Clustering / Unsupervised Methods

Jason Corso, Albert Chen

SUNY at Buffalo

Introduction

- Until now, we've assumed our training samples are "labeled" by their category membership.
- Methods that use labeled samples are said to be *supervised*; otherwise, they're said to be *unsupervised*.
- However:
    - Why would one even be interested in learning with unlabeled samples?
    - Is it even possible in principle to learn anything of value from unlabeled samples?

Why Unsupervised Learning?

1. Collecting and labeling a large set of sample patterns can be surprisingly costly.
   - E.g., videos are virtually free, but accurately *labeling* the video pixels is expensive and time consuming.

Why Unsupervised Learning?

1. Collecting and labeling a large set of sample patterns can be surprisingly costly.
   - E.g., videos are virtually free, but accurately *labeling* the video pixels is expensive and time consuming.
2. Extend to a larger training set by using *semi-supervised learning*.
   - Train a classifier on a small set of samples, then tune it up to make it run without supervision on a large, unlabeled set.
   - Or, in the reverse direction, let a large set of unlabeled data group automatically, then label the groupings found.

Why Unsupervised Learning?

1. Collecting and labeling a large set of sample patterns can be surprisingly costly.
   - E.g., videos are virtually free, but accurately *labeling* the video pixels is expensive and time consuming.
2. Extend to a larger training set by using *semi-supervised learning*.
   - Train a classifier on a small set of samples, then tune it up to make it run without supervision on a large, unlabeled set.
   - Or, in the reverse direction, let a large set of unlabeled data group automatically, then label the groupings found.
3. To detect the gradual change of pattern over time.

Why Unsupervised Learning?

1. Collecting and labeling a large set of sample patterns can be surprisingly costly.
   - E.g., videos are virtually free, but accurately *labeling* the video pixels is expensive and time consuming.
2. Extend to a larger training set by using *semi-supervised learning*.
   - Train a classifier on a small set of samples, then tune it up to make it run without supervision on a large, unlabeled set.
   - Or, in the reverse direction, let a large set of unlabeled data group automatically, then label the groupings found.
3. To detect the gradual change of pattern over time.
4. To find features that will then be useful for categorization.

Why Unsupervised Learning?

1. Collecting and labeling a large set of sample patterns can be surprisingly costly.
   - E.g., videos are virtually free, but accurately *labeling* the video pixels is expensive and time consuming.
2. Extend to a larger training set by using *semi-supervised learning*.
   - Train a classifier on a small set of samples, then tune it up to make it run without supervision on a large, unlabeled set.
   - Or, in the reverse direction, let a large set of unlabeled data group automatically, then label the groupings found.
3. To detect the gradual change of pattern over time.
4. To find features that will then be useful for categorization.
5. To gain insight into the nature or structure of the data during the early stages of an investigation.

Data Clustering

Source: A. K. Jain and R. C. Dubes. Alg. for Clustering Data, Prentiice Hall, 1988.

- What is data clustering?
  - Grouping of objects into meaningful categories
  - Given a **representation** of N objects, find k clusters based on a measure of **similarity**.

Data Clustering

Source: A. K. Jain and R. C. Dubes. Alg. for Clustering Data, Prentiice Hall, 1988.

- What is data clustering?
  - Grouping of objects into meaningful categories
  - Given a **representation** of N objects, find k clusters based on a measure of **similarity**.
- Why data clustering?
  - Natural Classification: degree of similarity among forms.
  - Data exploration: discover underlying structure, generate hypotheses, detect anomalies.
  - Compression: for organizing data.
  - Applications: can be used by any scientific field that collects data!

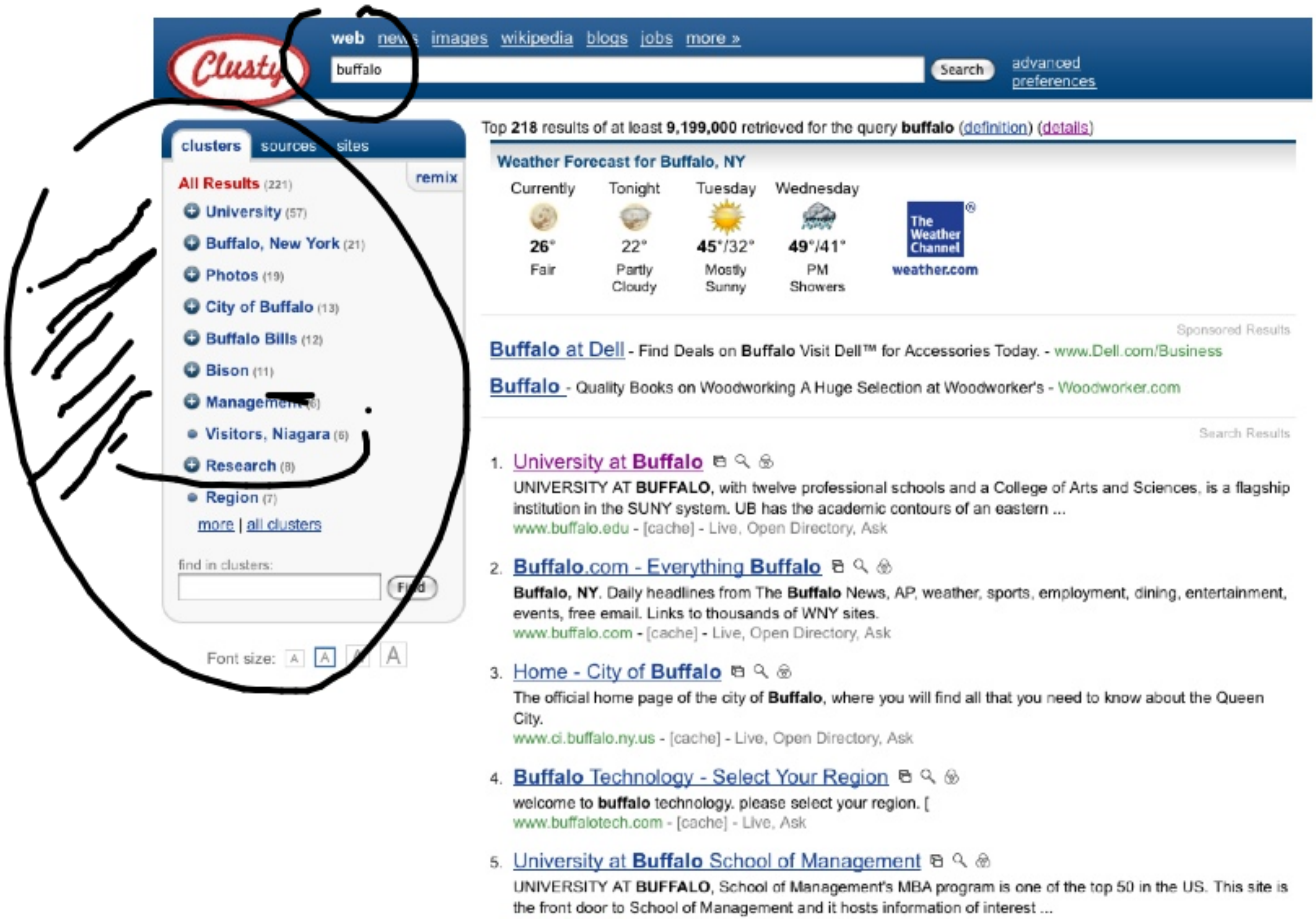
Data Clustering

Source: A. K. Jain and R. C. Dubes. Alg. for Clustering Data, Prentiice Hall, 1988.

- What is data clustering?
  - Grouping of objects into meaningful categories
  - Given a **representation** of N objects, find k clusters based on a measure of **similarity**.
- Why data clustering?
  - Natural Classification: degree of similarity among forms.
  - Data exploration: discover underlying structure, generate hypotheses, detect anomalies.
  - Compression: for organizing data.
  - Applications: can be used by any scientific field that collects data!
- Google Scholar: 1500 clustering papers in 2007 alone!

E.g.: Structure Discovering via Clustering

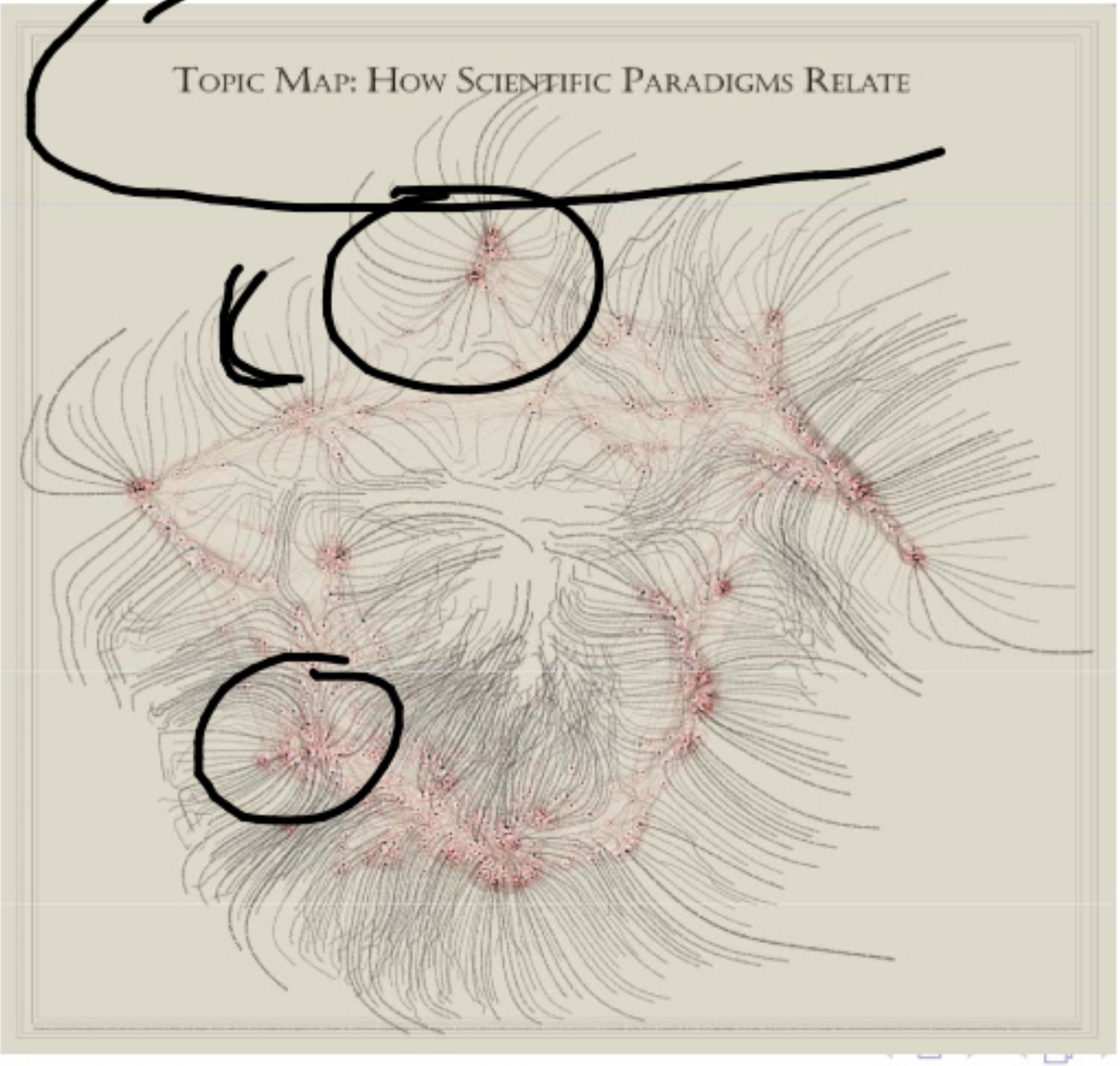
Source: http://clusty.com



E.g.: Topic Discovery

Source: Map of Science, Nature, 2006

- 800,000 scientific papers clustered into 776 topics based on how often the papers were cited together by authors of other papers



Data Clustering - Formal Definition

- Given a set of N unlabeled examples $D = x_1, x_2, ..., x_N$ in a d-dimensional feature space, D is partitioned into a number of disjoint subsets D_j's:

$$D = \cup_{j=1}^{k} D_j \quad \text{where } D_i \cap D_j = \emptyset, i \neq j \ , \tag{1}$$

where the points in each subset are similar to each other according to a given criterion ϕ.

Data Clustering - Formal Definition

- Given a set of N unlabeled examples $D = x_1, x_2, ..., x_N$ in a d-dimensional feature space, D is partitioned into a number of disjoint subsets D_j's:

$$D = \cup_{j=1}^{k} D_j \quad \text{where } D_i \cap D_j = \emptyset, i \neq j \quad , \tag{1}$$

  where the points in each subset are similar to each other according to a given criterion ϕ.

- A partition is denoted by

$$\pi = (D_1, D_2, ..., D_k) \tag{2}$$

  and the problem of data clustering is thus formulated as

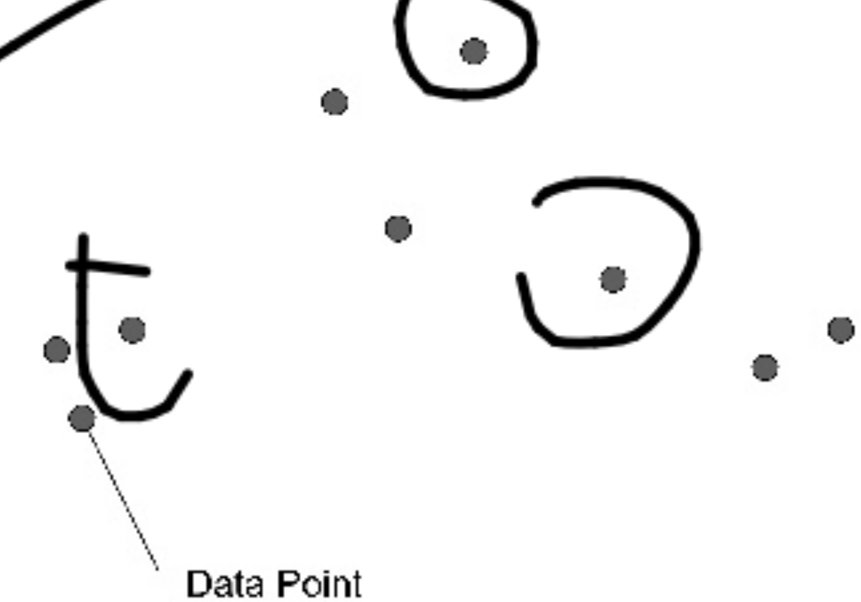
$$\pi^* = \operatorname*{argmin}_{\pi} f(\pi) \quad , \tag{3}$$

  where $f(\cdot)$ is formulated according to ϕ.

k-Means Clustering

Source: D. Aurthor and S. Vassilvitskii. k-Means++: The Advantages of Careful Seeding

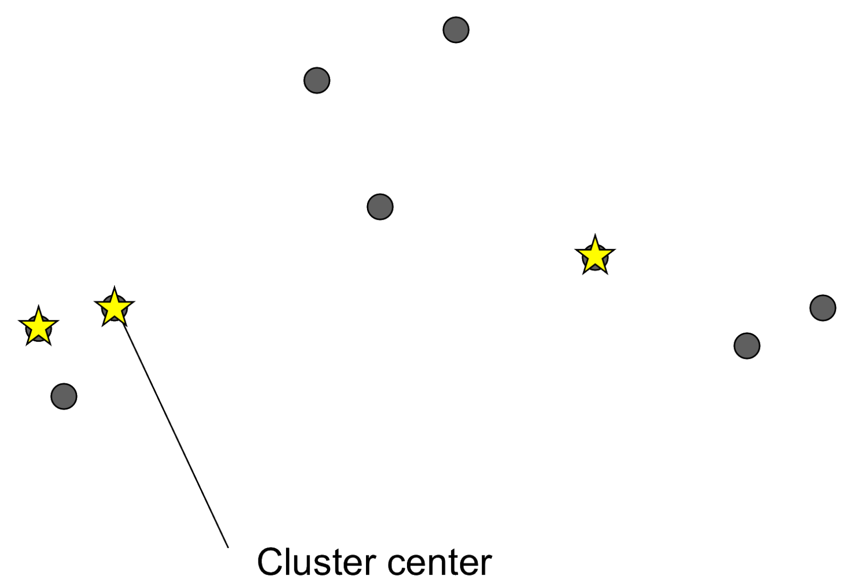
- Randomly initialize $\mu_1, \mu_2, ..., \mu_c$
- Repeat until no change in μ_i:
  - Classify N samples according to nearest μ_i
  - Recompute μ_i



k-Means Clustering

Source: D. Aurthor and S. Vassilvitskii. k-Means++: The Advantages of Careful Seeding

- Randomly initialize $\mu_1, \mu_2, ..., \mu_c$
- Repeat until no change in μ_i:
  - Classify N samples according to nearest μ_i
  - Recompute μ_i

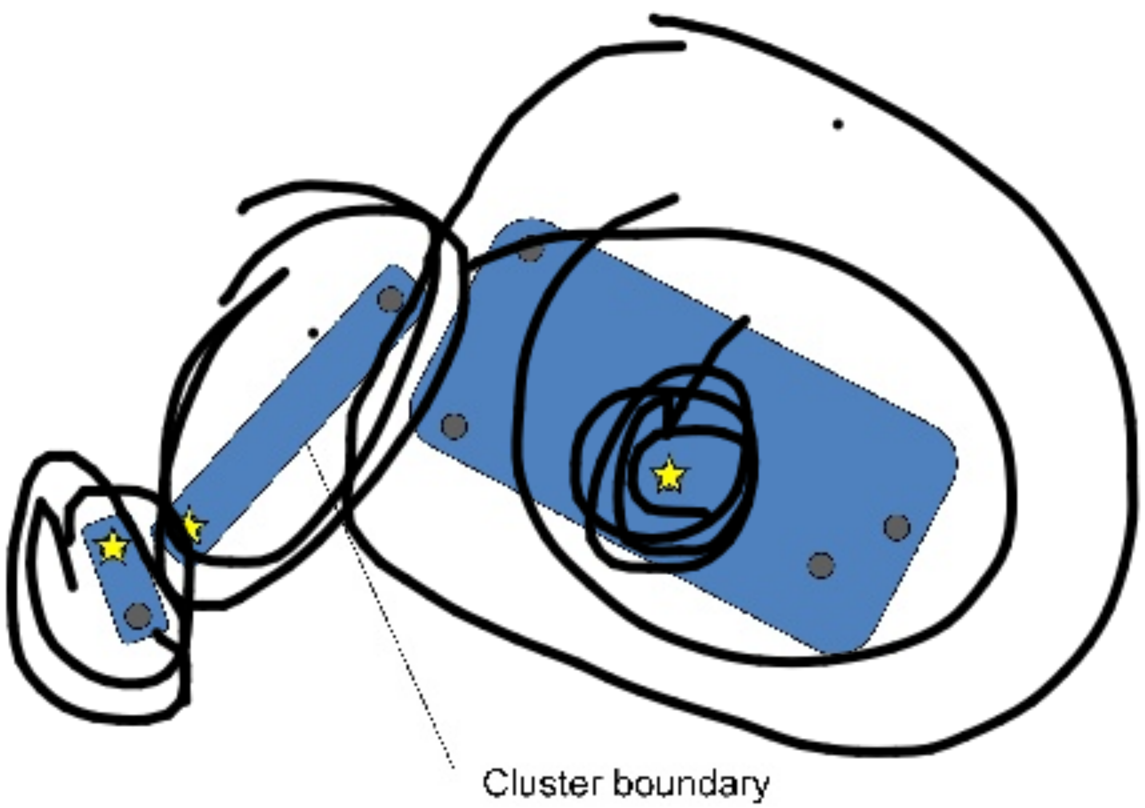


First choose *k* arbitrary centers

k-Means Clustering

Source: D. Aurthor and S. Vassilvitskii. k-Means++: The Advantages of Careful Seeding

- Randomly initialize $\mu_1, \mu_2, ..., \mu_c$
- Repeat until no change in μ_i:
  - Classify N samples according to nearest μ_i
  - Recompute μ_i

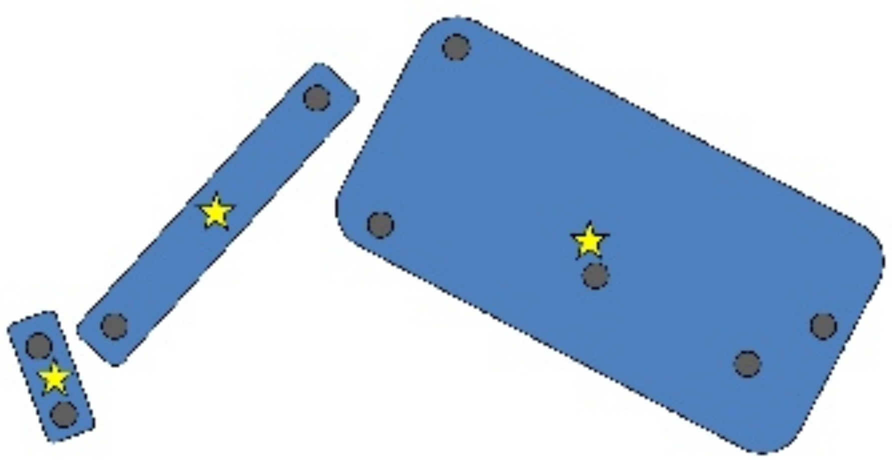


Assign points to closest centers

k-Means Clustering

Source: D. Aurthor and S. Vassilvitskii. k-Means++: The Advantages of Careful Seeding

- Randomly initialize $\mu_1, \mu_2, ..., \mu_c$
- Repeat until no change in μ_i:
  - Classify N samples according to nearest μ_i
  - Recompute μ_i

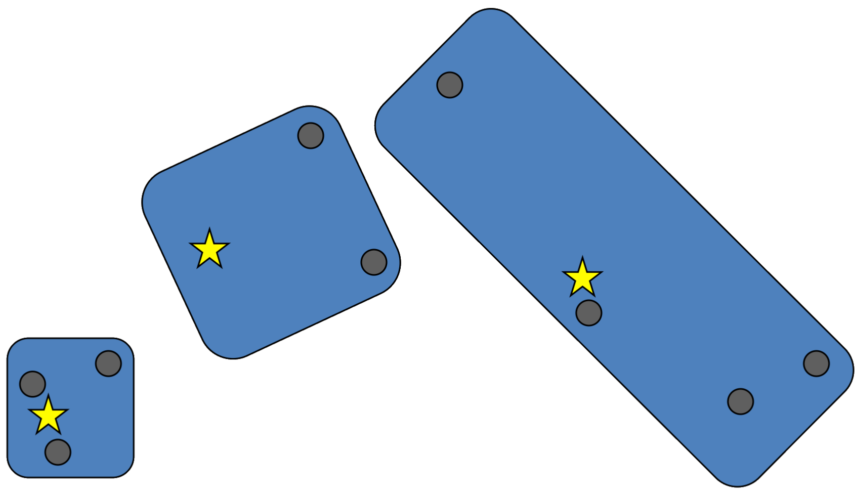


Recompute centers

k-Means Clustering

Source: D. Aurthor and S. Vassilvitskii. k-Means++: The Advantages of Careful Seeding

- Randomly initialize $\mu_1, \mu_2, ..., \mu_c$
- Repeat until no change in μ_i:
  - Classify N samples according to nearest μ_i
  - Recompute μ_i

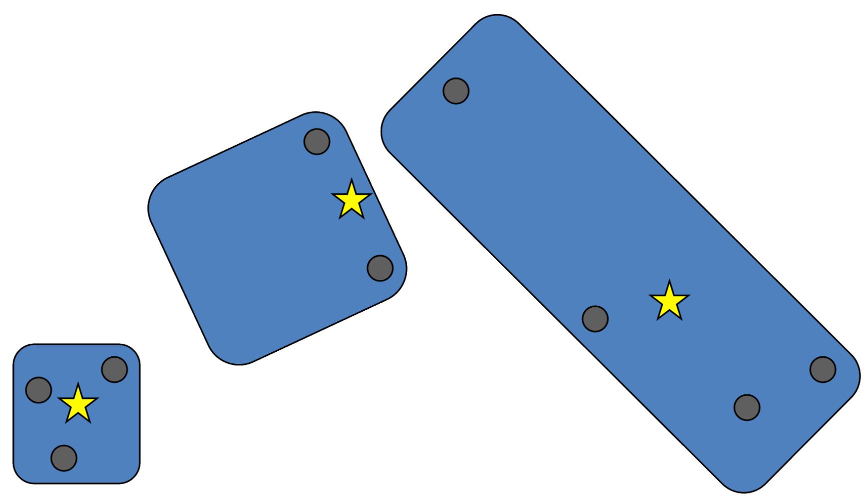


Assign points to closest centers

k-Means Clustering

Source: D. Aurthor and S. Vassilvitskii. k-Means++: The Advantages of Careful Seeding

- Randomly initialize $\mu_1, \mu_2, ..., \mu_c$
- Repeat until no change in μ_i:
  - Classify N samples according to nearest μ_i
  - Recompute μ_i

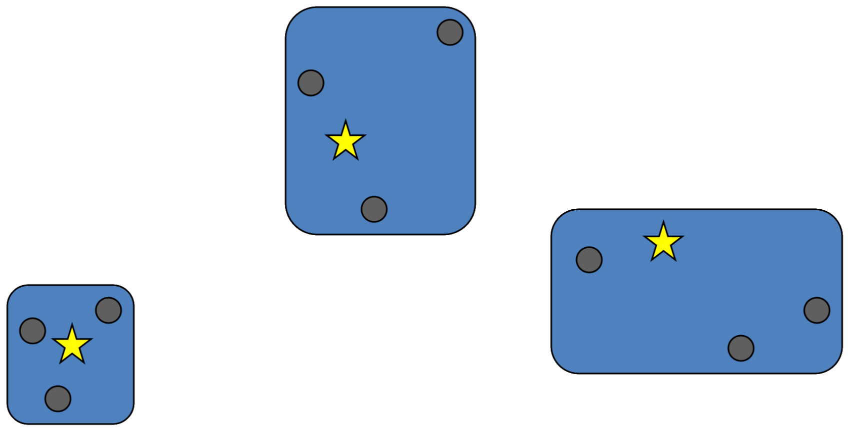


Recompute centers

k-Means Clustering

Source: D. Aurthor and S. Vassilvitskii. k-Means++: The Advantages of Careful Seeding

- Randomly initialize $\mu_1, \mu_2, ..., \mu_c$
- Repeat until no change in μ_i:
  - Classify N samples according to nearest μ_i
  - Recompute μ_i

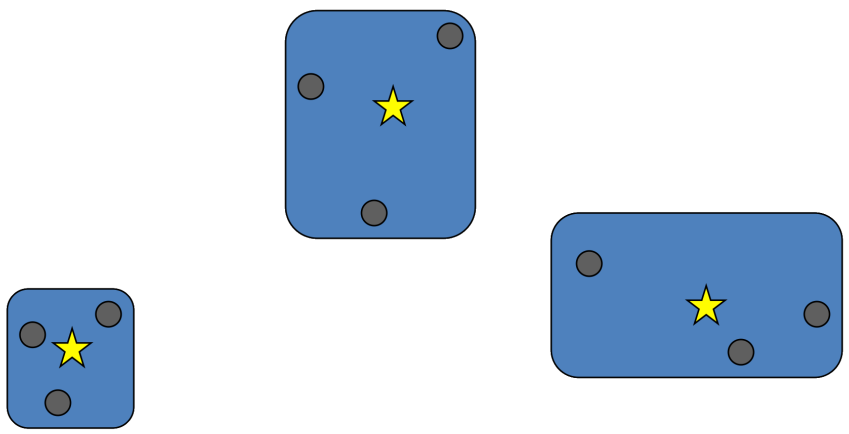


Assign points to closest centers

k-Means Clustering

Source: D. Aurthor and S. Vassilvitskii. k-Means++: The Advantages of Careful Seeding

- Randomly initialize $\mu_1, \mu_2, ..., \mu_c$
- Repeat until no change in μ_i:
  - Classify N samples according to nearest μ_i
  - Recompute μ_i

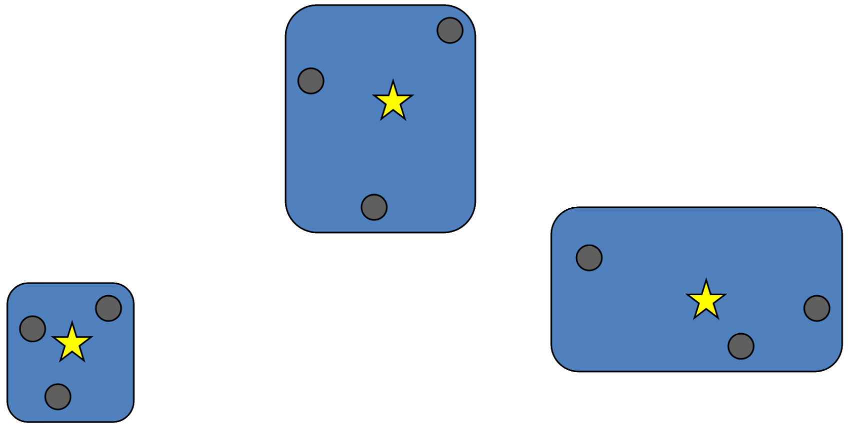


Recompute centers

k-Means Clustering

Source: D. Aurthor and S. Vassilvitskii. k-Means++: The Advantages of Careful Seeding

- Randomly initialize $\mu_1, \mu_2, ..., \mu_c$
- Repeat until no change in μ_i:
  - Classify N samples according to nearest μ_i
  - Recompute μ_i



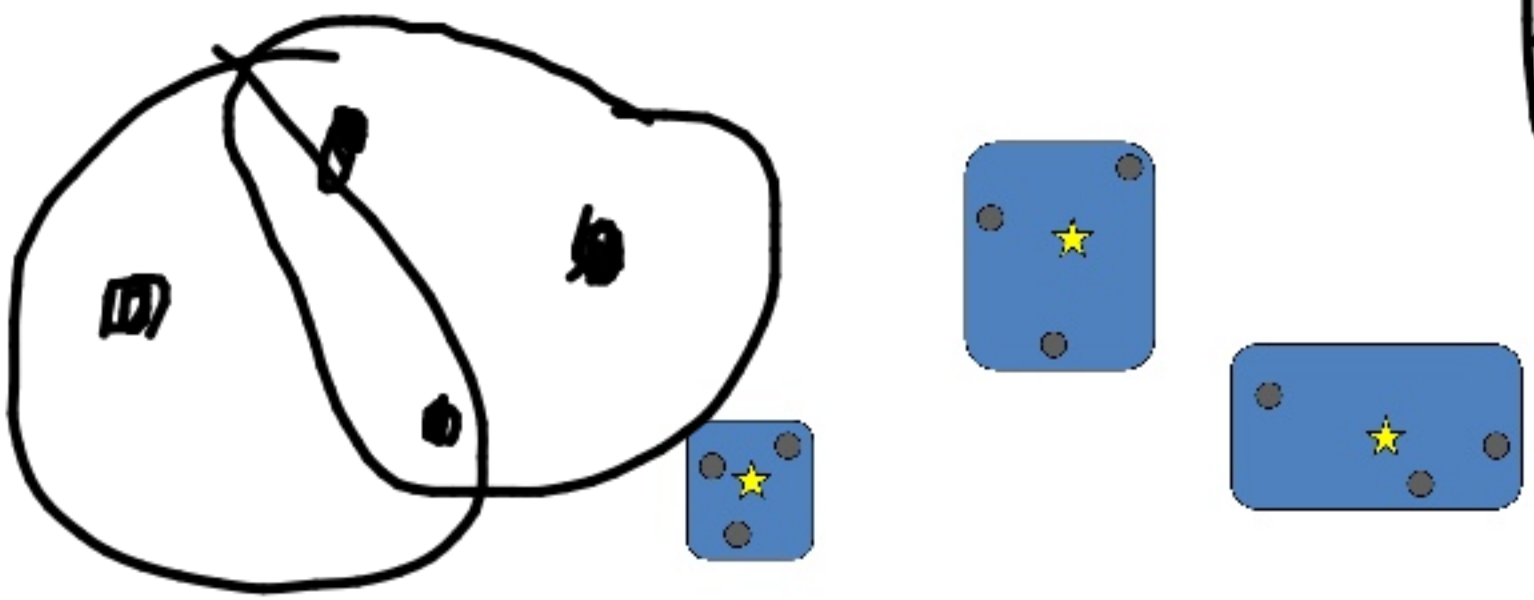
Points already assigned to nearest centers; Algorithm ends

k-Means Clustering

Source: D. Aurthor and S. Vassilvitskii. k-Means++: The Advantages of Careful Seeding

- Randomly initialize $\mu_1, \mu_2, ..., \mu_c$
- Repeat until no change in μ_i:
  - Classify N samples according to nearest μ_i
  - Recompute μ_i

k-Medoids



Points already assigned to nearest centers: Algorithm ends

k-Means++ Clustering

Source: D. Aurthor and S. Vassilvitskii. k-Means++: The Advantages of Careful Seeding

- Choose starting centers iteratively.
- Let $D(x)$ be the distance from x to the nearest existing center, take x as new center with probability $\propto D(x)^2$.
- Repeat until no change in μ_i:
  - Classify N samples according to nearest μ_i
  - Recompute μ_i
- (refer to the slides by D. Aurthor and S. Vassolvitskii for details)

User's Dilemma

Source: R. Dubes and A. K. Jain, Clustering Techniques: User's Dilemma, PR 1976

1. What is a cluster?
2. How to define pair-wise similarity?

User's Dilemma

Source: R. Dubes and A. K. Jain, Clustering Techniques: User's Dilemma, PR 1976

1. What is a cluster?
2. How to define pair-wise similarity?
3. Which features and normalization scheme?

User's Dilemma

Source: R. Dubes and A. K. Jain, Clustering Techniques: User's Dilemma, PR 1976

1. What is a cluster?
2. How to define pair-wise similarity?
3. Which features and normalization scheme?
4. How many clusters?

User's Dilemma

Source: R. Dubes and A. K. Jain, Clustering Techniques: User's Dilemma, PR 1976

1. What is a cluster?
2. How to define pair-wise similarity?
3. Which features and normalization scheme?
4. How many clusters?
5. Which clustering method?
6. Are the discovered clusters and partition valid?

User's Dilemma

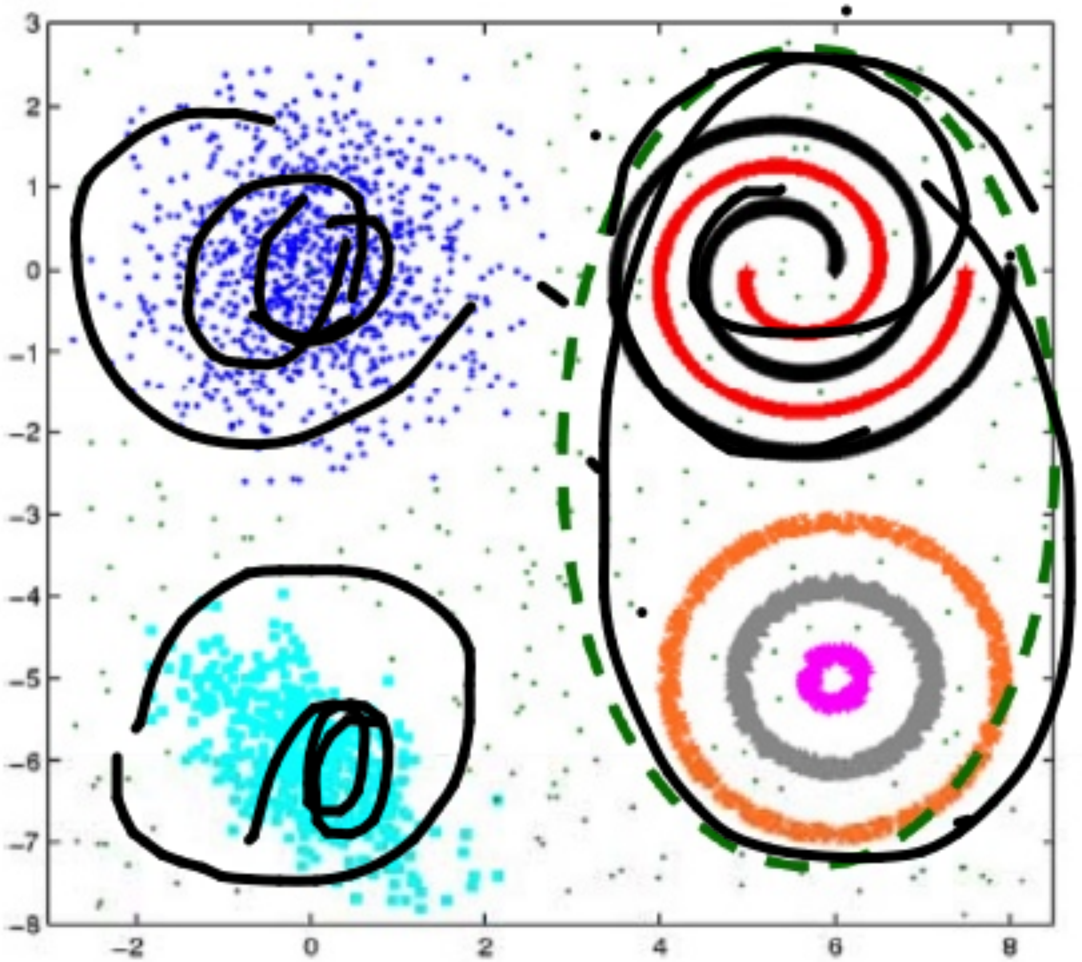
Source: R. Dubes and A. K. Jain, Clustering Techniques: User's Dilemma, PR 1976

1. What is a cluster?
2. How to define pair-wise similarity?
3. Which features and normalization scheme?
4. How many clusters?
5. Which clustering method?
6. Are the discovered clusters and partition valid?
7. Does the data have any clustering tendency?

Cluster Similarity?

Source: R. Dubes and A. K. Jain, Clustering Techniques: User's Dilemma, PR 1976

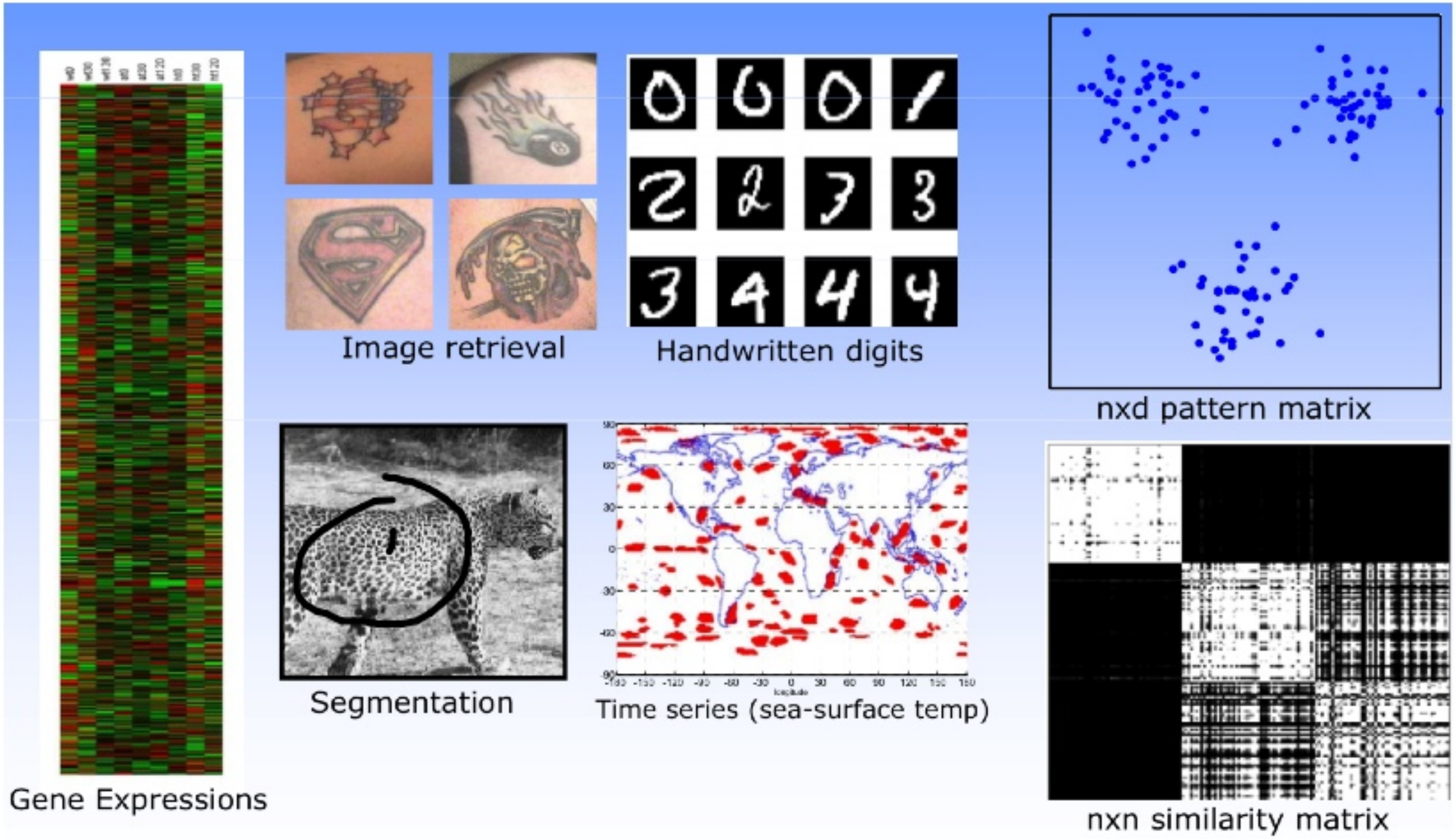
- Compact Clusters
  - Within-cluster **distance** < between-cluster connectivity
- Connected Clusters
  - Within-cluster **connectivity** > between-cluster connectivity
- Ideal cluster: **compact** and **isolated**.



Representation (features)?

Source: R. Dubes and A. K. Jain, Clustering Techniques: User's Dilemma, PR 1976

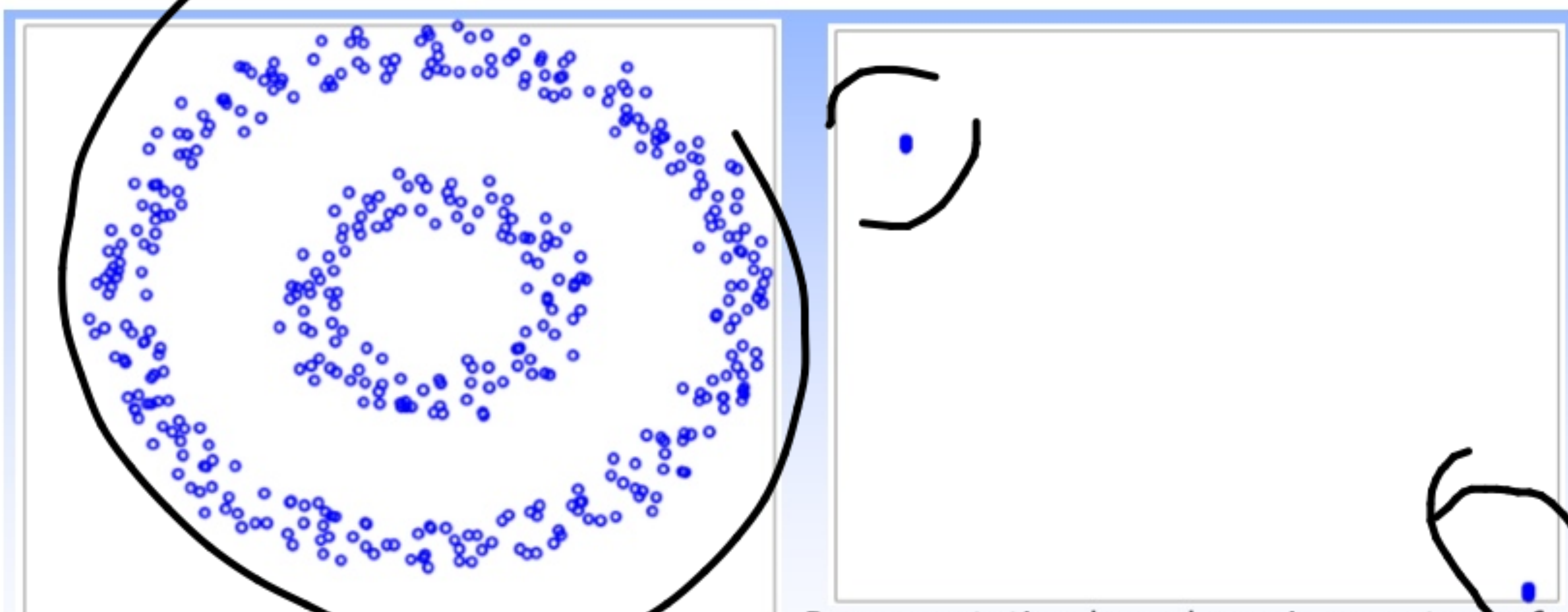
- There's no universal representation; they're domain dependent.



Good Representation

Source: R. Dubes and A. K. Jain, Clustering Techniques: User's Dilemma, PR 1976

- A good representation leads to compact and isolated clusters.



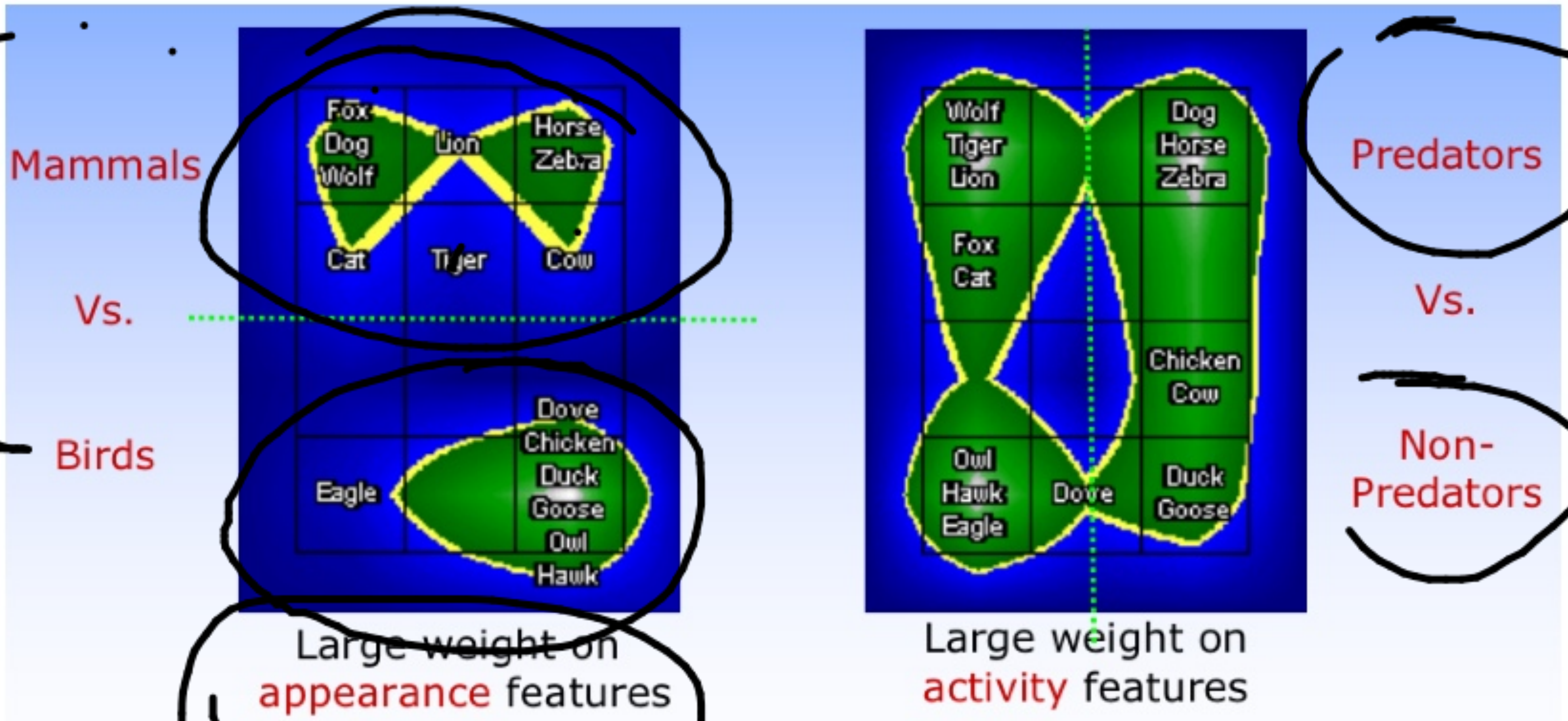
Points in given 2D space

Representation based on eigenvectors of RBF kernel

How do we weigh the features?

Source: R. Dubes and A. K. Jain, Clustering Techniques: User's Dilemma, PR 1976

- Two different meaningful groupings produced by different weighting schemes.

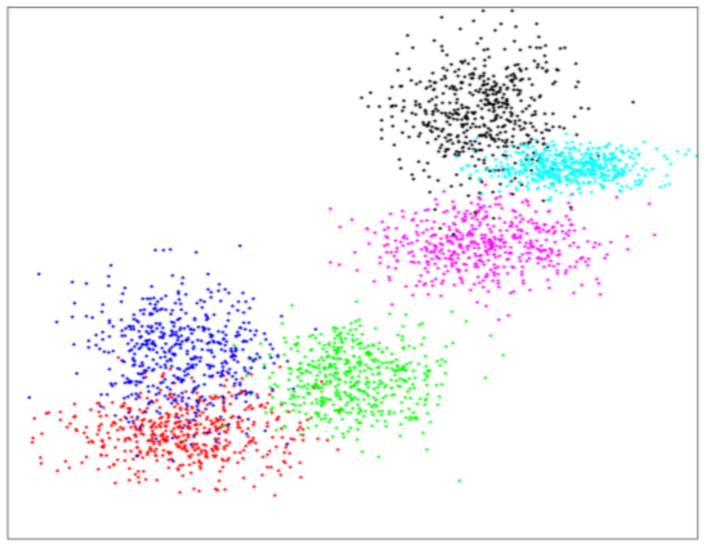


http://www.ofai.at/~elias.pampalk/kdd03/animals/

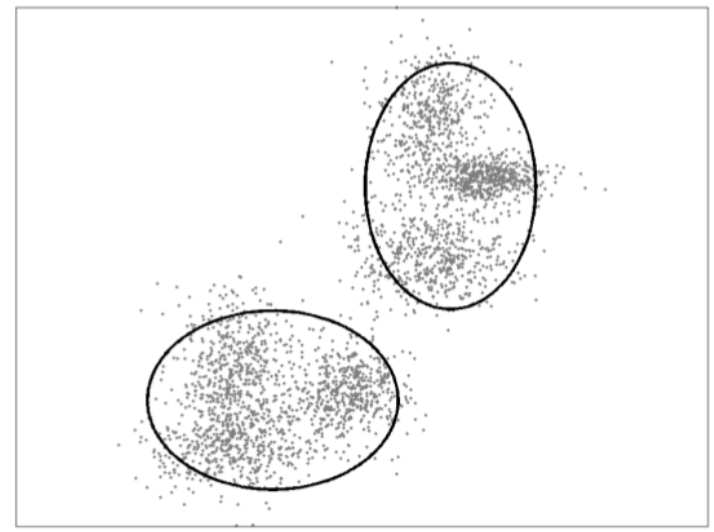
How do we decide the Number of Clusters?

Source: R. Dubes and A. K. Jain, Clustering Techniques: User's Dilemma, PR 1976

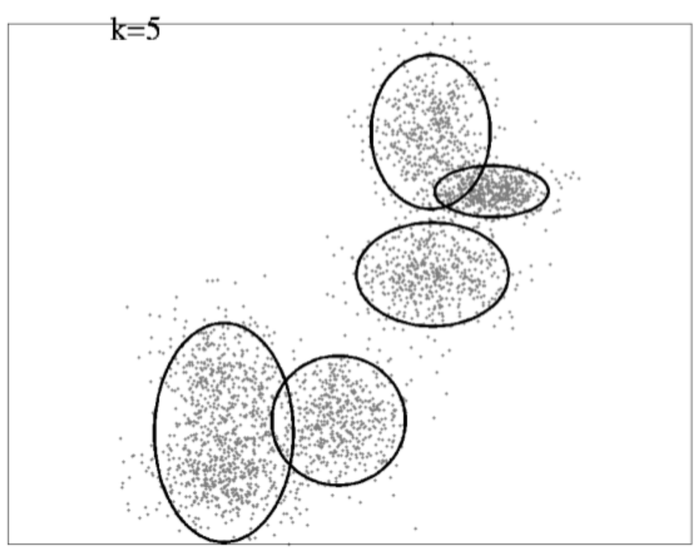
- The samples are generated by 6 independent classes, yet:



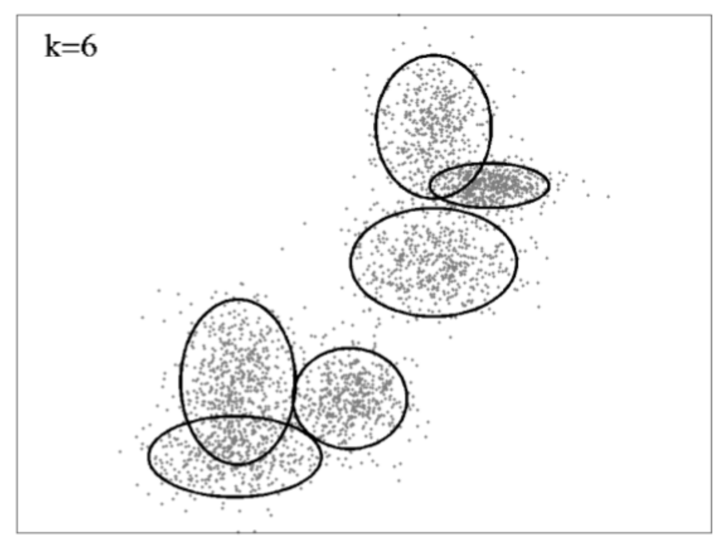
ground truth



$k = 2$



$k = 5$

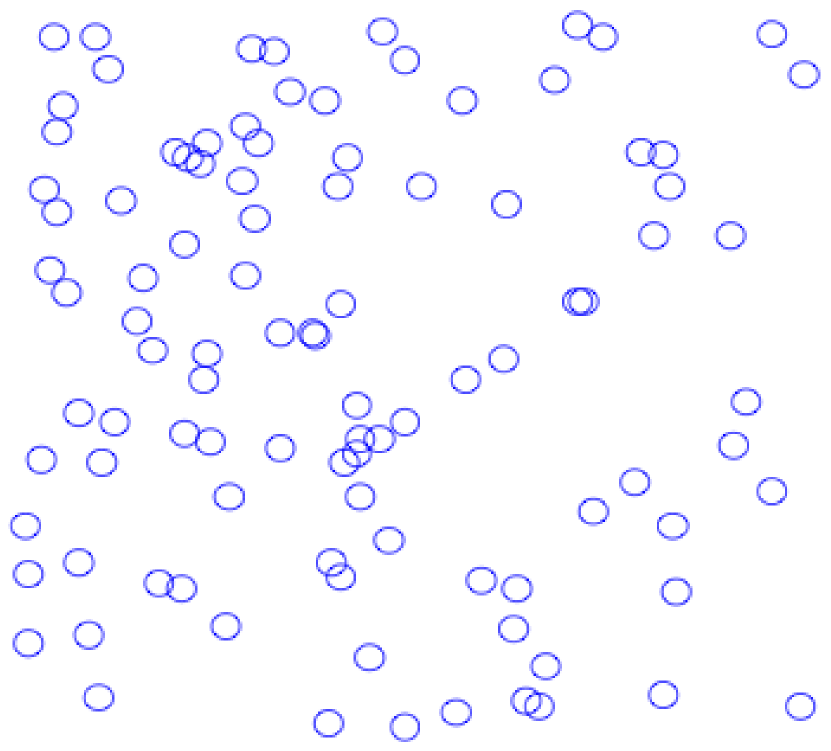


$k = 6$

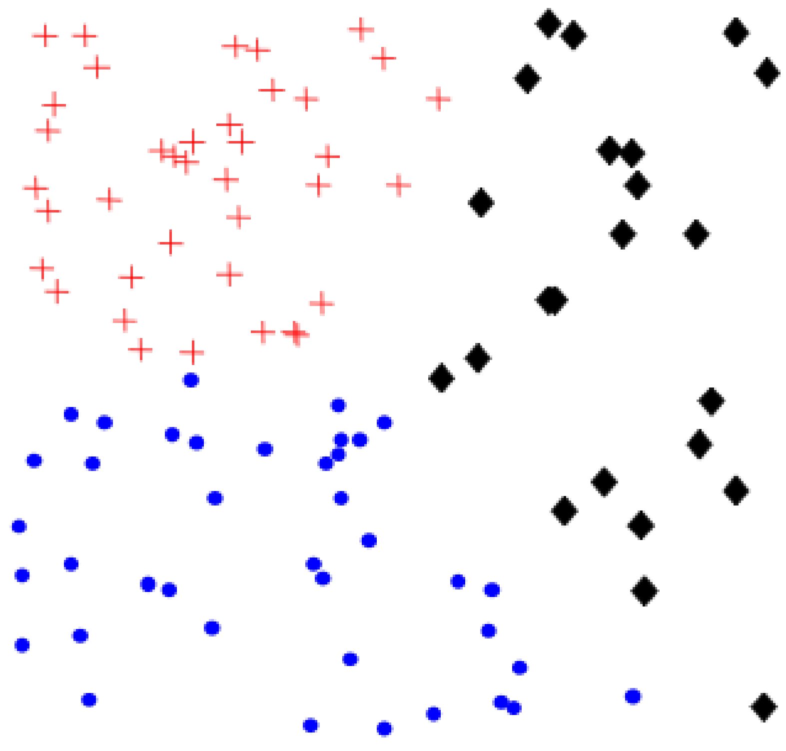
Cluster Validity

Source: R. Dubes and A. K. Jain, Clustering Techniques: User's Dilemma, PR 1976

- Clustering algorithms find clusters, even if there are no **natural** clusters in the data.



100 2D uniform data points

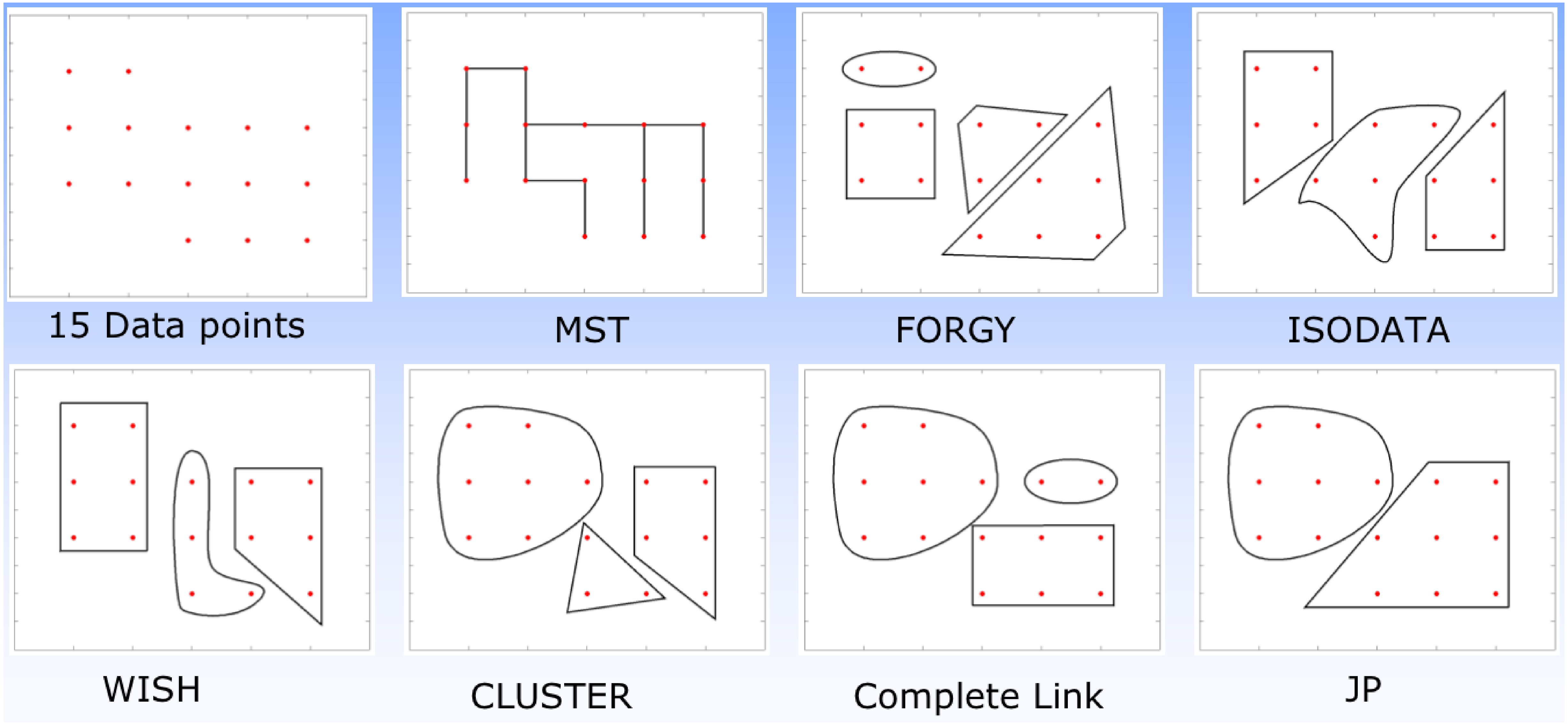


k-Means with k=3

Comparing Clustering Methods

Source: R. Dubes and A. K. Jain, Clustering Techniques: User's Dilemma, PR 1976

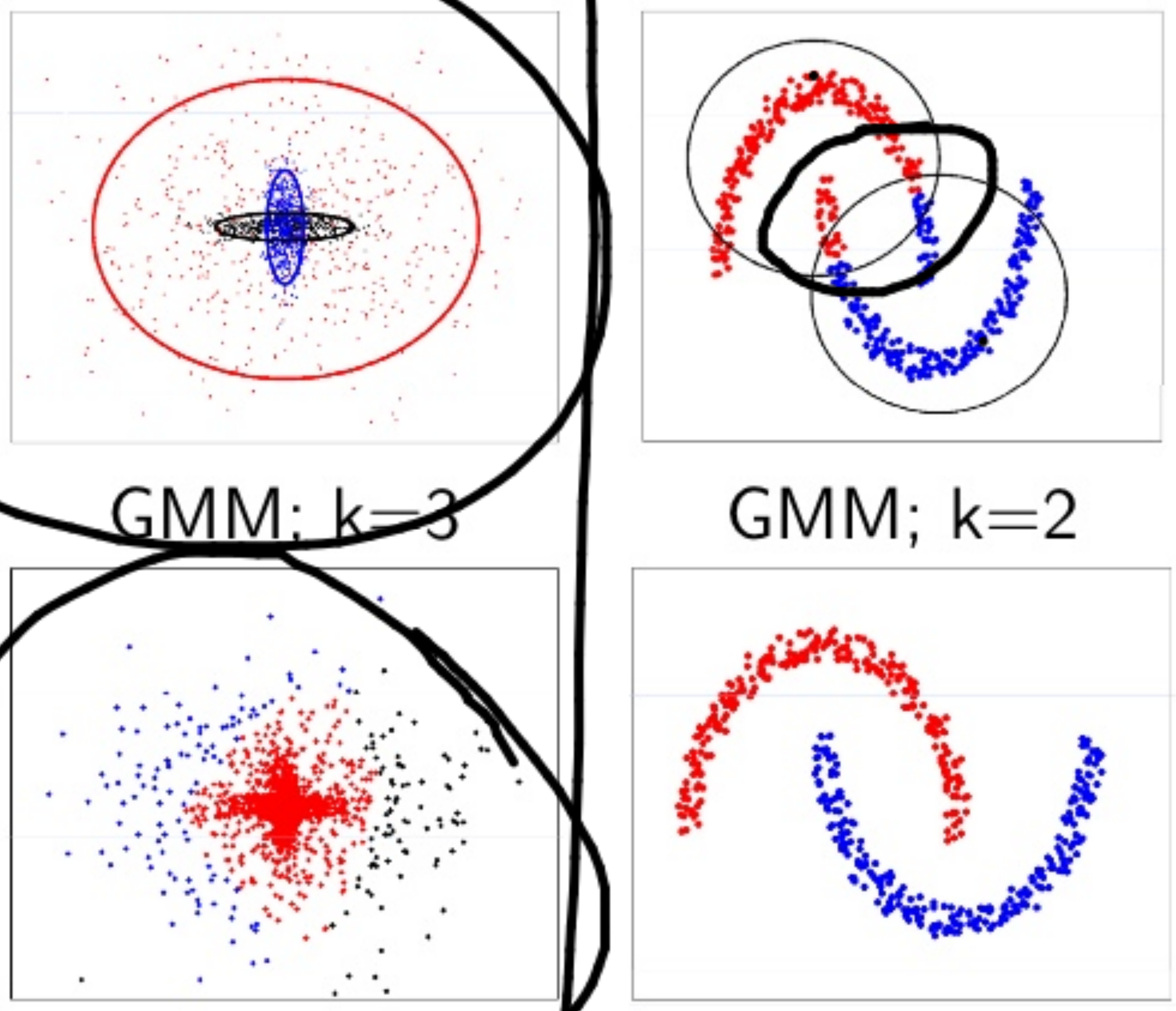
- Which clustering algorithm is the best?



There's no best Clustering Algorithm!

Source: R. Dubes and A. K. Jain, Clustering Techniques: User's Dilemma, PR 1976

- Each algorithm imposes a structure on data.
- Good fit between model and data $\Rightarrow$ success.



GMM; k=3 — GMM; k=2

Spectral; k=3 — Spectral; k=2

Gaussian Mixture Models

- Recall the Gaussian distribution:

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \tag{4}$$

Gaussian Mixture Models

- Recall the Gaussian distribution:

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \tag{4}$$

- It forms the basis for the important Mixture of Gaussians density.

Gaussian Mixture Models

- Recall the Gaussian distribution:

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \quad (4)$$

- It forms the basis for the important Mixture of Gaussians density.
- The Gaussian mixture is a **linear superposition of Gaussians** in the form:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad . \quad (5)$$

Gaussian Mixture Models

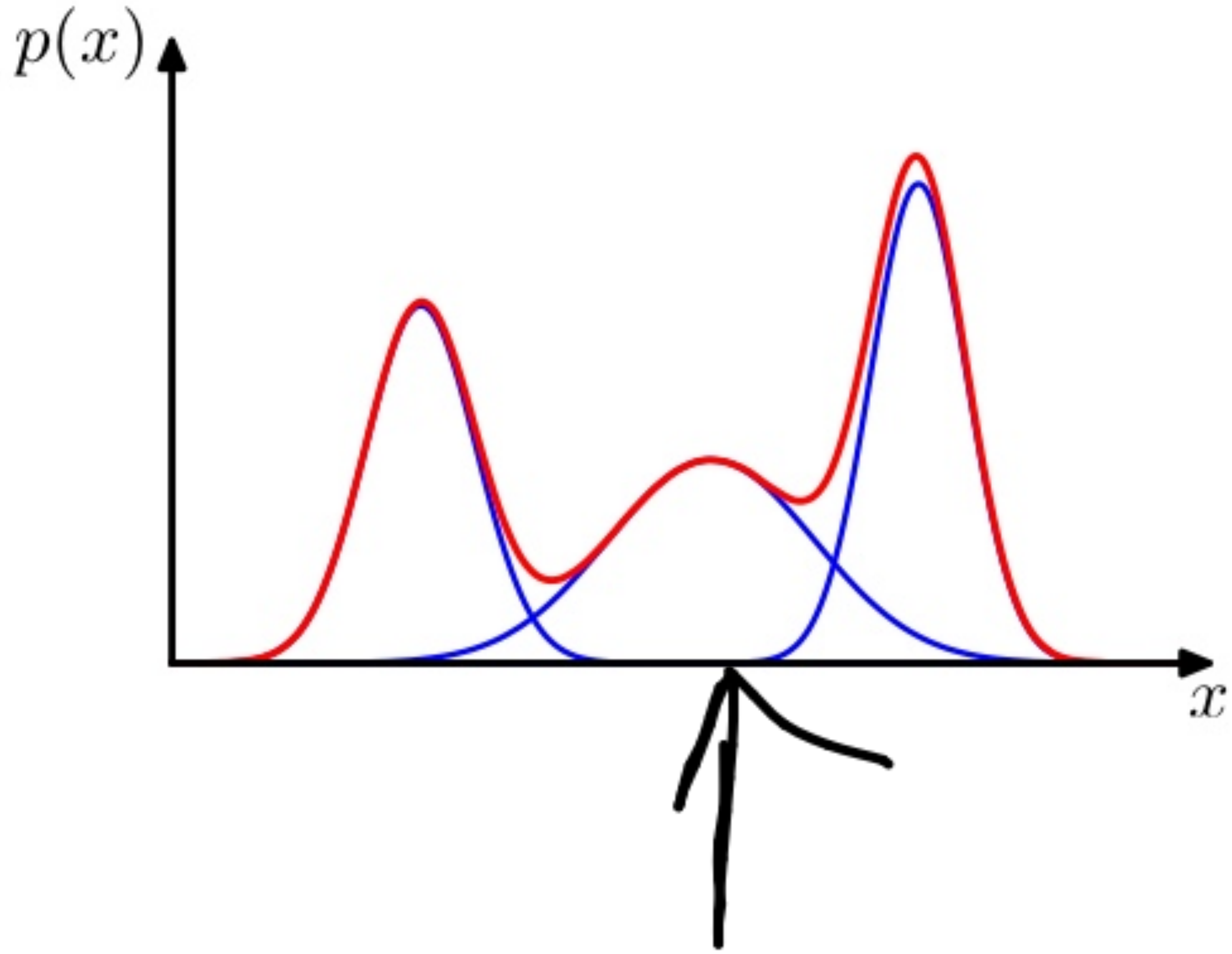
- Recall the Gaussian distribution:

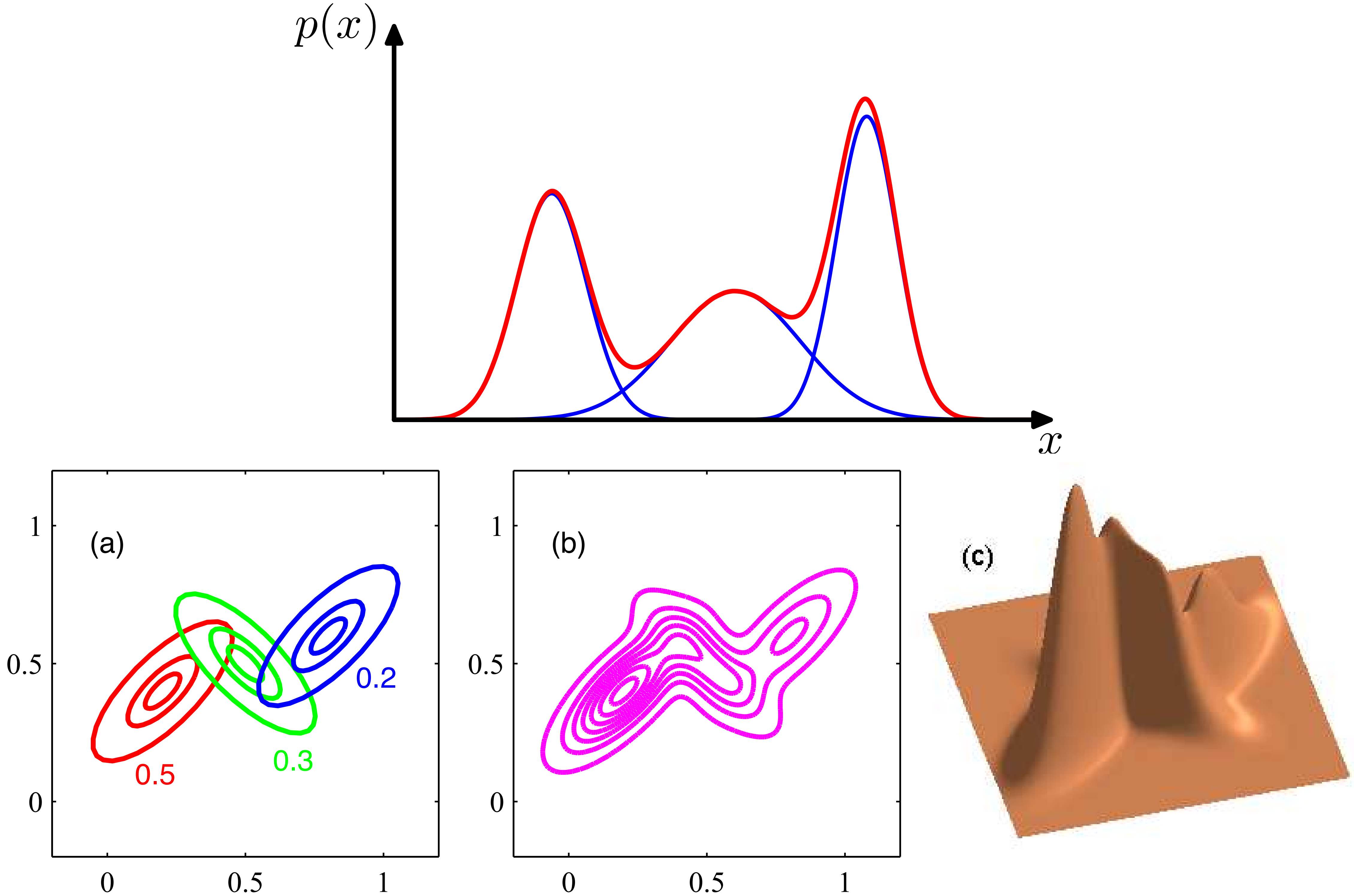
$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \quad (4)$$

- It forms the basis for the important Mixture of Gaussians density.
- The Gaussian mixture is a **linear superposition of Gaussians** in the form:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \ . \quad (5)$$

- The π_k are non-negative scalars called **mixing coefficients** and they govern the relative importance between the various Gaussians in the mixture density. $\sum_k \pi_k = 1$.


$p(x)$
x


$p(x)$
x
1
0.5
0
0
0.5
1
(a)
0.5
0.3
0.2
(b)
(c)