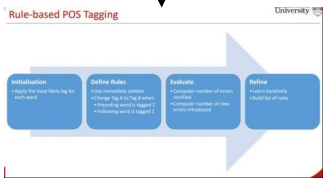
361 4



How to setup data for ML

Clustering and Classification

Split into Training and Testing Data

Supervised vs Unsupervised

Classification

Clustering

1. Training Process
2. Prediction
3. Evaluation

Also sometimes has validation data (Data tested during testing process)

Assigning categories to groups

Identifying meaningful patterns/clusters/ groups in observed data points

Has no effect on model

Aim is to reduce dissimilarity in a cluster

Maps Feature Vector to Label

K Nearest Neighbour = Classify by closest Neighbours

Hierarchical vs K-Means

K-Means = Optimisation Problem

Aim to minimise the within-cluster sum of squares (Variance)

Hierarchical =
1: Initialisation
2. Merge iteratively
3. Result

Choose K by:
- Prior knowledge of data space
- Use elbow method
- Run hierarchical clustering on subset of data

K = Amount of Categories

- Efficient
- Number of Clusters can influence "wrong" results
- Non deterministic

```
K = Choose
while (true) {
    For 1 to K: Assign point to closest centroid

    For 1 to K: Compute new centroid by avg points in each cluster

    If Centroids don't change
        Stop
}
```

- Deterministic
- Inefficient
- Multilevel Representation
- Flexible

Cluster Created by Splitting Data on Different Levels

Linkage = Distance
Single = 1 to 1
Complete = Furthest
Average = Average