Clustering in Digital Humanities

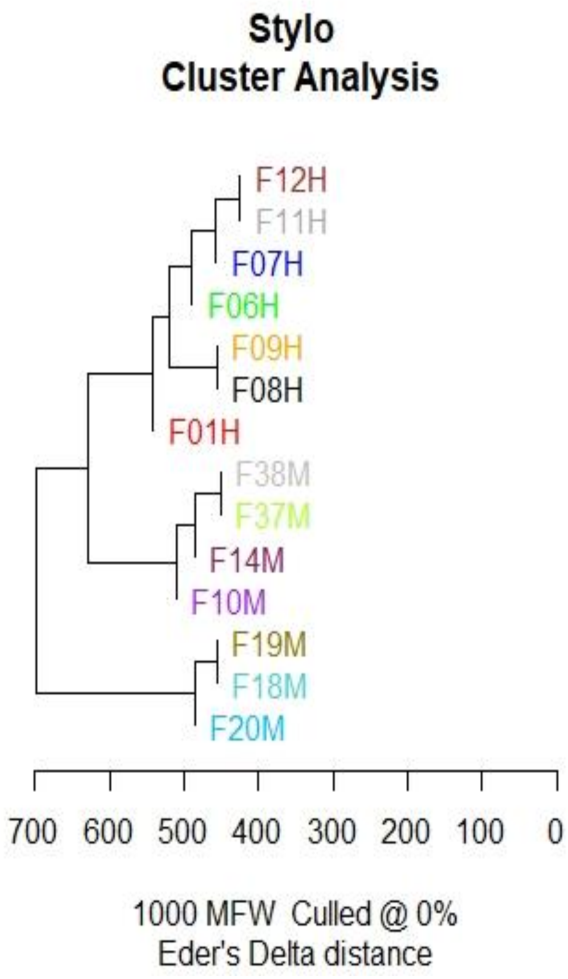
University of Tübingen, *Philosophische Fakultät*
Allgemeine Sprachwissenschat/Computerlinguistik,
Hauptseminar
Instructor Prof. Dr. John Nerbonne
Spring/Summer, 2022

Clustering

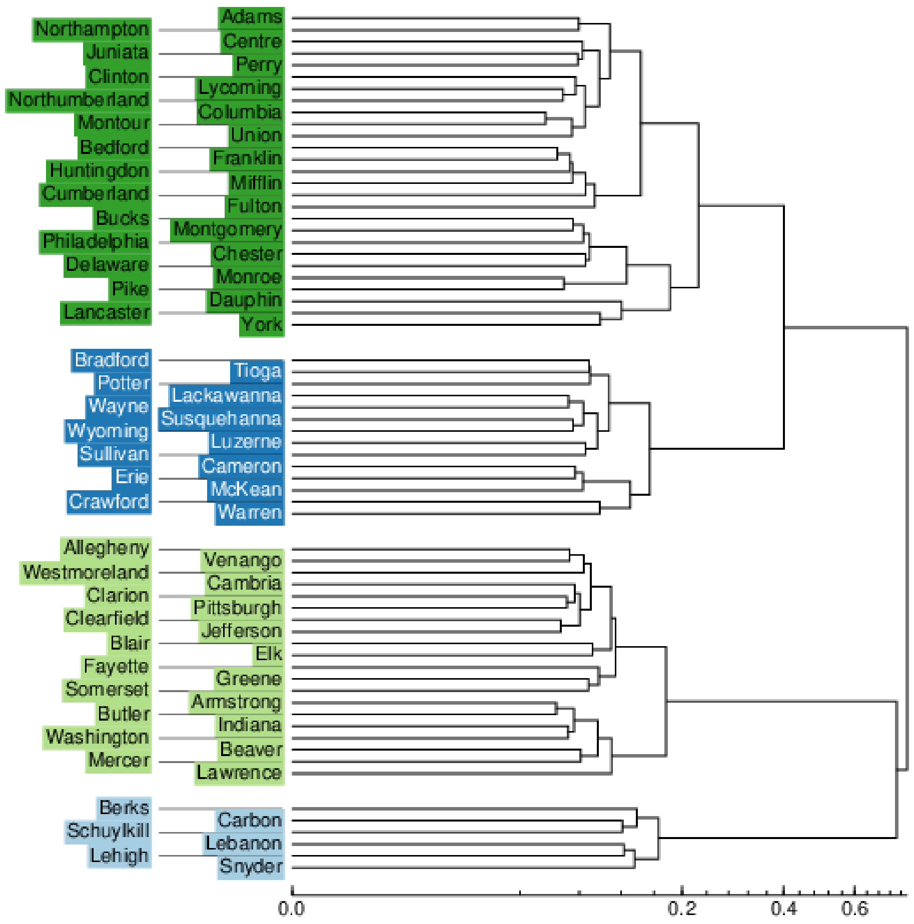
- Motivation
- General remarks, types
- Technique
- Quality

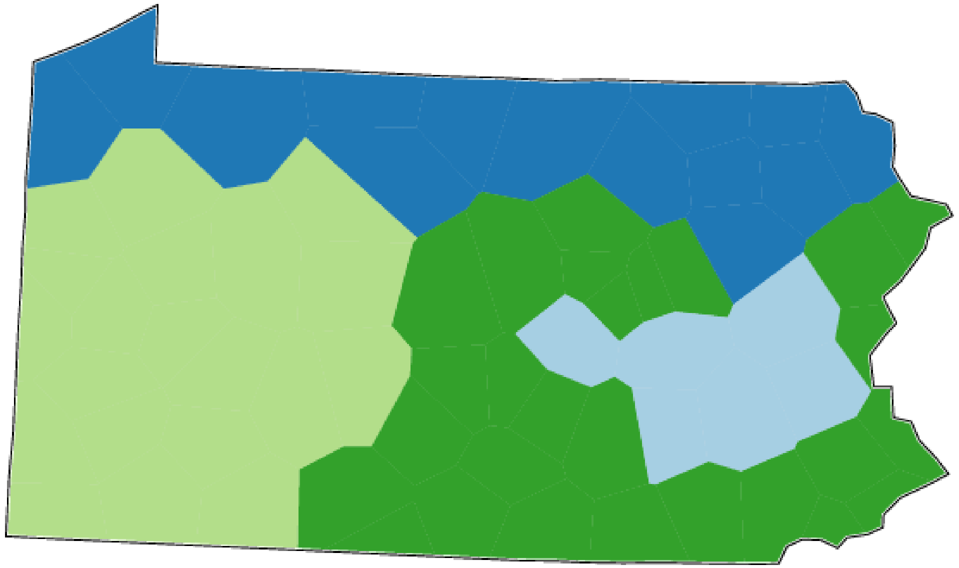
Clustering attractive for digital humanities

- Whenever we suspect the existence of groups (authors)
  - Similarly in dialectology, where sites are often categorized into regions!
- The results are presented in DENDROGRAMS (right)
  - 'H' Hamilton, 'M' Madison
- The branch length to the point of fusion indicates how different the fused varieties are
- One test of correctness in dialectology is the projection to geography



Projecting clusters to a map





- The clusters don’t have to be geographically coherent
- Usually a good sign if they are!
- From Gabmap: gabmap.nl

Clustering

- Motivation
- General remarks, types
- Technique
- Quality

Clustering – finding groups in data

- Clustering seeks similar groups
  - Technique in <u>exploratory</u> statistics (not hypothesis testing, not inferential)
  - Similarity/distance (in items and clusters) can be measured via Euclidean distance, Manhattan distance, …
    - In authorship attribution via Delta (based on MFWs, frequencies of most frequent words), in dialectology and historical linguistics, often based on (specialized) edit distance.

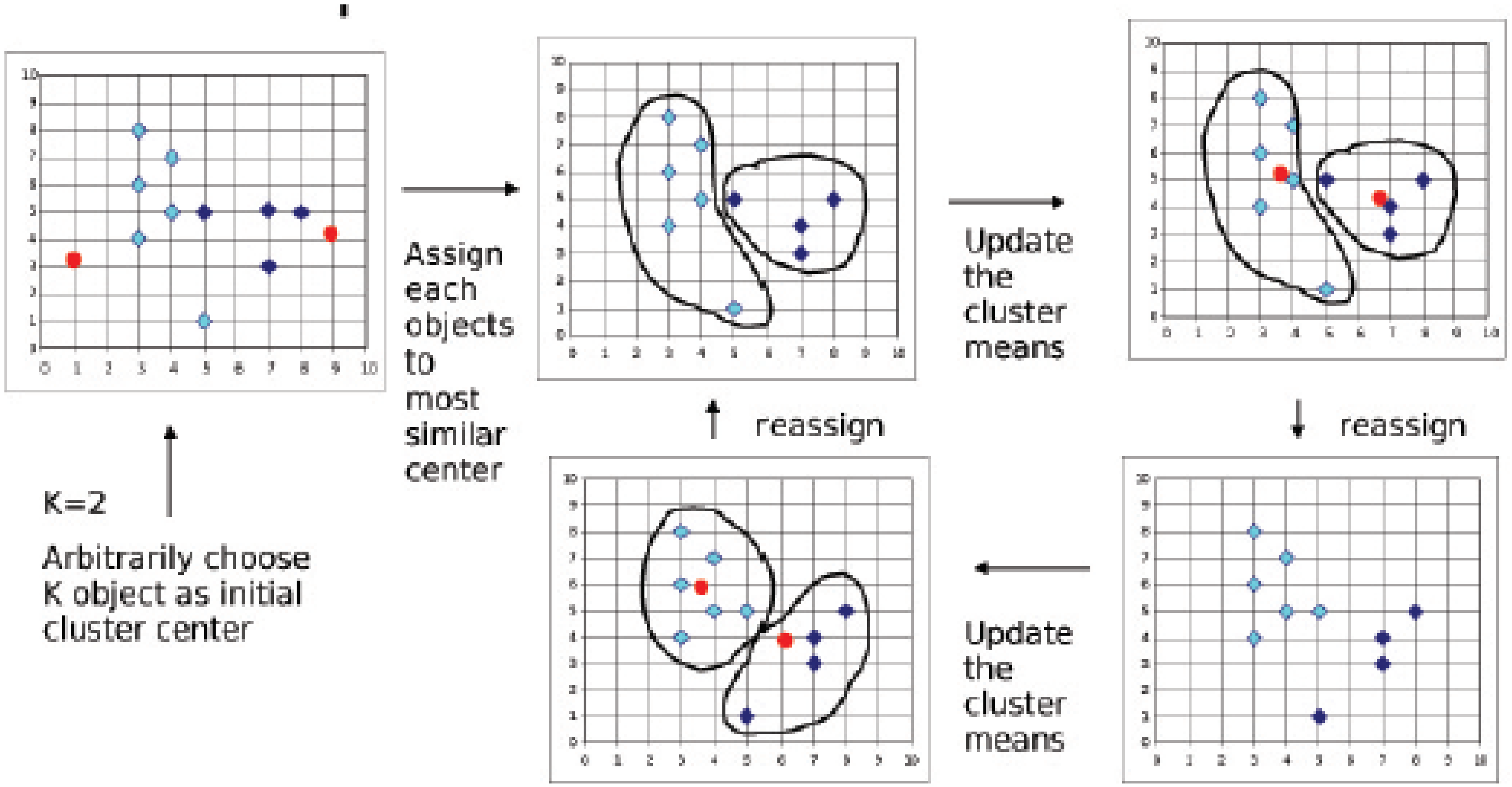
- Flat vs. hierarchical clustering
  - Flat clustering partitions data in (k) groups – no subgroups

- Hard vs. soft clustering
  - Hard: Each item belongs exactly one (immediate) supergroup
  - Soft: Items may belong to more than one supergroup

K-Means clustering

- Very popular in computational linguistics!
  - That’s why we discuss it here
- Hard clustering algorithm
- Starts by partitioning the input points into k initial sets (randomly)
- Calculates the mean point, or centroid, of each set
  - Note the need for a distance measure
- Constructs a new partition by associating each point with the closest centroid
- Repeats last two steps until the objects no longer switch clusters

K-Means (Sketch by Jelena Prokić)



K-means clustering in dialectology

- Not often used
  - Why not?
- Dialectologists find hierarchical structure in the data
  - Feldkirch is southern Baden, which is Baden, which is Alemannic, which is continental west Germanic, which is ….
- It is seen occasionally in dialectology (when there’s little interest in finer relations)
  - I don’t recommend it!
- Good candidate for use in authorship attribution if only the author is of interest
  - And not groups of authors, such as men vs. women, English vs. American, etc.
- Often used elsewhere in CL

Clustering

- Motivation
- General remarks, types
- Technique
- Quality

How to cluster hierarchically

- Input a distance table
  - Use points above the diagonal

	Grouw	Haarlem	Delft	Hattem	Lochem
Grouw	0	41	44	45	46
Haarlem	41	0	16	34	36
Delft	44	16	0	37	38
Hattem	45	34	37	0	20
Lochem	46	36	38	20	0

- Apply Johnson's algorithm
  - Iteratively,
    1. Select closest pair of data points
    2. Fuse the two points, reducing table size
       - Note that this requires assigning a distance from the new cluster to all other elements
  - Repeat until only one cluster is left

Reduction step in agglomerative clustering

- Determine closest pair in n×n table
  - Haarlem-Delft (16)

	Grouw	Haarlem	Delft	Hattem	Lochem
Grouw		42	44	46	47
Haarlem			16	36	38
Delft				38	40
Hattem					21
Lochem					

- Fuse these and reduce the table to (n-1) × (n-1)
  - Now we need to assign a distance from the fused element to all the others
  - For example, what's the new distance from Haarlem-Delft to Grouw?
  - Keeping it simple, we can use the mean of Haarlem-Grouw + Delft-Grouw , i.e. (42+44)/2 = 43

	Grouw	Haarlem & Delft	Hattem	Lochem
Grouw		**43**	46	47
Haarlem & Delft			**37**	**39**
Hattem				21
Lochem				

Pseudo-code for Johnson’s algorithm

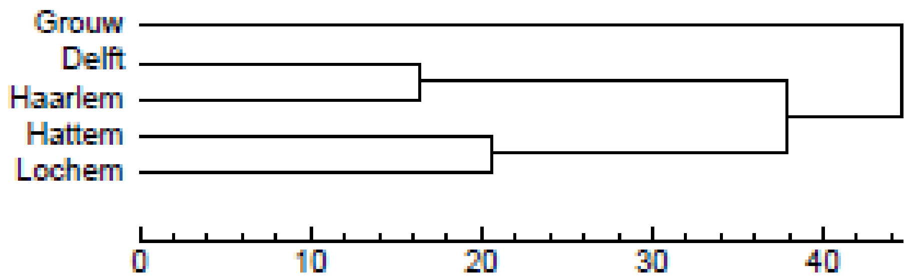
```
procedure cluster(DistanceMatrix,Cluster);
begin
  k:=number of elements;

  while elements or clusters are left that can be fused do begin
    k:=k+1;

    find pair (i,j) in DistanceMatrix that has smallest distance;
    store subclusters i and j in Cluster[k];
    distance between subclusters of Cluster[k]:=distance between i and j;

    delete rows and columns of i and j in DistanceMatrix;
    insert a row and a column of cluster k in the DistanceMatrix;
    calculate distances from cluster k to all remaining points;
  end;
end
```

Why?

Matrix Update

Notes on Johnson’s algorithm

- Keeping track of the distance between the elements being fused allows us to draw the dendrogram reflecting this
- Haarlem-Delft 16 diff.



- Lots of updating schemes!
  - OK: Farthest neighbor (complete link), unweighted mean (UPGMA), weighted mean (WPGMA), Ward’s method (minimize error)
  - Less useful: Nearest neighbor (single link), centroid methods (weighted & unweighted)
  - Prokić, Jelena, & John Nerbonne. 2008. "Recognizing groups among dialects." *Int. Journal of Humanities & Arts Computing* 2.1-2: 153-172.

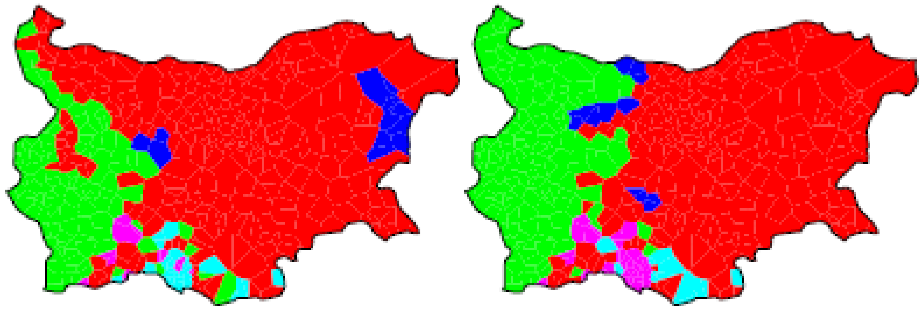
Clustering

- Motivation
- General remarks, types
- Technique
- Quality

Quality of Clustering

- There’s no perfect clustering algorithm
  - Kleinberg, Jon M. 2004. “An impossibility theorem for clustering” In: S. Thrun, S.Lawrence, & B. Schölkop (eds.), Advances in Neural Information Processing Systems 16: Proc. NIPS 16 (2003). Cambridge, MA: MIT Press.

- Clustering has a serious stability problem
  - A process is STABLE if small changes in input change the results only a little
  - Two Bulgarian datasets (*r=0.97*)
  - Clustered, projected to map



Promoting stability

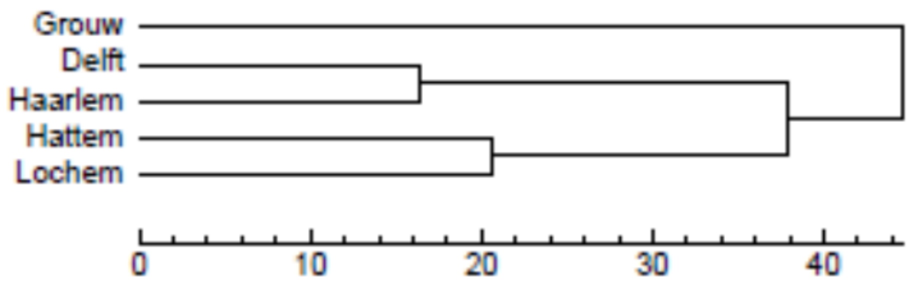
- There are several techniques that add stability
  - Jackknife
    - Cluster 10^n times, using different random subsets
  - Bootstrap (resampling)
    - Cluster 10^n times, resampling with replacement
  - Noisy clustering
    - Cluster 10^n times, adding different small amounts of noise

  - Nerbonne, John, et al. 2008. "Projecting dialect distances to geography: Bootstrap clustering vs. noisy clustering." *Data analysis, machine learning and applications*. Springer, Berlin, Heidelberg. 647-654. Avail. on JN's web site.

Measures of clustering quality

- Cophenetic correlation
  - How well do original distances (in input table) correlate with the distances assigned in the dendrogram?

	Grouw	Haarlem	Delft	Hattem	Lochem
Grouw	0	41	44	45	46
Haarlem	41	0	16	34	36
Delft	44	16	0	37	38
Hattem	45	34	37	0	20
Lochem	46	36	38	20	0



- The dendrogram systematically distorts the input distances, but in good clustering, the distortion is minimal, and the correlation is high (near 1)

A little information theory

- Since one of the cluster quality measures depends on information theory, we'll include a bit of that here.

- Entropy, etc. (other slides)

- Key: Entropy of a random variable:

Weighted average

$$H(v) = -\sum_{i=1}^{n} p(v_i) \log_2 p(v_i)$$

Number of (negative) bits needed to reduce uncertainty wrt one outcome

CL measures (external)

- Given a gold standard of what items the clusters should group, we can measure (for each cluster)
  - Its ENTROPY (to what extent does the cluster represent a single class?)

$$\mathrm{E}(S_r) = -\frac{1}{\log q}\sum_{i=1}^{q}\frac{n_r^i}{n_r}\log\frac{n_r^i}{n_r}$$

    - where S_r, n_r, n_r^i the number of elements of class i in S_r, $\max(n_r^i)$, the number of elements in the largest class
    - as above q the number of classes in the cluster

  - This simply applies the definition of the entropy of a random variable to $\frac{n_r^i}{n_r}$, then takes the mean over all classes. Recall

$$H(v) = -\sum_{i=1}^{n} p(v_i)\log_2 p(v_i)$$

Purity

- Its purity (to what extent is just one group label in the cluster?)

  - $P(S_r) = \frac{1}{n_r} \max(n_r^i),$

  - where (S_r) is the cluster, n_r the size of (S_r), n_r^i the number of elements of class i in S_r, $\max(n_r^i)$, the number of elements in the largest class

- Take the largest class (i) in the cluster, report its size relative to cluster size.
  - $0 \leq P(S_r) \leq 1$

Cluster quality

- Basically, the external CL measures reflect how uniform a given cluster is (purity, the extent to which one class dominates), and how divergent it is (the extent to which several classes are included).

  - Prokić, Jelena, and John Nerbonne. 2008. "Recognizing groups among dialects." *Int. Journal of Humanities and Arts Computing* 2.1-2:153-172.

Clustering in digital humanities

- The great advantage of clustering is that it yields groups of data items that can be compared to groups traditional scholarship
  - Authors of documents, regions in language areas, survey respondents, …
- Clustering can be seen in dendrograms
  - … and/or projected to maps if geography might be relevant
- Hierarchical clustering is preferred where theory suggests
  - Consensus among dialectologists
- Always report cluster quality, e.g., using cophenetic correlation
  - Purity & Entropy if you have gold standard data
- Big stability problem – if you cluster, use jackknife, bootstrap,or noisy clustering!