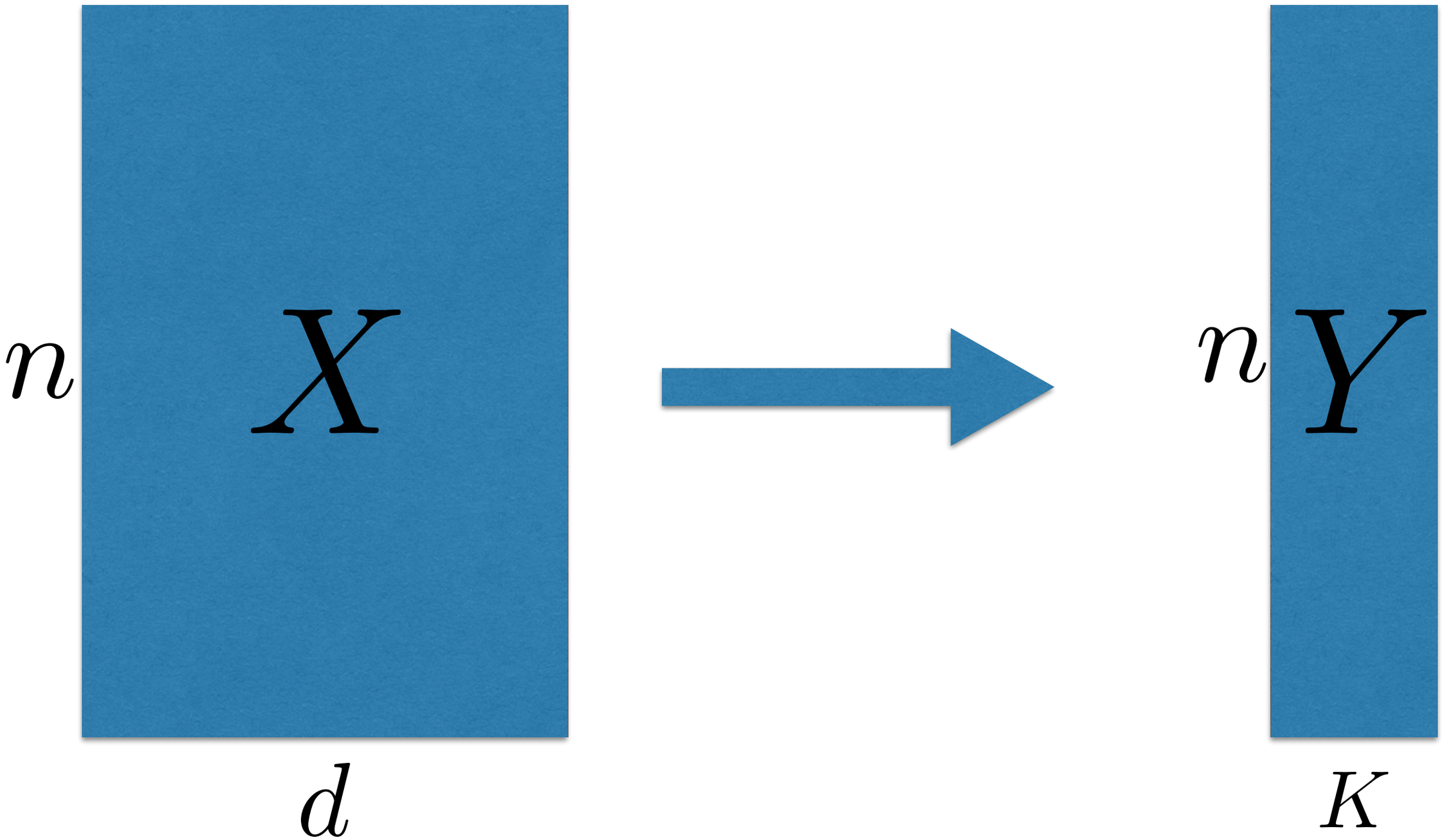
Machine Learning for Data Science (CS4786)
Lecture 8

Clustering

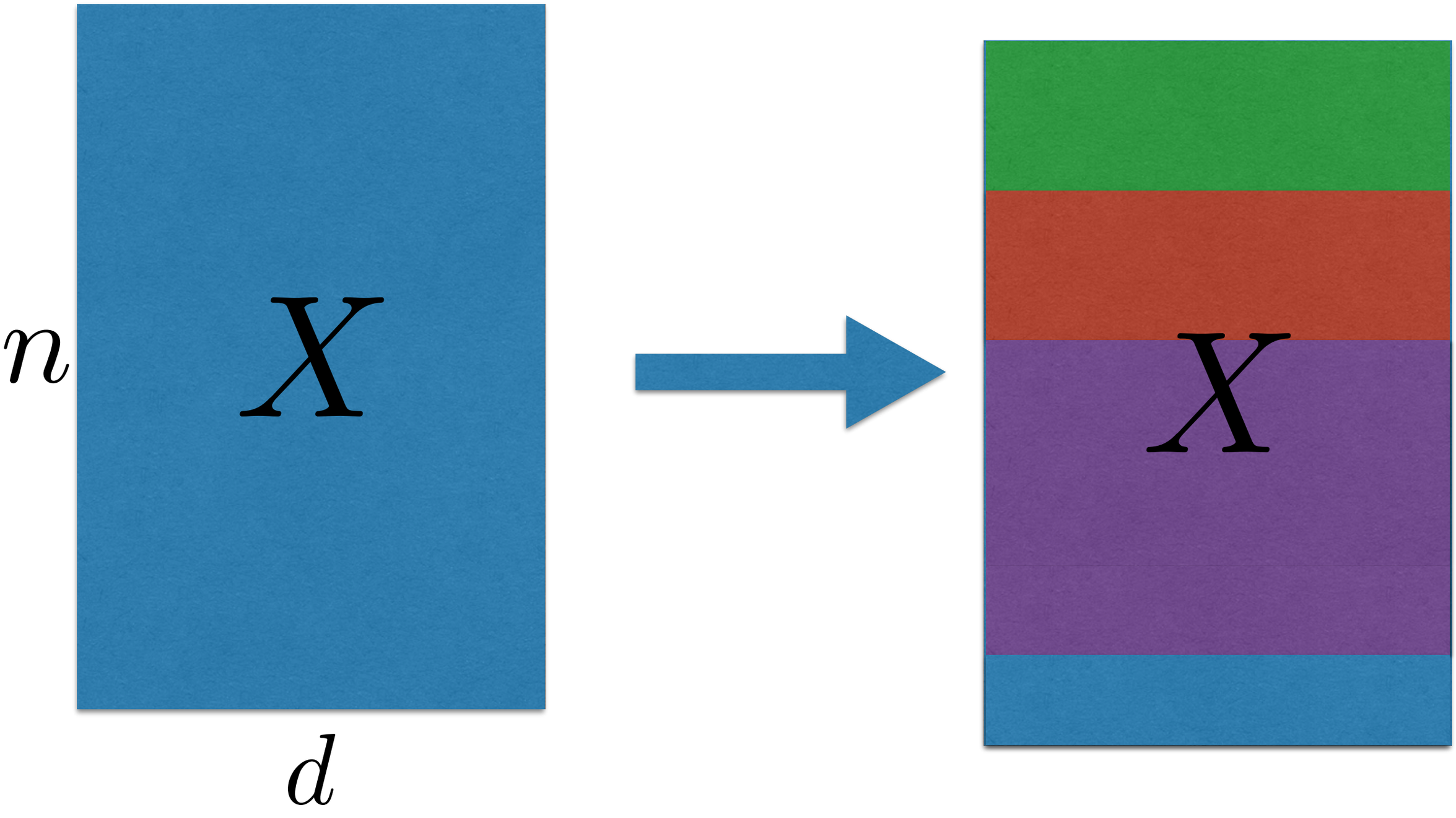
Feb 25, 2016

Course Webpage :
http://www.cs.cornell.edu/Courses/cs4786/2016sp/

Dimensionality Reduction



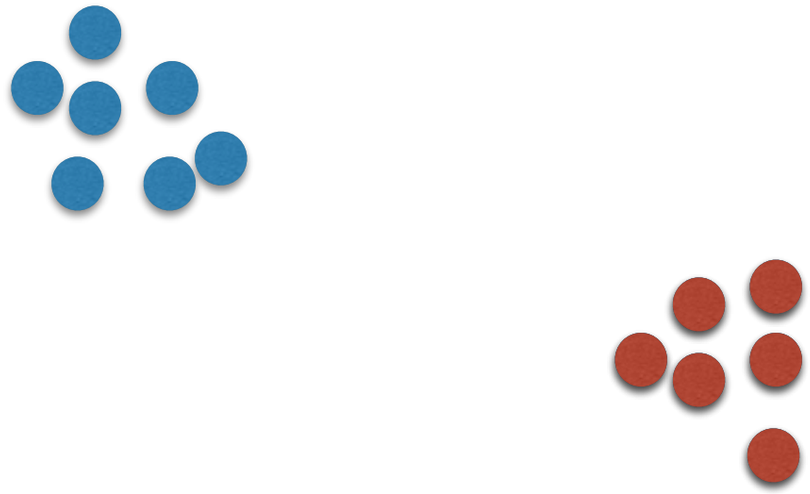
Clustering



Clustering

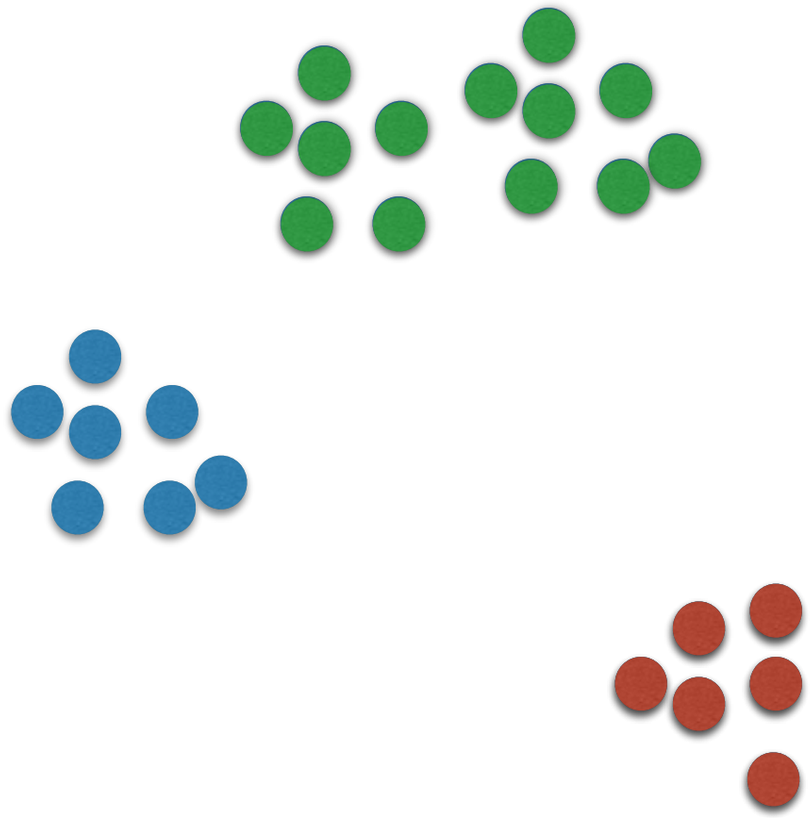
- Partition data into K disjoint groups
- Compression or Quantization
  - Compress n points into K representatives/groups
- Visualization or Understanding
  - Taxonomy: Animals Vs plants Vs Microbes, Science Vs Math Vs Social Sciences
  - Segmentation: different types of customers, students etc. Find natural groupings in data
- What this does not include: items belonging to more than one type

Examples



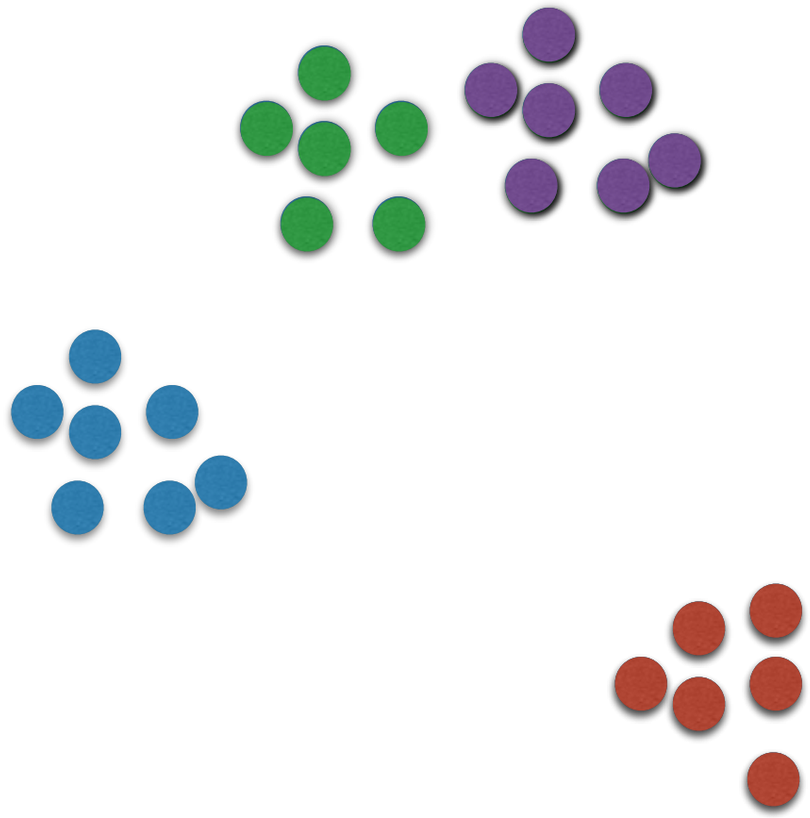
How many clusters?

Examples



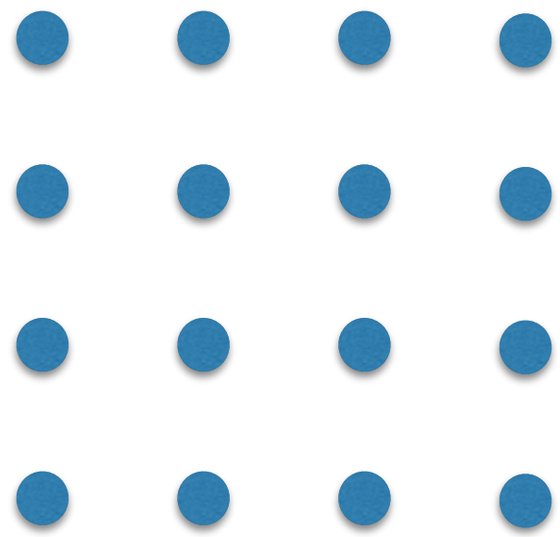
How many clusters?

Examples



How many clusters?

Examples



How many clusters?

Some Notations

- Kary clustering is a partition of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ into K groups
- For now assume the magical K is given to use
- Clustering given by $C_1, \ldots, C_K$, the partition of data points.
- Given a clustering, we shall use $c(\mathbf{x}_t)$ to denote the cluster identity of point $\mathbf{x}_t$ according to the clustering.
- Let n_j denote $|C_j|$, clearly $\sum_{j=1}^{K} n_j = n$.

Clustering Criterion

1. Minimize within-cluster scatter

$$M_1 = \sum_{j=1}^{K} \sum_{\mathbf{x}_s, \mathbf{x}_t \in C_j} \|\mathbf{x}_s - \mathbf{x}_t\|_2^2$$

2. Maximize between-cluster scatter

$$M_2 = \sum_{\mathbf{x}_s, \mathbf{x}_t : c(\mathbf{x}_s) \neq c(\mathbf{x}_t)} \|\mathbf{x}_s - \mathbf{x}_t\|_2^2$$

3. Minimize weighted within-cluster variance: $\mathbf{r}_j = \frac{1}{n_j} \sum_{\mathbf{x} \in C_j} \mathbf{x}$

$$M_3 = \sum_{j=1}^{K} n_j \sum_{\mathbf{x}_t \in C_j} \|\mathbf{x}_t - \mathbf{r}_j\|_2^2$$

4. Maximize smallest between-cluster distance

$$M_4 = \min_{\mathbf{x}_s, \mathbf{x}_t : c(\mathbf{x}_s) \neq c(\mathbf{x}_t)} \|\mathbf{x}_s - \mathbf{x}_t\|_2^2$$

5. Minimize largest within-cluster distance

$$M_4 = \max_{j \in [K]} \max_{\mathbf{x}_s, \mathbf{x}_t \in C_j} \|\mathbf{x}_s - \mathbf{x}_t\|_2^2$$

Clustering Criterion

6 Minimize within-cluster average scatter

$$M_6 = \sum_{j=1}^{K} \frac{1}{n_j} \sum_{\mathbf{x}_s, \mathbf{x}_t \in C_j} \|\mathbf{x}_s - \mathbf{x}_t\|_2^2$$

7 Minimize within-cluster variance: $\mathbf{r}_j = \frac{1}{n_j} \sum_{\mathbf{x} \in C_j} \mathbf{x}$

$$M_7 = \sum_{j=1}^{K} \sum_{\mathbf{x}_t \in C_j} \|\mathbf{x}_t - \mathbf{r}_j\|_2^2$$

How different are these various criterion?