What Comes After K-Means?

A Survey of Clustering

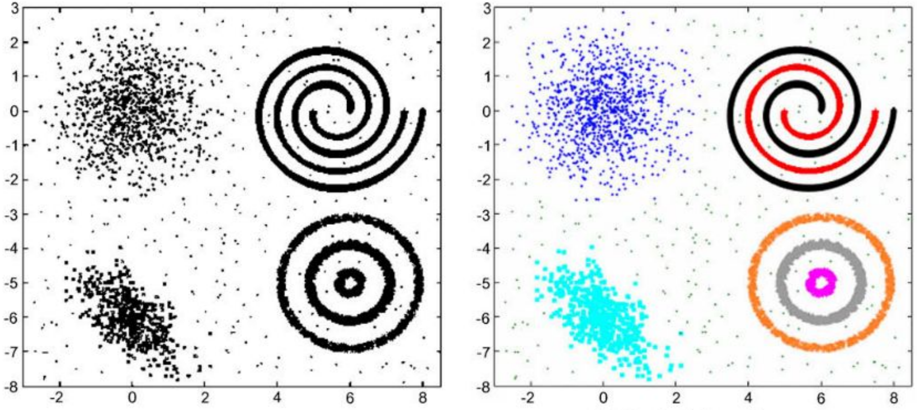
TEXAS STATE UNIVERSITY

The rising STAR of Texas

MEMBER **THE TEXAS STATE UNIVERSITY SYSTEM**

Data clustering: 50 years beyond K-means:

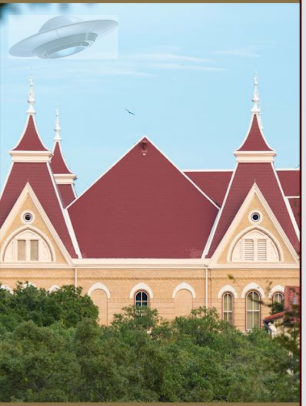
- ❖ Published by Anil K. Jain from Michigan State University in Pattern Recognition Letters 31, September 2009
- ❖ Emphasizes the importance of grouping data.
- ❖ Considers various algorithms on “tricky” clusters.
- ❖ Two types of algorithms:
  - – Partitional: builds clusters simultaneously
  - – Hierarchical: agglomerates or divides clusters.



(a) Input data (b) Desired clustering

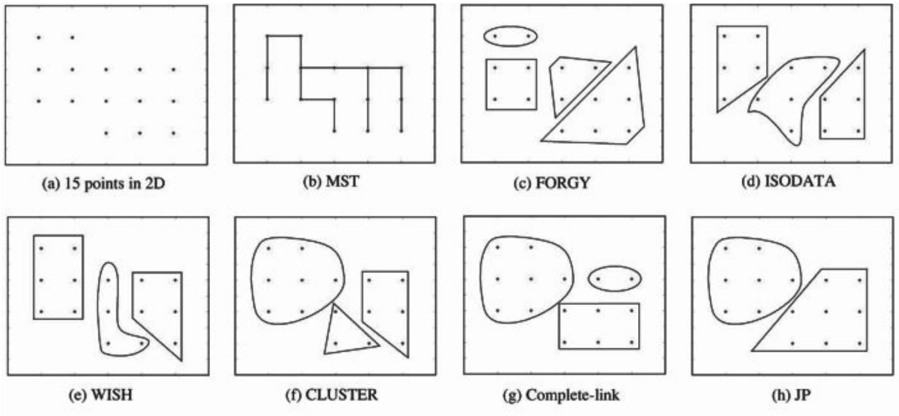
50 years beyond K-means, cont:

- K-Means, 1956: partitional clustering. Strongly dependent on heuristic or analytical measures to choose an effective K.
- Fuzzy C-Means, 1973: extension of K-Means which assign points to clusters probabilistically rather than fixedly.
- DBSCAN, 1996: cluster based on region density rather than closes to some centroid. Similar to Jarvis–Patrick algorithm.
- CLIQUE, 1998: density based clustering that estimates density on low dimension subspace of data.
- Bisecting K-Means, 2000: recursively divides cluster into 2 partitions at each step.
- X-Means, 2000: automatically fixes K.
- Dirichlet Process, 1973 and 2000: later becomes DP-Means.
- Ensemble Methods, 2000: applying several algorithms to one dataset and collecting the results into one conclusion.



So What's Missing?

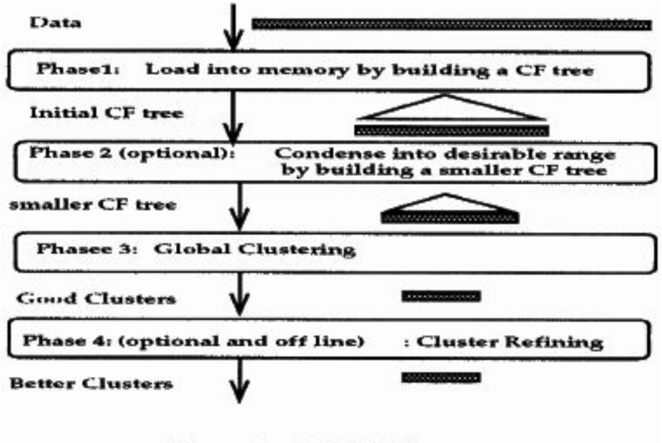
- ❖ Algorithms falter on sufficiently high dimension data.
- ❖ Datasets are getting much, much larger.
- ❖ The number of clusters is not always known or easily found.
- ❖ The "shape" of a cluster can still fool most algorithms.
- ❖ Fixed clustering is still the norm.
- ❖ "Noise" still commonly has to be cleaned out before clustering.
- ❖ Streaming rather than batch methods are uncommon but would be valuable for many real world applications.



TEXAS STATE UNIVERSITY
The rising STAR of Texas

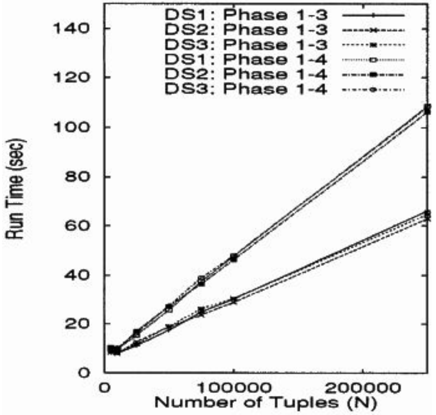
BIRCH: An Efficient Data Clustering Method for Very Large Databases:

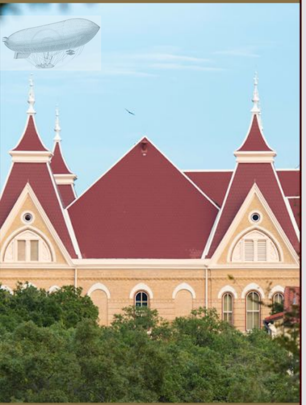
- ❖ Presented by Tian Zhang, Raghu Ramakrishnan, and Miron Livny at the ACM's SIGMOD in 1996.
- ❖ Specifically aimed to tackle big data within limited memory.
- ❖ Compares data points locally rather than globally.
- ❖ Can produce "good" results on a single scan of the data, but can optionally run for several iterations to improve the results.
- ❖ Captures the hierarchy of the dataset as a Clustering Feature Tree.



Pros and Cons of BIRCH:

- ❖ Built to work well in large data sets without a focus on high dimensional data.
- ❖ Can automatically compensate for noisy outliers without sacrificing speed.
- ❖ Strongly dependent on input parameters for speed and efficiency but not necessarily for producing good result.
- ❖ Has stayed relevant and in common usage despite being designed based on hardware which is no longer in usage.
- ❖ Can cluster regions of a data set based only on local data which makes the algorithm easily parallelizable.





The Global Kernel K-Means Clustering Algorithm:

- ❖ Presented to the IEEE Joint Conference on Neural Networks in 2008 by Grigorios Tzortzis and Aristidis Likas from the University of Ioannina.
- ❖ Kernel K-Means attempts to distinguish clusters which are not linearly separable. Global K-Means attempts to minimize the sensitivity of K-Means to its initialization parameters. This is both.
- ❖ Comes in “fast” and “near optimal” variety.
- ❖ Developed on MRI images.

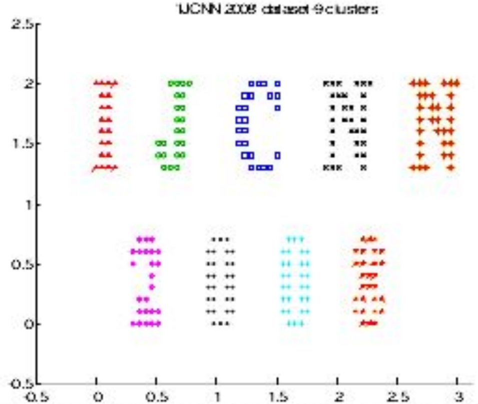


Fig. 2. Global kernel k-means, fast global kernel k-means and kernel k-

Pros of GK K-Means:

- ❖ "Near optimal" variety properly identifies clusters with "tricky" shapes.
- ❖ More deterministic performance than traditional K-Means.
- ❖ Produces solutions for cluster numbers of 1 up to K

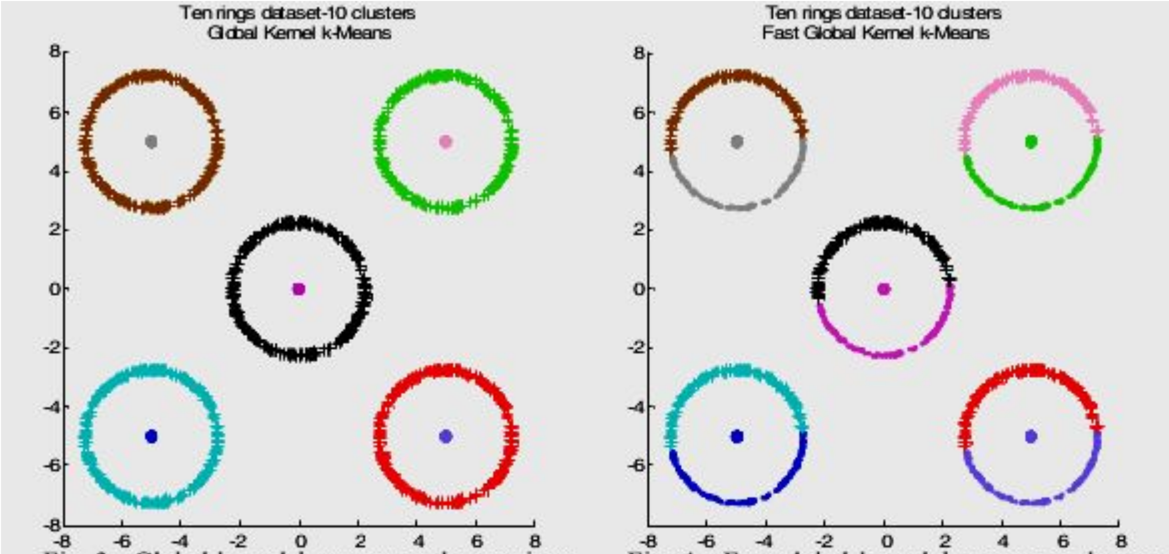


Fig. 3. Global kernel k-means on the ten rings dataset

Fig. 4. Fast global kernel k-means on the ten rings dataset

Cons of GK K-Means:

- Does not reduce or otherwise compensate for the dimensionality of the data.
- The testing data sets were extremely small by modern standards.
- The authors neglected to test their algorithm as a streaming clusterer which may have been more useful for the MRI application.
- The implementation was written in MATLAB giving it minimal portability.

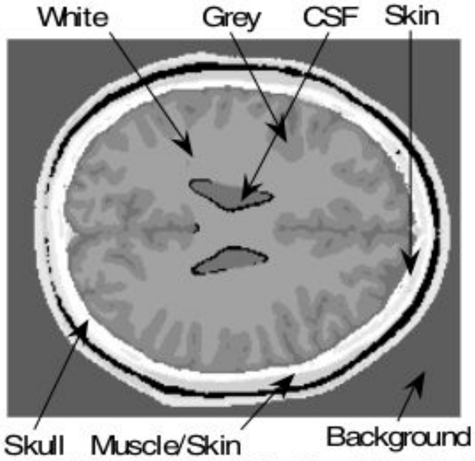


Fig. 11. Ground truth for slice 100. In black are the 3 tissues we ignore in our experiments.

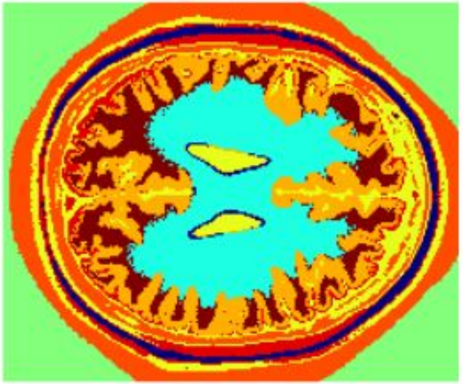
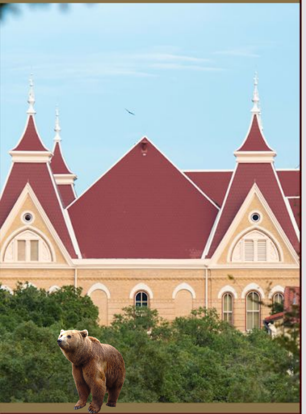
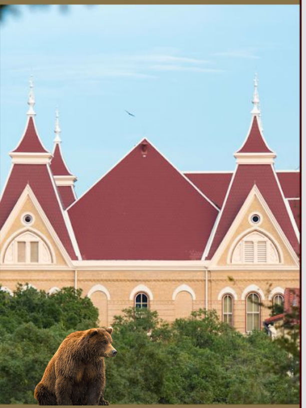


Fig. 14. Segmented tissues for slice 100 with fast global kernel k-means. In dark blue are the 3 tissues we ignore in our experiments.



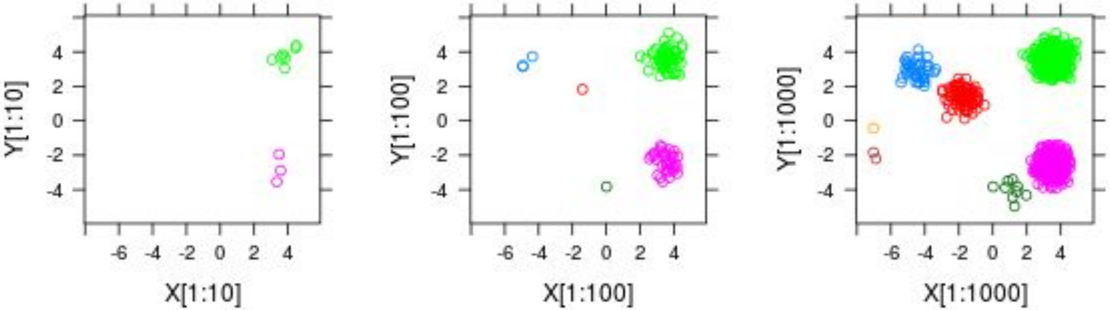
Dynamic Clustering via Asymptotics of the Dependent Dirichlet Process Mixture:

- Published in Advances in Neural Information Processing Systems 26 in 2013 by Trevor Campbell, Miao Liu, Brian Kulis, Jonathan P. How, and Lawrence Carin.
- Based on the Dirichlet process, which is roughly a random collection of random variables.
- "Non-parametric": the model grows as new data is added.
- Tested on a dataset of aircraft trajectories.
- Clusters can be created, eliminated, and altered as data is added.
- Promises high speed for time-sensitive applications.



Pros and Cons of Dynamic Means:

- ❖ Scalable number of clusters.
- ❖ Functions as a batch or streaming algorithm.
- ❖ Very fast execution.
- ❖ Guarantees cluster validity similar to K-Means, which is part good and part bad.
- ❖ Demonstrated fast and effective clustering on 400-dimensional data.
- ❖ Focuses exclusively on fixed clustering.
- ❖ Despite the high dimensionality of the data set, the total size of the data was still quite small.



The Dirichlet Mixture Model from Victor Veitch at
https://en.wikipedia.org/wiki/Dirichlet_process#/media/File:DP_clustering_simulation.png

A Density Based Algorithm for Discovering Clusters:

- ❖ Presented to the Association for the Advancement of Artificial Intelligence in 1996 by Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu from the University of Munich.
- ❖ Only takes on input parameter.
- ❖ Clusters by density of data points rather than by placing centroids.
- ❖ Can define clusters of arbitrary shapes.
- ❖ This is a partitioning algorithm, which was unusual by this point in time.

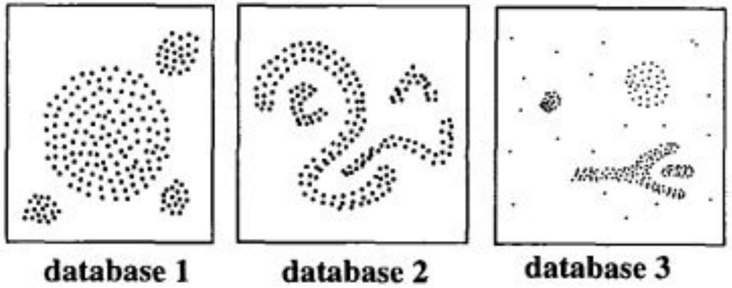


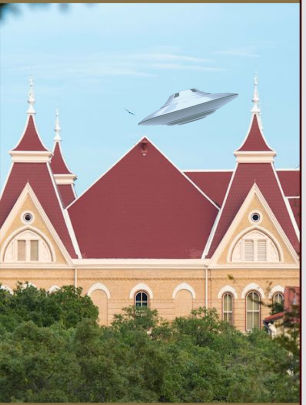
figure 1: Sample databases

TEXAS STATE UNIVERSITY

The rising STAR of Texas

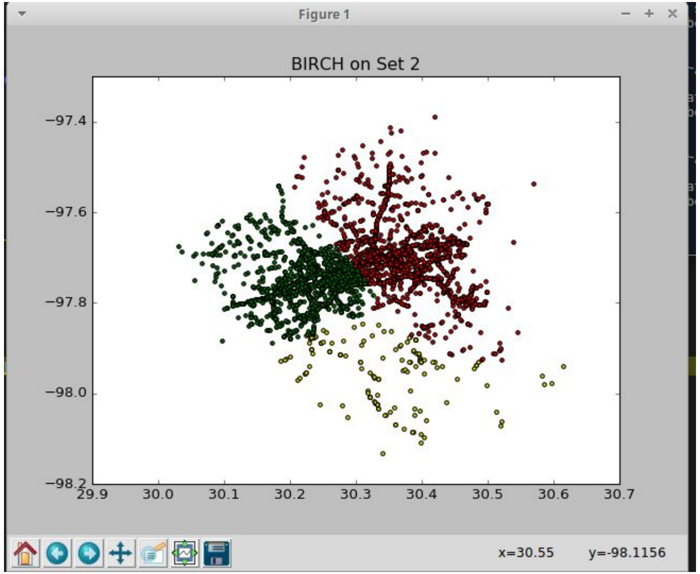
Pros and Cons of DBSCAN:

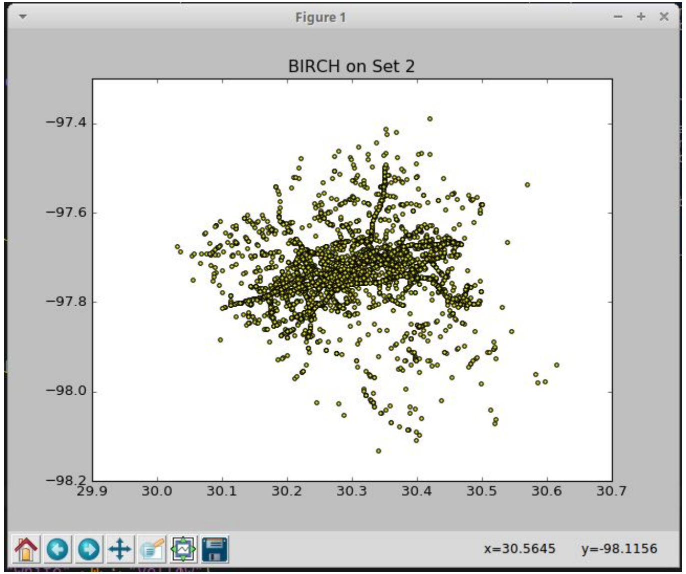
- ❖ Near linear scaling of time requirement as the dataset grows, which is very close to BIRCH's performance.
- ❖ Can find clusters of absolutely any shape, rather than find clusters fitting "several" models.
- ❖ Compensates for noisy data, but always includes outliers in some cluster.
- ❖ Commonly available in Python or other languages of choice.
- ❖ Does not balance clusters by their number of points. Rather the algorithm focuses only on the density of the points.
- ❖ "Neighborhood" consideration becomes more difficult in high dimension data.
- ❖ Not entirely deterministic.



Birch on Austin Traffic Data:

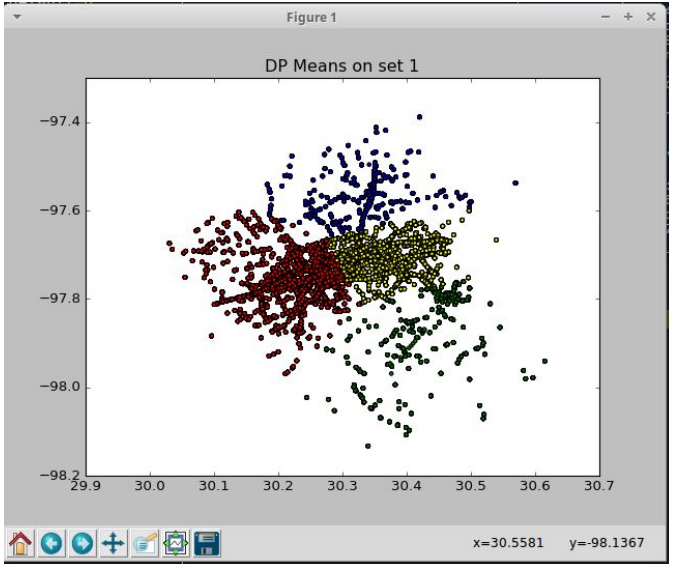
- ❖ Data from https://data.austintexas.gov/Transportation-and-Mobility/Real-Time-Traffic-Incident-Reports
- ❖ Altering the ‘threshold’ parameter can yield different results on the same data.

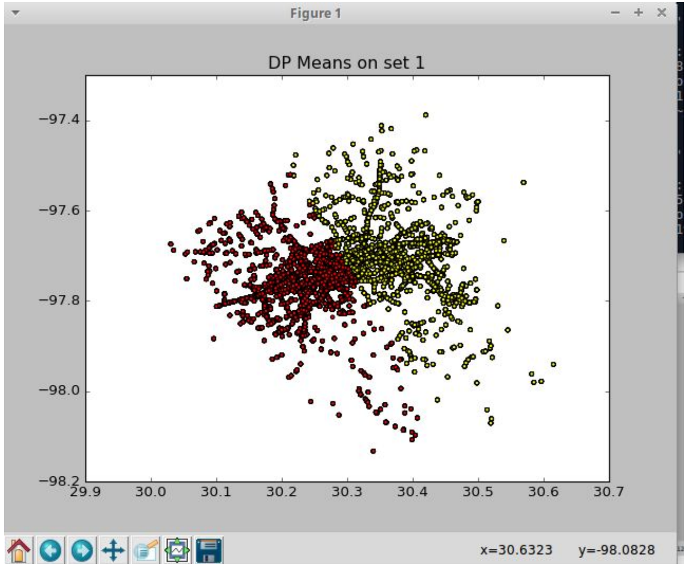




DP Means on Austin Traffic Data:

- Data from https://data.austintexas.gov/Transportation-and-Mobility/Real-Time-Traffic-Incident-Reports
- Slight alterations to the ‘lambda’ parameter can produce very different results.





TEXAS STATE UNIVERSITY
The rising STAR of Texas

Future Steps:

- ❖ Modern algorithms must work on billion scale data sets.
- ❖ They must also work high dimensional data which should include deep descriptors. We have found work-arounds for the “Curse of Dimensionality” but not a solution.
- ❖ Measures of distance and neighborhoods begin to break down at high dimensions. Perhaps clustering is possible without using one of these two techniques.
- ❖ Defining a cluster with a centroid makes an assumption about the underlying structure of the data.
- ❖ Parallelism is a necessity on modern data.
- ❖ Fuzzy clustering should always be a possibility.
- ❖ A good algorithm should rely as little on foreknowledge and input parameters as possible.

Questions or Comments?