HPE DSI 311
Introduction to Machine Learning

Summer 2021

Instructor: Ioannis Konstantinidis

UNIVERSITY of HOUSTON
DIVISION OF RESEARCH
HEWLETT PACKARD ENTERPRISE DATA SCIENCE INSTITUTE

Clustering

- K means
- Hierarchical
- Dendrograms

What is clustering?

Unsupervised Learning

Clustering is unsupervised classification: no predefined classes (only X, no y)

Cluster analysis

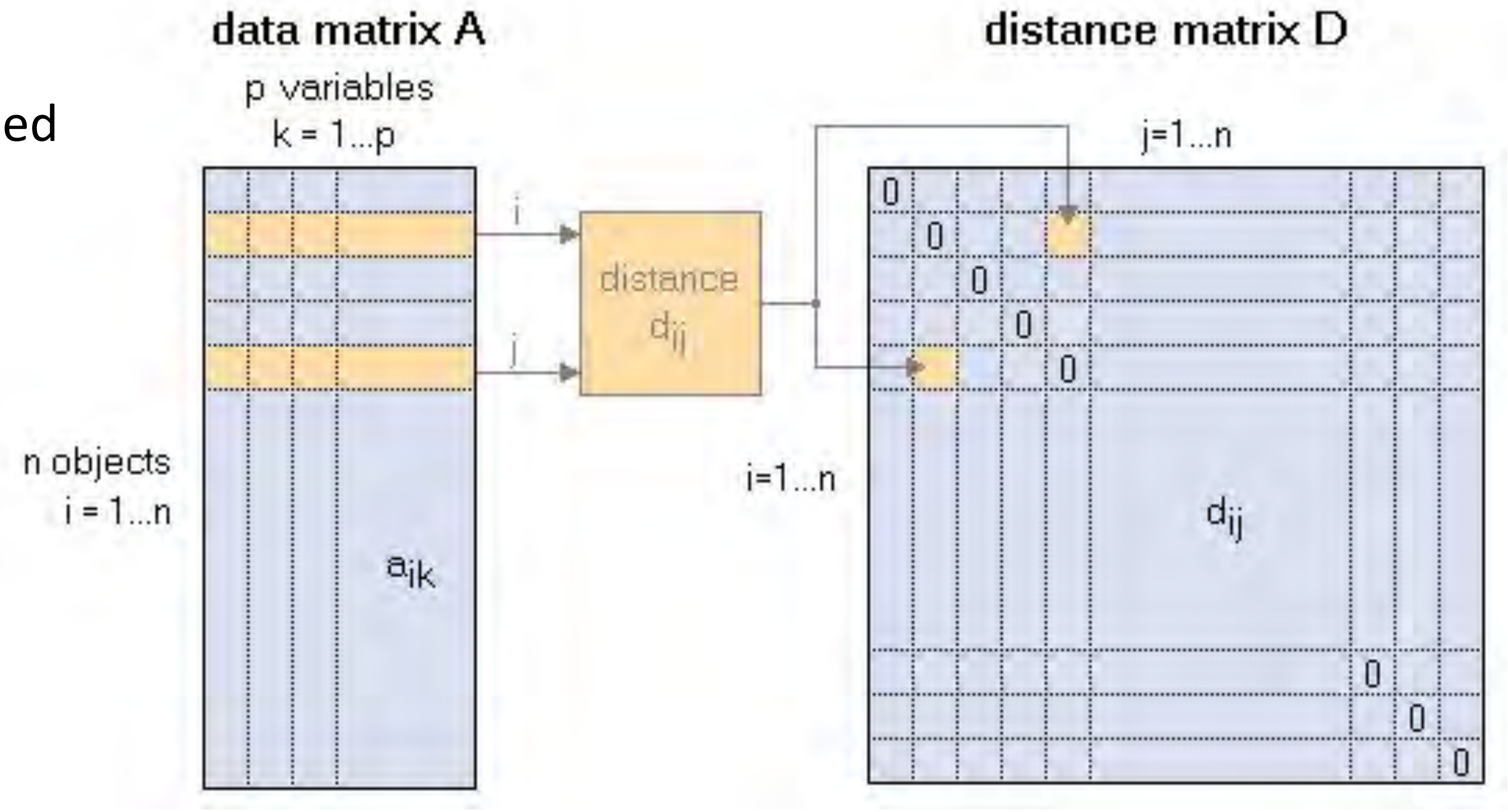
- Grouping a set of data objects into clusters
- Assign (instead of predict) values for y

Cluster: a collection of data objects

- Similar to one another within the same cluster
- Dissimilar to the objects in other clusters

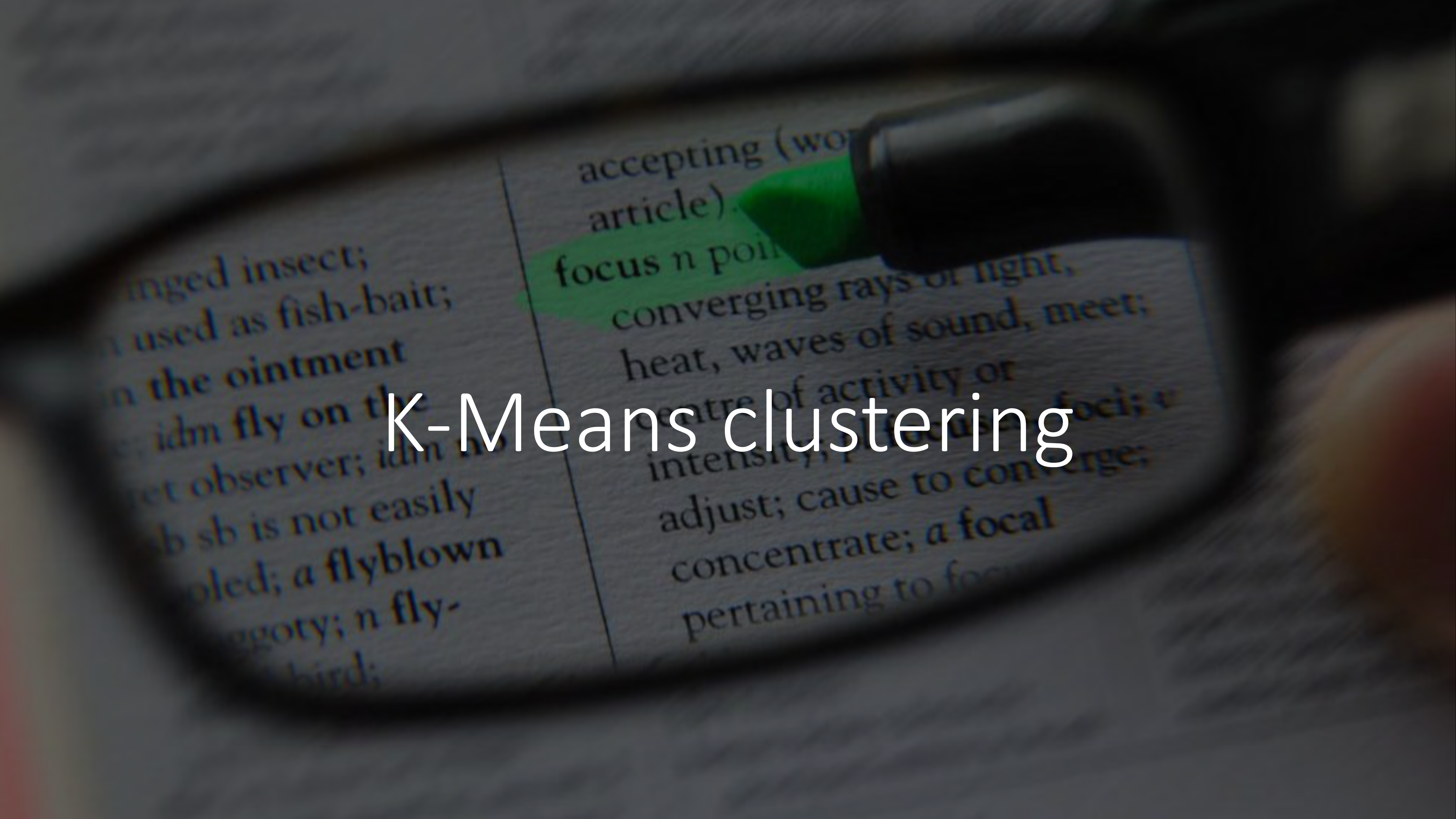
Similarity matrix

Dissimilarity is expressed as a distance function d(i, j)



Could be Minkowski, e.g.,

- Euclidean,
- manhattan



K-Means clustering

The k-Means Clustering Method

Given k, split samples into k groups of equal variance, minimizing a criterion known as the *inertia* or within-cluster sum-of-squares (WCSC):

$$\sum_{i=0}^{n} \min_{\mu_j \in C}(||x_i - \mu_j||^2)$$

Initialize: Choose k initial centroids.

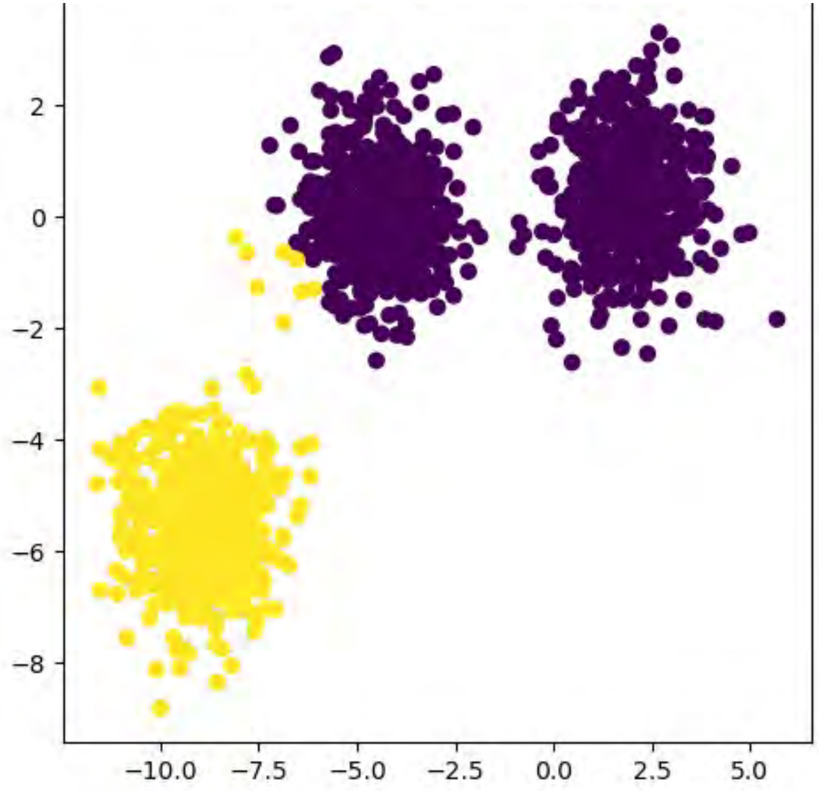
1. Assign each sample to its nearest centroid.
2. For each group created in Step 1, compute
    - the centroid of all of the samples assigned to that group.
    - the difference between the old and the new centroids for that group

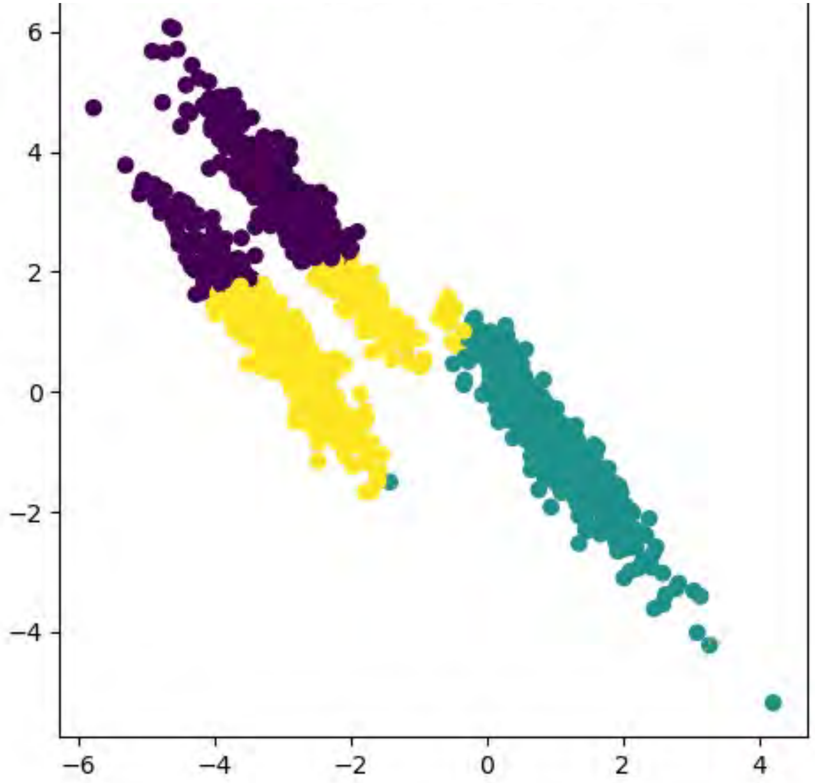
Repeat until the new centroids do not change significantly.

Cautions

- Given enough time, K-means will always converge, however this may be to a local minimum.
- This is highly dependent on the initialization of the centroids.
- As a result, the computation is often done several times, with different initializations of the centroids.

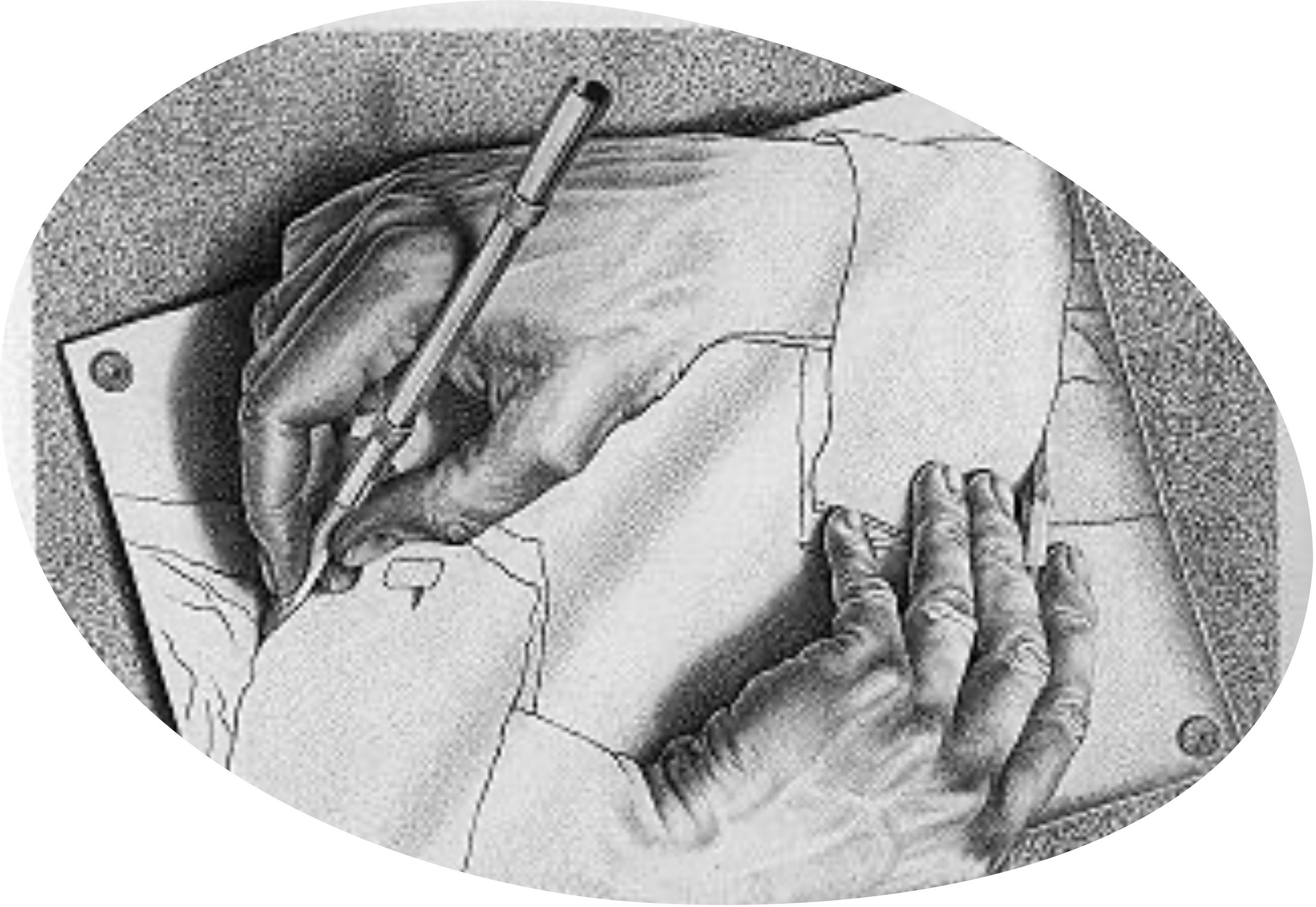
Failure mode examples





Kmeans()

- *n_clusters:* The number of clusters to form = the number of centroids to generate
- *Init:* Method for initialization
  - Default is ‘k-means++’ : selects initial cluster centers for k-mean clustering in a smart way to speed up convergence.
- *n_init=10:* Number of time the k-means algorithm will be run with different centroid seeds. The final results will be the best output of n_init consecutive runs in terms of inertia.



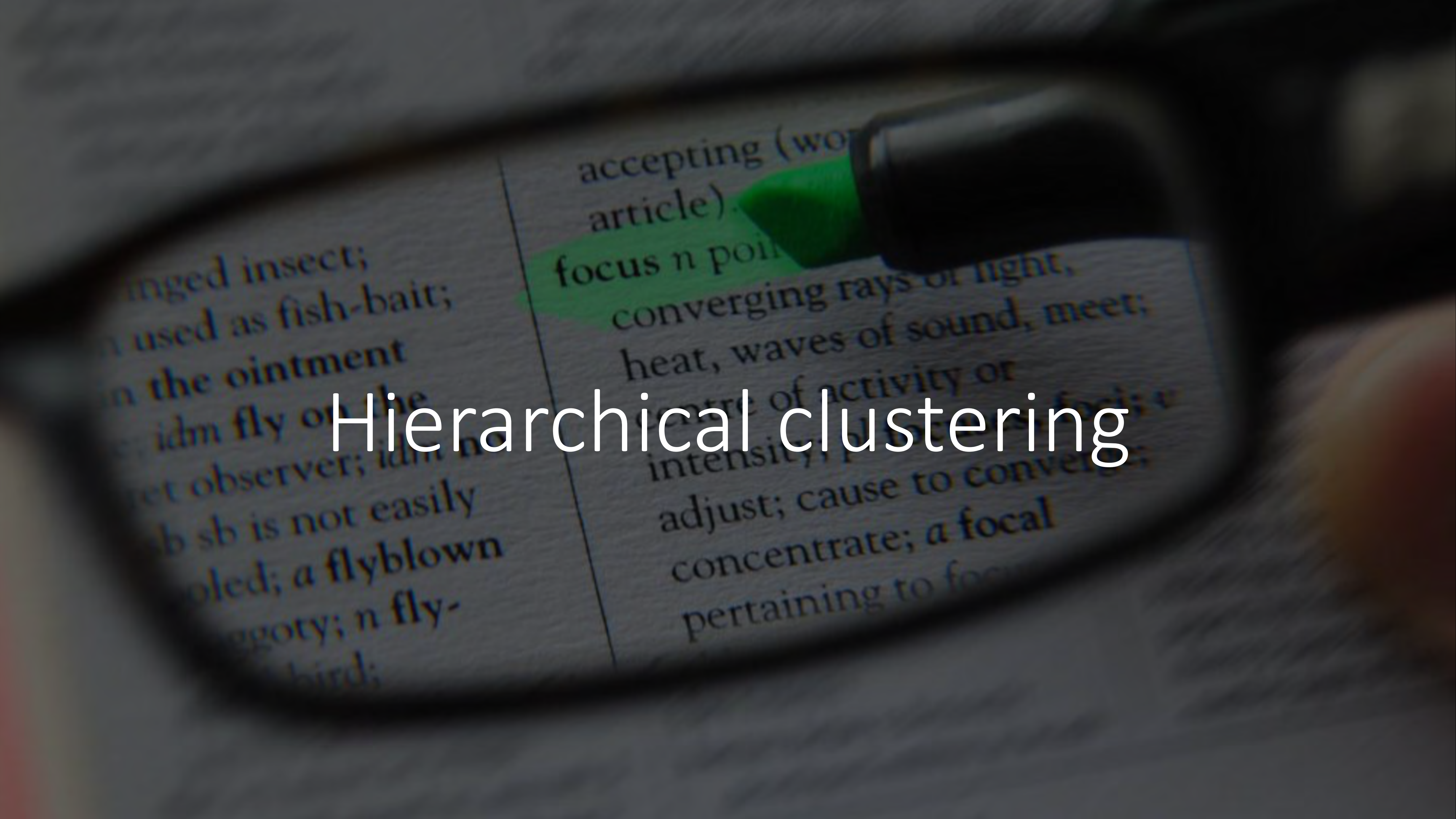
Hands-on Example:

K means

Hard Clustering vs Hierarchical Clustering

- Hard: Divide objects into a set number of groups (clusters)
- Hierarchical: organize clusters in a hierarchy

Hierarchical clustering

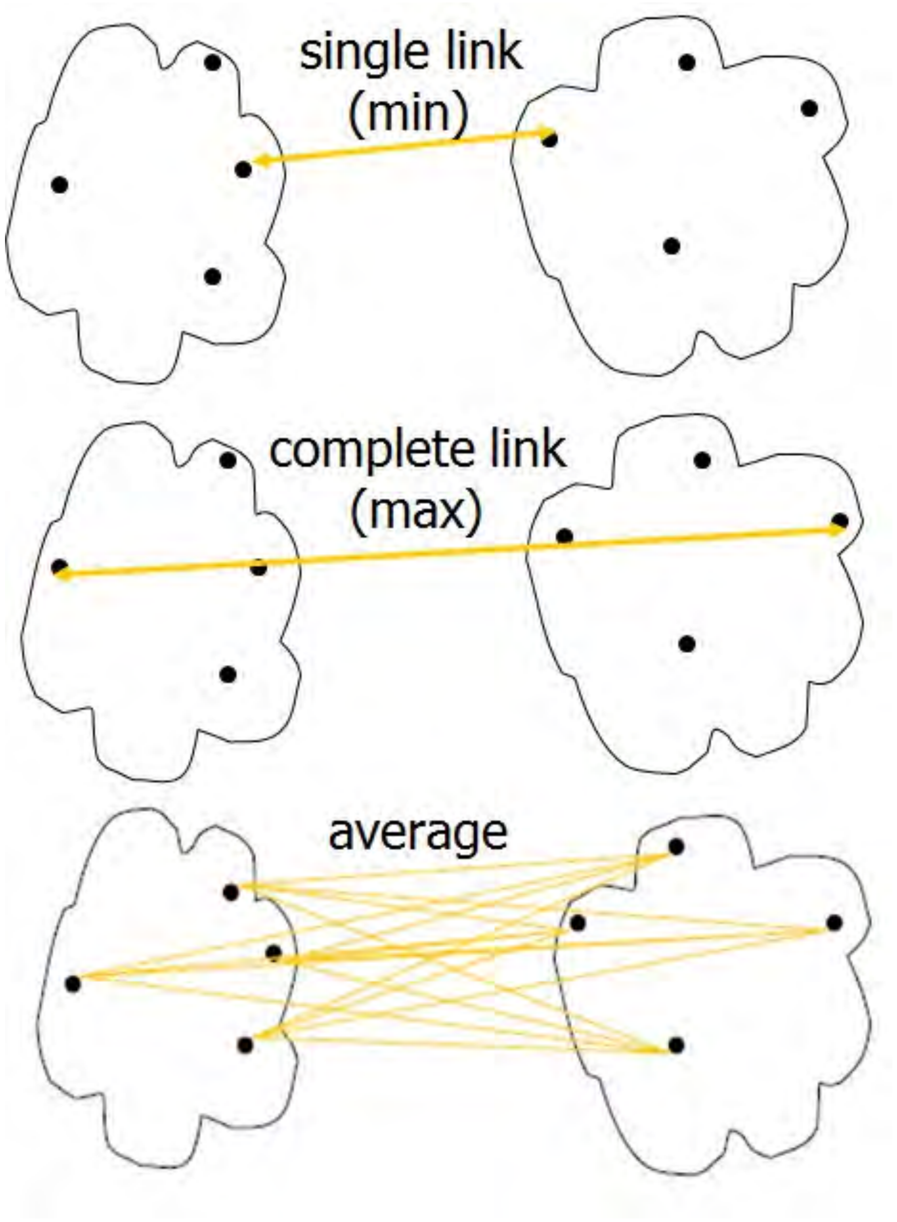


Hierarchical clustering

- Top-down approach:
  - start with **all samples** in the dataset as one cluster
  - **divide** that cluster into subclusters based on criteria
  - repeat for each subcluster until done
- Bottom-up approach:
  - starts with **every single sample** in the dataset as its own cluster
  - **merge** samples into superclusters based on criteria
  - repeat for each supercluster collection until done

Nesting criteria: Linkage

- Single link: the distance between the nearest pair of points from each cluster
- Complete link: the distance between the farthest pair of points from each cluster
- Average link: the mean distance between all the pairs of points from each cluster
- Ward: Same distance concept as K-Means

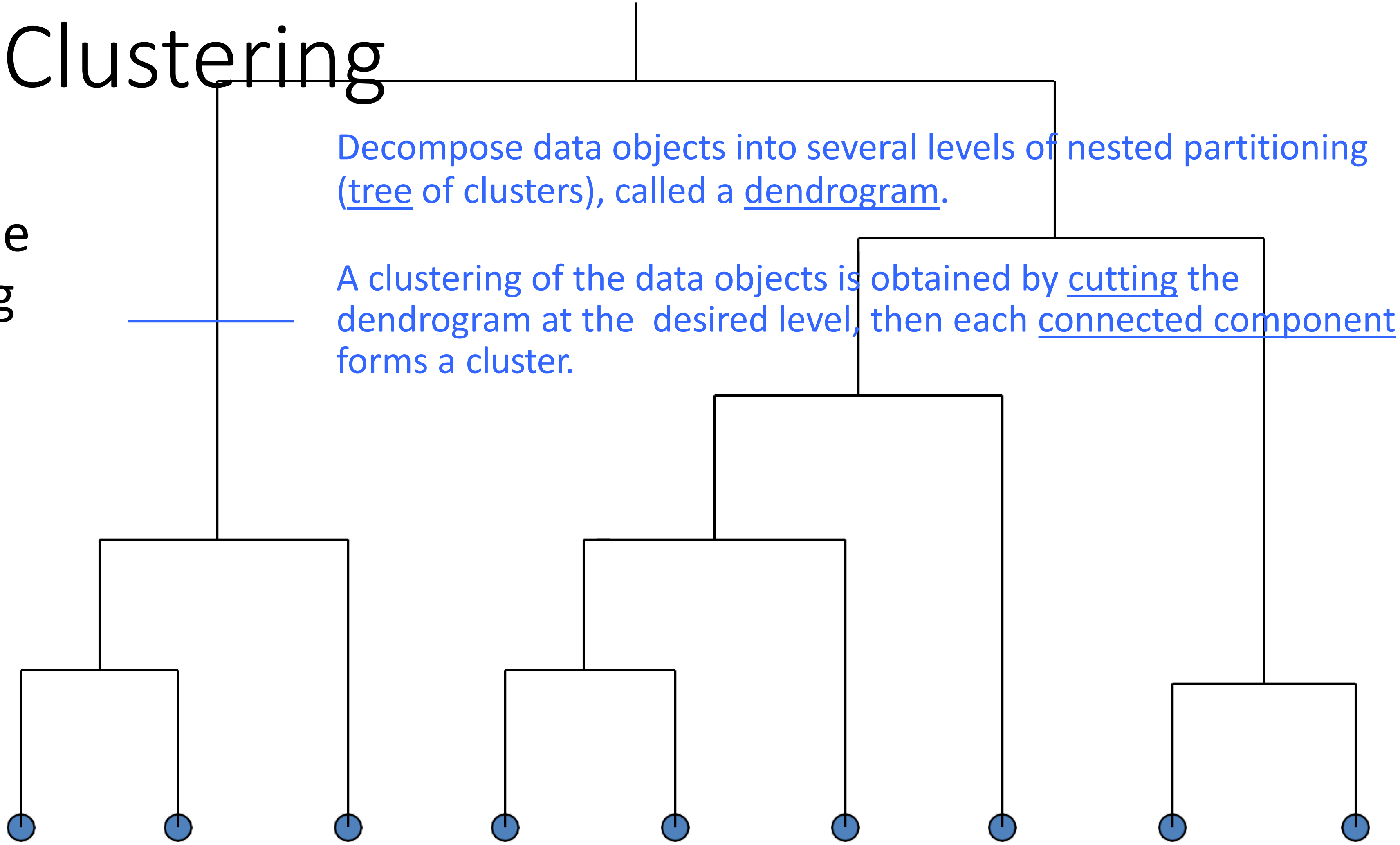


AgglomerativeClustering()

- *Affinity:* Metric used to compute the linkage. Can be “euclidean”, “l1”, “l2”, “manhattan”, “cosine”, or “precomputed”.
- *Connectivity:* Connectivity matrix. Defines for each sample the neighboring samples following a given structure of the data.
- *Linkage:* Which linkage criterion to use. The linkage criterion determines which distance to use between sets of observation. The algorithm will merge the pairs of cluster that minimize this criterion.

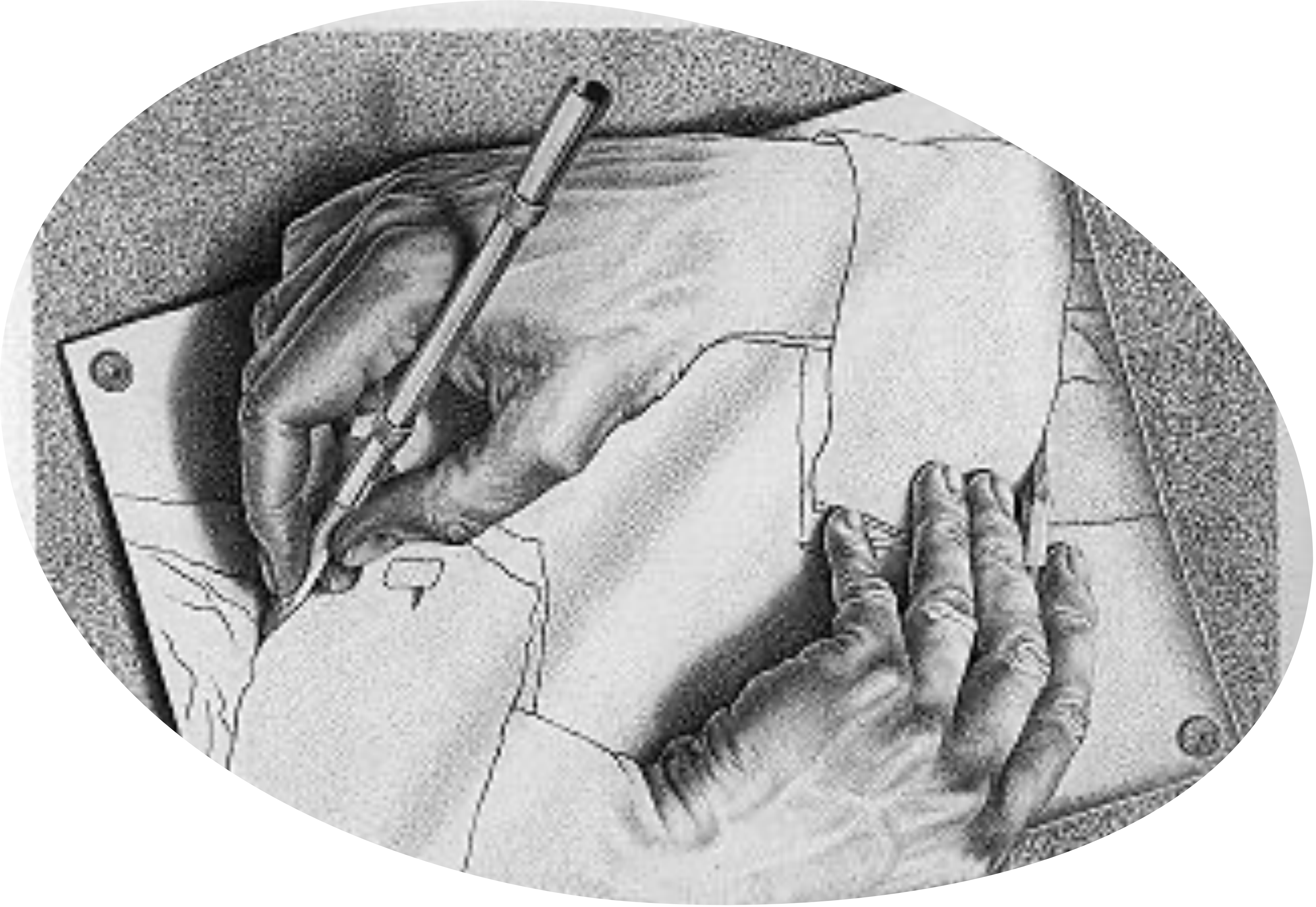
Hierarchical Clustering

- We can visualize the results of clustering as a dendrogram.
- This helps us in deciding when we want to stop clustering further (how "deep") by setting "depth" with some threshold.



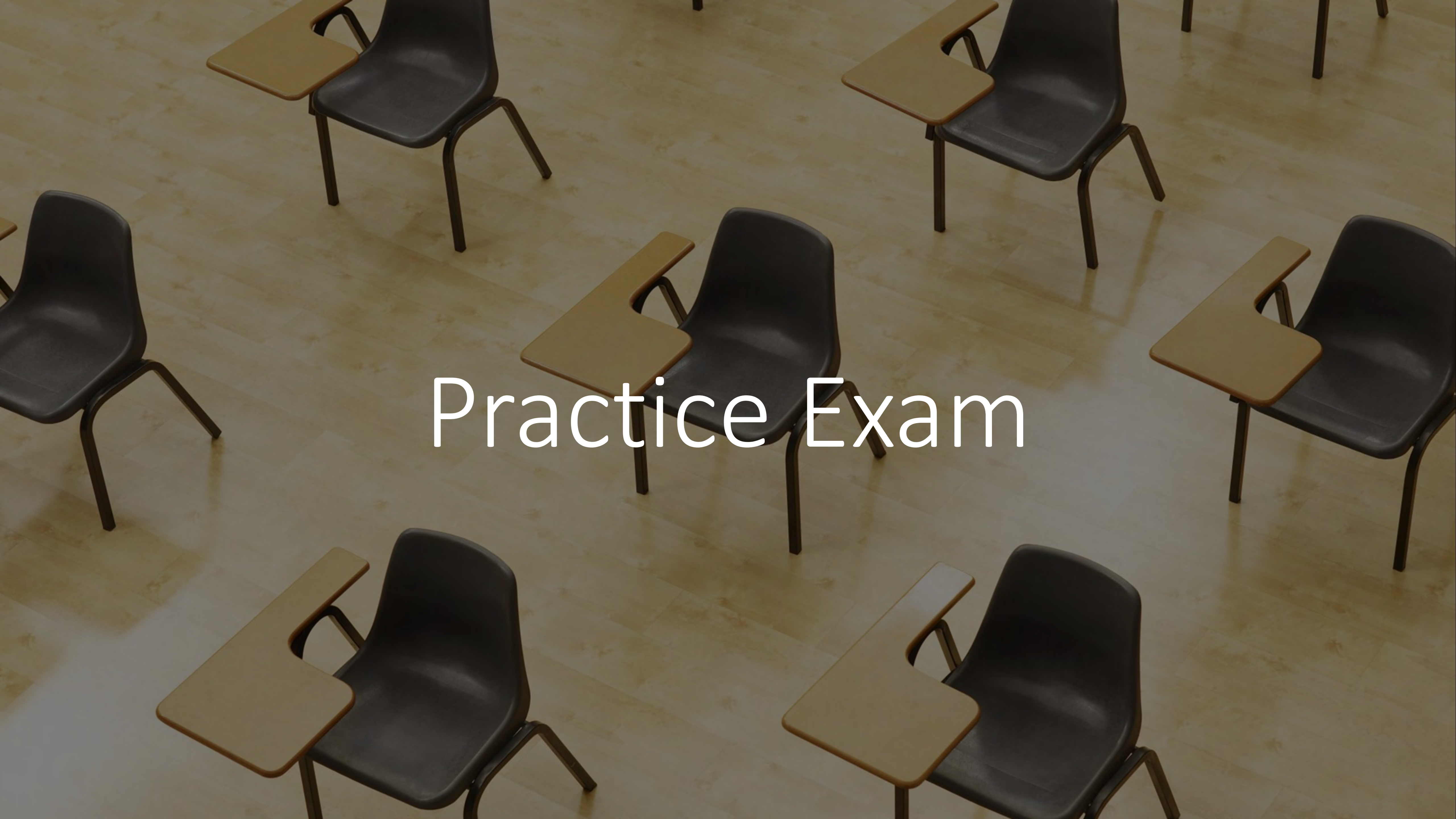
Decompose data objects into several levels of nested partitioning (tree of clusters), called a dendrogram.

A clustering of the data objects is obtained by cutting the dendrogram at the desired level, then each connected component forms a cluster.



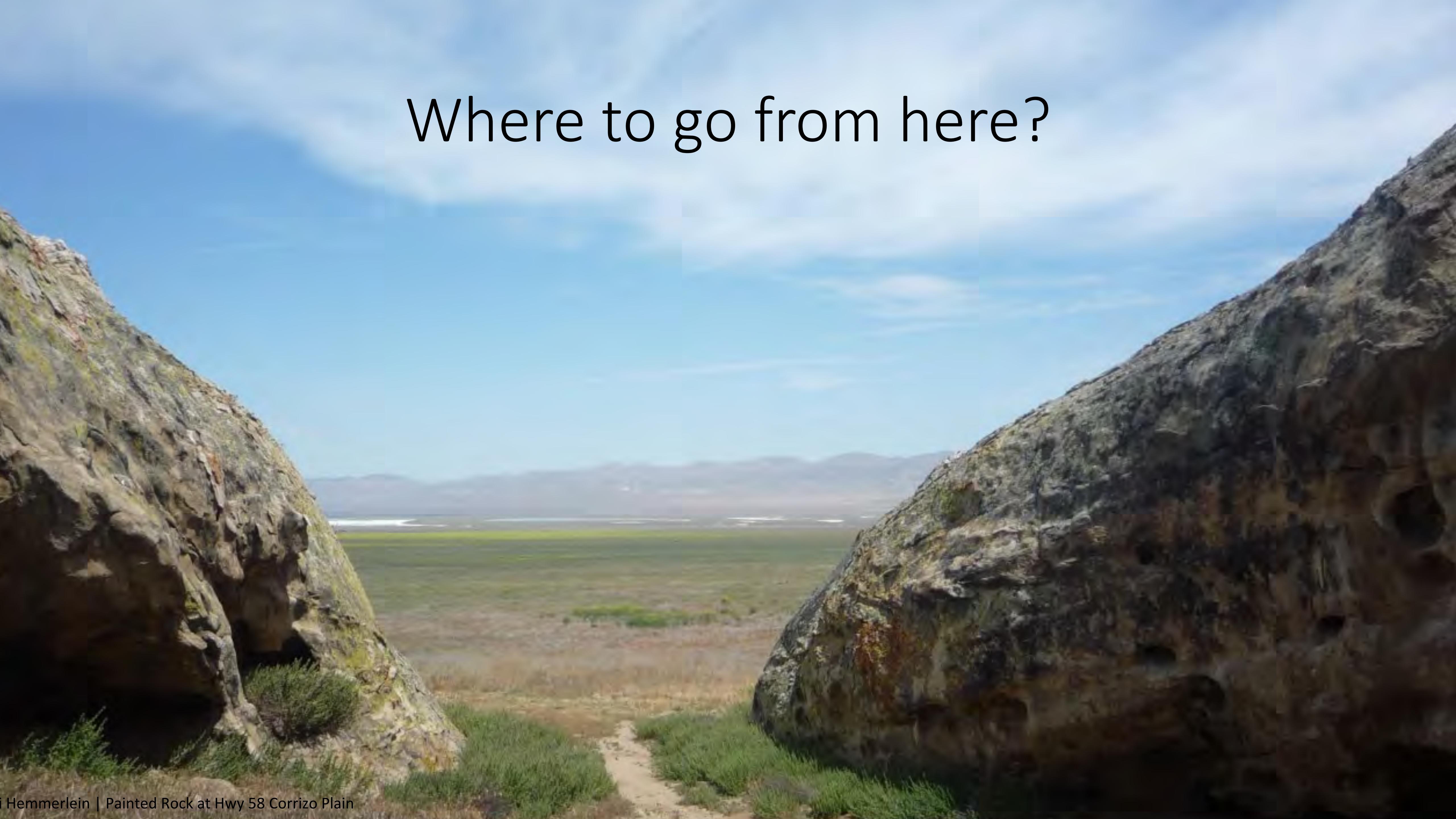
Hands-on Example:

Hierarchical Dendrograms



Practice Exam

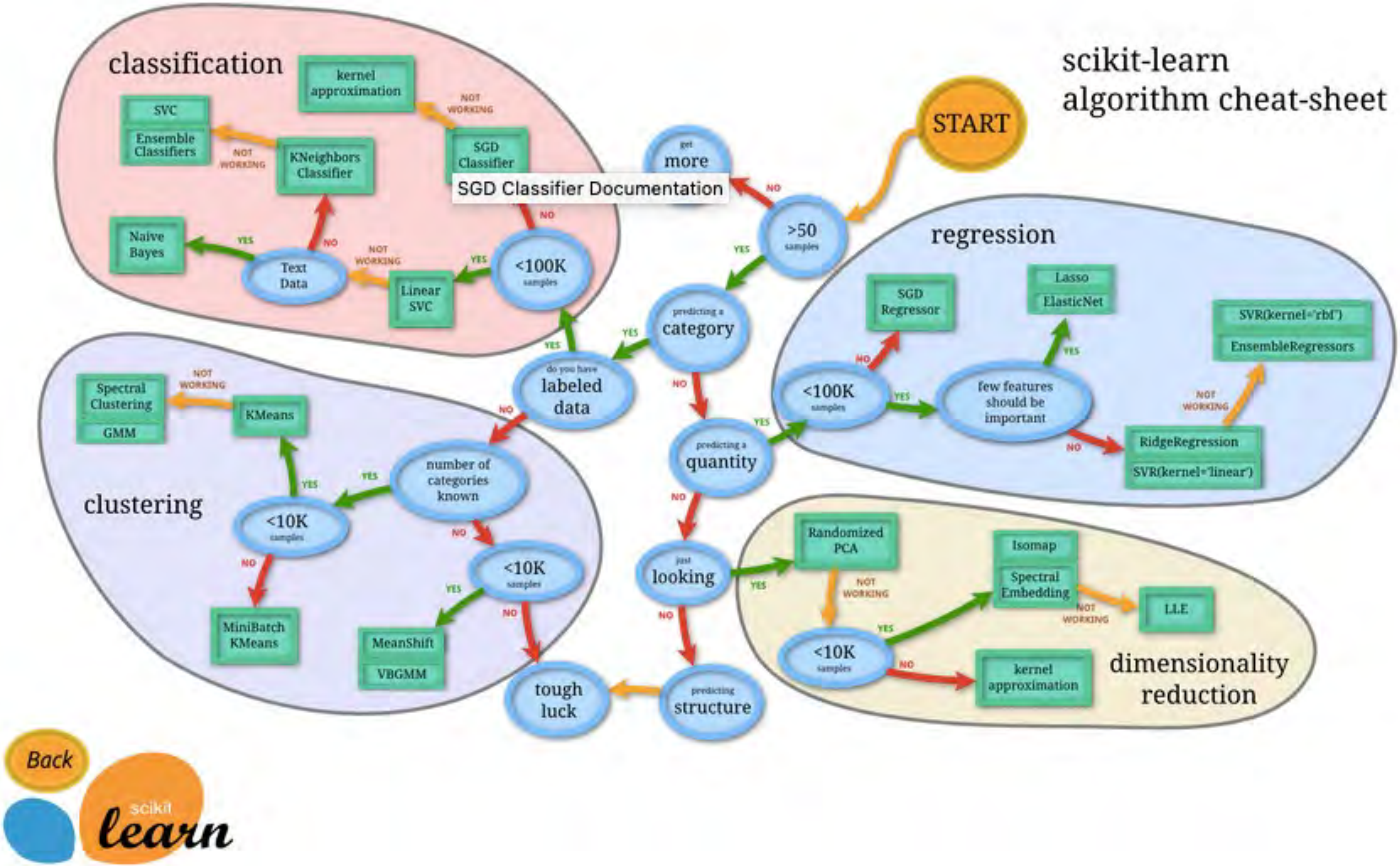
Where to go from here?



i Hemmerlein | Painted Rock at Hwy 58 Corrizo Plain

Course Feedback

Please fill out the feedback form in Moodle!


scikit-learn
algorithm cheat-sheet
START
>50 samples
NO
get more data
YES
predicting a category
YES
do you have labeled data
YES
<100K samples
YES
Linear SVC
NOT WORKING
Text Data
YES
Naive Bayes
NO
KNeighbors Classifier
NOT WORKING
SVC
Ensemble Classifiers
NO
SGD Classifier
NOT WORKING
kernel approximation
SGD Classifier Documentation
classification
NO
number of categories known
YES
<10K samples
YES
KMeans
NOT WORKING
Spectral Clustering
GMM
NO
MiniBatch KMeans
NO
<10K samples
YES
MeanShift
VBGMM
NO
tough luck
clustering
NO
predicting a quantity
YES
<100K samples
NO
SGD Regressor
YES
few features should be important
YES
Lasso
ElasticNet
NO
RidgeRegression
SVR(kernel='linear')
NOT WORKING
SVR(kernel='rbf')
EnsembleRegressors
regression
NO
just looking
YES
Randomized PCA
NOT WORKING
<10K samples
YES
Isomap
Spectral Embedding
NOT WORKING
LLE
NO
kernel approximation
dimensionality reduction
NO
predicting structure
Back
scikit learn

- https://developers.google.com/machine-learning/guides
- https://www.kaggle.com/learn/overview