

Unsupervised Learning: k-Means and Mixtures of Gaussians

CS6780 – Advanced Machine Learning
Spring 2015

Thorsten Joachims
Cornell University

Reading: Murphy 11.1 – 11.4.2

Supervised Learning vs. Unsupervised Learning

- Supervised Learning
 - Classification: partition examples into groups according to pre-defined categories
 - Regression: assign value to feature vectors
 - Requires labeled data for training
- Unsupervised Learning
 - Clustering: partition examples into groups when no pre-defined categories/classes are available
 - Outlier detection: find unusual events (e.g. hackers)
 - Novelty detection: find changes in data
 - Only instances required, but no labels

Clustering

- Partition unlabeled examples into disjoint subsets of *clusters*, such that:
 - Examples within a cluster are similar
 - Examples in different clusters are different
- Discover new categories in an *unsupervised* manner (no sample category labels provided).

Applications of Clustering

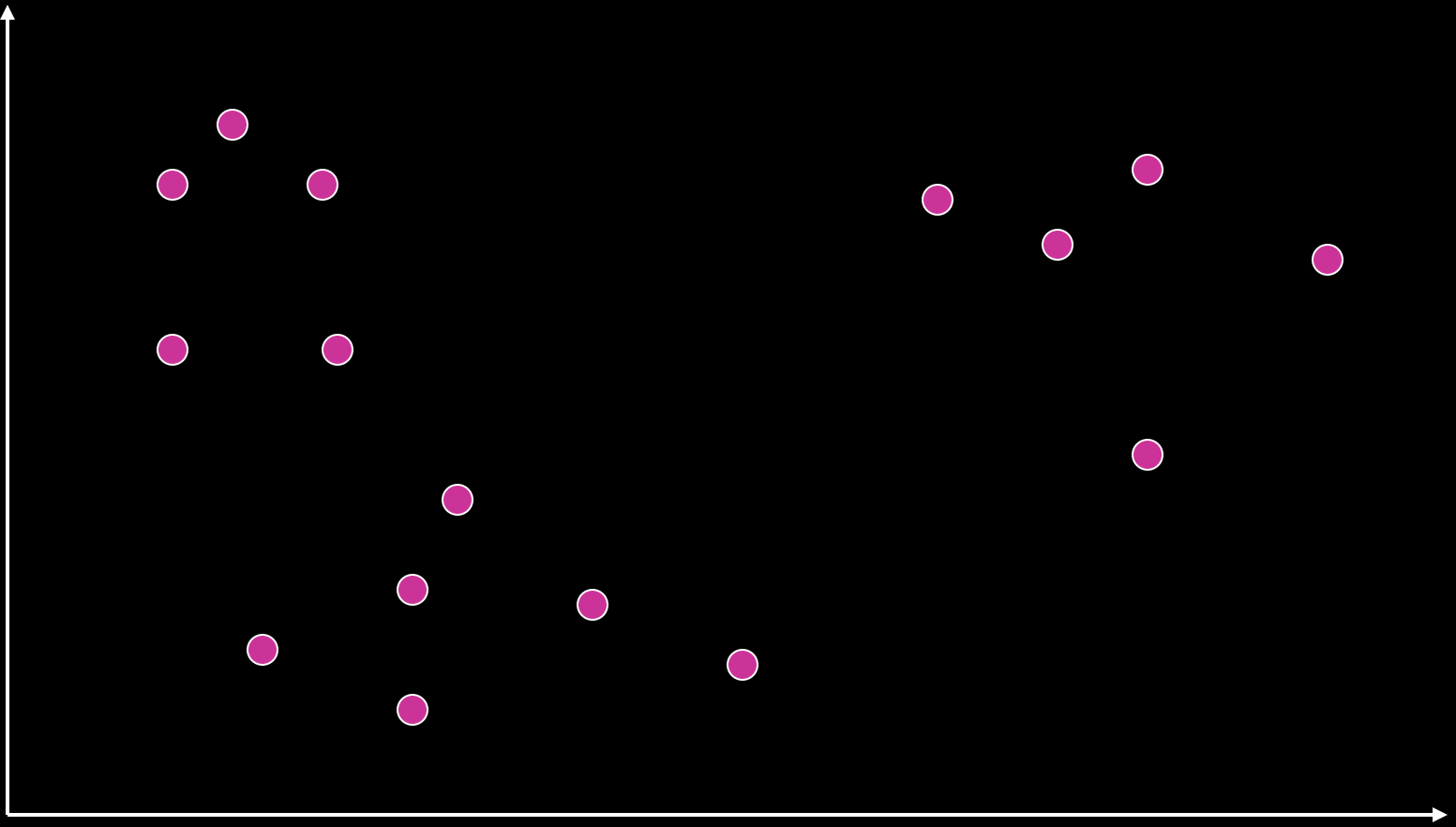
- Exploratory data analysis
- Cluster retrieved documents
 - to present more organized and understandable results to user → “diversified retrieval”
- Detecting near duplicates
 - Entity resolution
 - E.g. “Thorsten Joachims” == “Thorsten B Joachims”
 - Cheating detection
- Automated (or semi-automated) creation of taxonomies
 - e.g. Yahoo, DMOZ
- Compression

Applications of Clustering

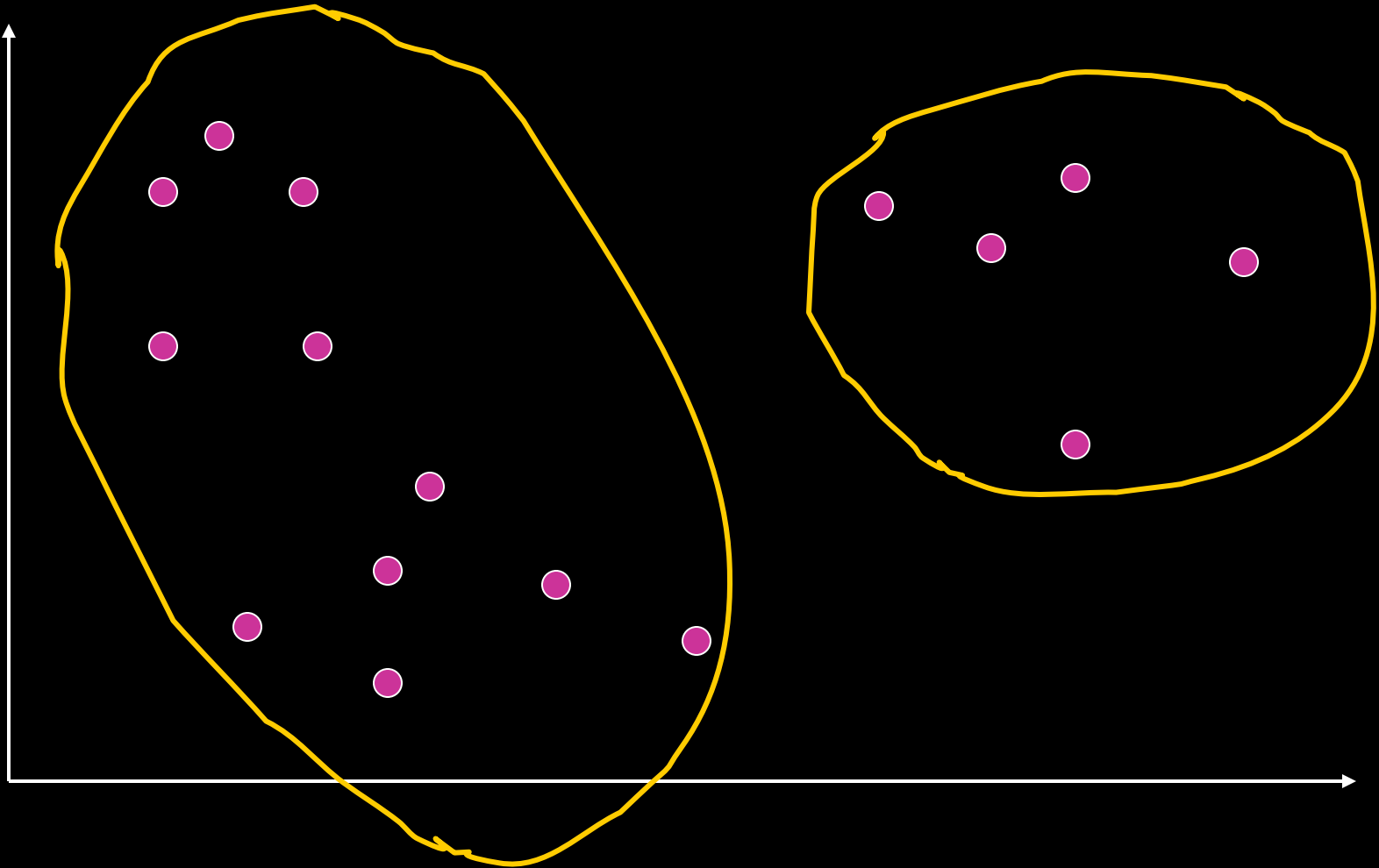
The screenshot shows a search engine interface with the following elements:

- Browser Tab:** svm - Carrot2 Clustering Engine
- Address Bar:** search.carrot2.org/stable/search?source=web&view=folders&skin=fancy-compact&query=svm&results=100&algorithm=lingo&EToolsDocu...
- Navigation Bar:** Includes icons for CMS, FacultyCenter, Brio, e-Shop, Finance, COLTS, DUS Wiki, UGrad, FacultyWiki, MLJ, Thermostat, GAPS, PhDApply, and Other. It also has links for About, New features!, Search feeds, Search plugins, Download, Carrot Search, and Contact.
- Search Bar:** Contains the text 'svm' and a 'Search' button with 'More options'.
- Left Sidebar (Folders):**
 - All Topics (95)
 - Support Vector Machines (17)
 - Support Vector Machine SVM (11)
 - Marketing (3)
 - Cards Gift (5)
 - Silvercorp Metals (5)
 - Magazine Ordinateur Individuel (4)
 - Manager (4)
 - SVM E-Business Solutions (4)
 - Secure Virtual Machine (4)
 - Training (4)
 - more | show all
- Main Content Area (Top 95 results of about 1350000 for svm):**
 - Support vector machine - Wikipedia, the free encyclopedia**
In machine learning, support vector machines (**SVMs**, also support vector networks) are supervised learning models with associated learning algorithms that ...
http://en.wikipedia.org/wiki/Support_vector_machine [Google, Wikipedia]
 - Gift Cards, Gas Gift Cards, Retail Gift Cards, Online Gift Cards | SVM**
SVM is the leader in sales, marketing and promotion of gas gift cards and many other forms of gift cards including service and retail gift cards.
<http://www.svmcards.com/> [Ask, Google, Teoma]
 - SVM: Summary for Silvercorp Metals Inc Ordinary - Yahoo! Finance**
View the basic **SVM** stock chart on Yahoo! Finance. Change the date range, chart type and compare Silvercorp Metals Inc Ordinary against other companies.
<http://finance.yahoo.com/q?s=SVM> [Google, Teoma]
 - SVM E-Marketing Solutions | Industrial Online Marketing Solutions**
SVM helps distributors and manufacturers leverage online marketing to increase sales, strengthen relationships with customers and measure marketing ROI.
<http://www.svmsolutions.com/> [Bing, Entireweb, Teoma, Yahoo, Google]
 - Silvercorp Metals Inc. (SVM) Stock - Seeking Alpha**
Up to date analysis of Silvercorp Metals Inc. (**SVM**) and its stock by hedge fund managers and industry experts. Find out what Silvercorp Metals Inc. is saying ...
<http://seekingalpha.com/symbol/svm> [Teoma, Google]
 - Society for Vascular Medicine : About SVM : Home**
Patient Information · Slides · Current Case Study · Case Study Archives · Sponsor the Case Study · Join **SVM** · Member Login · Find a Physician · Donate to **SVM** ...
<http://www.vascularmed.org/> [Teoma, Google]
 - SVM - Support Vector Machines**
SVM, support vector machines, **SVMC**, support vector machines classification, **SVMR**, support vector machines regression, kernel, machine learning, pattern ...
<http://www.support-vector-machines.org/> [Google, Teoma]
 - SVM-Struct Support Vector Machine for Complex Outputs**
SVMstruct is a Support Vector Machine (**SVM**) algorithm for predicting multivariate or structured outputs. It performs supervised learning by approximating a ...
http://svmlight.joachims.org/svm_struct.html [Google]
 - SVM Schweizerischer Verein für Mediation**
Verband zur Förderung aussergerichtlicher Konfliktbewältigung ...
<http://www.mediation-svm.ch/> [Ask, Bing, Entireweb, Yahoo]
 - SV Raumwelt Koch Mattershorn ...**
- Footer:** Query: svm -- Source: Web (95 results, 1758 ms) -- Clusterer: Lingo (188 ms) | v3.6.1-SNAPSHOT | build | 2012-06-22 22:55 © 2002-2012 Stanislaw Osinski, Dawid Weiss

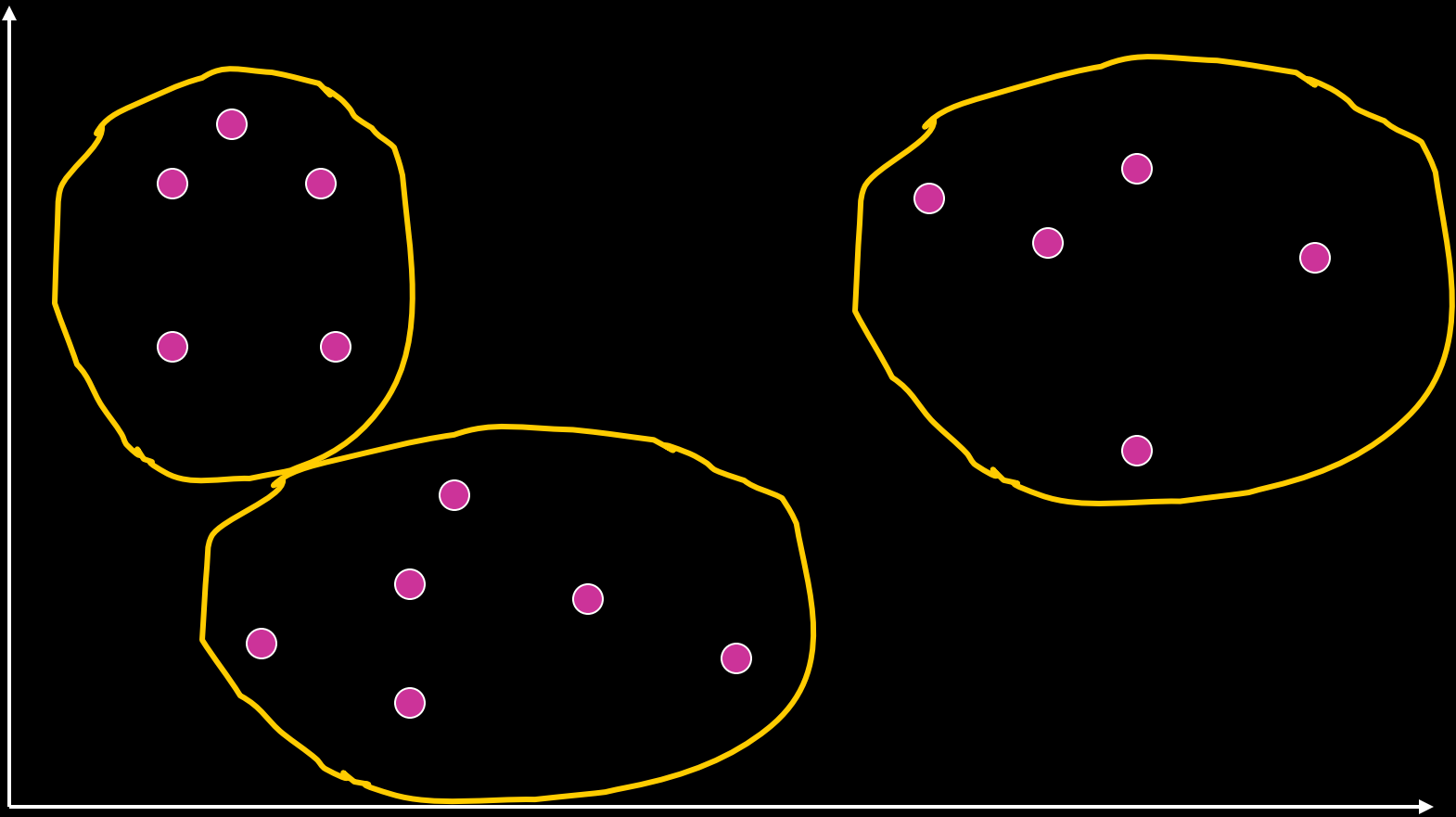
Clustering Example



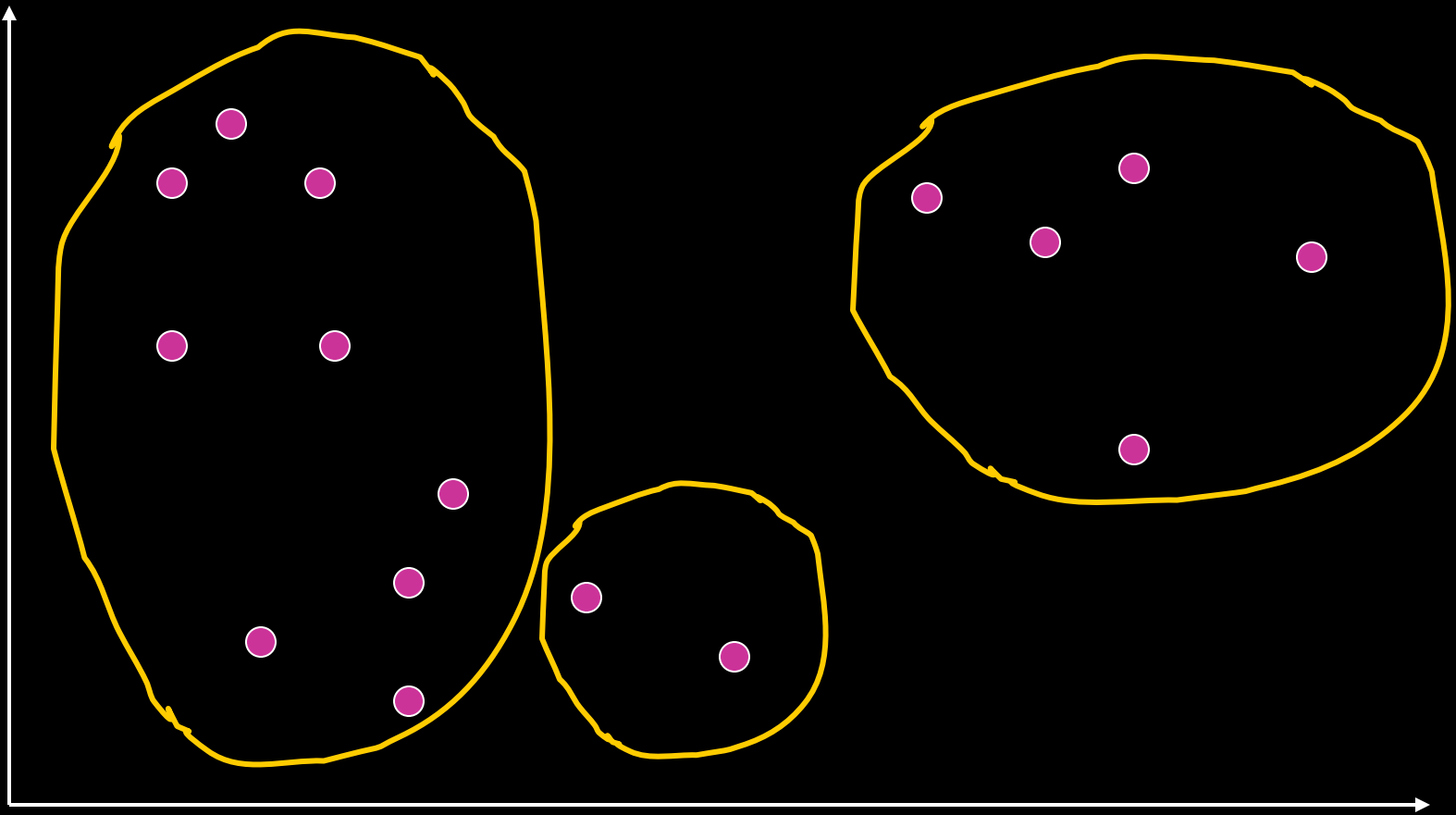
Clustering Example



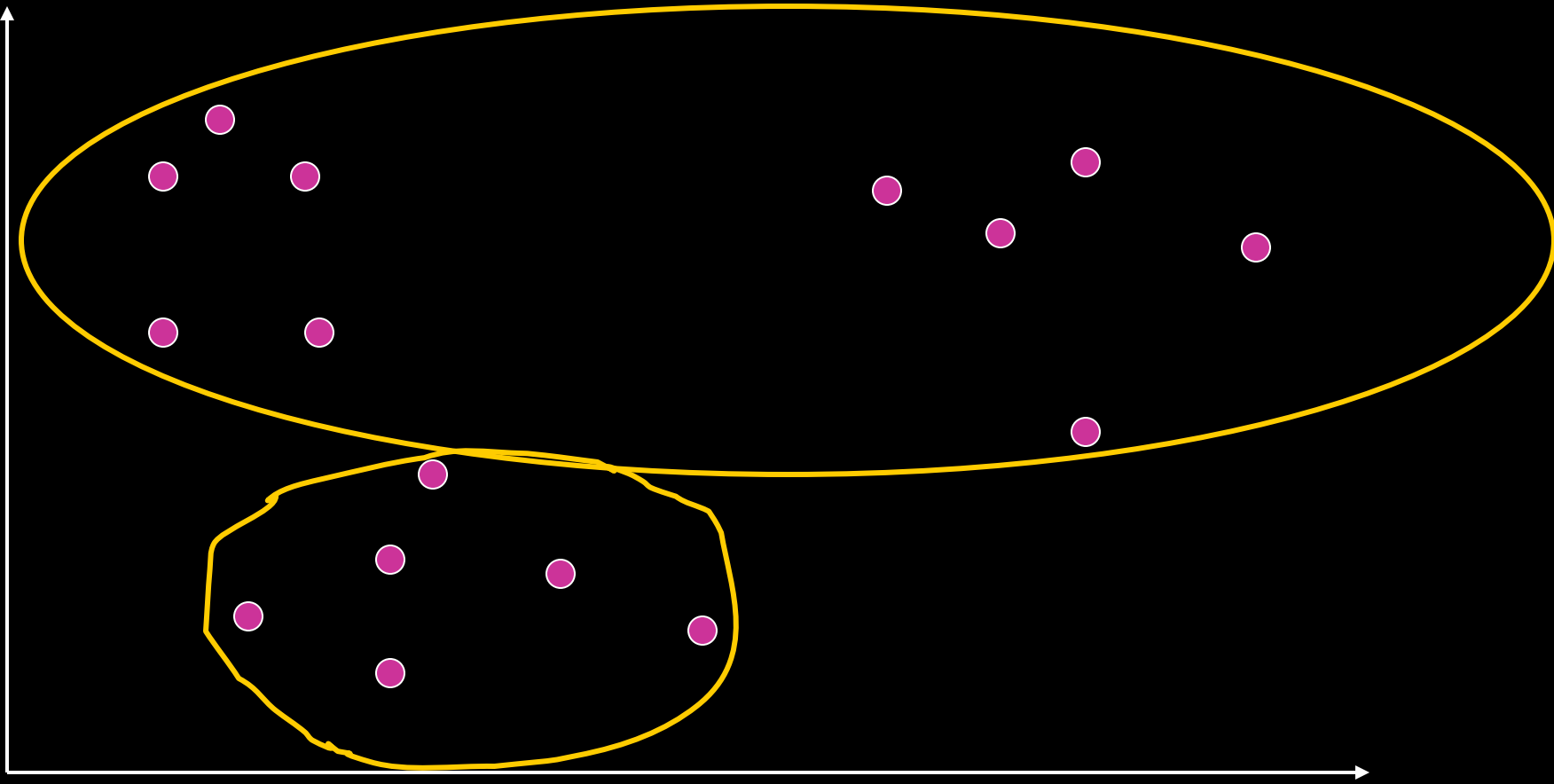
Clustering Example



Clustering Example



Clustering Example



Clustering Criterion

- Evaluation function that assigns a (usually real-valued) value to a clustering
 - Clustering criterion typically function of
 - within-cluster similarity and
 - between-cluster dissimilarity
- Optimization
 - Find clustering that maximizes the criterion
 - Global optimization (often intractable)
 - Greedy search
 - Approximation algorithms

Similarity (Distance) Measures

- Euclidian distance (L_2 norm):

$$L_2(\vec{x}, \vec{x}') = \sqrt{\sum_{i=1}^N (x_i - x'_i)^2}$$

- L_1 norm:

$$L_1(\vec{x}, \vec{x}') = \sqrt{\sum_{i=1}^N |x_i - x'_i|}$$

- Cosine similarity:

$$\cos(\vec{x}, \vec{x}') = \frac{\vec{x} * \vec{x}'}{\|\vec{x}\| \|\vec{x}'\|}$$

- Kernels

K-Means Algorithm

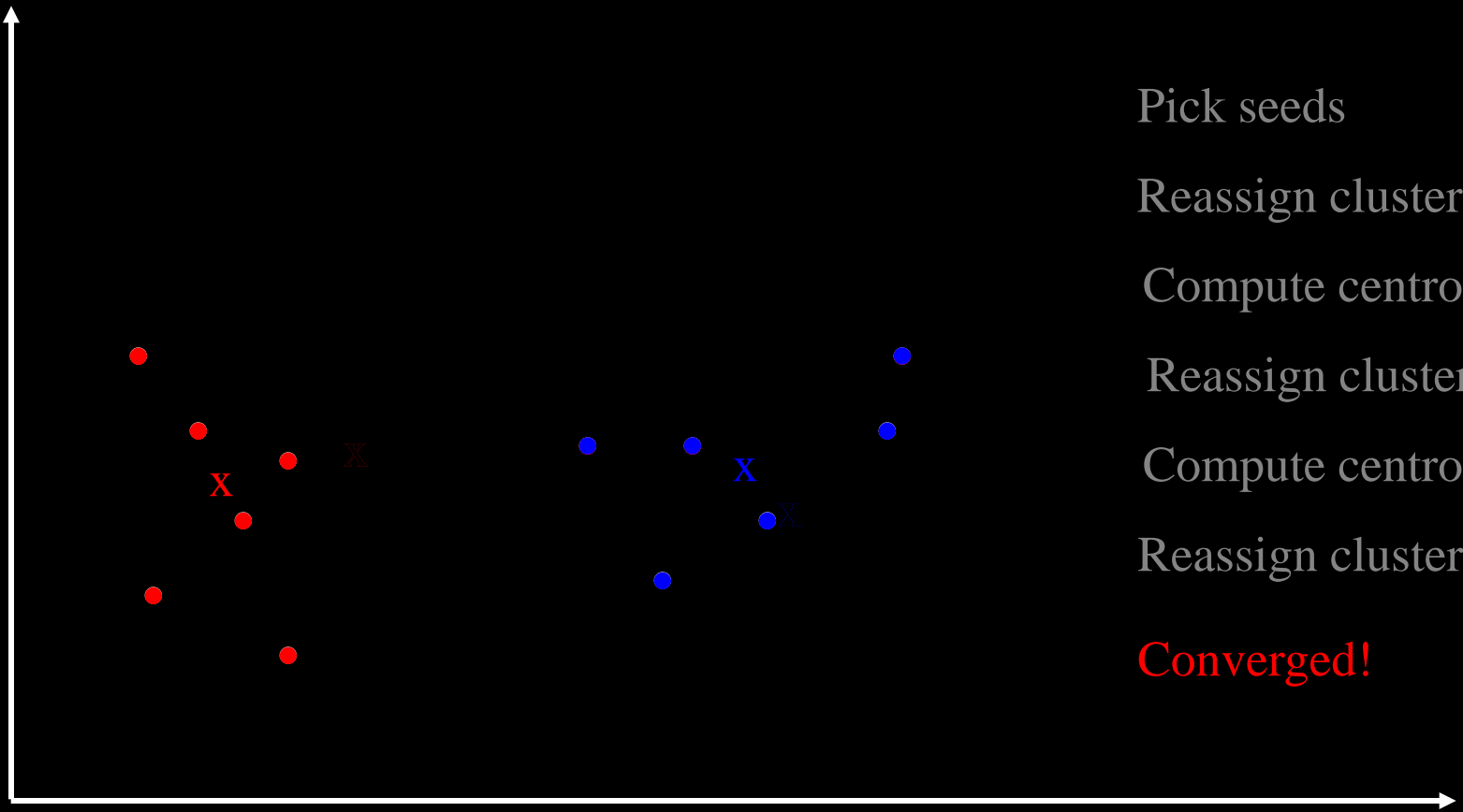
- Input: k = number of clusters, distance measure d
- Select k random instances $\{s_1, s_2, \dots, s_k\}$ as seeds.
- Until clustering converges or other stopping criterion:
 - For each instance x_i :
 - Assign x_i to the cluster c_j such that $d(x_i, s_j)$ is min.
 - For each cluster c_j //update the centroid of each cluster
 - $s_j = \mu(c_j)$

Note: Clusters represented via *centroids*

$$\vec{\mu}(c) = \frac{1}{|c|} \sum_{\vec{x} \in c} \vec{x}$$

K-means Example

(k=2)



Pick seeds

Reassign clusters

Compute centroids

Reassign clusters

Compute centroids

Reassign clusters

Converged!

Time Complexity

- Assume computing distance between two instances is $O(N)$ where N is the dimensionality of the vectors.
- Reassigning clusters for n points: $O(kn)$ distance computations, or $O(knN)$.
- Computing centroids: Each instance gets added once to some centroid: $O(nN)$.
- Assume these two steps are each done once for i iterations: $O(iknN)$.
- Linear in all relevant factors, assuming a fixed number of iterations.

Buckshot Algorithm

Problem

- Results can vary based on random seed selection, especially for high-dimensional data.
- Some seeds can result in poor convergence rate, or convergence to sub-optimal clusterings.

Idea: Combine HAC and K-means clustering.

- First randomly take a sample of instances of size $n^{1/2}$
- Run group-average HAC on this sample
- Use the results of HAC as initial seeds for K-means.
- Overall algorithm is efficient and avoids problems of bad seed selection.

Non-Hierarchical Clustering

- K-means clustering (“hard”)
- Mixtures of Gaussians and training via Expectation maximization Algorithm (“soft”)

Clustering as Prediction

- Setup
 - Learning Task: $P(X)$
 - Training Sample: $S = (\vec{x}_1, \dots, \vec{x}_n)$
 - Hypothesis Space: $H = \{h_1, \dots, h_{|H|}\}$ each describes $P(X|h_i)$ where h_i are parameters
 - Goal: learn which $P(X|h_i)$ produces the data
- What to predict?
 - Predict where new points are going to fall

Gaussian Mixtures and EM

- Gaussian Mixture Models

- Assume

$$P(X = \vec{x} | h_i) = \sum_{j=1}^k P(X = \vec{x} | Y = j, h_i) P(Y = j)$$

where $P(X = \vec{x} | Y = j, h) = N(X = \vec{x} | \vec{\mu}_j, \Sigma_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{1}{2\sigma_{ij}^2}(x-\mu_{ij})^2}$

and $h = (\vec{\mu}_1, \dots, \vec{\mu}_k, \Sigma_1, \dots, \Sigma_k)$.

- EM Algorithm

- Assume $P(Y)$ and k known and $\Sigma_i = 1$.

- REPEAT

- $$\vec{\mu}_j = \frac{\sum_{i=1}^n P(Y=j|X=\vec{x}_i, \vec{\mu}_1, \dots, \vec{\mu}_k) \vec{x}_i}{\sum_{i=1}^n P(Y=j|X=\vec{x}_i, \vec{\mu}_1, \dots, \vec{\mu}_k)}$$

- $$P(Y = j | X = \vec{x}_i, \vec{\mu}_1, \dots, \vec{\mu}_k) = \frac{P(X=\vec{x}_i|Y=j, \vec{\mu}_j)P(Y=j)}{\sum_{l=1}^k P(X=\vec{x}_i|Y=l, \vec{\mu}_l)P(Y=l)} = \frac{e^{-0.5(\vec{x}_i - \vec{\mu}_j)^2} P(Y=j)}{\sum_{l=1}^k e^{-0.5(\vec{x}_i - \vec{\mu}_l)^2} P(Y=l)}$$