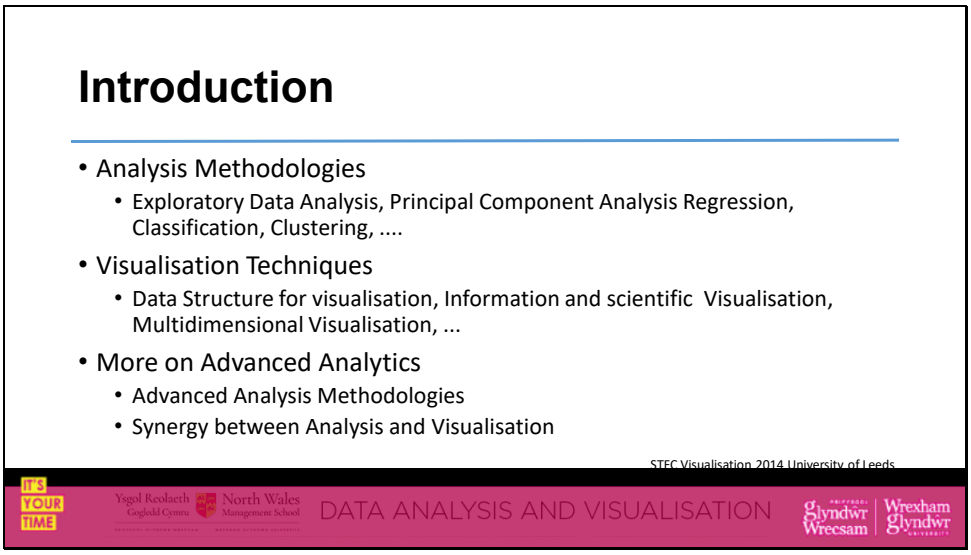
Slide 1

Welcome to week 6, more advanced analytics.

Slide 2



You have seen so far various.
Analysis Methodologies

Exploratory Data Analysis, Principal Component Analysis Regression, Classification, Clustering, ...

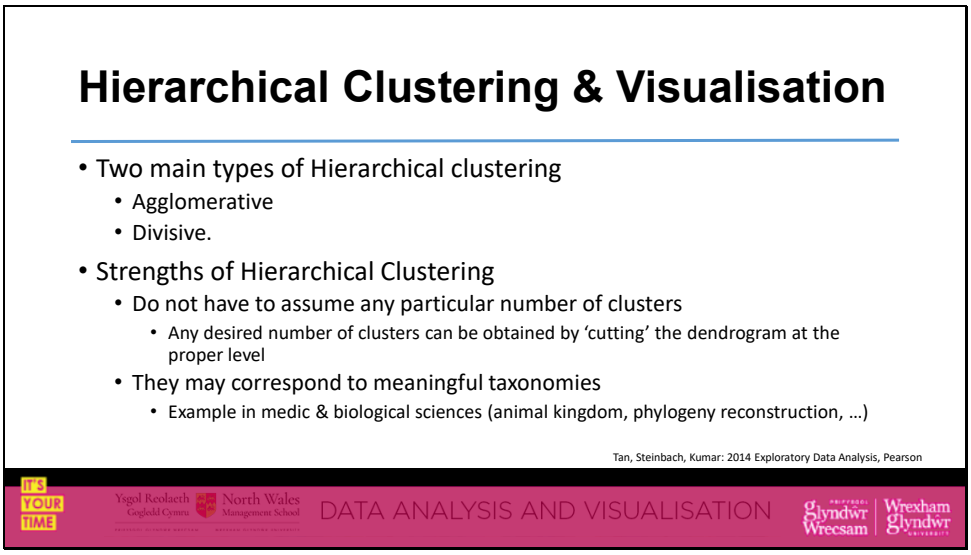
Visualization Techniques

Data Structure for visualization, Information and scientific visualization, Multidimensional Visualization, ...

And have solved few problems for each topic individually.

In this lecture we follow our discussion on More Advanced Analytics, especially we look at on more Advanced Analysis Methodologies such as Hierarchical Clustering and integration and Synergy between Analysis and Visualization to learn more about dendrogram that has been used especially in medical applications.

Slide 3

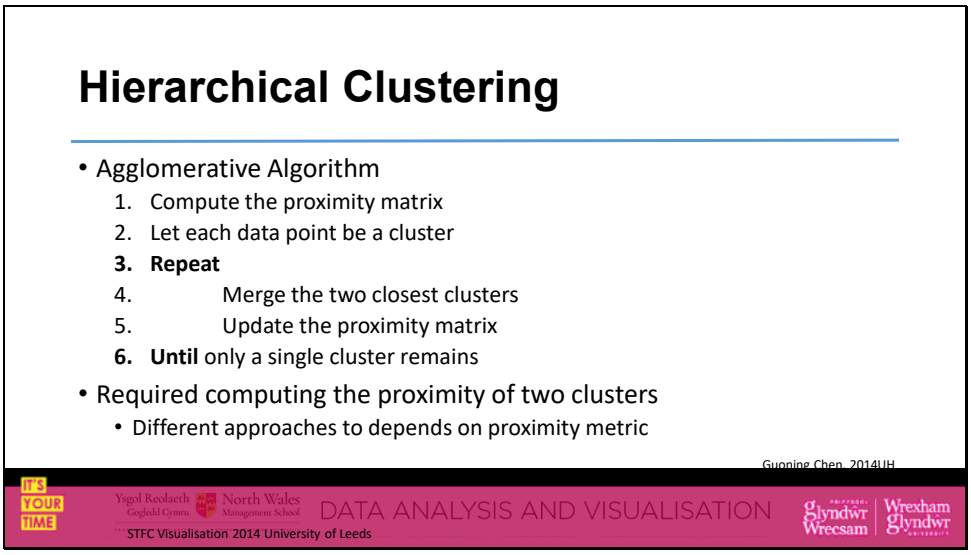


Two main types of hierarchical clustering
1. Agglomerative: Start with the points as individual clusters then at each step, merge the closest pair of clusters until only one cluster (or k clusters) left.
2. Divisive: Start with one, all-inclusive cluster, then at each step, split a cluster until each cluster contains an individual point (or there are k clusters)

Traditional hierarchical algorithms use a similarity or distance matrix and need Merge or split one cluster at a time.

Strengths of Hierarchical Clustering:
Do not have to assume any particular number of clusters.
Any desired number of clusters can be obtained by 'cutting' the dendrogram at the proper level.
They may correspond to meaningful taxonomies.
Examples in medic and biological sciences (e.g., animal kingdom, phylogeny reconstruction, …)

Slide 4



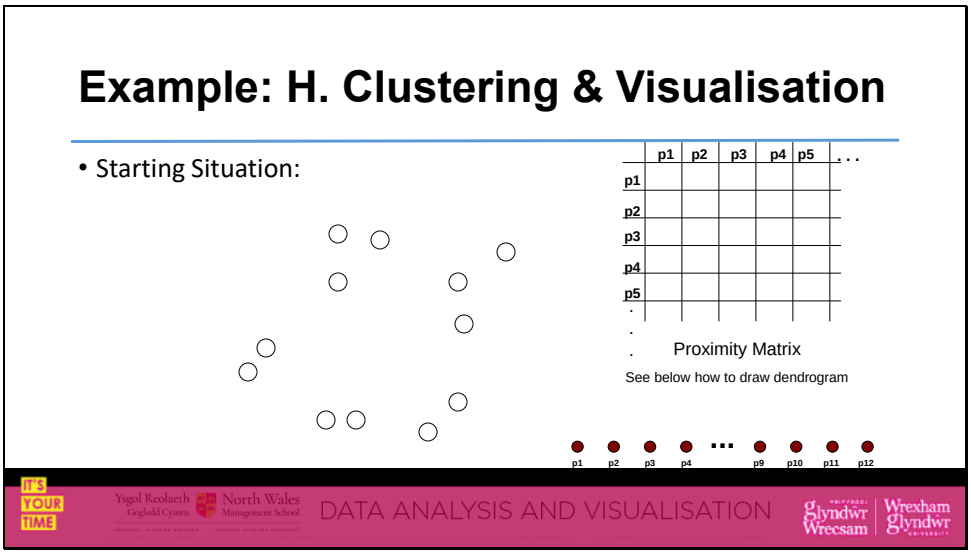
Agglomerative Algorithm is the most popular hierarchical clustering technique and basic Algorithm is straightforward.

Agglomerative Algorithm

- Compute the proximity matrix.
- Let each data point be a cluster.
- **Repeat**
- Merge the two closest clusters.
- Update the proximity matrix.
- **Until** only a single cluster remains

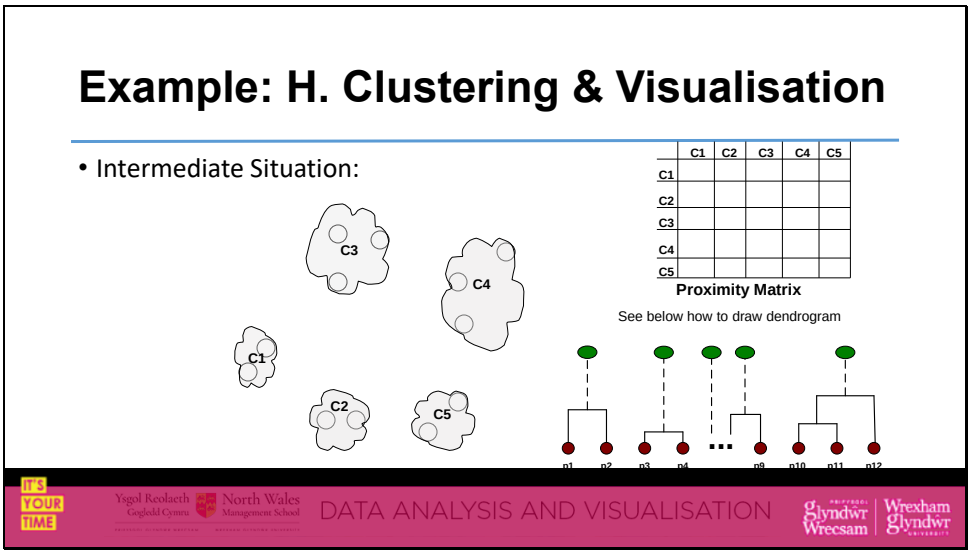
Key operation is the computation of the proximity of two clusters.
Different approaches to defining the distance between clusters distinguish the different algorithms.

Slide 5



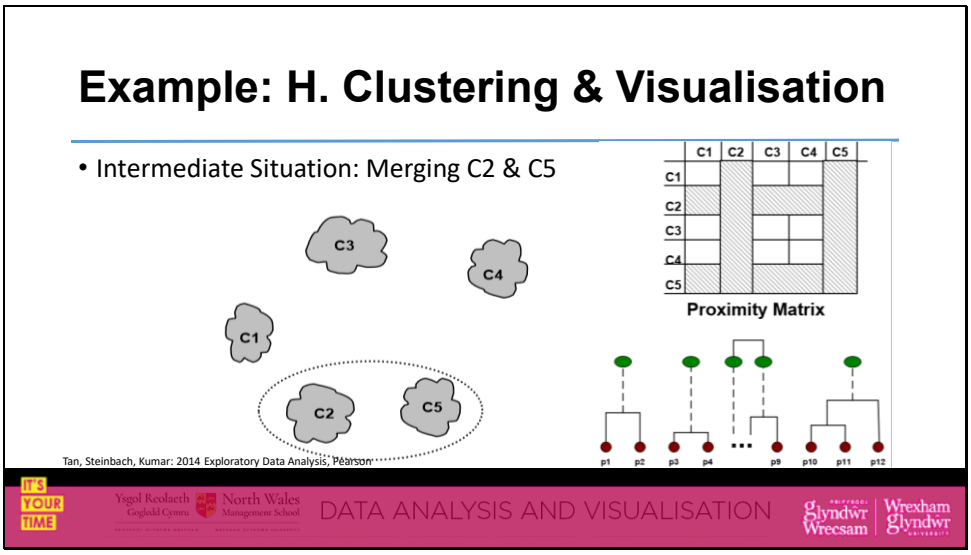
Consider this point in space.
Start with clusters of individual points and a proximity matrix.

Slide 6



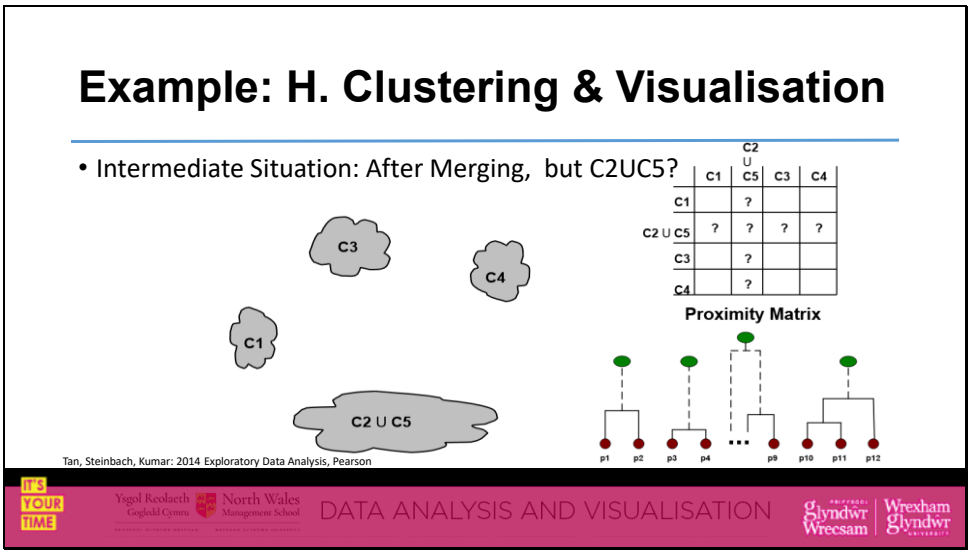
Intermediate Situation:
After some merging steps, we have 5 clusters out of these 12 points and now the proximity matrix is only for these 5 clusters.

Slide 7



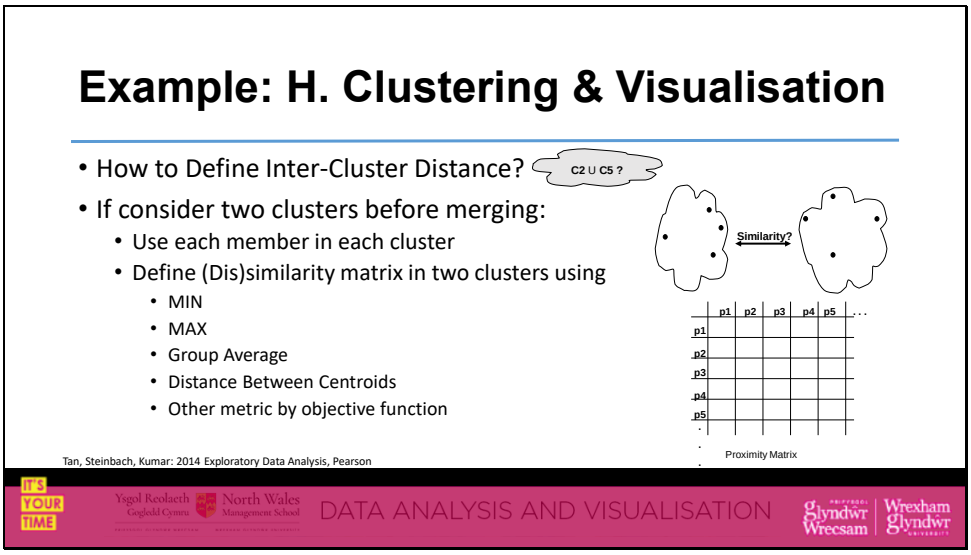
We want to merge the two closest clusters (C2 and C5) and update the proximity matrix.

Slide 8



But what is C2 U C5?
The question is “How do we update the proximity matrix?”

Slide 9



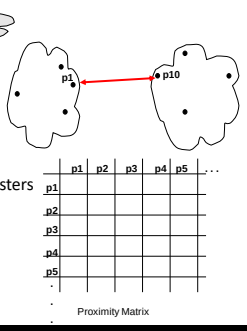
"How do we update the proximity matrix?" How to Define Inter-Cluster Distance C2UC5?
If consider two clusters before merging:
We can use each member in each cluster to define dissimilarity or similarity matrix in two clusters using metrics such as

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function.
- Ward's Method uses squared error.

Slide 10

Example: H. Clustering & Visualisation

- How to Define Inter-Cluster Distance?
- If consider two clusters before merging:
  - Use each member in each cluster
  - Define (Dis)similarity matrix in two clusters using
    - **MIN**
      - Uses distance between two closest points in the different clusters
      - In this case C2UC5 = Distance between p1 & p10 for example
      - It is shown by Red arrow



Tan, Steinbach, Kumar: 2014 Exploratory Data Analysis, Pearson

Proximity of two clusters is based on the two closest points in the different clusters.
Determined by only one pair of points, i.e., by one link in the proximity graph.
It is also called Single Link

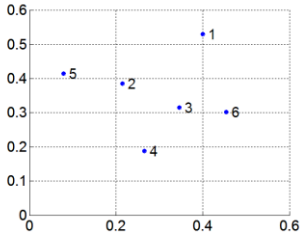
Slide 11

Example: H. Clustering & Visualisation

- How to Define Inter-Cluster Distance?
  - Obtain Single Link if these are your data points:
    - Calculate Distance Matrix

Distance Matrix

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

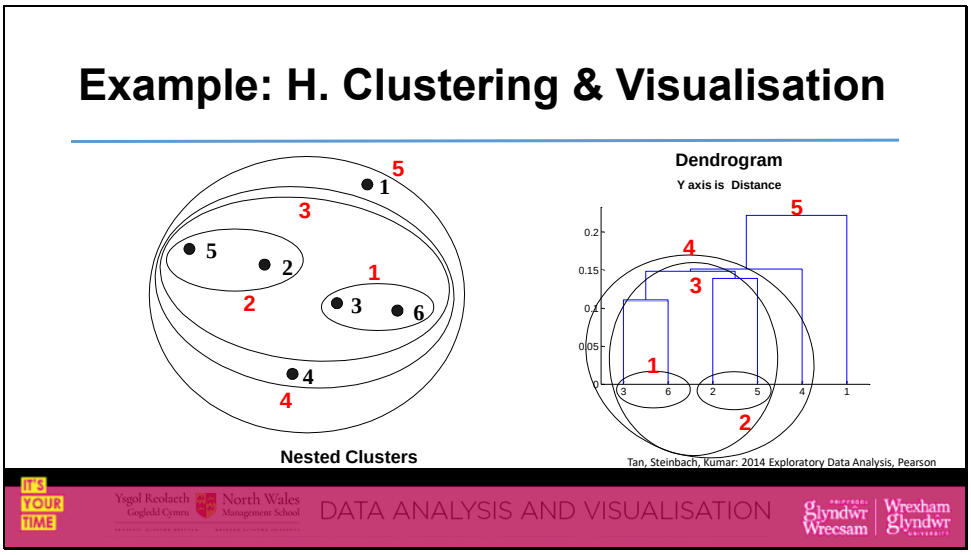


Animation next will show you how to cluster and draw a dendrogram

Tan, Steinbach, Kumar: 2014 Exploratory Data Analysis, Pearson

Example Single Link for data points below & calculate Distance Matrix.
In next slide will show you how to cluster and draw a dendrogram.

Slide 12

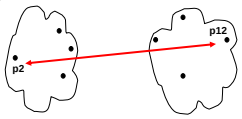


- First merge closest neighbors, i.e., merge points 3 & 6 to form cluster 1
- and merge point 2 & 5 to form cluster 2
- Then merge cluster 1 and 2 to form cluster 3. to calculate distance between them use the distance between closest point in two cluster, i.e., point 2 & 3 (see distance matrix) . see how dendrogram is forming.
- Now merge point 4 with cluster 3 to form cluster 4 (point 4 is closet to point 3) D = 0.15
- And finally merge 1 to all. Note point 1 is closet to point 3 D = 0.22

Slide 13

Example: H. Clustering & Visualisation

- How to Define Inter-Cluster Distance? C2 ∪ C5 ?
- If consider two clusters before merging:
  - Use each member in each cluster
  - Define (Dis)similarity matrix in two clusters using
    - Min
    - **Max**
      - Uses distance between two furthest points in the different clusters (see the figure above)
      - In this case C2UC5 = Distance between p2 & p12 for example
      - It is shown by Red arrow
      - Try to draw a dendogram for data shown in previous slide.



Tan, Steinbach, Kumar: 2014 Exploratory Data Analysis, Pearson

Proximity of two clusters is based on the two most distant points in the different clusters (see the figure above)

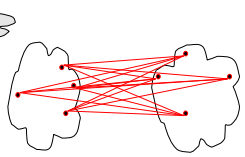
Determined by all pairs of points in the two clusters.
In this case C2UC5 = Distance between p2 & p12 for example
It is shown by Red arrow.

It is also called Complete Linkage.

Slide 14

Example: H. Clustering & Visualisation

- How to Define Inter-Cluster Distance?
- If consider two clusters before merging:
  - Use each member in each cluster
  - Define (Dis)similarity matrix in two clusters using
    - Min
    - Max
    - **Group Average**
      - Uses average distance among them
      - In this case C2UC5 = average distance of all red arrows in above
      - Try to draw a dendogram for data shown previously.



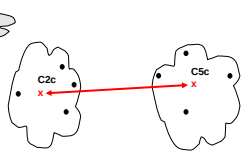
Tan, Steinbach, Kumar: 2014 Exploratory Data Analysis, Pearson

Proximity of two clusters is the average of pairwise proximity between points in the two clusters.
Need to use average connectivity for scalability since total proximity favors large clusters.

Slide 15

Example: H. Clustering & Visualisation

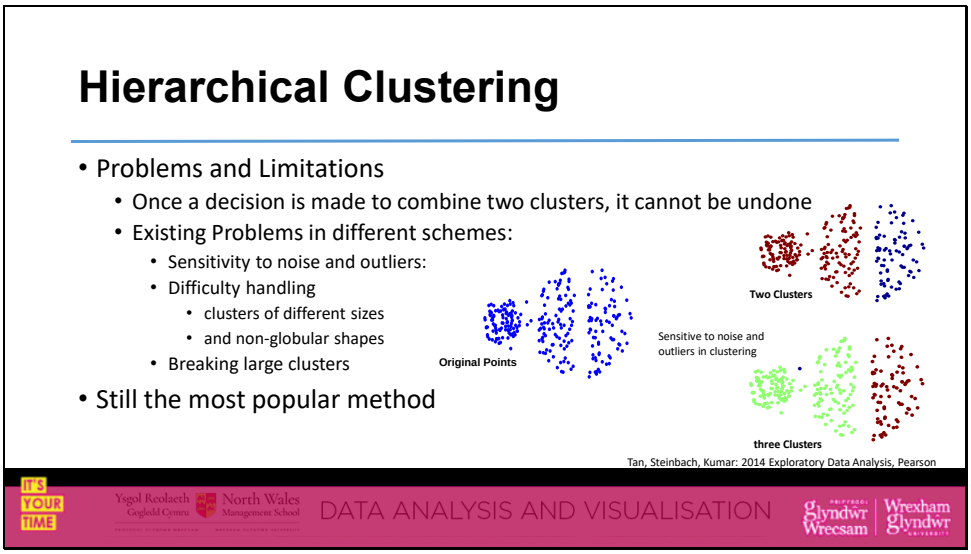
- How to Define Inter-Cluster Distance?
- If consider two clusters before merging:
  - Use each member in each cluster
  - Define (Dis)similarity matrix in two clusters using
    - Min
    - Max
    - Group Average
    - **Distance Between Centroids**
      - Obtain the center of each cluster
      - In this case C2UC5 = Distance between two centre C2c & C5c
      - It is shown by Red arrow



Tan, Steinbach, Kumar: 2014 Exploratory Data Analysis, Pearson

Distance Between Centroids
Obtain the center of each cluster.
In this case C2UC5 = Distance between two centers C2c & C5c
It is shown by Red arrow.

Slide 16



Once a decision is made to combine two clusters, it cannot be undone.
No global objective function is directly minimized.
Different schemes have problems with one or more of the following:
Sensitivity to noise and outliers
Difficulty handling clusters of different sizes and non-globular shapes
Breaking large clusters
the cluster plots demonstrate how sensitive they are to noise and outliers.

Despite its limitations it is still the most popular method.