

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

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

FODAVA Annual Meeting, Dec. 2012

(PIs: H. Park, A. Gray, J. Stasko, V. Koltchinskii, R. Monteiro)



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)

Challenges in Computational Methods for *High Dimensional Large-scale Data* on Visual Analytics System

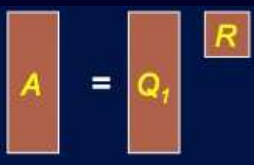
- **Data challenges**
 - Massive, High-dimensional, Nonlinear
 - Vast majority of data is unstructured
 - Noisy, errors and missing values are inevitable in real data set
 - Heterogeneous format/sources/reliability
 - Time varying, dynamic, ...
- **Visualization challenges**
 - **Screen Space and Visual Perception**
 - High dimensional data: Effective dimension reduction
 - Large data sets: Informative representation of data
 - **Speed**: necessary for real-time, interactive use
 - Scalable algorithms
 - Adaptive algorithms

Key Foundational Components for VA System Development

- **Dimension Reduction**
 - Dimension reduction with prior info/interpretability constraints
 - Manifold learning
- **Informative Presentation of Large Scale Data**
 - Sparse recovery by L_1 penalty
 - Clustering, semi-supervised clustering
 - Multi-resolution data approximation
- **Fast Algorithms**
 - Large-scale optimization/matrix decompositions
 - Adaptive updating algorithms for dynamic and time-varying data, and interactive vis.
- **Information Fusion**
 - Fusion of different types of data from various sources, vis. comparisons
- **Integration with DAVA systems**
 - **Testbed, Jigsaw, iVisClassifier, iVisClustering, VisIRR ..**

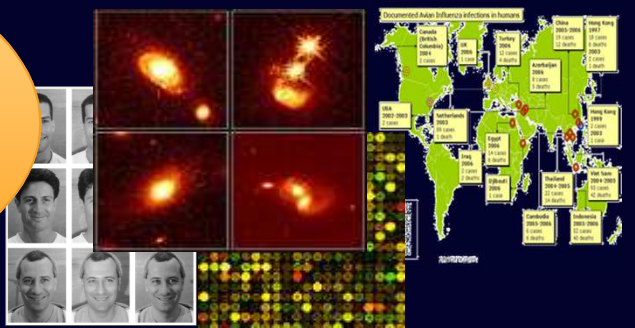
FODAVA Research Test Bed for Visual Analytics of *High Dimensional Data*

- Library of key computational methods for visual analytics of high dimensional large scale data
 - With visual representations and interactions
 - Easily accessible for DAVA researchers and readily available for applications
- Identifies effective methods for specific problems (evaluation)
- Modular: A base for specialized VA systems
(e.g. iVisClassifier, iVisClustering, VisIRR)

 FODAVA
Fundamental
Research

$$S_w = \sum_{1 \leq i \leq r} \sum_{j \in N_i} (a_j - c_i) (a_j - c_i)^T$$
$$S_b = \sum_{1 \leq i \leq r} \sum_{j \in N_i} (c_i - c) (c_i - c)^T$$
$$S_t = \sum_{1 \leq i \leq n} (a_i - c) (a_i - c)^T$$



Applications




Documented Asian Influenza Infections in Humans



FODAVA Research Testbed Software: Available at <http://fodava.gatech.edu/fodava-testbed-software>

- Supports various dimension reduction, clustering, and their visual representations and comparisons through alignments for high-dimensional data
- Application domains: document analysis, bioinformatics, seismic data analysis, healthcare, communications, computer vision, ...
- Language used: backend library in Matlab, GUI in JAVA (no need for Matlab installed)
- System support: Windows 32/64 bit, Linux 32/64 bit





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

SAMSI-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

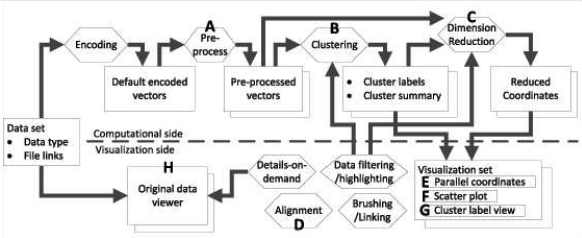
Related news

Latest News and Events

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

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.



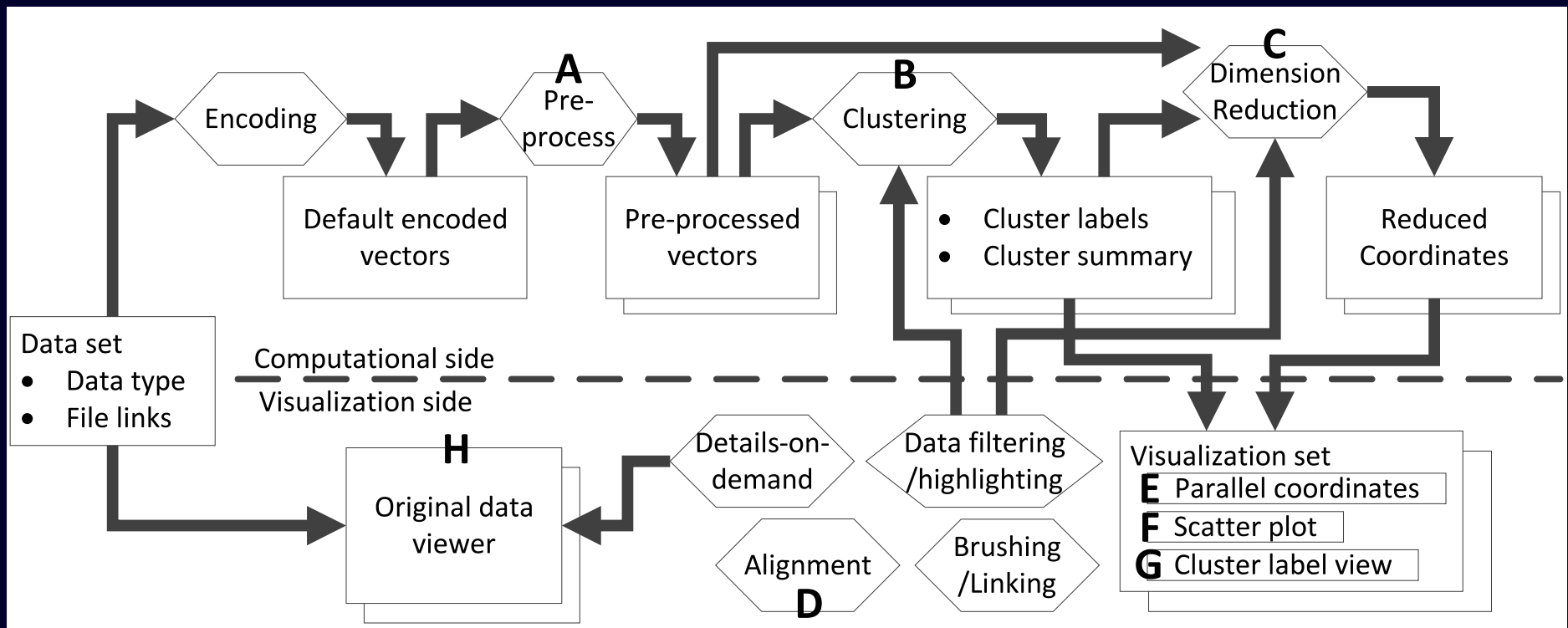
Testbed Modules

- Computational modules

- Vector encoding
- Pre-processing
- **Clustering**
- **Dimension reduction**

- Interactive visualization modules

- Parallel coordinates
- Scatter plot
- Cluster summary
- Brushing and Linking
- **Space alignment**
- Raw data view



Dimension Reduction

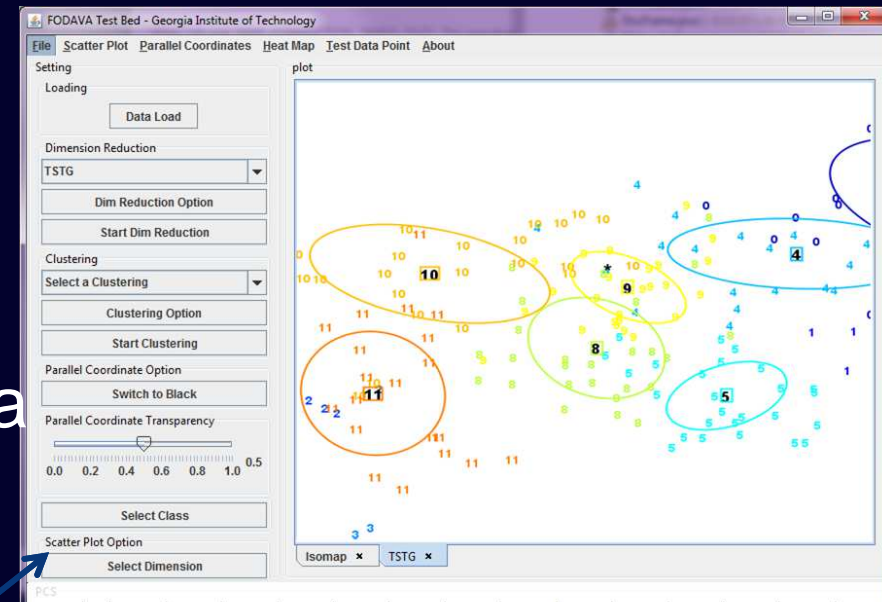
- Visualizes high-dimensional data by parallel coordinates and/or scatter plot
- Methods
 - Linear methods
 - PCA, FA, ProbPCA, **LDA**, OCM, NPE, LPP, LLTSA, NCA, MCML
 - Nonlinear methods
 - MDS, Isomap, LLE, LTSA, Sammon, HessLLE, MVU, LandMVU, KernPCA, GDA, DiffMaps, SPE, AutoEnc, LLC, ManiChart, CFA, GPLVM, SNE, T-SNE
- Provides initial parameters that can be changed interactively
- Can recursively apply dimension reduction on user-selected data
- Fast algorithms implemented

Clustering and Classification

- Generates cluster/class labels of data, which are color-coded in visualization.
- Methods
 - Clustering
 - Hierarchical clustering, *K*-means, spherical *K*-means, GMM, **NMF**, constrained *K*-means, DisCluster/Diskmeans [J. Ye]
 - Classification (on-going work)
 - *K*-nearest neighbors classifier, SVM, Logistic regression, Naïve Bayes
- Provides cluster summary
- Provides GUI for semi-supervision, e.g., must/cannot link
- Can hierarchically construct cluster structures

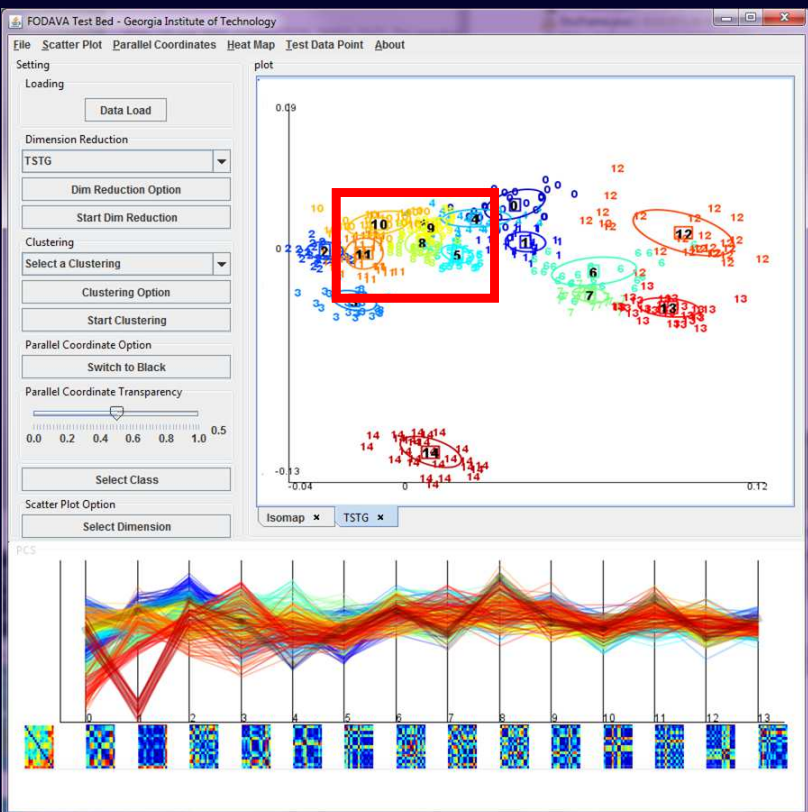
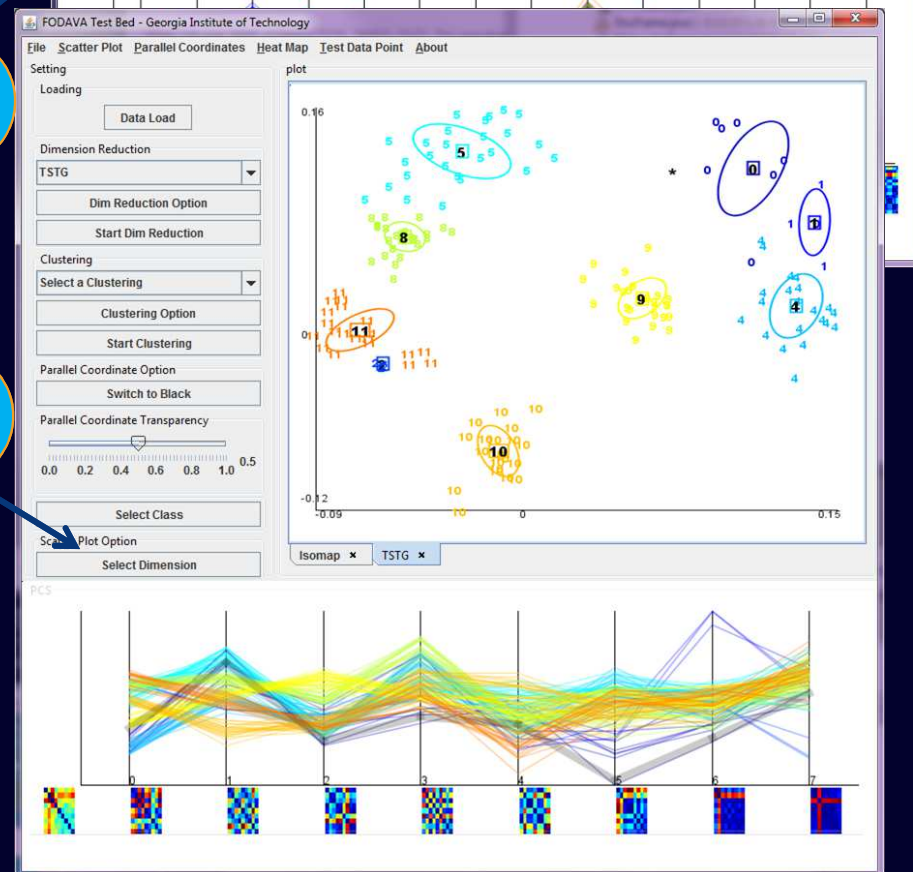
Computational Zoom-in

Computational zoom-in by recursive dimension reduction on selected data



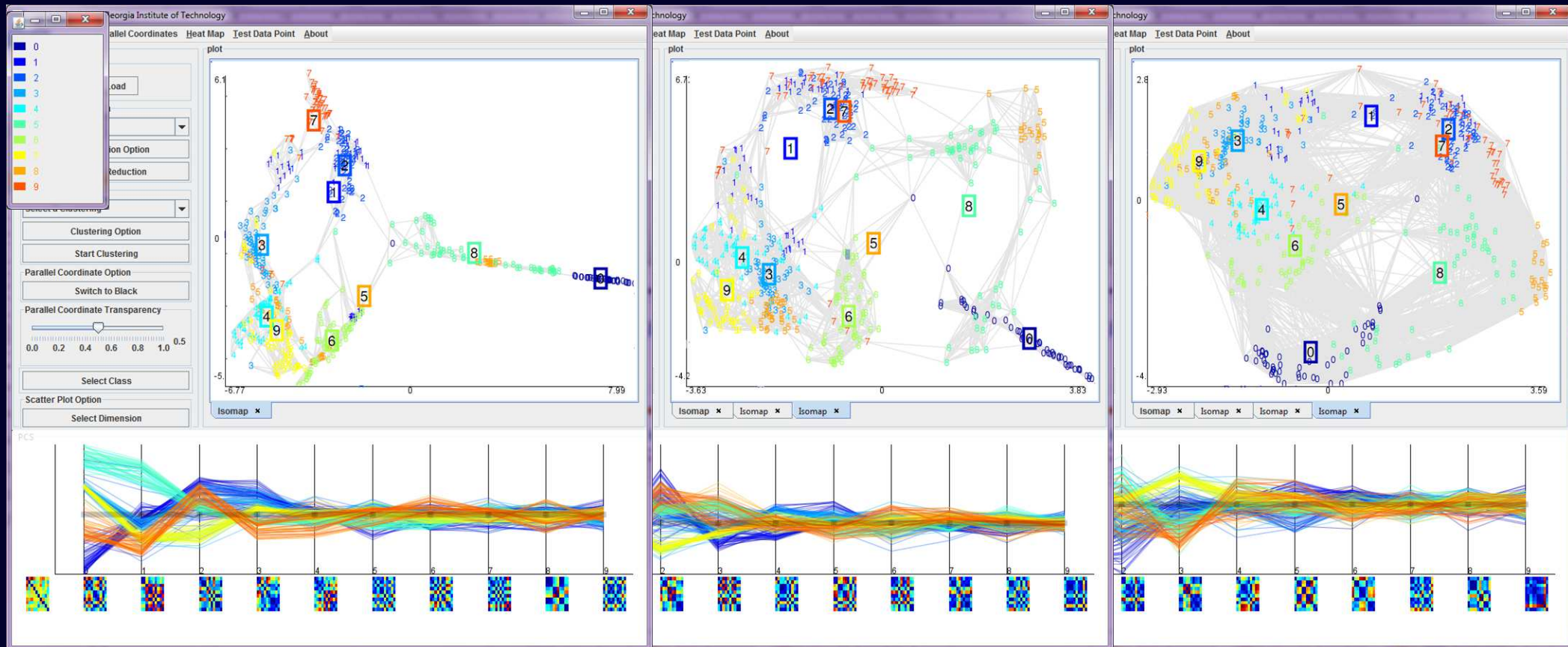
Normal Zoom-in

Comp. Zoom-in



Interactive Parameter Change

e.g. in Isomap k value in k -NN Graph
Controls the level of focus on locality



Isomap parameter input window

#Dimensions: 10 **K in K-NN: 7**

Show centroid Show ellipse

Confirm

Isomap parameter input window

#Dimensions: 10 **K in K-NN: 15**

Show centroid Show ellipse

Confirm

Isomap parameter input window

#Dimensions: 10 **K in K-NN: 40**

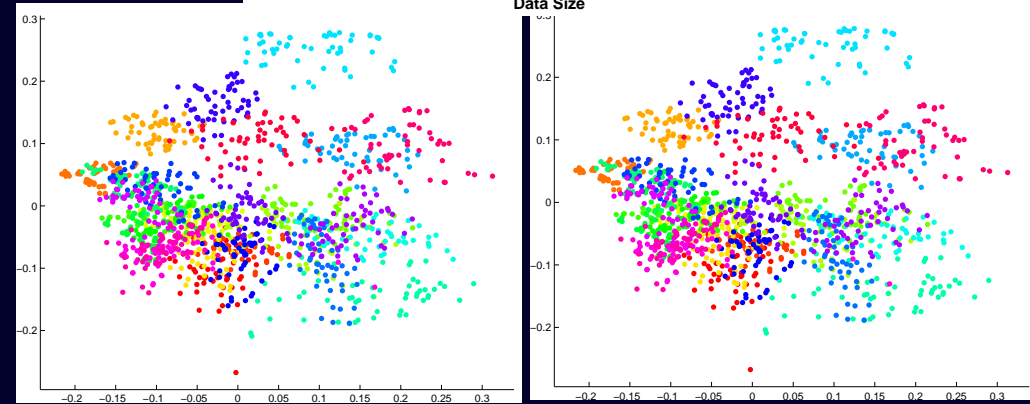
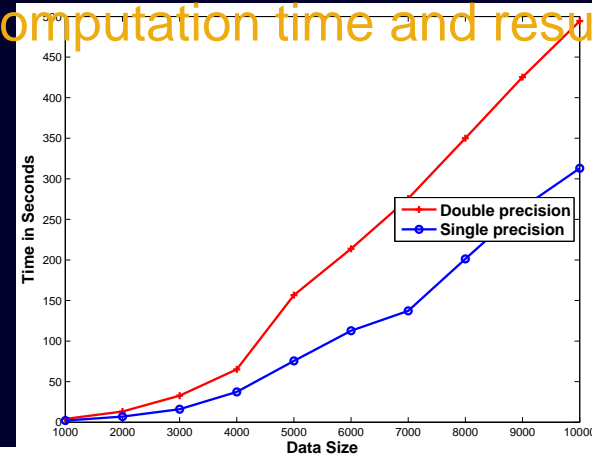
Show centroid Show ellipse

Confirm

Fast Comp. Modules for Interactive Vis.

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

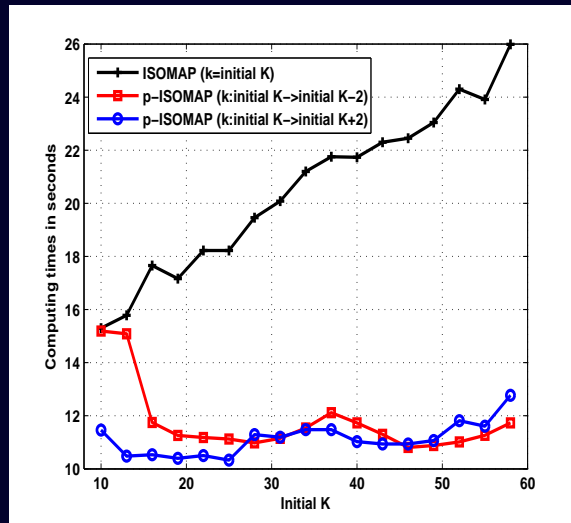
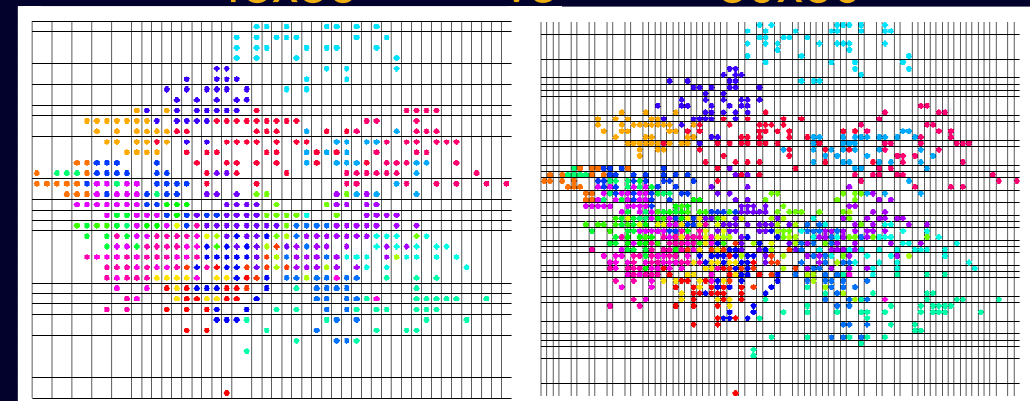
PCA timing: double vs single precision computation time and results



48x36

vs

80x60



p-Isomap computing time vs. k value in k -NN graph

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 08 SIAM Journal on Matrix Analysis and Applications, Kim and Park 11 SISC...)

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

- Why Nonnegativity Constraints?
 - Better Approx. vs. Better Representation/Interpretation
 - Nonnegative Constraints often *physically meaningful*
 - *Interpretation* of analysis results possible
- One of the Fastest Algorithms for NMF & theoretical convergence analysis
- Matlab codes publicly available (J. Kim and H. Park, IDCM08, SISC11)
<http://www.cc.gatech.edu/~hpark/nmfsoftware.php>
- NMF is better and faster than K-means in clustering
 - 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
 - SymNMF (Kuang, Ding, Park, SDM12), Sparse NMF for clustering (Kim and Park, Bioinfo., 07)

NMF for Clustering

- NMF more accurate and faster on document and image data

(Xu et al. 03; Pauca et al. 04; Li et al. 07; Kim & Park, 08; Ding et al. 10 ...)

– Clustering accuracy averaged over 20 runs:

	K	K-means	SphKmeans	NMF/ANLS	GTMF	Spectral	SymNMF
TDT2	4	.7994	.7978	.9440	.9150	.9093	.9668
	8	.6147	.6208	.8292	.8200	.7357	.8819
	16	.5286	.5305	.6709	.6812	.5959	.7635
Reuters	4	.5755	.5738	.7737	.7798	.7171	.8077
	8	.5170	.5049	.6747	.6758	.6452	.7343
	16	.3712	.3687	.4608	.5338	.5001	.6688

	TDT2	Reuters	COIL20	NIPS	ORL	PIE	<i>Overall</i>
K-means	0.6734	0.4289	0.6184	0.4650	0.6499	0.7384	<i>0.5957</i>
NMF/ANLS	0.8534	0.3770	0.6312	0.4877	0.7020	0.7912	<i>0.6404</i>
SymNMF	0.8979	0.5305	0.7258	0.5129	0.7798	0.7517	<i>0.6998</i>

– Problem sizes:

	TDT2	Reuters	COIL20	NIPS	ORL	PIE	20 Newsgroups
m	26618	12998	4096	17583	5796	4096	36982
n	8741	8095	1440	420	400	232	18669
k	20	20	20	9	40	68	20

On average, NMF is faster than k-means by a factor of at least 2

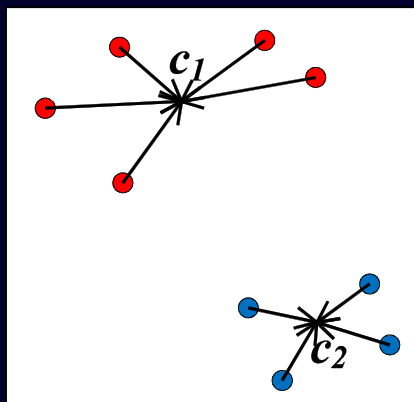
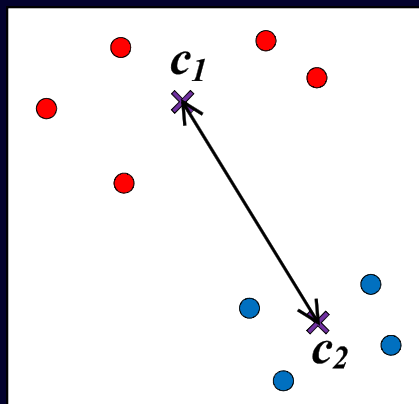
Linear Discriminant Analysis for 2D/3D Representation of Clustered Data

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

Max trace ($G^T S_b G$)

&

min trace($G^T S_w G$)



LDA/GSVD

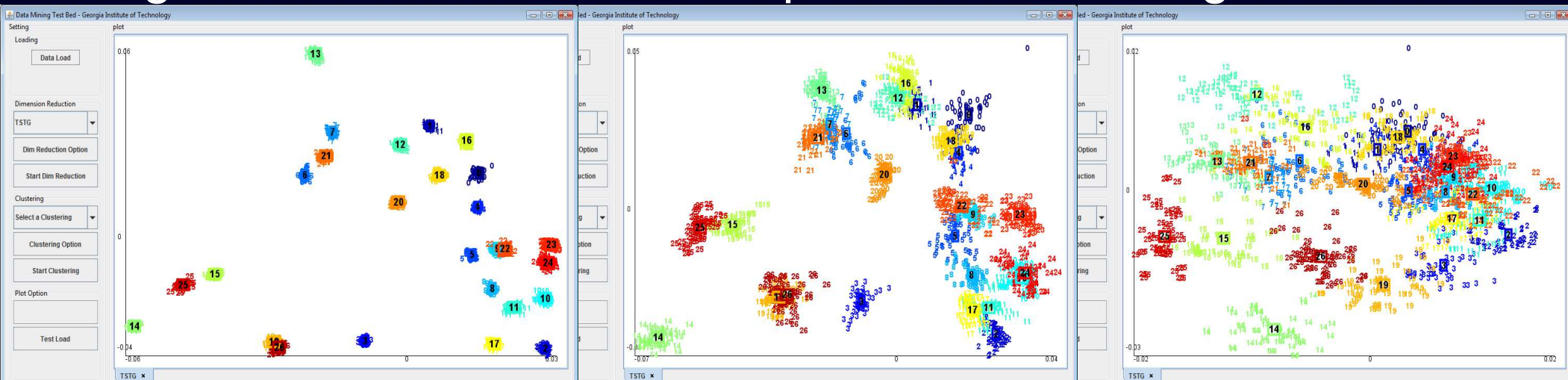
$$\alpha^2 H_b H_b^T x = \beta^2 H_w H_w^T x$$



max trace

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

- Regularization in LDA for Computational Zooming-in

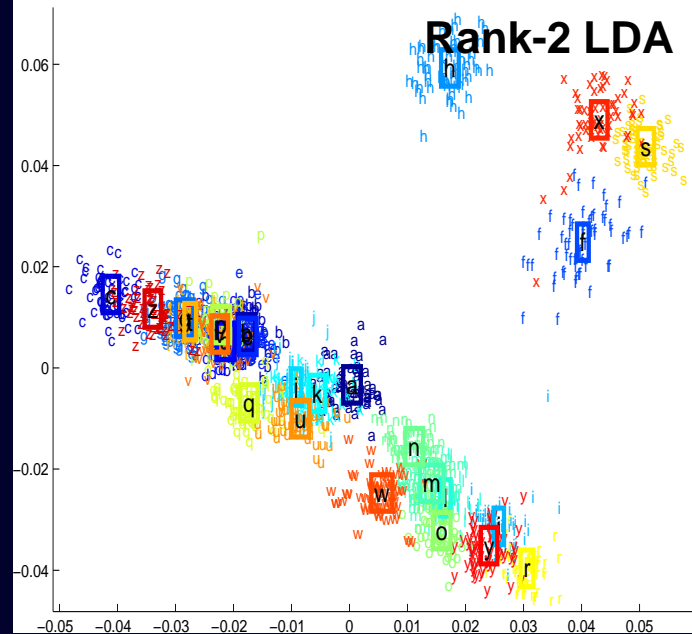
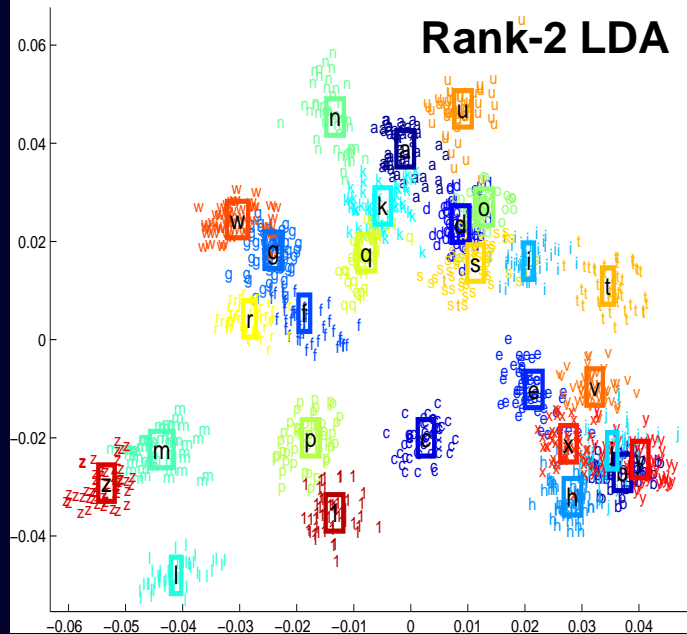
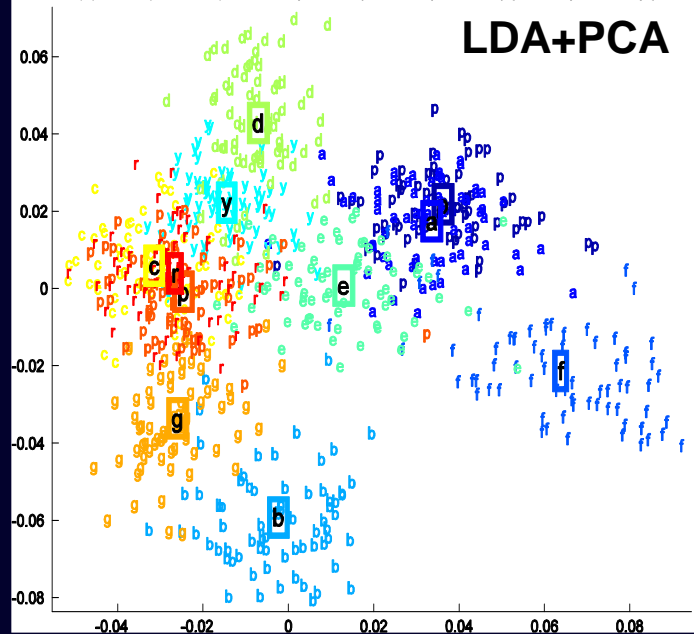
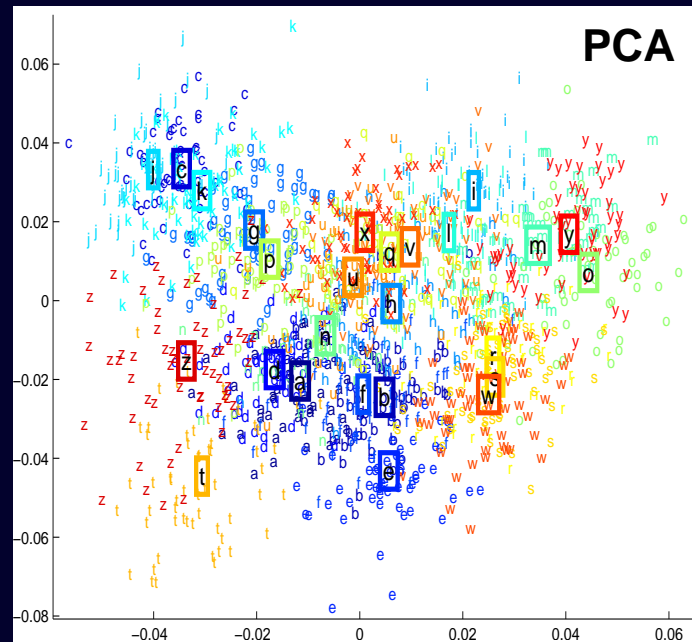
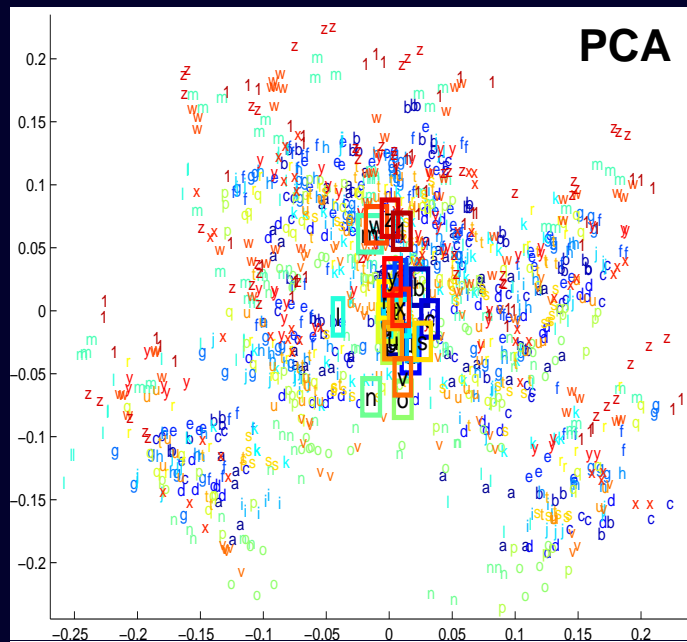
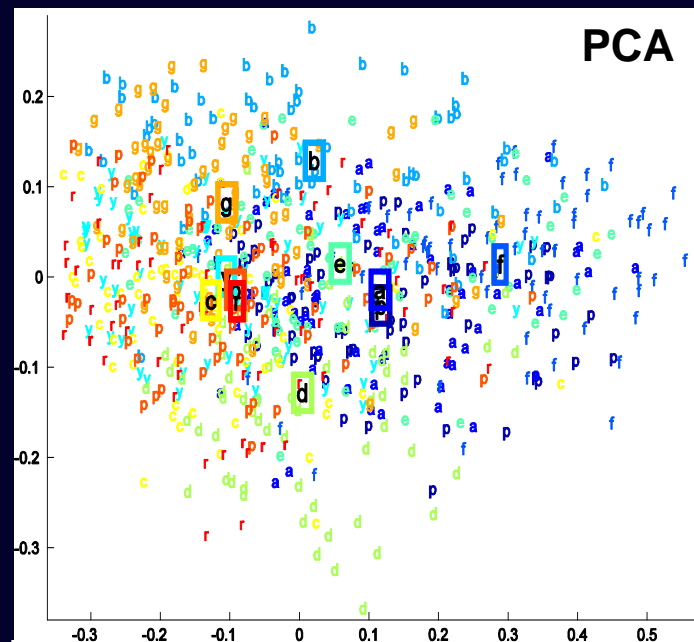


Small regularization



Large regularization

2D Visualization of Clustered Text, Image, Audio Data

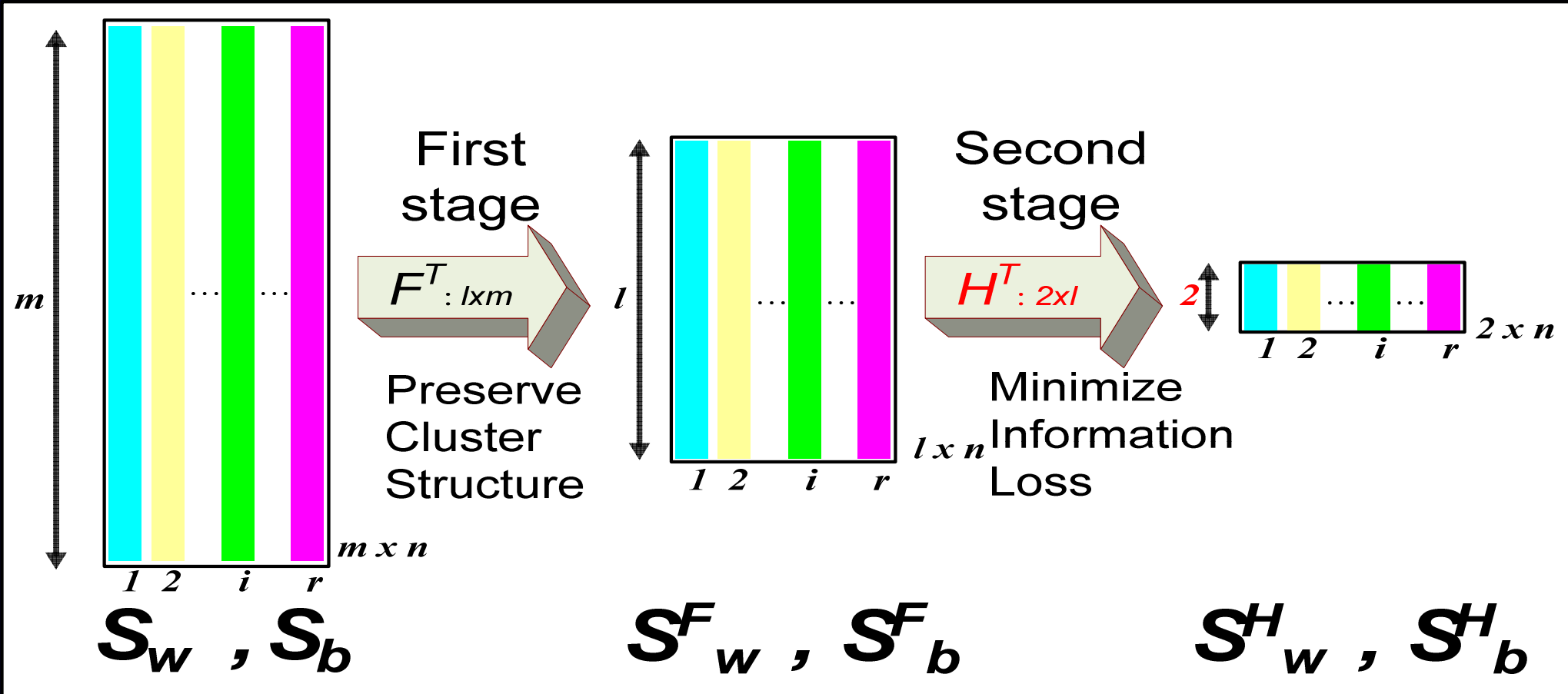


20news Data (Text)

Facial Data (Image)

Spoken Letters (Audio)

Two-stage Dimension Reduction for 2D Vis. of Clustered Data



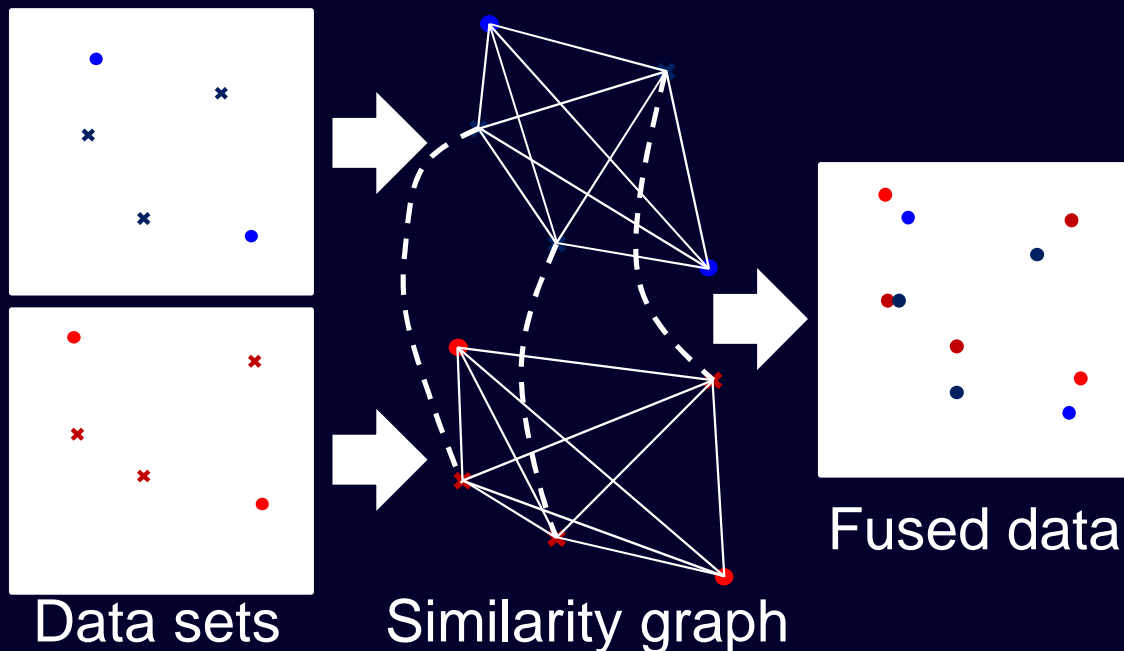
- LDA + LDA = Rank2 LDA
- LDA + PCA
- OCM + PCA
- OCM + Rank-2 PCA on S_b^F = Rank-2 PCA on S_b

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)**

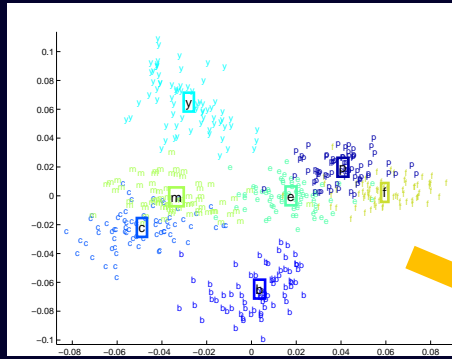


- **Procrustes analysis**

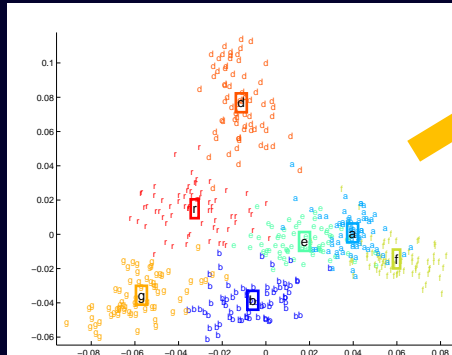
$$\min \| (A - \mu_A \mathbf{1}^T) - kQ(B - \mu_B \mathbf{1}^T) \|_F$$
$$Q^T Q = I$$

Fusion and Alignment in Testbed

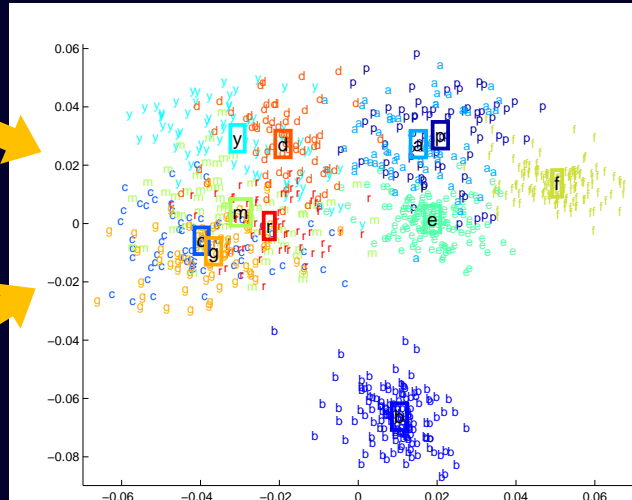
Data set 1



Data set 2



Fused

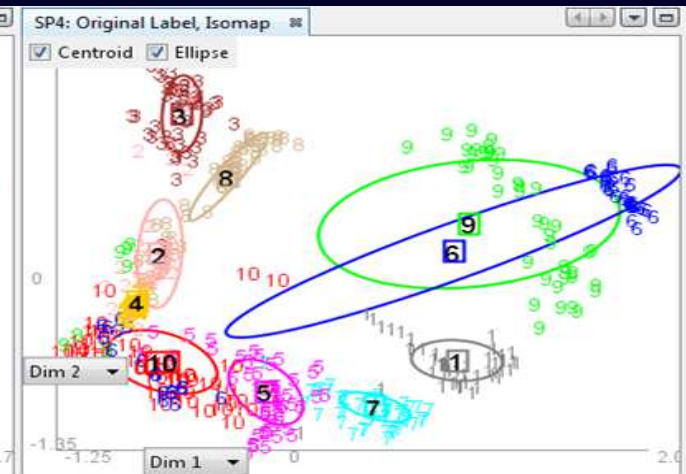
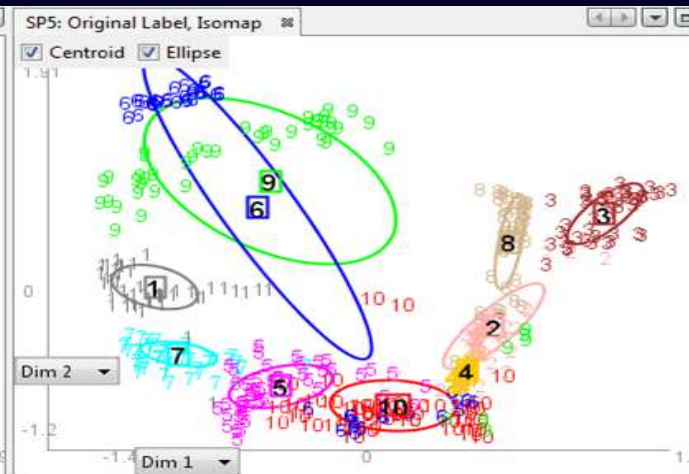
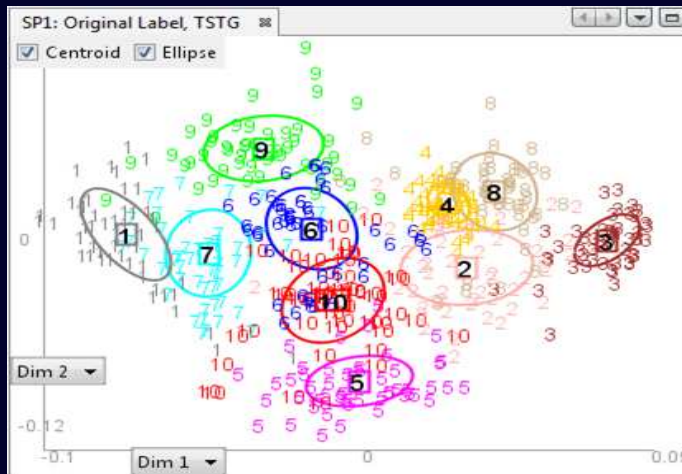


- Data set 1 only:
 - comp.sys.ibm.pc.hardware ('p'),
 - sci_crypt ('y'),
 - soc.religion.christian ('c'),
 - talk.politics.misc ('m')
- Data set 2 only:
 - comp.sys.mac.hardware ('a'),
 - sci.med ('d'), talk.religion.misc ('r'),
 - talk.politics.guns ('g')
- Shared:
 - rec.sport.baseball ('b'),
 - sci.electronics ('e'), misc.forsale ('f')

Reference

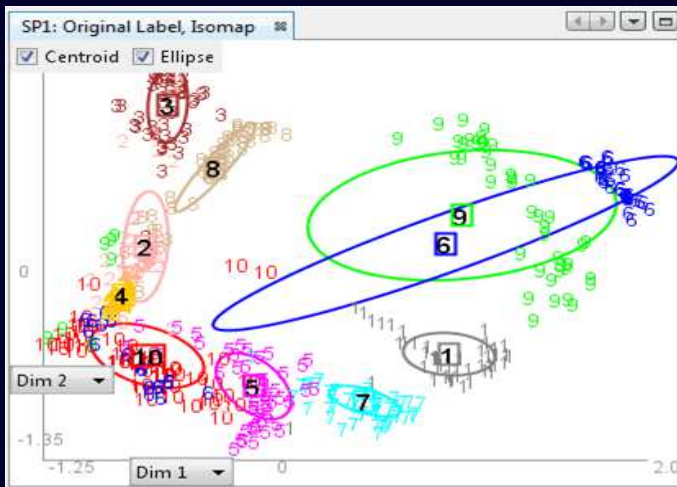
Aligned

Un-Aligned

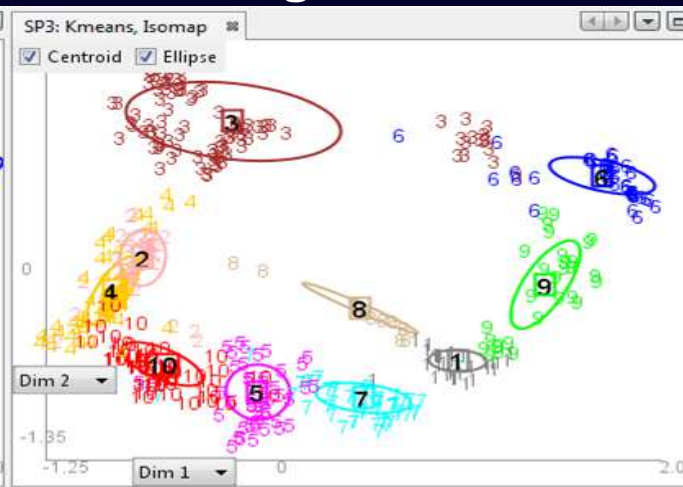


Cluster Alignment: Label Matching and Space Alignment

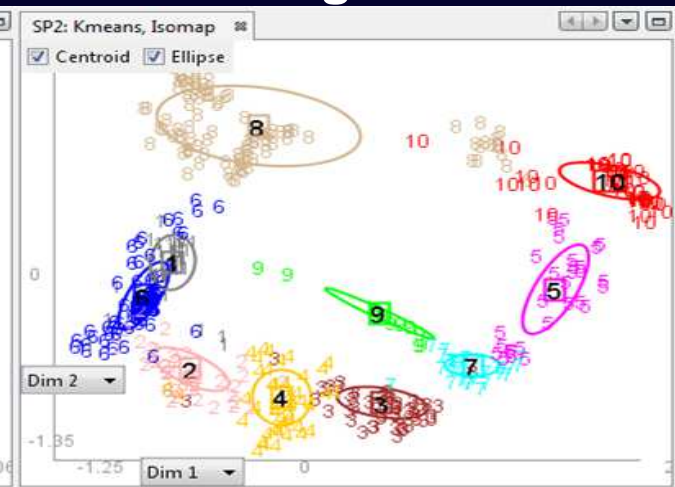
Reference



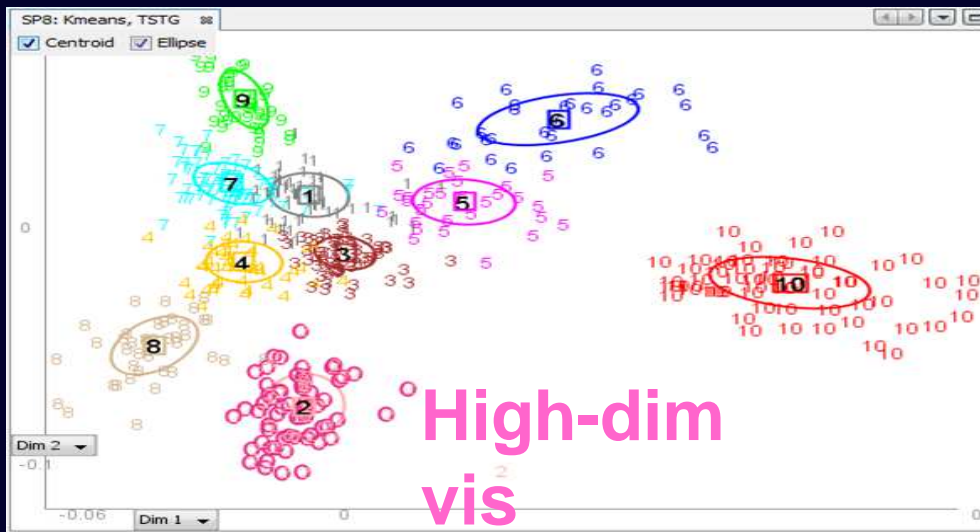
Aligned



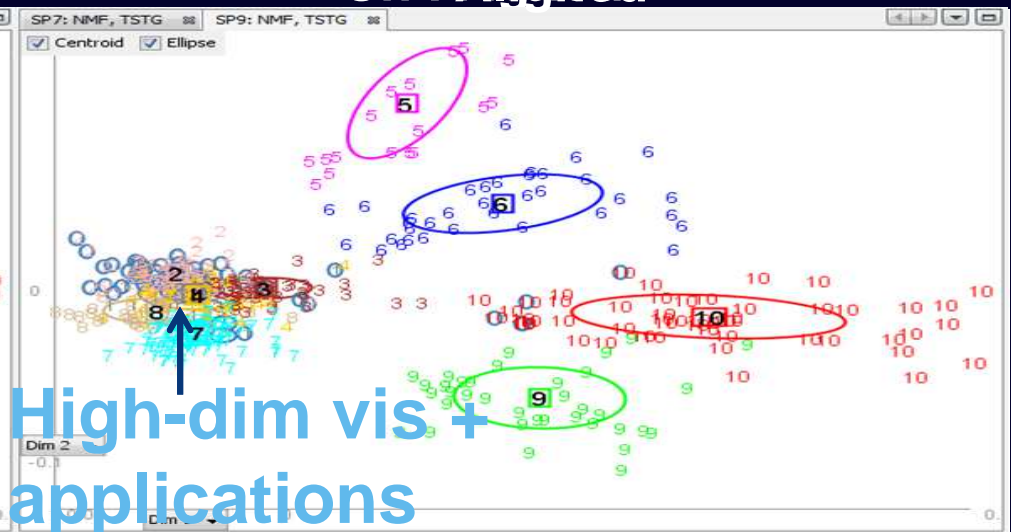
Un-Aligned



Reference



Un-Aligned



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

Testbed Overview

FODAVA Visual Testbed - Georgia Institute of Technology

File Edit View Navigate Source Refactor Run Debug Team Tools Window Help

Settings

Preprocessing

Pre-Normalization Centering

Document-specific

TF-IDF weighting

Filter Word Min Freq: 2

Preprocess

Clustering

NMF

Clusters: 10

Algorithm: HALS BPP

Max Iteration: 200

Clustering

Dimension Reduction

TSTG

Dimensions: 10

Method: LDA LDA+PCA OCM OCM+PCA

Regularization

Visualize

Alignment

View 1 View 4

Clustering Dimension Reduction

Align

SP1: Kmeans, PCA SP4: NMF, TSTG

Centroid Ellipse

Dim 2

Dim 1

Document View - infovis02--1173140.txt

Cluster-wise Representative Keywords

0 10 20 30 40 50 60 70 80 90 100

infovis98--729562.txt
 infovis97--636761.txt
 infovis96--559226.txt
 infovis10--164.txt
 infovis09--5290715.txt
 infovis09--5290714.txt
 infovis09--5290711.txt
 infovis09--5290703.txt
 infovis07--4376141.txt
 infovis05--1532152.txt
 infovis04--1382901.txt
 infovis04--1382889.txt
 infovis03--1249007.txt
 infovis02--1173155.txt
 infovis02--1173140.txt

s, the traffic was either much below its average rate or much above. In other words, the traffic was not smooth, not staying at all times close to its average. It was bursty on the cable running down a street, carrying the merged traffic of a small number of cable modem users in one section of a town. It was bursty on the core fiber of an internet service provider, carrying the merged traffic of thousands of users from all over the country. The internet was designed to accommodate the bursty traffic. The routers and switches that forward traffic from one place to the next were designed for burstiness, and internet service providers allocated traffic loads on the devices based on an assumption of burstiness. Recently, it was discovered that the old common wisdom is not true. Visualization

Image View - VP4-IL3-EX1.jpg

VP4-IL3-EX1.jpg
 VP3-IL4-EX1.jpg
 VP1-IL4-EX1.jpg
 VP1-IL3-EX1.jpg
 VP0-IL2-EX2.jpg
 VP4-IL1-EX3.jpg
 VP3-IL1-EX2.jpg
 VP2-IL0-EX3.jpg
 VP3-IL0-EX2.jpg
 VP2-IL1-EX2.jpg
 VP2-IL1-EX1.jpg
 VP1-IL1-EX2.jpg
 VP0-IL1-EX3.jpg
 VP0-IL0-EX3.jpg

PC1: Kmeans, PCA PC4: NMF, TSTG

LB1: Kmeans, PCA LB4: NMF, TSTG

1: graph, clusters, layout, edge, node, lat
 2: querying, interface, databases, multiple, s
 3: document, text, collections, informatio
 4: dimensions, parallelize, coordinated, mull
 5: treemaps, layout, hierarchical, ratio, w
 6: trees, hierarchy, node, genealogical, s
 7: collaboration, wikipedia, analytics, shared
 8: 3d, spatial, landscapes, animation, map
 9: networks, traffic, flow, social, node, layout
 10: designed, model, analytics, informatio

CSV View

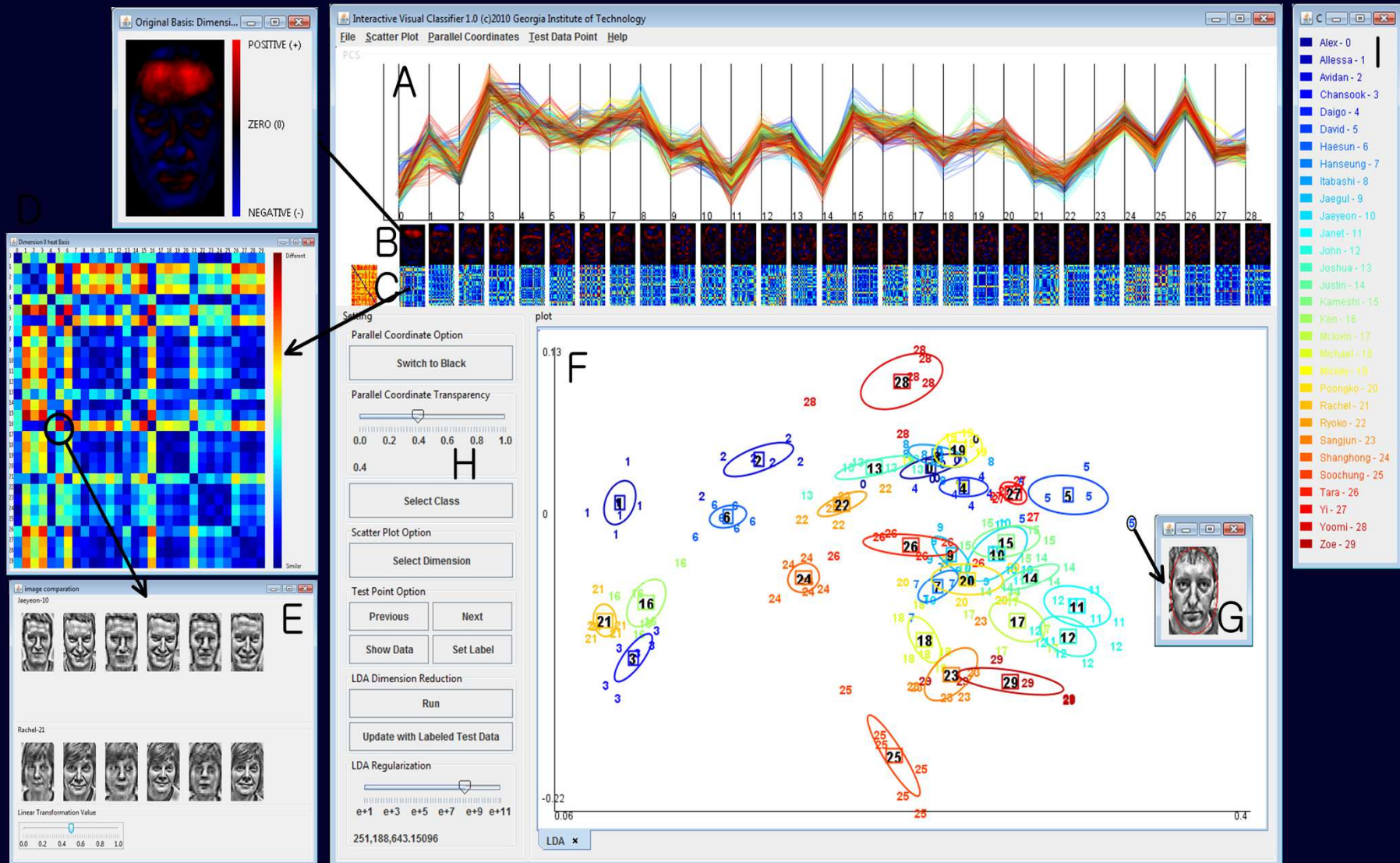
	dim 1	dim 2	dim 3	dim 4	dim 5
1	0.378	0.259	0.221	0.382	0
2	0.356	0.252	0.193	0.387	0
3	0.492	0.492	0.335	0.384	0.182
4	0.189	0.366	0.149	0.253	0.261

9 8 5

iVisClassifier

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

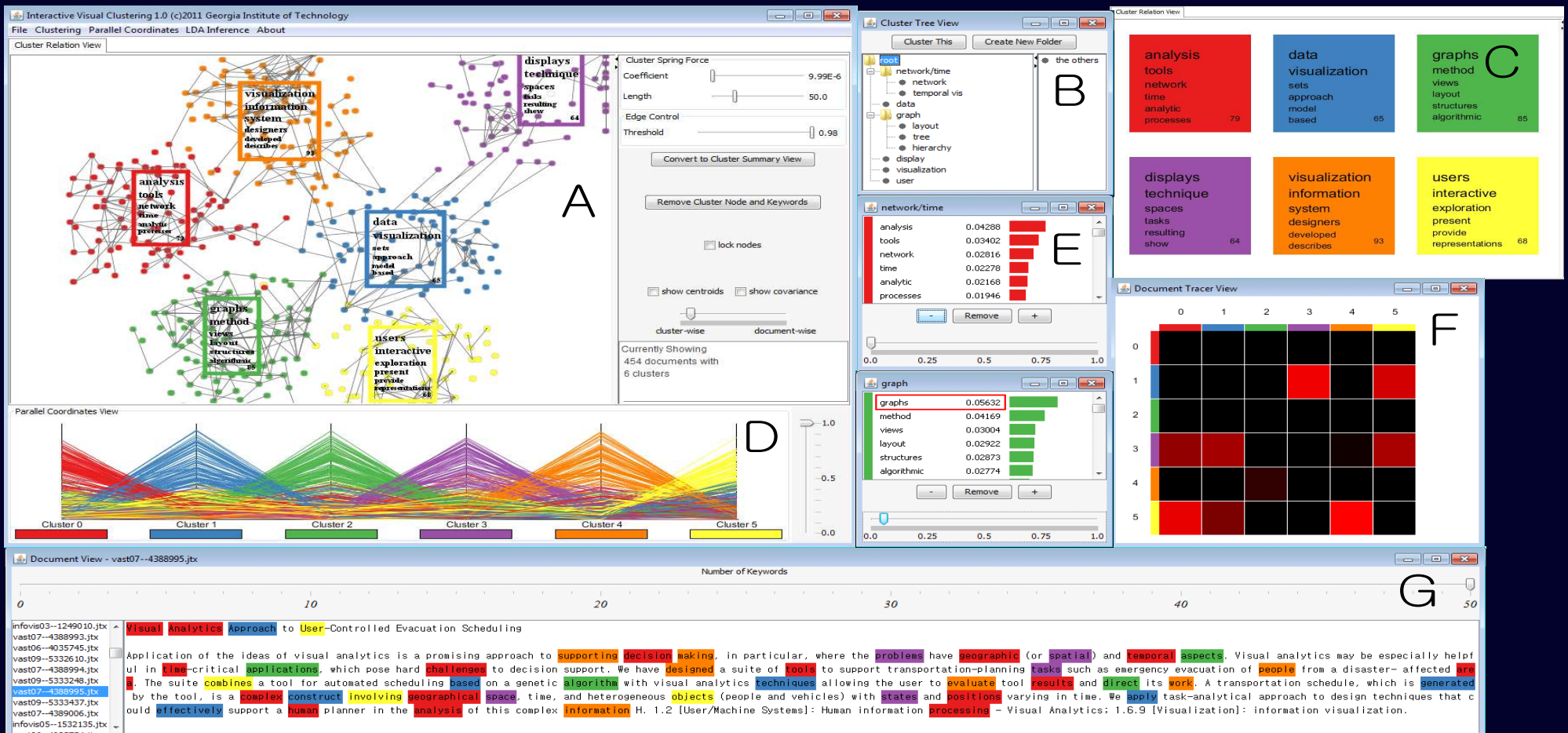
Interactive visual analytics system for classification of high-dim. data (image, text, etc) and search space reduction



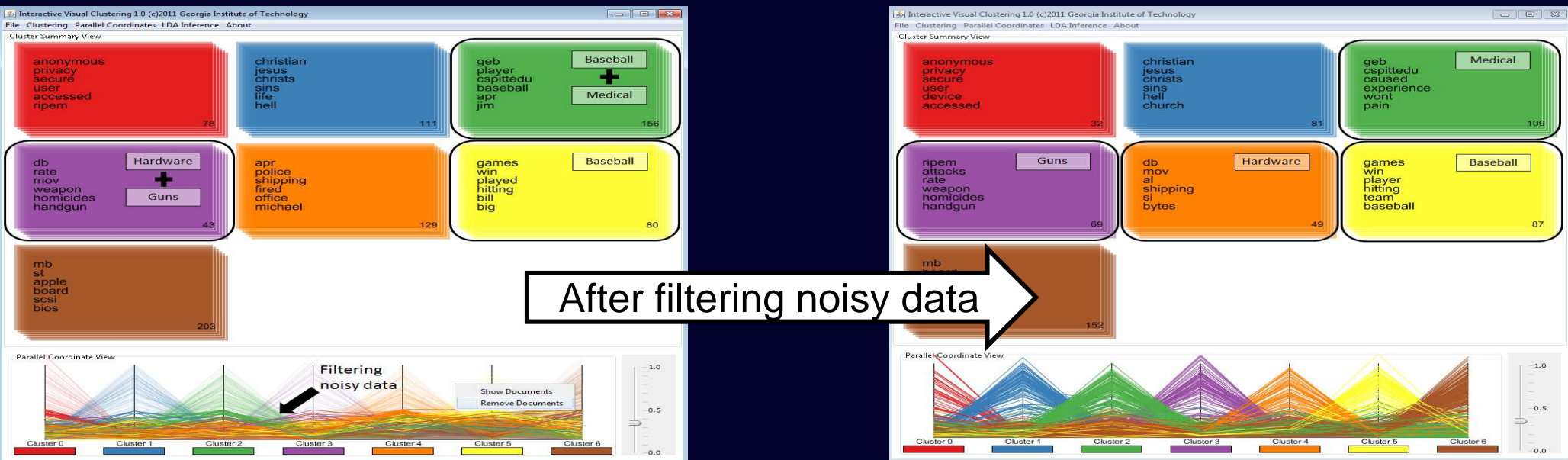
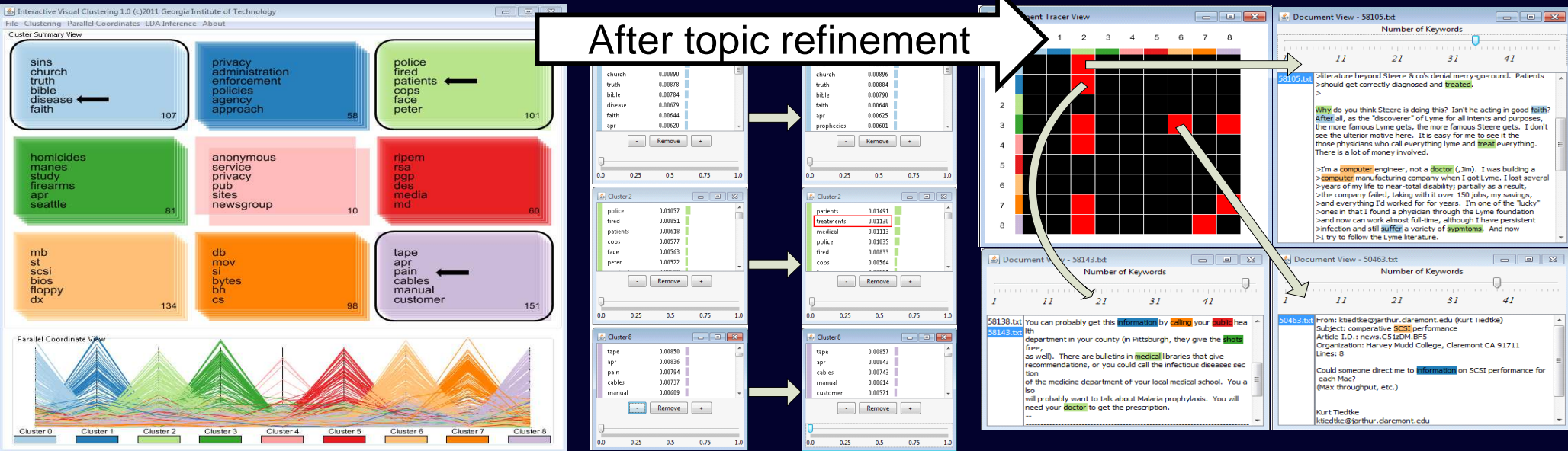
iVisClustering

(H. Lee, J. Kihm, J. Choo., J. Stasko, HP, EuroVis12)

- Interactive visual document clustering system using topic modeling
- Refines clusters and supports hierarchical cluster structure in an interactive way

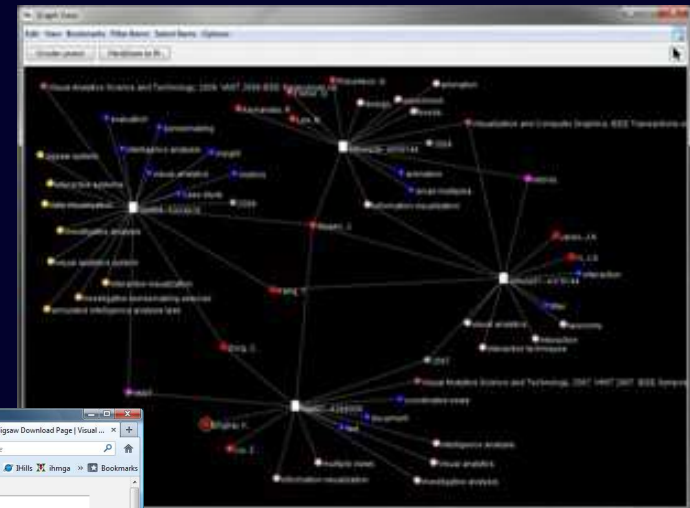


Key Interactions with LDA: Topic Refinement and Noisy Data Filtering



Jigsaw

- Combining computational text analysis (text mining) with interactive visualization
- Placed system on web in Fall '12 where anyone can download it <http://www.cc.gatech.edu/gvu/ii/jigsaw>
 - Created video tutorials
 - Many sample data sets provided
- Working on opening up architecture

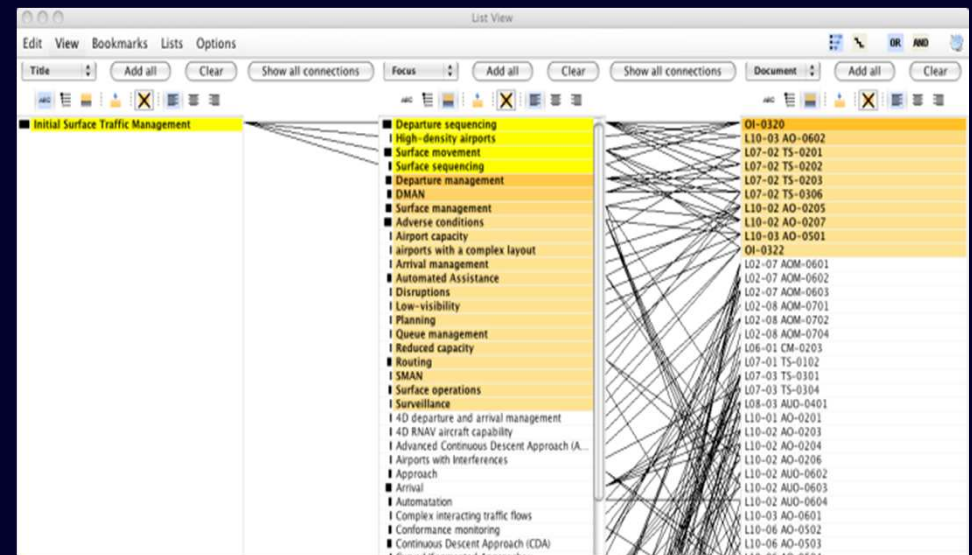
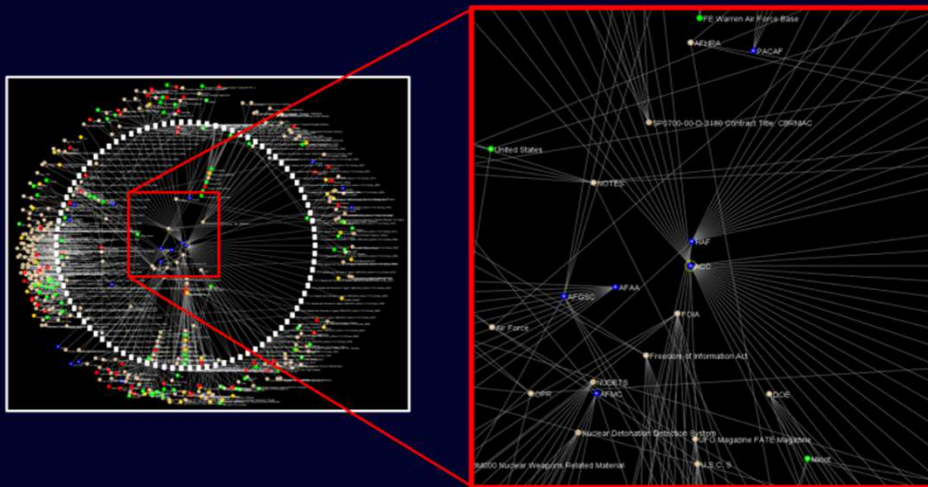


A screenshot of the Jigsaw website. The top part shows the home page with a navigation menu, a search bar, and a list of active downloads and papers. The bottom part shows the download page, which includes a registration form with fields for Name, Affiliation, Email, and a dropdown for 'How did you hear about Jigsaw?'. The website has a clean, professional layout with a white background and blue accents.

Case Study of System Usage

(Y. Kang & J. Stasko, VAST12)

- Interviewed six people who had been using Jigsaw for 2-14 months
 - fraud, law enforcement, intelligence analysis, research
- Understand how they are using Jigsaw, different domains
- Learn about strengths of system and its limitations



VisIRR: Visual Information Retrieval and Recommendation System for Document Discovery

Improves personalization and understandability via integrated visualizations of document retrieval and recommendation

- Visual IR: beyond Google-like keyword search:
 - See **more** 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

Related work

Commercial tools for researchers:

- Mendeley - www.mendeley.com – Free reference manager and PDF organizer
 - Offline client, personal homepage, social features (Community, follow researchers etc), recommendation engine (people/paper), plug in support. Naive collaborative filtering based recommendations, no visualization.
- Arnetminer - www.arnetminer.com – Academic researcher Social network search
 - Metrics (uptrend, longevity, diversity etc), Authorship Network. Author specific website. Research is beyond only authors.
- Microsoft academic – academic.research.microsoft.com – A free academic search engine
 - Innovative ways to explore academic publications, authors, conferences, journals, organizations and keywords, connecting millions of scholars, students, librarians, and other users, very rich visualization features. Very limited set of domains (only for computer science), No social features
- Google scholar – scholar.google.com – Search engine for scholarly articles.
 - A simple search interface to search all scholarly articles, multiple disciplines, multiple sources (books, patents, articles, university websites, etc). No recommendation, No visualization, Irrelevant search results, Very limited research specific information (number of citations alone).
- Braque.cc - informs researchers of others' research.
 - Academic launch. Not successful commercially.
- Exlibris - BxRecommenderSystems - <http://www.exlibrisgroup.com/category/bXOverview> - Discovery, Distribution and management of print/electronic and digital materials. Recommendations for librarians.
 - Official librarian tool, encompasses huge repository of data from all the university libraries, Web based tool, multi disciplinary recommendations. No author level information (h-index), journal/conference level (rating of the journal), paper specific information(citation etc). Even though recommends papers, does not provide the statistics that researchers relies up on.

Related work

Commercial conference management systems:

Web based software that supports organization of scientific conferences.

- Easychair
- Confmaster.net
- CMT – Microsoft academic conference management site
- Openconf
- PCS

Academic systems:

- C. Basu, H. Hirsh, W. Cohen, and C. Nevill-Manning. Technical paper recommendation: a study in combining multiple information sources. *Journal of AI Research*, pages 231–252, 2001
- D. Mimno and A. McCallum. Expertise modeling for matching papers with reviewers. In *Proc. KDD'07*, pages 500–509, 2007
- J. Goldsmith and R. Sloan. The conference paper assignment problem. In *Proc. AAI Workshop on Preference Handling for Artificial Intelligence*, 2007
- C. J. Taylor. On the optimal assignment of conference papers to reviewers. Technical Report MS-CIS-08-30, University of Pennsylvania, 2008
- Don Conry, Yehuda Koren, Naren Ramakrishnan: Recommender systems for the conference paper assignment problem. *RecSys 2009*: 357-360
- Naveen Garg, Telikepalli Kavitha, Amit Kumar, Kurt Mehlhorn, Julián Mestre: Assigning Papers to Referees. *Algorithmica* 58(1): 119-136 (2010)
- Chong Wang, David M. Blei: Collaborative topic modeling for recommending scientific articles. *KDD 2011*: 448-456

Our differentiators

- **Visual big-picture interface**
 - See **more** documents: utilize screen space limit better
 - See **relationships**: inter-paper, topic/clusters,
 - Relevance based on **full content**, not just a few keywords
- **Personalized and persistent**
 - **Feedback** from the user
 - **Machine learning** under the hood:
 1. 2-d projection
 2. topical clustering
 3. recommendation
 4. neighborhood graphs
 5. classification
- **Interactive speed**: Only possible due to **fast algorithms**

VisIRR

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

The screenshot displays the VisIRR application interface. On the left, there are two main configuration panels: 'Settings' and 'LB1: NMF, TSTG'. The 'Settings' panel includes options for 'Grouping Options' (NMF), 'Clusters' (10), 'Algorithm' (HALS), 'Max Iteration' (200), 'Visualization Options' (#Dimensions: 2, Option: 1), 'Regularization' (UI: Slider, Regularization Value: 10), and 'Recommendation Options' (Based on: Content, #Iterations: 3, Decaying Factor: 0.7). The 'LB1: NMF, TSTG' panel includes 'Recommended Documents' and 'Edges' (Content, Citation, Co-authorship). The main area shows a network visualization with 10 clusters (1-10) and 100 recommended items. A 'Document View' window is open on the right, showing a document snippet with highlighted keywords and an abstract. The bottom section displays a table of retrieved and recommended documents.

386 Items Retrieved				100 Items Recommended						
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...	9,193...	
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...	
4294824	Paper	Align human interactions with phenoms to identify causative genes and networks und...	Xuebing Wu, Qi and Lu, Rui Jiang	2009	Bioinformatics/computer ...	8	Motivatio...	Gene Netwo...	8,524...	Highly DisLike (1)
4408687	Paper	Improved genetic algorithm inspired by biological evolution	P. Kumar, D. Gospodaric, P. Bauer	2007	Soft Computing	3	The process...	Biological Ev...	7,091...	Highly Like (5)
1826631	Paper	Ontology-Based Support for Human Disease Study	Maja Hadzic, Elizabeth Chang	2005	Hawaii International Conf...	11	In this page...	Depressive ...	6,362...	
4291746	Paper	Identifying gene-disease associations using centrality on a literature mined gene-inter...	Arzucan Özgür, Thuy Vu, Günes Erkan, Dragomir ...	2008	Intelligent Systems in Mol...	26	Motivatio...	Candidate G...	6,218...	
4755093	Paper	Gene-disease relationship discovery based on model-driven data integration and data...	S. Yilmaz, P. Jonveaux, C. Bicep, L. Pierron, Malika...	2009	Bioinformatics/computer ...	2	Motivatio...	Data Integri...	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 Bl...	1	Constructi...	Colon Canc...	5,070...	
4234635	Paper	CDMiner: 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)
4428698	Paper	Medical ontologies to support human disease research and control	Maja Hadzic, Elizabeth Chang	2005	International Journal of ...	4	In this page...	Human Dis...	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 interactio...	Jianzhen Xu, Yongjin Li	2006	Bioinformatics/computer ...	51	Motivatio...	Cross Valid...	4,363...	
4755021	Paper	A Classifier-based approach to identify genetic similarities between diseases	Marc Schaub, Irene Kaplow, Marina Sirota, Chuon...	2009	Bioinformatics/computer ...	4	Motivatio...	Genetic Simil...	4,295...	No opinion (3)
6065805	Paper	Phenotypic categorization of genetic skin diseases reveals new relations between phe...	Ruslan Sadreyev, Jamison Feramisco, Hensin Tsa...	2009	Bioinformatics/computer ...	2	Motivatio...	Genetics, SKI...	4,240...	
4746056	Paper	Fast Mutation Operator Applied in Detector Generating Strategy	Xingbao Lu, Zixing Cai, Chixin Xiao	2008	International Conference ...	0	Inspired by ...	Artificial Im...	4,138...	
89623	Paper	An experimental evaluation of selective mutation	A. Offutt, Gregg Roethermel, Christian Zapf	1993	International Conference ...	64	Mutatio...	Experimen...	4,031...	
50643	Paper	BD Controller Optimization Based on the Self-Organization Genetic Algorithm with C...	Zhao, Jieyu, Zhuoan, Liu, Du, Hailiang, Wang, Su...	2007	Mexican International Co...	2	In this page...	Analysis of ...	3,864...	

Visualization Example of Queried Set

Keyword query, 'dimension reduction'

Query

- 1 : data,dimensional,reduction,dim
- 2 : algorithm,reduction,dimension,d
- 3 : analysis,dimension,reduction,d
- 4 : features,reduction,dimension,p
- 5 : dimension,reduction,dimension
- 6 : model,reduction,dimension,dim
- 7 : method,dimension,reduction,pr
- 8 : gene,expression,data,microarra
- 9 : power,performance,reduction,h
- 10 : image,dimension,reduction,fea

Computational zoom-in

Clear topics

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-I...	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	

Update Recommendation

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 recommended documents, with nodes representing documents and edges representing relationships. The nodes are colored and sized, and some are enclosed in rectangles. The interface includes a sidebar with settings for recommended documents, a list of retrieved items, and a list of recommended items.

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...	11.5787724...	
4417730	Paper	Visual Methods for Analyzing Time-Oriented Data	Wolfgang Aigner, Silvia Miksch, Wolf...	2008	IEEE Transactions on Visualizati...	36	Providing appropriate methods to facili...	Analytical Method, Dat...	10.3457933...	
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...	8.83127012...	
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...	6.85063852...	
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...	6.34965008...	
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	2002	IEEE Transactions on Visualizati...	365	Context and history visualization plays ...	Data Mining, Data Type...	5.96365931...	
2518079	Paper	Interactive Visual Analysis of Families of Function Graphs	Zoltan Konyha, Kresimir MatkovicMem...	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 Lammarsch, Wolfgang Aigner, Ale...	2009	International Conference on Inf...	1	Many real-world problems involve time...	Interactive Visualizati...	5.75226473...	
4408123	Paper	Trajectory-based visual analysis of large financial time series data	Tobias Schreck, Tatiana Telusova, Jo...	2007	SigKDD Explorations	14	Visual Analytics seeks to combine auto...	Applications of Visuals...	5.6162512...	
441773	Paper	A Visual Language for Internet-Based Data Mining and Data Visualization	Jatun Chattrachai, Vike Guo, Jam...	1999	Visual Languages/Human-Centri...	3	This paper describes a novel applica...	Data Mining, Data Visu...	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...	
476640	Doc	Descriptive Visual Analytics	Christine Chabot	2006	IEEE Computer Graphics and An...	3	Sketching Chabot addresses the open...	Business Intelligence, C...	4.0481464...	

Features

•Dynamic query-retrieval

- Keyword search on contents such as title, abstract, and keywords as well as author and venue fields
- Filtering on year, citation/reference count
- Different queries created either separately or jointly with their own visualization snapshots

•Interactive visualization

- Multiple visualizations via dimension reduction (for 2d coordinate) and clustering (for color-coded summary) on dynamically retrieved sets
- Support for easy comparison between views via clustering and dimension reduction alignment

•Preference feedback and recommendation

- Document preference assigned by users
- Recommendation performed based on document contents, citation, or co-authorship information
- Recommended items projected into the same space along with their predicted cluster labels

Large-scale Data Collection/Ingestion

•Data collection

- Starting with DBLP data set (432,605 data items)
- Data cleanup and missing value handling via Microsoft Academic Search API
- Title, author, year, venue, abstract, keywords, citation/reference count, and citation network info

•Data management

- Structured information stored in database
- Term-document information pre-computed
- Top K Cosine similarity pre-computed
- Citation network and co-authorship network pre-built
- Scalable streaming data handling with efficient update in $O(n)$

•Dynamic memory loading

- Document information dynamically loaded on the fly depending on user queries/interactions
- Cache-like memory management using “least recently used” approach

Graph-based Recommendation

• Various recommendation schemes

- Content-, co-authorship-, and citation-based recommendation supported

• Heat-kernel-based propagation algorithm

- Weighted graphs as an input (for content-based, k -NN cosine-similarity graph)
- User preference propagated efficiently on large-scale sparse graphs

$$r_\alpha = \alpha \sum_k (1 - \alpha)^k f W^k$$

- r_α is a recommendation score vector with a control parameter α , and f is a user-assigned rating, and W is an input graph.

• Embedding on existing visualization

- Out-of-sample embedding into previously computed dimension reduction
- Color-coding using k -NN classification on previous clusters

Summary

- Foundational algorithms for visual representations of high dimensional, large scale, heterogeneous data
(dimension reduction, clustering, space alignment)
- Fast algorithms for real time interaction
- Development of VA testbed
- Development of proof-of-concept VA system