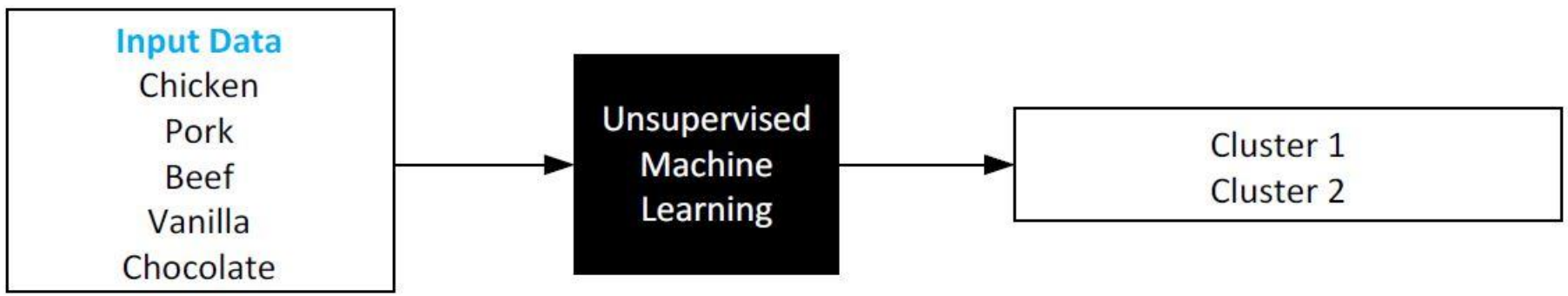
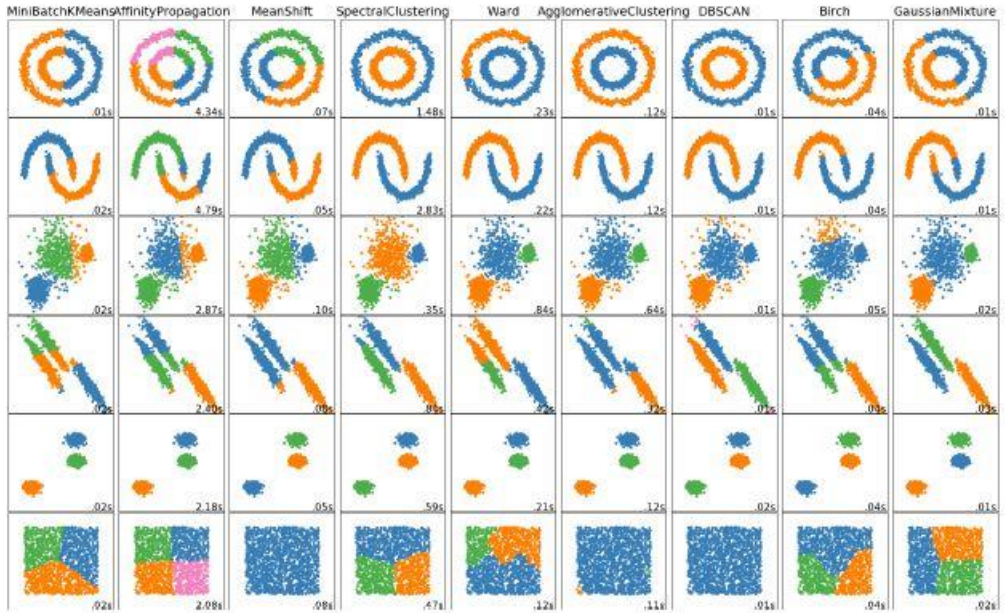
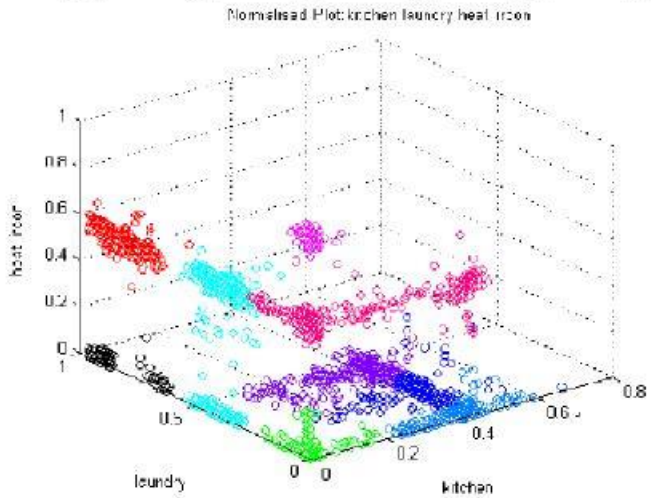
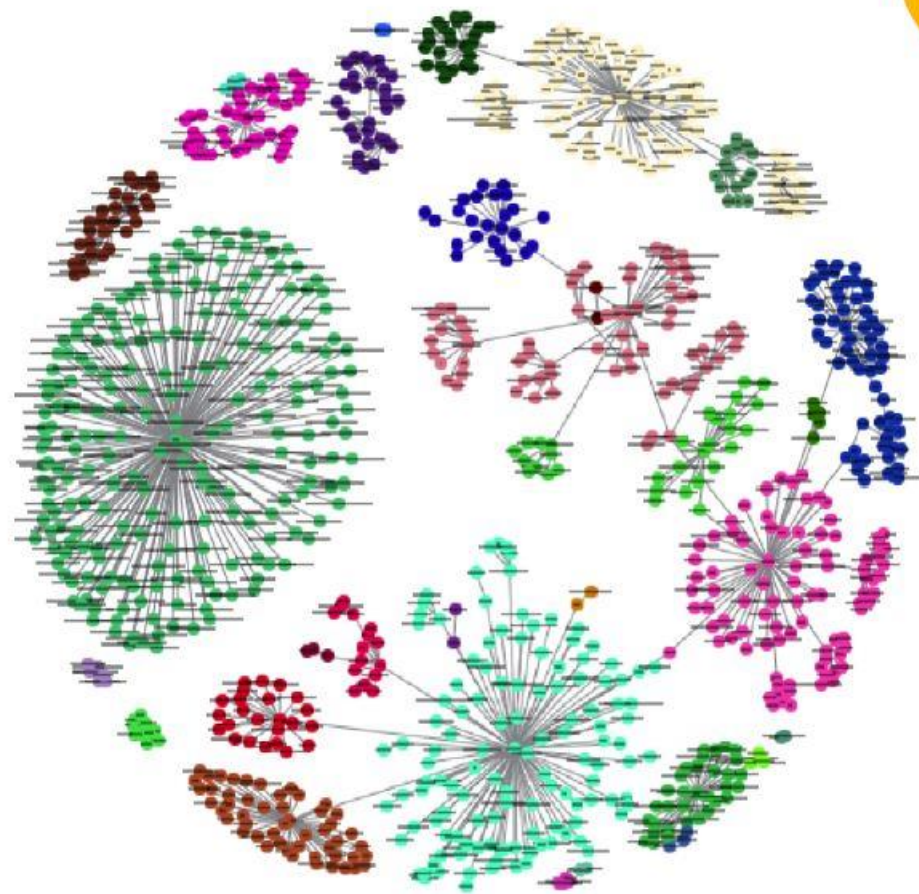
Unsupervised Learning

- Unsupervised learning is concerned with finding interesting clusters of input data. It does so without any help of data labeling.
- It does this by creating interesting transformations of the input data
- It is very important in data analytics when trying to understand data
- Examples in the Business world:
  - Customer Segmentation




MiniBatchKMeans
AffinityPropagation
MeanShift
SpectralClustering
Ward
AgglomerativeClustering
DBSCAN
Birch
GaussianMixture


Normalised Plot: kitchen laundry heat iron
heat iron
laundry
kitchen



Goal of Clustering

- The goal of clustering is if given a set of data points, we can classify each data point into a specific group or cluster.
- We can use clustering analysis to gain some valuable insights from our data by seeing what groups the data points fall into naturally by using our clustering algorithms.
- *A cluster refers to a collection of data points aggregated together because of certain similarities.*
- There are several types of clustering metods which we'll now discuss.

Types of Clustering Algorithms

There are many types of clustering algorithms, some far more widely used than others,

- K-Means Clustering
- Agglomerative Hierarchical Clustering
- Mean-Shift Clustering
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
- Expectation–Maximization (EM) Clustering using Gaussian Mixture Models (GMM)

K-Means Clustering

K-Means Clustering Algorithm

- K-Means is perhaps the most popular clustering algorithm in existence.
- It is extensively used in real world applications
- It's relatively simple, intuitive and great for beginners to conceptualize some machine learning concepts.

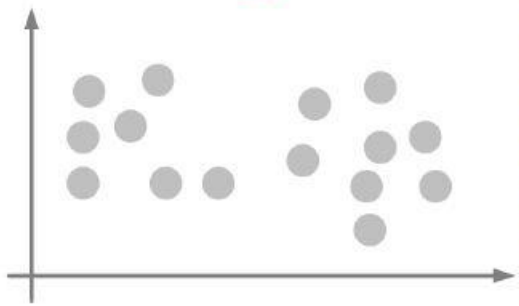
K-Means Clustering – Step 1

Choose the number of clusters you wish to identify by choosing k

- Choosing K can be done either intuitively or via the **Silhouette Method or the Elbow method**

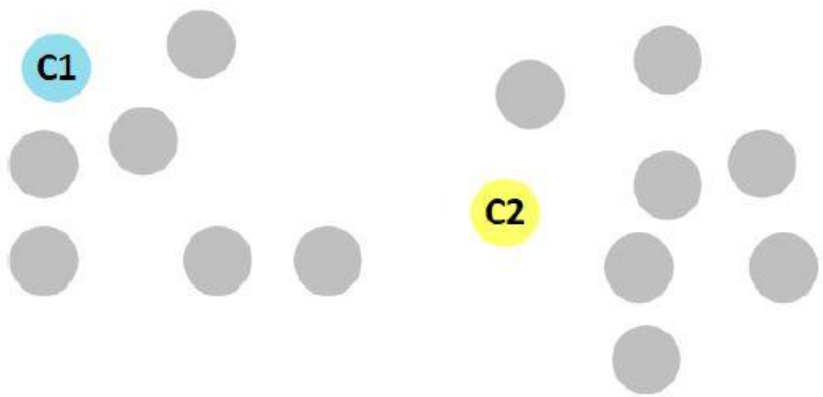
- You can choose K intuitively by:
  - **Understanding the data domain** – e.g. if you're trying to cluster amongst newspaper articles you might have quite a few types, whereas if you're clustering on customer types, it might be better to start with a smaller k (less than 5)
  - **Exploring it Visually** to see if we can spot natural clusters

Let's try k=2



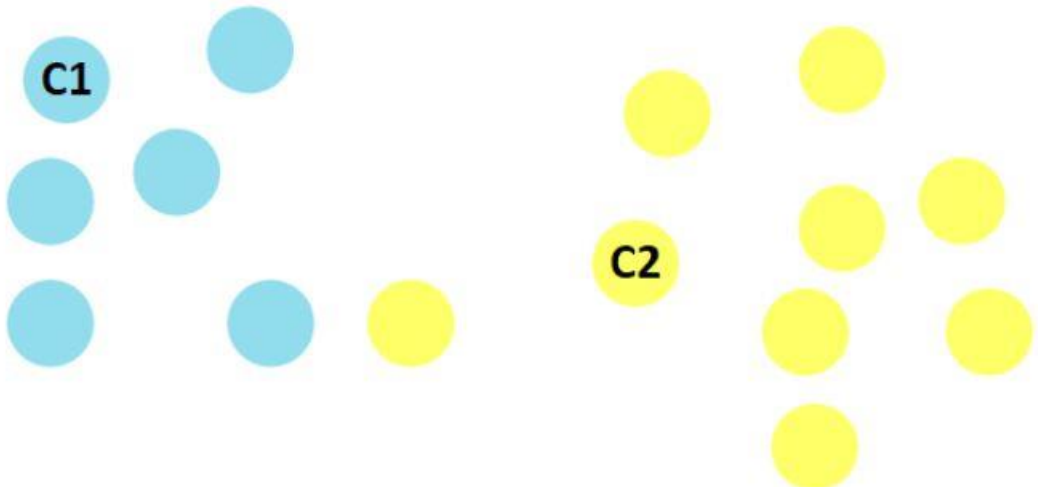
K-Means Clustering – Step 2

- Once we have **k (which was equal to 2)**, we select k random points from our dataset and use these are centroids.
- Here we have c1 (blue) and c2 (yellow) that represent the centroid of these two clusters



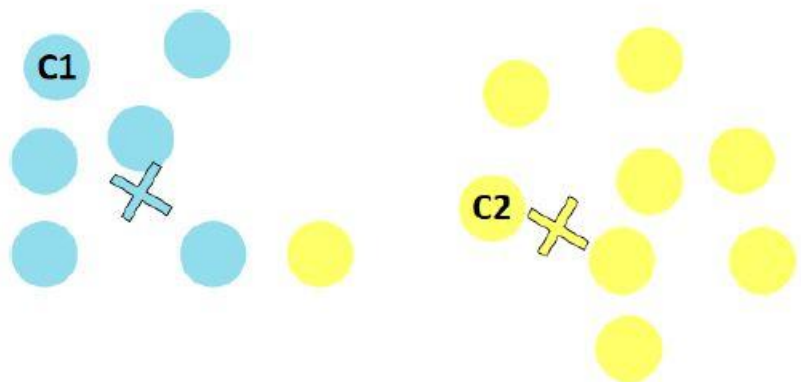
K-Means Clustering – Step 3

- Assign all the points to the closest cluster centroid
- So all points closest to C1 get assigned to the blue cluster and all points closest to C2 get assigned to the yellow cluster.



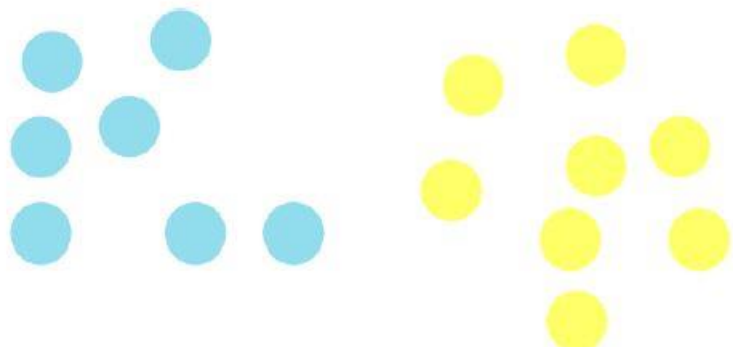
K-Means Clustering – Step 4

- We now compute the centroid of the newly formed clusters.
- The blue and yellow crosses represent the center of the newly formed clusters



K-Means Clustering – Step 5

- **Repeat Step 3 & 4** – We now use the new centroids (the yellow and blue crosses) as the cluster centers and then assign the closest points to each centroid's cluster.
- We keep repeating this step until the newly formed clusters stop changing, all points remain in the same cluster and/or the number of specified iterations is reached.



K-Means Clustering Algorithm Advantages

- Relatively simple to implement.
- Scales to large data sets.
- Guarantees convergence.
- Can warm-start the positions of centroids.
- Easily adapts to new examples.
- Generalizes to clusters of different shapes and sizes, such as elliptical clusters.

K-Means Clustering Algorithm Disadvantages

- We still need to choose K manually
- It is dependent on initial values
- Can run into problems when clustering varying sizes and density
- Sensitive to outliers
- Doesn't scale well with large number of dimensions (can be mitigated by using PCA)
- Only works for numeric values, as such categorical values will either have to be translated into some numerical meaning (e.g. high, medium, low can be mapped to 3,2,1 this can't always work for categories like types of fruit. Alternatively, we can use K-Medians, or K-Modes to alleviate this issue..