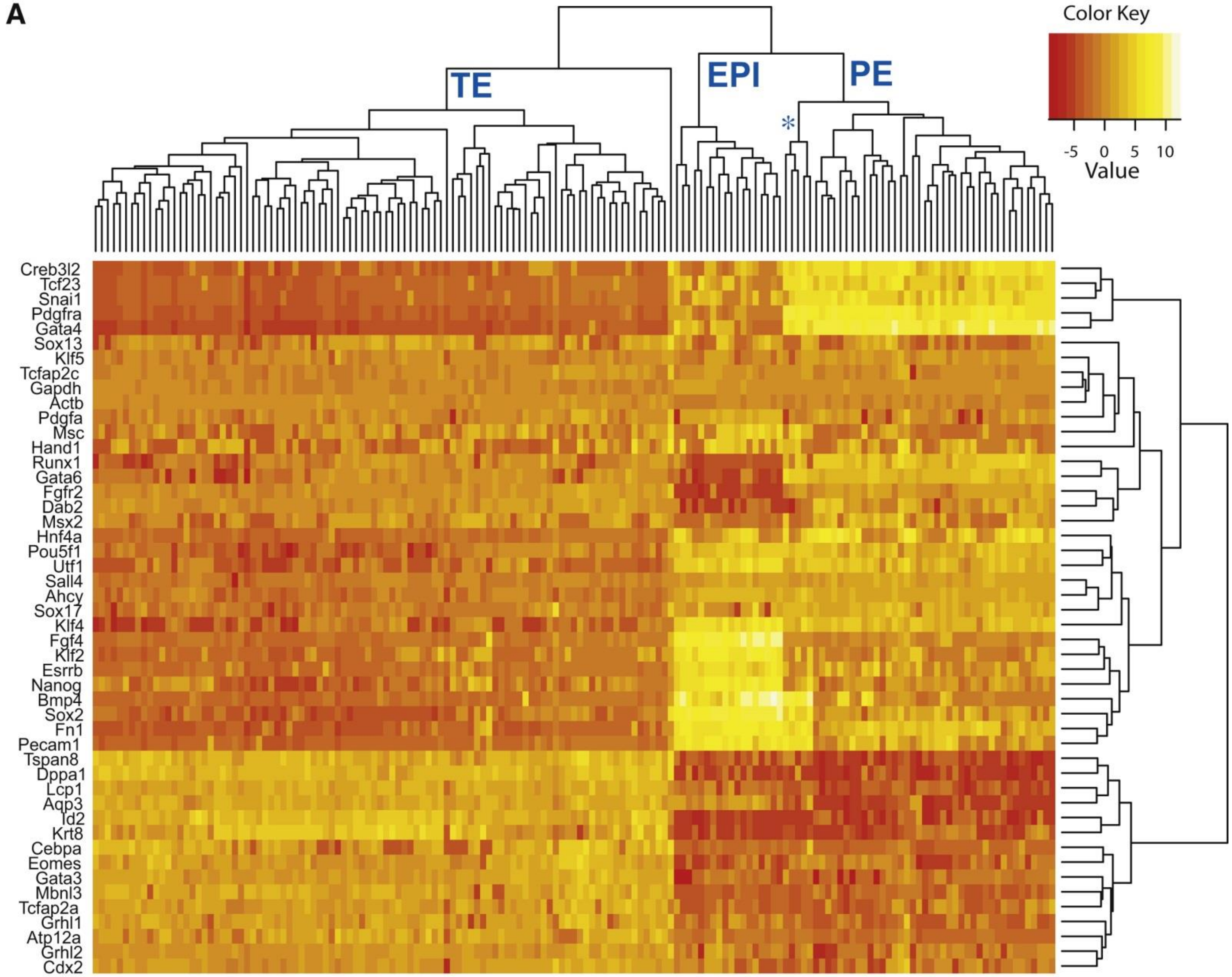
Clustering high-dimensional data

Magnus Rattray and Alexis Boukouvalas
Faculty of Biology, Medicine and Health
University of Manchester

Fig. 1A of Guo *et al.* shows clusters of cells and genes

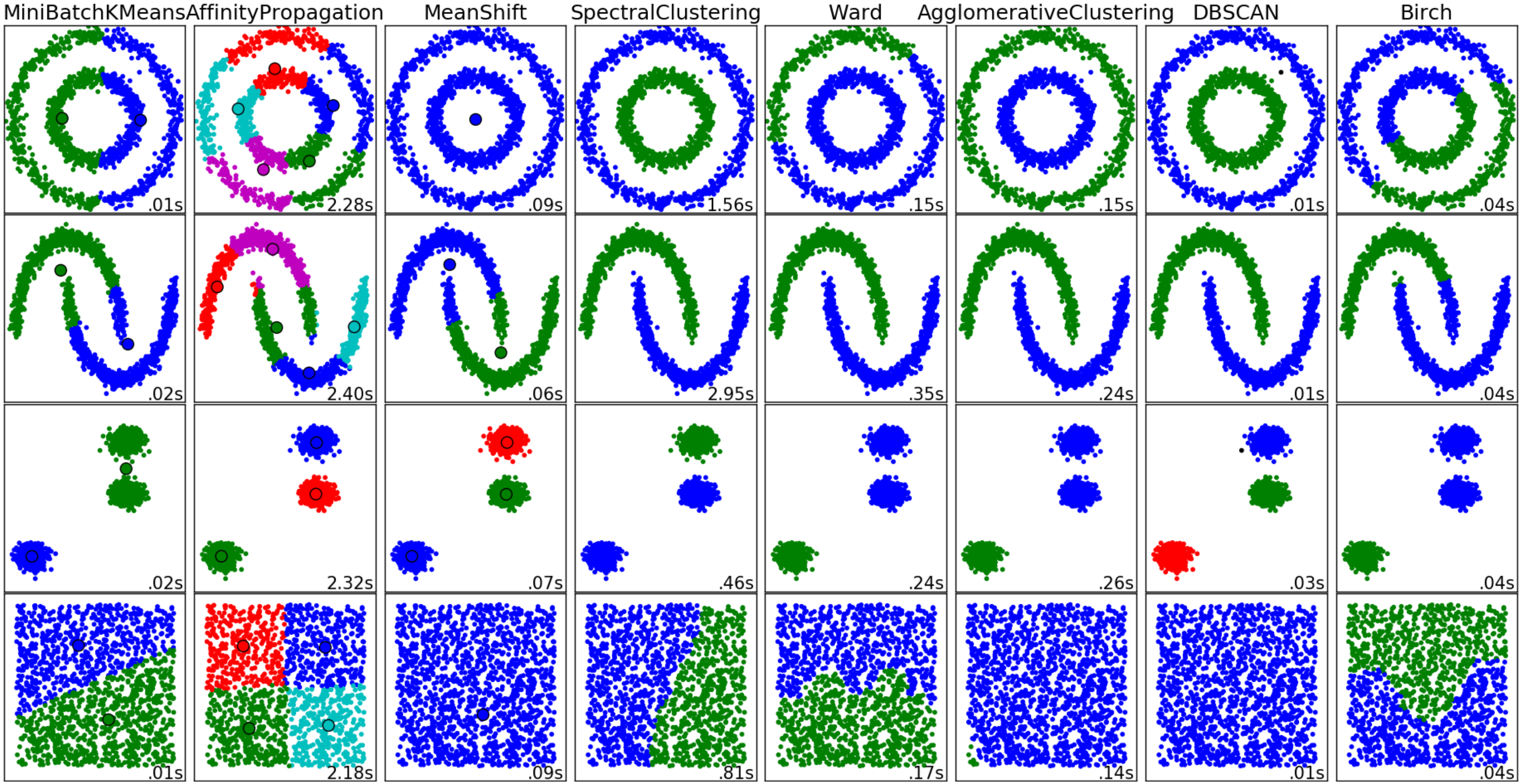


Popular approaches to clustering

- Agglomerative hierarchical clustering
  - Progressively merge closest items/groups
  - No need to define particular number of clusters
- K-means clustering
  - Identifies K groups of similar items
  - Maximizes within-group similarity
- Model-based clustering
  - Learn a model to best explain the data
  - Allows for soft *probabilistic* clustering
  - Bayesian methods to determine optimal K

Popular approaches to clustering

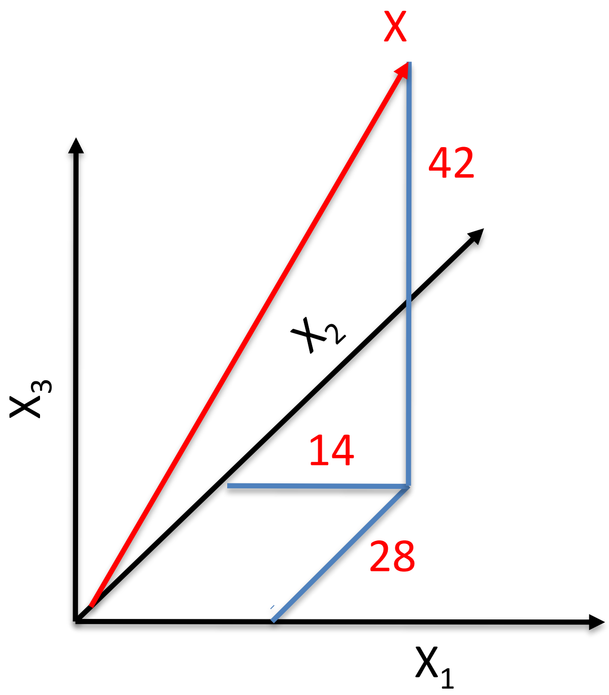
Many more – http://scikit-learn.org/stable/modules/clustering.html



Similarities and distances

- Many clustering algorithms require a quantity representing *similarity* or *distance*
- A common choice is the **Euclidean distance**
- This is the distance between two data vectors $X = [X_1, X_2, ..., X_D]$ and $Y = [Y_1, Y_2, ..., Y_D]$

Recall - data represented as vectors

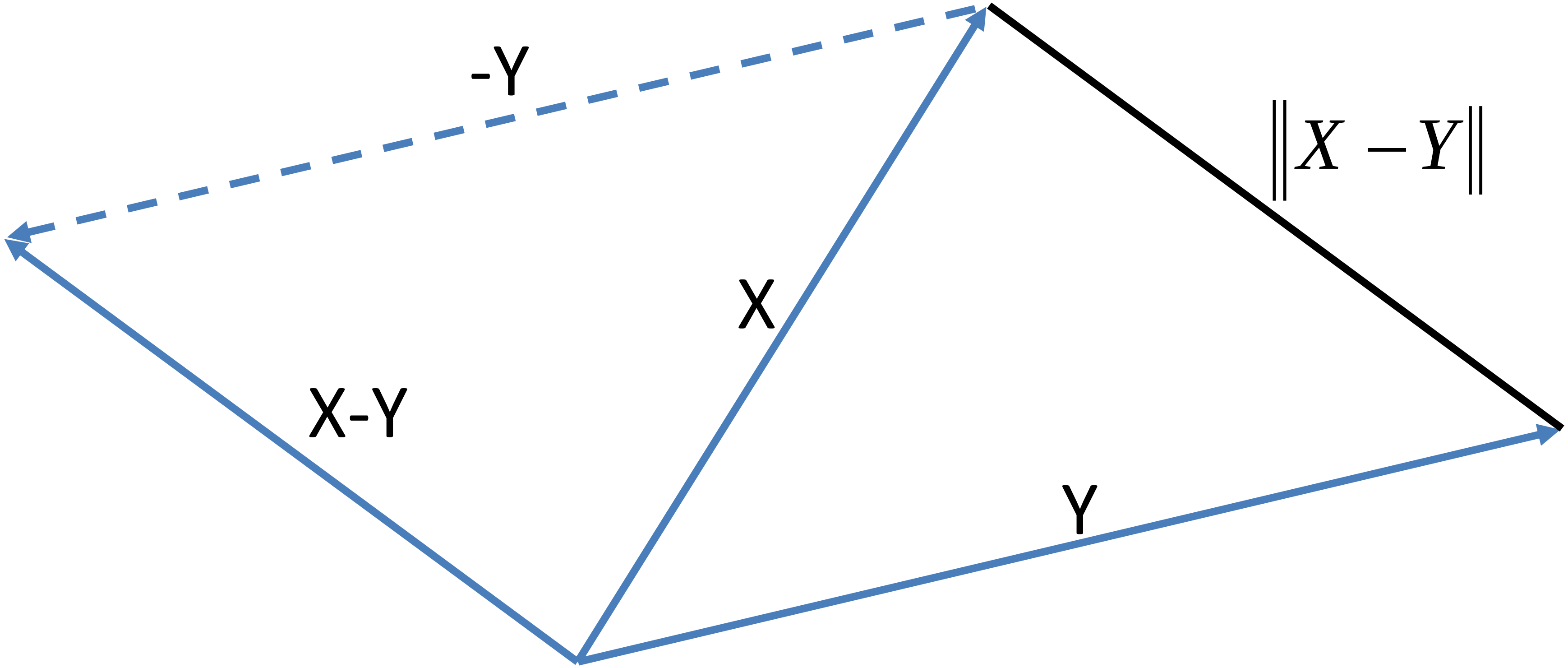


$X = [X_1, X_2, X_3]$

$X = [14,28,42]$

X_1 is value of feature 1, X_2 is value of feature 2 etc.

Euclidean distance between vectors



$$D(X,Y) = \|X - Y\| = \sqrt{\sum_{i=1}^{D} (X_i - Y_i)^2}$$

Euclidean distance

Given data vectors $X = [X_1, X_2, ..., X_D]$ and $Y = [Y_1, Y_2, ..., Y_D]$ the squared distance is:

$$D^2(X,Y) = (X_1 - Y_1)^2 + (X_2 - Y_2)^2 + ... + (X_D - Y_D)^2$$

The distance is the square root of that,

$$D(X,Y) = \|X - Y\| = \sqrt{\sum_{i=1}^{D}(X_i - Y_i)^2}$$

Be careful - some algorithms (e.g. k-means) use the squared distance

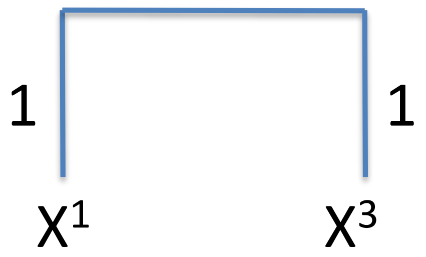
Agglomerative hierarchical clustering

- Very popular approach – especially in biology
- Progressively merge closest data points or clusters of data points
- Requires definition of distance between clusters, e.g. Average Linkage is mean distance between items in each cluster

https://en.wikipedia.org/wiki/Hierarchical_clustering

Average linkage clustering - example

$D(X^n,X^m)$	X^1	X^2	X^3	X^4
X^1	-	3	2	8
X^2	3	-	3	8
X^3	2	3	-	5
X^4	8	8	5	-



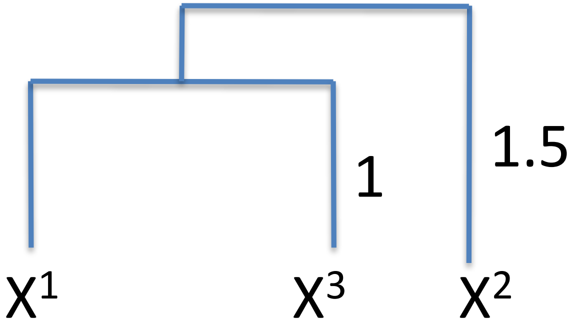
$$D(X^{13},X^2)=\frac{D(X^1,X^2)+D(X^3,X^2)}{2}=\frac{3+3}{2}=3$$

$$D(X^{13},X^4)=\frac{D(X^1,X^4)+D(X^3,X^4)}{2}=\frac{8+5}{2}=6.5$$

	X^{13}	X^2	X^4
X^{13}	-	3	6.5
X^2	3	-	8
X^4	6.5	8	-

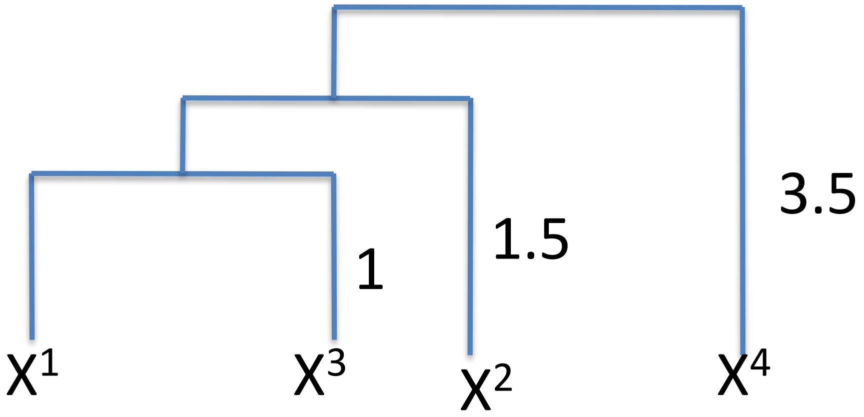
Average linkage clustering - example

	X13	X2	X4
X13	-	3	6.5
X2	3	-	8
X4	6.5	8	-



$$D(X^{123},X^{4})=\frac{D(X^{1},X^{4})+D(X^{2},X^{4})+D(X^{3},X^{4})}{3}=\frac{8+8+5}{3}=7$$

	X123	X4
X123	-	7
X4	7	-



Many different versions

- Average linkage
  - Distance between clusters is average distance between items in each cluster
- Complete linkage
  - Distance between clusters is distance between furthest items in each cluster
- Single linkage
  - Distance between clusters is distance between closest items in each cluster
- Ward linkage
  - Choose splits to minimize sum of squared

K-means clustering

- Optimisation-based method
- Partition data into clusters $k = 1 \ldots K$
- Cluster centre $\mu_k = Mean_{n \in cluster(k)}(X^n)$
- Find centres which minimize objective: sum of within cluster squared distances to centres

$$E = \sum_{k=1}^{K} \sum_{n \in cluster(k)} D^2(X^n, \mu_k)$$

K-means clustering: EM algorithm

Initialize – e.g. select K random points as centres μ_k
Iterate:

1) Identify closest centre for every data point
2) Assign points sharing same centre as a cluster, say $n \in cluster(k)$ for each $n = 1 \ldots N$
3) Compute mean of data in each cluster
   $m_k = Mean_{n\hat{I}\ cluster(k)}(X^n)$

Stop, when centres no longer change

K-means clustering: EM algorithm

- Some nice online demos you can try

  http://util.io/k-means

  http://syskall.com/kmeans.js/

Assessing performance

- Sometimes we know the desired answer, e.g. where data classes are known
- It is useful to then assess the performance of different clustering algorithms, to better understand their properties
- Many metrics have been proposed:

http://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation

In the lab you will use the **Adjusted Rand Index**

https://en.wikipedia.org/wiki/Rand_index

Week 11 lab instructions

- Look through the Iris dataset worked example notebook (IrisClustering.ipynb)

Exercise 1: Use k-means clustering on the Guo *et al.* 64-cell stage data and use PCA to visualize the clustering

Exercise 2: reproduce the two hierarchical clusterings shown in Figures 1A of the paper

In each case assess the performance for different parameter choices