08 – DATA MINING

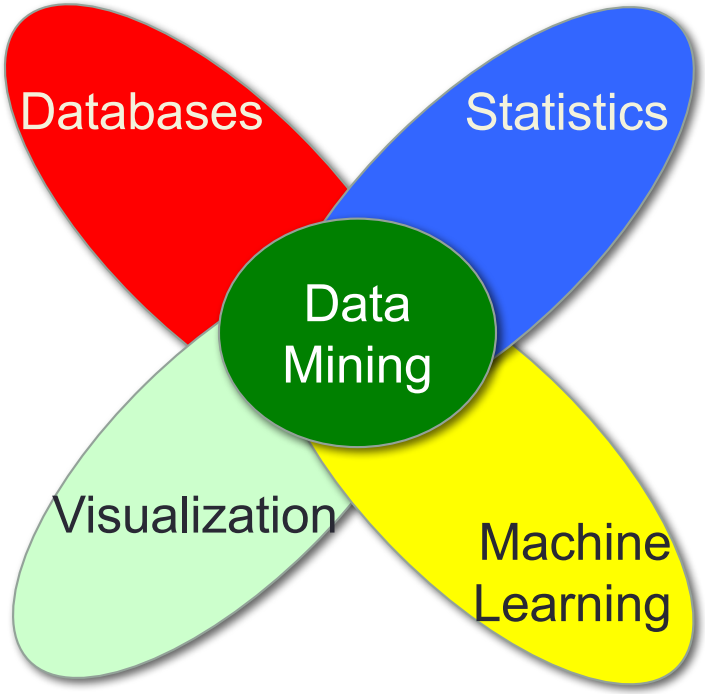
CS 0155 – Spring 2017

Data Witchcraft

Alexandros Labrinidis – http://labrinidis.cs.pitt.edu
University of Pittsburgh

Data Mining definition

- What is data mining?
  - Computational process to discover patterns in large data sets



Data Mining definition (cont)

Must produce novel and interesting patterns!

[Source: http://dilbert.com/strips/comic/1996-04-17/]

A bit of history

- Let us go back 160 years ago (1854) in London, England
  - On 31 August 1854, after several other outbreaks had occurred elsewhere in the city, a major outbreak of cholera struck Soho.
  - Over the next three days, 127 people on or near Broad Street died. In the next week, three quarters of the residents had fled the area.
  - By 10 September, 500 people had died and the mortality rate was 12.8 percent in some parts of the city. By the end of the outbreak, 616 people had died.

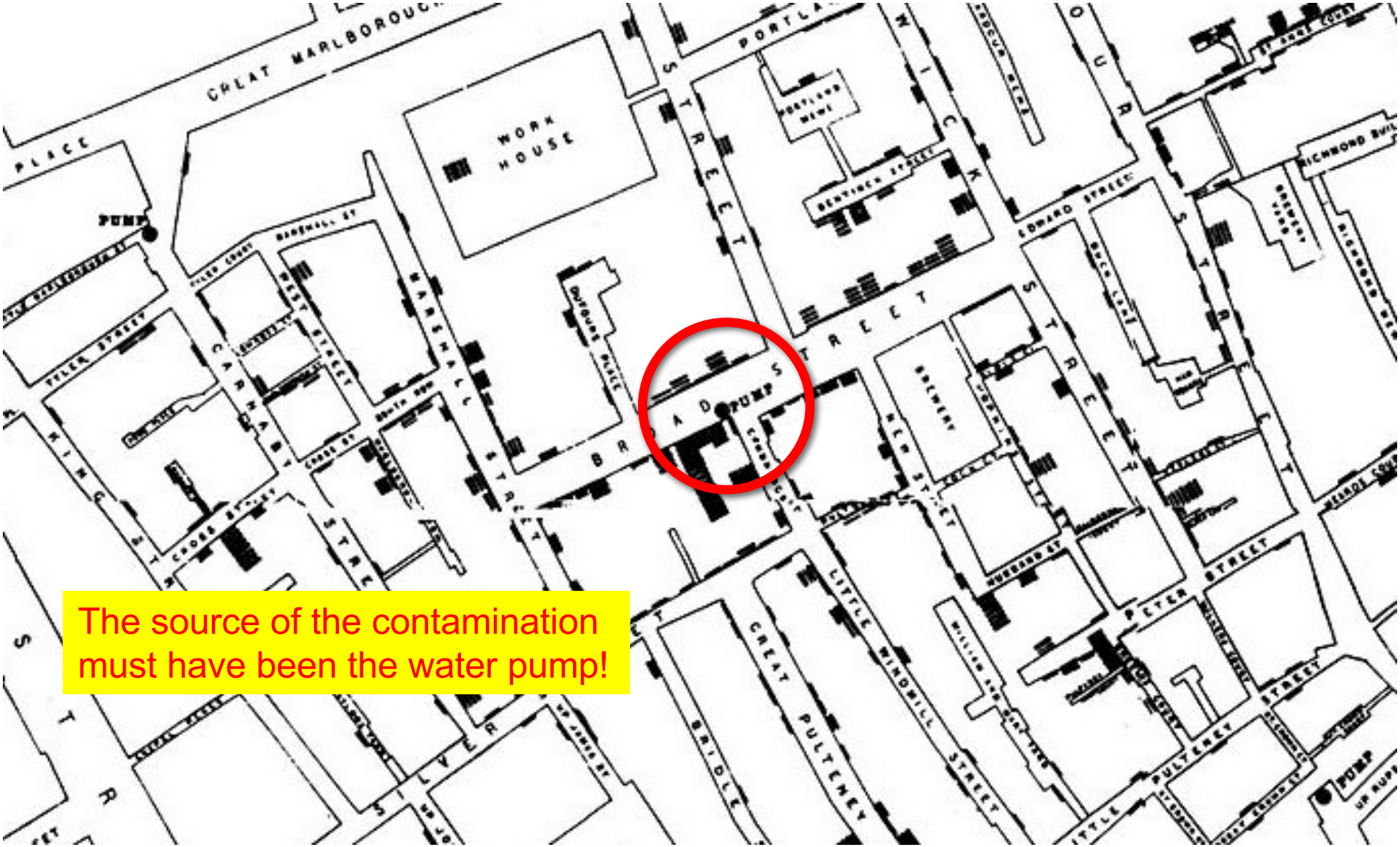
[Source: http://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak]

1854 Cholera Outbreak

- John Snow (a physician) wanted to investigate cause
  - Was skeptic of the *Miasma theory*, of “bad air”
  - Note that the germ theory was not created until 1861 by Louis Pasteur
  - Snow studied spread of disease
  - Made a map of fatalities

Map of Soho with fatalities

Map of Soho with fatalities (zoom)



In Snow's own words

On proceeding to the spot, I found that nearly all the deaths had taken place within a short distance of the [Broad Street] pump. There were only ten deaths in houses situated decidedly nearer to another street-pump. In five of these cases the families of the deceased persons informed me that they always sent to the pump in Broad Street, as they preferred the water to that of the pumps which were nearer. In three other cases, the deceased were children who went to school near the pump in Broad Street...

With regard to the deaths occurring in the locality belonging to the pump, there were 61 instances in which I was informed that the deceased persons used to drink the pump water from Broad Street, either constantly or occasionally...

The result of the inquiry, then, is, that there has been no particular outbreak or prevalence of cholera in this part of London except among the persons who were in the habit of drinking the water of the above-mentioned pump well.

I had an interview with the Board of Guardians of St James's parish, on the evening of the 7th inst [September 7], and represented the above circumstances to them. In consequence of what I said, the handle of the pump was removed on the following day.

—John Snow, *letter to the editor of the Medical Times and Gazette*

socrative by MasteryConnect Data Collection Question

- **Question**: Pick a random number from 1 to 10

- **Possible Answers**:
  - 1
  - 2
  - 3
  - 4
  - 5
  - 6
  - 7
  - 8
  - 9
  - 10

DATA MINING TASKS

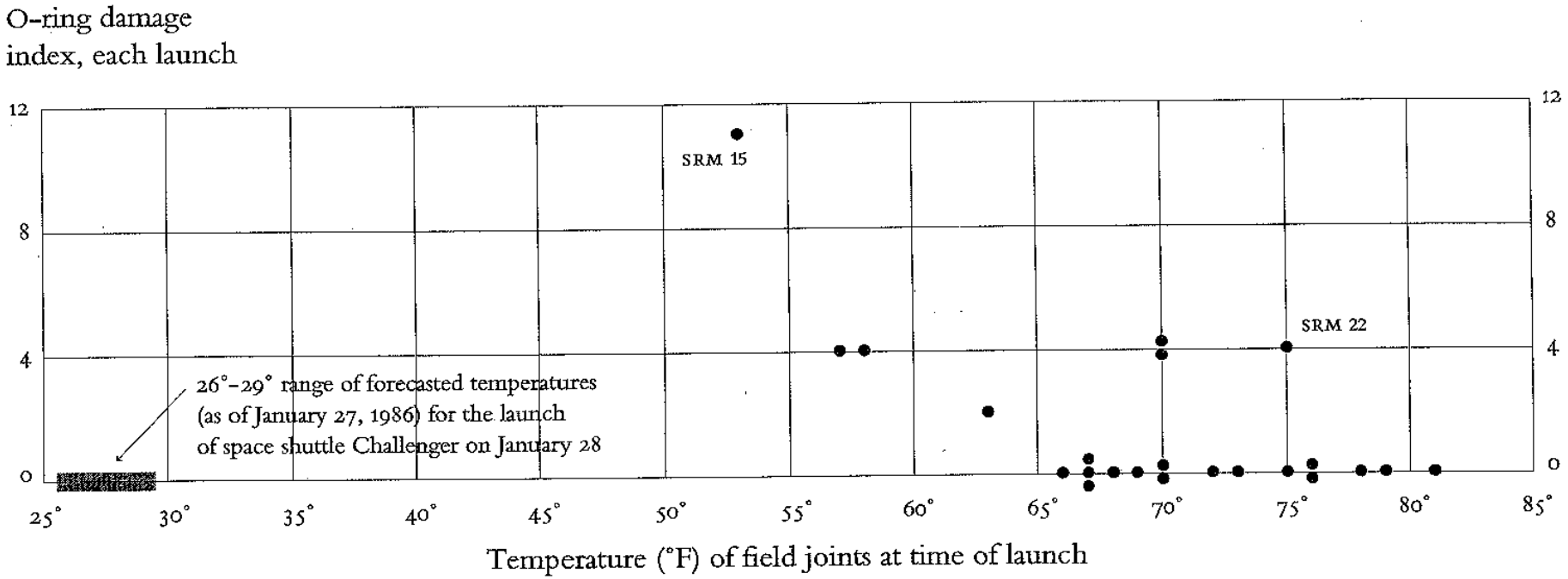
Typical Data Mining Tasks

- Anomaly Detection
- Association Rule Learning
- Clustering
- Classification
- Regression
- Summarization

Source: http://en.wikipedia.org/wiki/Data_mining

Anomaly Detection

- Anomaly Detection (Outlier/change/deviation detection)
  The identification of unusual data records, that might be interesting or data errors that require further investigation.



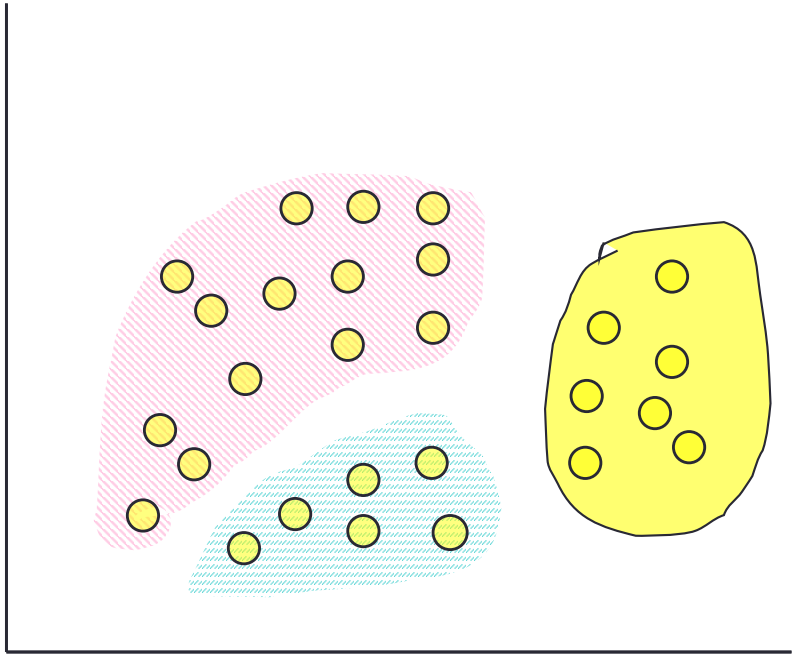
Source: Visual and Statistical Thinking: Displays of Evidence for Making Decisions, Edward Tufte, 1997

Association Rule Learning

- Association rule learning (Dependency modeling)
  Searches for relationships between variables.
  - For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes.
  - This is sometimes referred to as market basket analysis.

Clustering

- Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
  - Members of a cluster should be more “alike” among each other, than to members of other clusters
  - Clustering is one type of **unsupervised learning**



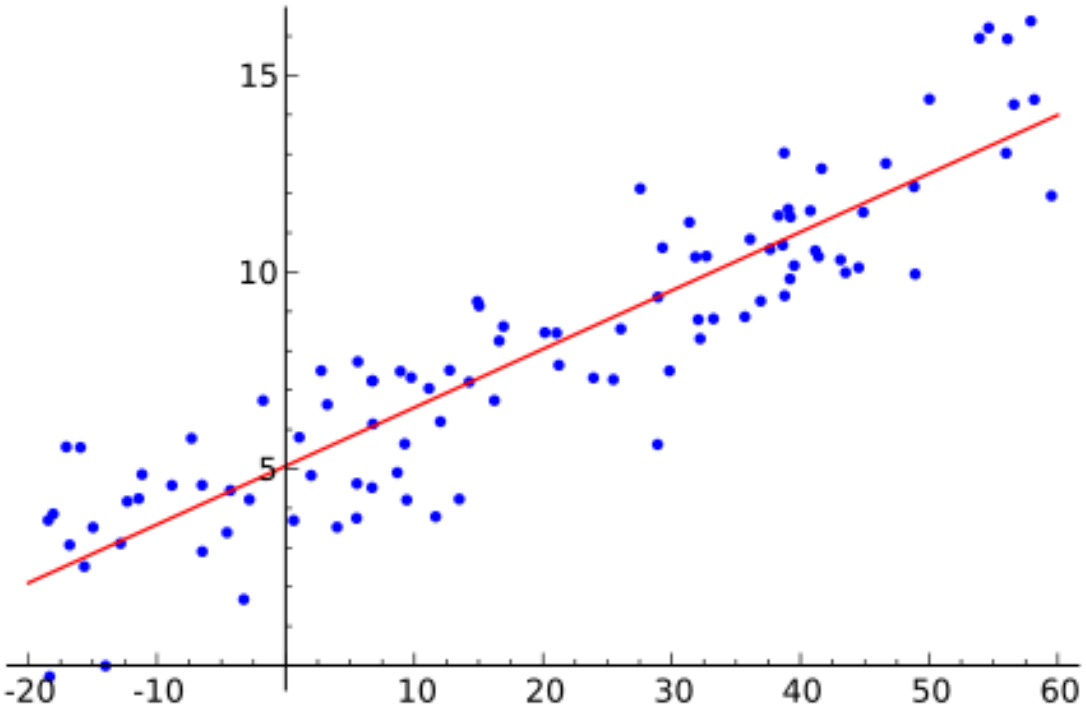
Classification

- Classification – is the task of generalizing known structure to apply to new data.
  - In other words: learn a method to predict a label for new data from pre-labeled (classified) data
  - Classification is one type of **supervised learning**

- Examples:
  - an e-mail program classifies an e-mail as "legitimate" or as "spam".
  - a lender classifies its customers as credit-worthy or credit-risky.
  - a credit card company identifies fraudulent transactions.
  - a phone company identifies which customer would abandon contract for another carrier.
  - a security agency identifies potential *evil-doers*.
  - personalized medicine – will drug work for specific patient?

Regression

- Regression – attempts to find a function which models the data with the least error.



Source: http://en.wikipedia.org/wiki/Regression_analysis

Summarization

- Summarization –
  providing a more compact representation of the data set, including **visualization** and report generation.

- Examples:
  - Document summarization (e.g., snippets in Gmail)
  - Choosing one representative member from each cluster
  - Choosing a few representative data points through *sampling*
- Question:
  - What would be a trivial (but simple) summarization of a set of numbers?
- Answer:
  - computing the average or the median

socrative by MasteryConnect Understanding Question

- **Question**:
  - Which of the following CANNOT be a summarization operation for a data set consisting of just numbers?

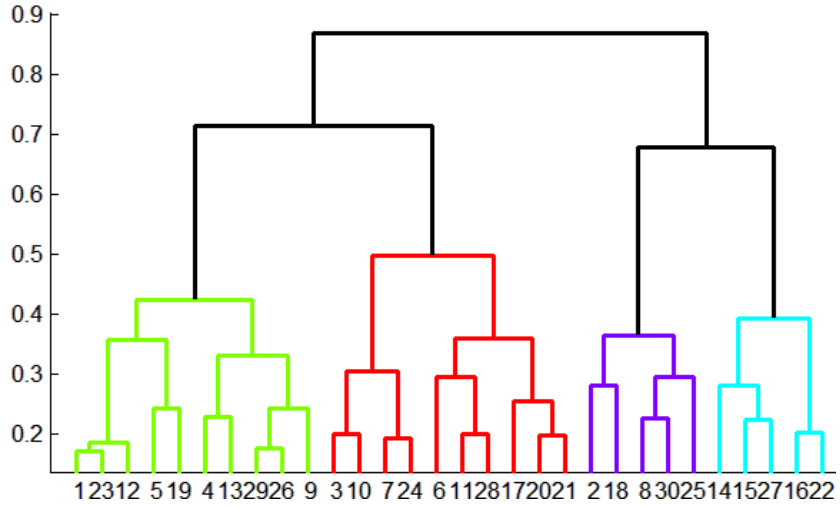
- **Possible Answers**:
  - Computing the average
  - Computing the maximum
  - Computing the median
  - Computing the minimum
  - Multiplying all numbers by 5

CLUSTERING

Source: http://www.mmds.org

Overview: Methods of Clustering

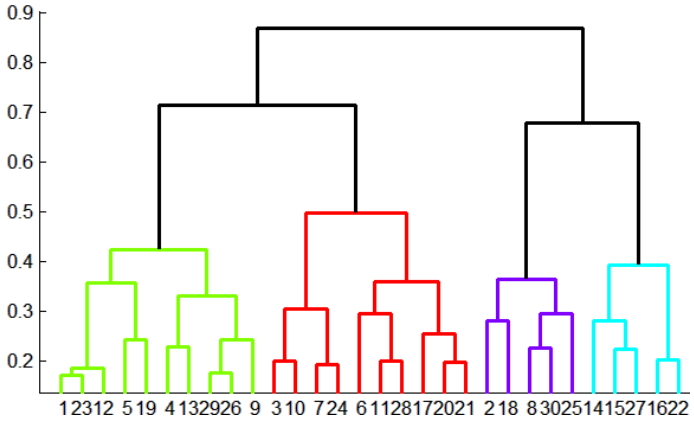
- **Hierarchical:**
  - **Agglomerative** (bottom up):
    - Initially, each point is a cluster
    - Repeatedly combine the two "nearest" clusters into one
  - **Divisive** (top down):
    - Start with one cluster and recursively split it



- **Point assignment:**
  - Maintain a set of clusters
  - Points belong to "nearest" cluster

Source: http://www.mmds.org

Hierarchical Clustering



- **Key operation:**
  Repeatedly combine two nearest clusters

- **Three important questions:**
  1) How do you represent a cluster of more than one point?
  2) How do you determine the "nearness" of clusters?
  3) When to stop combining clusters?

Source: http://www.mmds.org

Hierarchical Clustering

Key operation: Repeatedly combine two nearest clusters

(1) How to represent a cluster of many points?

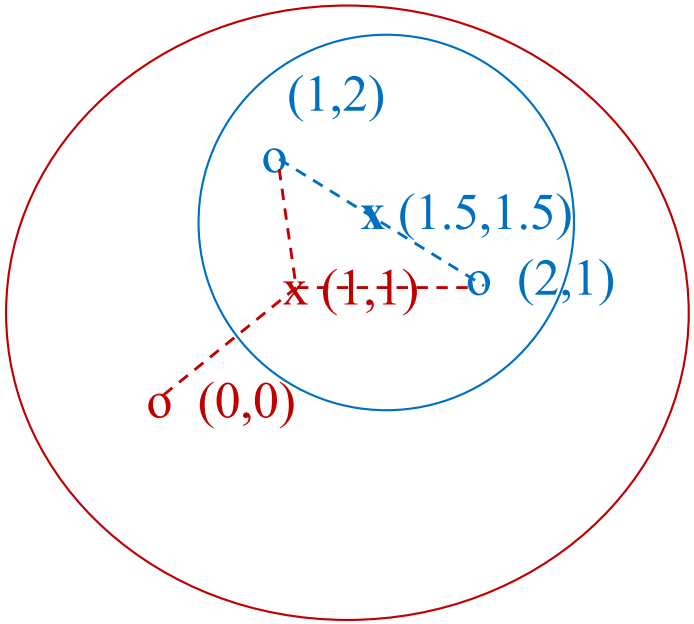
- **Key problem:** As you merge clusters, how do you represent the "location" of each cluster, to tell which pair of clusters is closest?
- **Euclidean case:** each cluster has a ***centroid*** = average of its (data)points

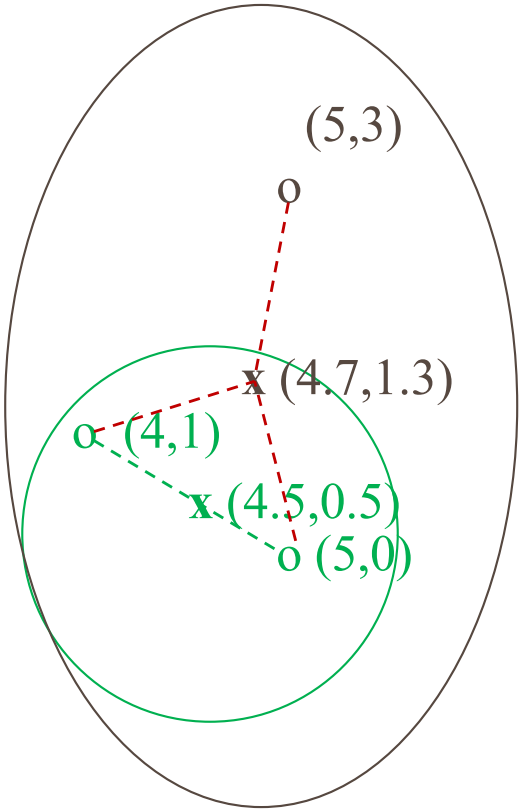
(2) How to determine "nearness" of clusters?

- Measure cluster distances by distances of centroids

Source: http://www.mmds.org

Example: Hierarchical clustering

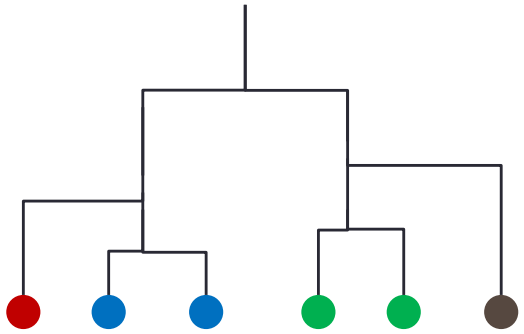




Data:

o … data point

x … centroid



Dendrogram

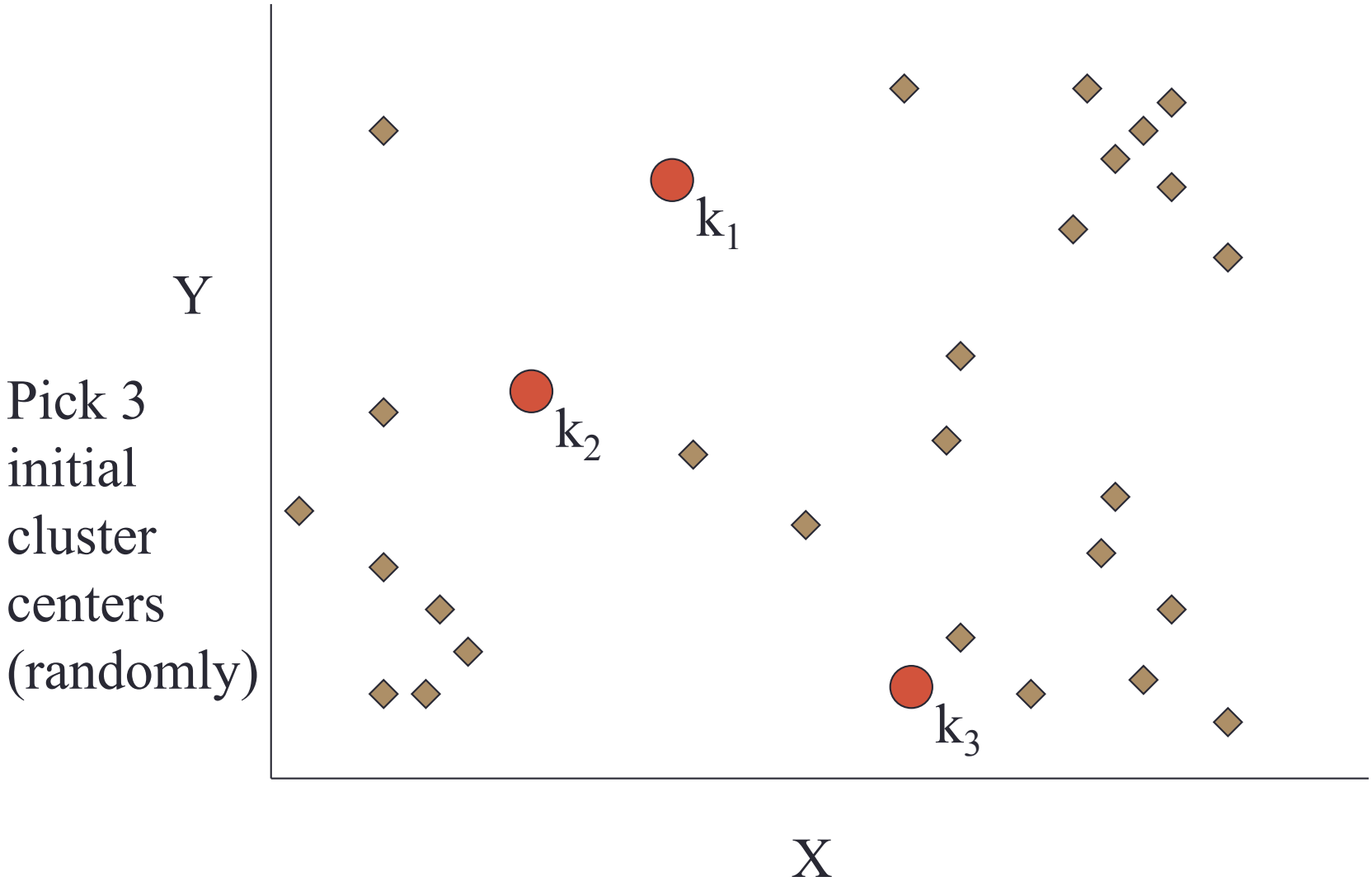
K-MEANS CLUSTERING

Simple Clustering: K-means

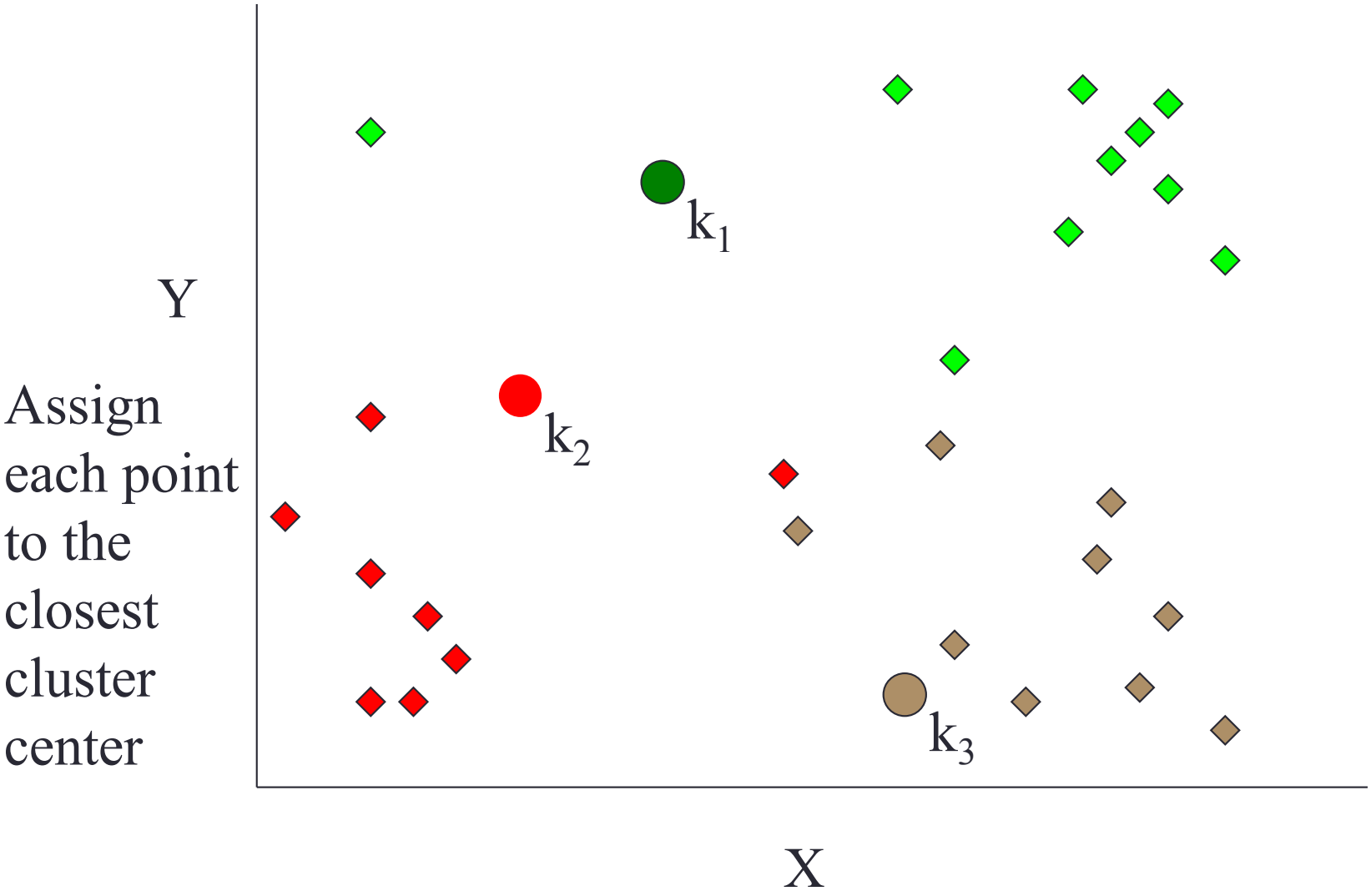
Works with numeric data only

1) Specify K, i.e., how many clusters to generate
2) Pick K cluster centers (at random)
3) Assign every item to its nearest cluster center (e.g. using Euclidean distance)
4) Move each cluster center to the mean of its assigned items
5) Repeat steps 3, 4 until convergence (change in cluster assignments less than a threshold)

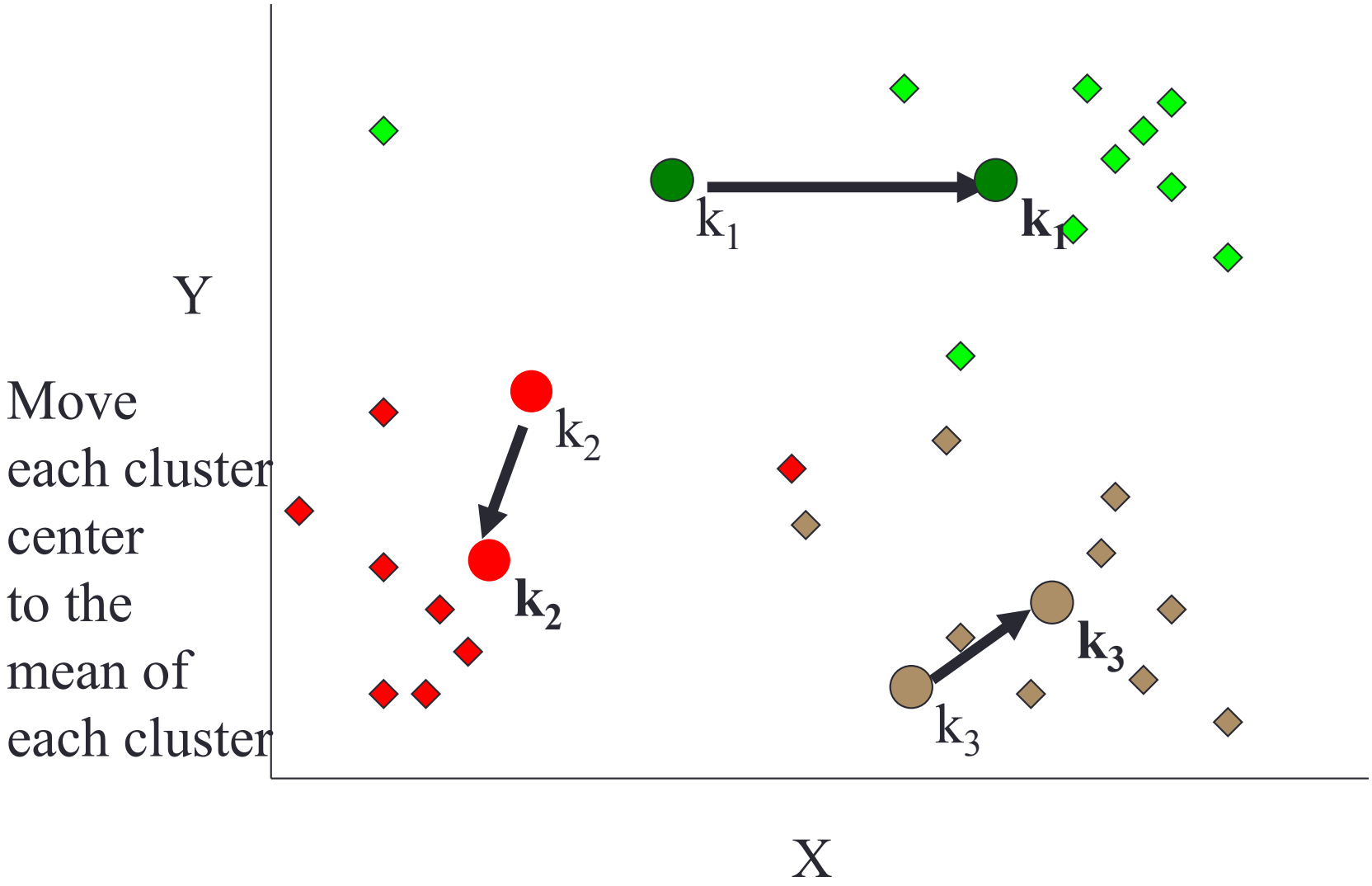
K-means example, step 1



K-means example, step 2



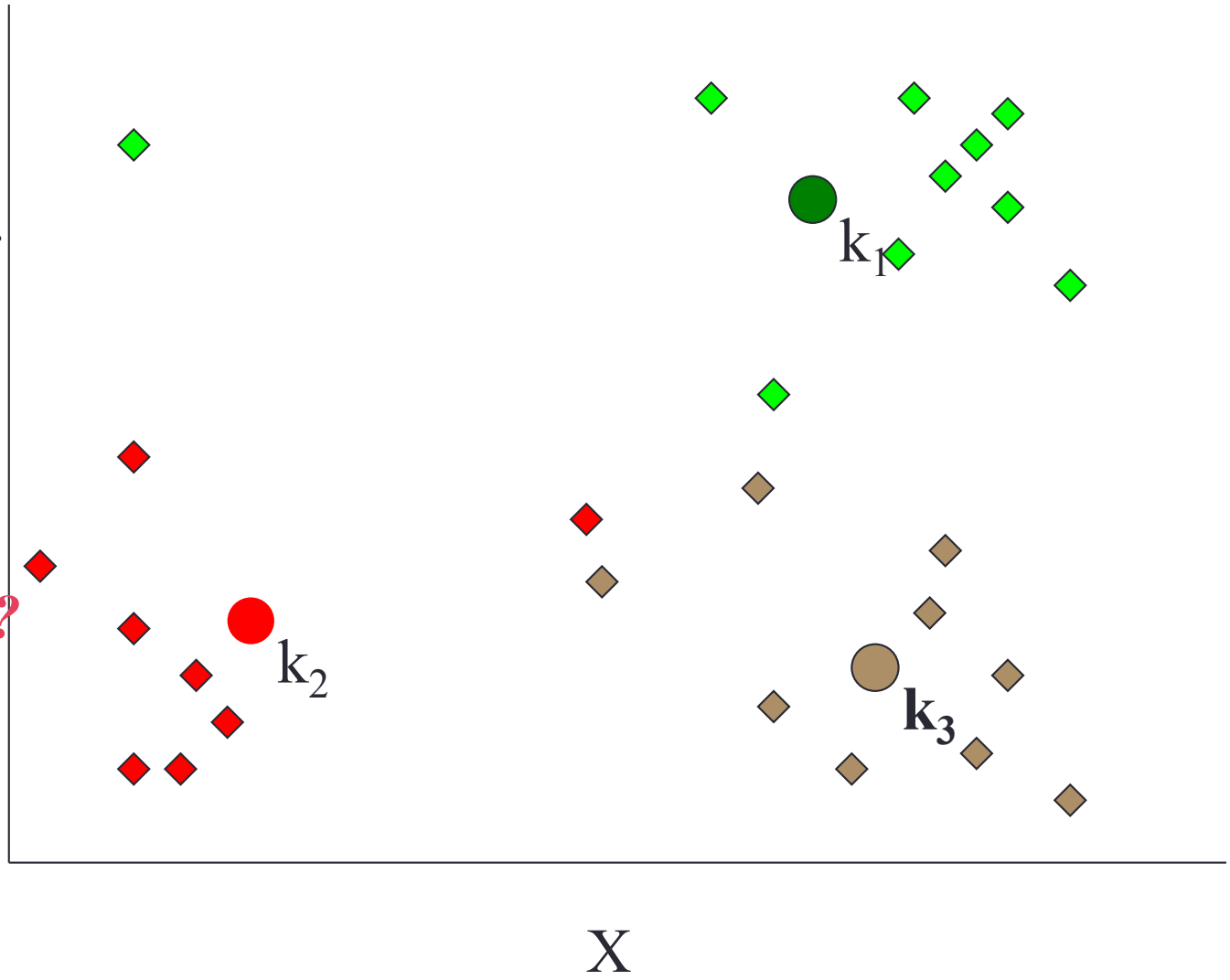
K-means example, step 3



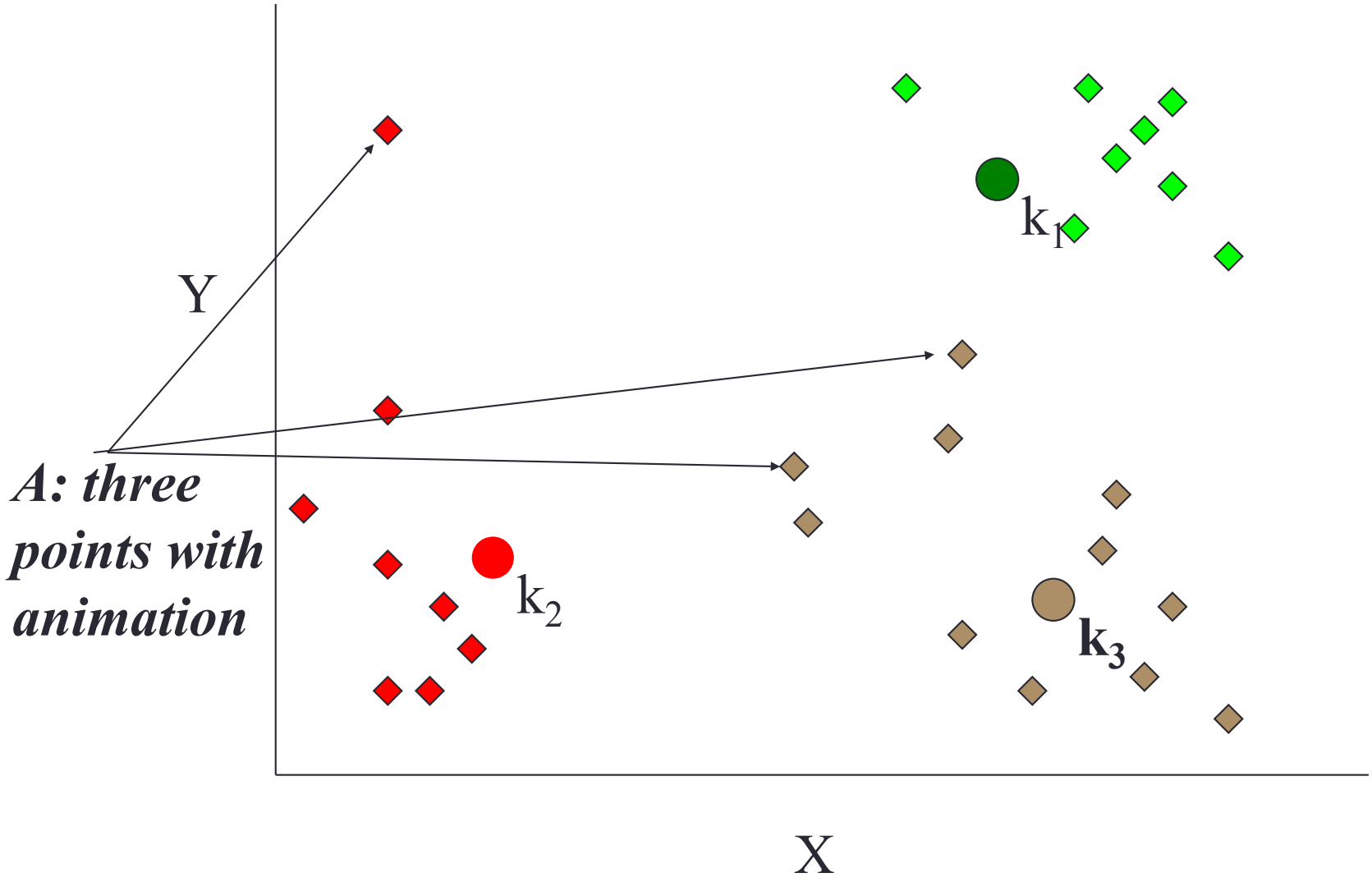
K-means example, step 4

Reassign points closest to a different new cluster center

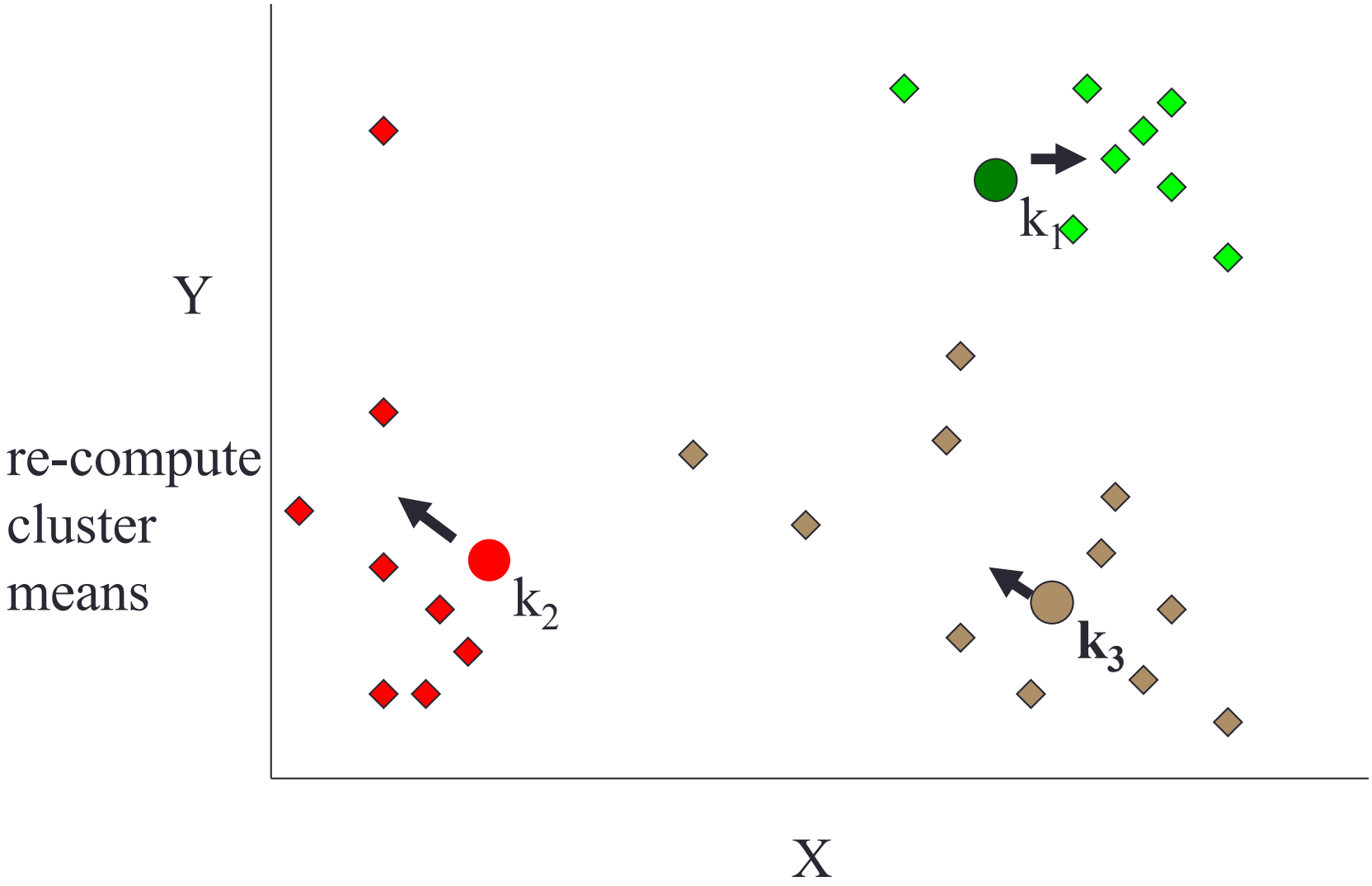
Q: Which points are reassigned?



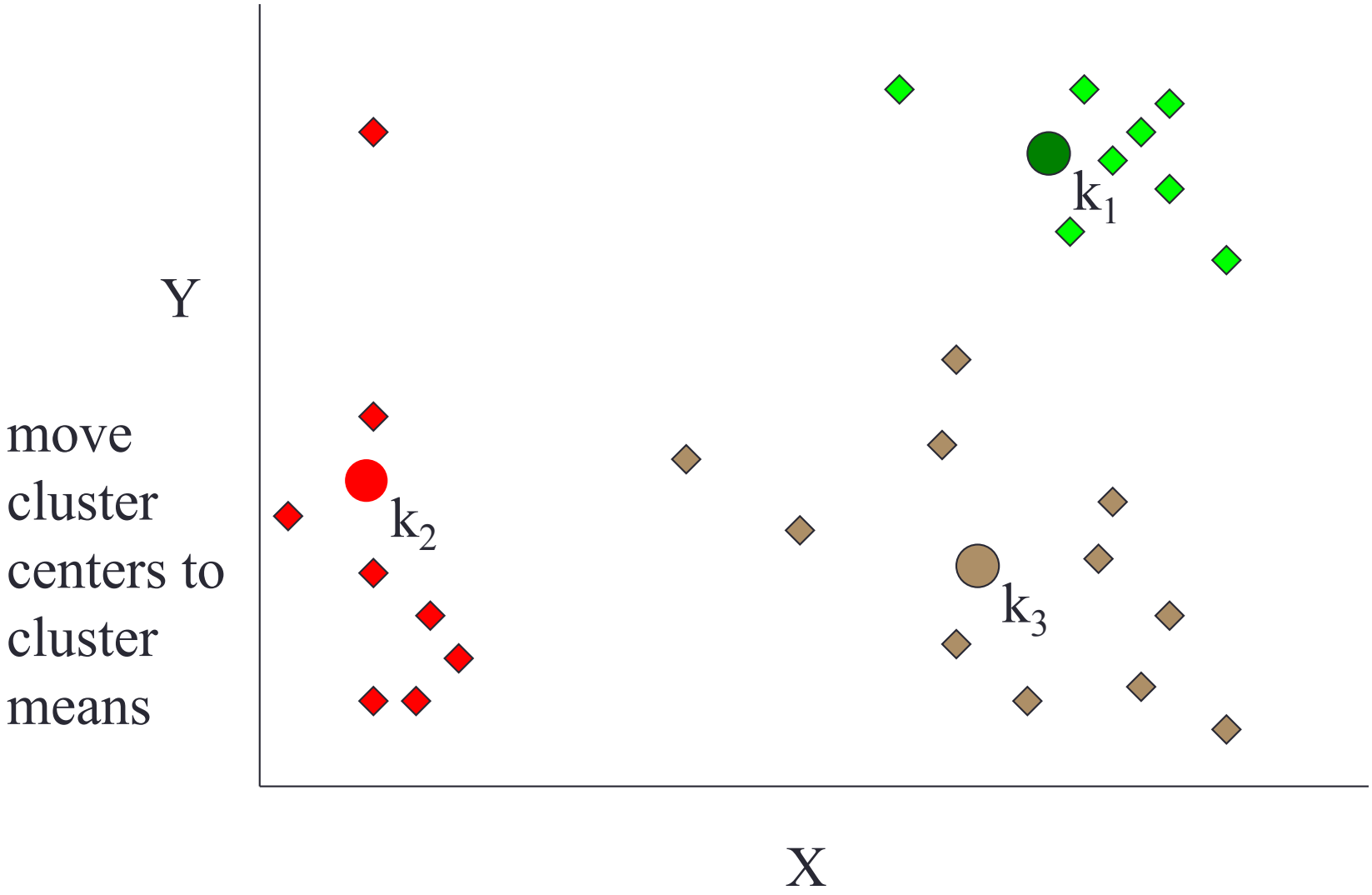
K-means example, step 4 …



K-means example, step 4b

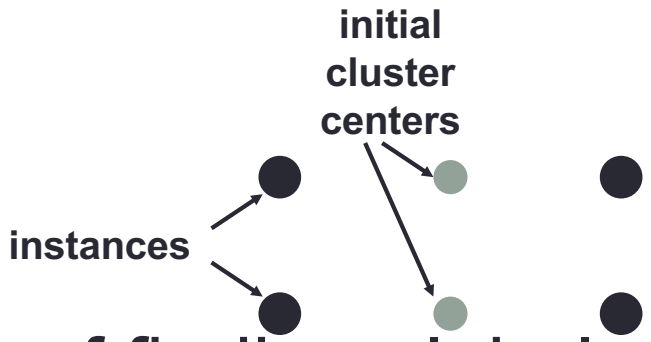


K-means example, step 5



Discussion

- Result can vary significantly depending on initial choice of seeds
- Can get trapped in local minimum
  - Example:



- To increase chance of finding global optimum: restart with different random seeds

K-Means Visualizations

- http://stanford.edu/class/ee103/visualizations/kmeans/kmeans.html
- Visualizing K-Means algorithm with D3.js http://tech.nitoyon.com/en/blog/2013/11/07/k-means/
- http://www.naftaliharris.com/blog/visualizing-k-means-clustering/
- http://www.bytemuse.com/post/k-means-clustering-visualization/

K-means clustering summary

Advantages

- Simple
- Understandable
- Items automatically assigned to clusters

Disadvantages

- Must pick number of clusters before hand
- All items forced into clusters
- Too sensitive to outliers

K-means variations

- **K-medoids** – instead of mean, use medians of each cluster
  - Mean of 1, 3, 5, 7, 9 is **5**
  - Mean of 1, 3, 5, 7, 1009 is **205**
  - Median of 1, 3, 5, 7, 1009 is **5**
  - Median **advantage**: not affected by extreme values
- For large databases, use sampling

socrative by MasteryConnect Understanding Question

- **Question**:
  - What does the K, in the K-means clustering algorithm, stand for?

- **Possible Answers**:
  - K comes from the name of the inventor of the algorithm
  - K sounds like C from Cluster
  - K is the exact number of clusters to generate
  - K is the minimum number of clusters to generate (actual number could be higher)
  - K is the maximum number of clusters to generate (actual number could be lower)

DBSCAN ALGORITHM

DBSCAN Algorithm

https://en.wikipedia.org/wiki/DBSCAN

- **DBSCAN** requires two parameters:

1. ε (eps) -- used to define ε-neighborhoods, and
2. the minimum number of points required to form a dense region (minPts).

- Start with a not-yet-visited **arbitrary point**; retrieve its ε-neighborhood,
  - If ε-neighborhood contains sufficiently many points, a cluster is started.
  - Otherwise, the point is labeled as noise.
- If a point is found to be a dense part of a cluster, its ε-neighborhood is also part of that cluster. Hence, all points that are found within the ε-neighborhood are added, as is their own ε-neighborhood when they are also dense.
- Process continues until the density-connected cluster is completely found.
- Then, a **new unvisited point** is retrieved and processed, leading to the discovery of a further cluster or noise.

DBSCAN Visualization

- http://www.naftaliharris.com/blog/visualizing-dbscan-clustering/

socrative by MasteryConnect Understanding Questions

- **Questions**:
- (Q3) **How many clusters** will DBSCAN create, if **e=2** (radius) and **min_points = 3**?
- (Q4) How many points will DBSCAN label as **noise**, if **e=2** (radius) and **min_points = 3**?
- (Q5) **How many clusters** will DBSCAN create, if **e=2** (radius) and **min_points = 5**?
- (Q6) How many points will DBSCAN label as **noise**, if **e=2** (radius) and **min_points = 5**?
- (Q7) **How many clusters** will DBSCAN create, if **e=5** (radius) and **min_points = 2**?