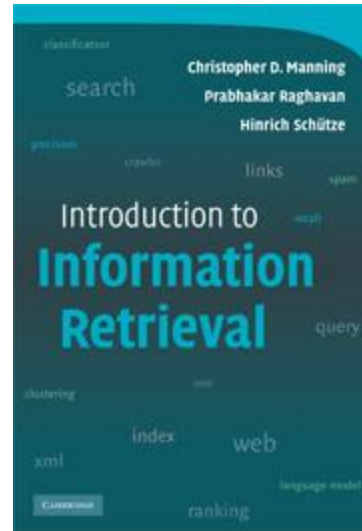# Information Retrieval and Organisation

Chapter 16

Flat Clustering

Dell Zhang

Birkbeck, University of London

What Is Text Clustering?

- Text Clustering =

  - Grouping a set of documents into classes of similar documents.

    - Documents within a cluster should be similar.
    - Documents from different clusters should be dissimilar.

- Classification vs. Clustering

  - Classification: *supervised* learning

    - Labeled data are given for training

  - Clustering: *unsupervised* learning

    - Only unlabeled data are available
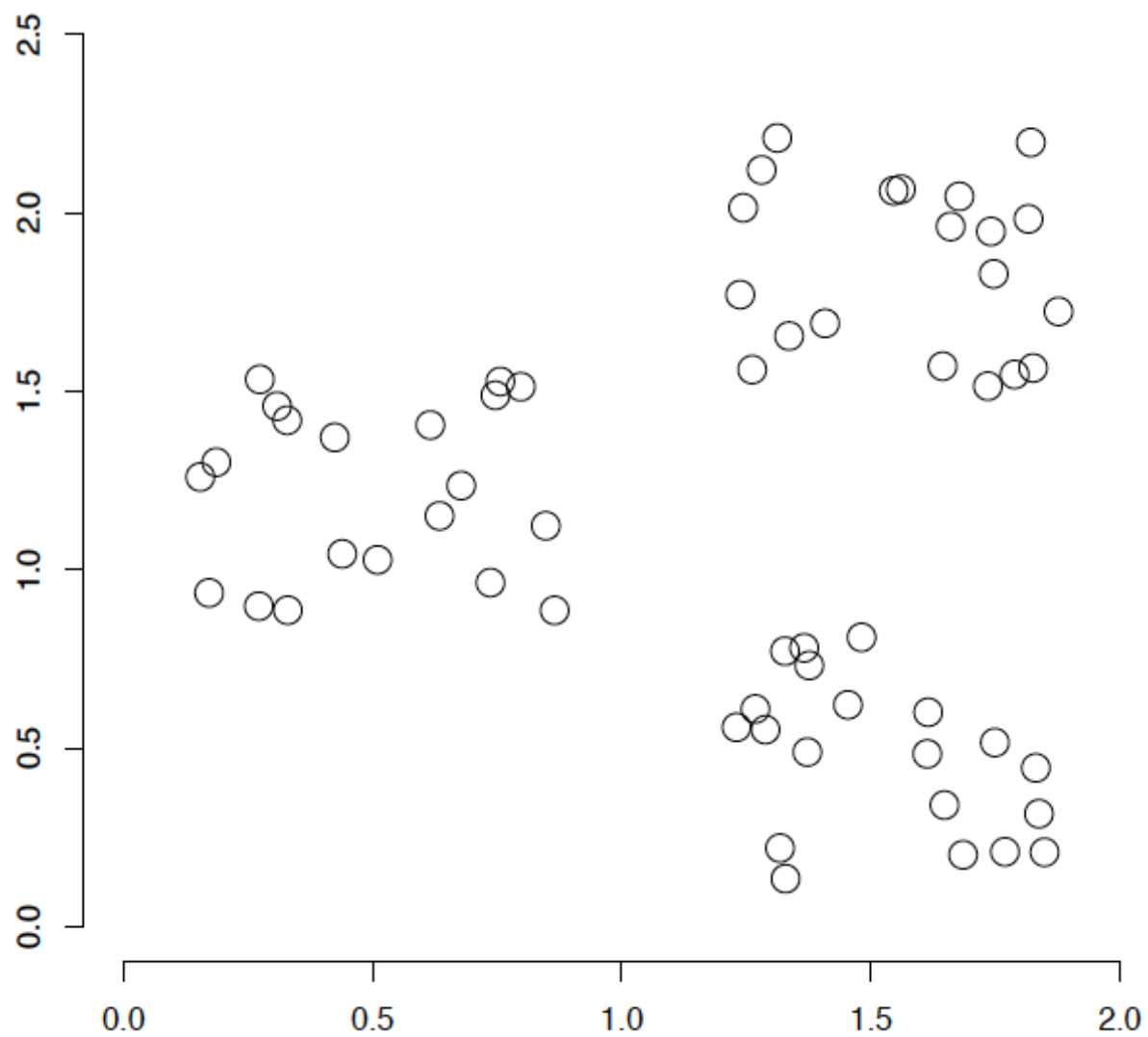
A data set with clear cluster structure
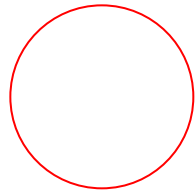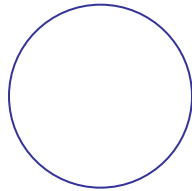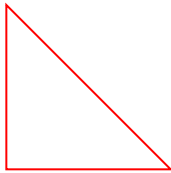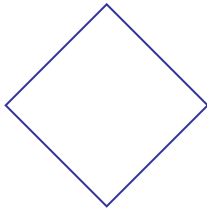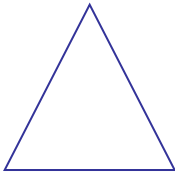
The Cluster Hypothesis

Documents in the same cluster behave similarly with respect to relevance to information needs.

Why Text Clustering?

- **To improve retrieval recall**

  - When a query matches a doc d, also return other docs in the cluster containing d. Hope if we do this, the query "car" will also return docs containing "automobile".

- **To improve retrieval speed**

  - Cluster Pruning: consider only documents in a small number of clusters as candidates for which we compute similarity scores.

Why Text Clustering?

- To improve user **interface**
    - Navigation/Visualization of <u>document collections</u>
    - Navigation/Visualization of <u>search results</u>

Searching + Browsing

File Edit Tools Visualizations Help

Selected 2 clusters
Selected 11 documents
Querying the database for gisting terms: please wait...

israeli israel
iraq iraqi
israeli israel
saudi fbi
liberia monrovia
iraq iraqi
bosnia bosnian
israel bosnia
ireland protestant
bosnia bosnian
freemen fbi
immigrants budget
iraq iraqi
fbi tobacco
budget megawati
ireland ira
twa fbi
korea korean
nichols mcveigh
church tobacco
cuba cuban
cadets citadel
olympic fbi
milosevic belgrade
kaczynski unabomber
colombia samper
russian chechen
india budget
russian china
russian yeltsin
valujet simpson
cuba cuban
simpson nicole
korea korean
refugees zaire
china rouge
india korea
hurricane bertha
china pope
kevorkian fleiger
teresa india
china kong
simpson nicole
auction onassis
iraqi iraq

CARTIA

Introduction
Insights
Eagle Eye View
Using the Map
Analysis Tools
Searching the Map
About ThemeScape

OATH
SEX
BLUMENTHAL   ASSERTED
ABSOLUTELY
EMAIL   BLACK DOG
PACKAGES

http://news.google.co.uk/

Google™ **News** U.K.

Search News   Search the Web

Search and browse 4,500 news sources updated continuously.

Business »

HSBC Creates $5 Billion Fund to Boost Credit Access (Update1)
Bloomberg - **5 hours ago**
By Rebecca Keenan Dec. 7 (Bloomberg) -- HSBC Holdings Plc, Europe's largest bank, created a $5 billion fund to increase access to credit for small and medium-sized businesses.
HSBC starts $5 billion lending fund for small, medium businesses MarketWatch
HSBC makes £1bn credit available BBC News
The Press Association - Shropshire Star - Sky News - Times Online
all 145 news articles » ✉LON:HSBA - HCS

ITV.com

More "world coordination" on rates needed
Reuters UK - **6 Dec 2008**
LONDON (Reuters) - More international coordination of macro-economic policy is needed, Bank of England Deputy Governor John Gieve said on Saturday.
Brown poised for banks showdown Financial Times
Mortgage rate rip-off: banks stand accused This is Money
Times Online - Telegraph.co.uk - Independent - Scotsman
all 518 news articles »

Telegraph.co.uk

Sun's shining on tracker customers but be prepared for the rainy days
guardian.co.uk - **17 hours ago**
Hundreds of thousands of homeowners with a tracker mortgage stand to save a fortune in the long run, if they resist the urge to splurge, and instead overpay their loan every month.
The £1bn rip-off on trackers Times Online
Some lenders need to hear the rates call This is Money
Independent - Telegraph.co.uk - Scotsman - Daily Mail
all 671 news articles »

BBC News

Show more stories   Show fewer stories

Sci/Tech »

Parents warned on fake consoles
BBC News - **5 Dec 2008**
Parents are being warned about fake imported Nintendo games consoles which could be a fire hazard and pose a danger to their children's safety.
Fake Cheap Nintendo DS Consoles Could Kill ITProPortal
Warning issued over dangerous consoles Inquirer
Idealo Product News - Reg Hardware - Computeractive - Reuters UK
all 74 news articles »

Reuters

One giant leap for 'teddy-nauts'
BBC News - **5 Dec 2008**
When the UK government said it was thinking of ending its long-standing ban on astronauts, surely it wasn't thinking of putting teddies in space?
These teddy bears have the right stuffing Telegraph.co.uk
Teddy bears launched into space Digital Spy
Times Online - Daily Mail - Channel 4 News - Metro
all 109 news articles »

The Age

Minister checks on science exams
BBC News - **4 Dec 2008**
By Gary Eason A government minister has said he is calling in copies of GCSE and A-level science exam papers to satisfy himself they are not being "dumbed down".
Britain's big challenges will be met by doctors of innovation Times Online
Sheffield to train engineering leaders for Britain's future Sheffield Telegraph
Imperial College London - UK Trade & Investment - University of Southampton - Reuters UK
all 147 news articles »

BBC News

Show more stories   Show fewer stories

Sport »

• Rock-bottom Baggies held to 1-1 draw by Portsmouth
guardian.co.uk - **21 minutes ago**
Peter Crouch's shot from outside the box levelled the match at the Hawthorns. Photograph: Andrew Yates/AFP/Getty Images West Brom remain rooted to the bottom of the Premier League after failing to take advantage of a first-half lead to draw 1-1 at home ...
West Brom 1-1 Portsmouth BBC Sport
West Brom Remain Bottom After Portsmouth Draw Goal.com
SkySports - Portsmouth News - The Sun - The Press Association
all 132 news articles »

Goal.com

Entertainment »

Ricky Gervais: Jonathan Ross is devastated by phone scandal
Now Magazine Online - **1 hour ago**
Ricky Gervais has revealed that he doesn't think Jonathan Ross deserved to be suspended from his BBC job for 12 weeks. The chat show host was taken off air after he and pal Russell Brand, 33, left prank messages for Andrew Sachs, 78, during a Radio 2 ...
Jonathan Ross and the British Comedy Awards: wot no Woss? Times Online
Just five weeks after quitting the BBC in disgrace, Russell Brand ... Daily Mail
guardian.co.uk - The Press Association - Telegraph.co.uk - BBC News
all 219 news articles »

Sky News

http://clusty.com/

MP3

Search

advanced
preferences

clusters   sources   sites

Cluster **Player** contains **53** documents.

All Results (248)

➕ **Download** (73)

🔴 **Player** (53)

  ➕ **Downloads, Free MP3 Software** (6)

  ➕ **Zune, Microsoft** (5)

  ⚫ **Media Player** (4)

  ⚫ **Reviews, Portable MP3 Players** (4)

  ⚫ **Flash MP3 Player** (4)

  ⚫ **MP3 Player Accessories** (3)

  ⚫ **Windows That Tags, Rips, And Burns** (2)

  ⚫ **Entertainment** (2)

  ⚫ **Zen, Muvo, Digital MP3 Players** (2)

  ⚫ **Music Jukebox** (2)

  more

➕ **Photos** (35)

➕ **MP3 Converter** (23)

Search Results

1. WINAMP.COM | Winamp.com: Media **Player**, Skins, Plug-ins, Videos, Songs ...

The popular **MP3 player** can play MODs as well.
www.winamp.com - [cache] - MSN, Open Directory, Wisenut, Ask

2. Portable **MP3 Players** - **MP3 Player** Reviews - Best **MP3 Players**

Portable **MP3 Players** - **MP3**.com offers **MP3 player** reviews of the best **MP3 players** available. ... Archos 604-WiFi The do-everything Archos 604-WiFi is completely stacked with awesome video ...
www.mp3.com/hardware.php - [cache] - MSN

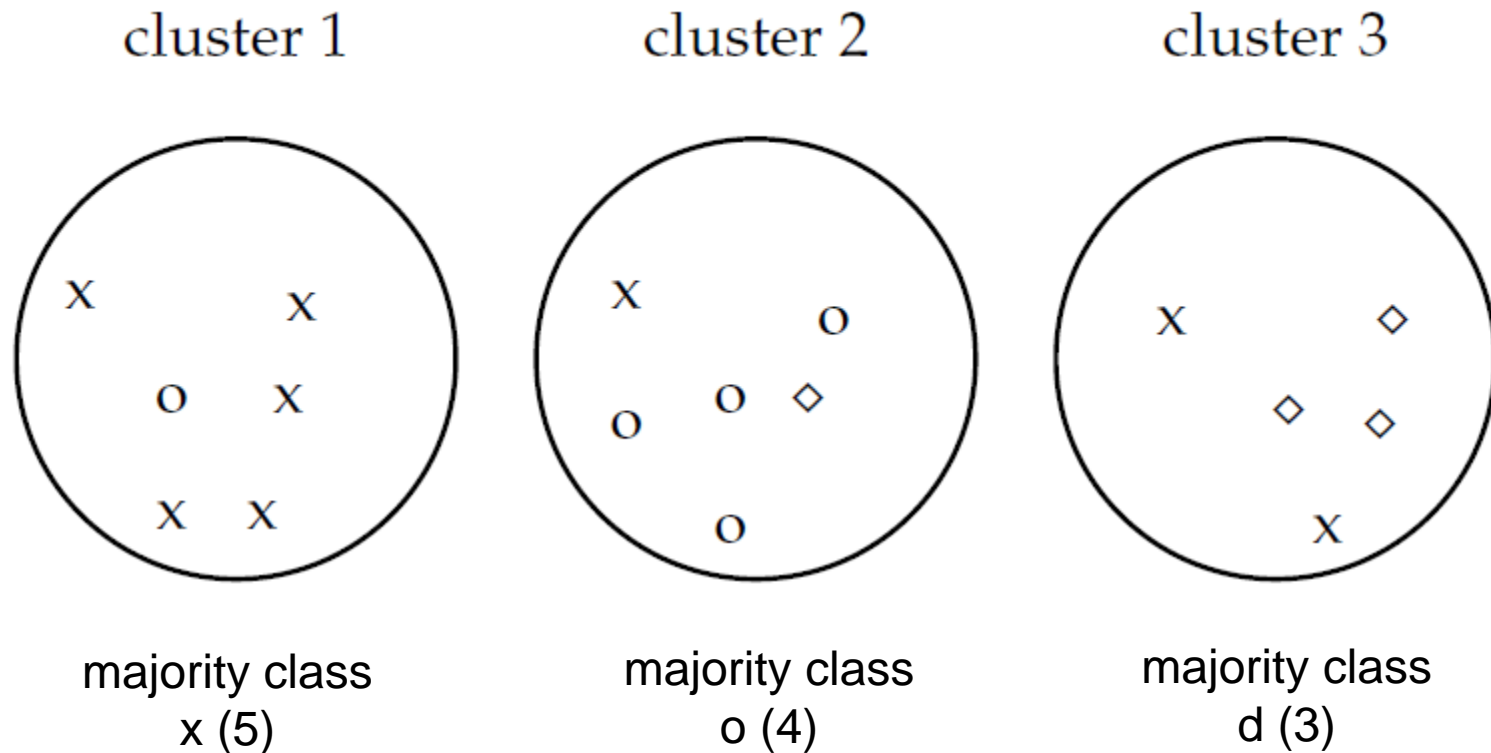3. Zune **MP3 player** has a lot of catching up to do

Nov 12, 2006 - Microsoft's sturdy Zune foot soldiers will march out in an improbable mission to topple Apple's world champion iPod **MP3 player**. Zunes echoing iPod's design but steeped in Microsoft functionality instead of Apple panache will debut in stores on ... technology; built-in FM tuners and three-inch (7.6-centimeter) screens. Microsoft was trying to set Zune apart from other **MP3 players** by promoting the ability of the devices to connect wirelessly with each other so users could share music or ... sales revenues. Apple has focused on consumers in a strategy that made iPods a "cultural fetish" and the most popular **MP3 player** on the planet, according to analysts. "It is an effort by Microsoft to turn the labels and artists into ...
news.yahoo.com/s/afp/20061112/tc_afp/afplifestyleitinternet - [cache] - Yahoo! News

What Clustering Is Good?

- **External Criteria**
  - Consistency with the latent classes in *gold standard* (*ground truth*) data.
    - Purity
    - Normalized Mutual Information
    - Rand Index
    - Cluster F Measure
- **Internal Criteria**
  - High *intra-cluster* similarity
  - Low *inter-cluster* similarity

What Clustering Is Good?

■ Purity

cluster 1   cluster 2   cluster 3

majority class x (5)   majority class o (4)   majority class d (3)

$$\text{purity} = (5+4+3)/17 = 0.71$$

What Clustering Is Good?

- ## Rand Index (RI)
  - The percentage of decisions (on document-pairs) that are correct

{**A**, **A**, **B**}
and
{**B**, **B**}

	A_1	A_2	B_1	B_2	B_3
A_1		*tp*	*fp*	*tn*	*tn*
A_2			*fp*	*tn*	*tn*
B_1				*fn*	*fn*
B_2					*tp*
B_3					

$$\text{RI} = (TP+TN)/(TP+FP+FN+TN) = 6/10 = 0.6$$

Issues in Clustering

- ## Similarity/Distance between docs
  - ## Ideally: semantic
  - ## Practically: statistical
    - ### e.g., cosine similarity or Euclidean distance
      - #### For text clustering, the doc vectors usually need to be length normalized.

- ## Membership of docs
  - ## Hard: each doc belongs to exactly one cluster
  - ## Soft: A doc can belong to more than one cluster
    - ### e.g., you may want to put a pair of sneakers in two clusters: (i) sports apparel and (ii) shoes.

Issues in Clustering

- Number of clusters
  - Fixed in advance
  - Discovered from data
- Structure of clusters
  - Flat (partition)
    - e.g., kMeans.
  - Hierarchical (tree)
    - e.g., HAC.

K-Means Algorithm

K-MEANS($\{\vec{x}_1, \ldots, \vec{x}_N\}, K$)

  1    $(\vec{s}_1, \vec{s}_2, \ldots, \vec{s}_K) \leftarrow$ SELECTRANDOMSEEDS($\{\vec{x}_1, \ldots, \vec{x}_N\}, K$)

  2   **for** $k \leftarrow 1$ **to** K

  3   **do** $\vec{\mu}_k \leftarrow \vec{s}_k$

  4   **while** stopping criterion has not been met

  5   **do for** $k \leftarrow 1$ **to** K

  6   **do** $\omega_k \leftarrow \{\}$

  7   **for** $n \leftarrow 1$ **to** N

  8   **do** $j \leftarrow \arg\min_{j'} |\vec{\mu}_{j'} - \vec{x}_n|$

  9           $\omega_j \leftarrow \omega_j \cup \{\vec{x}_n\}$ *(reassignment of vectors)*

10   **for** $k \leftarrow 1$ **to** K

11   **do** $\vec{\mu}_k \leftarrow \frac{1}{|\omega_k|} \sum_{\vec{x} \in \omega_k} \vec{x}$ *(recomputation of centroids)*

12   **return** $\{\vec{\mu}_1, \ldots, \vec{\mu}_K\}$

Time Complexity

- Computing distance between two docs is $O(m)$ where m is the dimensionality of the vectors.

- Reassigning clusters: $O(kn)$ distance computations, or $O(knm)$.

- Computing centroids: Each doc gets added once to some centroid: $O(nm)$.

- Assume these two steps are each done once for i iterations: $O(iknm)$.

Stopping Criterion

- Fixed number of iterations

- Convergence: to reach a state in which clusters don't change

  - k-means is proved to converge

  - k-means usually converges quickly, i.e., the number of iterations needed for convergence is typically small.

K-Means – Example

(*k* = 2)
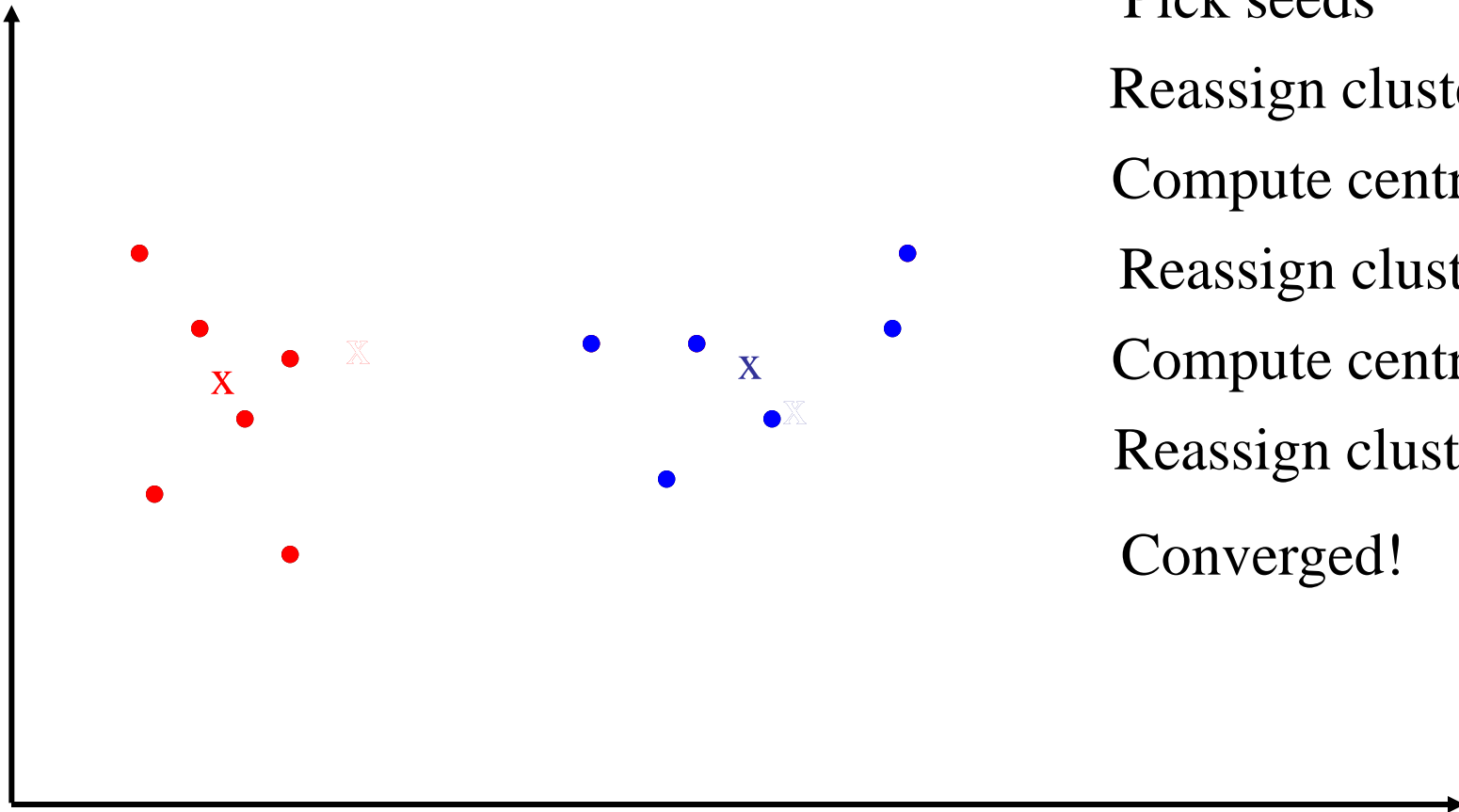
Pick seeds

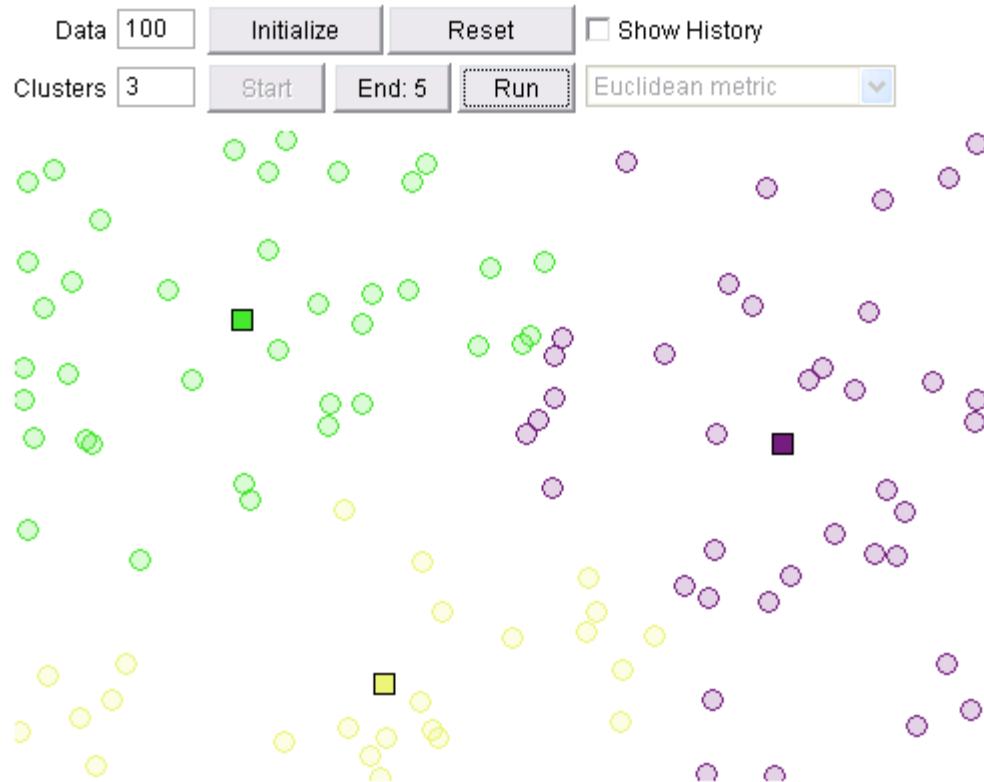Reassign clusters

Compute centroids

Reassign clusters

Compute centroids
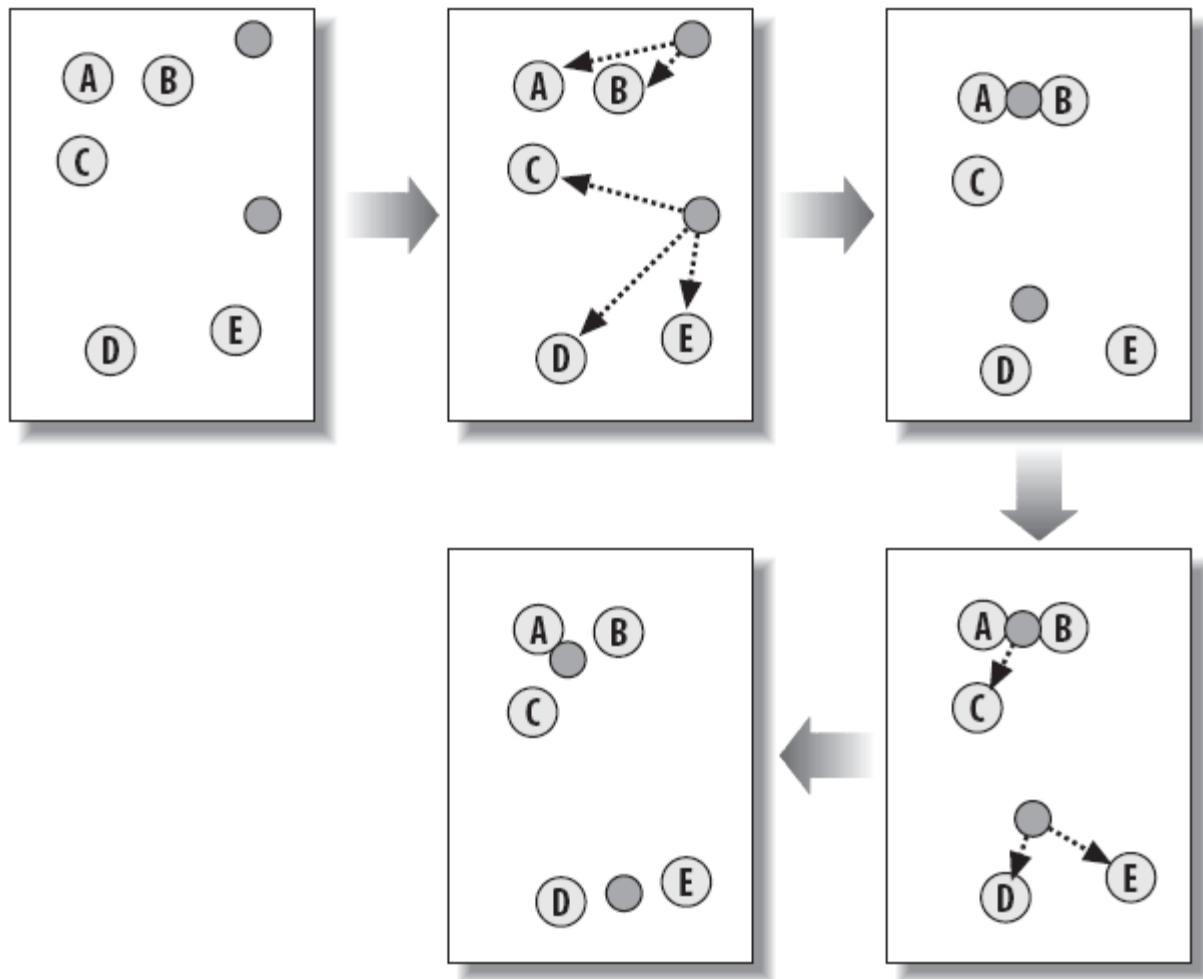
Reassign clusters

Converged!

K-Means – Demo

- http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletKM.html
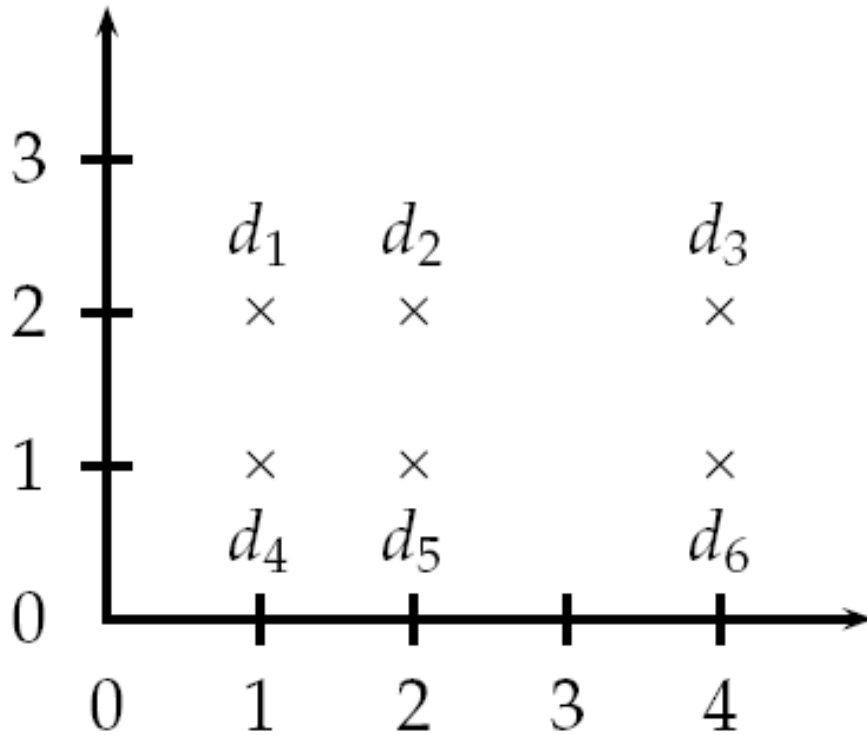
kMeans – Exercise

Digital Camera	Megapixel	Zoom
A	1	8
B	3	8
C	2	6
D	1.5	1
E	4	2

kMeans – Exercise

Seed Choice

- *k*-means (with *k*=2)
  - For seeds d_2 and d_5, the algorithm converges to $\{\{d_1, d_2, d_3\}, \{d_4, d_5, d_6\}\}$, a suboptimal clustering.
  - For seeds d_2 and d_3, the algorithm converges to $\{\{d_1, d_2, d_4, d_5\}, \{d_3, d_6\}\}$, the global optimum.

Seed Choice

- Problem
  - The outcome of clustering in k-means depends on the initial seeds.
    - Some seeds can result in poor convergence rate, or convergence to sub-optimal clustering.
- Solution
  - Excluding outliers from the seed set.
  - Trying out multiple sets of random starting points and choosing the clustering with the lowest cost.
  - Obtaining good seeds from another method
    - e.g., hierarchical clustering, k-means++