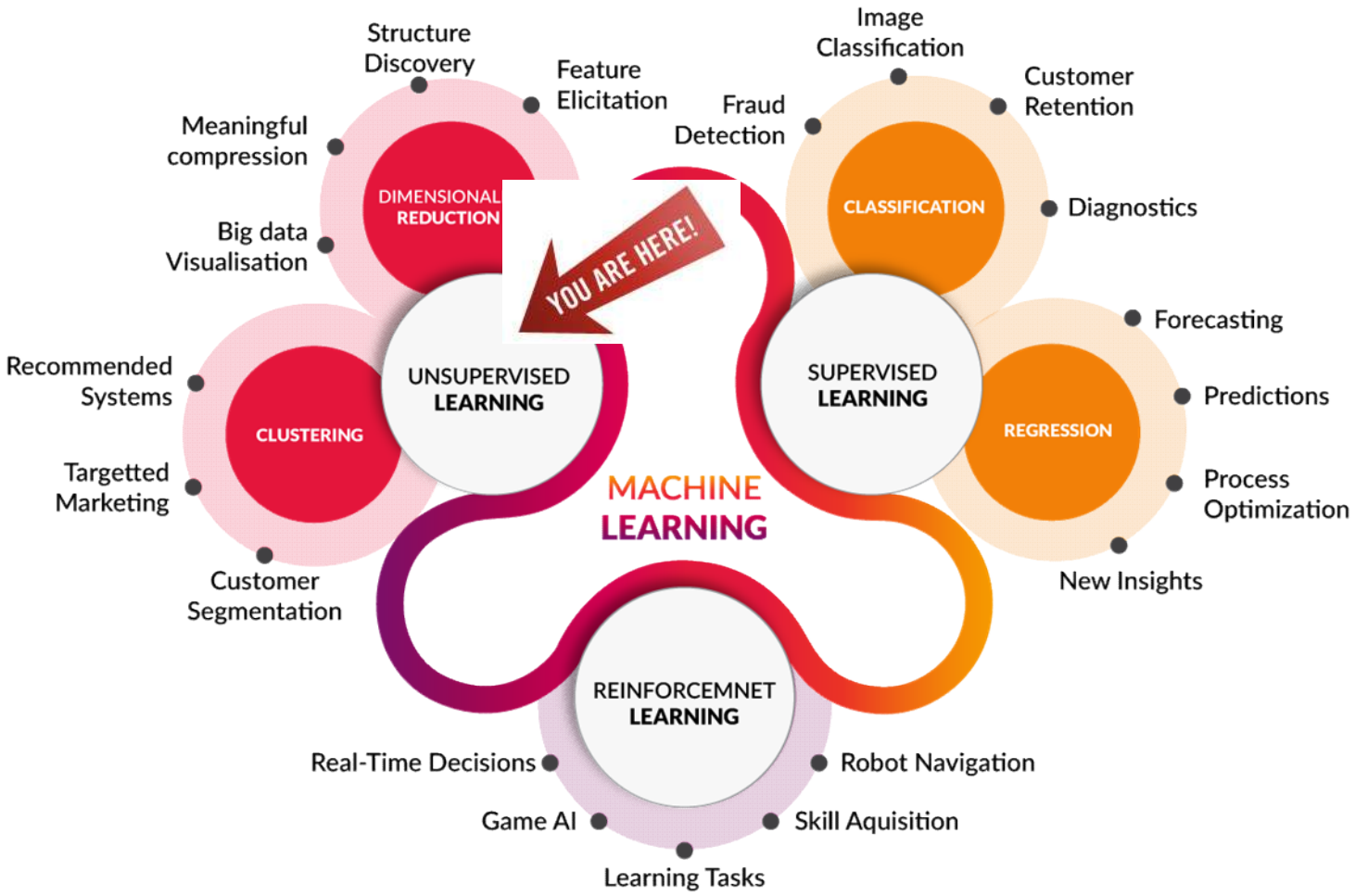
COMP 472: Artificial Intelligence
Machine Learning
Unsupervised Learning

- Russell & Norvig: *not much really*

Today

1. Introduction to ML
2. Naïve Bayes Classification
   a. Application to Spam Filtering
3. Decision Trees
4. (Evaluation
5. Unsupervised Learning)
6. Neural Networks
   a. Perceptrons
   b. Multi Layered Neural Networks

Types of Machine Learning



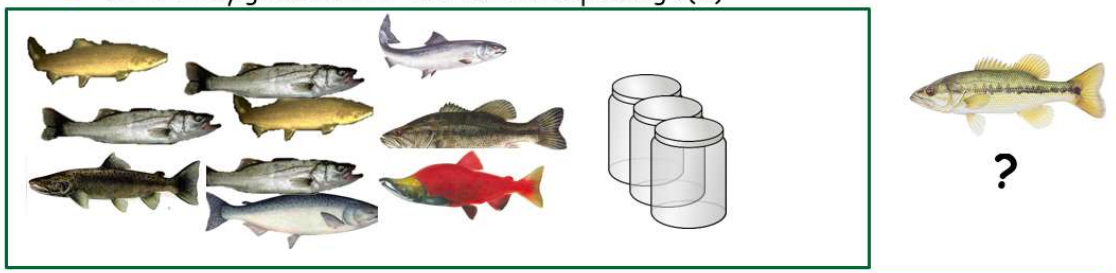
http://www.cognub.com/index.php/cognitive-platform/

Remember this slide?

Types of Learning

- Supervised learning
  - We are given a training set of (X, f(X)) pairs
  - X = <color, length>

- Unsupervised learning
  - We are only given the Xs - not the corresponding f(X)



4

unlabeled不会分辨出这个鱼是哪种鱼，而是把鱼根据特性group together

Unsupervised Learning

- Learn without labeled examples
  - i.e. X is given, but not f(X)

而是每一类中，里面的object尽量接近，不同的类别差距尽量大

small nose	big teeth	small eyes	moustache	f(X) = ?

not given

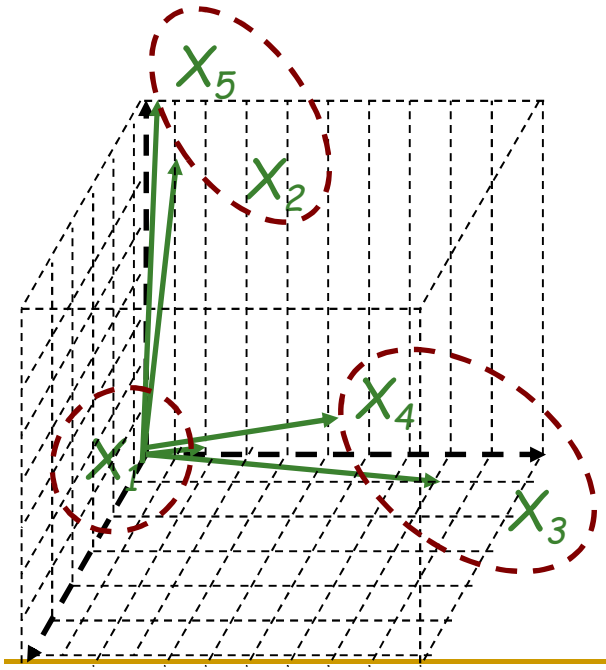
- Without a f(X)
  - you can't really identify/label a test instance
  - but you can:
    - Cluster/group the features of the test data into a number of groups

区别

    - Discriminate between these groups without actually labeling them

Clustering 分类归并

- Represent each instance as a vector <a_1, a_2, a_3,…, a_n>
- Each vector can be visually represented in a n dimensional space



	a_1	a_2	a_3	Output
X_1	1	0	0	?
X_2	1	6	0	?
X_3	8	0	1	?
X_4	6	1	0	?
X_5	1	7	1	?

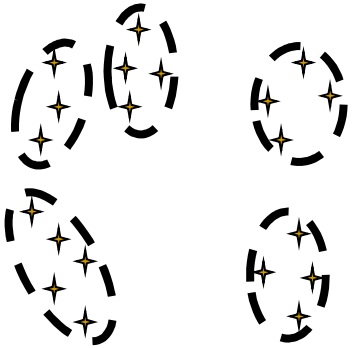
x5x2一组，x4x3一组，因为他们满足了同一组的distance minimize,不同组的distance maximize

k-means Clustering

就是把每个instance在n dimensional space上表达成一个点

1. Represent each instance as a point on a n dimensional space
2. Partition points into k regions such that:
   - distance between points within a region is minimized
   - distance between points across regions is maximized

把points分为k组，确保同一region距离最小，不同region距离最大



- Naturally works well with features with numerical values
  - where distance between points can be measured by the Euclidean distance
- Needs modifications for categorical values

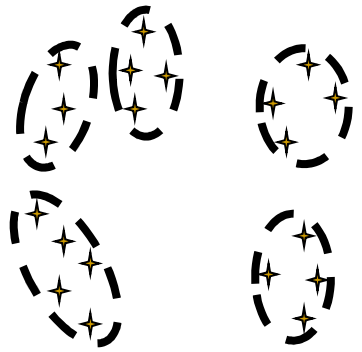
如果instance的feature都用数字表示，很好使，distance用Euclidean distance计算

  - which have no order
    - eg. "Honda", "Audi", "BMW", "Ferrari", "Nissan", "Lamborghini"
  - needs domain-specific distance measure

```
dist(Honda,Nissan)=1
dist(Honda,Audi)=3
dist(ferrari,lamborghini)=1
```

我们这里定义的是贵的距离近，便宜的距离近，不同价格区间的距离远
因为他们之间没有具体数字，需要一些特定领域的距离测量手段

k-means Clustering



- User selects how many clusters they want (the value of k)

1. Place k points into the space (eg. at random). These points represent initial group centroïds.
2. Assign each data point x_n to the nearest centroïd.
3. When all data points have been assigned, recalculate the positions of the k centroïds as the average of the cluster
4. Repeat Steps 2 and 3 until none of the data instances change group.

Euclidean Distance

- To find the nearest centroïd…
  - typical metric is the Euclidean distance
  - Euclidean distance between 2 pts:

p就是未分配的点的坐标 $p = (p_1, p_2, \ldots, p_n)$

q是中心点的坐标 $q = (q_1, q_2, \ldots, q_n)$

n=几取决于是几维空间

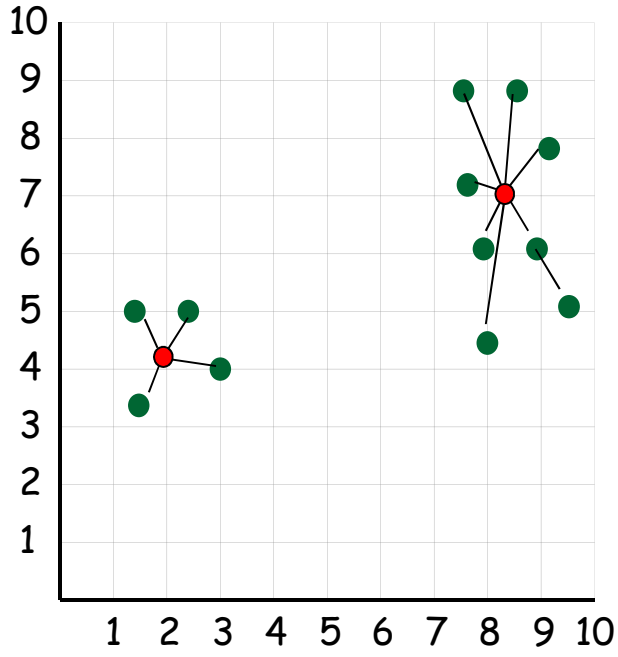
$$d = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

d=相减的平方和

- To compute the next generation of centroïds…
  - take mean of all points in the cluster in each dimension
  - mean of 2 points: 两个点之间的mean

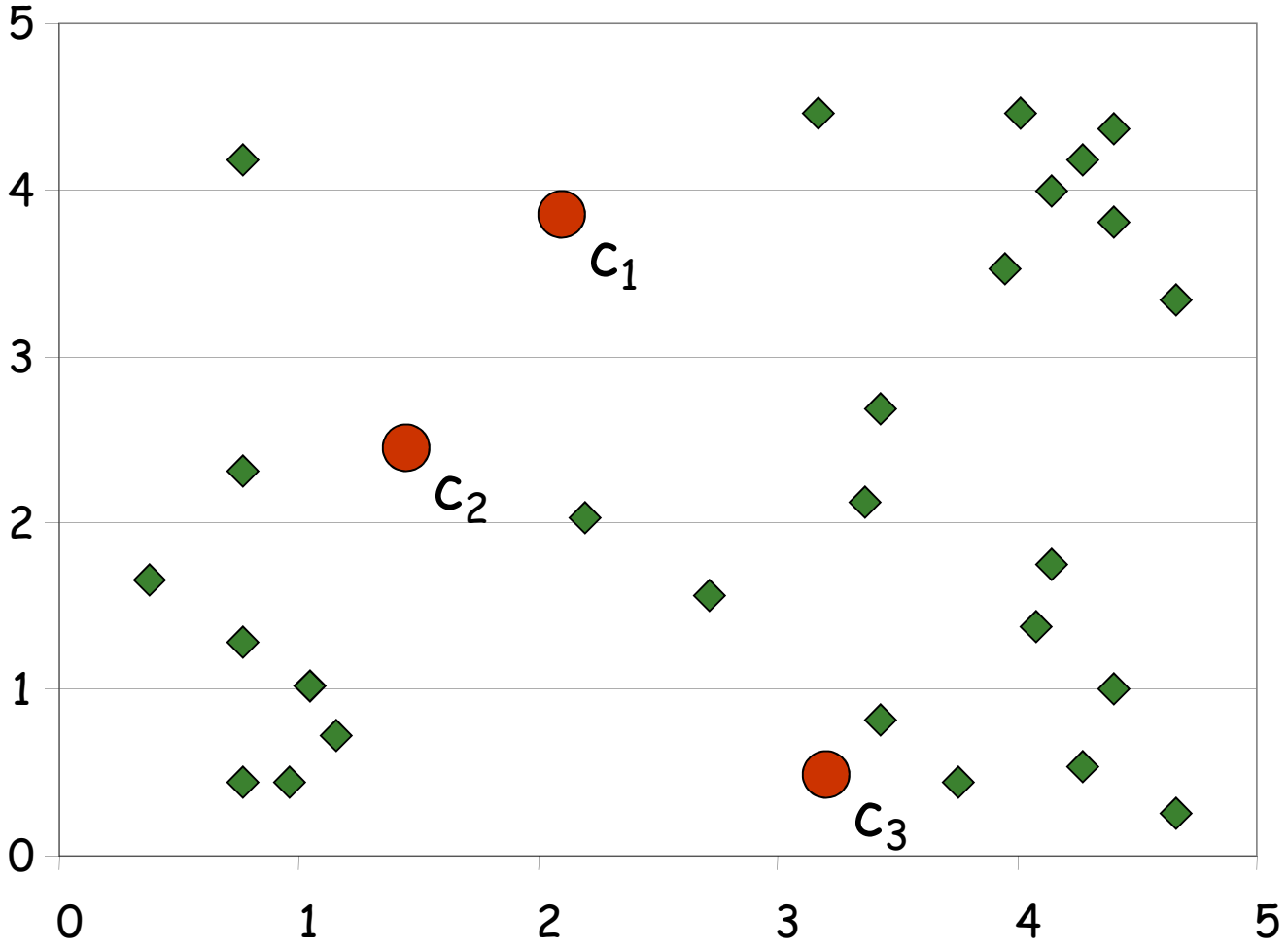
$$p = (p_1, p_2, \ldots, p_n)$$
$$q = (q_1, q_2, \ldots, q_n)$$

$$c = (\frac{p_1 + q_1}{2}, \frac{p_2 + q_2}{2}, \ldots, \frac{p_n + q_n}{2})$$



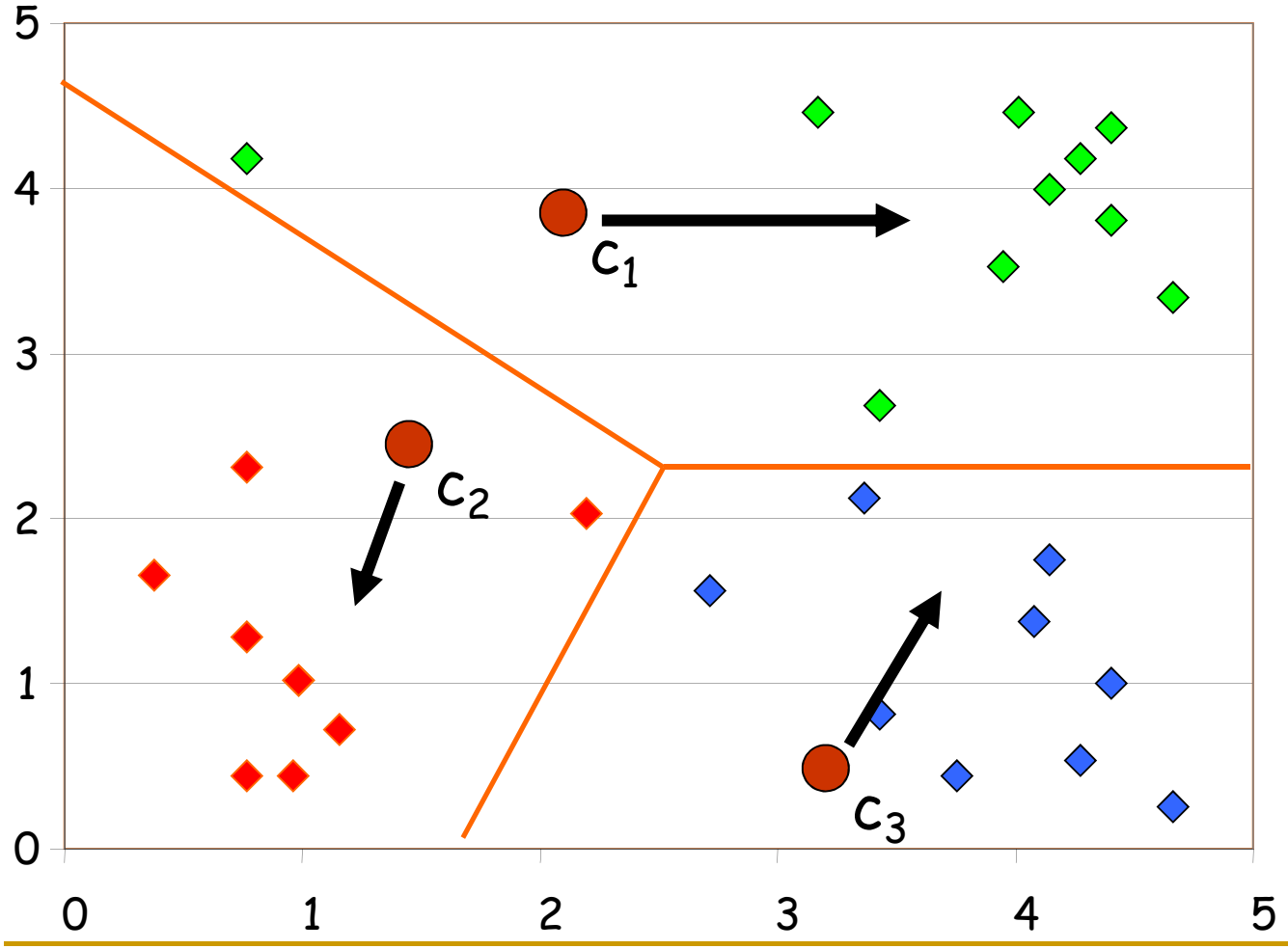
Example (in 2-D… i.e. 2 features)

initial 3 random centroïds



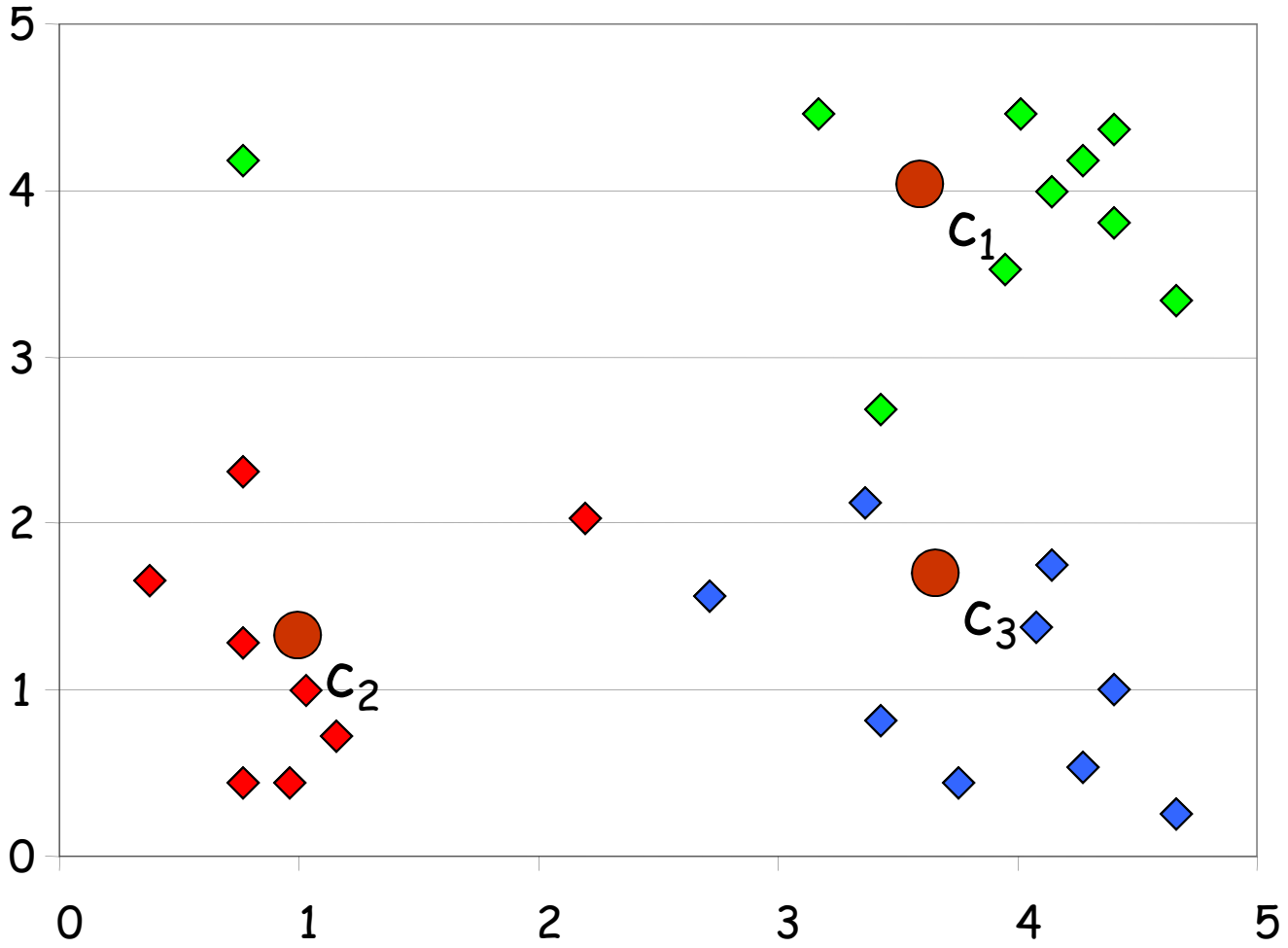
Example

partition data points to closest centroïd



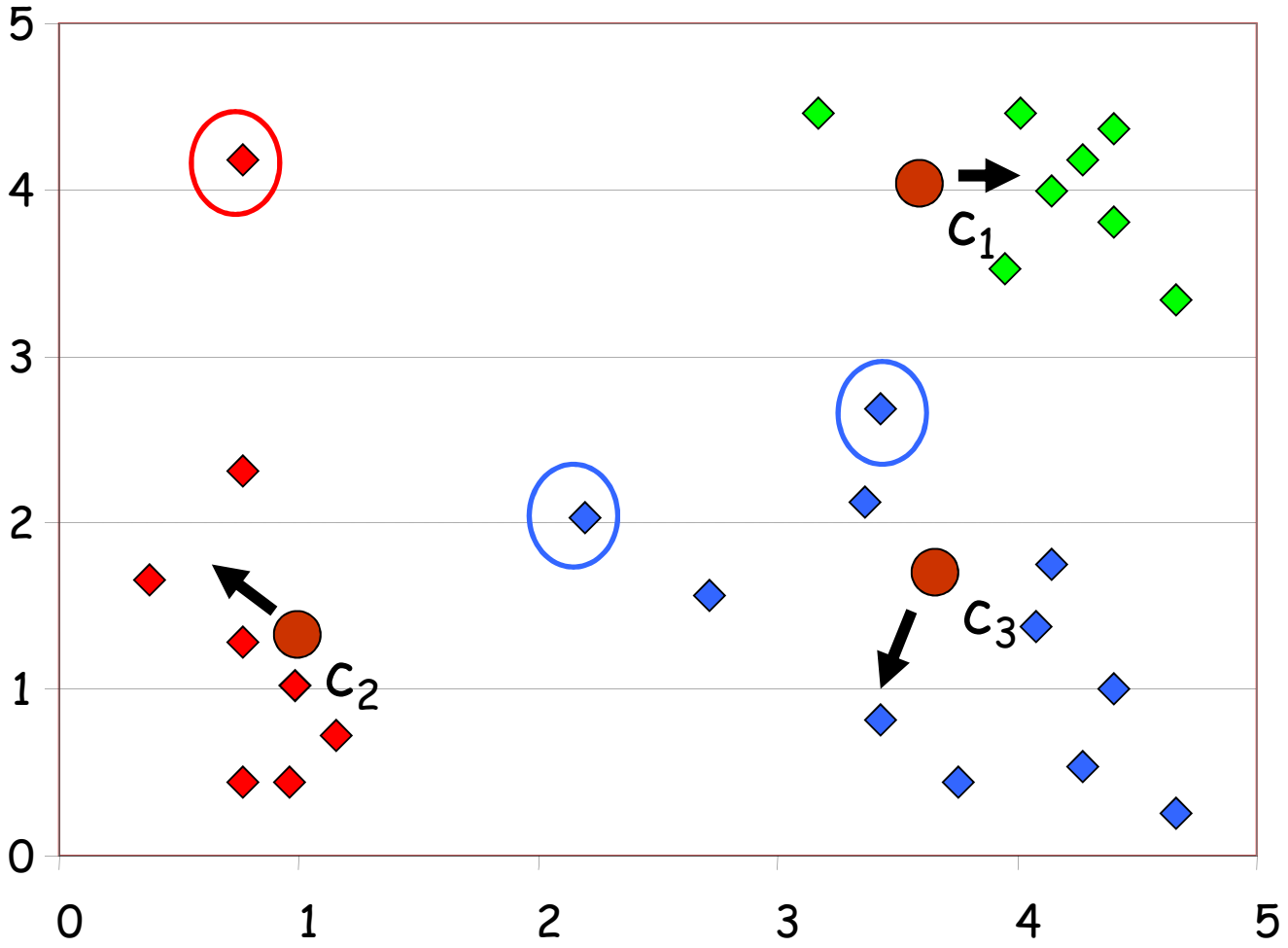
Example

re-compute new centroïds

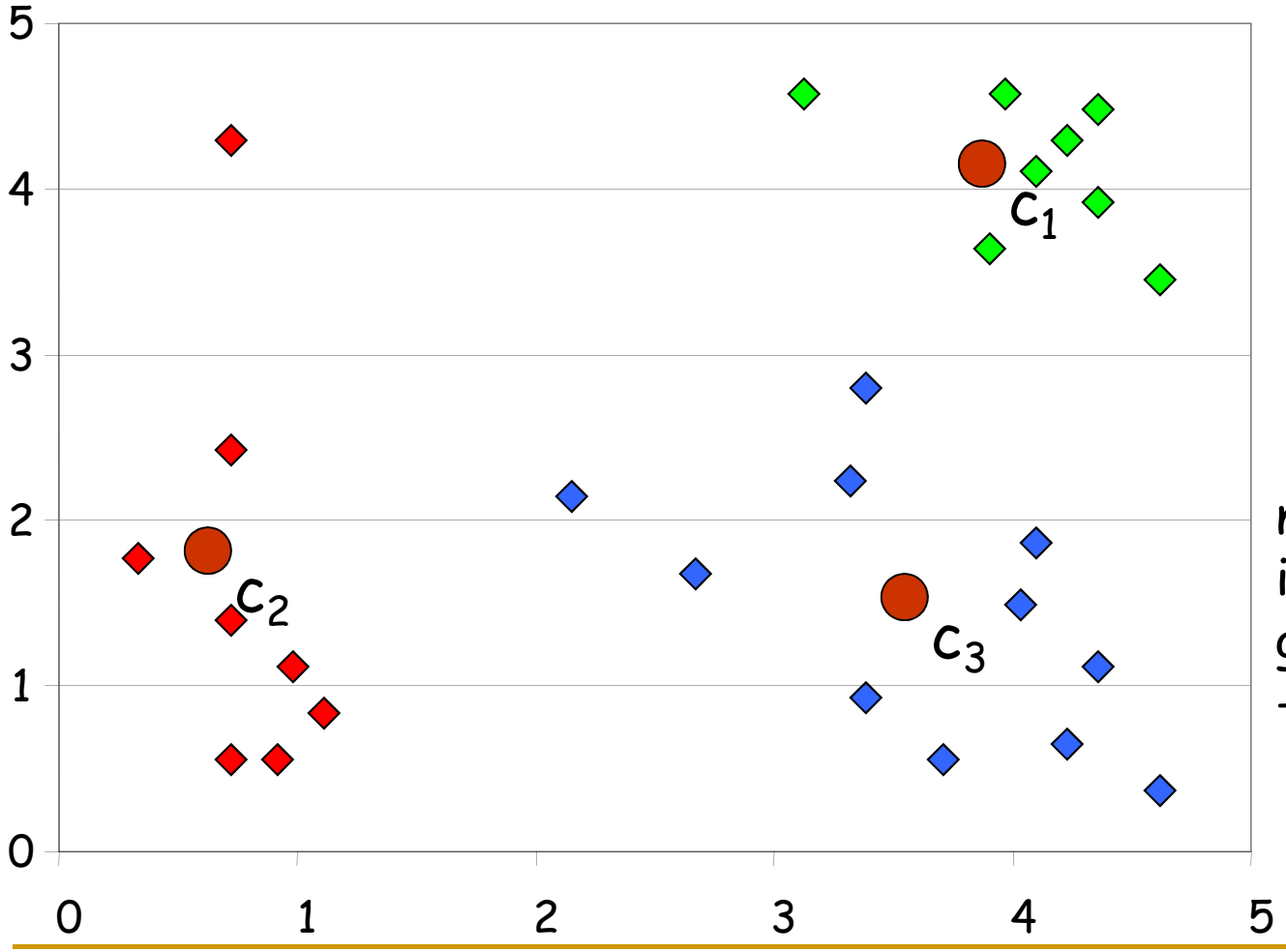


Example

re-assign data points to new closest centroïds



Example



none of the data instances change group
--> stop

Notes on k-means

- negatives:
  - does not guarantee to converge to the global optimum 不能确保最优解，因为你随机选的点不一定好用
  - very sensitive to initial choice of centroids
    - many find useless clusters… 对初始点很敏感，可以多跑几遍求最优解
  - user must set initial k
    - not easy to do… 必须想好一开始的k，
- but converges very fast!
  但是收敛得很快
- many other clustering algorithms…

Today

1. Introduction to ML ✓
2. Naïve Bayes Classification ✓
   a. Application to Spam Filtering ✓
3. Decision Trees ✓
4. (Evaluation ✓
5. Unsupervised Learning) ✓
6. Neural Networks
   a. Perceptrons
   b. Multi Layered Neural Networks

Up Next

1. Introduction to ML
2. Naïve Bayes Classification
   a. Application to Spam Filtering
3. Decision Trees
4. (Evaluation
5. Unsupervised Learning)
6. Neural Networks
   a. Perceptrons
   b. Multi Layered Neural Networks