A Tiny Taste of Machine Learning

Clustering

- Partition examples into groups (clusters) such that examples in a group are more similar to each other than to examples in other groups
- Unlike classification, there is not typically a "right answer"
  - Answer dictated by feature vector and distance metric, not by a ground truth label

Photo by Jan Willem

Optimization Problem

optimalization

$$variability(c) = \sum_{e \in c} distance(mean(c), e)^2$$

$$dissimilarity(C) = \sum_{c \in C} variability(c)$$

all clusters

- Why not divide variability by size of cluster?
  - Big and bad worse than small and bad
- Is optimization problem finding a C that minimizes *dissimilarity(C)*?
  - No, otherwise could put each example in its own cluster
- Need a constraint, e.g.,
  - Minimum between clusters
  - Number of clusters

K-means Clustering

- Constraint: exactly k non-empty clusters
- Use a greedy algorithm to find an approximation to minimizing objective function

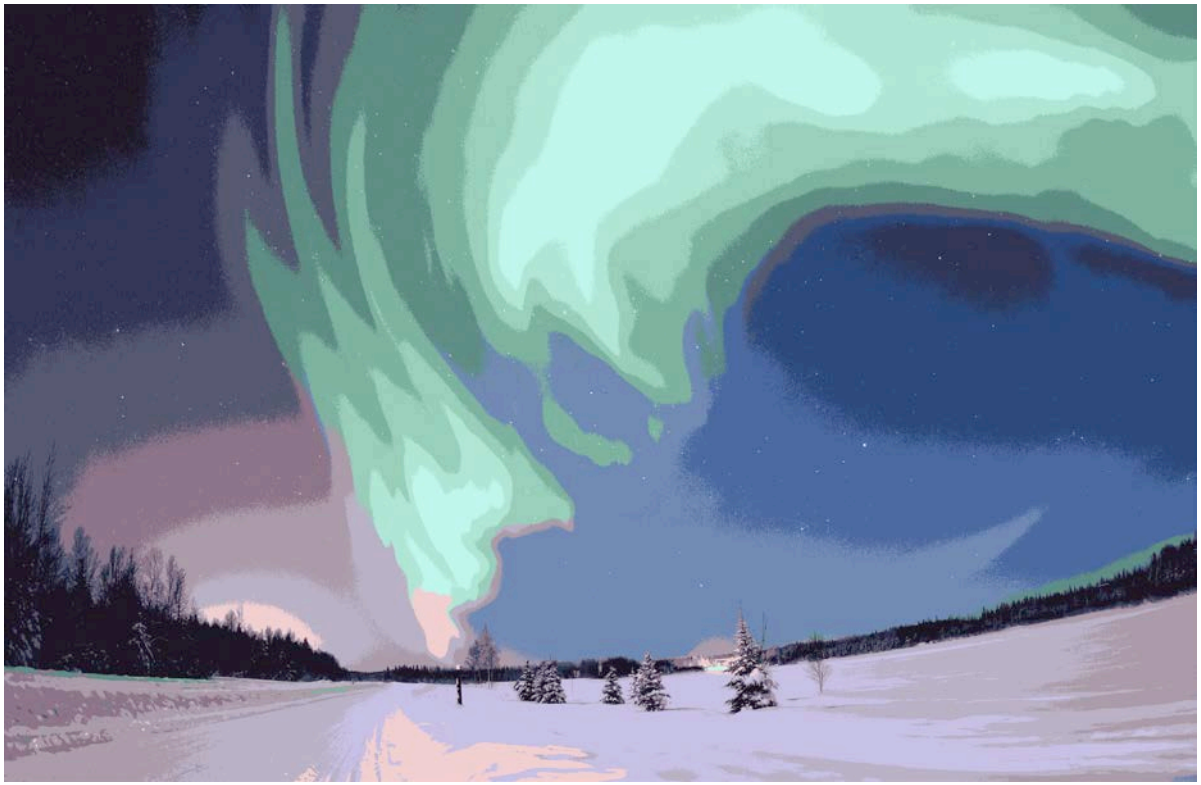
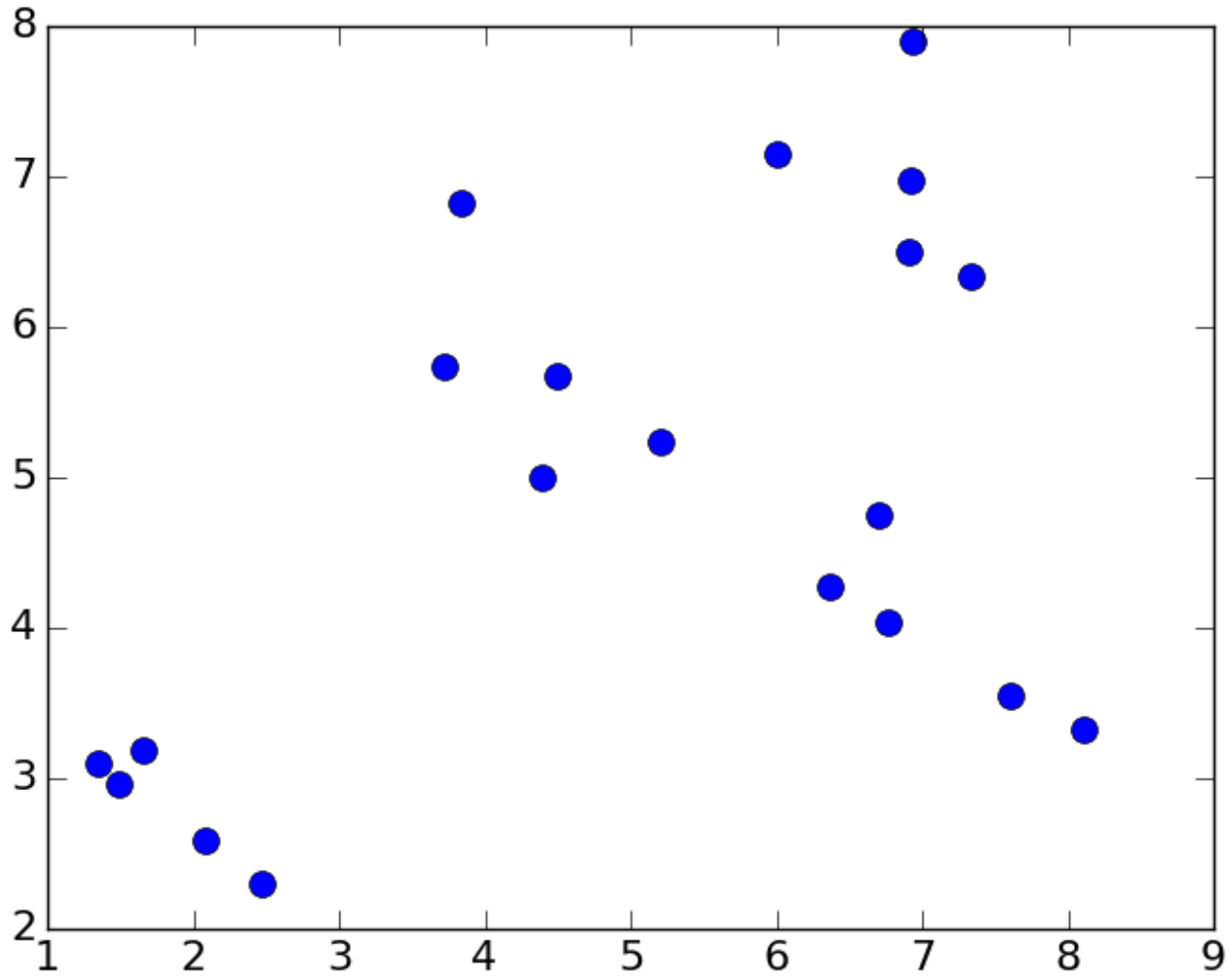


Image by Joshua Strang

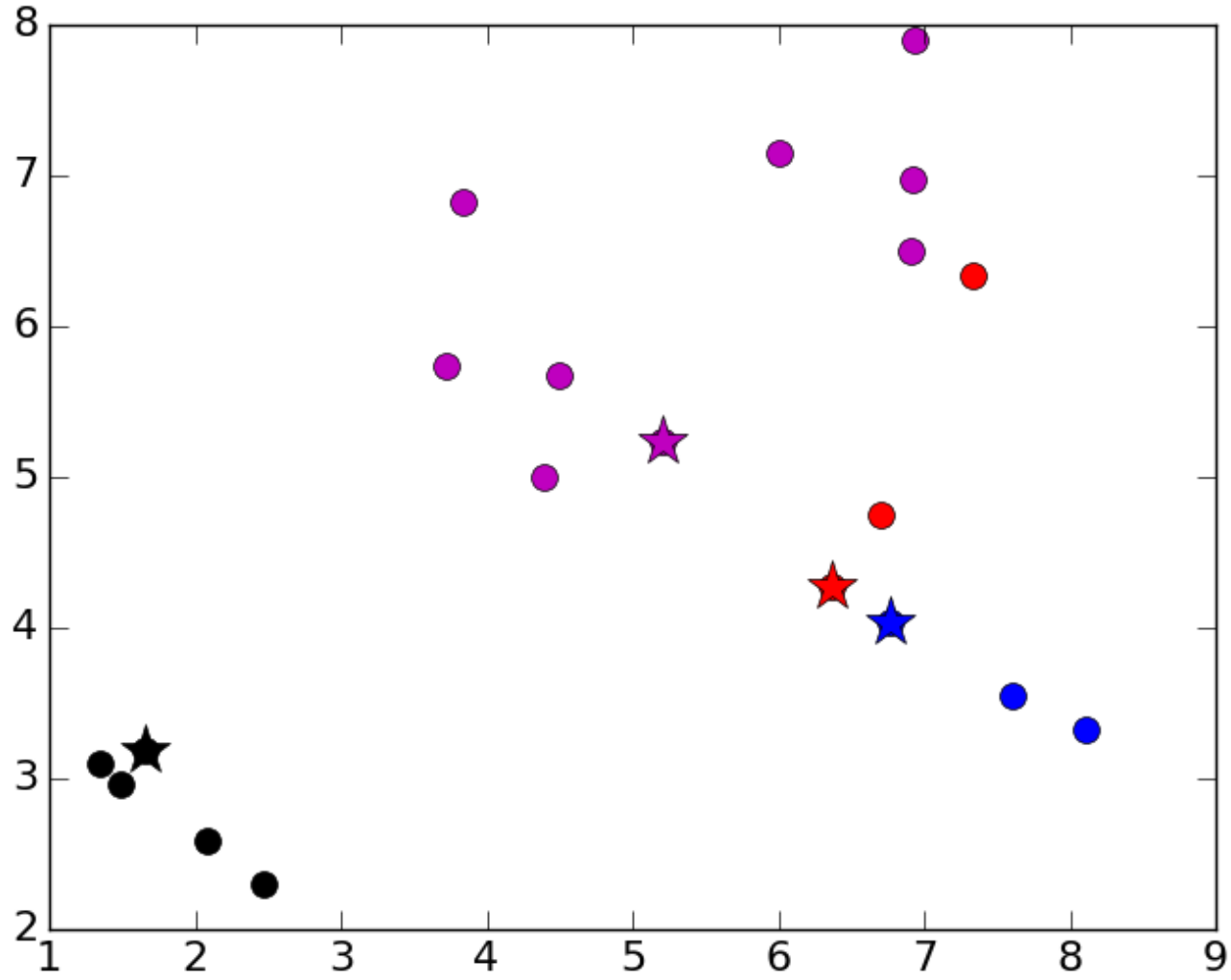
Algorithm

```
randomly chose k examples as initial centroids
while true:
    create k clusters by assigning each
        example to closest centroid
    compute k new centroids by averaging
        examples in each cluster
    if centroids don't change:
        break
```

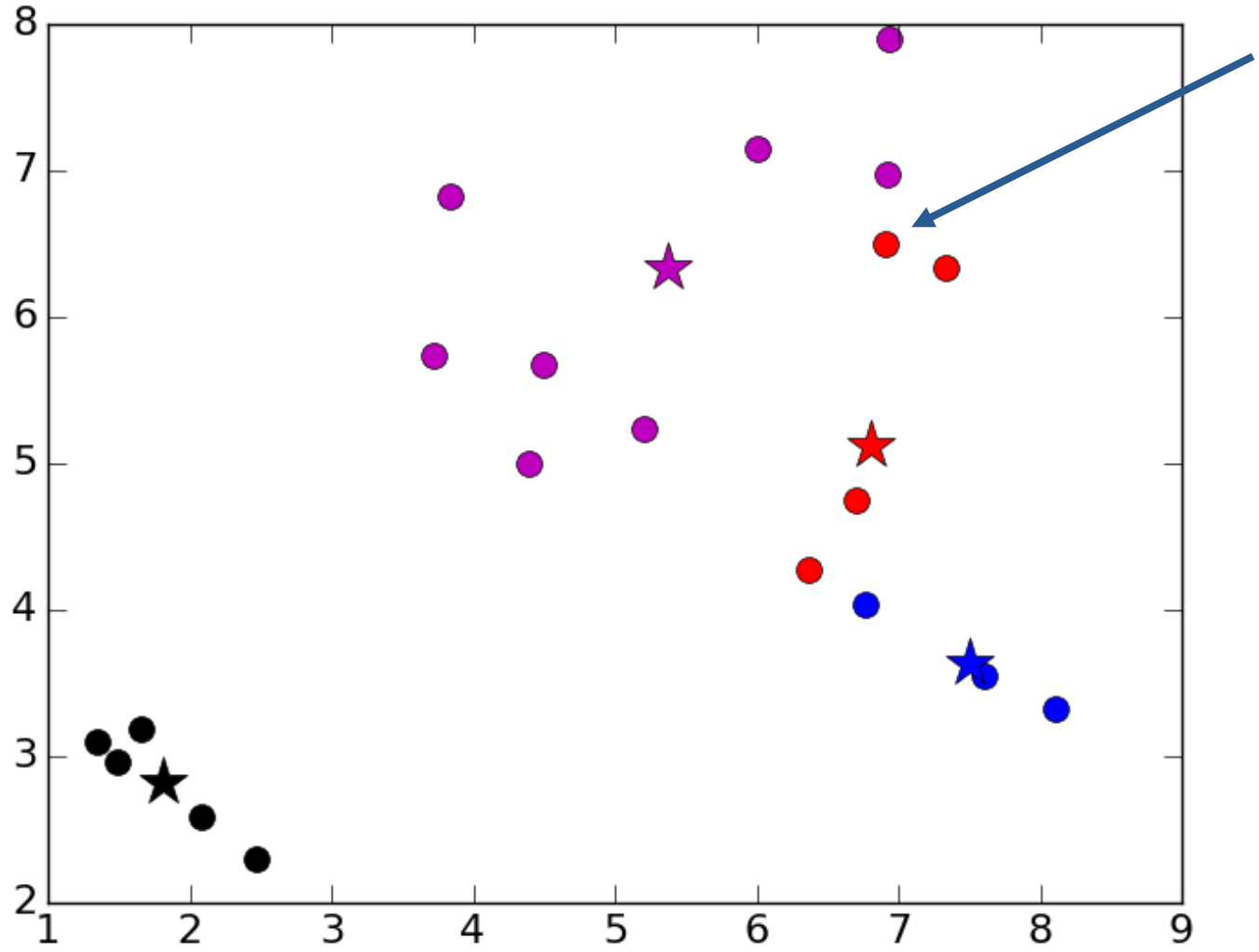
An Example



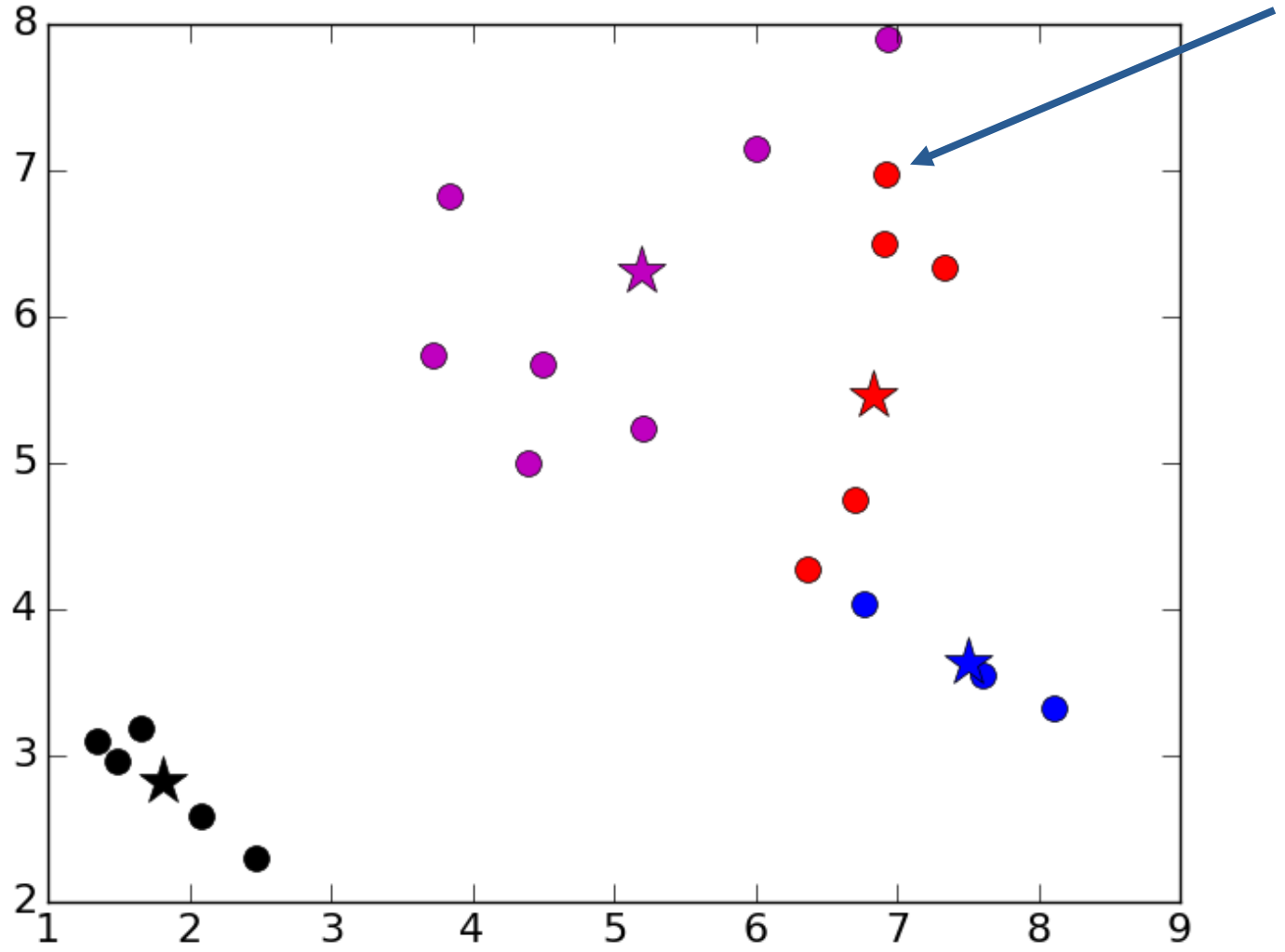
K = 4, Initial Centroids



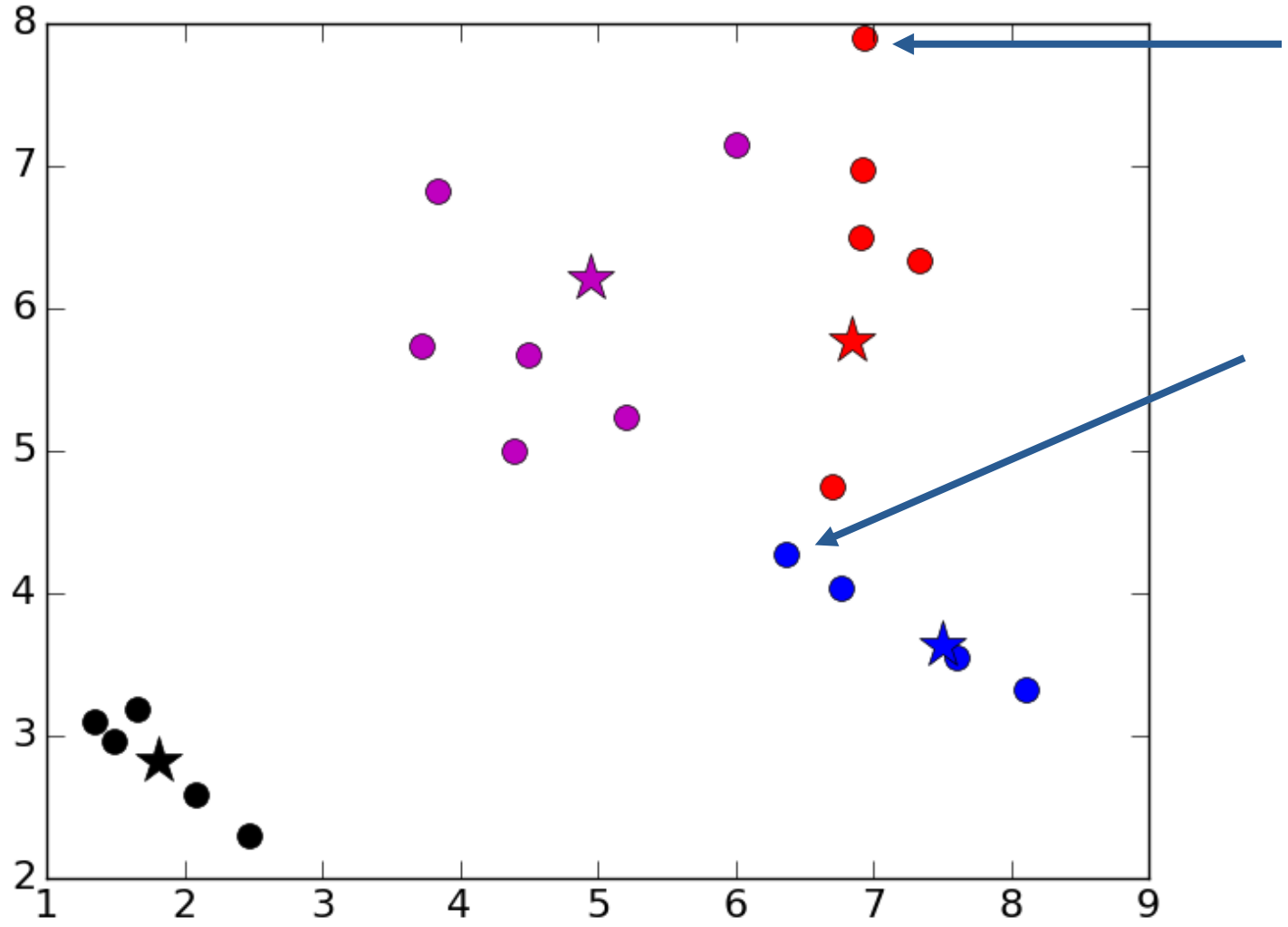
Iteration 1



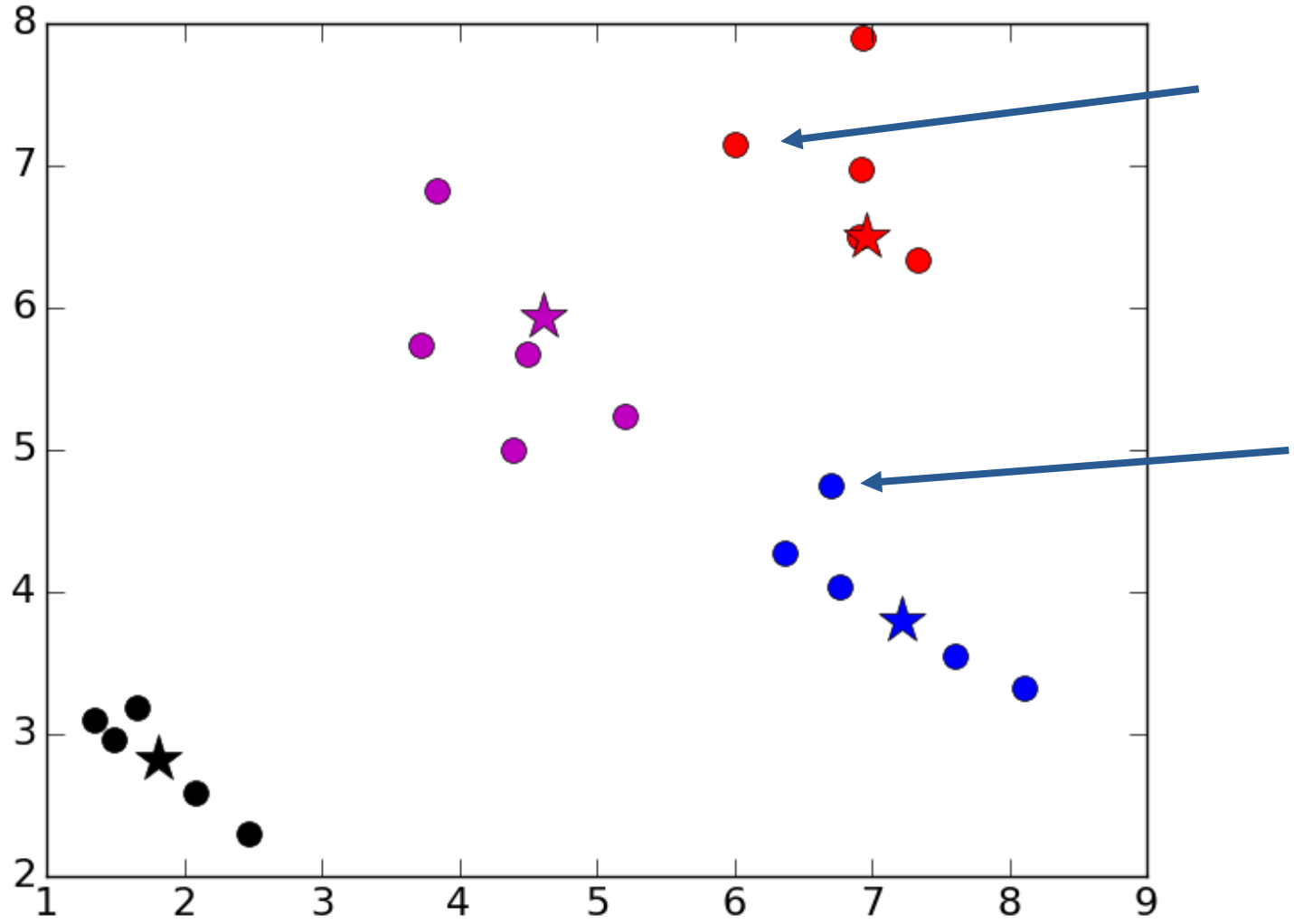
Iteration 2



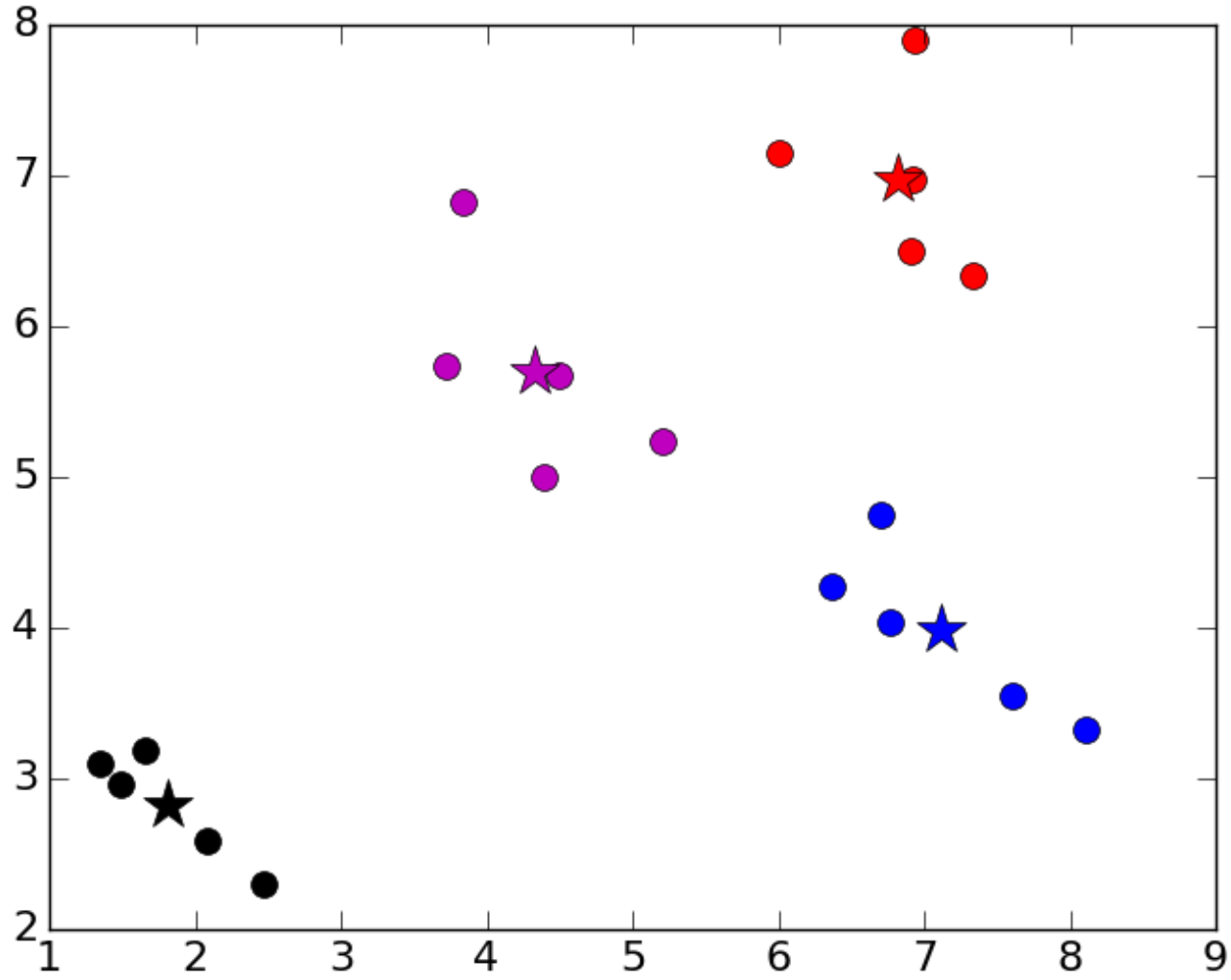
Iteration 3



Iteration 4

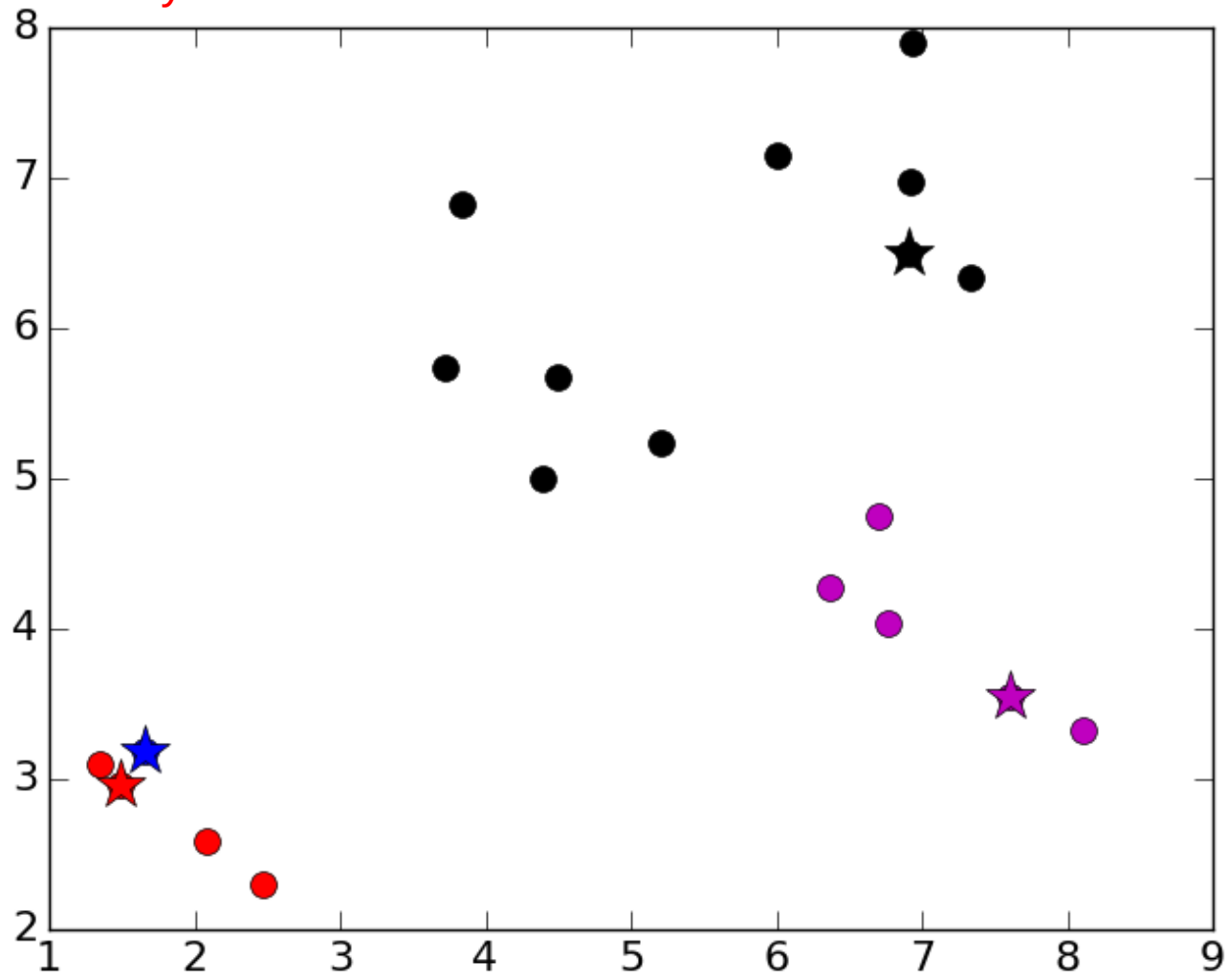


Iteration 5

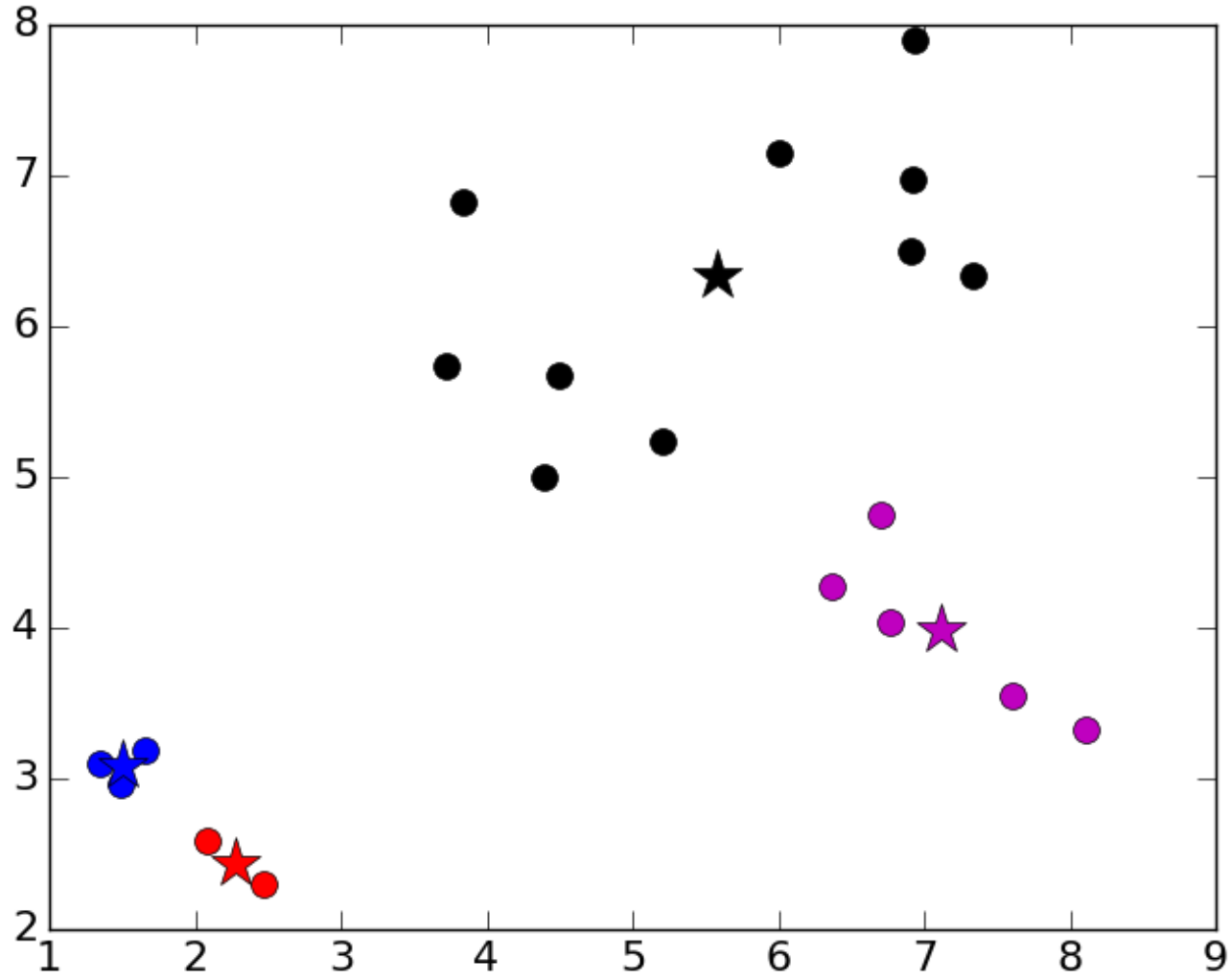


Unlucky Initial Centroids

only the local minum



Converges On



Mitigating Dependence on Initial Centroids

```
best = kMeans(points)
for t in range(numTrials):
    C = kMeans(points)
    if dissimilarity(C) < dissimilarity(best):
        best = C
return best
```

A Pretty Example

- Use k-means to cluster groups of pixels in an image by their color
- Get the color associated with the centroid of each cluster, i.e., the average color of the cluster
- For each pixel in the original image, find the centroid that is its nearest neighbor
- Replace the pixel by that centroid

k = 16

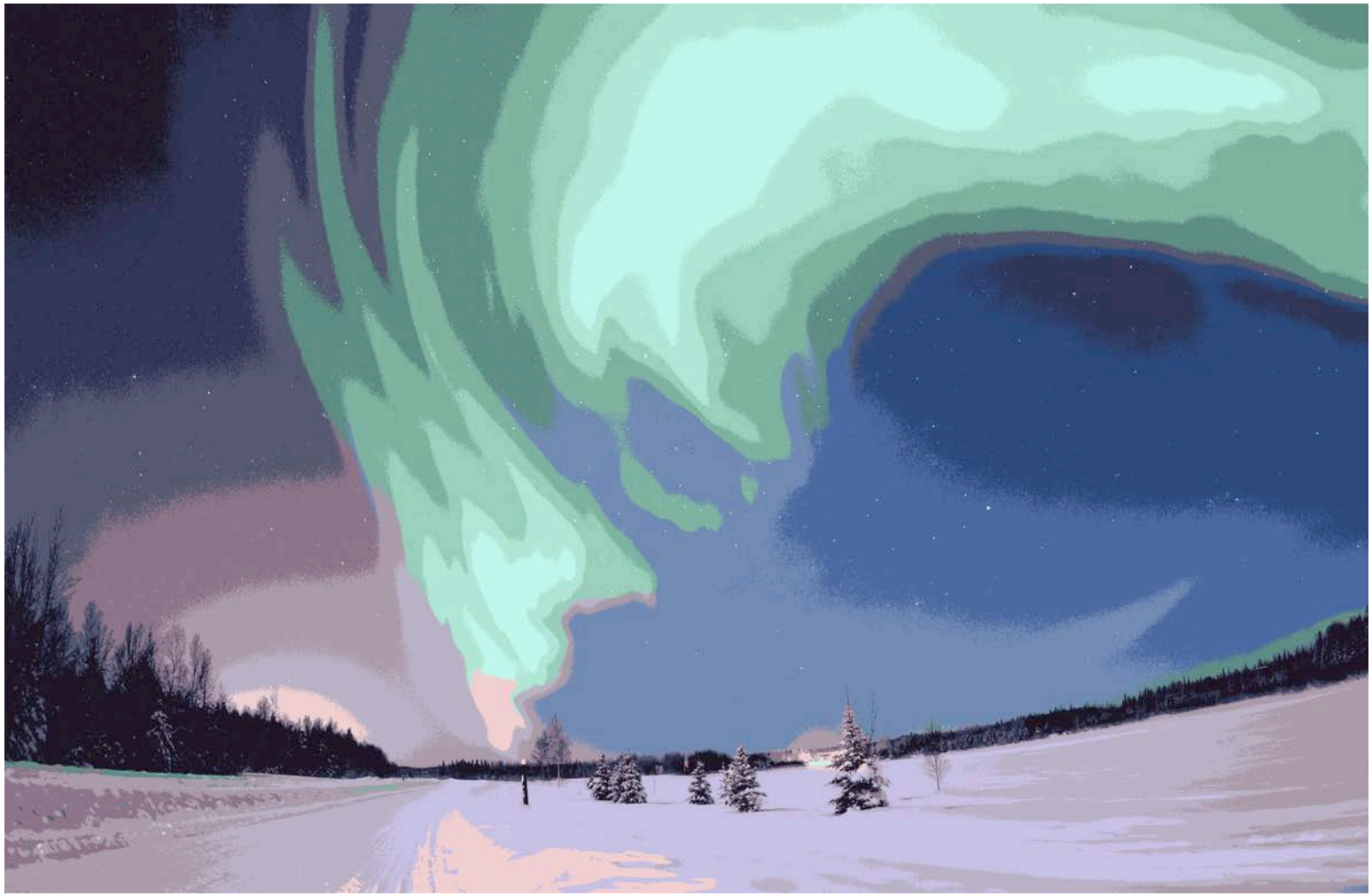


Image by Joshua Strang

Wrapping Up Machine Learning

- Use data to build statistical models that can be used to
  - Shed light on system that produced data
  - Make predictions about unseen data
- Supervised learning
- Unsupervised learning
- Feature engineering
- Goal was to expose you to some important ideas
  - Not to get you to the point where you could apply them
  - Much more detail, including implementations, in text