50 years beyond K-mea

Authored By: Anil K. Jain

What's After

eans

BIRCH

Presented by: Zhang, Ramakrish

K-Means?

Dynamic

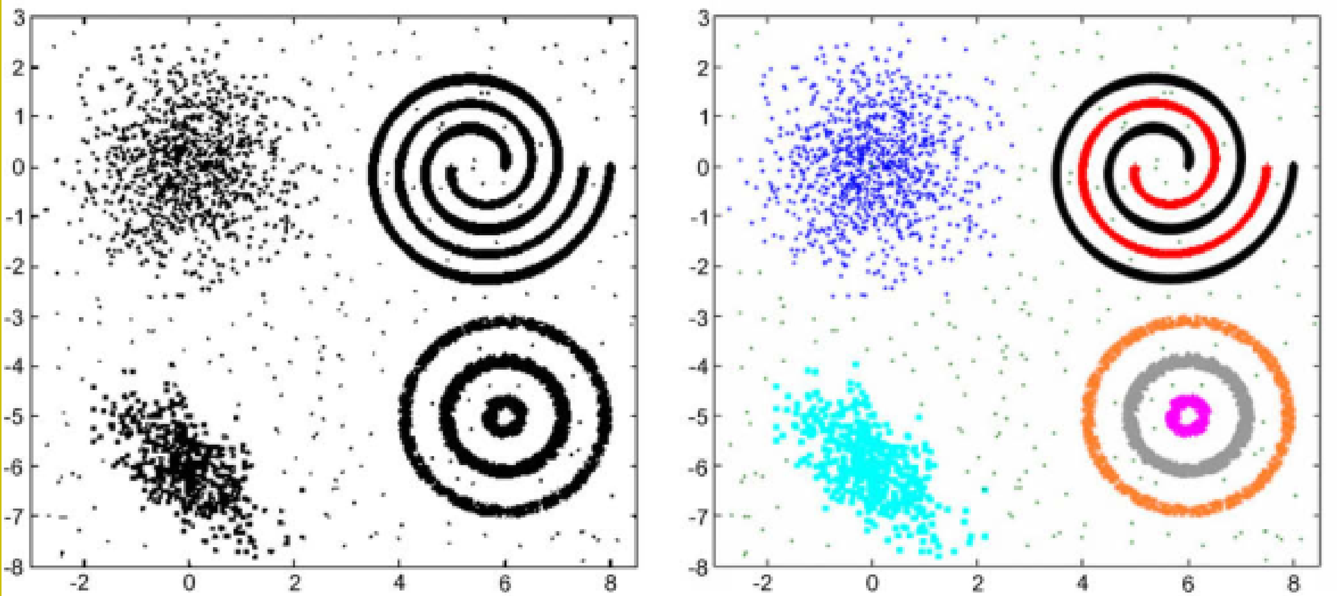
shnan, and Livny

Presented by: Camp

Clustering

mpbell, Liu, Kulis, How, and Carin

- Authored By: Anil K. Jain
- Published in Pattern Recognition Letters 31, 2009
- Emphasizes the importance of grouping data.
- Two types of algorithms:
  - Partitional: builds clusters simultaneously
  - Hierarchical: joins or divides clusters.



(a) Input data (b) Desired clustering

- Presented by: Zhang, Ramakrish
- SIGMOD 1996
- Aimed to handle big data withi
- Compares data points locally ra
- Can produce good results on a
- Captures hierarchy of the data

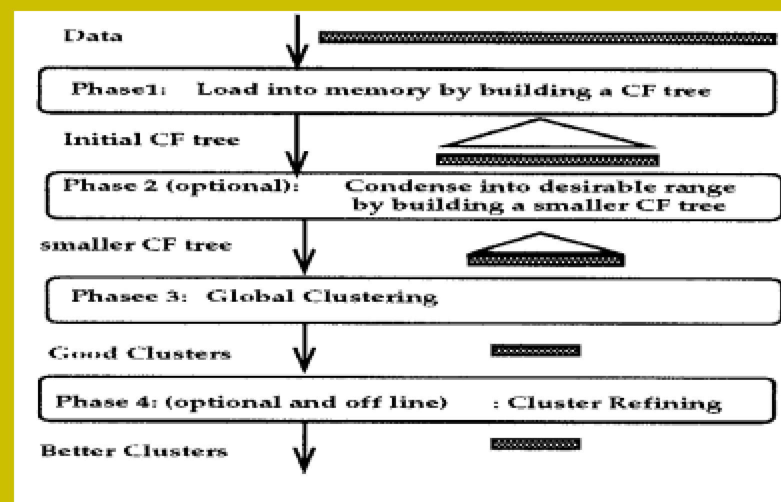


Figure 1: *BIRCH Overview*

Fi

shnan, and Livny

hin limited memory.

rather than globally.

a single scan of data.

a as a CF Tree.

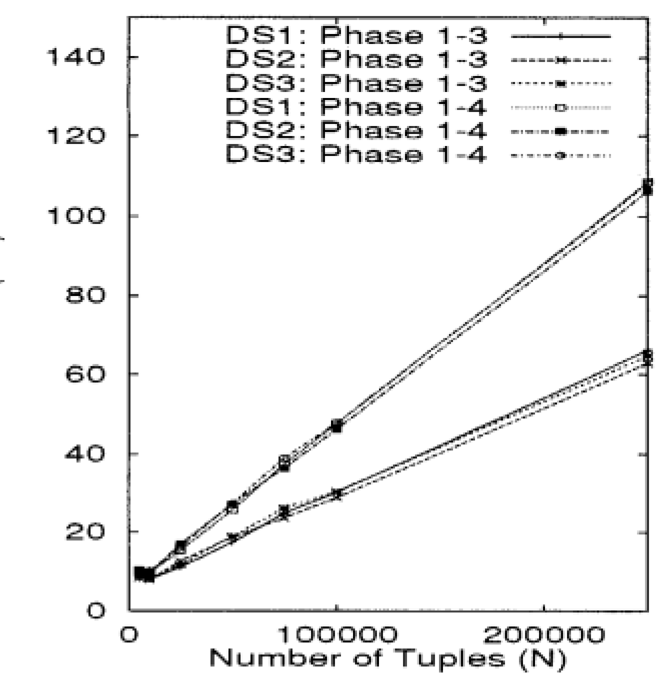
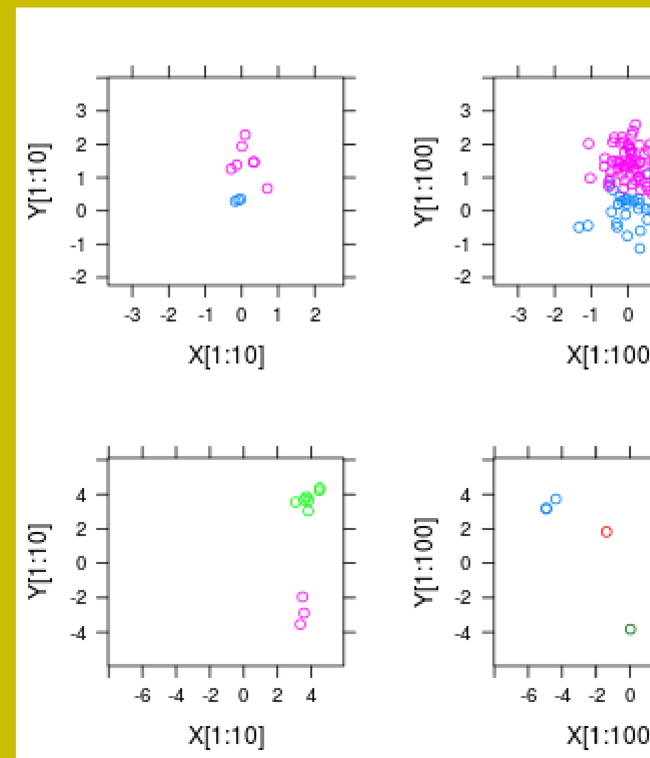


Figure 4: *Scalability wrt. Increasing* n_l, n_h

- Presented by: Campb
- ANIPS, 2013
- Based on Dirichlet P
- Non-parametric: mod
- Clusters can be crea
- Promises high speed



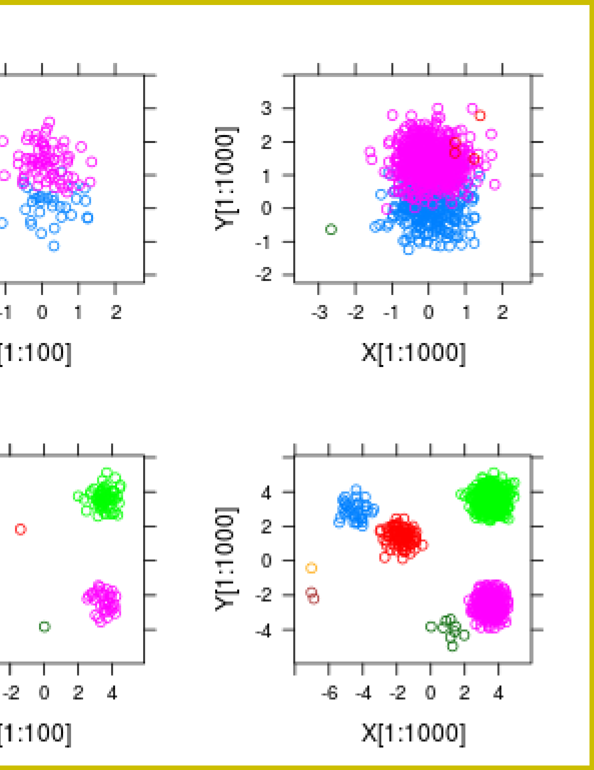
https://en.wikipedia.org/wiki/Dirichlet_process#/m mpbell, Liu, Kulis, How, and Carin.

t Processes

nodel grows with data.

reated, eliminated, and altered.

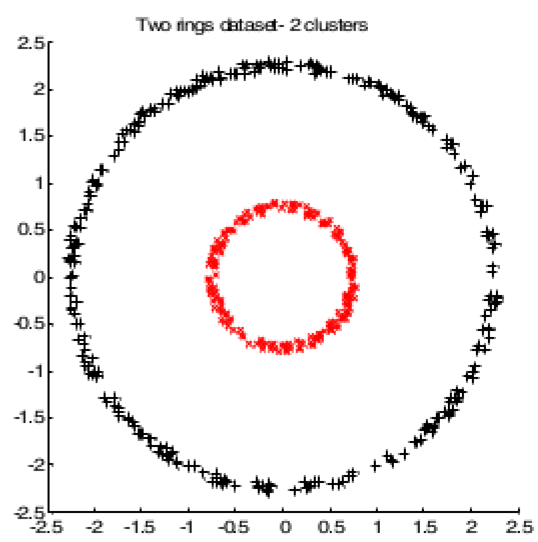
eed for time-sensitive applications.



ss#/media/File:DP_clustering_simulation.png

Global Kernel K-Means

- Presented by: Tzortzis and Likas
- IEEE Joint Conference on Neural Networks in 2008
- Kernel K-Means finds non linear separable clusters.
- Global K-Means minimizes sensitivity to initialization.
- Comes in “fast” and “near optimal” variety.



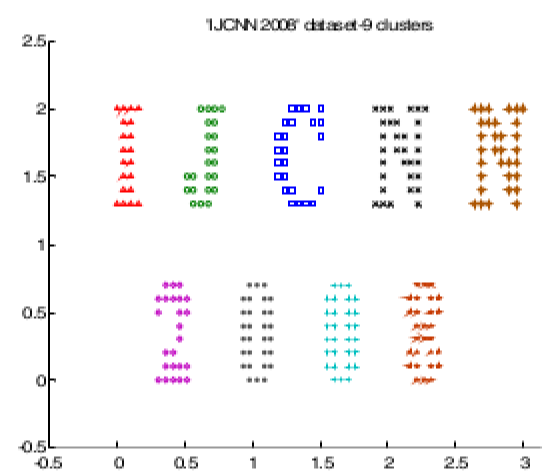


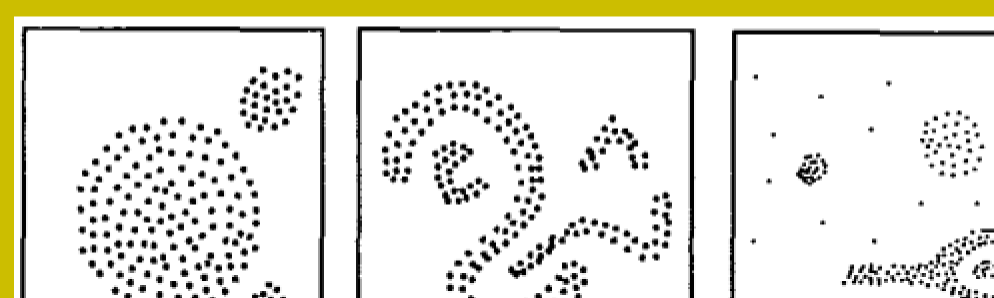
Fig. 2. Global kernel k-means, fast global kernel k-means

s

neans and kernel *k*-

A Density Bas

- Presented by: Ester, Kriegel, Sa
- Association for the Advancement
- Only takes one input parameter.
- Clusters by density of points rat
- Can define clusters of arbitrary
- This is a partitioning algorithm.



https://en.wikipedia.org/wiki/Dirichlet_process#/m

ased Algorithm

Sander and Xu.

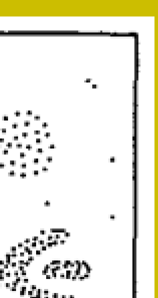
nt of AI in 1996

er.

rather than centroids.

y shapes.

m.



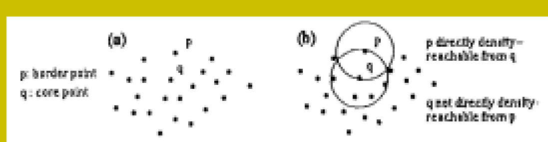
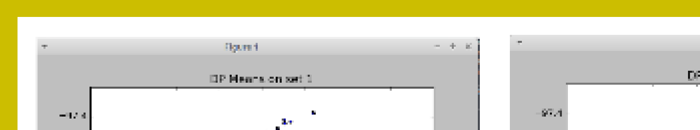


figure 2: core points and border points

So What

- Focus on Billion-Scale
- Solve the "Curse of D
- Distance measures bre
- Centroids make assum
- Parallelism is a must.
- Fuzzy clustering shoul
- **Good algorithms do no**



ss#/media/File:DP_clustering_simulation.png

Comes Next?

ale data.

Dimensionality” rather than avoiding.

break down at high dimension

umptions about structure of data.

st.

ould be a possibility.

not make assumptions about data.

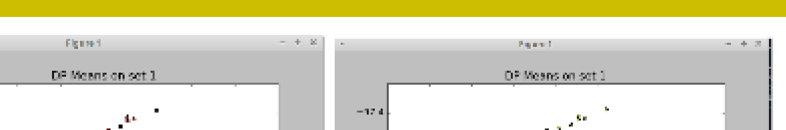


Fig. 1. Global kernel k-means, fast global kernel k-means and kernel k-means (run with minimum clustering error) on the two rings dataset.

Fig. 2. Global kernel k-means, fast global kernel k-means
means (run with minimum clustering error) on the 'IJCNN 200

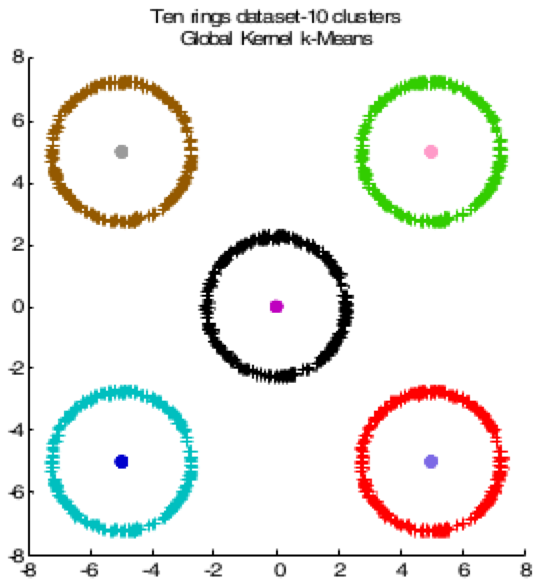


Fig. 3. Global kernel k-means on the ten rings dataset.

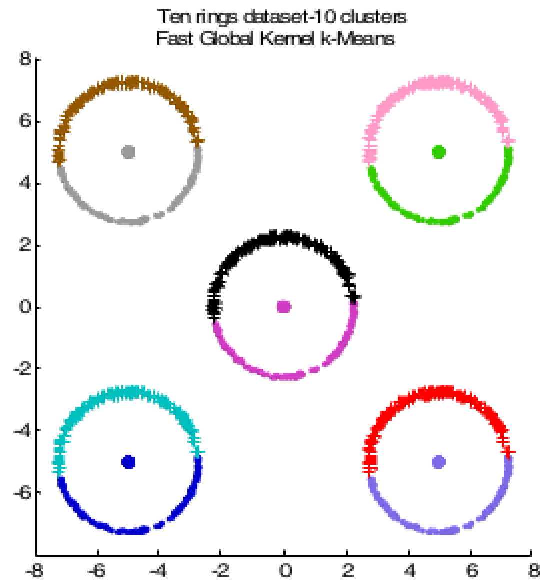


Fig. 4. Fast global kernel k-means on the ten rings dataset.

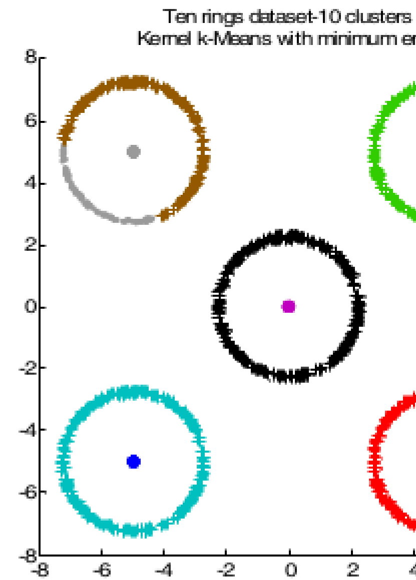
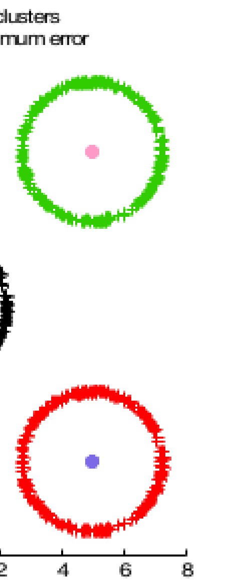


Fig. 5. Kernel k-means (the ru
mum clustering error) on the ten ring

eans and kernel k-
N 2008' logo.



he run with mini-
n rings dataset.

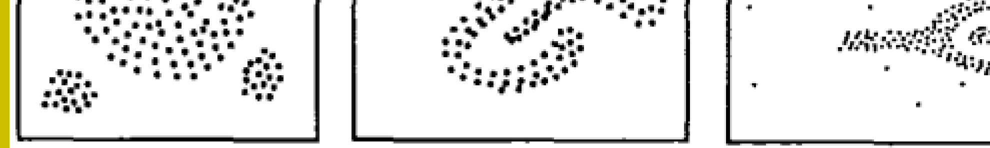
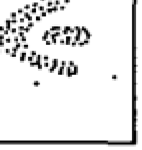


figure 1: Sample databases

Presenter: Gentry Atkinson

Faculty Mentors: Dr. Tesic and Dr. Tami



e 3

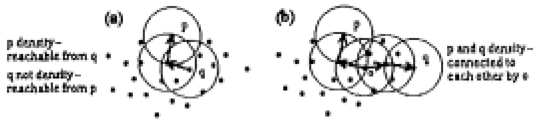
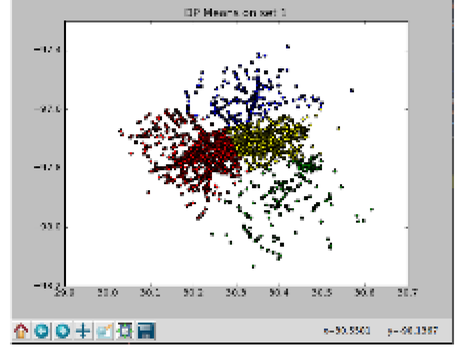
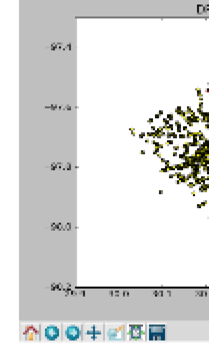


figure 3: density-reachability and density-connectivity



Lambda = 0.3
4 clusters

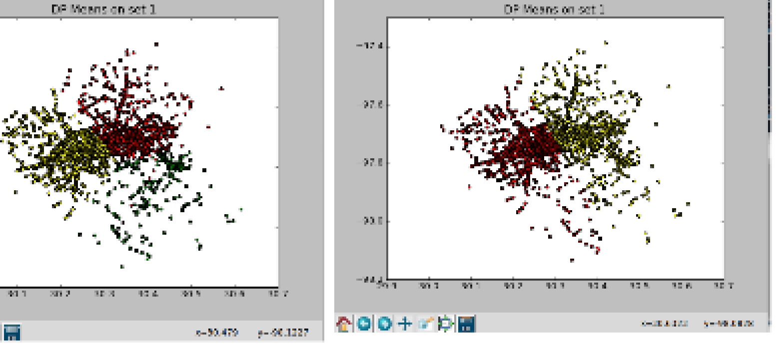


Lambda = 0.4
3 clusters

A demonstration showing a slight sh

finding 2, 3, or 4 clusters on Austin

mir



= 0.4
s

Lambda = 0.5
2 clusters

shift in the DP Means input parameter
stin traffic accidents.