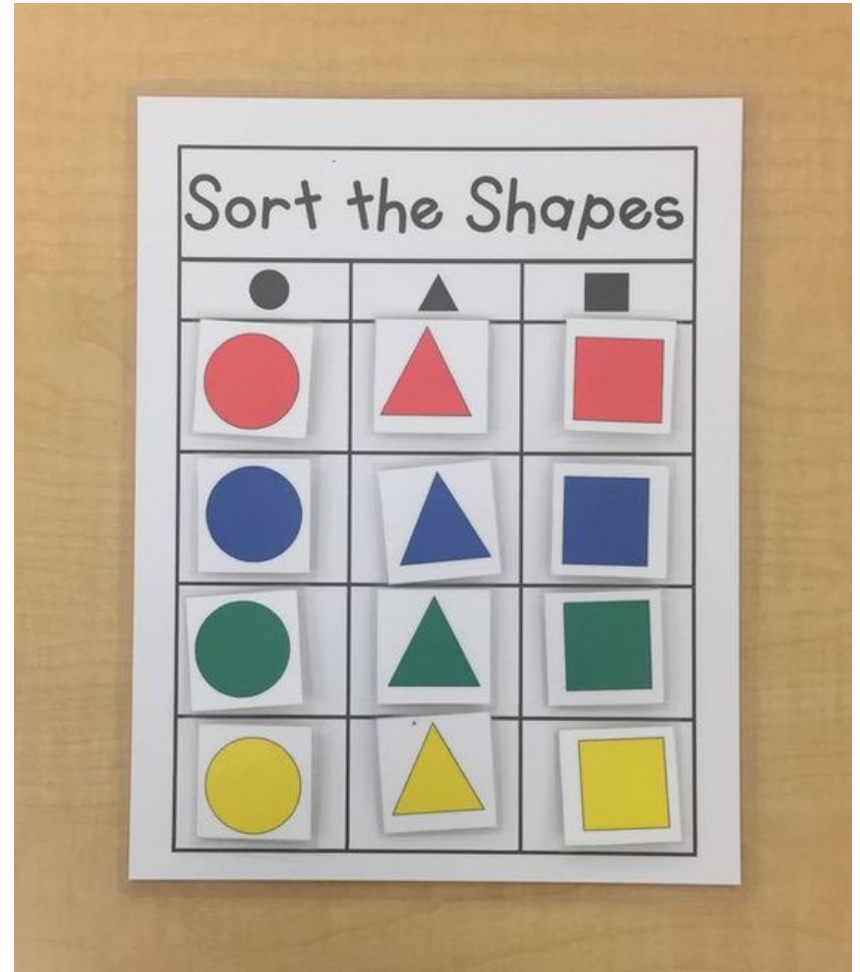


Clustering

Dr Ellis Patrick

School of Mathematics and Statistics, Usyd
The Westmead Institute for Medical Research

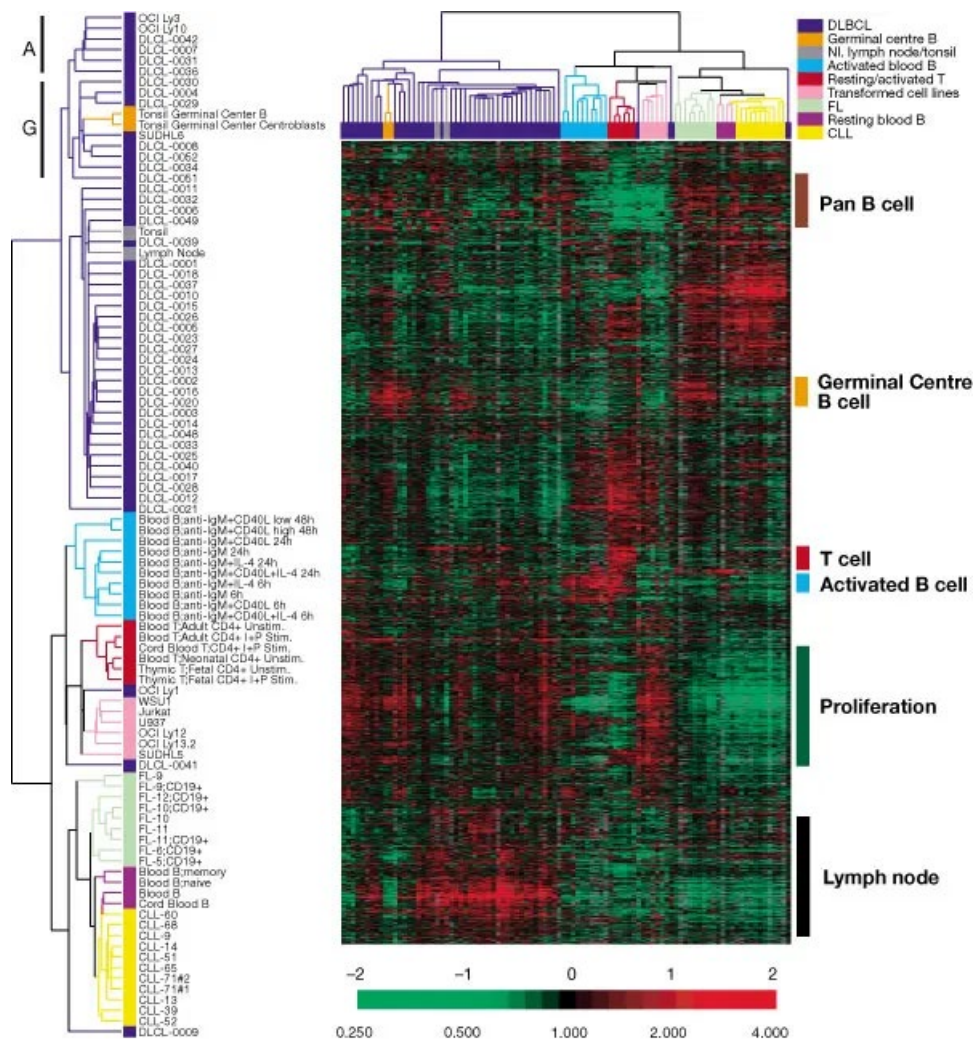


Bioinformatics

As an interdisciplinary field of science, **bioinformatics** combines biology, computer science, information engineering, mathematics and statistics to analyze and interpret biological data.

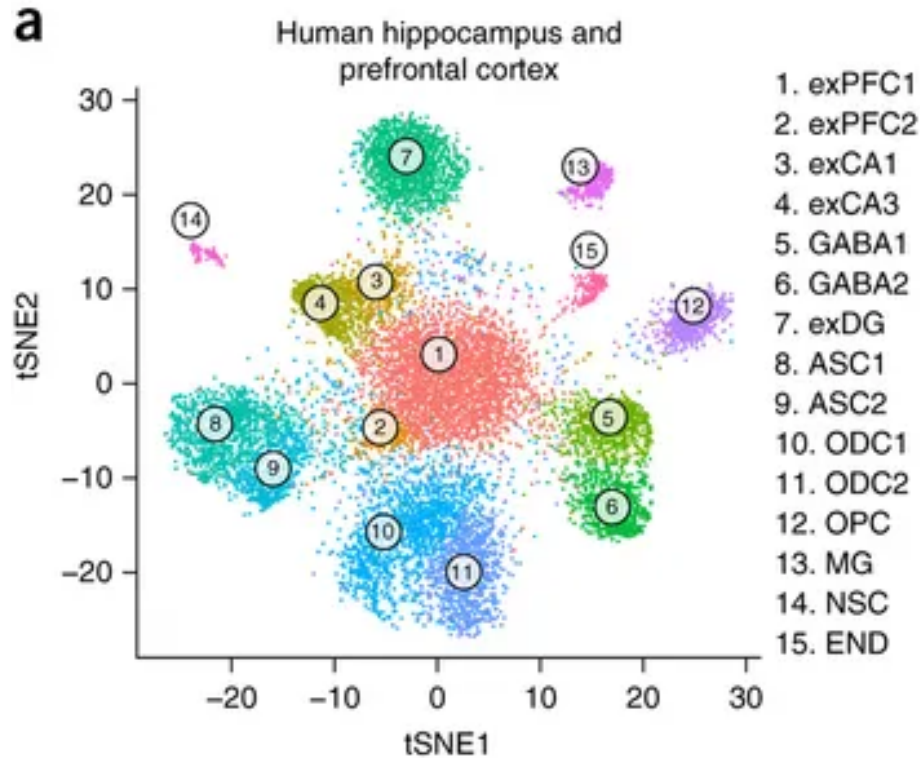
– “Wikipedia”

Clustering 20,000 genes



Alizadeh, Ash A., et al. "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling." *Nature* 403. 6769 (2000): 503.

Clustering 100,000 cells



Breast cancer

- 130 patients. Consider 10 variables:

node.positive

hormonal.therapy

chemotherapy.treatment

age.at.diagnosis

Ethnicity

tumor.size..mm

TumorGrading

radiation.treatment

Progesterone.Receptor.status

EstrogenReceptorStatus

Breast cancer

- 130 patients. Consider 10 variables:

node.positive

hormonal.therapy

chemotherapy.treatment

age.at.diagnosis

Ethnicity

tumor.size..mm

TumorGrading

radiation.treatment

Progesterone.Receptor.status

EstrogenReceptorStatus

Breast cancer

- 130 patients. Consider 10 variables:

node.positive

hormonal.therapy

chemotherapy.treatment

age.at.diagnosis

Ethnicity

tumor.size..mm

TumorGrading

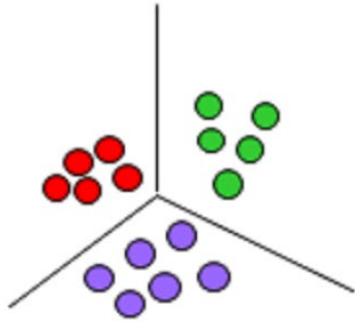
radiation.treatment

Progesterone.Receptor.status

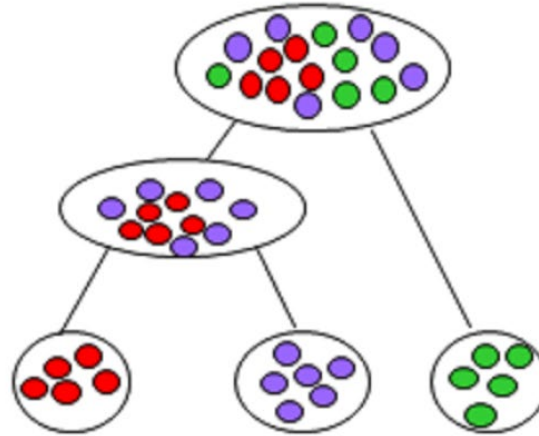
EstrogenReceptorStatus

Unsupervised clustering

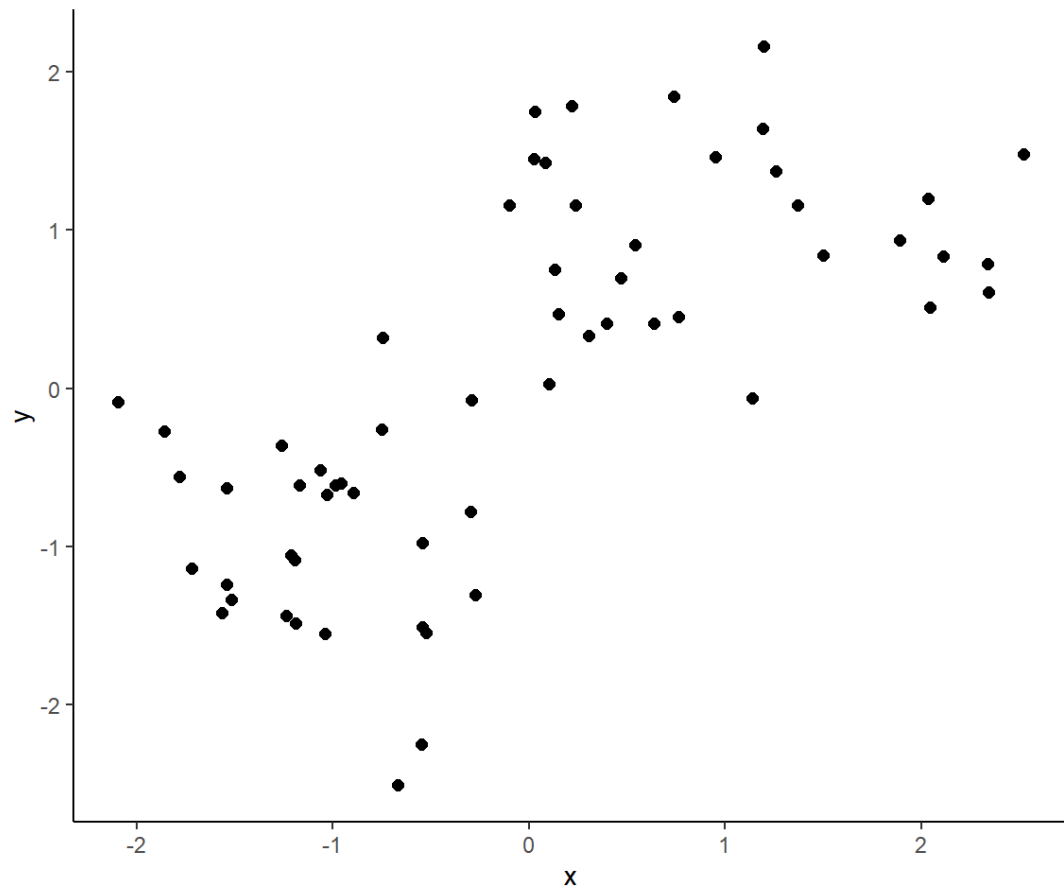
Partitioning



Hierarchical



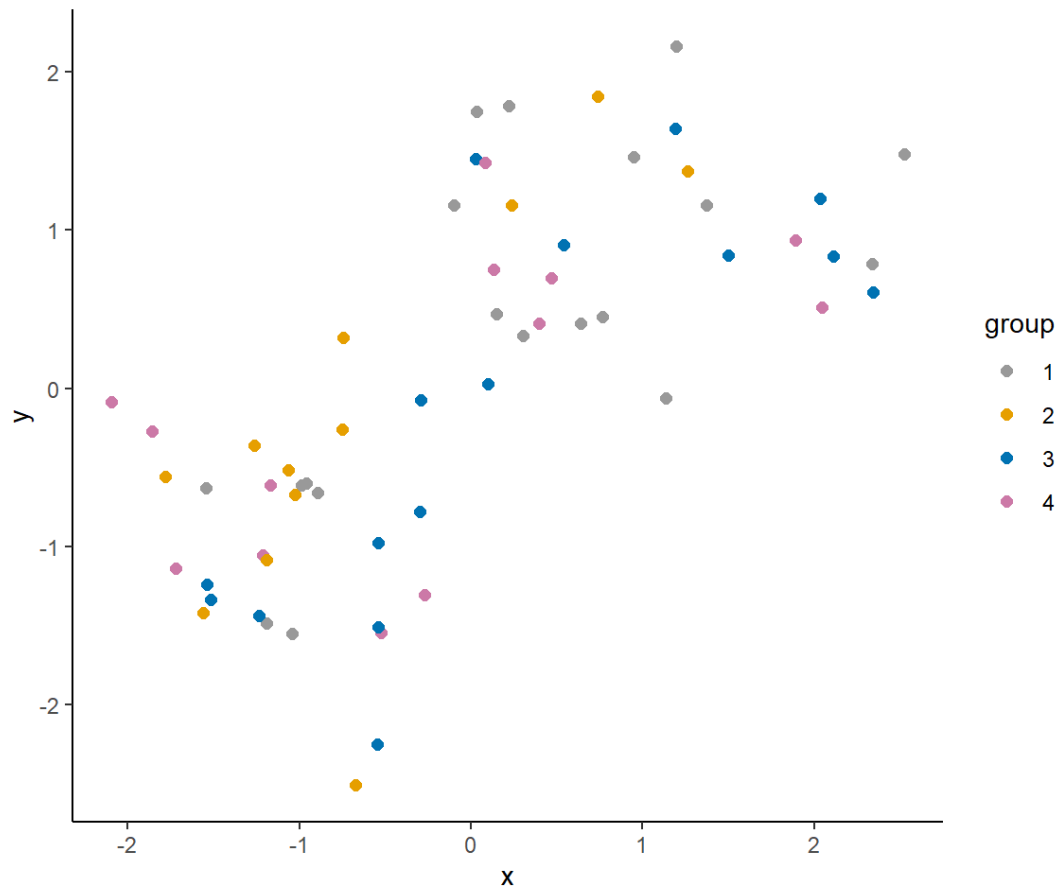
K-means clustering

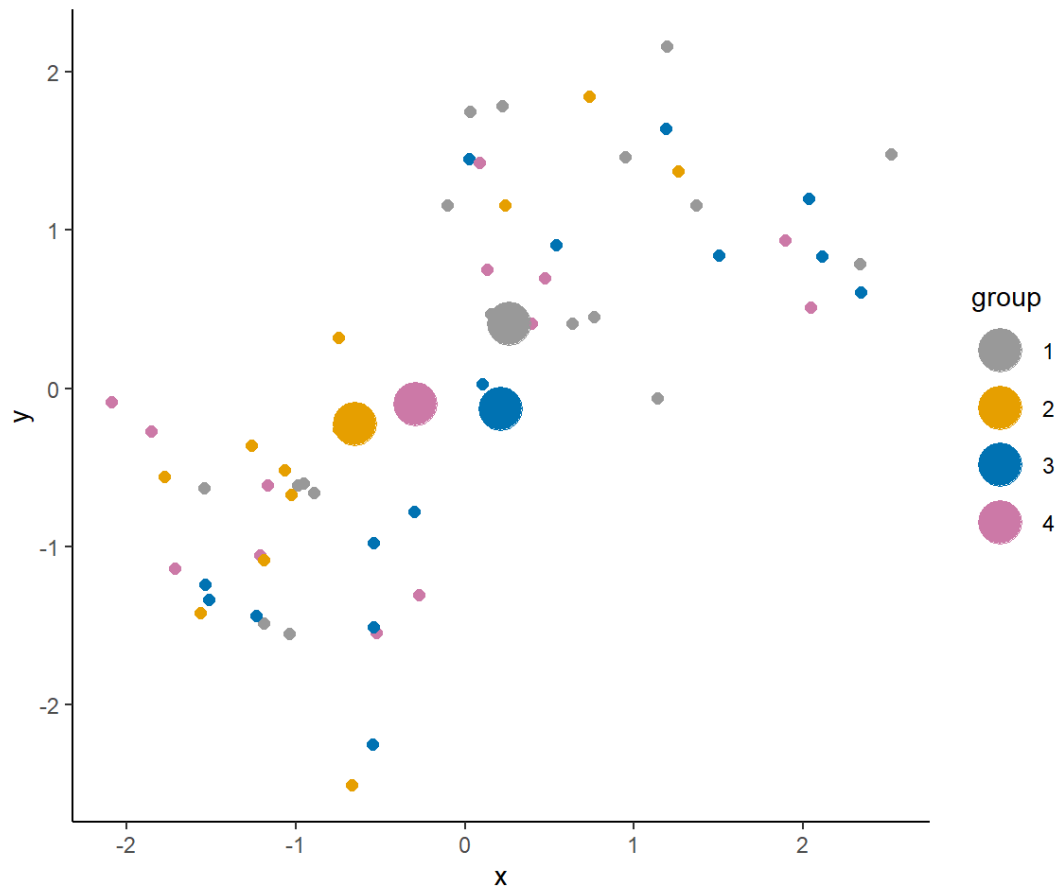


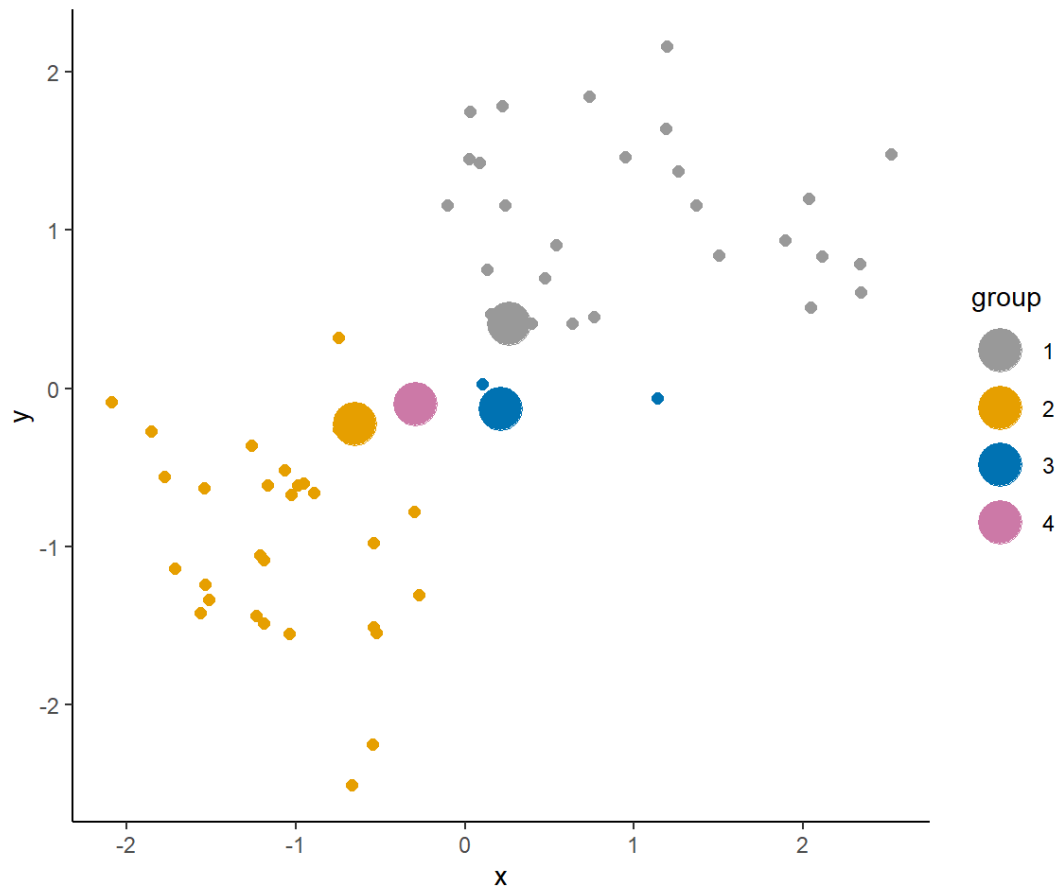
K-means clustering

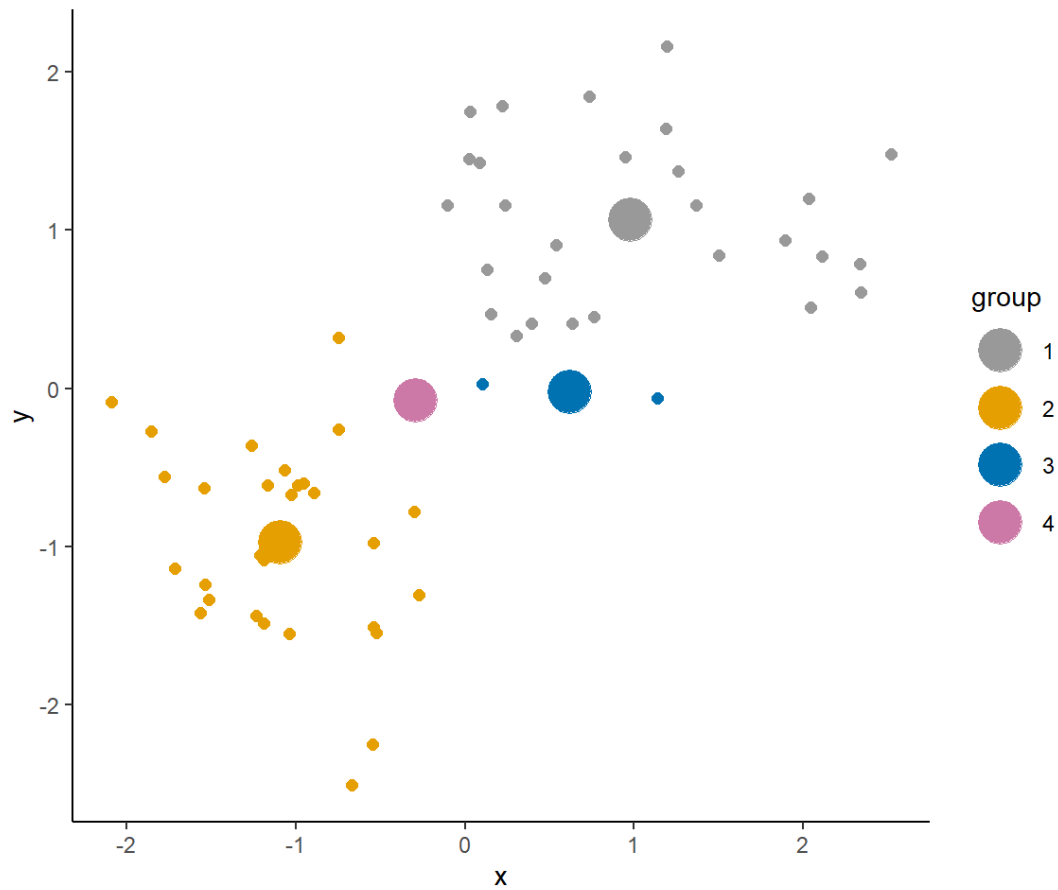
Choose K and then...

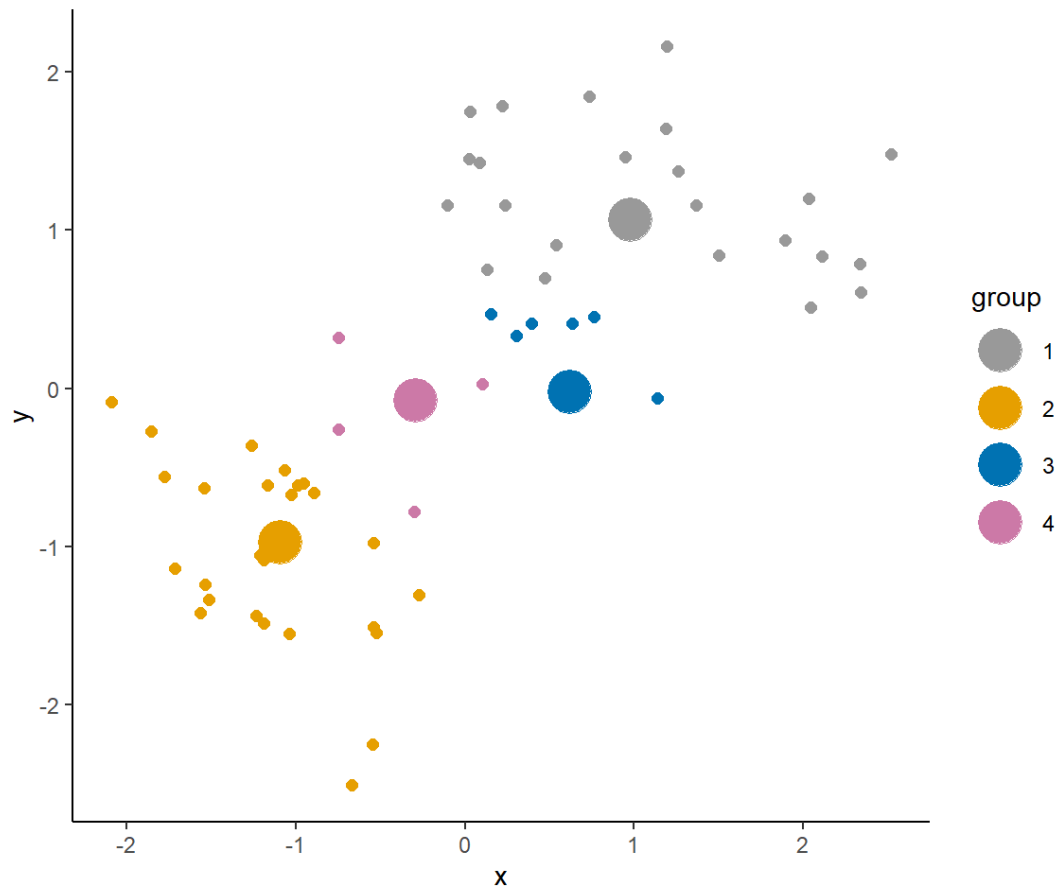
1. Randomly assign a number, from 1 to K , to each of the observations. (These serve as initial cluster assignments for the observations).
2. Iterate until the cluster assignments stop changing.
 - a) For each of the K clusters, compute the cluster centroid. The k th cluster centroid is the vector of the p covariate means for the observations in the k th cluster.
 - b) Assign each observation to the cluster whose centroid is closest (where closest is deemed using Euclidean distance).

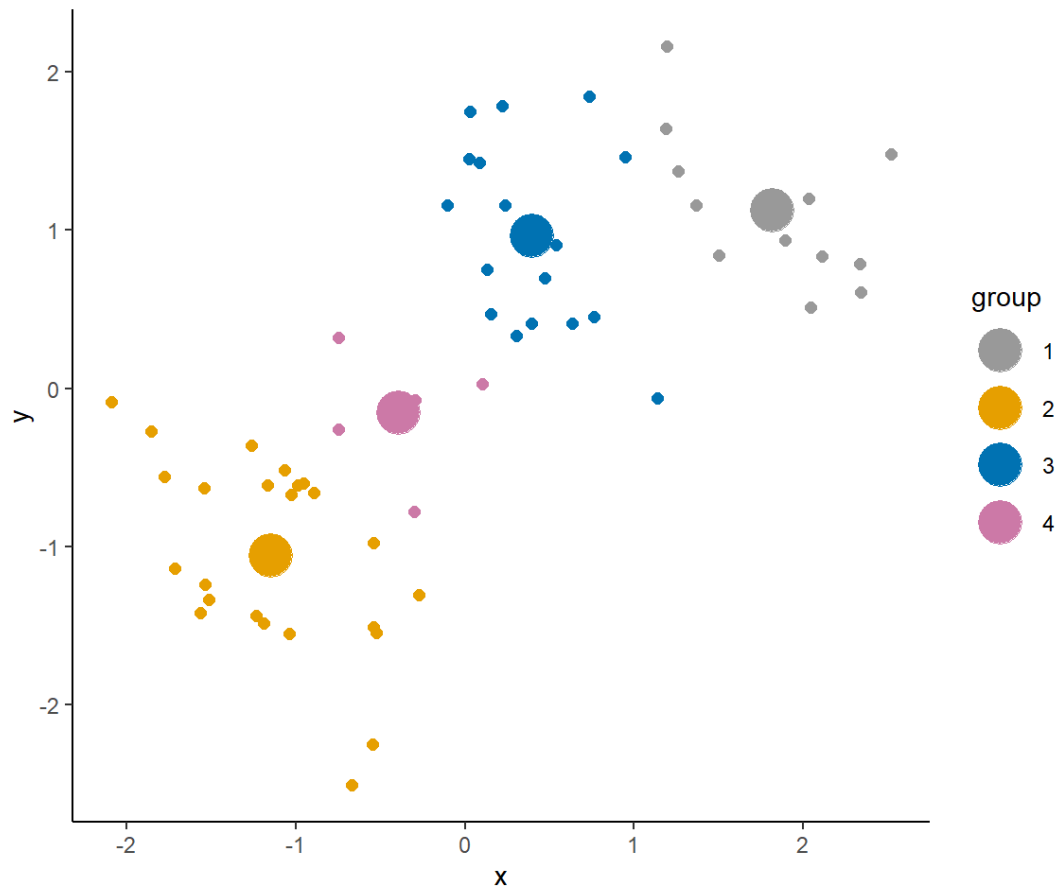


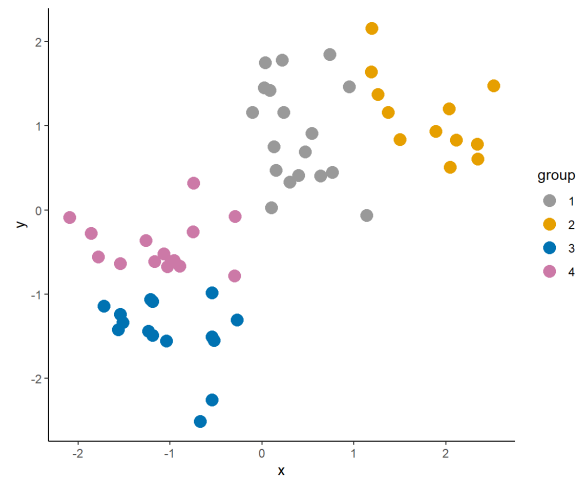
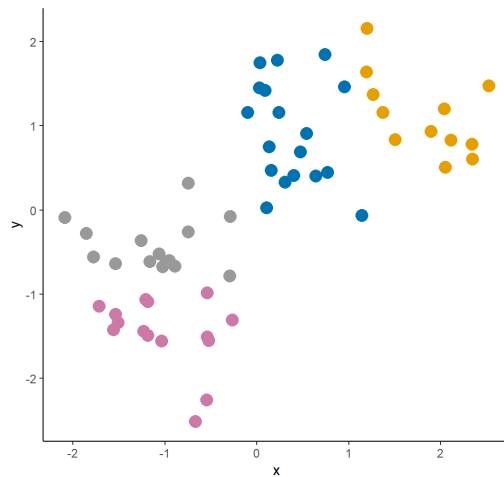
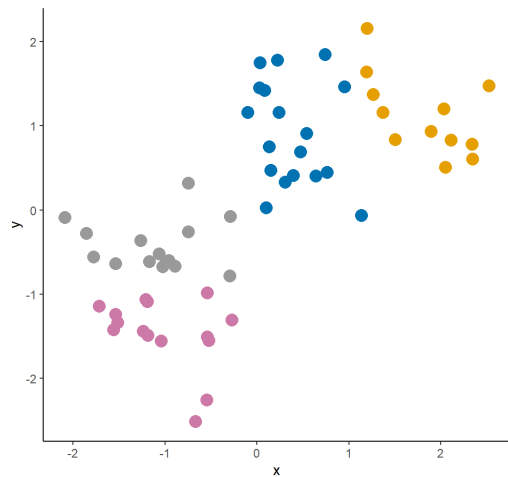
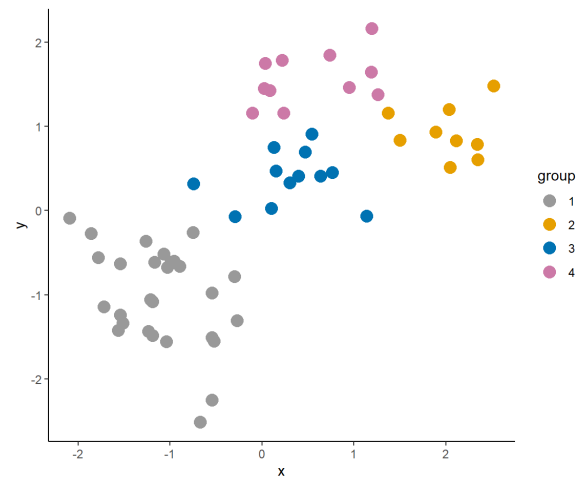
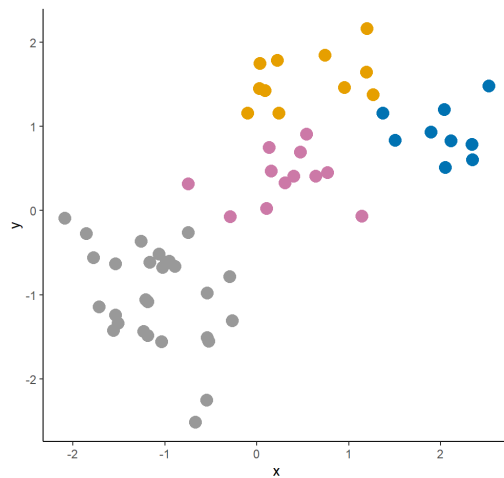
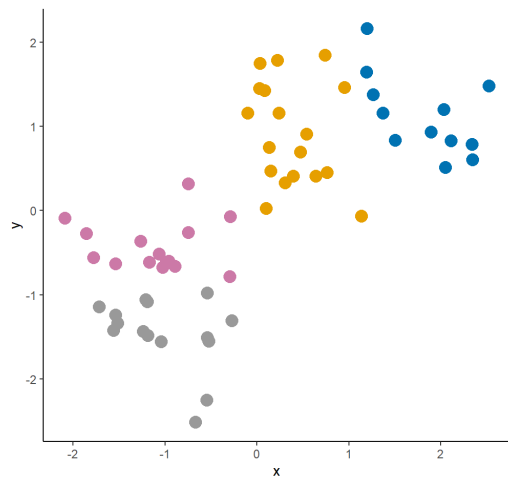


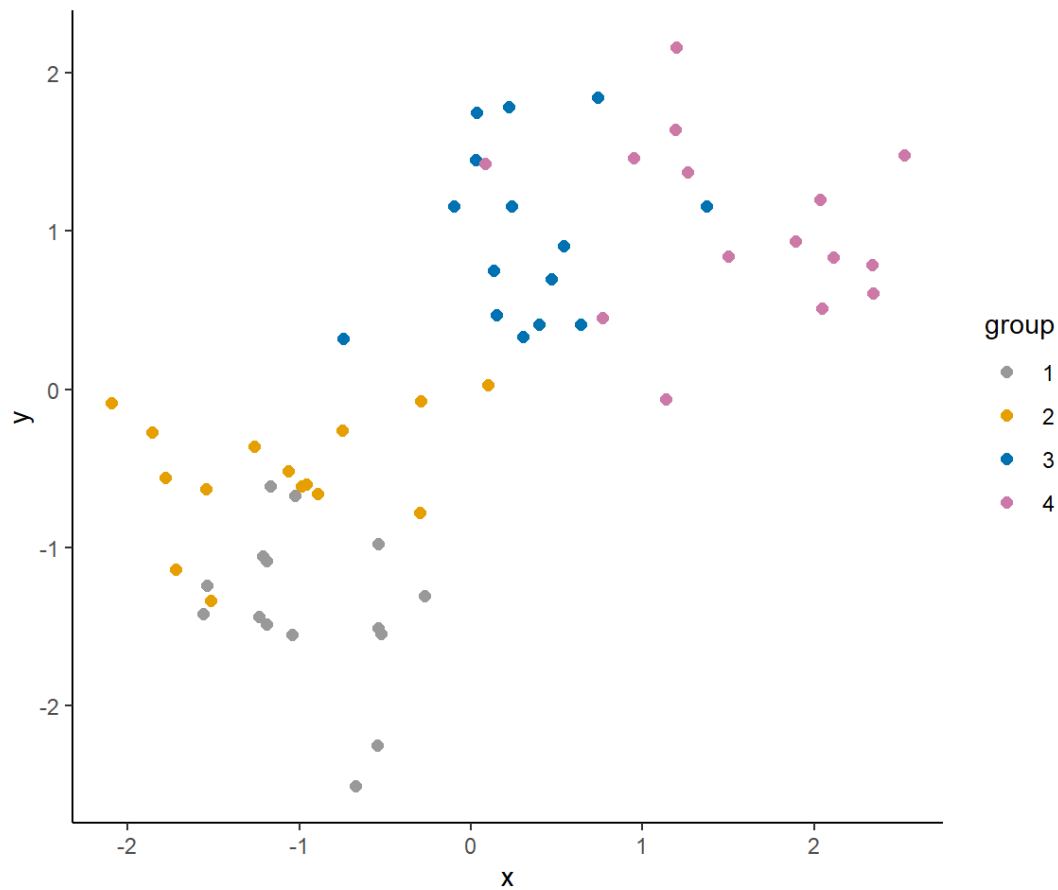












Code

```
km <- kmeans(x, 2)
```

```
km$cluster
```

```
##    [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2  
2 2 2 2 2 2 2 2 2 2 2 2  
##   [38] 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1  
1 1 1 1 1 1 1 1 1 1 1 1  
##   [75] 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
1 1
```

Hierarchical clustering

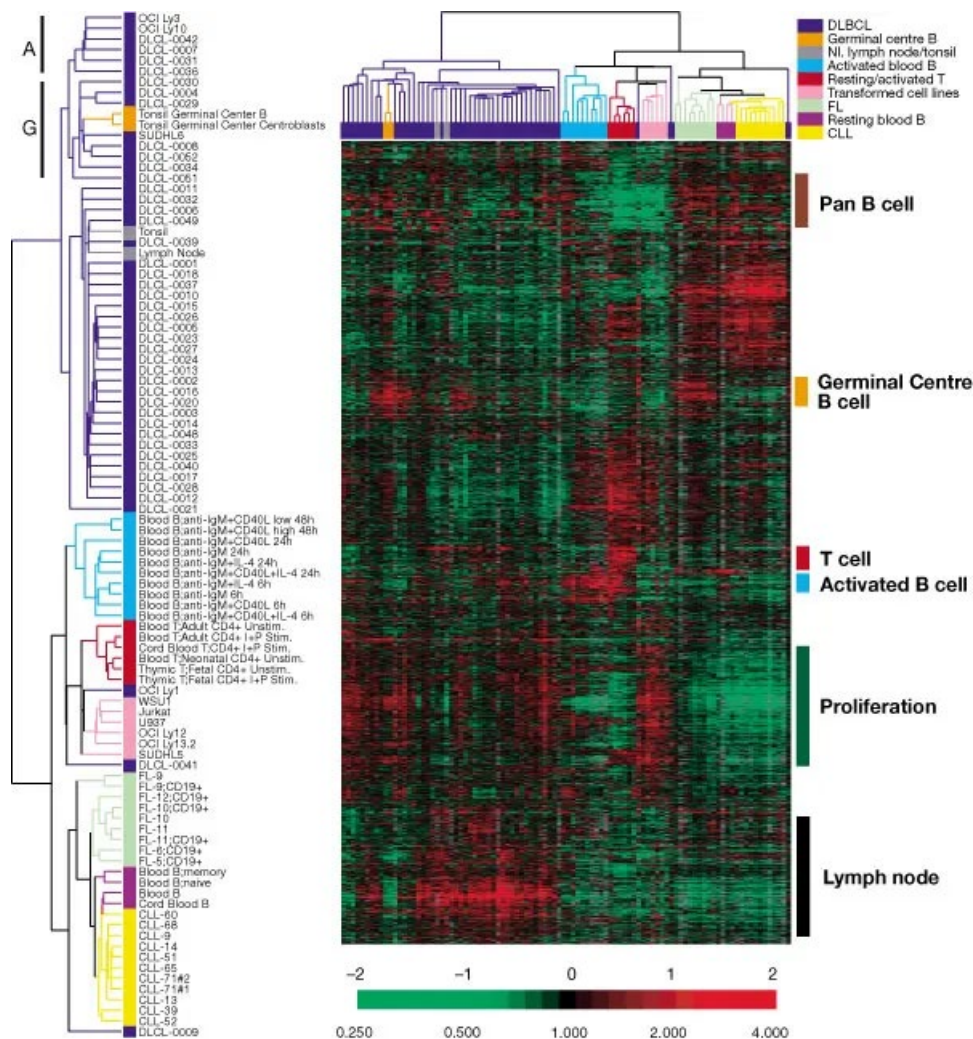
Hierarchical clustering

Hierarchical clustering methods produce a tree or dendrogram.

They avoid specifying how many clusters are appropriate by providing a partition for each k obtained from cutting the tree at some level.

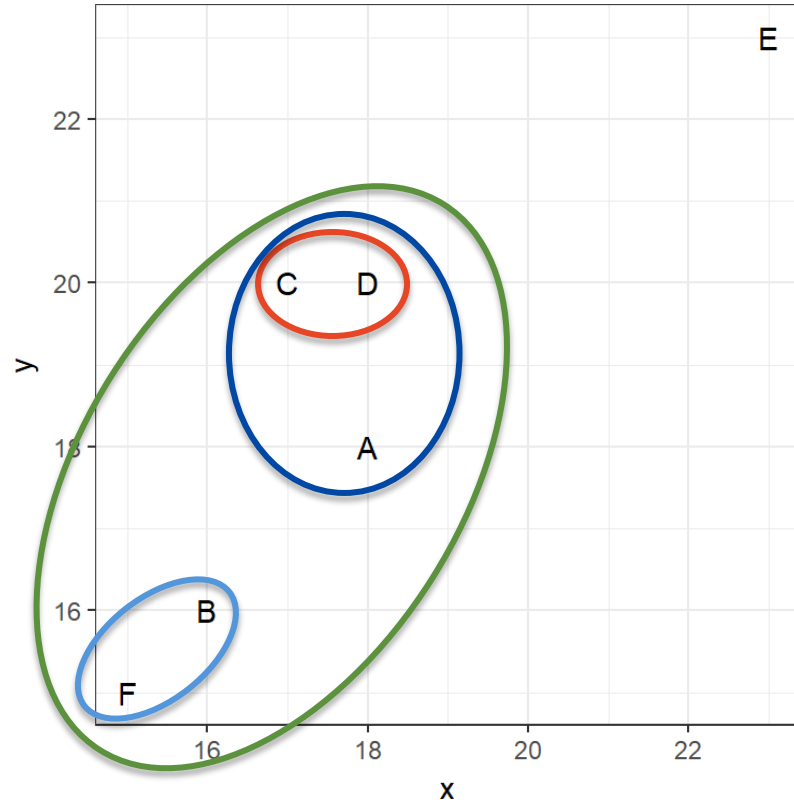
The tree can be built in two distinct ways

1. bottom-up: agglomerative clustering.
2. top-down: divisive clustering.

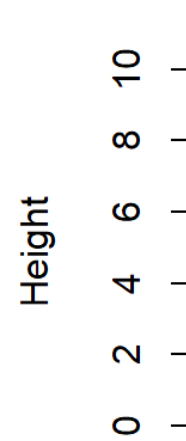


Alizadeh, Ash A., et al. "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling." *Nature* 403. 6769 (2000): 503.

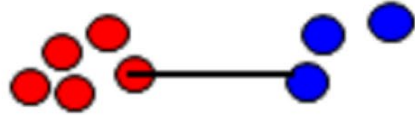
Example 1



Cluster Dendrogram



Between cluster similarity measures



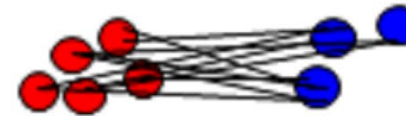
Single (minimum)



Complete (maximum)

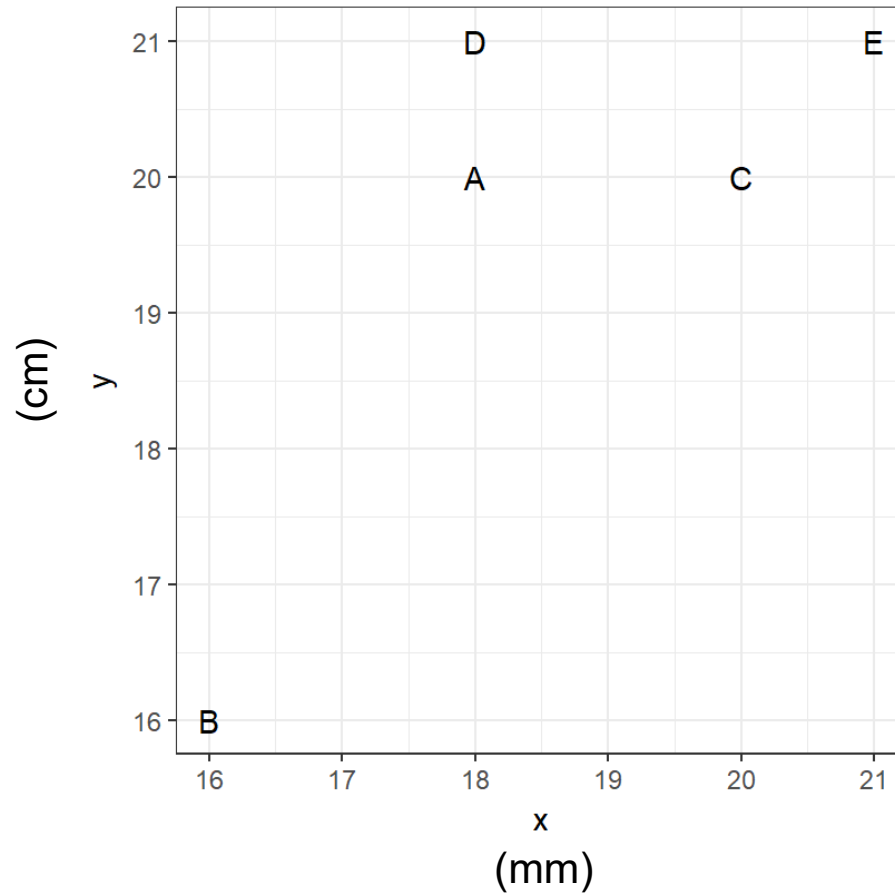


Distance between centroids



Average (Mean) linkage

Example 2



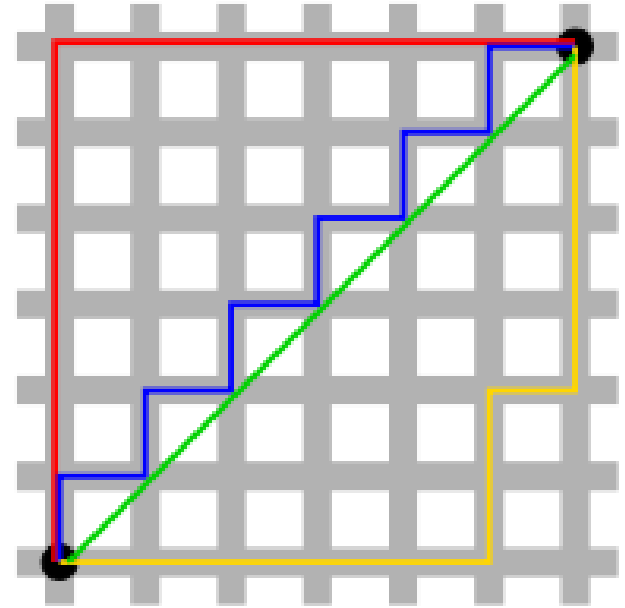
Distance metrics

- Euclidean

If $\mathbf{x} = \{x_1, x_2\}$ and $\mathbf{y} = \{y_1, y_2\}$

$$D(\mathbf{x}, \mathbf{y}) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

- Manhattan



Green = 8.49

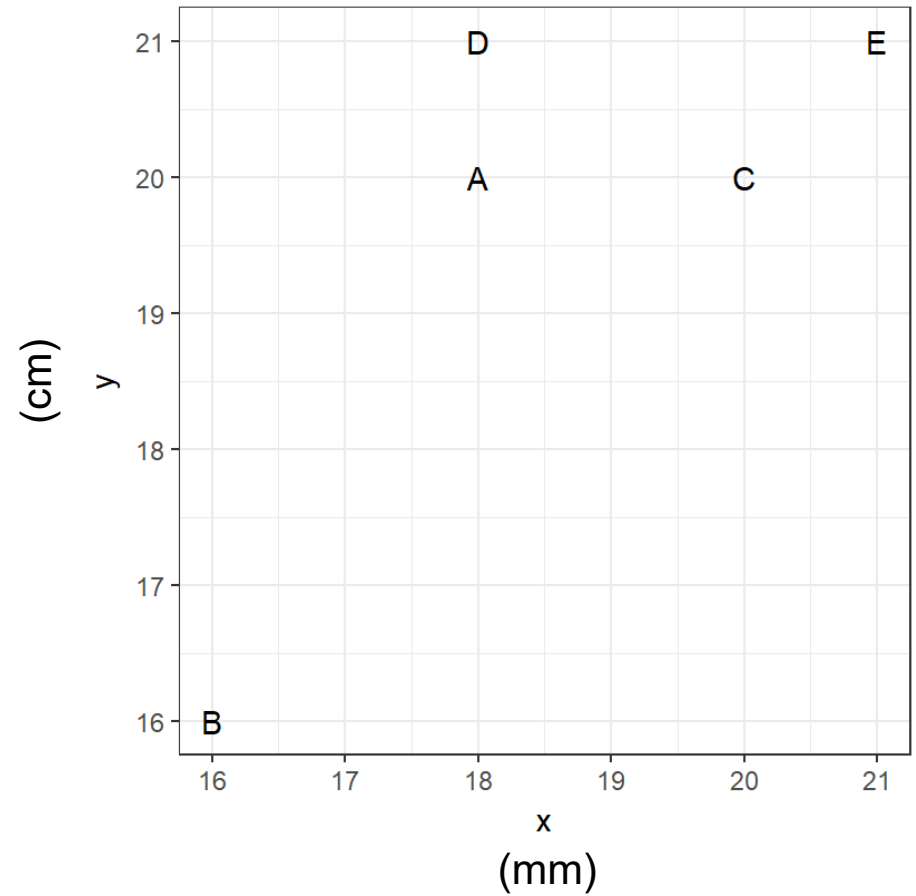
Red, blue, yellow = 12

Code

```
hc <- x |>  
  dist(method = "euclidean") |>  
  hclust(method = "complete")  
  
cutree(hc, k = 2)
```

```
##    [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1  
1 1 1 1 1 1 1 1 1 1 1 1 1  
##   [38] 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2  
2 1 2 2 2 2 2 2 2 2 2 2 2  
##   [75] 1 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2  
2 2
```

$$D(x, y) = \sqrt{\frac{(x_1 - x_2)^2}{scale_x^2} + \frac{(y_1 - y_2)^2}{scale_y^2}}$$

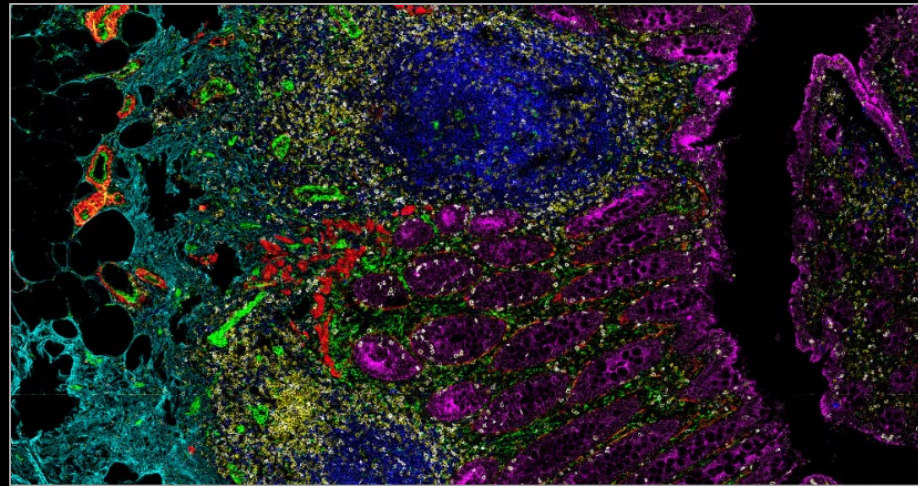
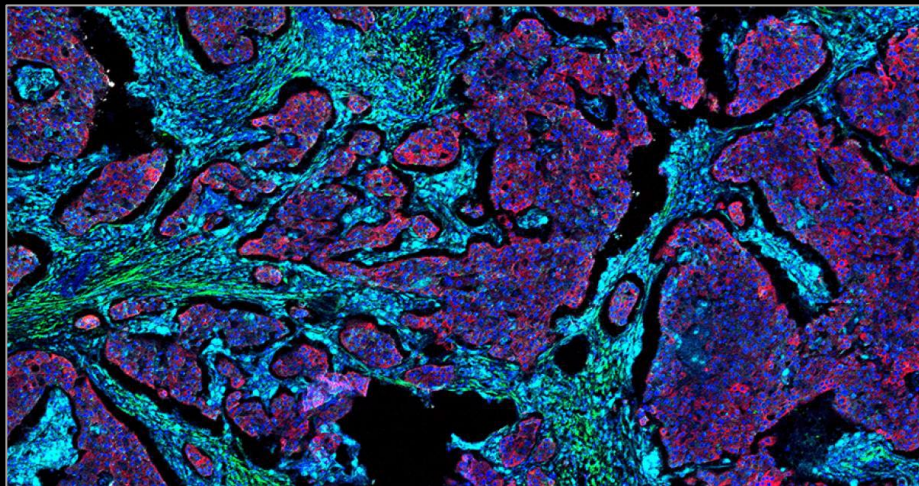


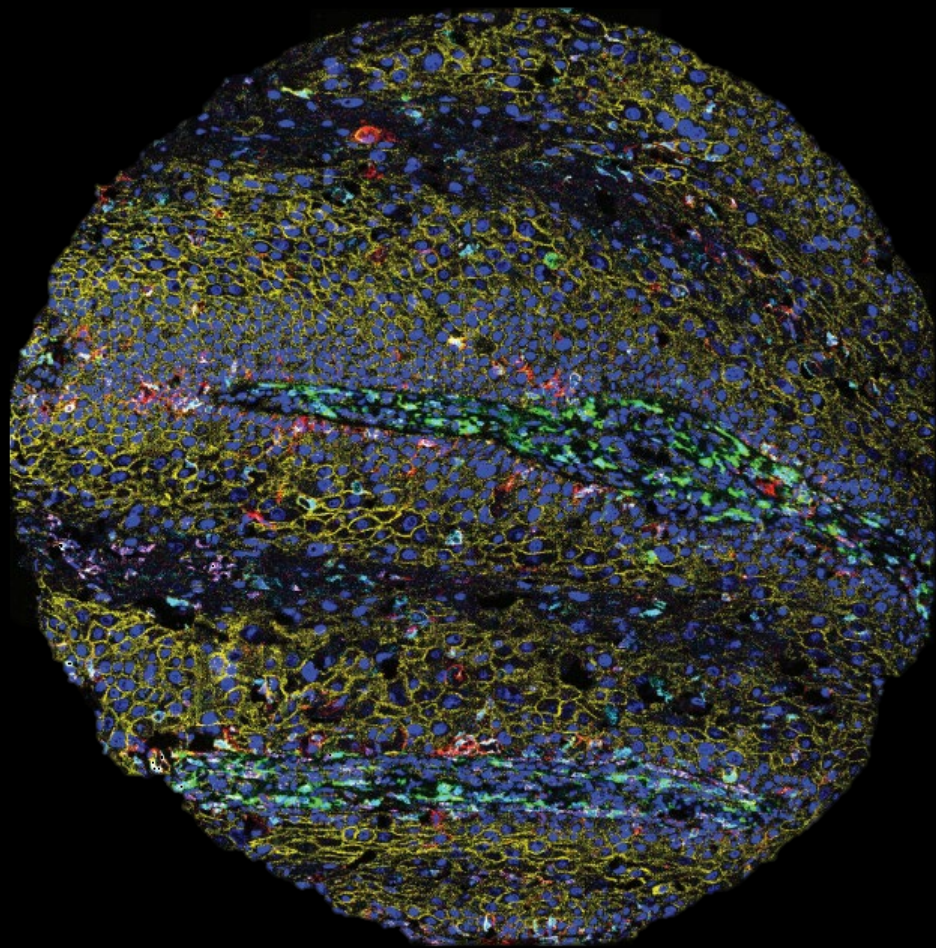
Code

```
hc <- x |>  
  scale() |>  
  dist(method = "euclidean") |>  
  hclust(method = "complete")  
  
cutree(hc, k = 2)
```

```
##   [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1  
1 1 1 1 1 1 2 1 1 1 1 1 2  
##   [38] 1 1 1 1 1 1 1 1 1 1 1 2 2 1 2 2 2 2 2 2 2 2 2  
2 2 2 2 2 2 2 2 2 2 2 2 2  
##   [75] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2  
2 2
```

Highly multiplexed in situ imaging cytometry





FXIIIa

CD66

CD68

Tim3

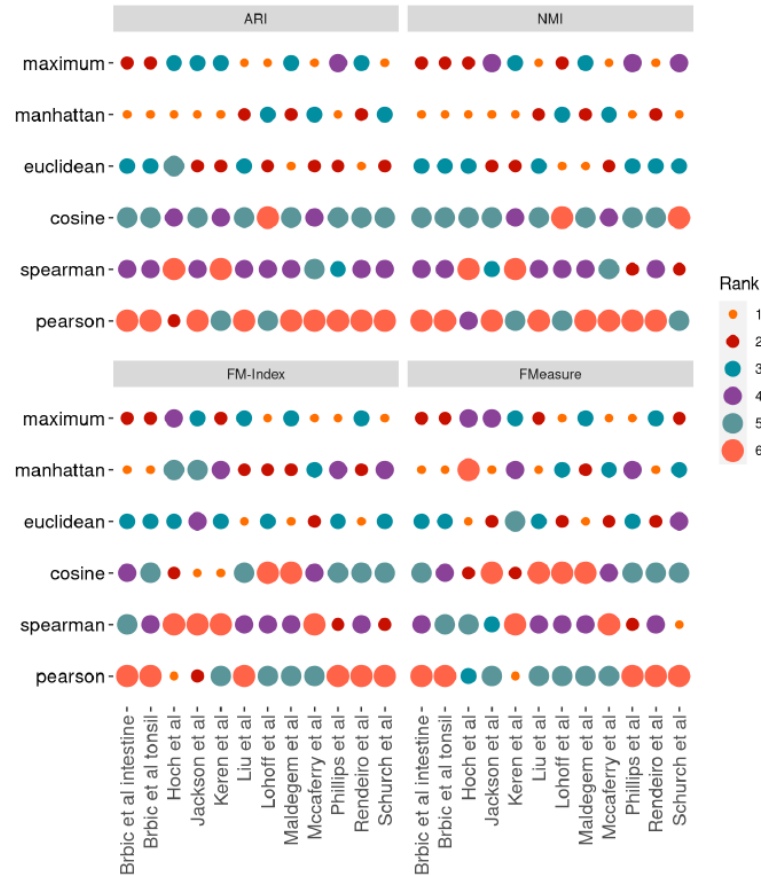
HLA-DR

DNA

Clustering highly multiplexed imaging data



Elijah Willie



When and why?

When and why?

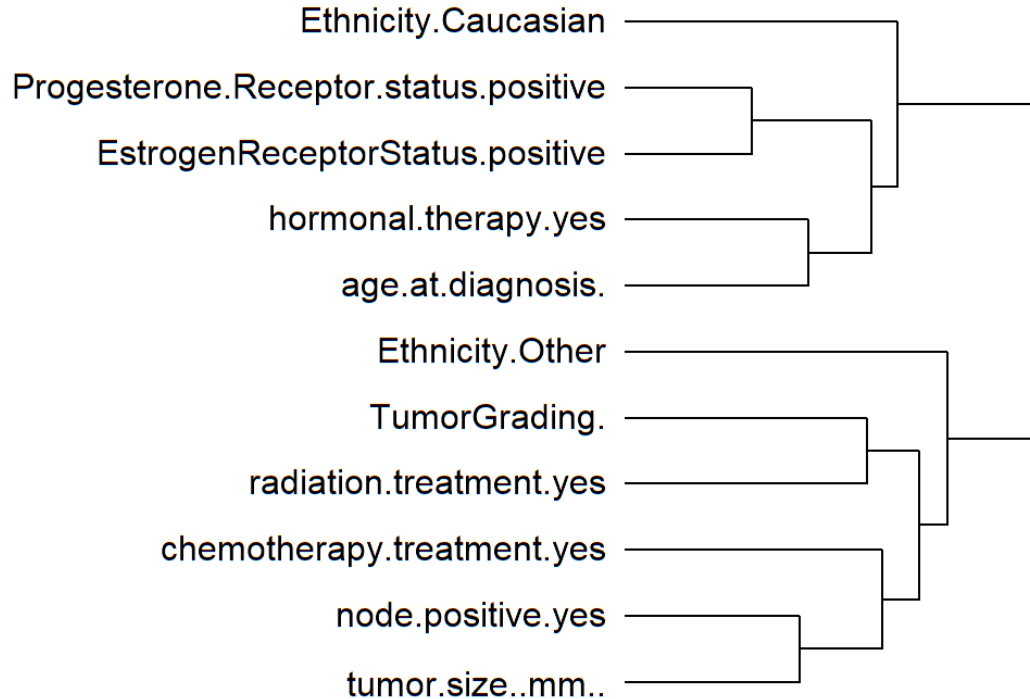
Some advantages of hierarchical clustering:

1. Don't need to know how many clusters you're after.
2. Can cut hierarchy at any level to get any number of clusters.
3. Easy to interpret hierarchy for particular applications.
4. Deterministic.

Some advantages of k-means clustering:

1. Can be much faster than hierarchical clustering, depending on data.
2. Nice theoretical framework.
3. Can incorporate new data and reform clusters easily.

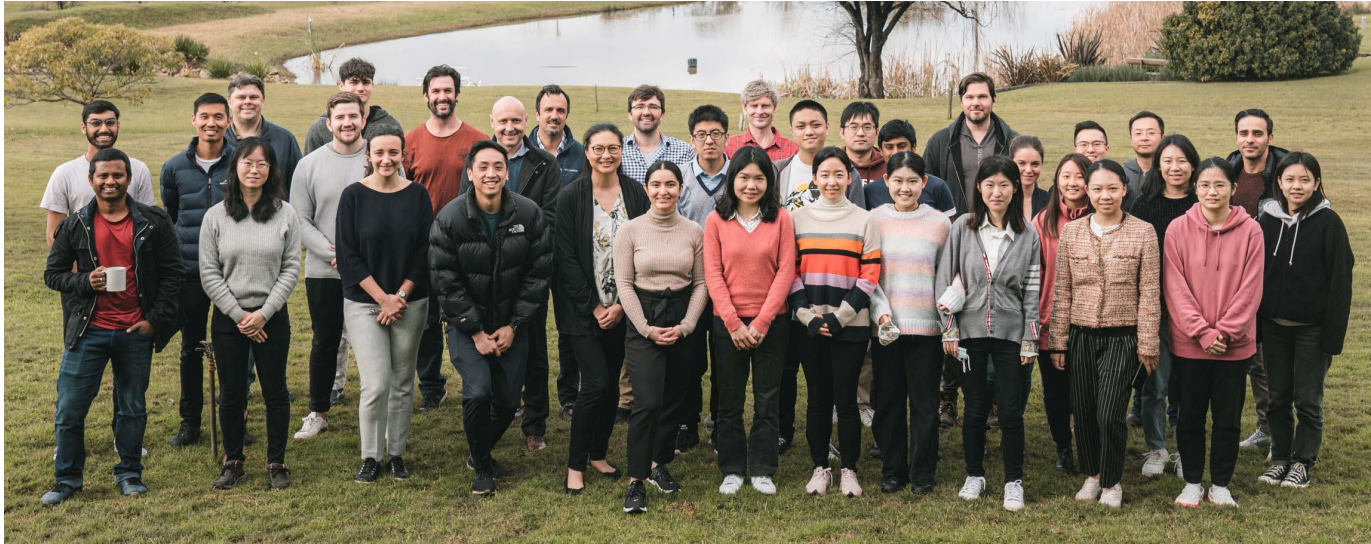
Quantitative clustering



Sydney Precision Data Science Centre

Precision Bioinformatics Cluster

We share an interest in developing statistical and computational methodologies to tackle the foremost significant challenges posed by modern biology and medicine.



Find out more: <http://www.maths.usyd.edu.au/bioinformatics/>

Shiny apps: <http://shiny.maths.usyd.edu.au/>

Github: <https://github.com/SydneyBioX>

Thanks!