# Data Mining and Machine Learning



## Clustering I



Peter Jančovič



UNIVERSITY OF BIRMINGHAM

# Data Mining

- Objective of Data Mining is to find structure and patterns in large, abstract data sets
  – Is the data homogeneous or does it consist of several separately identifiable subsets?
  – Are there patterns in the data?
  – If so, do these patterns have an intuitive interpretation?
  – Are there correlations in the data?
  – Is there redundancy in the data?





UNIVERSITY OF BIRMINGHAM

# Partitioning data into "clusters"

- In this lecture we will start to develop tools to understand the structure of data that can be partitioned into (more or less) distinct subsets
- Can think of these subsets as arising from distinct "sources"
- We will consider three different techniques:
  - Clustering
  - Multi-modal statistical modelling (Gaussian Mixture Models – GMMs)
  - Decision trees

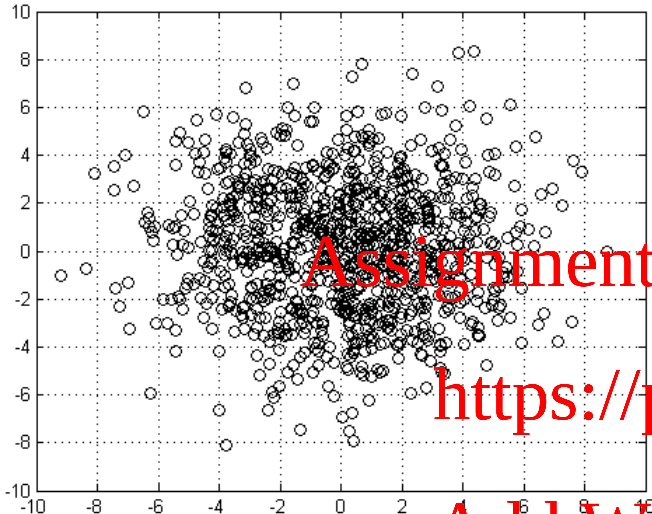


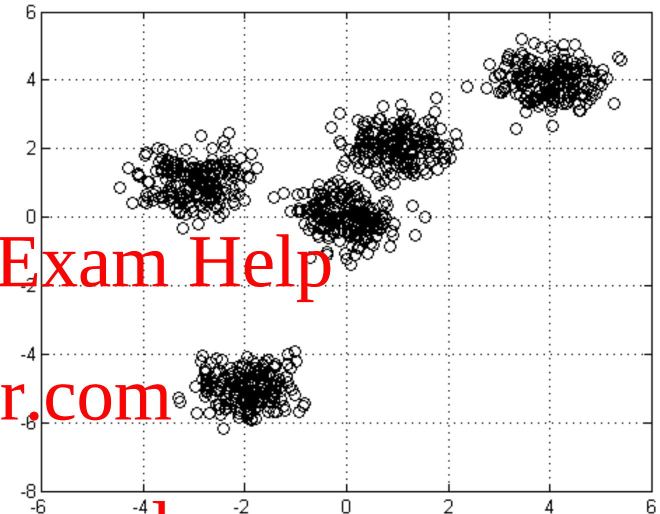
# Clustering - Objectives

- To explain the motivation for clustering
- To introduce the ideas of distance and distortion
- To describe agglomerative and divisive clustering
- To explain the relationships between clustering and decision trees

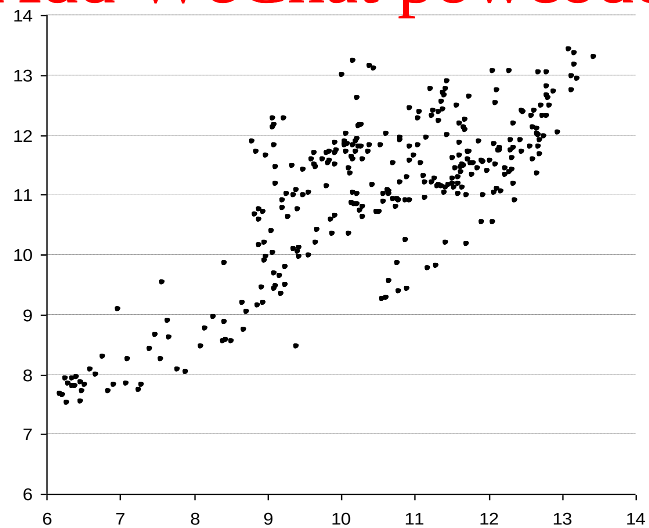


# What does the data look like?






# Structure of data

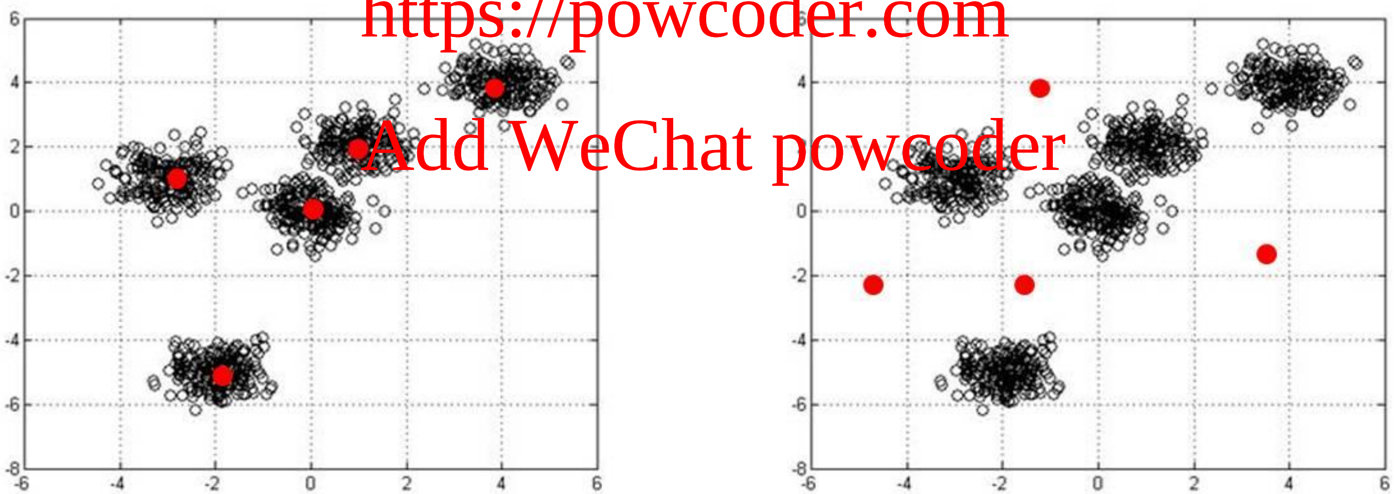
- Typical real data is <u>not</u> uniformly distrubuted
- It has <u>structure</u>
- Variables might be correlated
- The data might be grouped into natural ‘clusters’ – it may have been generated by several different “sources”
- The purpose of cluster analysis is to find this underlying structure <u>automatically</u>



# Clusters and centroids

- Assume clusters are spherical - determined by <u>centres</u>
- Cluster centres are called <u>centroids</u>
- Questions: How many centroids do we need? Where should we put them?








UNIVERSITY OF BIRMINGHAM

# Distance

- A function $d(x,y)$ defined on pairs of points $x$ and $y$ is called a <u>distance</u> or <u>metric</u> if it satisfies:
  - $d(x,y) \geq 0$ and $d(x,y) = 0$ if and only if $x = y$
  - $d(x,y) = d(y,x)$ for all points $x$ and $y$ (<u>symmetry</u>)
  - $d(x,z) \leq d(x,y) + d(y,z)$ for all points $x$, $y$ and $z$ (<u>triangle inequality</u>)



# Example metrics

- The most common metric is the Euclidean metric
- If $x = [x_1, x_2, \ldots, x_N]$ and $y = [y_1, y_2, \ldots, y_N]$ then:



$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_N - y_N)^2}$$



- This  is normal distance in Euclidean space



- There are lots of others, but focus on this one

# The $L^p$ Metrics

- Euclidean distance is sometimes called the $L^2$-metric

$$d_2(x,y)=\left[\sum_{n=1}^{N}(x_n-y_n)^2\right]^{\frac{1}{2}}$$



- It is one of a family of metrics called the $L^p$-metrics

$$d_p(x,y)=\left[\sum_{n=1}^{N}(x_n-y_n)^p\right]^{\frac{1}{p}}$$

# Special $L^p$ metrics

- $p=1$ – the 'City Block' metric

$$d_1(x,y)=\left[\sum_{n=1}^{N}|x_n-y_n|\right]$$



- $p=\infty$

$$d_\infty(x,y)=\max_{n=1,\ldots,N}|x_n-y_n|$$

# Unit sphere

- For a metric $d$ defined on $N$ dimensional space, the <u>unit sphere</u> is the set of vectors $\boldsymbol{x}$ such that $d(\boldsymbol{x},\boldsymbol{0}) = 1$

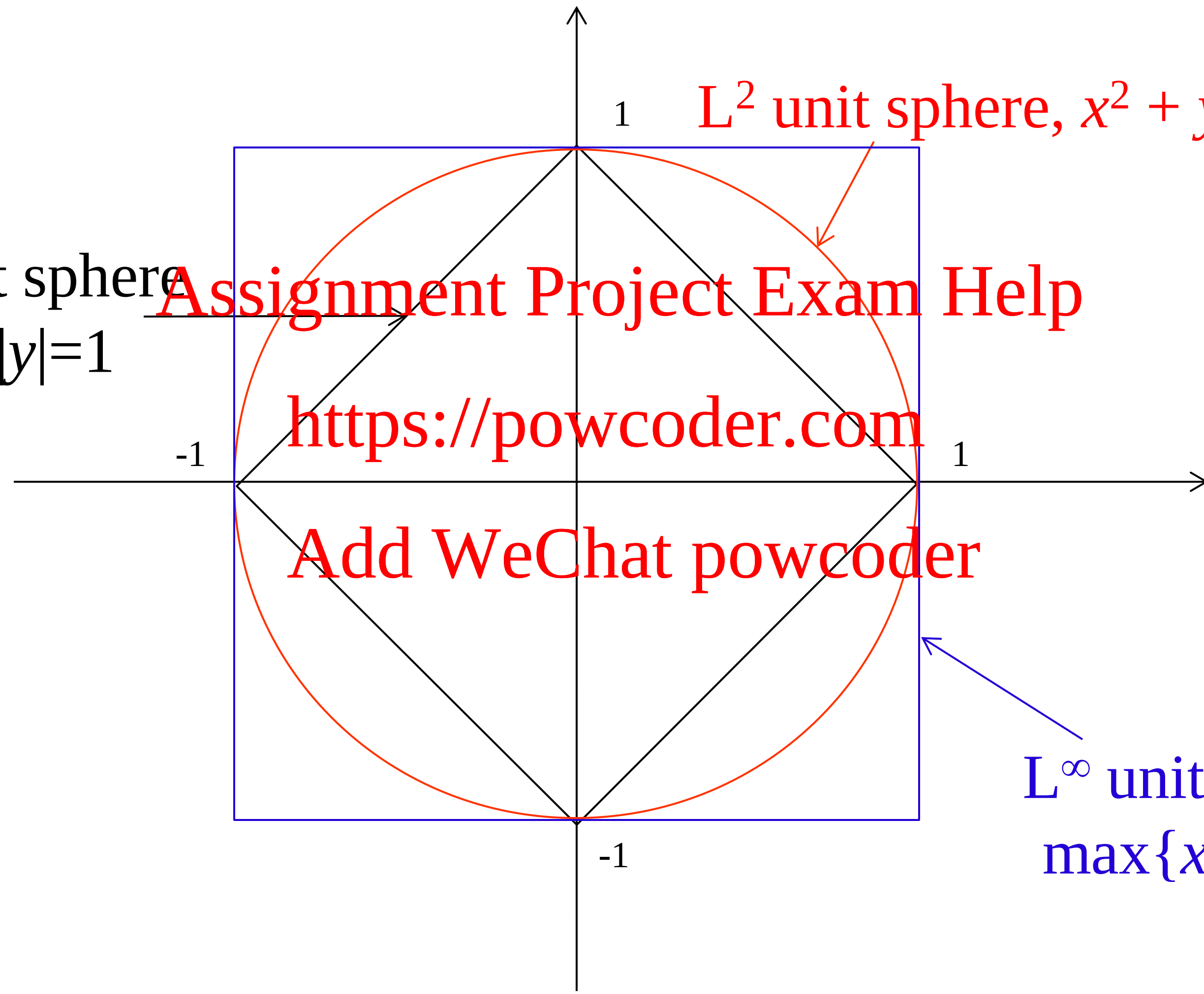


$$S_d = \{x : d(x, 0) = 1\}$$



- What do the unit spheres in 2D look like for these metrics?

# Example Unit Spheres (2D)






# Distortion

- Distortion is a measure of how well a set of centroids models a set of data
- Suppose we have:
  - data points $y_1, y_2, \ldots, y_T$
  - centroids $c_1, \ldots, c_M$
- For each data point $y_t$ let $c_{i(t)}$ be its closest centroid
- In other words: $d(y_t, c_{i(t)}) = \min_m d(y_t, c_m)$



# Distortion

- The <u>distortion</u> for the centroid set $C = c_1, \ldots, c_M$ is defined by:



$$Dist(C) = \sum_{t=1}^{T} d\left(y_t, c_{i(t)}\right)$$



- In other words, the distortion is the sum of distances between each data point and its nearest centroid



- The task of clustering is to find a centroid set $C$ such that the distortion $Dist(C)$ is <u>minimised</u>

# Types of Clustering

- We will start with two types of cluster analysis:
  - Agglomerative clustering, or ‘bottom-up’ hierarchical clustering
  - Divisive clustering, or ‘top-down’ clustering
- In the next lecture we will focus on a more sophisticated clustering method called *k*-means clustering

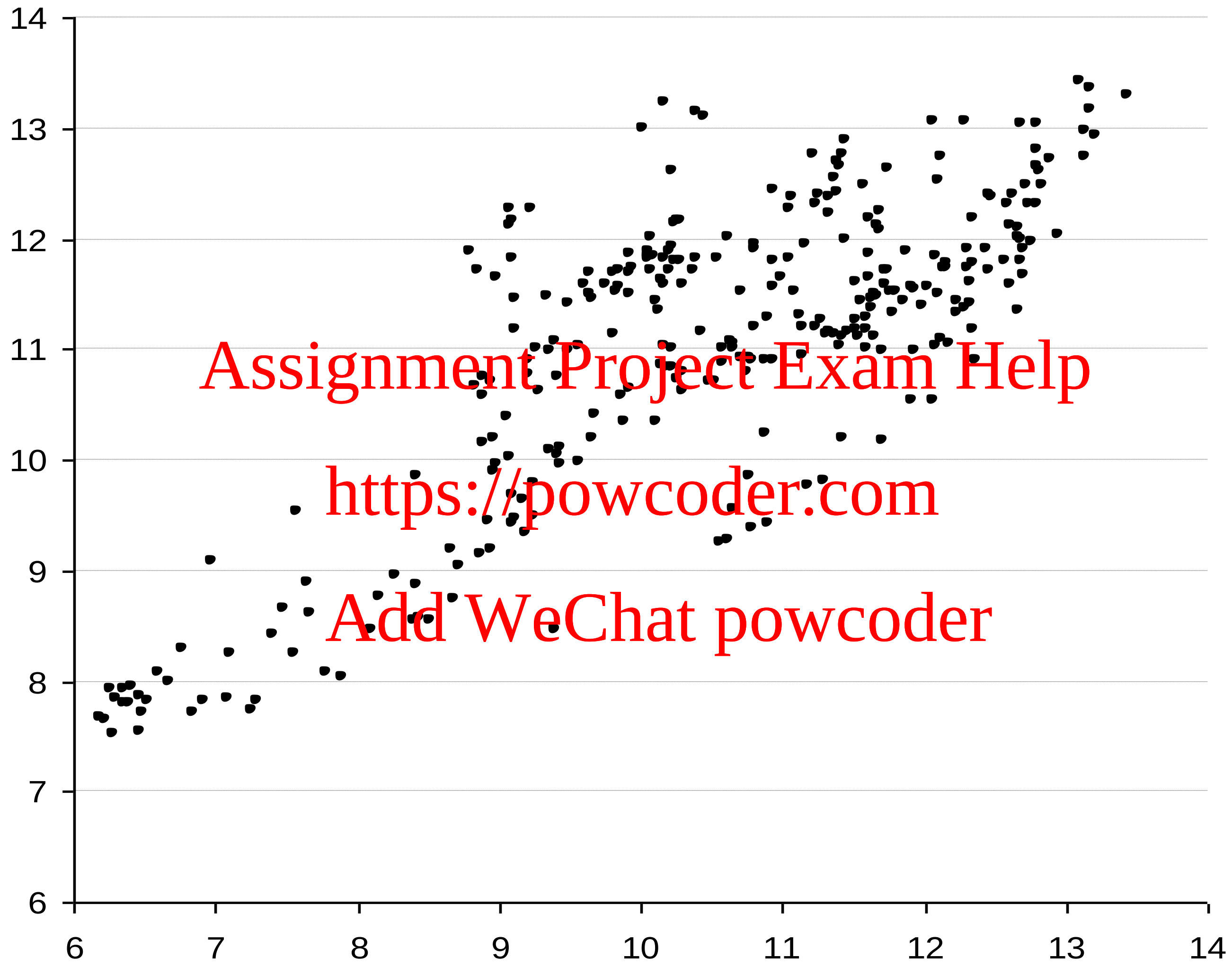


# Agglomerative clustering

- Agglomerative clustering begins by assuming that each data point belongs to its own, unique, 1 point cluster – each point is a centroid
- Clusters are then combined until the required number of centroids is obtained
- The simplest agglomerative clustering algorithm is one which, at each stage, combines the two closest centroids into a single centroid

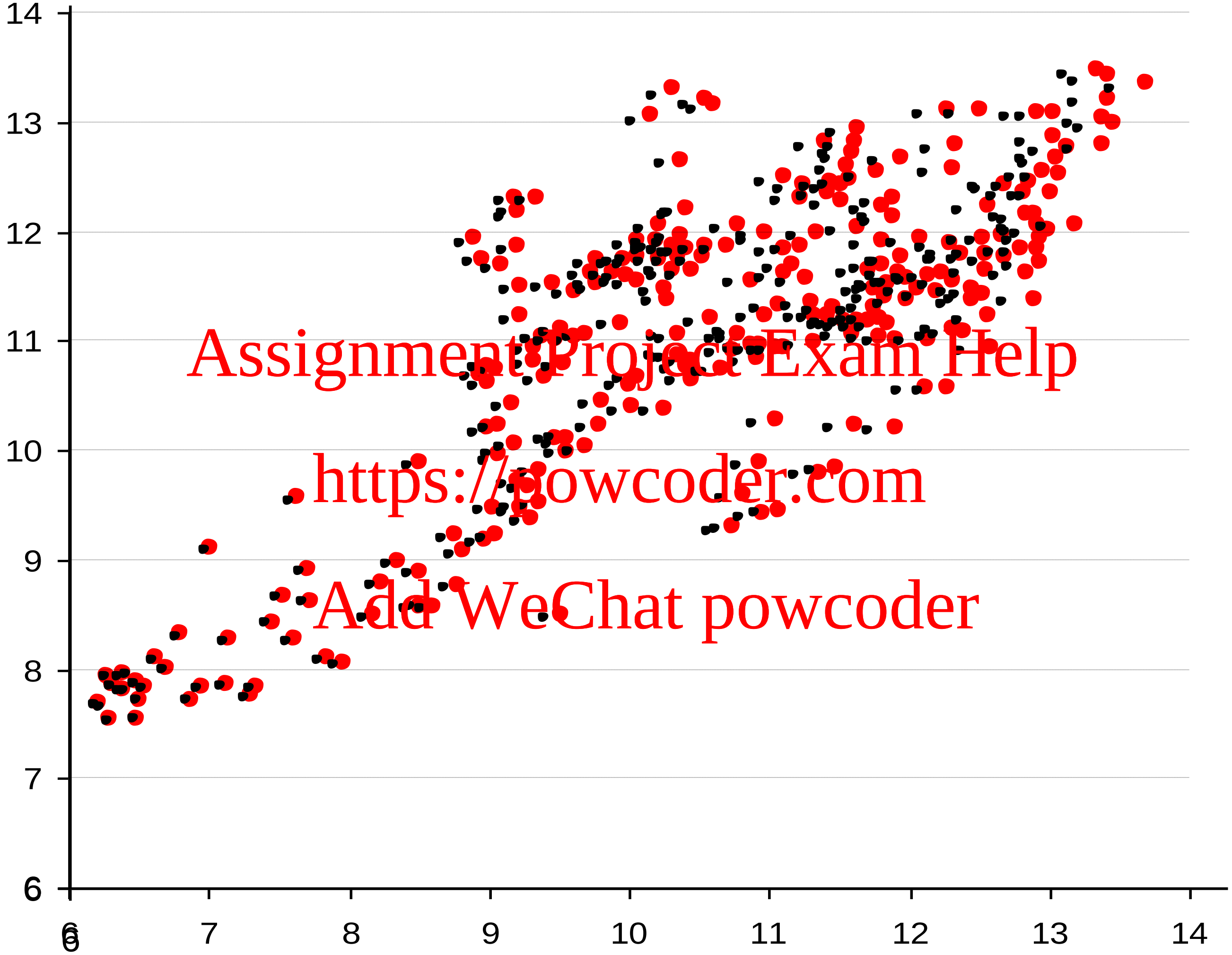


# Original data (302 points)



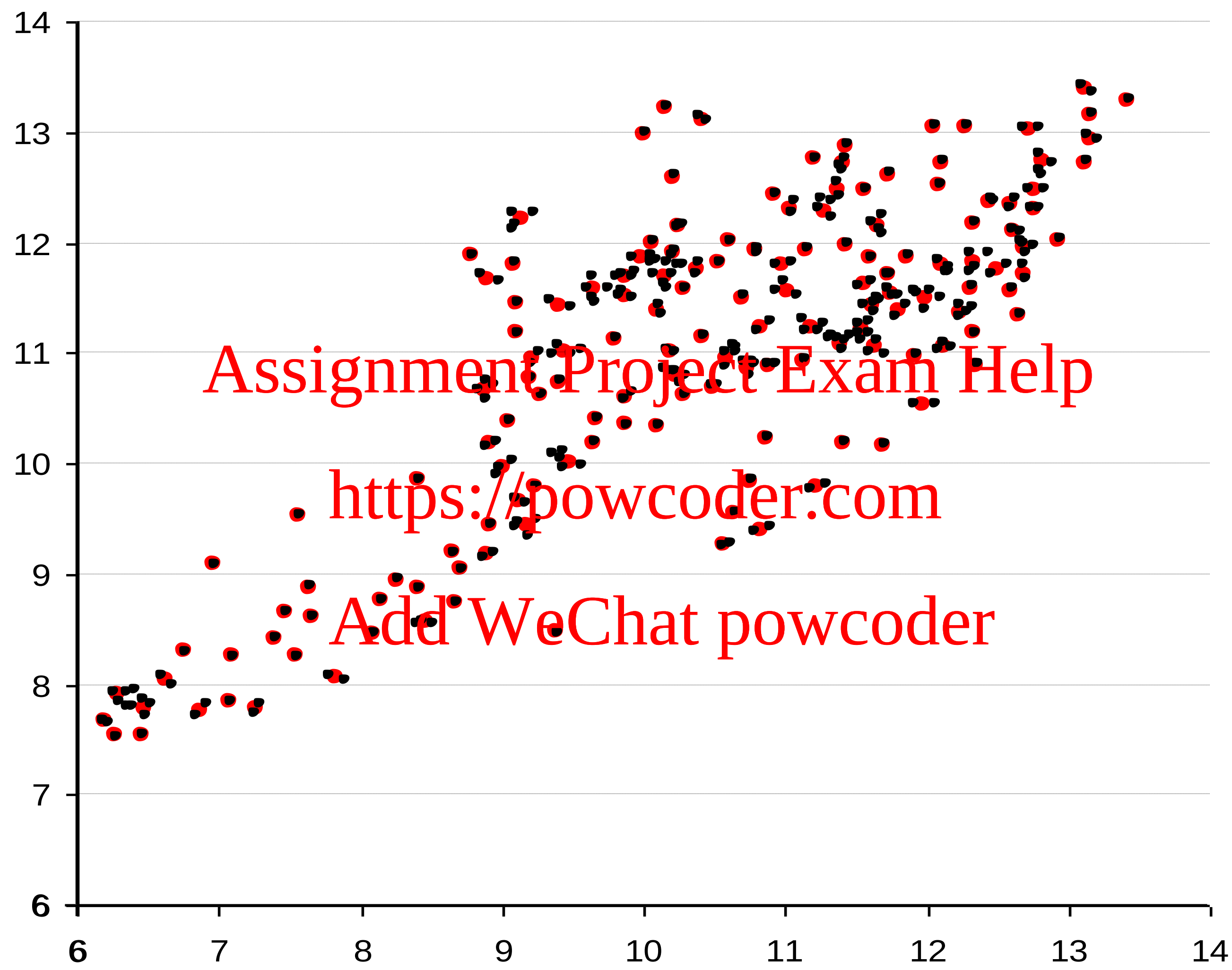



# 252 centroids



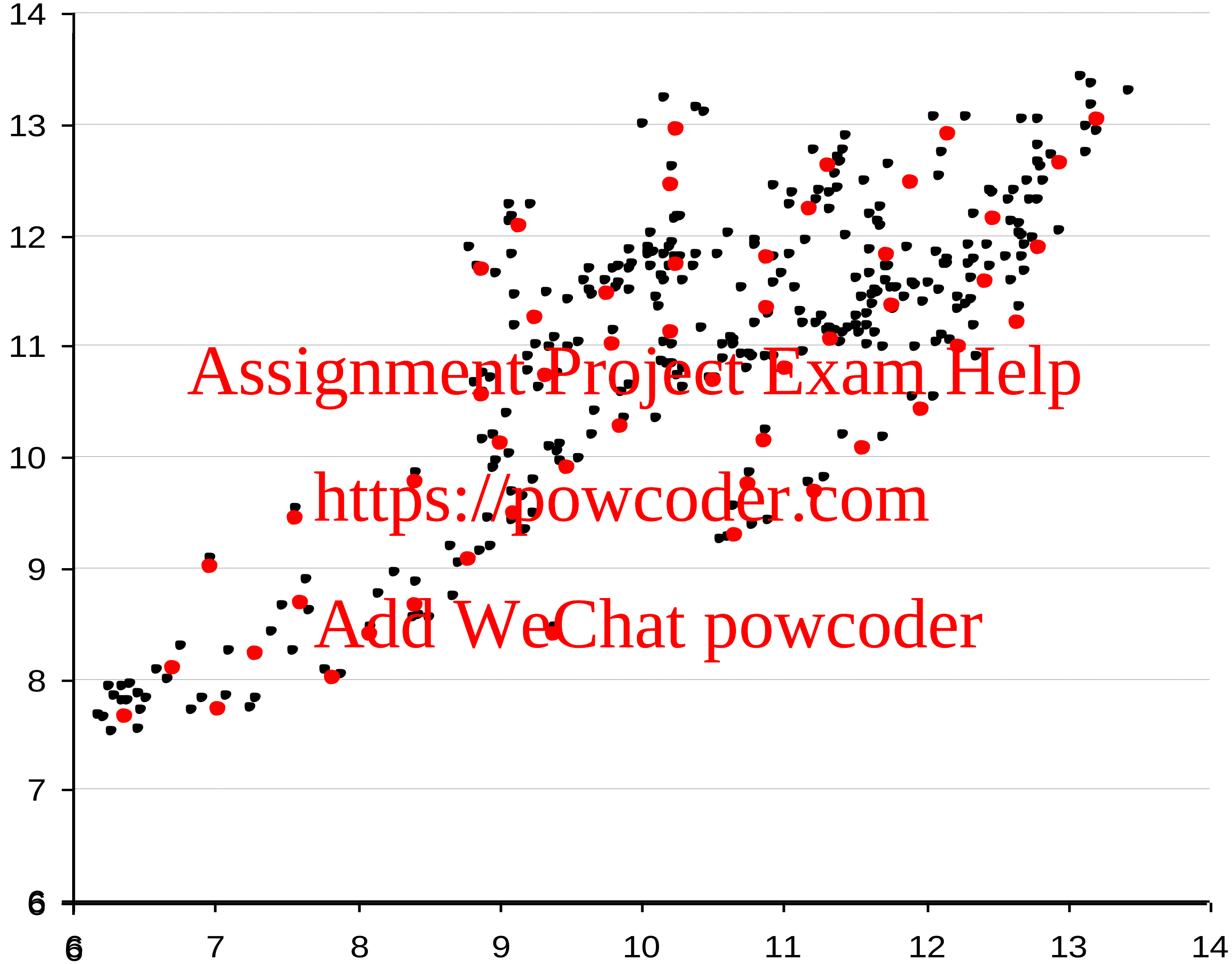



# 152 centroids



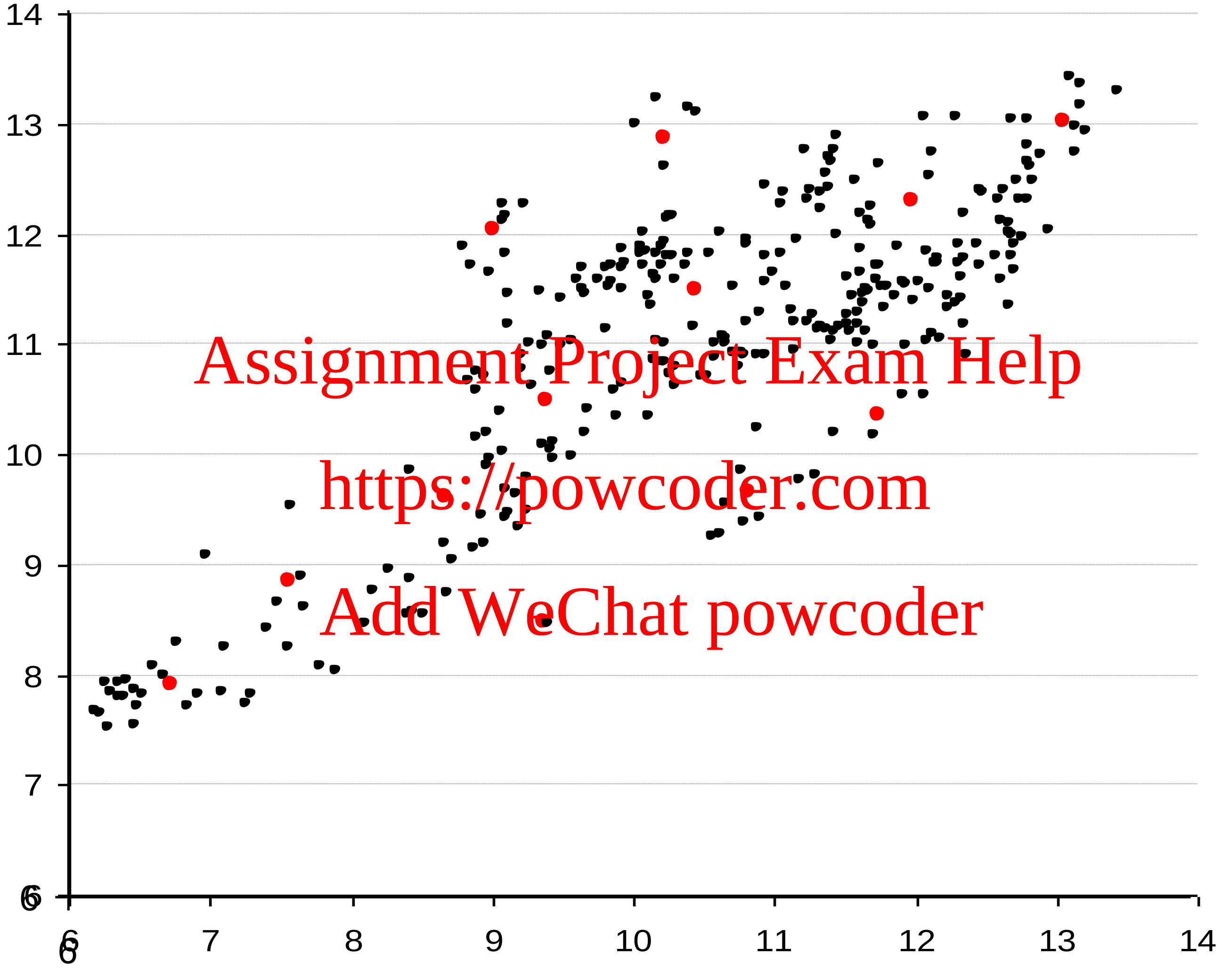



# 52 centroids






# 12 centroids








UNIVERSITY OF BIRMINGHAM

# Optimality of agglomerative clustering

- The result of agglomerative clustering is not optimal
- Generally it does not result in a set of centroids $C$ such that

$$Dist(C) = \min_B Dist(B)$$

- For example,
  – Outliers may be given their own centroids
  – Dense clusters may be given too few centroids

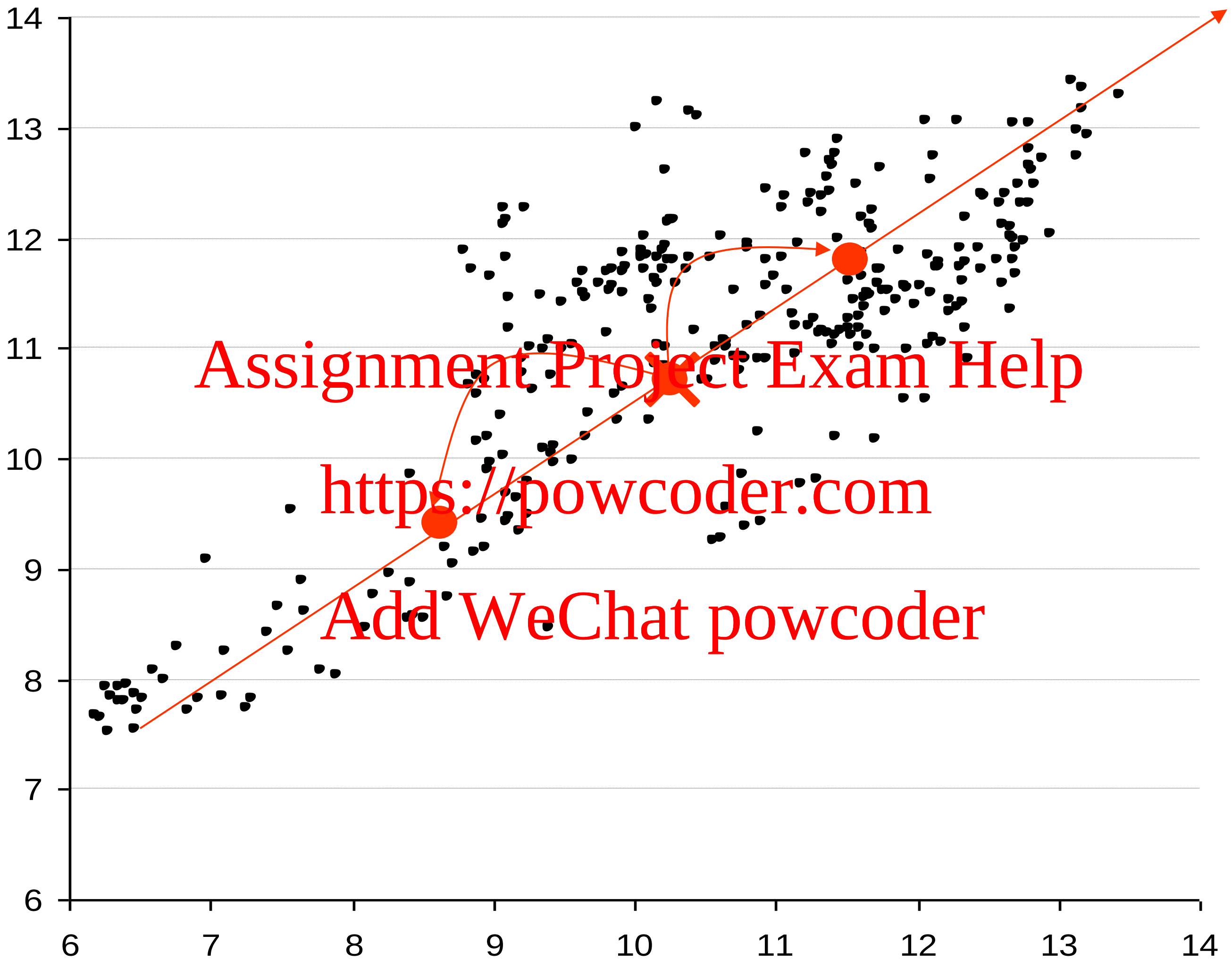


# Divisive Clustering

- Divisive clustering begins by assuming that there is just one centroid – typically in the centre of the set of data points
- That point is replaced with 2 new centroids
- Then each of these is replaced with 2 new centroids
- …



# Original data (302 points)

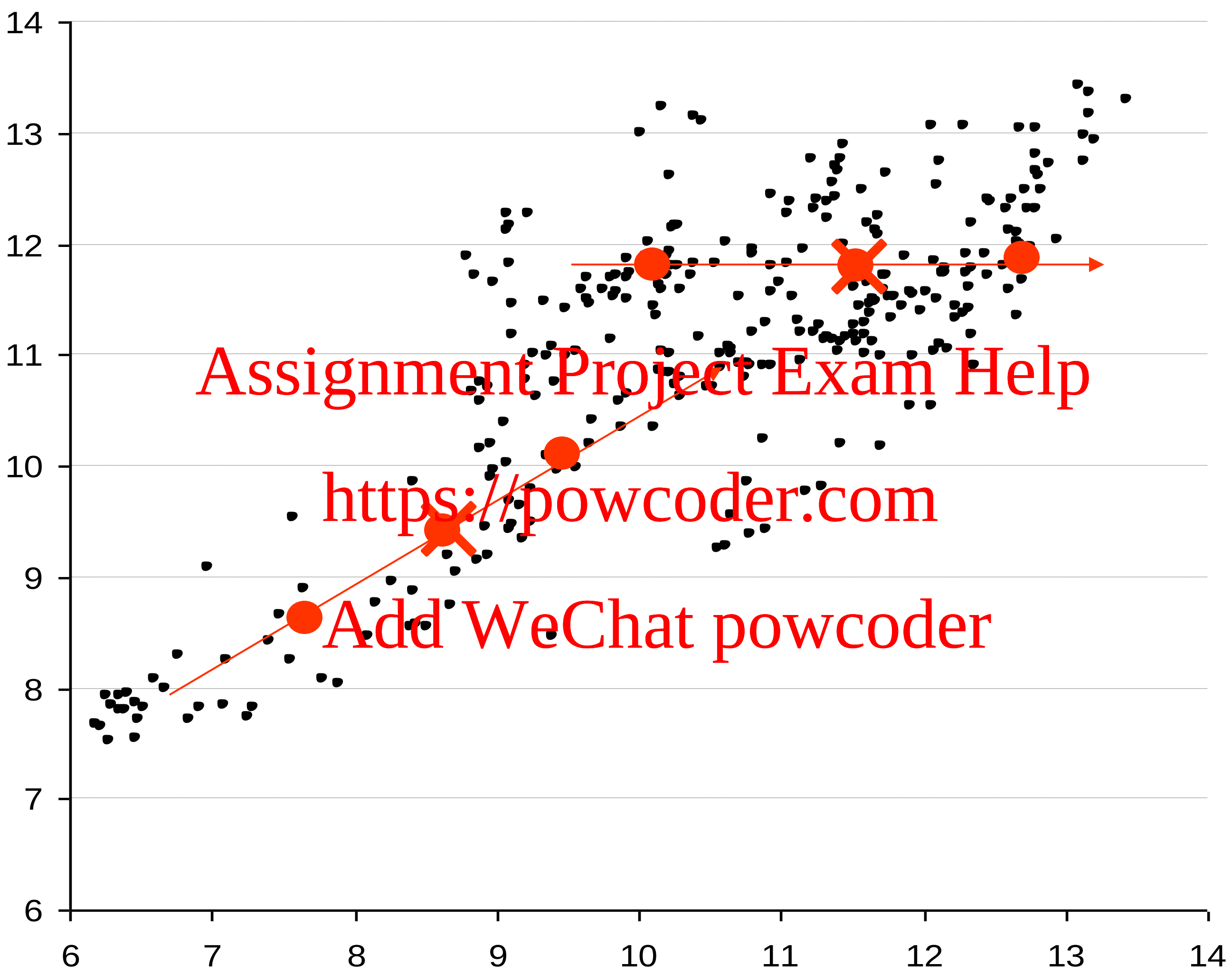



Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

# Original data (302 points)






# Optimality of divisive clustering

- The result of agglomerative clustering is not optimal
- Generally it does not result in a set of centroids $C$ such that



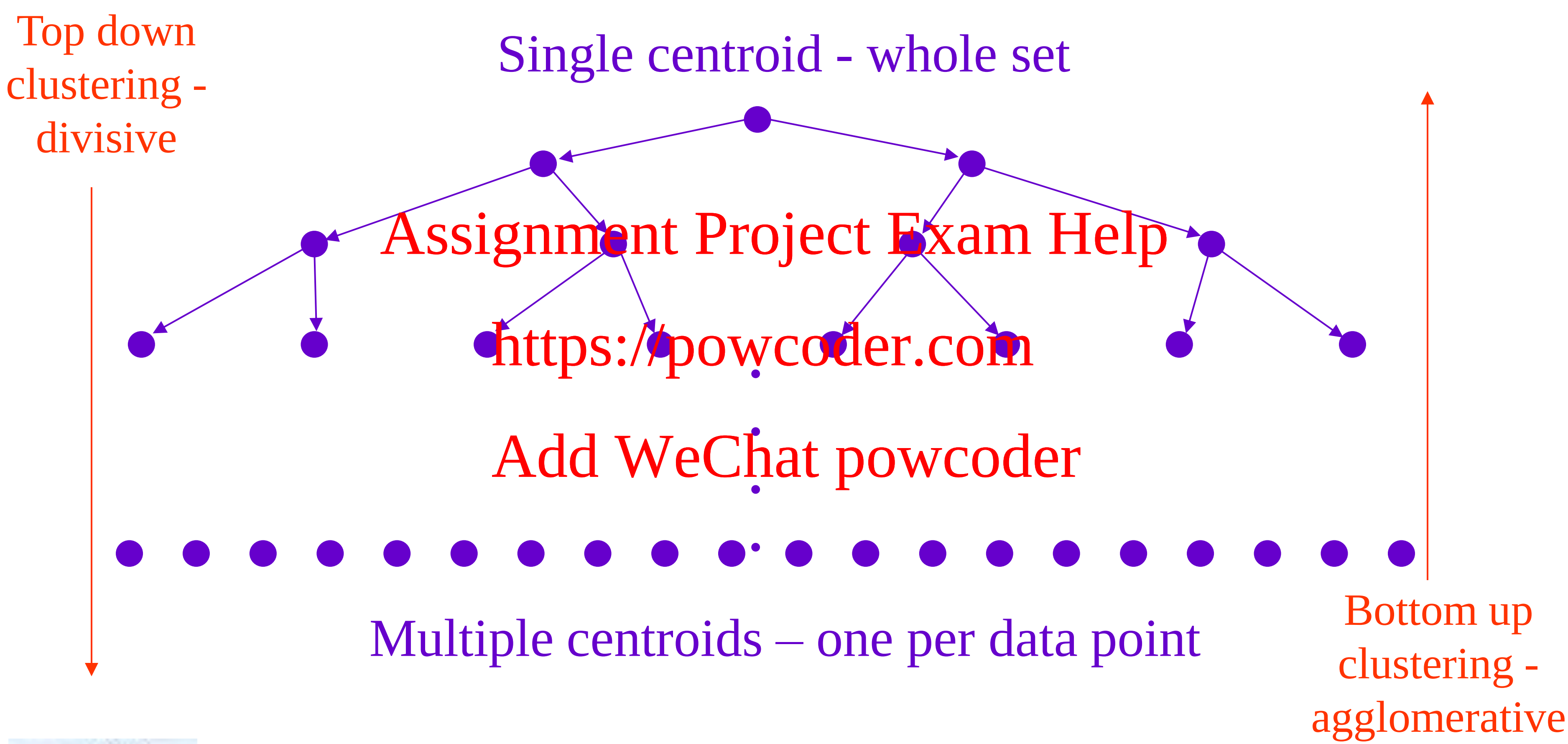
$$Dist(C) = \min_B Dist(B)$$



- Sequential decision making is normally suboptimal
  - Decisions are not reversible
  - If a point goes to a particular half of a partition it will never be re-allocated to the other half
  - Probably not how a human would do it



# Decision tree interpretation








UNIVERSITY OF BIRMINGHAM

# Optimality

- An ‘optimal’ set of centroids is one which minimises the distortion
- In general, neither method gives optimal sets of centroids
- A more principled approach would be to think of distortion as a function of the centroid set and minimize it



# Notation and method

- $N$ dimensional space
- $T$ data points $X = \{x_1,\ldots,x_T\}$



- $K$ centroids $C = \{c_1,\ldots,c_K\}$



- Calculate

$$\frac{d}{dc_k^n} Dist(C)$$



for each $k$ and $n$, set to zero and solve

# Summary

- Distance metrics and distortion
- Agglomerative clustering
- Divisive clustering
- Decision tree interpretation