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

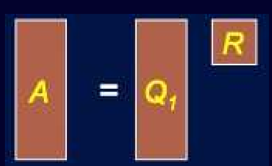
Key Foundational Components for VA System Development

- 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)

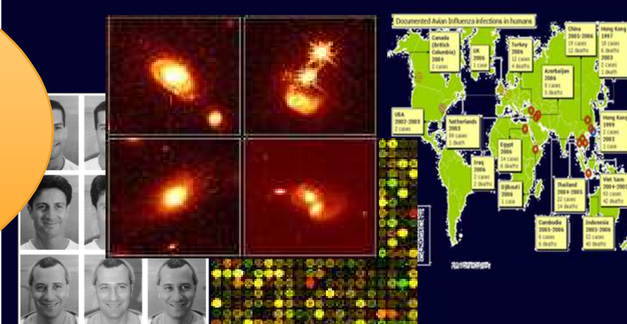


FODAVA Fundamental Research

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

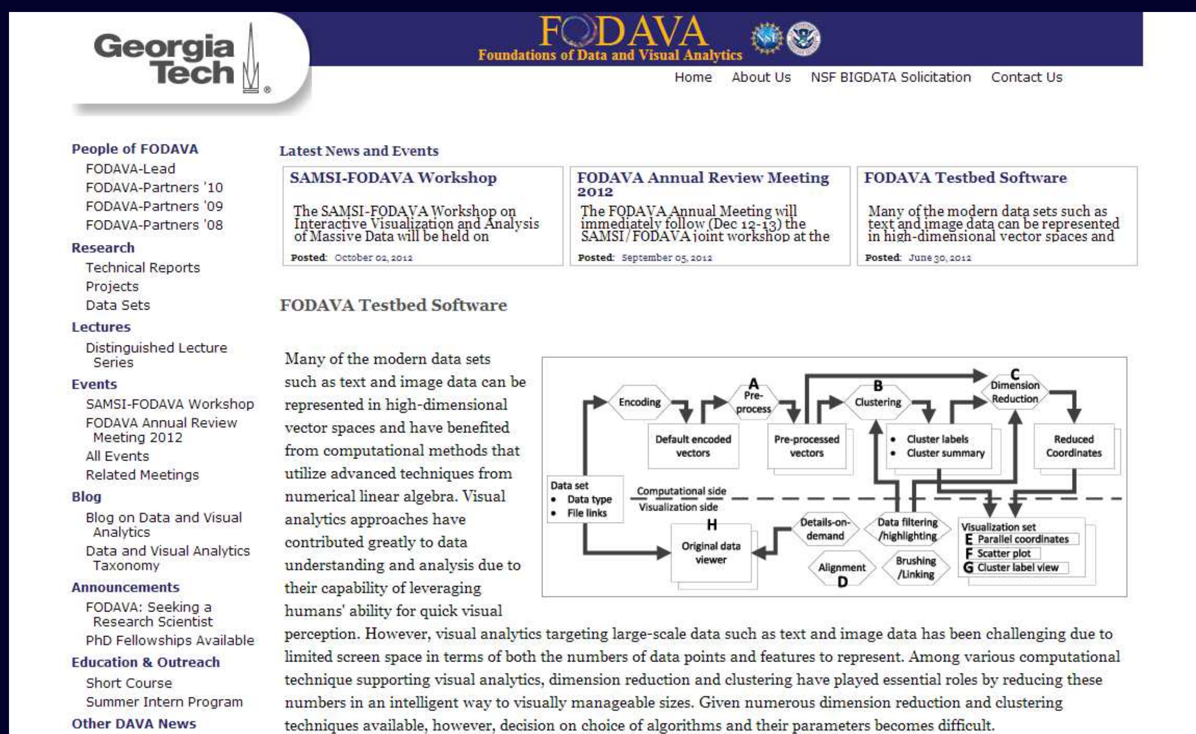
Test Bed

Applications



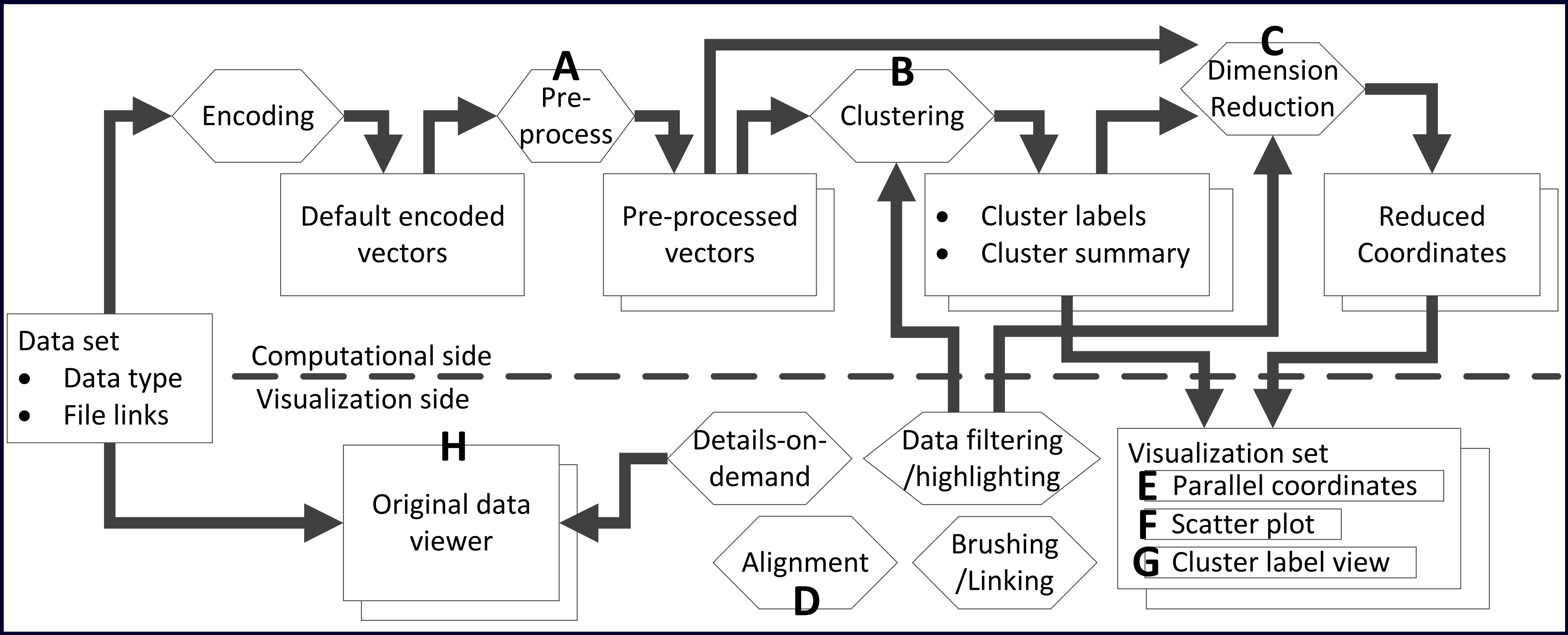
FODAVA Research Testbed Software: Available at `http://fodava.gatech.edu/fodava-testbed-software`

- 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



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

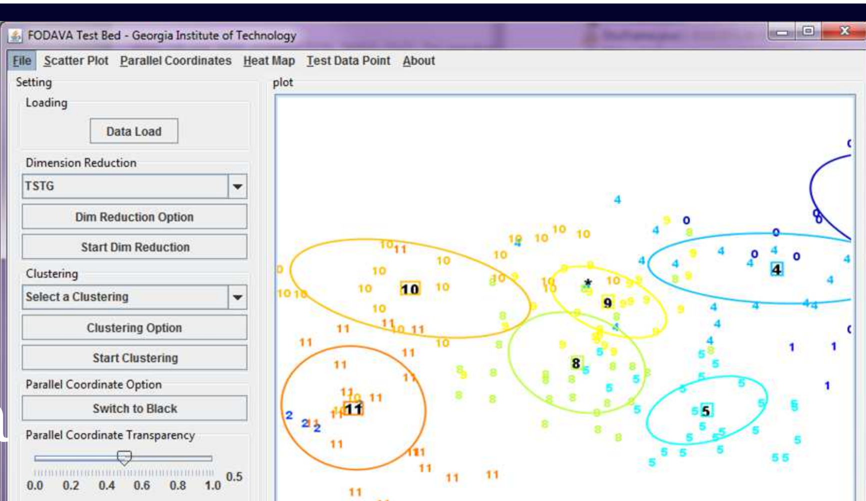
- Visualizes high-dimensional data
  by parallel coordinates and/or scatter plot
- Methods
  - Linear methods
    - PCA, FA, ProbPCA, LDA, OCM, NPE, LPP, LLTSA, NCA, MCML
  - Nonlinear methods
    - MDS, Isomap, LLE, LTSA, Sammon, HessLLE, MVU, LandMVU, KernPCA, GDA, DiffMaps, SPE, AutoEnc, LLC, ManiChart, 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

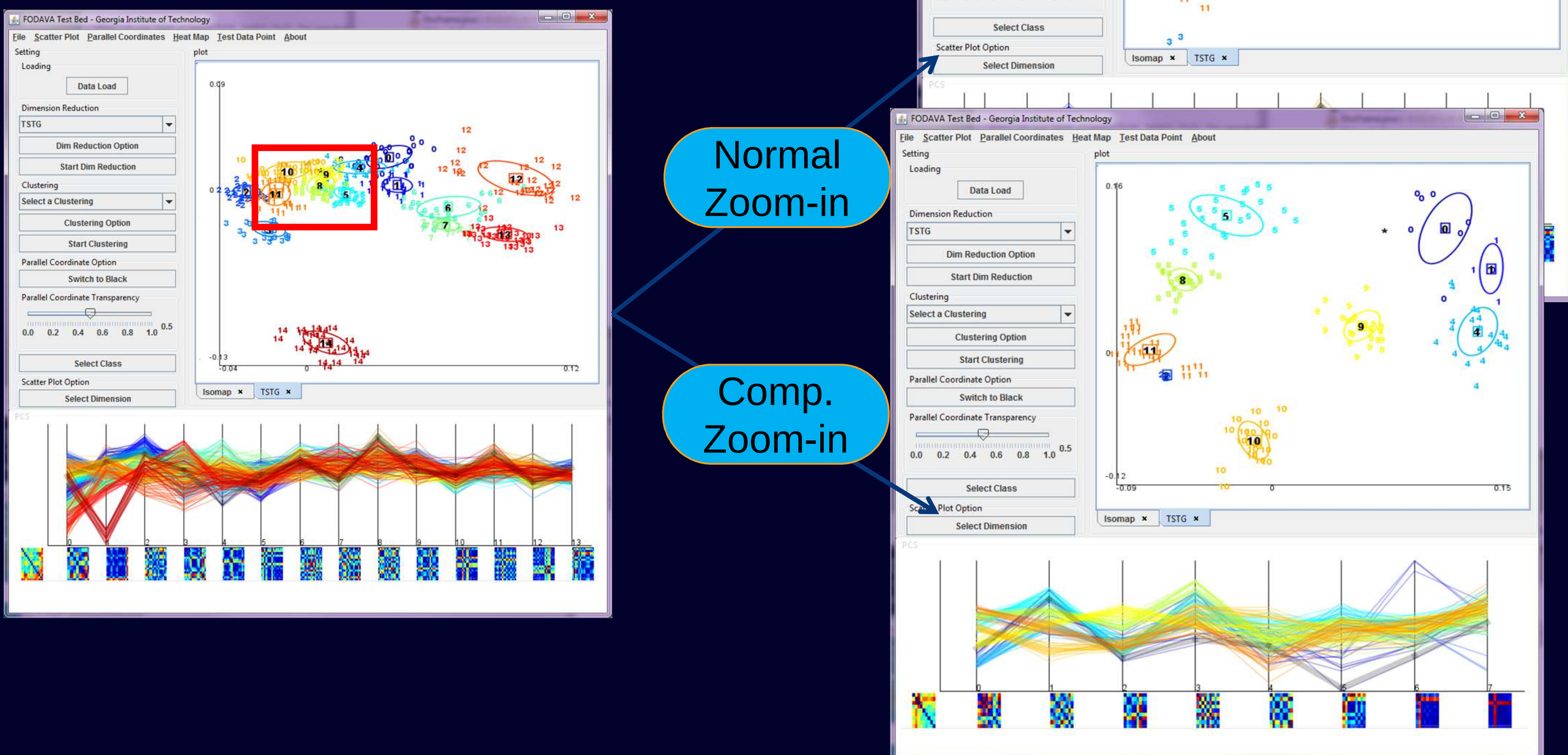
Clustering and Classification

- 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

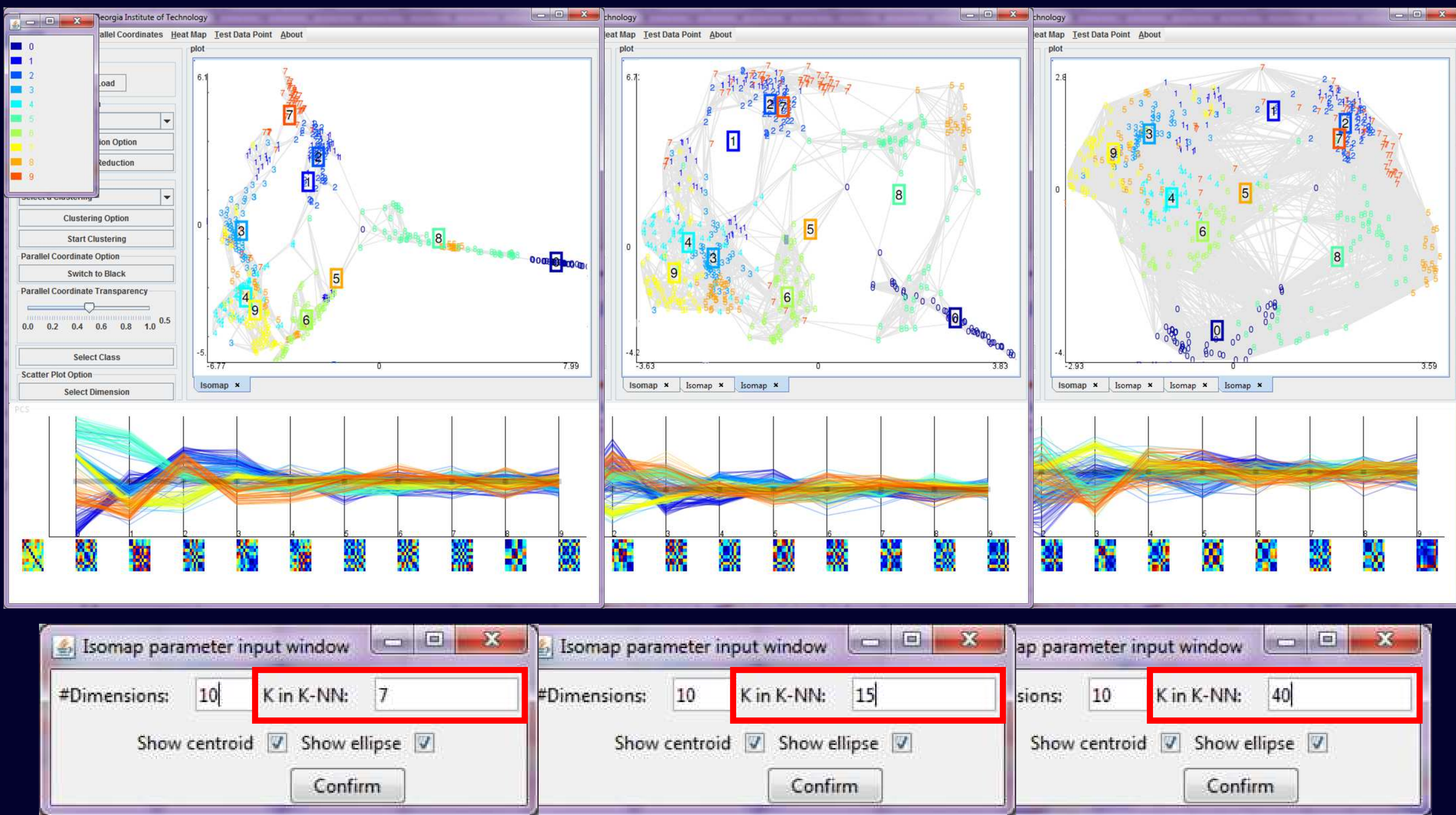




Interactive Parameter Change

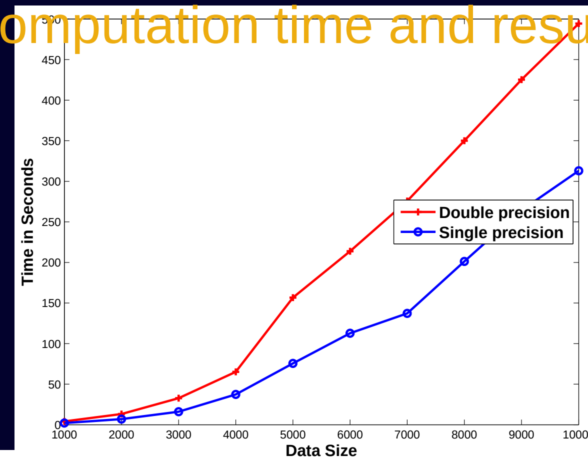
e.g. in Isomap *k* value in *k*-NN Graph

Controls the level of focus on locality

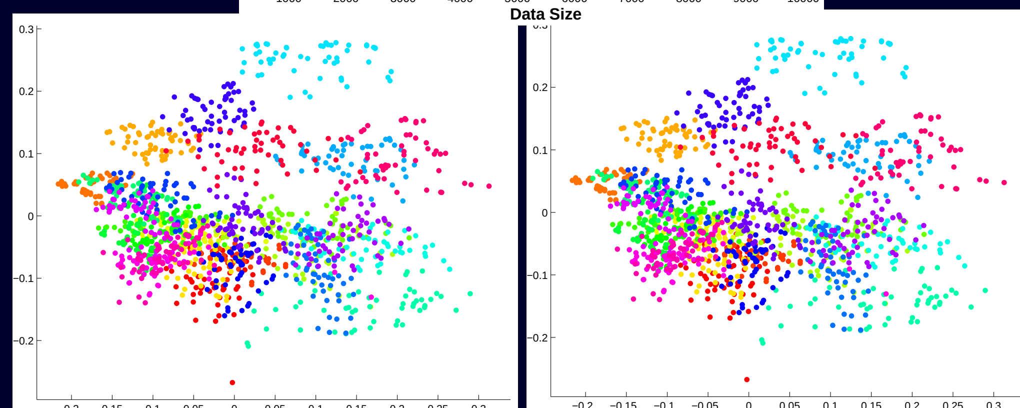


Fast Comp. Modules for Interactive Vis.

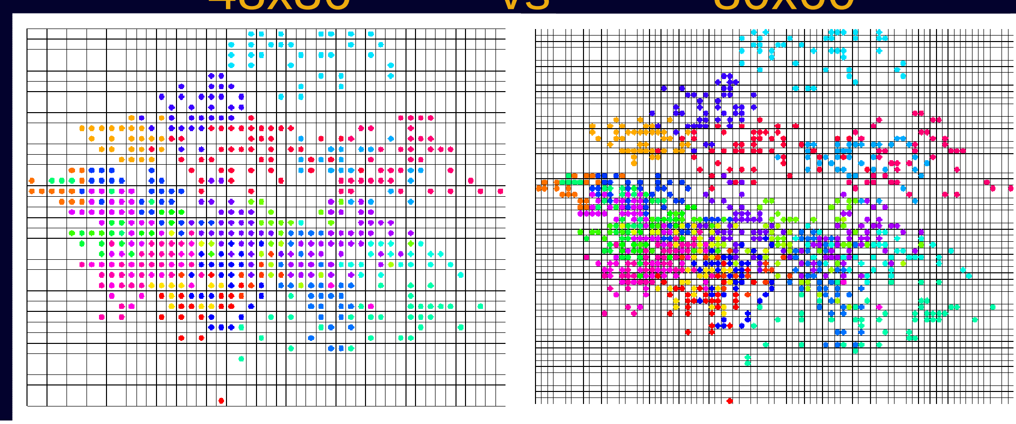
PCA timing: double vs single precision computation time and results

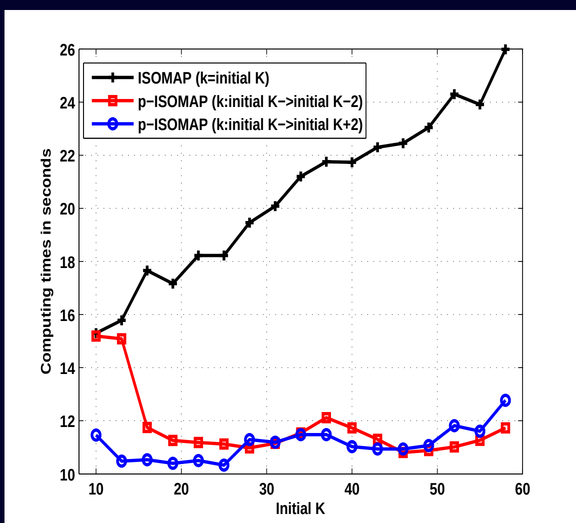


- Essential for real-time interaction
- Let computational precision be governed by visual precision/resolution
- Hierarchical refinement
- Adaptive algorithms



48x36 vs 80x60





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

Key Computational Methods

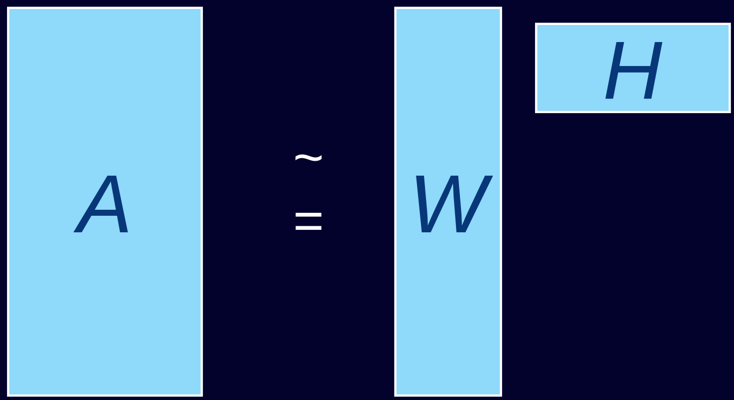
- NMF (Nonnegative Matrix Factorization) and its variations:
  for dimension reduction and clustering

- 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…)



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

$W>=0, H>=0$

- 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	GNMF	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	*Overall*
K-means	0.6734	0.4289	0.6184	0.4650	0.6499	0.7384	*0.5957*
NMF/ANLS	0.8534	0.3770	0.6312	0.4877	0.7020	**0.7912**	*0.6404*
SymNMF	**0.8979**	**0.5305**	**0.7258**	**0.5129**	**0.7798**	0.7517	*0.6998*

  - Problem sizes:

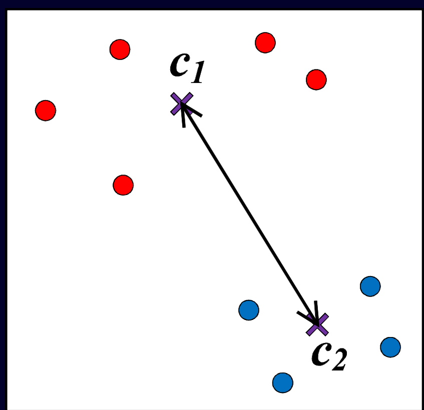
	TDT2	Reuters	COIL20	NIPS	ORL	PIE	20 Newsgroups
m	26618	12998	4096	17583	5796	4096	36982
n	8741	8095	1440	420	400	232	18669
k	20	20	20	9	40	68	20

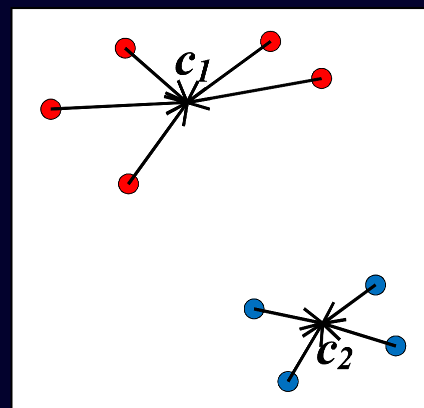
On average, NMF is faster than k-means by a factor of at least 2

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($G^T S_w G$)





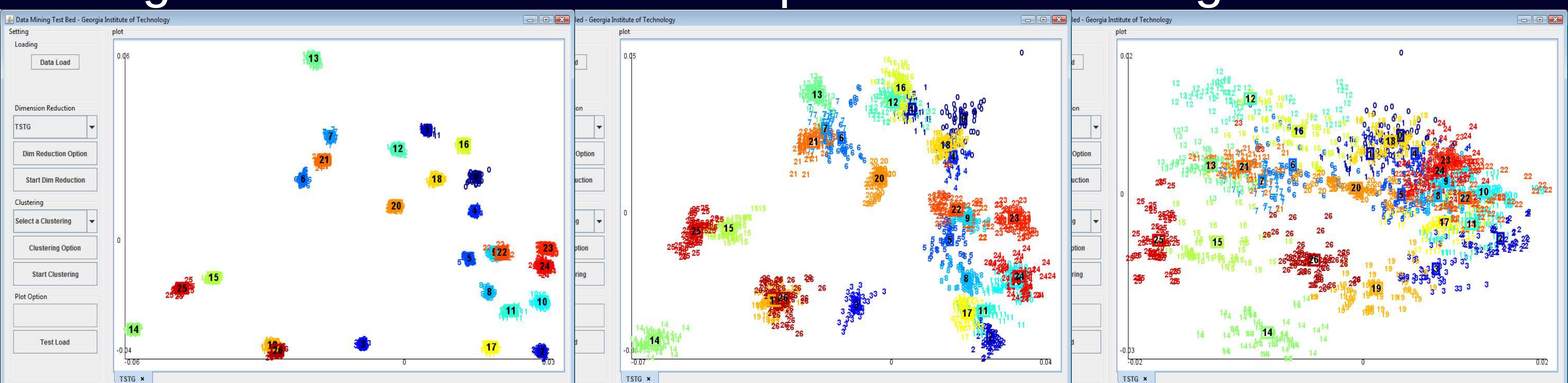
LDA/GSVD

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

max trace

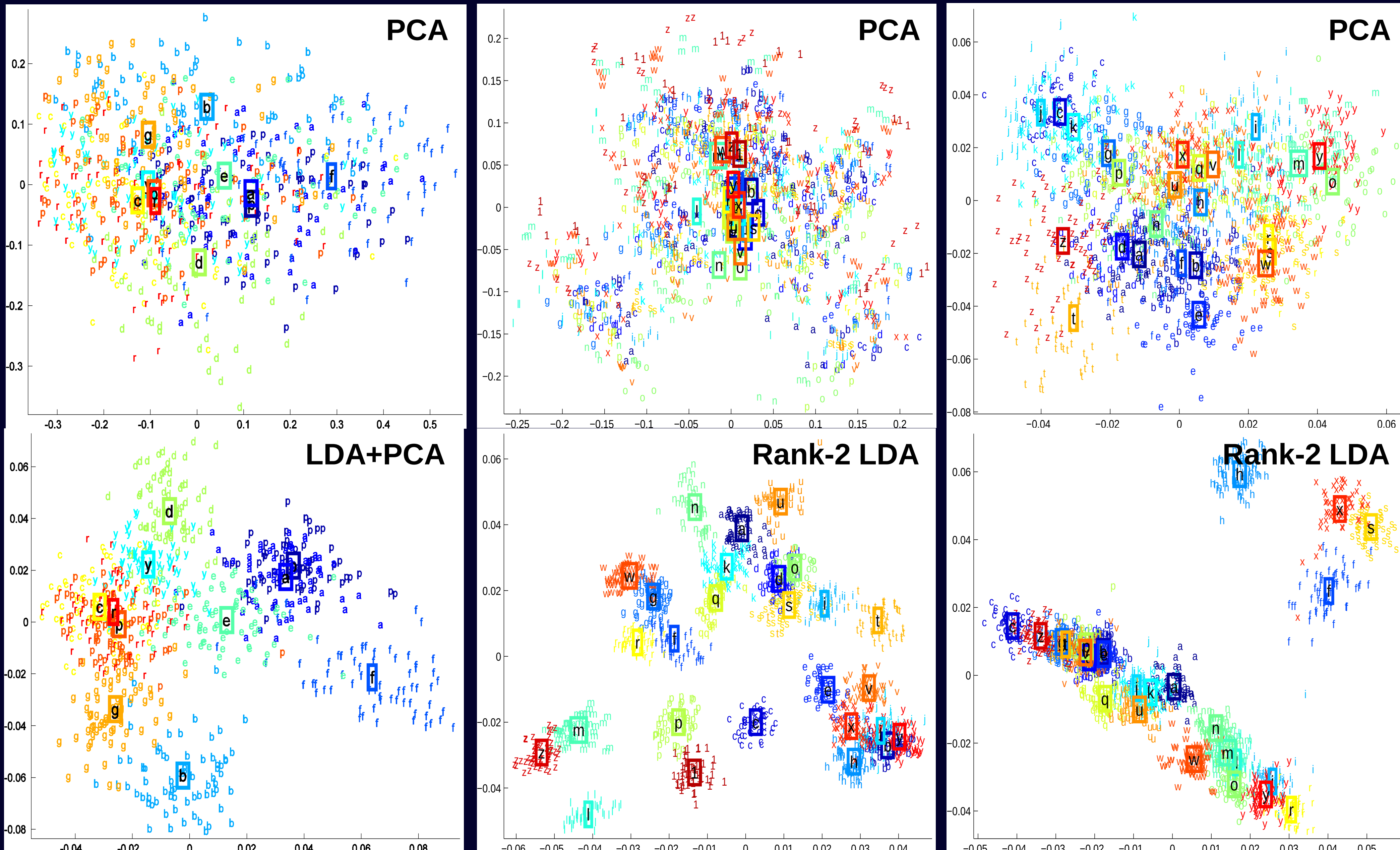
$$(G^T (S_w + \mu I) G)^{-1} (G^T S_b G)$$

- Regularization in LDA for Computational Zooming-in



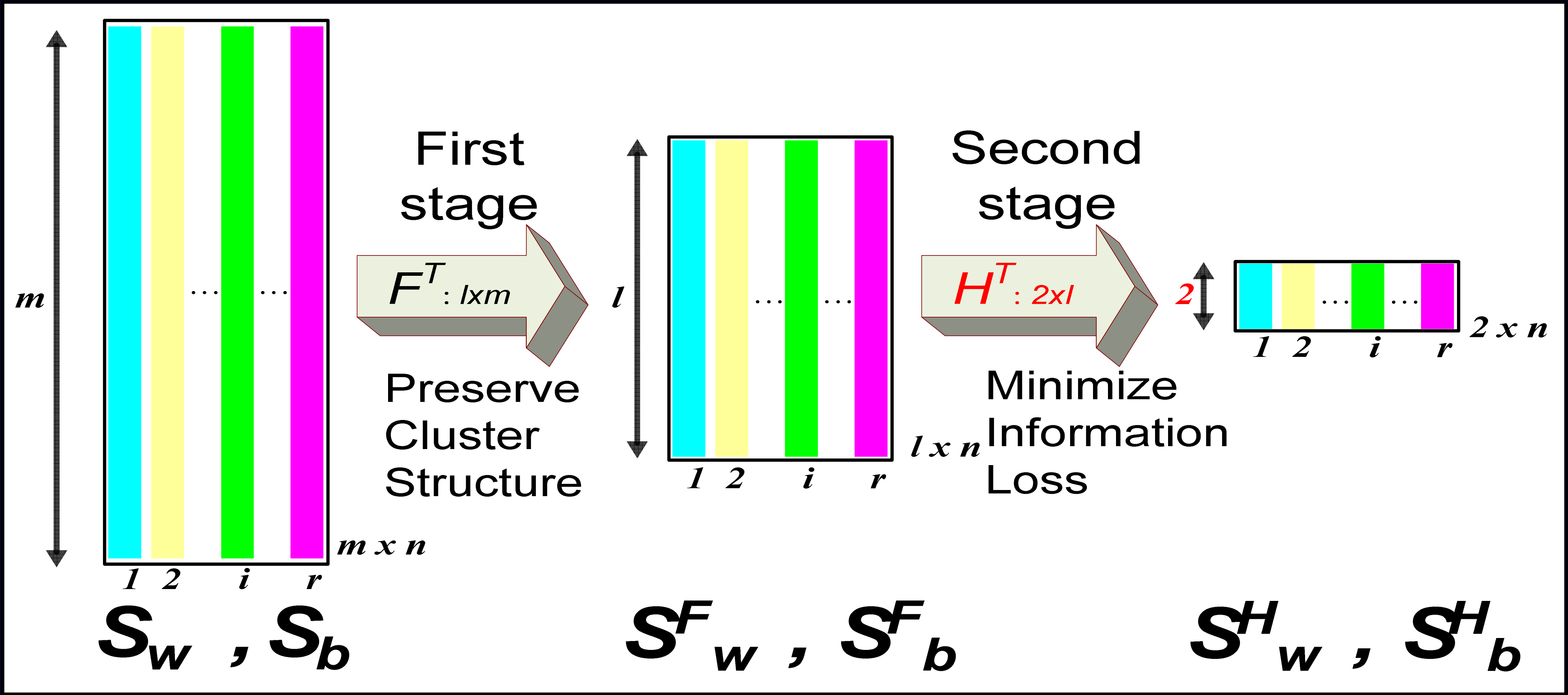
Small regularization → Large regularization

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

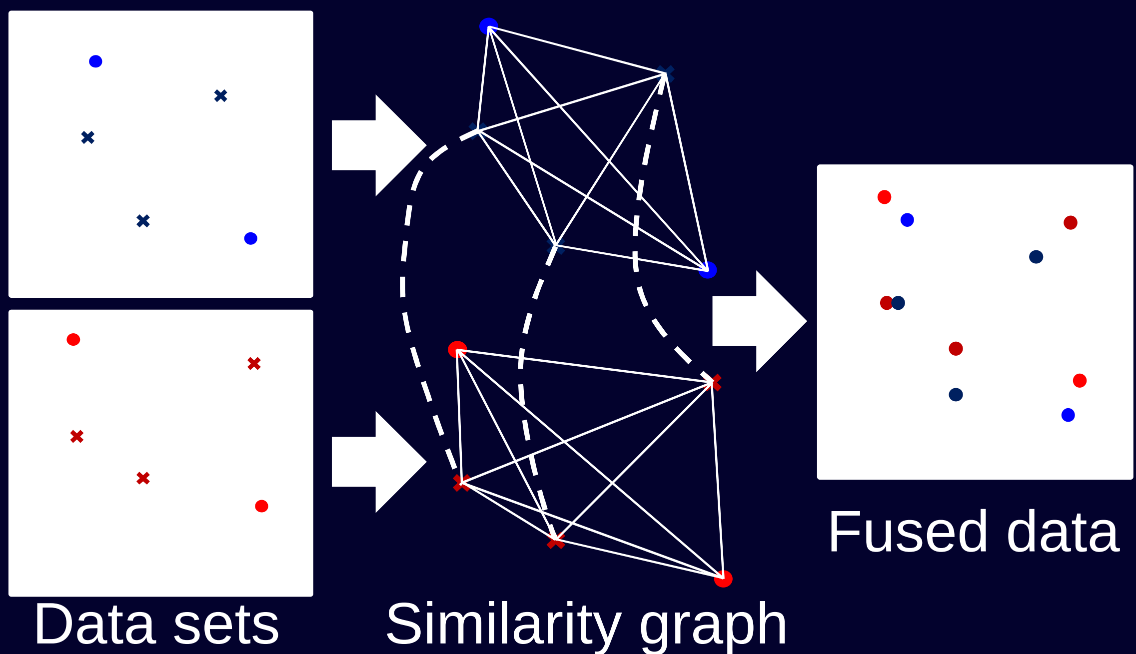


- LDA + LDA = Rank2 LDA
- LDA + PCA
- OCM + PCA
- OCM + Rank-2 PCA on S^F_b = Rank-2 PCA on S_b

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)**



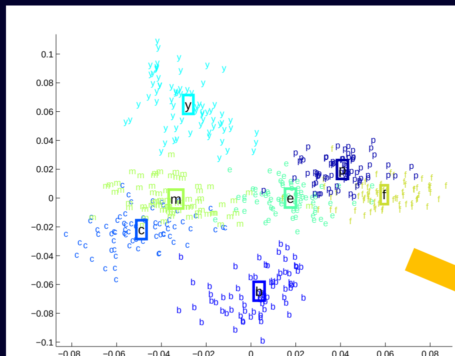
- **Procrustes analysis**

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

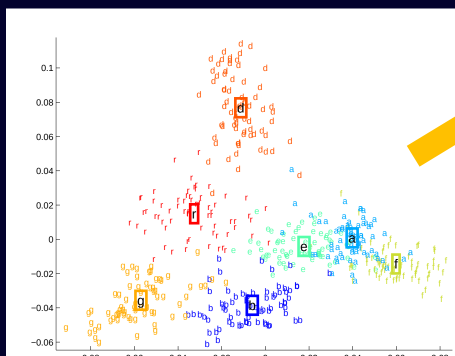
$$Q^T Q = I$$

Fusion and Alignment in Testbed

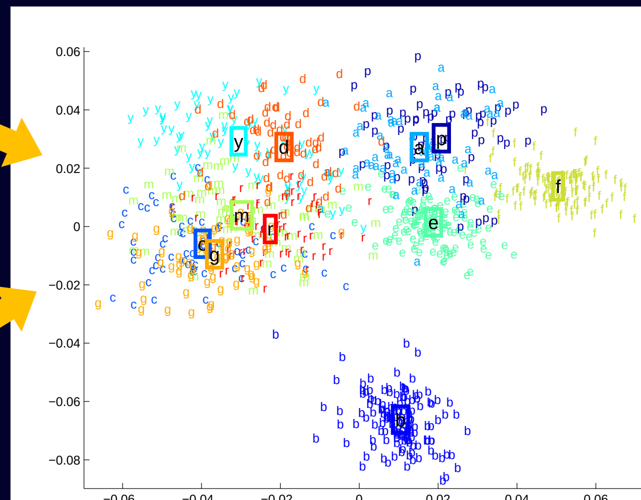
Data set 1



Data set 2

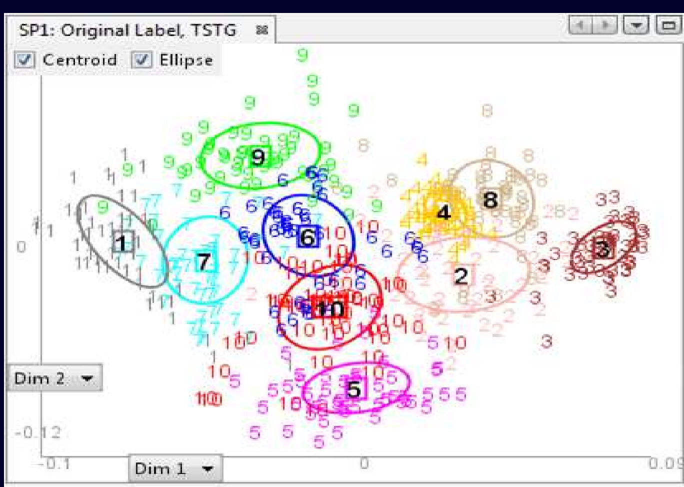


Fused

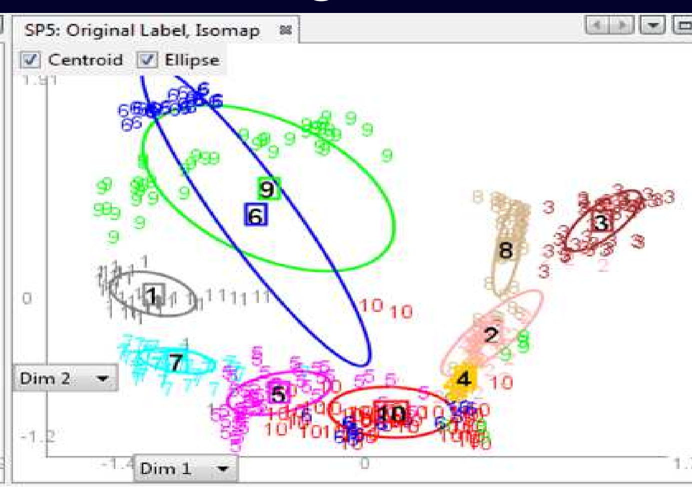


- **Data set 1 only:** **comp.sys.ibm.pc.hardware ('p'), sci_crypt ('y'), soc.religion.christian ('c'), talk.politics.misc ('m')**
- **Data set 2 only:** **comp.sys.mac.hardware ('a'), sci.med ('d'), talk.religion.misc ('r'), talk.politics.guns ('g')**
- **Shared: rec.sport.baseball ('b'), sci.electronics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

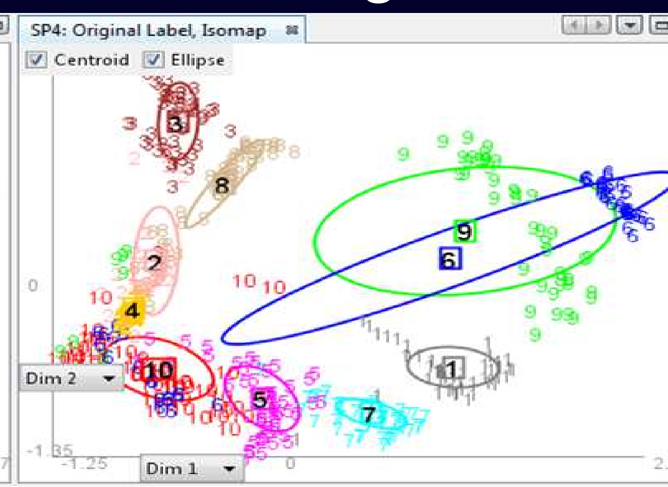
Reference



Aligned

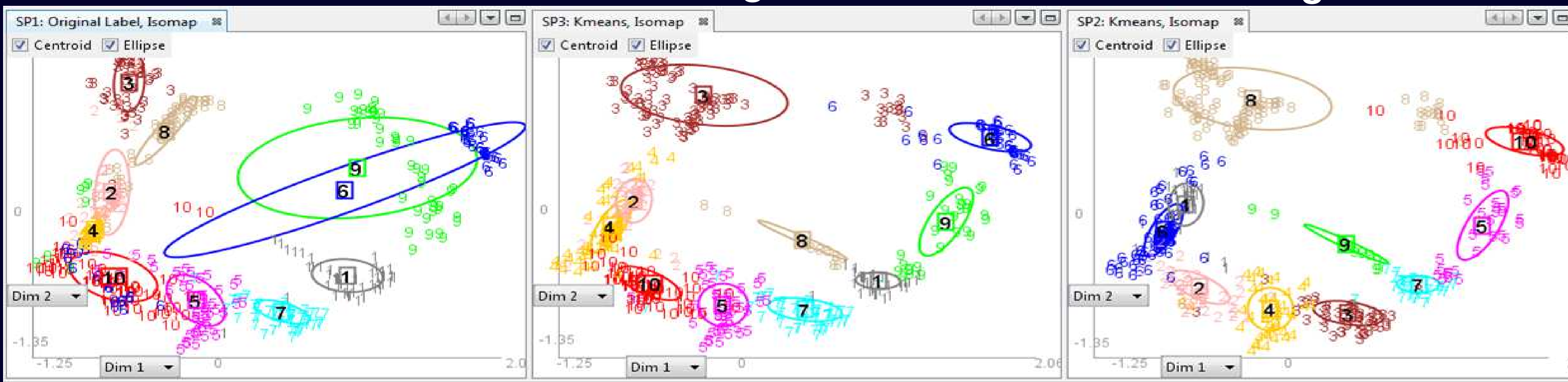


Un-Aligned

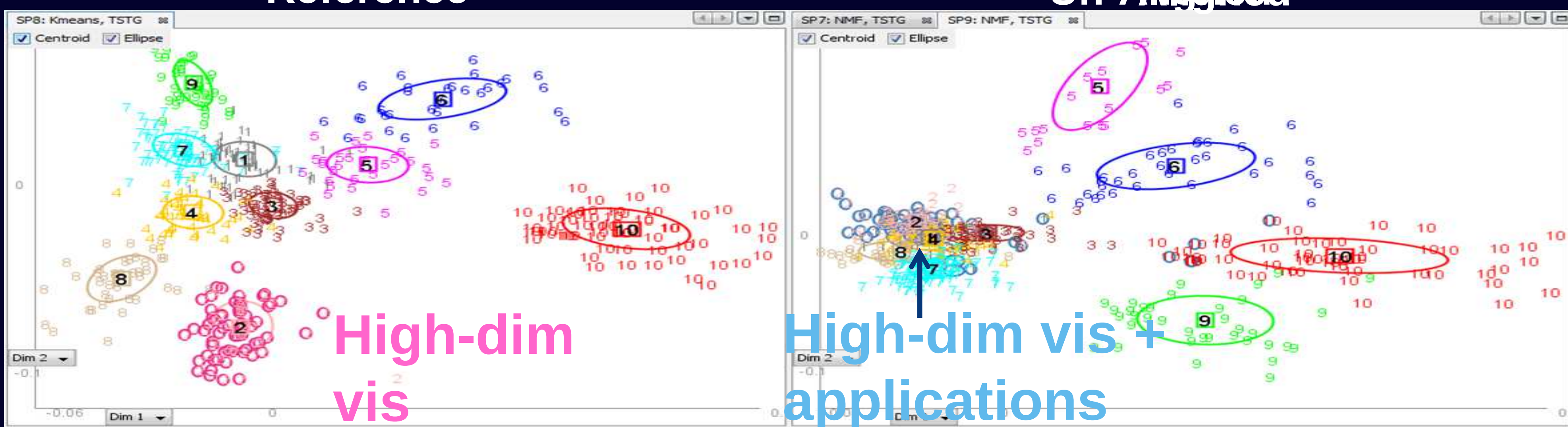


Cluster Alignment: Label Matching and Space Alignment

Reference | Aligned | Un-Aligned



Reference | Un-Aligned / Aligned

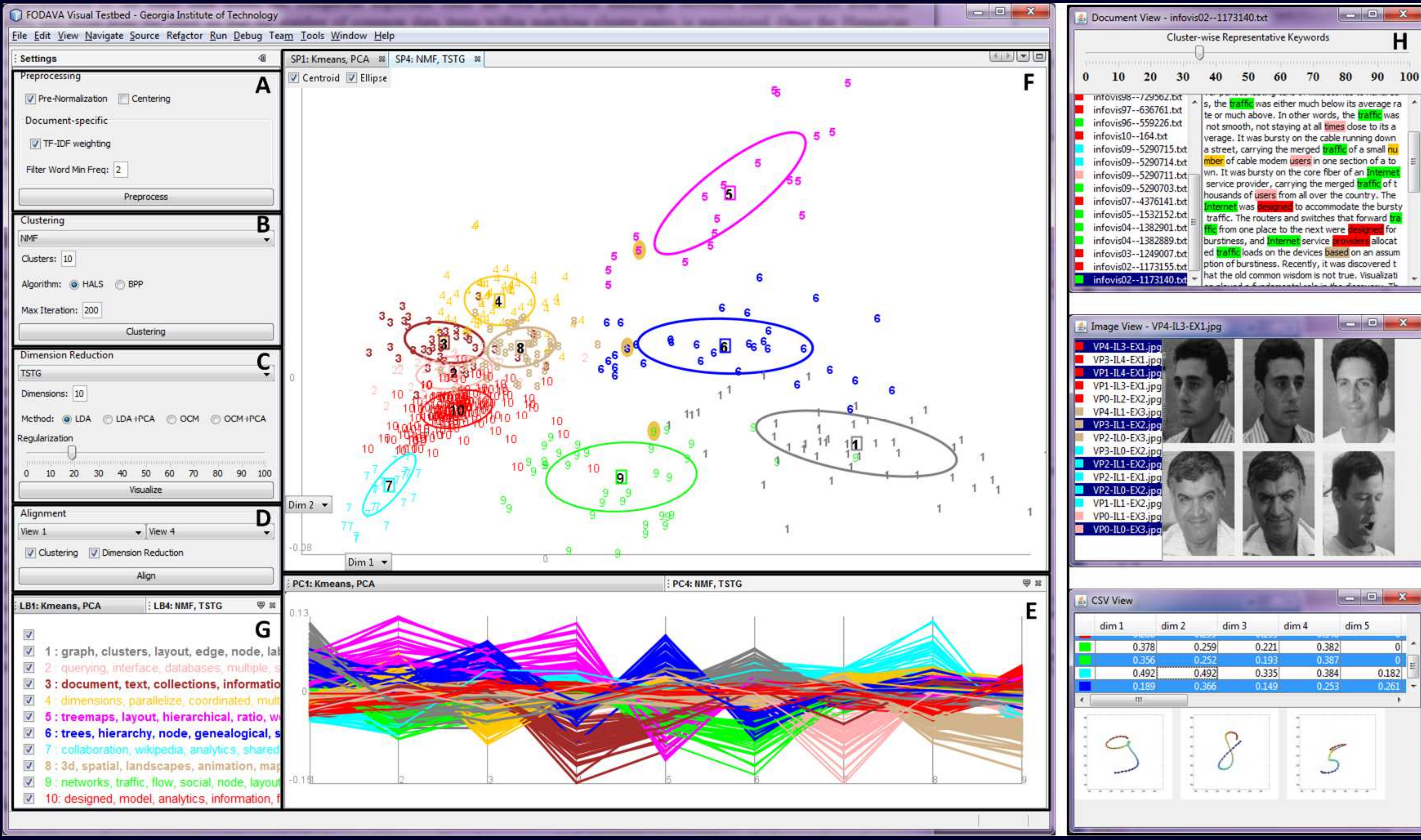


High-dim vis

High-dim vis + applications

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

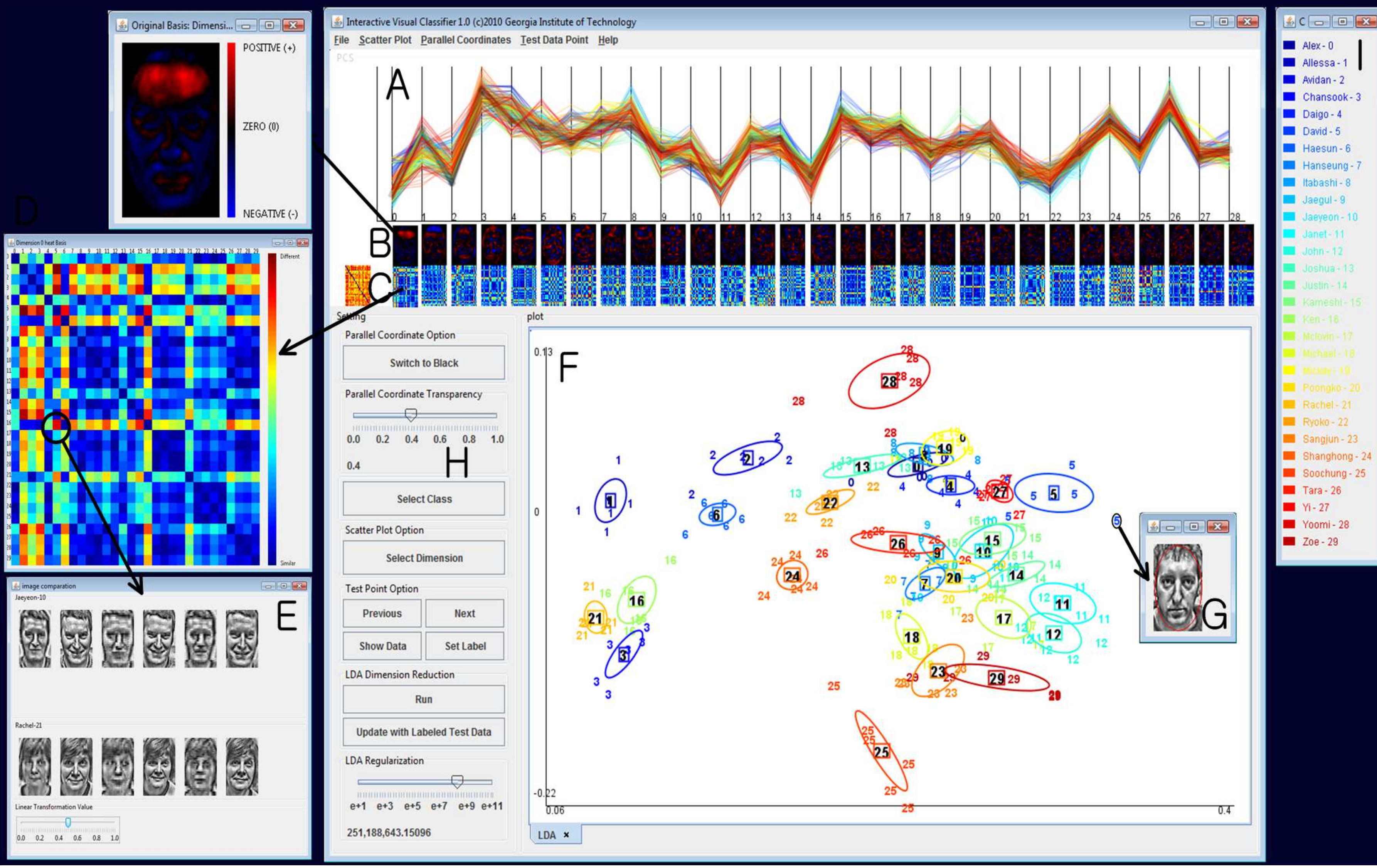
Testbed Overview



iVisClassifier

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

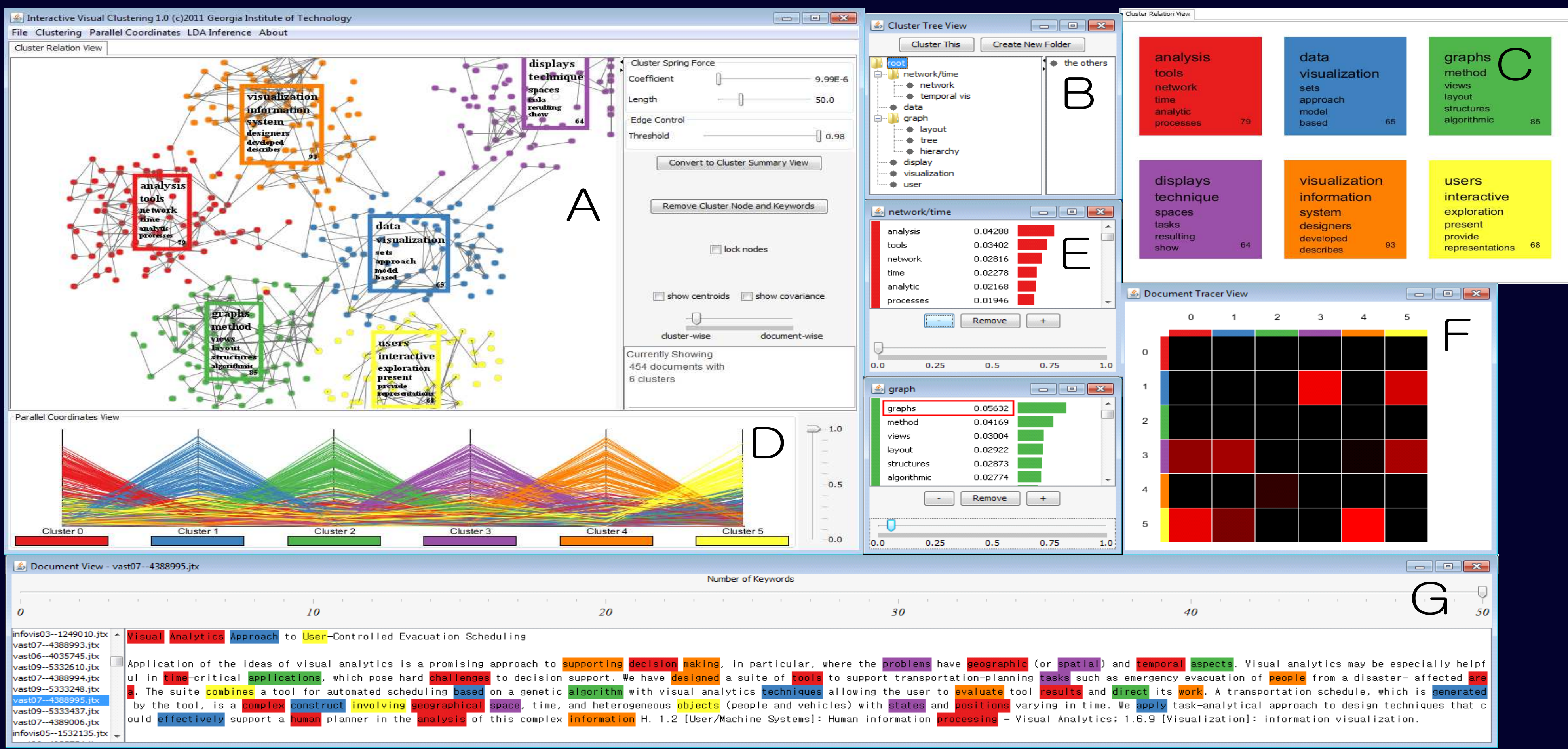
Interactive visual analytics system for **classification** of high-dim. data (image, text, etc) and **search space reduction**



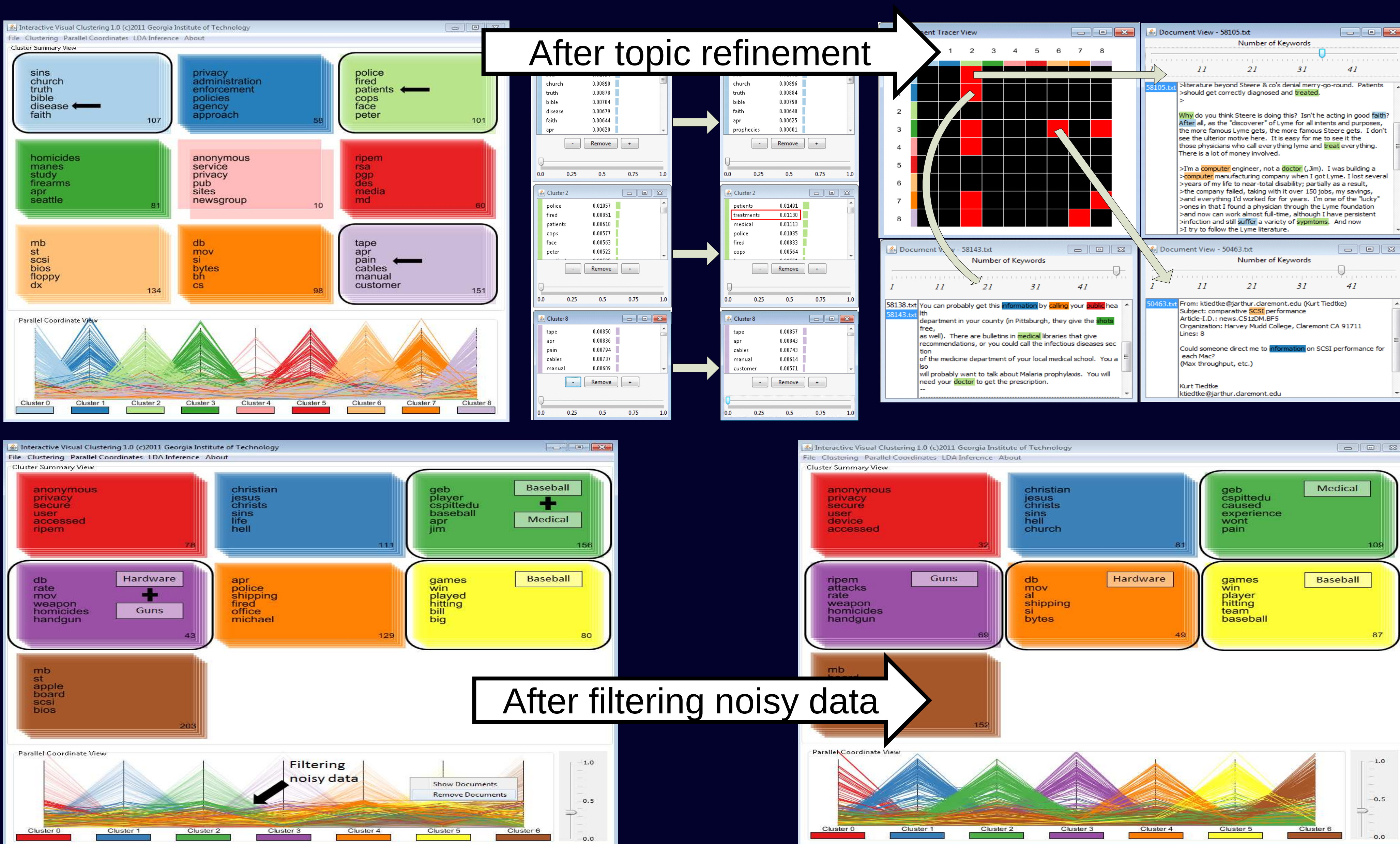
iVisClustering

(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

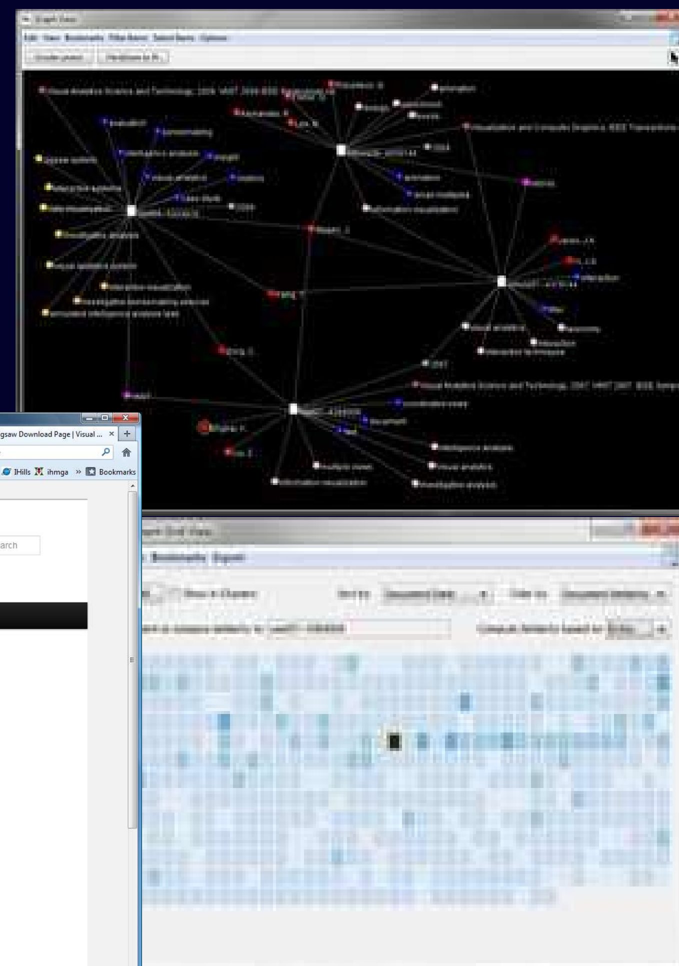


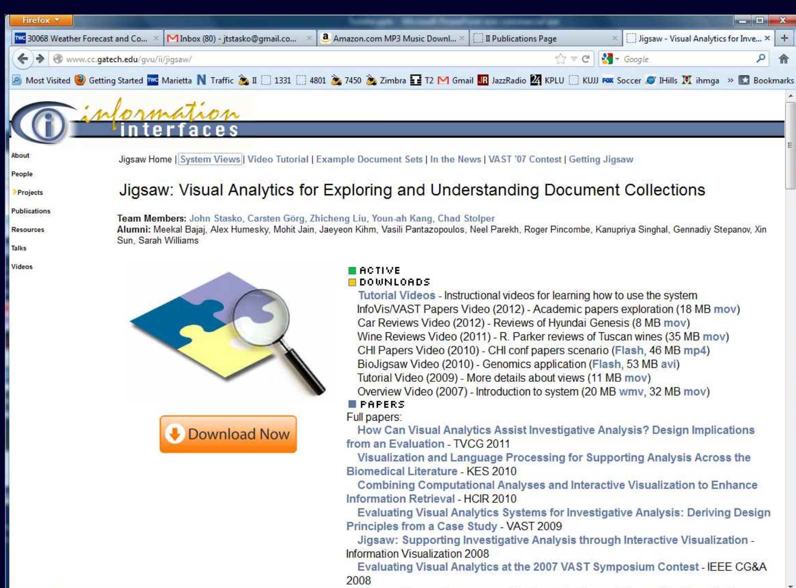
Key Interactions with LDA:
Topic Refinement and Noisy Data Filtering

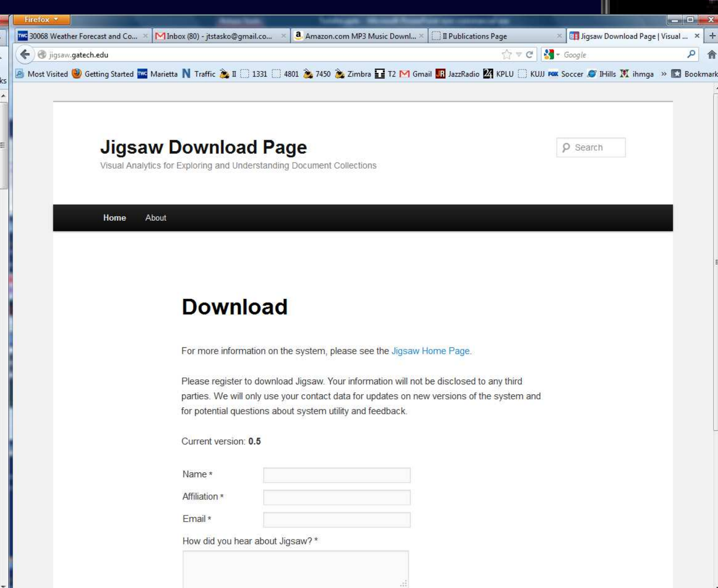


Jigsaw

- 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

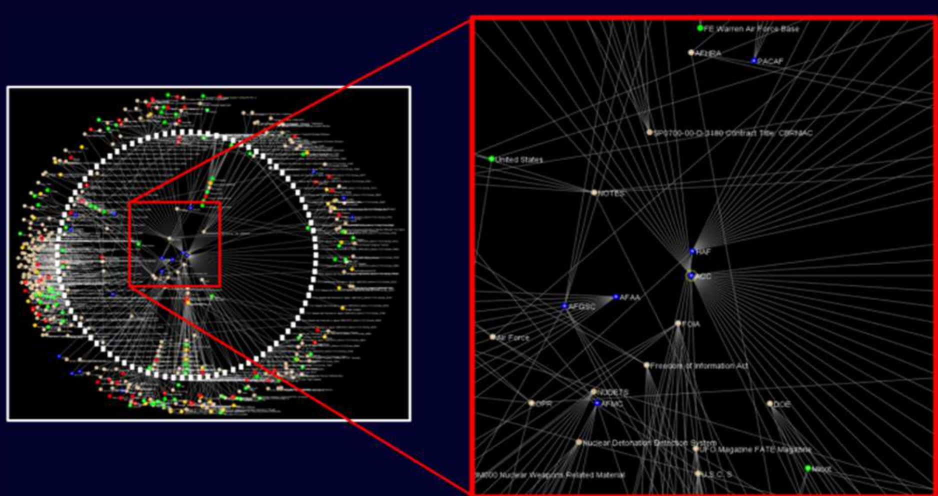


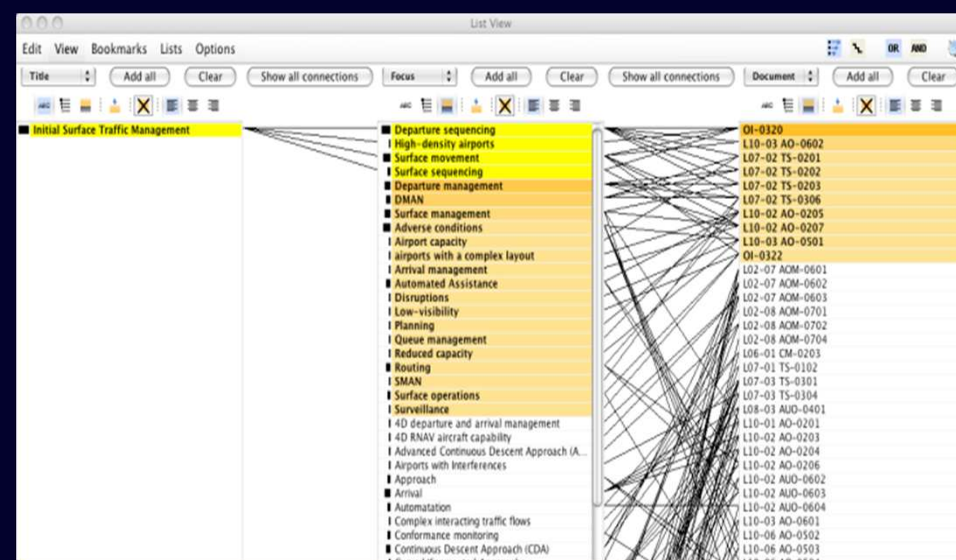




Case Study of System Usage
(Y. Kang & J. Stasko, VAST12)

- 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. AAAI 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)

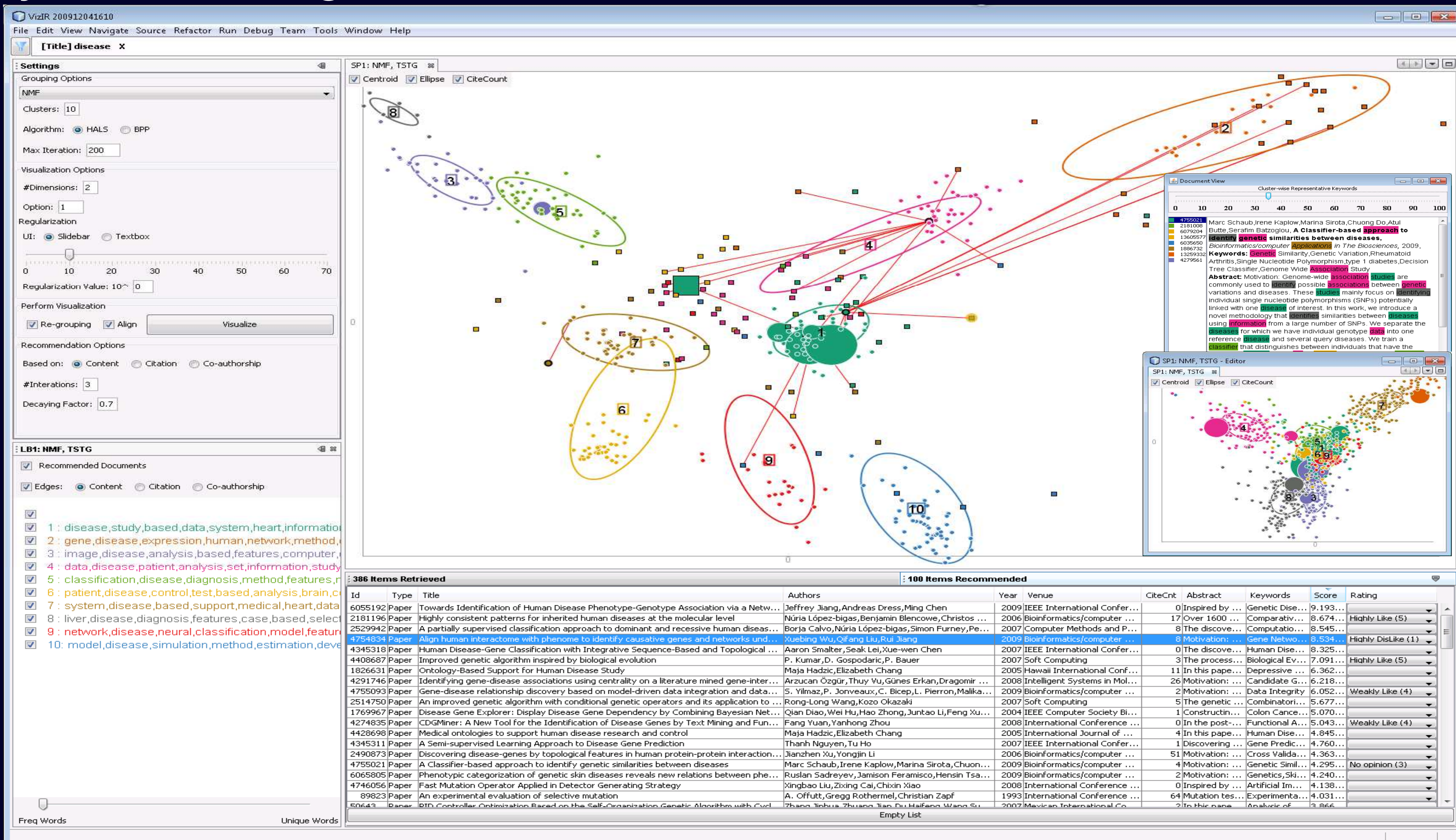
•Chong Wang, David M. Blei: Collaborative topic modeling for recommending scientific articles. KDD 2011: 448-456

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**

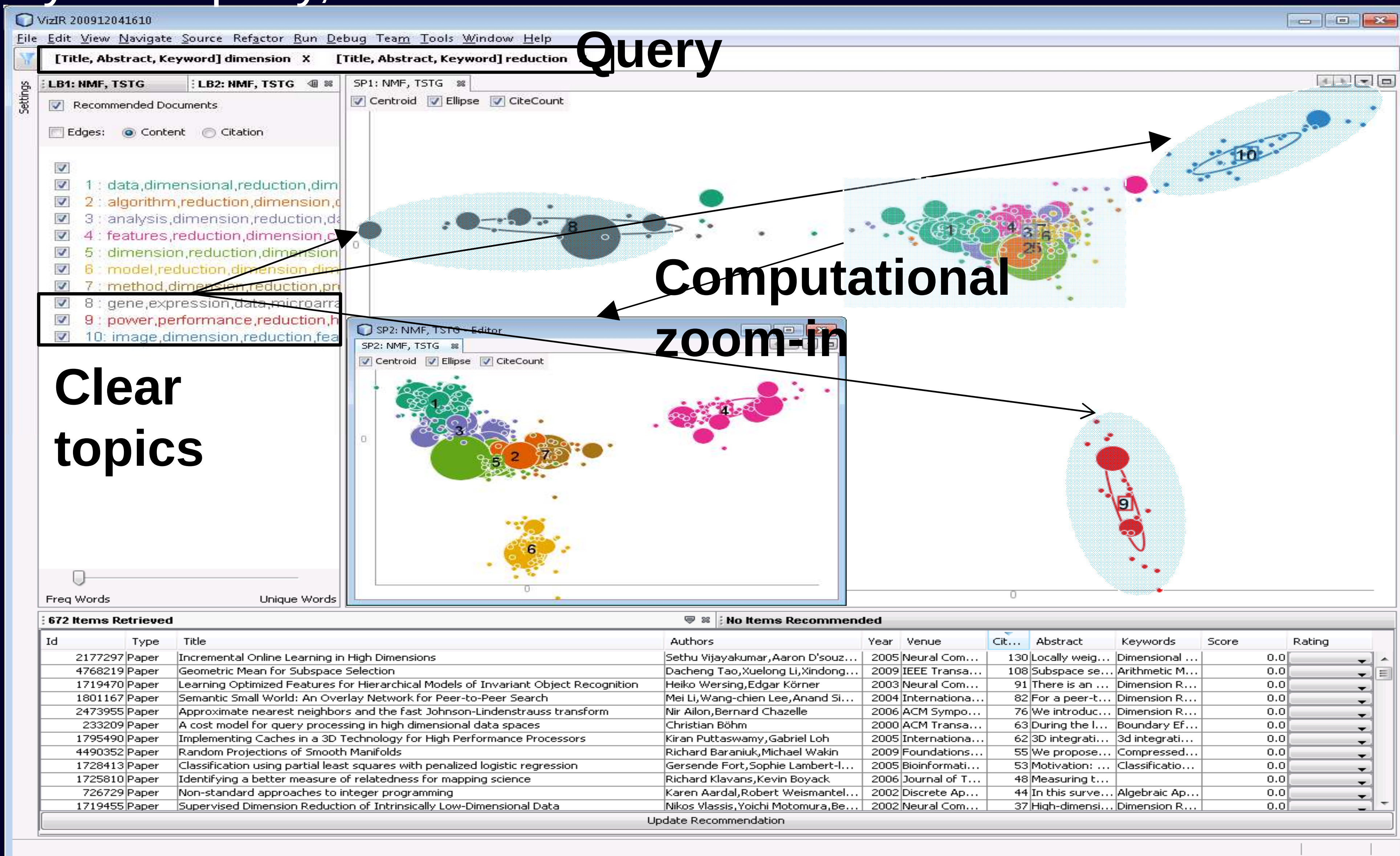
VisIRR

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



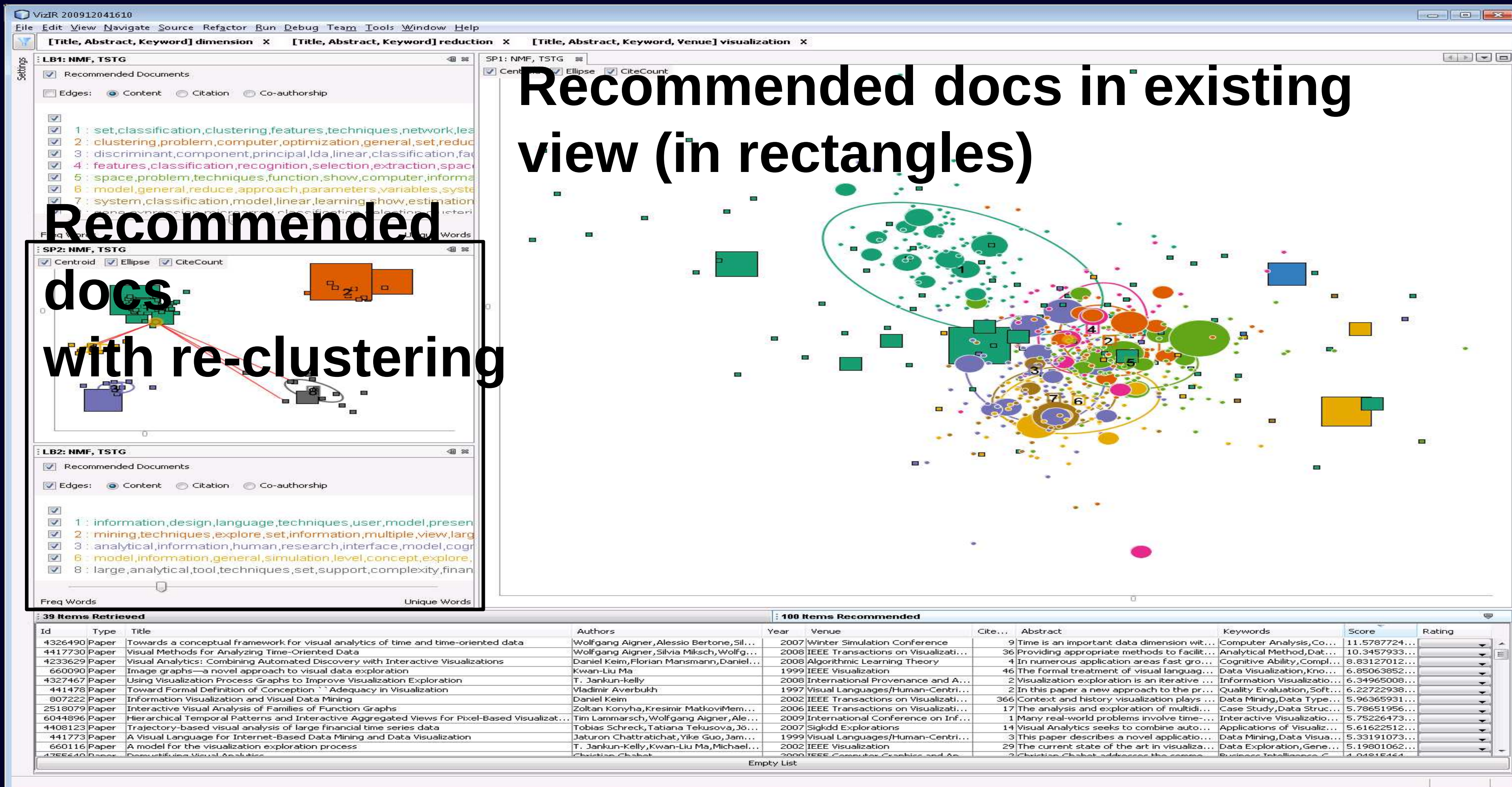
Visualization Example of Queried Set

Keyword query, ‘dimension reduction’



Recommendation Example

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



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