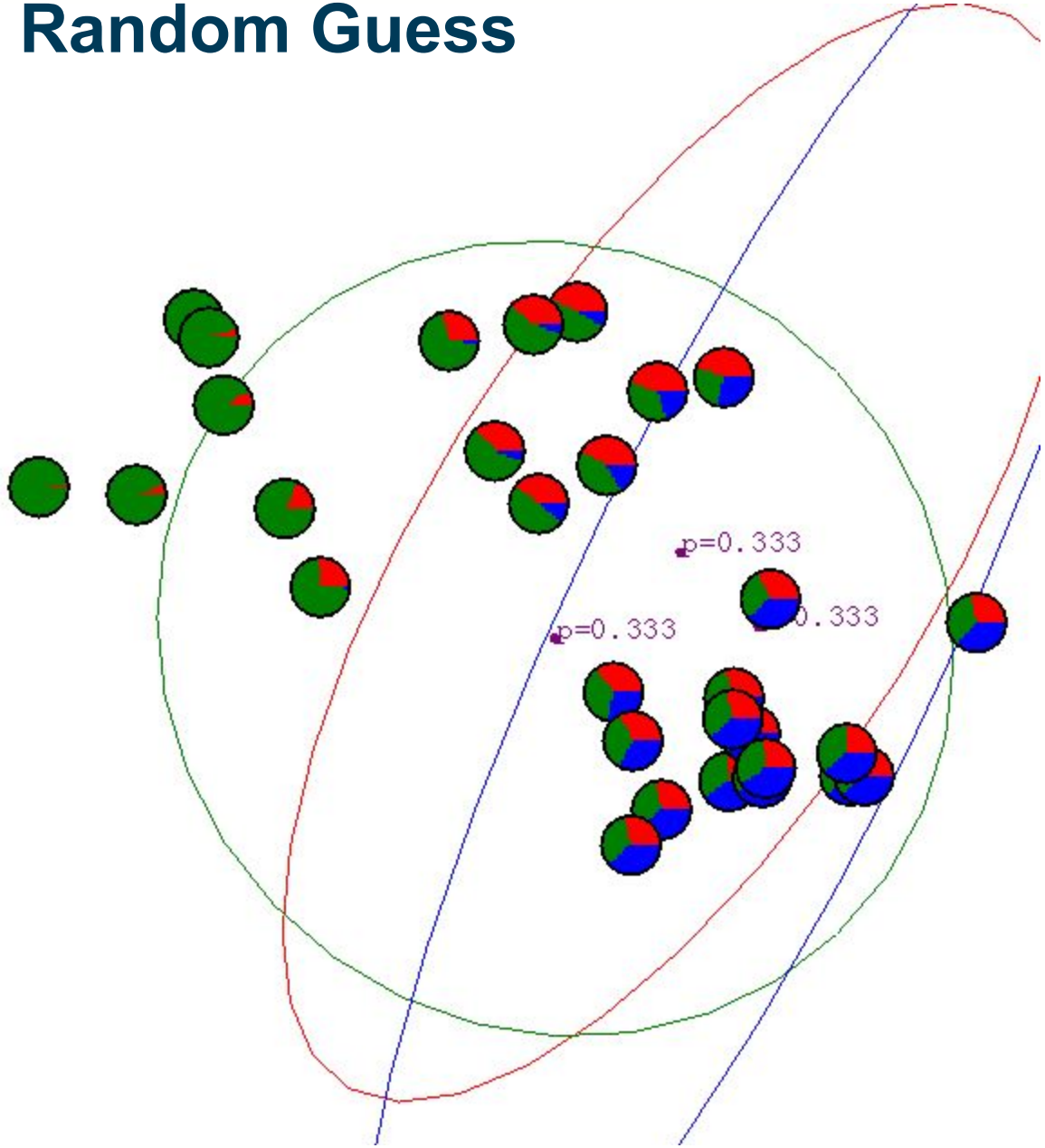


Image credit Prof A Moore, Carnegie Mellon University

1st Iteration

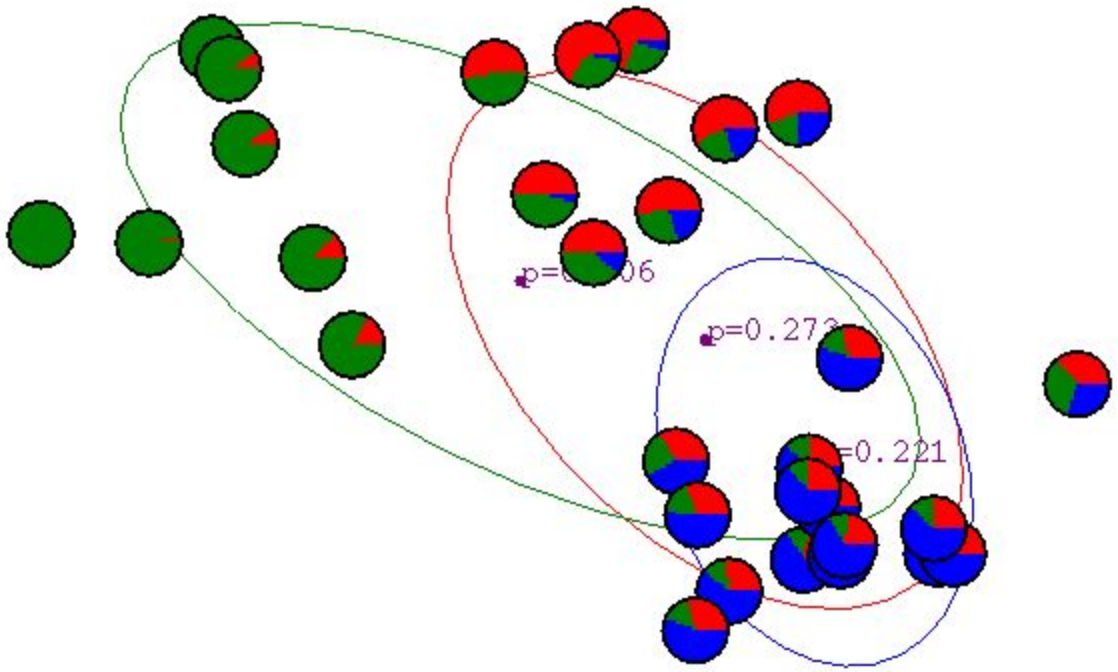


Image credit Prof A Moore, Carnegie Mellon University

2nd Iteration

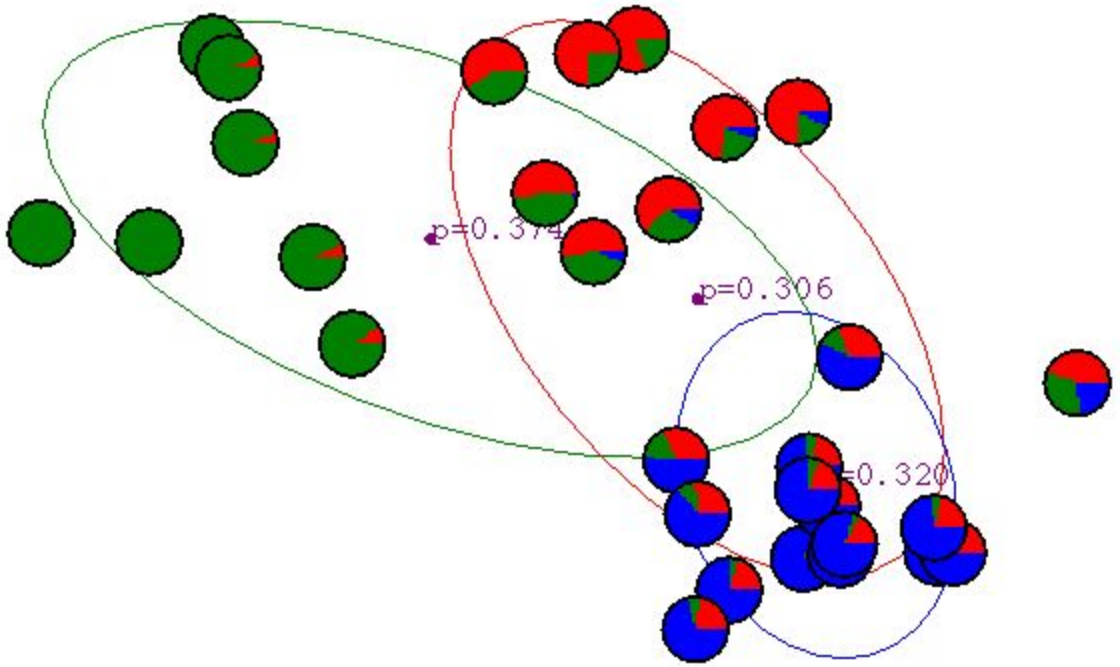


Image credit Prof A Moore, Carnegie Mellon University

3rd Iteration

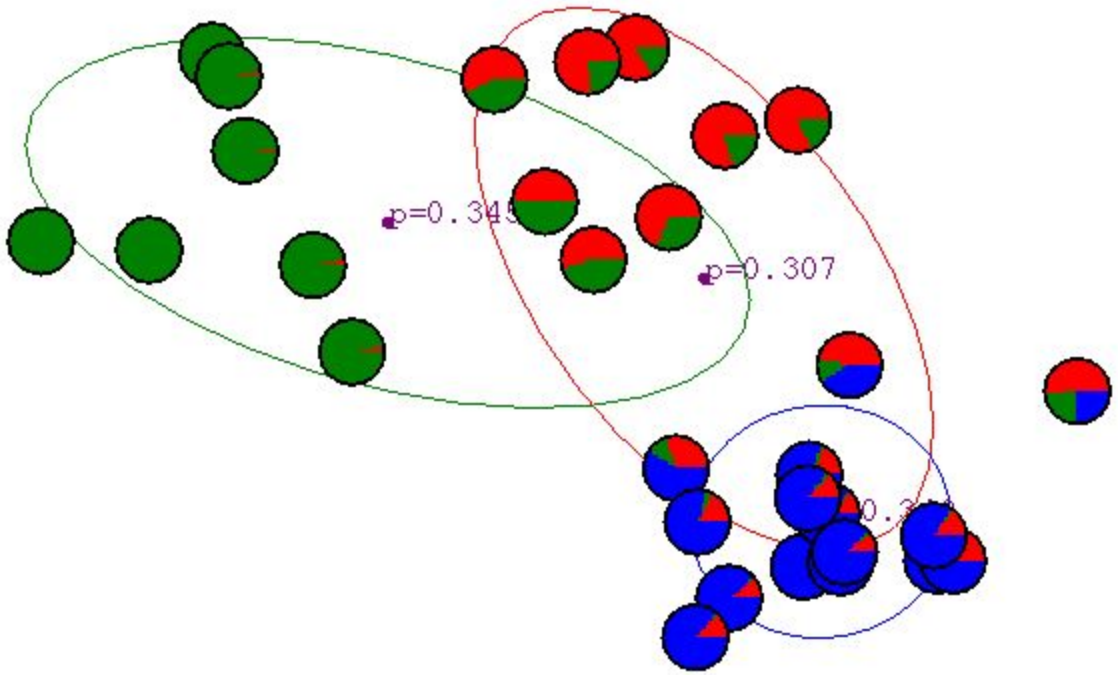


Image credit Prof A Moore, Carnegie Mellon University

4th Iteration

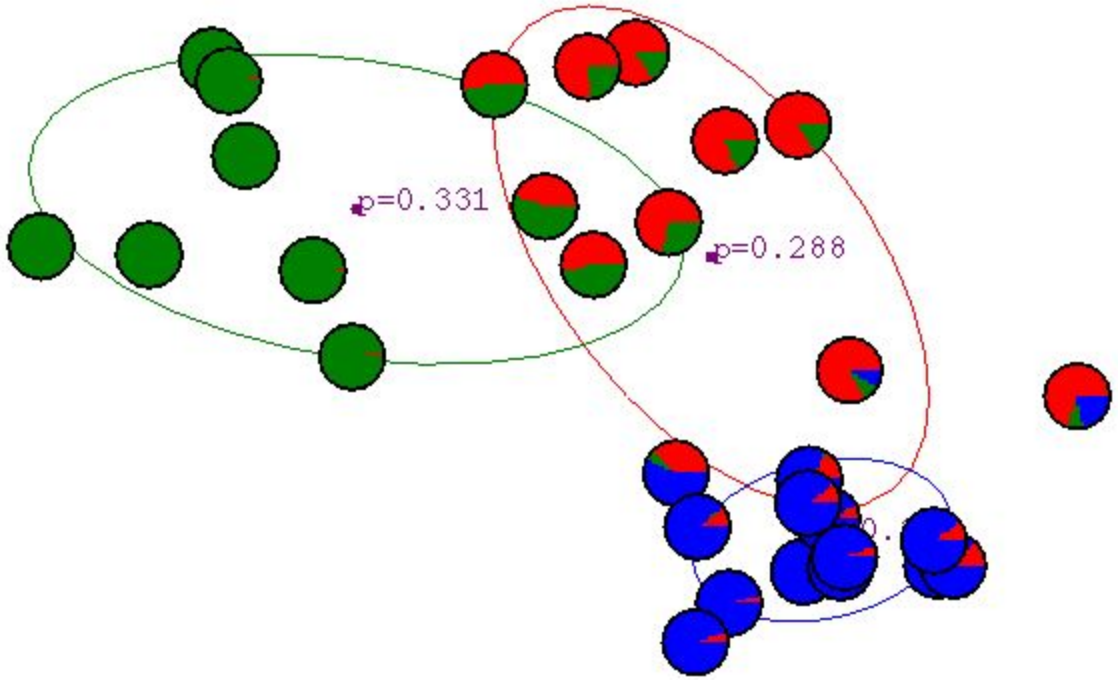


Image credit Prof A Moore, Carnegie Mellon University

5th Iteration

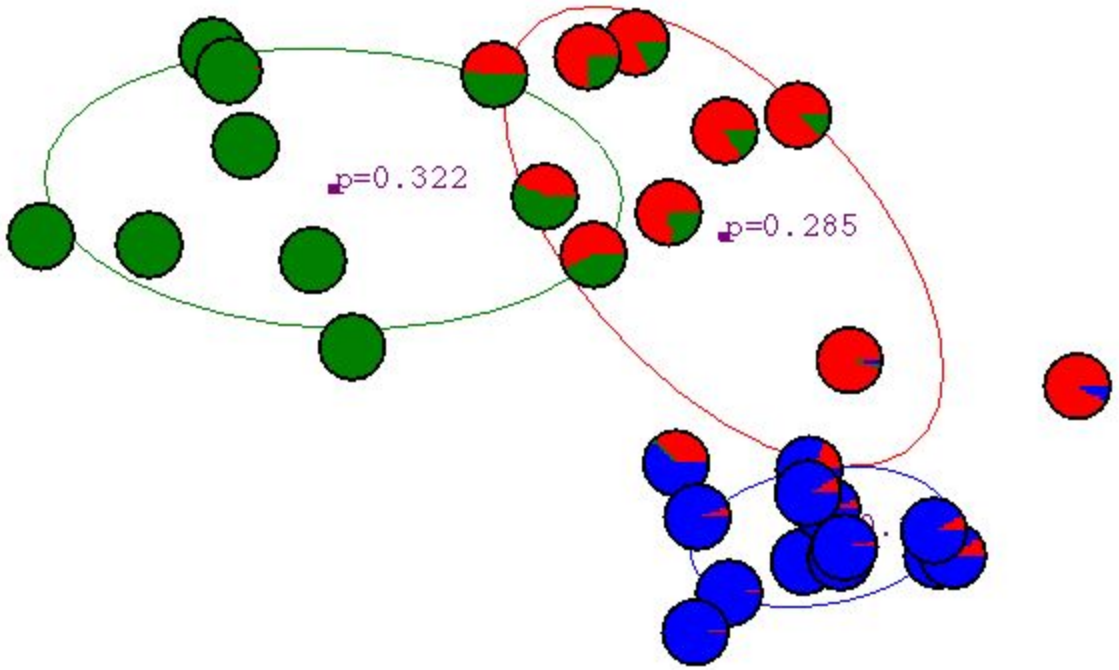


Image credit Prof A Moore, Carnegie Mellon University

6th Iteration…

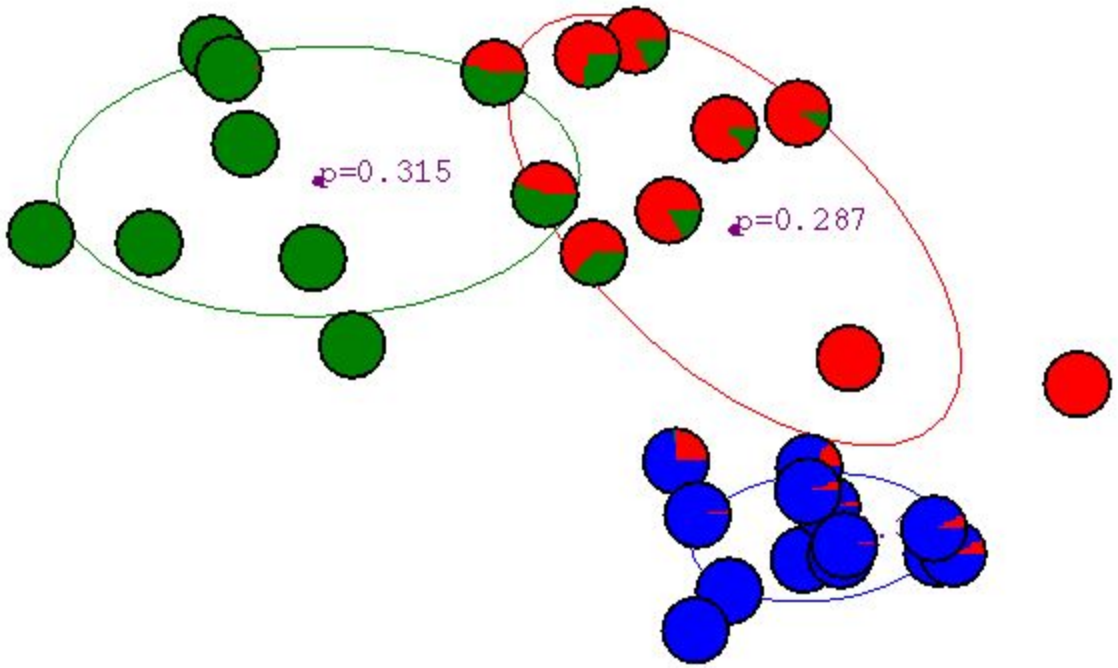


Image credit Prof A Moore, Carnegie Mellon University

…20th Iteration

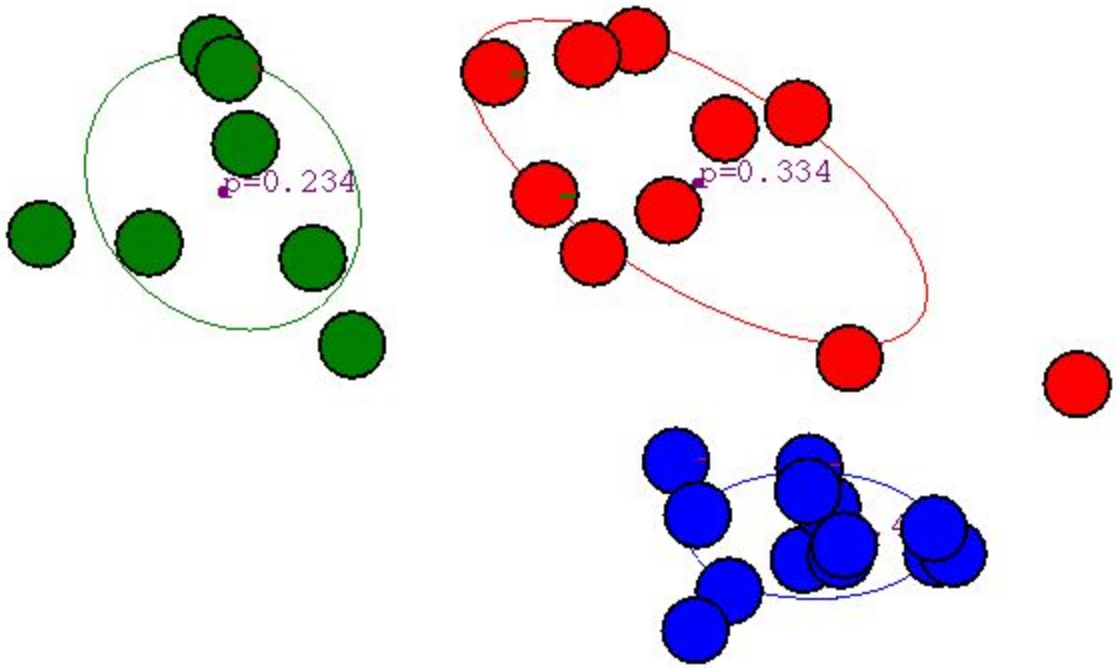


Image credit Prof A Moore, Carnegie Mellon University

University of Glasgow

Gaussian Mixture Model (GMM)

Bio Assay Data

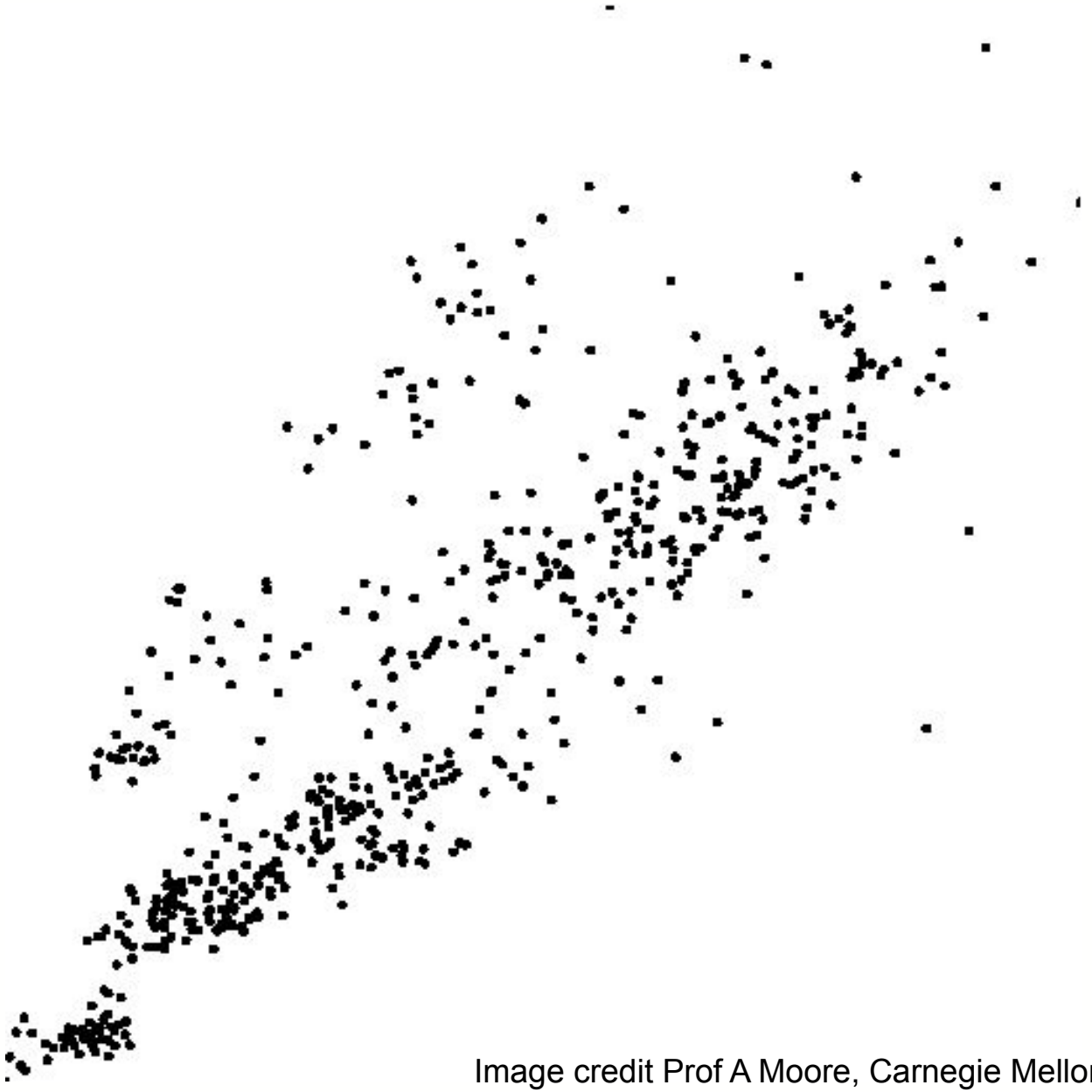
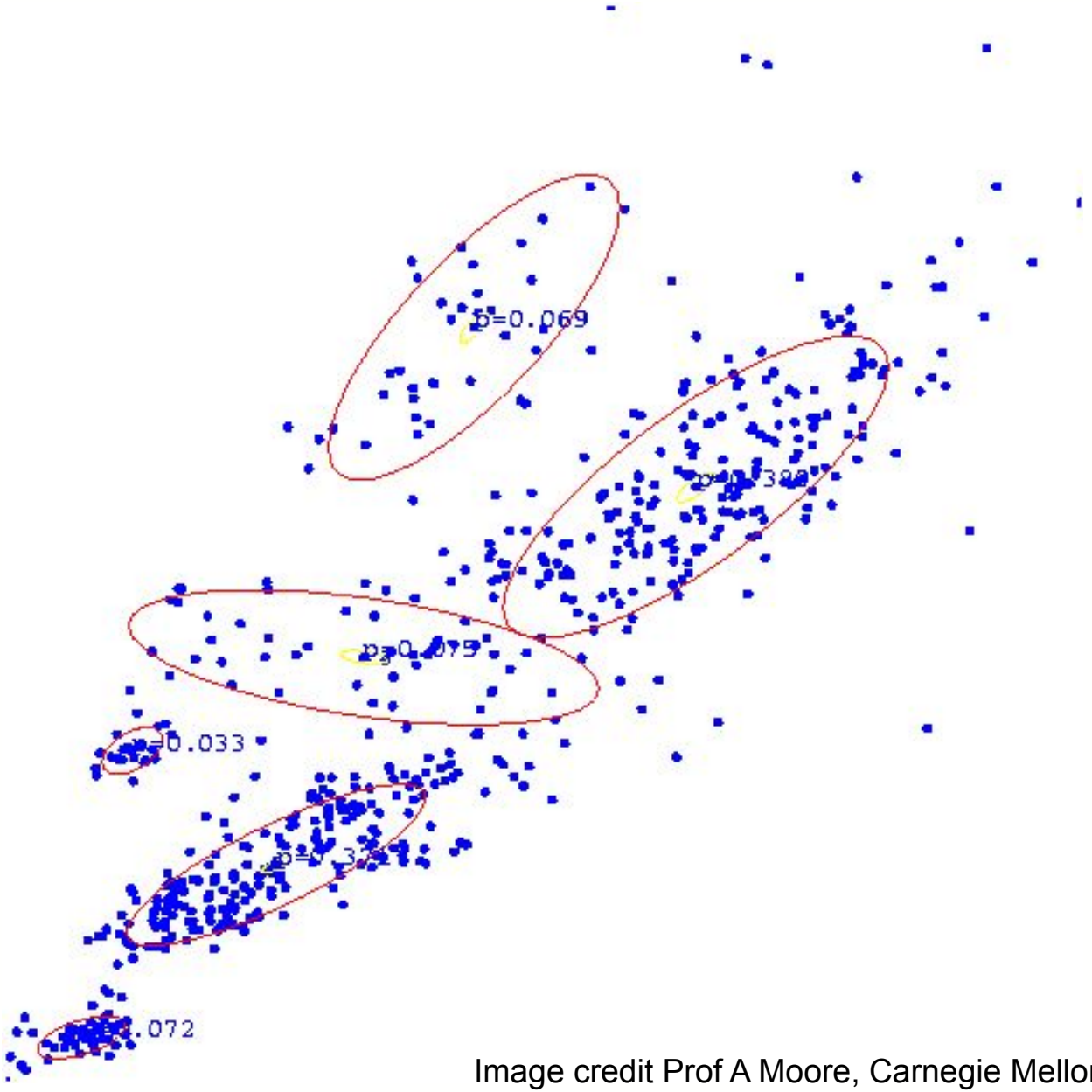


Image credit Prof A Moore, Carnegie Mellon University

Gaussian Mixture Model (GMM)

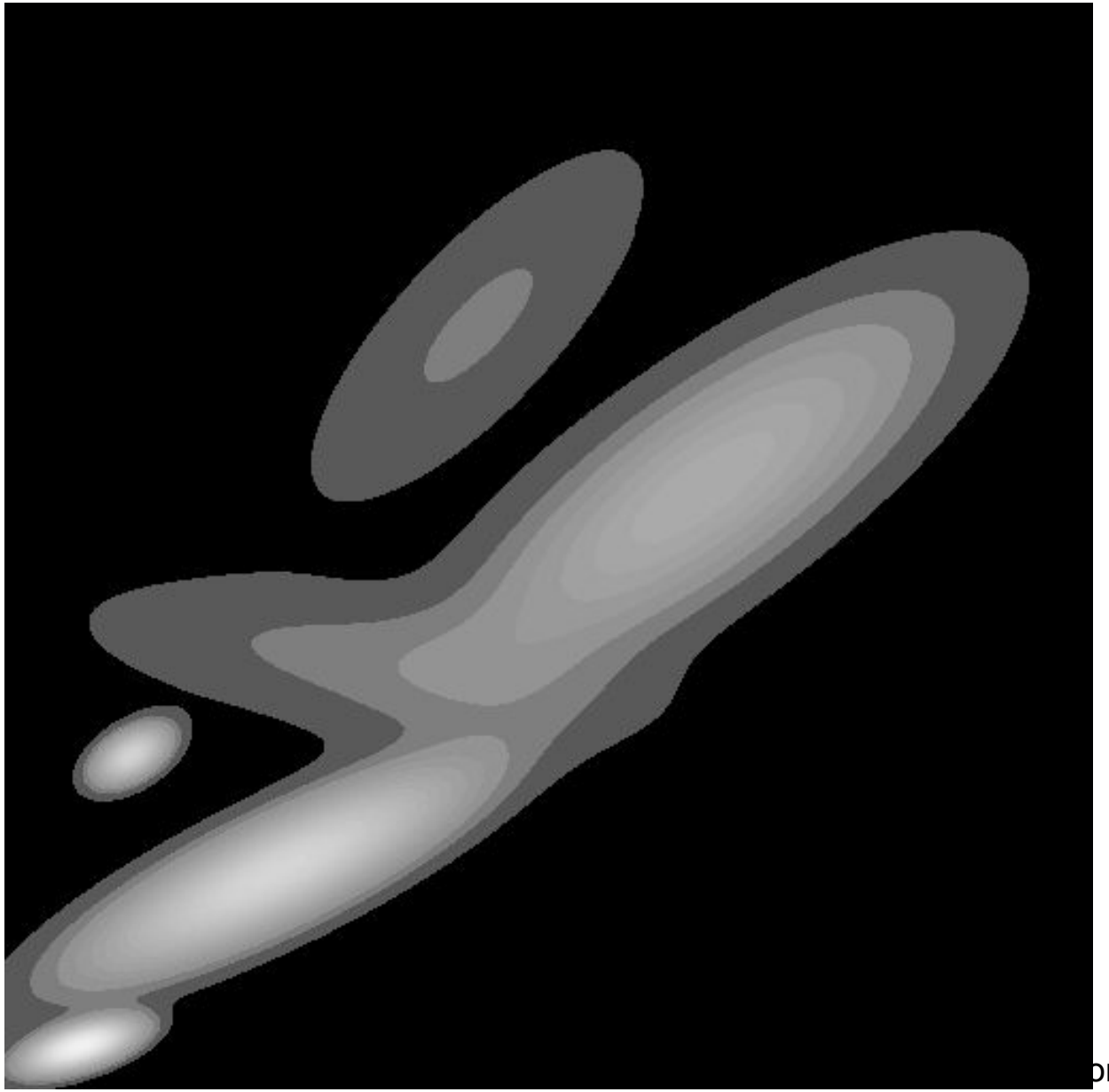


Bio Assay Data

Image credit Prof A Moore, Carnegie Mellon University

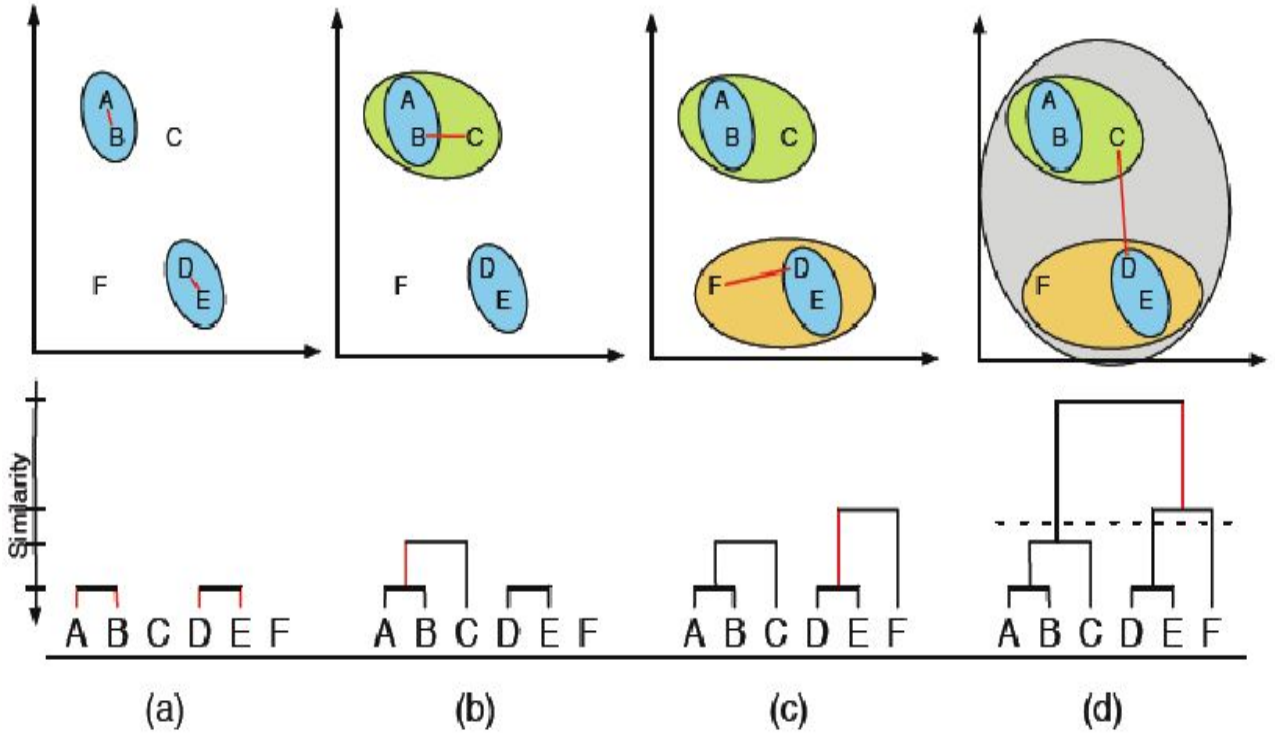
University of Glasgow

Gaussian Mixture Model (GMM)



Bio Assay Data

Hierarchical Clustering (HC)

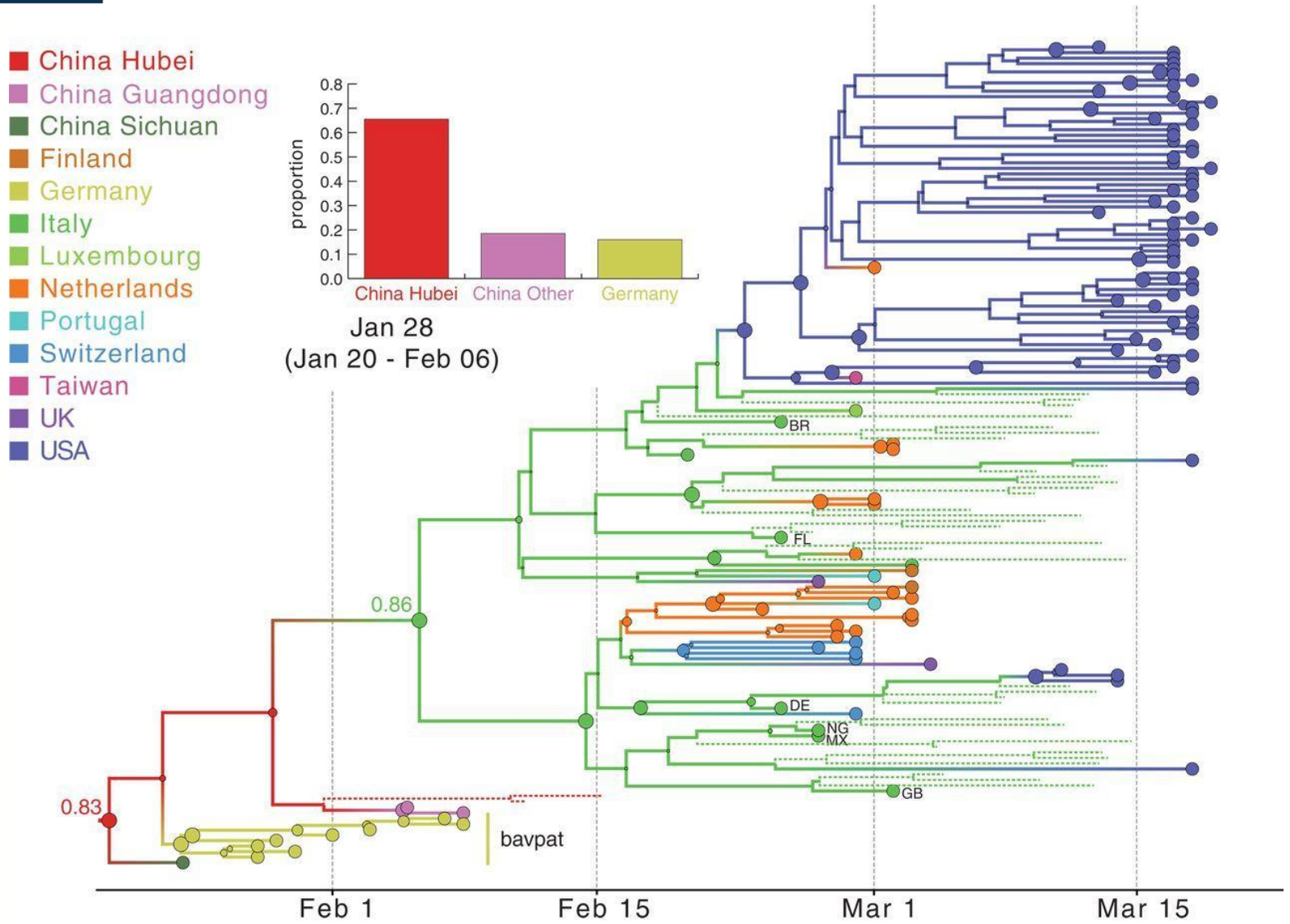


- Consider each data point as separate cluster.
- Consecutively merge clusters until all clusters are connected or specified number of clusters are obtained.

Dendrograms illustrate distance between clusters.

No assumption on number of clusters

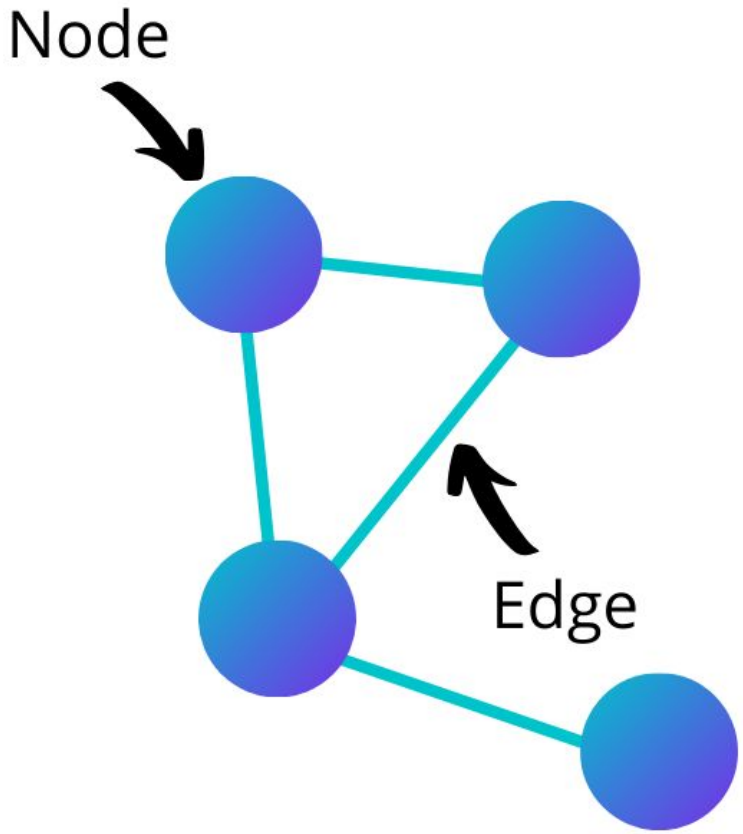
Hierarchical Clustering (HC)

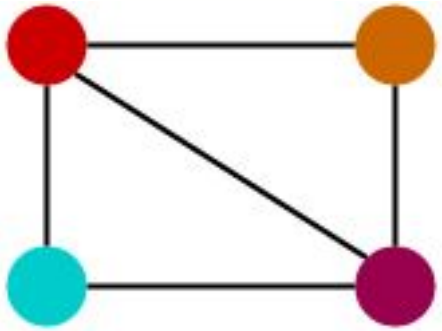


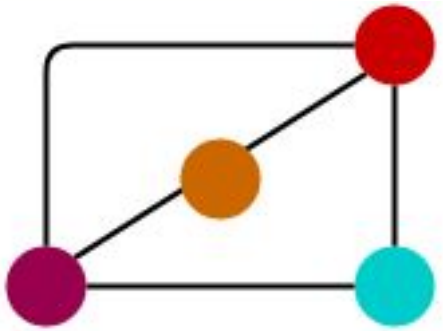
DOI: 10.1126/science.abc8169

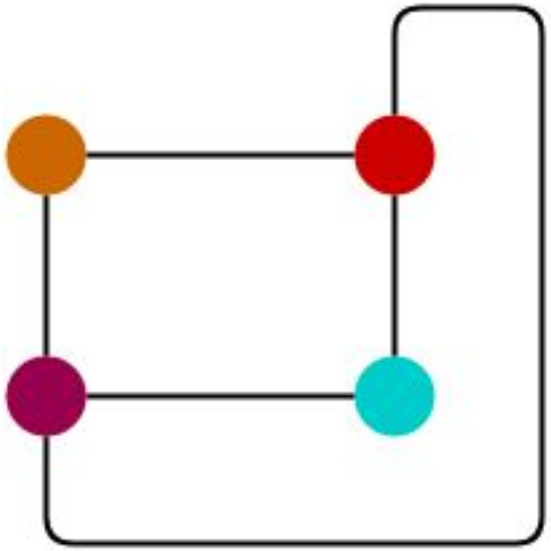
University of Glasgow

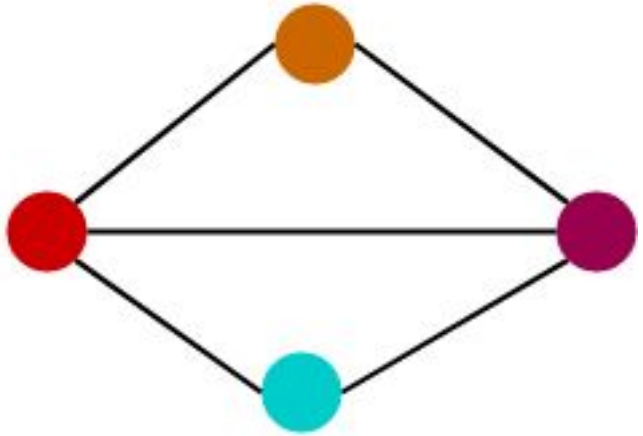
Graph Theory





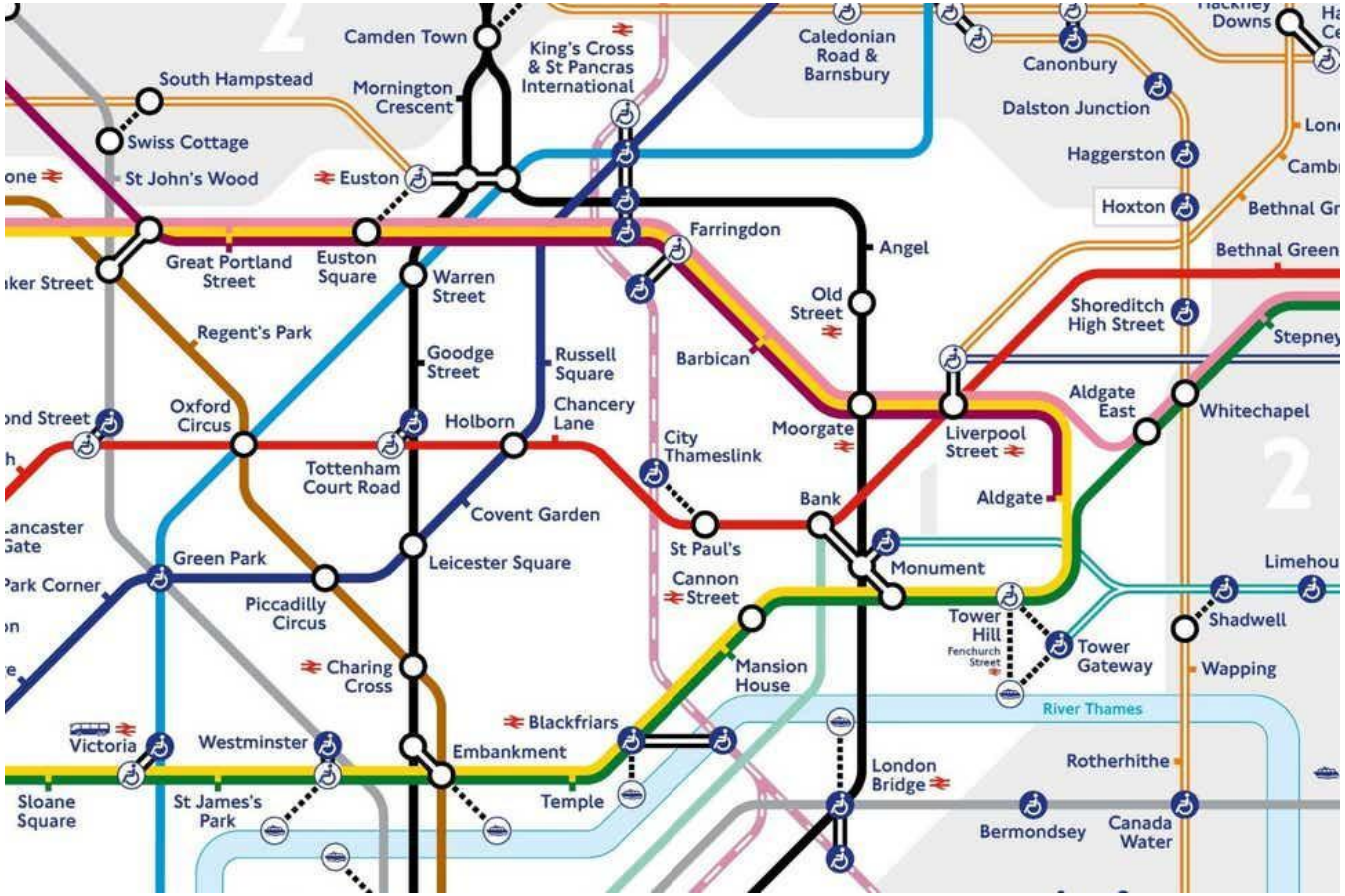






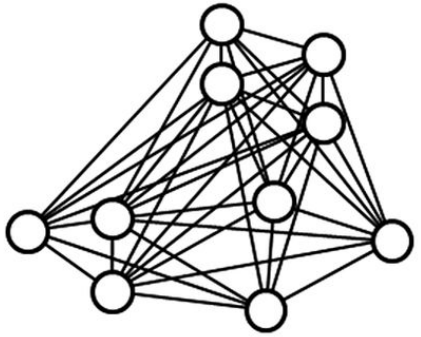
University of Glasgow

Making Sense of Arbitrary Shapes

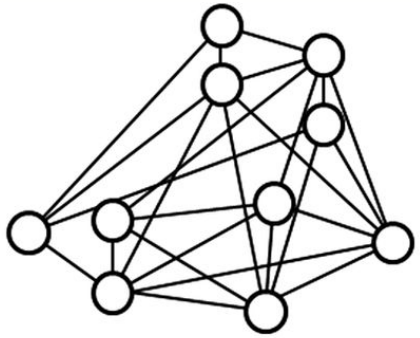


University of Glasgow

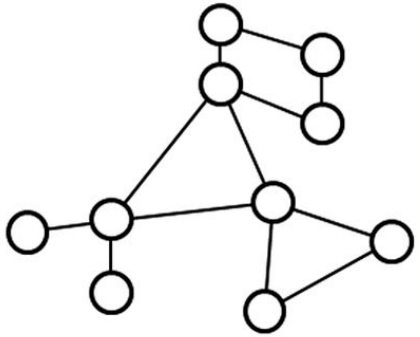
Graph-based clustering



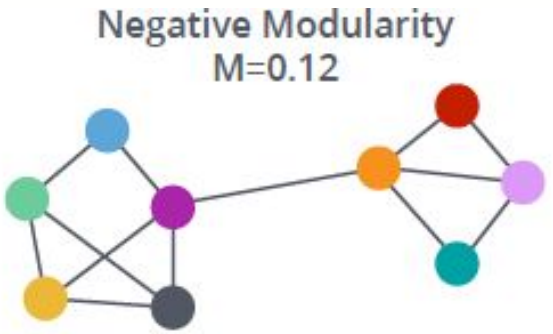
Modularity ≈ -1
(almost complete network)

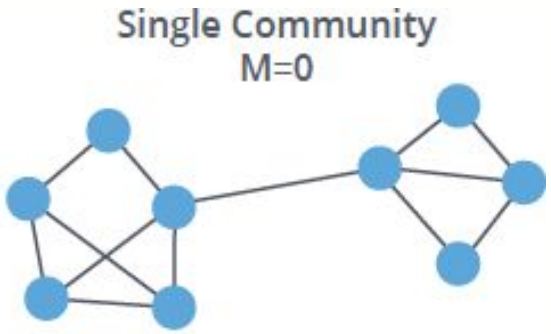


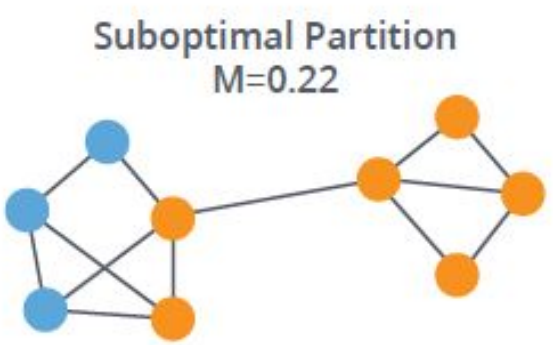
Modularity ≈ 0
(random network)

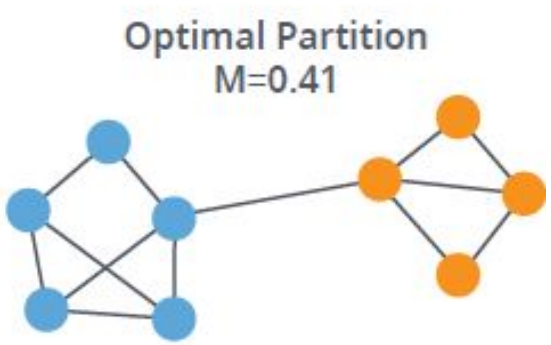


Modularity ≈ 1
(modular network)





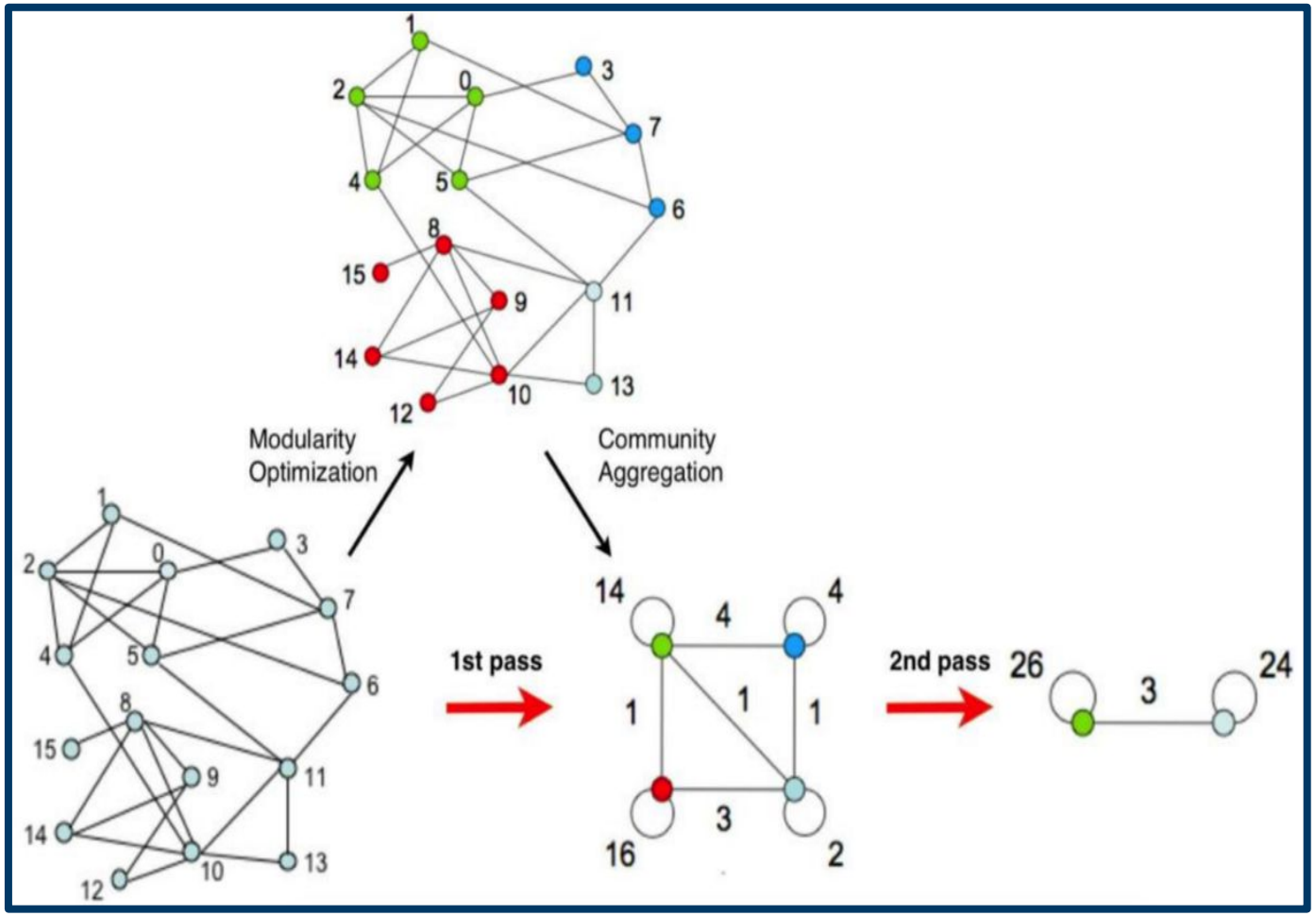




Louvain Clustering

- Each data point represented as a node
- Similarity between two data points represented by an edge
- Assign nodes to different clusters by Modularity
- Clusters repeatedly combined until no improvement in modularity

Resolution parameter in modularity function determines number of clusters.



University of Glasgow

What make model selection for clustering challenging?

- No labels
- Outcomes are subjective
- Hard to define model performance (cluster quality)
- Sensitive to different clustering algorithms and different feature spaces

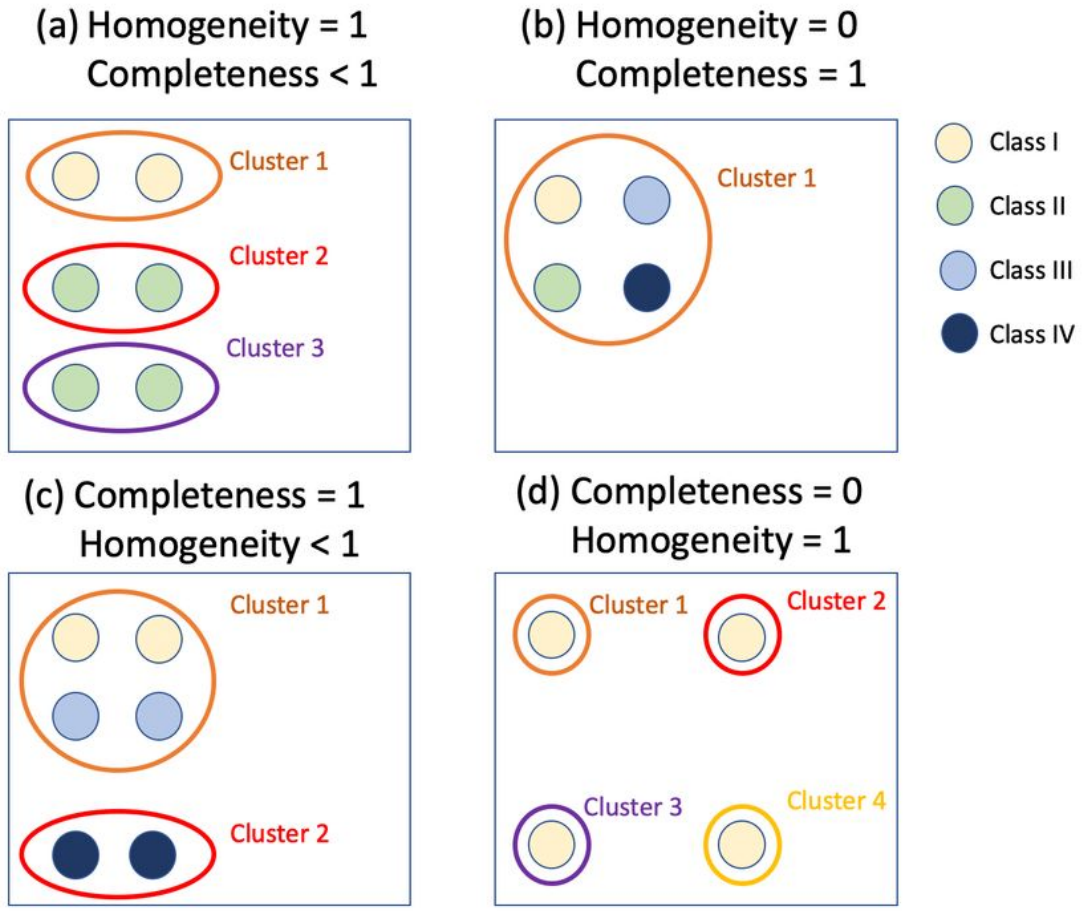
How do we evaluate then?

V-measure

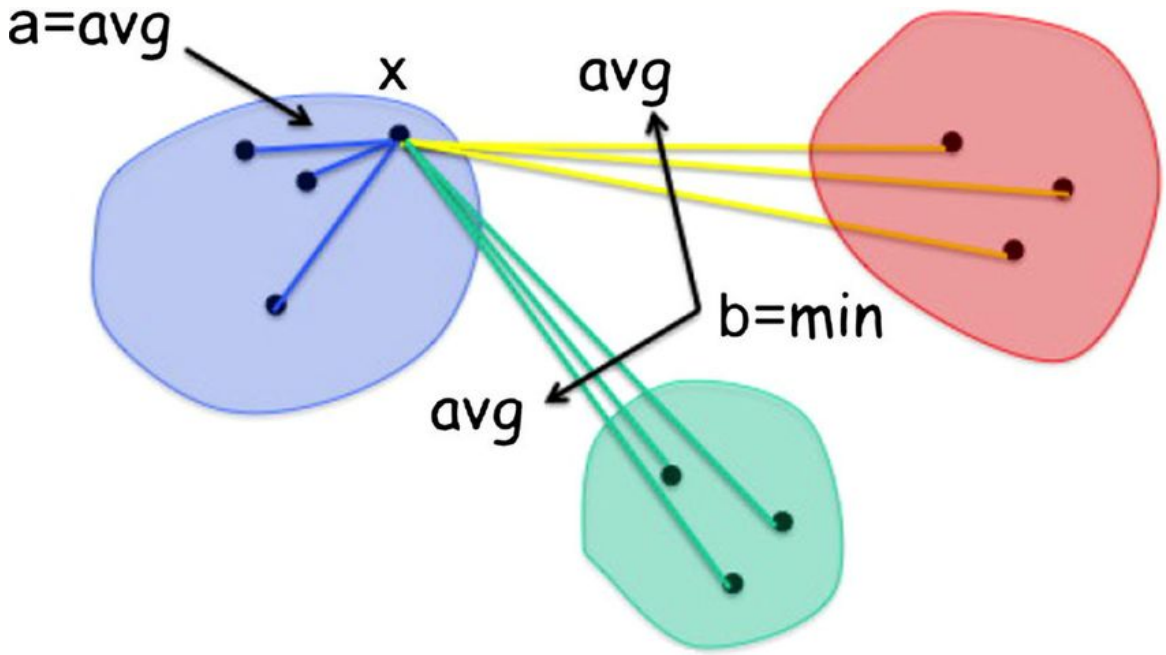
Given ground truth labels: Use them!

- homogeneity: each cluster contains only members of a single class
- completeness: members of a given class are assigned to the same cluster

V-measure is the average (harmonic mean) of the two metrics



Silhouette Score



If the ground truth labels are not known:

Evaluate on structure

- -1: incorrect clustering
- ≈0: overlapping clusters.
- +1: highly dense clustering

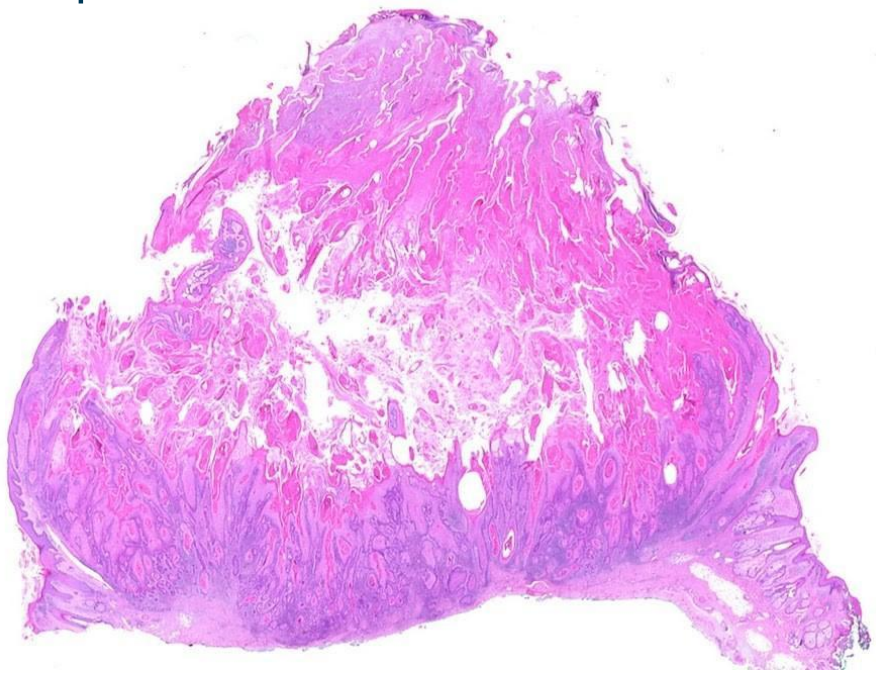
Higher score means clusters are dense and well separated

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

a. 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

Introduction to digital pathology

- Cancer diagnosed by biopsy
- Removed tissue analysed under digital microscope
- 20Gb on average ~100000x100000 pixels



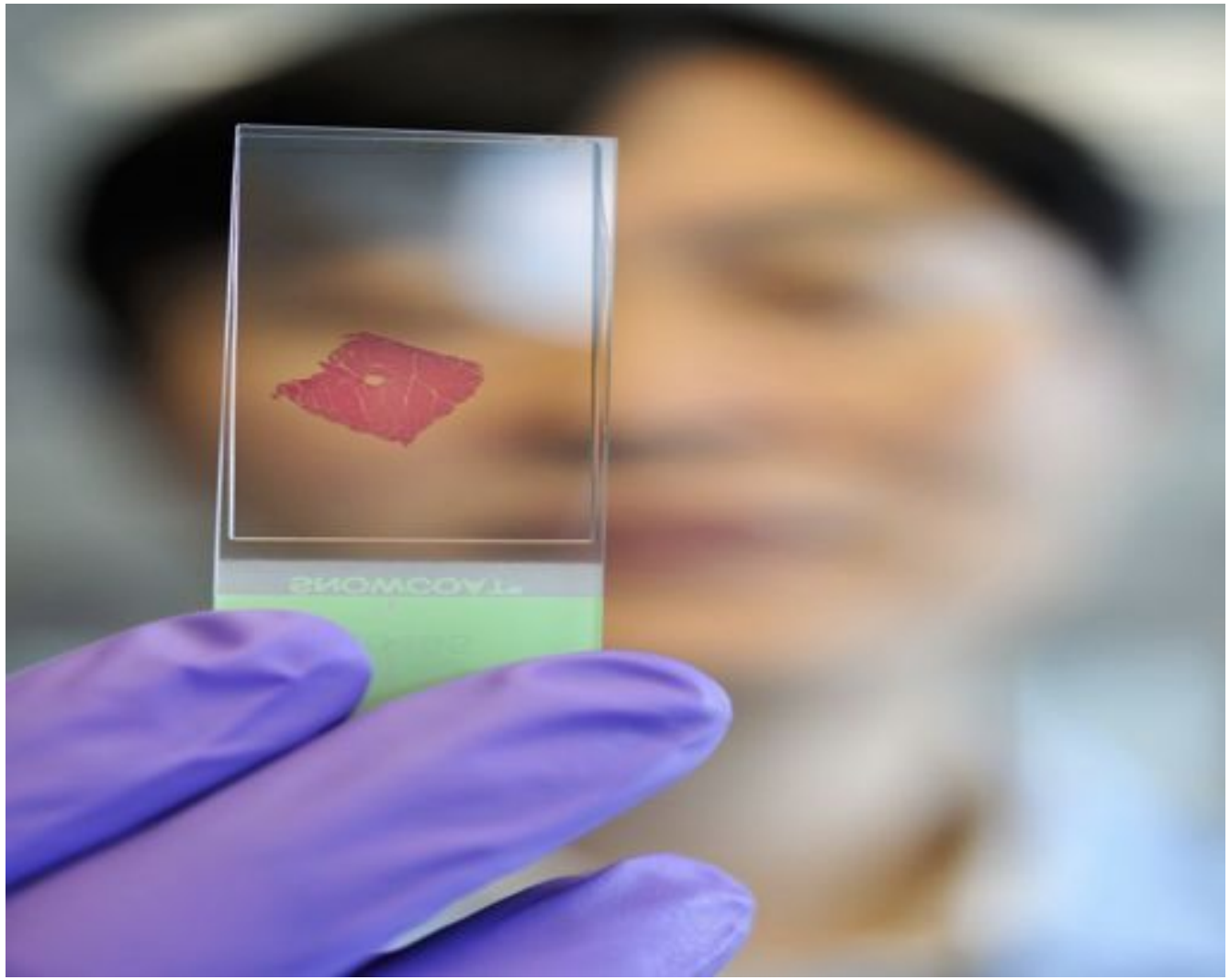
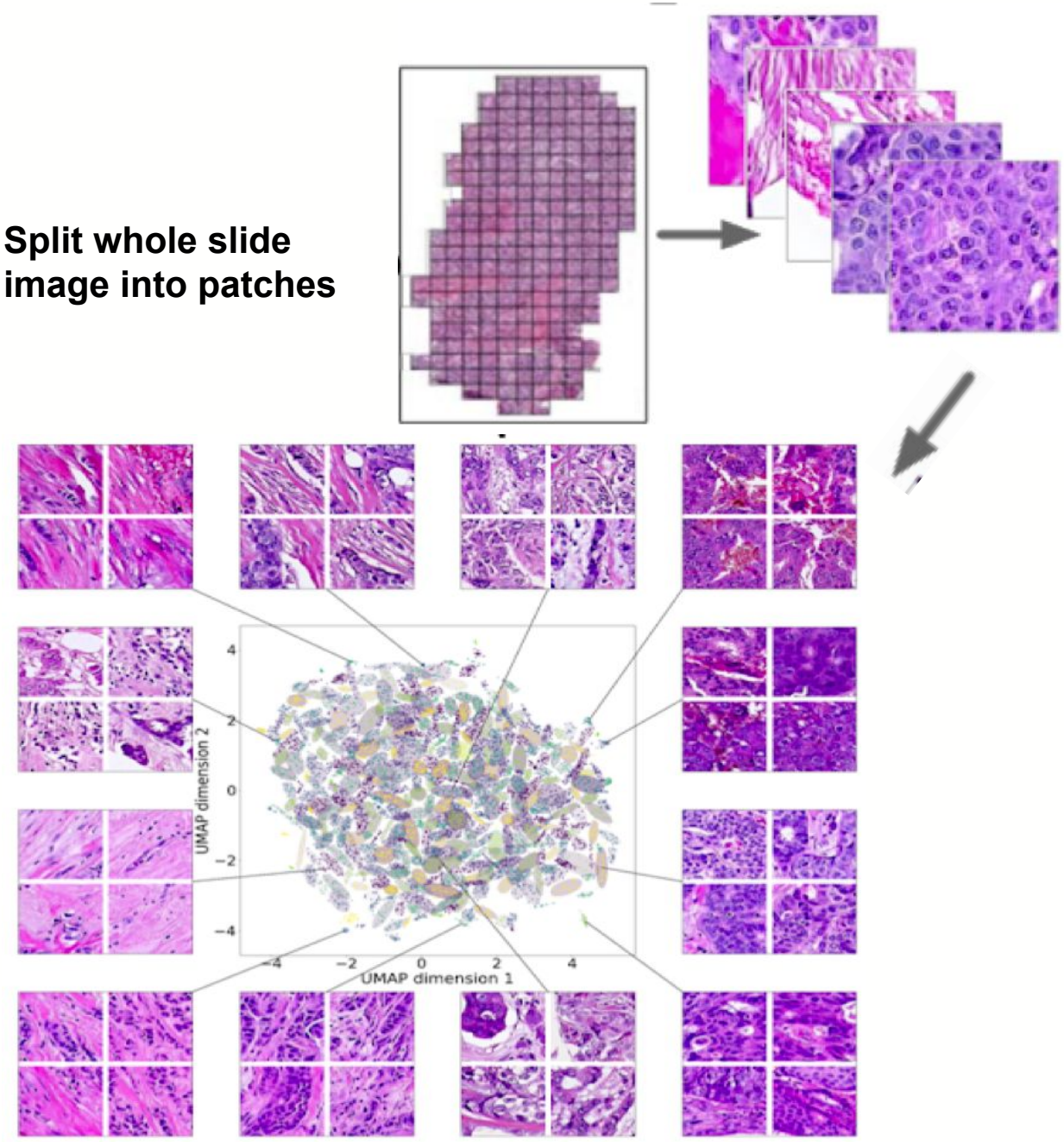


Image Processing

- Break whole slide images (WSIs) into small patches
- (Dimensionality reduction/feature extraction)
- Cluster tissue patches

Clustering gives statistical summary of visual features

Split whole slide image into patches

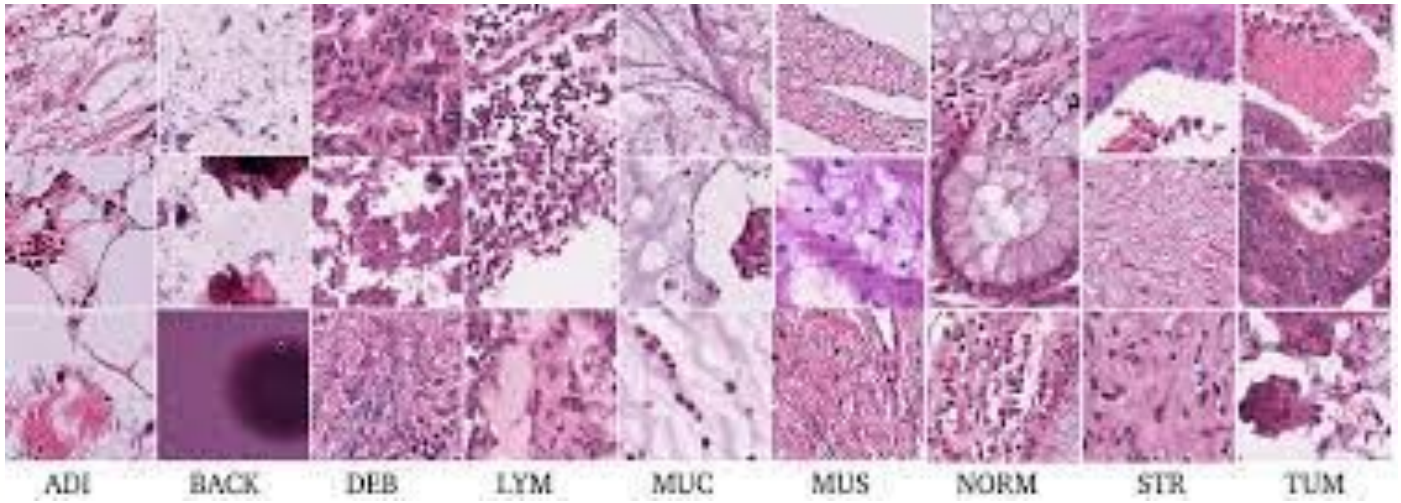


Group similar patches into the same clusters

Your task: model selection for tissue patch clustering

Dataset:

- 5,000 colorectal cancer tissue patches.
- 9 tissue types:
  - Adipose (ADI)
  - background (BACK)
  - debris (DEB)
  - lymphocytes (LYM)
  - mucus (MUC)
  - smooth muscle (MUS)
  - normal colon mucosa (NORM)
  - cancer-associated stroma (STR)
  - colorectal adenocarcinoma epithelium (TUM)



Your task

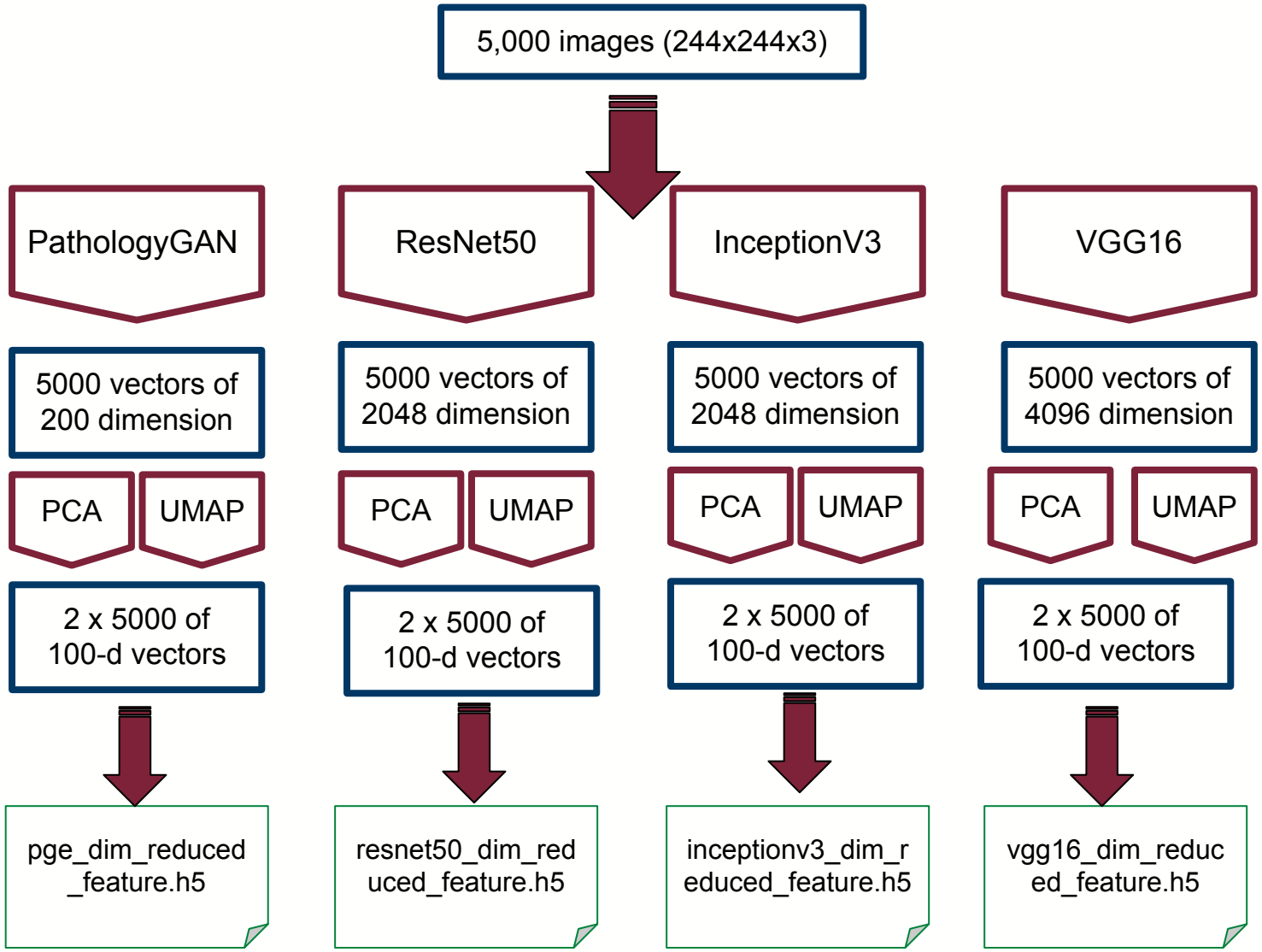
- Select appropriate clustering algorithms (Kmeans, GMM, HC, and Louvain, or HDBScan, Leiden or any other algorithm)
- Apply to a cancer dataset
- Assess model performance (Silhouette score/V measure)

Feature extraction and preprocessing has been done for you

- **PathologyGAN:**
  - state-of-the-art model for tissue images
  - Trained on unlabelled data
- **ResNet50/InceptionV3/VGG16:**
  - Popular CNN classifiers
  - Trained on ImageNet dataset and achieve 74.9%, 77.9%, 71.3% accuracy

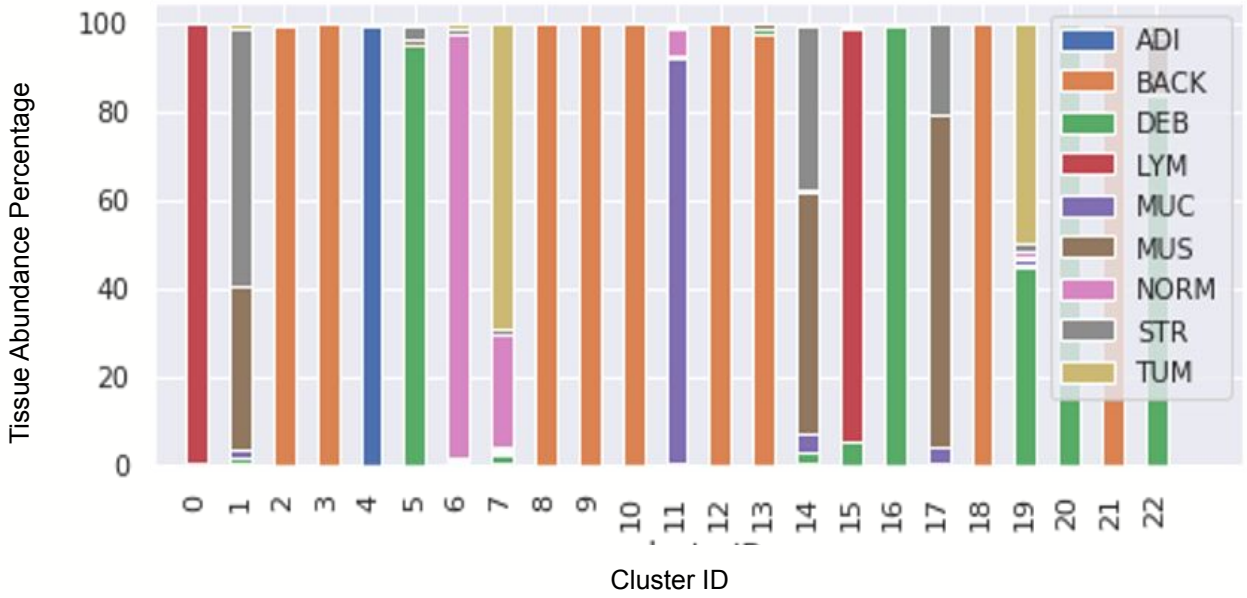
Dimensionality reduction methods reduce each representation size to 100

- **PCA:** the first 100 PCs with highest variance are obtained
- **UMAP:** 100 umap components



Example of cluster performance report

Measure A				
Representation	**Kmeans**	**GMM**	**DBScan**	**Louvain**
PathologyGAN	?	?	?	?
ResNet50	?	?	?	?
InceptionV3	?	?	?	?
VGG16	?	?	?	?



University of Glasgow

Report summary

Objective: test clustering algorithms and evaluate performance on real data

Include in your report:

1. **Introduction**
   Clustering/background/data
2. **Methodology**:
   Theory and intuition behind each algorithm and representation
3. **Experimental framework:**
   Parameter searching and evaluation
4. **Results**
   Report and discuss cluster qualities according to both qualitative and quantitative measures
5. **Conclusion**

University of Glasgow

Summary of Methods

K-means

- Fast, simple, guaranteed convergence, scales well with dataset size
- Manual number of clusters, sensitive to outliers, scales badly with dimensions

Hierarchical Clustering

- Simple, guaranteed convergence, scales well, no predetermined number of clusters, gives different resolutions of clustering
- Slow, limited meaningful clustering, can infer artificial relationships

Gaussian Mixture Model

- Handles outliers, more informative, guaranteed convergence
- Slow, same issues as k-means, computation time scales badly with dataset size

Louvain Community Detection

- Handles outliers, more informative, scales well with dataset size and dimension, no assumption on shape
- Computationally intensive for small datasets, pre-determined resolution

University of Glasgow

Summary

Model Selection

- What does *best* mean?
- Consider what outcomes are most important

Clustering

- Different models give very different results
- Use prior knowledge to select appropriate model
- Evaluate clusters based on structure or labels

Summary and extra hints!

1. The objective of this case study is **to test different clustering algorithms on 4 different deep neural network-based representations** extracted from colorectal tissue patches by reporting the cluster qualities according to both intrinsic and extrinsic measures
2. In your report, you are expected to present…
   - Introduction to tasks/backgrounds/data
   - Methodology:

     Theory and Intuition behinds each algorithm and representation
   - Experimental framework:

     Parameter searching and evaluation
   - Result discussion
   - Conclusion

University of Glasgow
THANK YOU