Artificial Intelligence

Machine Learning (2)

Unsupervised Learning

Nacim Ihaddadene
Junia ISEN / M1 / 2024-2025

Machine learning problems

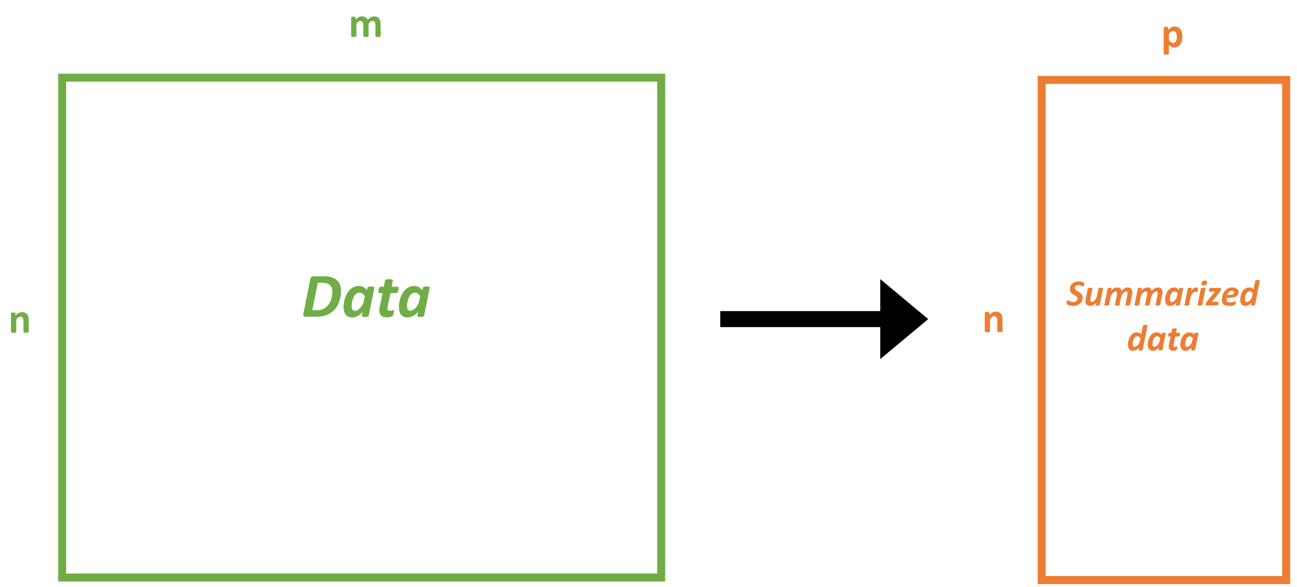
	Supervised learning	Unsupervised learning
Discrete data	**Classification**	**Clustering**
Continuous data	**Regression**	**Dimensionality reduction**

Dimensionality Reduction

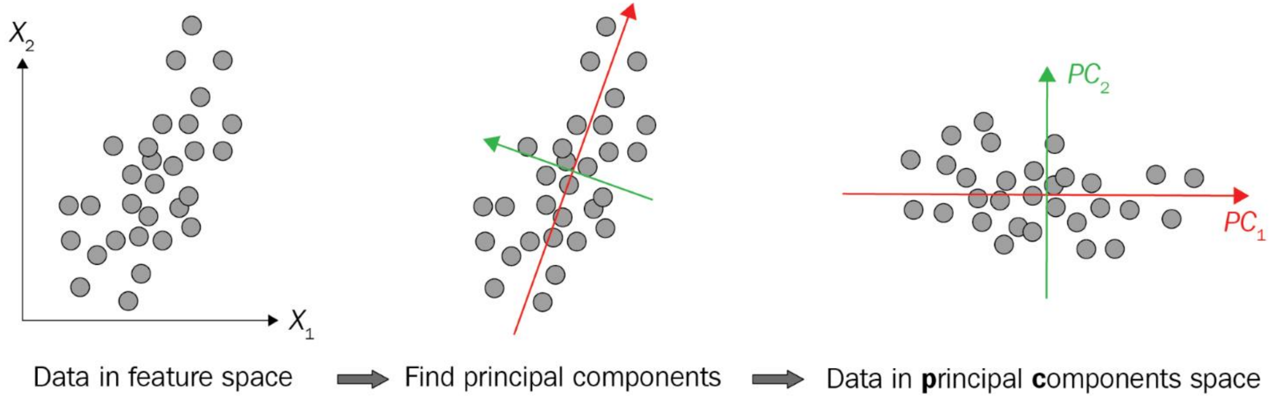
- summarization of data (n examples) with many dimensions (m attributes) by a smaller set of (p) derived (synthetic, composite) dimension

Why ?

- Data compression (with less loss of information)
- Structure Discovery
- Reducing training time and cost
- Effective visualization

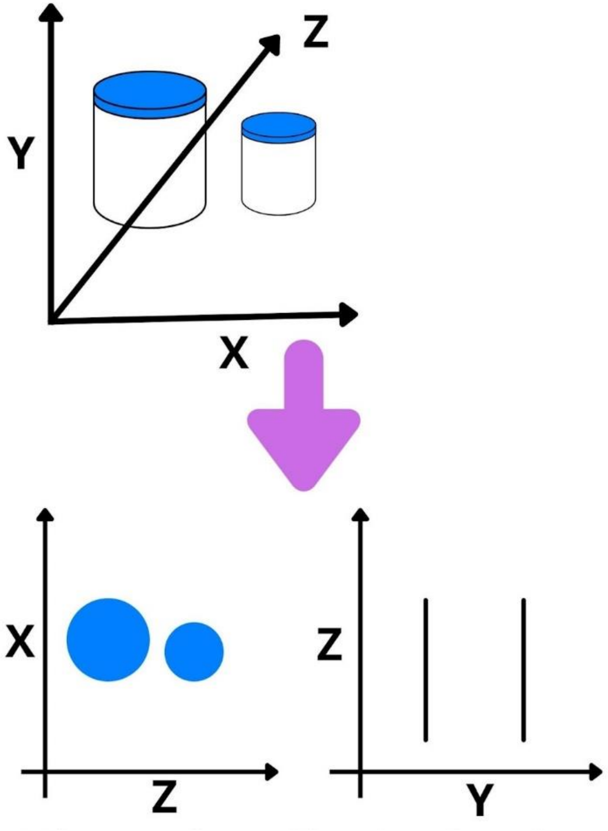


Dimensionality Reduction



Data in feature space → Find principal components → Data in **p**rincipal **c**omponents space

EXAMPLE 1



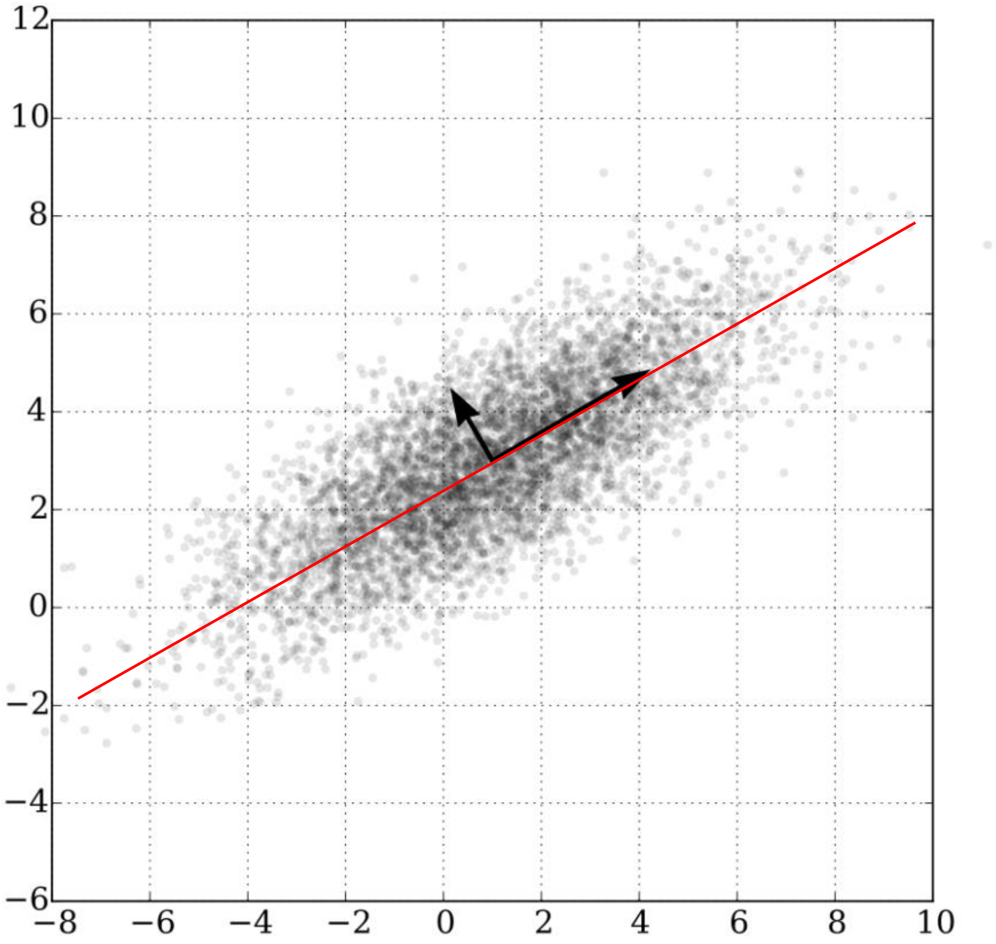
EXAMPLE 2

Dimensionality Reduction

- Multiple methods : **PCA,** ICA, LLE, Isomap, …
- Manifold Learning

Example :

https://scikit-learn.org/stable/auto_examples/manifold/plot_lle_digits.html



Tutorial

Apply PCA algorithm to IRIS dataset…

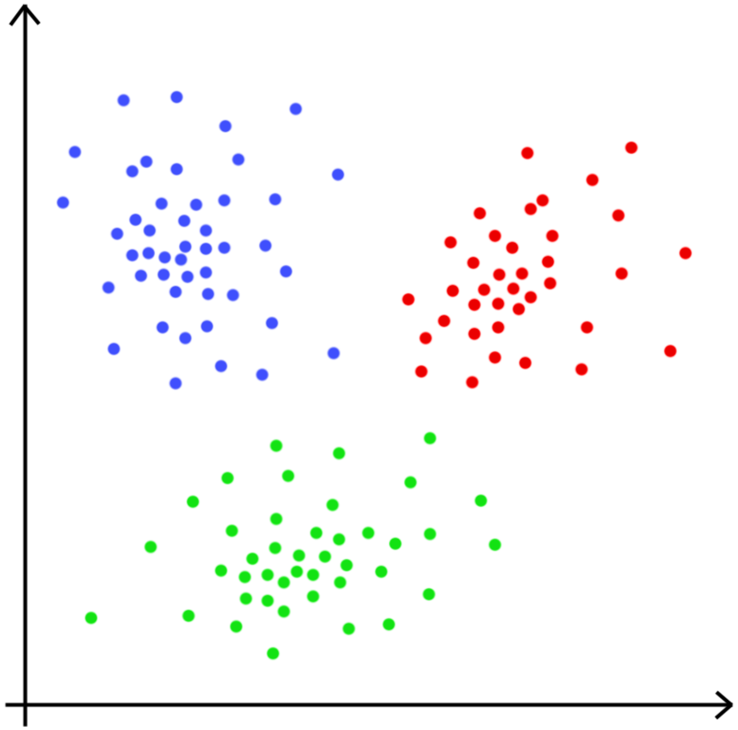
Machine learning problems

	Supervised learning	Unsupervised learning
Discrete data	**Classification**	**Clustering**
Continuous data	**Regression**	**Dimensionality reduction**

Intelligence is also the ability to recognize similar objects and group them!

The Problem of Clustering

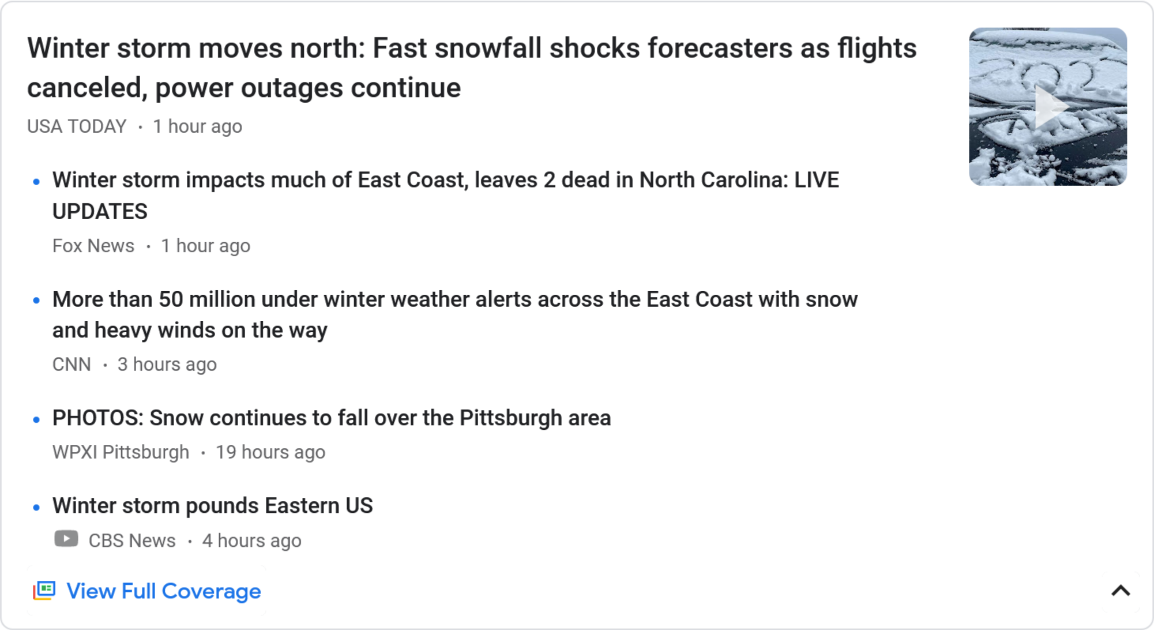
Given a set of unlabeled items (in n-dimensions), with a notion of distance between items, group the points into some number of clusters, so that members of a cluster are in some sense as nearby as possible.



Data without labels

Labeled Data

Example : Clustering news



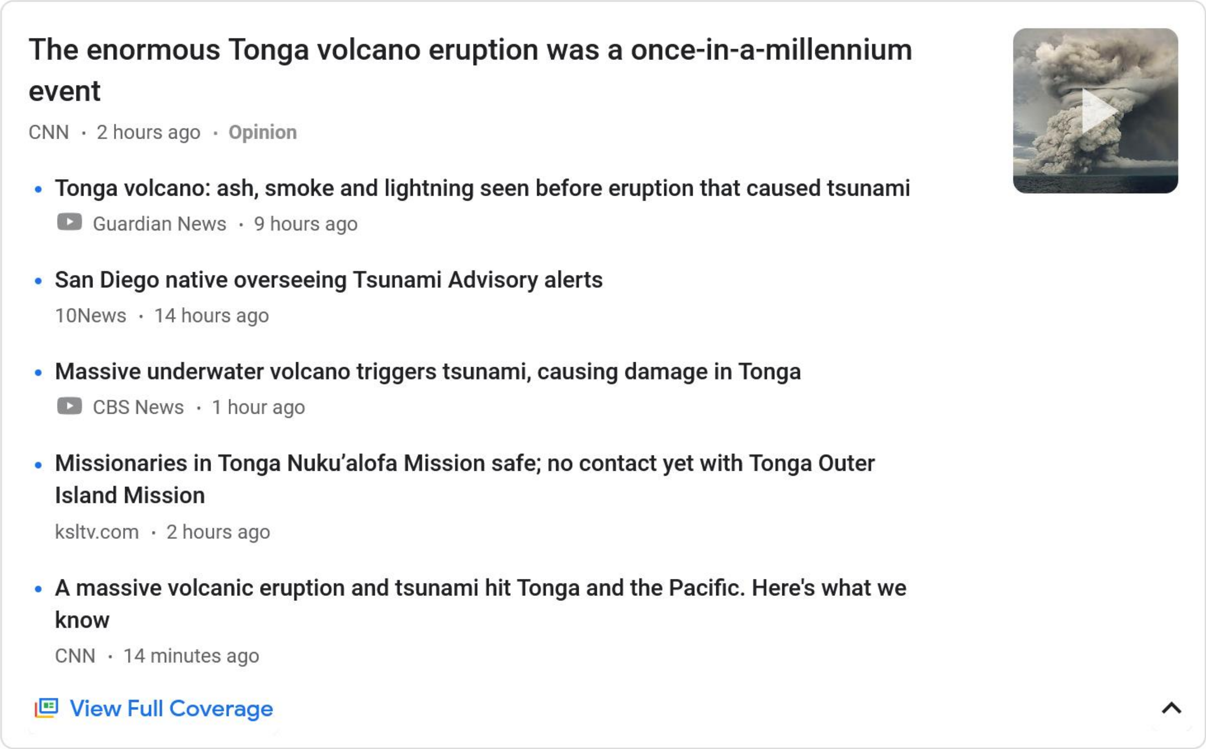
Winter storm moves north: Fast snowfall shocks forecasters as flights canceled, power outages continue

USA TODAY · 1 hour ago

- Winter storm impacts much of East Coast, leaves 2 dead in North Carolina: LIVE UPDATES
  Fox News · 1 hour ago
- More than 50 million under winter weather alerts across the East Coast with snow and heavy winds on the way
  CNN · 3 hours ago
- PHOTOS: Snow continues to fall over the Pittsburgh area
  WPXI Pittsburgh · 19 hours ago
- Winter storm pounds Eastern US
  CBS News · 4 hours ago

View Full Coverage

Winter Storm



The enormous Tonga volcano eruption was a once-in-a-millennium event

CNN · 2 hours ago · Opinion

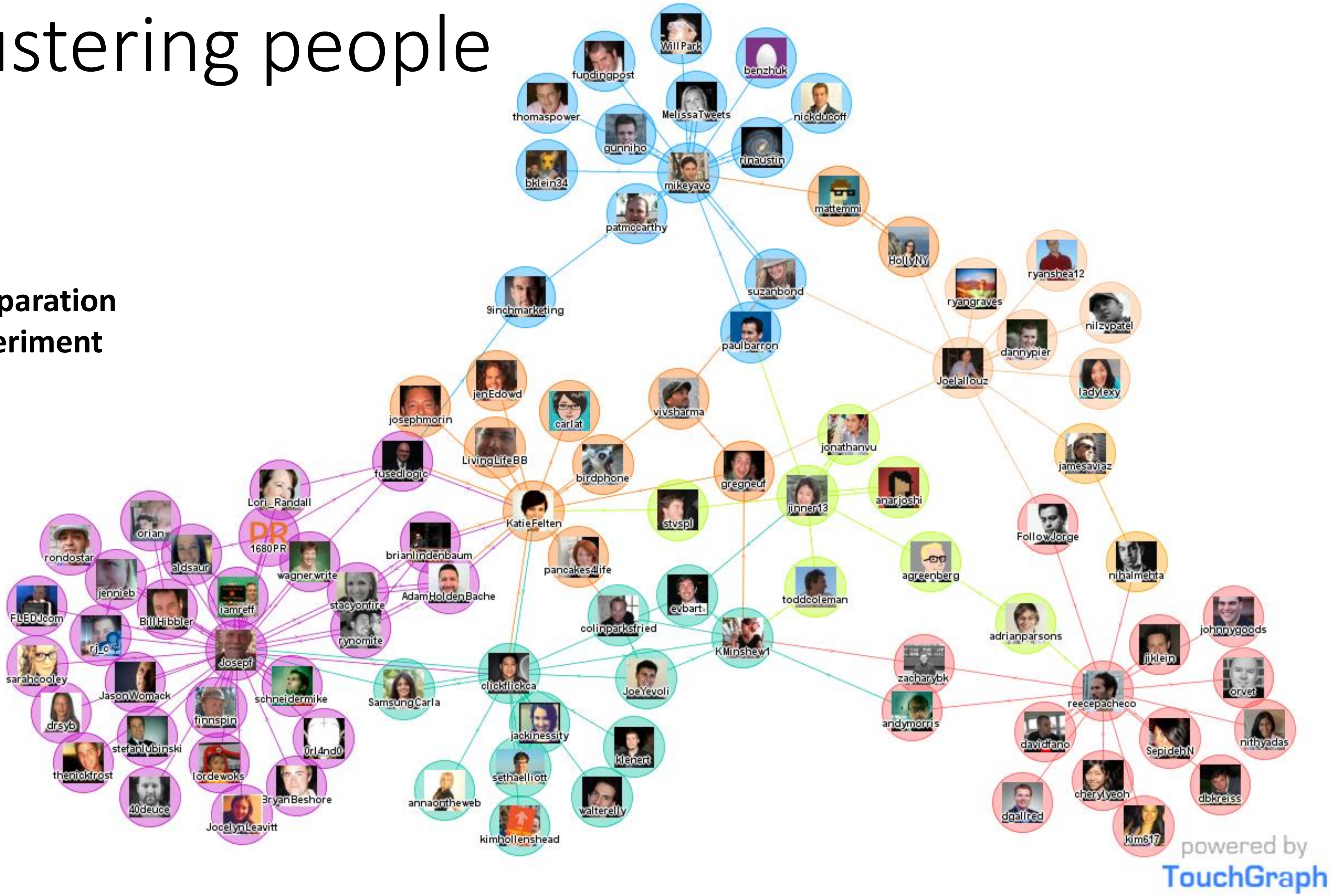
- Tonga volcano: ash, smoke and lightning seen before eruption that caused tsunami
  Guardian News · 9 hours ago
- San Diego native overseeing Tsunami Advisory alerts
  10News · 14 hours ago
- Massive underwater volcano triggers tsunami, causing damage in Tonga
  CBS News · 1 hour ago
- Missionaries in Tonga Nuku'alofa Mission safe; no contact yet with Tonga Outer Island Mission
  ksltv.com · 2 hours ago
- A massive volcanic eruption and tsunami hit Tonga and the Pacific. Here's what we know
  CNN · 14 minutes ago

View Full Coverage

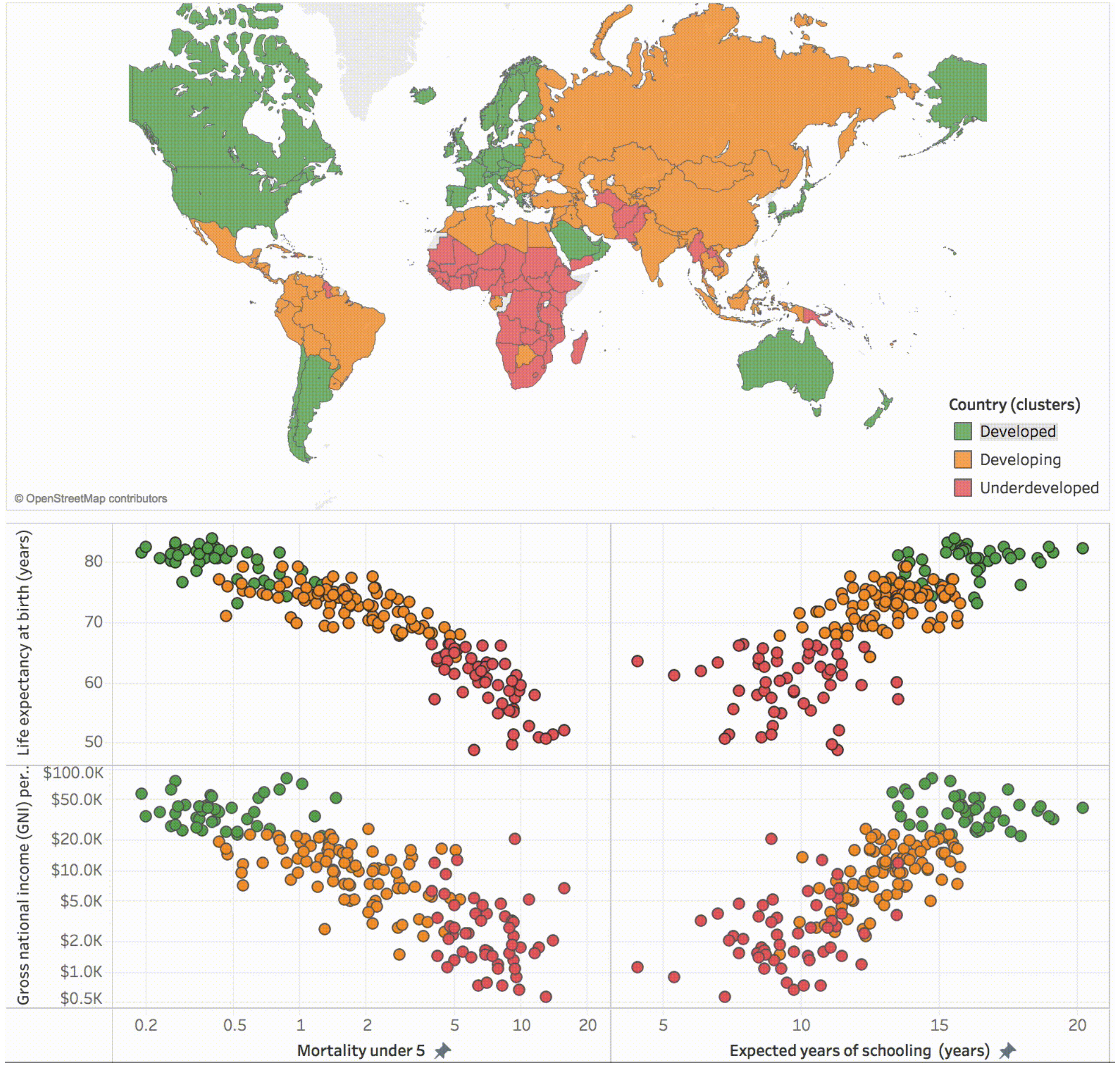
Tonga volcano Tsunami

Example : Clustering people

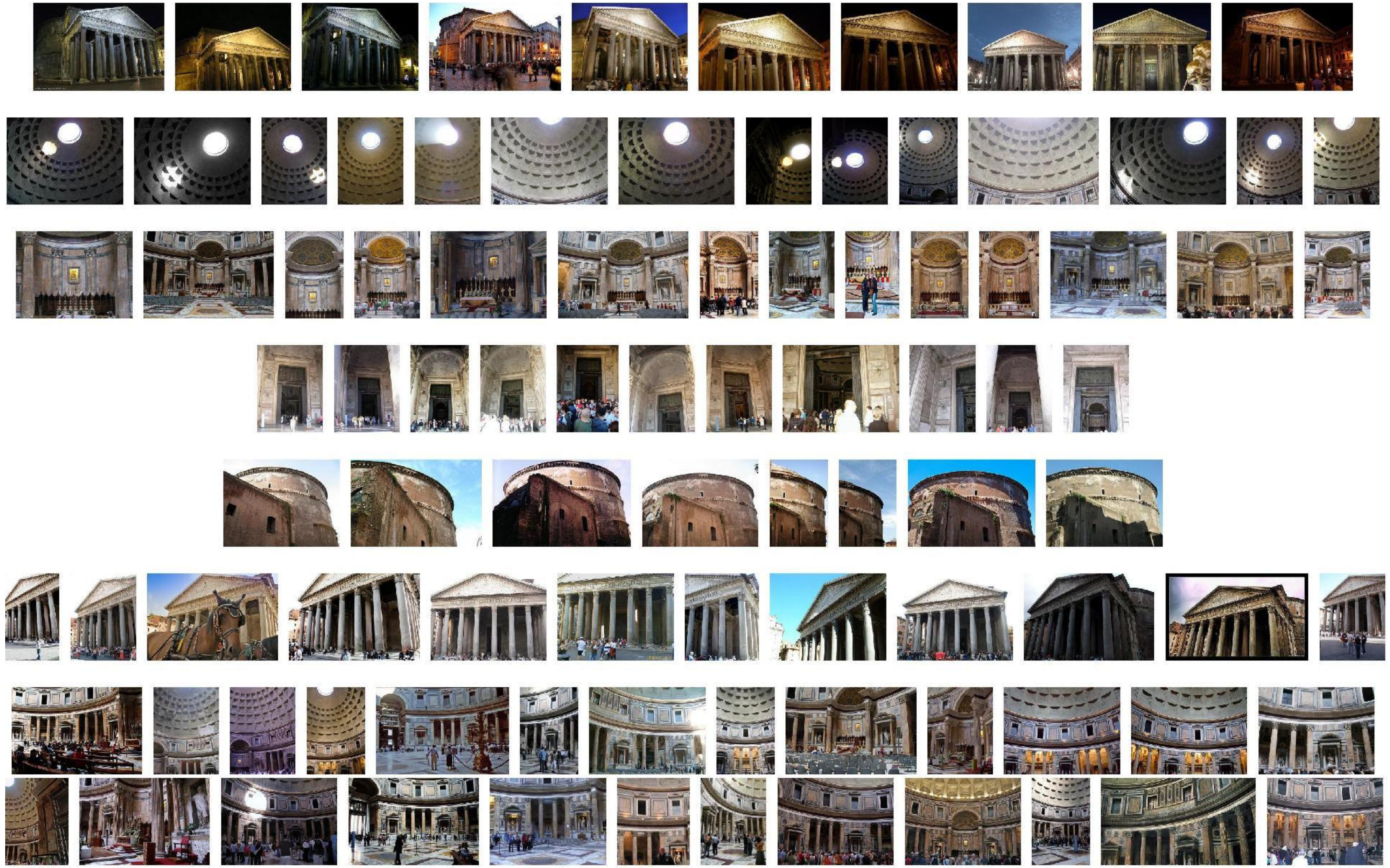
- **Six degrees of separation**
- **Small-world experiment**



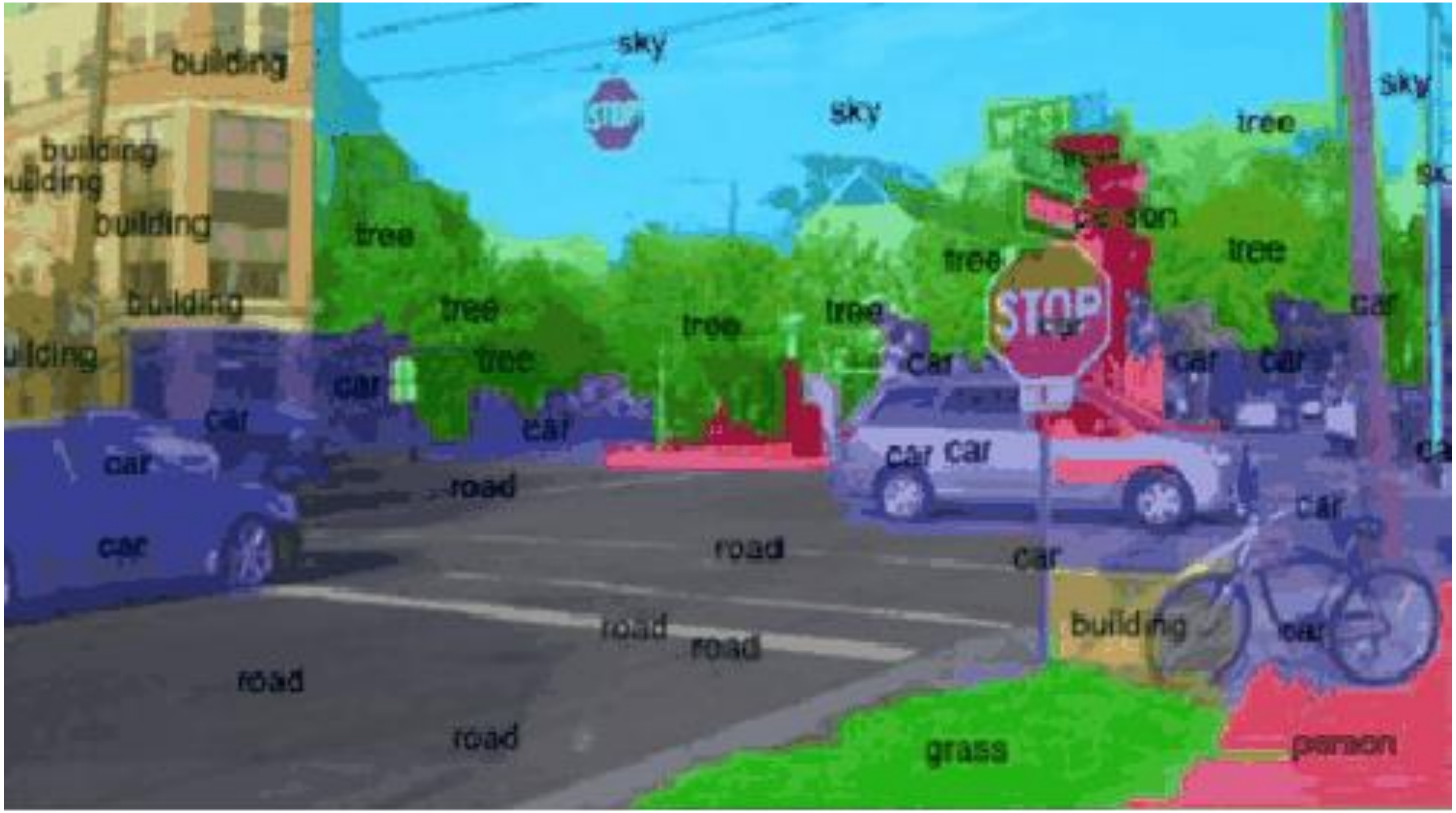
Example : Clustering countries



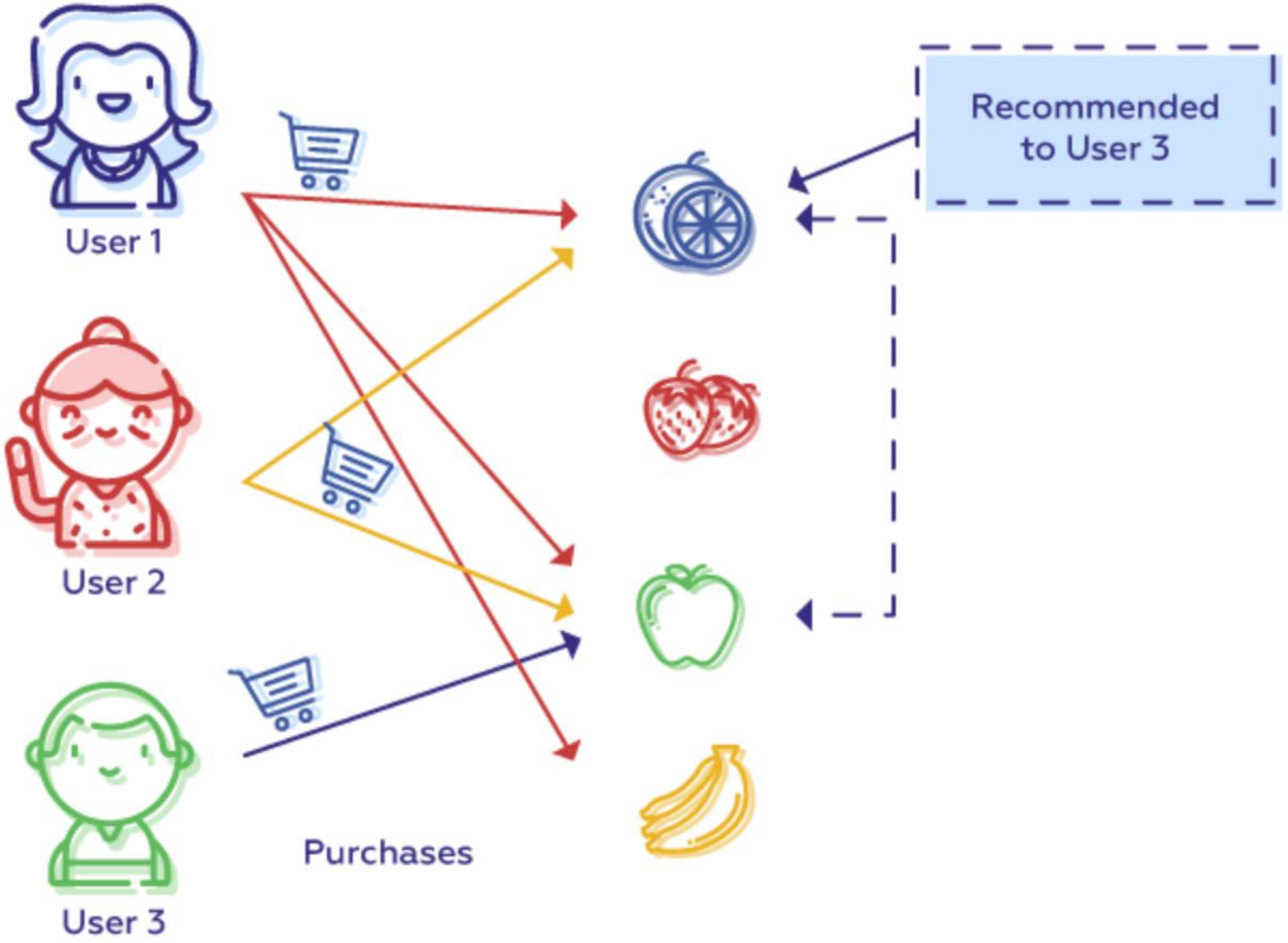
Example : Clustering images



Example : Clustering pixels



Example : Clustering items for recommendation

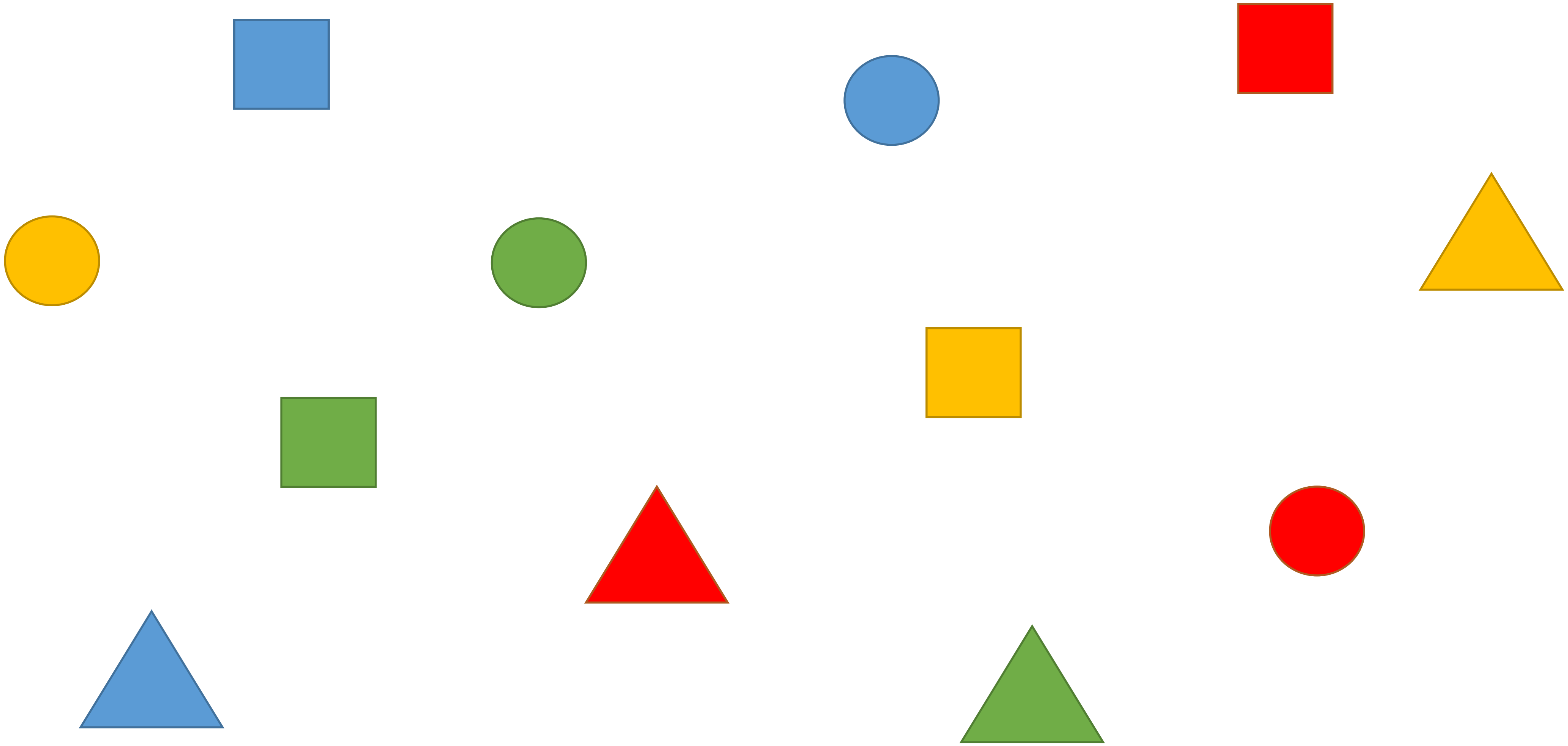


Examples

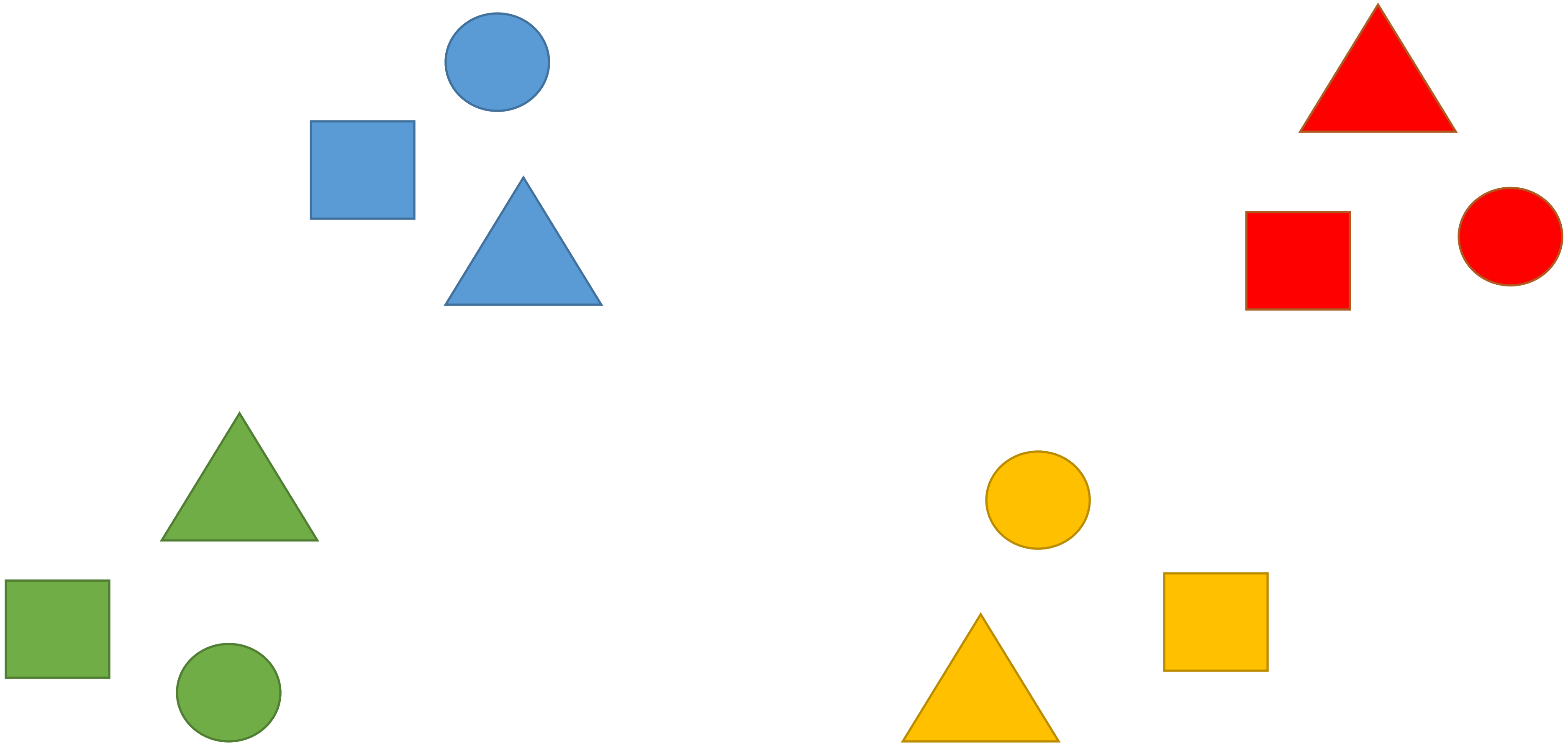
- What is the distance/similarity between :

- Two articles in news feeds ?
- Two images in a gallery ?
- Two pixels in a photo ?
- Two shows in a VOD service ?
- Two products in an e-commerce website ?
- Two persons in a social network ?
- Two ADN sequences ?

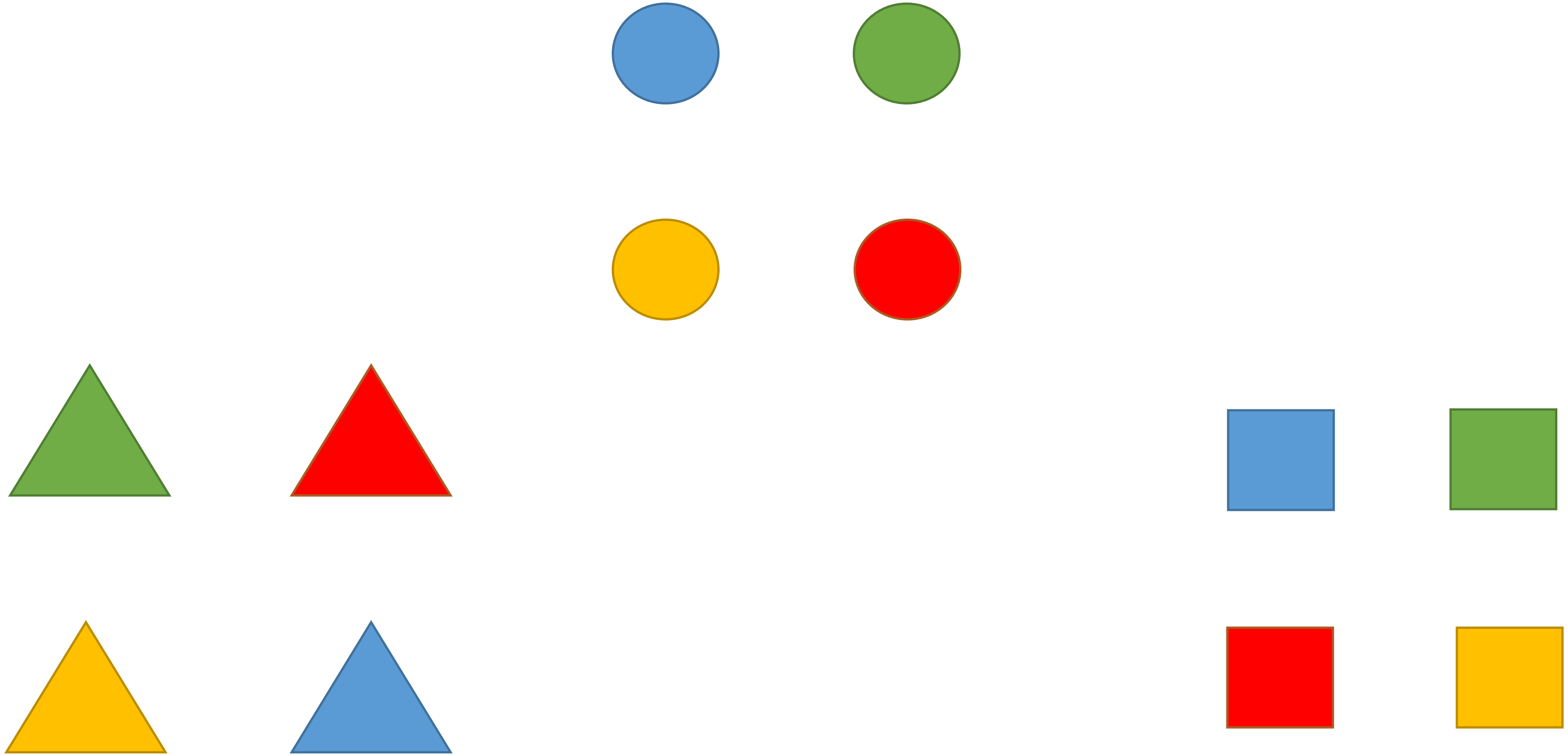
Grouping items with different colors and shapes



Based on color similarity → 4 groups



Based on shape similarity → 3 groups



The distance function

Distance axioms :

- The distance from a point to itself is null: d(x,x) = 0
- Positivity: d(x,y) >= 0
- Symmetry: d(x,y) = d(y,x)
- Triangle inequality: d(x,z) <= d(x,y) + d(y,z)

- Simplest case: one numeric attribute A

Distance(X,Y) = A(X) – A(Y)

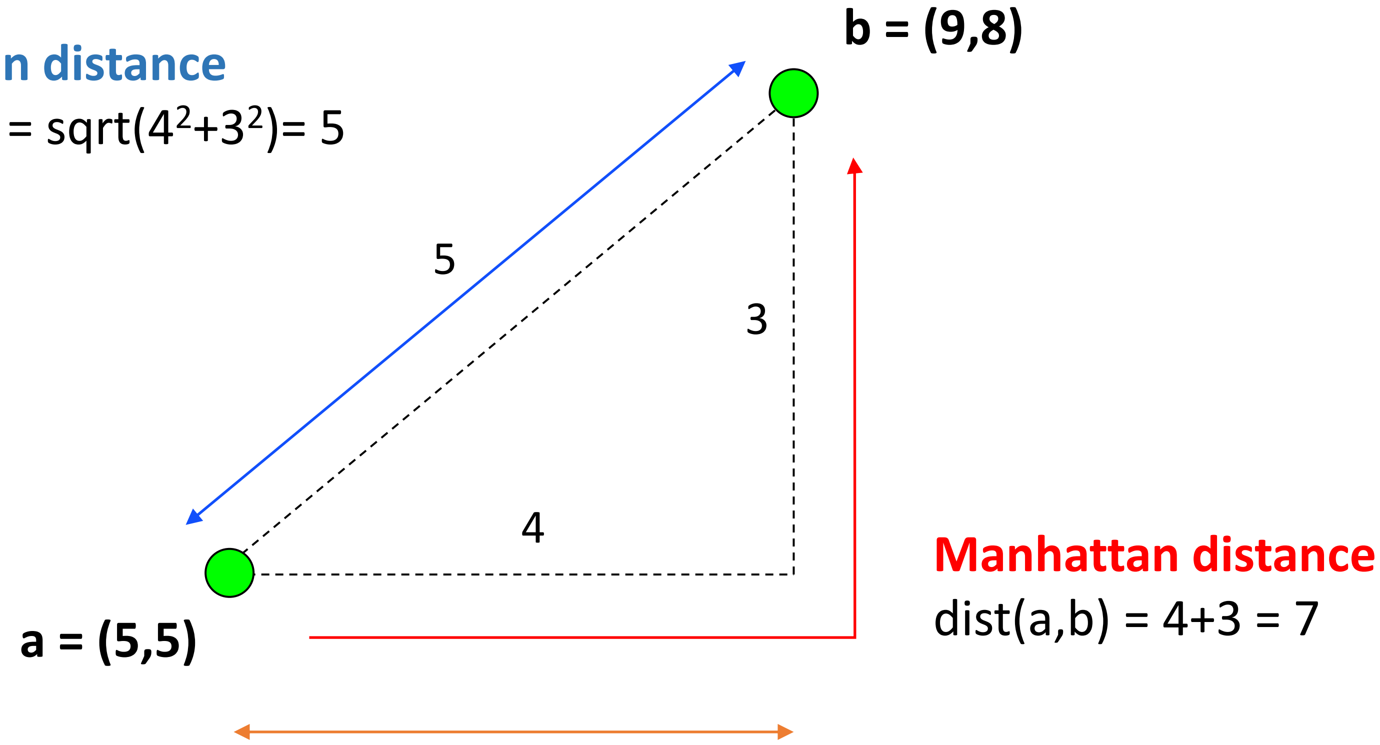
- Several numeric attributes:

Distance(X,Y) = Euclidean distance between X,Y

- Nominal attributes:

distance is set to 1 if values are different, 0 if they are equal

Examples of Euclidean Distances



Euclidean distance
dist(a,b) = sqrt(4^2+3^2)= 5

Manhattan distance
dist(a,b) = 4+3 = 7

Tchebychev distance
dist(a,b) = max(4,3) = 4

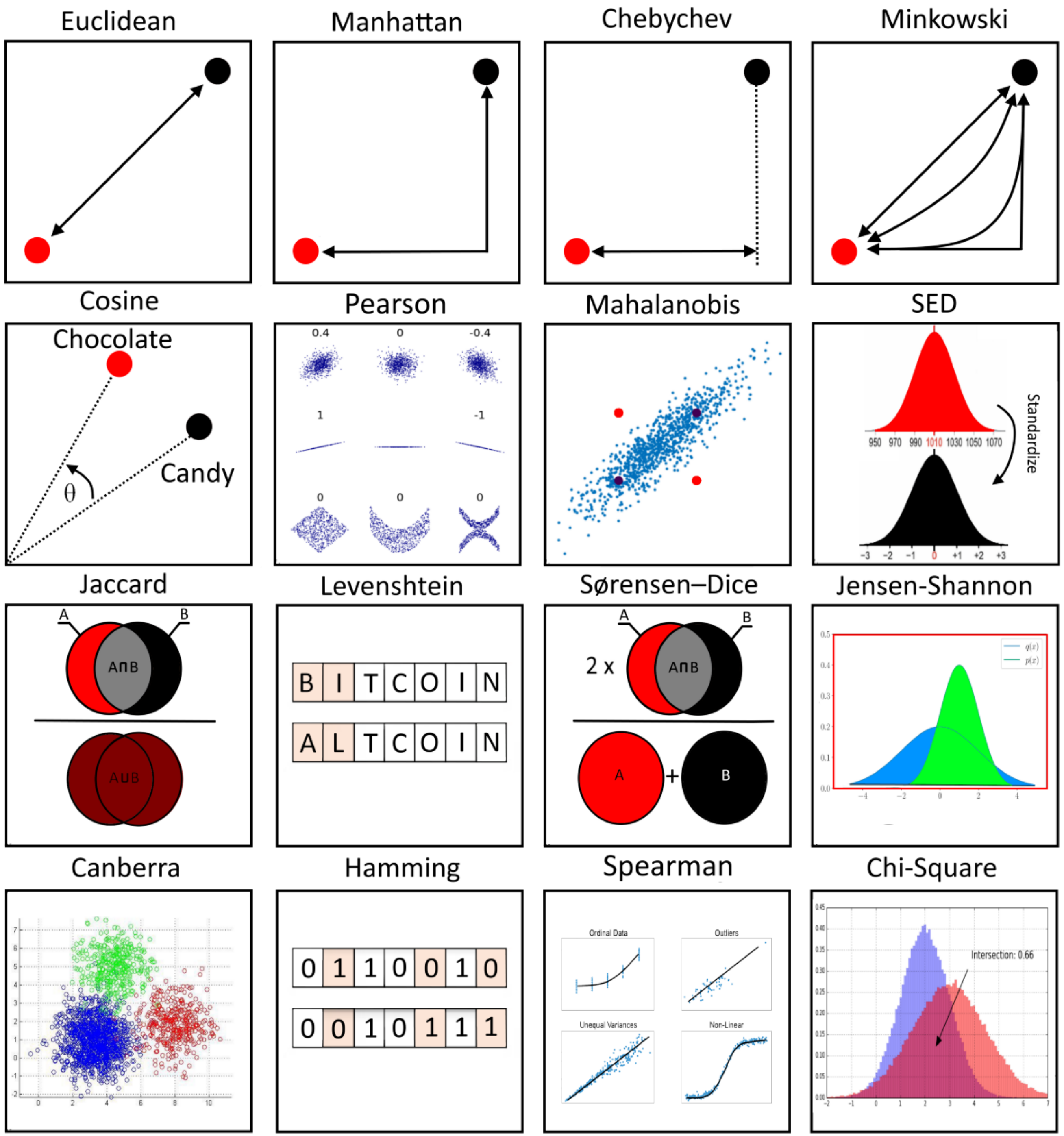
Distance / Similarity

How instances and samples
are related or close
to each other ?

Different ways to measure
depending on the nature of data
and the problems

https://docs.scipy.org/doc/scipy/reference/spatial.distance.html

You can also define your own
distance for your specific problem


Euclidean

Manhattan

Chebychev

Minkowski

Cosine

Pearson

Mahalanobis

SED

Jaccard

Levenshtein

Sørensen–Dice

Jensen-Shannon

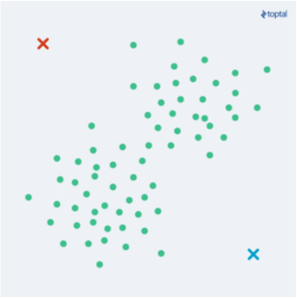
Canberra

Hamming

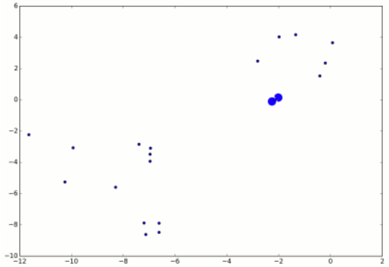
Spearman

Chi-Square

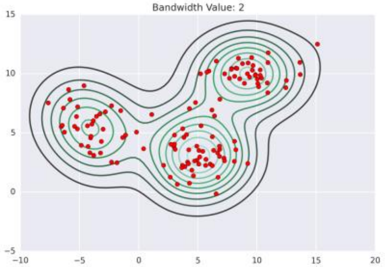
Some Clustering Algorithms



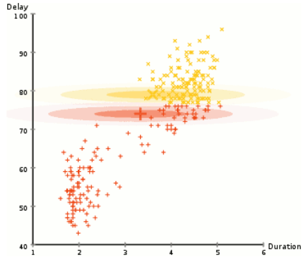
- K-means
  - Fix K. Iteratively re-assign points to the nearest cluster



- Agglomerative/Hierarchical clustering
  - Each point is a cluster. Iteratively merge the closest clusters



- Mean-shift clustering
  - Based on Kernel Density Estimation (KDE)



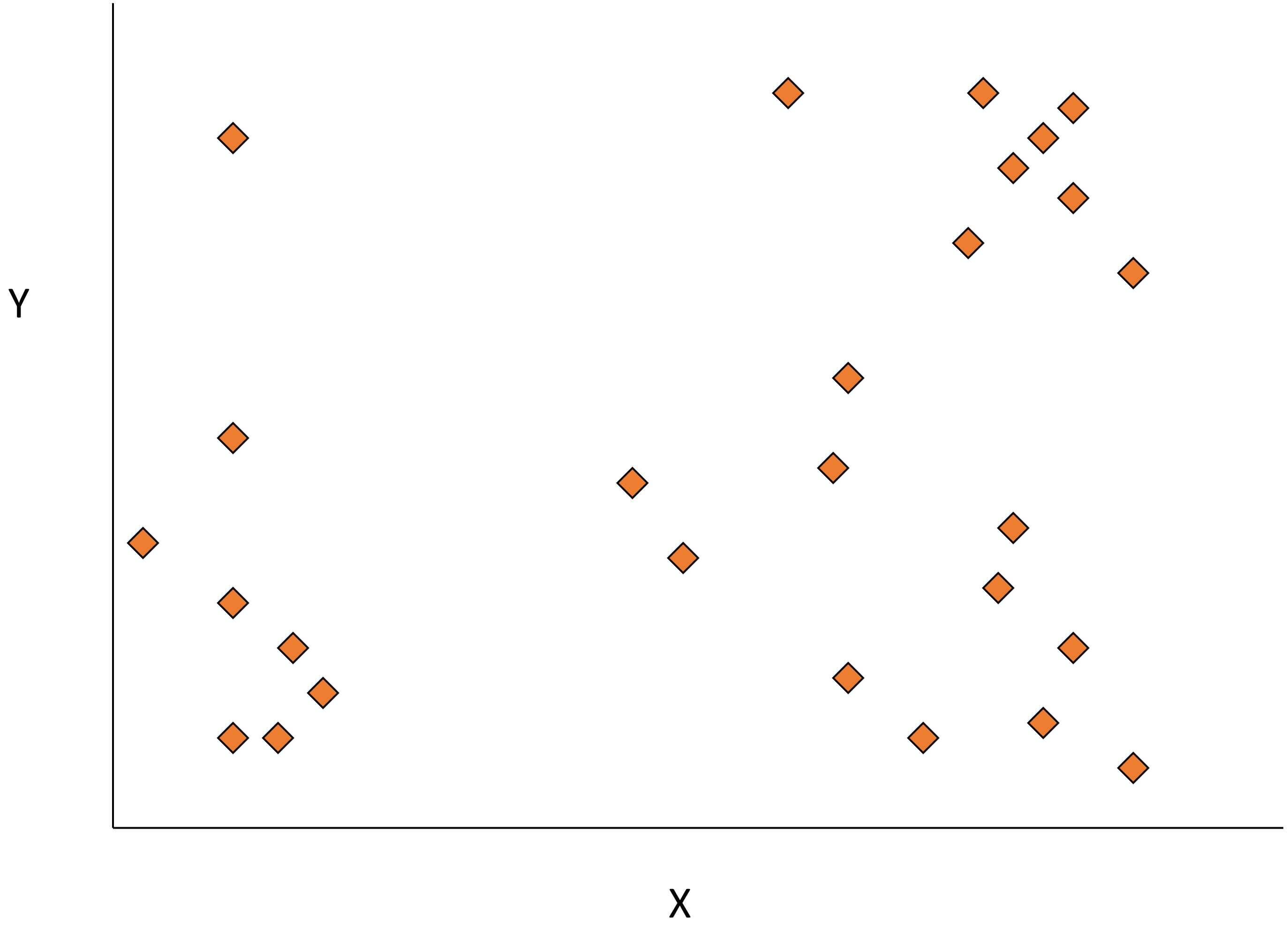
- EM Algorithm
  - Expectation of likelihood, Maximizing parameters
- And many others…

Simple Clustering: K-means

- Works with numeric data only
- Pick a number (K) of cluster centers (at random)
- Assign every item to its nearest cluster center (e.g. using Euclidean distance)
- Move each cluster center to the mean of its assigned items
- Repeat steps 2,3 until convergence (change in cluster assignments less than a threshold)

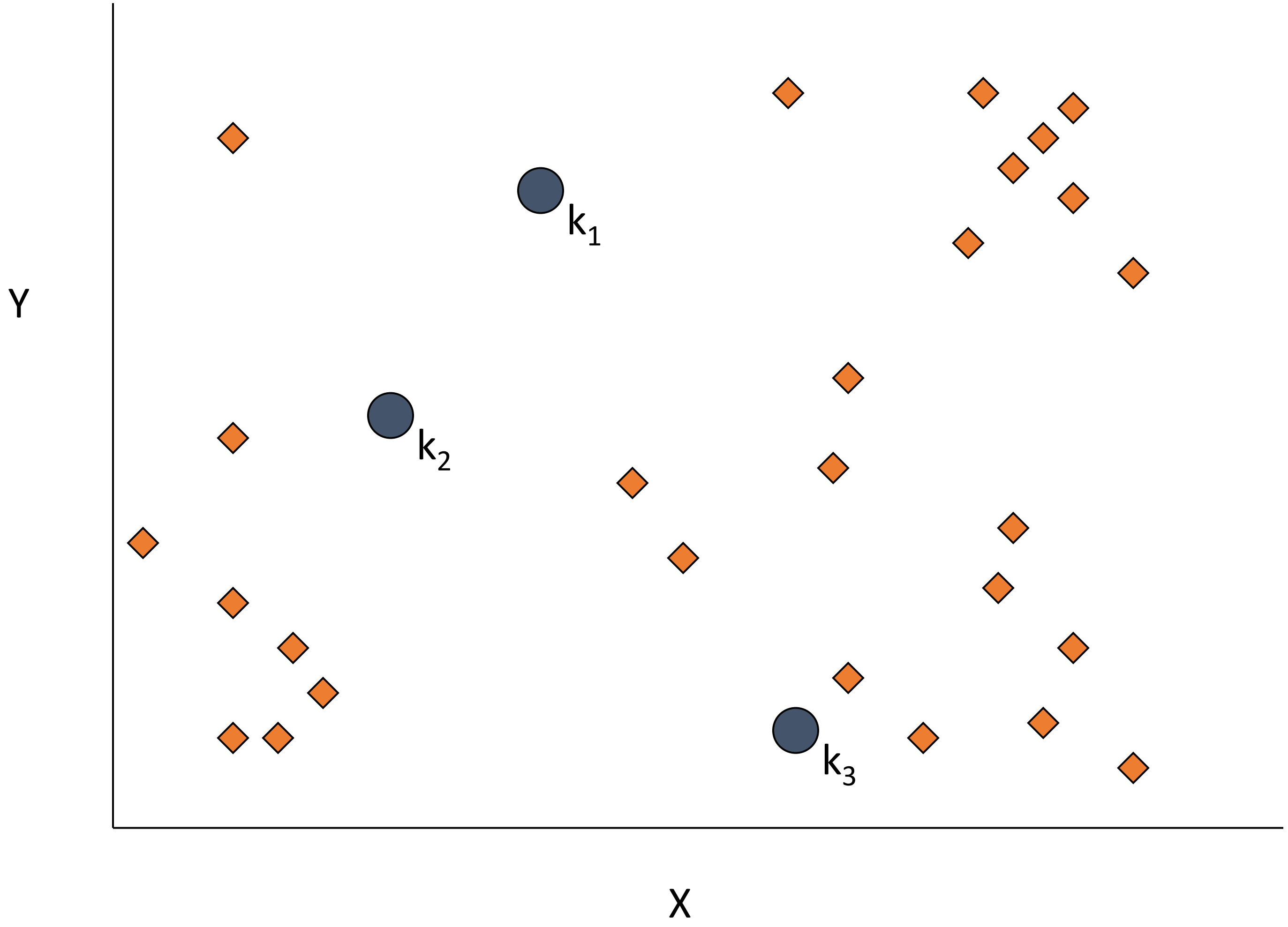
K-means example

Data
without labels



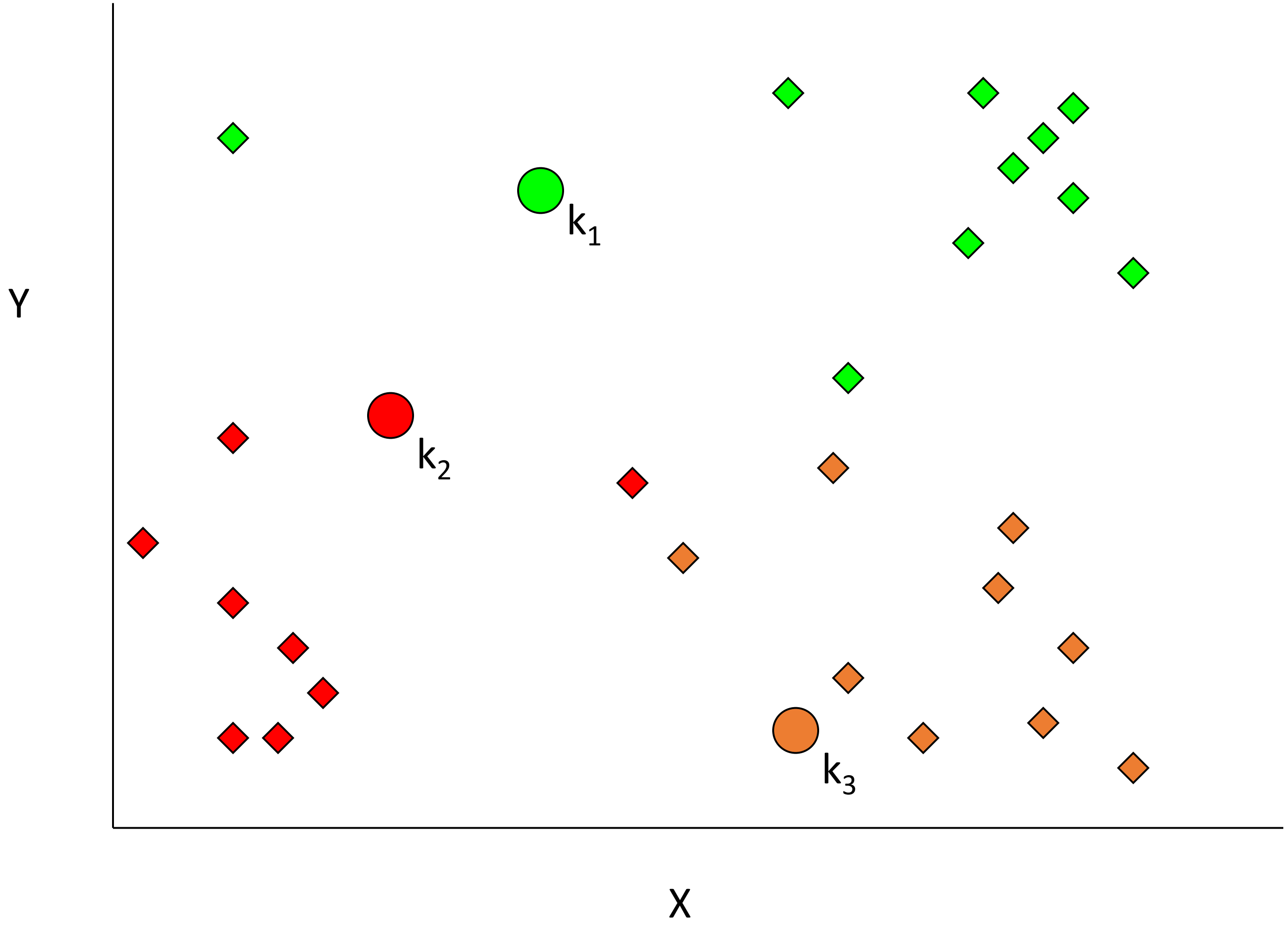
K-means example, step 1

Pick 3 initial cluster centers (randomly)



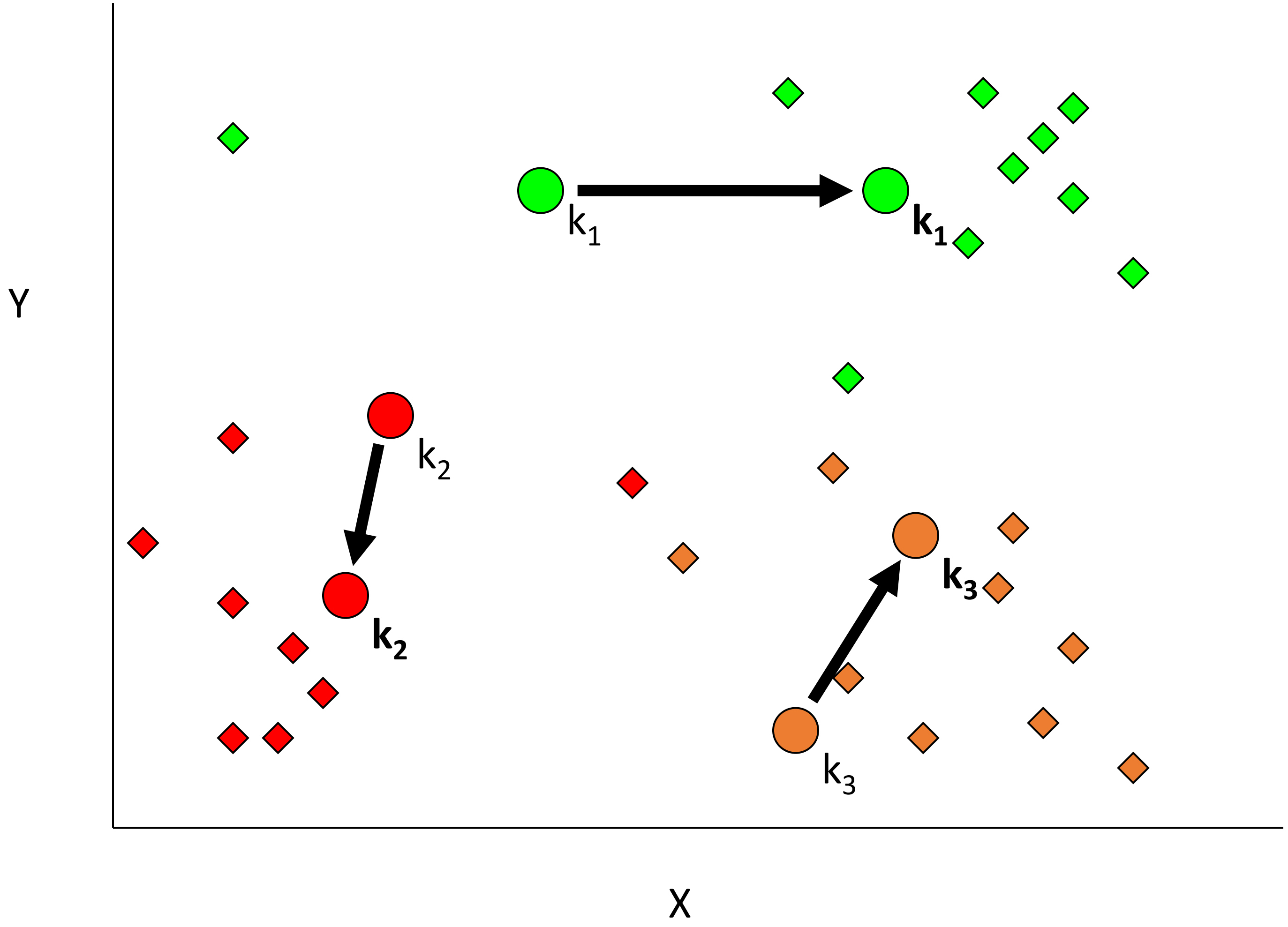
K-means example, step 2

Assign each point to the closest cluster center



K-means example, step 3

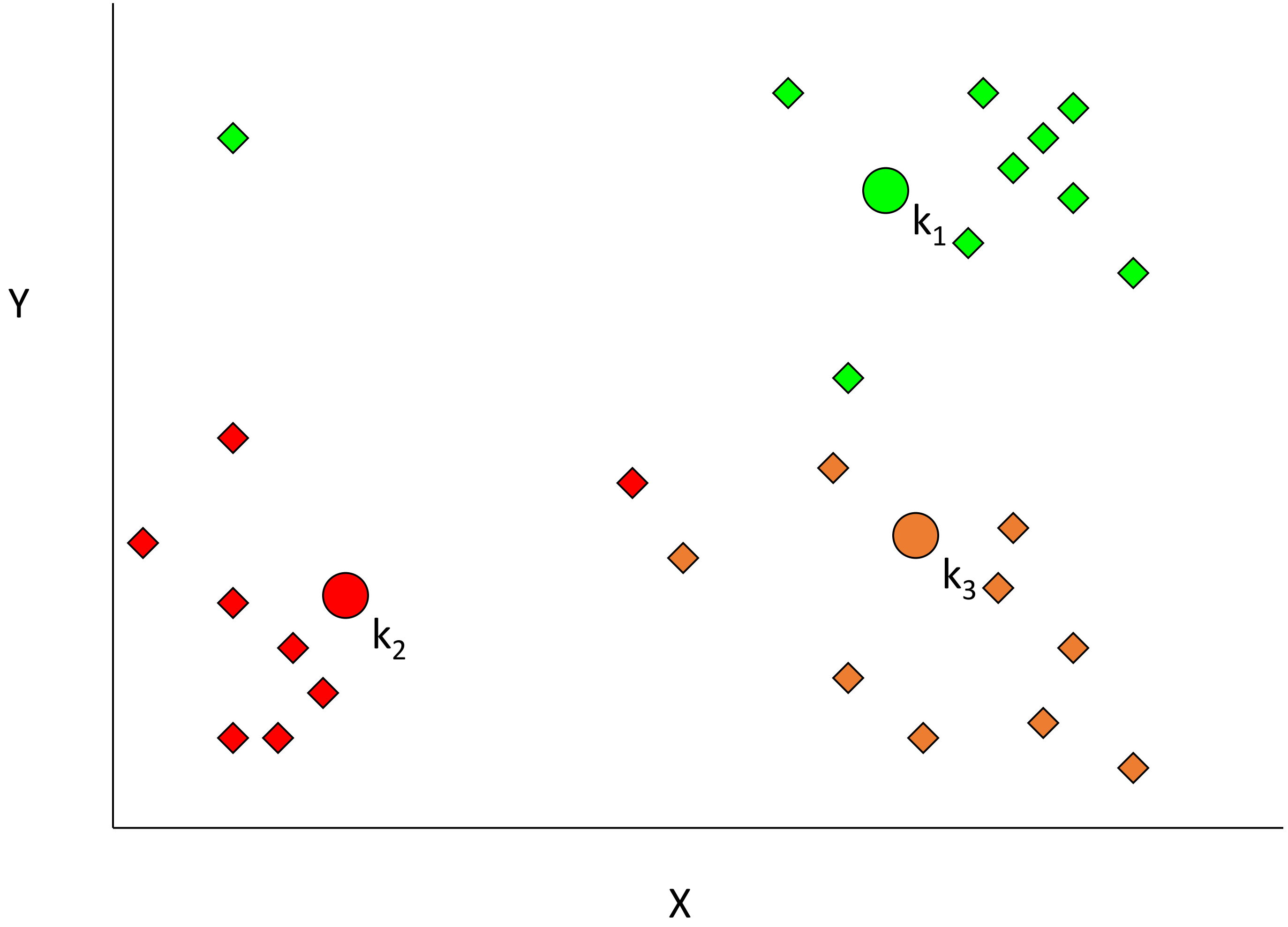
Move each cluster center to the mean of each cluster



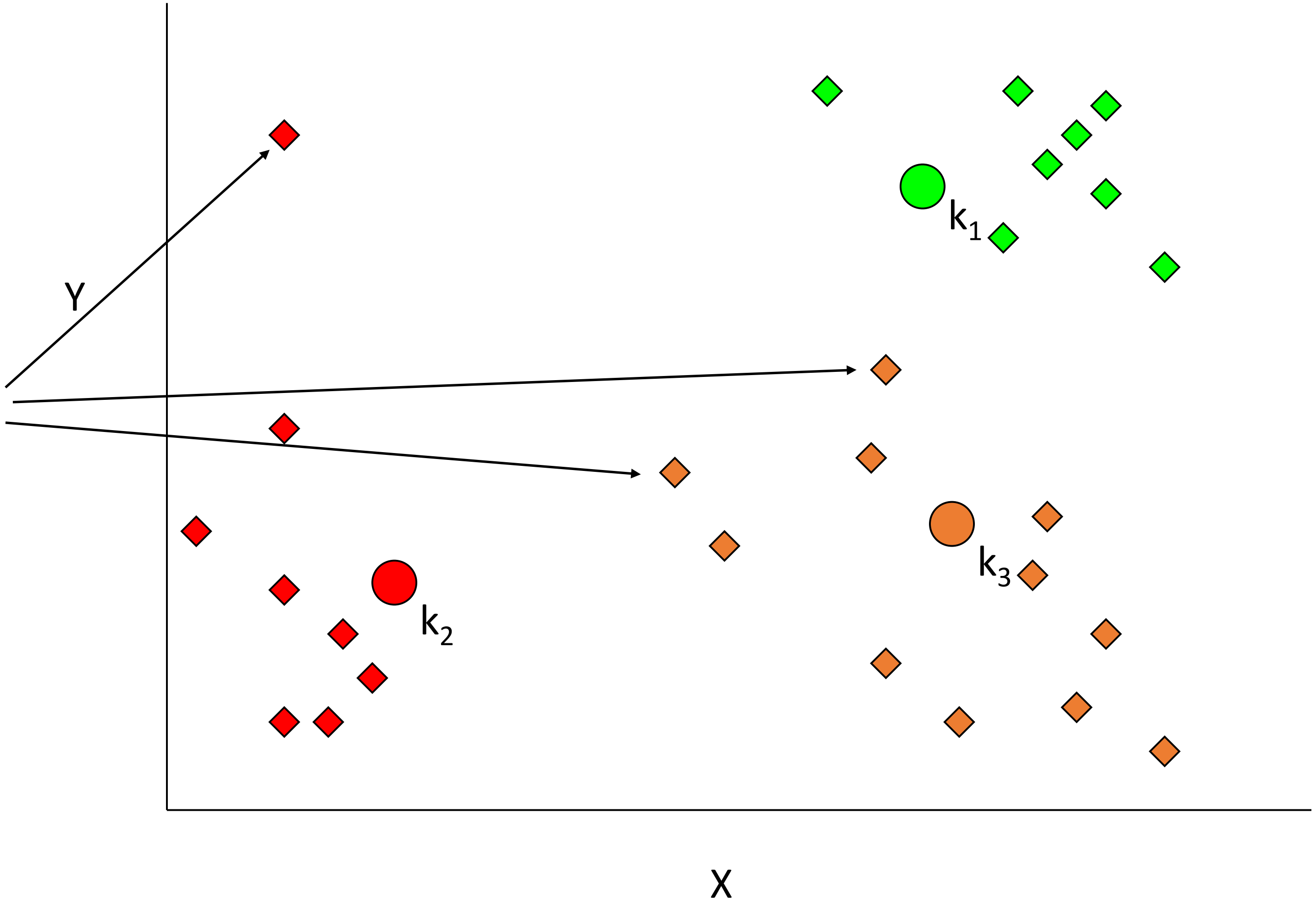
K-means example, step 4

Reassign points closest to a different new cluster center

Q: Which points are reassigned?



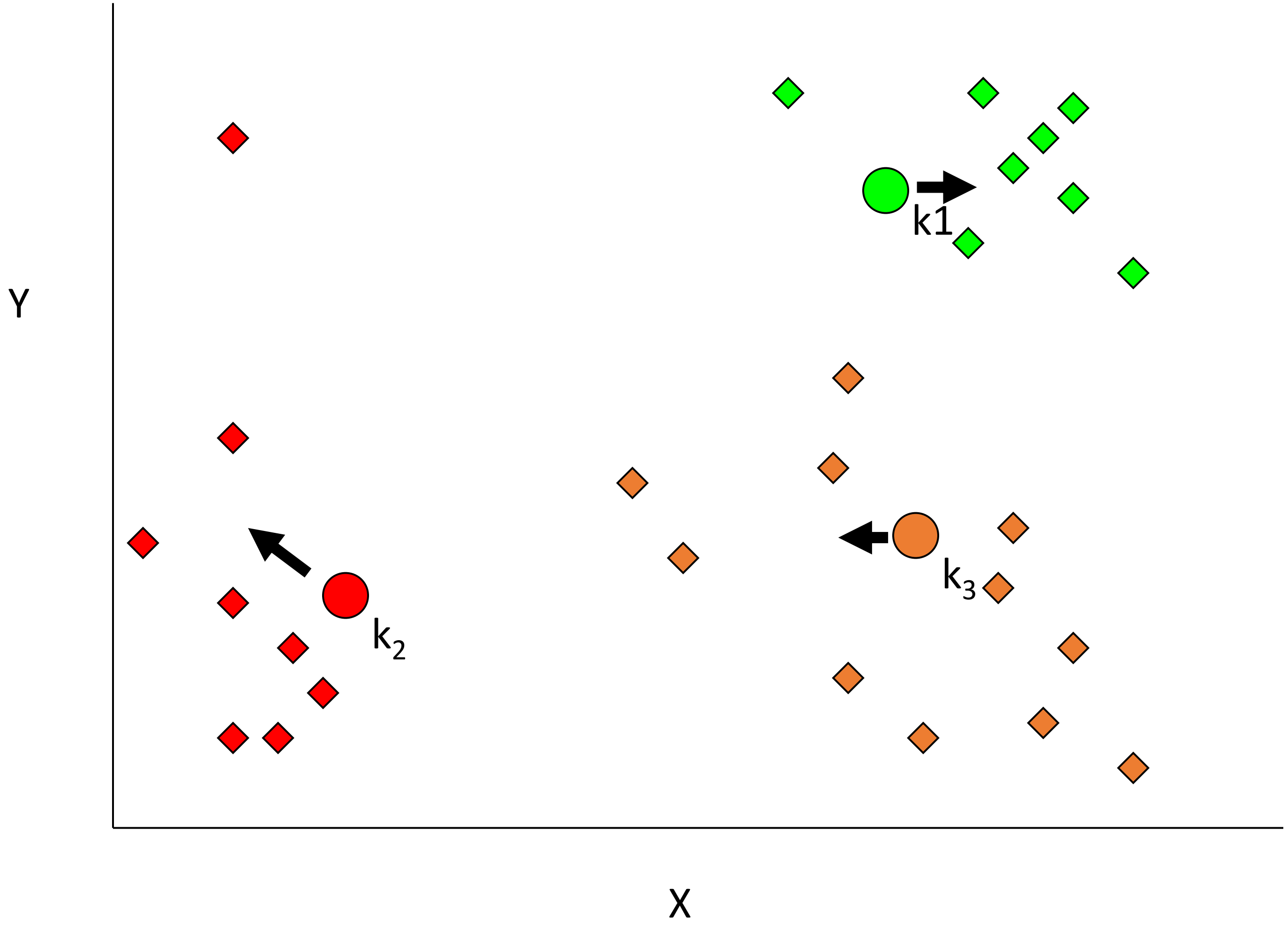
K-means example, step 4 …



for these three points ?

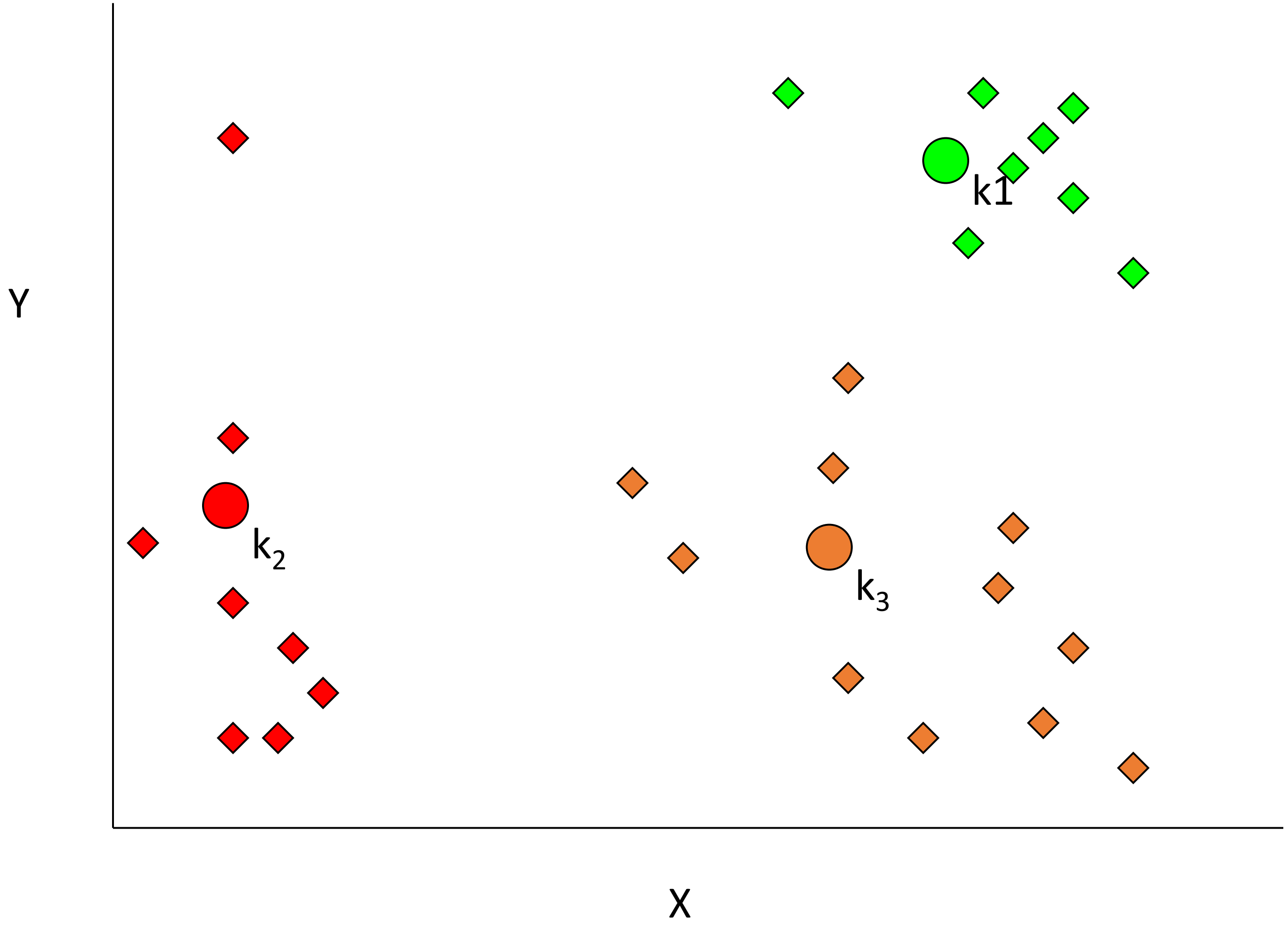
K-means example, step 4b

re-compute cluster means

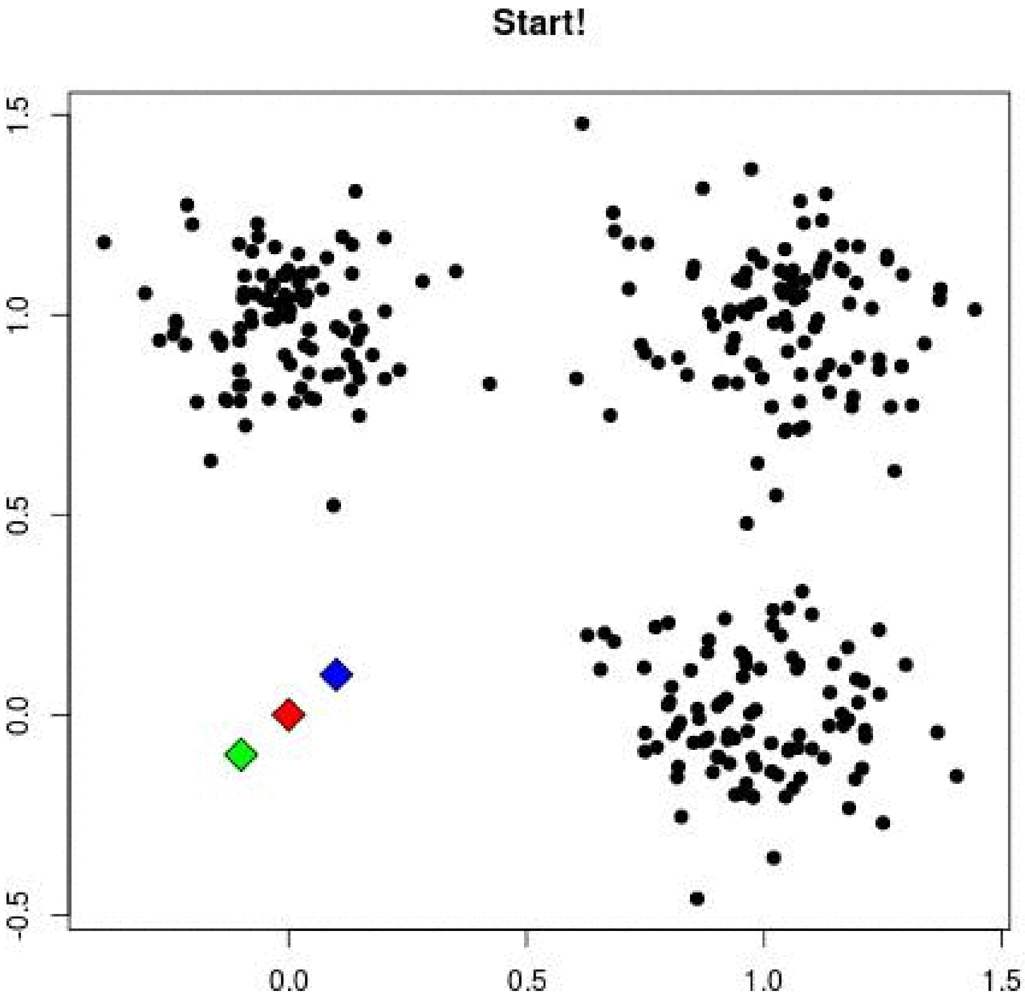


K-means example, step 5

move cluster centers to cluster means



K-means example, iterate…



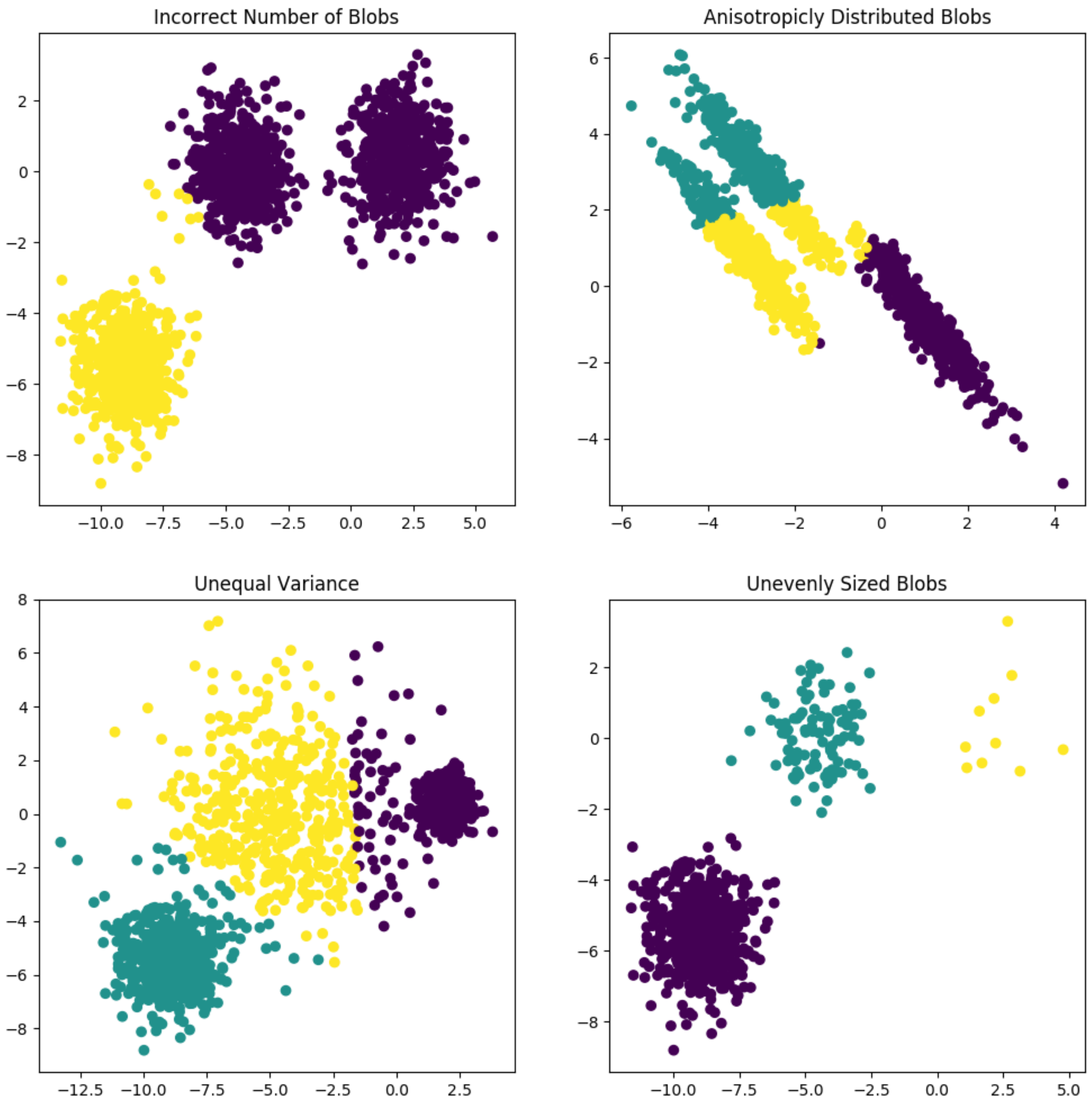
K-Means pros and cons

- **Pros**
  - Finds cluster centers that minimize conditional variance
  - Simple and fast
  - Easy to implement
  - …
- **Cons**
  - Need to choose K
  - Sensitive to outliers
  - All clusters have the same parameters
  - …

k-means assumptions

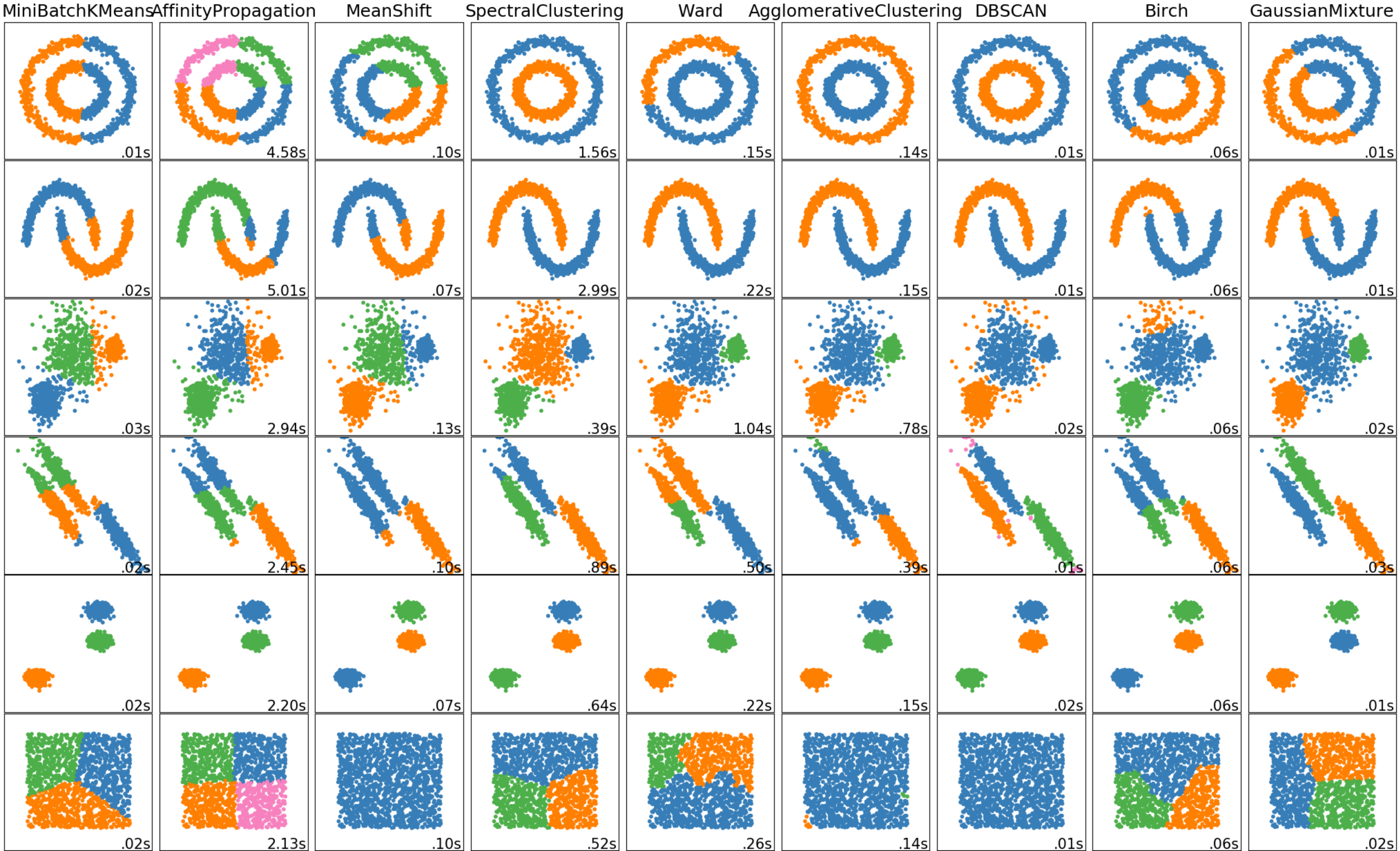
Situations where k-means will produce unintuitive and possibly unexpected clusters

Importance of dataviz

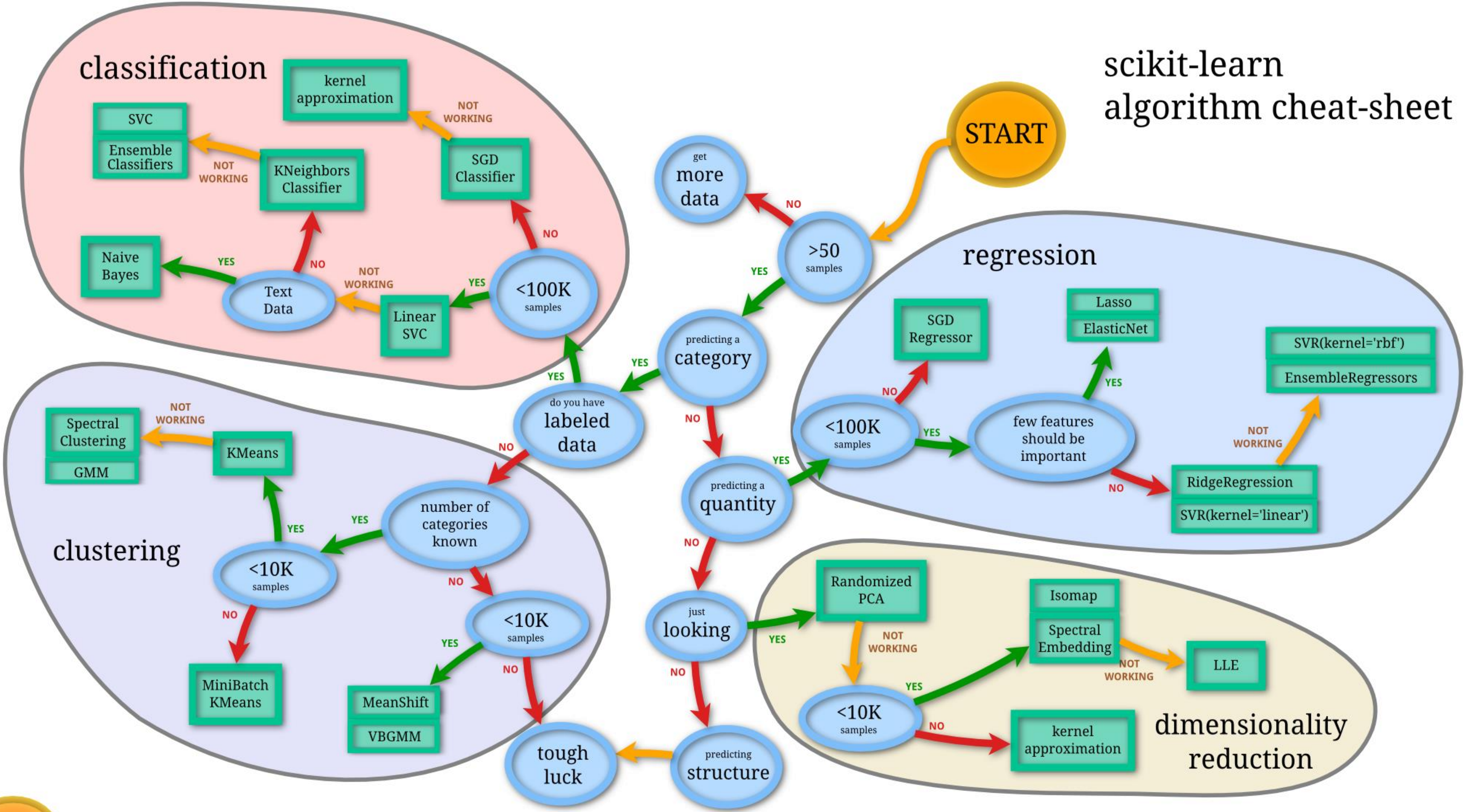


- Clustering in 2D looks easy
- Clustering small amounts of data look easy too

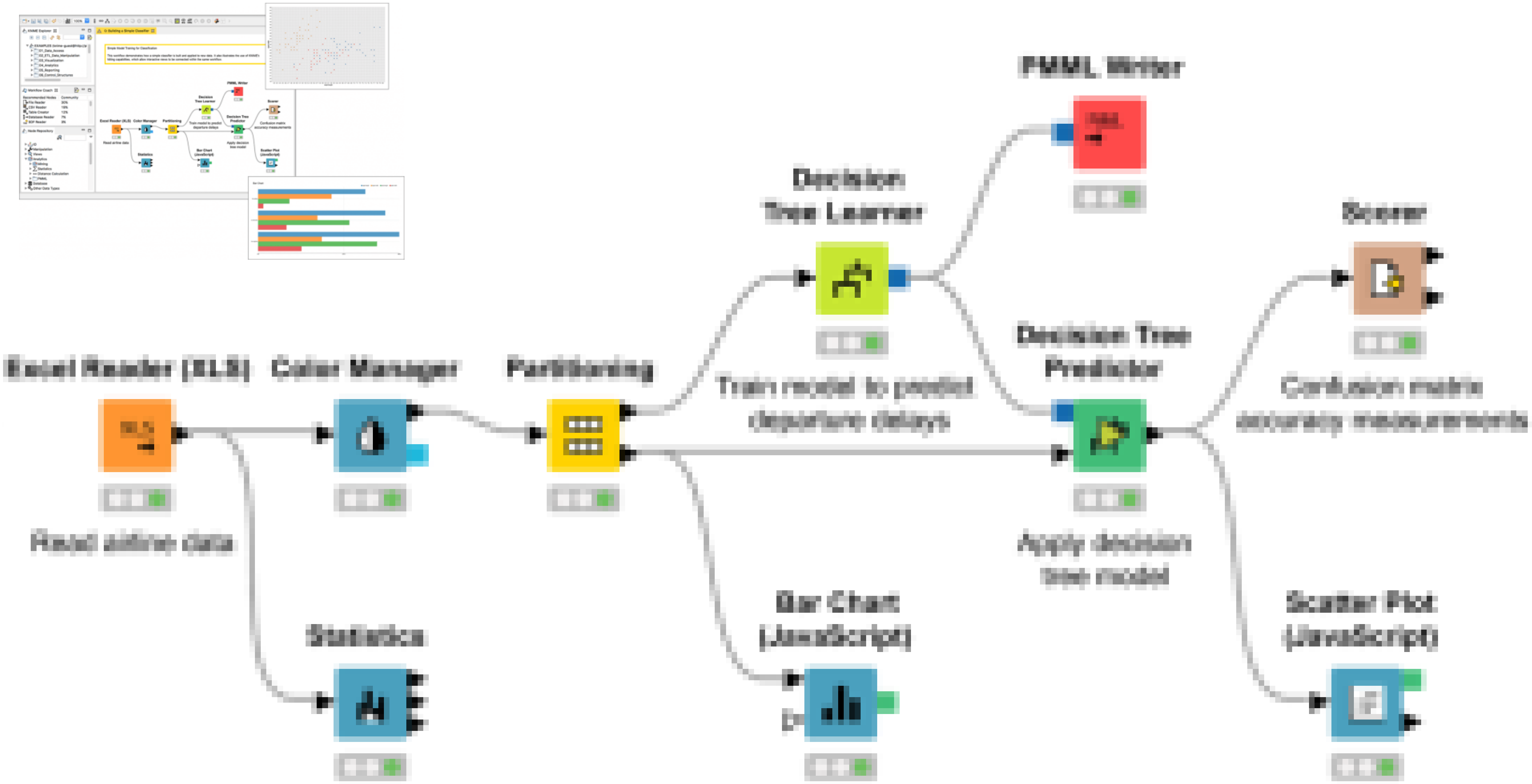
- Many applications involve more than 2D (Ex. > 10000 dimensions) with huge amounts of data


MiniBatchKMeans
AffinityPropagation
MeanShift
SpectralClustering
Ward
AgglomerativeClustering
DBSCAN
Birch
GaussianMixture
.01s
4.58s
.10s
1.56s
.15s
.14s
.01s
.06s
.01s
.02s
5.01s
.07s
2.99s
.22s
.15s
.01s
.06s
.01s
.03s
2.94s
.13s
.39s
1.04s
.78s
.02s
.06s
.02s
.02s
2.45s
.10s
.89s
.50s
.39s
.01s
.06s
.03s
.02s
2.20s
.07s
.64s
.22s
.15s
.02s
.06s
.01s
.02s
2.13s
.10s
.52s
.26s
.14s
.01s
.06s
.02s

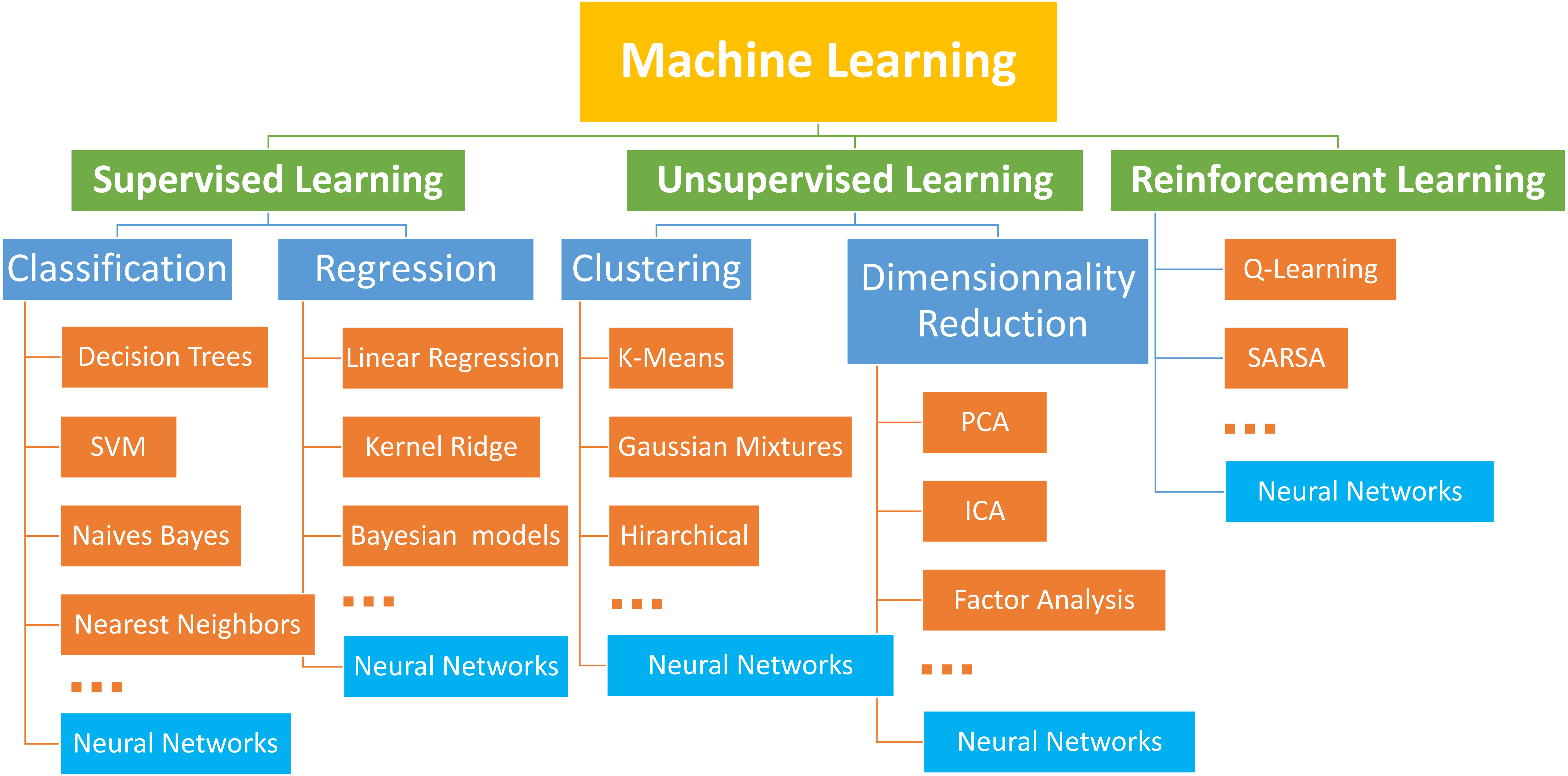
scikit-learn algorithm cheat-sheet



Back



https://www.knime.com/



- Machine Learning
  - Supervised Learning
    - Classification
      - Decision Trees
      - SVM
      - Naives Bayes
      - Nearest Neighbors
      - ...
      - Neural Networks
    - Regression
      - Linear Regression
      - Kernel Ridge
      - Bayesian models
      - ...
      - Neural Networks
  - Unsupervised Learning
    - Clustering
      - K-Means
      - Gaussian Mixtures
      - Hirarchical
      - ...
      - Neural Networks
    - Dimensionnality Reduction
      - PCA
      - ICA
      - Factor Analysis
      - ...
      - Neural Networks
  - Reinforcement Learning
    - Q-Learning
    - SARSA
    - ...
    - Neural Networks