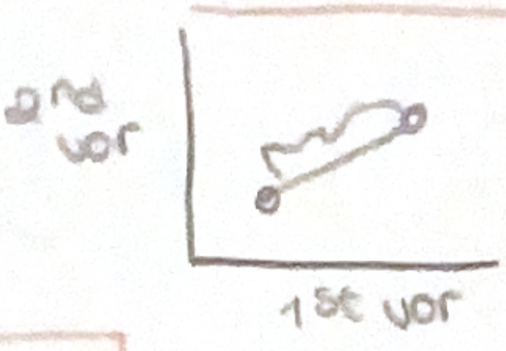
Unsupervised Learning

Clustering → Flat (Centroid, k-means)
↓ → Hierarchical (Bottom up, agglomerative)
Top down, divisive
Density (Dense regions, DBScan)

Distance Metric: Euclidean Distance



- Variables might be on different scales
- Convert to z score or standardize

$$\frac{\#\text{vars with matching value for samples}}{\#\text{vars}}$$

Matching coefficient
- Used when clustering observations are 0 or 1
- Count number of variables with matching values

Jaccard's Coefficient
Measure doesn't count matching zero entries

Centroid Algorithms

* K-means
  1. Pick a k
  2. Initialize k points (means)
  3. Categorize each point to closest mean
  4. Repeat until clusters don't change
- Can't handle outliers (use k-medoids)
- Features are needed to be scaled

* K-medoids
Uses real data points instead of means

Centroid based ones cannot handle arbitrary shaped clusters

Density-Based

* DBSCAN
- Can handle arbitrary shaped clusters
- Algorithm checks number of points near a point
  * If #points > threshold, accept the point as core point
  * Core points and border points form a cluster together

Hierarchical-Clustering

- Agglomerative
- Divisive
- Isolation Forest