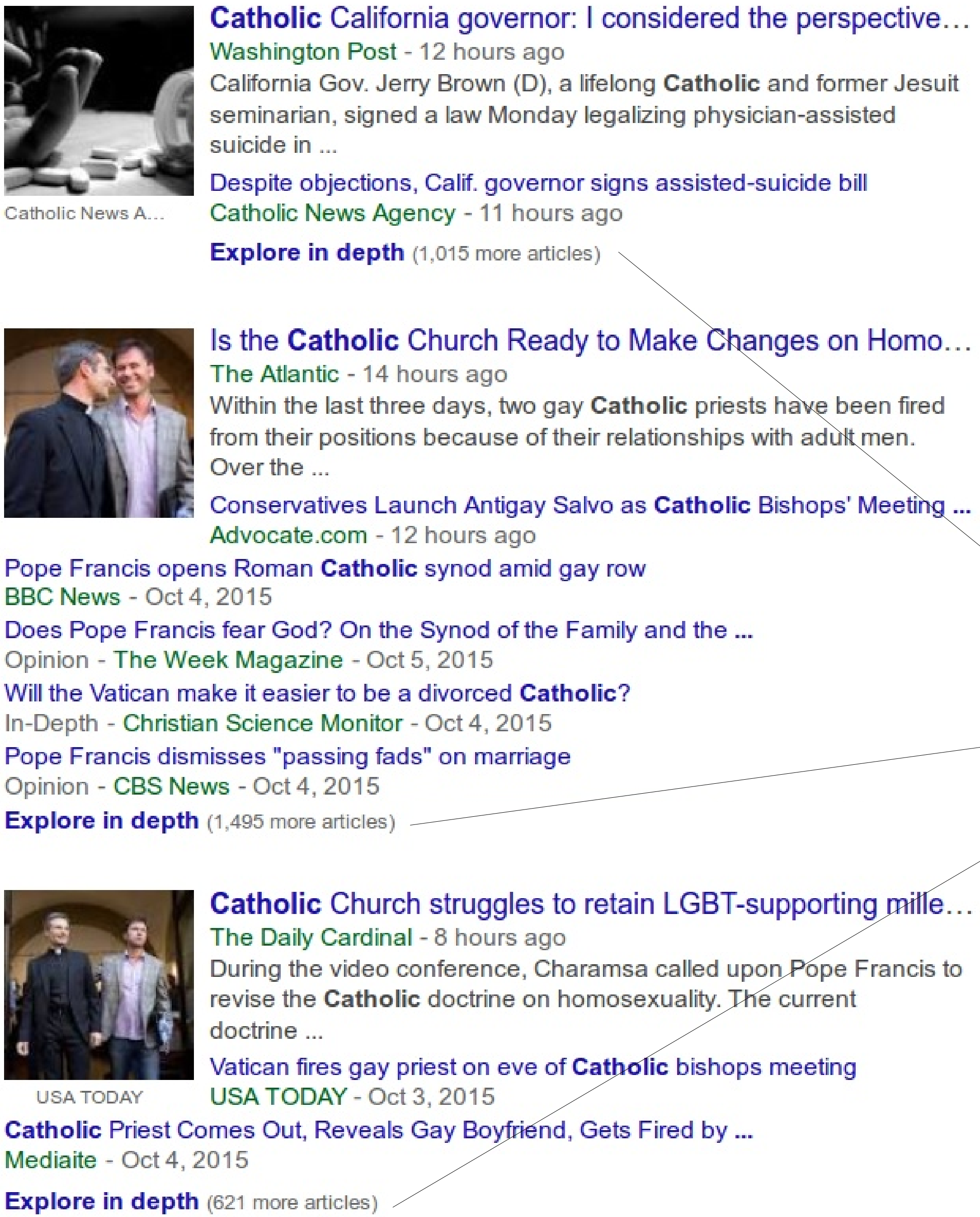
Clustering

Class	Algorithmic Methods of Data Mining
Program	M. Sc. Data Science
University	Sapienza University of Rome
Semester	Fall 2015
Lecturer	Carlos Castillo http://chato.cl/

Sources:

- Mohammed J. Zaki, Wagner Meira, Jr., Data Mining and Analysis: Fundamental Concepts and Algorithms, Cambridge University Press, May 2014. Part 3. [download]
- Evimaria Terzi: Data Mining course at Boston University http://www.cs.bu.edu/~evimaria/cs565-13.html



Catholic California governor: I considered the perspective…
Washington Post - 12 hours ago
California Gov. Jerry Brown (D), a lifelong **Catholic** and former Jesuit seminarian, signed a law Monday legalizing physician-assisted suicide in ...

Despite objections, Calif. governor signs assisted-suicide bill
Catholic News Agency - 11 hours ago

Catholic News A…

Explore in depth (1,015 more articles)

Is the **Catholic** Church Ready to Make Changes on Homo…
The Atlantic - 14 hours ago
Within the last three days, two gay **Catholic** priests have been fired from their positions because of their relationships with adult men. Over the ...

Conservatives Launch Antigay Salvo as **Catholic** Bishops' Meeting ...
Advocate.com - 12 hours ago

Pope Francis opens Roman **Catholic** synod amid gay row
BBC News - Oct 4, 2015

Does Pope Francis fear God? On the Synod of the Family and the ...
Opinion - The Week Magazine - Oct 5, 2015

Will the Vatican make it easier to be a divorced **Catholic**?
In-Depth - Christian Science Monitor - Oct 4, 2015

Pope Francis dismisses "passing fads" on marriage
Opinion - CBS News - Oct 4, 2015

Explore in depth (1,495 more articles)

Catholic Church struggles to retain LGBT-supporting mille…
The Daily Cardinal - 8 hours ago
During the video conference, Charamsa called upon Pope Francis to revise the **Catholic** doctrine on homosexuality. The current doctrine ...

Vatican fires gay priest on eve of **Catholic** bishops meeting
USA TODAY - Oct 3, 2015

USA TODAY

Catholic Priest Comes Out, Reveals Gay Boyfriend, Gets Fired by ...
Mediaite - Oct 4, 2015

Explore in depth (621 more articles)

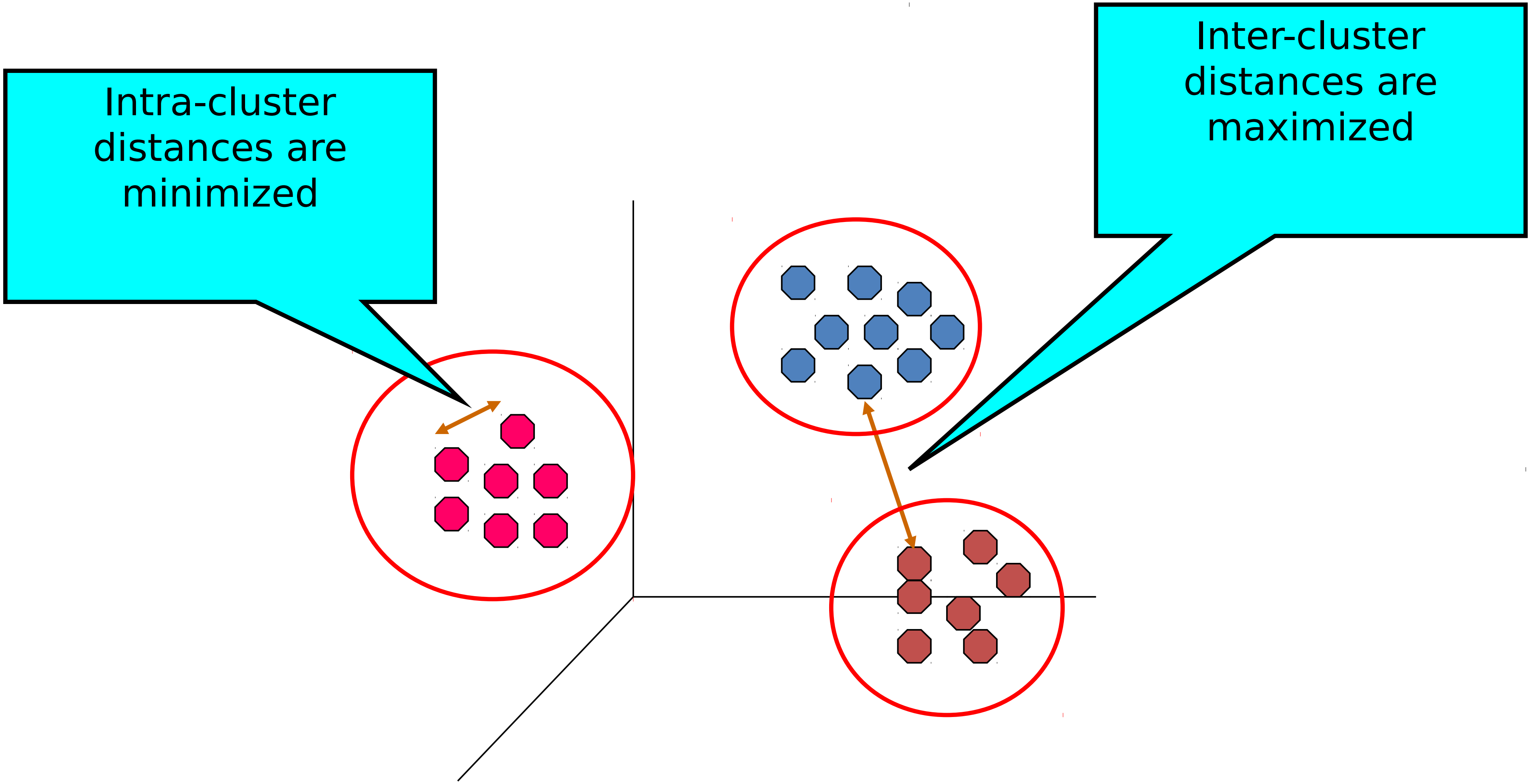
Why these sizes?

Why 3 groups instead of 2?

Clustering

- Given a set of elements (e.g. documents)
- Group similar elements together
- So that:
  - Inside a group, elements are similar
  - Across groups, elements are different

What is clustering?

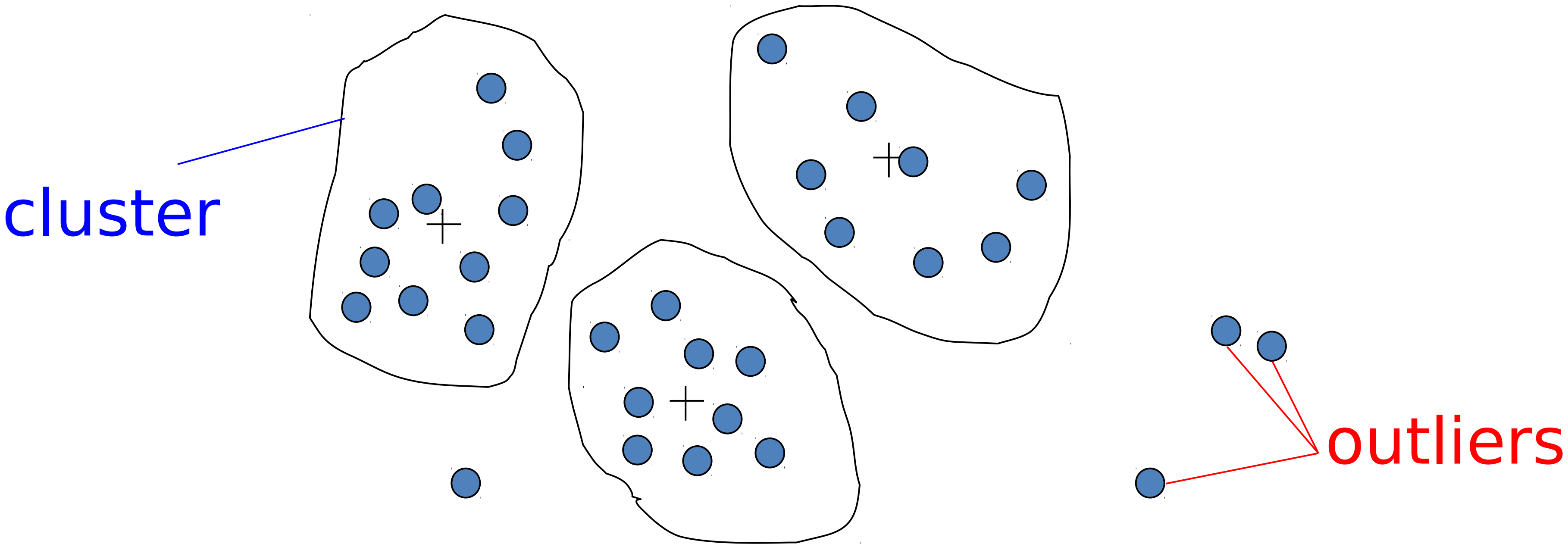


Intra-cluster distances are minimized

Inter-cluster distances are maximized

Outliers

- **Outliers** are **objects that do not belong to any cluster** or form clusters of very small cardinality



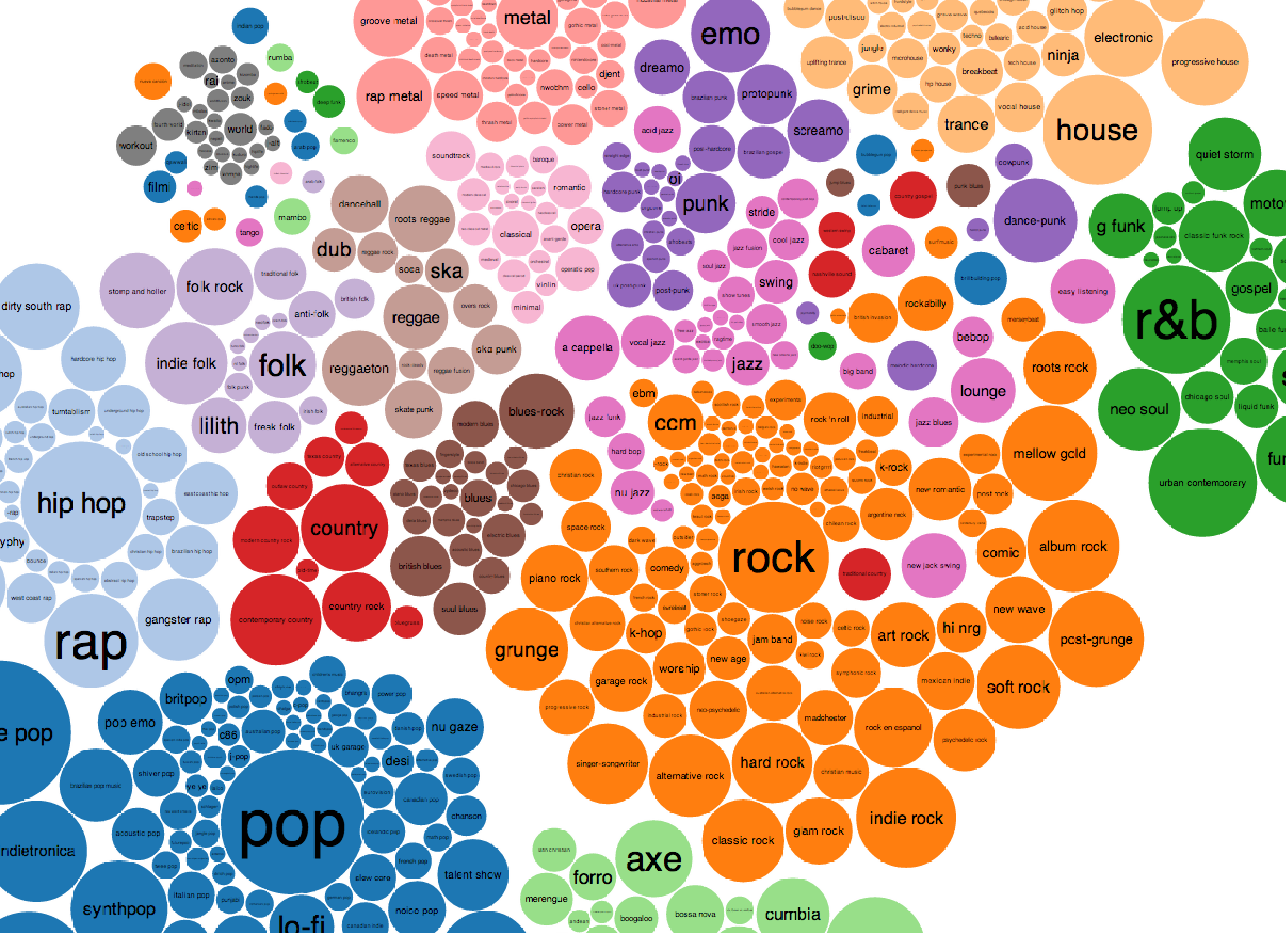
- In some applications we are interested in discovering outliers, not clusters (outlier analysis)

Why do we cluster?

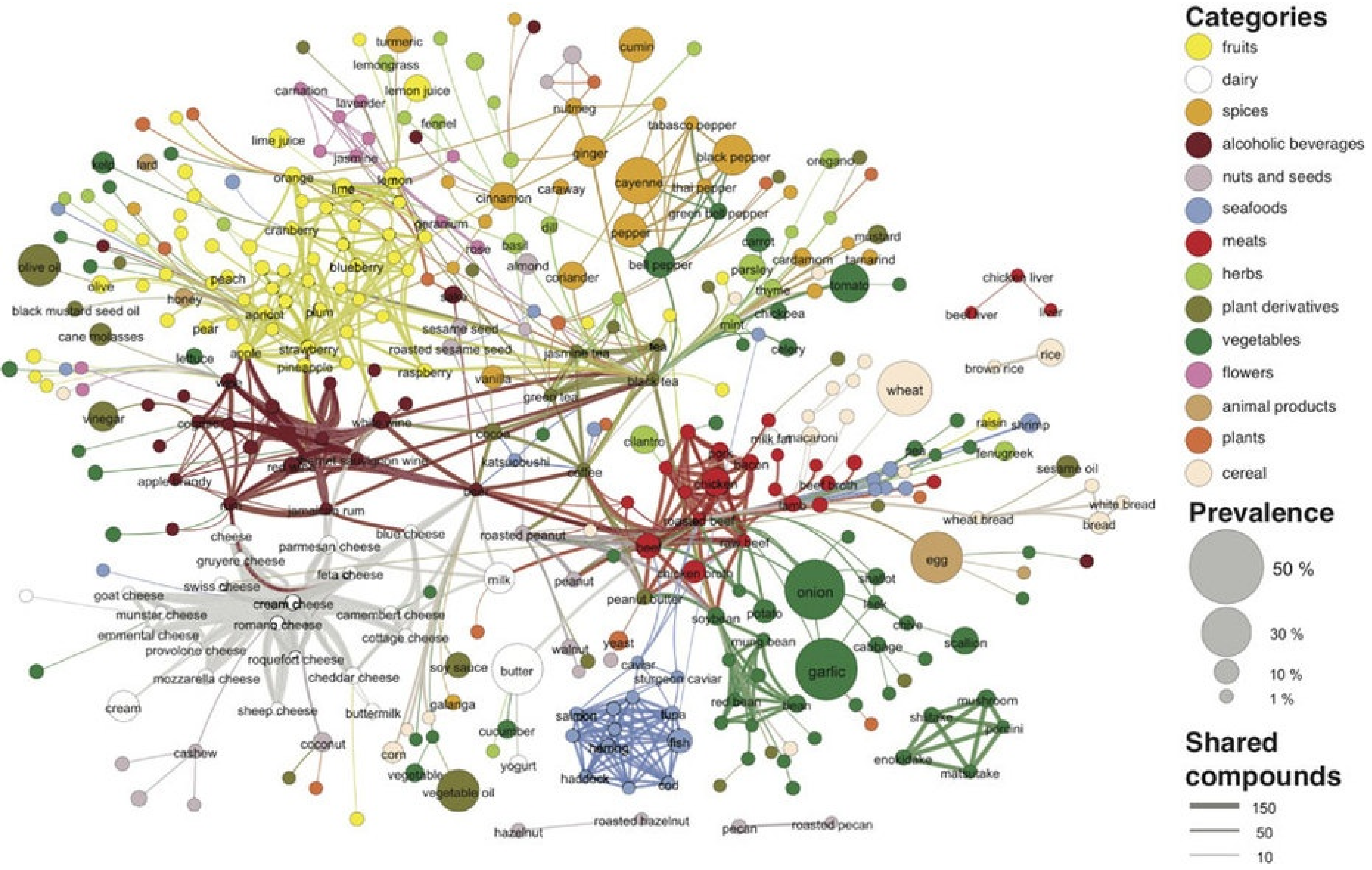
- Clustering results are used:
  - As a stand-alone tool to get insight into data distribution
    - Visualization of clusters may unveil important information
  - As a preprocessing step for other algorithms
    - Efficient indexing or compression often relies on clustering

Applications

- Image Processing
  - cluster images based on their visual content
- Web
  - Cluster groups of users based on their access patterns on webpages
  - Cluster webpages based on their content
- Bioinformatics
  - Cluster similar proteins together (similarity wrt chemical structure and/or functionality etc)
- Many more…



http://musicmachinery.com/2013/09/22/5025/

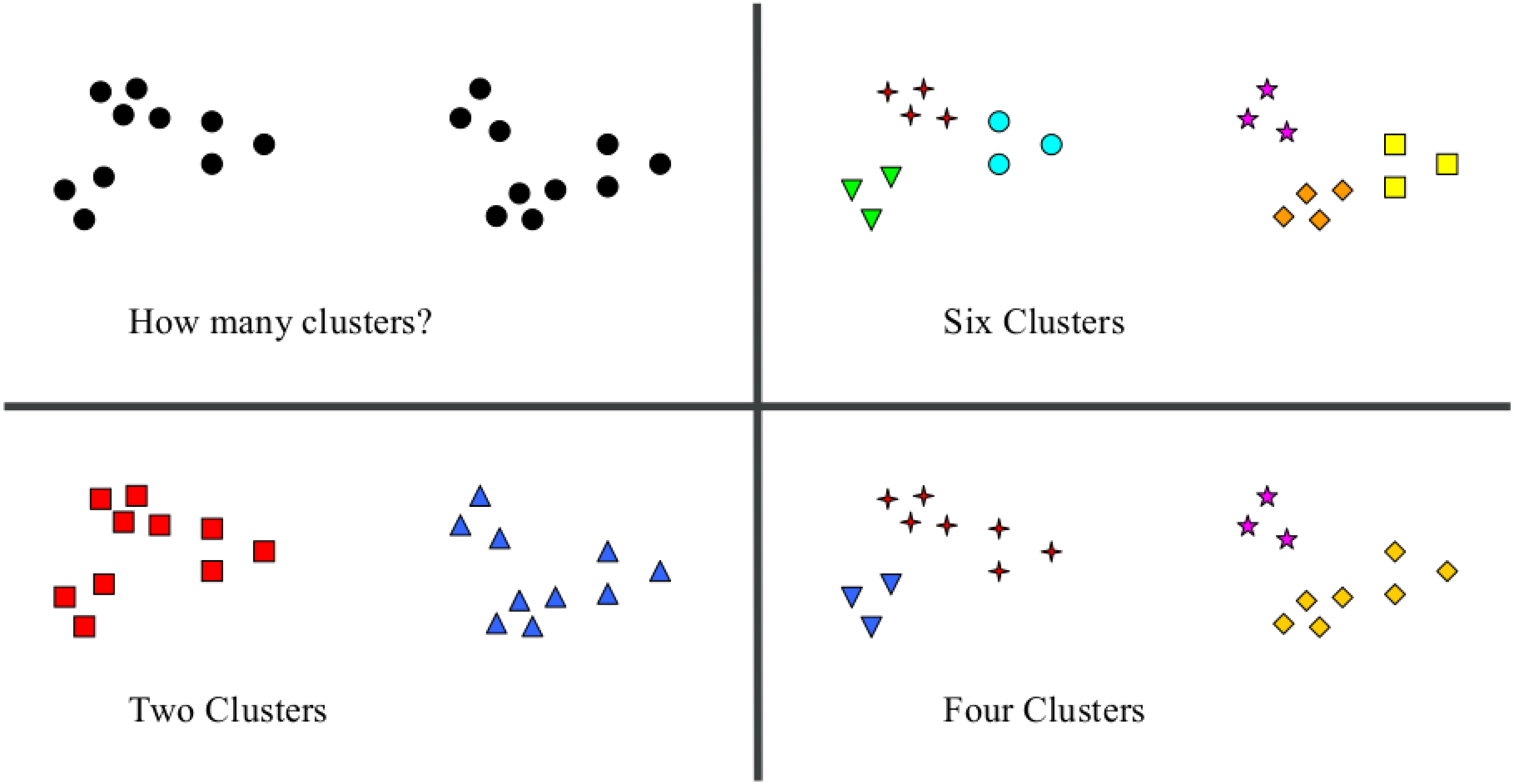


http://www.nature.com/articles/srep00196/figures/2

Clustering questions

- How many clusters?
  - Given as input or determined by algorithm
- How good is a clustering?
  - Intra similarity, inter similarity, number of clusters
- Can an element belong to > 1 cluster?
  - Hard clustering vs Soft clustering

How many clusters?



How many clusters?

Six Clusters

Two Clusters

Four Clusters

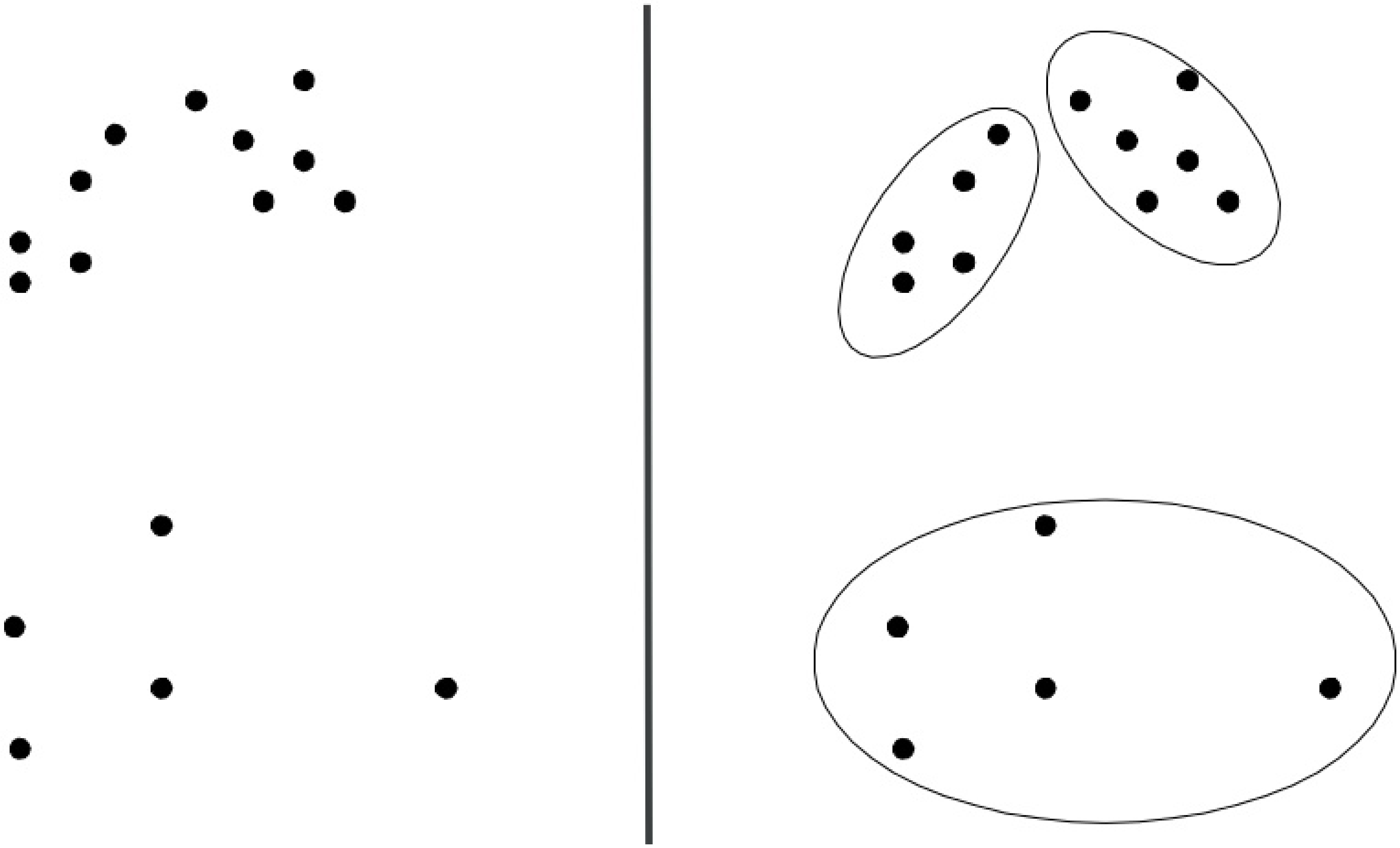
Types of clusterings

- Partitional
  - each object belongs in exactly one cluster
- Hierarchical
  - a set of nested clusters organized in a tree

Partitional algorithms

- partition the n objects into k clusters
  - each object belongs to exactly one cluster
  - the number of clusters k is given in advance

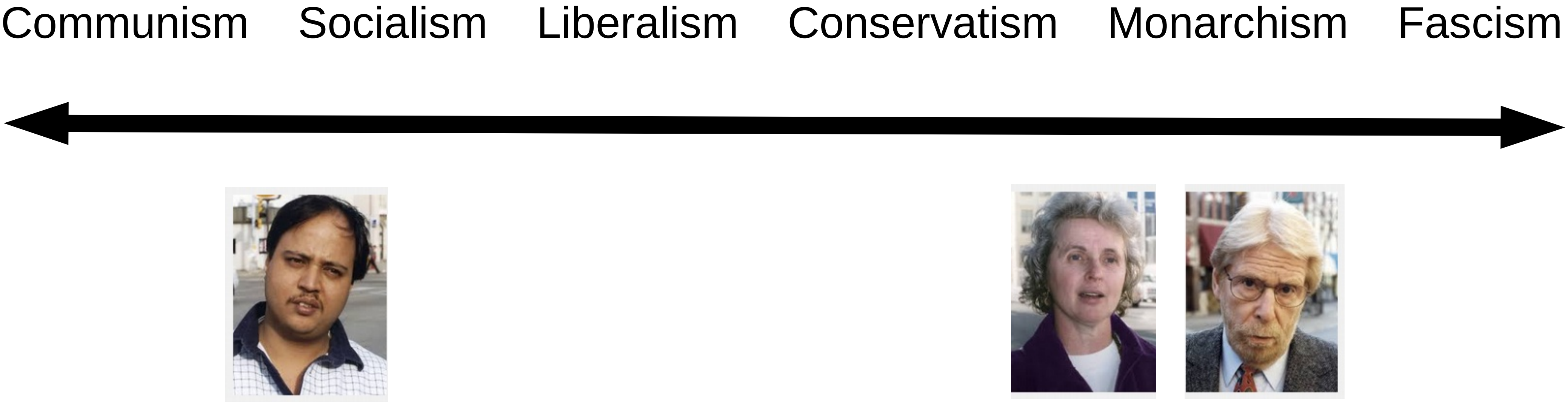
Partitional clustering



Original points

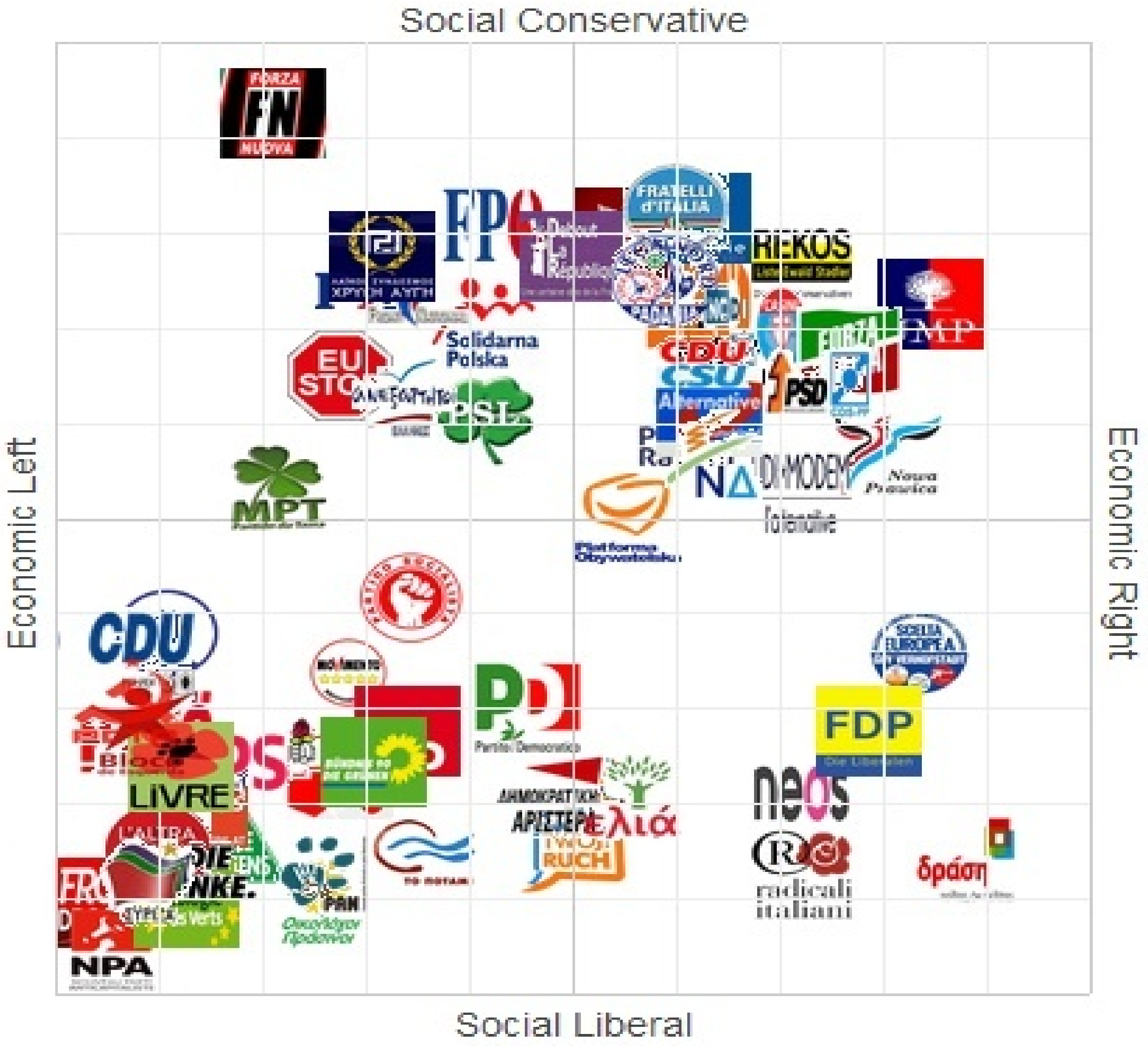
Partitional clustering

Example: 1-dimensional clustering



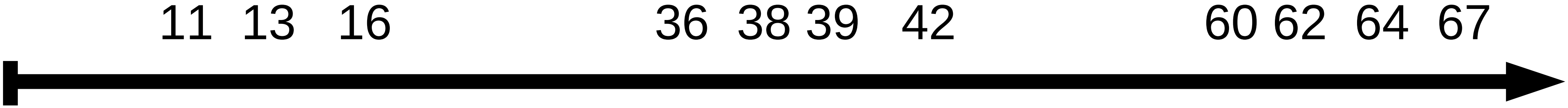
Communism Socialism Liberalism Conservatism Monarchism Fascism

Parenthesis: 2D political spectrum



http://www.termometropolitico.it/119350_dai-modelli-collocazione-nello-spazio-politico-test-per-elezioni-europee-2014.html

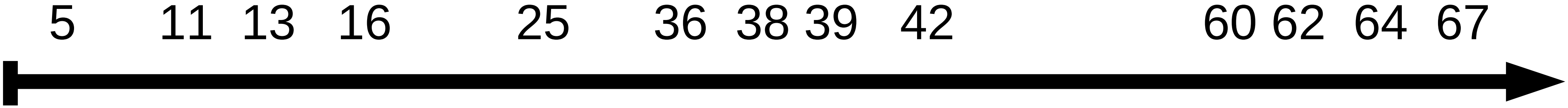
1 dimensional clustering



11 13 16 36 38 39 42 60 62 64 67

How would you cluster this data? Why?

1 dimensional clustering



What about now, how would you cluster?

Two very important metrics

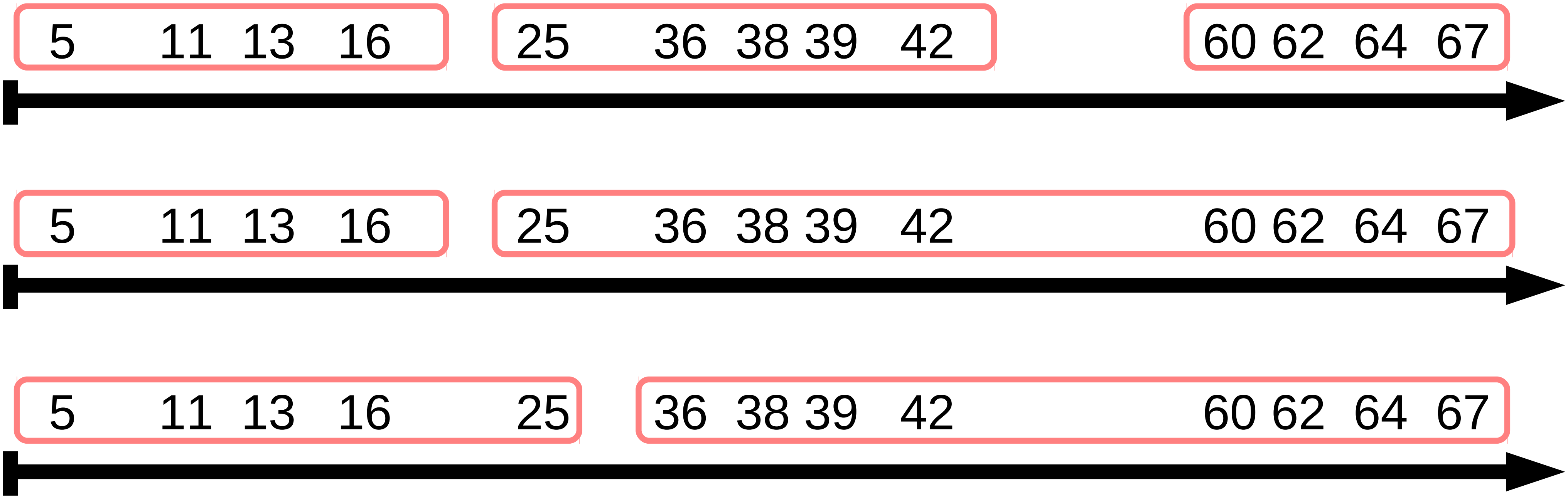
- Minimum **inter**-cluster distance

  (should be large) $\min_{i,j} \min_{u \in C_i, v \in C_j} d(u,v)$

- Maximum **intra**-cluster distance

  (should be small) $\max_{i} \max_{u \in C_i, v \in C_i} d(u,v)$

1 dimensional clustering



5 11 13 16

25 36 38 39 42

60 62 64 67

5 11 13 16

25 36 38 39 42 60 62 64 67

5 11 13 16 25

36 38 39 42 60 62 64 67

Exercise:

For each of these 3 clusterings:

- *Compute minimum inter-cluster distance.*
- *Compute maximum intra-cluster distance.*

http://chato.cl/2015/data_analysis/exercise-answers/clustering_exercise_01_answer.txt