

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

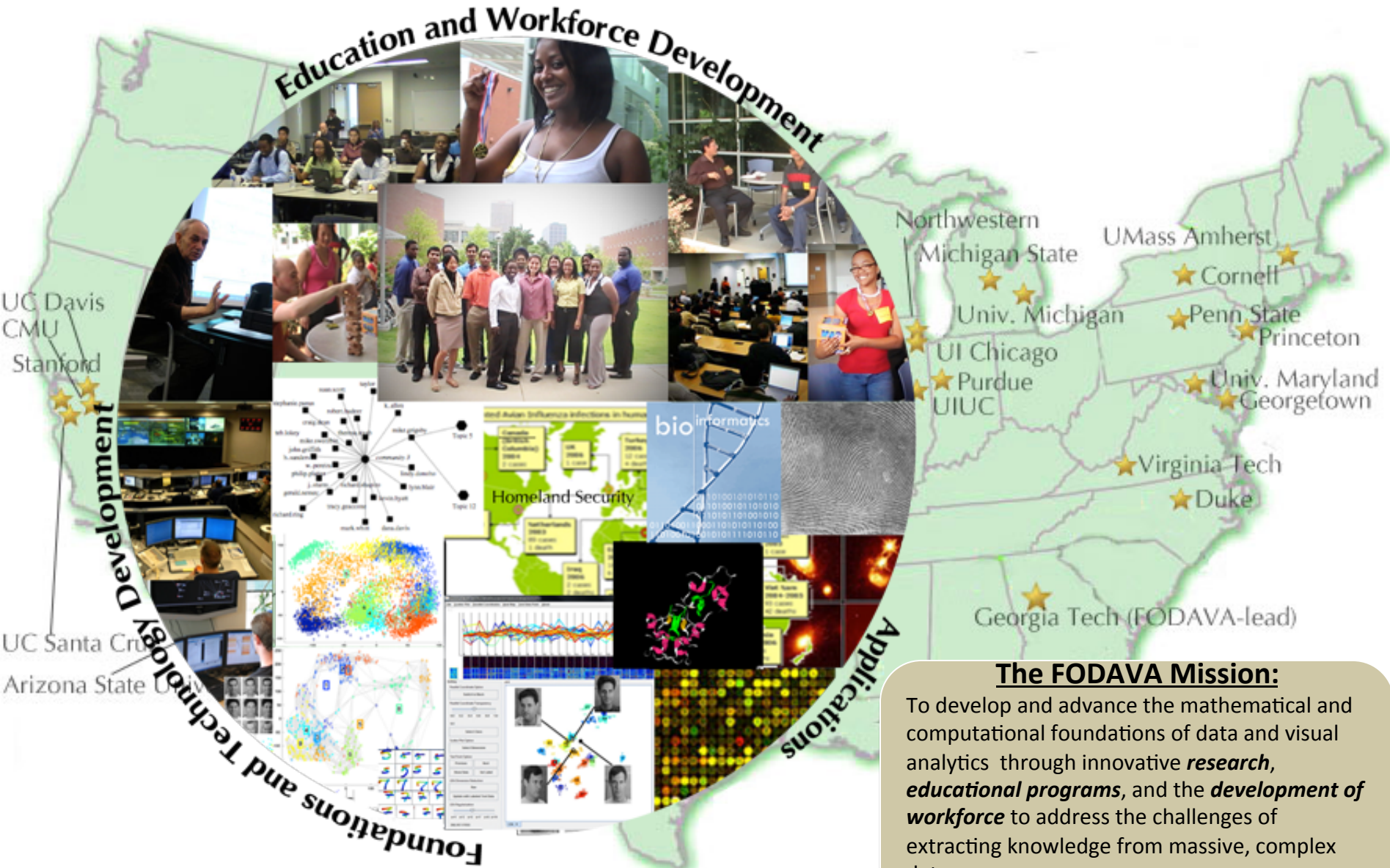
Haesun Park  
School of Computational Science and Engineering  
Georgia Institute of Technology  
Atlanta, GA, U.S.A.

CSol 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)

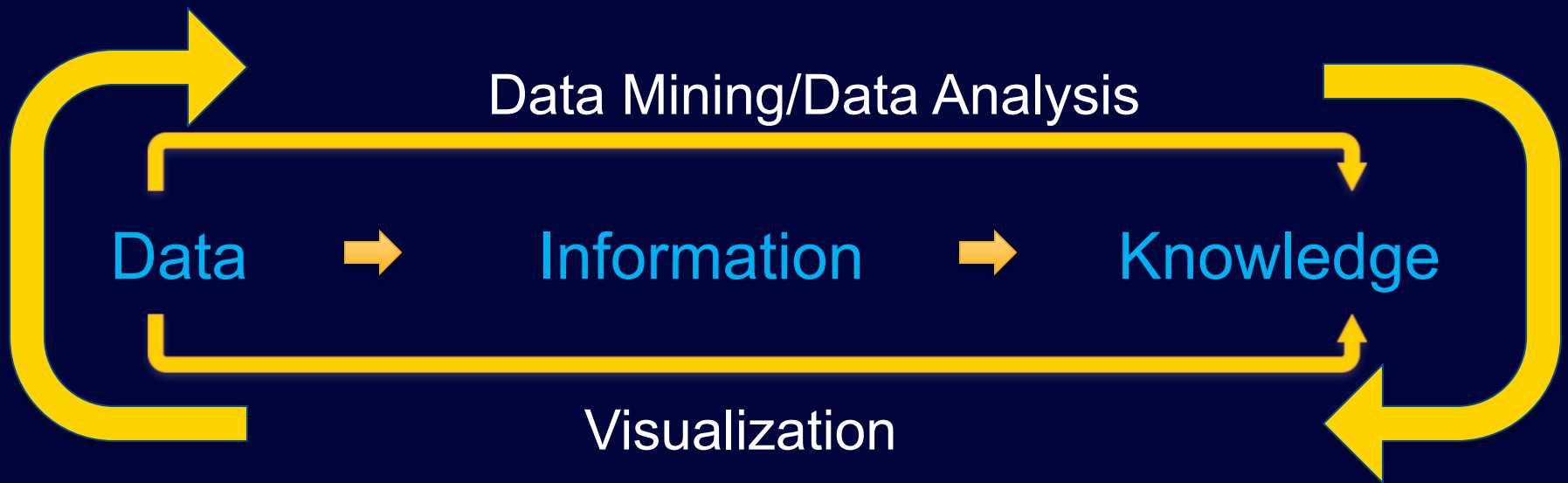


### The FODAVA Mission:

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

**Georgia Tech**

**FODAVA**  
Foundations of Data and Visual Analytics

Home About Us NSF BIGDATA Solicitation Contact Us

**People of FODAVA**  
FODAVA-Lead  
FODAVA-Partners '10  
FODAVA-Partners '09  
FODAVA-Partners '08

**Research**  
Technical Reports  
Projects  
Data Sets

**Lectures**  
Distinguished Lecture Series

**Events**  
SAMS-FODAVA Workshop  
FODAVA Annual Review Meeting 2012  
All Events  
Related Meetings

**Blog**  
Blog on Data and Visual Analytics  
Data and Visual Analytics Taxonomy

**Announcements**  
FODAVA: Seeking a Research Scientist  
PhD Fellowships Available

**Education & Outreach**  
Short Course  
Summer Intern Program

**Other DAVA News**

**Latest News and Events**

**SAMS-FODAVA Workshop**  
The SAMS-FODAVA Workshop on Interactive Visualization and Analysis of Massive Data will be held on  
Posted: October 01, 2012

**FODAVA Annual Review Meeting 2012**  
The FODAVA Annual Meeting will immediately follow (Dec 12-13) the SAMS-FODAVA's joint workshop at the  
Posted: September 03, 2012

**FODAVA Testbed Software**  
Many of the modern data sets such as text and image data can be represented in high-dimensional vector spaces and  
Posted: June 30, 2012

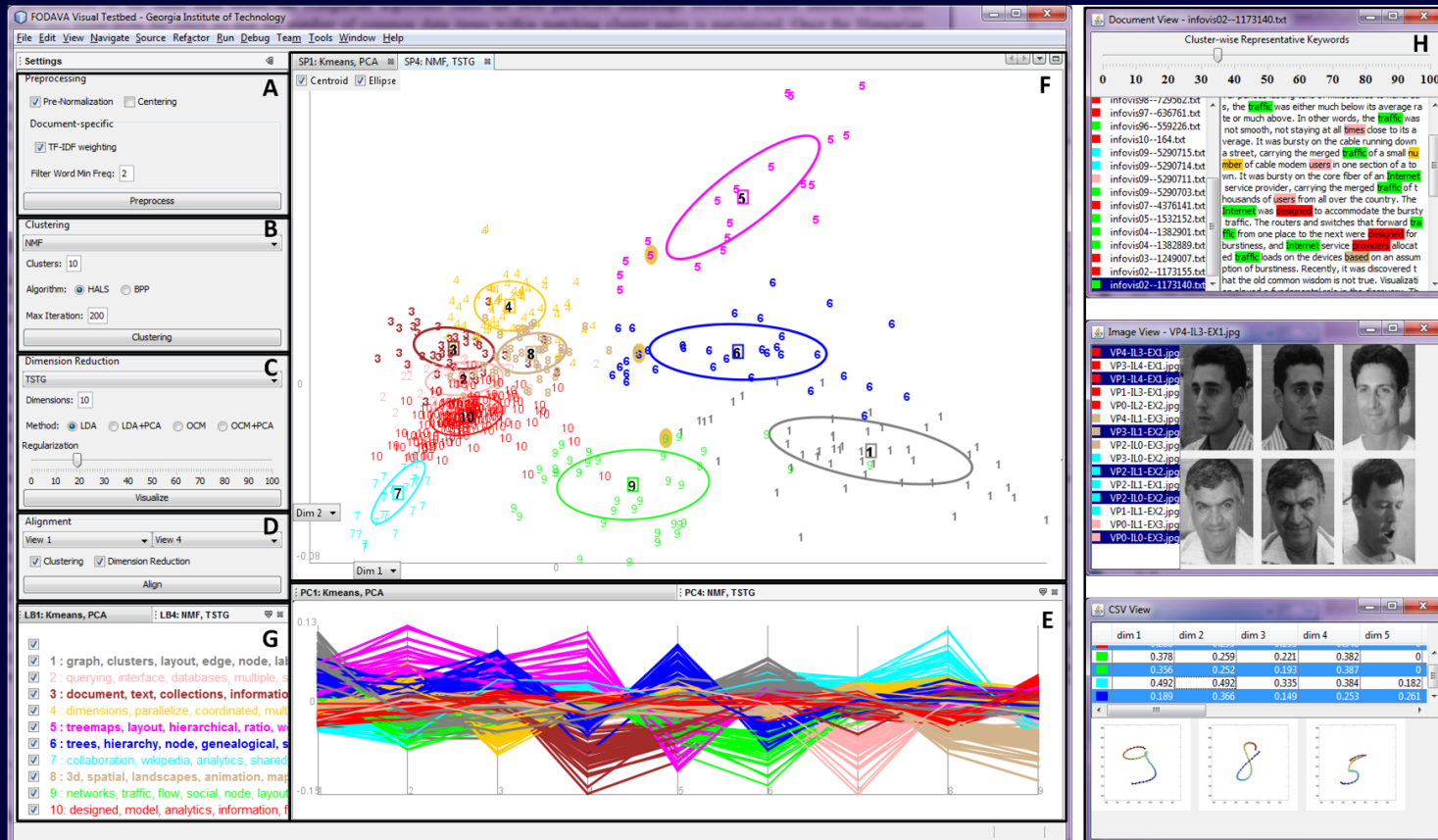
**FODAVA Testbed Software**

Many of the modern data sets such as text and image data can be represented in high-dimensional vector spaces and have benefited from computational methods that utilize advanced techniques from numerical linear algebra. Visual analytics approaches have contributed greatly to data understanding and analysis due to their capability of leveraging humans' ability for quick visual perception. However, visual analytics targeting large-scale data such as text and image data has been challenging due to limited screen space in terms of both the numbers of data points and features to represent. Among various computational techniques supporting visual analytics, dimension reduction and clustering have played essential roles by reducing these numbers in an intelligent way to visually manageable sizes. Given numerous dimension reduction and clustering techniques available, however, decision on choice of algorithms and their parameters becomes difficult.

**Flowchart Description:**  
The flowchart illustrates the data processing pipeline. It starts with a 'Data set' (Data type, File links) which goes to 'Encoding' (Default encoded vectors). This leads to 'Pre-process' (Pre-processed vectors). From there, it branches into 'Clustering' (Cluster labels, Cluster summary) and 'Dimension Reduction' (Reduced Coordinates). The 'Clustering' and 'Dimension Reduction' steps feed into 'Visualization side' components: 'Original data viewer', 'Details-on-demand', 'Data filtering /highlighting', and 'Alignment'. The 'Visualization side' also includes 'Visualizer set' (Parallel coordinates, Scatter plot, Cluster label view). The 'Computational side' includes 'Encoding', 'Pre-process', 'Clustering', and 'Dimension Reduction'.

# 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

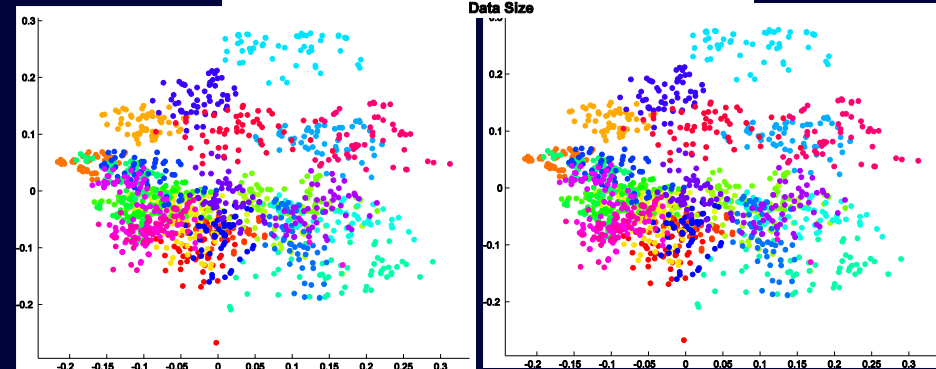
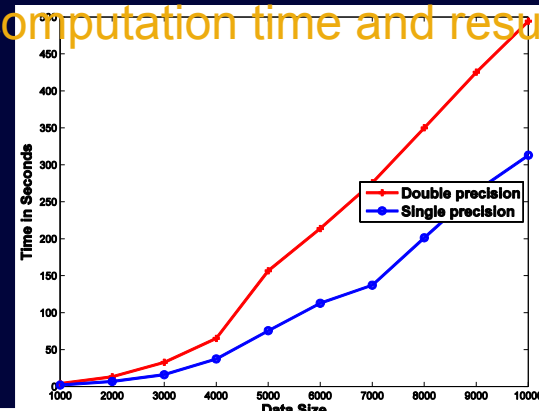




# Fast Comp. Modules for Interactive Vis.

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

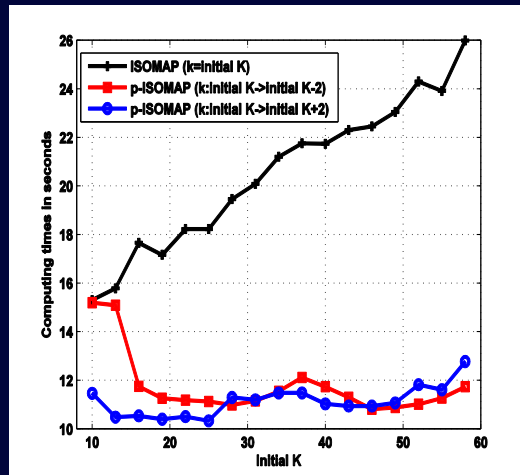
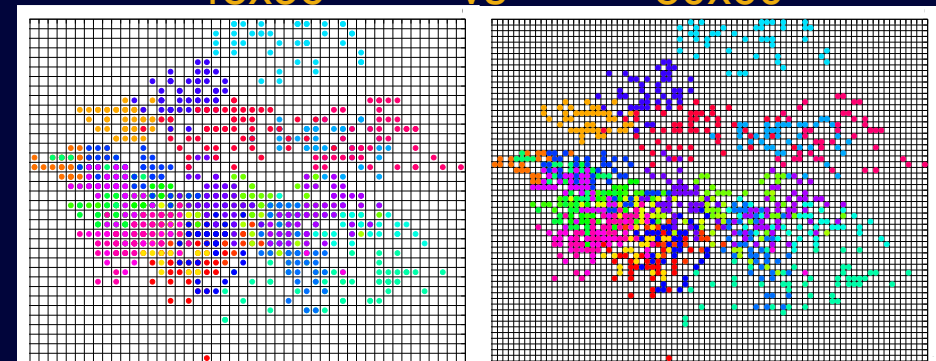
PCA timing: double vs single precision computation time and results



48x36

vs

80x60



p-Isomap computing time vs. # of nearest neighbors

## • 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, ...)

$$A \approx WH$$
$$\rightarrow \min \|A - WH\|_F$$
$$W \geq 0, H \geq 0$$

- 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  $\min_{H \geq 0} \|WH - A\|_F$
  - fixing  $H$ , solve  $\min_{W \geq 0} \|H^T W^T - A^T\|_F$
- \* Software available at [www.cc.gatech.edu/~hpark](http://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.
  - $\min \sum_{1 \leq i \leq n} \| a_i - w_{\sigma_i} \|^2 \rightarrow \min \| A - WH \|_F^2$   
 $\sigma_i = j$  when  $i$ -th point is assigned to  $j$ -th cluster ( $j \in \{1, \dots, k\}$ )

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)

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

- Obj. fun. of K-means and NMF are related when  $H \in E$  and  $A \geq 0$ , 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

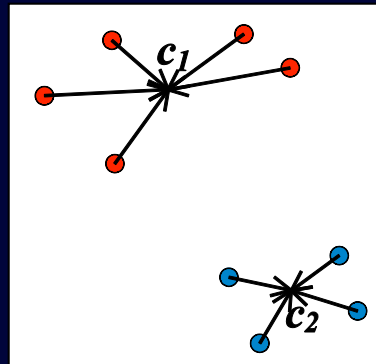
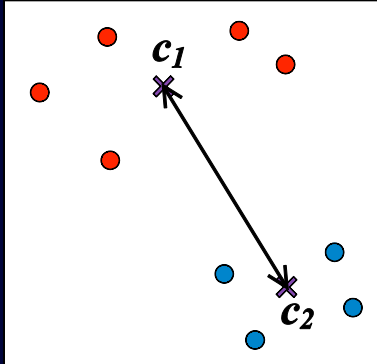
# Linear Discriminant Analysis for 2D/3D Representation of Clustered Data

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

Max trace ( $G^T S_b G$ )

&

min trace ( $G^T S_w G$ )



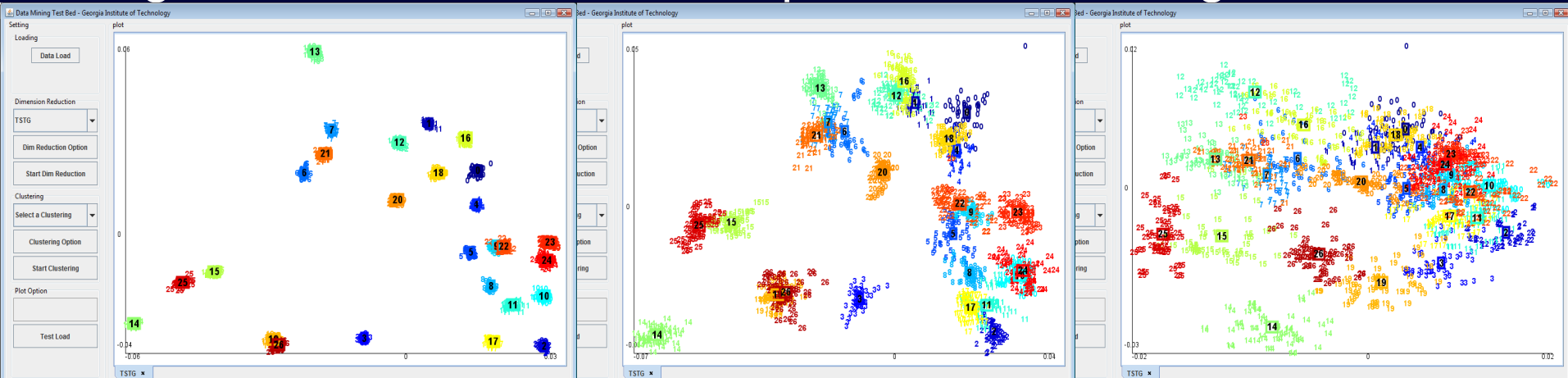
**LDA/GSVD**

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

max trace

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

- Regularization in LDA for Computational Zooming-in

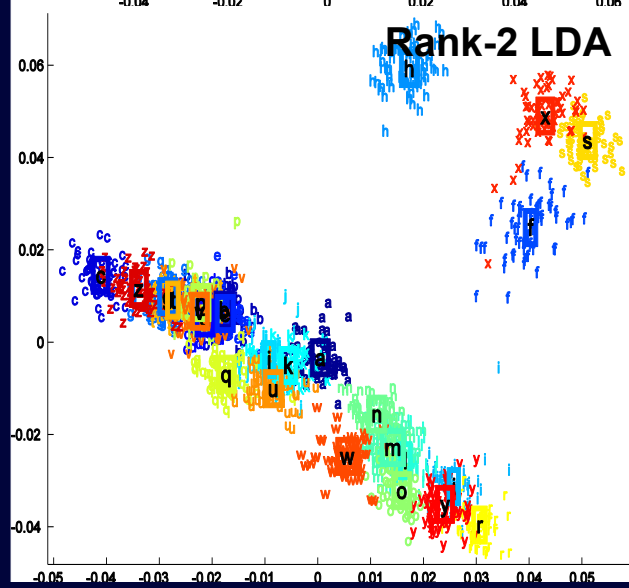
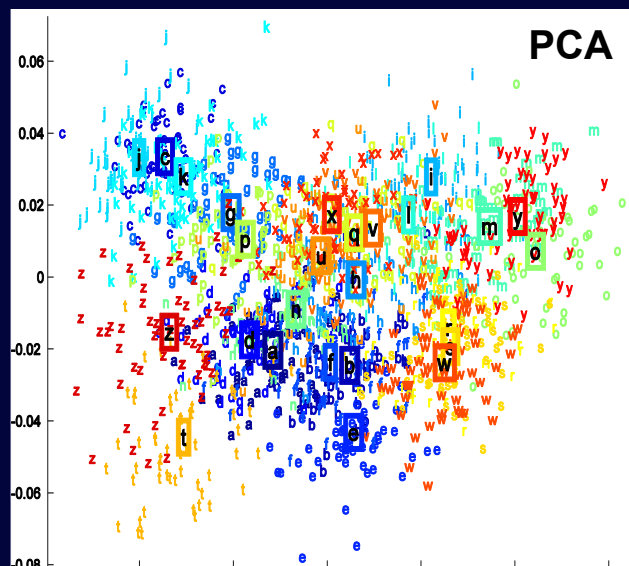
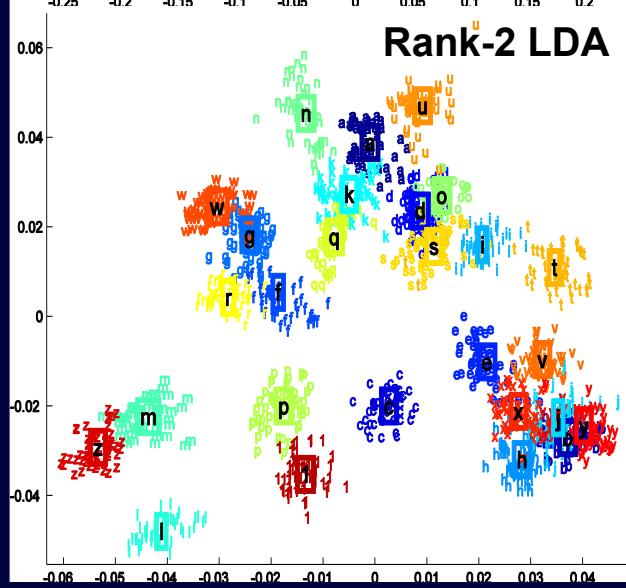
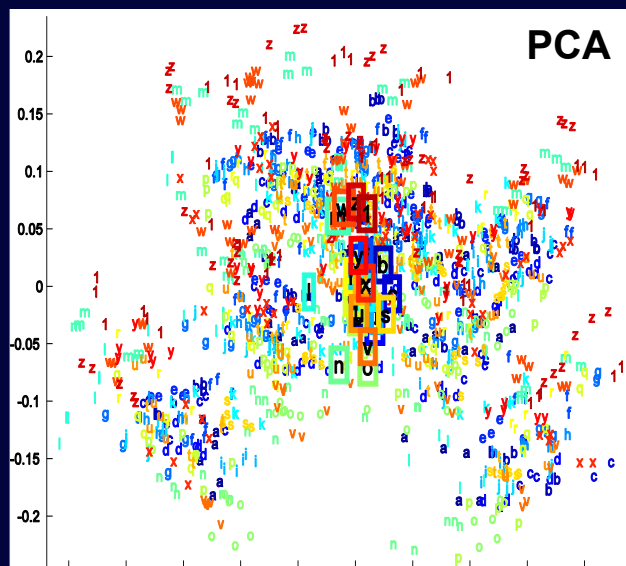
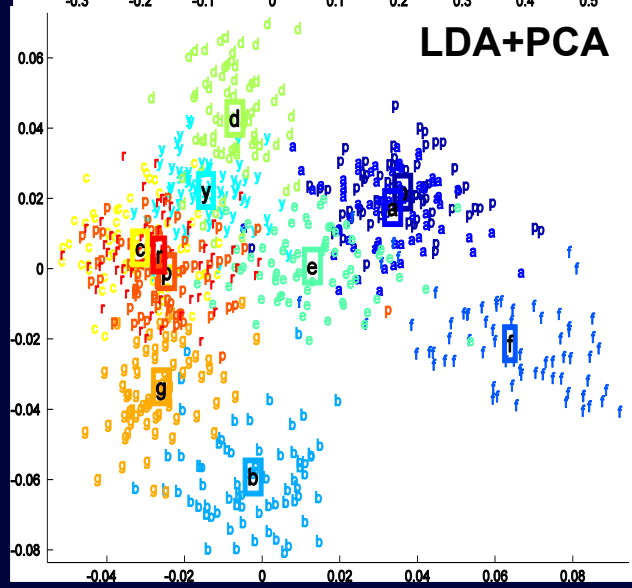
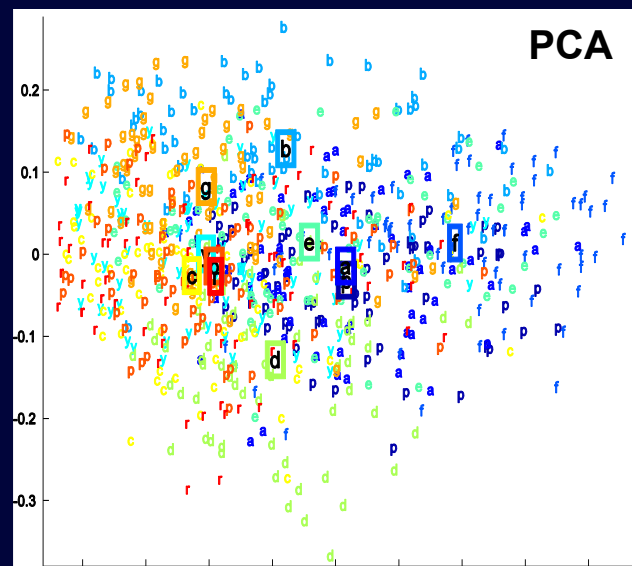


Small regularization



Large regularization

# 2D Visualization of Clustered Text, Image, Audio Data



20news Data (Text)

Facial Data (Image)

Spoken Letters (Audio)

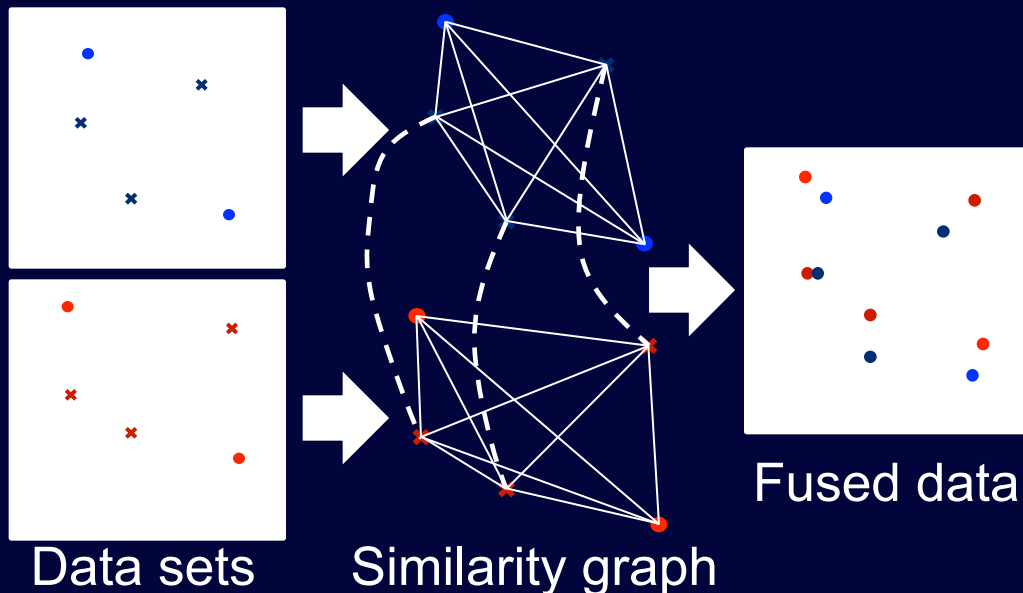


# Information Fusion and Visual Comparisons based on Space Alignment

(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**

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

# Space Alignment by Orthogonal Procrustes

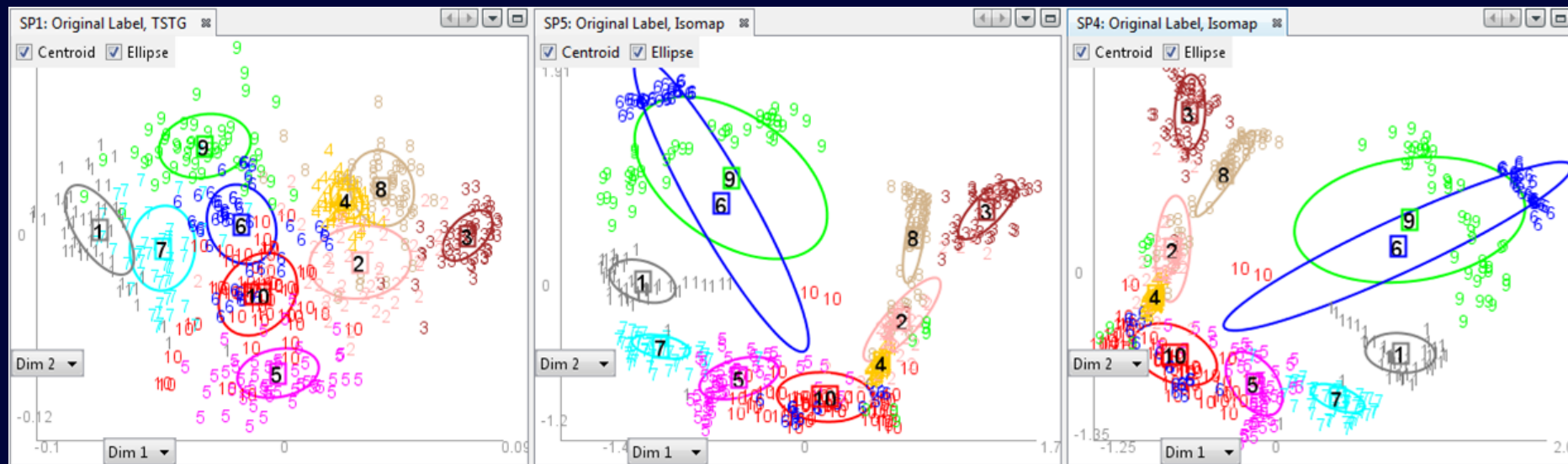
$$\min \| (A - \mu_A \mathbf{1}^T) - kQ(B - \mu_B \mathbf{1}^T) \|_F, \text{ where } Q^T Q = I$$

## Alignment of Dimension Reduction Results

Reference

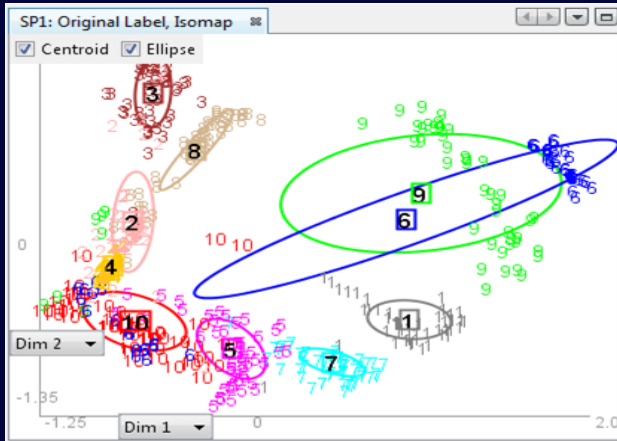
Aligned

Un-Aligned

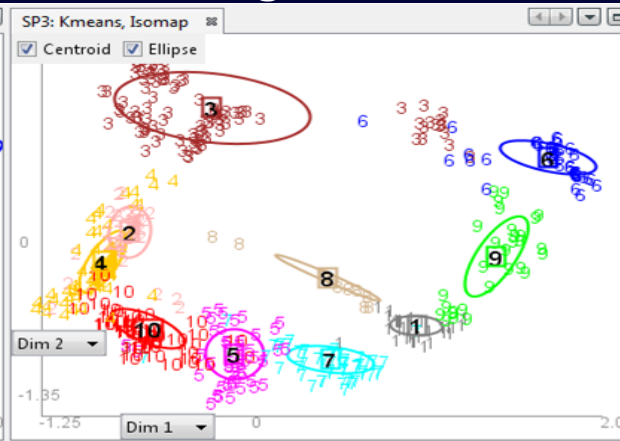


# Cluster Alignment: Label Matching and Space Alignment

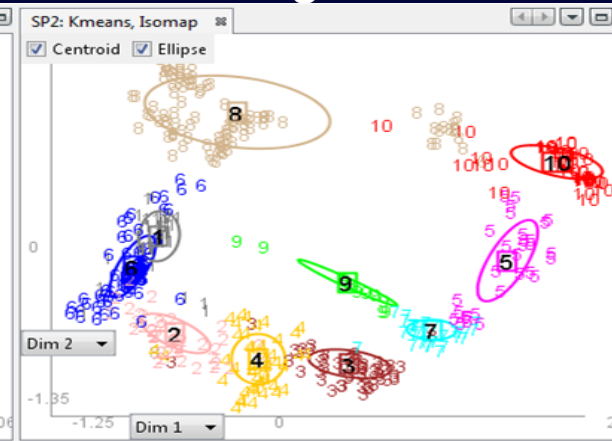
Reference



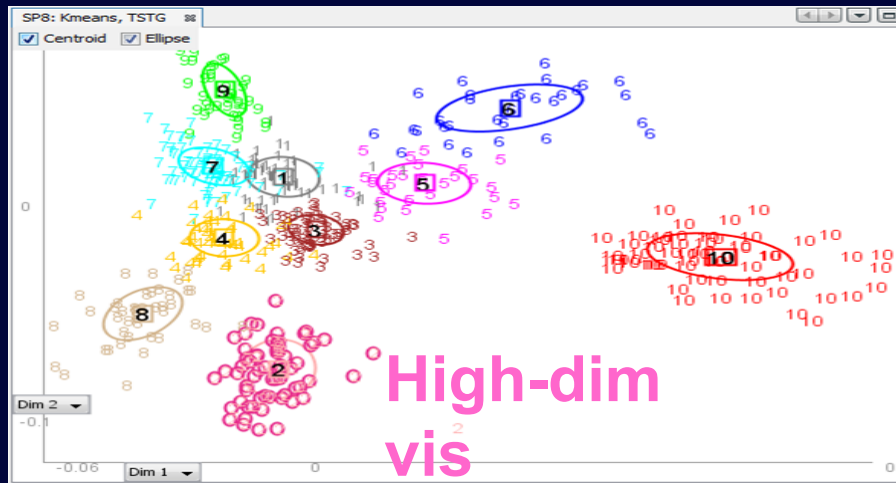
Aligned



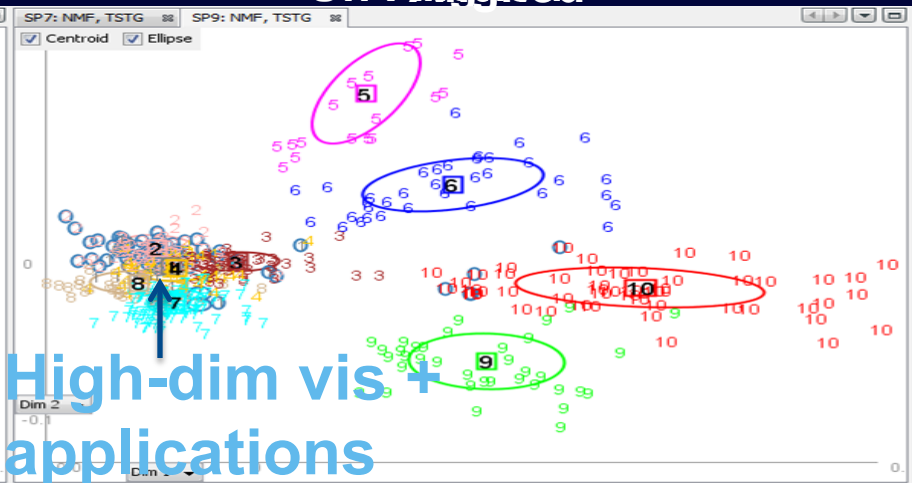
Un-Aligned



Reference



Un-Aligned

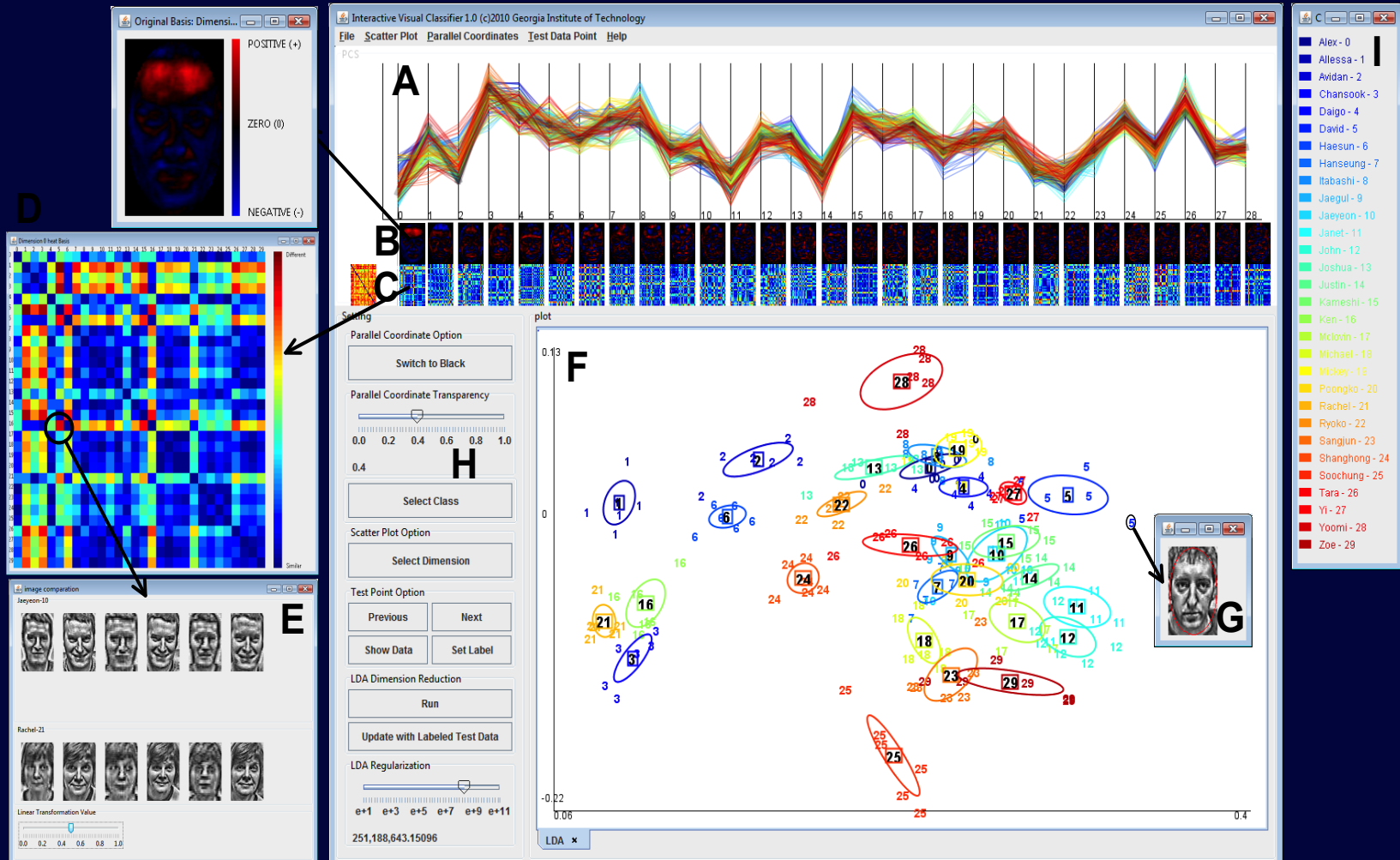


- InfoVis and VAST paper data set
- Help refine cluster results and obtain consensus clustering

# 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

The screenshot displays the VisIRR application interface. At the top, the title bar shows 'VisIR 200912041610'. The main window is titled '[Title] disease x'. The interface is divided into several sections:

- Settings Panel (Left):** Includes 'Grouping Options' (NMF, Clusters: 10, Algorithm: HALS/BPP, Max Iteration: 200), 'Visualization Options' (#Dimensions: 2, Option: 1), 'Regularization' (UI: Slidebar/Textbox, Regularization Value:  $10^{-0}$ ), 'Perform Visualization' (Re-grouping, Align, Visualize), and 'Recommendation Options' (Based on: Content/Citation/Co-authorship, #Interactions: 3, Decaying Factor: 0.7).
- Main Visualization:** A network graph with nodes and edges, showing clusters of documents. Nodes are color-coded and numbered (1-10). Edges represent relationships between documents.
- Document List (Bottom Left):** A table of retrieved documents with columns for ID, Type, Title, Authors, Year, and Venue. It lists 386 items.
- Document List (Bottom Right):** A table of recommended documents with columns for CiteCnt, Abstract, Keywords, Score, and Rating. It lists 100 items.
- Document View (Right):** A detailed view of a selected document, showing its abstract, keywords, and a smaller version of the network visualization.

The document list includes the following entries:

ID	Type	Title	Authors	Year	Venue	CiteCnt	Abstract	Keywords	Score	Rating
6055192	Paper	Towards Identification of Human Disease Phenotype-Genotype Association via a Netw...	Jeffrey Jiang, Andreas Dress, Ming Chen	2009	IEEE International Confer...	0	Inspired by ...	Genetic Dise...	9,190...	
2181196	Paper	Highly consistent patterns for inherited human diseases at the molecular level	Núria López-bigas, Benjamin Blencowe, Christos ...	2006	Bioinformatics/computer ...	17	Over 1600 ...	Comparativ...	8,674...	Highly Like (5)
2529942	Paper	A partially supervised classification approach to dominant and recessive human diseas...	Borja Calvo, Núria López-bigas, Simon Furney, Pe...	2007	Computer Methods and P...	8	The discove...	Computatio...	8,545...	
4754834	Paper	Align human interactome with phenome to identify causative genes and networks und...	Xuebing Wu, Qifang Liu, Rui Jiang	2009	Bioinformatics/computer ...	8	Motivation...	Gene Netwo...	8,534...	Highly DisLike (1)
4345318	Paper	Human Disease-Gene Classification with Integrative Sequence-Based and Topological ...	Aaron Smalzer, Seok Leu, Xue-wen Chen	2007	IEEE International Confer...	0	The discove...	Human Dise...	8,325...	
4408687	Paper	Improved genetic algorithm inspired by biological evolution	P. Kumar, D. Gospodaric, P. Bauer	2007	Soft Computing	3	The proces...	Biological Ev...	7,091...	
1826631	Paper	Ontology-Based Support for Human Disease Study	Maja Hadzic, Elizabeth Chang	2005	Hawaii International Conf...	11	In this paper...	Depressive ...	6,362...	
4291746	Paper	Identifying gene-disease associations using centrality on a literature mined gene-inter...	Arzuhan Ozgur, Thuy Vu, Günes Erkan, Dragomir ...	2008	Intelligent Systems in Mol...	26	Motivation...	Candidate G...	6,218...	
4755093	Paper	Gene-disease relationship discovery based on model-driven data integration and data...	S. Vilmaz, P. Jonveaux, C. Bicep, L. Pierron, Malka...	2009	Bioinformatics/computer ...	2	Inspired by ...	Data Integrity	6,052...	Weakly Like (4)
2514750	Paper	An improved genetic algorithm with conditional genetic operators and its application to ...	Rong Long Wang, Kozo Okazaki	2007	Soft Computing	5	The genetic ...	Combinator...	5,677...	
1769967	Paper	Disease Gene Explorer: Display Disease Gene Dependency by Combining Bayesian Net...	Qian Diao, Wei Hu, Hao Zhong, Juntao Li, Feng Xu...	2004	IEEE Computer Society Bi...	1	Constructio...	Colon Canc...	5,070...	
4274835	Paper	CDGMiner: A New Tool for the Identification of Disease Genes by Text Mining and Fun...	Fang Yuan, Yanhong Zhou	2008	International Conference ...	0	In the post...	Functional A...	5,043...	Weakly Like (4)
4286998	Paper	Medical ontologies to support human disease research and control	Maja Hadzic, Elizabeth Chang	2005	International Journal of ...	4	In this paper...	Human Dise...	4,845...	
4345311	Paper	A Semi-supervised Learning Approach to Disease Gene Prediction	Thanh Nguyen, Tu Ho	2007	IEEE International Confer...	1	Discovering ...	Gene Predic...	4,760...	
2490873	Paper	Discovering disease-genes by topological features in human protein-protein interaction...	Jianzhen Xu, Yongjin Li	2006	Bioinformatics/computer ...	51	Motivation...	Cross Valid...	4,363...	
4755021	Paper	A Classifier-based approach to identify genetic similarities between diseases	Marc Schaub, Irene Kaplow, Marina Sirota, Choon...	2009	Bioinformatics/computer ...	4	Motivation...	Genetic Simil...	4,295...	No opinion (3)
6058005	Paper	Phenotypic categorization of genetic skin diseases reveals new relations between phe...	Ruslan Sadreyev, Jamison Ferrasico, Hensin Tsa...	2009	Bioinformatics/computer ...	2	Motivation...	Genetics, SKI...	4,240...	
4746056	Paper	Fast Mutation Operator Applied in Detector Generating Strategy	Xingbao Liu, Zixing Cai, Chixin Xiao	2008	International Conference ...	0	Inspired by ...	Artificial Im...	4,136...	
69623	Paper	An experimental evaluation of selective mutation	A. Offutt, Gregg Rothmel, Christian Zapf	1993	International Conference ...	64	Mutation tes...	Experimenta...	4,031...	
52663	Paper	SDS Control: Optimization Based on the Self-Organization Genetic Algorithm with Cyl...	Zhenan Jinhua, Zhuoan Jiang, Du Haiyan, Liang Su...	2002	Medical, Informational Co...	2	In this, name...	Analysis of ...	3,064...	

# Visualization Example of Queried Set

Keyword query, 'dimension reduction'

The screenshot displays the VizIR interface with a dimensionality reduction visualization. The main window shows a scatter plot of data points, with several clusters highlighted by light blue ellipses and labeled with numbers 1 through 10. A zoomed-in view of cluster 4 is shown in a separate window titled 'SP2: NMF, TSTG - Editor'. The zoomed-in view shows a dense cluster of points, with a red ellipse highlighting a sub-cluster containing points 9 and 10. The interface includes a menu bar, a toolbar, and a list of retrieved items at the bottom.

**Query**

**Computational zoom-in**

**Clear topics**

**672 Items Retrieved**

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'

The screenshot displays the VizIR 200912041610 interface. The main window shows a network visualization of documents, with nodes represented by colored circles and squares. A large green circle is highlighted, and a red rectangle is drawn around it. The text "Recommended docs in existing view (in rectangles)" is overlaid on this area. On the left, there are two panels for "Recommended Documents" (LB1 and LB2) with checkboxes for "Edges" (Content, Citation, Co-authorship) and a list of keywords. The bottom panel shows a list of 100 recommended items with columns for Id, Type, Title, Authors, Year, Venue, Cite..., Abstract, Keywords, Score, and Rating.

**Recommended docs in existing view (in rectangles)**

**Recommended docs with re-clustering**




Id	Type	Title	Authors	Year	Venue	Cite...	Abstract	Keywords	Score	Rating
4326490	Paper	Towards a conceptual Framework for visual analytics of time and time-oriented data	Wolfgang Aigner, Alessio Bertone, Sil...	2007	Winter Simulation Conference	9	Time is an important data dimension wit...	Computer Analysis,Co...	11.578724...	
4117730	Paper	Visual Methods for Analyzing Time-Oriented Data	Wolfgang Aigner, Silvia Ribsch, Wolf...	2008	IEEE Transactions on Visualizati...	36	Providing appropriate methods to Facilit...	Analytical Method, Dat...	10.3495793...	
4233629	Paper	Visual Analytics: Combining Automated Discovery with Interactive Visualizations	Daniel Keim, Florian Mansmann, Daniel...	2008	Algorithmic Learning Theory	4	In numerous application areas fast gro...	Cognitive Ability, Compl...	6.83127012...	
660090	Paper	Image graphs—a novel approach to visual data exploration	Kwan-Liu Ma	1999	IEEE Visualization	46	The Formal treatment of visual languag...	Data Visualization, Kno...	6.85063052...	
4327467	Paper	Using Visualization Process Graphs to Improve Visualization Exploration	T. Jankun-Kelly	2008	International Provenance and A...	2	Visualization exploration is an iterative ...	Information Visualizati...	6.34960008...	
441478	Paper	Toward Formal Definition of Conception "Adequacy in Visualization	Vladimir Averbukh	1997	Visual Languages/Human-Centri...	2	In this paper a new approach to the pr...	Quality Evaluation, Soft...	6.22722938...	
807222	Paper	Information Visualization and Visual Data Mining	Daniel Keim	2006	IEEE Transactions on Visualizati...	366	Context and history visualization plays ...	Data Mining, Data Type...	5.96365931...	
2518079	Paper	Interactive Visual Analysis of Families of Function Graphs	Zoltan Konyha, Kresimir Matkovic/Mem...	2006	IEEE Transactions on Visualizati...	17	The analysis and exploration of multidi...	Case Study, Data Struc...	5.78651956...	
6044896	Paper	Hierarchical Temporal Patterns and Interactive Aggregated Views for Pixel-Based Visualizat...	Tim Lammarisch, Wolfgang Aigner, Ale...	2009	International Conference on Inf...	1	Many real-world problems involve time...	Interactive Visualizatio...	5.75226473...	
4408123	Paper	Trajectory-based visual analysis of large financial time series data	Tobias Schreck, Tatiana Telusova, Jo...	2009	SigKDD Explorations	14	Visual Analytics seeks to combine auto...	Applications of Visualiz...	5.61622512...	
441773	Paper	A Visual Language for Internet-Based Data Mining and Data Visualization	Jatunon Chattratchai, Yike Guo, Jam...	1999	Visual Languages/Human-Centri...	3	This paper describes a novel applicatio...	Data Mining, Data Visua...	5.33191073...	
660116	Paper	A model for the visualization exploration process	T. Jankun-Kelly, Kwan-Liu Ma, Michael...	2002	IEEE Visualization	29	The current state of the art in visualiza...	Data Exploration, Gene...	5.19801062...	
4786440	Paper	Compositing Visual Analytics	Justin Chabot	2000	IEEE Computer Graphics and An...	2	Chabot, Chabot, address the screen...	Business Intelligence, C...	4.04816464...	



# 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



Home About Us NSF BIGDATA Solicitation Contact Us

**People of FODAVA**  
FODAVA-Lead  
FODAVA-Partners '10  
FODAVA-Partners '09  
FODAVA-Partners '08

**Research**  
Technical Reports  
Projects  
Data Sets

**Lectures**  
Distinguished Lecture Series

**Events**  
All Events  
Related Meetings

**Blog**  
Blog on Data and Visual Analytics  
Data and Visual Analytics Taxonomy

**Announcements**  
FODAVA: Seeking a Research Scientist  
PhD Fellowships Available

**Education & Outreach**  
Short Course  
Summer Intern Program


**Other DAVA News**  
Related news

**Latest News and Events**

<b>SAMSI-FODAVA Workshop</b> The SAMSI-FODAVA Workshop on Interactive Visualization and Analysis of Massive Data will be held on December Posted: October 02, 2012	<b>FODAVA Annual Review Meeting 2012</b> The FODAVA Annual Meeting will immediately follow (Dec 12-13) the SAMSI/FODAVA joint workshop at the Posted: September 05, 2012	<b>FODAVA Testbed Software</b> Many of the modern data sets such as text and image data can be represented in high-dimensional vector spaces and have Posted: June 30, 2012
--	--	---

**About FODAVA**

Enormous amounts of data are being generated every day in health care, computational biology, homeland security, commerce, and many other areas. Analyzing these massive and complex data sets is essential to achieve new discoveries, but extremely difficult. An emerging research field known as data and visual analytics is concerned with synthesizing information and deriving insight from massive, dynamic, ambiguous and possibly conflicting digital data for increased understanding and effective decision making.



The Foundations on Data Analysis and Visual Analytics (FODAVA) research initiative is dedicated to both defining the foundations of the data and visual analytics fields and advancing the state-of-the-art. Established in 2008, the FODAVA initiative is a collaborative effort funded jointly by the National Science Foundation (NSF) and the Department of Homeland Security (DHS).

The Georgia Institute of Technology, as the FODAVA-Lead institution will lead and coordinate this new initiative. It will perform foundational research in massive data

**About fodava.gatech.edu**




Our goal is to keep you informed on the progress of the FODAVA initiative while being maintained as a base for further education and outreach to the data and visual analytics community.

Read more [about FODAVA](#) and view a presentation on FODAVA's [Research, Education and Community Building!](#)

**Recently updated**

- [Department of Homeland Security](#)
- [National Science Foundation](#)
- [Research, Education and Community Building](#)

**Links**



Home About Us NSF BIGDATA Solicitation Contact Us

**People of FODAVA**  
FODAVA-Lead  
FODAVA-Partners '10  
FODAVA-Partners '09  
FODAVA-Partners '08

**Research**  
Technical Reports  
Projects  
Data Sets

**Lectures**  
Distinguished Lecture Series

**Events**  
All Events  
Related Meetings

**Blog**  
Blog on Data and Visual Analytics  
Data and Visual Analytics Taxonomy

**Announcements**  
FODAVA: Seeking a Research Scientist  
PhD Fellowships Available

**Education & Outreach**  
Short Course  
Summer Intern Program

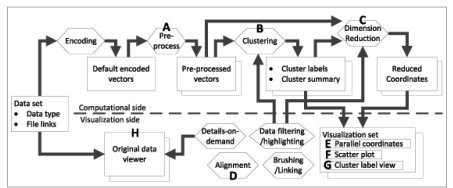
**Other DAVA News**  
Related news

**Latest News and Events**

<b>SAMSI-FODAVA Workshop</b> The SAMSI-FODAVA Workshop on Interactive Visualization and Analysis of Massive Data will be held on December Posted: October 02, 2012	<b>FODAVA Annual Review Meeting 2012</b> The FODAVA Annual Meeting will immediately follow (Dec 12-13) the SAMSI/FODAVA joint workshop at the Posted: September 05, 2012	<b>FODAVA Testbed Software</b> Many of the modern data sets such as text and image data can be represented in high-dimensional vector spaces and have Posted: June 30, 2012
--	--	---

**FODAVA Testbed Software**

Many of the modern data sets such as text and image data can be represented in high-dimensional vector spaces and have benefited from computational methods that utilize advanced techniques from numerical linear algebra. Visual analytics approaches have contributed greatly to data understanding and analysis due to their capability of leveraging humans' ability for quick visual perception. However, visual analytics targeting large-scale data such as text and image data has been challenging due to limited screen space in terms of both the numbers of data points and features to represent. Among various computational technique supporting visual analytics, dimension reduction and clustering have played essential roles by reducing these numbers in an intelligent way to visually manageable sizes. Given numerous dimension reduction and clustering techniques available, however, decision on choice of algorithms and their parameters becomes difficult.



The FODAVA testbed system is an interactive visual testbed system for dimension reduction and clustering in a large-scale high-dimensional data analysis. The testbed system enables users to apply various dimension reduction and clustering methods with different settings, visually compare the results from different algorithmic methods to obtain rich knowledge for the data and tasks at hand, and eventually choose the most appropriate path for a collection of algorithms and parameters.

The testbed can load image, raw text, and vector-encoded data types. It offers 4 different clustering and 17 different dimension reduction methods. Furthermore, the FODAVA testbed system is implemented in a flexible and modular way so

# 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!