Clustering

Jeff Leek

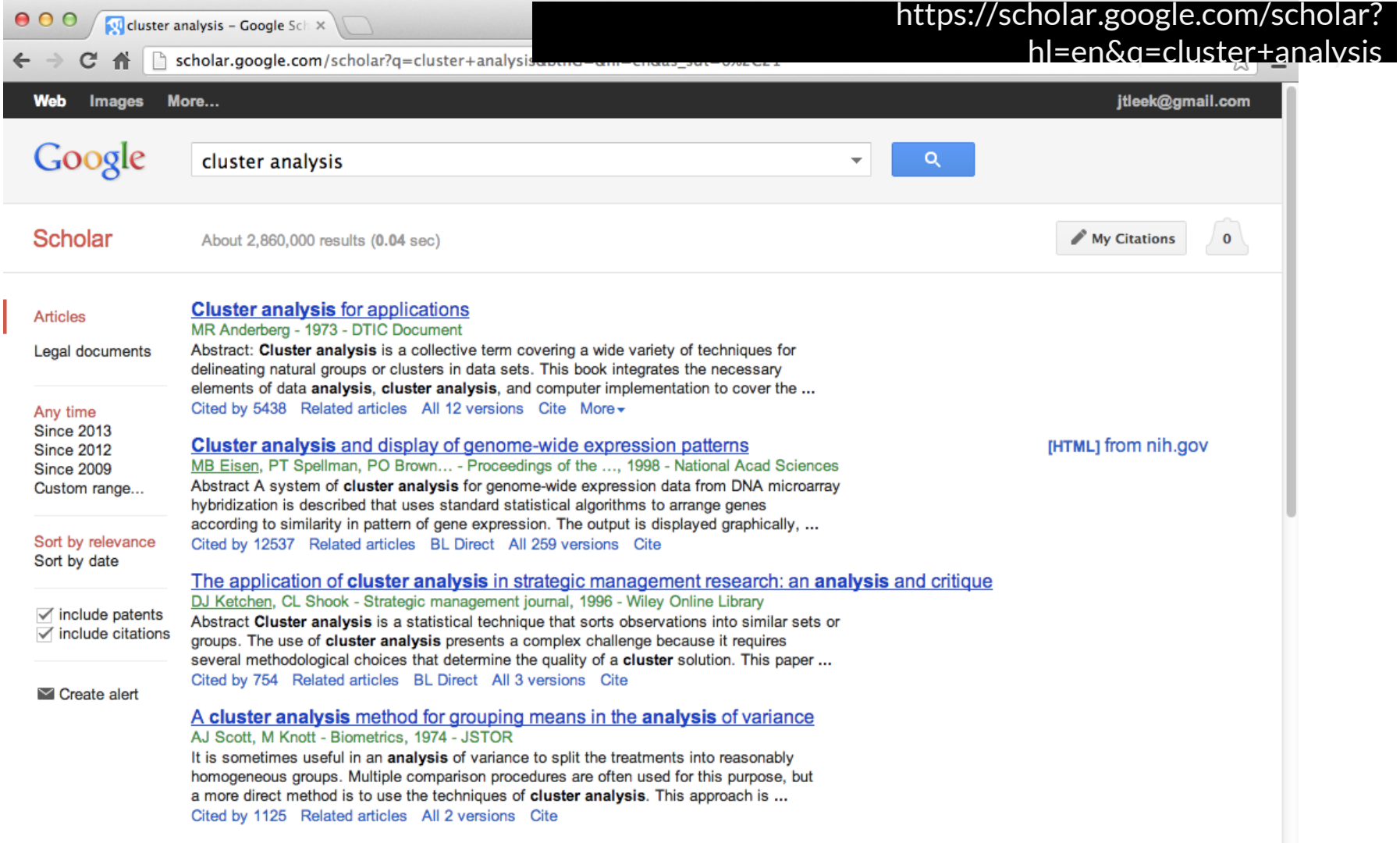
@jtleek

www.jtleek.com

Key ideas

- Find “close” samples/genes/etc.
- Put them into groups

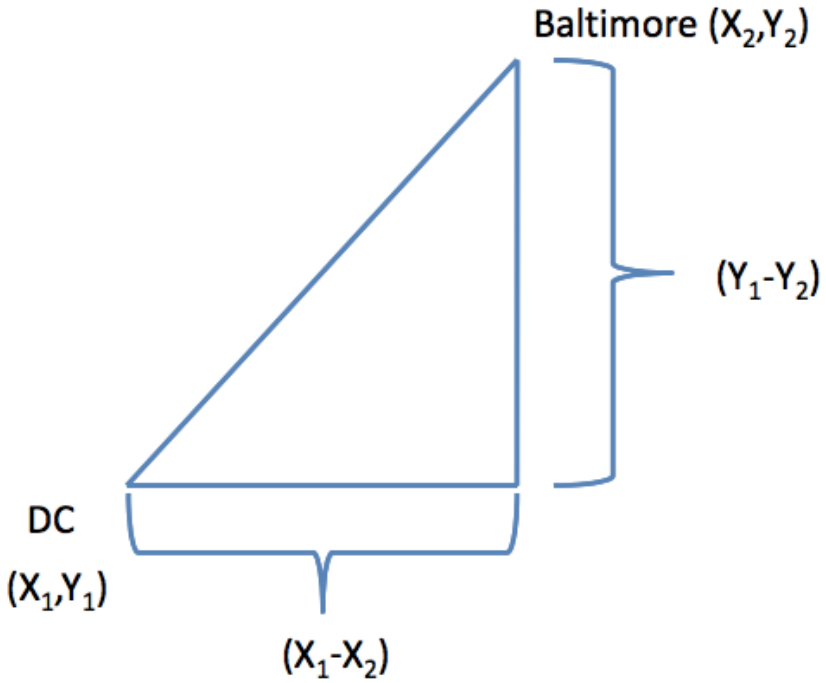
Clustering is important


cluster analysis - Google Sch
scholar.google.com/scholar?q=cluster+analysis
https://scholar.google.com/scholar?hl=en&q=cluster+analysis
Web
Images
More...
jtleek@gmail.com
Google
cluster analysis
Scholar
About 2,860,000 results (0.04 sec)
My Citations
0
Articles
Legal documents
Any time
Since 2013
Since 2012
Since 2009
Custom range...
Sort by relevance
Sort by date
include patents
include citations
Create alert
Cluster analysis for applications
MR Anderberg - 1973 - DTIC Document
Abstract: Cluster analysis is a collective term covering a wide variety of techniques for delineating natural groups or clusters in data sets. This book integrates the necessary elements of data analysis, cluster analysis, and computer implementation to cover the ...
Cited by 5438 Related articles All 12 versions Cite More
Cluster analysis and display of genome-wide expression patterns
[HTML] from nih.gov
MB Eisen, PT Spellman, PO Brown... - Proceedings of the ..., 1998 - National Acad Sciences
Abstract A system of cluster analysis for genome-wide expression data from DNA microarray hybridization is described that uses standard statistical algorithms to arrange genes according to similarity in pattern of gene expression. The output is displayed graphically, ...
Cited by 12537 Related articles BL Direct All 259 versions Cite
The application of cluster analysis in strategic management research: an analysis and critique
DJ Ketchen, CL Shook - Strategic management journal, 1996 - Wiley Online Library
Abstract Cluster analysis is a statistical technique that sorts observations into similar sets or groups. The use of cluster analysis presents a complex challenge because it requires several methodological choices that determine the quality of a cluster solution. This paper ...
Cited by 754 Related articles BL Direct All 3 versions Cite
A cluster analysis method for grouping means in the analysis of variance
AJ Scott, M Knott - Biometrics, 1974 - JSTOR
It is sometimes useful in an analysis of variance to split the treatments into reasonably homogeneous groups. Multiple comparison procedures are often used for this purpose, but a more direct method is to use the techniques of cluster analysis. This approach is ...
Cited by 1125 Related articles All 2 versions Cite

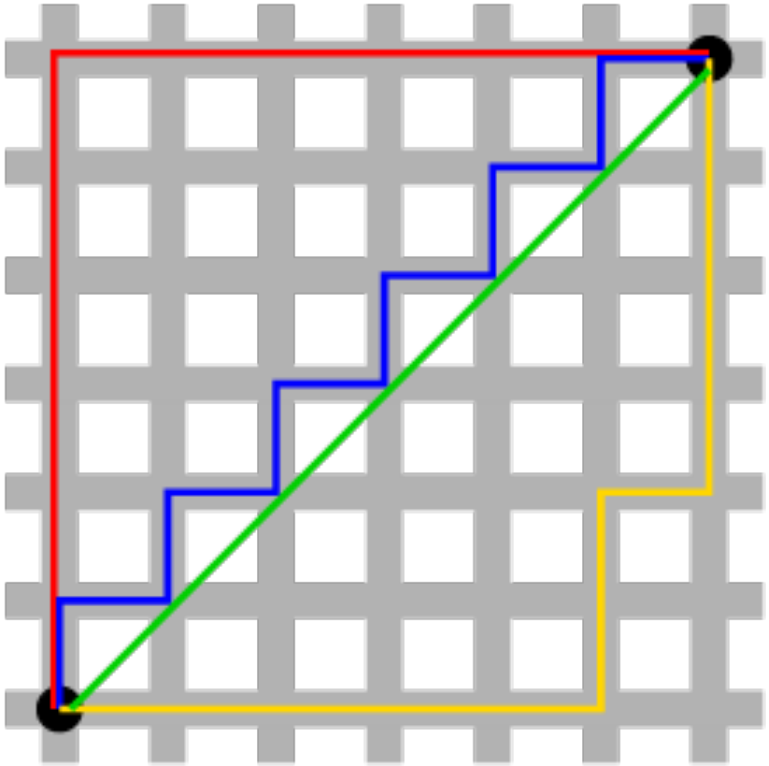
Defining distance

http://rafalab.jhsph.edu/688/lec/lecture5-clustering.pdf

$$\sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2}$$



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

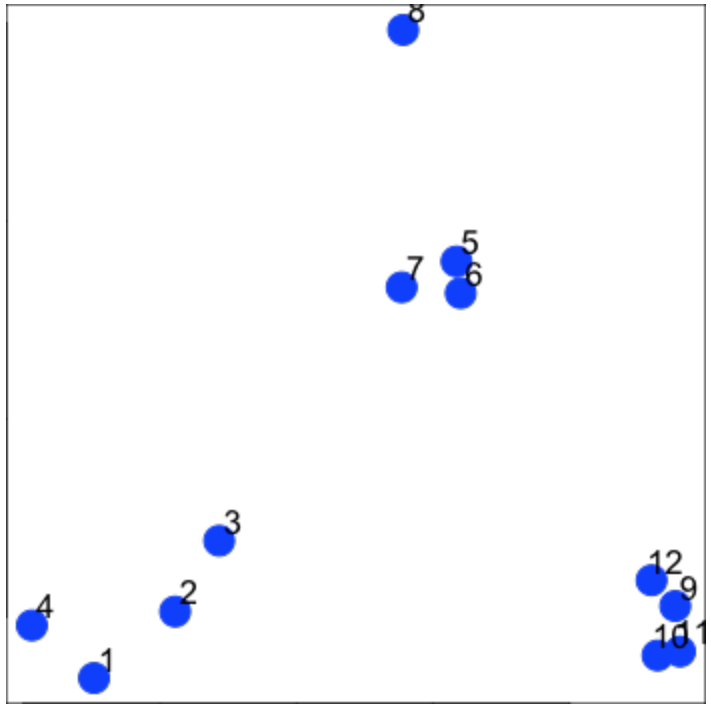


$$|X_1 - X_2| + |Y_1 - Y_2|$$

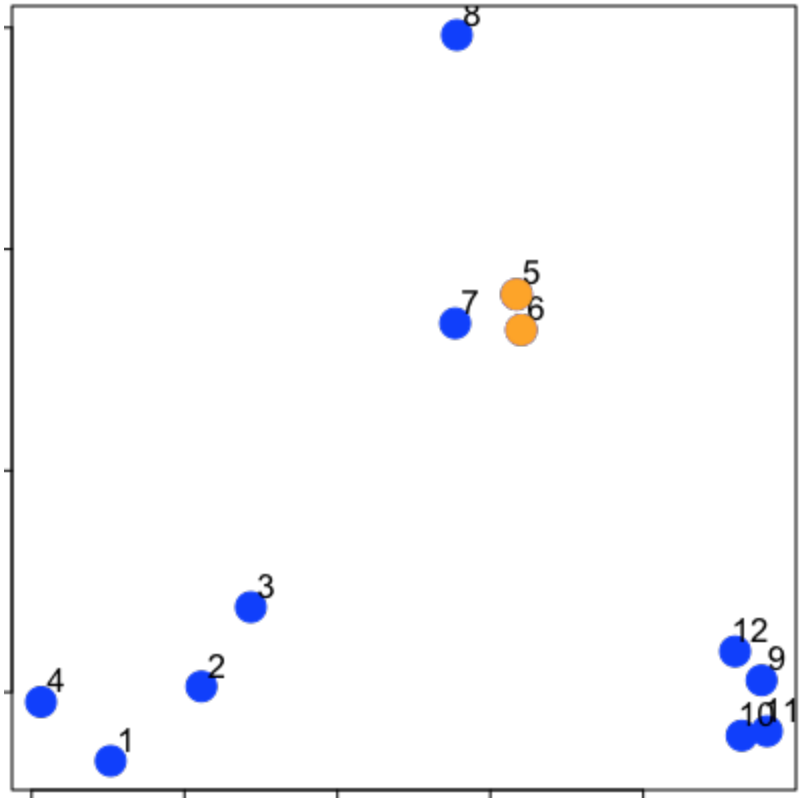
Hierarchical clustering

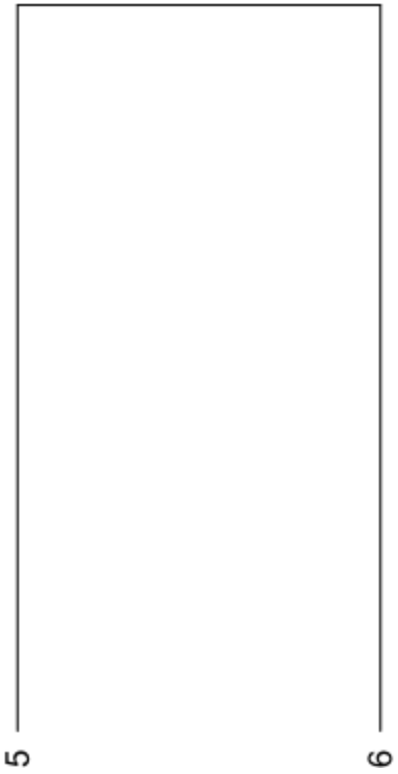
- Find “closest” points
- Merge
- Repeat

https://github.com/jtleek/jhsph753and4/blob/master/lectures/05_02_clustering/

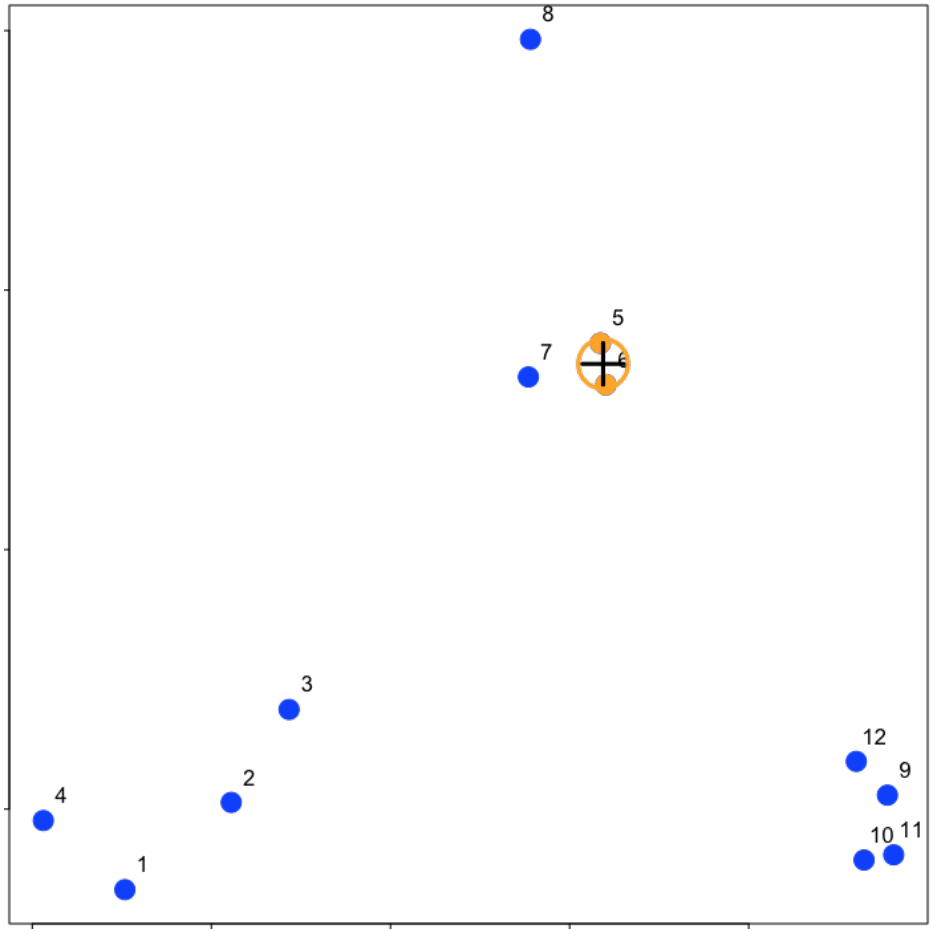


https://github.com/jtleek/jhsph753and4/blob/master/lectures/05_02_clustering/

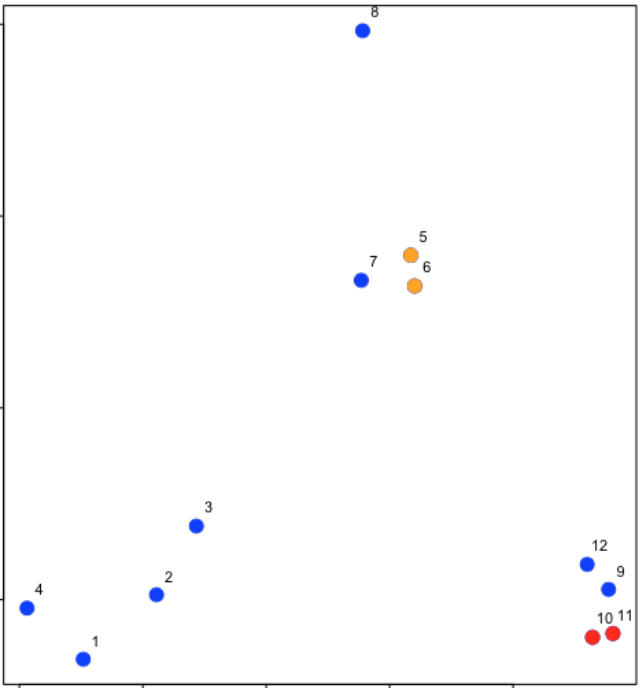


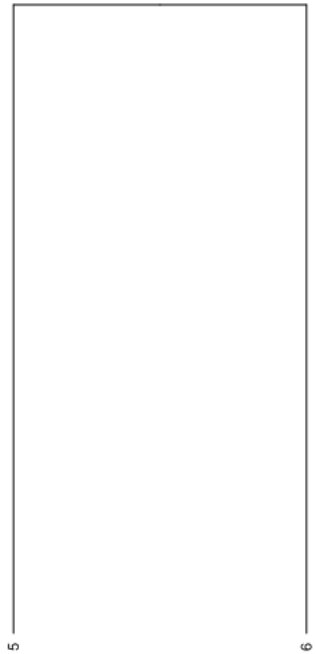


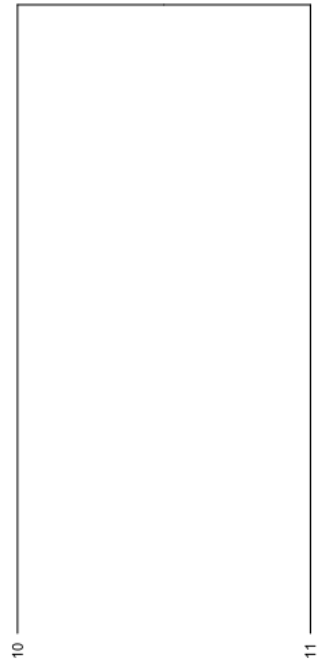
https://github.com/jtleek/jhsph753and4/blob/master/lectures/05_02_clustering/



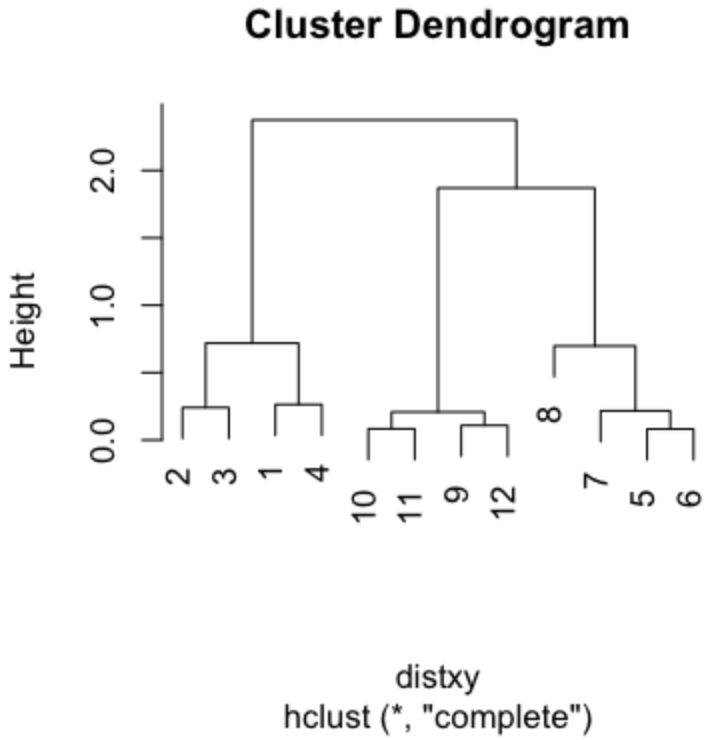
https://github.com/jtleek/jhsph753and4/blob/master/lectures/05_02_clustering/







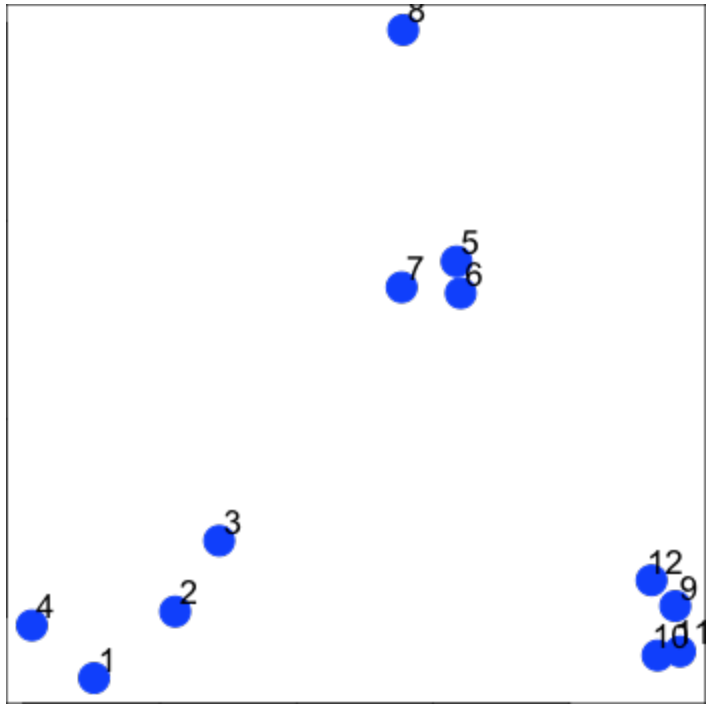
https://github.com/jtleek/jhsph753and4/blob/master/lectures/05_02_clustering/



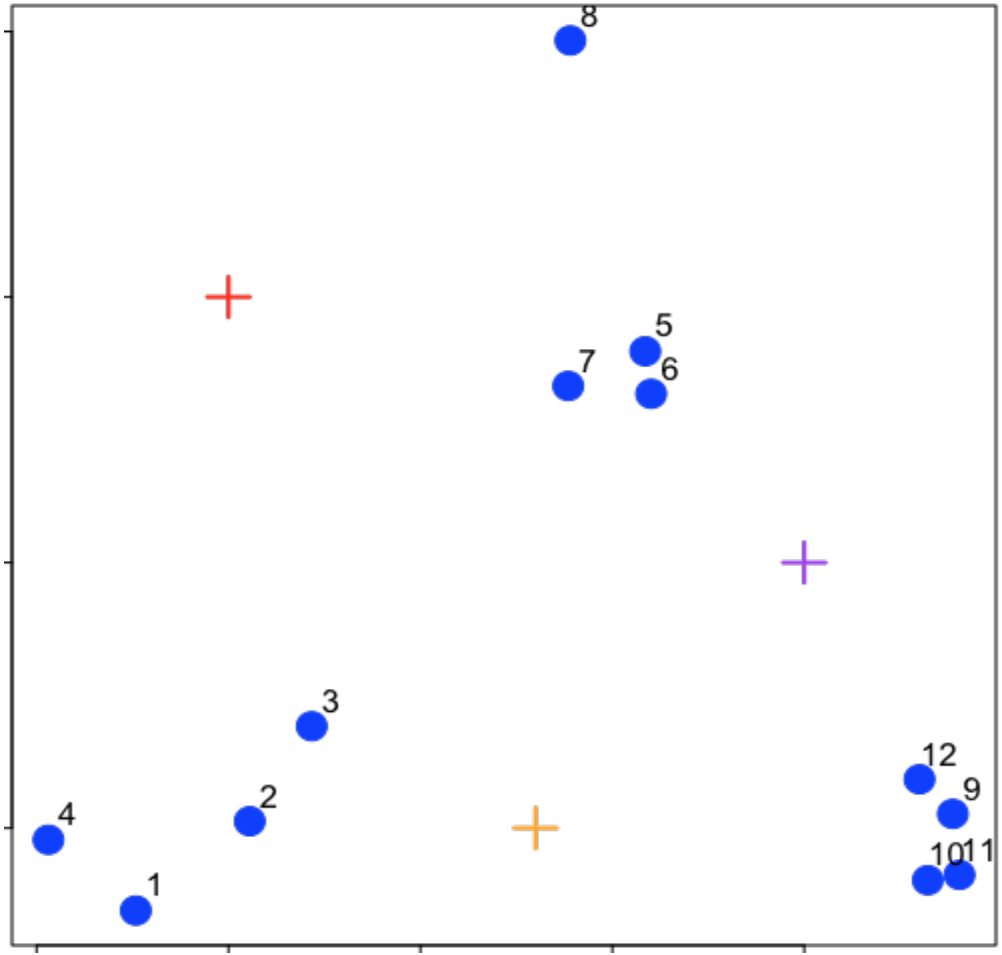
K-means clustering

- Initialize cluster “centers”
- Assign values
- Update centers
- Reassign values
- Repeat

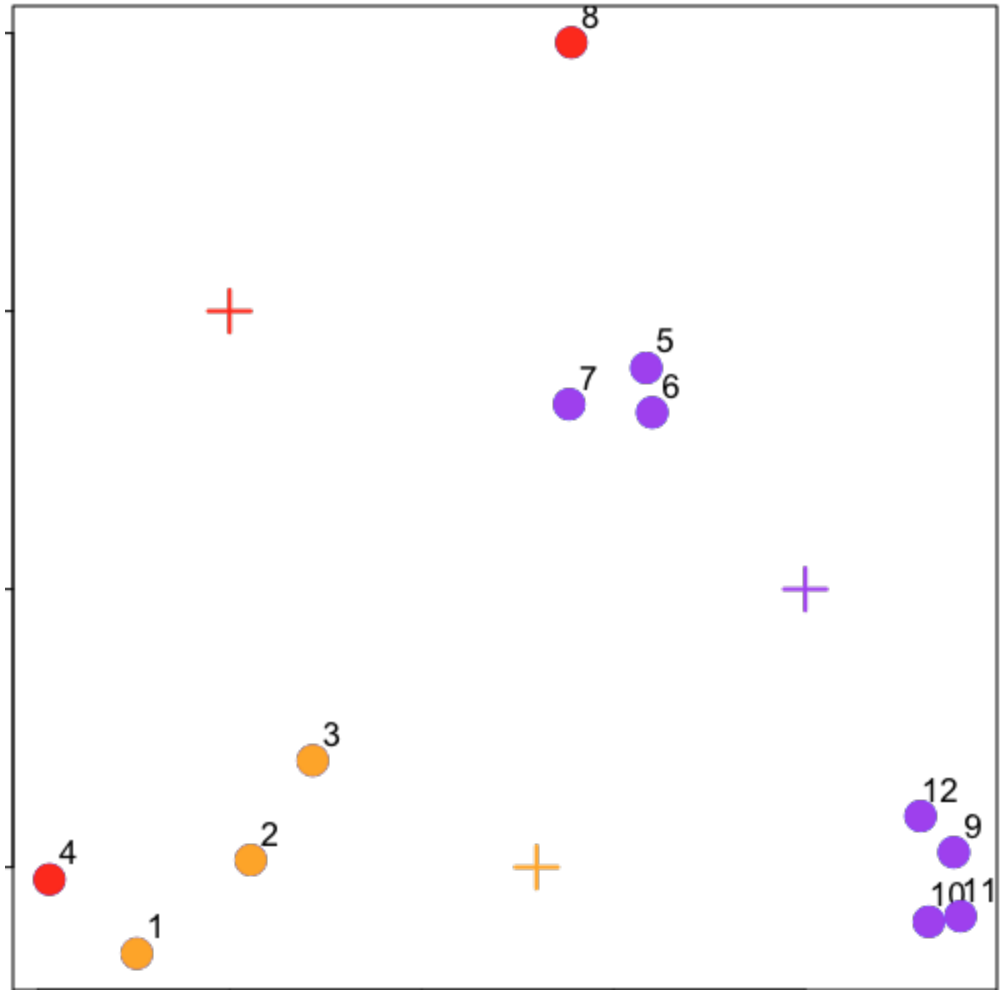
https://github.com/jtleek/jhsph753and4/blob/master/lectures/05_02_clustering/



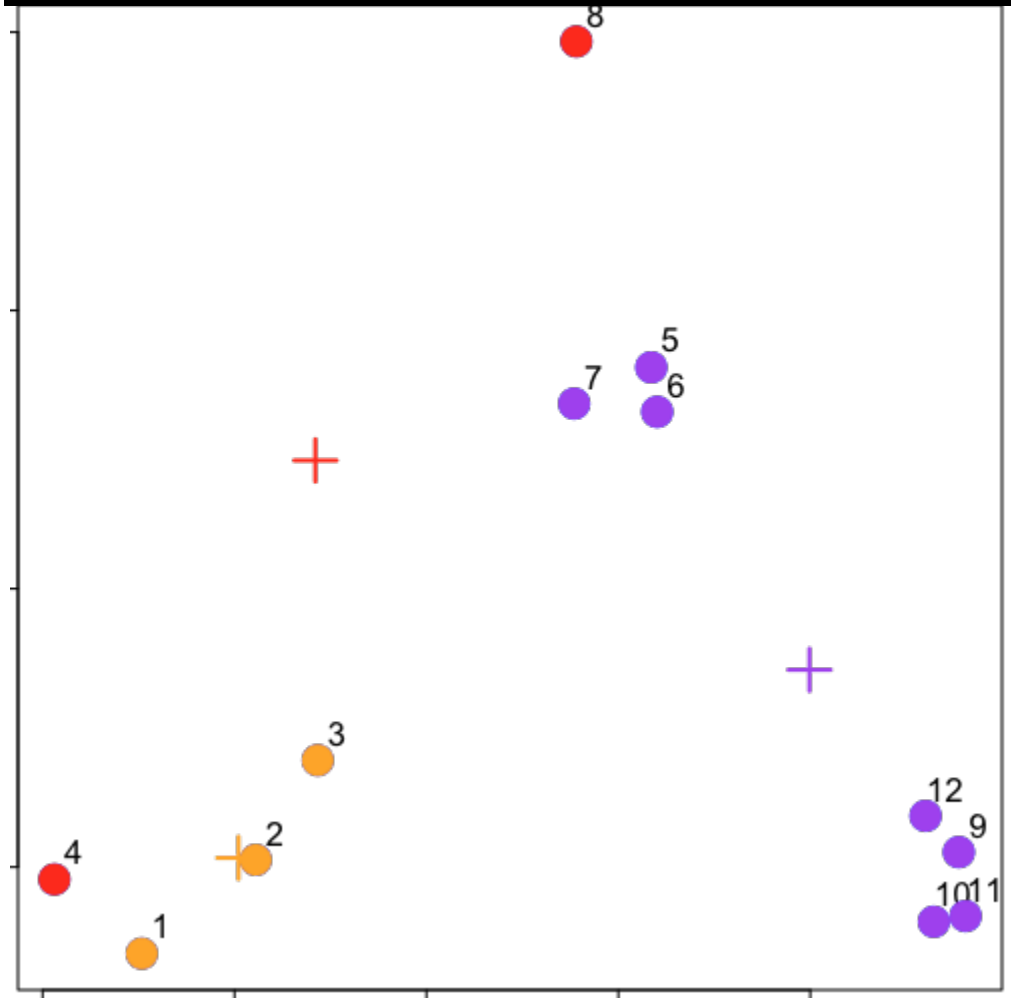
https://github.com/jtleek/jhsph753and4/blob/master/lectures/05_02_clustering/



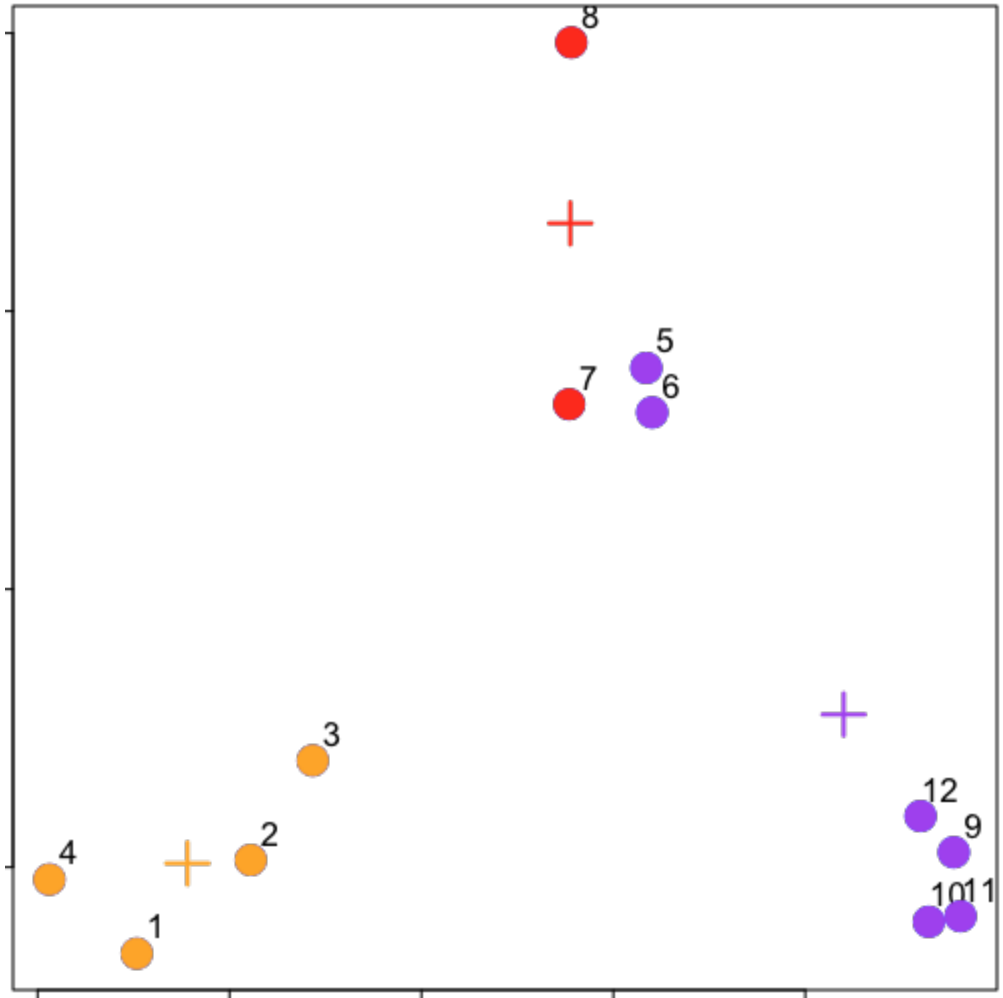
https://github.com/jtleek/jhsph753and4/blob/master/lectures/05_02_clustering/



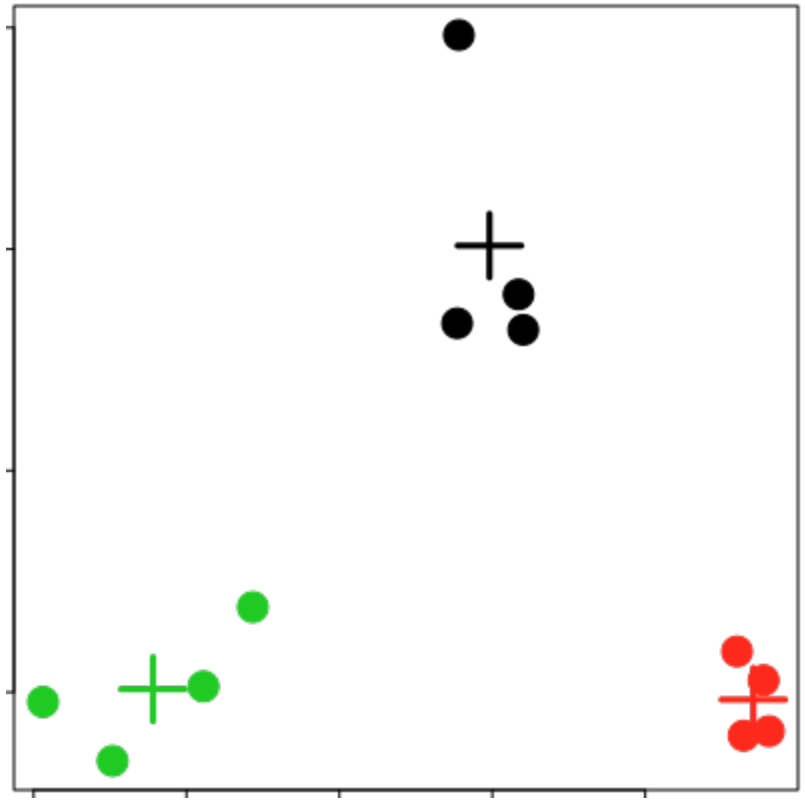
https://github.com/jtleek/jhsph753and4/blob/master/lectures/05_02_clustering/



https://github.com/jtleek/jhsph753and4/blob/master/lectures/05_02_clustering/



https://github.com/jtleek/jhsph753and4/blob/master/lectures/05_02_clustering/



Notes

- Can be useful for exploring multivariate relationships
- Things that have a bigger than expected impact
  - Scaling
  - Outliers
  - Starting values (k-means)
- Selecting the number of clusters isn’t trivial
- Better to visualize!
- Widely overutilized/overinterpreted

Further resources

- http://stat.ethz.ch/education/semesters/SS_2006/CompStat/sk-ch2.pdf
- http://www.cbcb.umd.edu/~hcorrada/PracticalML/
- Rafa's Distances and Clustering Video
- Elements of statistical learning
- Vadim's lecture notes
- http://www.public.iastate.edu/~maitra/stat501/lectures/ModelBasedClustering.pdf
- http://www.ics.uci.edu/~smyth/courses/cs274/readings/fraley_raftery.pdf