Clustering: unsupervised learning

goal: find patterns/structure in unlabeled data

idea: assume data consits of multiple blobs

↳ determine
- what are blob parameters
- which points belong to which blob

cluster = blob

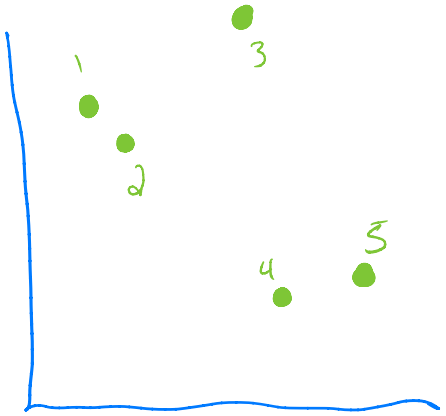
applications:
- summarize dataset by summarizing clusters
- example: cluster shoppers and see where each cluster tends to shop

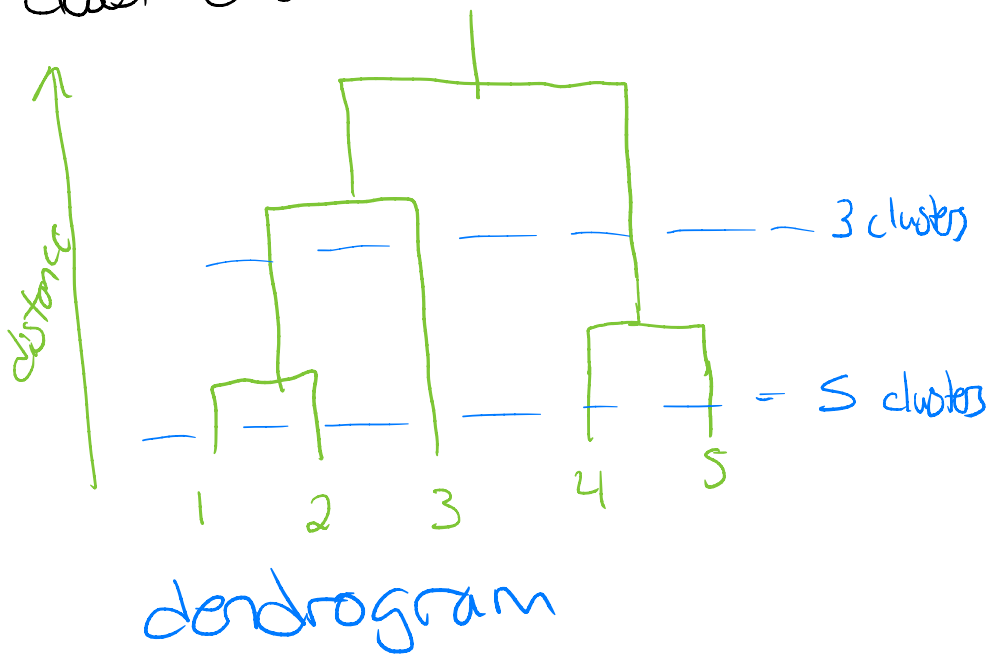
Hierarchical Clustering

- 2 Types
  - agglomerative clustering
    - ↳ merge items/clusters that are similar

Algorithm:

1) choose distance metric and inter-cluster distance
2) assign each point to its own cluster
3) Repeat until clustering satisfactory (or until single cluster)
   a) compute distances between clusters
   b) merge two closest clusters into one





dendrogram

Intercluster Distance:

- distance between data points does not give us definition for distance between clusters
- single-link clustering:
  - choose distance between closest elements (greatest similarity)
  - tends to yield extended clusters
- complete-link clustering:
  - choose maximum distance between element of each (least similarity)
  - tends to yield rounded clusters
- group average clustering:
  - use average of distances between elements
  - tends to yield rounded clusters

- divisive clustering
  - need spliting method
    - ↳ natural split tends to be more application specific
- Either agglomerative / divisive
  - no right place to stop
    - ↳ typically show whole dendrogram
  - scaling of features is important
    - rescale
    - normalize to 0 mean, unit variance
    - whitening