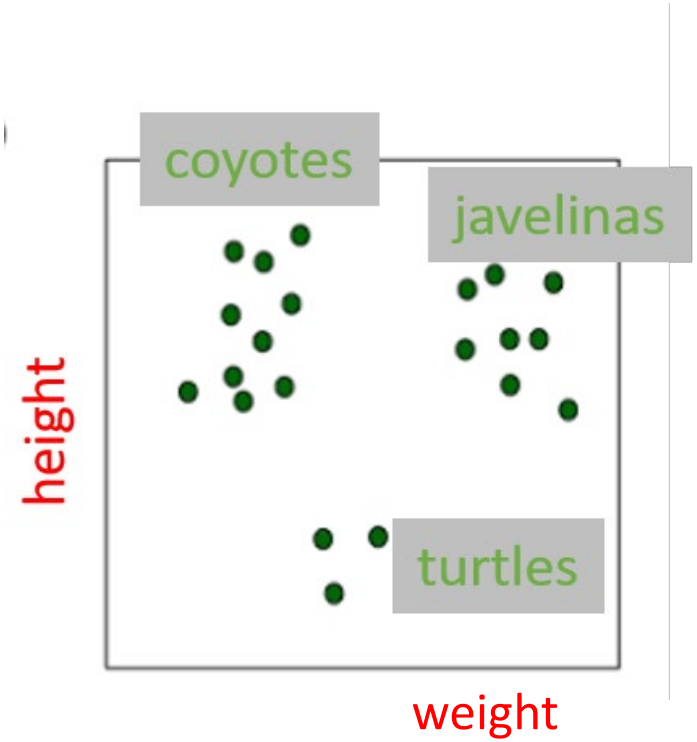
EEE511
Unsupervised Learning

Learning Paradigms

- Supervised learning
  predict target value y given input features x
- Unsupervised learning
  - Understand patterns of data represented in features x without explicitly provided label y, anomaly detection
  - Use a small set of prototypes to characterize what's in the dataset (e.g., vector quantization, data compression)
  - Use a small number of parameters to capture relevant properties of the data (e.g., PCA, estimation of principal subspaces of received signals)
  - Clustering: objects are grouped together into clusters if they are similar

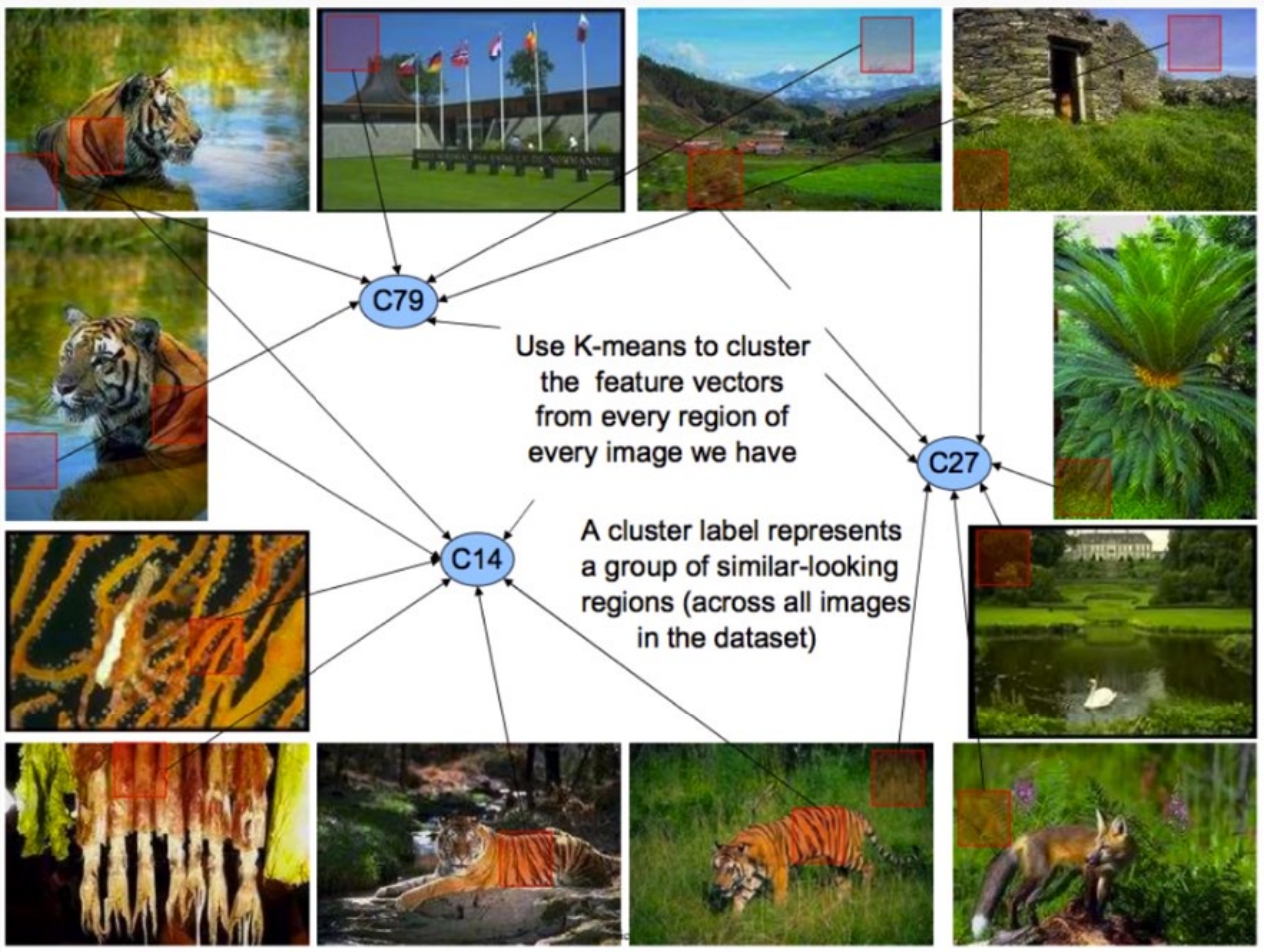


- **Clustering** described data by “groups”
- Labels can be added post hoc and be used for **classification**

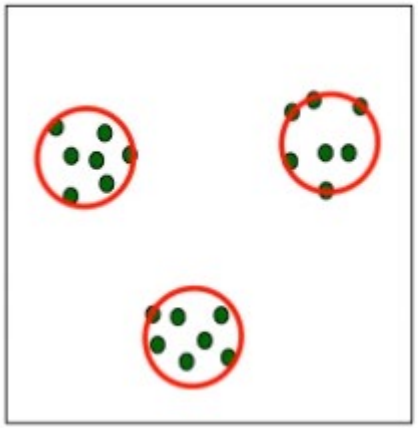
Clustering and Data Compression

Clustering is related to vector quantization

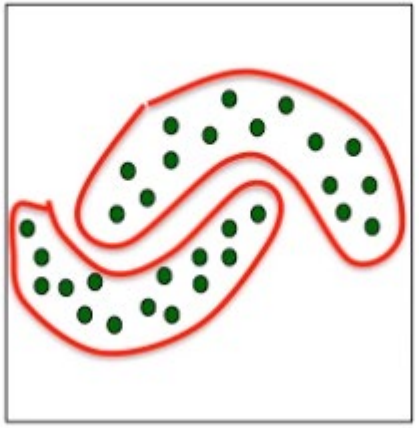
- Dictionary of vectors (the cluster centers)
- Each original value represented using a dictionary index
- Each center 'claims" a nearby region (Voronoi region)



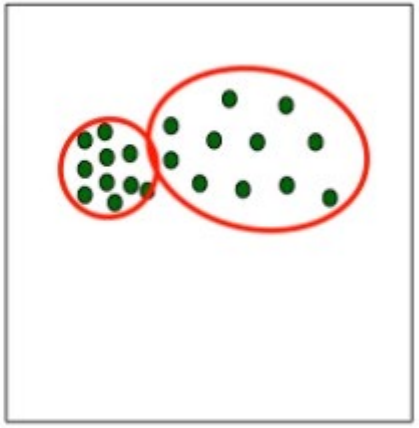
Clusters identified by



Location



Shape/connectivity



Density

Hierarchical clustering (connectivity-based)

- More related objects are nearby
- Connect/merge objects to form clusters based on distance
- At different distances, different clusters will form
- A dendrogram representation (x axis being the objectives grouped into non-overlapping clusters, y axis being the distance at which clusters merge)
- Do not scale well (O(n^3) for agglomerative, O(2^(n-1) for divisive), provide a good theoretical foundation for cluster analysis but considered obsolete

DBSCAN (Density-based spatial clustering of applications with noise)

- a density-based non-parametric clustering method
- points are grouped together if closely packed together
- outliers are those lie alone in low-density regions
- Schubert, Erich, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu. "DBSCAN revisited, revisited: why and how you should (still) use DBSCAN." ACM Transactions on Database Systems (TODS) 42, no. 3 (2017): 1-21.

OPTICS (ordering points to identify the clustering structure)

- finding density-based clusters in spatial data
- the basic idea is similar to DBSCAN, additional advantage of detecting meaningful clusters in data of varying density
- Points are ordered such that spatially closest points become neighbors in the ordering. Each point is associated with a special distance representing the density for a cluster used for placing two points in the same cluster (using a dendrogram).

Summary of popular clustering algorithms

K-means
Expectation-Maximization (EM)
Hierarchical agglomerative clustering (HAC)
DBSCAN
Optics
…