Clustering

Social Data Mining

Afra Alishahi
March 20, 2017

Supervised Learning

- **Classification**
  - Movie/book/restaurant reviews: good vs. bad
  - Emails: spam vs. not spam
- **Regression**
  - Predict height based on weight and gender
  - Predict income based on education, specialization and country

➡ We look for a pre-specified structure in data

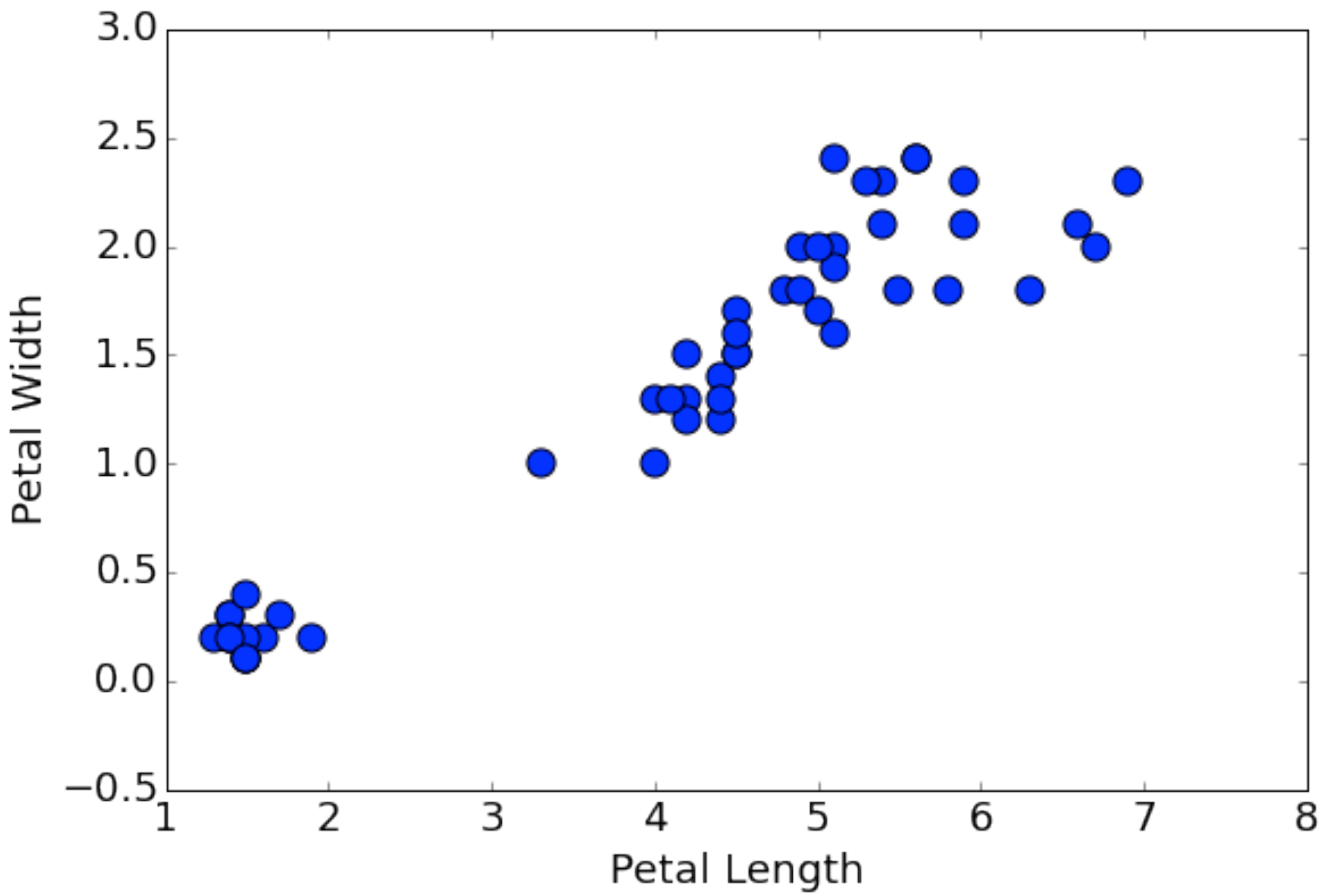
- Training data: feature sets annotated with labels or numbers

Exploratory Data Analysis

- Sometimes the structure of data is not known in advance
  - Emails: work vs. family vs. friends vs. advertisement vs. ...?
  - Shapes: square vs. circle vs. triangle vs. ...?
  - Types of questions asked in a forum
- We have a number of observation points, but no pre-defined set of labels attached to them.

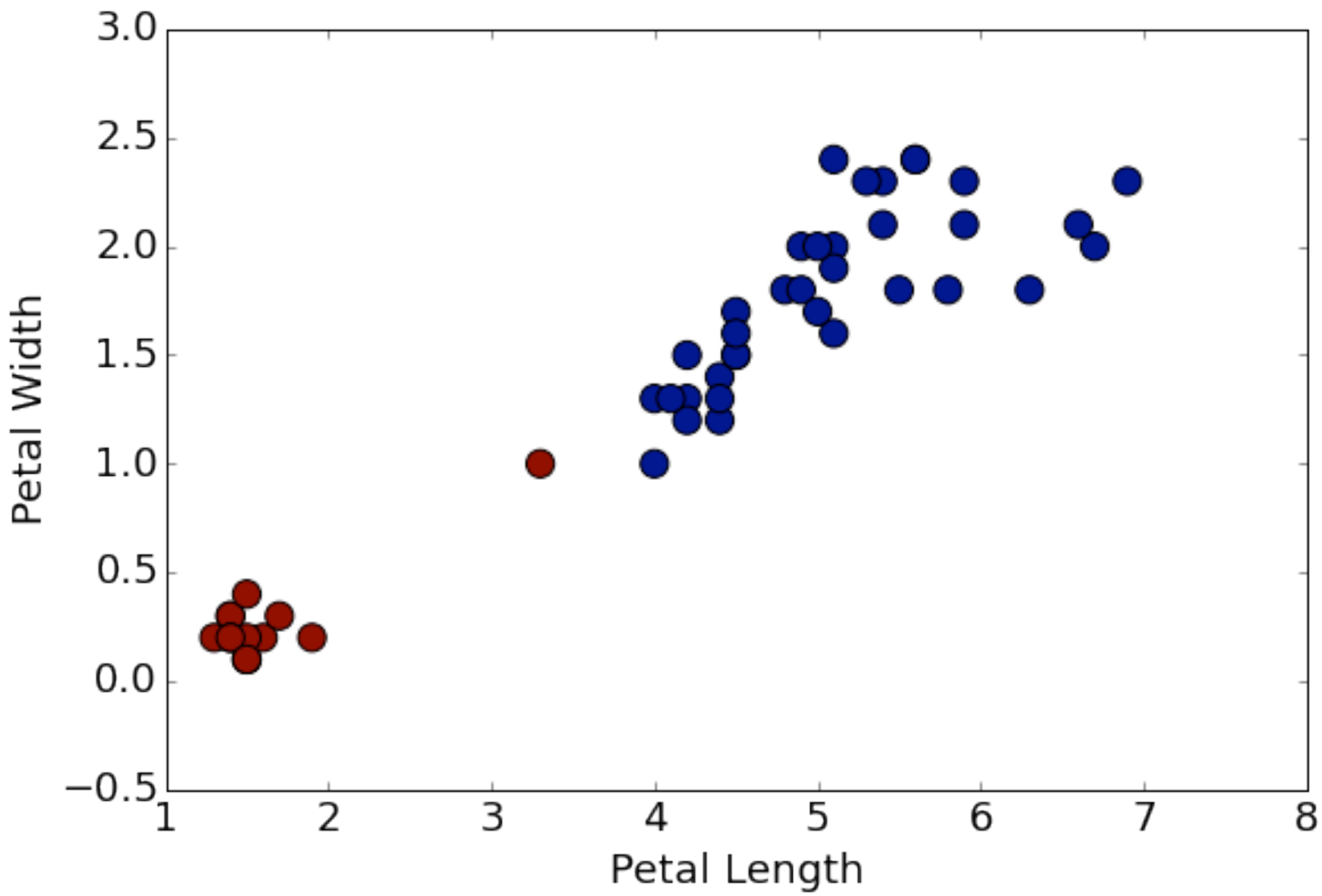
Clustering

- The Iris dataset:



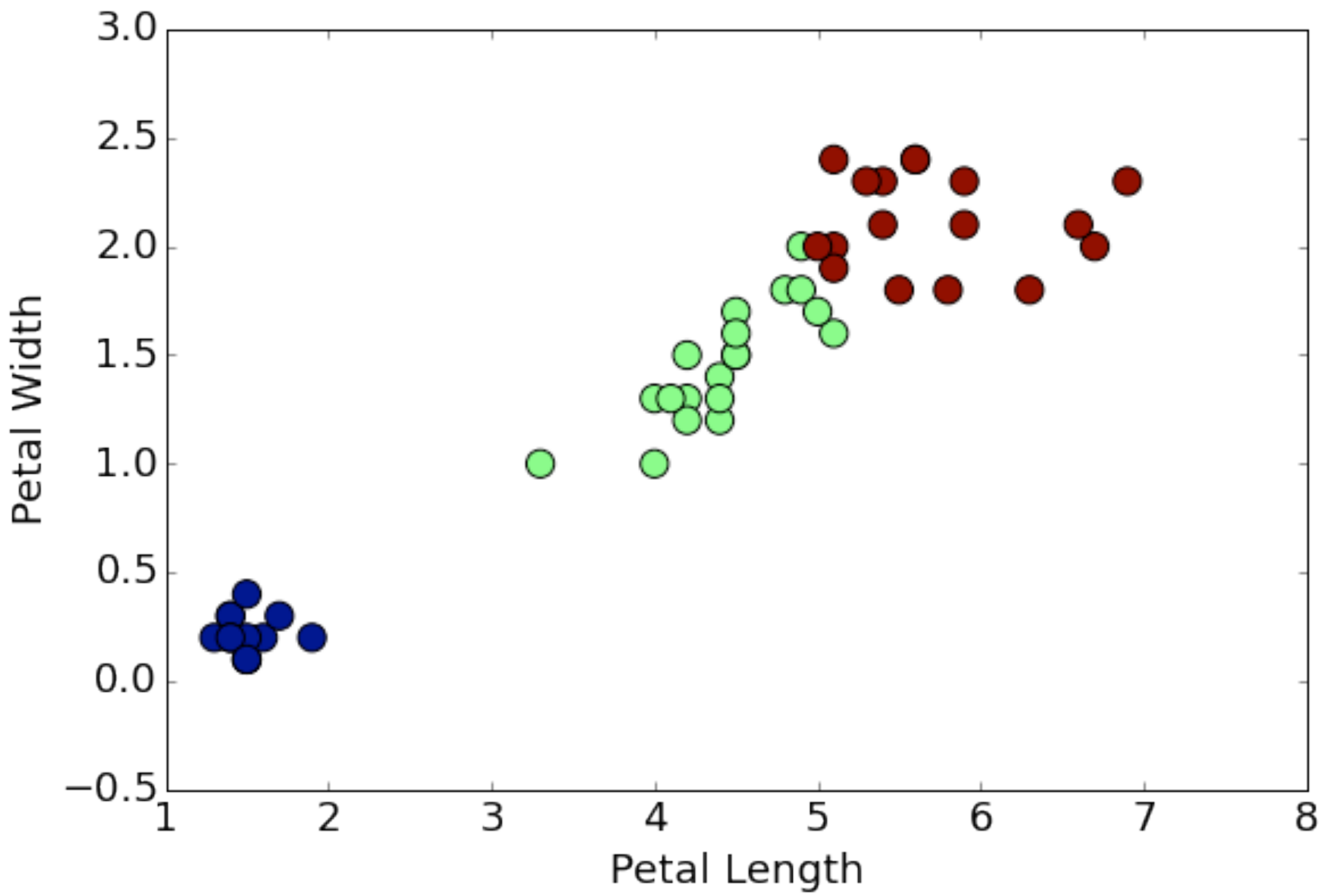
Clustering

- **Two** clusters:



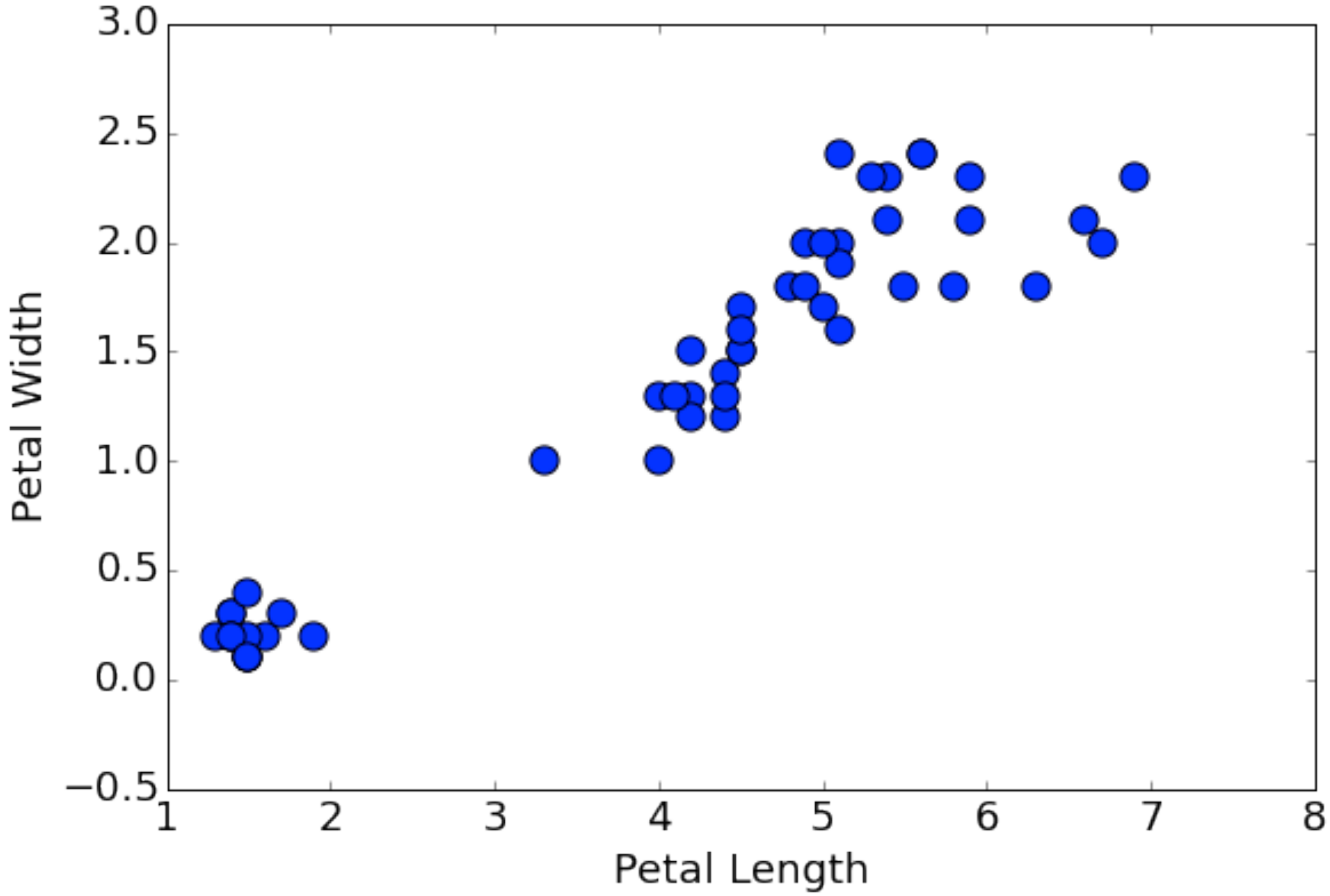
Clustering

- **Three** clusters:



Clustering

- How do we group the observation points together?

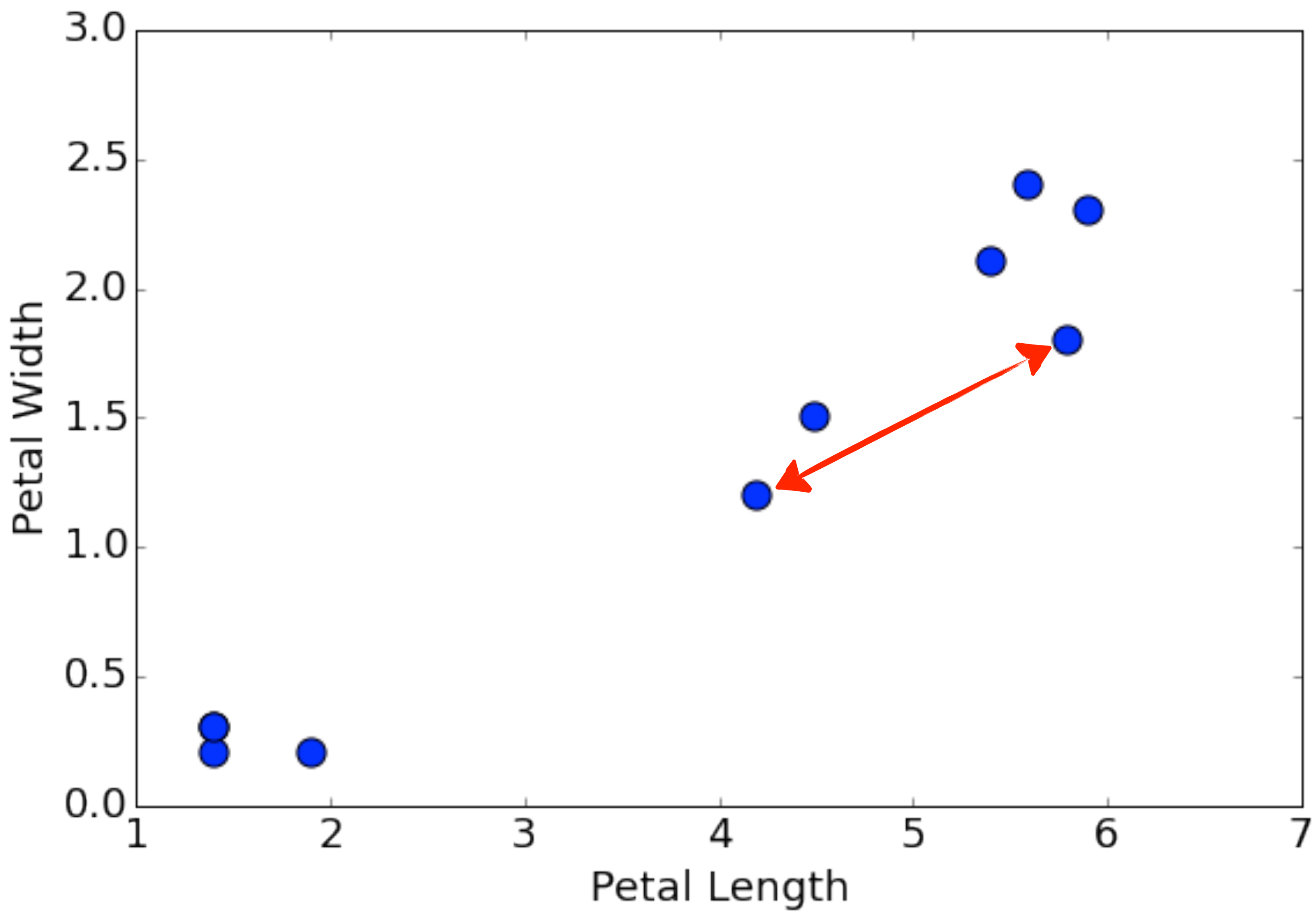


What Makes a Cluster "Good"?

- Clusters should be coherent
  - Members of the same cluster must be as close/similar to each other as possible
  - Clusters must be as distant/dissimilar from each other as possible

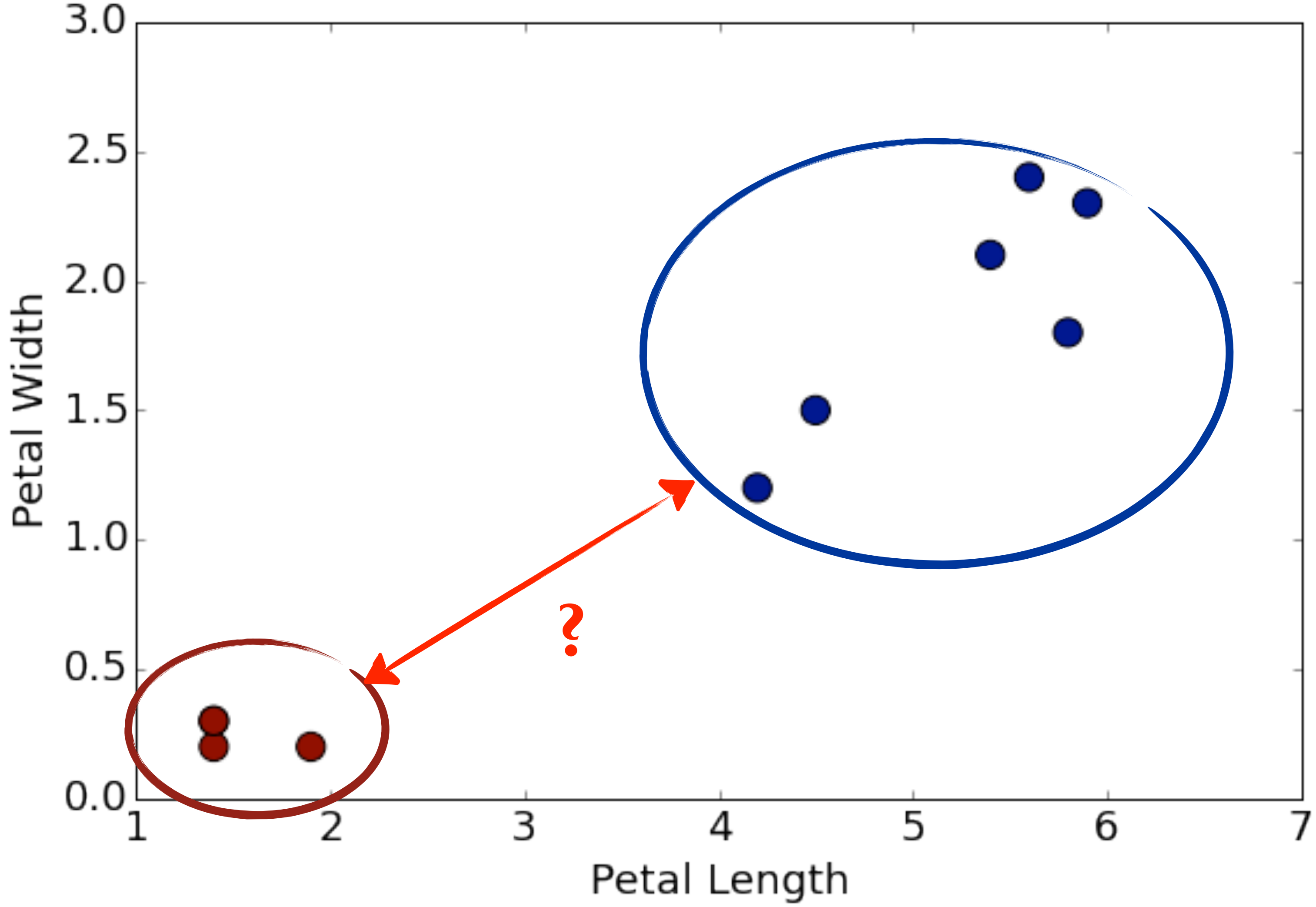
- Needed machinery:
  - Distance between two data points
  - Distance between two clusters

Distance between Data Points

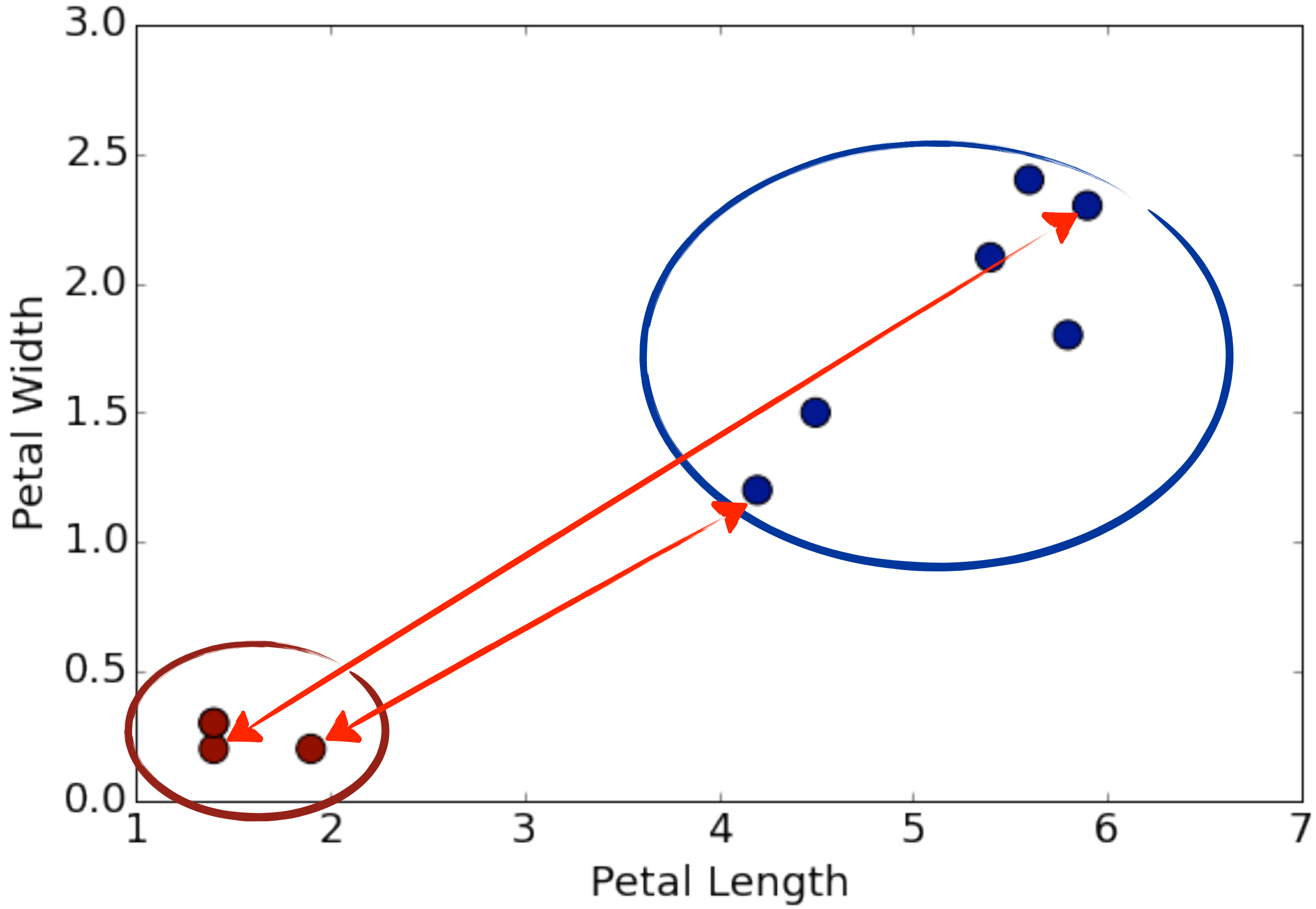


For numerical features, Euclidean distance is a good measurement.

Distance between Clusters

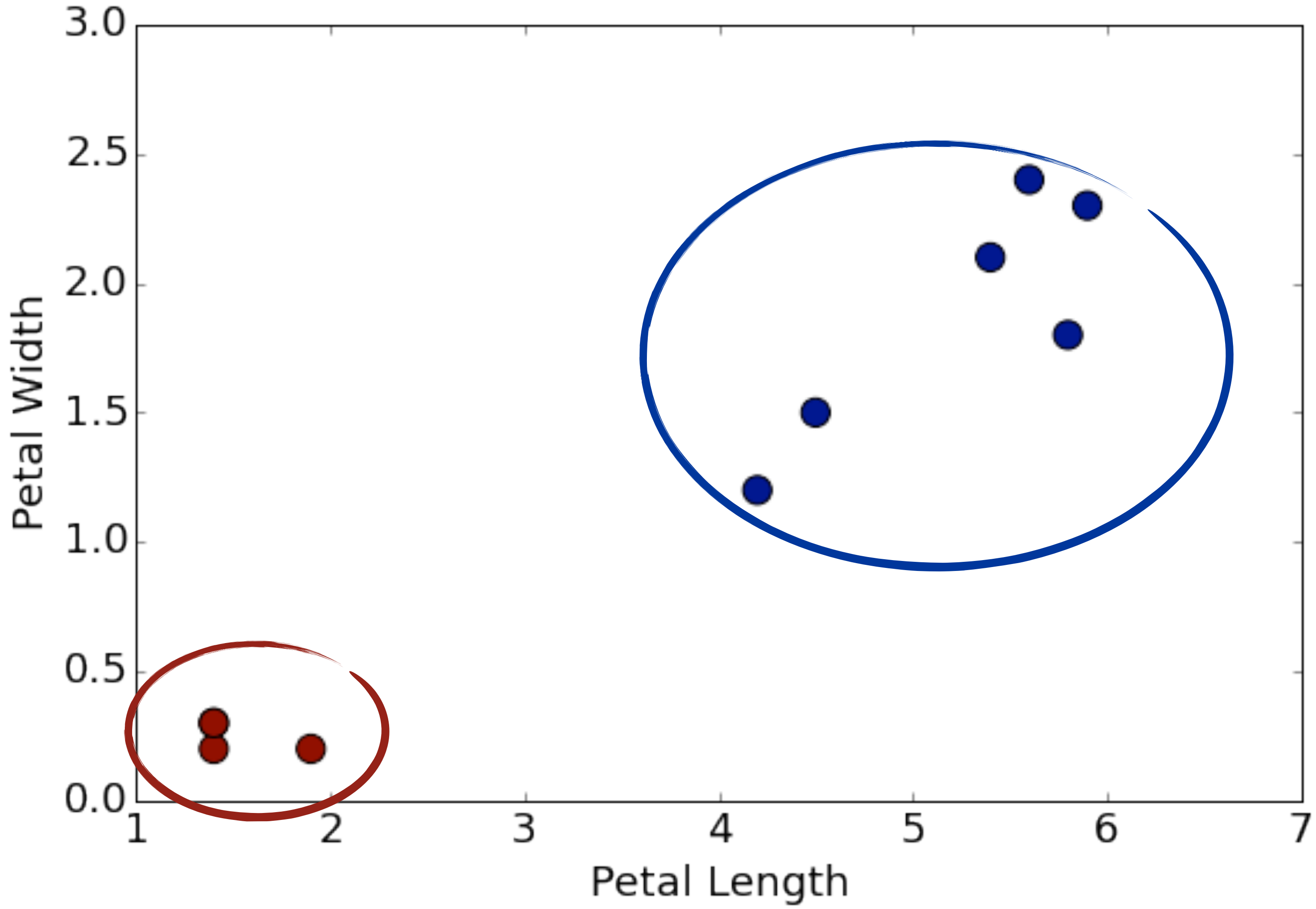


Distance between Clusters

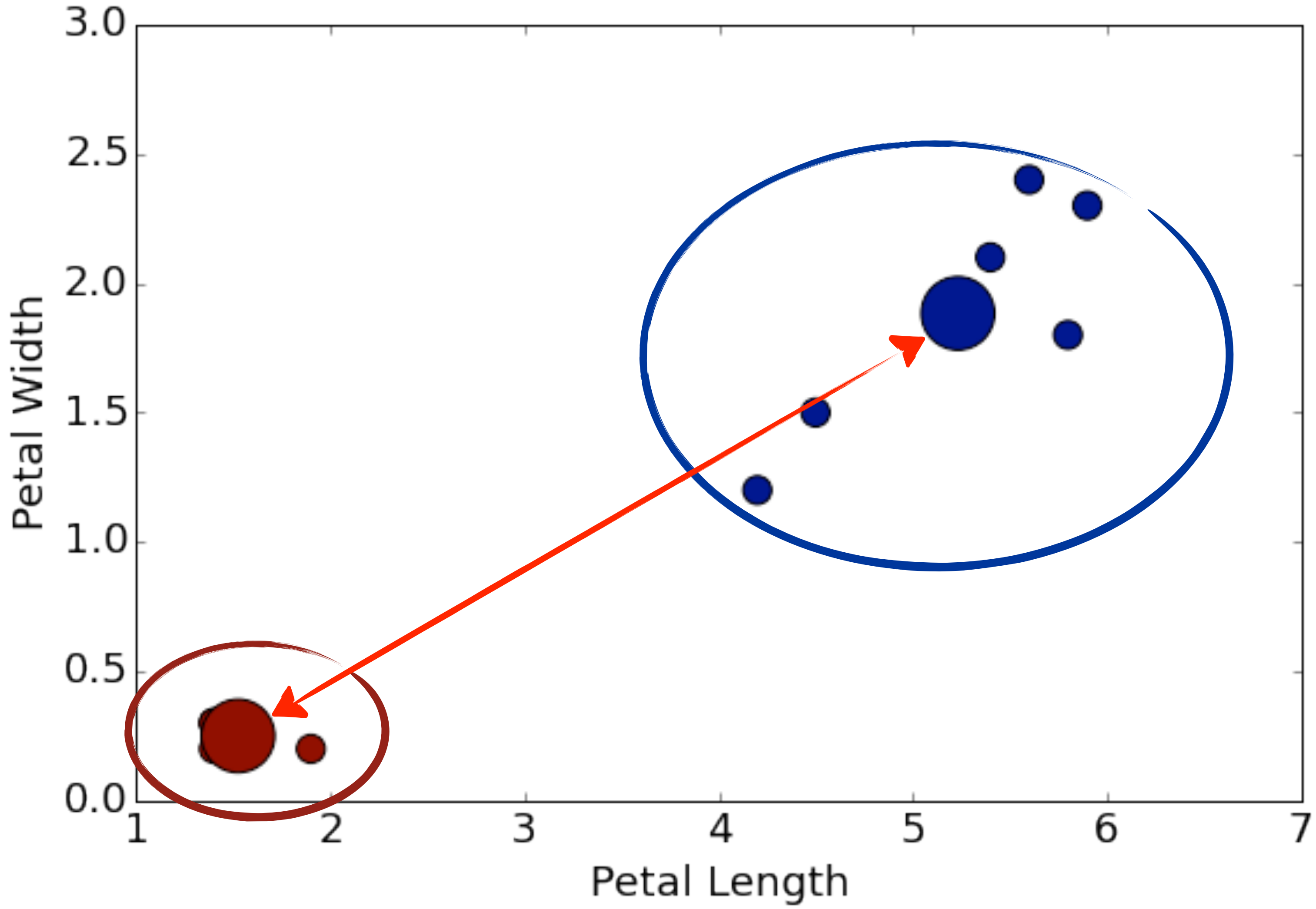


- **Single link:** distance btw the two most similar members
- **Complete link:** distance btw the two least similar members

Cluster Centroids

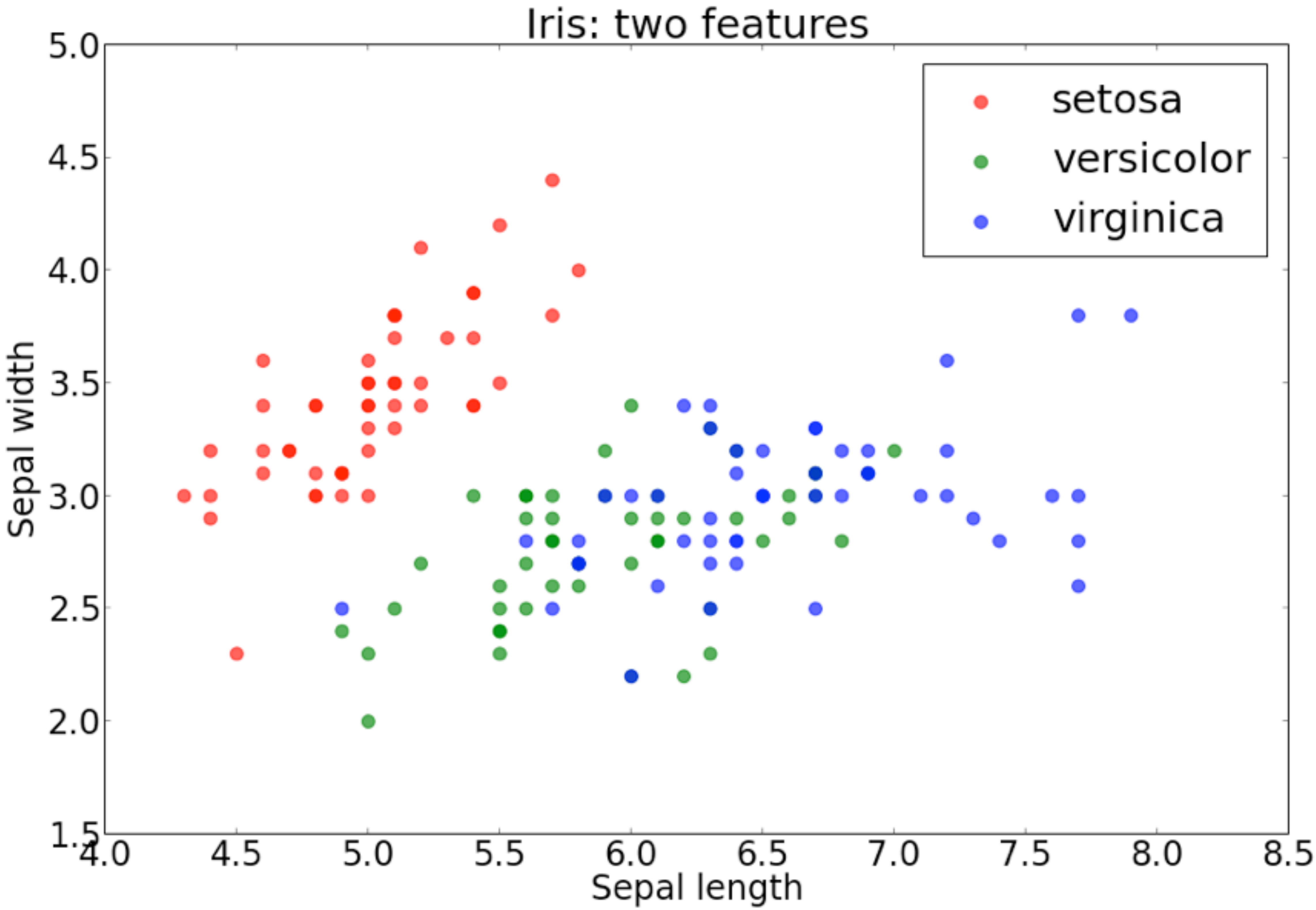


Cluster Centroids

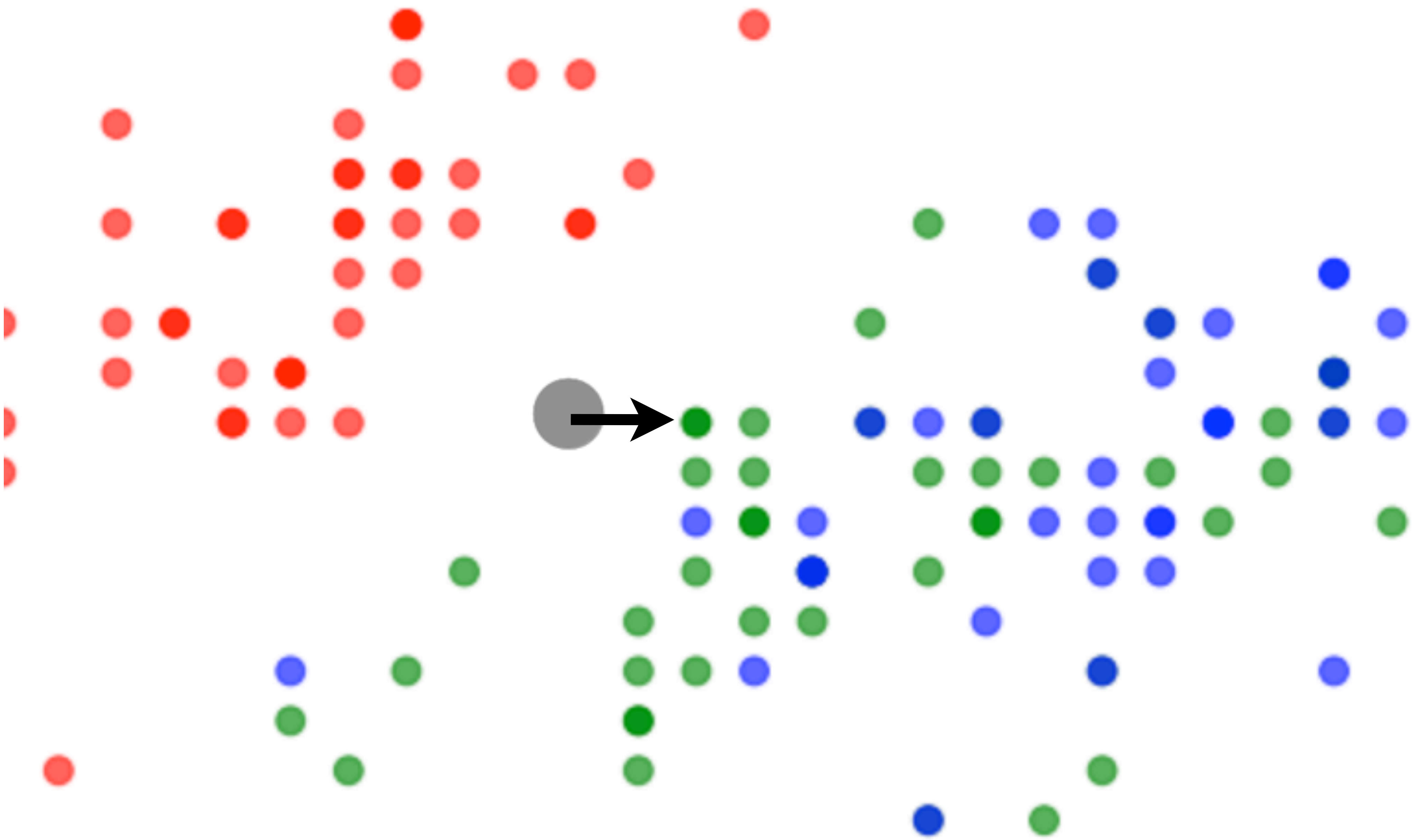


Cluster centroid: $$\mu_k = \frac{1}{||k||} \sum_{x \in k} x$$

Remember KNN Classification?



Remember KNN Classification?



The same idea can be used to find coherent clusters in the data.

K-means Clustering Algorithm

- Given:
  - a dataset $X = \{x_1, \cdots, x_n\}$
  - a distance measure $d(x_i, x_j)$

- Randomly assign data points to K clusters
- repeat
  - calculate cluster centroids

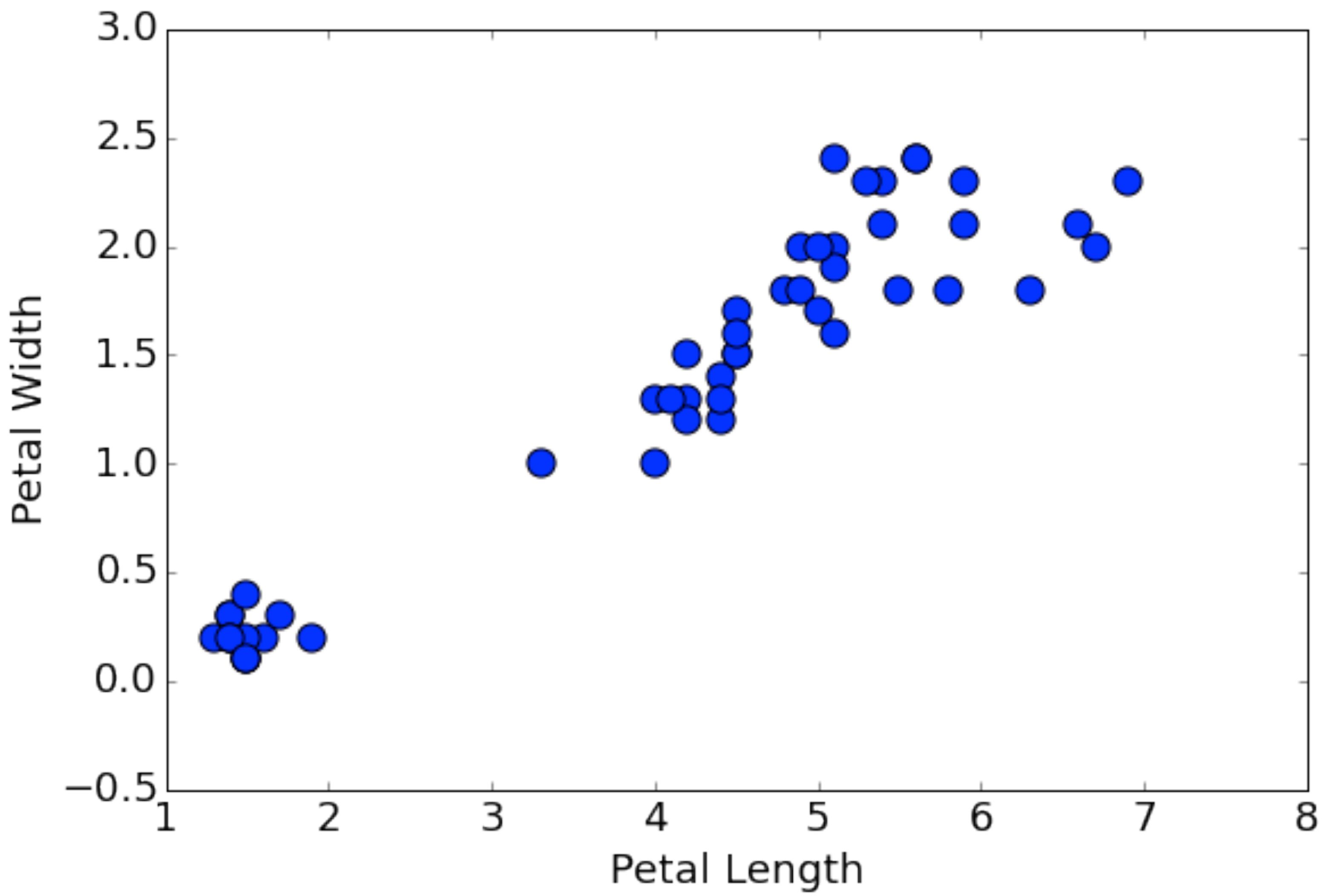
$$\mu_k = \frac{1}{||k||} \sum_{x \in k} x$$

  - assign each data point to the cluster with the closest centroid

$$k = \{x | \forall k', d(x, \mu_{k'}) \leq d(x, \mu_k)\}$$

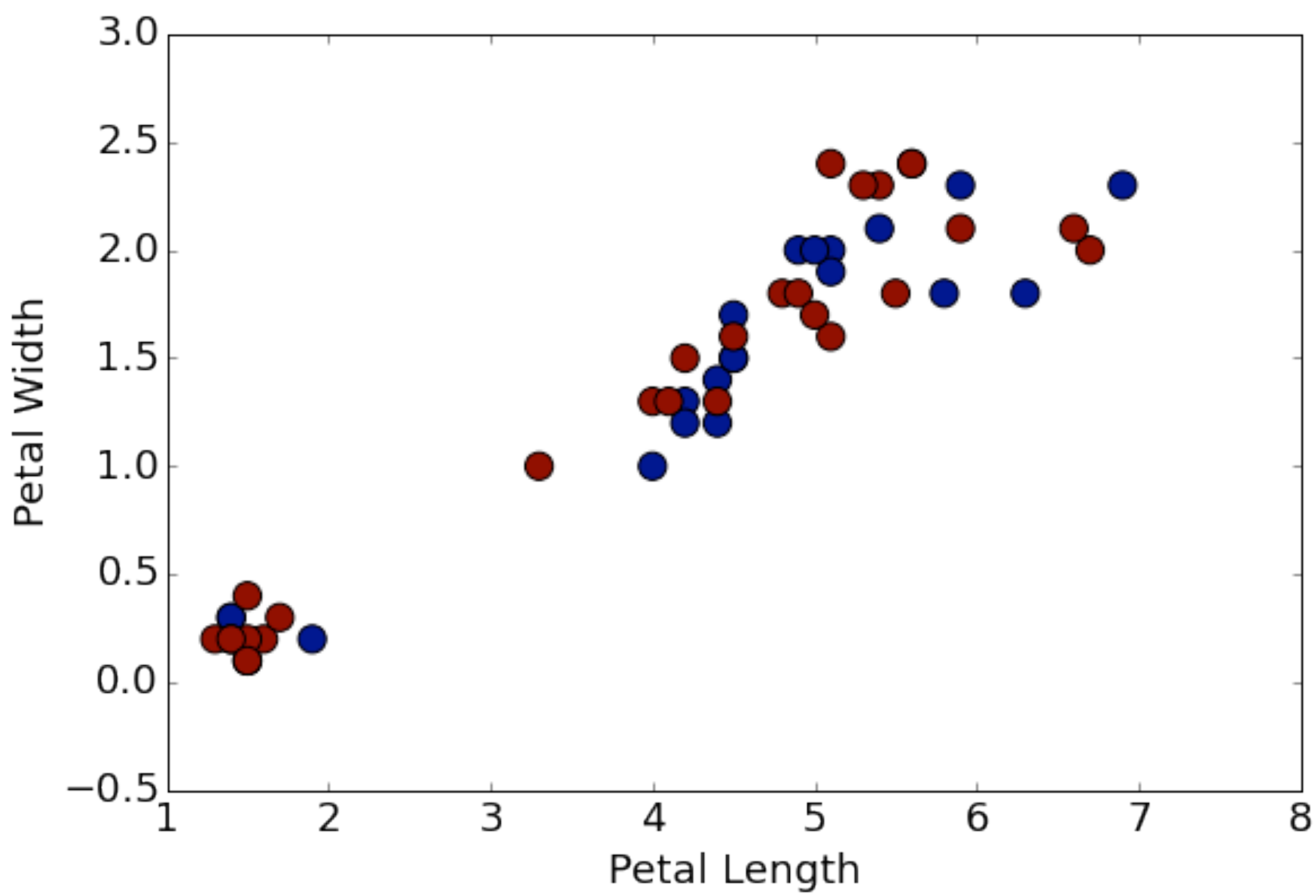
K-means Clustering Algorithm

- The Iris dataset:



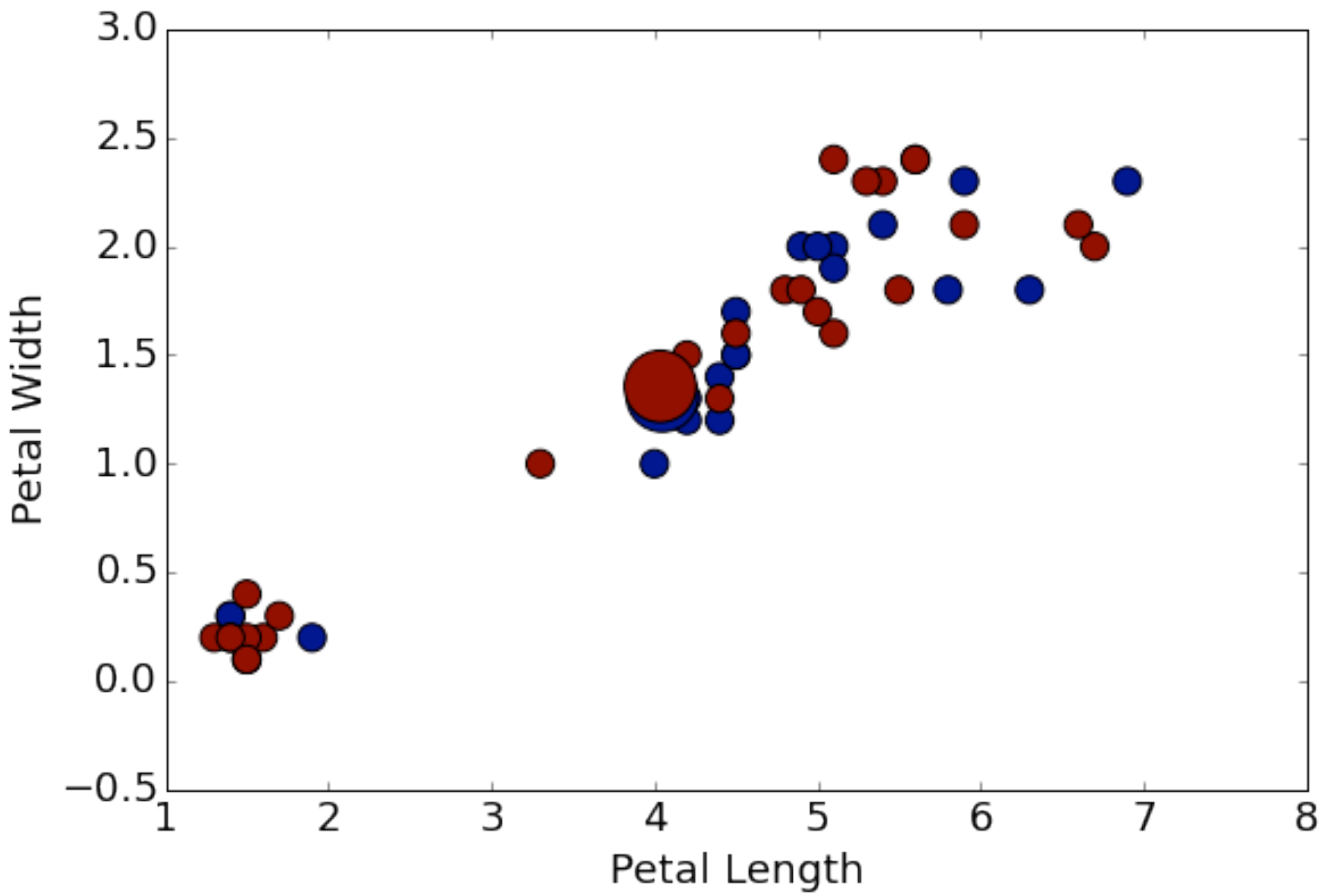
K-means Clustering Algorithm

- Randomly assign points to two clusters:



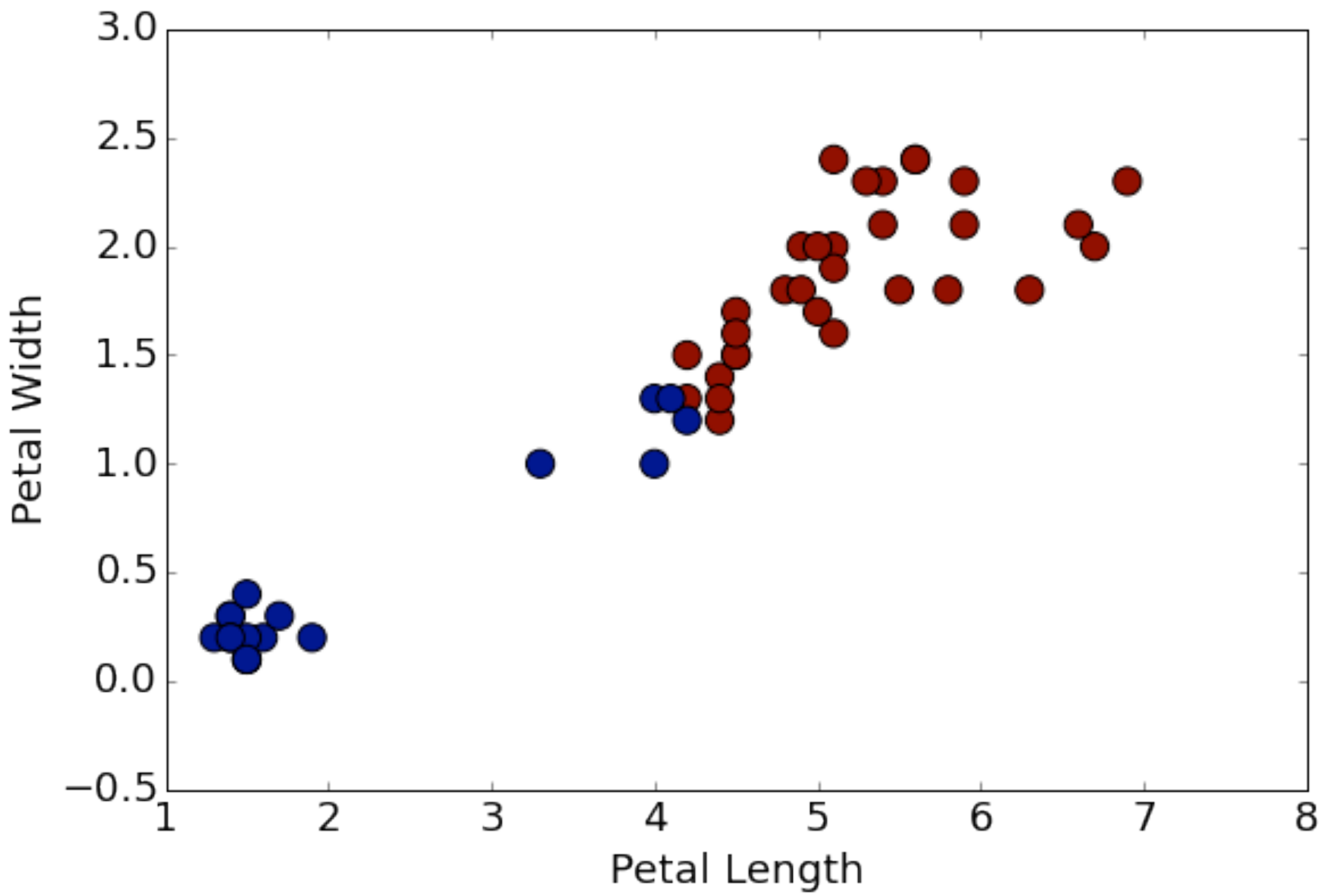
K-means Clustering Algorithm

- Calculate the centroids:



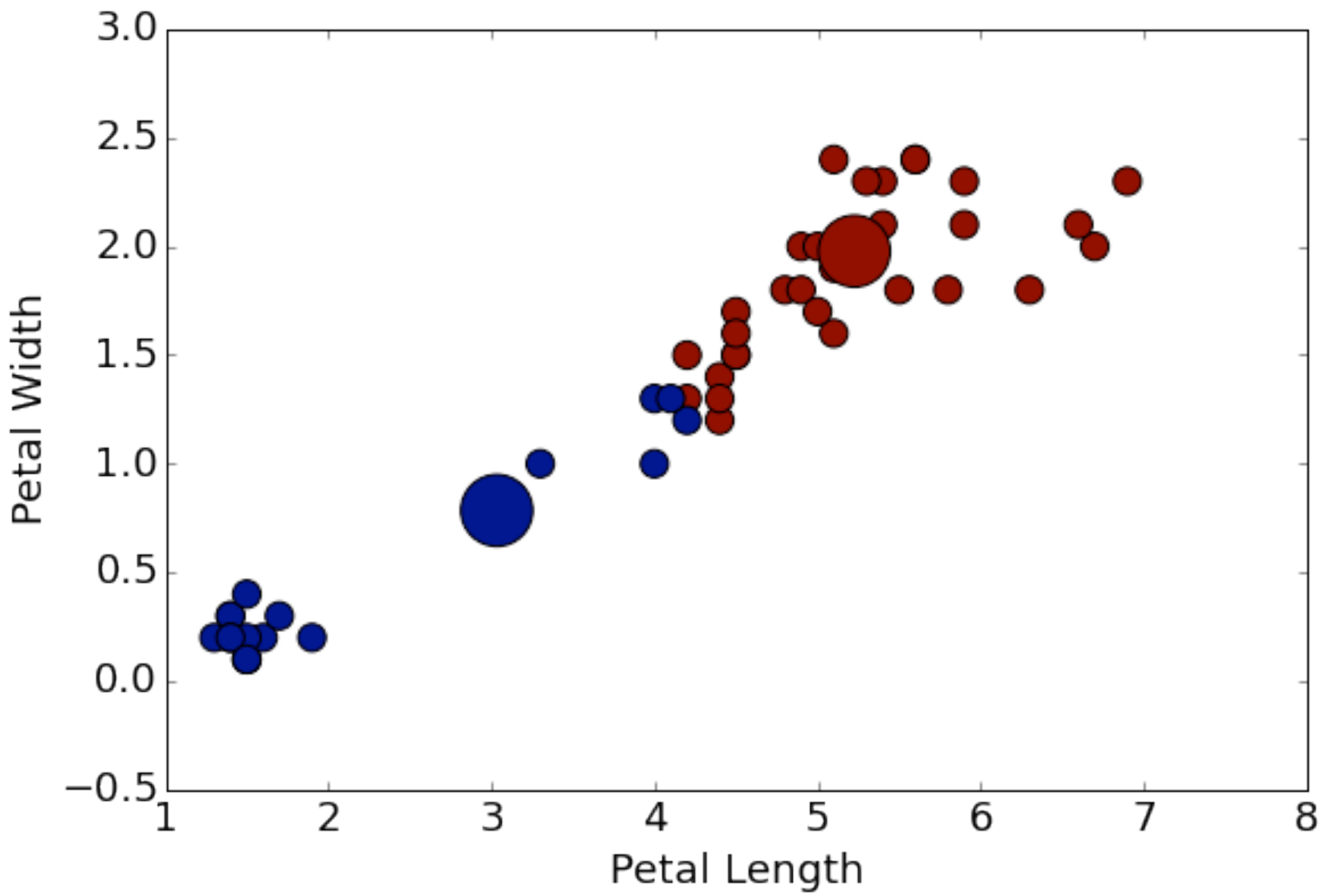
K-means Clustering Algorithm

- Re-assign the data points to the clusters:



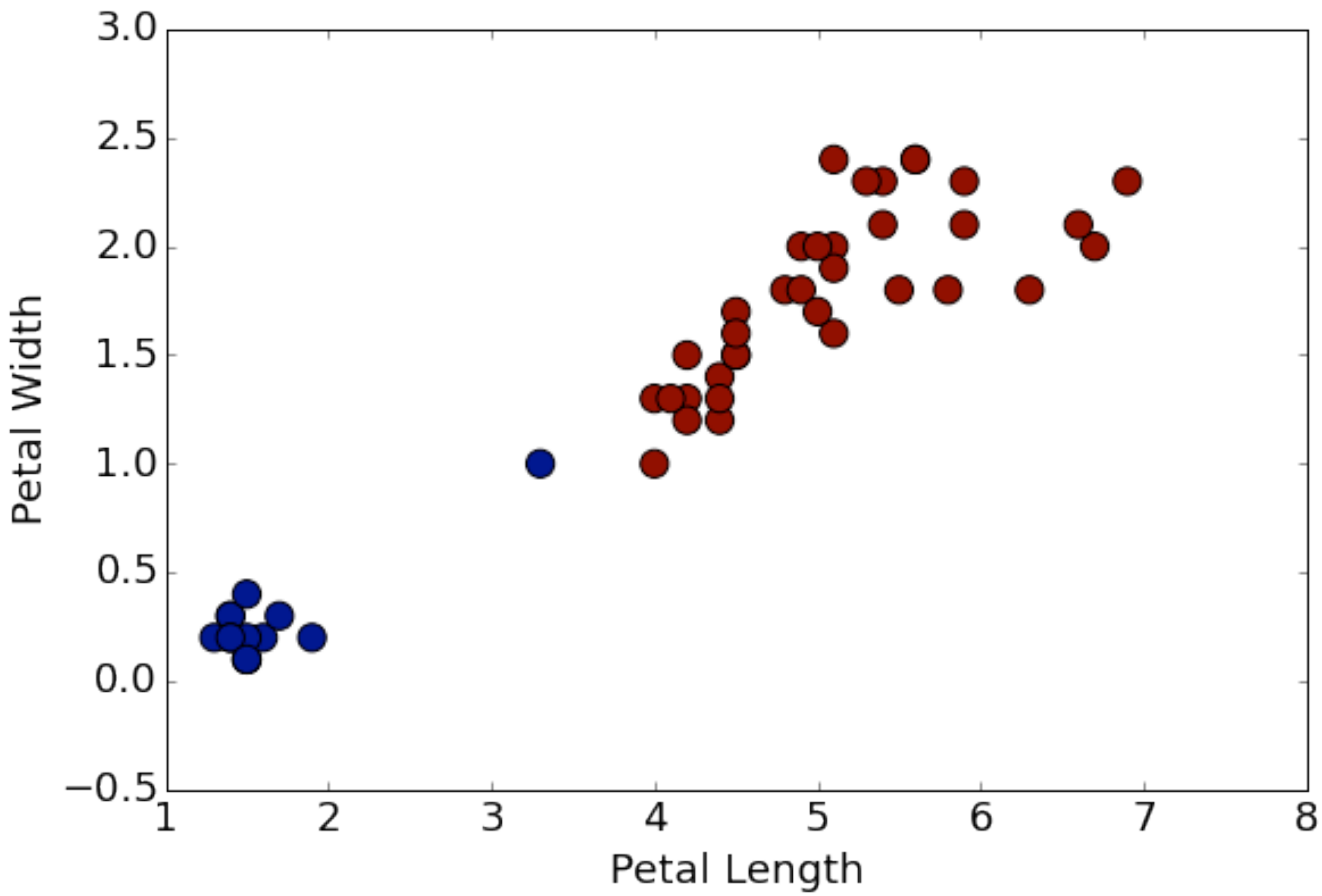
K-means Clustering Algorithm

- Re-calculate the centroids:



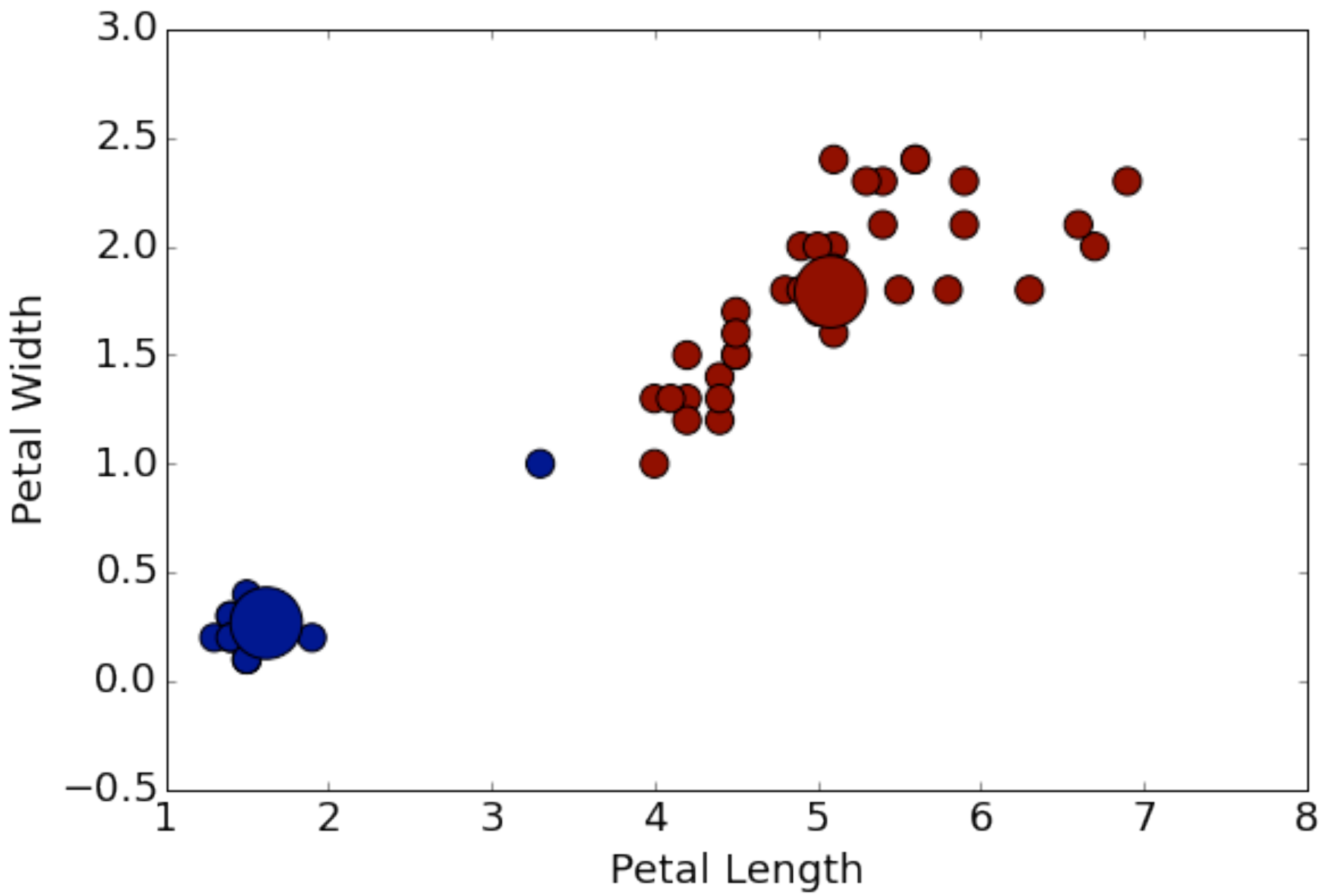
K-means Clustering Algorithm

- And again …



K-means Clustering Algorithm

- … and again.



Objective Function

Objective function for k clusters: members of a cluster must be as close to its centroid as possible

$$J = \sum_{k=1}^{K} \sum_{x \in k} ||x - \mu_k||^2$$

- J: distortion measure
- K: number of clusters
- $x \in k$: members of cluster k
- μ_k: centroid of cluster k

Practical Questions

- When to stop?
  - There are no more changes in the cluster structure/membership
  - The objective function reaches a certain level
- How many clusters?
  - An informed guess?
  - Trying different numbers to see which one yields the best value for some objective function
- Can we speed up the algorithm?
  - Update the centroids incrementally
  - Select the initial centroids wisely (how?)

Evaluating Cluster Quality

- How do we evaluate the quality of the induced clusters?
  - There is no gold standard to compare our clusters against, therefore no precision/recall estimates
  - We can use an objective function as a measure of coherence
- If the induced clusters are used by another task, the performance in that task can be an indicator of the quality of clusters

Remember Naive Bayes?

- Supervised technique for probabilistic classification: choose the most probable class based on the observed features

$$P(\text{Lime}|x_1, x_2) = \frac{P(x_1, x_2|\text{Lime})P(\text{Lime})}{P(x_1, x_2)}$$

- Simplifying assumption: assume features are independent

$$P(x_1, x_2|\text{Lime}) = P(x_1|\text{Lime})P(x_2|\text{Lime})$$

- Prior and likelihood probabilities are estimated based on training data

Unsupervised Naive Bayes

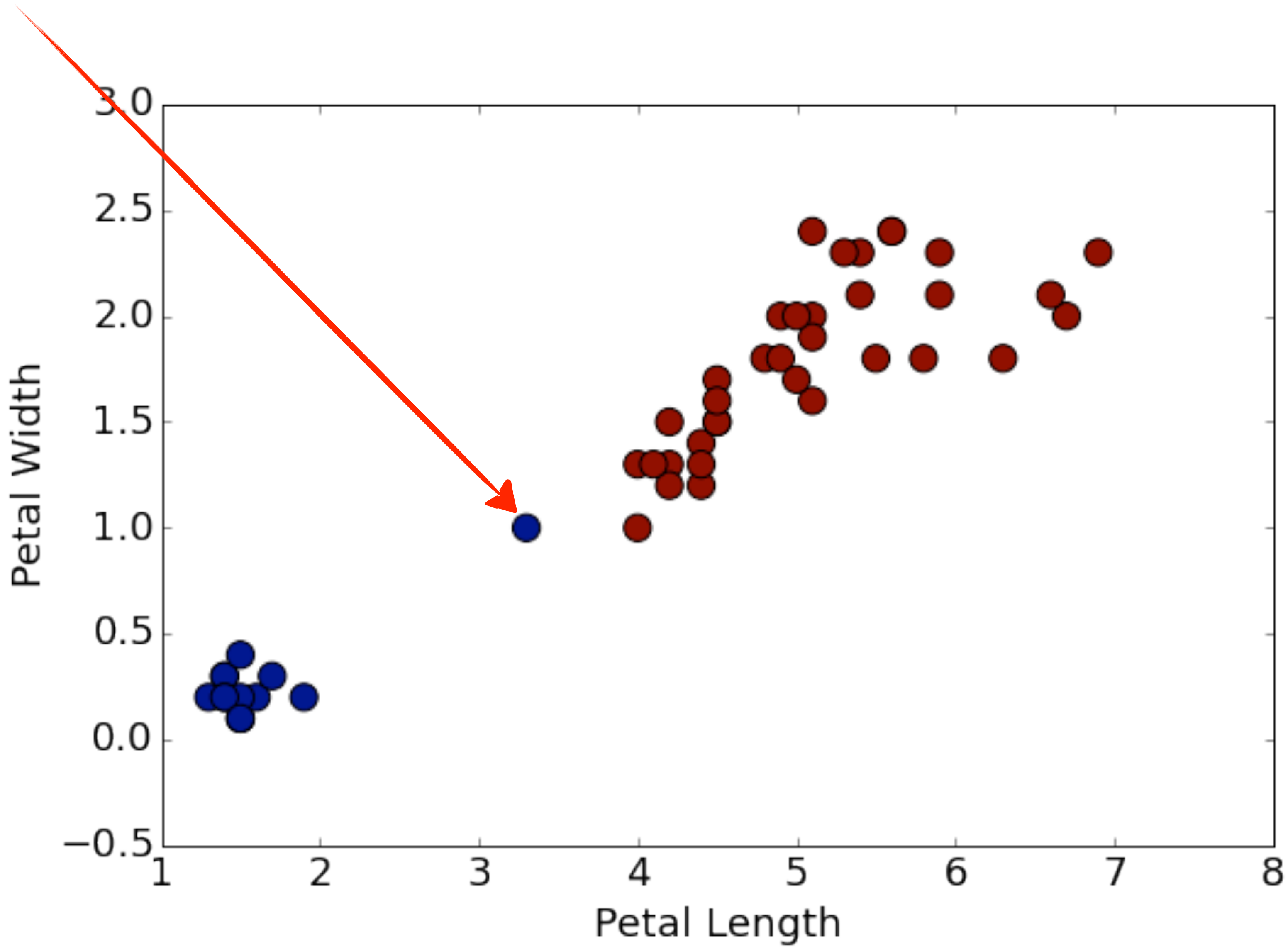
$$y_{\text{pred}} = \arg\max_y P(y) \prod_{j=1}^{J} P(x_j | y)$$

The Expectation-Maximization (EM) Algorithm

1. Guess initial cluster labels for each data point
2. Iterate until convergence:
    a. M-step: compute prior and likelihood probabilities from the labeled data
    b. E-step: re-cluster each example using the estimated probabilities

Hard vs. Soft Clustering

- Could this point also belong to the other cluster?



Soft Clustering

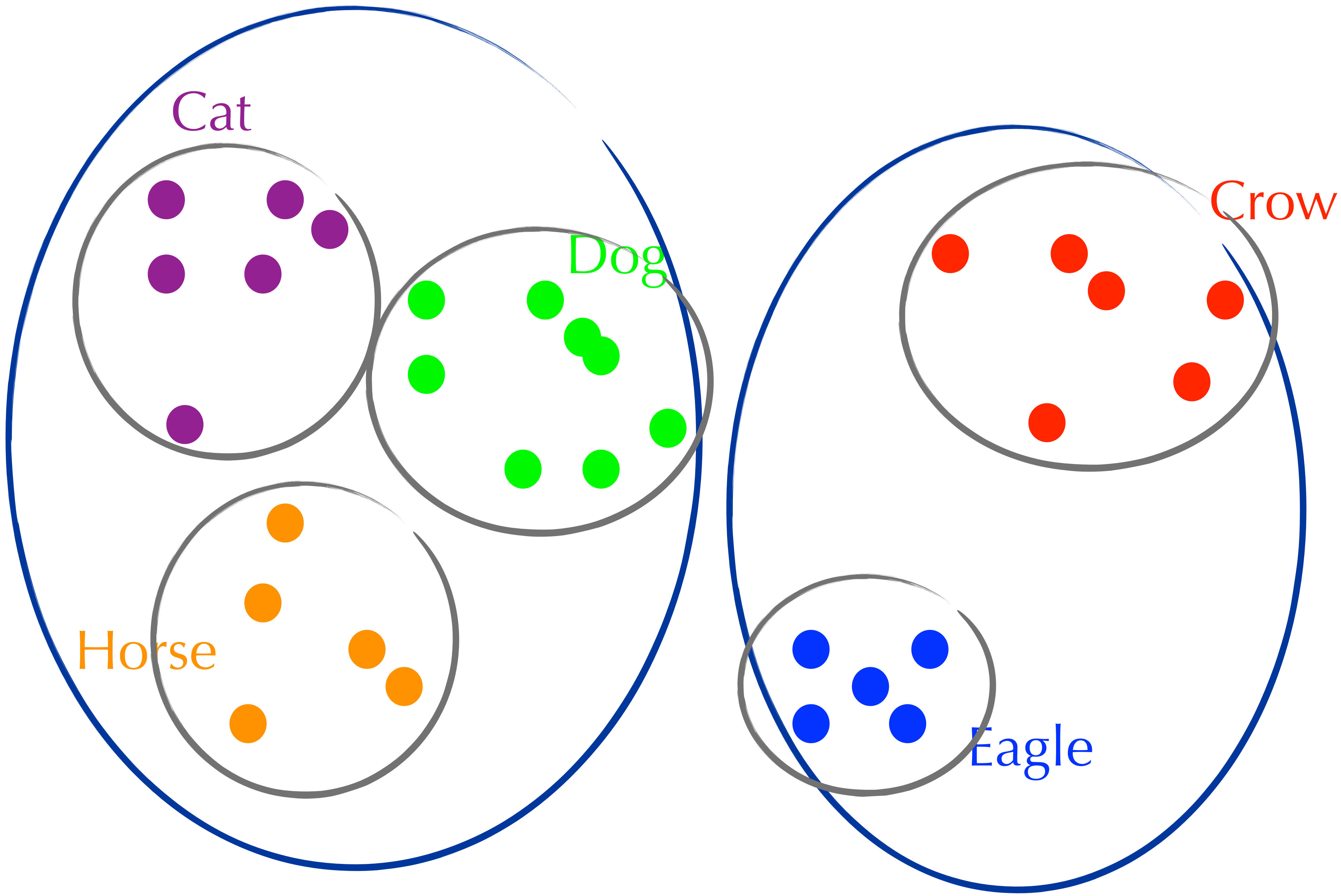
- We can estimate a "membership degree", or a probability that a point belongs to a cluster.

  h_{ik}: the probability that the data point x_i belongs to cluster k

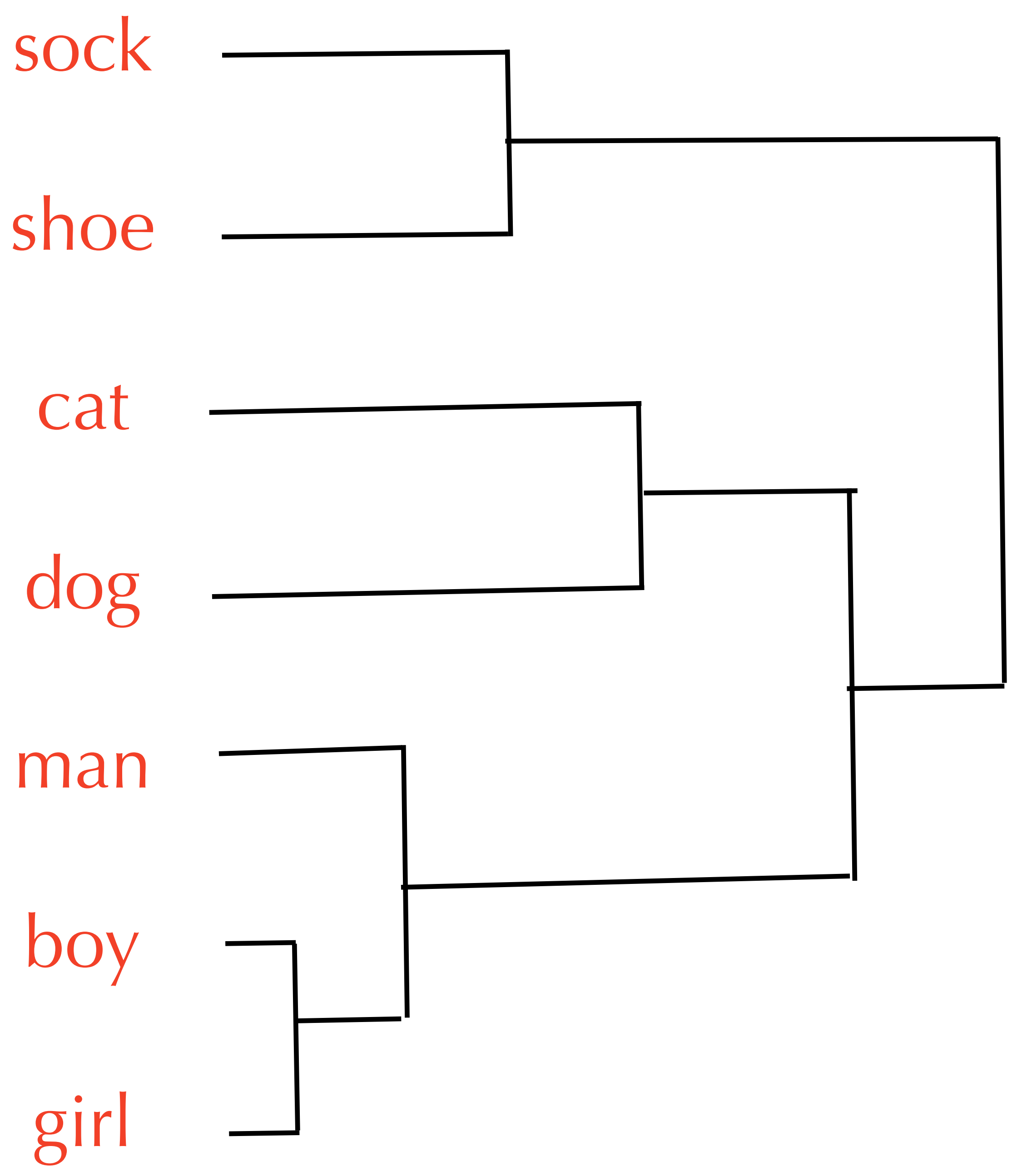
- This probability can show the reliability of a prediction:
  - The flower looks kind of like an iris, but I'm only 73% sure.

- ... or it can show an actual multi-membership:
  - An email can be 65% related to work and 35% related to friends
  - A document can be 82% about politics and 18% about science

Flat vs. Hierarchical Clustering



Hierarchical Clustering



Many Cognitive Tasks are Unsupervised

- Image processing:
  - Recognizing edges, texture, shadows, ...
  - Estimating distance, overlap, spatial relations, ...
  - Identifying objects
- Formation of concepts:
  - categorizing visual entities (e.g., furniture, humans, food) based on their features (shape, color, size, movement, etc.)
  - categorizing relations (e.g., causal movement, manner of motion, change of state) based on their participants
- Processing language
  - Lexical categories, sentence structure, ...

Questions?