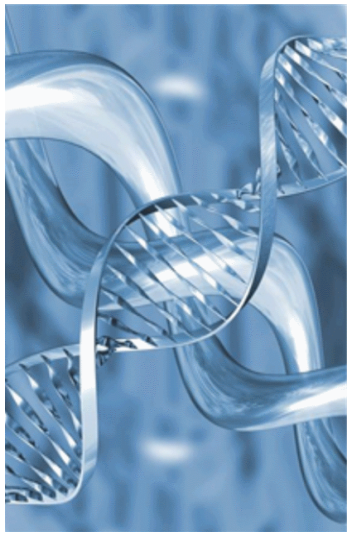
PENNSTATE

Exploiting Sparsity in the Statistical Analysis of Gene Expression Data

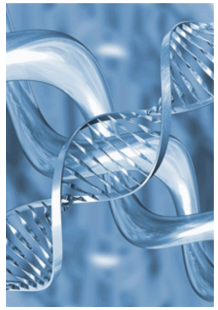
Anirban Chatterjee
Padma Raghavan
Francesca Chiaromonte

The Pennsylvania State University



Supported by NSF through grants NSF-STHEC 0444345 and NSF CSR-SMA 0720749.
(*Chatterjee, Bhowmick, Raghavan, Textmining at SDM 2008, longer version under review)*

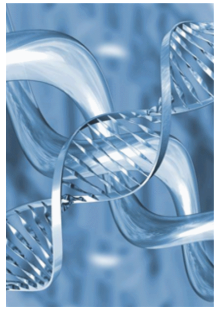
PENNSTATE



Outline

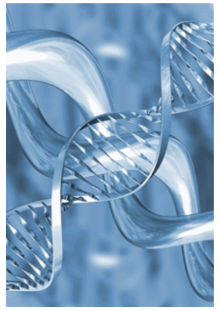
- Introduction
- Force-directed Graph Embedding
- Feature Subspace Transformation (FST)
- FST-K-Means Clustering
- Clustering Gene Expression Data
- Summary

PENNSTATE



Introduction

- Scientific dataset sizes growing rapidly with high throughput instruments, especially, in *Life Sciences*.
- FST-K-Means: Fast yet improved clustering by utilizing sparsity and structure. (Chatterjee, Bhowmick, Raghavan, Textmining at SDM 2008, longer version under review)
  - SPARSITY
    - An mxn matrix is sparse if the number of nonzeros is O(m) or O(n).
    - Many feature values are either zero or numerically insignificant.
    - Observed data are sparse and high-dimensional.
  - STRUCTURE
    - Similarity between observations indicates a relationship.
    - Relationship induces a structure in the data.
- Feature subspace transformation: Combines sparsity and structure through embedding in high-dimensional feature space.



Force-directed Graph Embedding

Fruchterman & Reingold Graph Drawing
(Prior work)

Attractive force

Repulsive force

$$\Delta d_{fr} = \sum \frac{u - v}{d_{uv}} fa_{uv} + \sum \frac{u - v}{d_{uv}} fr_{uv}$$

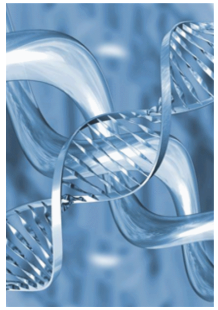
Calculated for all vertex pairs (***O(V²)***)

Calculated for vertices connected by an edge (***O(E)***)

where, d_{uv} = Euclidean distance between document *u* and *v*

$$fa_{uv} = \frac{d_{uv}^2}{k} \qquad fr_{uv} = -\frac{k^2}{d_{uv}} \qquad k = C\sqrt{\frac{area}{number\ \ of\ \ vertices}}$$

T.M.J. Fruchterman and E.M. Reingold. *Graph drawing by force-directed placement*. Software Practice and Experience, 21(11):1129--1164, November 1991.



Force-directed Graph Embedding

(Our modifications)

$$\Delta d_{fr} = \sum \frac{u - v}{d_{uv}} fa_{uv} * \frac{w_{uv}}{iter_i}$$

Edge weight

Current iteration count

Calculated for vertices connected by an edge (***O(E)***)

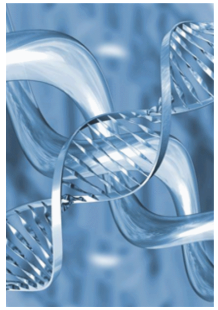
where, d_{uv} = Euclidean distance between document *u* and *v*

$$fa_{uv} = \frac{d_{uv}^2}{k} \qquad k = C\sqrt{\frac{area}{number \quad of \quad vertices}}$$

During embedding only non-zero term entries are modified,

i.e. entity vectors modified only in active dimensions or terms.

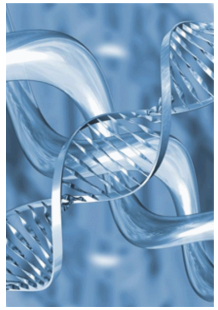
Computational Cost of Modified Embedding



- Gain in time complexity.
- Fruchterman-Reingold computational costs using attractive + repulsive forces is ***O(V² + E)***
- Our approach
  - No repulsive force calculation
  - Reduced costs: ***O(E)***

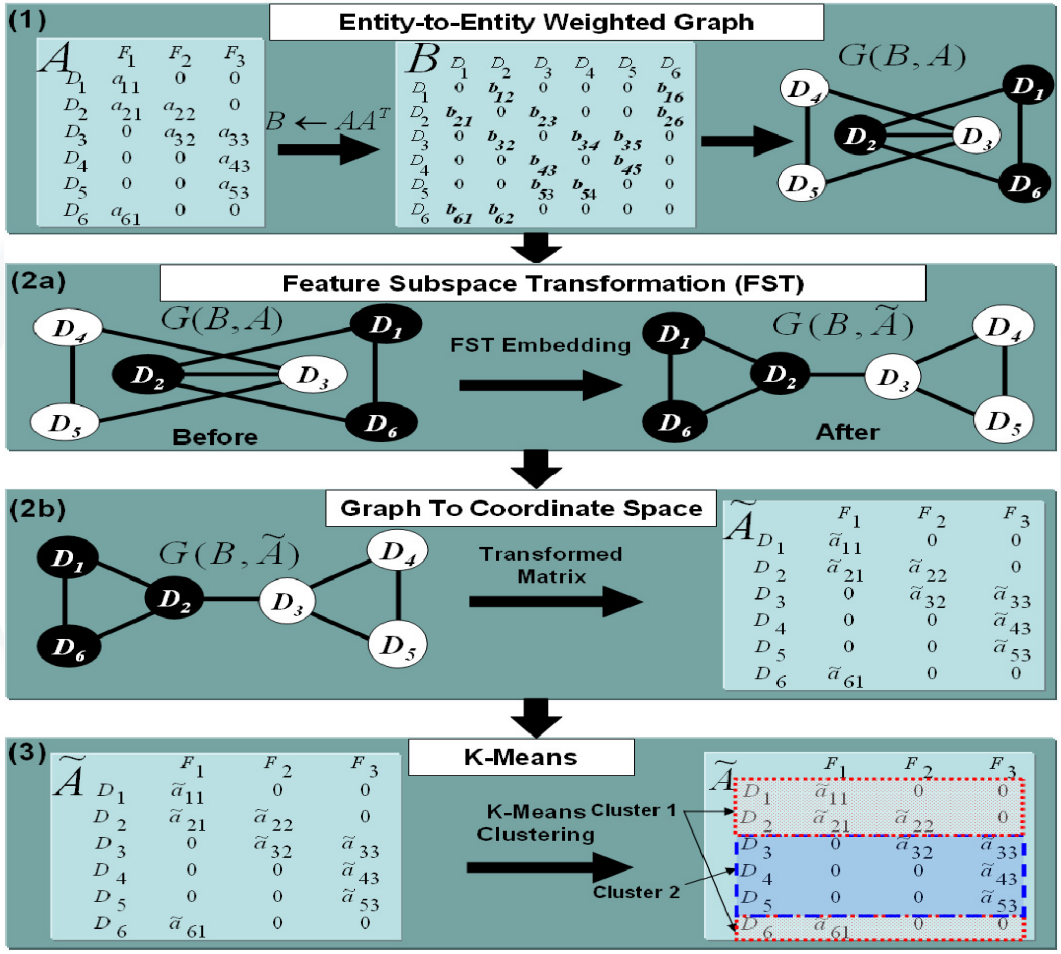
PENNSTATE

FST: Feature Subspace Transformation

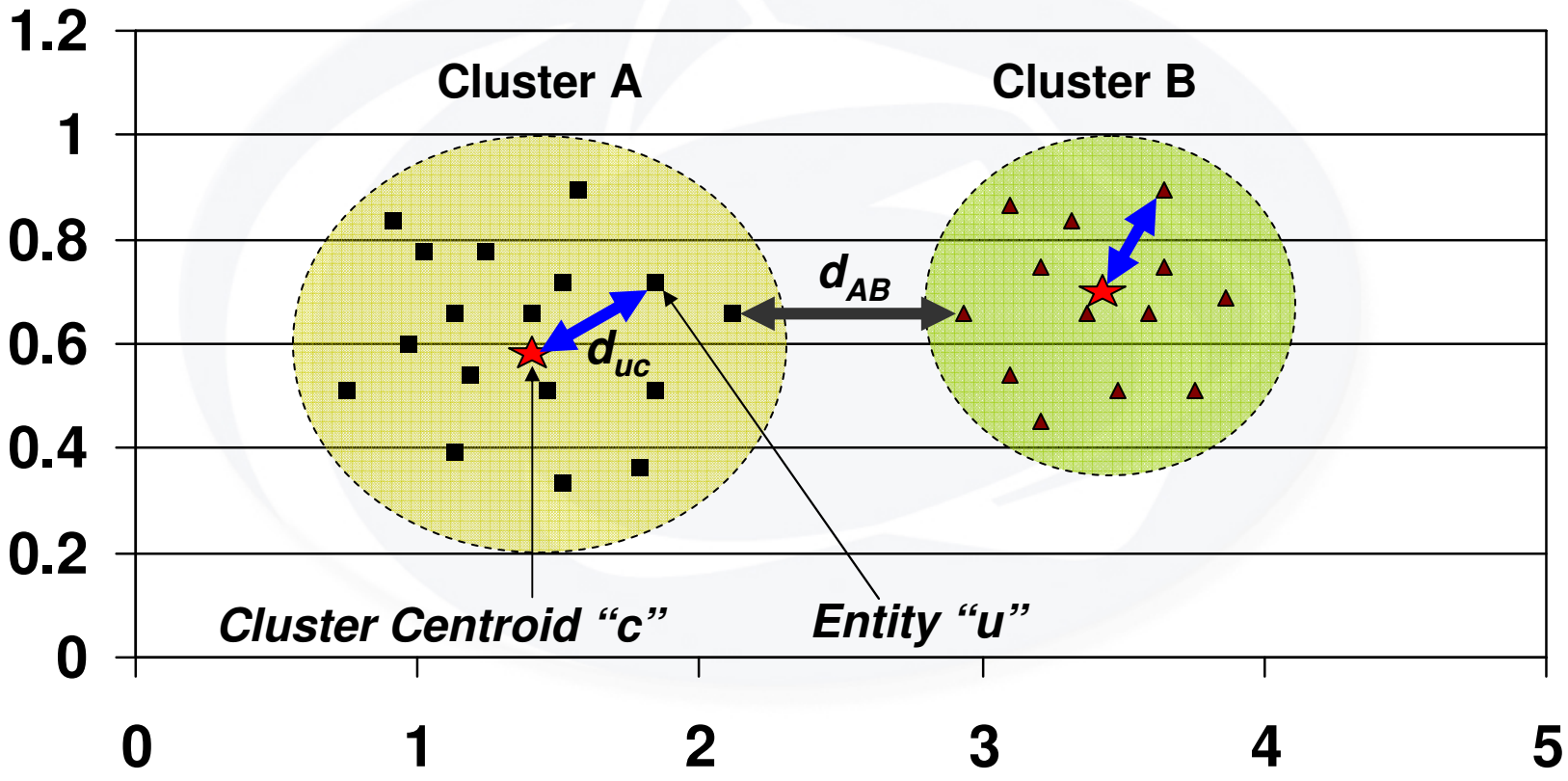


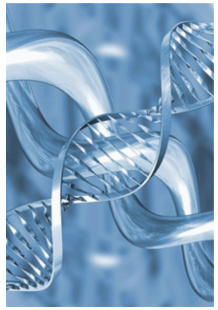
- Entity-Feature matrix
  - An ***N*x*R*** sparse matrix ***A*** of N entities and R features.
  - Each entry a_{ij} is the number of times feature t_j appears in entity d_i.
- Entity Graph
  - An undirected graph ***G = (V,E)*** with N entities and E edges between these entities.
  - Edge weight $\boldsymbol{e_{ij}}$ is the **number of common features** between the two entities.

FST: Main Steps



Cluster Quality Metrics





Quality Metrics: Internal

- External Quality Metric
  - Accuracy

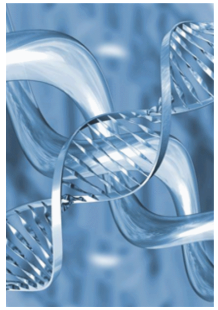
$$P = \frac{\varepsilon}{N}$$

  - where,

ε → Number of correctly classified documents

N → Total number of documents

PENNSTATE



Quality Metrics: Internal

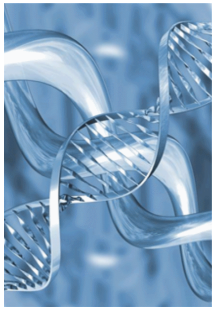
- Internal Quality Metrics
  - Measure of intra-cluster cohesiveness

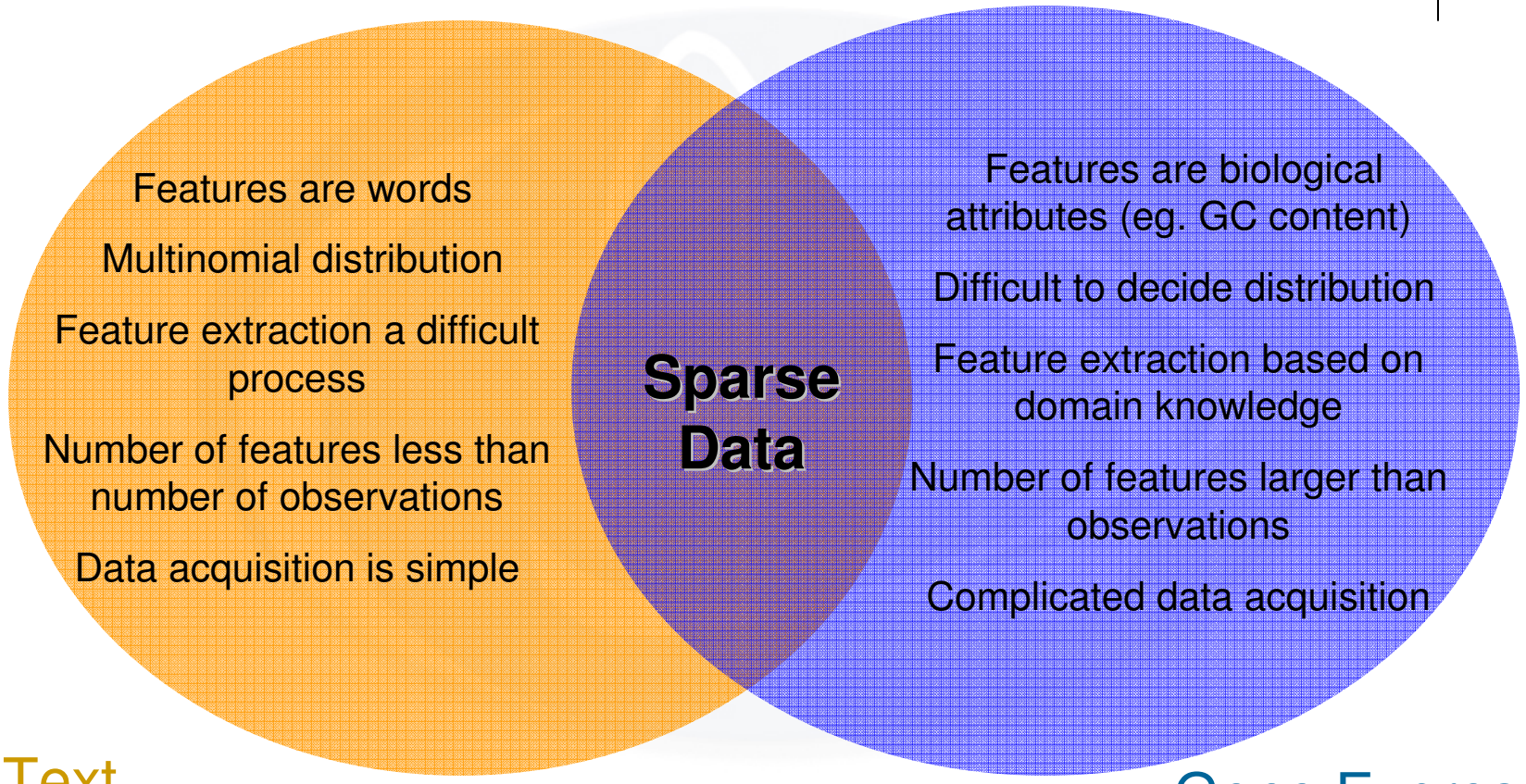
$$J(M_1, M_2 \cdots M_k) = \sum_{M_{ij} \in M_1} \left\| M_{ij} - w_1 \right\|^2 + \cdots + \sum_{M_{ij} \in M_k} \left\| M_{ij} - w_k \right\|^2$$

  - where,
    - M_c → Set of documents belonging to cluster ***c***
    - w_k → Centroid of cluster $\mathbf{M_c}$
    - M_{ij} → **jth** document of cluster $\mathbf{M_i}$
    - k → Total number of clusters

S. M. Savaresi, D. L. Boley, S. Bittanti, and G. Gazzaniga. Cluster selection in divisive clustering algorithms. In *Inproceeding SIAM Data Mining Conference*, Arlington, VA,2002

Text and Gene Datasets

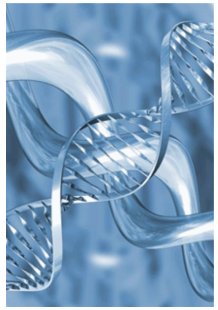




Text Datasets

Gene Expression Datasets

PENNSTATE

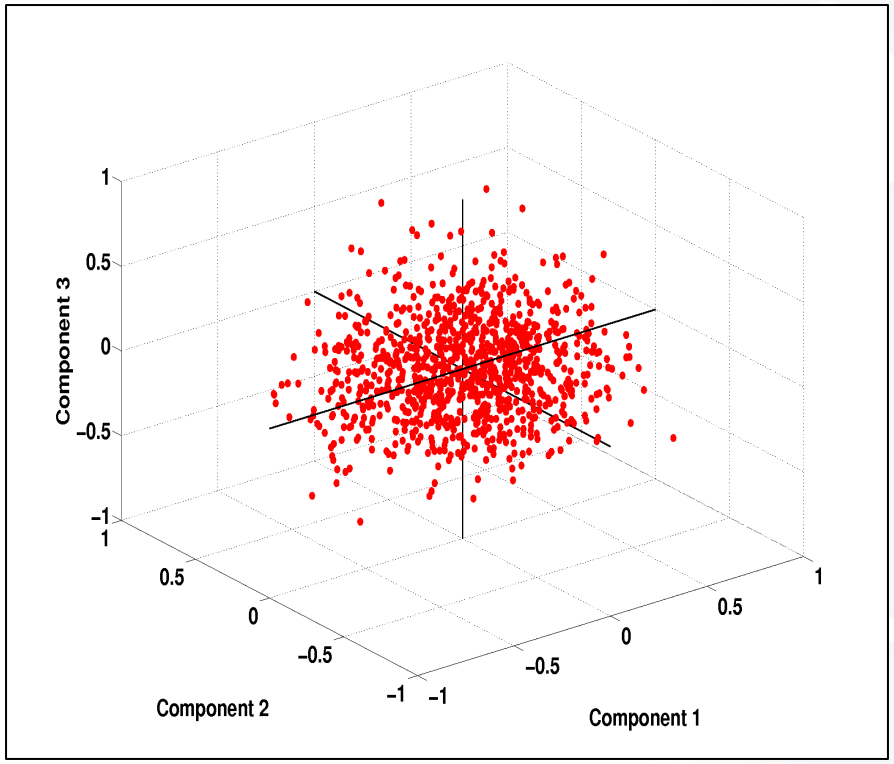


Datasets

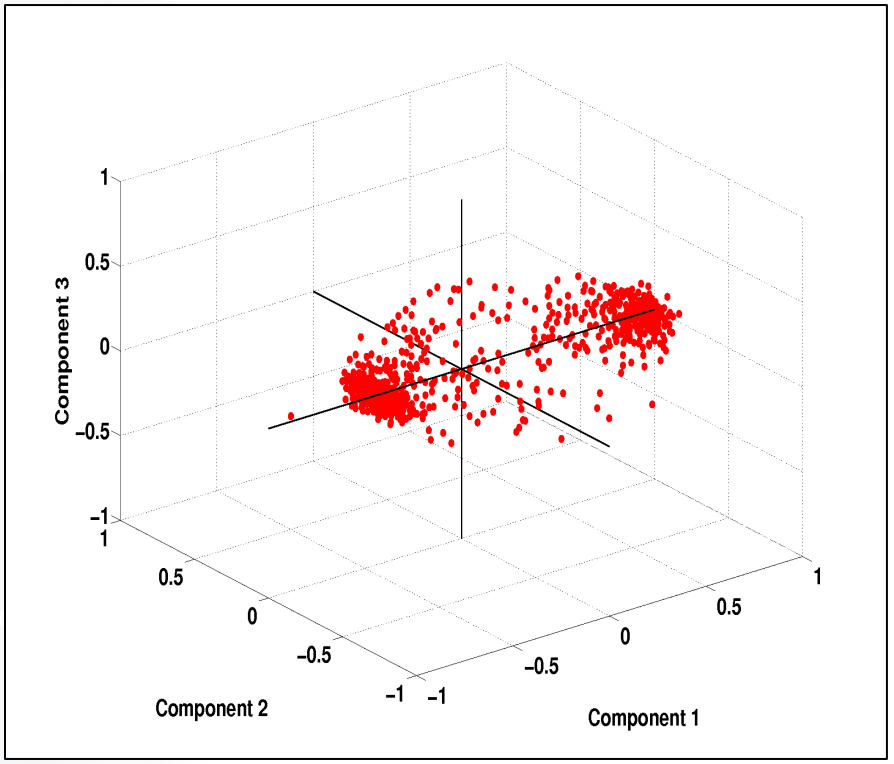
Name	Samples	Features	Source(Type)
adult_a2a	2,265	123	UCI(Census)
australian	690	14	UCI(Credit Card)
breast-cancer	683	10	UCI(Census)
dna	2,000	180	Statlog(Medical)
splice	1,000	60	Delve(Medical)
180txt	180	19,698	SMART(Text)
300txt	300	53,914	SMART(Text)
20news	1,061	16,127	Yahoo Newsgroup(Text)

Table 1: Test suite of datasets

FST-K-Means on DNA SPLICE



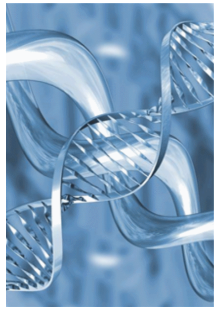
(a)



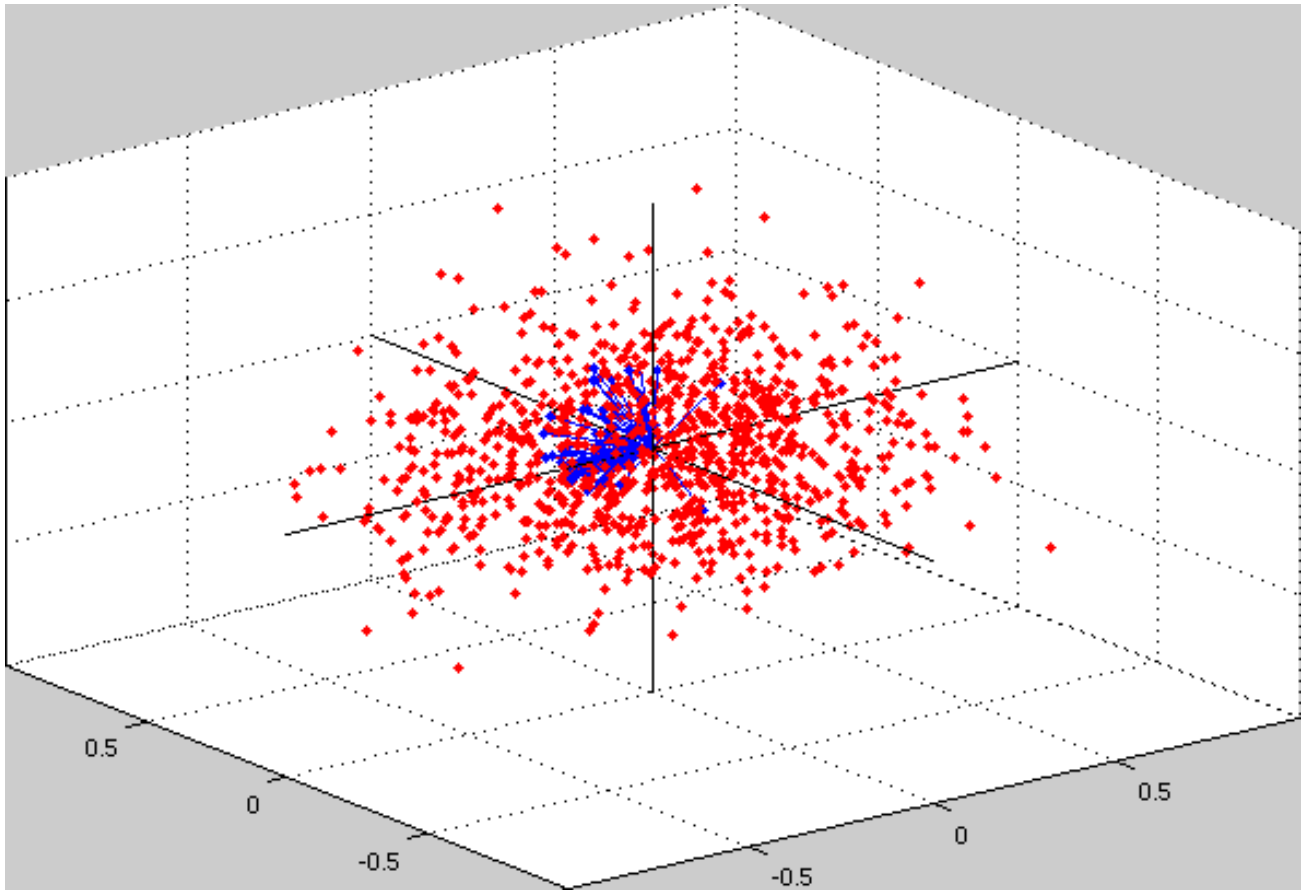
(b)

Figure 2: (a) Original data projected to first three principal components (b) Embedded data projected to first three principal components

PENNSTATE

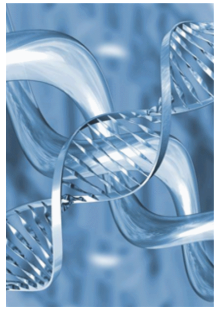


FST-K-Means on DNA SPLICE



We observe an improvement of *25.27%* in clustering accuracy relative to K-Means

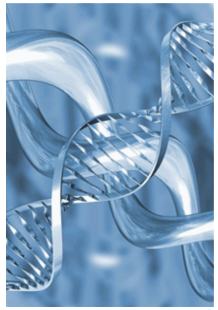
PENNSTATE



Classification Accuracy

Datasets	Classification Accuracy (P)		
	K-Means	MLKM	FST-K-Means
adult_a2a	70.60	52.49	**74.17**
australian	**85.51**	74.20	85.36
breast-cancer	**93.70**	69.69	83.16
dna	**72.68**	70.75	70.75
splice	55.80	53.20	**69.90**
180txt	73.33	91.67	**91.67**
300txt	78.67	64.33	**95.00**
20news	46.74	54.85	**73.70**

Table 2: Accuracy of classification of K-Means, MLKM and FST-K-Means

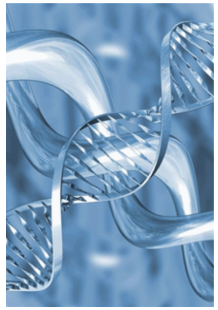


Cluster Cohesiveness

Datasets	Cluster Cohesiveness (J)		
	K-Means	MLKM	FST-K-Means
Adult_a2a	24,013	**16,665**	16,721
australian	4,266	3,034	**2,638**
breast-cancer	2,475	2,203	**1,366**
dna	84,063	**65,035**	65,545
splice	31,883	31,618	**31,205**
180txt	25,681	**23,776**	24,131
300txt	47,235	**44,667**	45,052
20news	3,851,900	3,483,591	**3,341,400**

Table 3: Cluster cohesiveness of K-Means, MLKM and FST-K-Means

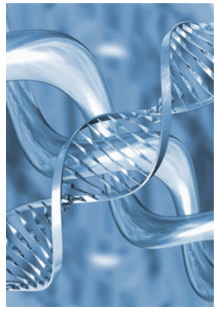
X Chromosome Inactivation (XCI) Data



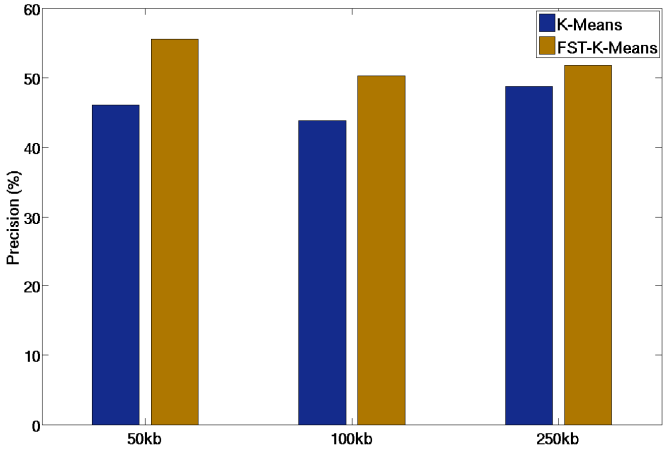
Dataset	Genes (E/I)	Features	Non-zeros
50kb	399 (346/53)	248	28480
100kb	365 (318/47)	248	42655
250kb	315 (278/37)	248	57444

Results: XCI

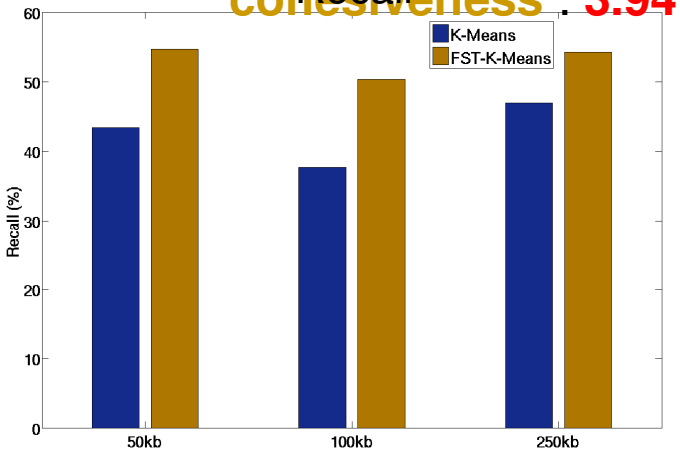
Average "Improvements":
accuracy : 24.40%.
precision : 13.83%.
recall : 25.04%.
cohesiveness : 3.94%.



Precision



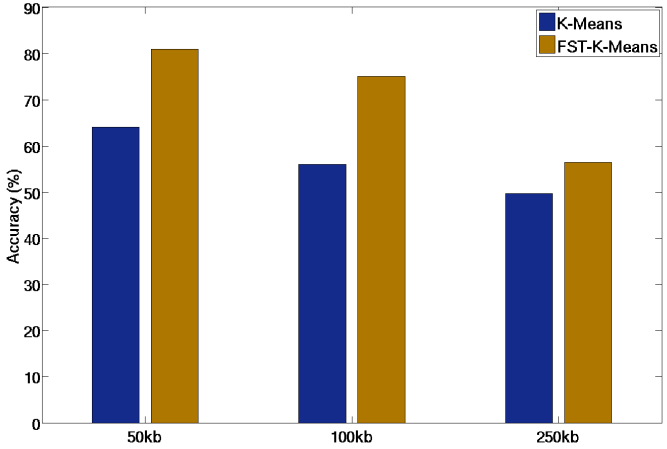
Recall



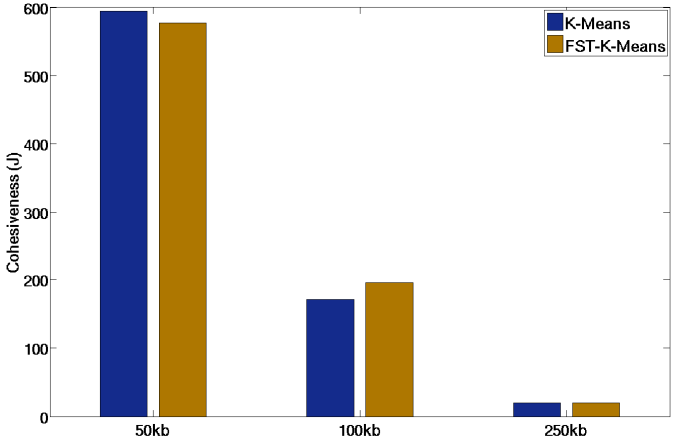
FST-K-Means

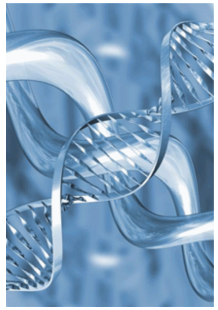
K-Means

Accuracy



Cohesiveness





Why does FST work?

Optimal clustering and bounds on cohesiveness:

Y = centered data matrix

(Mean of data A subtracted from each entity a_i)

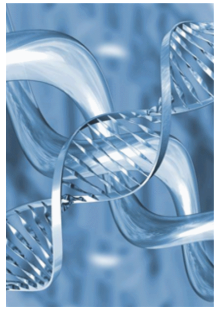
$N\bar{y}^2$ = trace of $\mathrm{Y}^{\mathrm{T}}\mathrm{Y}$

λ_i = *i-th* principal eigenvalue of $\mathrm{Y}^{\mathrm{T}}\mathrm{Y}$

$$N\bar{y}^2 - \sum_{i=1}^{k-1}(\lambda_i) \leq J \leq N\bar{y}^2$$

C.Ding and X.He, K-means clustering via principal component analysis, pages 225–232, ACM Press, 2004.

PENNSTATE



Cohesiveness after FST-K-Means

FST-K-Means

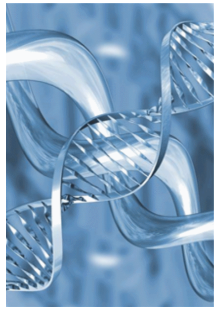
Cluster Cohesiveness: FST-K-Means				
Datasets	Lower Bound	Min	Max	Upper Bound
Adult_a2a	15,250	15,866	17,208	17,380
australian	2,442	2,638	3,008	3,493
breast-cancer	747	1,366	1,366	5,169
dna	63,890	65,525	65,865	67,190
splice	30,389	31,205	31,205	31,942
180txt	23,188	23,765	24,178	25,290
300txt	43,708	44,800	45,194	46,512

K-Means

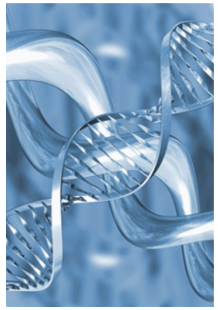
Cluster Cohesiveness: K-Means				
Datasets	Lower Bound	Min	Max	Upper Bound
Adult_a2a	15,250	**24,013**	24,409	17,380
australian	2,442	**4,266**	4,458	3,493
breast-cancer	747	2,475	2,475	5,169
dna	63,890	**84,062**	84,123	67,190
splice	30,389	31,882	31,884	31,942
180txt	23,188	**25,651**	25,730	25,290
300txt	43,708	**47,220**	47,288	46,512

FST-K-Means satisfies the optimality bounds while K-Means fails to do so.

PENNSTATE



Questions?



Thank You

Email: achatter@cse.psu.edu