Visual Analytics for Large-scale High Dimensional Data: from Algorithms to Software Systems

Haesun Park
School of Computational Science and Engineering
Georgia Institute of Technology
Atlanta, GA, U.S.A.

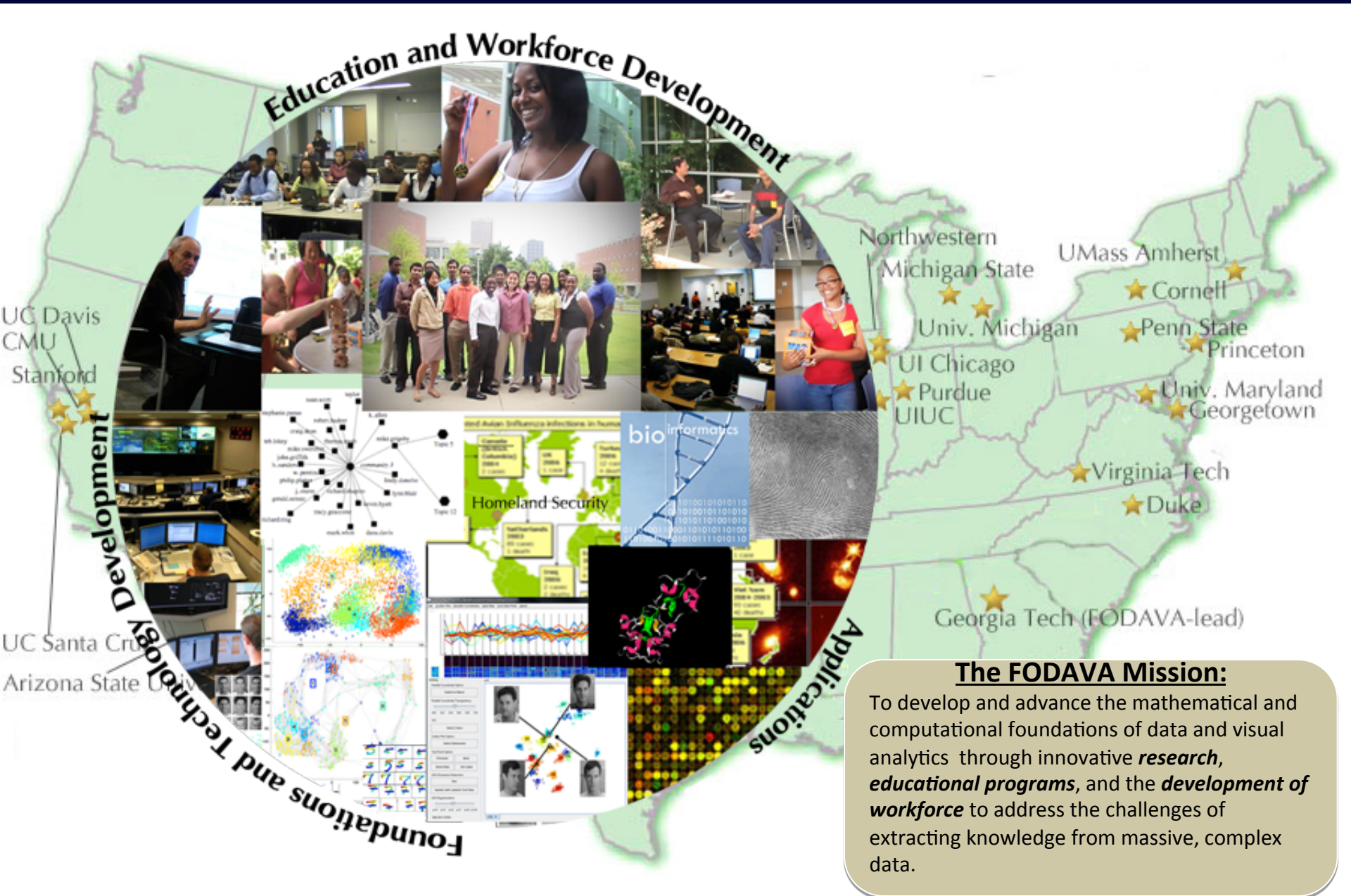
CSoI Big Data Workshop, March 18, 2013

This work is supported in part by NSF and DHS.

Contributors

- Jaegul Choo (Georgia Tech)
- Changhyun Lee (Georgia Tech)
- Hanseung Lee (Univ. of Maryland)
- Zhicheng Liu (Stanford University)
- Fuxin Li (Georgia Tech)
- Yunlong He (Georgia Tech)
- Jaeyeon Kihm (Cornell University)
- Jingu Kim (Nokia)
- Da Kuang (Georgia Tech)
- Sen Yang (Arizona State University)
- Ed Clarkson (Georgia Tech Research Institute)
- Polo Chau (Georgia Tech)
- Alexander Gray (and many of his students, Georgia Tech)
- Vladimir Koltchinskii (Georgia Tech)
- Renato Monteiro (Georgia Tech)
- John Stasko (Georgia Tech)
- Jieping Ye (Arizona State University)

FODAVA
Foundations of Data and Visual Analytics

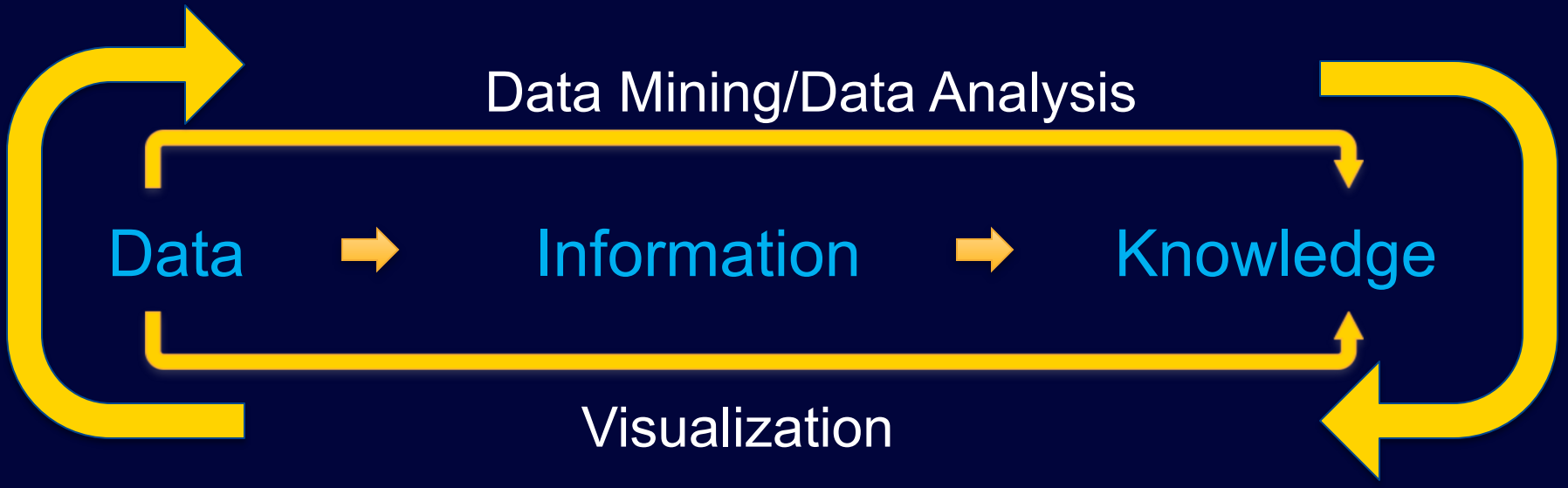


The FODAVA Mission:

To develop and advance the mathematical and computational foundations of data and visual analytics through innovative ***research***, ***educational programs***, and the ***development of workforce*** to address the challenges of extracting knowledge from massive, complex data.

Data and Visual Analytics

The Science of Analytical Reasoning facilitated by Automated Methods for Data Analysis and Interactive Visual Interfaces. (based on Thomas and Cook, Illuminating the Path: the research and development agenda for visual analytics, 2005)



"Solving a problem simply means representing it so that the solution is obvious."

Herbert Simon, 96

Challenges and our Approaches for Interactive Analysis of *High Dimensional Large-scale Data*

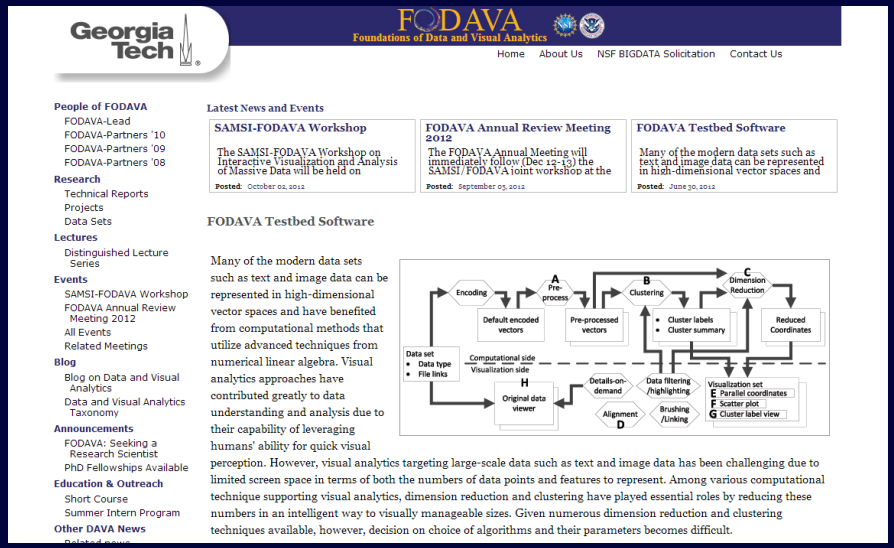
- Challenges:
  - Data are Massive, High-dimensional, Nonlinear, Unstructured, Imperfect, Heterogeneous, Time-varying, …
  - Limited Screen Space and Limited Visual Perception
  - Need for real-time Interaction
- Our Approaches:
  - Scalable and Robust algorithms:

    works even when parts of the data are missing
  - Integrated analysis: Representation of heterogeneous data on one map
  - Fast Interaction: scalable, real-time, adaptive, on-line algorithms
  - Severe dimension reduction: but key info preserved as much
  - Informative representation of large volume of data

FODAVA Research Test-bed for Visual Analytics of High Dimensional Data

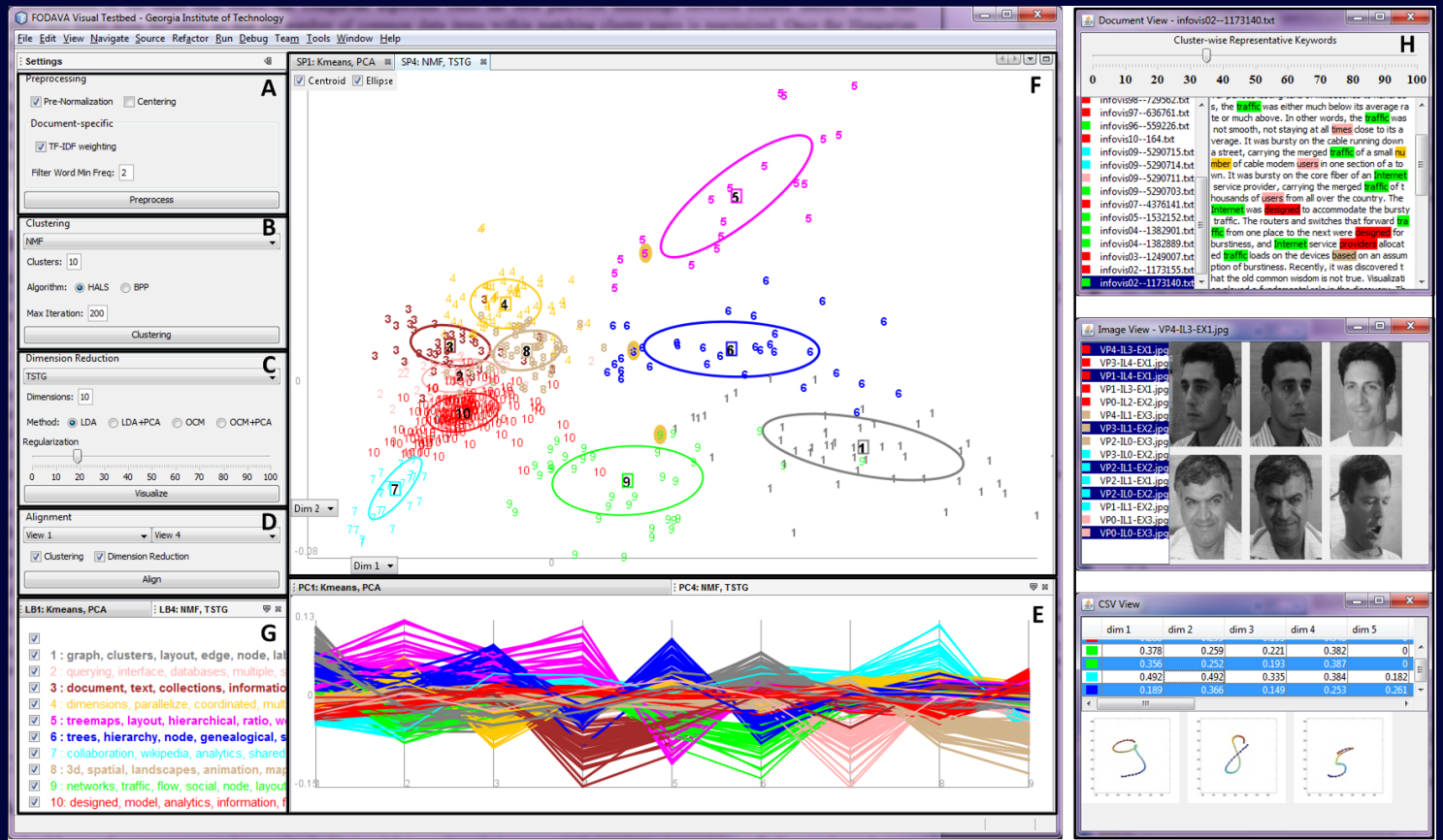
http://fodava.gatech.edu/fodava-testbed-software

- Library of key computational methods for visual analytics of high dimensional data
- Modular: A base for specialized VA systems
- Supports various dimension reduction, clustering, and their visual representations and comparisons through alignments
- Application domains: document analysis, bioinformatics, healthcare, computer vision, …
- Languages: backend library in Matlab, GUI in JAVA (no need for Matlab installed)
- System support: Windows 32/64 bit, Linux 32/64 bit



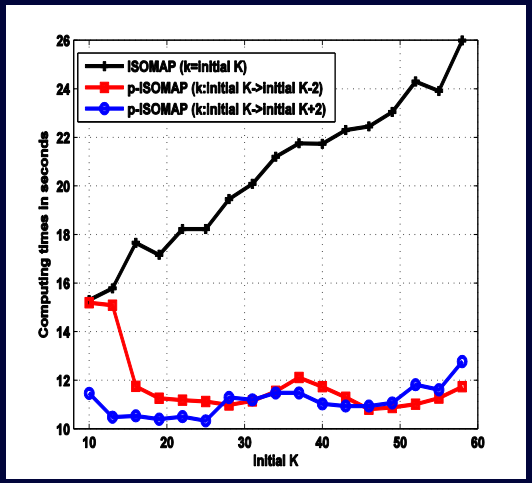
Testbed Modules and Overview

- Computational modules
  - Vector encoding
  - Pre-processing
  - Clustering
  - Dimension reduction
- Interactive visualization modules
  - Scatter plot
  - Parallel coordinates
  - Cluster summary
  - Raw data view
  - Brushing and Linking
  - Space alignment



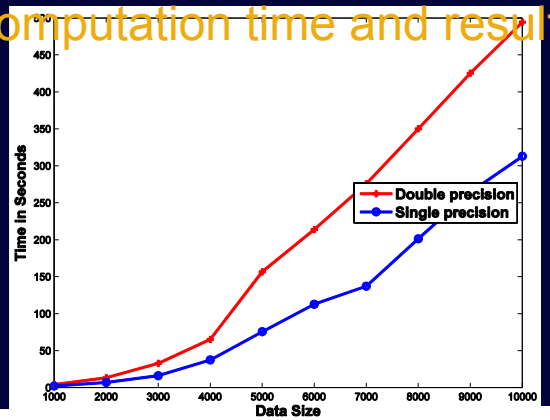
Fast Comp. Modules for Interactive Vis.

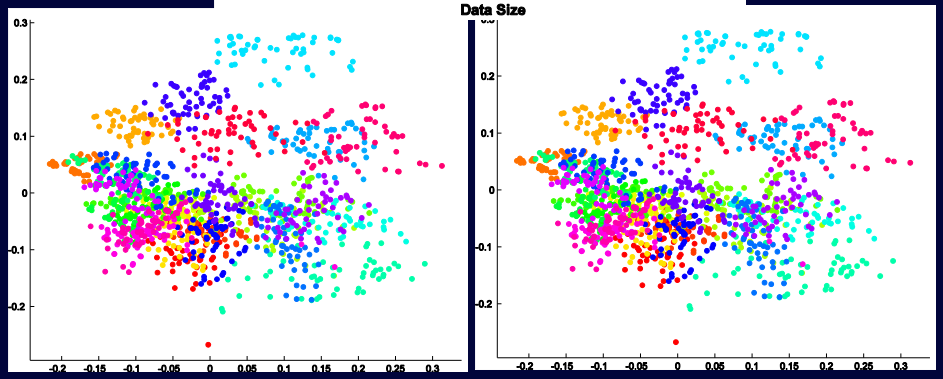
- Essential for real-time interaction
- Let computational precision be governed by visual precision/ resolution
- Hierarchical refinement
- Adaptive algorithms



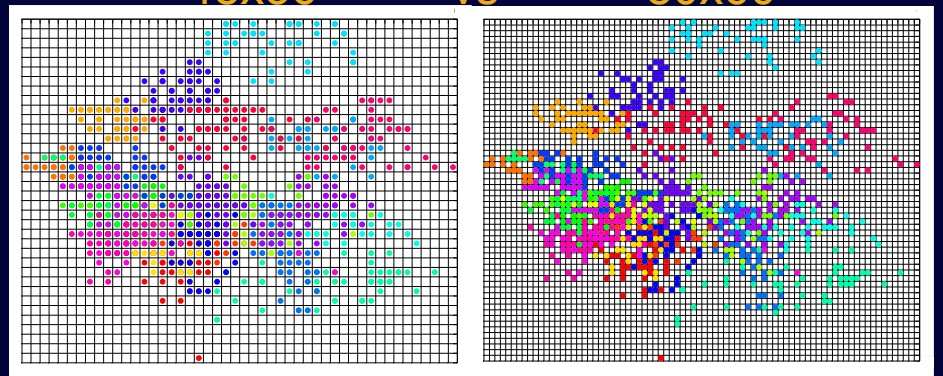
p-Isomap computing time vs. *# of nearest neighbors*

PCA timing: double vs single precision computation time and results





48x36 vs 80x60



- **Dimension Reduction**
  - Linear and Nonlinear methods: PCA, FA, ProbPCA, LDA, OCM, NPE, LPP, LLTSA, NCA, MCML, MDS, Isomap, LLE, LTSA, Sammon, HessLLE, MVU, LandMVU, KernPCA, GDA, DiffMaps, SPE, AutoEnc, LLC, ManiChart, CFA, GPLVM, SNE, T-SNE
  - Recursive dimension reduction: apply dimension reduction on user-selected data

- **Clustering**
  - Hierarchical clustering, *K*-means, spherical K-means, GMM, NMF, constrained K-means, DisCluster/ DisKmeans [J. Ye]
  - Cluster summary for document data
  - Semi-supervised clustering
  - Color-coded cluster/class labels

- **Classification**
  - *K*-nearest neighbors classifier, SVM, Logistic regression, Naïve Bayes

Key Computational Methods

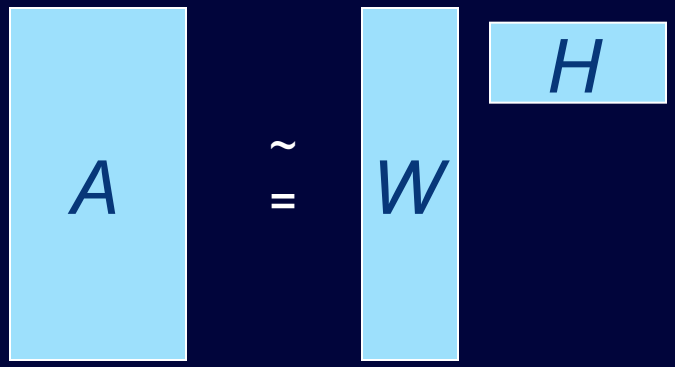
- NMF (Nonnegative Matrix Factorization) and its variations:
  for dimension reduction and clustering

- LDA/GSVD (Linear Discriminant Analysis) and its variations:
  for informative 2D representation of
  clustered and large scale data

- Orthogonal Procrustes and MDS (Multi-Dimensional Scaling):
for space alignment and comparisons of visual representations

Nonnegative Matrix Factorization (NMF)

(Paatero&Tappa 94, Lee&Seung NATURE 99, Pauca et al. SIAM DM 04, Hoyer 04, Lin 05, Berry 06, Kim and Park 06 Bioinformatics, Kim and Park 08 SIAM Journal on Matrix Analysis and Applications, …)



$\rightarrow min \; || A - WH ||_F$

$W>=0, H>=0$

- Why Nonnegativity Constraints?
  - Better Approx. vs. Better Representation/Interpretation
  - *Nonnegative Constraints often physically meaningful, interpretable*

- Fast Algorithms for NMF, with theoretical convergence (J. Kim and H. Park, IDCM08)

NMF/ANLS: Iterate the following with Active Set-type Method (ANLS/BPP)

fixing W , solve $\min_{H>=0} || WH - A||_F$

fixing H , solve $\min_{W>=0} || H^T W^T - A^T||_F$

** Software available at www.cc.gatech.edu/~hpark*

- NMF variants developed for clustering, topic modeling, and graph clustering (sNMF, tNMF, SymNMF, hierarchicalNMF, BMF for recommender system,...)

NMF and K-means

- Clustering and Lower Rank Approximation are related.
  - NMF for Clustering: (Ding et al. SDM 05; Kim & Park, TR 08)
  - Document (Xu et al. SIGIR 03), Image (Cai et al. ICDM 08), Microarray (Kim & Park, Bio 07), etc.
  - min $\sum_{1\le i\le n} \| a_i - w_{\sigma i} \|^2$ ➡ min $\| A - WH \|_F^2$
    $\sigma_i = j$ when i-th point is assigned to j-th cluster ($j \in \{1, \ldots, k\}$)

K-means: W: k cluster centroids, h_i: cluster membership indicator

NMF: W: basis vectors for rank-k approx., h_i: k-dim rep. of a_i

Sparse NMF (for sparse H) (H. Kim and Park, Bioinformatics, 07)

$\min_{W,H} \{ \|A - WH\|_F^2 + \eta \|W\|_F^2 + \beta \sum_{1\le j\le n} \|H(:, j)\|_1^2 \}$, $\forall$ i,j, $W_{ij}, H_{ij} \ge 0$

- Obj. fun. of K-means and NMF are related when H $\in$ E and A $\ge$ 0, but their performances may be very different.

NMF for Clustering

#clusters	2	6	10	14	18
K-means	0.7867	0.5137	0.4191	0.4529	0.3403
NMF/ANLS	0.9257	0.6934	0.5568	0.5654	0.43130

#clusters	2	6	10	14	18
K-means	0.8099	0.7295	0.7015	0.6675	0.6675
NMF/ANLS	0.9990	0.8717	0.7436	0.7021	0.7160
SNMF/ANLS	0.9991	0.8770	0.7512	0.7269	0.7278

Clustering Accuracy on Reuters-21578 and TCT2

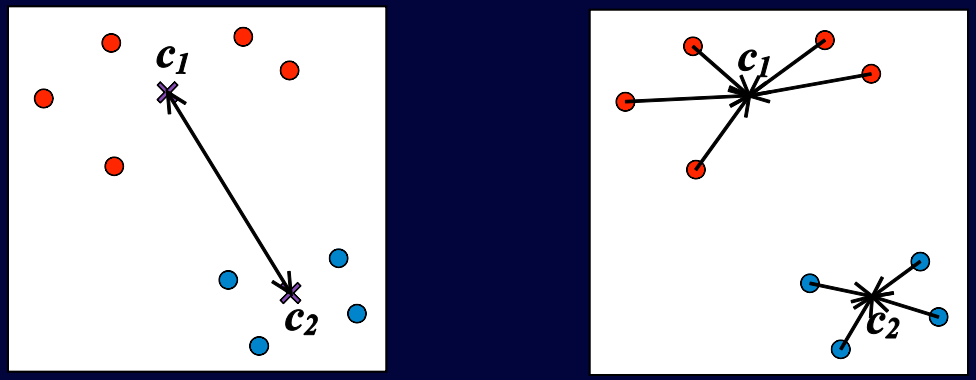
NMF is faster by factor of 2 at least

Data set	TDT2	Reuters	NIPS	ORL	Ext YaleB
Dimension	26,618	12,998	17,583	69x84	56x64
# data points	8,741	8,095	420	400	2,414
# clusters	20	20	9	40	38
Kmeans	0.6734	0.4289	0.4650	0.6499	0.0944
Ker. Kmeans	0.6789	0.3455	0.5071	0.6858	0.1692
NMF	0.8534	0.3770	0.4877	0.7020	0.1926
GNMF	0.8077	0.4441	0.4894	0.7282	0.2109
SymNMF	0.8979	0.5305	0.5129	0.7798	0.2307

Linear Discriminant Analysis for 2D/3D Representation of Clustered Data

(J. Choo, S. Bohn, HP, VAST09)

Max trace ($G^T S_b G$) **&** min trace($G^T S_w G$)



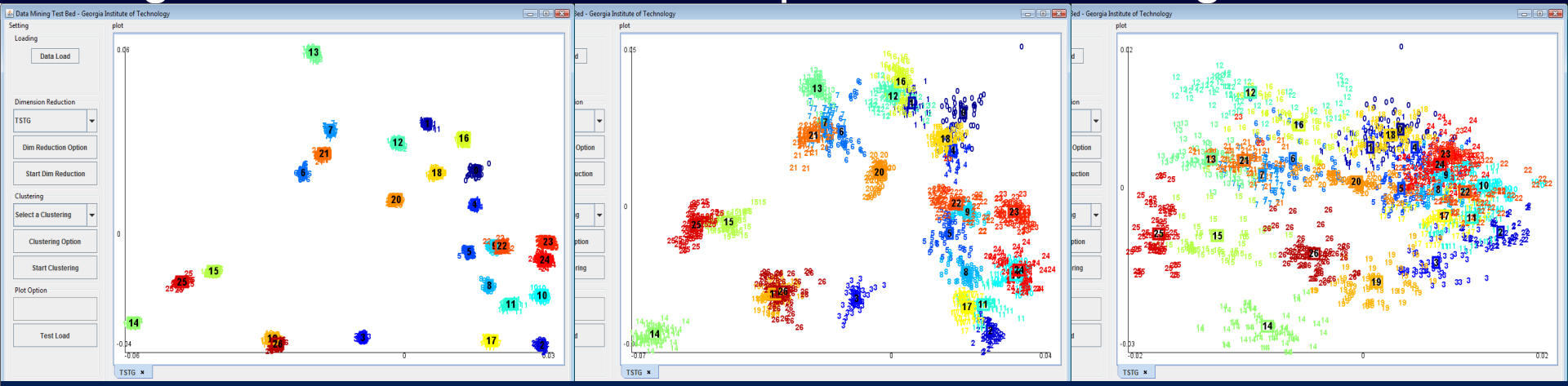
LDA/GSVD

$$\alpha^2 H_b H_b^T x = \beta^2 H_w H_w^T x$$

max trace

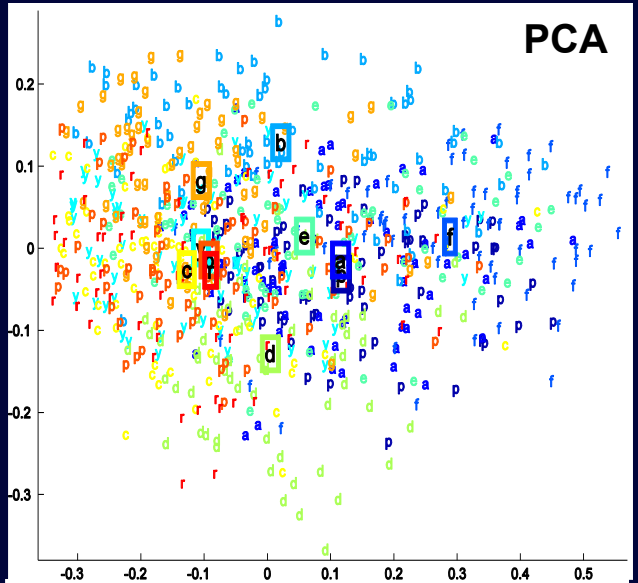
$$(G^T (S_w + \mu I) G)^{-1} (G^T S_b G)$$

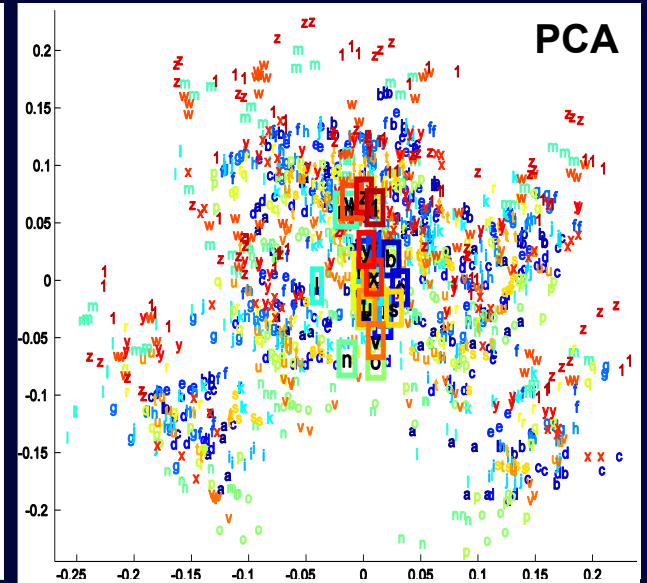
- Regularization in LDA for Computational Zooming-in

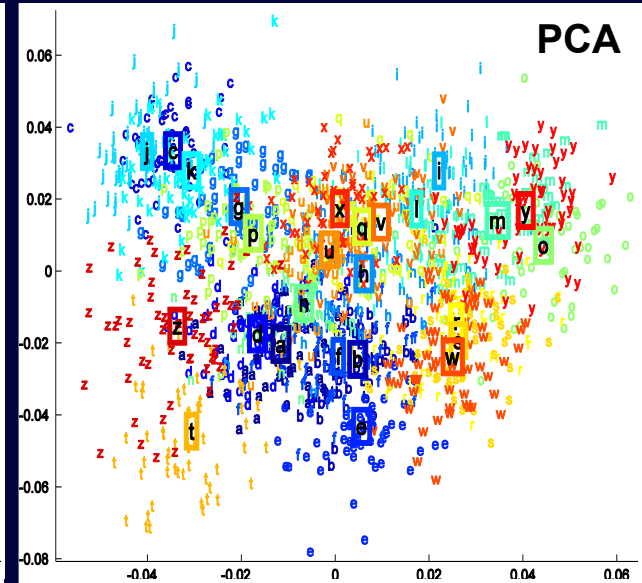


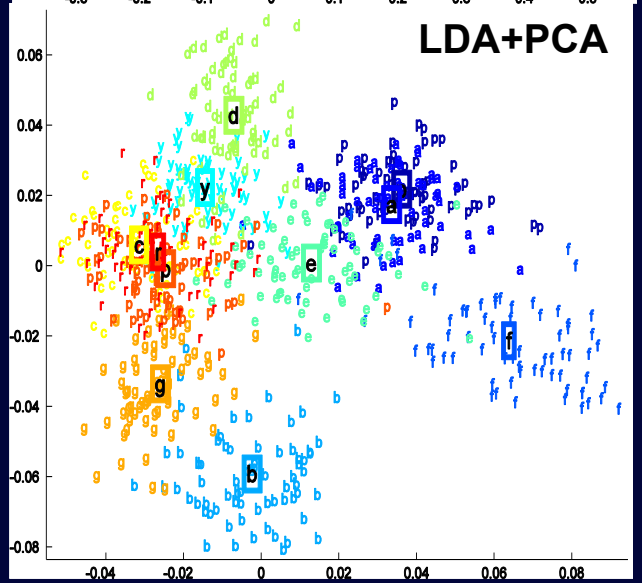
Small regularization → Large regularization

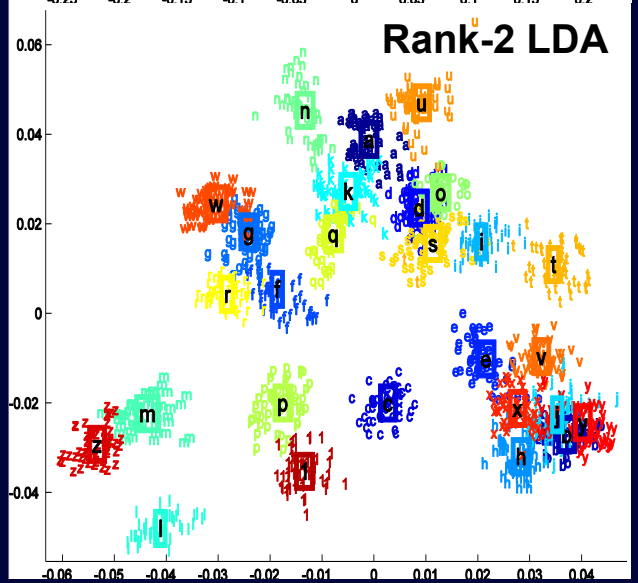
2D Visualization of Clustered Text, Image, Audio Data

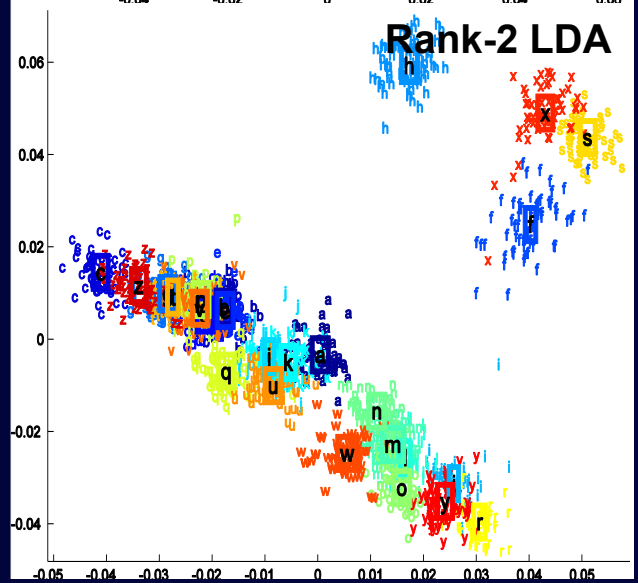












20news Data *(Text)*

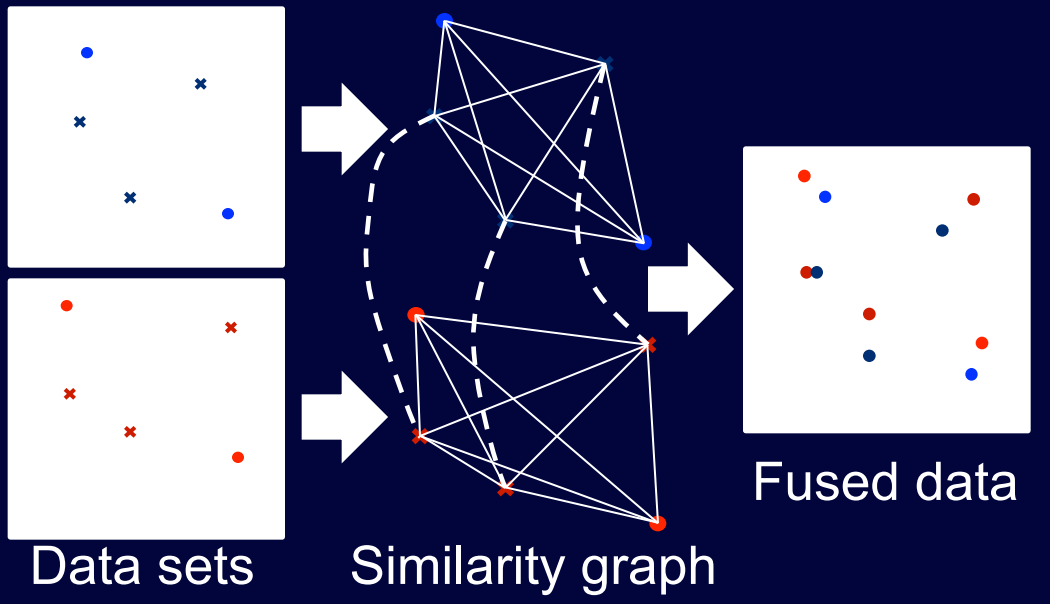
Facial Data *(Image)*

Spoken Letters *(Audio)*

Information Fusion and Visual Comparisons based on Space Alignment (J. Choo, S. Bohn, G. Nakamura, A. White, HP)

- Want: Unified visual representations of different results
- Assume: Reference correspondence information between data pairs or cluster correspondence
- Two conflicting criteria:
  maximize alignment and minimize deformation

- **Graph embedding approach (MDS)**



- **Procrustes analysis**

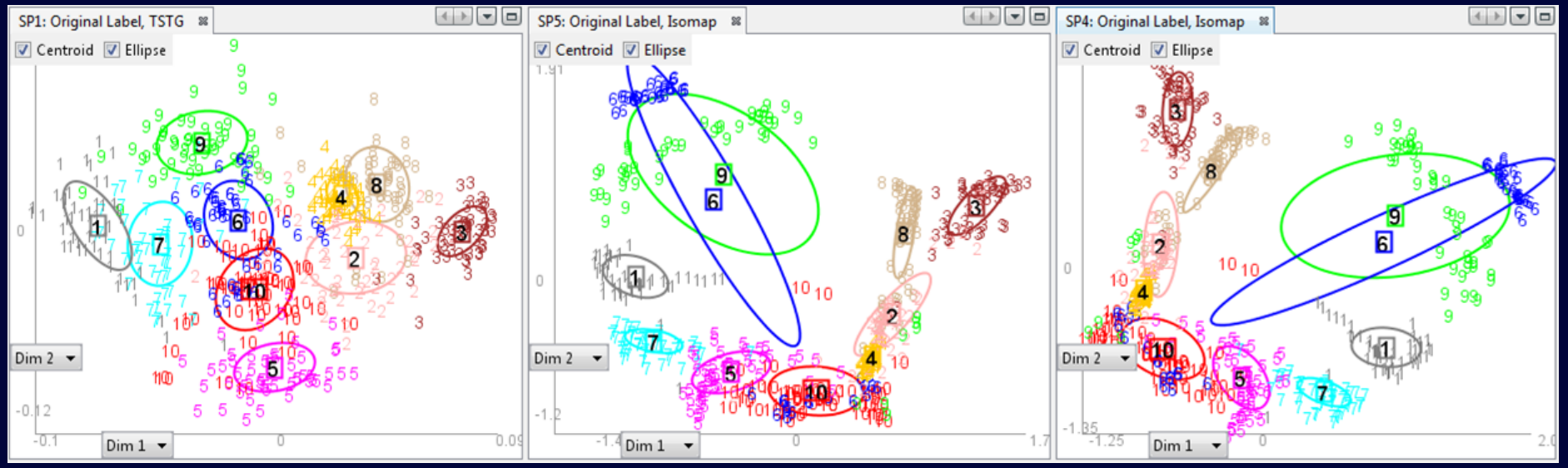
$$\min \| (A-\mu_A 1^T)-kQ(B-\mu_B 1^T) \|_F$$
$$Q^T Q = I$$

Space Alignment by Orthogonal Procrustes

$$\min \| (A-\mu_A 1^T)-kQ(B-\mu_B 1^T) \|_F \text{ , where } Q^T Q=I$$

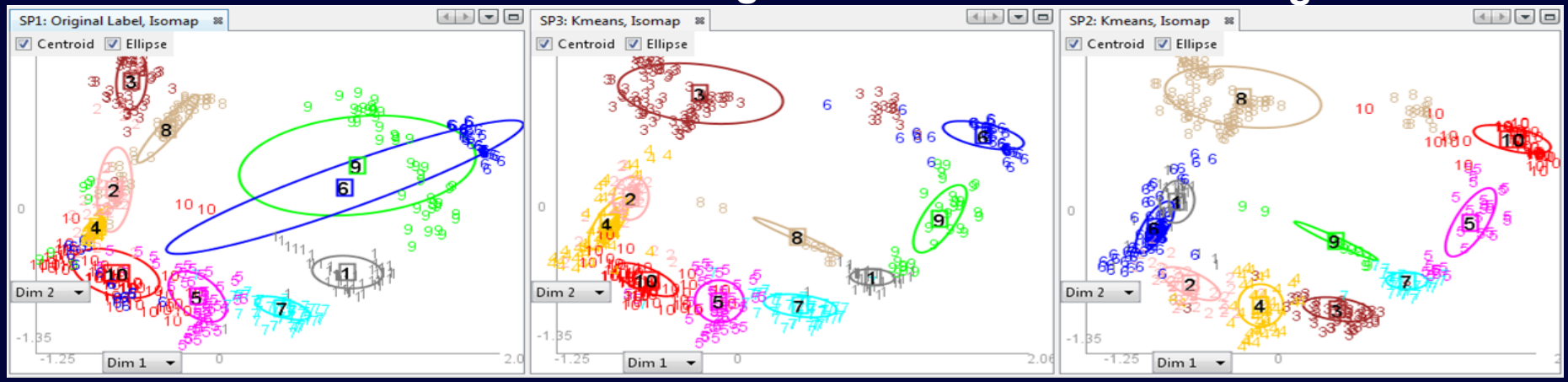
Alignment of Dimension Reduction Results

Reference **Aligned** **Un-Aligned**

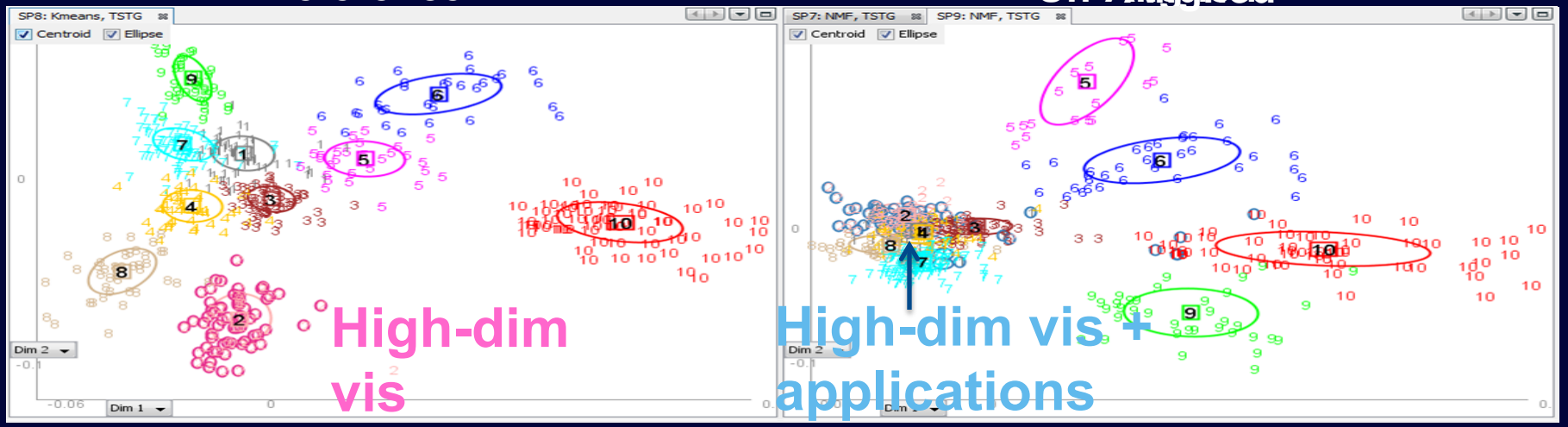


Cluster Alignment: Label Matching and Space Alignment

Reference | Aligned | Un-Aligned



Reference | Un-Aligned

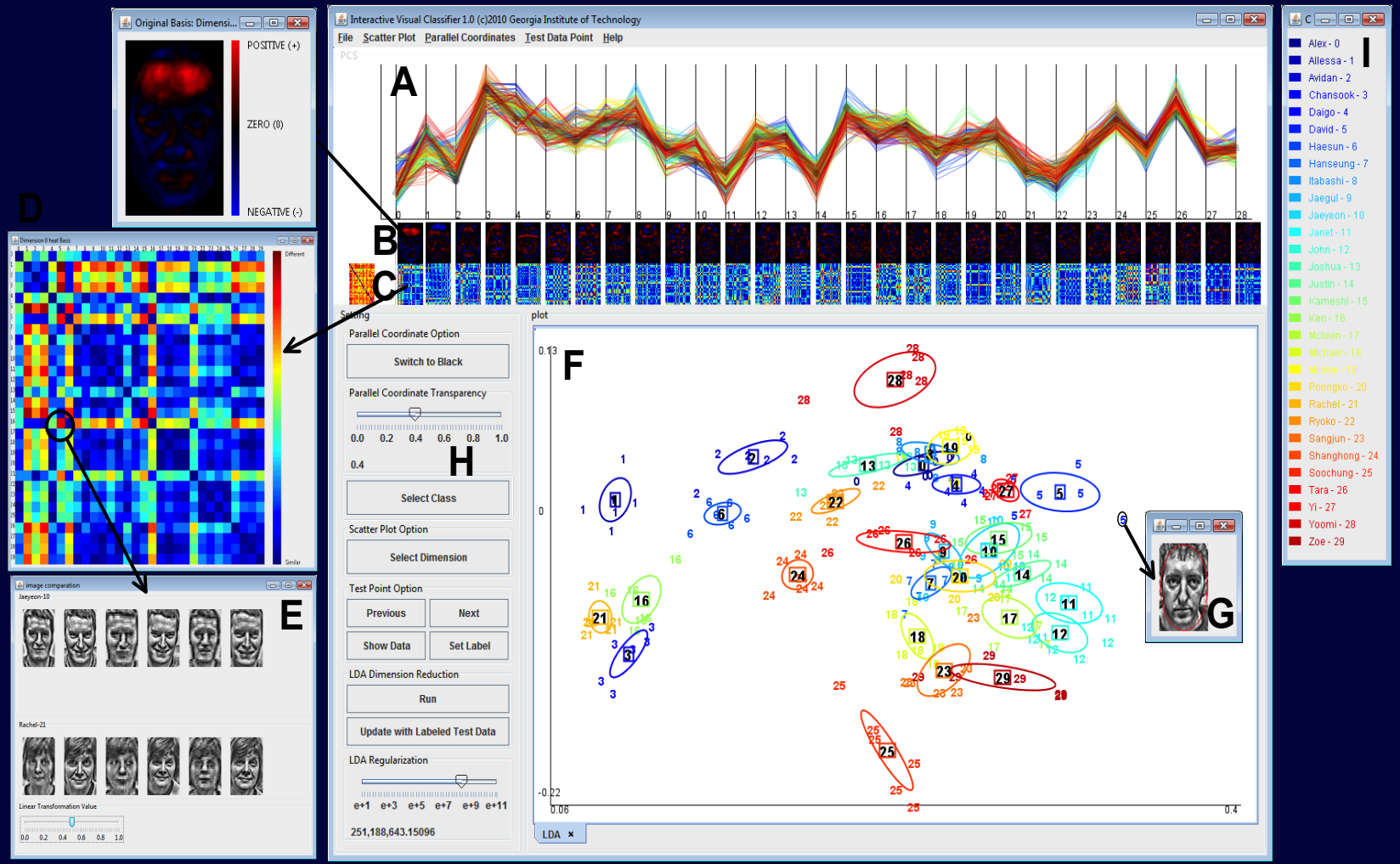


- InfoVis and VAST paper data set
- Help refine cluster results and obtain consensus clustering

iVisClassifier

(J. Choo, H. Lee, J. Kihm, HP, VAST10)

Interactive visual analytics system for **classification** of high-dim. data (image, text, etc) and **search space reduction**



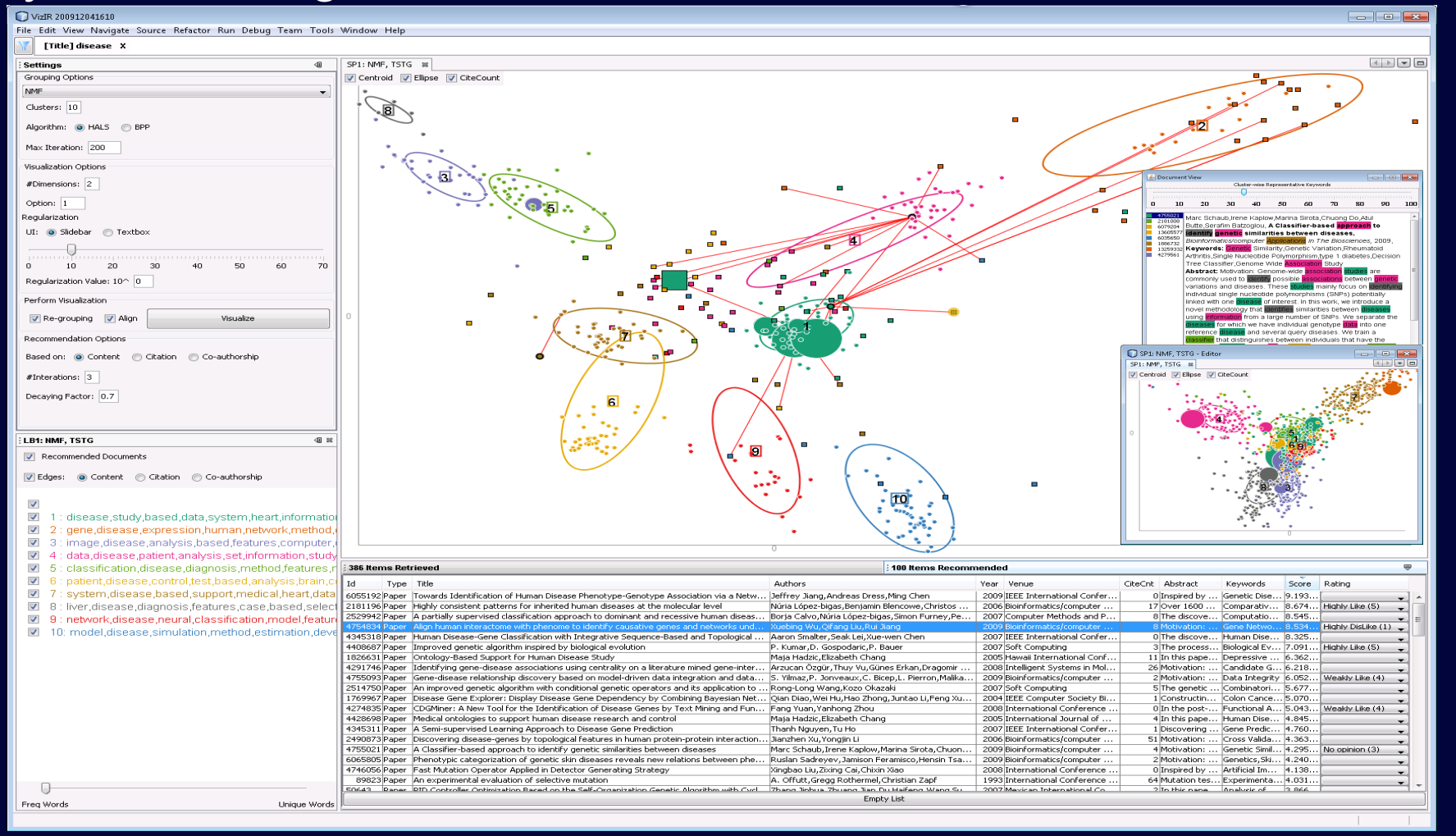
VisIRR: Visual Information Retrieval and Recommendation System for Document Discovery

Our differentiators:

- Improves personalization and understandability via integrated visualizations of document retrieval and recommendation
- Visual IR: beyond Google-like keyword search:
  - See **more** relevant documents
  - See **relationships**: topical, inter-document
  - Whole **content**-based, not keyword-based
- Visual Recommendation: enables discovery
  - **Personalized** based on user feedback, persistent
  - Understand **"why"** due to visualized relationships
- Only possible due to **new/fast ML** algorithms

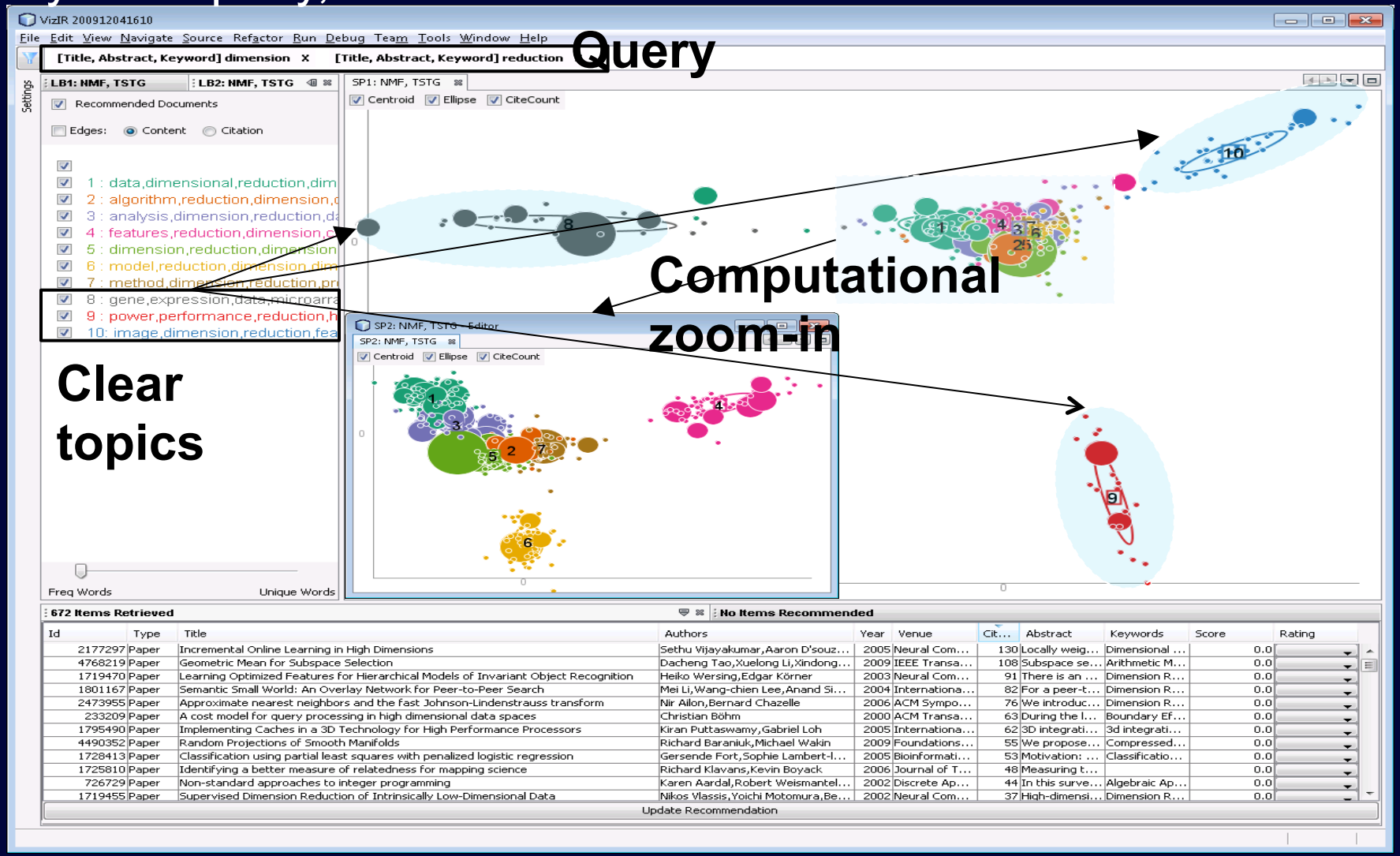
VisIRR

An interactive visual information retrieval and recommender system for large-scale document data



Visualization Example of Queried Set

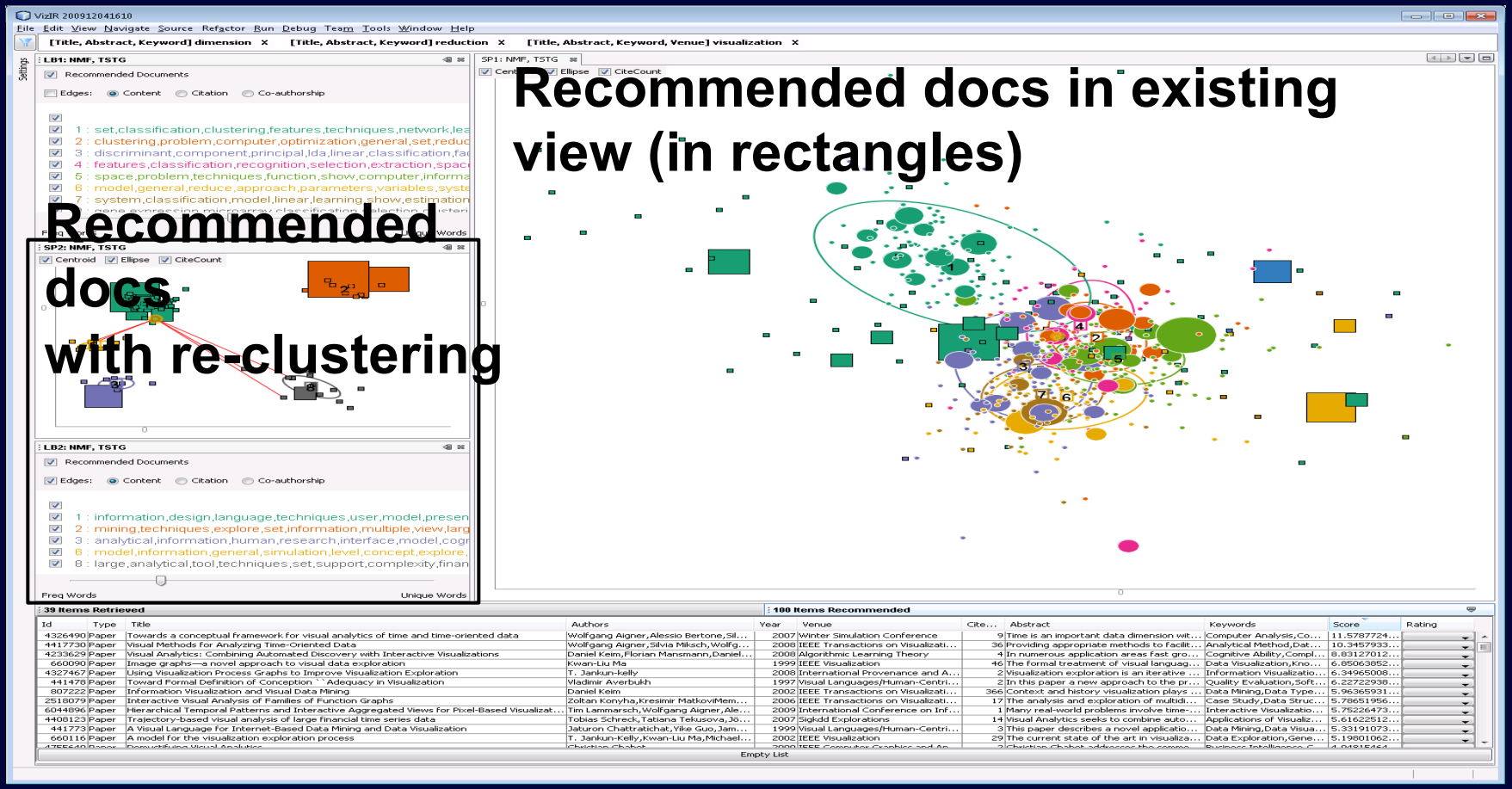
Keyword query, ‘dimension reduction’



Id	Type	Title	Authors	Year	Venue	Cit...	Abstract	Keywords	Score	Rating
2177297	Paper	Incremental Online Learning in High Dimensions	Sethu Vijayakumar,Aaron D'souz...	2005	Neural Com...	130	Locally weig...	Dimensional ...	0.0	
4768219	Paper	Geometric Mean for Subspace Selection	Dacheng Tao,Xuelong Li,Xindong...	2009	IEEE Transa...	108	Subspace se...	Arithmetic M...	0.0	
1719470	Paper	Learning Optimized Features for Hierarchical Models of Invariant Object Recognition	Heiko Wersing,Edgar Körner	2003	Neural Com...	91	There is an ...	Dimension R...	0.0	
1801167	Paper	Semantic Small World: An Overlay Network for Peer-to-Peer Search	Mei Li,Wang-chien Lee,Anand Si...	2004	Internationa...	82	For a peer-t...	Dimension R...	0.0	
2473955	Paper	Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform	Nir Ailon,Bernard Chazelle	2006	ACM Sympo...	76	We introduc...	Dimension R...	0.0	
233209	Paper	A cost model for query processing in high dimensional data spaces	Christian Böhm	2000	ACM Transa...	63	During the l...	Boundary Ef...	0.0	
1795490	Paper	Implementing Caches in a 3D Technology for High Performance Processors	Kiran Puttaswamy,Gabriel Loh	2005	Internationa...	62	3D integrati...	3d integrati...	0.0	
4490352	Paper	Random Projections of Smooth Manifolds	Richard Baraniuk,Michael Wakin	2009	Foundations...	55	We propose...	Compressed...	0.0	
1728413	Paper	Classification using partial least squares with penalized logistic regression	Gersende Fort,Sophie Lambert-l...	2005	Bioinformati...	53	Motivation: ...	Classificatio...	0.0	
1725810	Paper	Identifying a better measure of relatedness for mapping science	Richard Klavans,Kevin Boyack	2006	Journal of T...	48	Measuring t...		0.0	
726729	Paper	Non-standard approaches to integer programming	Karen Aardal,Robert Weismantel...	2002	Discrete Ap...	44	In this surve...	Algebraic Ap...	0.0	
1719455	Paper	Supervised Dimension Reduction of Intrinsically Low-Dimensional Data	Nikos Vlassis,Yoichi Motomura,Be...	2002	Neural Com...	37	High-dimensi...	Dimension R...	0.0	

Recommendation Example

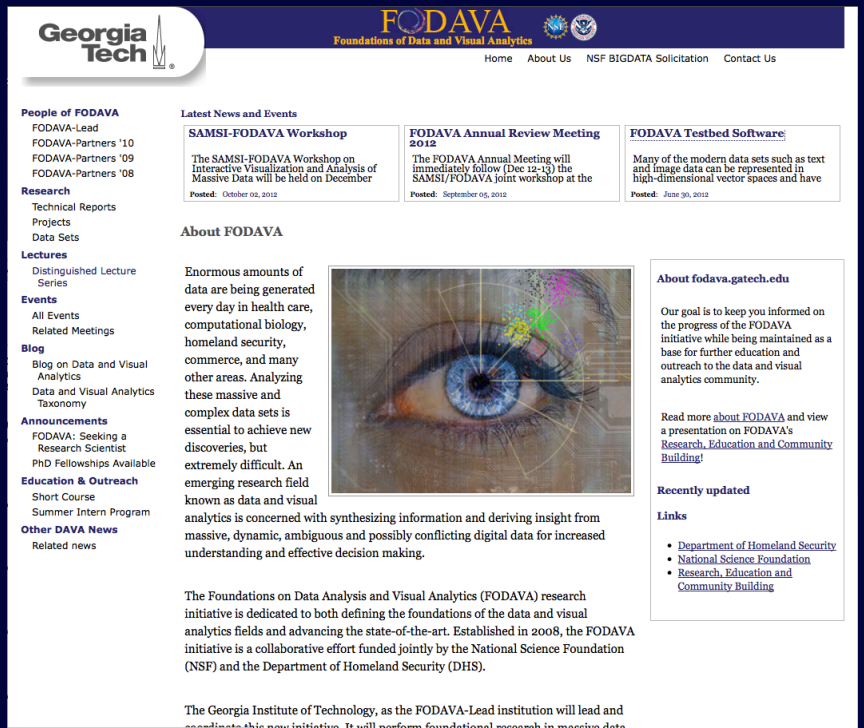
Preference-assigned item as ‘highly like’ :
‘Enhancing the visualization process with principal component analysis to support the exploration of trends’

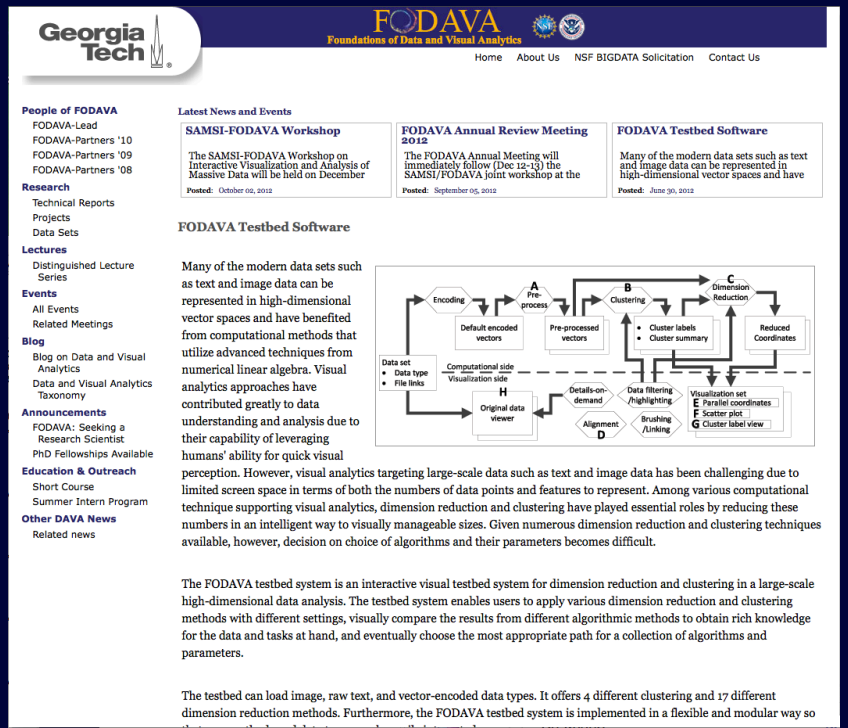


FODAVA Website

http://fodava.gatech.edu

- Dissemination of FODAVA results to user communities
  - FODAVA Tech Reports/Software
- FODAVA meeting/lecture materials
- Data Sets
- DAVA Taxonomy and course material
- DAVA community events and meeting information





Concluding Remarks

- Data and Visual Analytics is especially important for data understanding and question generation
- For Big data analytics, more integrated research that tie automated algorithms and interactive visualization needed

- Our Contributions:
  - Foundational algorithms for visual representations of high dimensional, large scale, heterogeneous data
  - Fast algorithms for real time interaction
  - Development of VA testbed and other VA systems

Thank you!