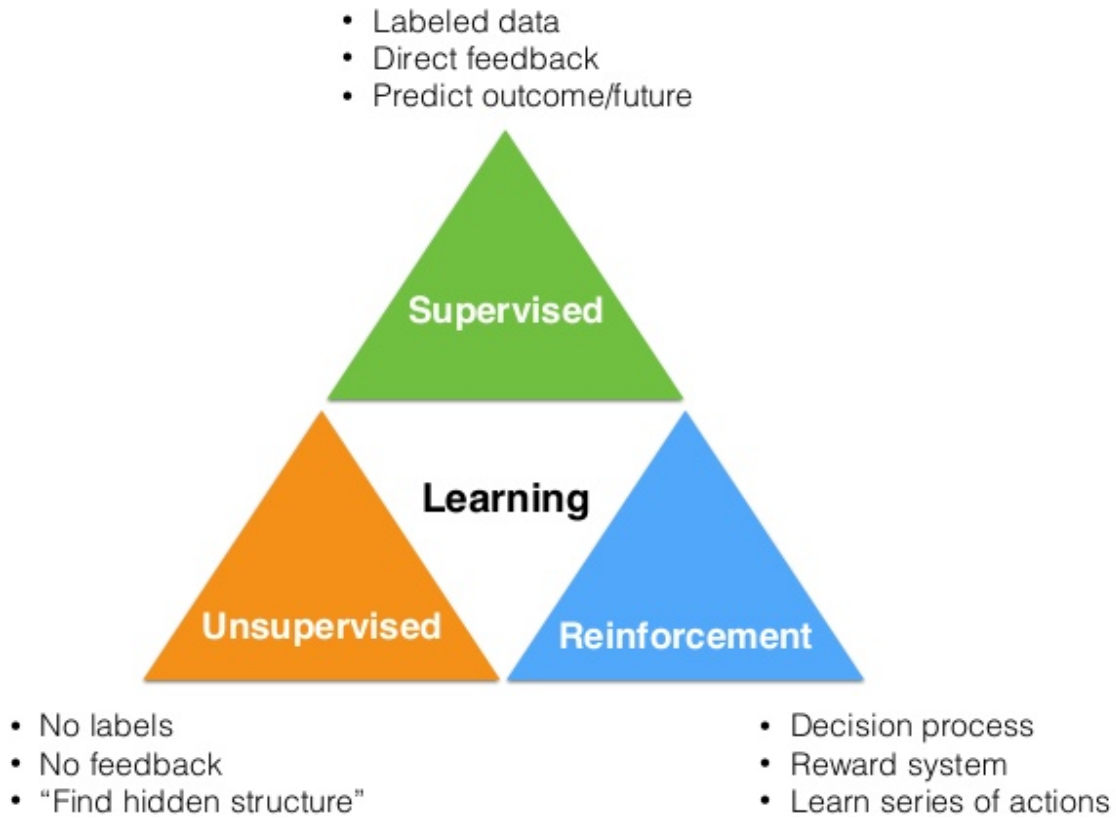
Introduction to Machine Learning

Clustering and Python ML Libraries

Types of machine learning



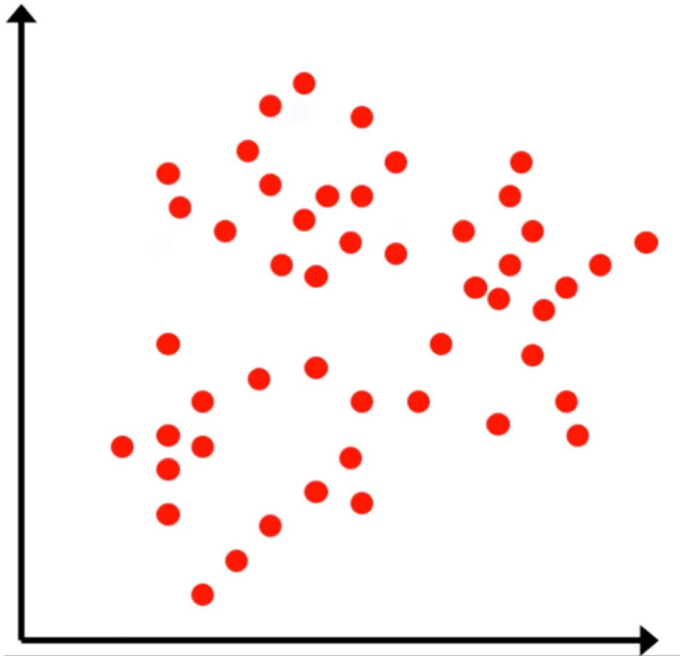
Supervised Learning

Unsupervised Learning

Reinforcement Learning

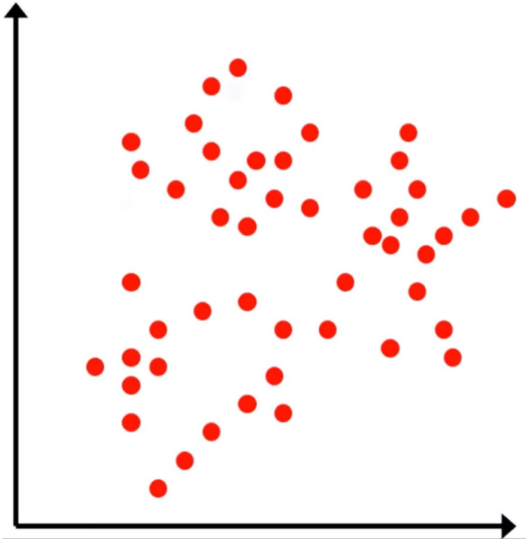
K-Mean clustering

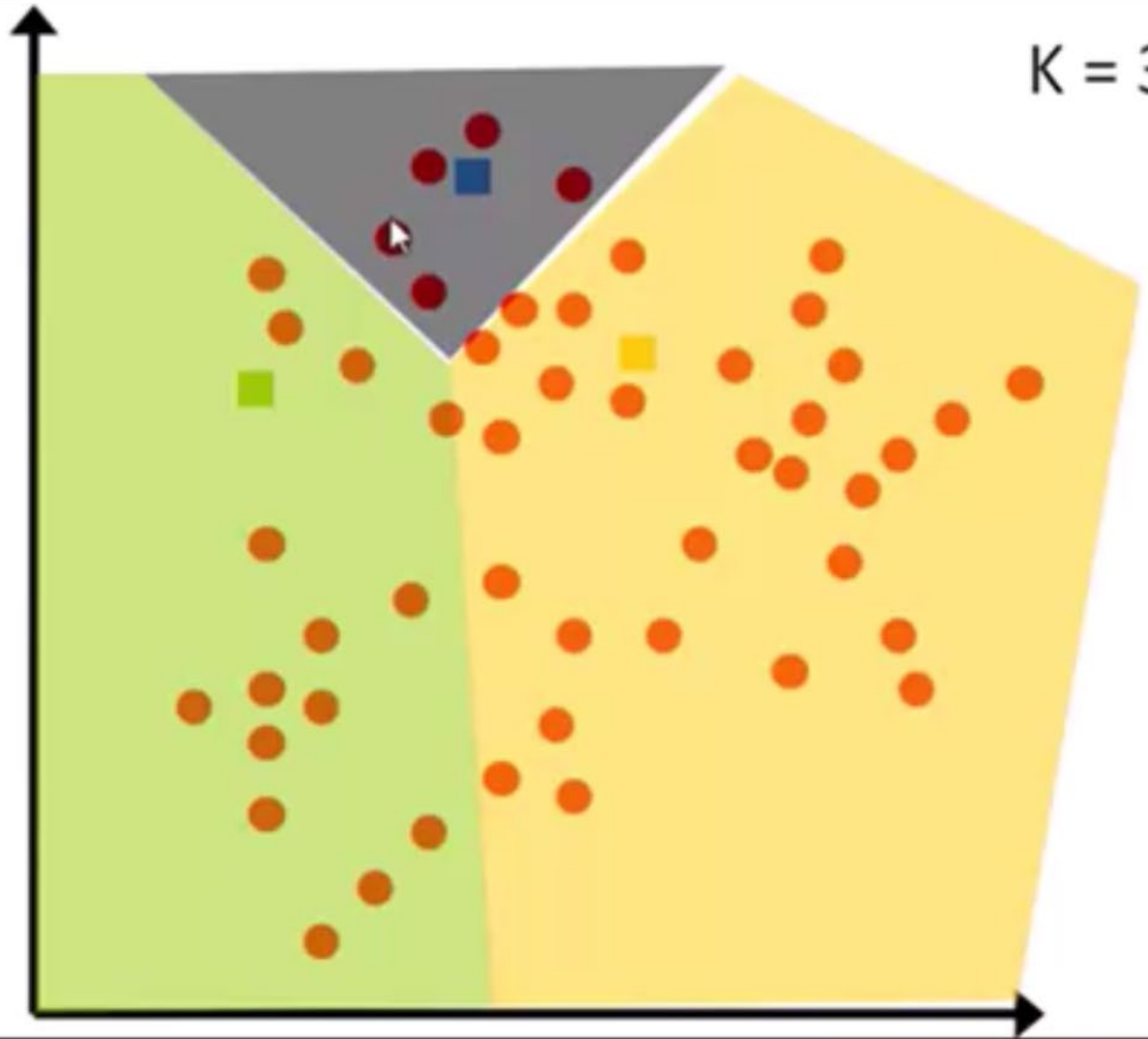
- Unsupervised learning
- Data is just examples x_1, .., x_N

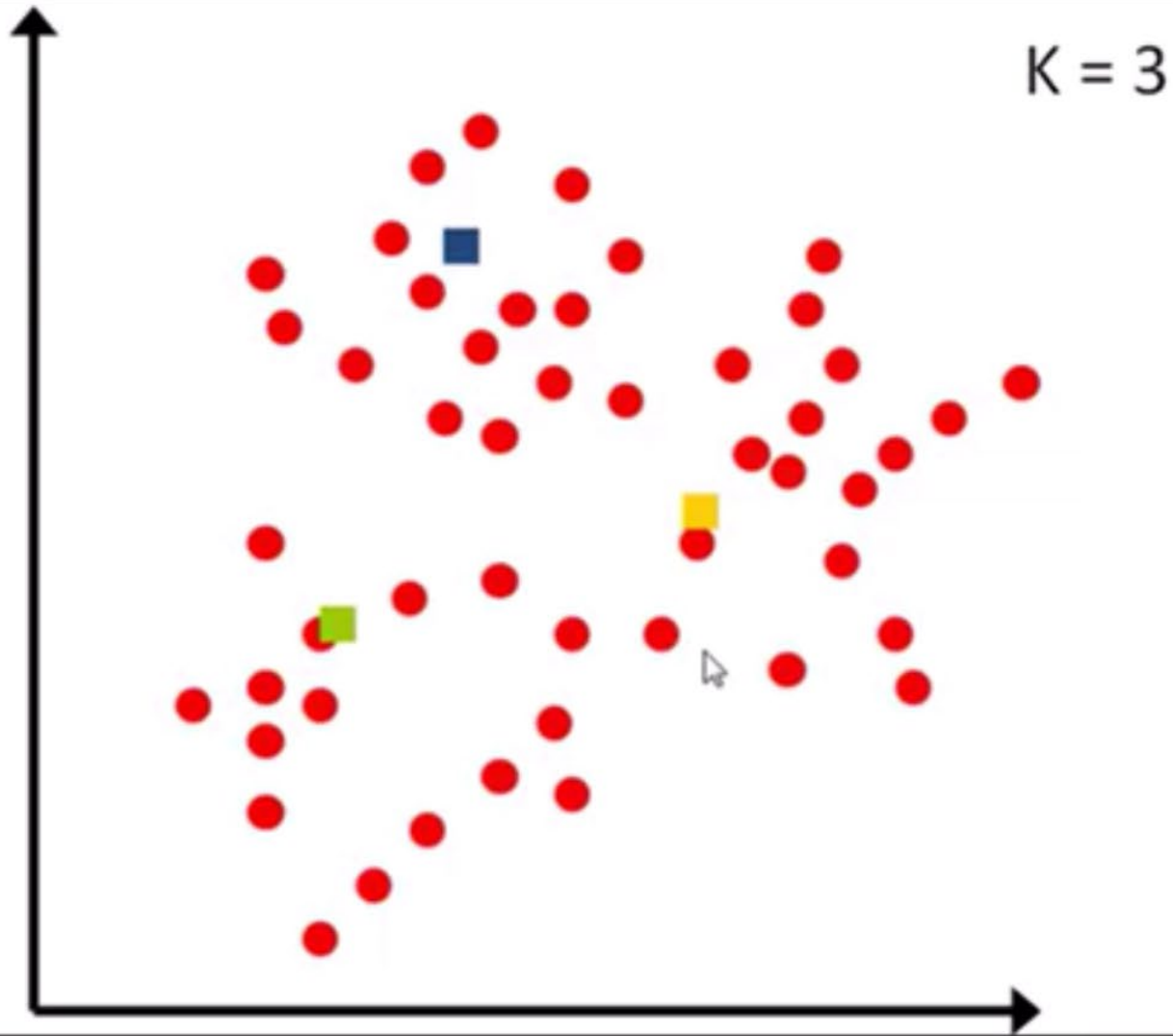


Algorithm 4 K-MEANS($\mathbf{D}$, K)

```
for k = 1 to K do
    μ_k ← some random location          // randomly initialize center for kth cluster
end for
repeat
    for n = 1 to N do
        z_n ← argmin_k ||μ_k − x_n||    // assign example n to closest center
    end for
    for k = 1 to K do
        X_k ← { x_n : z_n = k }         // points assigned to cluster k
        μ_k ← MEAN(X_k)                 // re-estimate center of cluster k
    end for
until μs stop changing
return z                                // return cluster assignments
```




K = 3

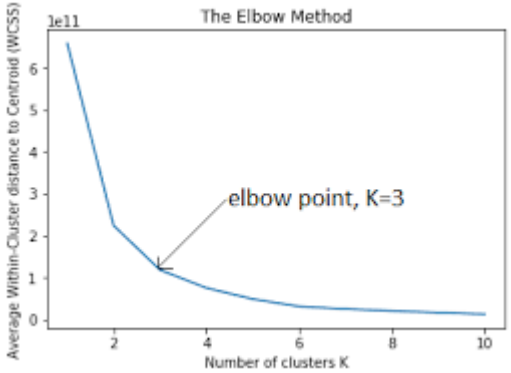

K = 3

Will it converge (stop changing)?

- Will k-means converge (stop changing)?

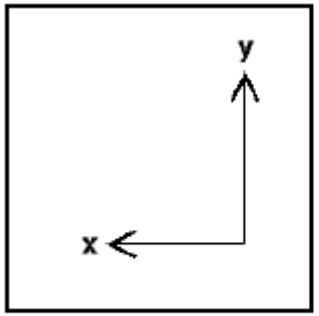
K-Means hyperparameters

- K
  - Choose using elbow method

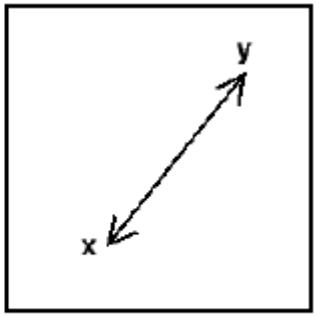


K-Means hyperparameters

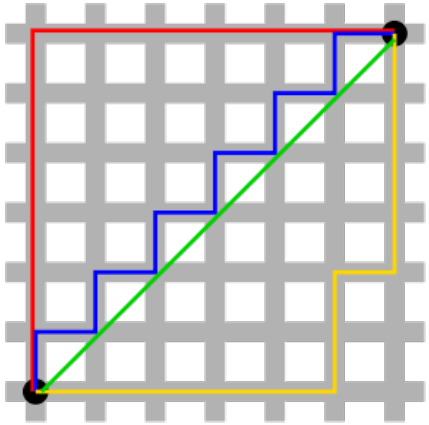
- Distance
  - Euclidean
  - Manhattan



Manhattan



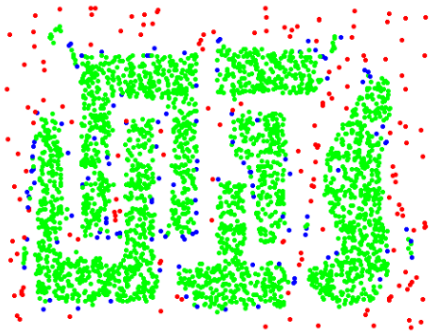
Euclidean



Alternative: Density-based clustering (DBSCAN)

- ε-neighborhood: objects within distance ε
- Three types of points
  - Core: ε-neighborhood contains ≥MinPts
  - Border: not core but inside neighborhood of a core point
  - Noise: any other point

Original Points

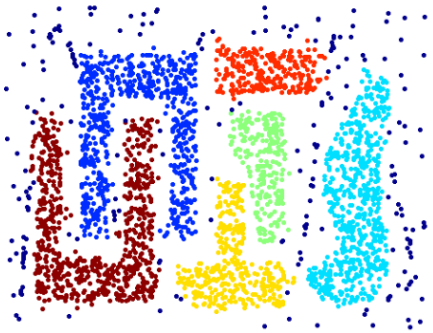


Point types: core, border and outliers

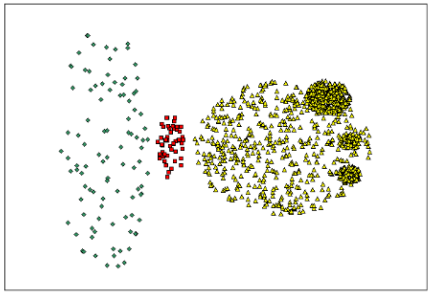
Alternative: Density-based clustering (DBSCAN)

- Cluster
  - Core and all points density-reachable from core
  - Density-reachable: accessible through chain of ε-neighborhoods

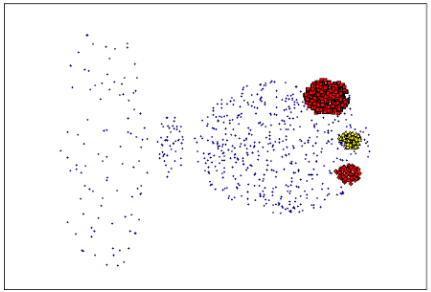
Original Points



Clusters



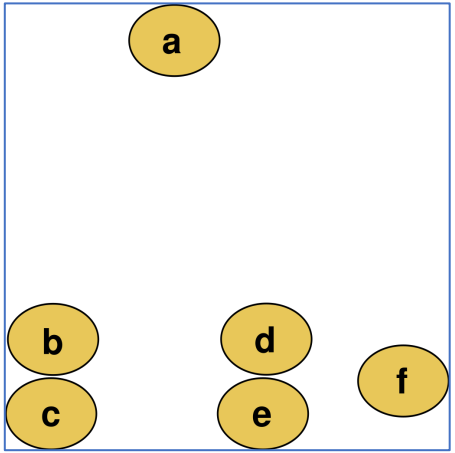
(MinPts=4, Eps=9.92).

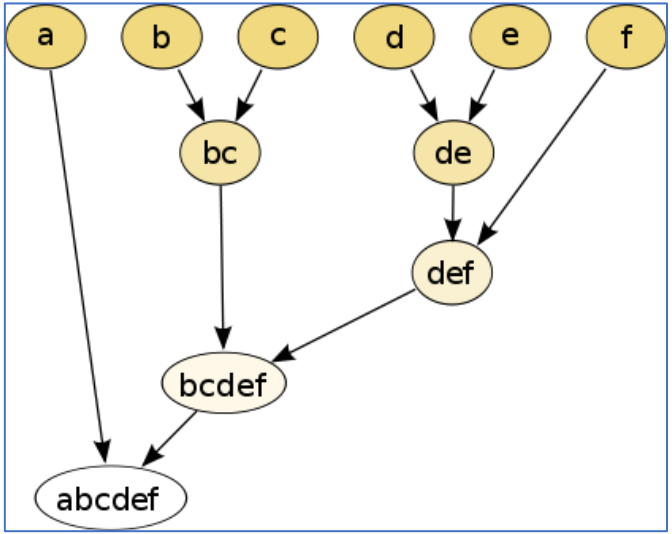


(MinPts=4, Eps=9.75)

Alternative: Hierarchical clustering

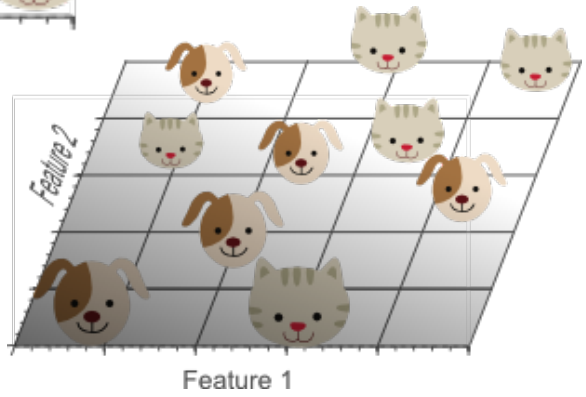
- Build hierarchy (binary tree) of clusters using greedy method
- Top down (divisive): Start with one cluster, split
- Bottom up (agglomerative): Start with #clusters=#data points, merge
- Result is dendrogram

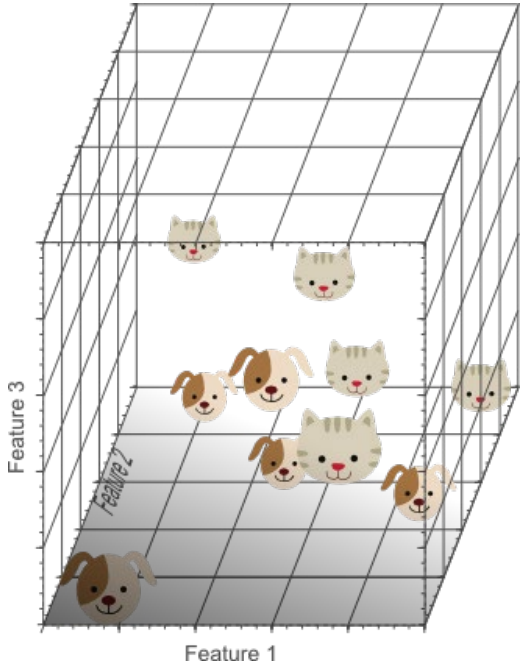




The Curse of Dimensionality

- Cost

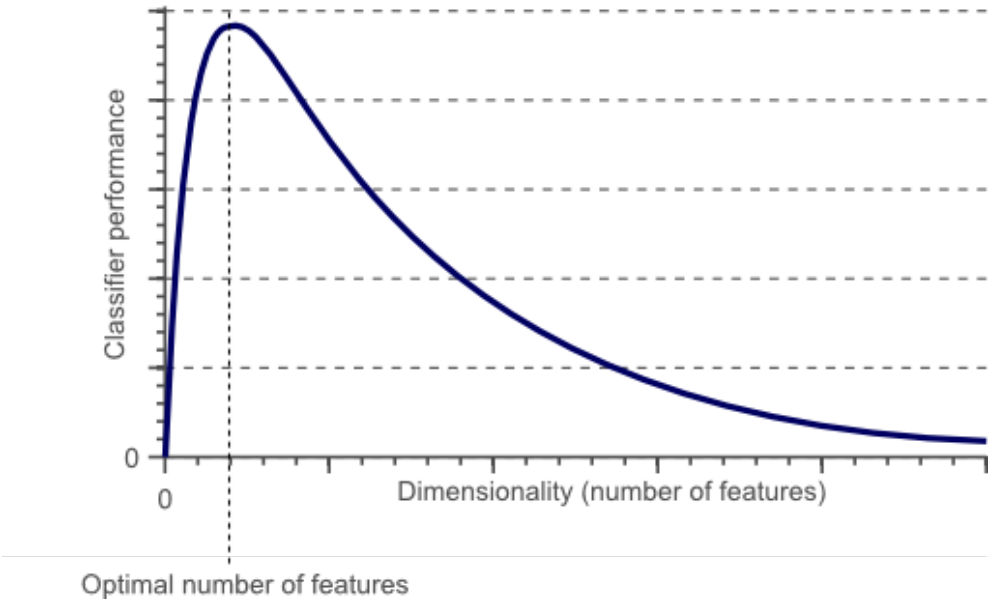




The Curse of Dimensionality

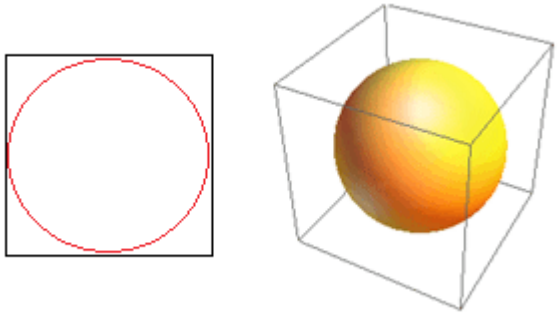
- ML performance

The Curse of Dimensionality



The Curse of Dimensionality

- Cost
- Classifier performance
- Mathematical phenomena



Sklearn (scikit-learn)