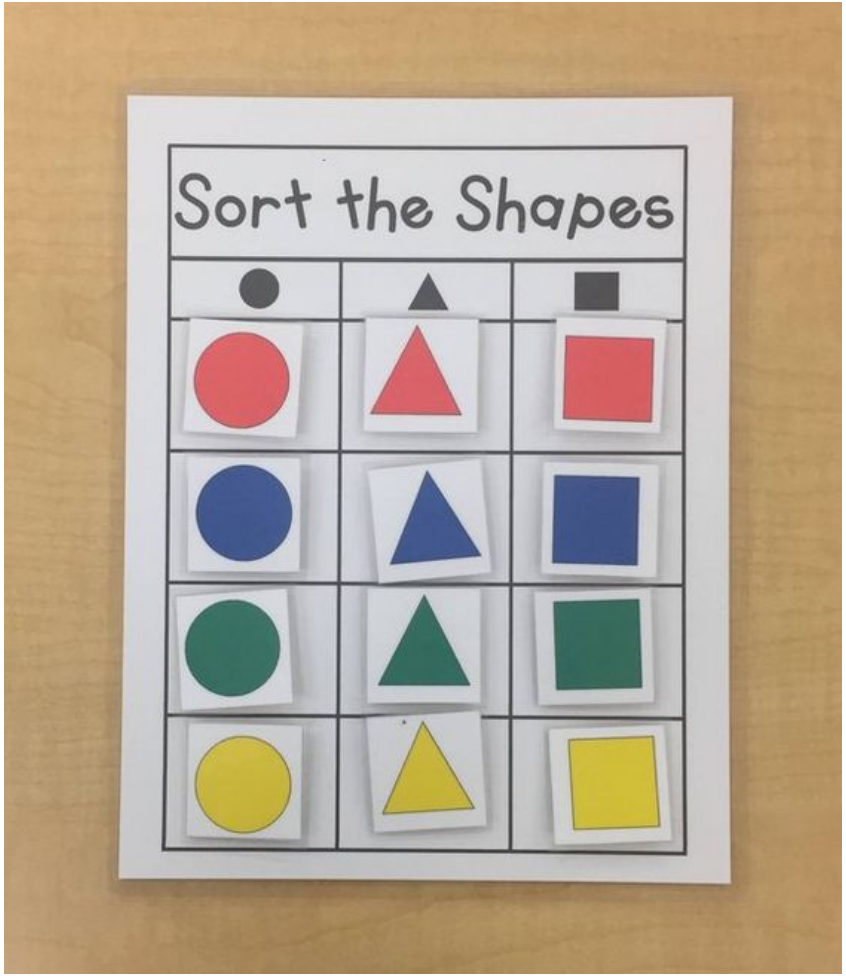
Clustering

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



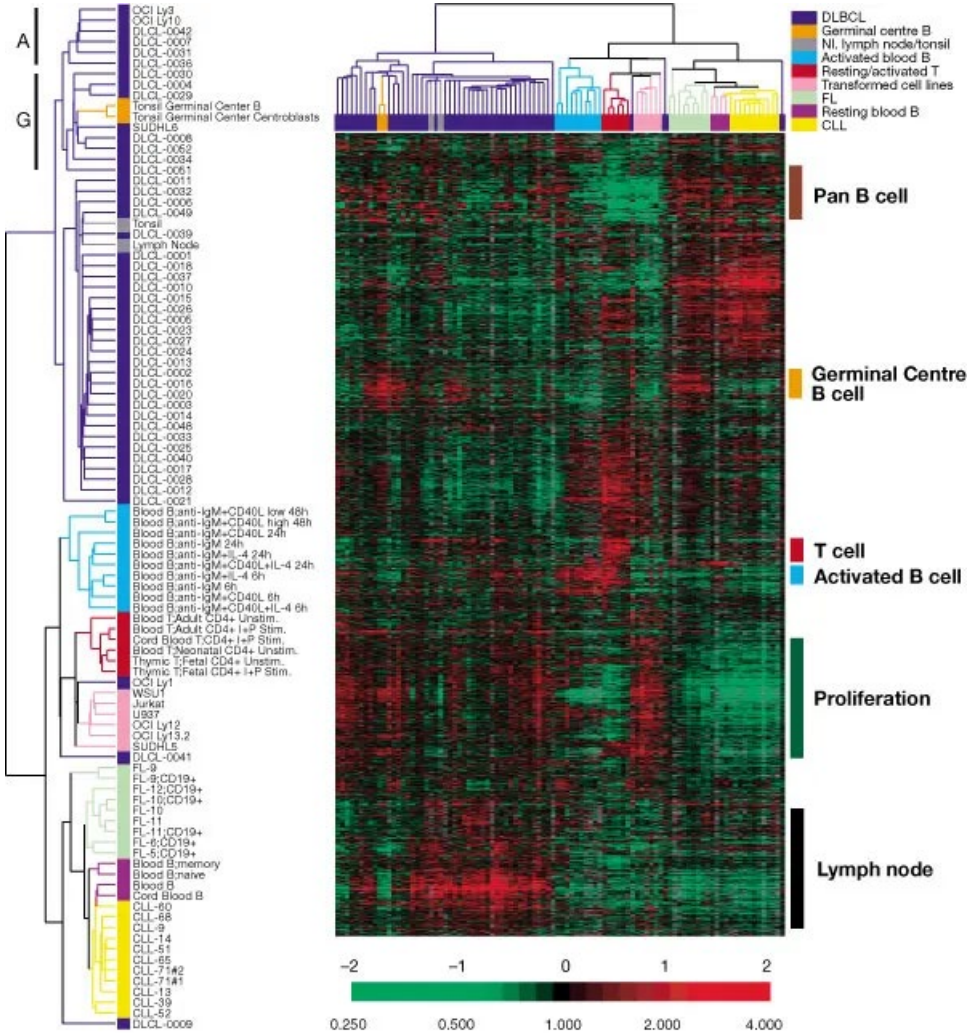
THE UNIVERSITY OF SYDNEY

Bioinformatics

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

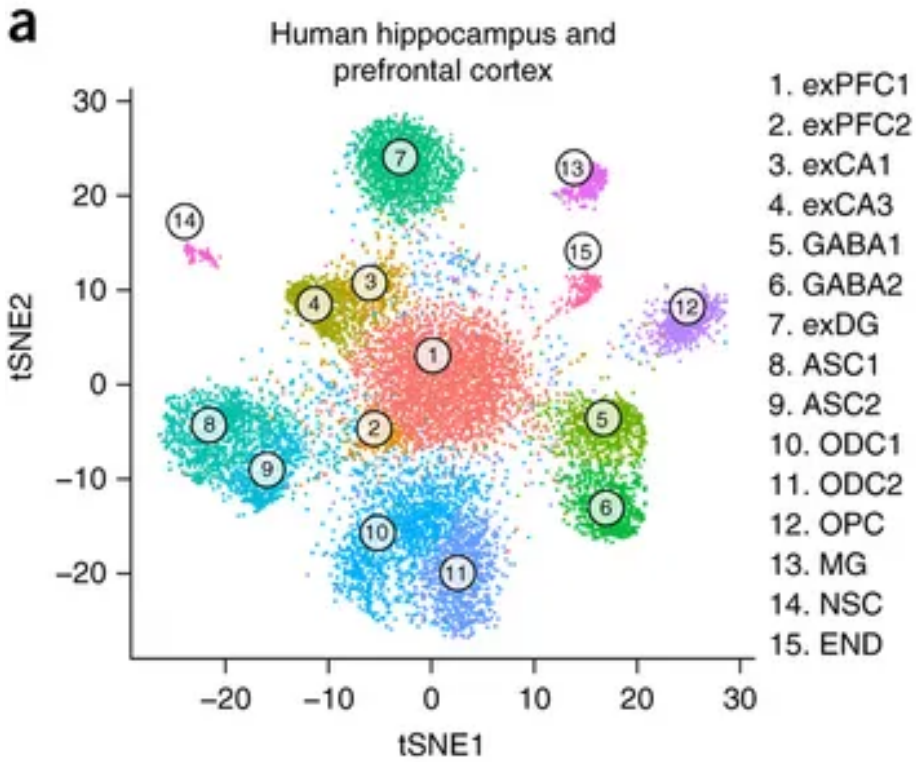
– “Wikipedia”

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



Habib, Naomi, et al. "Massively parallel single-nucleus RNA-seq with DroNc-seq." *Nature methods* 14.10 (2017): 955.

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		

Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. (PMID:17157792)

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		

Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. (PMID:17157792)

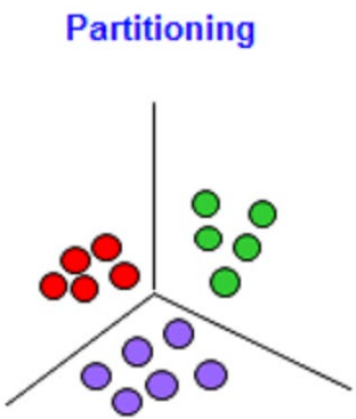
Breast cancer

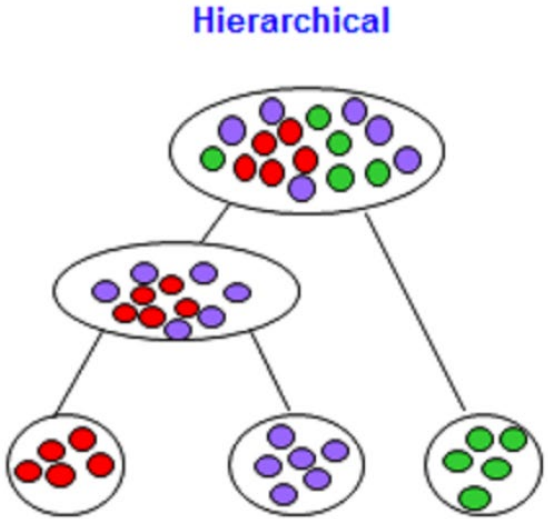
- 130 patients. Consider 10 variables:

node.positive	hormonal.therapy	chemotherapy.treatment
age.at.diagnosis	Ethnicity	tumor.size..mm
TumorGrading	radiation.treatment	Progesterone.Receptor.status
EstrogenReceptorStatus		

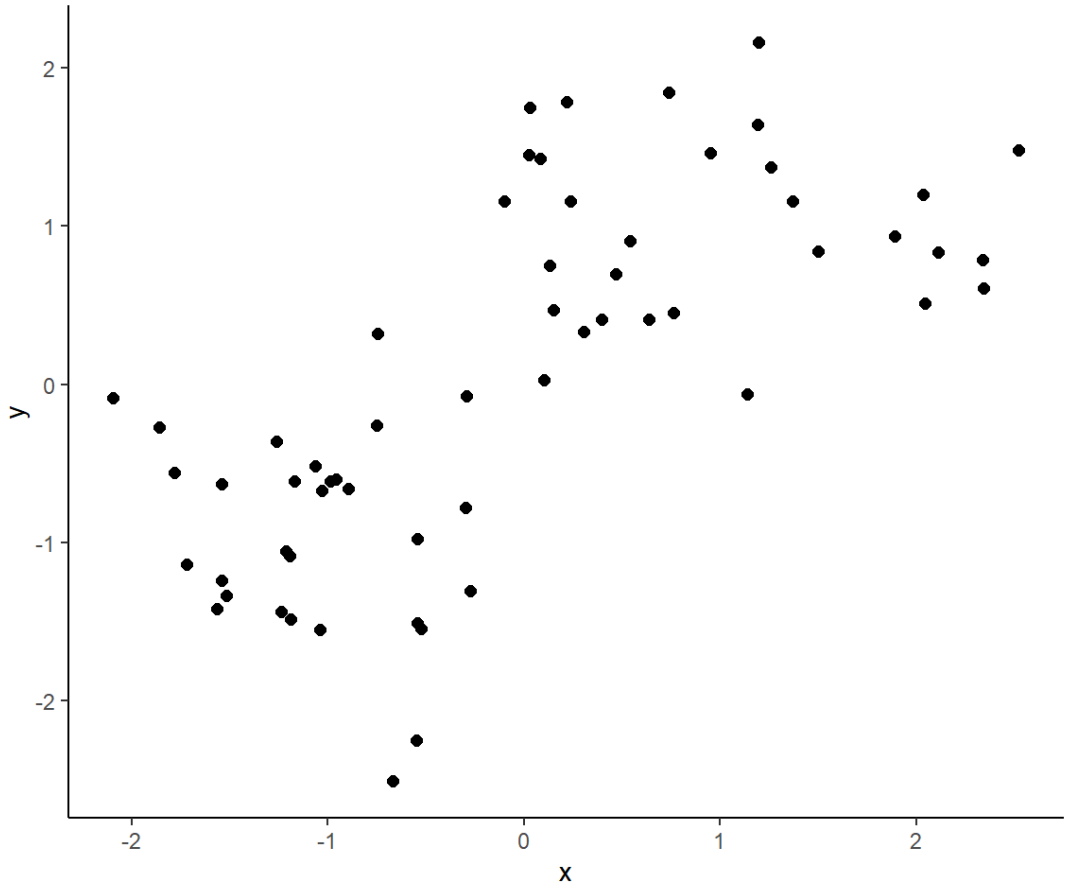
Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. (PMID:17157792)

Unsupervised clustering





K-means clustering


2
1
0
-1
-2
y
-2
-1
0
1
2
x

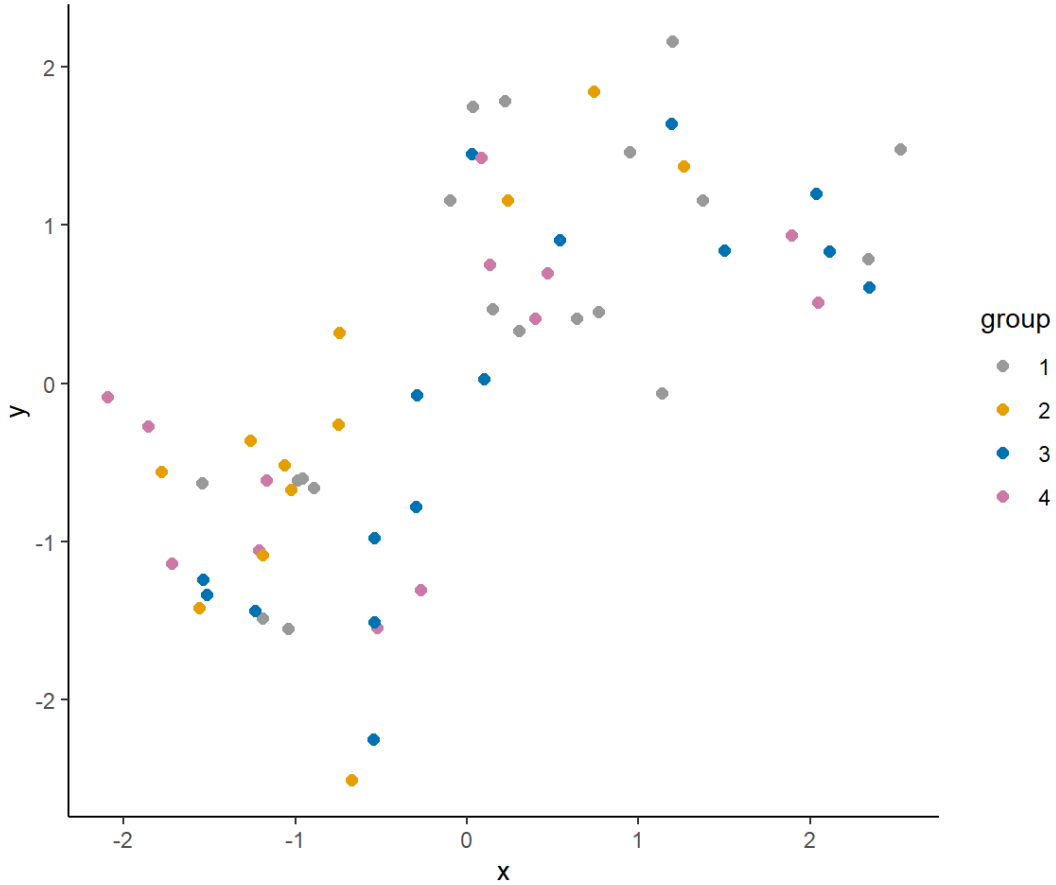
K-means clustering

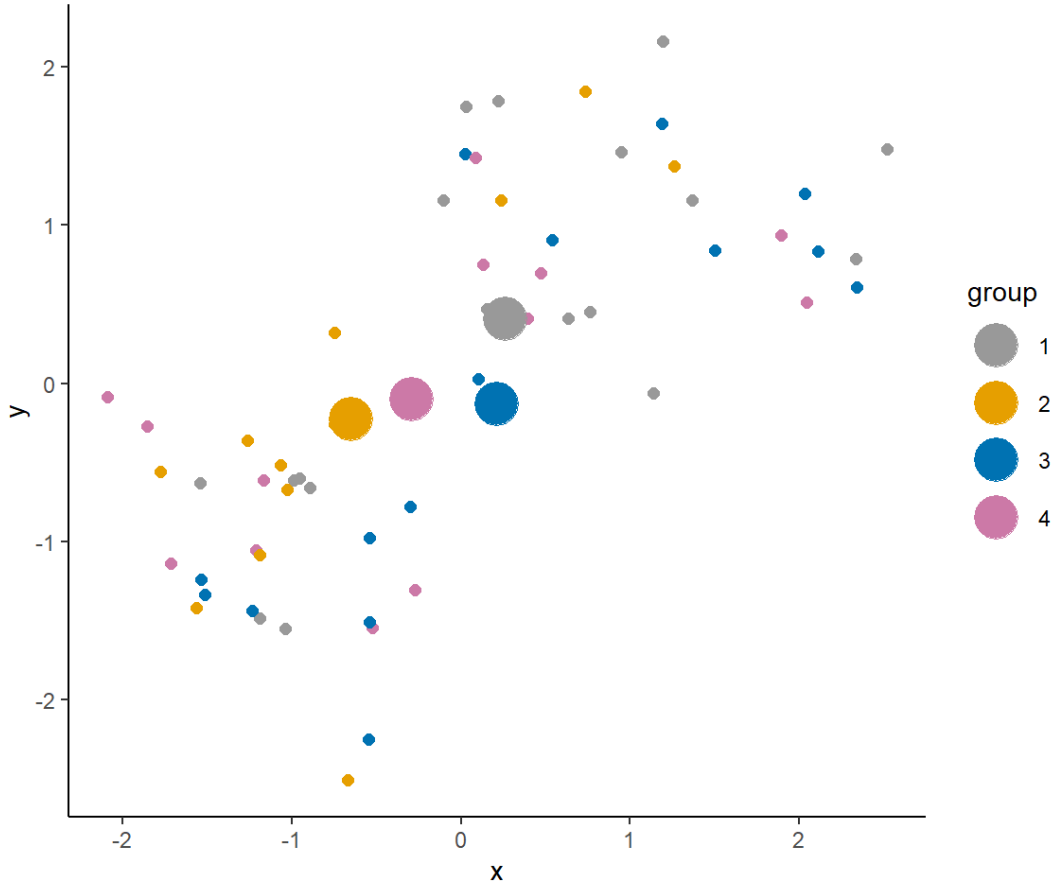
Choose K and then...

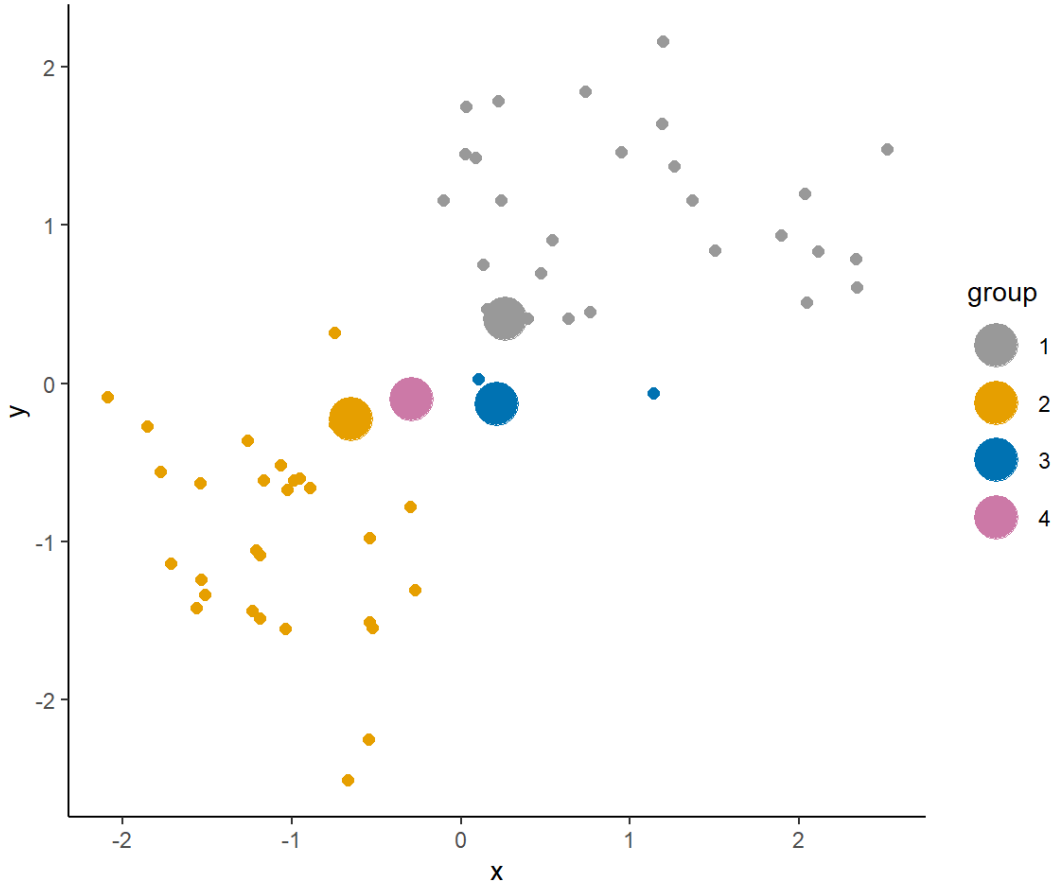
1. Randomly assign a number, from 1 to K, to each of the observations. (These serve as initial cluster assignments for the observations).

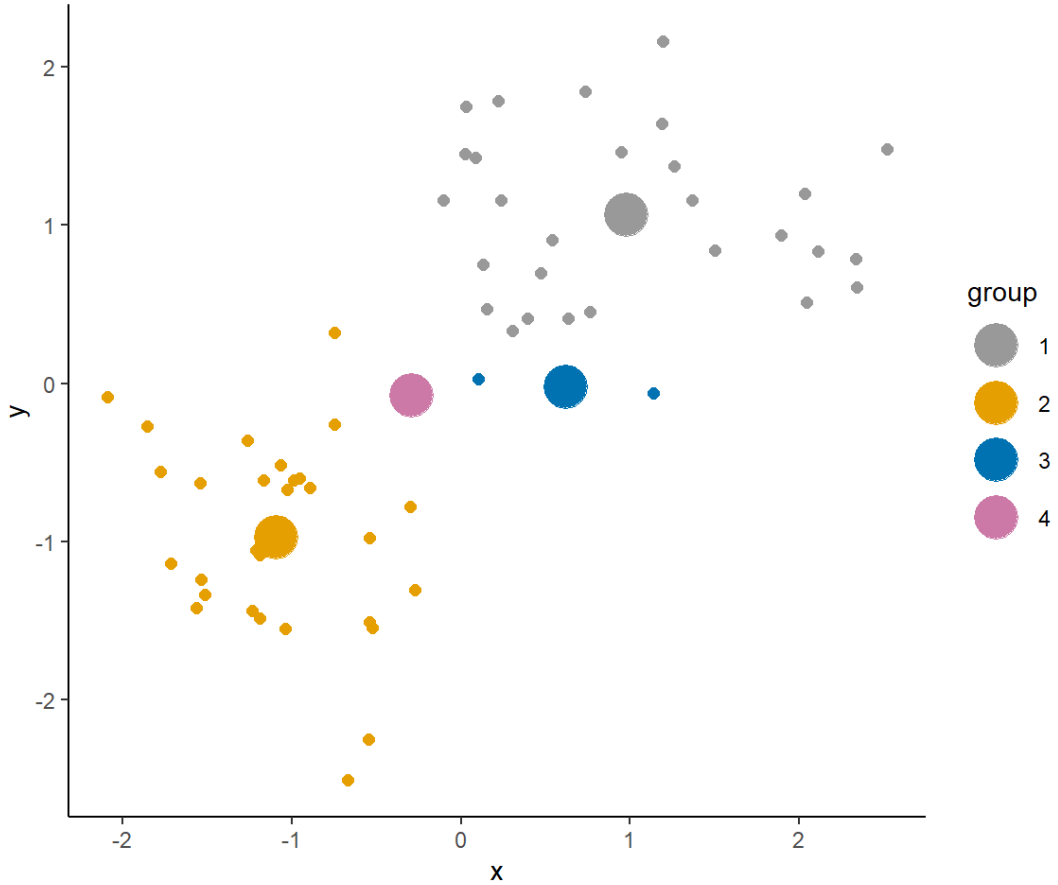
2. Iterate until the cluster assignments stop changing.

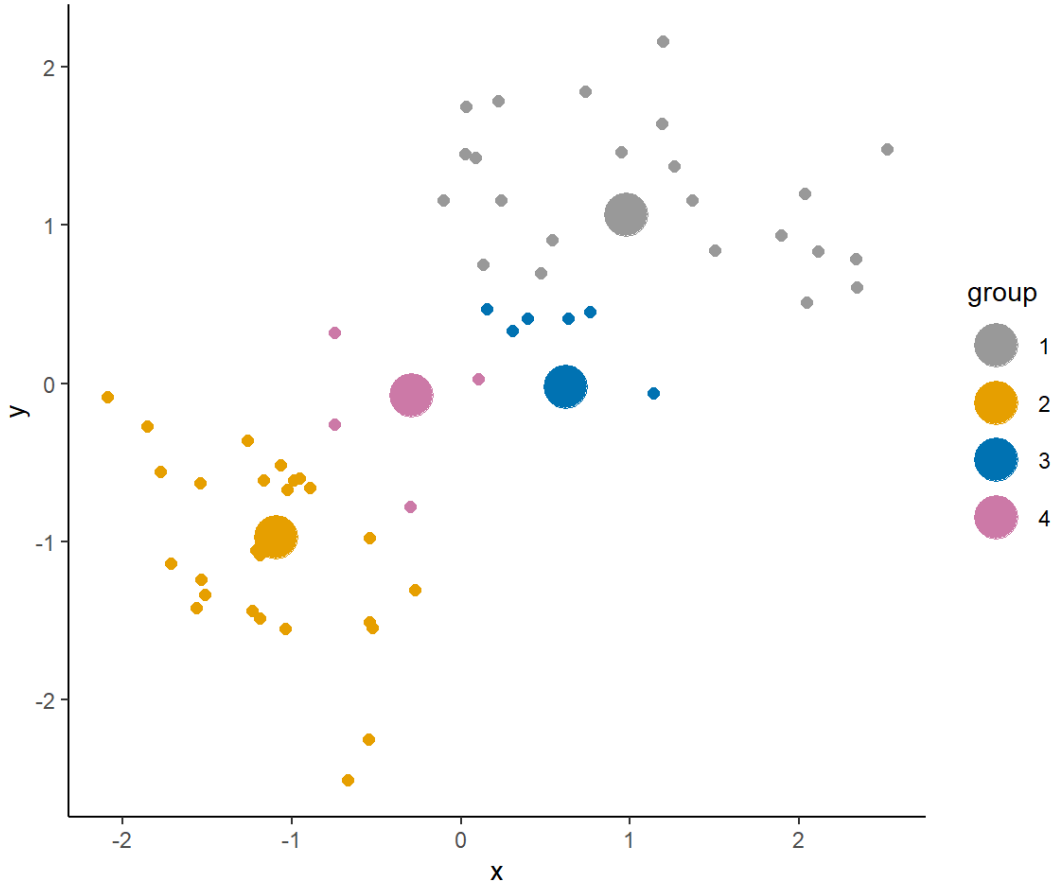
    a) For each of the K clusters, computer the cluster centroid. The kth cluster centroid is the vector of the p covariate means for the observations in the kth cluster.
    b) Assign each observation to the cluster whose centroid is closest (where closest is deemed using Euclidean distance).

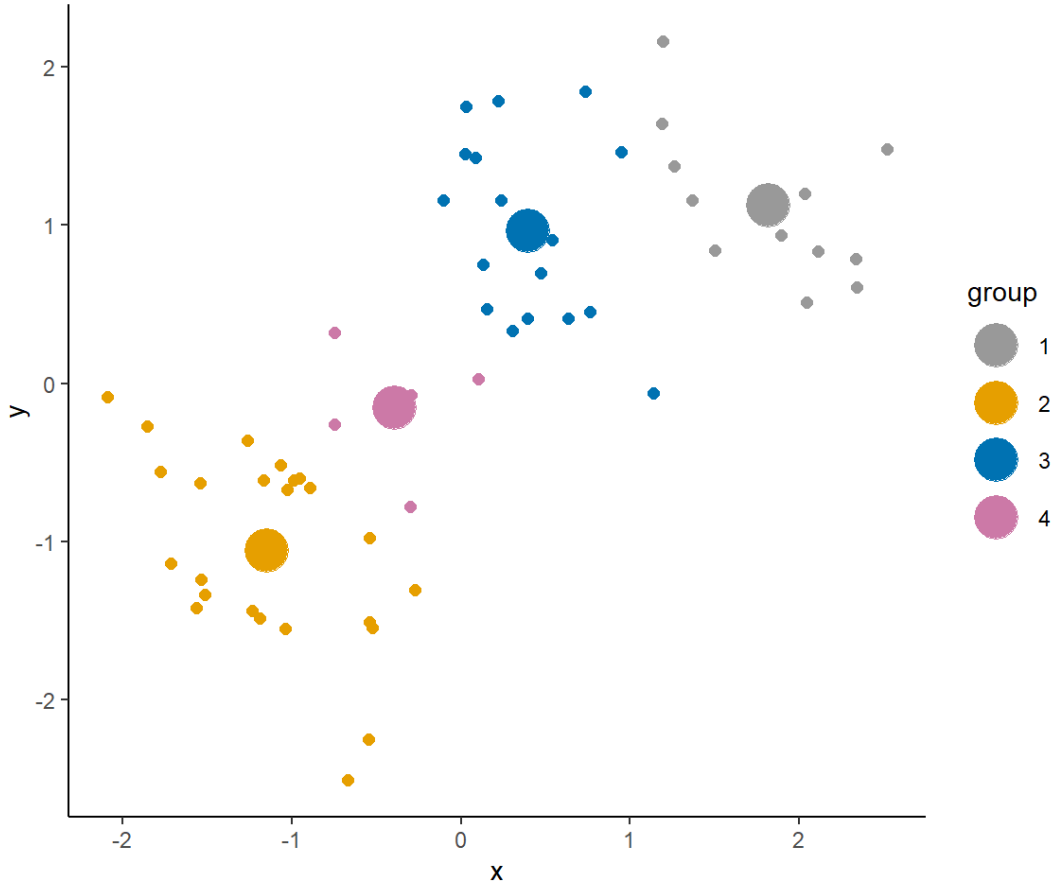

2
1
0
-1
-2
y
-2
-1
0
1
2
x
group
1
2
3
4

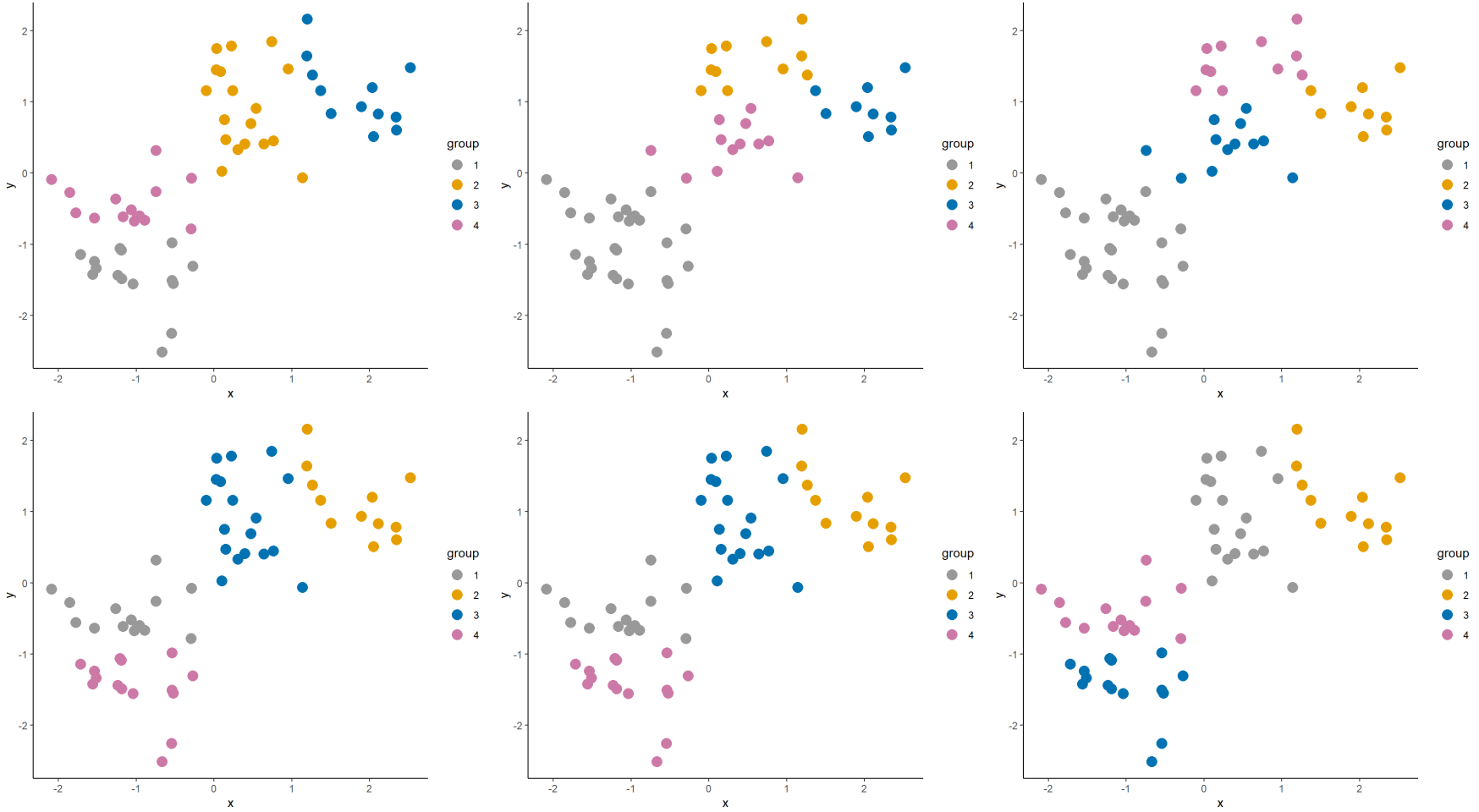

2
1
0
-1
-2
y
-2
-1
0
1
2
x
group
1
2
3
4

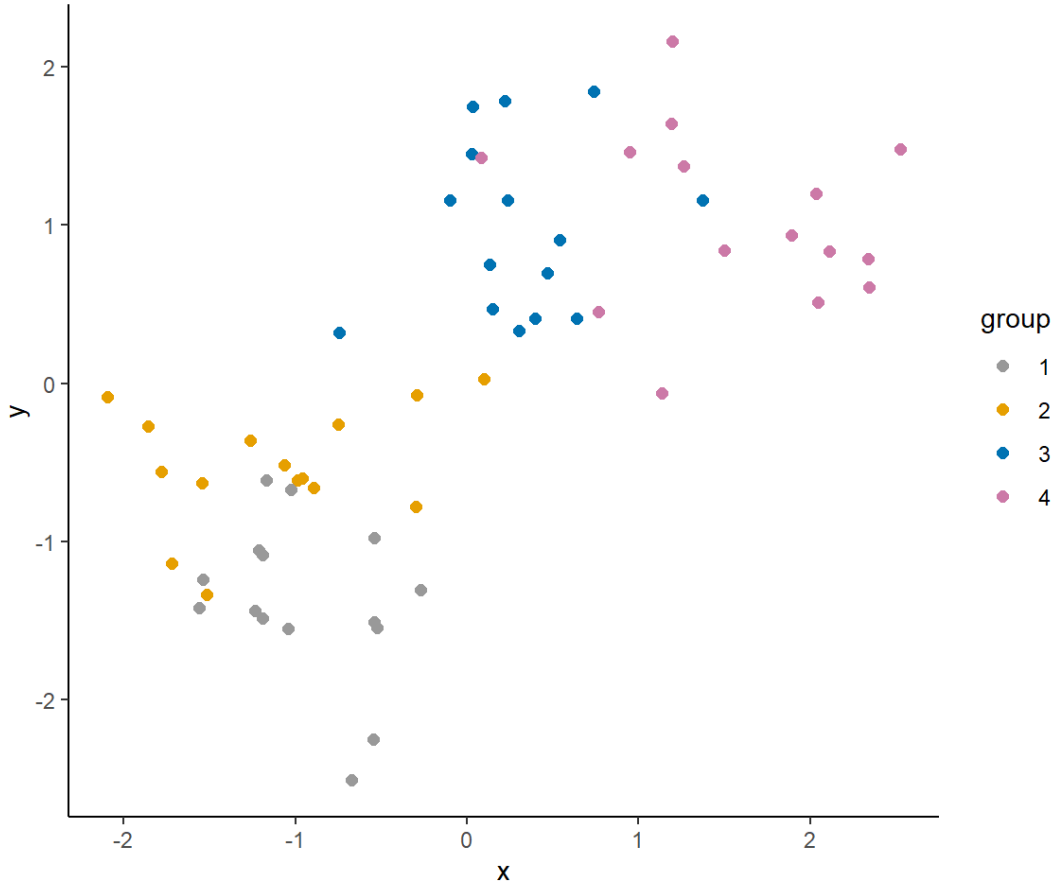

2
1
0
-1
-2
y
-2
-1
0
1
2
x
group
1
2
3
4


2
1
0
-1
-2
y
-2
-1
0
1
2
x
group
1
2
3
4


2
1
0
-1
-2
y
-2
-1
0
1
2
x
group
1
2
3
4


2
1
0
-1
-2
y
-2
-1
0
1
2
x
group
1
2
3
4


y
x
group
1
2
3
4


2
1
0
-1
-2
y
-2
-1
0
1
2
x
group
1
2
3
4

Hierarchical clustering

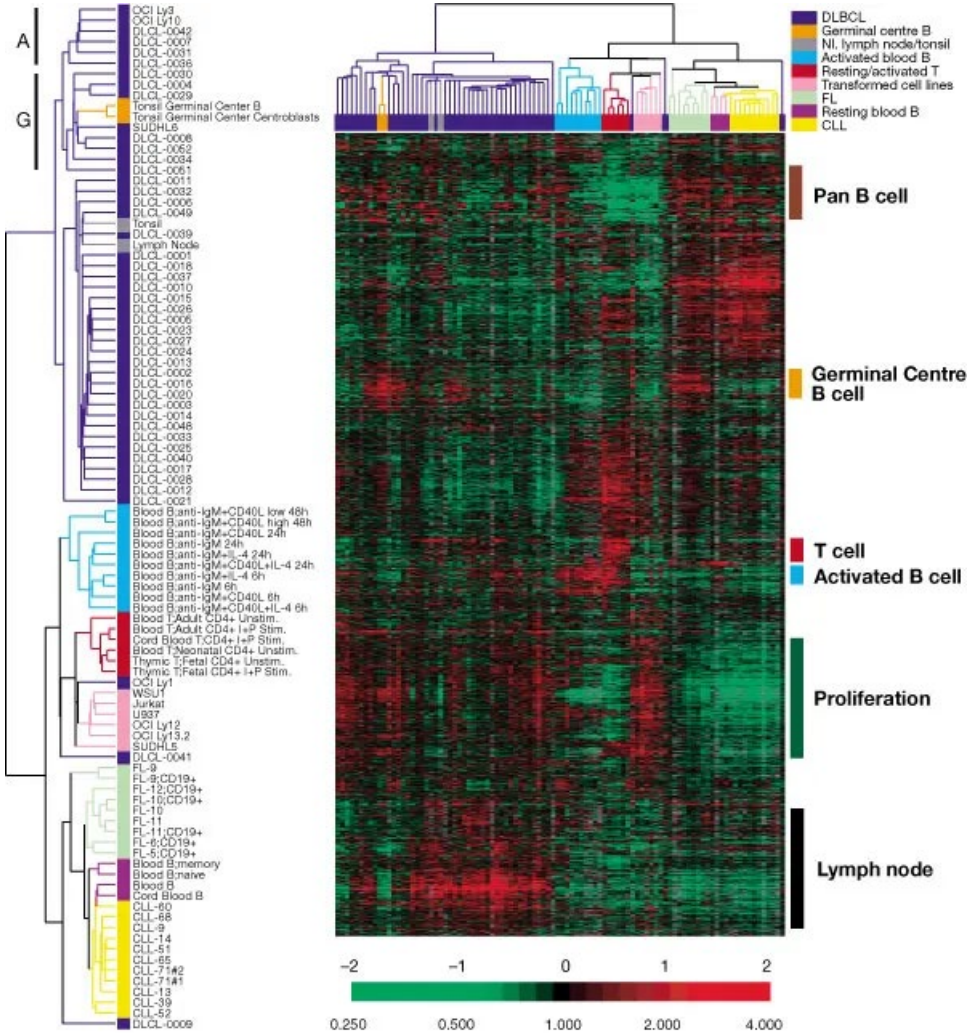
Hierarchical clustering

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.

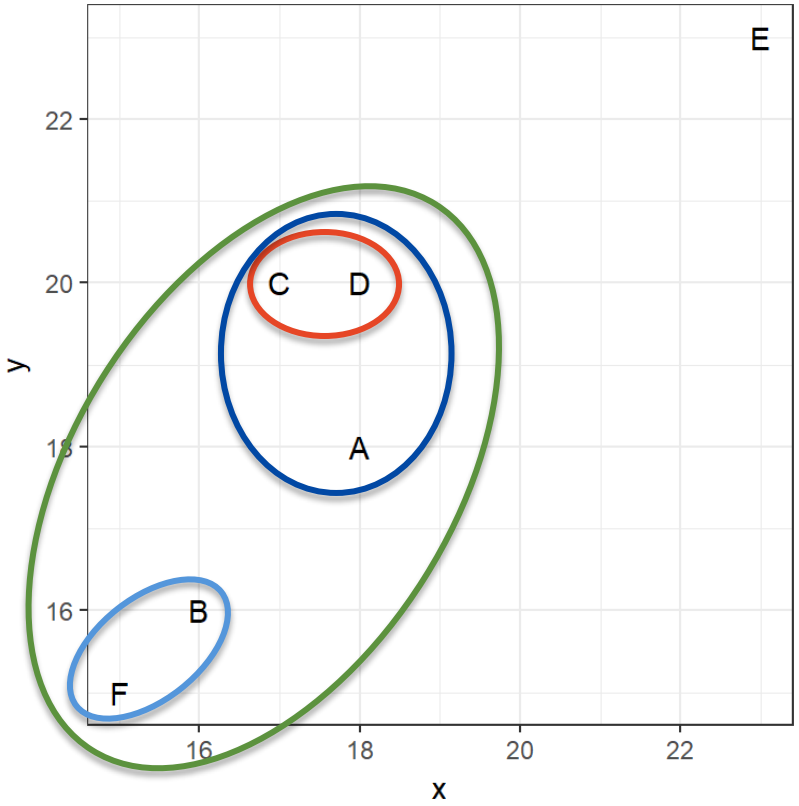
The tree can be built in two distinct ways

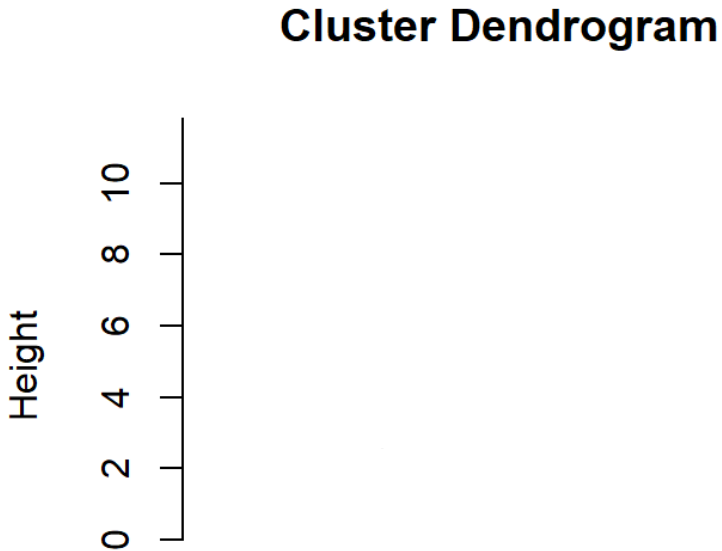
1. bottom-up: agglomerative clustering.
2. top-down: divisive clustering.



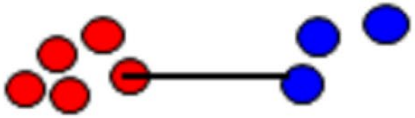
Alizadeh, Ash A., et al. "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling." *Nature* 403. 6769 (2000): 503.

Example 1


E
C
D
A
B
F
22
20
18
16
y
16
18
20
22
x


Cluster Dendrogram
Height
0
2
4
6
8
10

Between cluster similarity measures



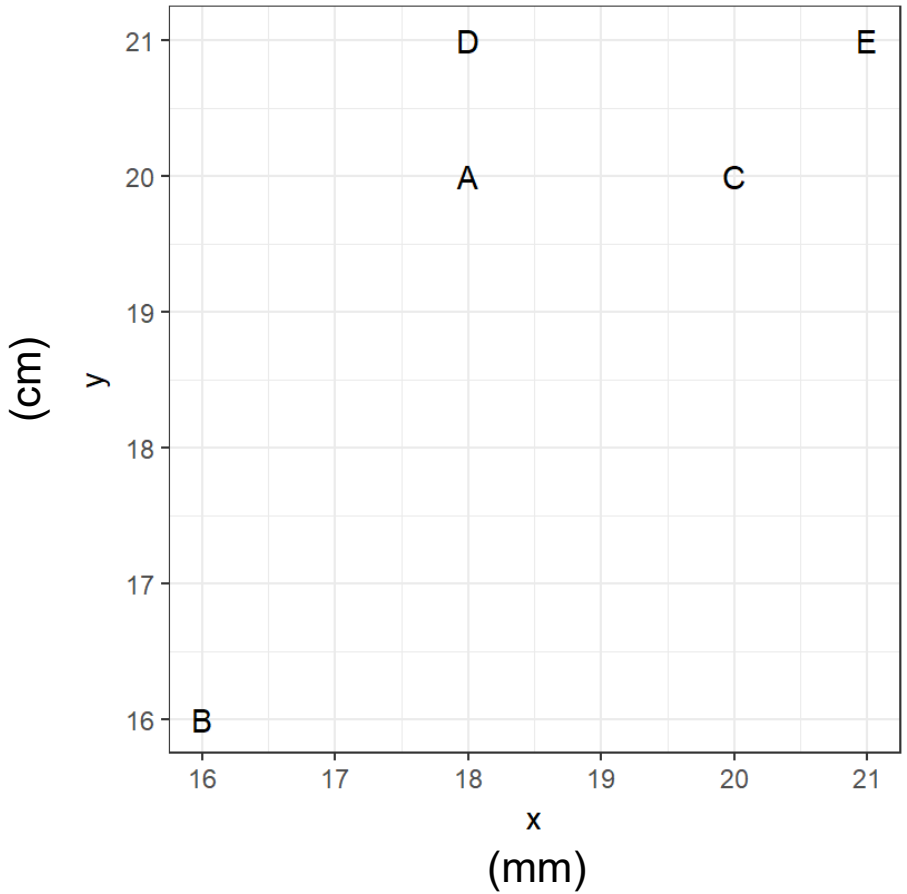
Single (minimum)

Complete (maximum)

Distance between centroids

Average (Mean) linkage

Example 2



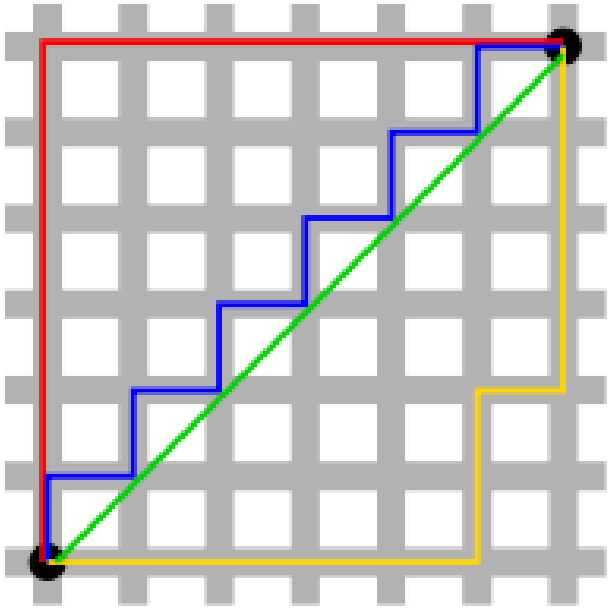
Distance metrics

- Euclidean

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

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

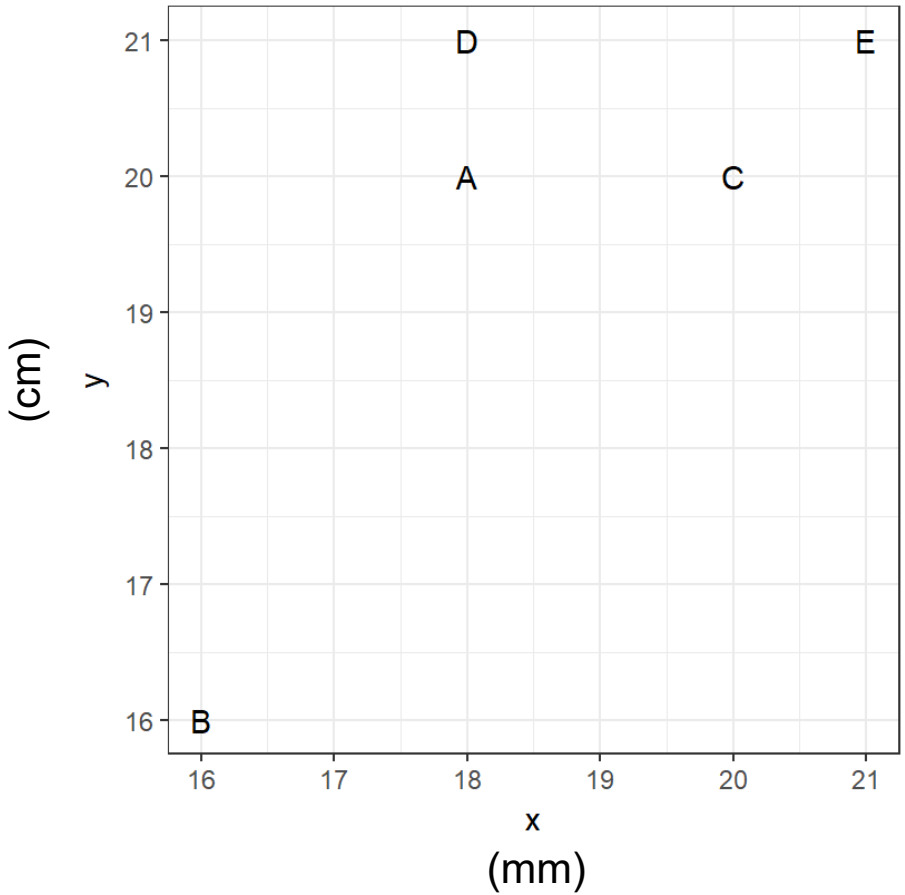
- Manhattan



Green = 8.49

Red, blue, yellow = 12

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



When and why?

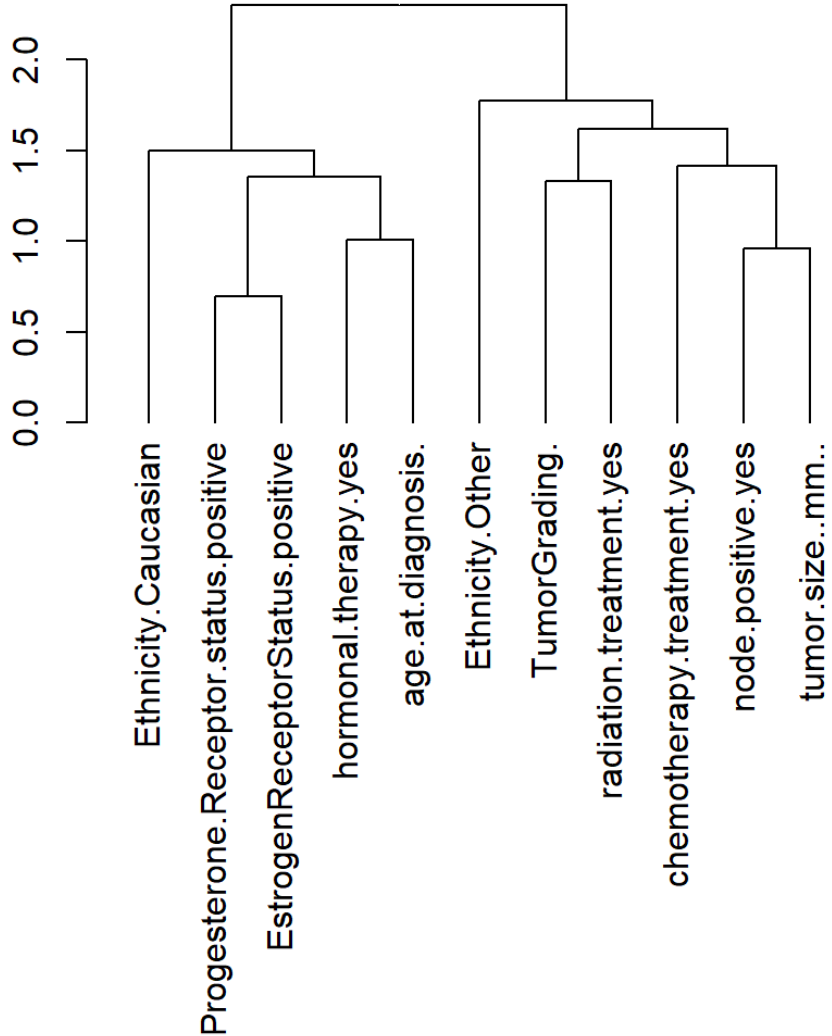
When and why?

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.


0.0
0.5
1.0
1.5
2.0
Ethnicity.Caucasian
Progesterone.Receptor.status.positive
EstrogenReceptorStatus.positive
hormonal.therapy.yes
age.at.diagnosis.
Ethnicity.Other
TumorGrading.
radiation.treatment.yes
chemotherapy.treatment.yes
node.positive.yes
tumor.size..mm..

Sydney Precision Bioinformatics Research Alliance

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

Row 2: John Ormerod; Jean Yang; Samuel Mueller; Garth Tarr; Rachel Wang
Row 1: Ellis Patrick; Pengyi Yang

Thanks!

AMED3002

Interrogating biomedical and health data

A unit in both the Applied Medical Science major and the Data Science major.

Your data science skills could cure cancer

Develop the analytical skills required to work with data obtained in the medical and diagnostic sciences.

Biotechnological advances have given rise to an explosion of original and shared public data relevant to human health. These data, including the monitoring of expression levels for thousands of genes and proteins simultaneously, together with multiple databases on biological systems, now promise exciting, ground-breaking discoveries in complex diseases. Critical to these discoveries will be our ability to unravel and extract information from these data.

Study at Westmead

Immerse yourself in the largest medical precinct in the country

The undergraduate courses at Westmead are harnessing the expertise and specialised facilities of the research institutes and hospitals of the area to educate our future scientific leaders. The translational research hub at The University of Sydney Westmead campus is one of the few places in the world where your learning and application can occur in a medical research and hospital environment.