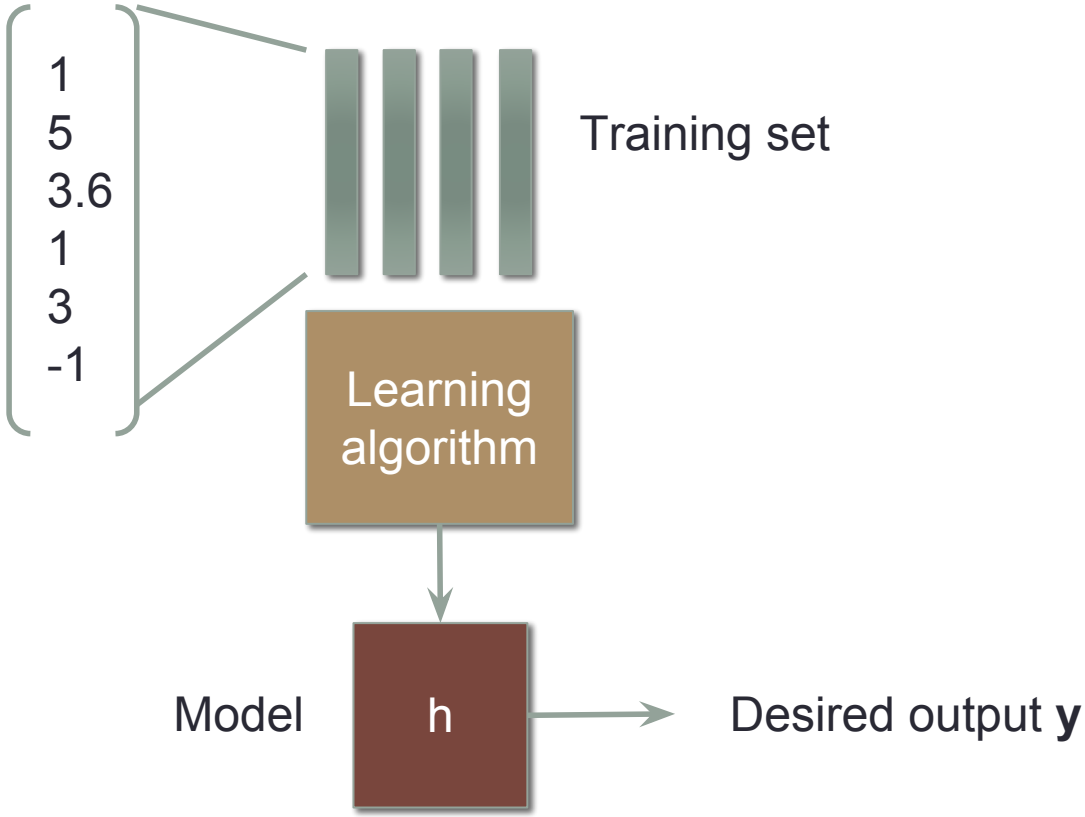
K-means

with hyperparameter tuning

How do we learn from data?



Types of machine learning

1. Supervised learning

   Learn a model F from pairs of (**x**,y)

2. Unsupervised learning

   Discover the hidden structure in unlabeled data **x** (no y)

3. Reinforcement learning

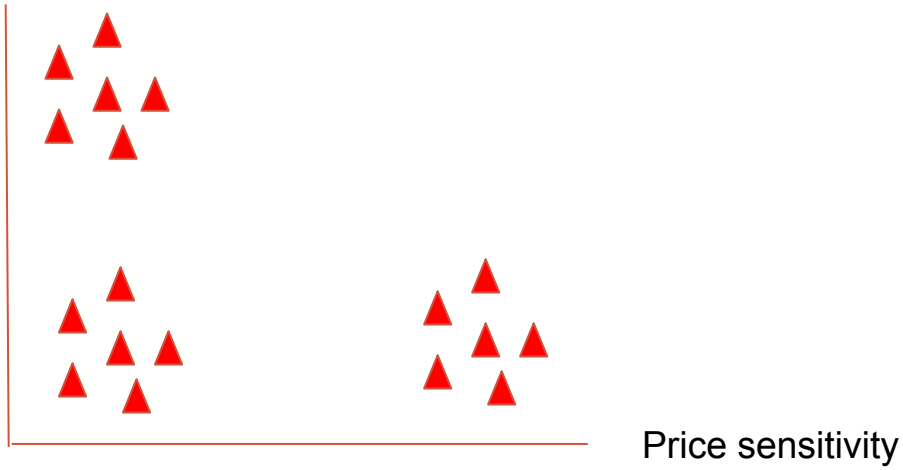
   Train an agent to take appropriate actions in an environment by maximizing rewards

Our first unsupervised learning

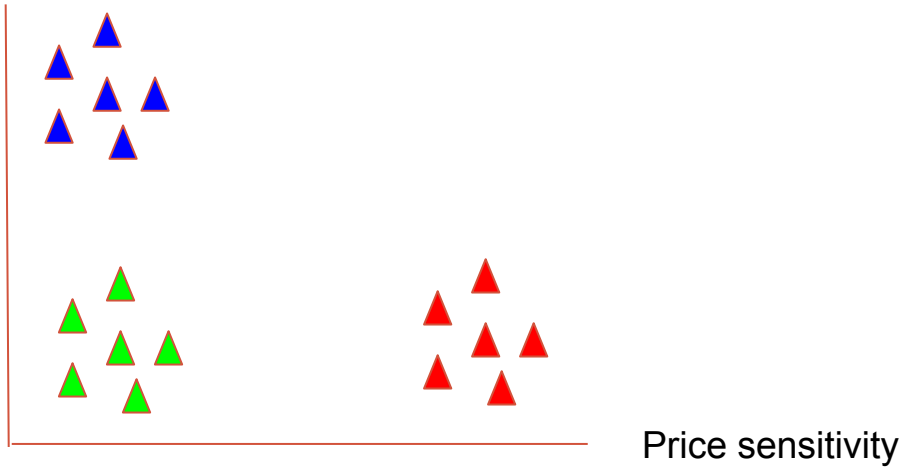
Discover the hidden structure in unlabeled data X (no y)

- Customer/product segmentation
- Data analysis for ...
- Identify number of speakers in a meeting recording
- Helps supervised learning in some task

Example - Customer analysis

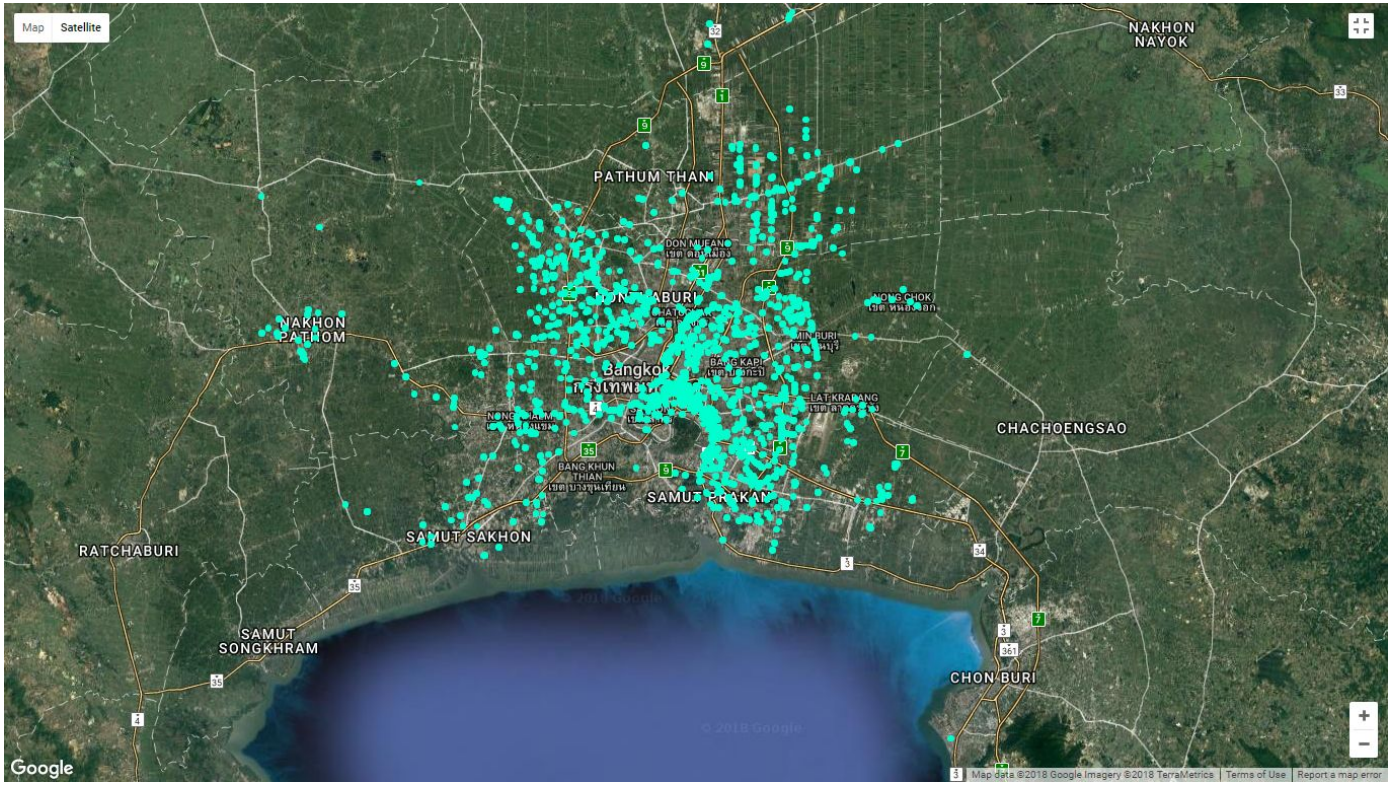


Example - Customer analysis



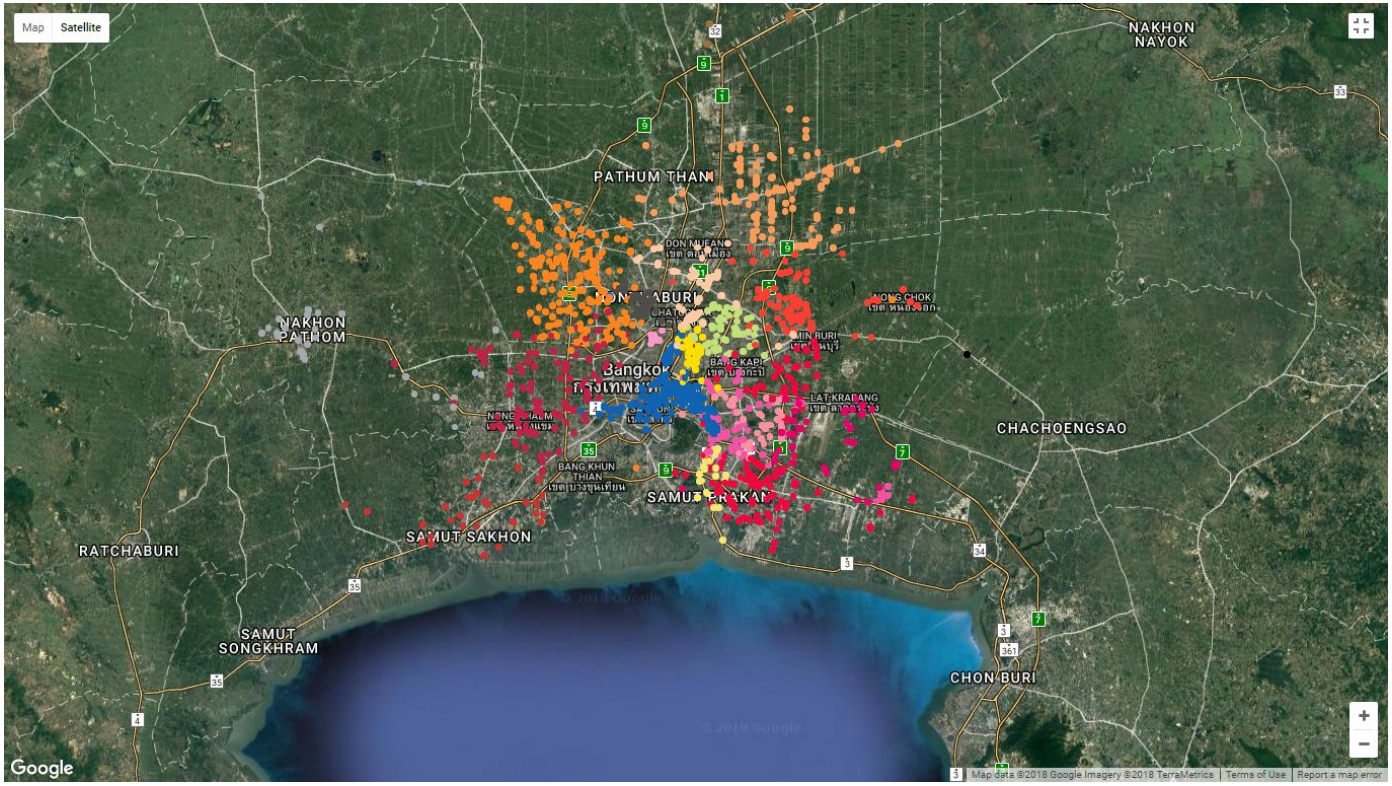
Example - Real Estate segmentation in Thailand

What should be the input feature of this?

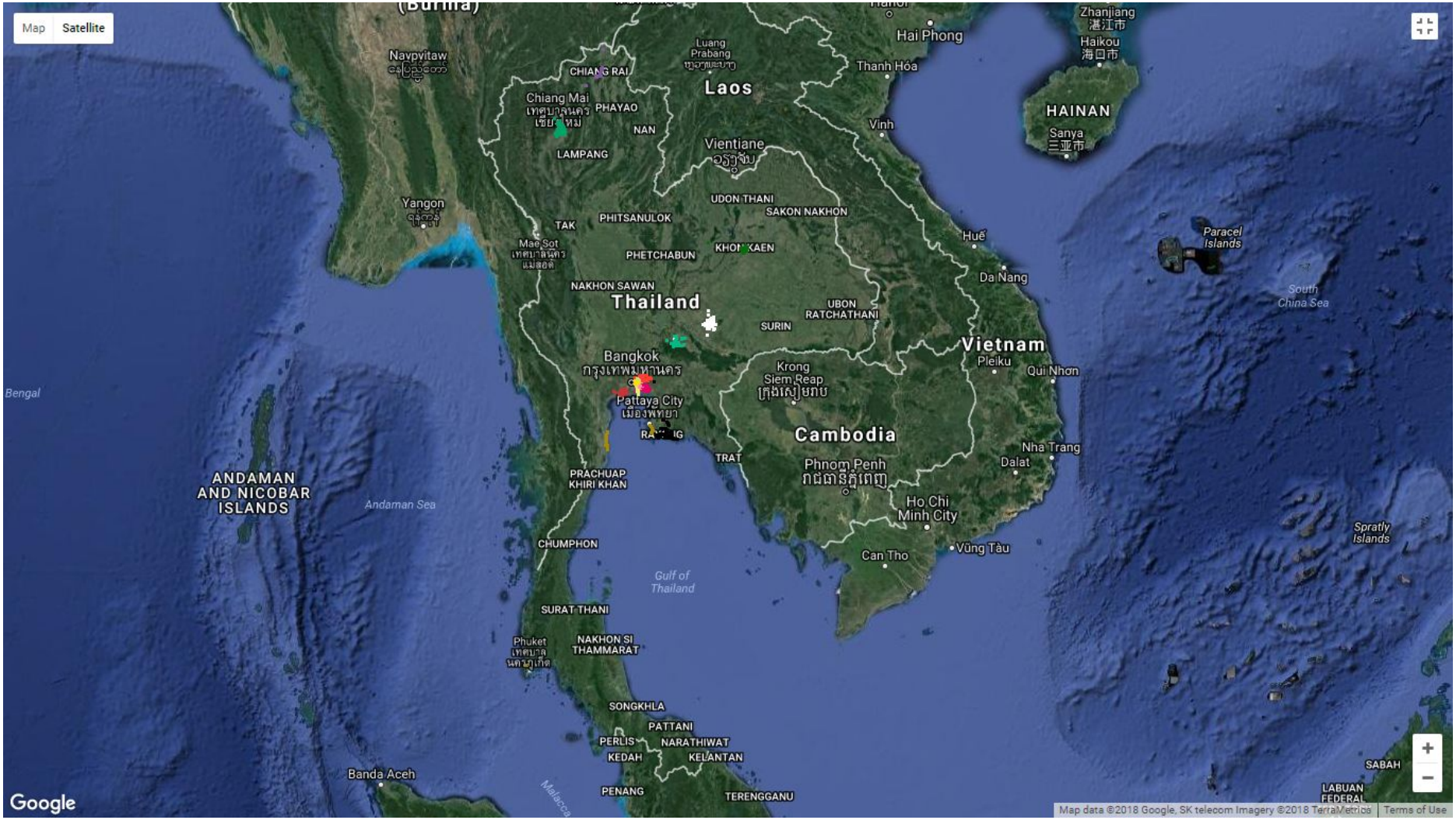


Example - Real Estate segmentation in Thailand

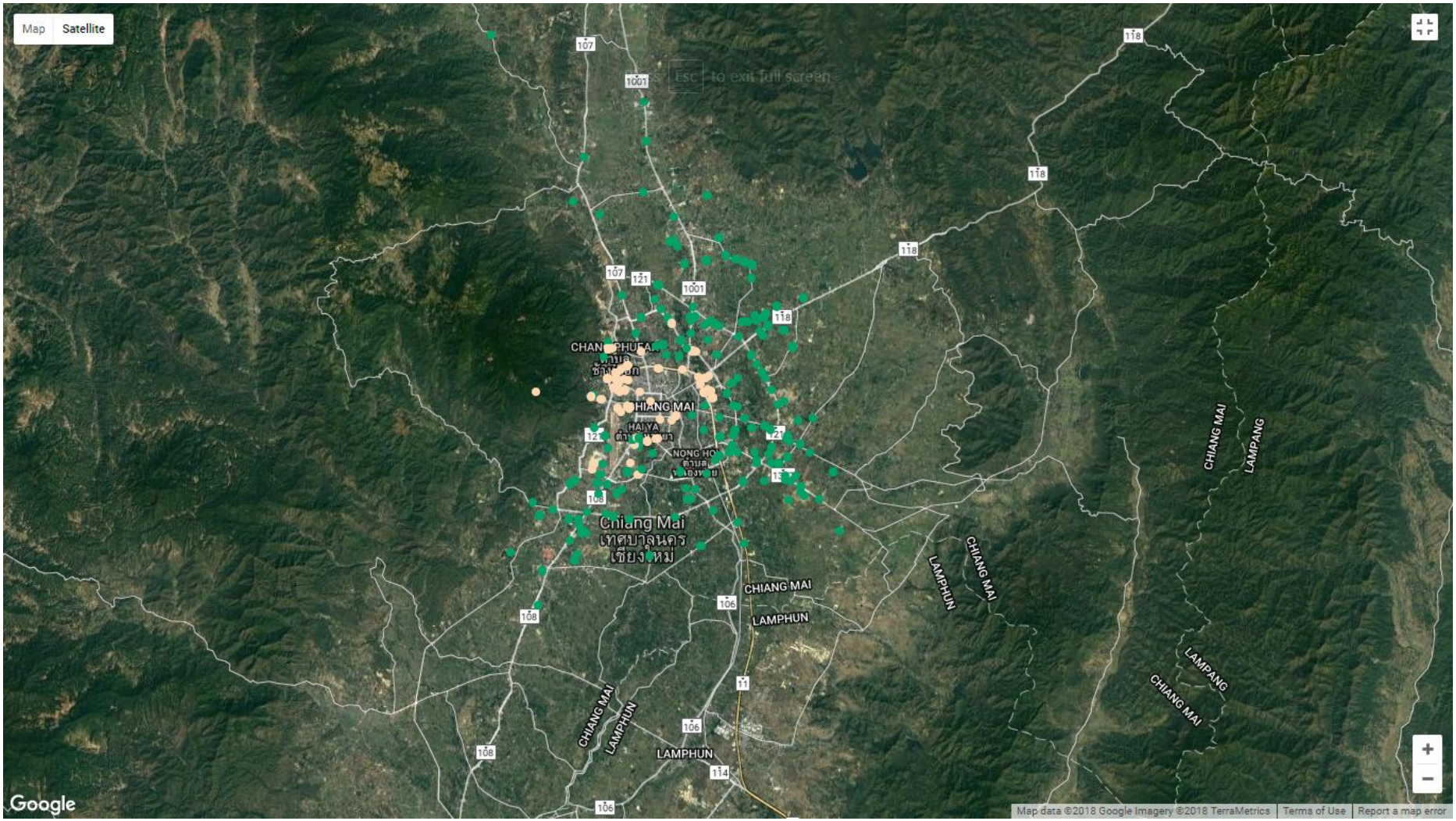
What should be the input feature of this?



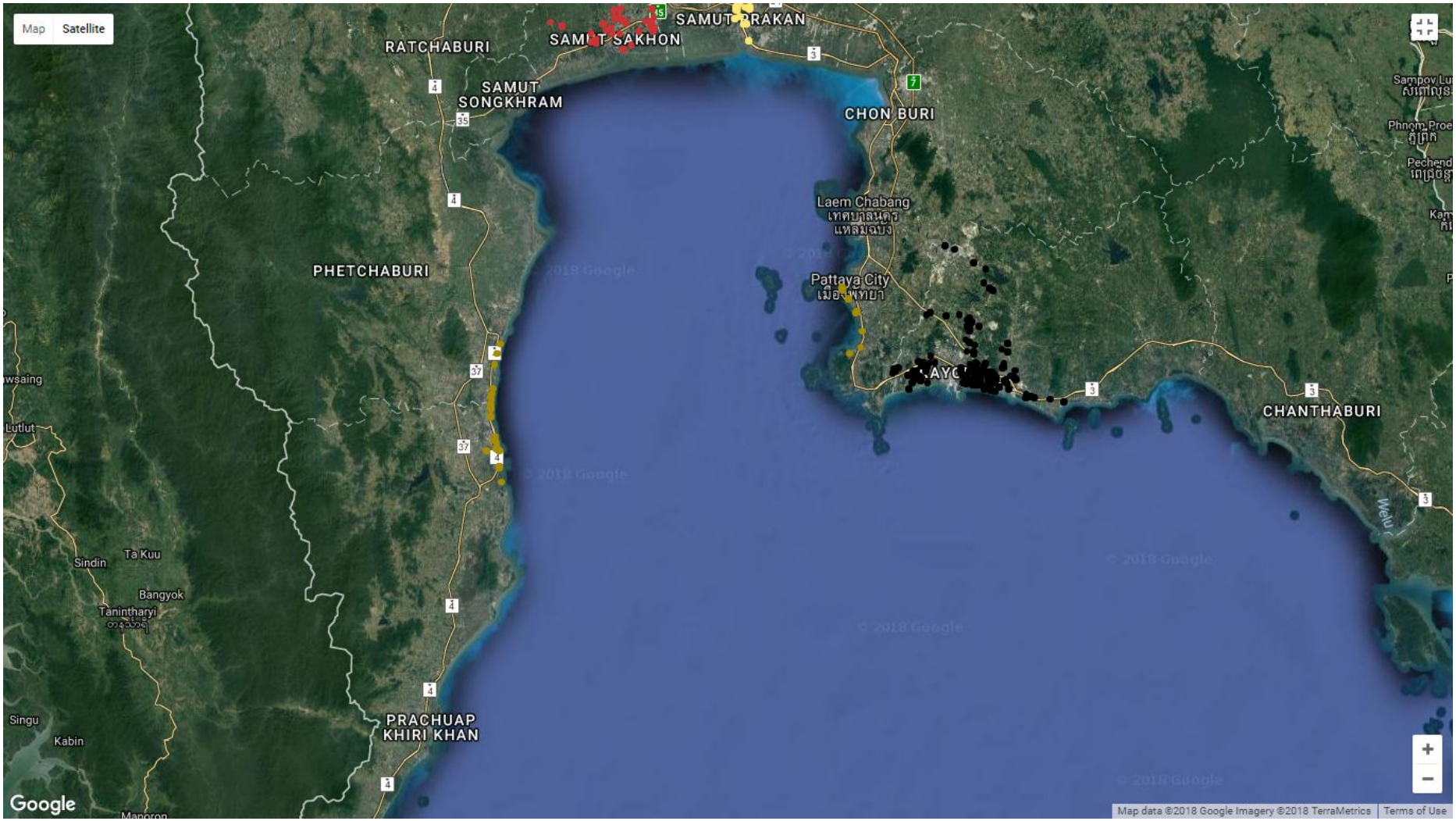
Example - Real Estate segmentation in Thailand



Example - Real Estate segmentation in Thailand



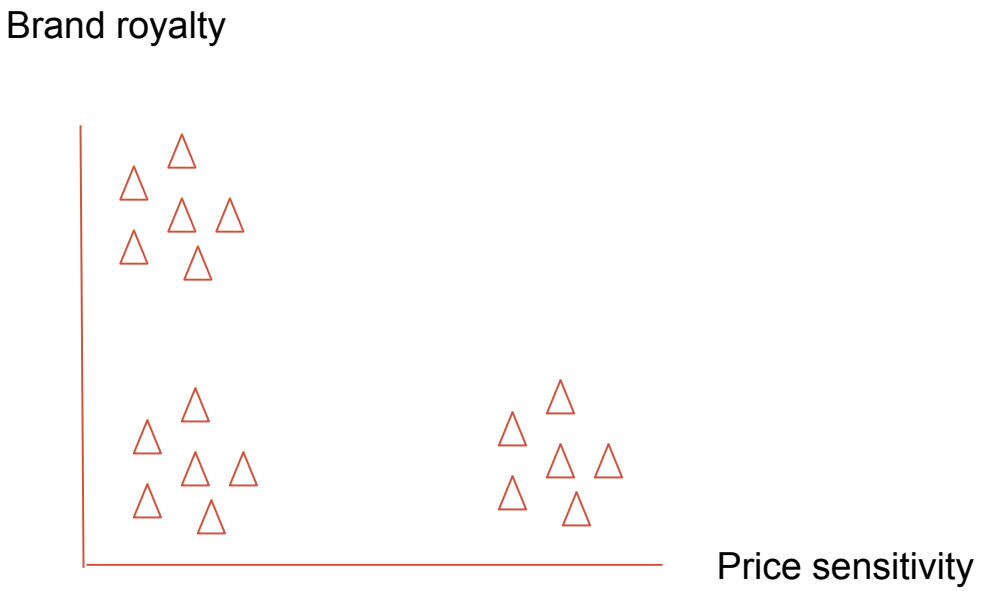
Example - Real Estate segmentation in Thailand



K-mean clustering

Clustering - task that tries to automatically discover groups within the data

Too hard…



K-mean clustering

Clustering - task that tries to automatically discover groups within the data

Too hard…

Easier if we know the grouping beforehand (supervised)

How?

Nearest Neighbour classification

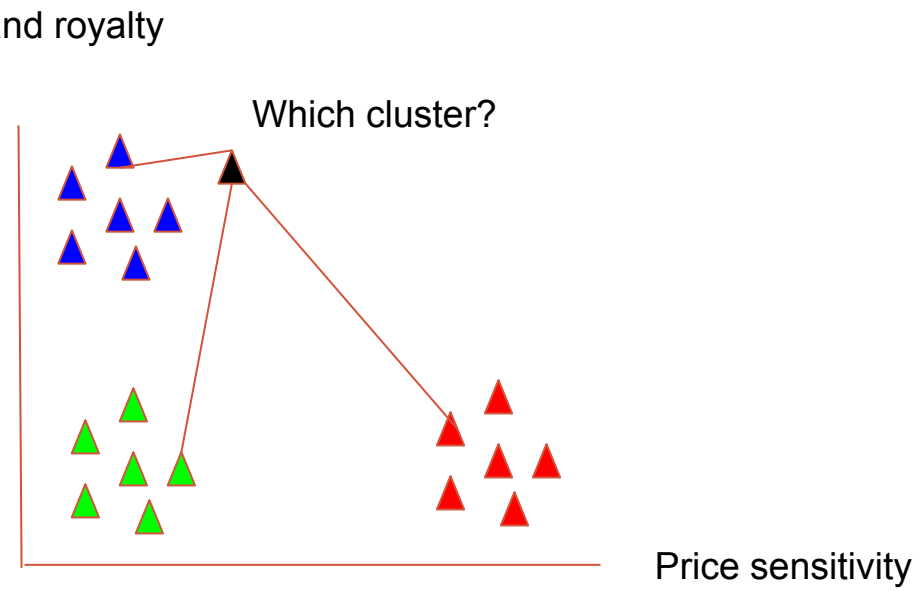
Find the closest training data, assign the same label as the training data

Given query data

For every point in the training data

Compute the distance with the query

Assign label of the smallest distance



K-Nearest Neighbour (kNN) classification

Nearest Neighbour is susceptible to noise in the training data

Use a voting scheme instead

K-Nearest Neighbour (kNN) classification

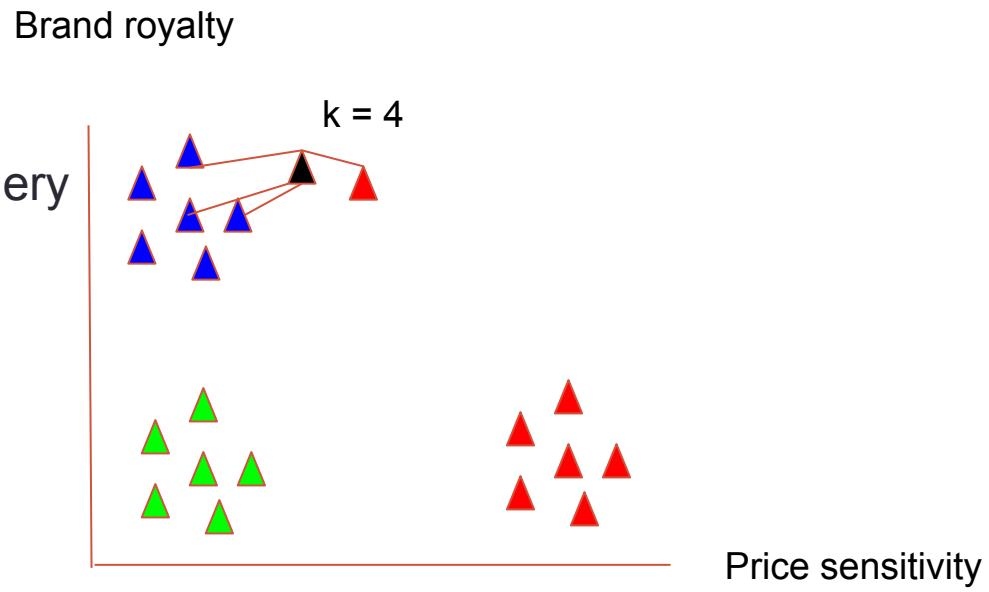
Nearest Neighbour is susceptible to noise in the training data

Use a voting scheme instead

Given query data

For every point in the training data

- Compute the distance with the query
- Find the K closest data points
- Assign label by voting



K-Nearest Neighbour (kNN) classification

Nearest Neighbour is susceptible to noise in the training data

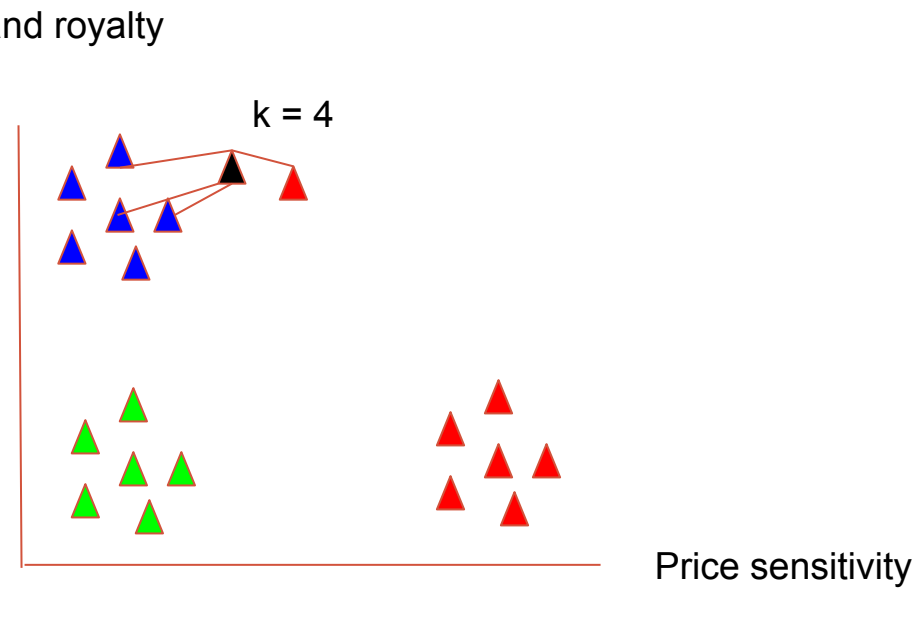
Use a voting scheme instead

Given query data

For every point in the training data

- Compute the distance with the query
- Find the K closest data points
- Assign label by voting

The votes can be weighted by the inverse distance (weighted k-NN)

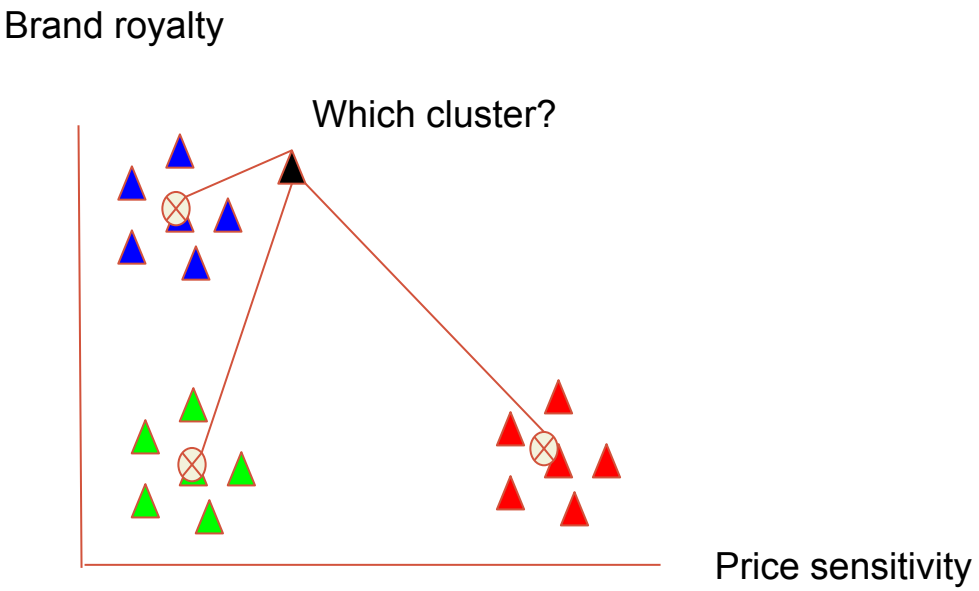


Centroids

Basically the representative of the cluster

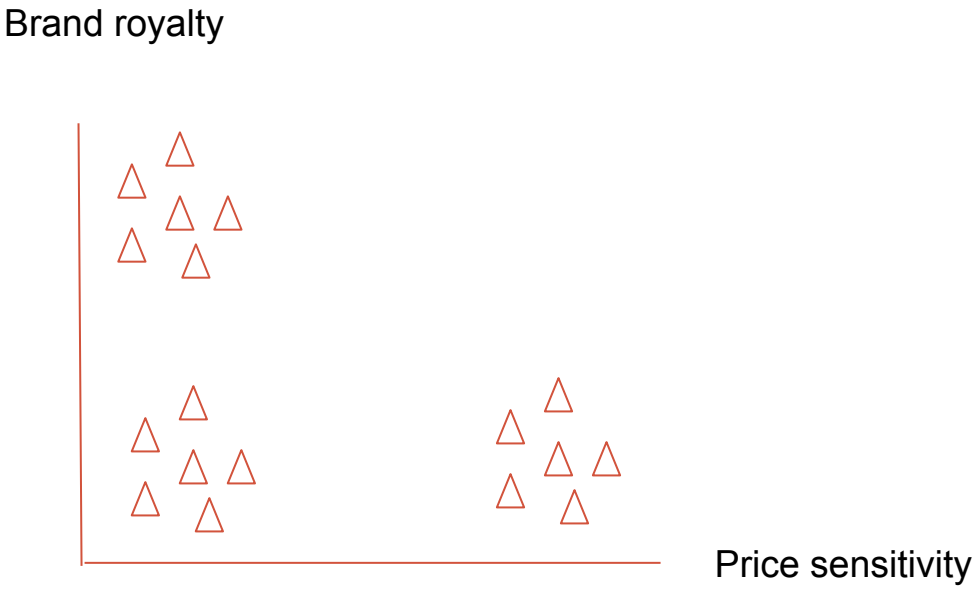
Find the mean location of the cluster by averaging

Can use mode or median depending on the data

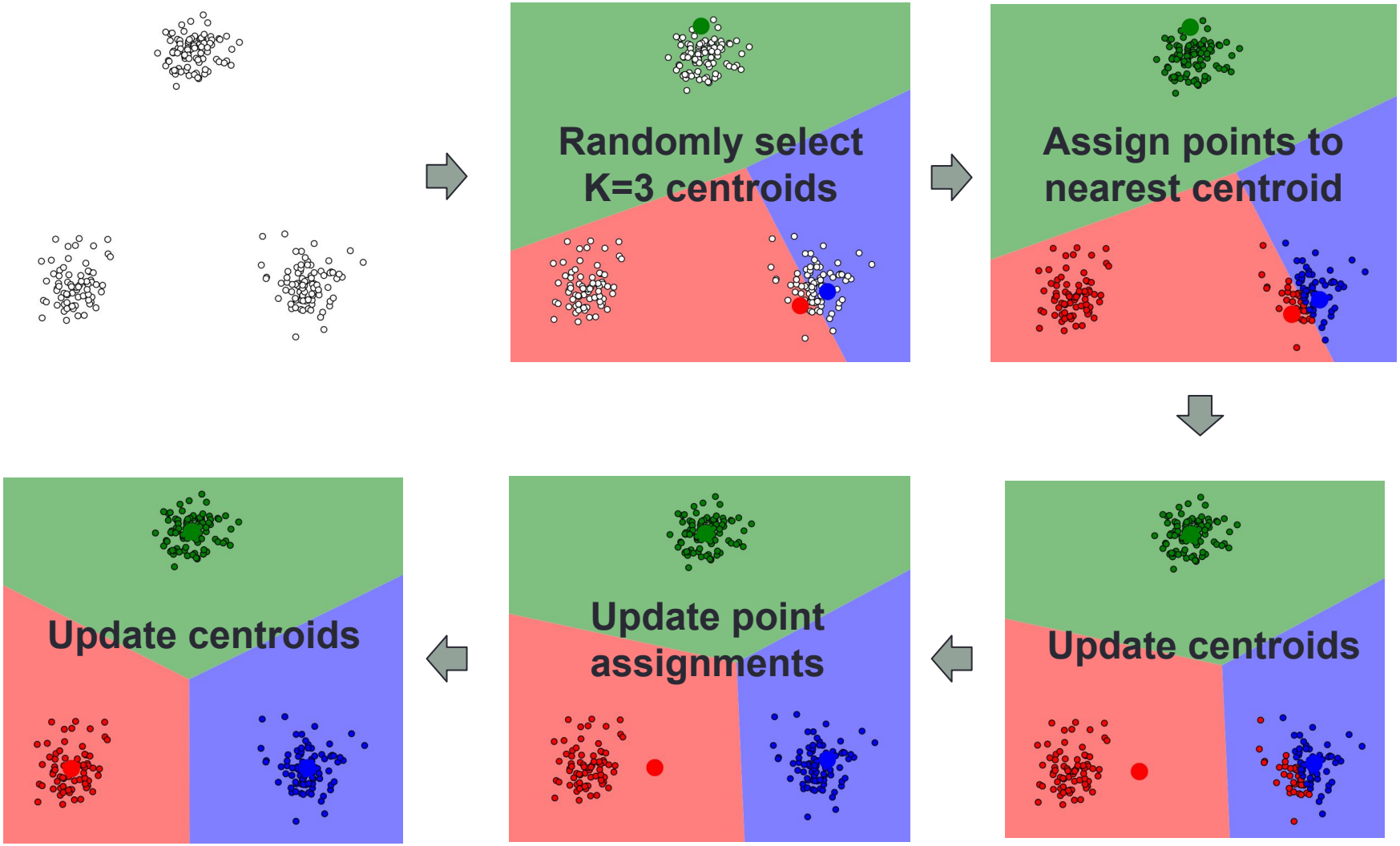


K-mean clustering

1. Randomly k centroids by picking from data points
2. Assign each data points to centroids
3. Update centroids for each cluster
4. Repeat 2-3 until centroids does not change



An Illustration Of K-Mean Clustering



From https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

Characteristics of K-means

- **The number of clusters, *K*, is specified in advance.**
- **Always converge to a (local) minimum.**
  - **Poor starting centroid locations can lead to incorrect minima.**

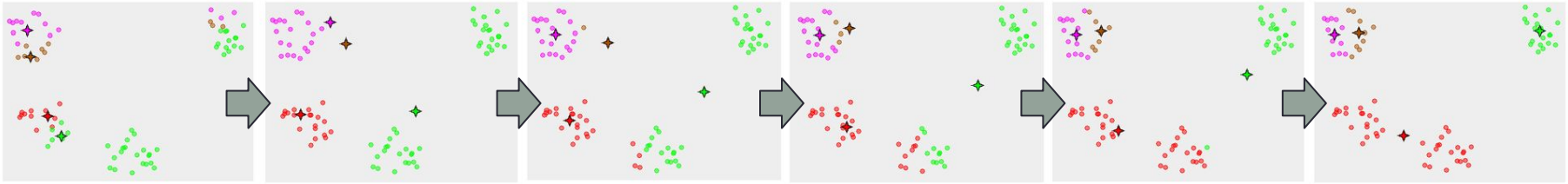
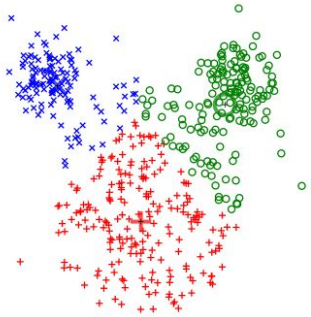
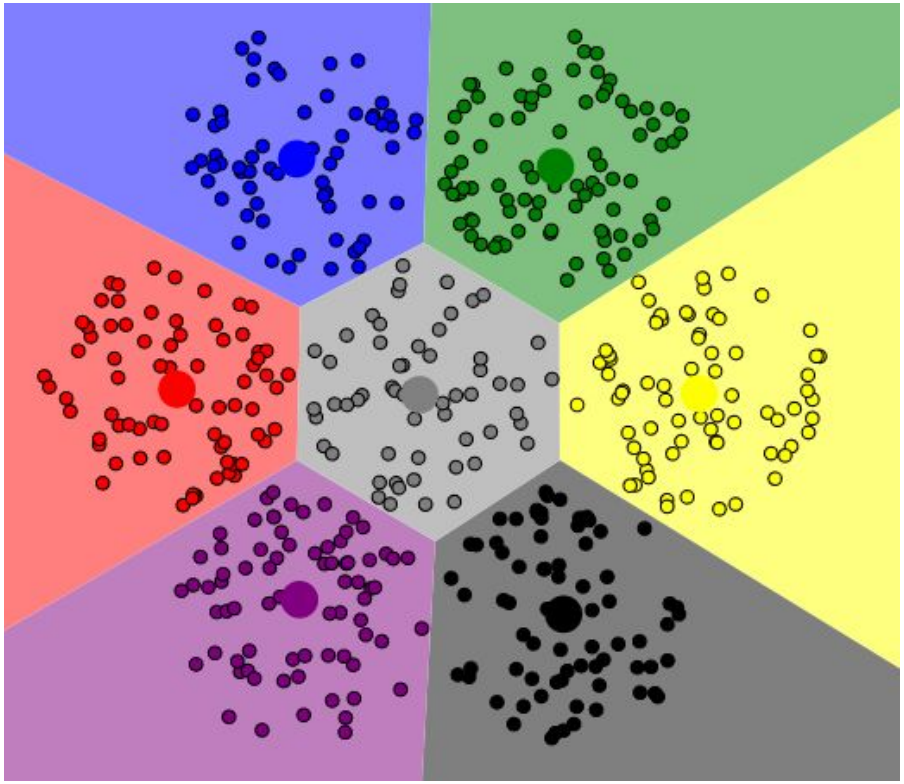


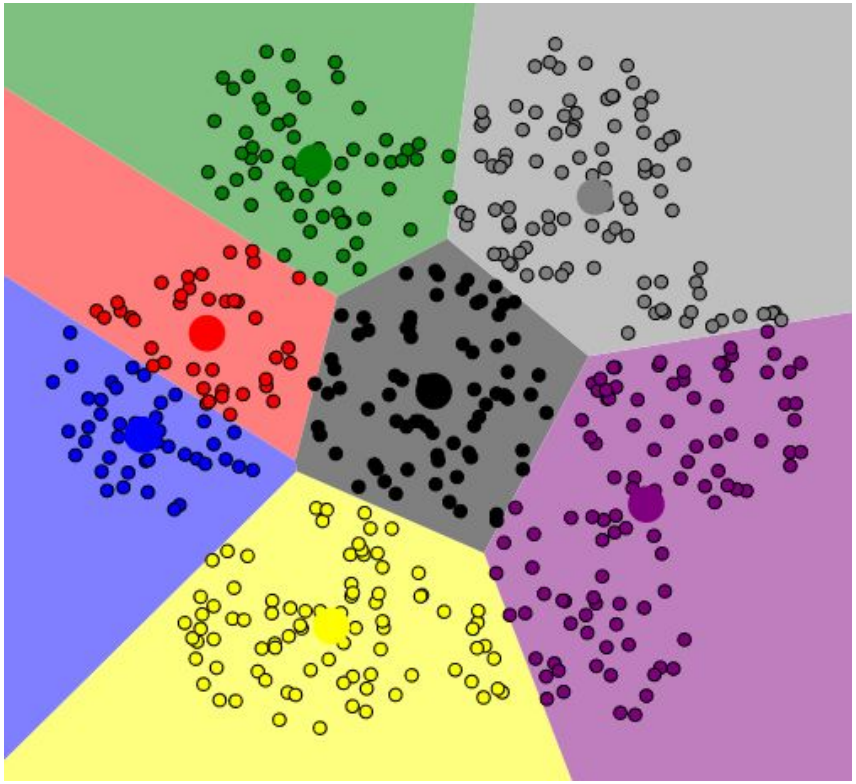
Image from https://en.wikipedia.org/wiki/K-means_clustering

- **The model has several implicit assumptions:**
  - **Data points scatter around cluster's centers.**
  - **Boundary between adjacent clusters is always halfway between the cluster centroids.**



Effect of bad initializations

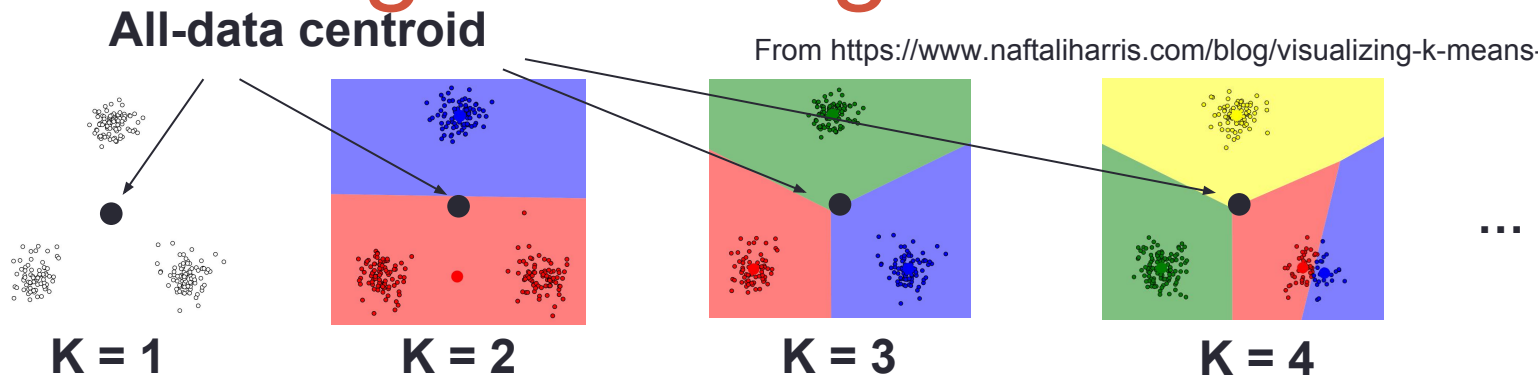




Solution, try different randomization and pick the best

From https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

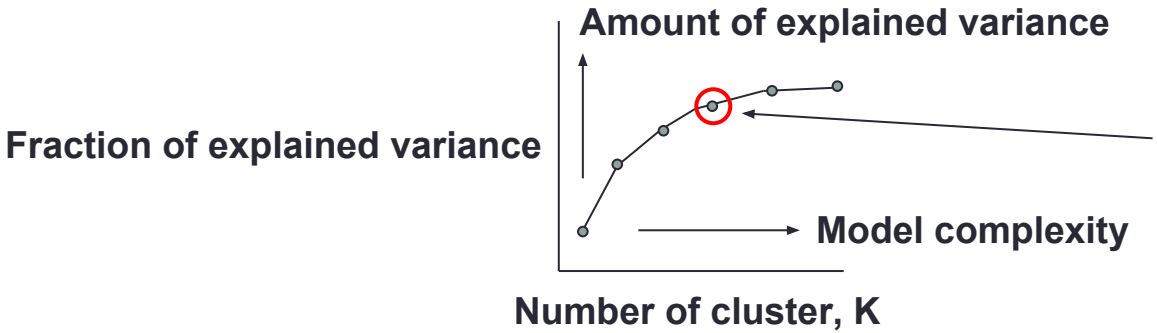
Selecting K - Using Elbow method



fraction of explained variance = $\frac{\text{between-cluster variance}}{\text{all-data variance}}$

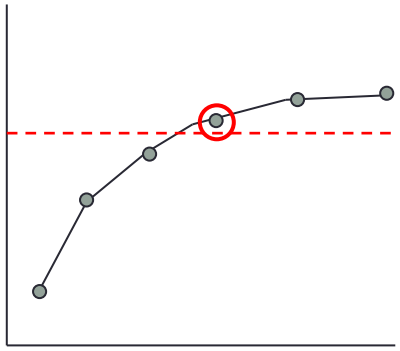
between-cluster variance = $\sum_{i=1}^{K} \frac{n_i(M_i - M)^2}{N-1}$, where n_i = size of i^{th} cluster,
M_i = centroid of i^{th} cluster, and
M = all-data centroid.

all-data variance = $\sum_{i=1}^{N} \frac{(x_i - M)^2}{N-1}$, where x_i = i^{th} data point and N = # of data.



The elbow method chooses K where increasing complexity doesn't yield much in return.

Selecting K - other methods



Choose minimal K that explains at least 95% of the all-data variance.

K = 2
K = 3
K = 4
⋮

→ Training K-mean Clustering Model → Testing / Cross-validation →

K	Accuracy
2	50%
3	68%
4	83%
⋮	⋮

Choose K that maximizes certain objective (e.g. accuracy on testing data)

Best method

Lab

Hyperparameter tuning

K-means

- Effect of initialization
- Effect of k
- Effect of features

Hyperparameter

Parameter - a variable in the model that the model automatically learns from data

Hyperparameter - a variable in the model that you set

How to set?

Use validation set

Three sets: training, test, and validation.