Clustering Big Data

Anil K. Jain
(with Radha Chitta and Rong Jin)

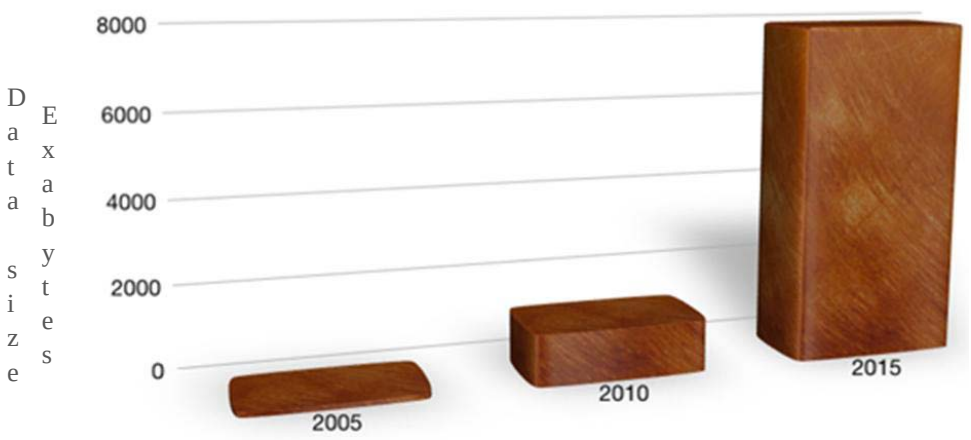
Department of Computer Science
Michigan State University
November 29, 2012

Outline

- Big Data
- How to extract “information”?
- Data clustering
- Clustering Big Data
- Kernel K-means & approximation
- Summary

How Big is Big Data?

- Big is a fast moving target: kilobytes, megabytes, gigabytes, terabytes (10^{12}), petabytes (10^{15}), exabytes (10^{18}), zettabytes (10^{21}),.....
- Over 1.8 zb created in 2011; ~8 zb by 2015



Source: IDC's Digital Universe study, sponsored by EMC, June 2011
http://idcdocserv.com/1142
http://www.emc.com/leadership/programs/digital-universe.htm

As of June 2012

Nature of Big Data: Volume, Velocity and Variety

Big Data on the Web

~900 million users, 2.5 billion content items, 105 terabytes of data each half hour, 300M photos and 4M videos posted per day

Over 225 million users generating over 800 tweets per second

http://techcrunch.com/2012/08/22/how-big-is-facebooks-data-2-5-billion-pieces-of-content-and-500-terabytes-ingested-every-day/
http://royal.pingdom.com/2012/01/17/internet-2011-in-numbers/
http://www.dataversity.net/the-growth-of-unstructured-data-what-are-we-going-to-do-with-all-those-zettabytes/

Big Data on the Web

Over 50 billion pages indexed and more than 2 million queries/min

Google News

Articles from over 10,000 sources in real time

flickr from YAHOO!

~4.5 million photos uploaded/day

YouTube

48 hours of video uploaded/min; more than 1 trillion video views

No. of mobile phones will exceed the world's population by the end of 2012

What to do with Big Data?

- Extract information to make decisions
- Evidence-based decision: data-driven vs. analysis based on intuition & experience
- Analytics, business intelligence, data mining, machine learning, pattern recognition
- Big Data computing: IBM is promoting Watson (Jeopardy champion) to tackle Big Data in healthcare, finance, drug design,..

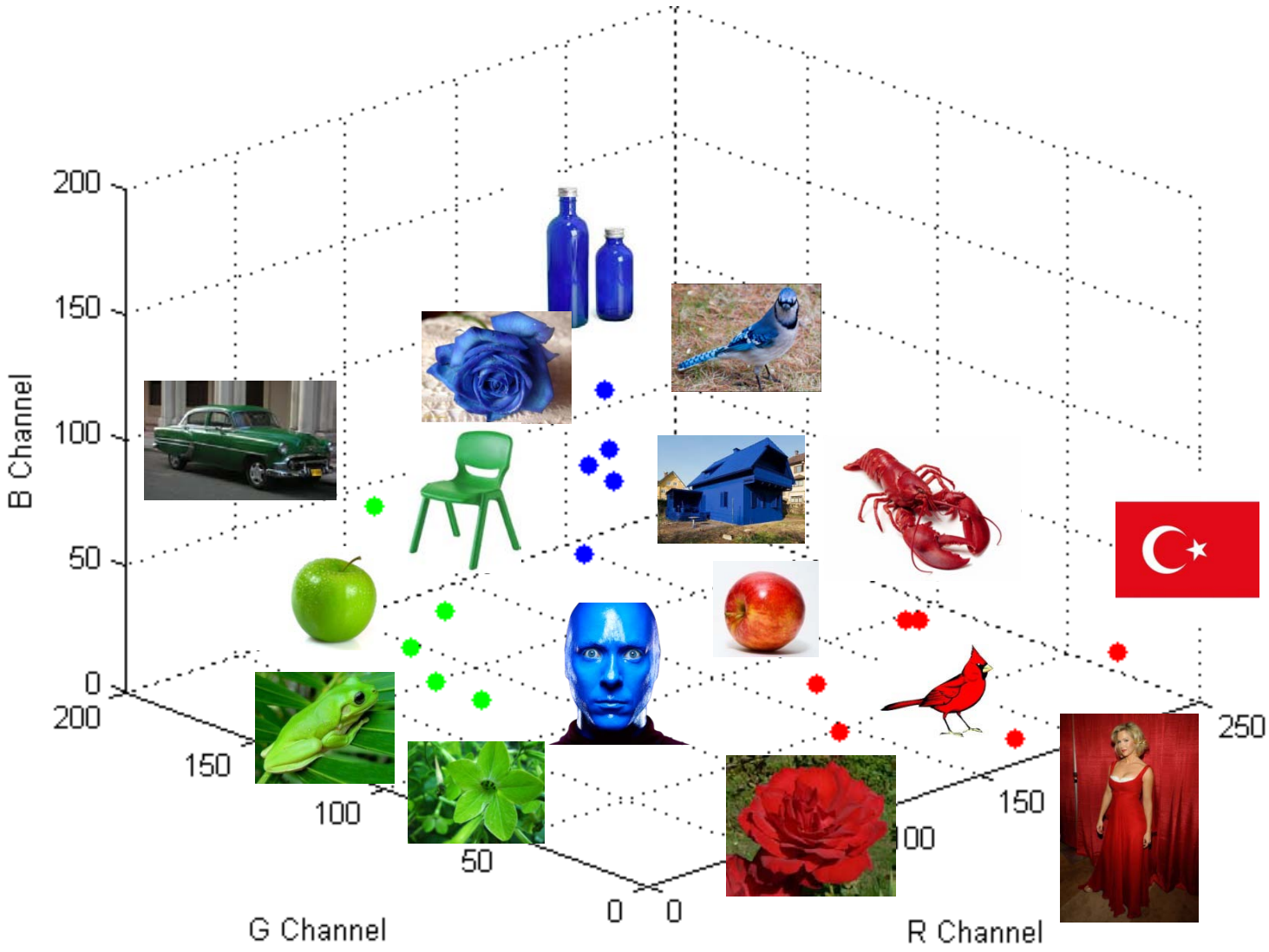
Steve Lohr, “Amid the Flood, A Catchphrase is Born”, NY Times, August 12, 2012

Decision Making

- Data Representation
  - Features and similarity
- Learning
  - Classification (labeled data)
  - Clustering (unlabeled data)

Most big data problems have unlabeled objects

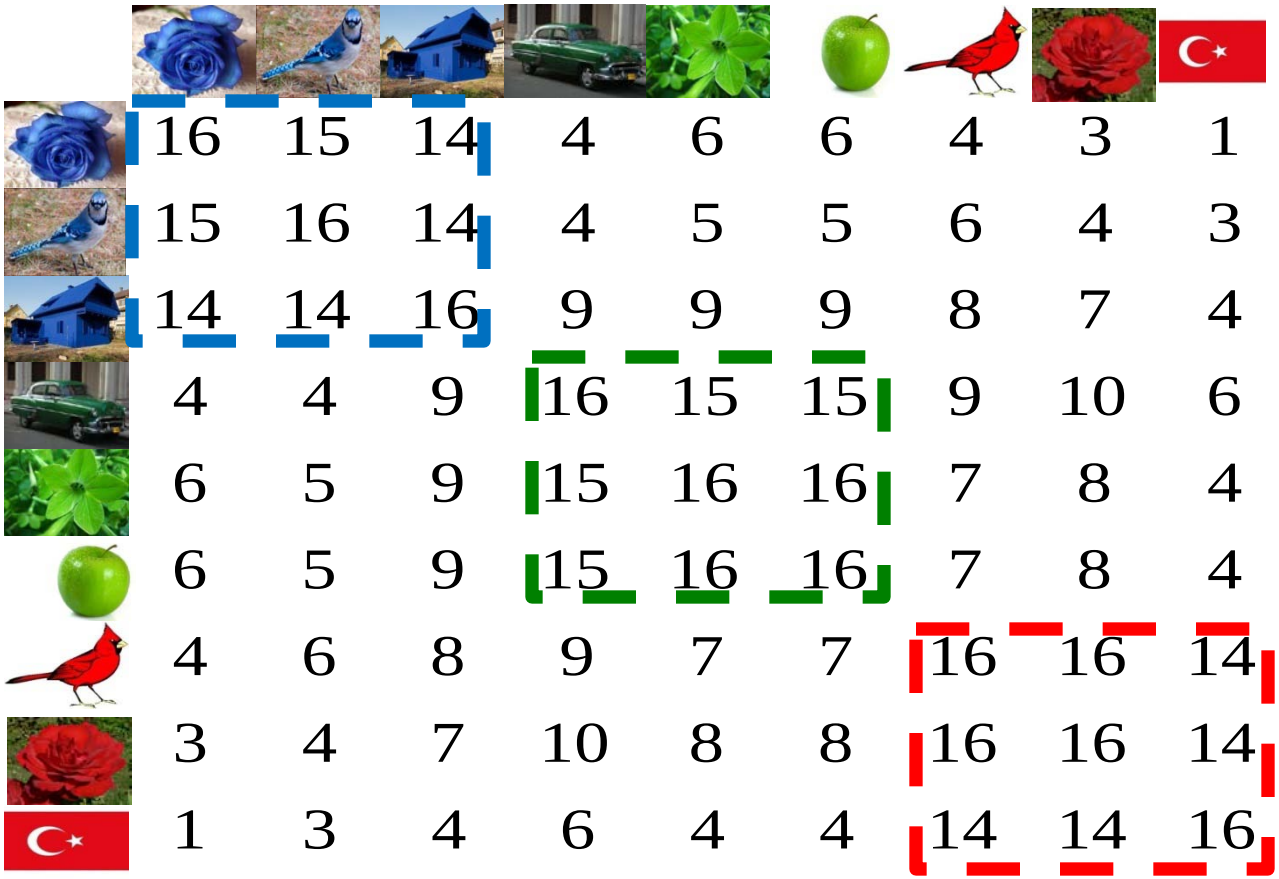
Pattern Matrix



n x d pattern matrix

Similarity Matrix

Polynomial kernel: $K(\mathbf{x},\mathbf{y}) = \left(\mathbf{x}^T\mathbf{y} + 1\right)^4$



16	15	14	4	6	6	4	3	1
15	16	14	4	5	5	6	4	3
14	14	16	9	9	9	8	7	4
4	4	9	16	15	15	9	10	6
6	5	9	15	16	16	7	8	4
6	5	9	15	16	16	7	8	4
4	6	8	9	7	7	16	16	14
3	4	7	10	8	8	16	16	14
1	3	4	6	4	4	14	14	16

n x n similarity matrix

Classification

Dogs

Cats

Given a training set of labeled objects, learn a decision rule

Clustering

Given a collection of (unlabeled) objects, find meaningful groups

Semi-supervised Clustering

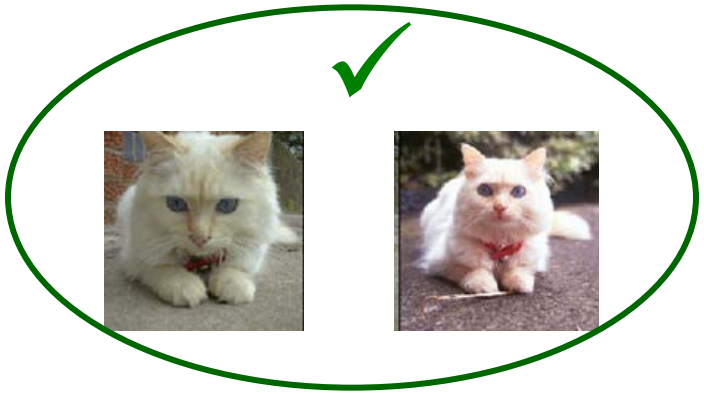
Supervised

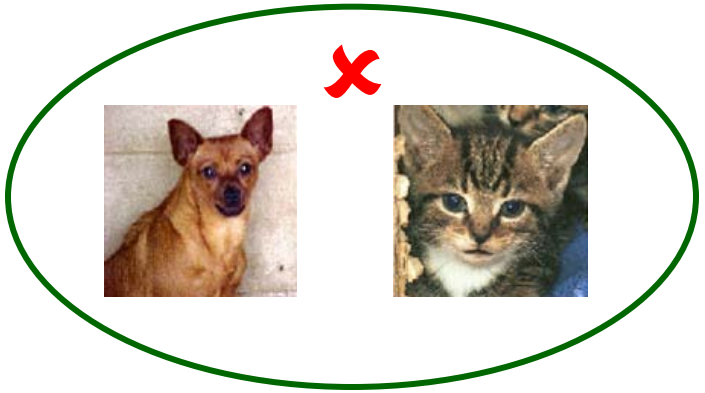
Dogs

Cats

Unsupervised

Semi-supervised





Pairwise constraints improve the clustering performance

What is a cluster?

“A group of the same or similar elements gathered or occurring closely together”

Galaxy clusters

Birdhouse clusters

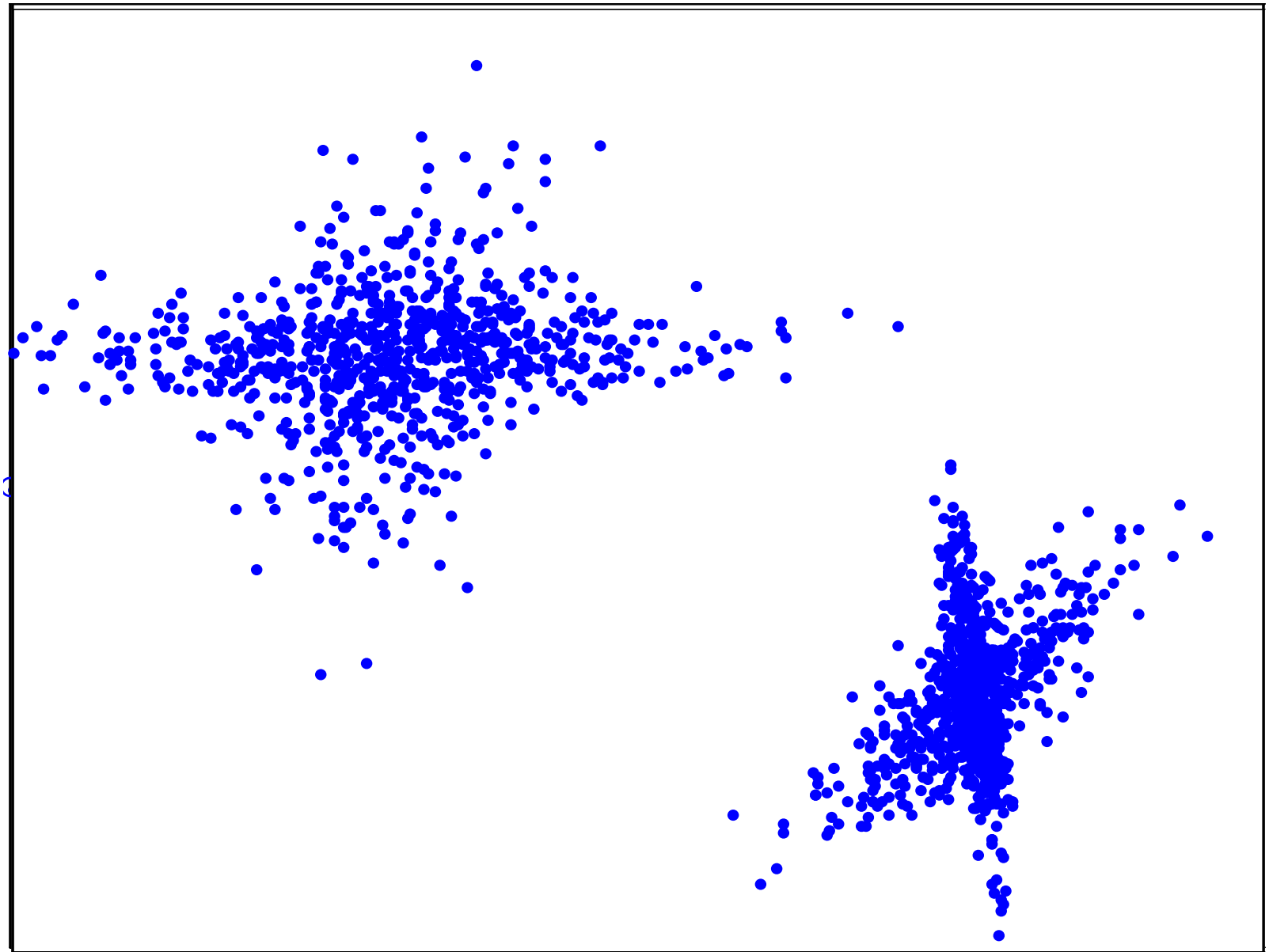
Cluster munition

Cluster computing

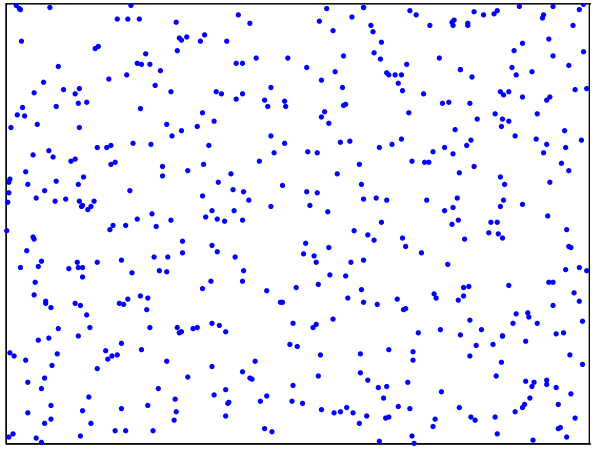
Cluster lights

Hongkeng Tulou cluster

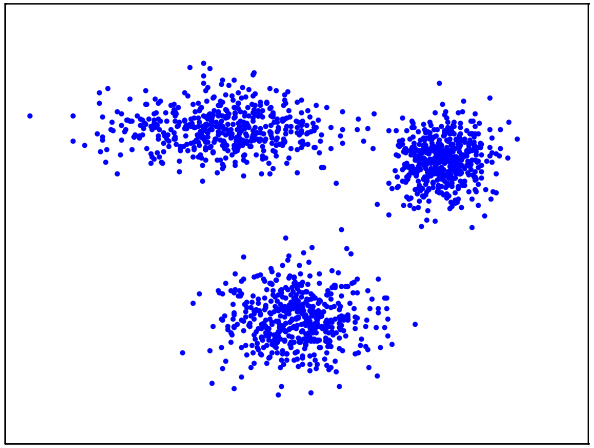
Clusters in 2D

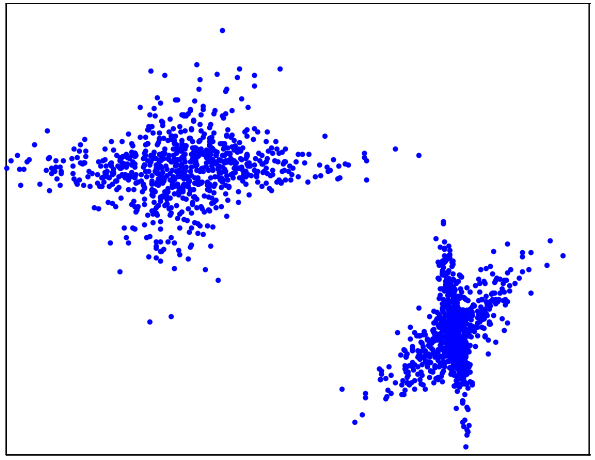


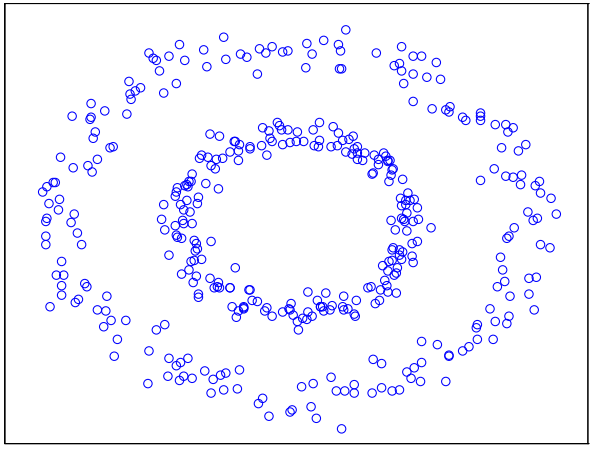
Challenges in Data Clustering

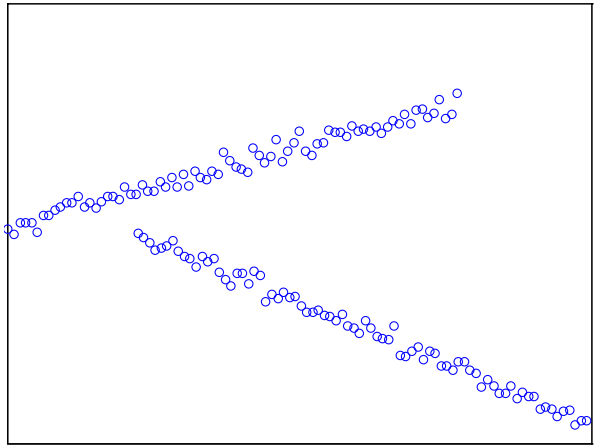


- Measure of similarity
- No. of clusters
- Cluster validity
- Outliers



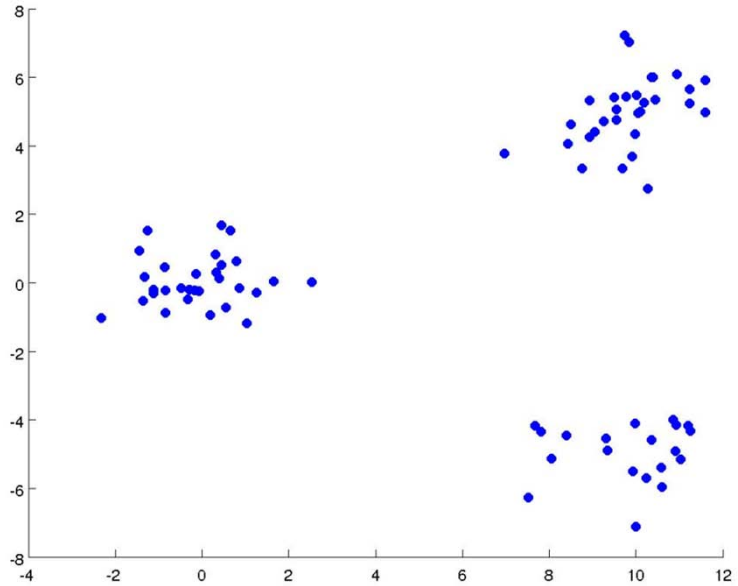


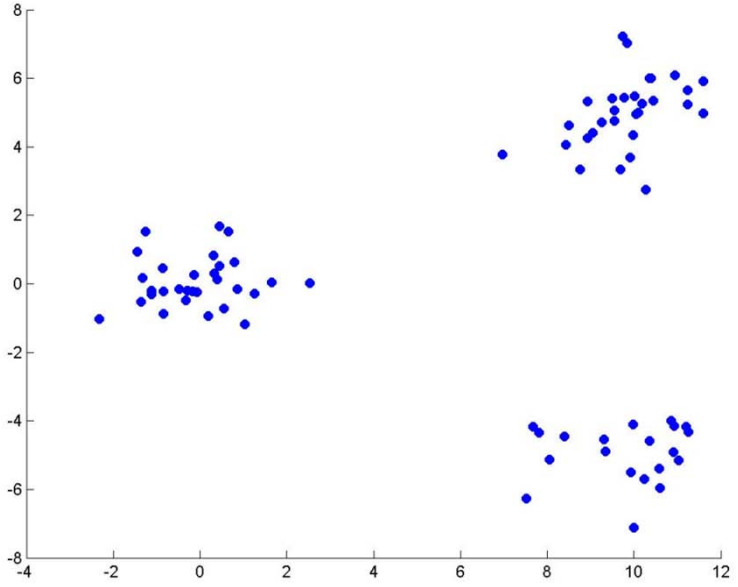




Data Clustering

Organize a collection of n objects into a partition or a hierarchy (nested set of partitions)





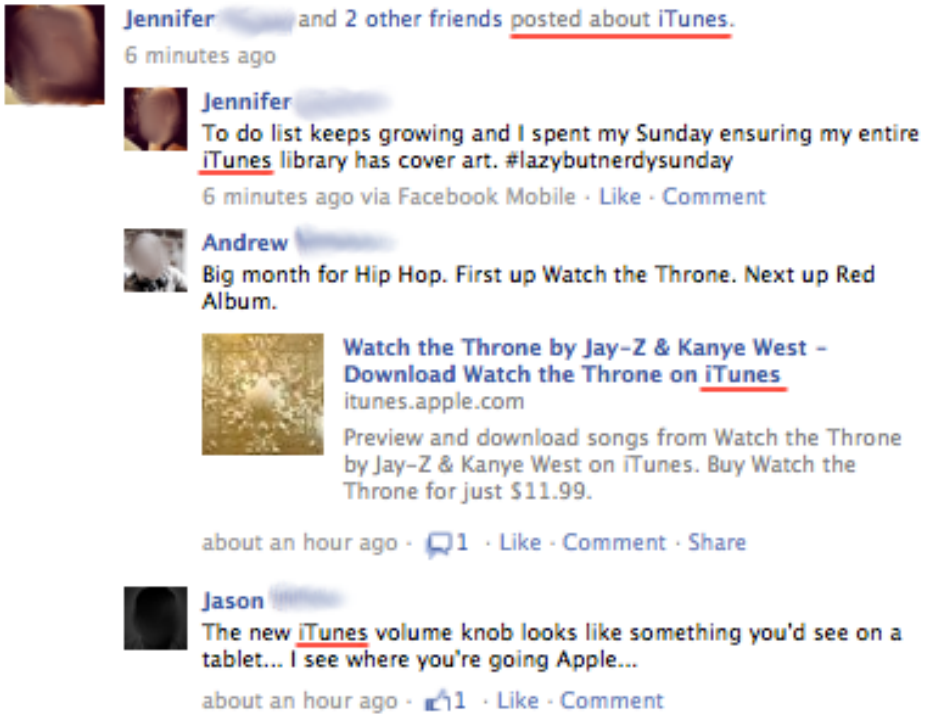
“Data clustering” returned ~6,100 hits for 2011 (Google Scholar)

Clustering is the Key to Big Data Problem

- Not feasible to “label” large collection of objects
- No prior knowledge of the number and nature of groups (clusters) in data
- Clusters may evolve over time
- Clustering provides efficient browsing, search, recommendation and organization of data

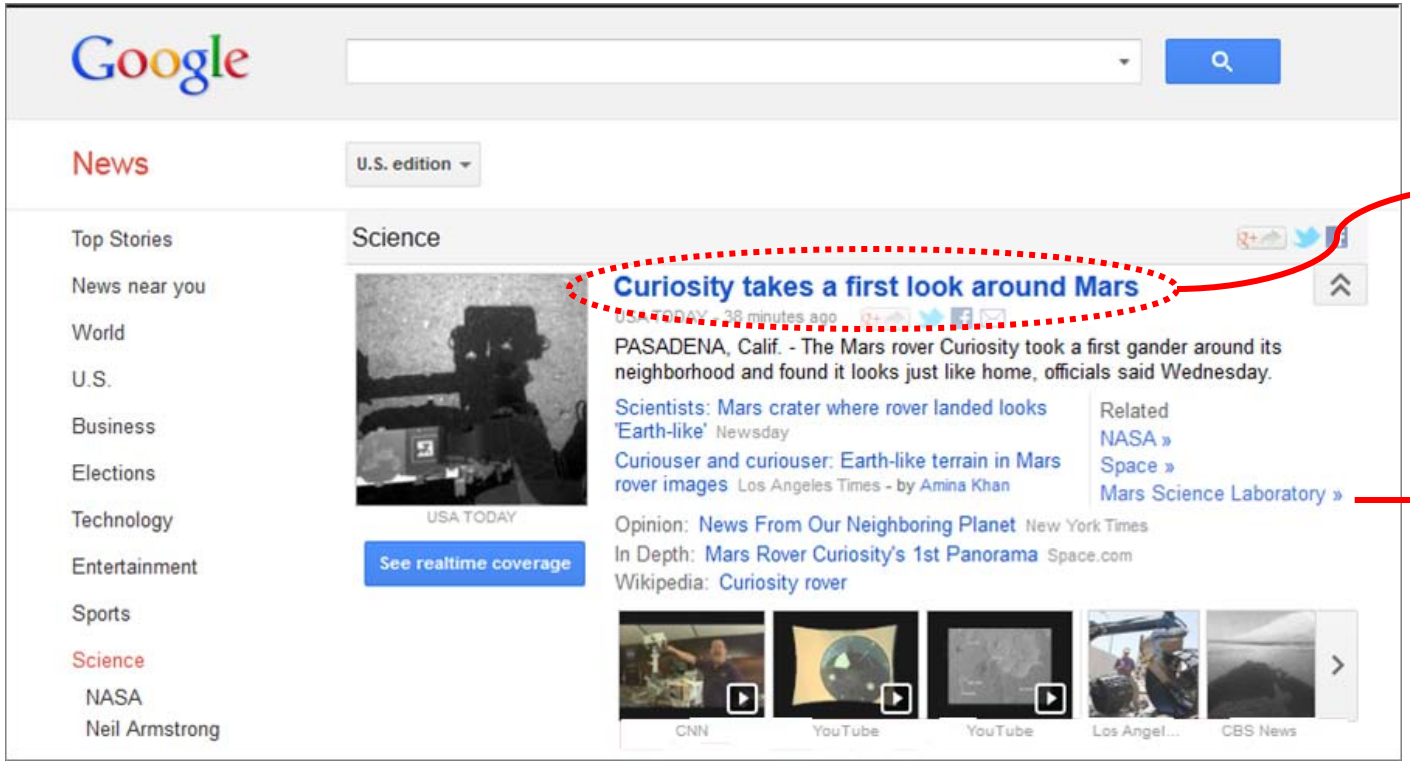
Clustering Users on Facebook

- ~300,000 status updates per minute on tens of thousands of topics
- Cluster users based on topic of status messages



http://www.insidefacebook.com/2011/08/08/posted-about-page/
http://searchengineland.com/by-the-numbers-twitter-vs-facebook-vs-google-buzz-36709

Clustering Articles on Google News

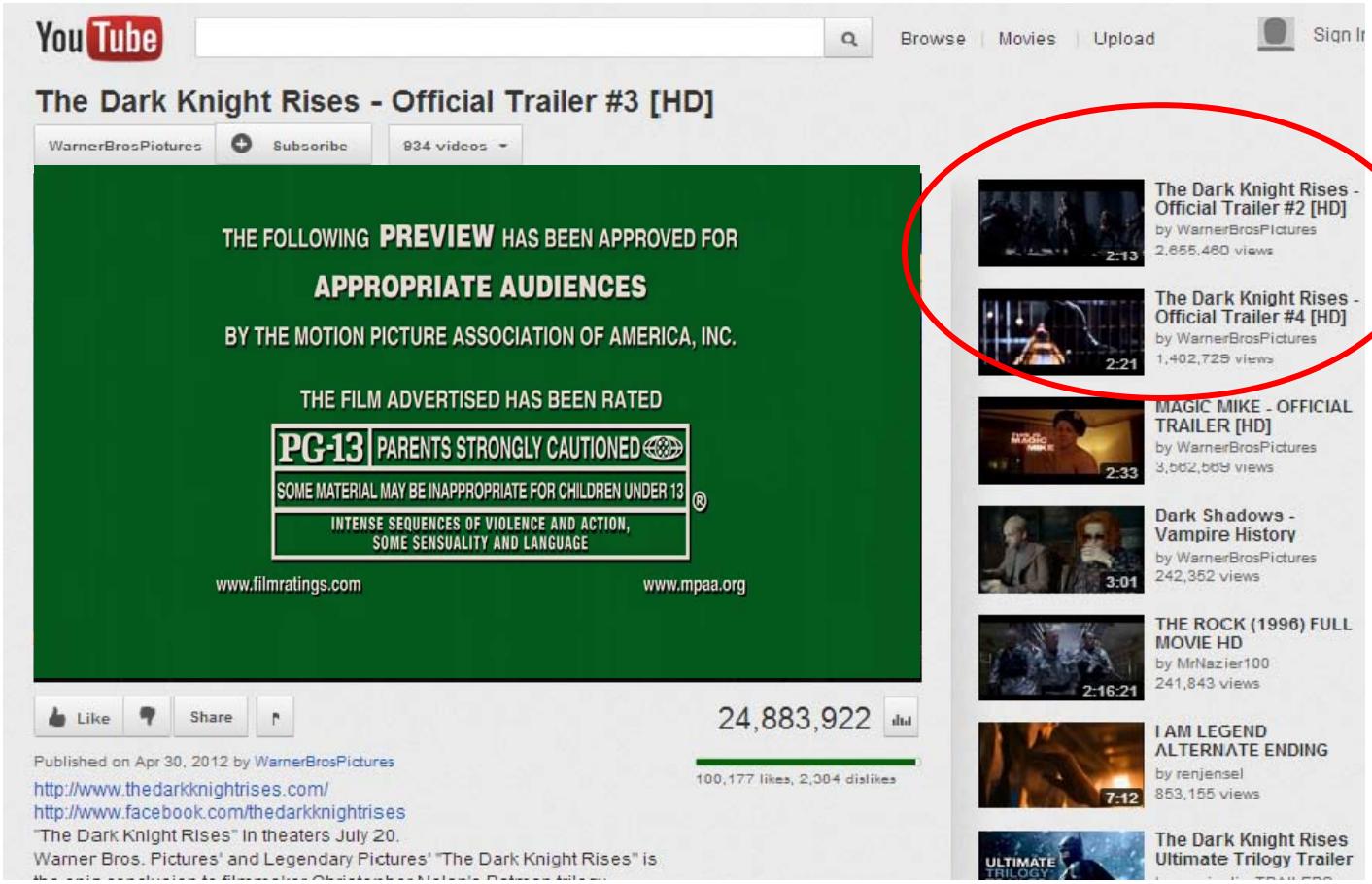


Topic cluster

Article Listings

http://blogoscoped.com/archive/2006-07-28-n49.html

Clustering Videos on Youtube



- Keywords
- Popularity
- Viewer engagement
- User browsing history

http://www.strutta.com/blog/blog/six-degrees-of-youtube

Clustering for Efficient Image retrieval

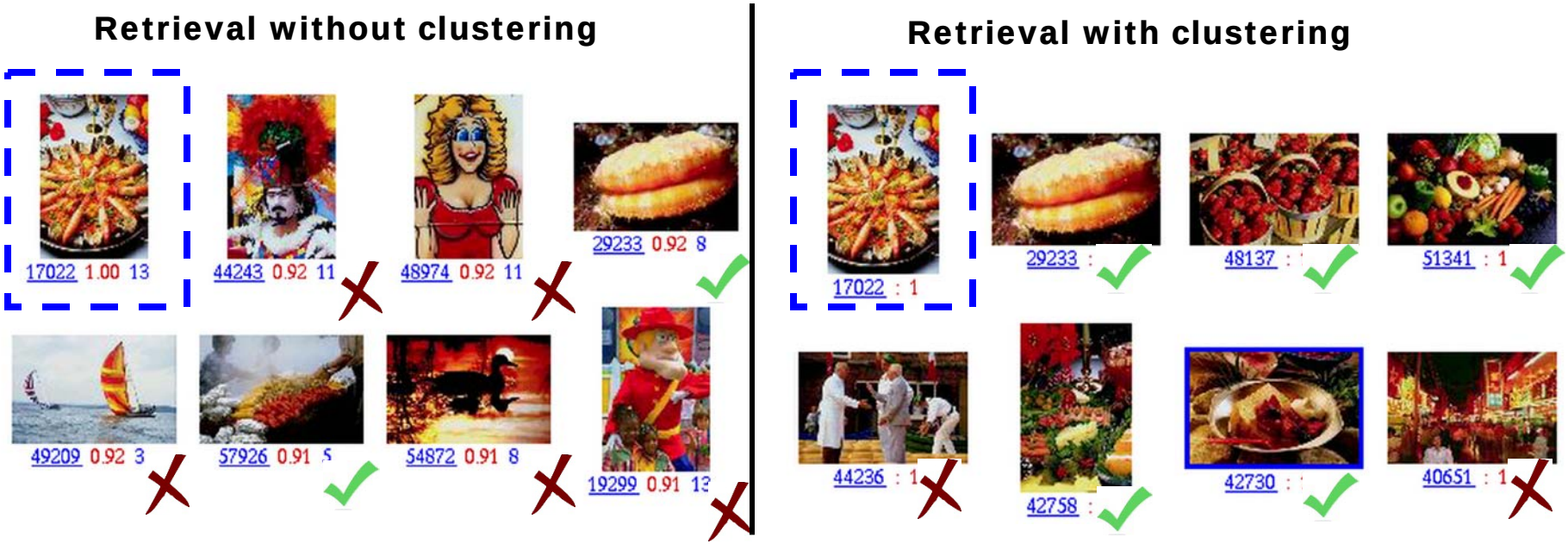


Fig. 1. Upper-left image is the query. Numbers under the images on left side: image ID and cluster ID; on the right side: Image ID, matching score, number of regions.

Retrieval accuracy for the “food” category (average precision):

Without clustering: **47%**

With clustering: **61%**

Chen et al., “CLUE: cluster-based retrieval of images by unsupervised learning,” IEEE Tans. On Image Processing, 2005.

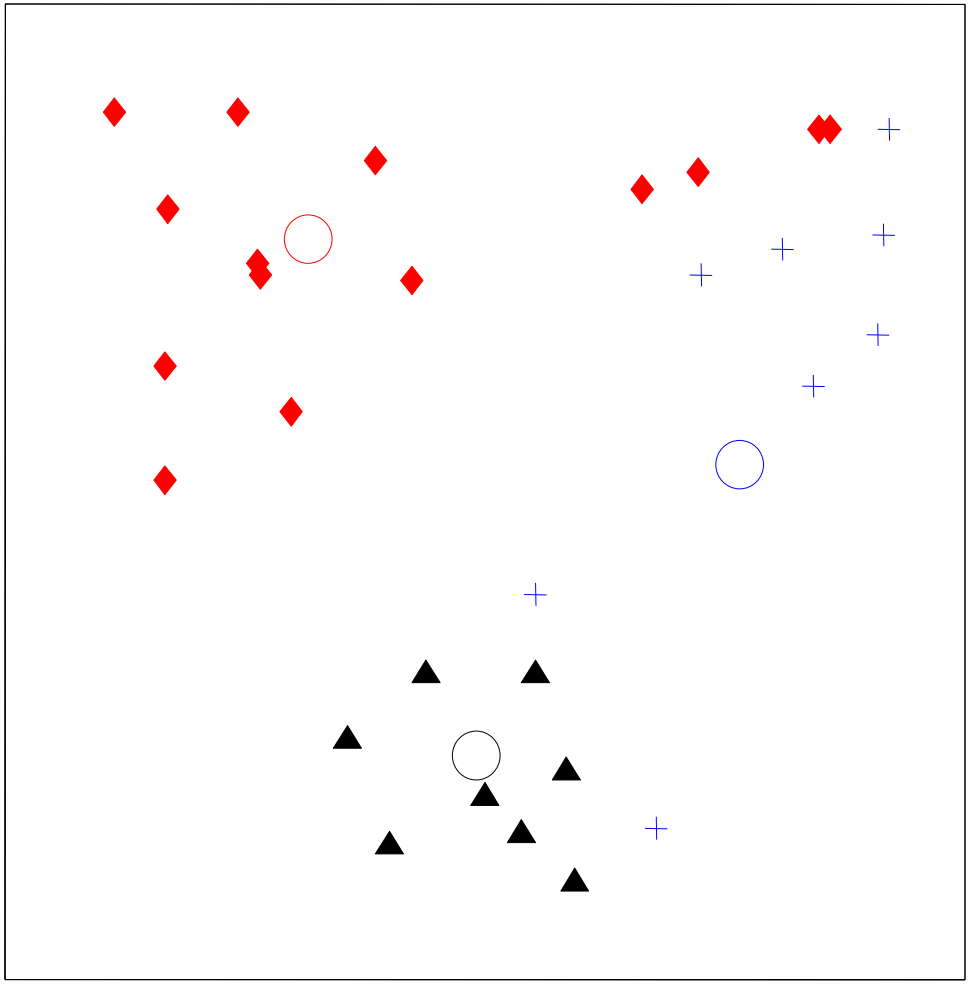
Clustering Algorithms

Hundreds of clustering algorithms are available; many are “admissible”, but no algorithm is “optimal”

- K-means
- Gaussian mixture models
- Kernel K-means
- Spectral Clustering
- Nearest neighbor
- Latent Dirichlet Allocation

A.K. Jain, “Data Clustering: 50 Years Beyond K-Means”, PRL, 2011

K-means Algorithm

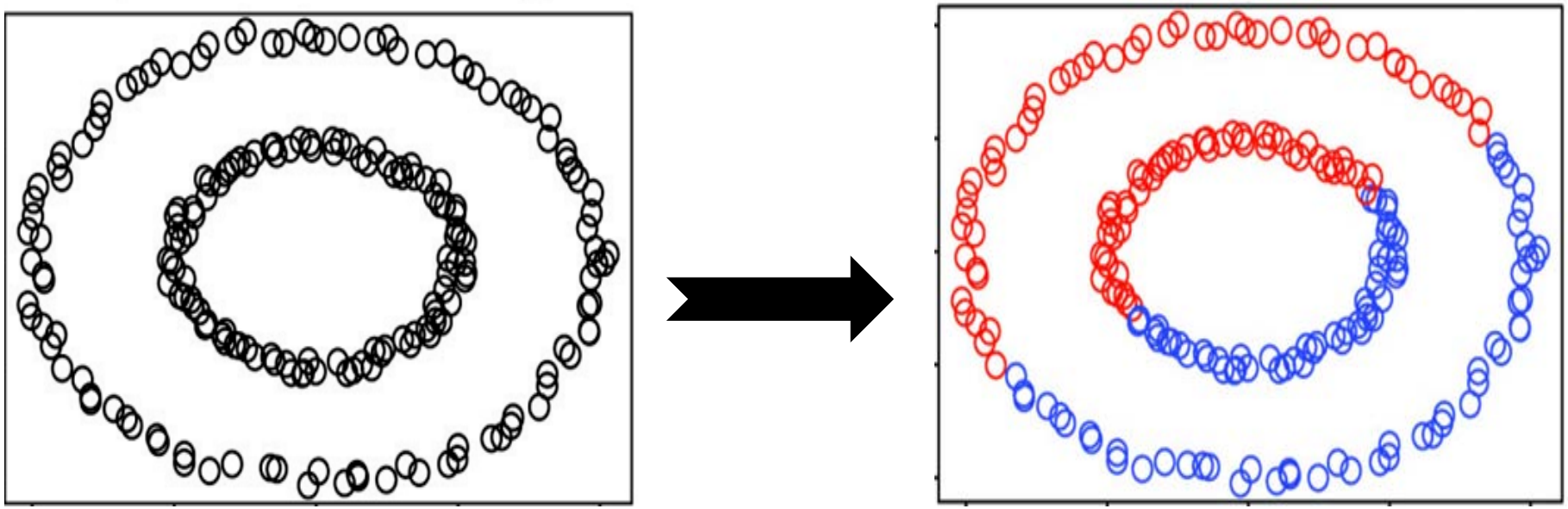


Repeat: Assign points to the nearest cluster center
Recompute the cluster centers

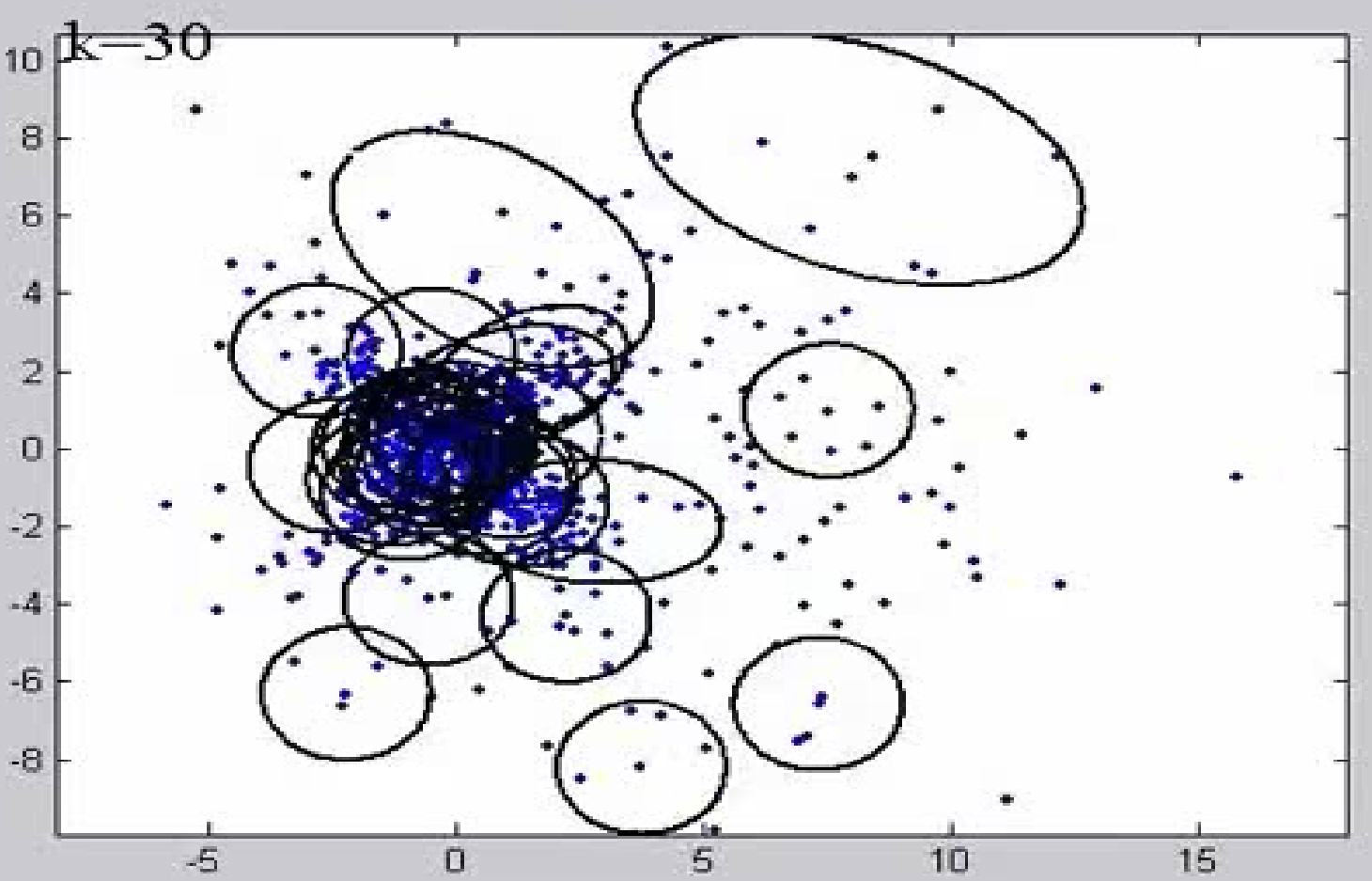
K-means: Limitations

Prefers “compact” and “isolated” clusters

$$\min \sum_{i=1}^{n} \sum_{k=1}^{K} u_{ik} \|x_i - c_k\|^2$$



Gaussian Mixture Model

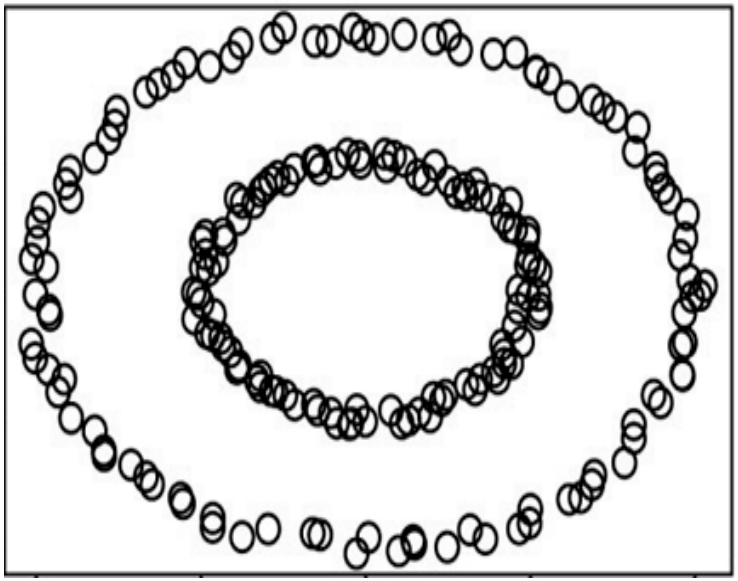


Figueiredo & Jain, “Unsupervised Learning of Finite Mixture Models”, PAMI, 2002

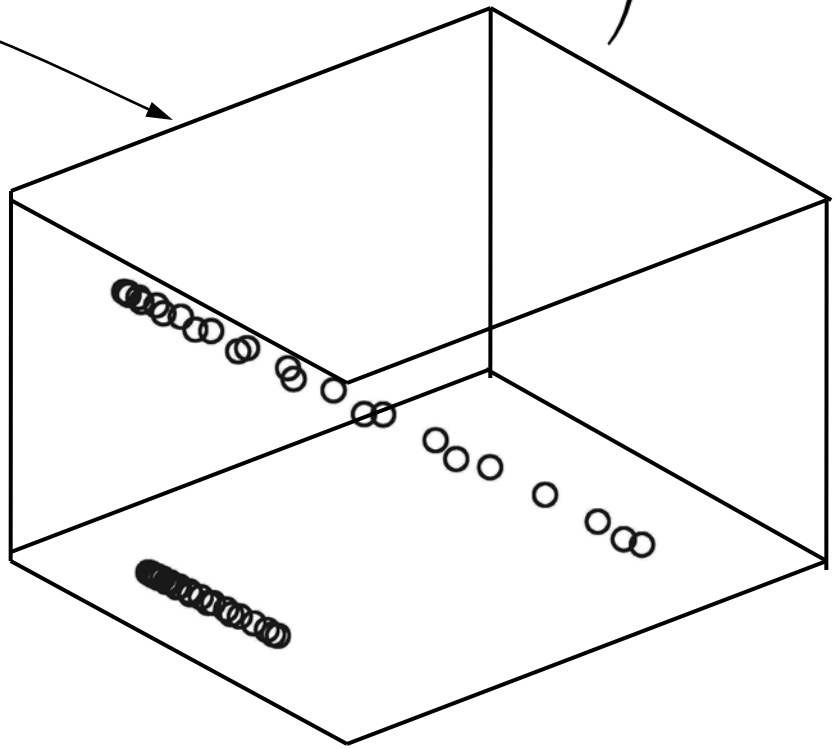
Kernel K-means

Non-linear mapping to find clusters of arbitrary shapes

$$\min Trace\left(\sum_{i=1}^{n}\sum_{k=1}^{K} u_{ik}\left(\phi(x_i) - c_k{}^{\phi}\right)\left(\phi(x_i) - c_k{}^{\phi}\right)^T\right)$$



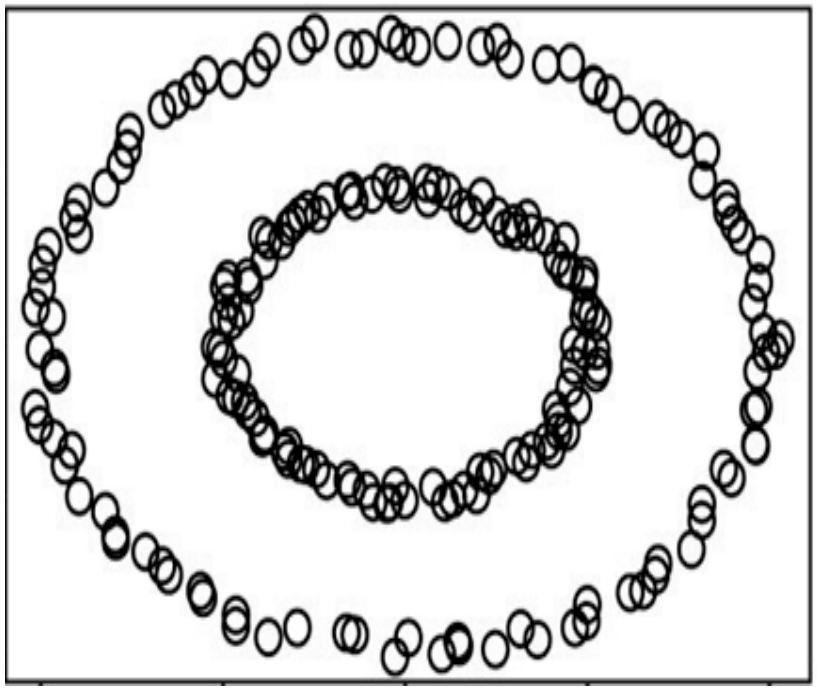
ϕ

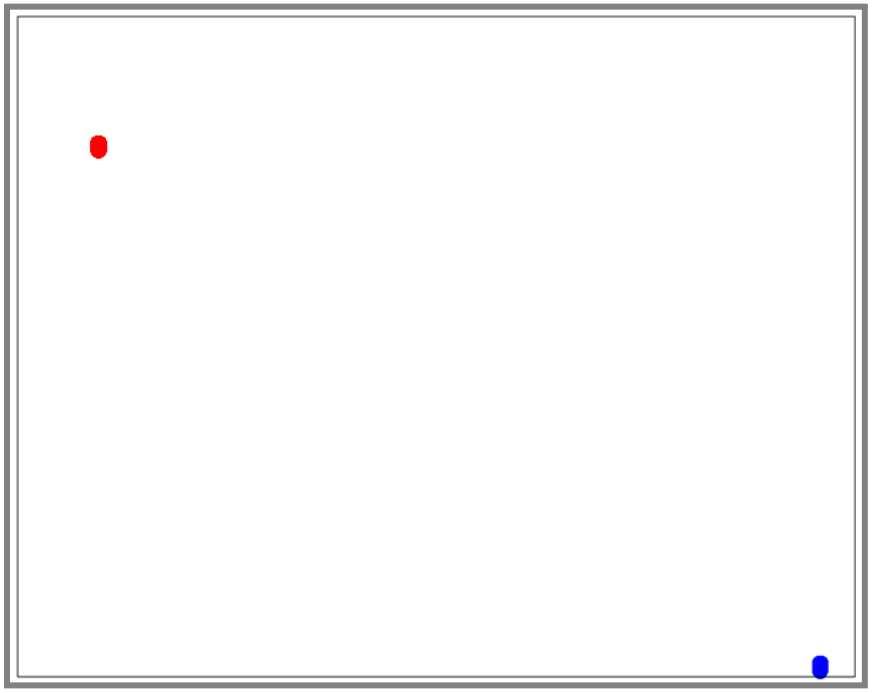


$\phi(x, y) = (x^2, \sqrt{2}xy, y^2)^T \quad K(a, b) = \phi(a)^T\phi(b)$

Polynomial kernel representation

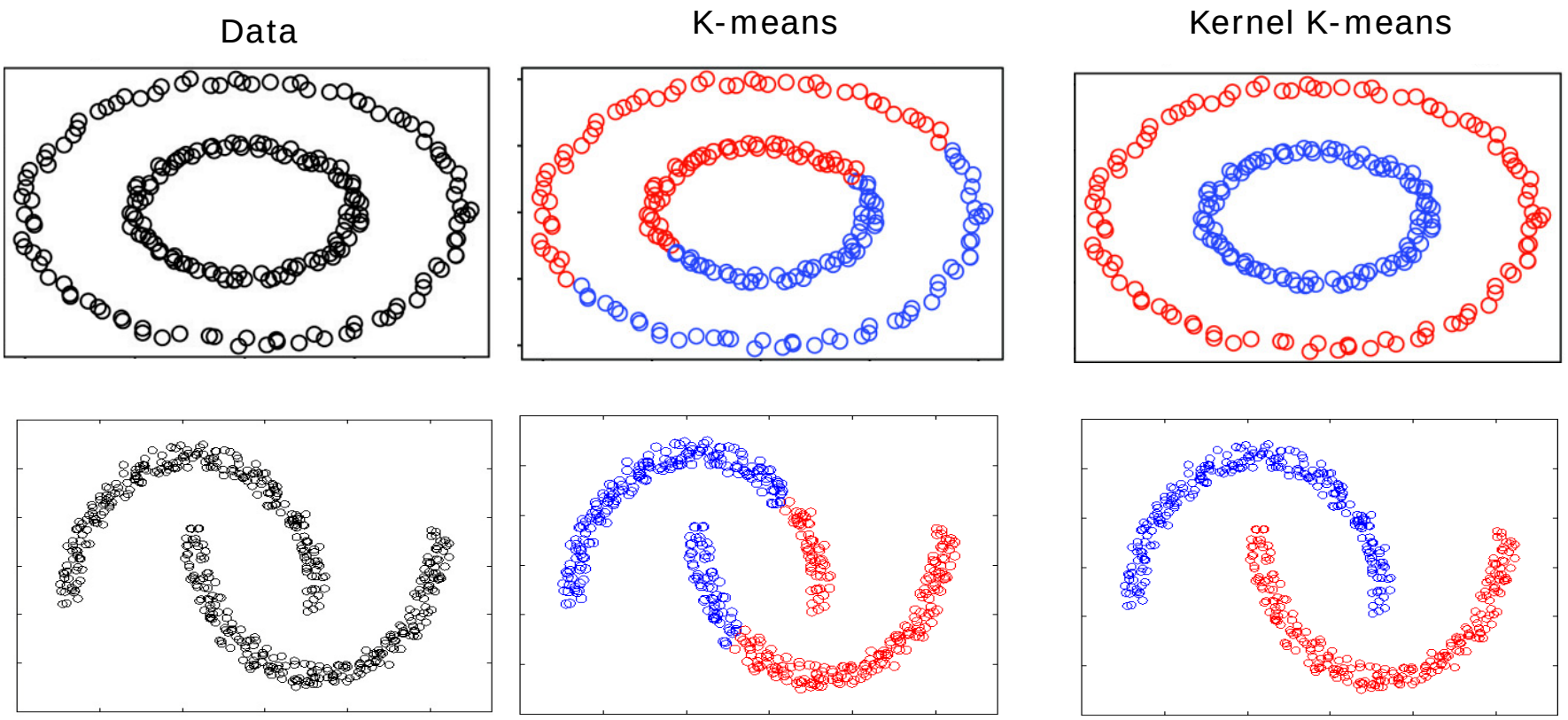
Spectral Clustering





Represent data using the top K eigenvectors of the kernel matrix; equivalent to Kernel K-means

K-means vs. Kernel K-means



Kernel clustering is able to find “complex” clusters

How to choose the right kernel? RBF kernel is the default

Kernel K-means is Expensive

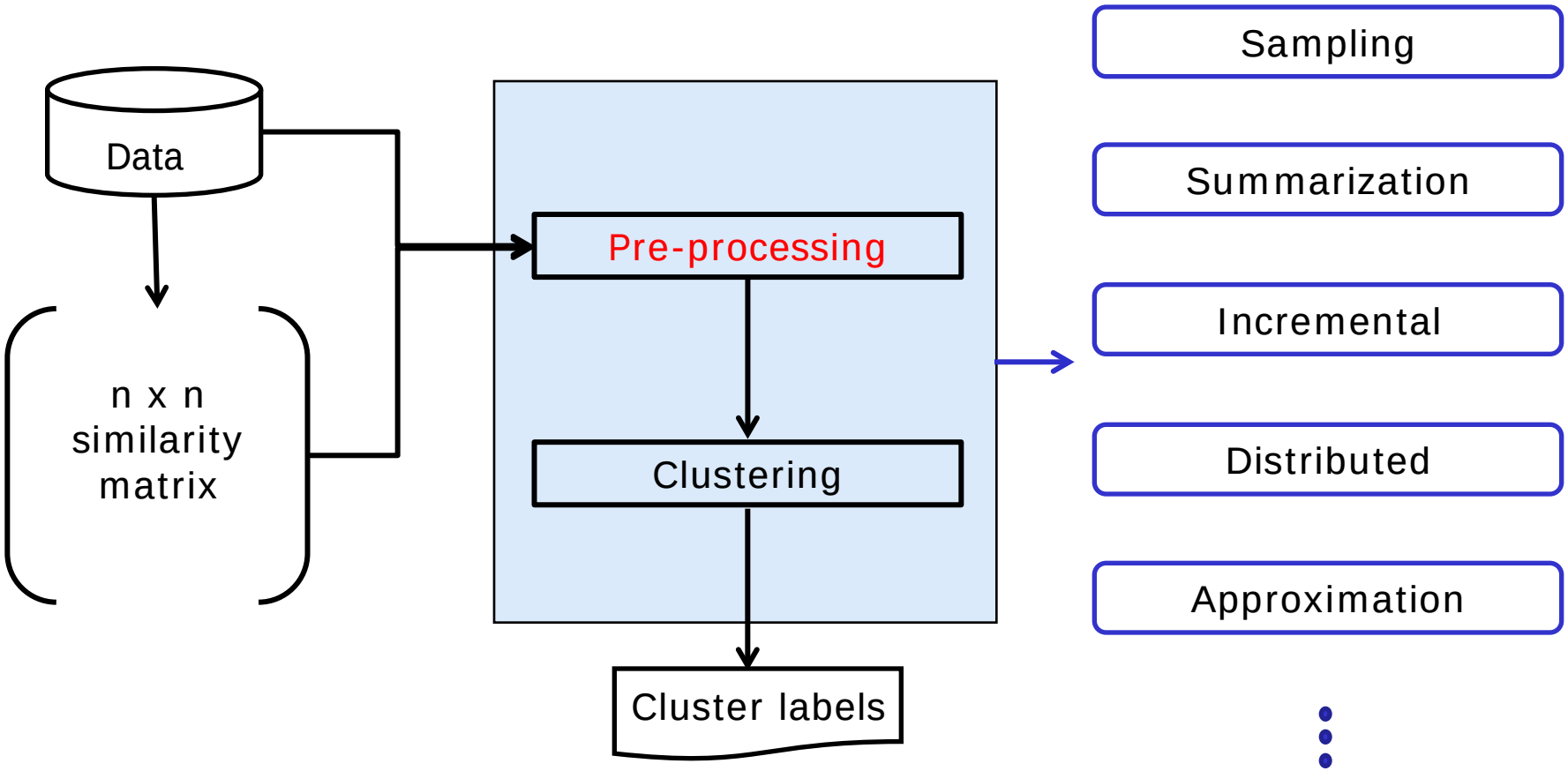
No. of Objects (n)	No. of operations	
	K-means	Kernel K-means
	O(nKd)	$O(n^2K)$
1M	10^{13} (6412*)	10^{16}
10M	10^{14}	10^{18}
100M	10^{15}	10^{20}
1B	10^{16}	10^{22}

d = 10,000; K=10

* Runtime in seconds on Intel Xeon 2.8 GHz processor using 40 GB memory

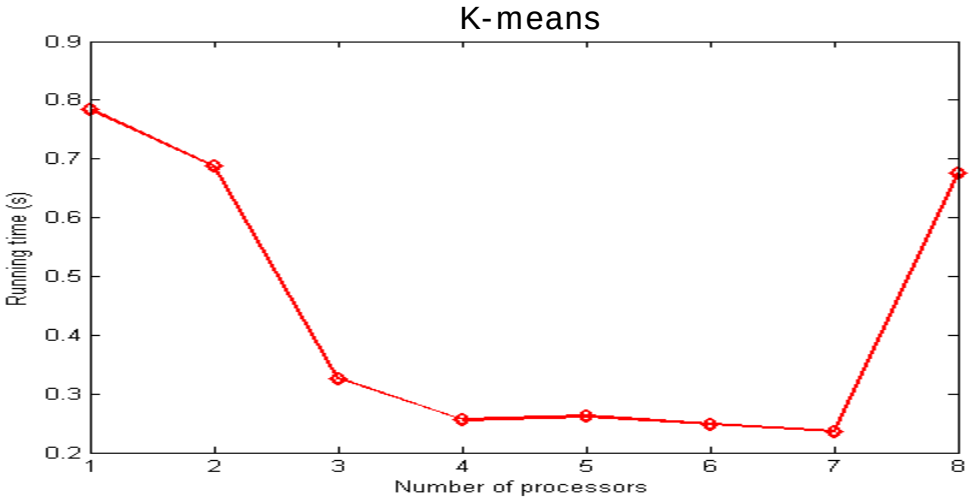
A petascale supercomputer (IBM Sequoia, June 2012) with ~1 exabyte memory is needed to run kernel K-means on 1 billion points!

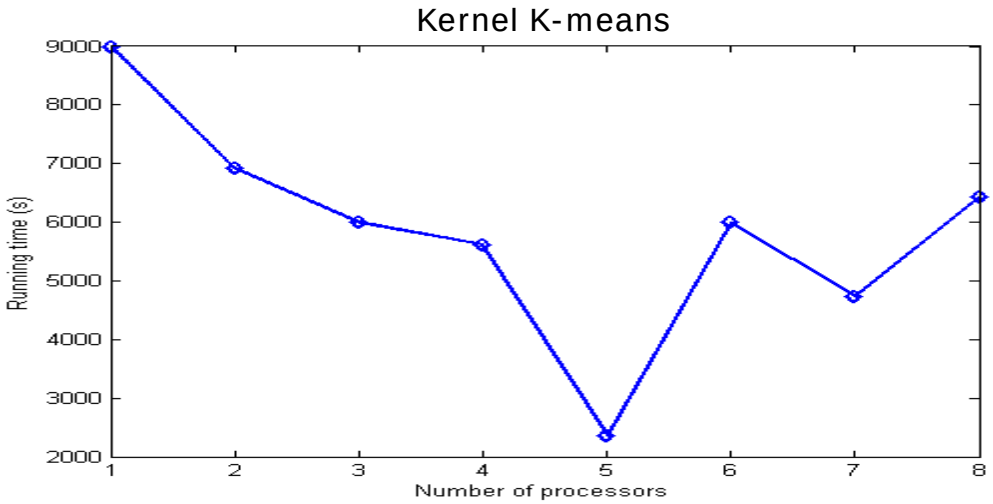
Clustering Big Data



Distributed Clustering

Clustering 100,000 2-D points with 2 clusters on 2.3 GHz quad-core Intel Xeon processors, with 8GB memory in intel07 cluster



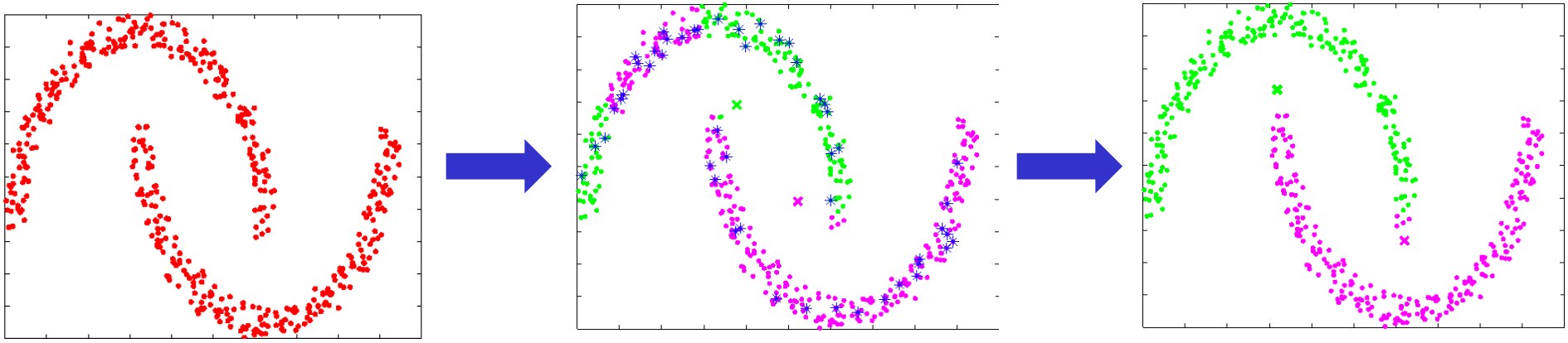


Number of processors	Speedup	
	K-means	Kernel K-means
2	1.1	1.3
3	2.4	1.5
4	3.1	1.6
5	3.0	3.8
6	3.1	1.9
7	3.3	1.5
8	1.2	1.5

Network communication cost increases with the no. of processors

Approximate kernel K-means

Tradeoff between clustering accuracy and running time



Randomly sample m points from the dataset

Empirically approximate the m-dimensional space

Compute the kernel similarity matrices $K_A\ (m \times m)$ and $K_B\ (n \times m)$

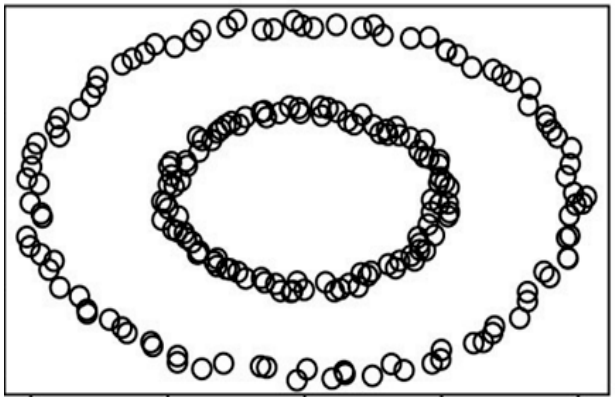
Linear runtime and memory complexity

Given K_A and K_B, find k clusters

(equivalent to running K-means on $K_B K_A^{-1} K_B^T$)

Chitta, Jin, Havens & Jain, Approximate Kernel k-means: solution to Large Scale Kernel Clustering, *KDD*, 2011

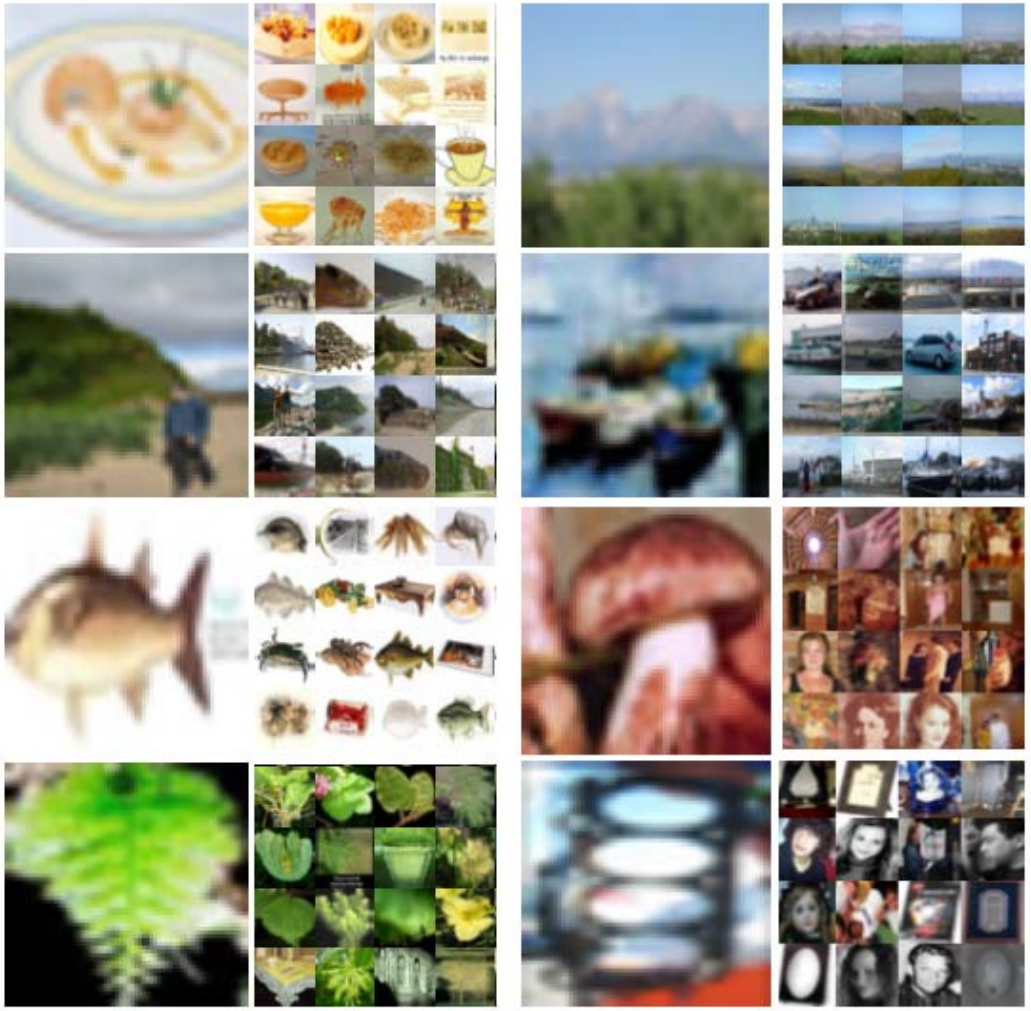
Approximate Kernel K-Means



No. of objects (n)	Running time (seconds)			Clustering accuracy (%)		
	Kernel K-means	Approximate kernel K-means (m=100)	K-means	Kernel K-means	Approximate kernel K-means (m=100)	K-means
10K	3.09	0.20	0.03	100	93.8	50.1
100K	320.10	1.18	0.17	100	93.7	49.9
1M	-	15.06	0.72	-	95.1	50.0
10M	-	234.49	12.14	-	91.6	50.0

2.8 GHz processor, 40 GB

Tiny Image Data set

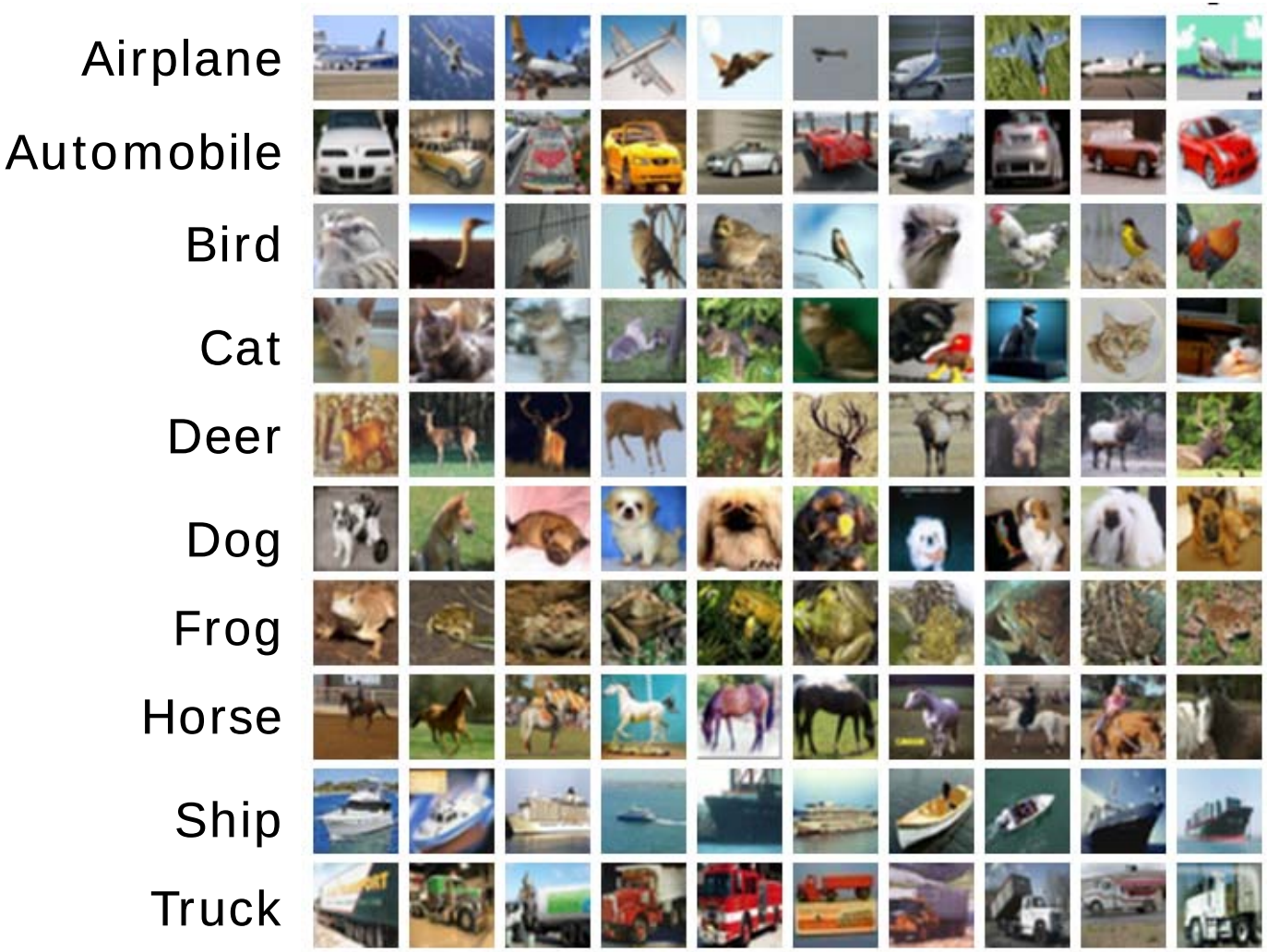


~80 million 32x32 images from ~75K classes (bamboo, fish, mushroom, leaf, mountain,...); image represented by 384-dim. GIST descriptors

Fergus et al., *80 million tiny images: a large dataset for non-parametric object and scene recognition*, PAMI 2008

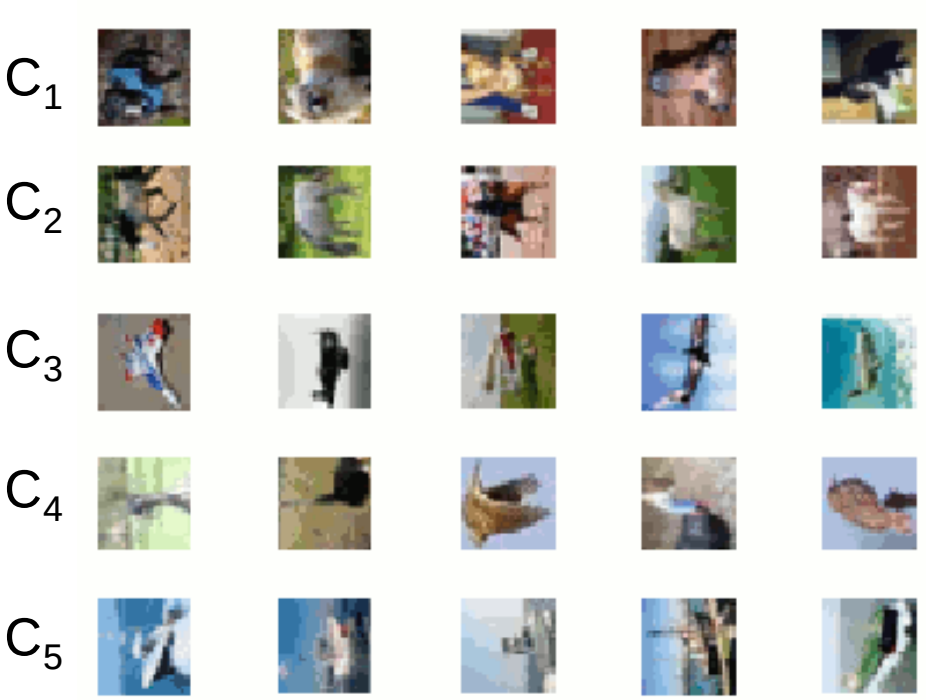
Tiny Image Data set

10-class subset (CIFAR-10): 60K manually annotated images



Krizhevsky, *Learning multiple layers of features from tiny images*, 2009

Clustering Tiny Images



Example Clusters

Average clustering time (100 clusters)	
Approximate kernel K-means (m=1,000)	8.5 hours
K-means	6 hours

2.3GHz, 150GB memory

Clustering Tiny Images

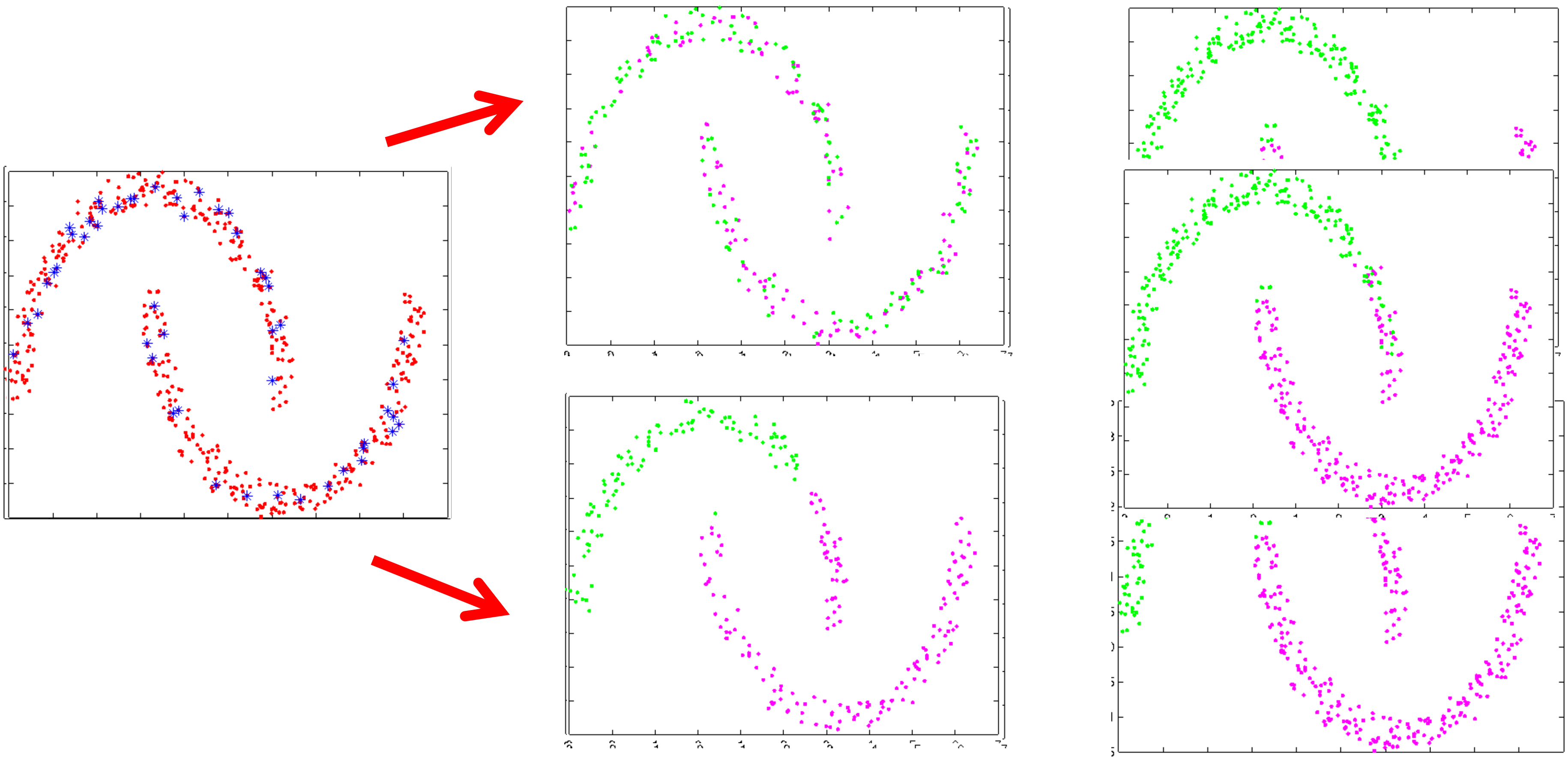
Best Supervised Classification Accuracy on CIFAR-10: 54.7%

Clustering accuracy	
Kernel K-means	29.94%
Approximate kernel K-means (m = 5,000)	29.76%
Spectral clustering	27.09%
K-means	26.70%

Ranzato et. Al., Modeling pixel means and covariances using factorized third-order boltzmann machines, CVPR 2010
Fowlkes et al., *Spectral grouping using the Nystrom method*, PAMI 2004

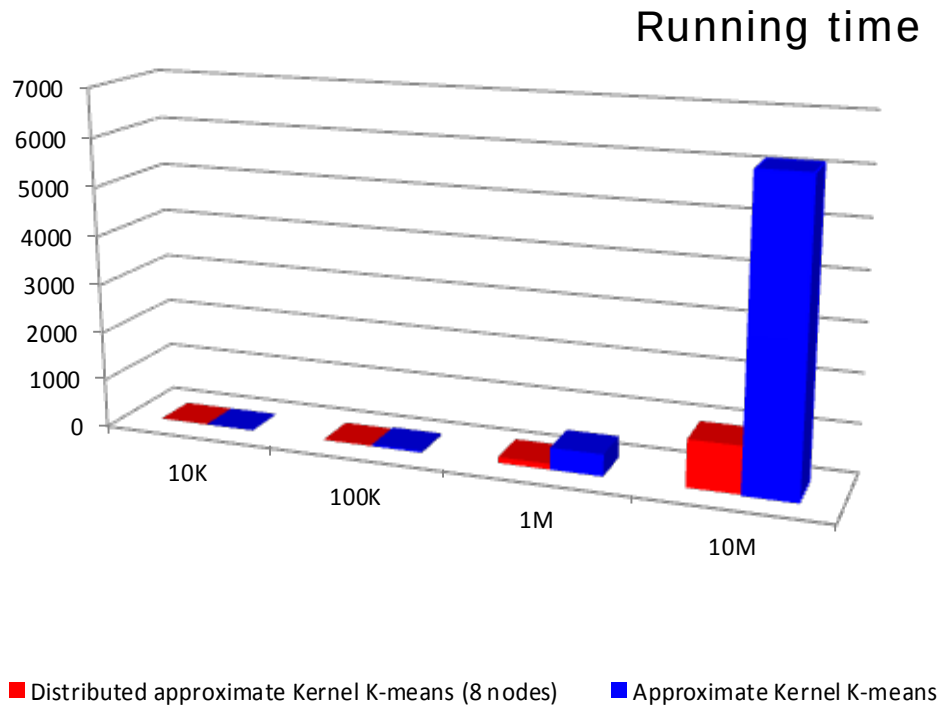
Distributed Approx. Kernel K-means

For better scalability and faster clustering



Distributed Approximate kernel K-means

2-D data set with 2 concentric circles



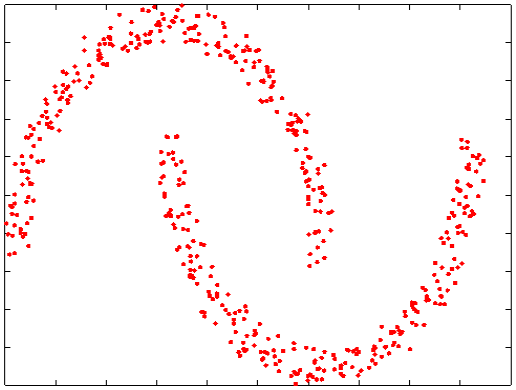
2.3 GHz quad-core Intel Xeon processors, with 8GB memory in the intel07 cluster

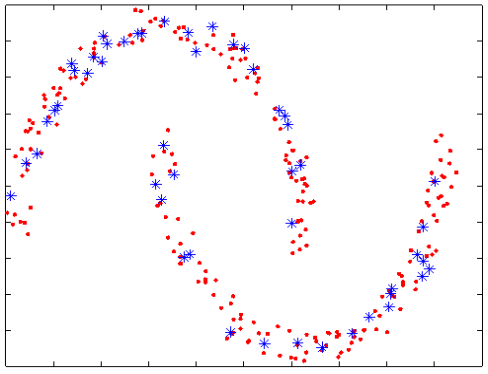
Size of data set	Speedup
10K	3.8
100K	4.8
1M	3.8
10M	6.4

Limitations of Approx. kernel K-means

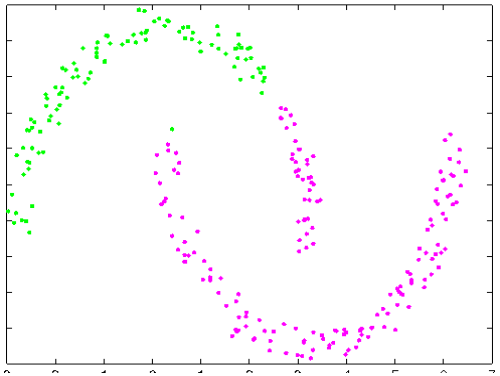
Clustering data with more than 10 million points will require terabytes of memory!

Sample and Cluster Algorithm (SnC)

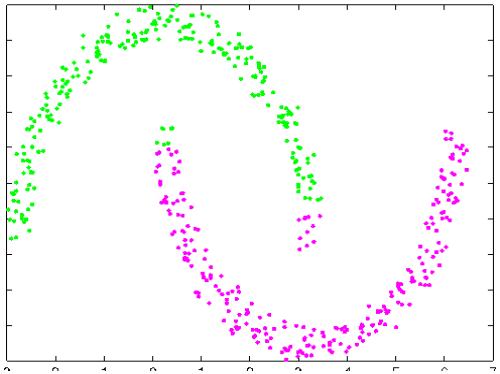




Sample s points from data



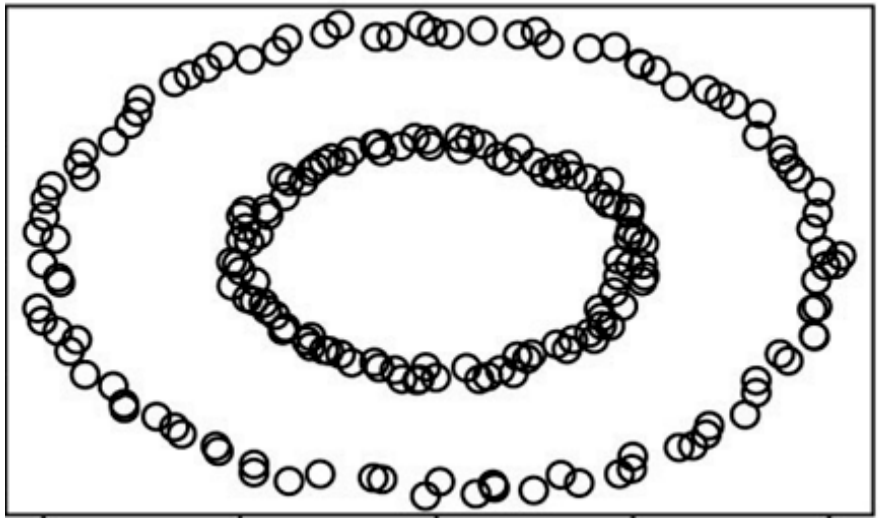
Run approximate kernel K-means on the s points



Assign remaining points to the nearest cluster center

Clustering one billion points

Sample and Cluster (s = 1 million, m = 100)



Running time			Average Clustering Accuracy	
K-means	SnC	SnC –distributed (8 cores)	K-means	SnC
53 minutes	1.2 hours	45 minutes	50%	85%

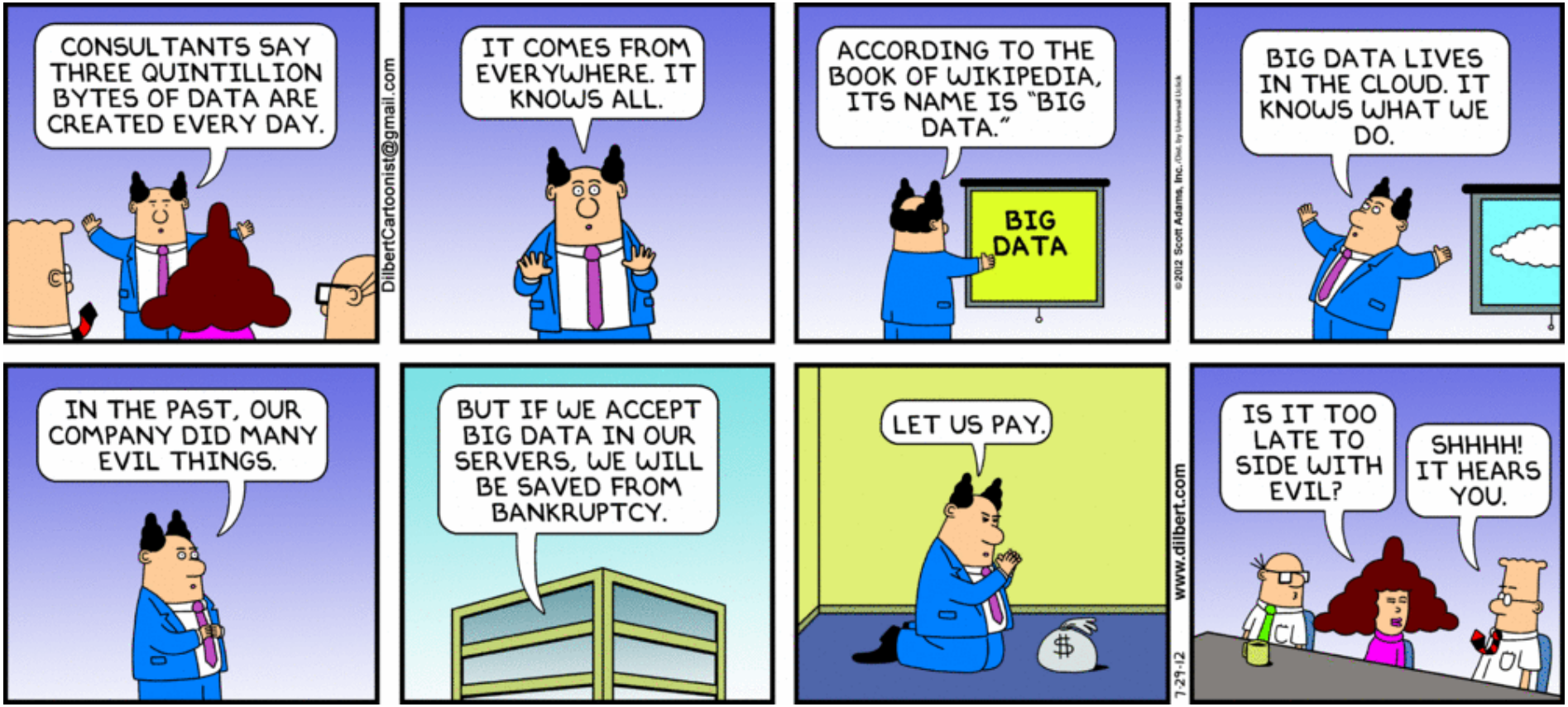
Clustering billions of points

- Work in progress
  - Application to real data sets
  - Yahoo! AltaVista Web Page Hyperlink Connectivity Graph (2002) containing URLs and hyperlinks for over 1.4 billion public web pages
- Challenges
  - Graph Sparsity: Reduce the dimensionality using random projection, PCA
  - Cluster Evaluation: No ground truth available, internal measures such as link density of clusters

Summary

- Clustering is an exploratory technique; used in every scientific field that collects data
- Choice of clustering algorithm & its parameters is data dependent
- Clustering is essential for “Big Data” problem
- Approximate kernel K-means provides good tradeoff between scalability & clustering accuracy
- Challenges: Scalability, very large no. of clusters, heterogeneous data, streaming data, validity

Big Data



http://dilbert.com/strips/comic/2012-07-29/