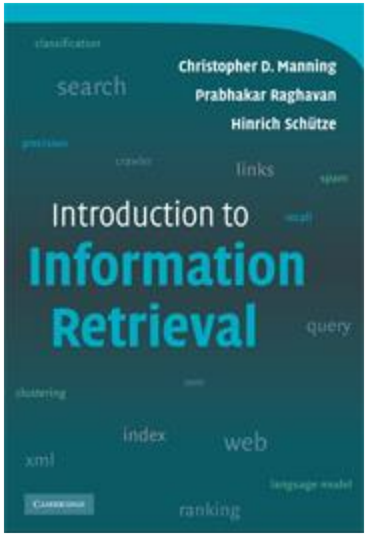
Information Retrieval and Organisation



Chapter 16

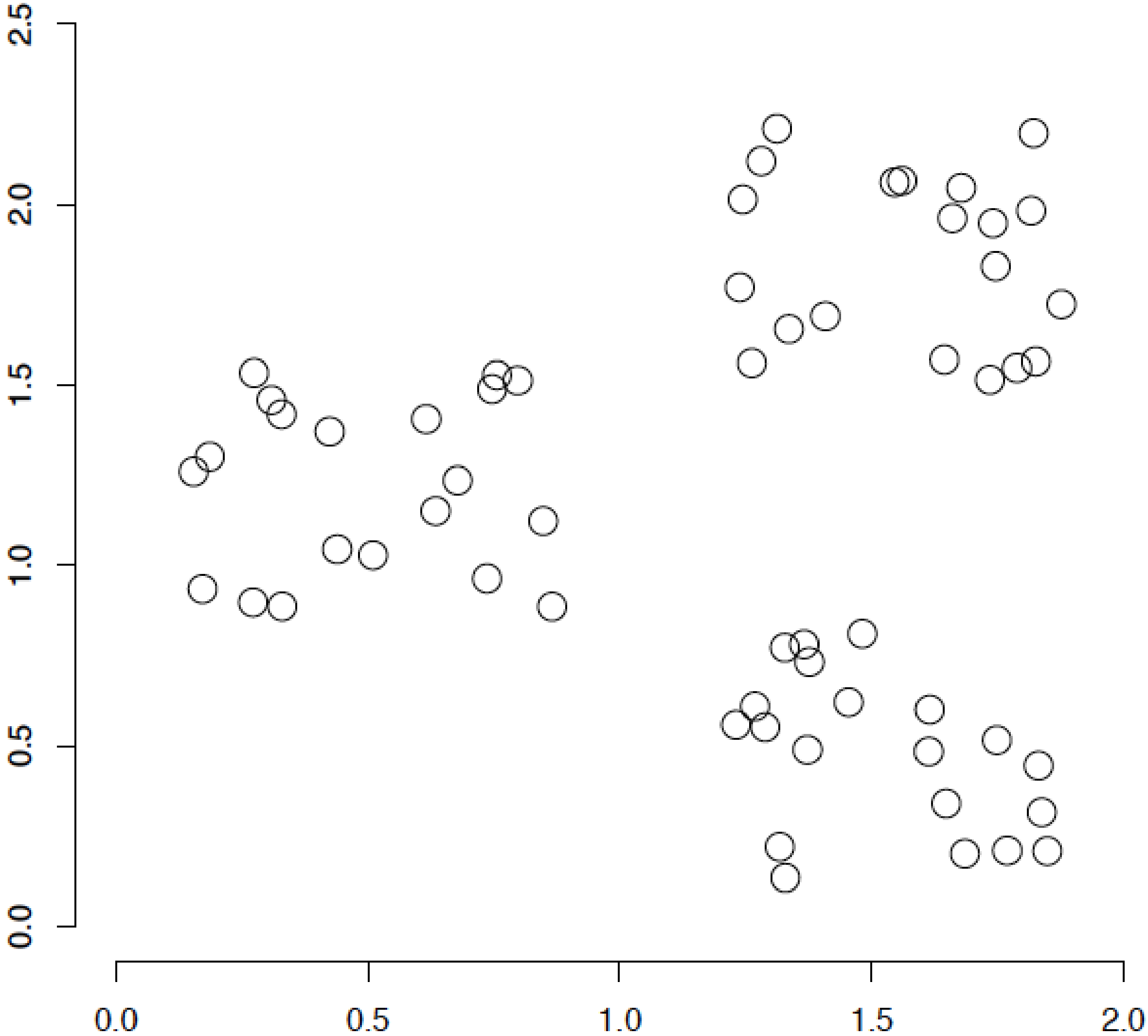
Flat Clustering

Dell Zhang

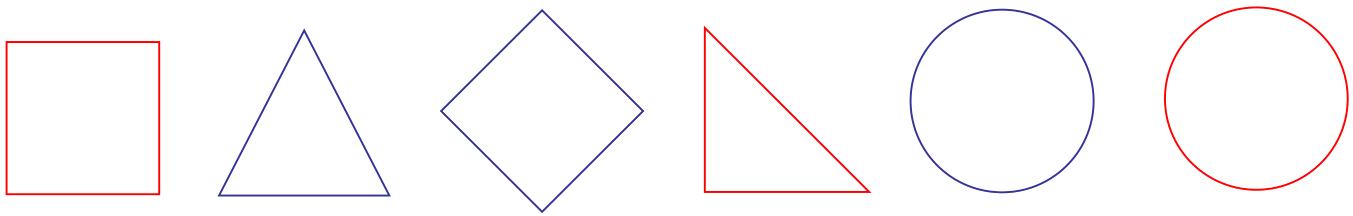
Birkbeck, University of London

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".

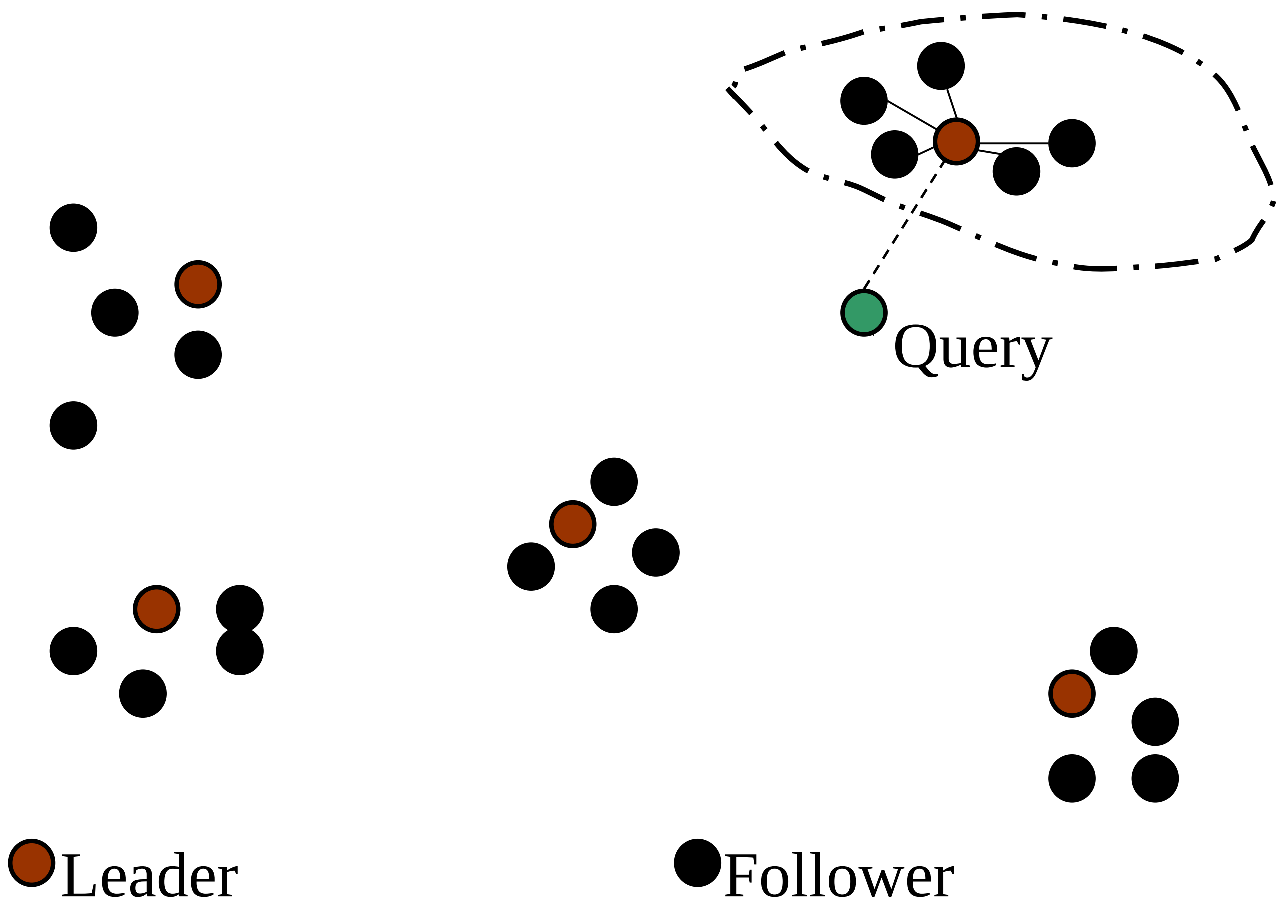
Why Text Clustering?

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

Cluster Pruning

- **Chapter 7 Section 7.1.6**
- Preprocessing
  - Pick $\sqrt{N}$ *docs* at random: call these *leaders*
  - For every other doc, pre-compute nearest leader
    - Docs attached to a leader: its *followers;*
    - <u>Likely</u>: each leader has ~ $\sqrt{N}$ followers.
- Query Processing
  - Given query *Q*, find its nearest *leader L.*
  - Seek *K* nearest docs from among *L*'s followers.

Cluster Pruning



Cluster Pruning

- Why use random sampling?
  - Fast
  - Leaders reflect data distribution
- More sophisticated clustering techniques later

Cluster Pruning

- General Variants
  - Have each follower attached to *b1*=3 (say) nearest leaders.
  - From query, find *b2*=4 (say) nearest leaders and their followers.
  - Can recurse on leader/follower construction.

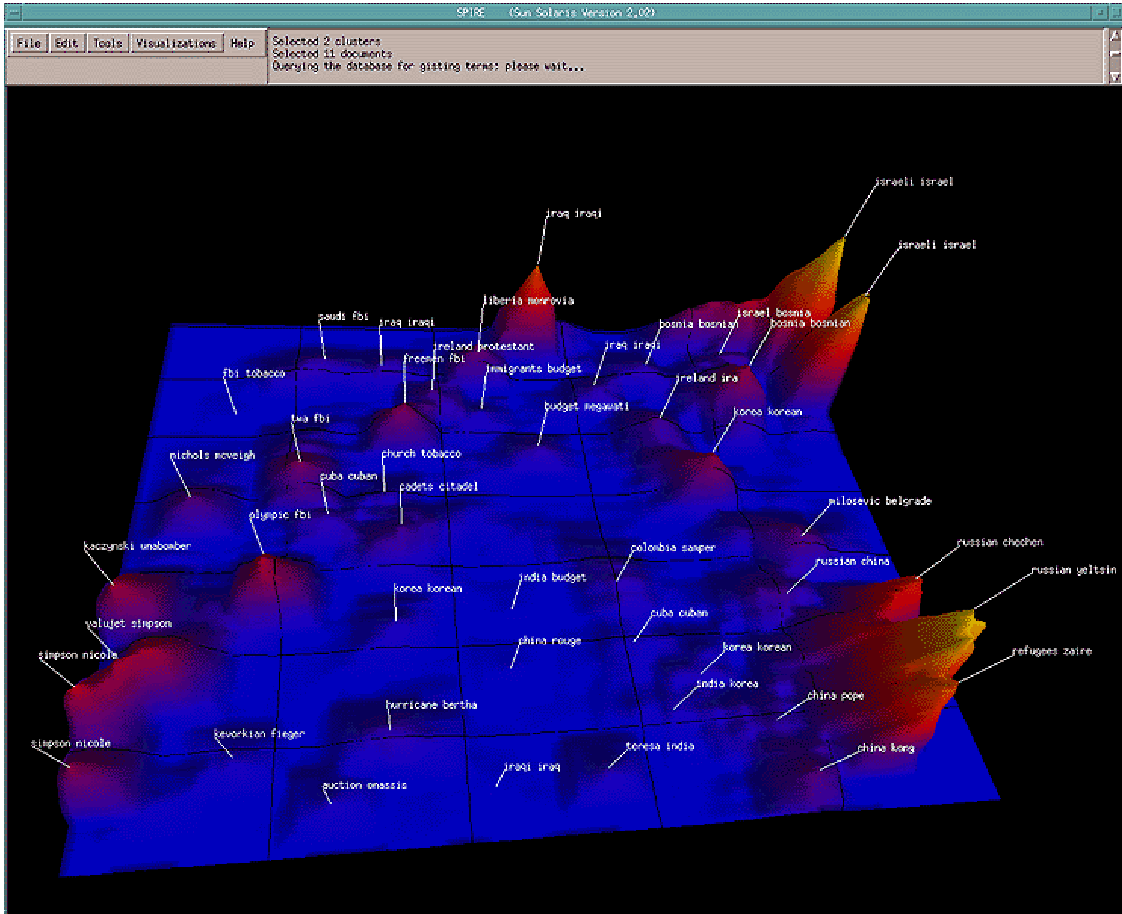
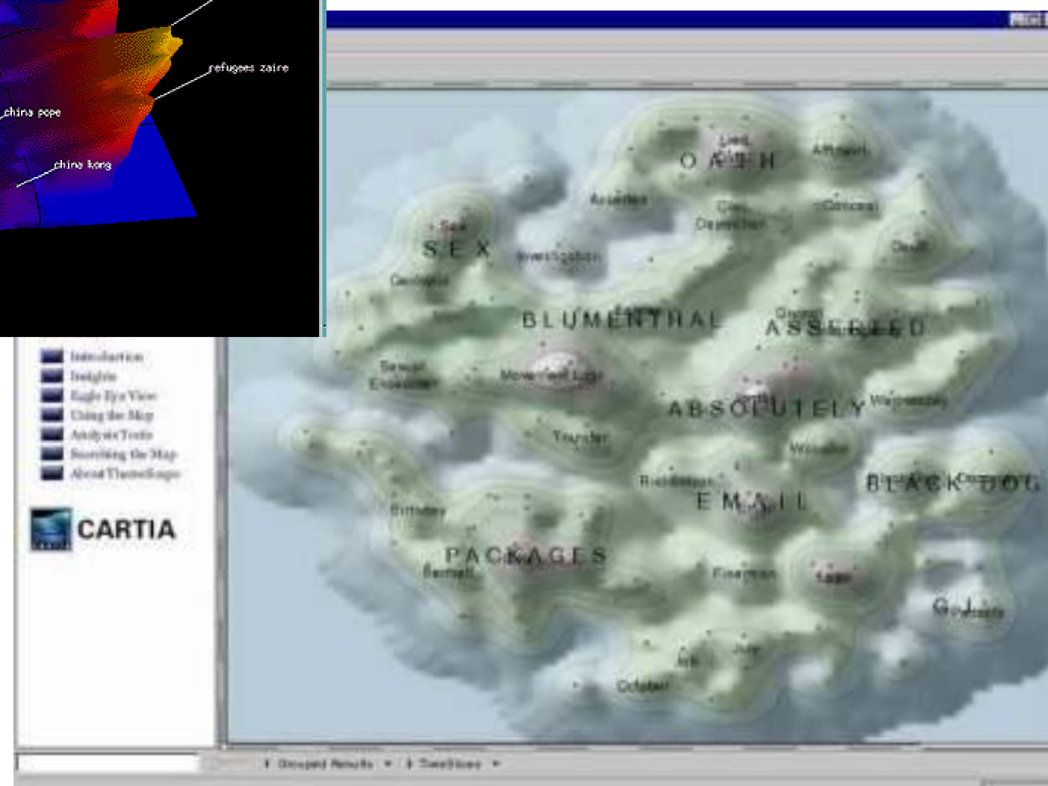
Cluster Pruning

- Exercises
  - To find the nearest leader in step 1, how many cosine computations do we do?
    - Why did we have $\sqrt{N}$ in the first place?
  - What is the effect of the constants *b1, b2* on the previous slide?
  - Devise an example where this is *likely to* fail – i.e., we miss one of the *K* nearest docs.
    - *Likely* under random sampling.

Why Text Clustering?

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

Searching + Browsing


SPIRE (Sun Solaris Version 2.02)
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
liberia monrovia
saudi fbi
iraq iraqi
israel bosnia
bosnia bosnian
bosnia bosnian
ireland protestant
freemen fbi
iraq iraqi
immigrants budget
ireland ira
fbi tobacco
budget megawati
twa fbi
korea korean
nichols mcveigh
church tobacco
cuba cuban
cadets citadel
milosevic belgrade
olympic fbi
kaczynski unabomber
colombia samper
russian chechen
russian china
russian yeltsin
india budget
korea korean
cuba cuban
valujet simpson
china rouge
refugees zaire
simpson nicole
korea korean
india korea
china pope
hurricane bertha
simpson nicole
kevorkian fieger
teresa india
china kong
iraqi iraq
auction onassis

CARTIA
OATH
SEX
BLUMENTHAL
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 »

edit

HSBC Creates $5 Billion Fund to Boost Credit Access (Update1)

ITV.com

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.

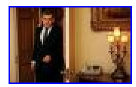
HSBC starts $5 billion lending fund for small, medium businesses MarketWatch

HSBC makes £1bn credit available BBC News

The Press Association - Shropshire Star - Sky News - Times Online

all 145 news articles » LON:HSBA - HCS

More "world coordination" on rates needed



Telegraph.co.uk

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 »

Sun's shining on tracker customers but be prepared for the rainy days

BBC News

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 »

Show more stories | Show fewer stories

Sci/Tech »

edit

Parents warned on fake consoles

Reuters

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.

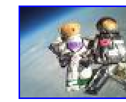
Fake Cheap Nintendo DS Consoles Could Kill ITProPortal

Warning issued over dangerous consoles Inquirer

Idealo Product News - Reg Hardware - Computeractive - Reuters UK

all 74 news articles »

One giant leap for 'teddy-nauts'



The Age

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 »

Minister checks on science exams

BBC News

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

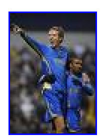
all 147 news articles »

Show more stories | Show fewer stories

Sport »

edit

• Rock-bottom Baggies held to 1-1 draw by Portsmouth



Goal.com

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 »

Entertainment »

edit

Ricky Gervais: Jonathan Ross is devastated by phone scandal

Sky News

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 **...**

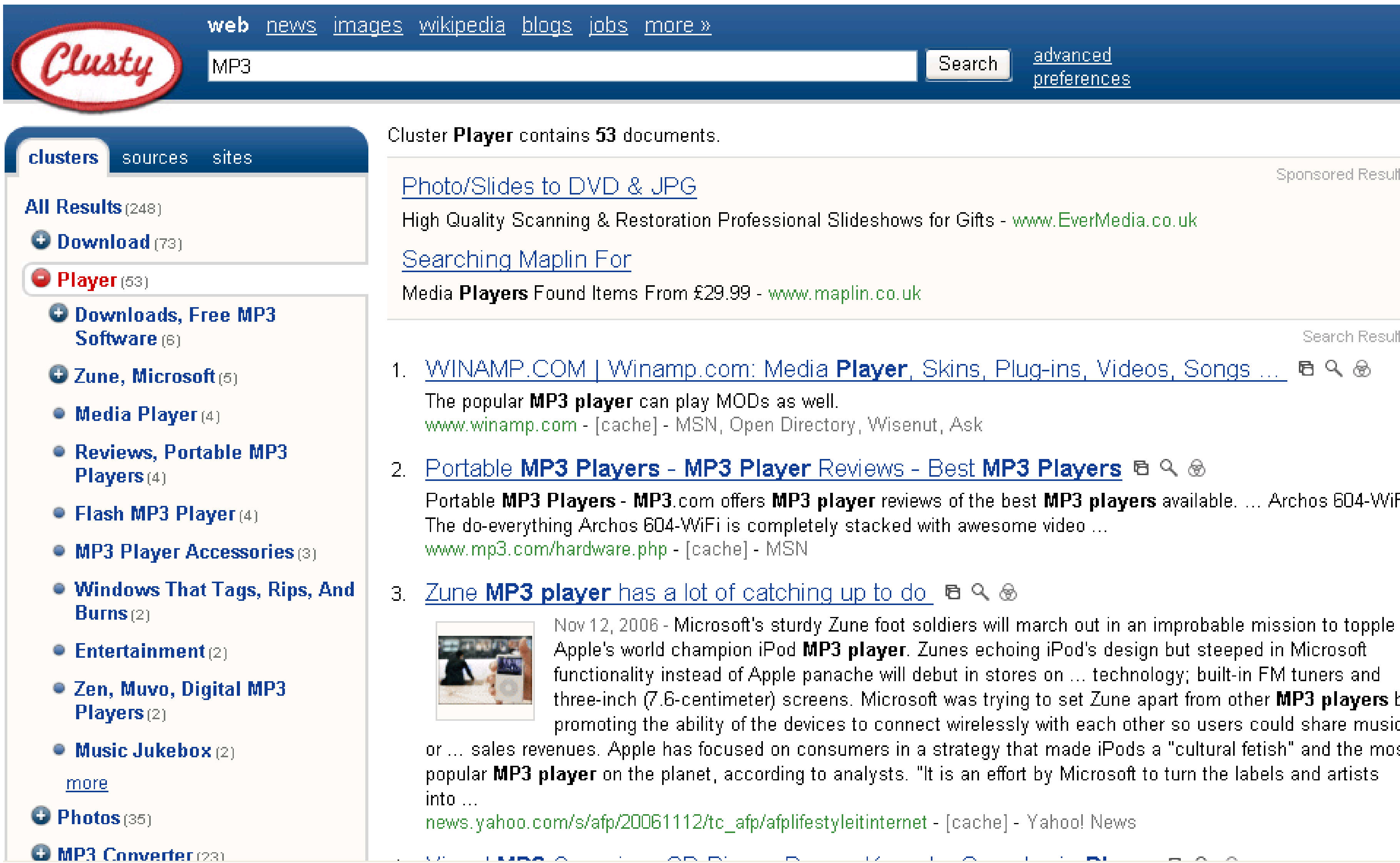
Jonathan Ross and the British Comedy Awards: wot no Woss? Times Online

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 »

http://clusty.com/

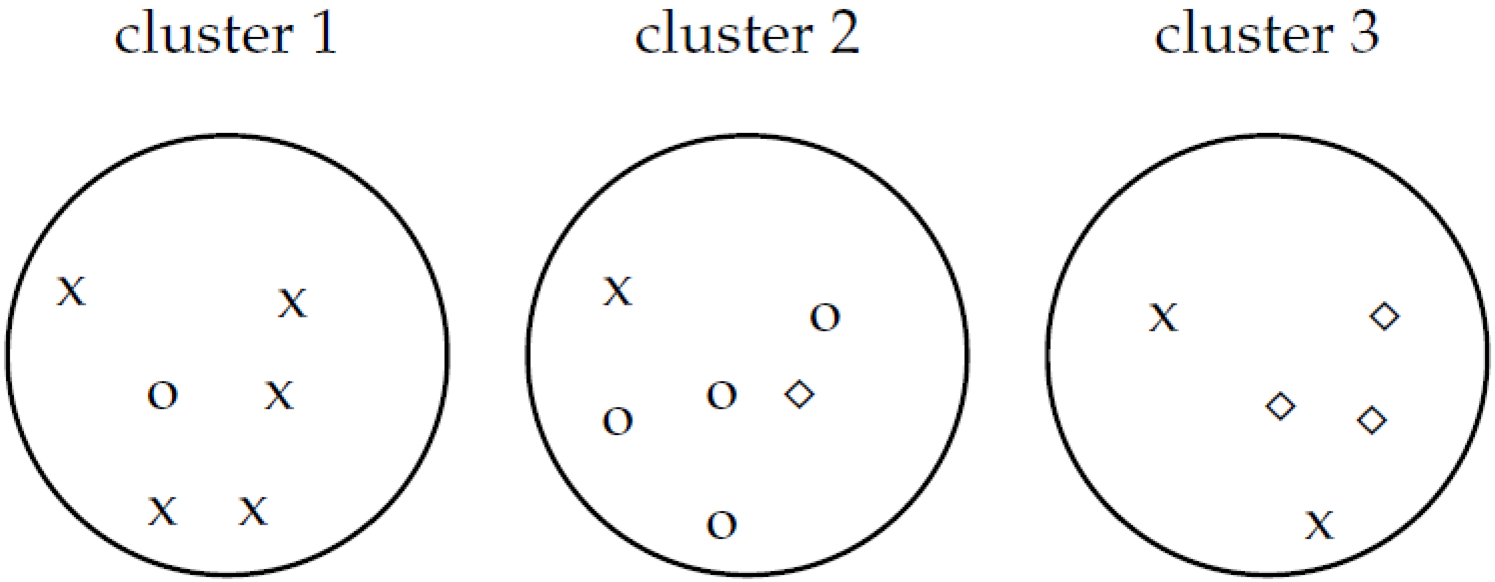


What Clustering Is Good?

- 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



majority class
x (5)

majority class
o (4)

majority class
d (3)

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					

$$RI = (TP+TN)/(TP+FP+FN+TN) = 6/10 = 0.6$$

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

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

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

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

while stopping criterion has not been met

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

    do $\omega_k \leftarrow \{\}$

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

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

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

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

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

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

Time Complexity

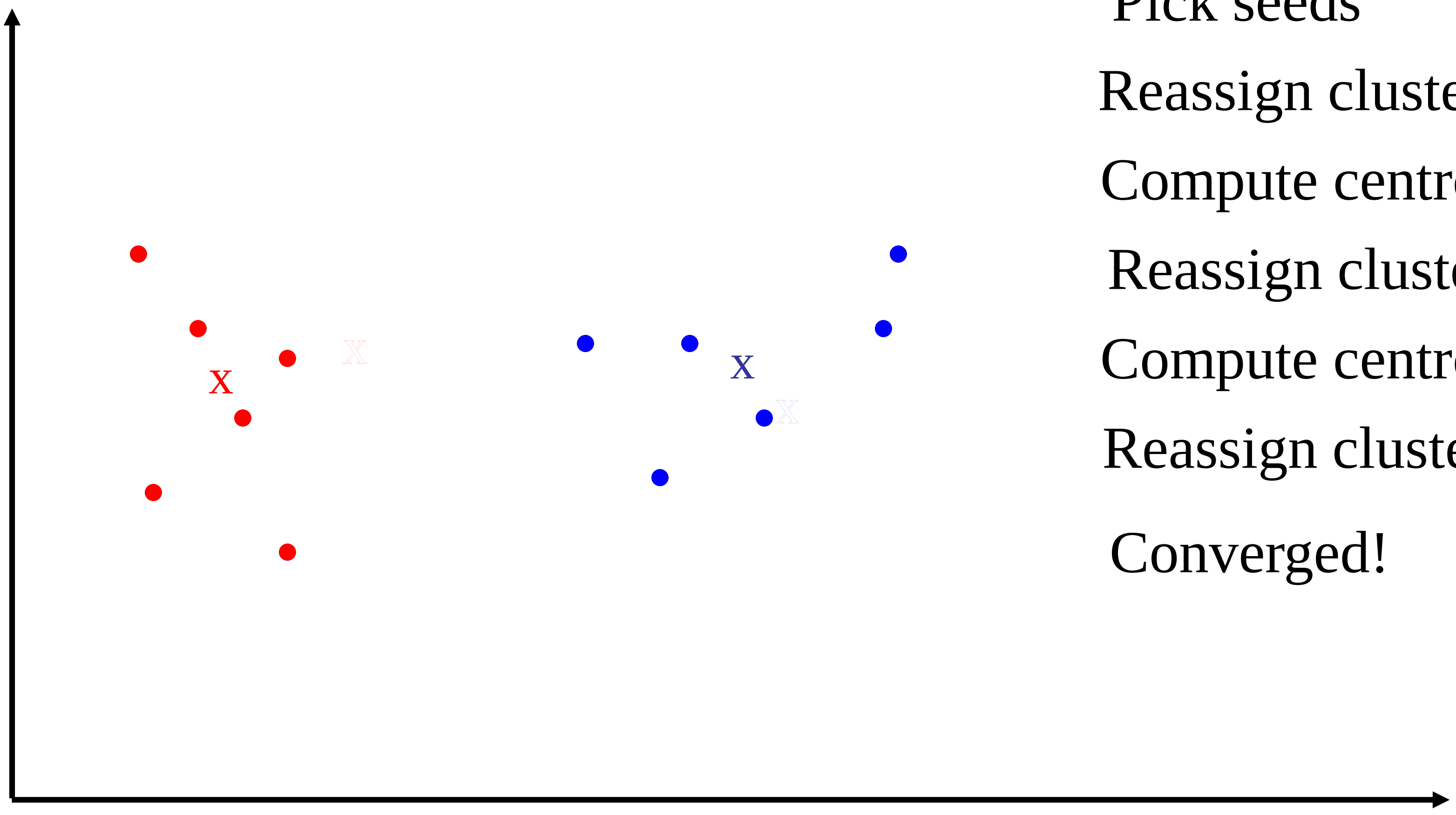
- Computing distance between two docs is $O(m)$ where m is the dimensionality of the vectors.
- Reassigning clusters: $O(kn)$ distance computations, or $O(knm)$.
- Computing centroids: Each doc gets added once to some centroid: $O(nm)$.
- Assume these two steps are each done once for i iterations: $O(iknm)$.

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

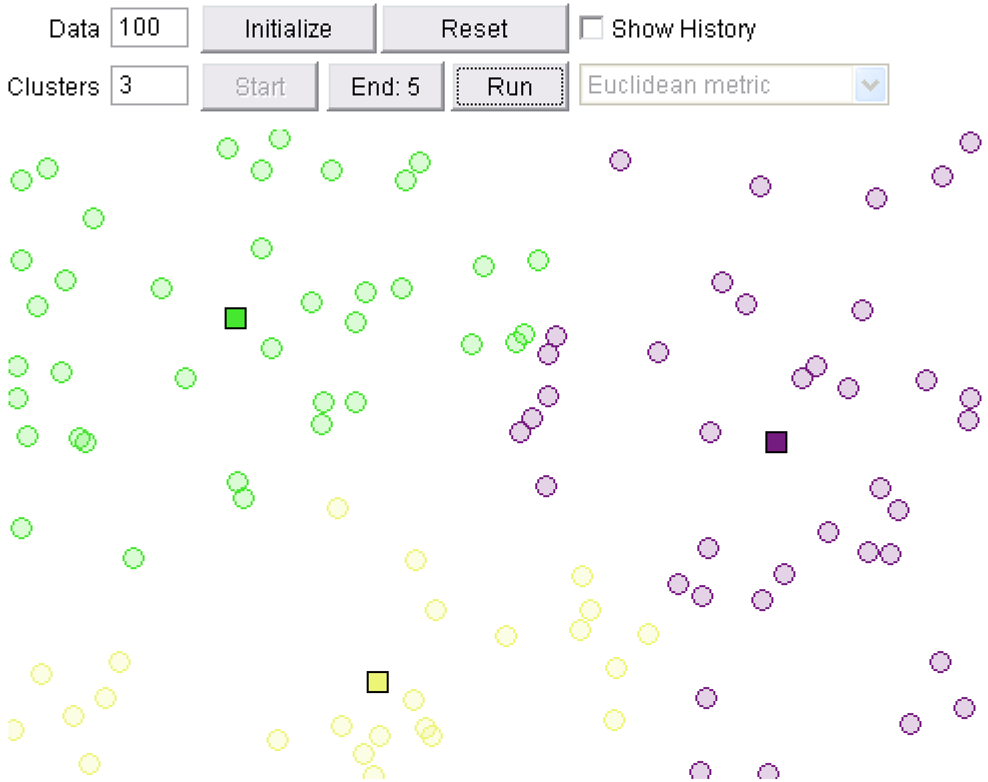
Compute centroids

Reassign clusters

Converged!

K-Means – Demo

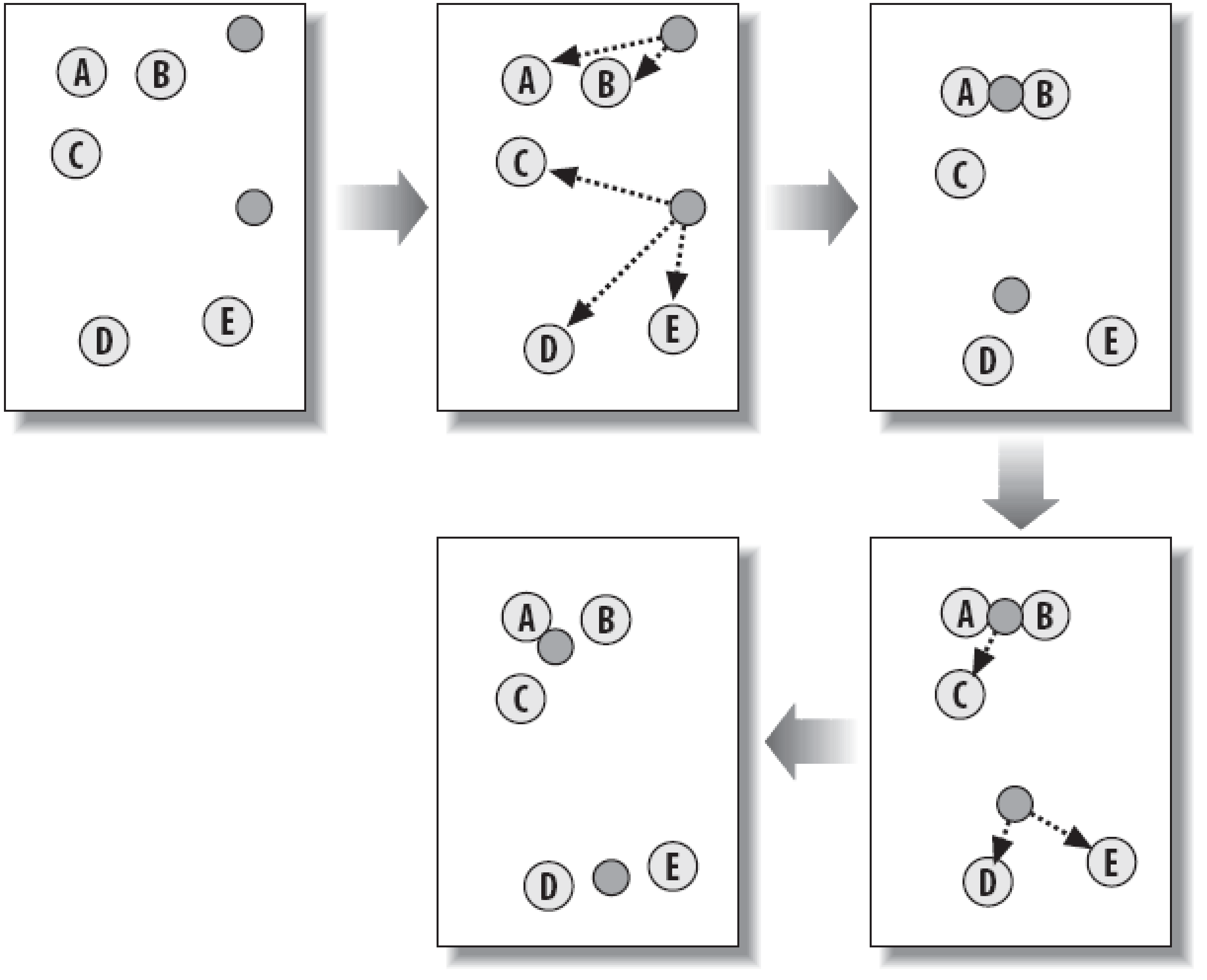
- http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletKM.html



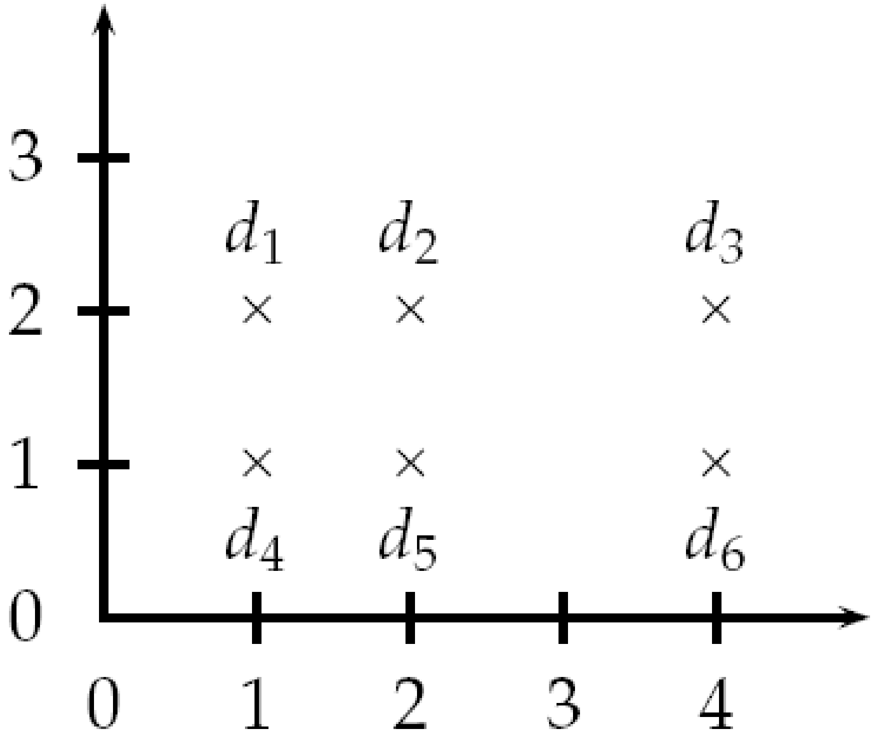
*k*Means – Exercise

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

*k*Means – 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++