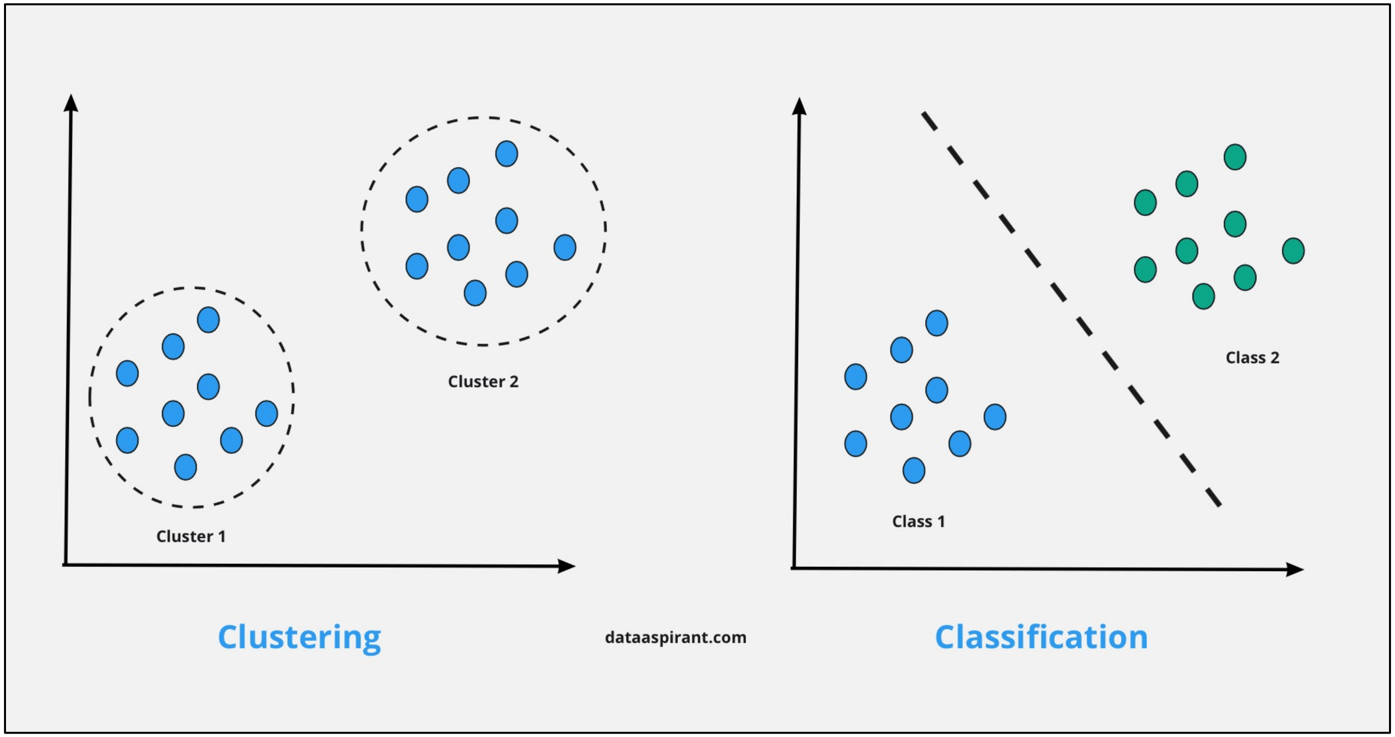
Data mining

Finding patterns in data: Dimensionality reduction

Many algorithms

- PCA
- Factor analysis
- Independent component analysis (ICA)
- Singular value decomposition (SVD – closely related to PCA)
- Non-negative matrix factorization (NMF)
- demixed PCA (dPCA)
- Linear discriminant analysis (LDA)
- Bespoke statespace analyses

Finding patterns in data: Clustering



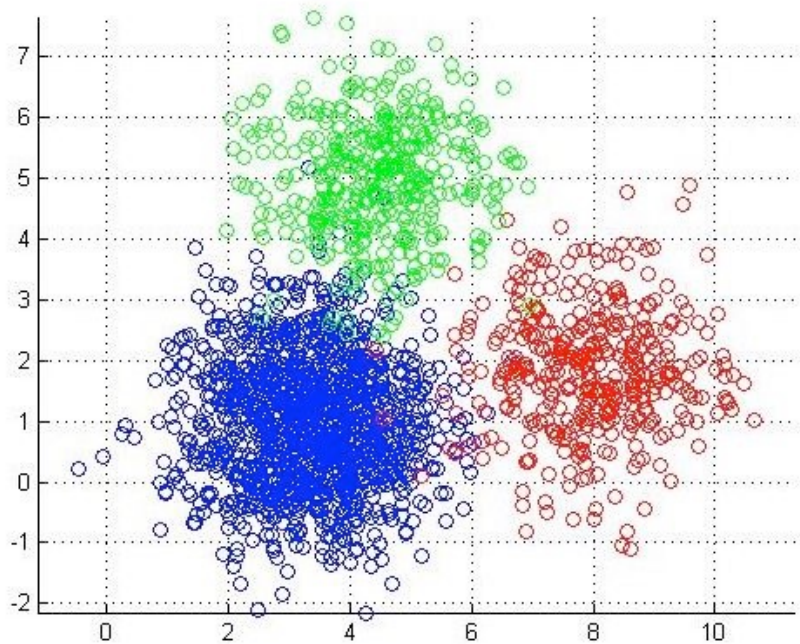
Finding patterns in data: Clustering

Even more algorithms!

K-means

K-means

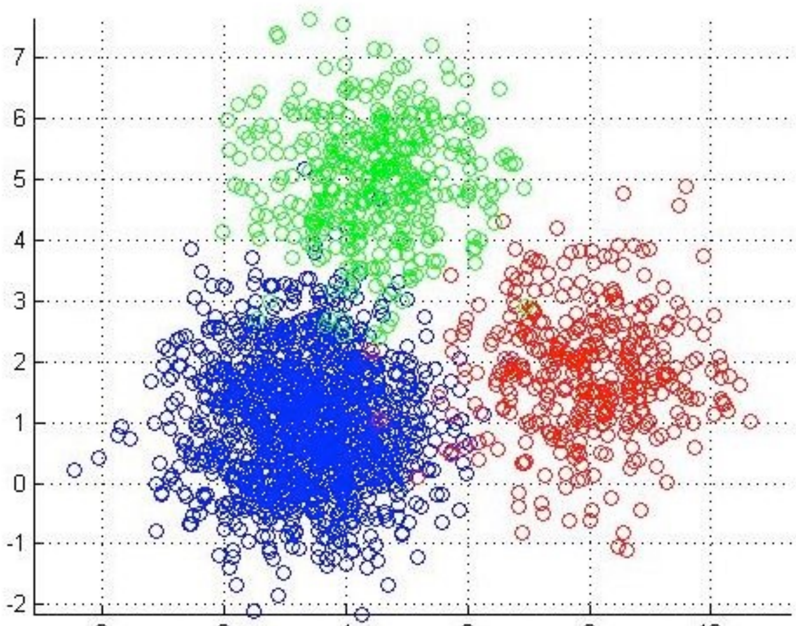
- K = the number of clusters (set by the experimenter)
- Minimizes the total sum of squared distances from each point to its respective cluster center (in n-dimensional space)

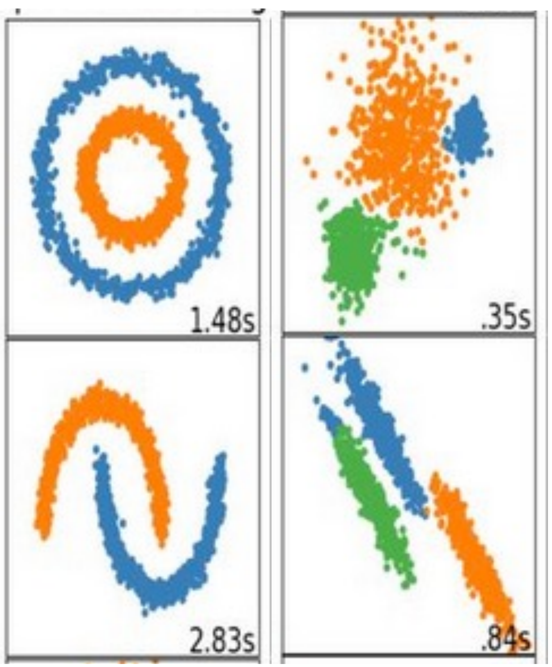


Finding patterns in data: Clustering

Even more algorithms!

K-means

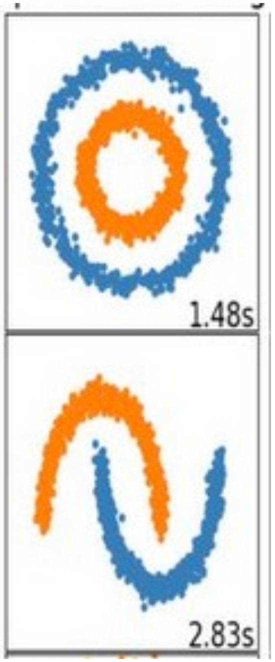




Finding patterns in data:
Clustering

Even more algorithms!

K-means
Spectral Clustering – *group based on graph distances*
Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
- *group based on distances between nearest points*



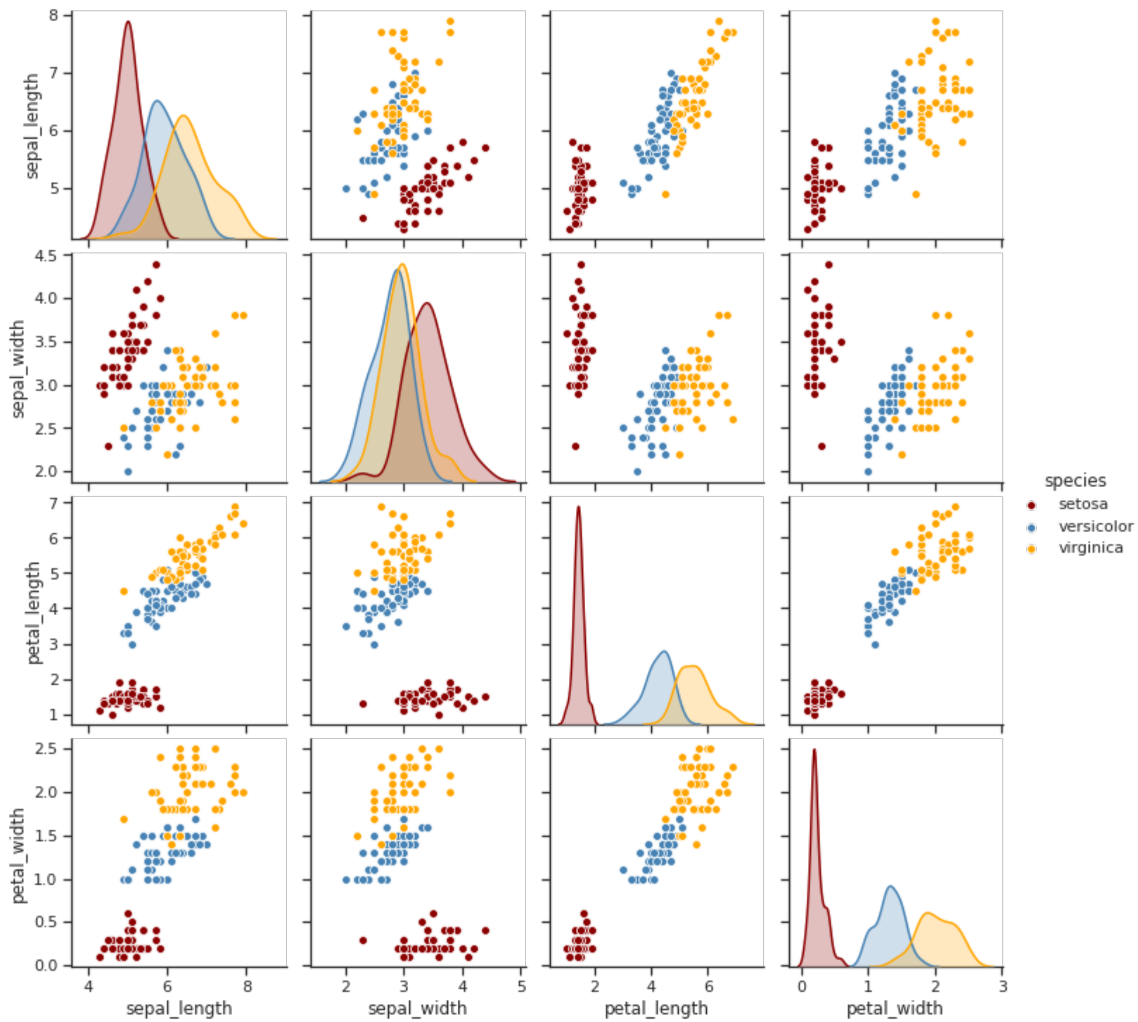
Finding patterns in data: Clustering

Even more algorithms!

K-means
Spectral Clustering
Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
Gaussian Mixture Models (GMM) using Expectation Maximization (EM)

Gaussian Mixture Models (GMM) using Expectation Maximization (EM)

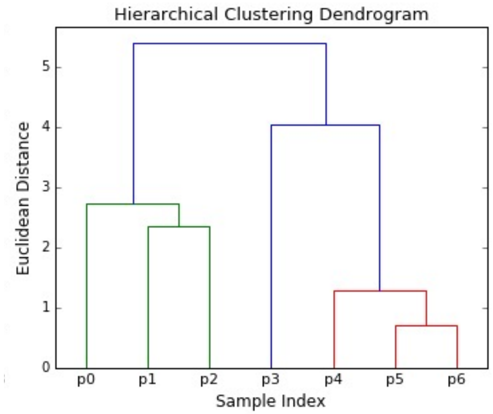
- "soft" clustering (=assigns probabilities)
- tries to assign data to different Gaussian distributions

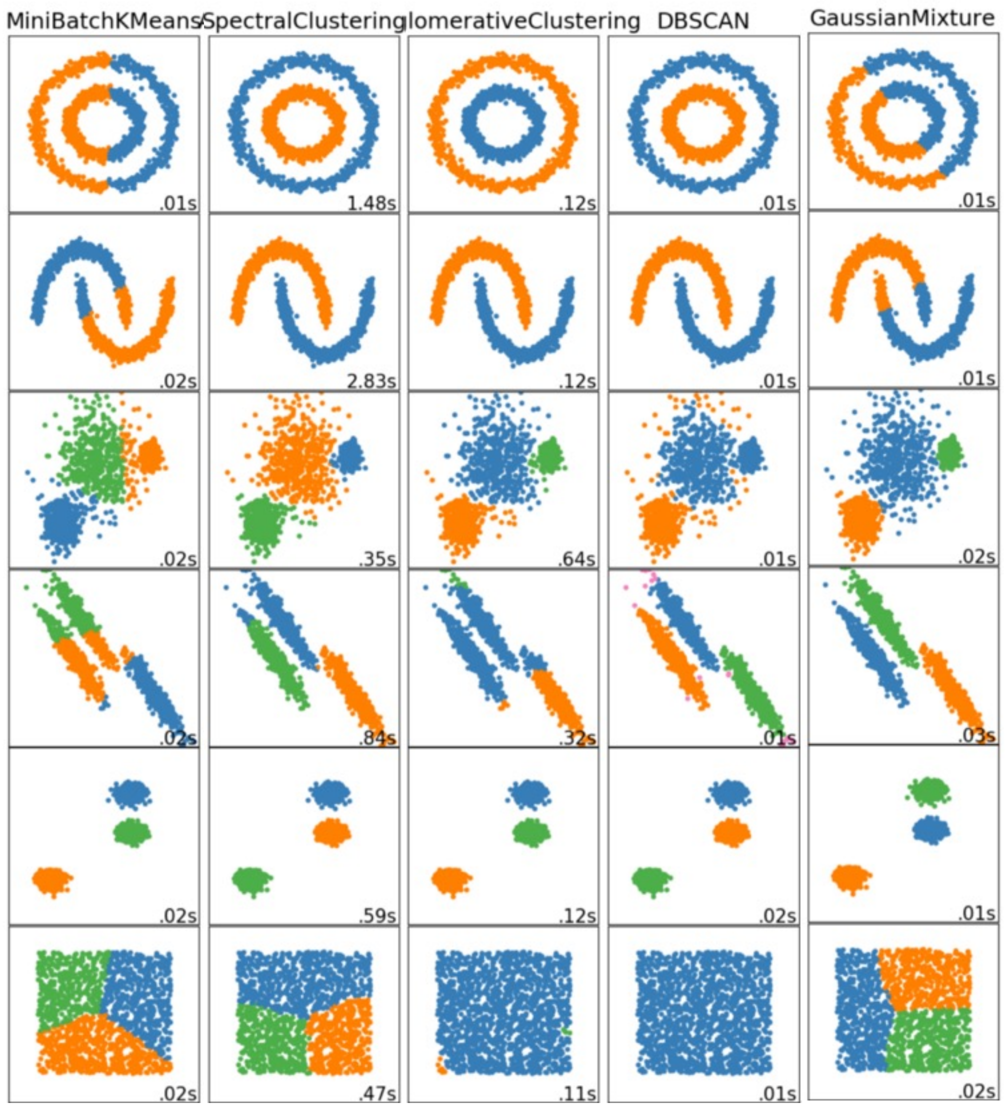


Finding patterns in data:
Clustering

Even more algorithms!

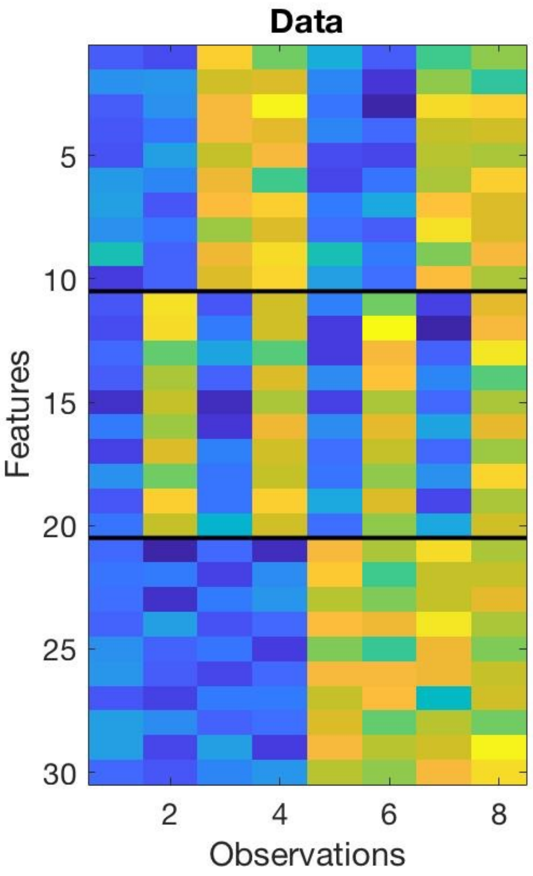
K-means
Spectral Clustering
Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
Gaussian Mixture Models (GMM) using Expectation Maximization (EM)
Hierarchical clustering algorithms
- "bottom up" (assign each point to a cluster -> successively merge)
- "top down" (start with one cluster -> split)





https://scikit-learn.org/stable/modules/clustering.html

K-means - demo



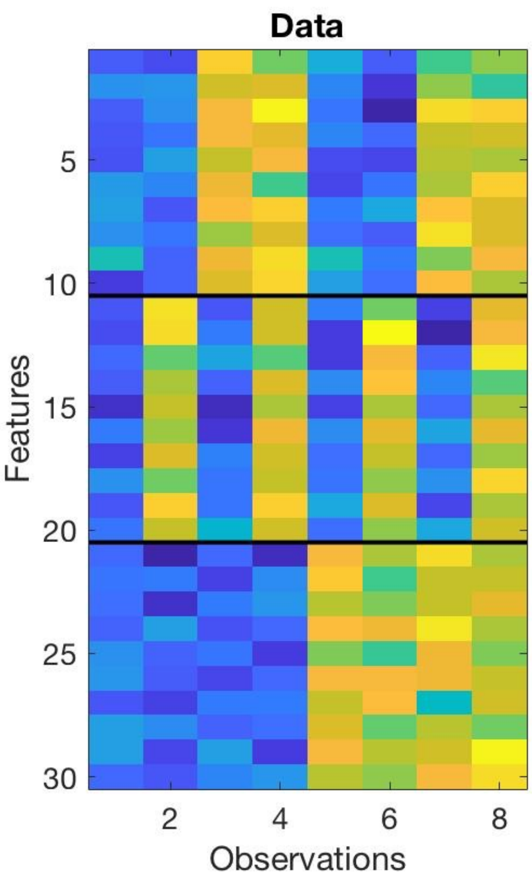

```
% Let's do it again with a more complex data set
% Additional patterns
pattern2 = [5 10 5 10 5 10 5 10];
pattern3 = [5 5 5 5 10 10 10 10];

% and create two more populations that follow different patterns
pop2 = []; pop3 = [];
for k = 1:10
    for j = 1:8
        noise = normrnd(0,var); % noise should be independent for this simulation
        pop2(k,j) = pattern2(j)+noise;
        noise = normrnd(0,var);
        pop3(k,j) = pattern3(j) + noise;
    end
end
pop = [pop1;pop2;pop3]; % Our full feature matrix is all of these subpopulations together
```

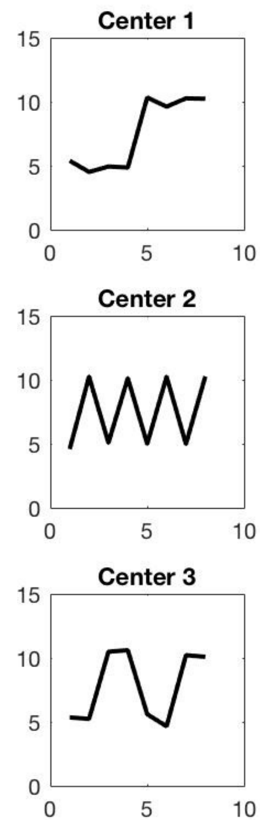
K-means - demo

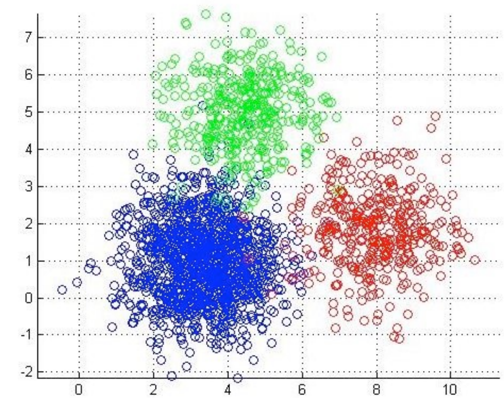
30 observations in 8-dimensions

Matlab: `[idx,centers] = kmeans(pop,3); %first run kmeans with 3 clusters`

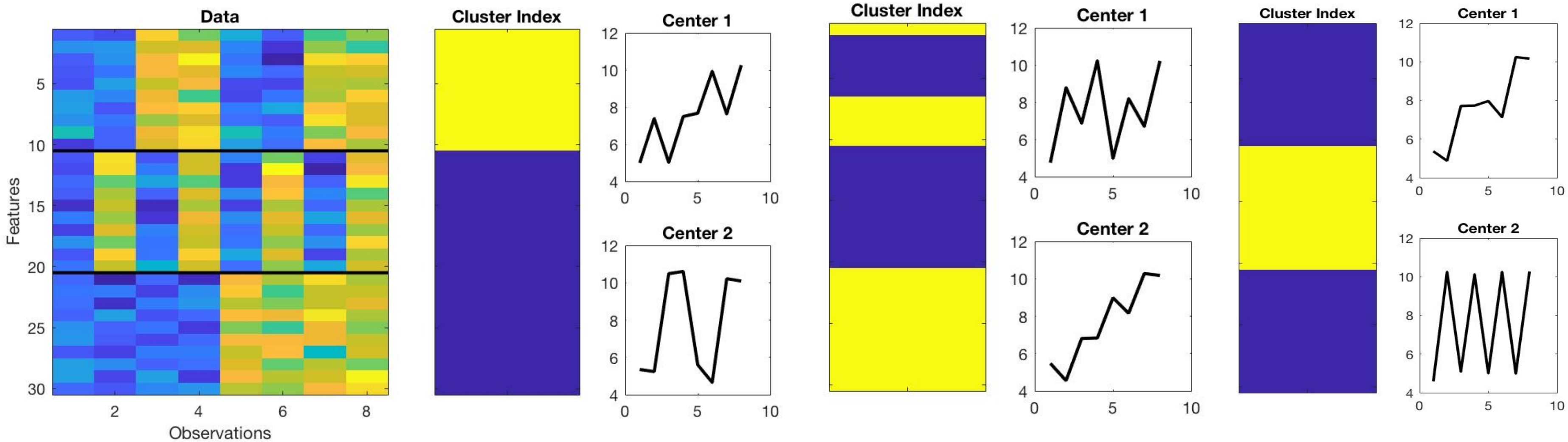
R: `pop_cluster <- pop %>% kmeans(centers = 3)`







K-means - demo

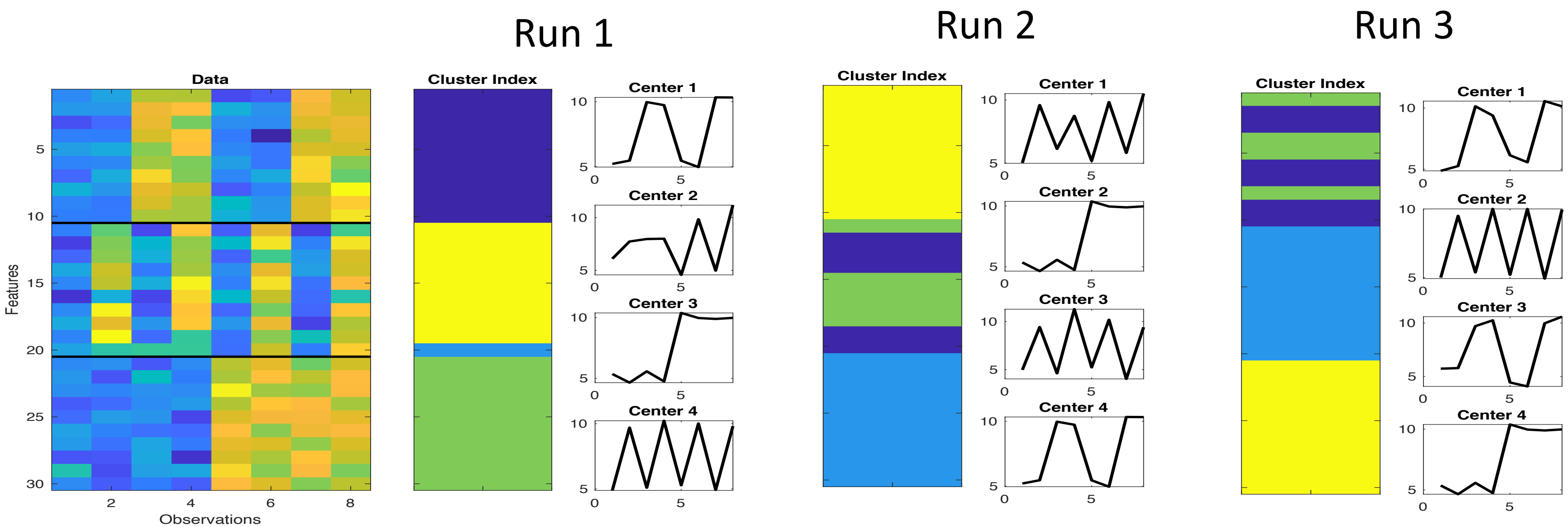


Considerations:

K-means optimization can be inconsistent

-> run with a random seed multiple times, find the global optimal solution

K-means - demo



Considerations:

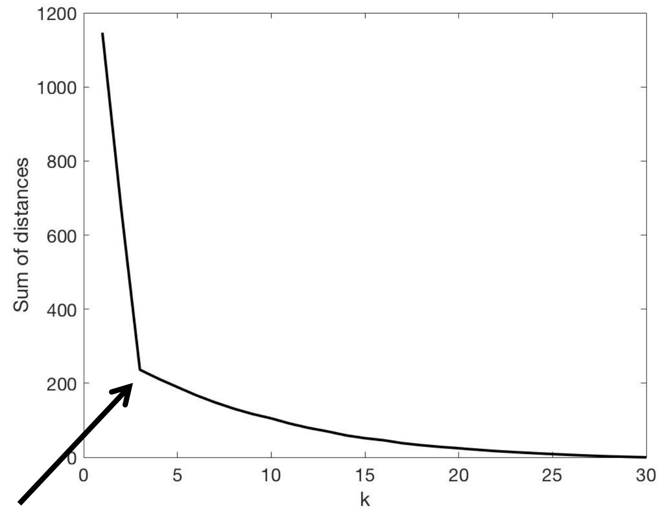
K-means optimization can be inconsistent

-> run with a random seed multiple times, find the global optimal solution

K is unknown

-> elbow method

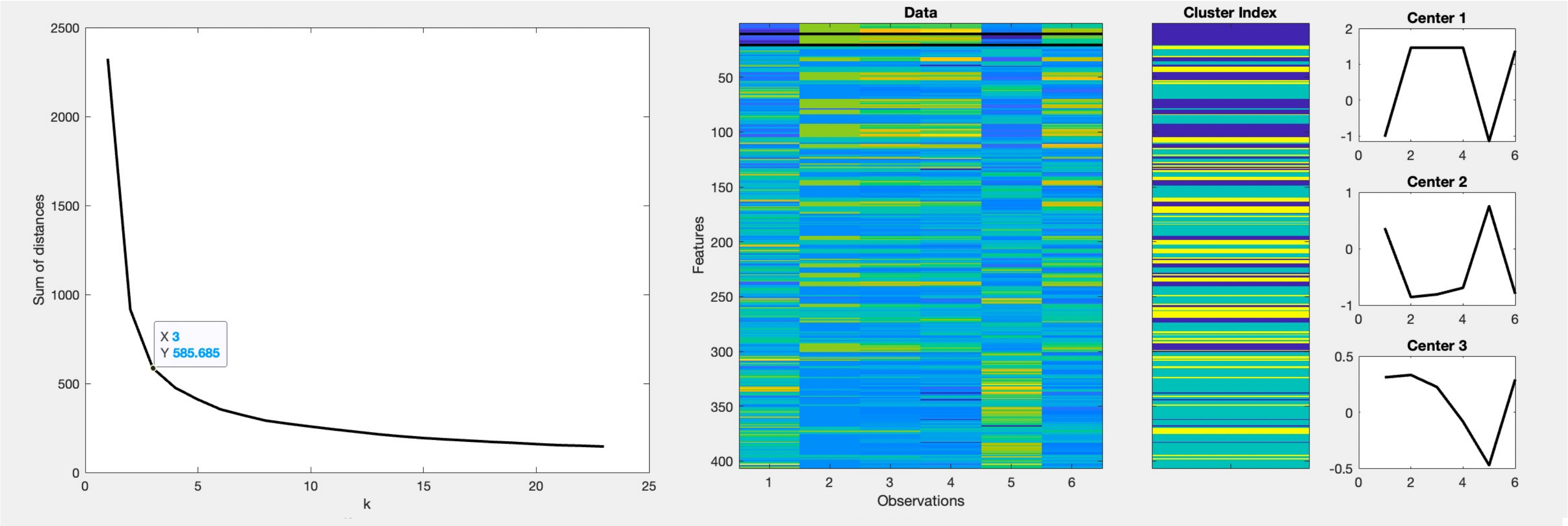
Find K with the elbow method:



Elbow
at k=3

```
numiter = 1000; %to find the optimal solution for each value of k, we'll need to rerun the algo
numk = length(pop(:,1)); %we'll test each value of k until we have 1 per feature (e.g. neuron)
distances = [];
for k = 1:numk
    d = [];
    for iter = 1:numiter
        [~,~,sumd] = kmeans(pop,k); %sumd is the sum of distances to each cluster center
        d(iter) = sum(sumd); %the total distance is the sum of sumd
    end
    distances(k) = min(d); %the minimum distance is the optimal solution for that k
end
```

K-means



<u>Considerations:</u>

K-means optimization can be inconsistent

-> run with a random seed multiple times, find the global optimal solution

K is unknown

-> elbow method

K-means struggles with irregular data (e.g. unequal numbers or variance)

A note on model comparison

The situation:

You have a lot of data, and want to find explanatory variables

The problem:

Adding more variables will always add explanatory power

But – 1. it may be a small improvement

2. it may be *overfitting*

What to do:

Formal model comparisons estimate model fit with penalties for increasing number of parameters:

Akaike Information Criterion (AIC) & *Bayesian Information Criterion (BIC)*

Also remember to hold out data when data mining!

Homework #9

(second part)

HW9: Data mining

You have recorded pupil responses in a subject viewing different images.
The data are saved in *data.txt*, which includes 500 trials where each trial is 1200ms long

1. Plot the mean pupil response over all trials
2. Do PCA across trials (Hint: each PC should be1200 elements long, and there should be 500 of them). How much variance does the first PC account for? How many components account for >=90% of the variance?
3. Plot the first principal component.
4. Run k-means 10 times with k=2. For each run, plot the cluster centers you obtain.