

Big Data Visual Analytics: **Machine Learning** **Meets** **Visualization**

Jaegul Choo

Assistant Professor

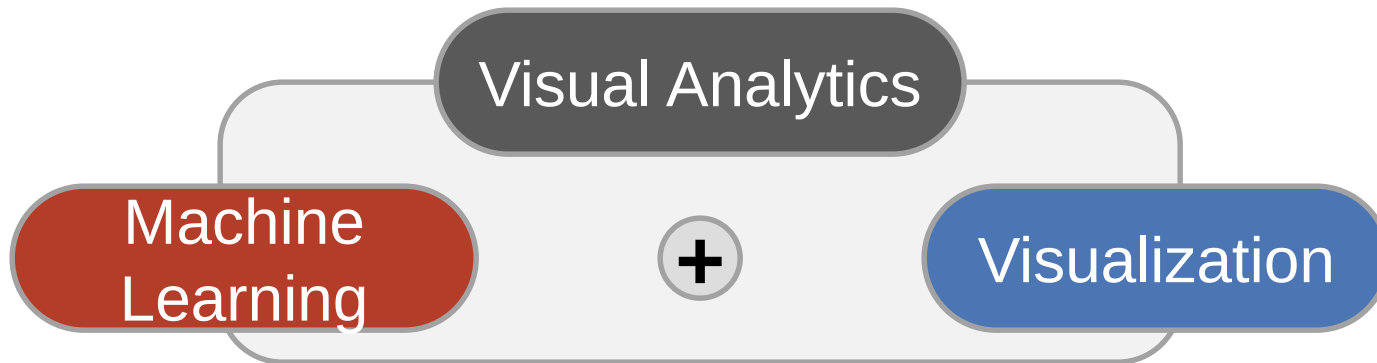
Dept. of Computer Science and Engineering

Korea University

About Me

Google 'Jaegul Choo'

- ▶ Assistant Professor at Computer Science dept. in Korea Univ.
- ▶ B.S. (2001) in Electrical Engineering at SNU
- ▶ M.S. (2009) and Ph.D (2013) at Georgia Tech
- ▶ Main Research



- ▶ Main Expertise: **Dimension Reduction** and **Clustering**
- ▶ Published >50 research articles (>300 citations)

Two Approaches for Data Analysis

Machine
Learning

Visualization

Automated

Interactive (human in the loop)

Clearly defined tasks

Exploratory analysis

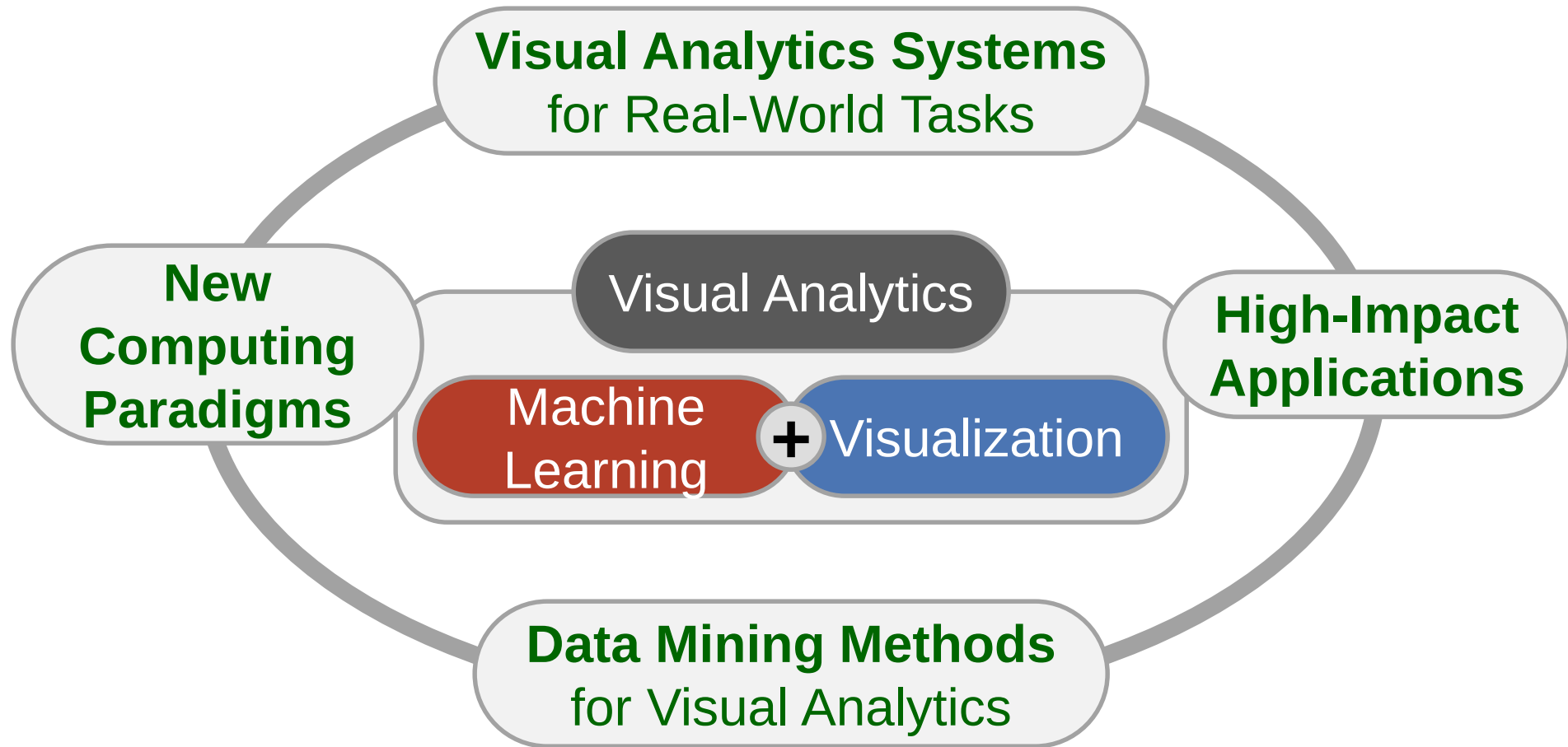
Fast computation

Deeper understanding

>Millions of data items

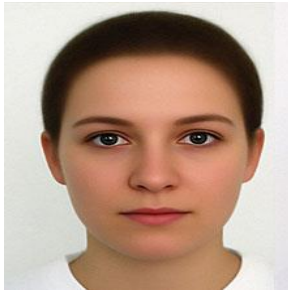
Thousands of data items

My Research: True Integration of Both Worlds

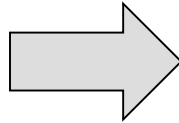


High-Dimensional Data Images

- ▶ Serialized/rasterized pixel values

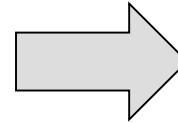


Raw images



5	34	78
3	80	63
58	24	45

Pixel values



5
3
58
34
80
24
63
45
63

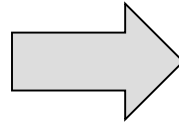
Serialized
pixels

High-Dimensional Data Images

► Serialized pixel values

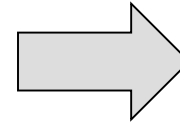


Raw images



5	24	70	
22	86	15	
3	49	54	67
5	14	78	36

Pixel values



5	22
3	49
58	14
34	86
80	54
24	78
63	15
45	67
63	36

...

Serialized
pixels

► Huge dimensions

- 640x480 image size → 307,200 dimensions

High-Dimensional Data Documents

- ▶ Bag-of-words vector
 - Document 1 = “John likes movies. Mary likes too.”
 - Document 2 = “John also likes football.”

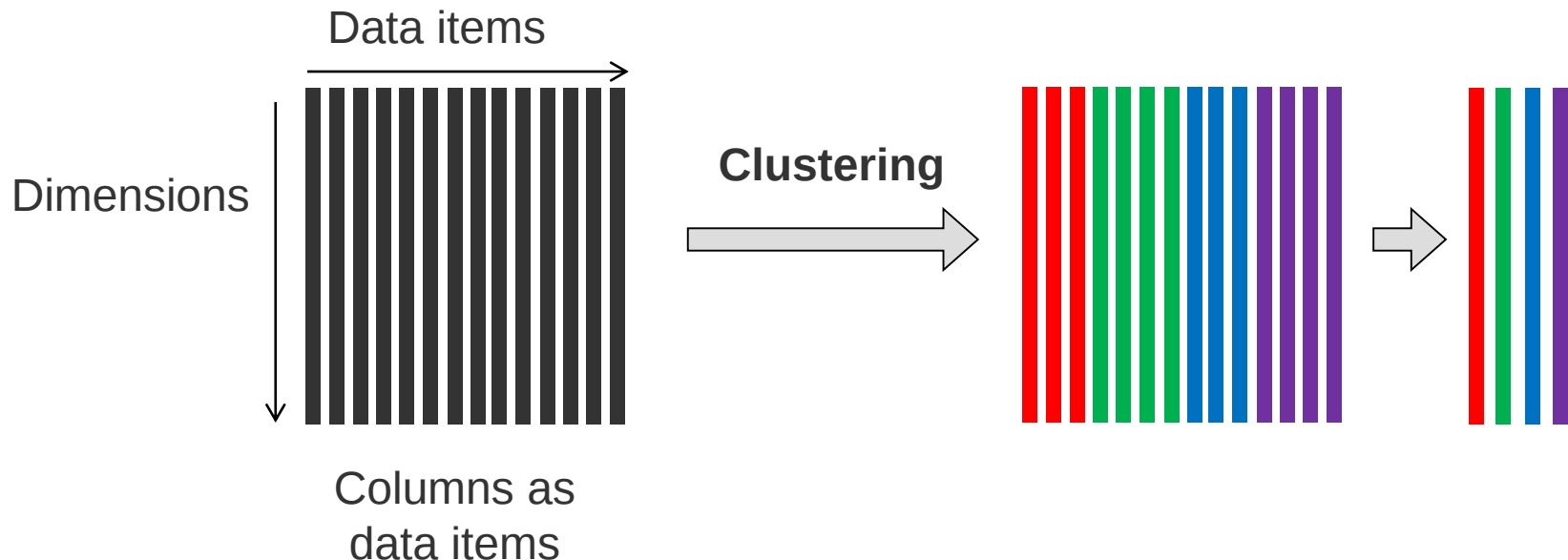
Vocabulary	Doc 1	Doc 2
John	1	1
likes	2	1
movies	1	0
also	0	1
football	0	1
Mary	1	0
too	1	0

...

Visualizing High-Dimensional Data

Clustering

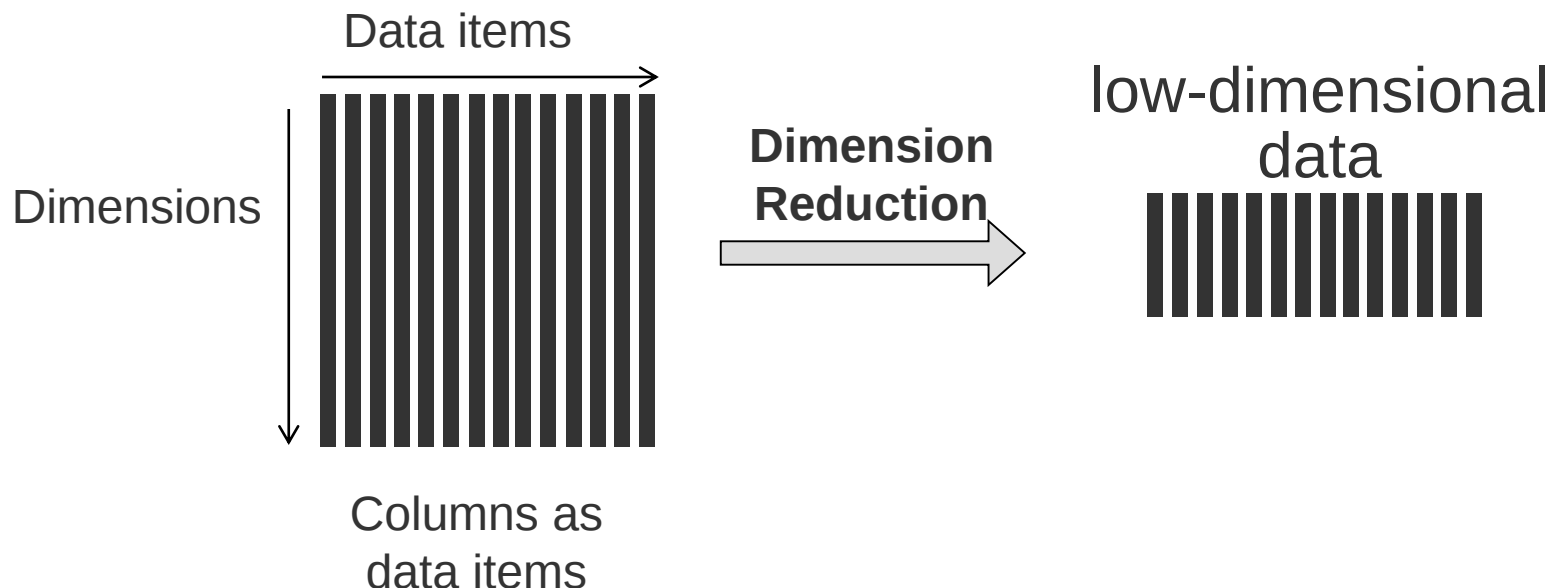
- ▶ Goal: Provide the **Groupings** of High-Dimensional Data
- ▶ *K*-means, hierarchical clustering, Gaussian mixture model, nonnegative matrix factorization, semi-supervised methods, ...



Visualizing High-Dimensional Data

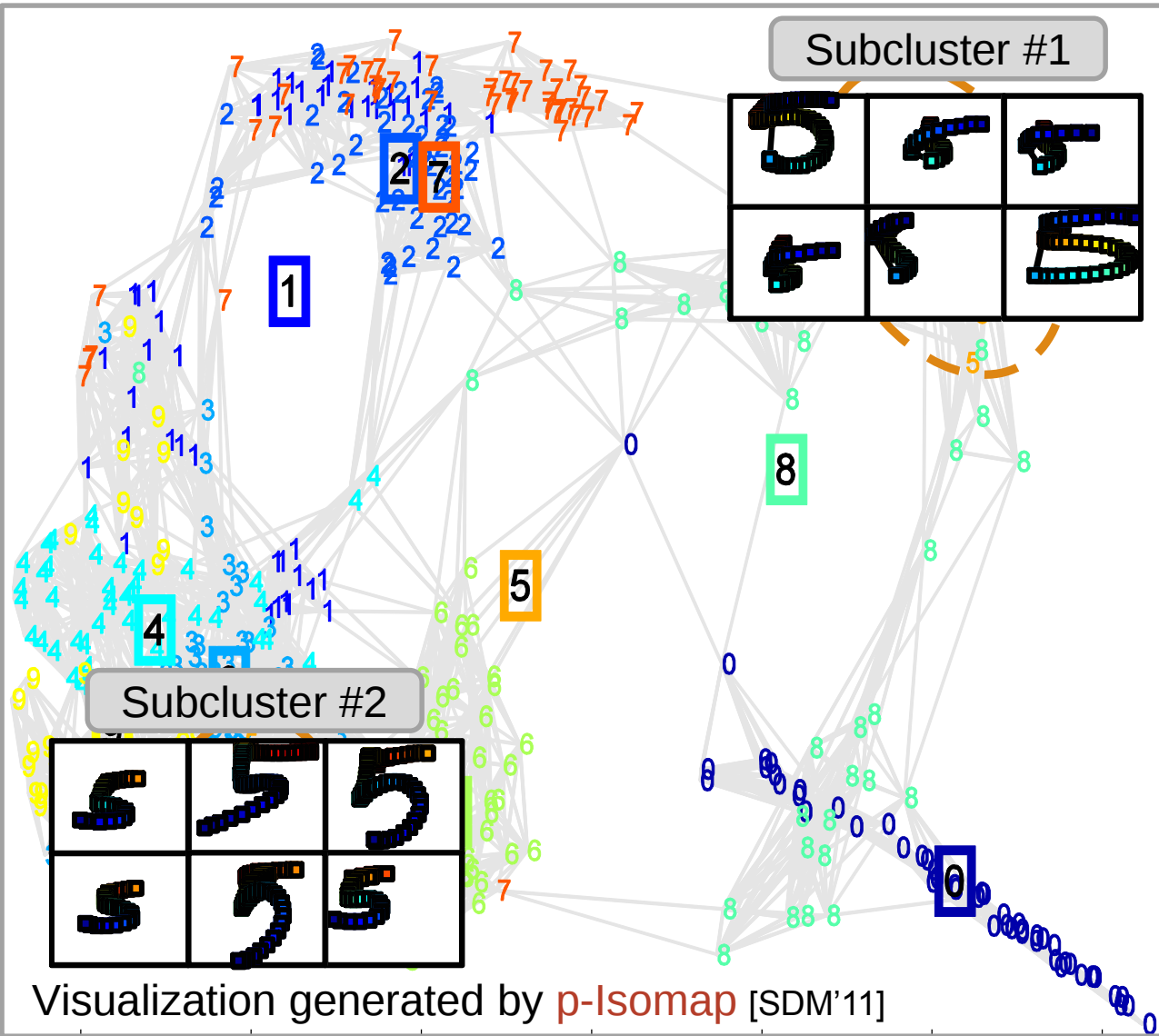
Dimension Reduction

- ▶ Goal: Provide the **2D mapping** of High-Dimensional Data
- ▶ Principal component analysis, multidimensional scaling, linear discriminant analysis, t-distributed stochastic neighbor embedding, ...



Visual Insight to Machine Learning

Handwritten Digit Recognition



Subclusters in
digit '5'

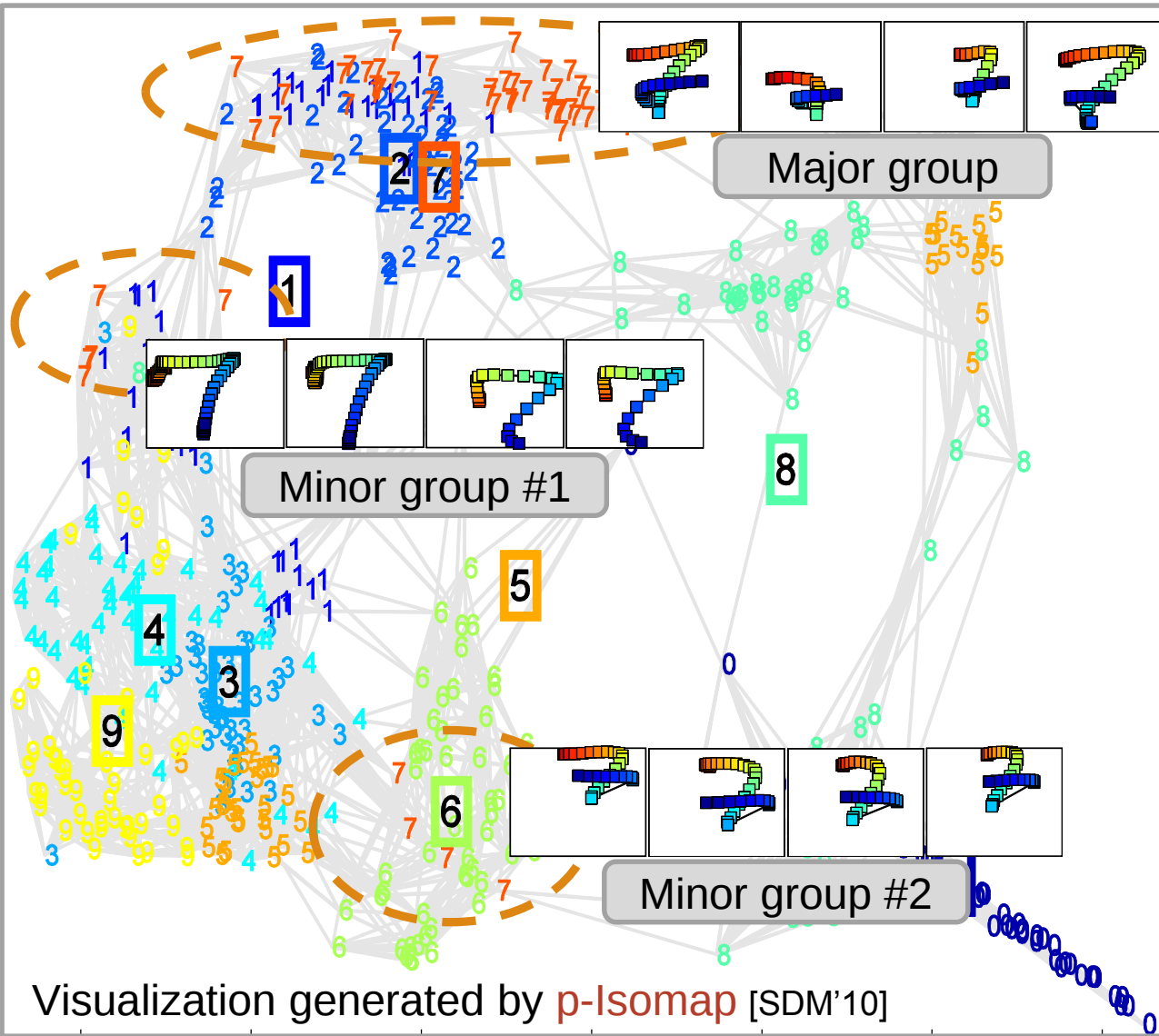


Handling them as
separate clusters



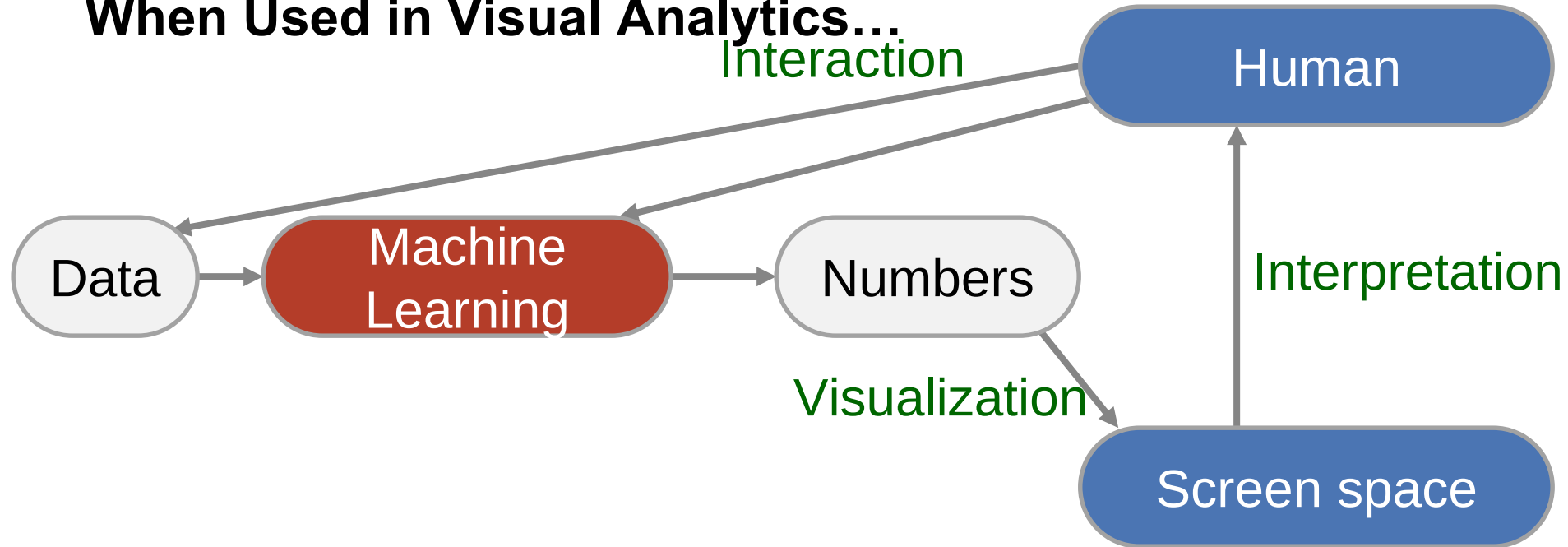
Better prediction
(89% → **93%**)

Visual Insight to Machine Learning Handwritten Digit Recognition



Challenges in Machine Learning + Visualization

When Used in Visual Analytics...

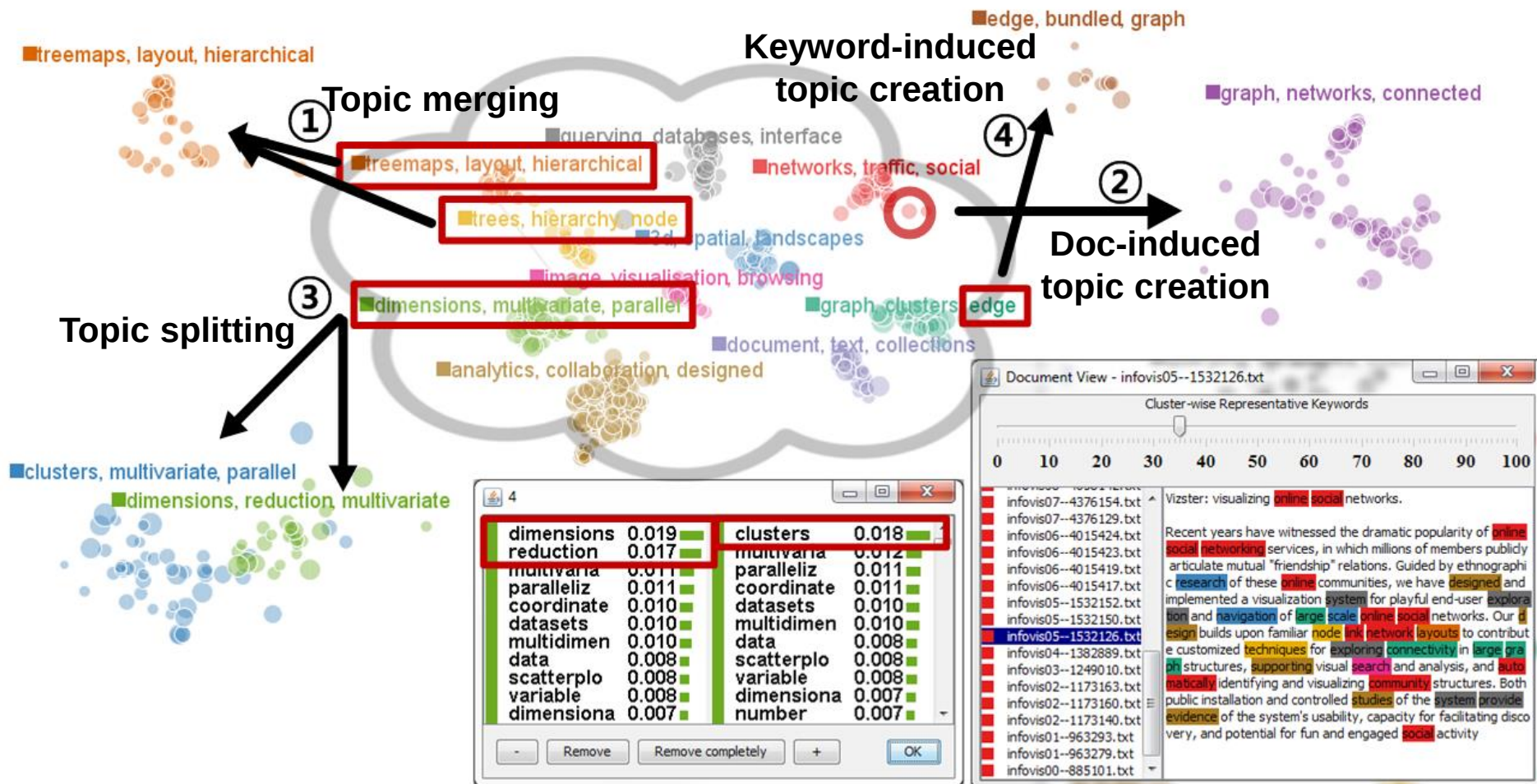


Machine learning methods should be

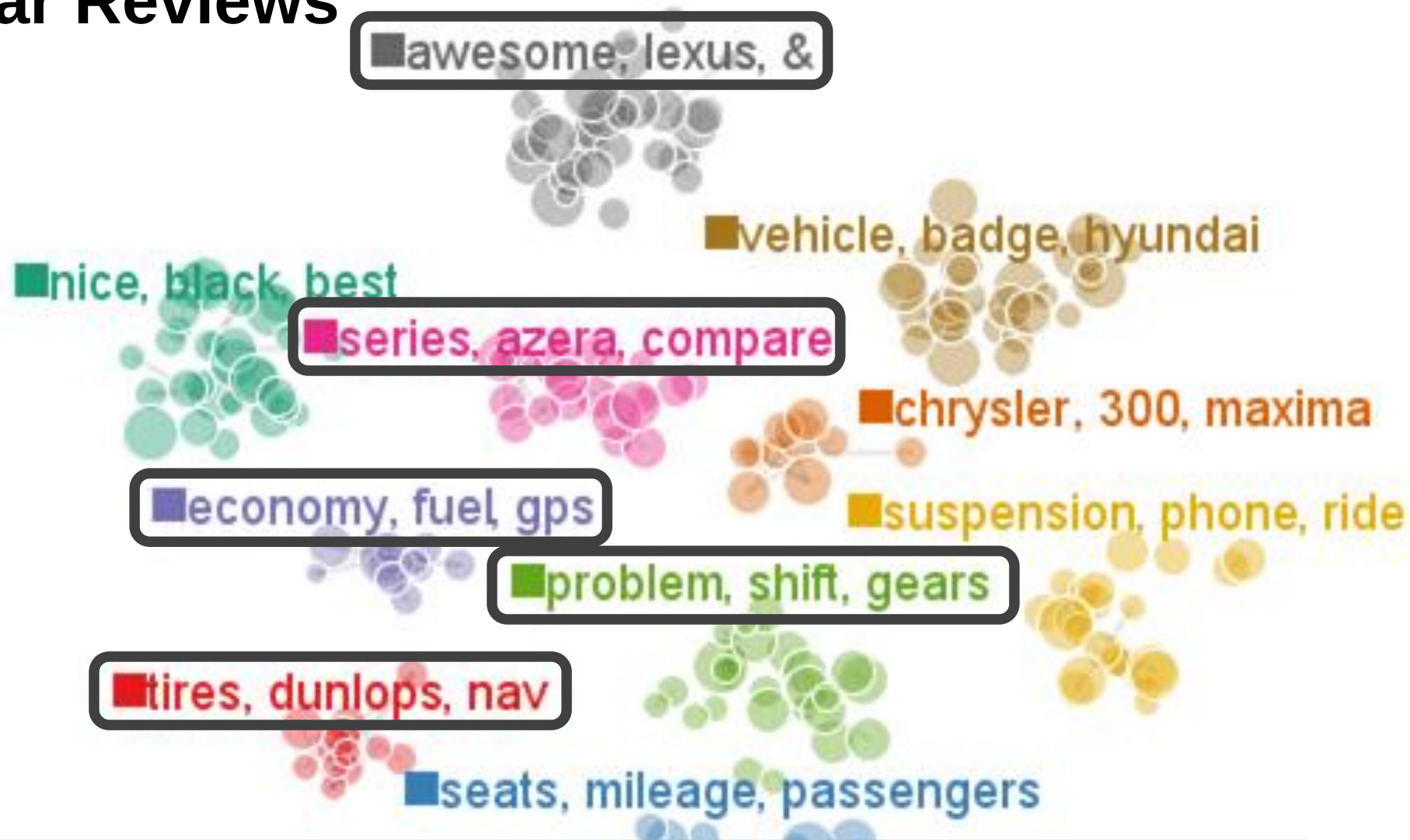
- More interpretable
- More user-interactive
- Real-time responsive, i.e., faster

UTOPIAN: User-Driven Topic Modeling Based on Interactive NMF

[TVCG 2013]



Visualization Example: Car Reviews



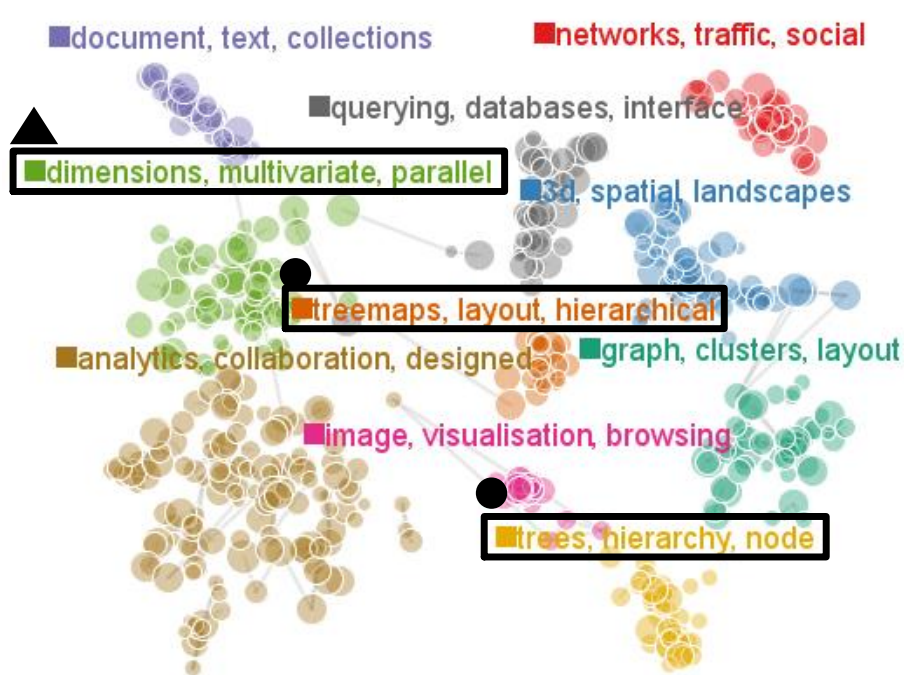
Topic summaries are **NOT** perfect.

➔ UTOPIAN allows **user interactions** for improving them.

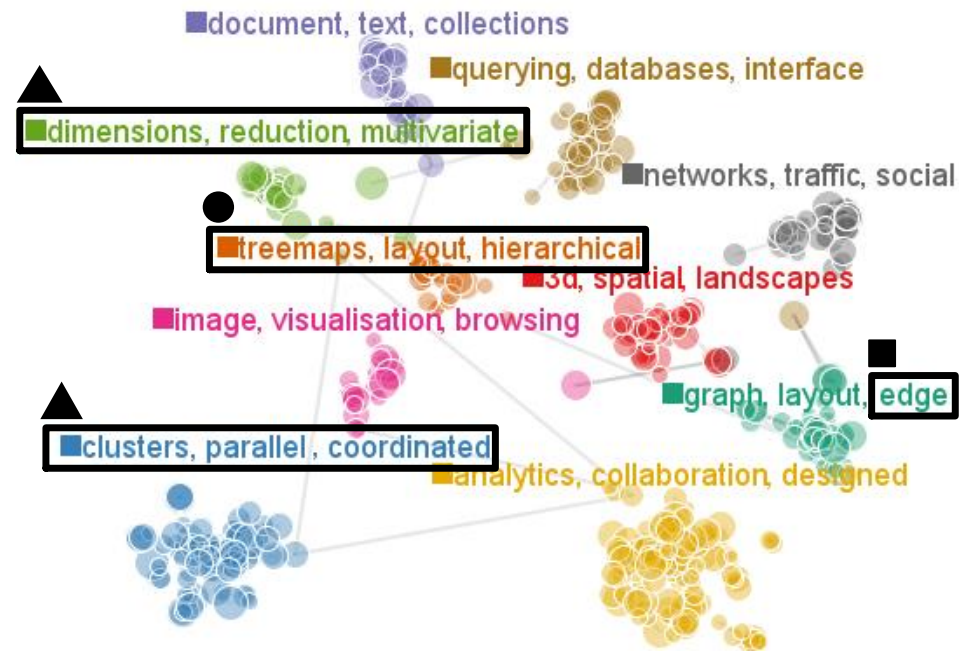
Interaction Demo Video

<http://tinyurl.com/UTOPIAN2013>

InfoVis-VAST Paper Data



Before interaction



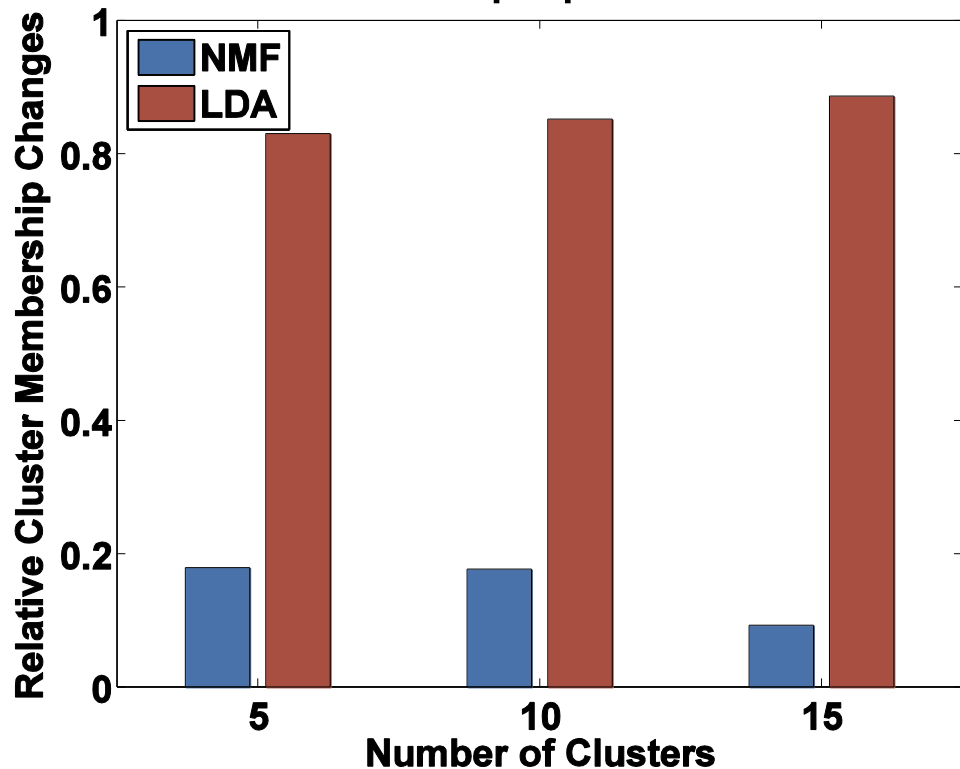
After topic splitting (triangle)
and topic merging (circle)

Why NMF (instead of LDA)?

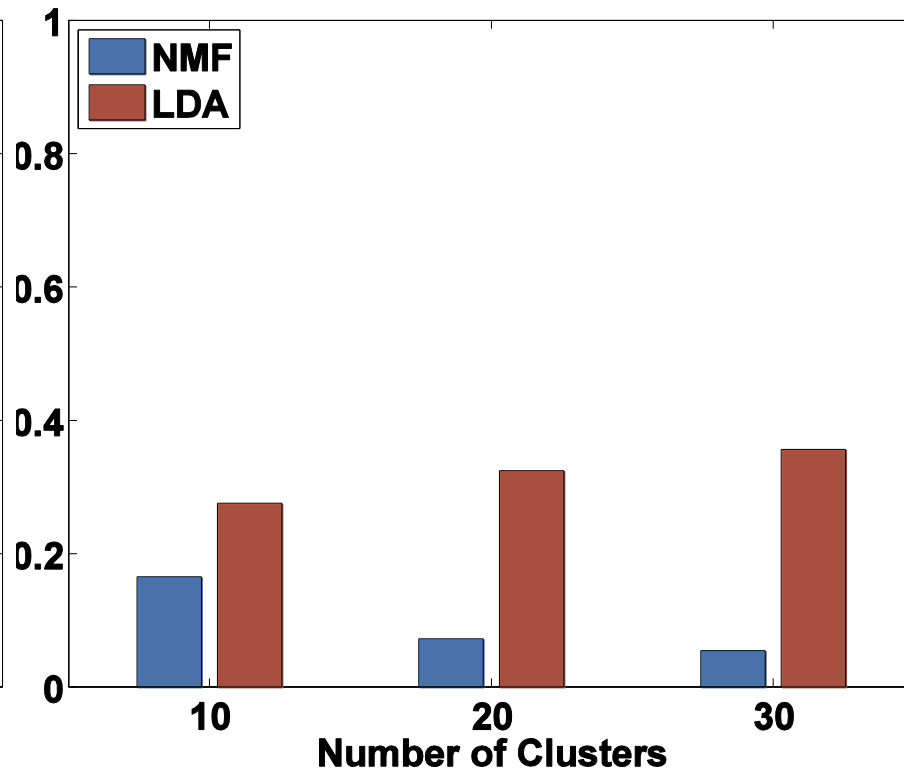
Consistency from Multiple Runs

Documents' topic membership changes among 10 runs

InfoVis/VAST paper data set



20 newsgroup data set

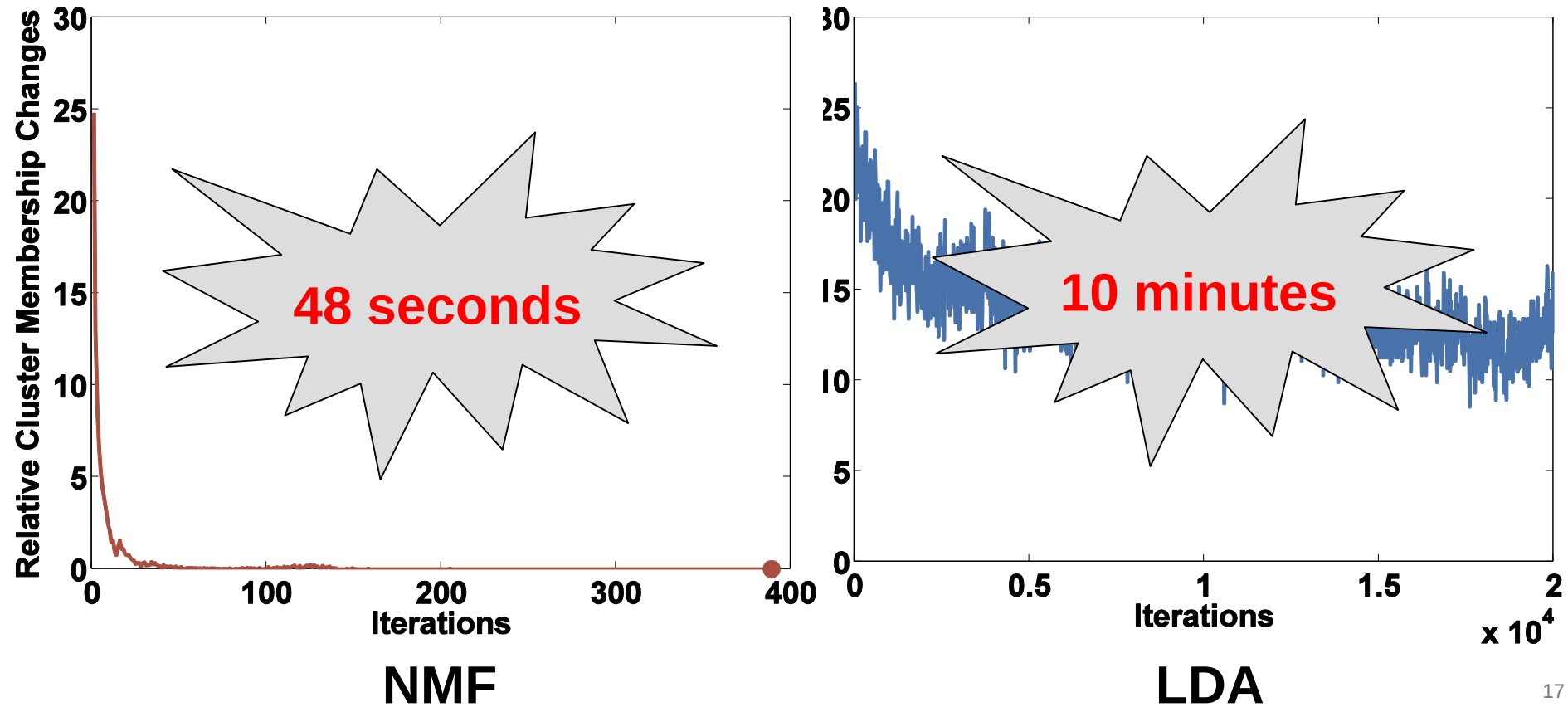


Why NMF (instead of LDA)?

Empirical Convergence

Documents' topic membership changes between iterations

InfoVis/VAST paper data set



NMF vs. LDA

Topic Summary (Top Keywords)

NMF InfoVis/VAST paper data set

Run	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
#1	visualization design	information user	analysis system	graph layout	visual analytics	data sets	color weaving
#2	visualization design	information user	analysis system	graph layout	visual analytics	data sets	color weaving

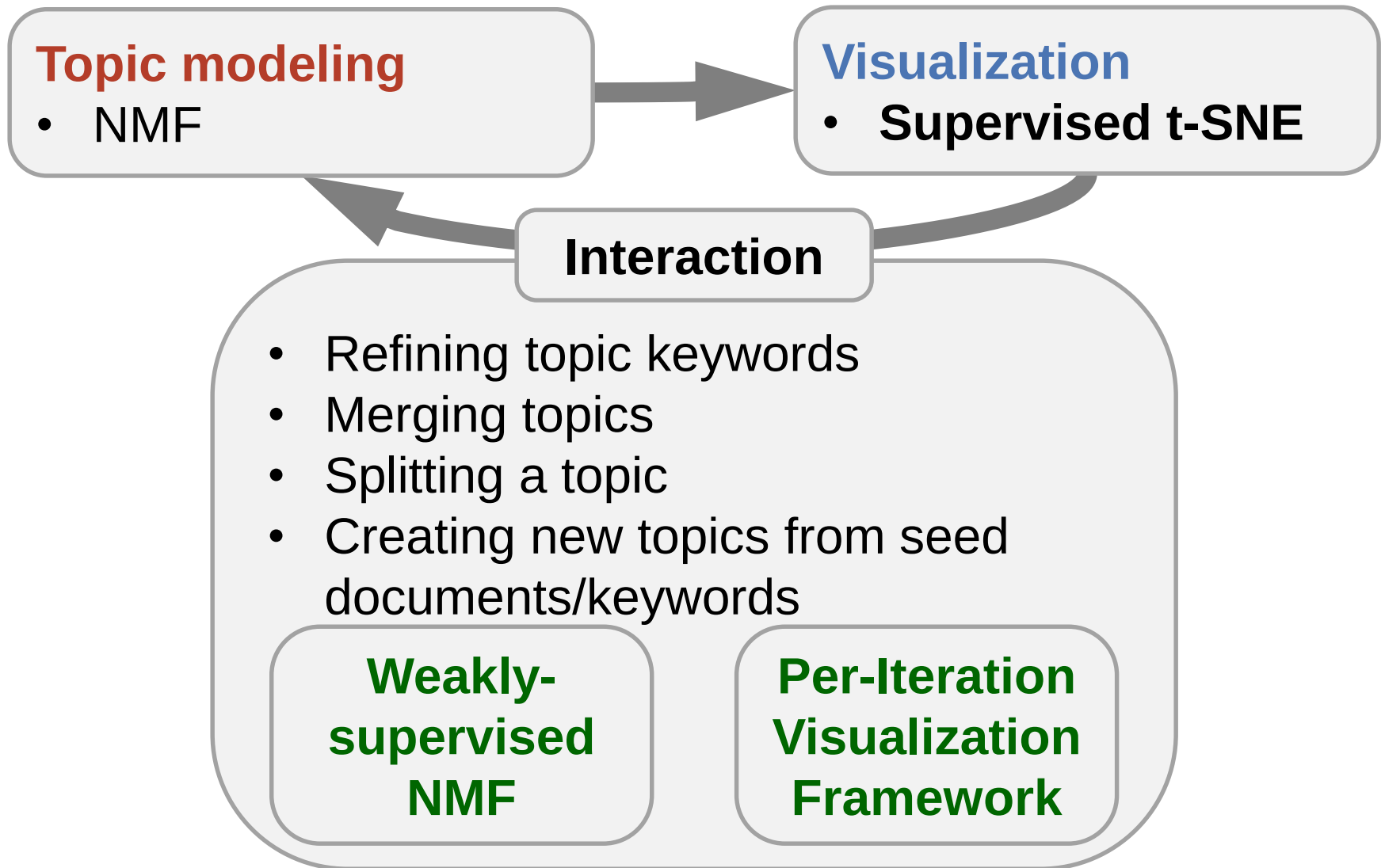
LDA

Run	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
#1	document similarities	knowledge edge	query collaborative	social tree	measures multivariate	tree animation	dimension treemap
#2	document query	analysts scatterplot	spatial collaborative	text document	multidimen- sional high	tree aggregation	dimension treemap

- ▶ Topics are more consistent in NMF than in LDA.
- ▶ Topic quality is comparable between NMF and LDA.

UTOPIAN

Interactions and Key Techniques



Supervised t-SNE: Visualizing documents

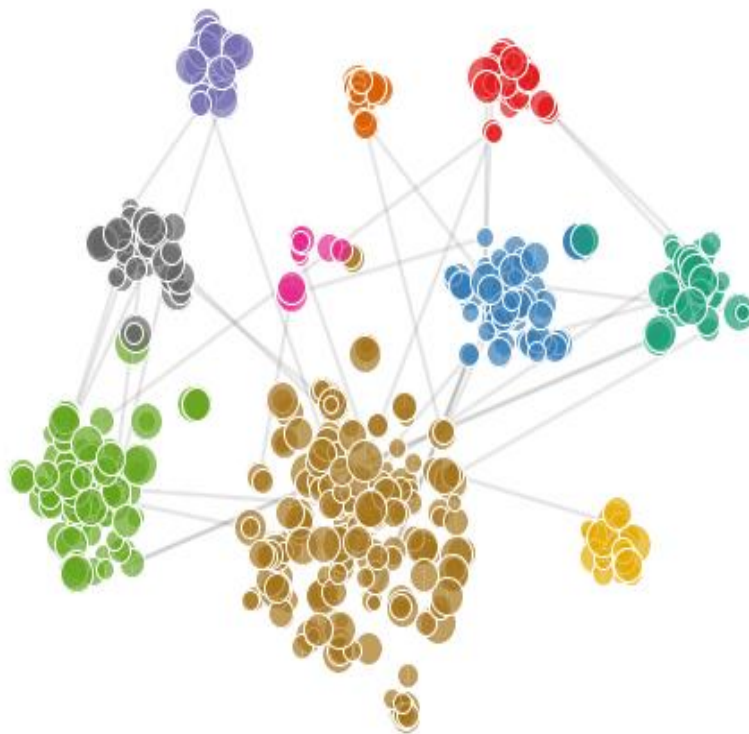
Original t-SNE

- Documents do not have clear topic clusters.



Supervised t-SNE

- $d(x_i, x_j) \leftarrow \alpha \cdot d(x_i, x_j)$ if x_i and x_j belong to the same topic.
(e.g., $\alpha = 0.3$)



Weakly Supervised NMF: Supporting user interactions

Weakly supervised NMF [DMKD 2014]

$$\min_{W \geq 0, H \geq 0} \|A - WH\|_F^2 + \alpha \|(W - W_r)M_W\|_F^2 + \beta \|M_H(H - D_H H_r)\|_F^2$$

W_r, H_r : reference matrices for W and H (user-input)

M_W, M_H : diagonal matrices for weighting/masking columns and rows of W and H

► Algorithm: block-coordinate descent framework

PIVE: (Per-Iteration Visualization Environment)

https://youtu.be/zURFA9P5E_s

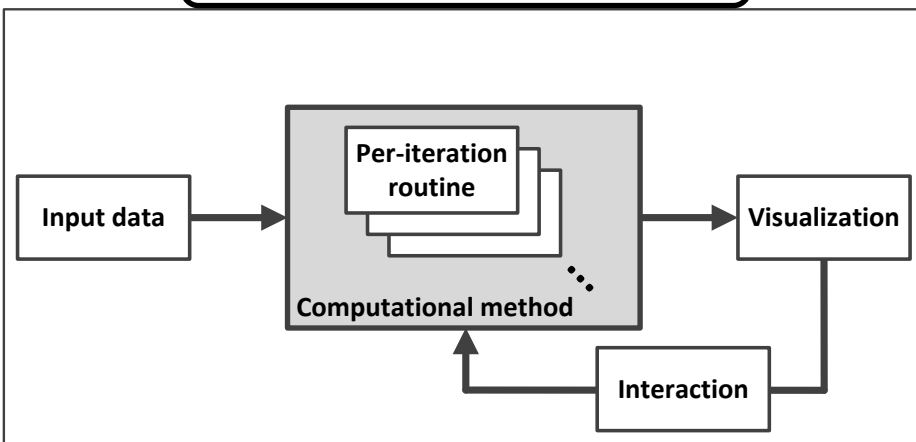
Motivation

- ▶ Many algorithms are iterative methods.

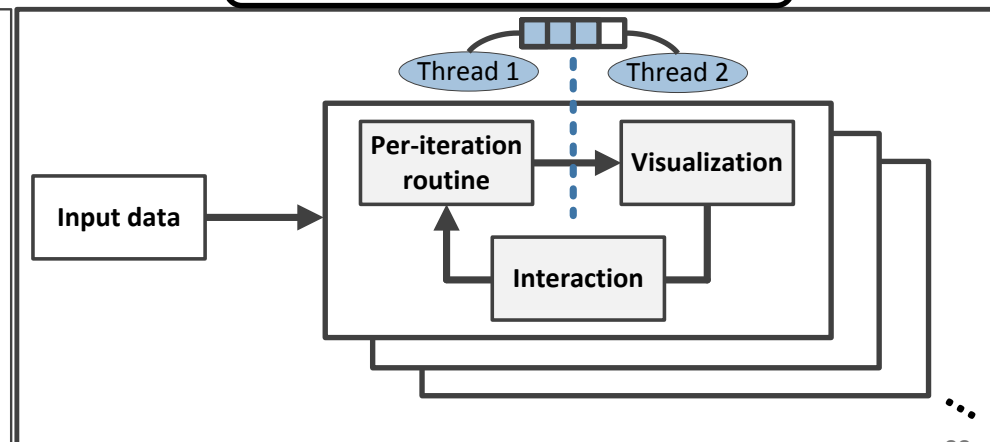
PIVE

- ▶ Integration methodology of iterative methods for **Real-Time** interactive visualization [Choo et al., VAST'14 Poster]

Standard approach



PIVE approach



Compare and Contrast: Joint Topic Discovery

[KDD'15]

Formulation

$$\min_{W \geq 0, H \geq 0} \quad 1/n_1 \|A_1 - W_1 H_1\|_F^2 + 1/n_2 \|A_2 - W_2 H_2\|_F^2 + \alpha \|W_{1,c} - W_{2,c}\|_F^2 + \beta \|W_{1,d}^T W_{2,d}\|_F^2$$

where $W_i = [W_{i,c} \ W_{i,d}]$

2000-2005

driven gener interesting tempor restrict deriv vertic profil
dure induct optim mine associ rule
confin discov implic maxim frequent
mine sequenc disjunct updat tidset
decreas closemin depth list discov
invers structur streammonoton geometr window gener

Common
topics in DM

outlier sequenc dataset motif mino
theoret applic duplic detect
weight seri chang event system inform point anomali
spam intrus vidio
non attribut geograph robust
measur subspac geo relationship optim spatial
distanc hierarchi cluster
region

2006-2008

techniqu partit data
larg scale graph databas
design corre represent
bipartit paramet compact adapt
short strategi duplic recommend veri dynam
rank commun neural
academ framework bayesian
evolut dynam sensor social
research propag predict behavior latent group
onlin network
traffic

Compare and Contrast: Joint Topic Discovery

[KDD'15]

Formulation

$$\min_{W \geq 0, H \geq 0} \frac{1}{n_1} \|A_1 - W_1 H_1\|_F^2 + \frac{1}{n_2} \|A_2 - W_2 H_2\|_F^2 + \alpha \|W_{1,c} - W_{2,c}\|_F^2 + \beta \|W_{1,d}^T W_{2,d}\|_F^2$$

where $W_i = [W_{i,c} \ W_{i,d}]$

VAST

data semant reduc
stage subspac framework base
high propos dimension
reduct cluster space imag
visual method analysi
astronom browser

applic
concept environ
exist support analyt
evalu insight knowledg synthesi
process visual analyst area
makemodel system decis
reason develop

Common
topics

effect applic paper result
present develop framework
larg structur approach interact method set
design visual
explor base queri

tool paper
interact effect featur discoveri
nugget knowledg similar explor
interfac dataset nm
user present
system task base
manag

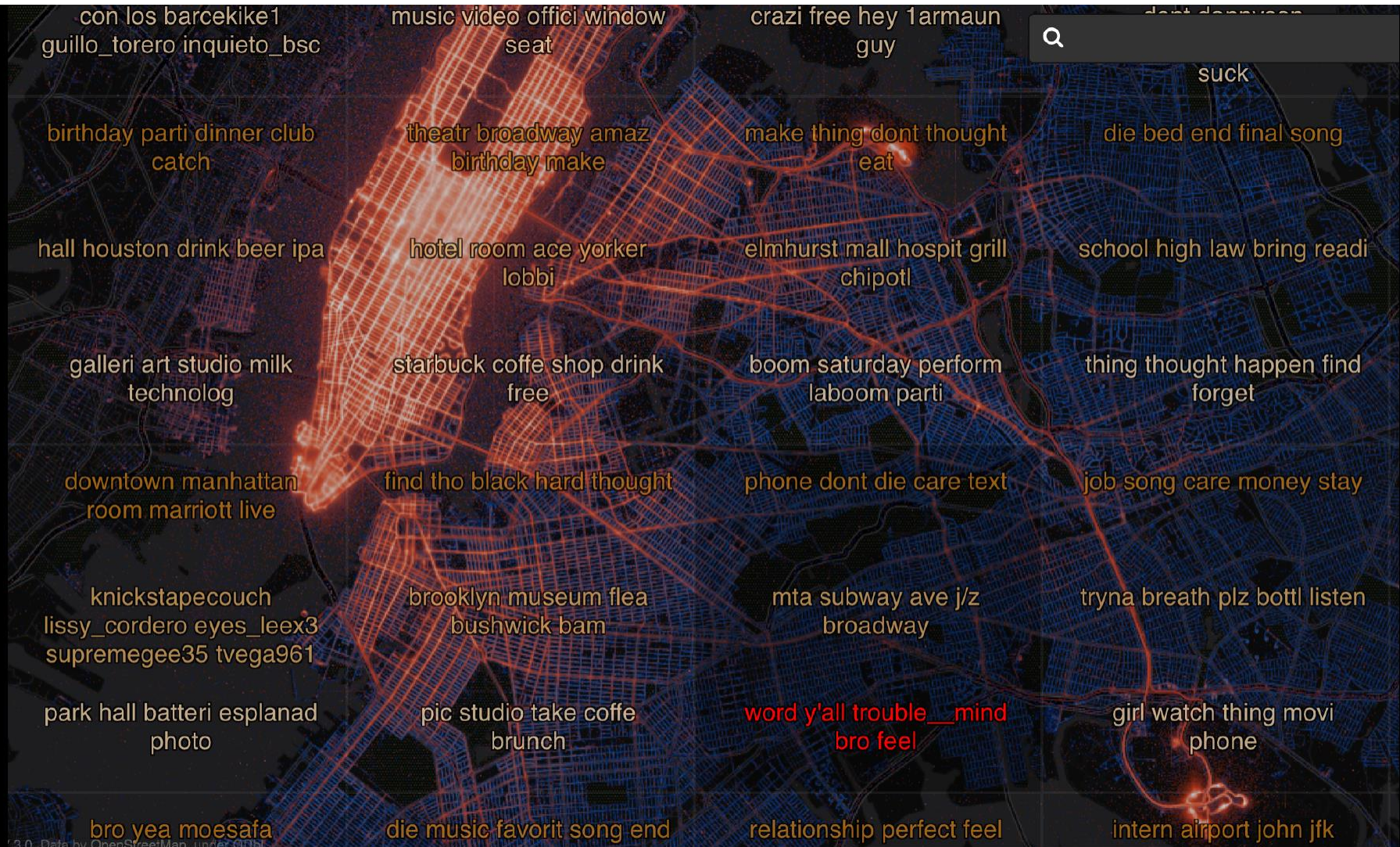
InfoVis

base dynam direct larg
class network edge product layout
cluster draw vertic algorithm method
graph task structur node interact

map
repres found high sideabil
region blend color experi
compon particip space
textur displai method encod
valu weav inform

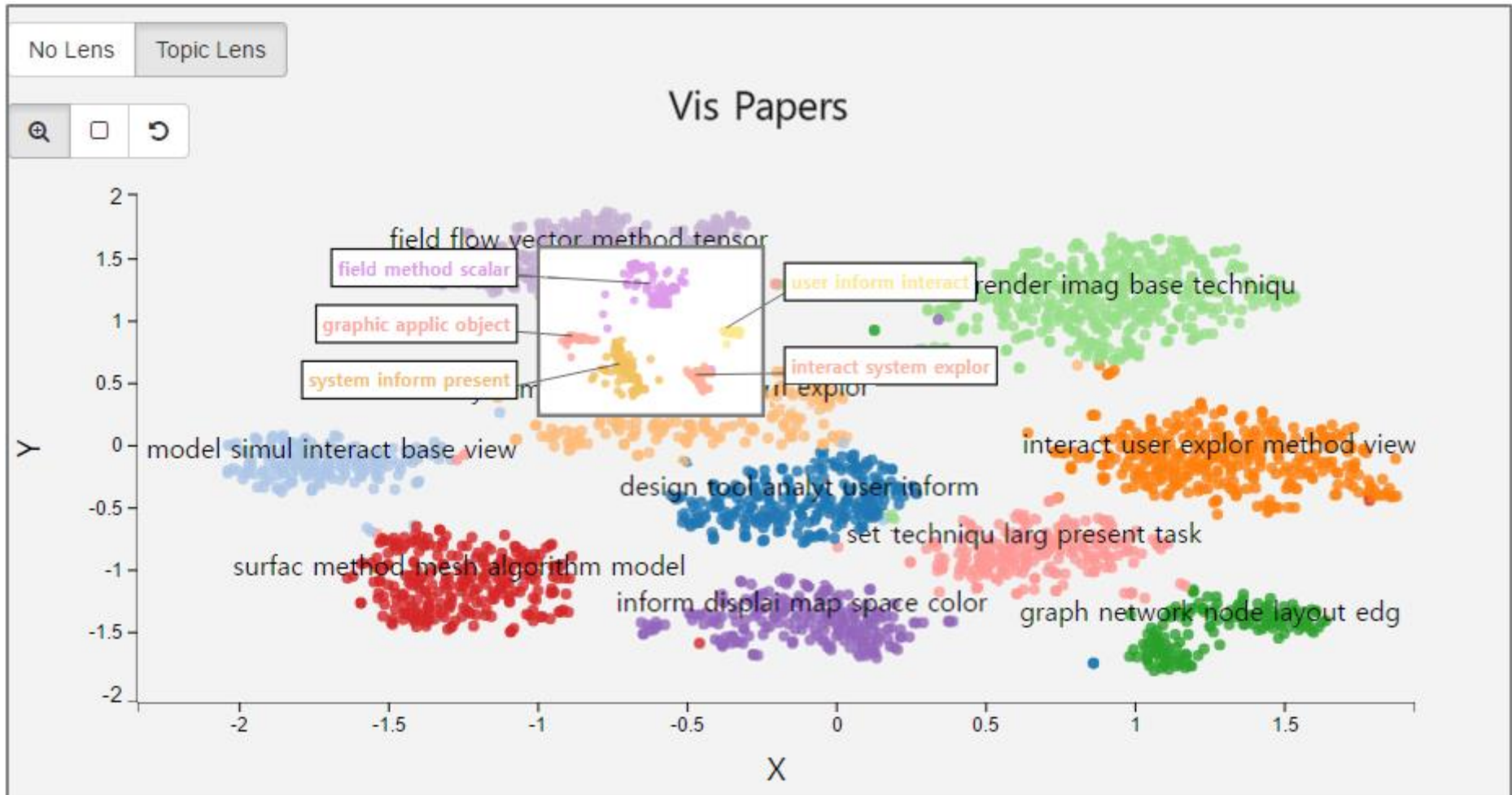
Geospatio-Temporal Topic Modeling

<http://aperture.xdataonline.com/#/>



TopicLens: Efficient Multi-Level Visual Topic Exploration

[Under review]



TopicLens: Efficient Multi-Level Visual Topic Exploration

[Under submission]

Key aspects of backend topic modeling and dimension reduction methods

- ▶ Real-time response

- How can we ensure real-time response against highly-dynamic user interactions such as lens?

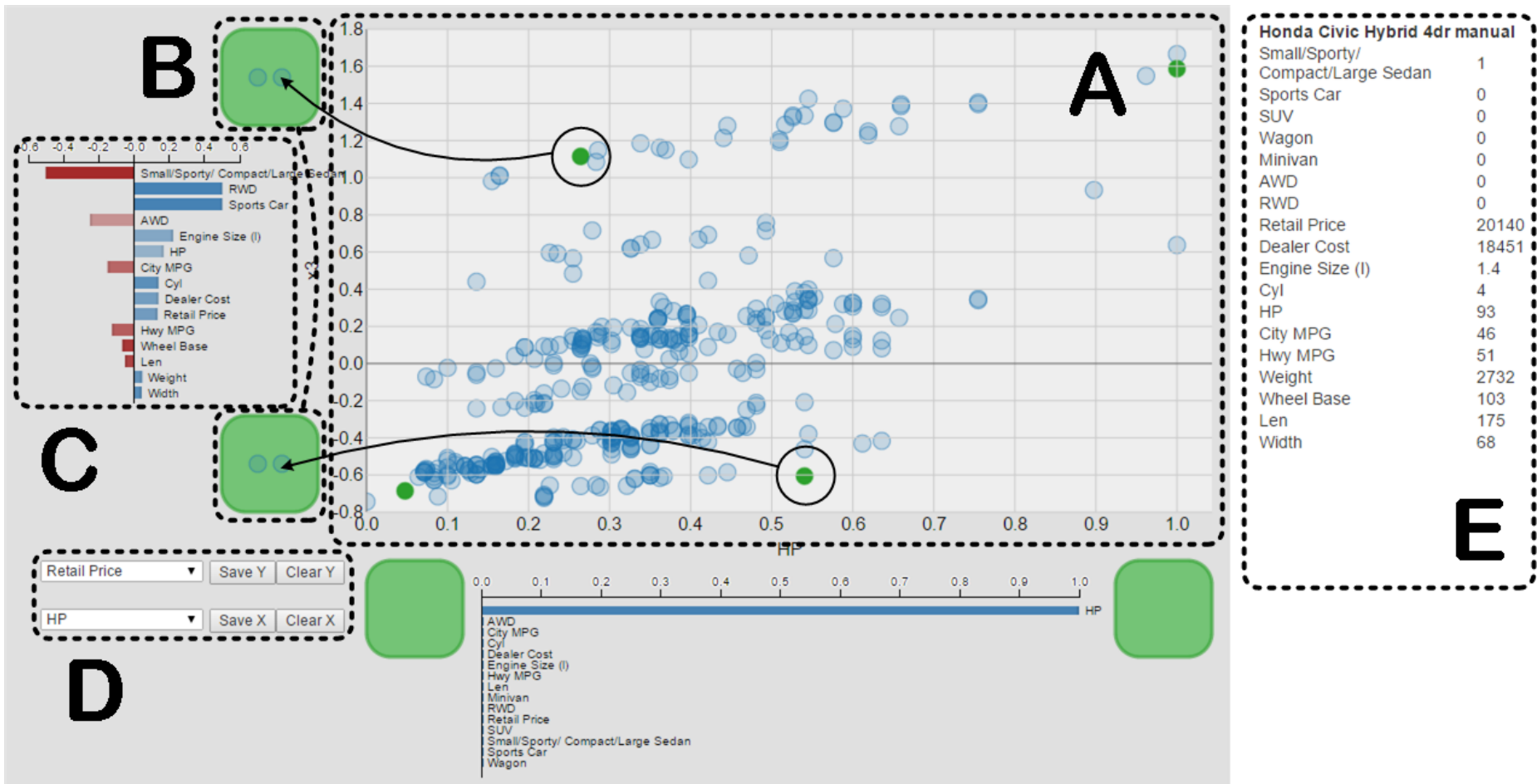
- ▶ Continuity and consistency with previous results

- How can we allow users to maintain the continuity and consistency between the previous and the new results?

InterAxis: Steering Scatterplot Axes via Observation-Level Interaction

[TVCG'15]

<http://www.cc.gatech.edu/~hkim708/InterAxis/>



ConceptVector: Building User-Driven Concepts via Word Embedding

[Under review]

<http://conceptvector.org/>

Concept Name: Immigration-related

Concept Type: Unipolar

Save

Go Back

Positive Words Input

Advanced Settings

Please type a few words for the concepts you are looking for.

immigration x citizenship x naturalization x asylum x nationality x deportation x visa x visas x extradition x custody x immigrants x

undocumented x migrants x Add a positive words

Click to add suggested words.

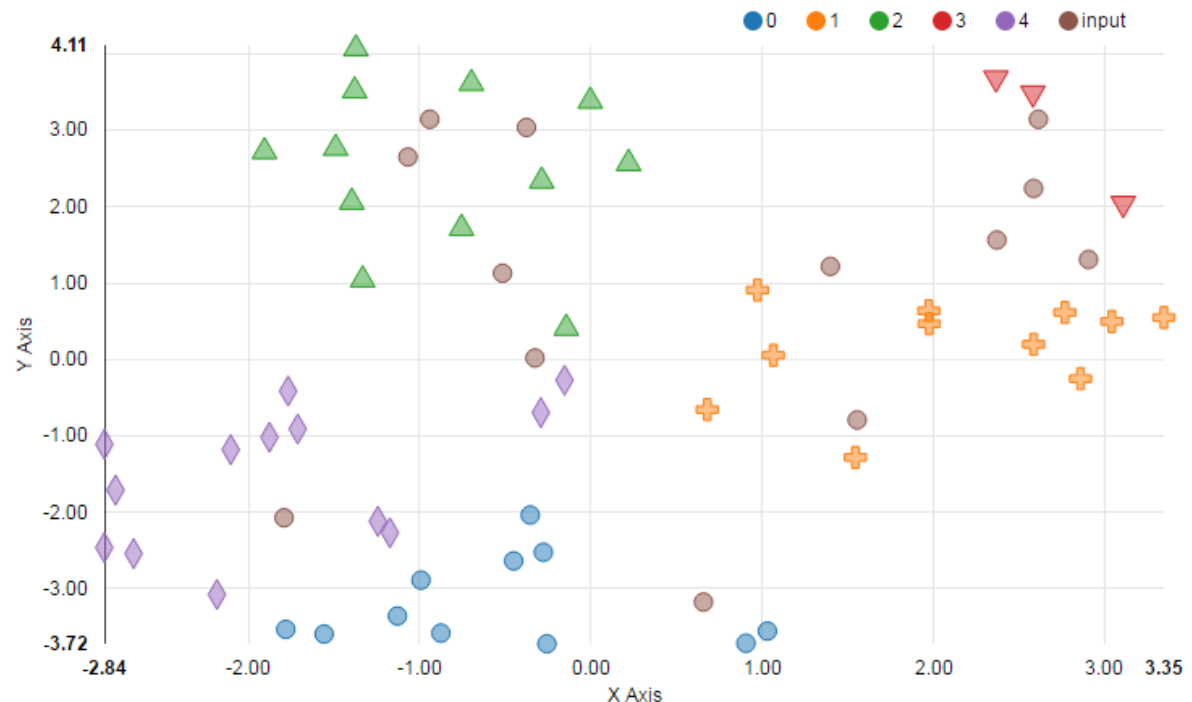
Group0 extradite warrant trial pleaded prosecution
extradited retrial defendant defendants
questioning acquitted

Group1 granting waiver revoked barring
enforcement ins permits revoke deny permit
granted

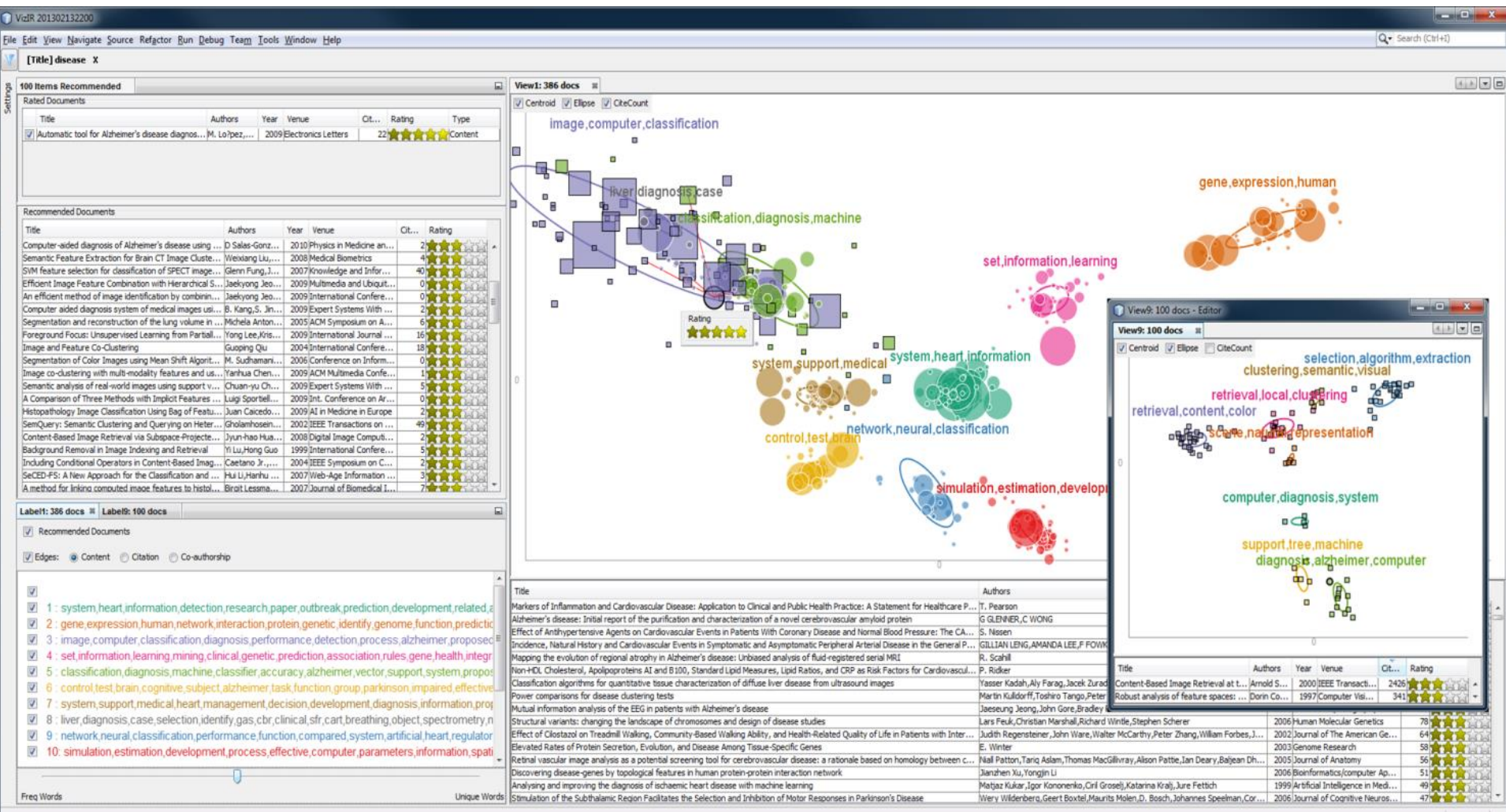
Group2 seekers immigrant migrant citizens
mexicans illegals foreigners deportations
emigrants deport fishermen haitians

Group3 passport ethnicity identity

Group4 freed arrest expulsion prisoner jail amnesty
sentenced suspects detention imprisoned
convicts detainee prison



VisIRR: Information Retrieval and Personalized Recommender System



VisIRR Demo

Citation-based Recommendation

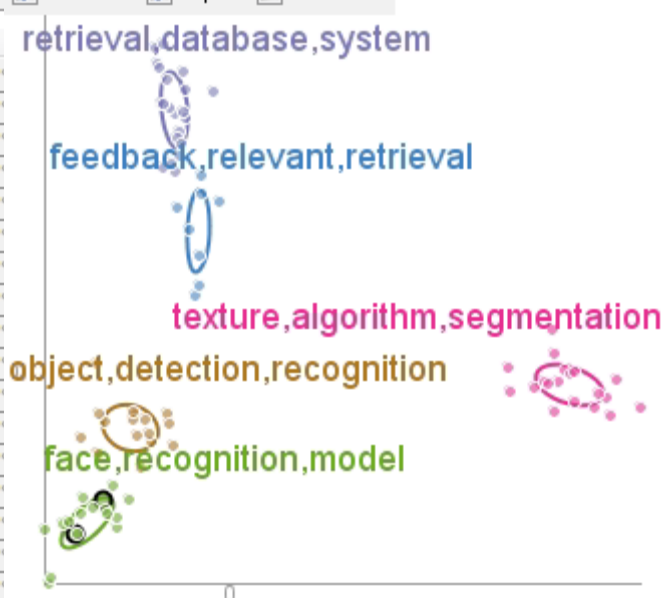
Preference-assigned item as 'highly like' :

'Automatic Classification System for the Diagnosis of Alzheimer Disease Using Component-Based SVM Aggregations'

Recommended Documents

Title	Authors	Year	Venue	Cit...	Rating
Asymmetric Bagging and Random Subspace for Support...	Dacheng Tao,...	2006	IEEE Transactions on ...	168	★★★★★
Face Recognition Using Component-Based SVM Classific...	Jennifer Huan...	2002	Support Vector Machines	24	★★★★★
Face Recognition: Features Versus Templates	Roberto Brun...	1993	IEEE Transactions on ...	1208	★★★★★
Example-Based Object Detection in Images by Compon...	Anuj Mohan,...	2001	IEEE Transactions on ...	424	★★★★★
Face Recognition with Support Vector Machines: Global ...	Bernd Heisele...	2001	International Confere...	168	★★★★★
Component-based Face Detection	Bernd Heisele...	2001	Computer Vision and P...	85	★★★★★
Texture Features for Browsing and Retrieval of Image ...	B. Manjunath...	1996	IEEE Transactions on ...	1498	★★★★★
Texture analysis and classification with tree-structured ...	Tianhorng Ch...	1993	IEEE Transactions on ...	648	★★★★★
SIMPLICity: Semantics-Sensitive Integrated Matching fo...	James Wang,...	2001	IEEE Transactions on ...	876	★★★★★
Content-Based Image Retrieval at the End of the Early ...	Arnold Smeul...	2000	IEEE Transactions on ...	2426	★★★★★
Image retrieval using color and shape	Anil Jain,Adit...	1996	Pattern Recognition	455	★★★★★
Boosting Image Retrieval	Kinh Tieu,Paul...	2000	Computer Vision and P...	340	★★★★★
Support vector machine active learning for image retrieval	Simon Tong,E...	2001	ACM Multimedia Confe...	543	★★★★★
Multi-class relevance feedback content-based image re...	Jing Peng	2003	Computer Vision and I...	21	★★★★★
Hierarchical Mixtures of Experts and the EM Algorithm	Michael Jorda...	1994	Neural Computation	1269	★★★★★
Incorporate Support Vector Machines to Content-Base...	Pengyu Hong...	2000	International Confere...	86	★★★★★
On Combining Classifiers	Josef Kittler,...	1998	IEEE Transactions on ...	2176	★★★★★

☒ Centroid ☒ Ellipse ☐ CiteCount

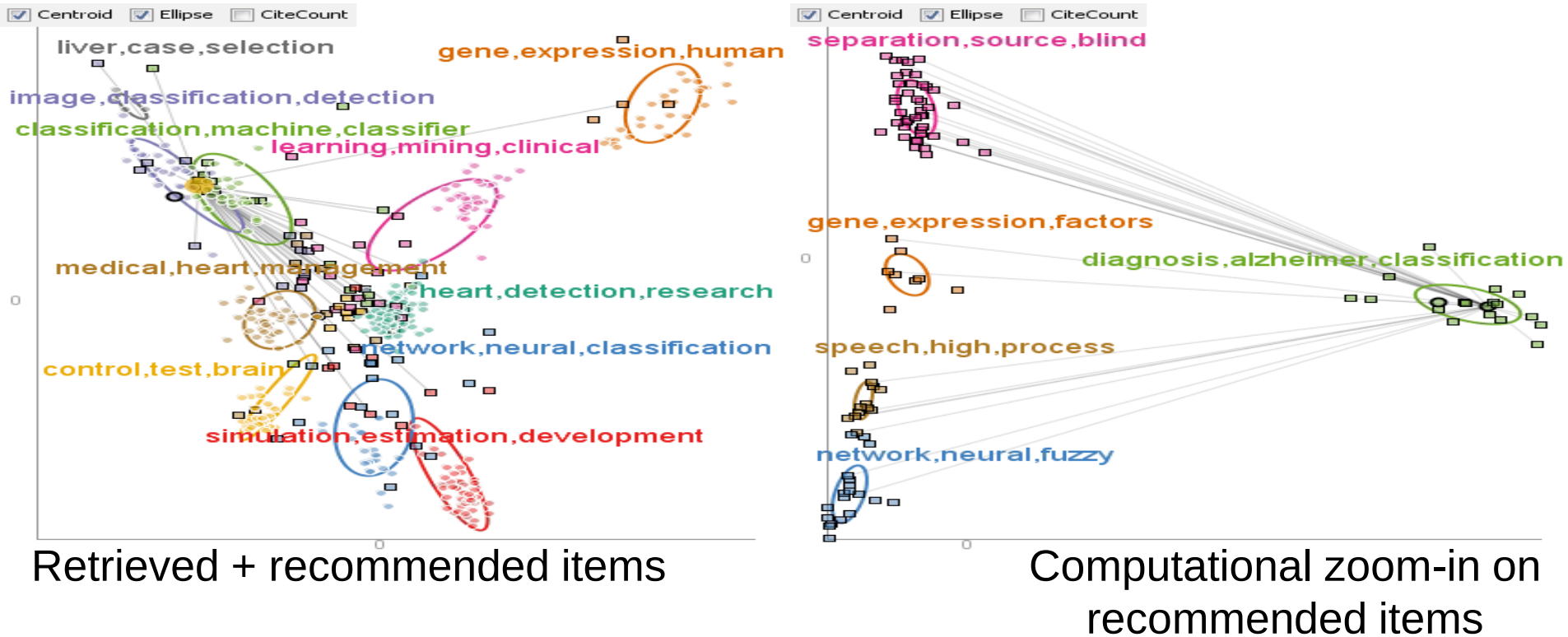


Most of the recommended items are highly cited.
Computational zoom-in shows sub-areas relevant to the article.

VisIRR Demo

Co-authorship-based Recommendation

Preference-assigned item as 'highly like' :
'Automatic Classification System for the Diagnosis of Alzheimer Disease Using Component-Based SVM Aggregations'



It shows other areas of the authors of this paper.

Perception- and Screen Space-Driven Integration Framework

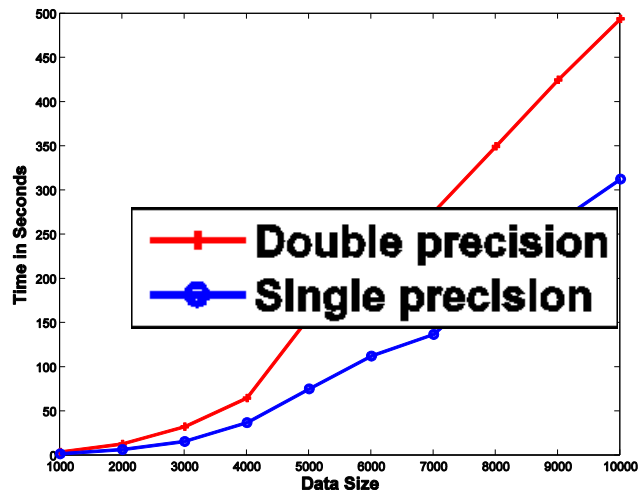
[CG&A, 2013]

Motivation

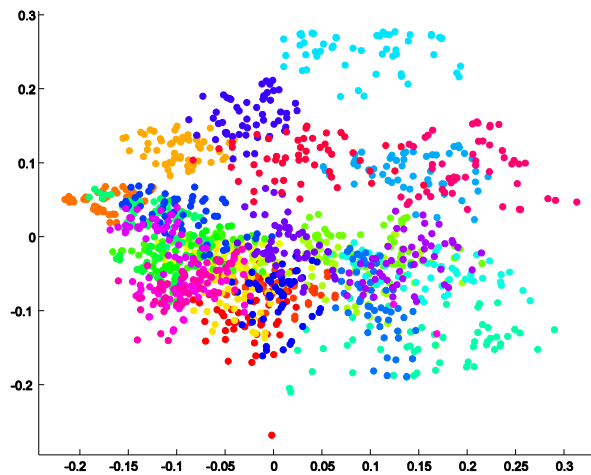
- ▶ Humans and computer screens **do not** require **high precision**.

Approach

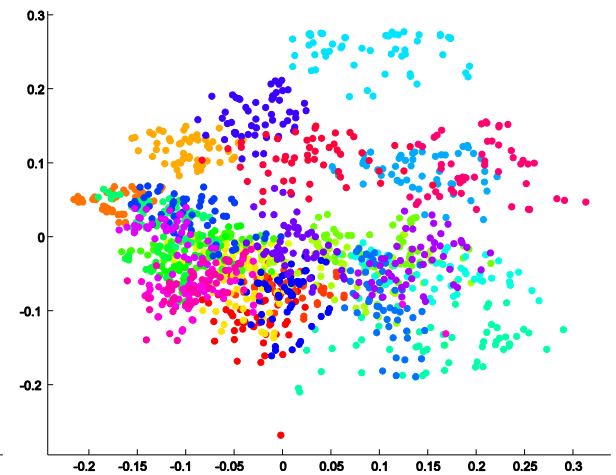
- ▶ **Approximate computing**



Computing time
vs. data size



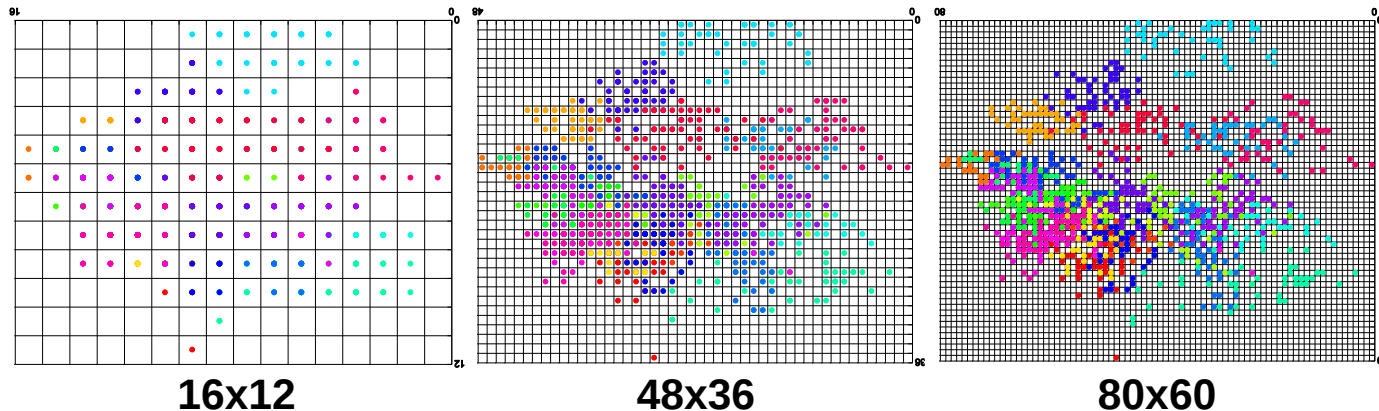
Double-precision PCA



Single-precision PCA

New Computing Paradigms for Visual Analytics

Adaptive hierarchical refinement



- Leveraging ideas from other literatures, e.g., wavelet



Images src: http://www.cse.lehigh.edu/~spletzer/rip_f06/lectures/lec013_Pyramids.pdf

Thank you! Jaegul Choo jchoo@korea.ac.kr

Collaborators from academia, industry, and the government

A. Endert, A. Gray, A. White, B. Drake, B. Dilkina, B. Kwon, C. Görg, C. Reddy, C. Lee, C. Stolper, D. Lee, E. Clarkson, E. Fujimoto, F. Li, G. Nakamura, H. Park, H. Pileggi, H. Lee, H. Zha, H. Kim, J. Eisenstein, J. Shim, J. Park, J. Kihm, J. Yi, J. Ye, J. Kang, J. Stasko, J. Turgeson, K. Joo, M. Hu, P. Walteros, P. Chau, R. Sadana, R. Decuir, R. Boyd, S. Yang, S. Bohn, S. Muthiah, T. Liu, W. Zhuo, Y. Han, Z. Liu, ...

Selected Papers

- ▶ InterAxis: Observation-level Interactive Axis Steering for Scatterplots of Multi-Dimensional Data Visualization, **TVCG**, 2015
- ▶ VisOHC: Designing Visual Analytics for Online Health Communities, **TVCG**, 2015
- ▶ Simultaneous Discovery of Common and Discriminative Topics via Joint Nonnegative Matrix Factorization, **KDD**, 2015
- ▶ To Gather Together for a Better World: Understanding and Leveraging Communities in Micro-lending Recommendation, **WWW**, 2014
- ▶ Understanding and Promoting Micro-finance Activities in Kiva.org, **WSDM**, 2014
- ▶ Weakly Supervised Nonnegative Matrix Factorization for User-Driven Clustering, **DMKD**, 2014
- ▶ Document Topic Modeling and Discovery in Visual Analytics via Nonnegative Matrix Factorization, **TVCG**, 2013
- ▶ Screen space- and Perception-based Framework for Efficient Computational Algorithms in Large-scale Visual Analytics, **CG&A**, 2013
- ▶ Heterogeneous Data Fusion via Space Alignment Using Nonmetric Multidimensional Scaling, **SDM**, 2012
- ▶ iVisClassifier: An Interactive Visual Analytics System for Classification based on Supervised Dimension Reduction, **VAST**, 2010
- ▶ p-ISOMAP: An Efficient Parametric Update for ISOMAP for Visual Analytics, **SDM**, 2009