



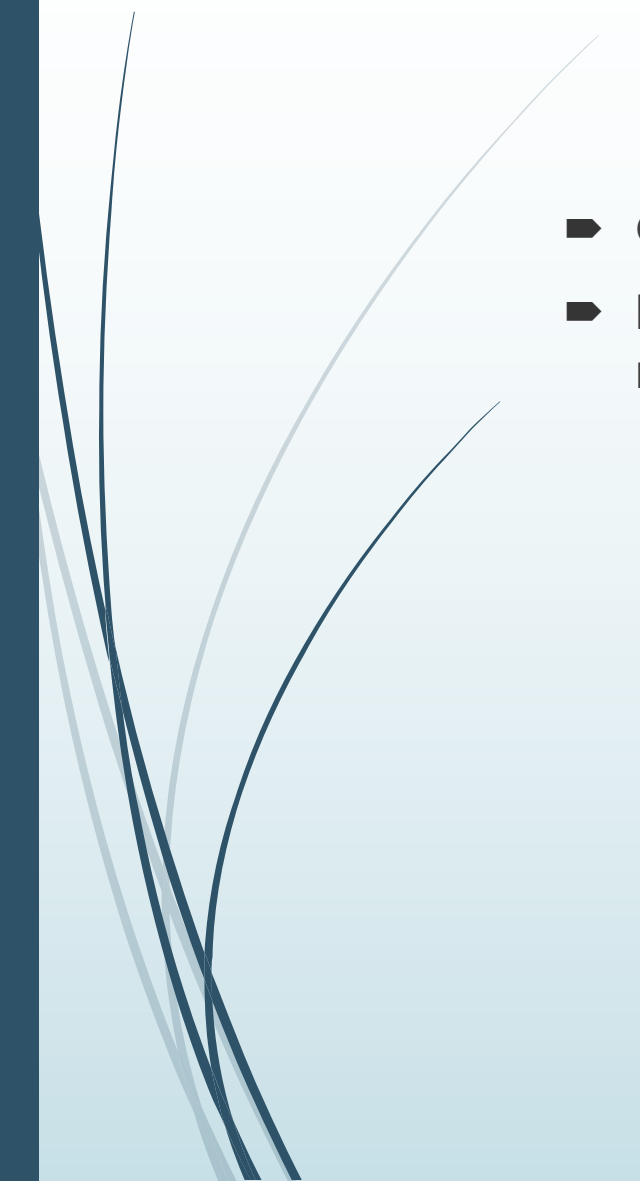
Data clustering: 50 years beyond K-means

Anil K. Jain- Pattern Recognition Letters 31 (2010) 651–666

Research Review by- William Steimel



Motivation

- Over 4500 Citations
 - I have not studied Unsupervised Learning yet but would like to understand it more
- 



Abstract

- “Cluster analysis is the formal study of methods and algorithms for grouping, or clustering, objects according to measured or perceived intrinsic characteristics or similarity”
 - Example: In the scientific kingdom animals organisms are grouped based on: domain, kingdom, phylum, class
- Clustering is a form of unsupervised learning as the data is unlabeled
- One of the most popular clustering algorithms is the K-Means algorithm which was proposed in 1955
 - It is still widely used despite all the time that has passed
- This paper provides an overview on clustering, summarizes well known clustering methods, presents challenges with design of clustering algorithms, as well as future directions of clustering.



1. Introduction

- Advances in sensing and storage technology have created many high-volume, high-dimensional datasets
 - We now create petabytes of data through E-mails, blogs, transaction data, web pages, sensor technology
 - Many of these datasets are unstructured which poses a challenge for analysis
- Due to this large volume of data, procedures are needed to automatically understand, process, and summarize the data
- Data analysis techniques can be partitioned into two types including:
 - Exploratory/Descriptive – No predetermined hypothesis/or models but would like to understand the data characteristics
 - Confirmatory/Inferential – Aimed at confirming whether a predetermined hypothesis/model is valid/true



1. Introduction

- In pattern recognition, we generally want to predict a result on unseen test data given some training data.
 - Supervised Learning (Classification) - Labeled data
 - Unsupervised Learning (Clustering) – Unlabeled data
- Clustering is often considered a more challenging problem than Classification



2. Data Clustering

- Cluster analysis is a “a statistical classification technique for discovering whether the individuals of a population fall into different groups by making quantitative comparisons of multiple characteristics.”
- Clustering is largely subjective- What is the notion of similarity?
 - Clusters can differ in shape, size, and density
- Due to high dimensionality of data and the many application areas that clustering is applied to 1000's of clustering algorithms were developed and continue to be developed



2.1 Why Clustering ?

- Cluster analysis has applications in most disciplines that work with multivariate data – some examples below
 - Image segmentation
 - Document Grouping
 - Customer segmentation
 - Grouping of services in workforce management and planning
 - Genome data
- Clustering is used for three main purposes
 - To find Underlying structure in data (Gain Insight into the data)
 - Natural Classification (Identify degree of similarity between objects)
 - Compression (Method for organizing and summarizing data through clusters)



2.2 Historical Developments

- Clustering has multidisciplinary origins and has been developed through a combination of fields:
 - Taxonomists
 - Social Scientists
 - Psychologists
 - Biologists
 - Statisticians
 - Mathematicians
 - Engineers
 - Computer Scientists
 - Medical Researchers
- Clustering algorithms can be broken out into two groups:
 - Hierarchical- Recursively find nested clusters either in a divisive(top down) or agglomerative (bottom up) method
 - Partitional – Find all clusters simultaneously as a partition of the data and do not create hierarchical structure
 - K-means is the most popular partitional clustering algorithm

2.3 K-means algorithm

Let $X = \{x_i\}, i = 1, \dots, n$ be the set of n d – dimensional points to be clustered into a set of K clusters, $C = \{c_k, k = 1, \dots, K\}$.

K-Means Intuition:

- K-means finds a partition (split) where the squared error is minimized between the empirical mean of the cluster and the points of the cluster

- μ_k – mean of cluster c_k

- The squared error between μ_k and the points in cluster c_k

- Is defined as

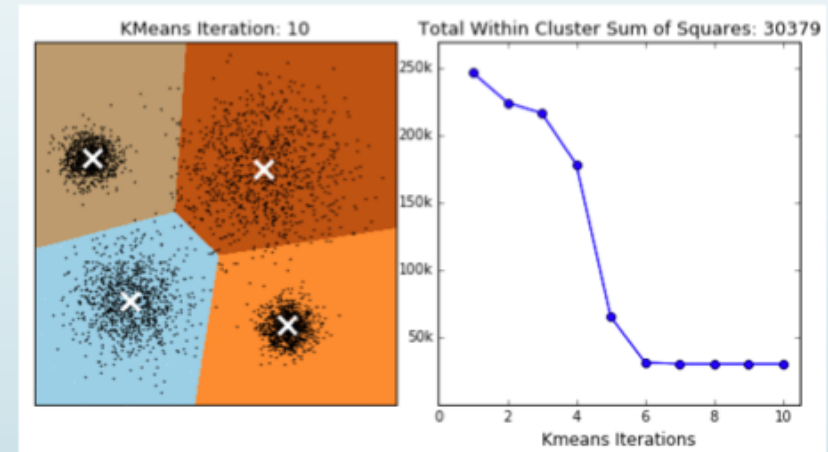
$$J(c_k) = \sum_{x_i \in c_k} \|x_i - \mu_k\|^2.$$

- The goal of K-means is to minimize the sum of squared error over all K clusters.

$$J(C) = \sum_{k=1}^K \sum_{x_i \in c_k} \|x_i - \mu_k\|^2.$$

Steps:

1. Select an initial partition with K clusters; repeat steps 2 and 3 until cluster membership stabilizes.
2. Generate a new partition by assigning each pattern to its closest cluster center.
3. Compute new cluster centers.





2.4 Parameters of K-means

- The K-means algorithm requires three user-specified parameters
 - K- Number of clusters (Most Important)
 - Cluster initialization (How to initiate clusters)
 - Distance metric (Euclidean Distance is most common which leads to spherical or ball shaped clusters)
- K-means is heavily influenced by the number of clusters, initialization method, and distance metric

2.6 Major Approaches to clustering

- As mentioned previously there are thousands of clustering methods so this research then reviews the major approaches for clustering
 - **Centroid (Partitioning Based algorithms)** - Centroid or Partitioning based Clustering essential create partitions or clusters based on a centroid
 - K-means and variants
 - **Hierarchical Clustering** - Hierarchical clustering aims to group data points in an order of rank
 - Divisive Clustering, Agglomerative Clustering
 - **Density Based Clustering** - Density based clustering differ from centroid based as the goal is to model points into clusters based on density instead of distance to a central point.
 - DB Scan, Jarvis-Patrick Algorithm
 - **Probabilistic Clustering** - approach to clustering based on how probable a data point belongs to a cluster based on modeling of statistical data distributions.
 - EM Algorithm, Latent Dirichlet Allocation, Pachinko Allocation Model undirected graphical modeling
 - **Spectral Clustering** – represents data as nodes in a weighted graph



3. User's Dilemma

- Although there are many different clustering algorithms and they have been very successful there are also challenges associated with clustering as the definition of a cluster is vague, and defining an appropriate objective function and similarity measure is difficult.
 - (a) What is a cluster?
 - (b) What features should be used?
 - (c) Should the data be normalized?
 - (d) Does the data contain any outliers?
 - (e) How do we define the pair-wise similarity?
 - (f) How many clusters are present in the data?
 - (g) Which clustering method should be used?
 - (h) Does the data have any clustering tendency?
 - (i) Are the discovered clusters and partition valid?

The next section of this paper will discuss these challenges in further detail

3.1 Data Representation

- Data Representation is one of the most important factors that influence the performance of clustering algorithms
- If the representation/choice of features is good it is much easier to find the clusters

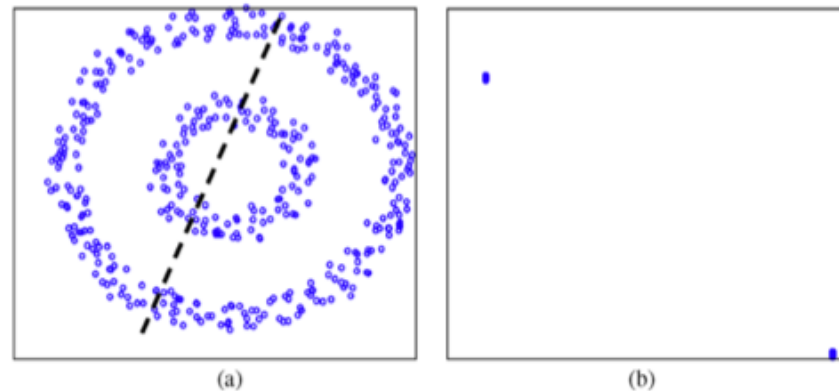


Fig. 5. Importance of a good representation. (a) "Two rings" dataset where K-means fails to find the two "natural" clusters; the dashed line shows the linear cluster separation boundary obtained by running K-means with $K=2$. (b) a new representation of the data in (a) based on the top 2 eigenvectors of the graph Laplacian of the data, computed using an RBF kernel; K-means now can easily detect the two clusters.

3.2 Purpose of Grouping

- Purpose of grouping is also important to consider when clustering – How do you as the user want to group the data?
- This section references an example dataset with 16 animals and 13 features.
 - When the animals are grouped by appearance – Animals were clustered into mammals / birds
 - When the animals were grouped by activity – animals were clustered into predators/ non-predators

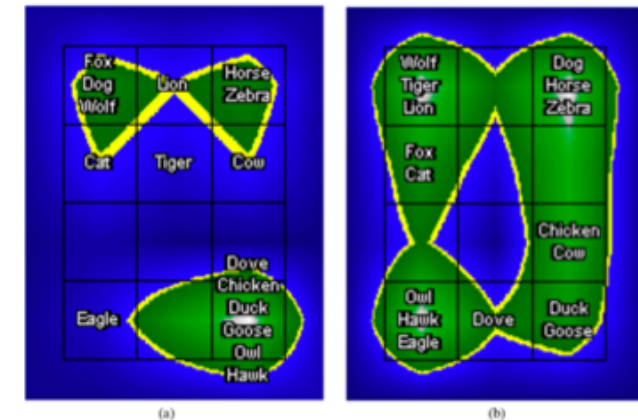


Fig. 6. Different weights on features result in different partitioning of the data. Sixteen animals are represented based on 13 Boolean features related to appearance and activity. (a) Partitioning with large weights assigned to the appearance based features; (b) a partitioning with large weights assigned to the activity features. The figures in (a) and (b) are excerpted from Pampalk et al. (2003), and are known as “heat maps” where the colors represent the density of samples at a location; the warmer the color, the larger the density.

3.3 Number of clusters

- The number of Clusters is a parameter determined by user
- Automatically determining the number of clusters is a difficult problem
 - Although methods/heuristics exist it is not easy to decide which value of K will return the most meaningful clusters

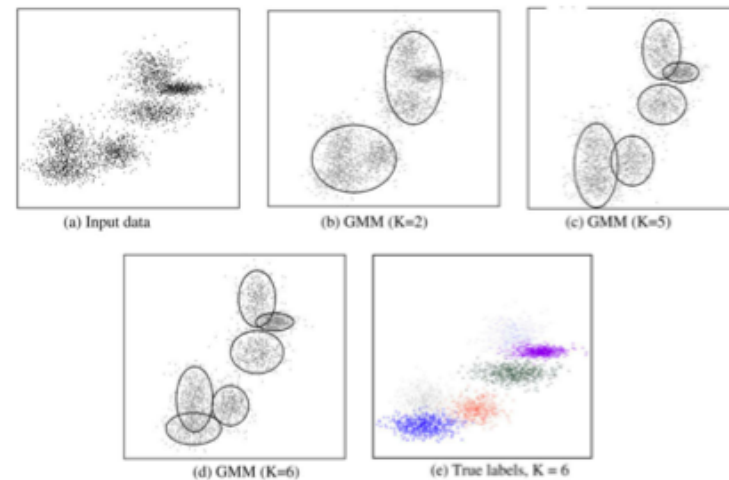


Fig. 7. Automatic selection of number of clusters, K . (a) Input data generated from a mixture of six Gaussian distributions; (b)-(d) Gaussian mixture model (GMM) fit to the data with 2, 5, and 6 components, respectively; and (e) true labels of the data.

3.4 Cluster Validity

- "Clustering Algorithms tend to find clusters in the data irrespective of whether any clusters are present"
- This means that we must be aware whether our dataset actually has "clustering tendency" before even performing clustering.
- The example on the right shows a dataset with no natural clustering after K-means with $K = 3$

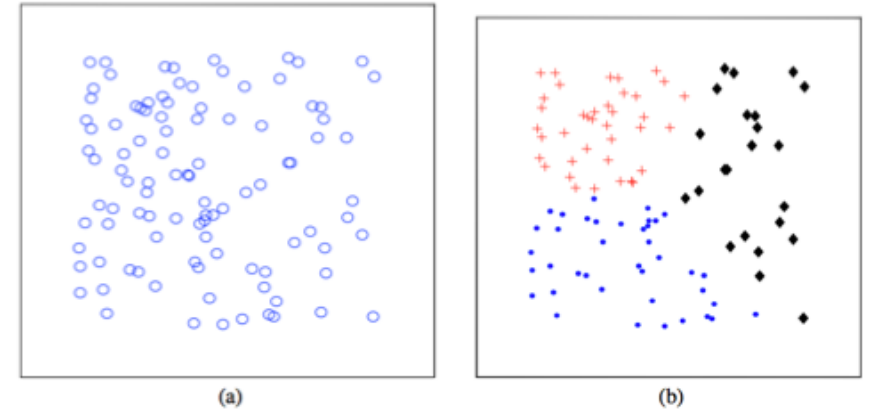


Fig. 8. Cluster validity. (a) A dataset with no "natural" clustering; (b) K-means partition with $K = 3$.

3.5 Comparing Clustering Algorithms

- Different clustering algorithms can often have completely different clustering behavior even on the same data.

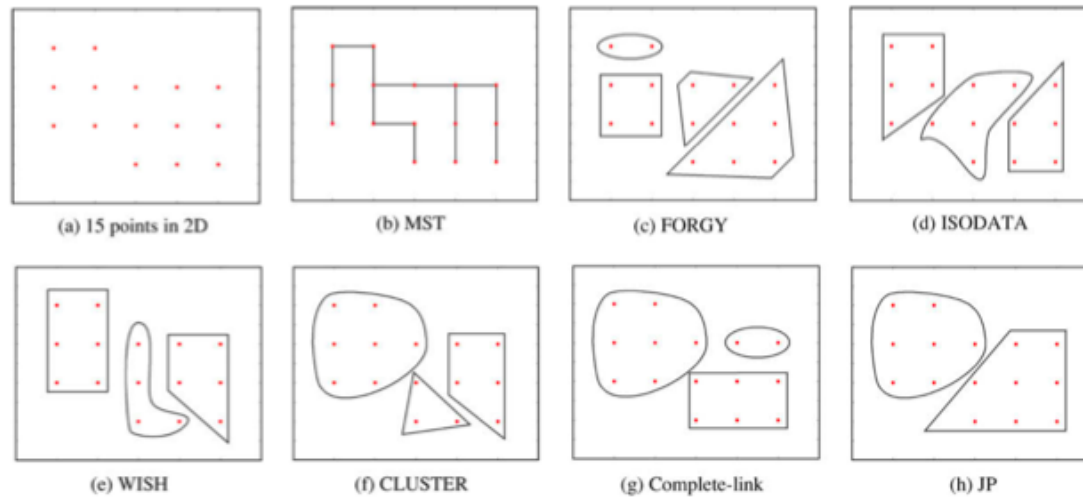
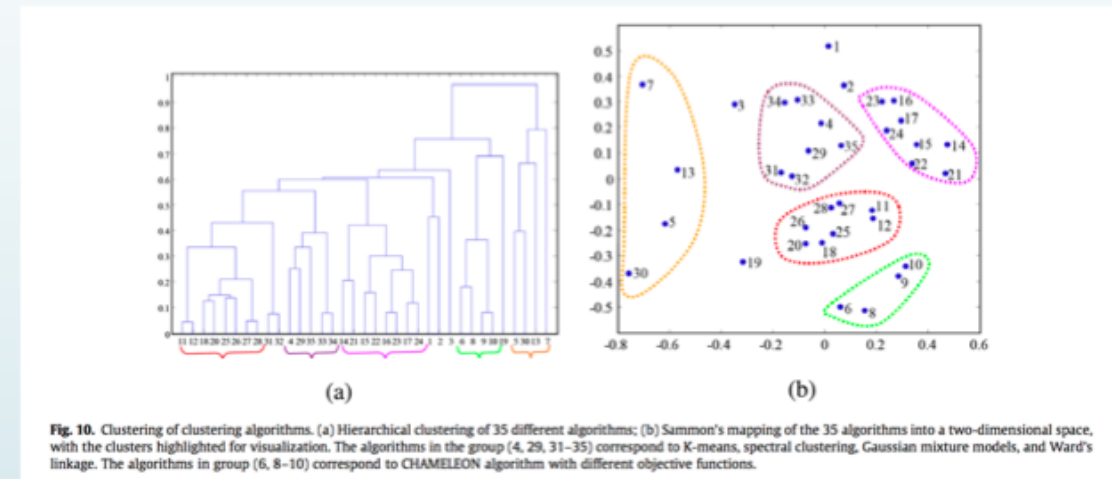


Fig. 9. Several clusterings of fifteen patterns in two dimensions: (a) fifteen patterns; (b) minimum spanning tree of the fifteen patterns; (c) clusters from FORGY; (d) clusters from ISODATA; (e) clusters from WISH; (f) clusters from CLUSTER; (g) clusters from complete-link hierarchical clustering; and (h) clusters from Jarvis-Patrick clustering algorithm. (Figure reproduced from [Dubes and Jain \(1976\)](#).)

3.5 Comparing Clustering Algorithms

- A researcher tried to group Clustering Algorithms together by how they partitioned data.
- This study clustered 35 clustering algorithms into 5 groups based on their behavior on 12 different datasets
- “There is no best clustering algorithm”



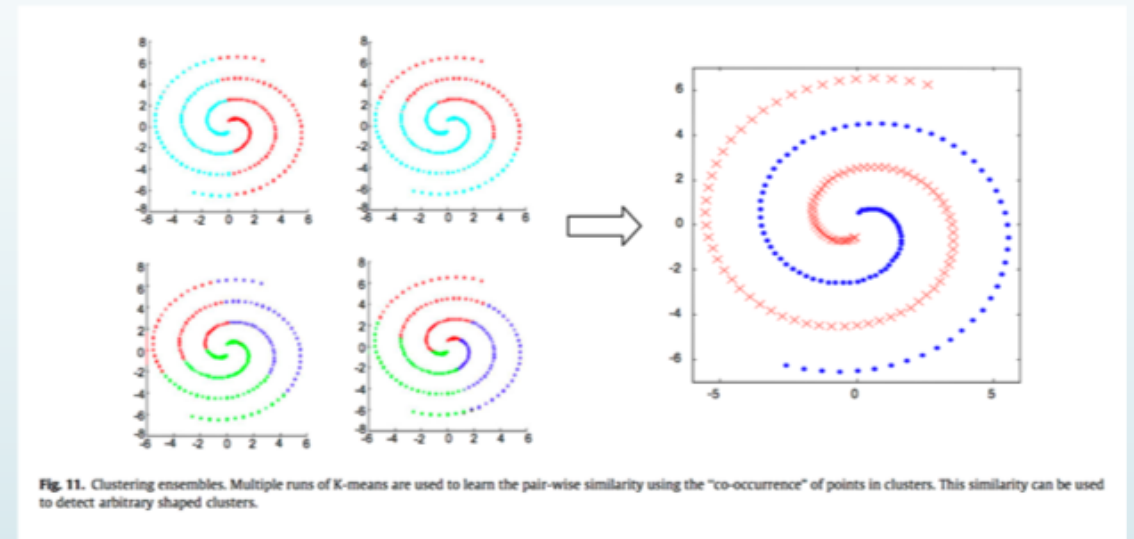


4. Trends in Data Clustering

- This section summarizes some of the recent trends in data clustering
 - Clustering Ensembles
 - Semi-Supervised Clustering
 - Large Scale Clustering
 - Multiway Clustering
 - Heterogeneous data


4.1 Clustering Ensembles

- The idea of clustering ensembles comes from the success of ensemble methods for supervised learning
 - Running clustering algorithm multiple times on data, can generate multiple partitions of the same data
 - The partitions are then combined with pair-wise similarity or other similarity measure
- Many methods exist for generating cluster ensembles:
 - Applying Different clustering algorithms
 - Applying the same clustering algorithm with different values of parameters or initializations
 - Combining of different data representations and clustering algorithms





4.2 Semi Supervised Clustering

- “Any external or side information available along with the $n \times d$ pattern matrix or the $n \times n$ similarity matrix can be extremely useful in finding a good partition of data. Clustering algorithms that utilize such side information are said to be operating in a semi-supervised mode (Chapelle et al., 2006).”
- 

4.3 Large-Scale Clustering

- “Large-scale data clustering addresses the challenge of clustering millions of data points that are represented in thousands of features. “
 - Example - Image data: can take 30h to answer one query when an image database size increases past 10 million
- There are a number of clustering algorithms to handle large-size datasets efficiently:
 - Efficient Nearest Neighbor (NN) Search:
 - Data Summarization:
 - Distributed Computing:
 - Incremental Clustering:
 - Sampling-based methods:

Table 1
Example applications of large-scale data clustering.

Application	Description	# Objects	# Features
Document clustering	Group documents of similar topics (Andrews et al., 2007)	10^6	10^4
Gene clustering	Group genes with similar expression levels (Lukashin et al., 2003)	10^5	10^2
Content-based image retrieval	Quantize low-level image features (Philbin et al., 2007)	10^9	10^2
Clustering of earth science data	Derive climate indices (Steinbach et al., 2003)	10^5	10^2



4.3 Large-Scale Clustering

- **Efficient Nearest Neighbor (NN) Search** – Algorithms for efficient NN search are either tree-based or random projection
- **Data Summarization** – Improves clustering by summarizing the dataset into a small subset and applying clustering algorithms to this subset
- **Distributed Computing** – Divides each step of data clustering into independent procedures that can be executed concurrently (at the same time)
- **Incremental Clustering** – Algorithms are designed to perform a single pass over data points and arrange data into classification tree incrementally
- **Sampling-based methods** – subsample a large dataset and perform clustering on the smaller dataset which is then transferred to the larger dataset (Cure Algorithm)



4.4 Multi-way clustering

- “Co-clustering (Hartigan, 1972; Mirkin, 1996) aims to cluster both features and instances of the data (or both rows and columns of the $n \times d$ pattern matrix) simultaneously to identify the subset of features where the resulting clusters are meaningful according to certain evaluation criterion. “
- Co-clustering is popular in the field of bioinformatics for gene clustering or document clustering
- Co-clustering was then extended to multi-way clustering in 2005 where the goal is to cluster objects simultaneously by clustering their heterogeneous components

4.5 Heterogeneous data

- " Heterogeneous data refers to the data where the objects may not be naturally represented using a fixed length feature vector. "
 - **Rank Data:** Clustering users whose rankings are similar for set of n items – based on ranking data
 - **Dynamic Data** – Changes over the course of time and arrives continuously, as the data gets modified the clustering needs to adjust
 - Ex: Blogs, webpages, network packets, stock market, retail chain, credit card transactions
 - This creates additional requirements as the clustering algorithm must be adaptive to changes in the data distributions, and must be able to rapidly process continuously arriving data
 - Many Clustering models use extensions of simple algorithms like K-means, K-medoid, fuzzy c-means, and density-based clustering to meet dynamic data's processing needs
 - **Graph Data:** Objects such as chemical compounds are best represented as graphs. Graph clustering features are extracted on patterns like frequent subgraphs, shortest paths, cycles, and tree-based patterns
 - **Relational Data:** The goal is to partition a large graph (network) into subgraphs based on their link structure and node attributes



5. Conclusion

- Main takeaways from this paper include:
 - Organizing data into groups comes naturally to many scientific fields
 - Clustering is popular and new algorithms continue to be developed
 - Clustering is a useful exploratory tool: its output only suggest hypothesis
 - There is no single clustering algorithm that dominates others
 - “Clustering is in the eye of the beholder” – Method for clustering greatly involves the user or applications needs