Big Data Visual Analytics: Machine Learning Meets Visualization

Jaegul Choo

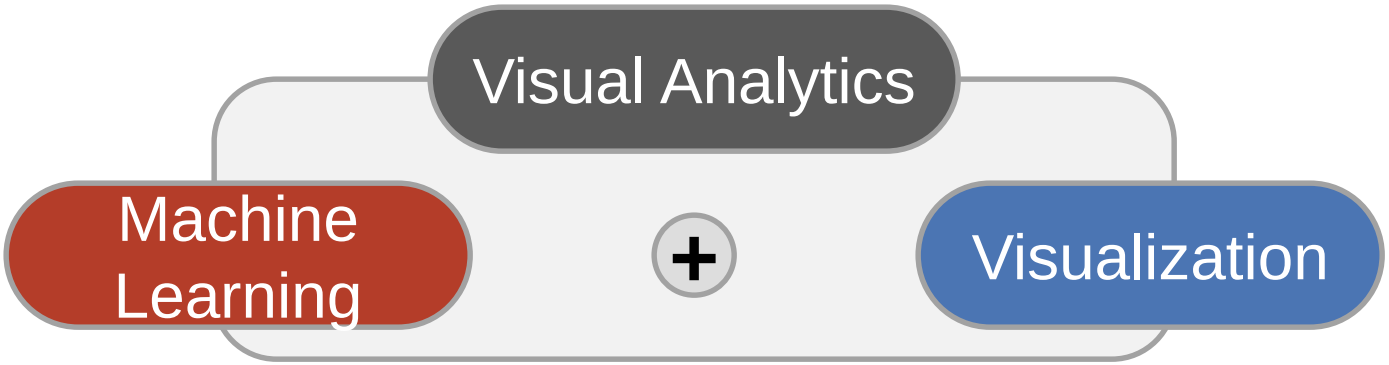
Assistant Professor

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)

High-Dimensional Data

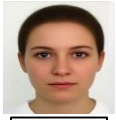
Images

- 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

22	86	15
49	54	67
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.”

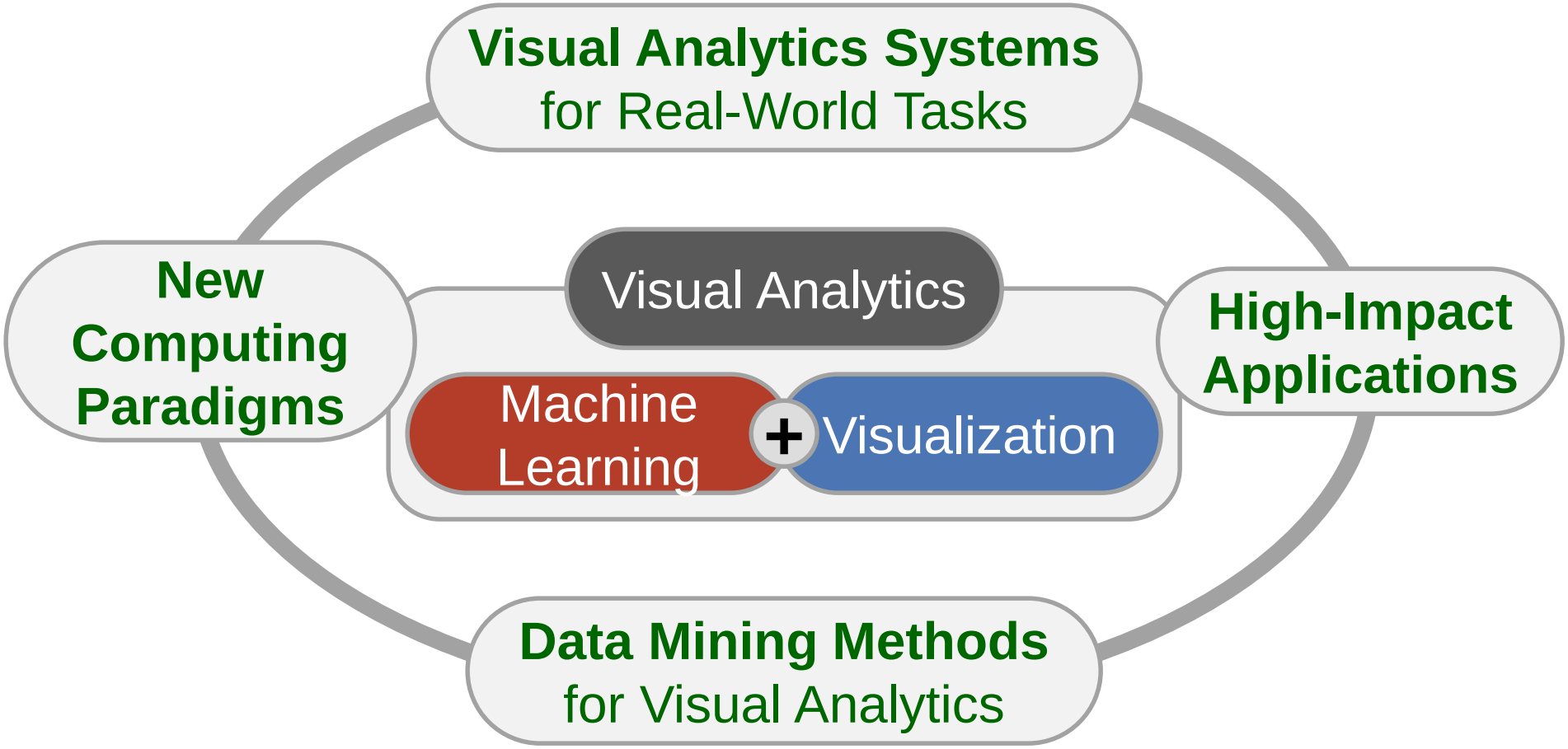
Vocabulary	Doc 1	Doc 2
John	1	1
likes	2	1
movies	1	0
also	0	1
football	0	1
Mary	1	0
too	1	0

...

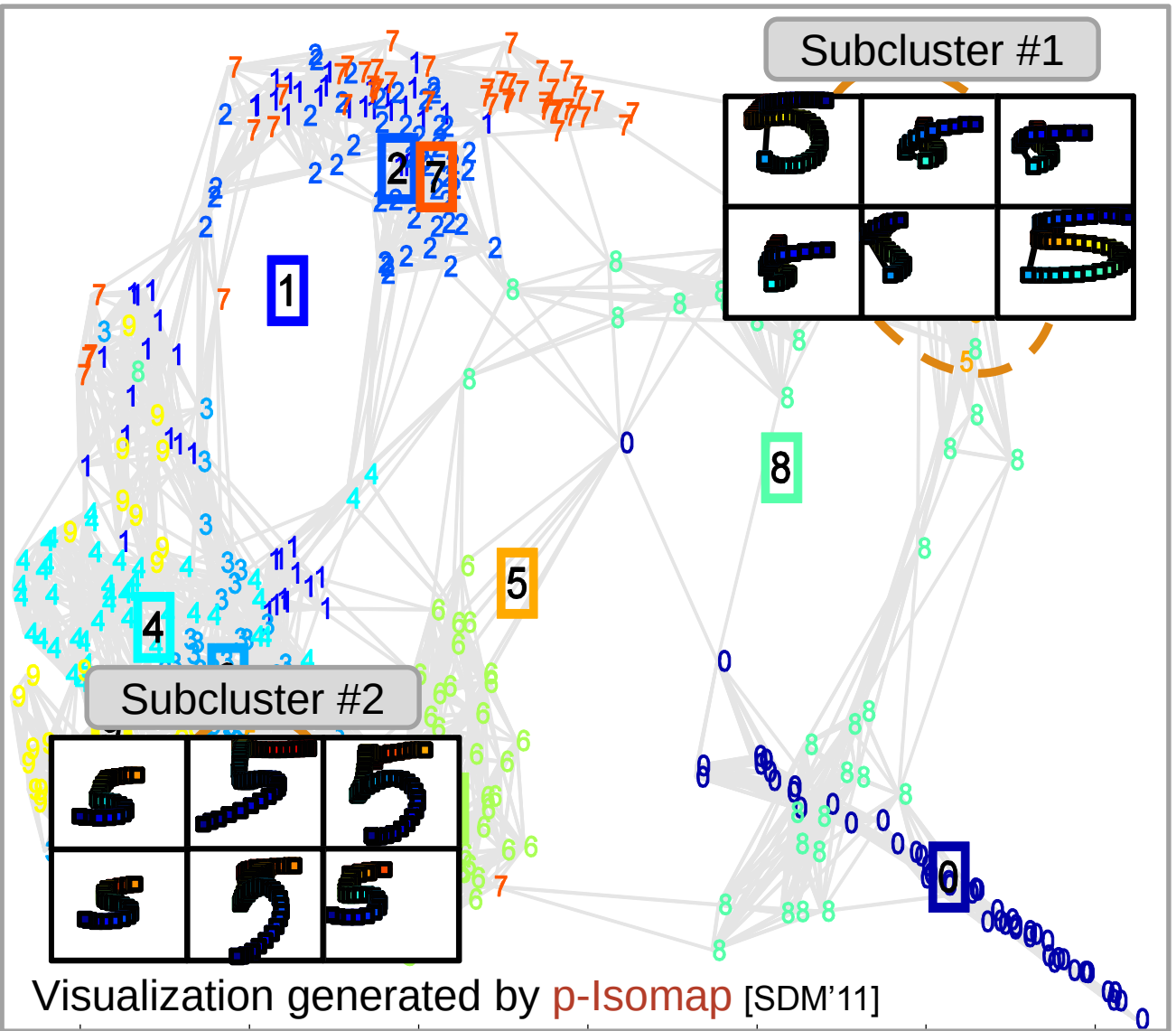
Two Approaches for Data Analysis

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



Visual Insight to Machine Learning
Handwritten Digit Recognition



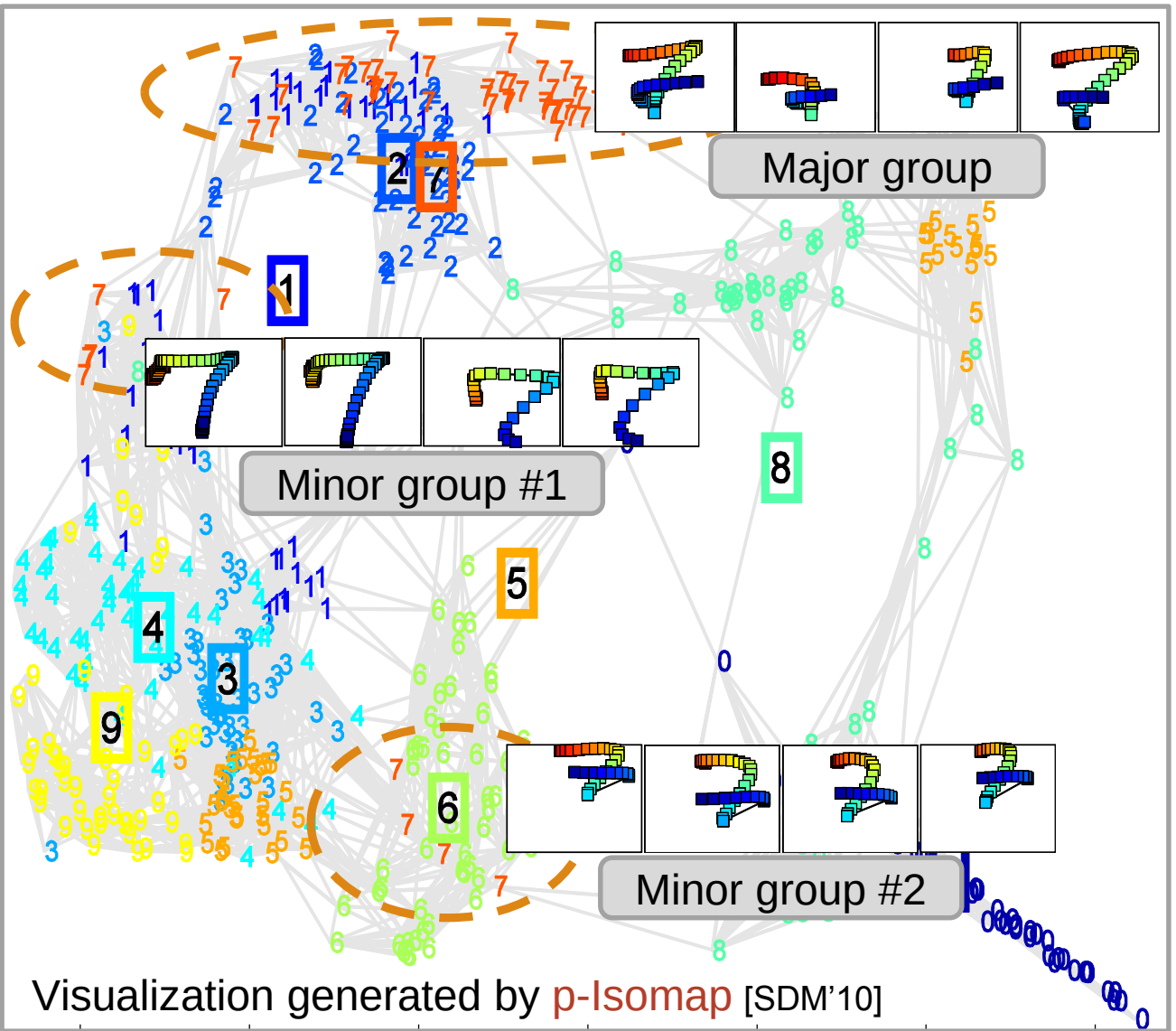
Visualization generated by p-Isomap [SDM'11]

Subclusters in digit '**5**'

Handling them as separate clusters

Better prediction (89%→ 93%)

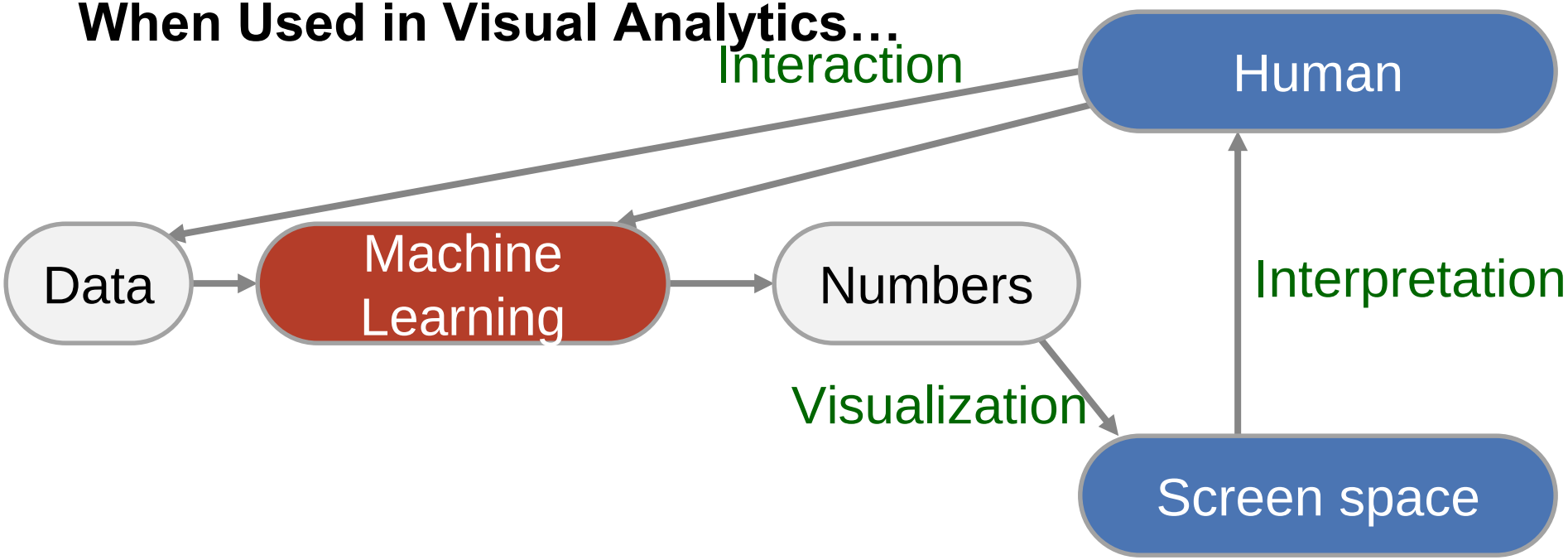
Visual Insight to Machine Learning
Handwritten Digit Recognition



Visualization generated by p-Isomap [SDM'10]

Challenges in Machine Learning + Visualization

When Used in Visual Analytics…

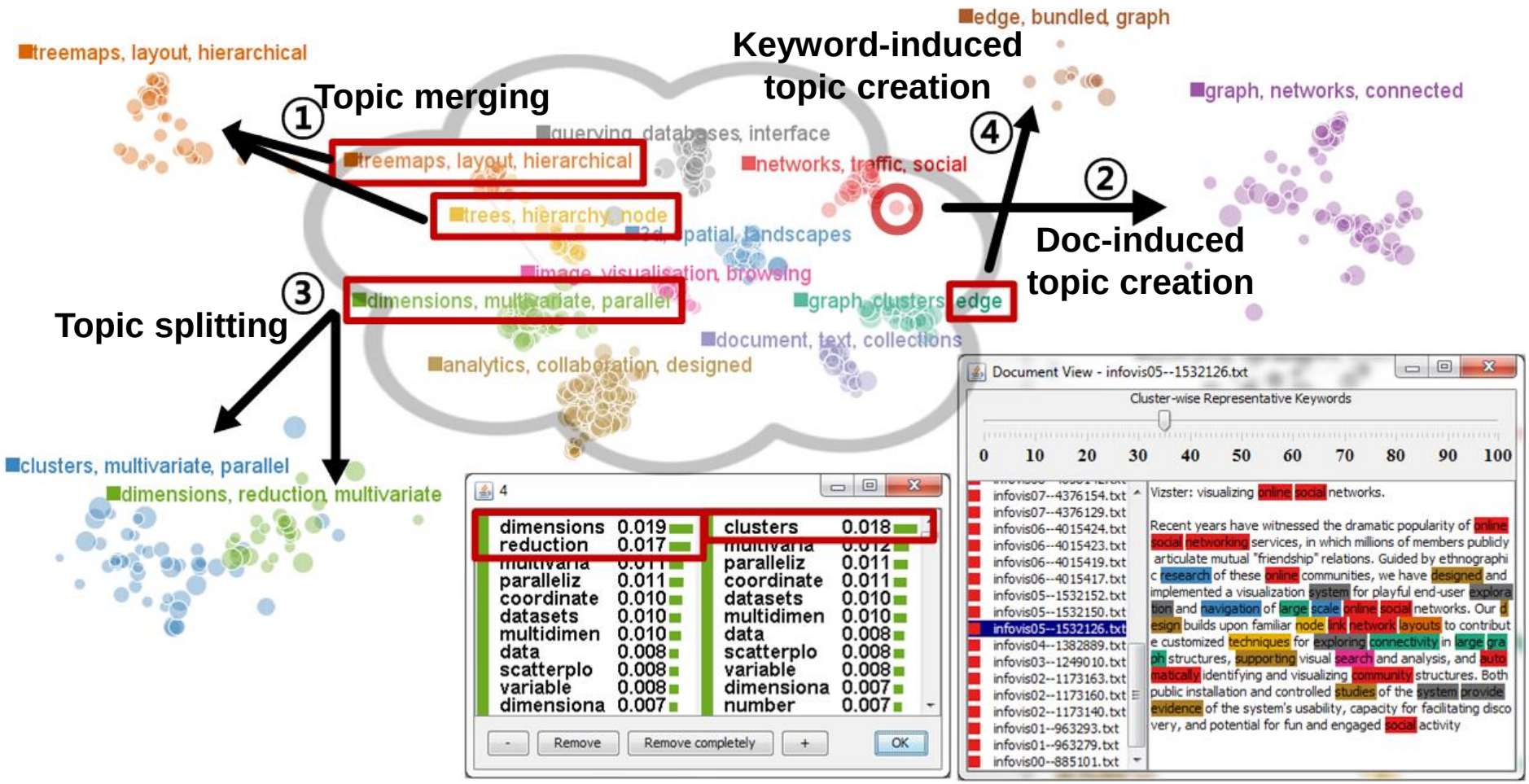


Machine learning methods should be

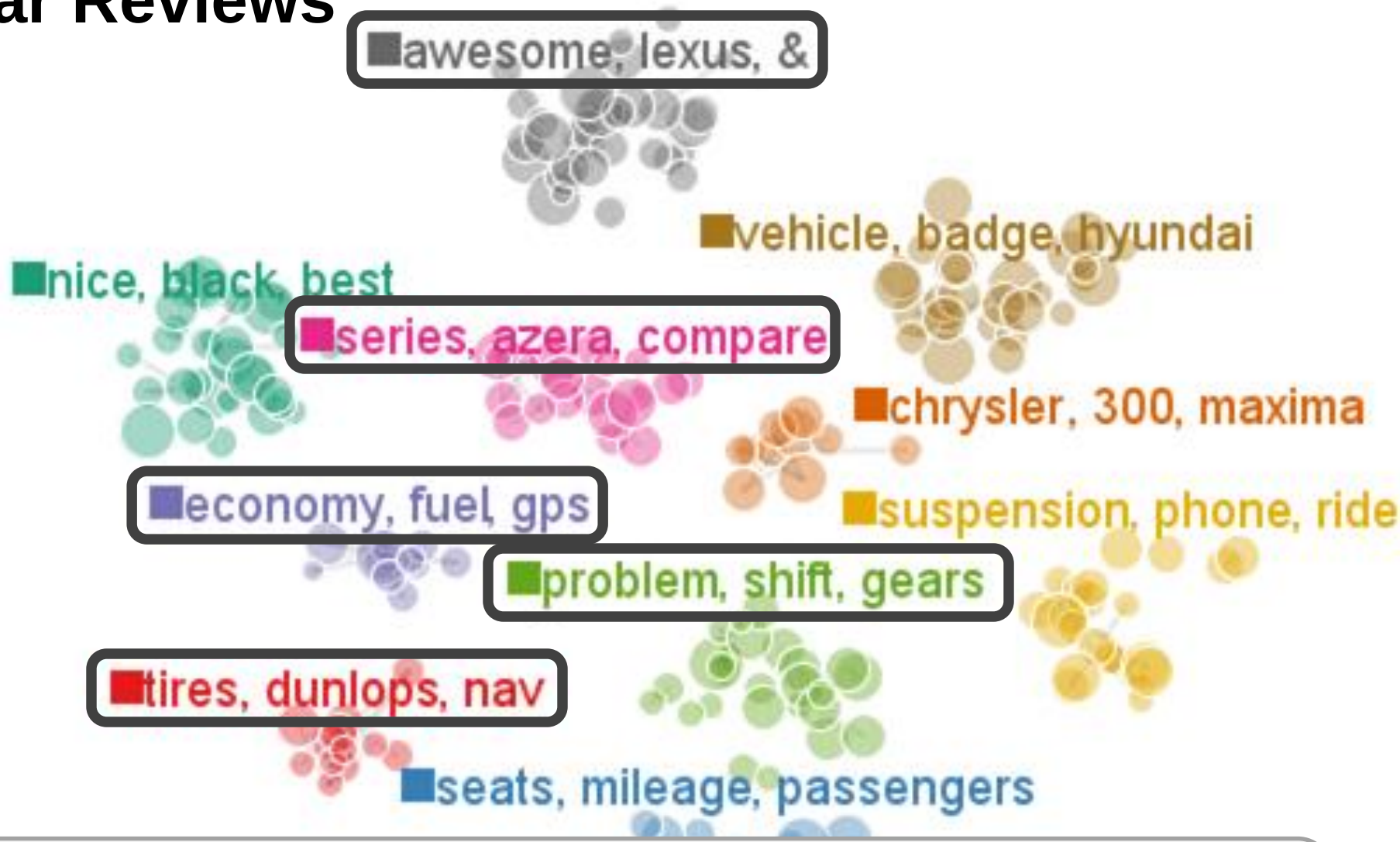
- More interpretable
- More user-interactive
- Real-time responsive, i.e., faster

UTOPIAN: User-Driven Topic Modeling Based on Interactive NMF

[TVCG 2013]



Visualization Example:
Car Reviews



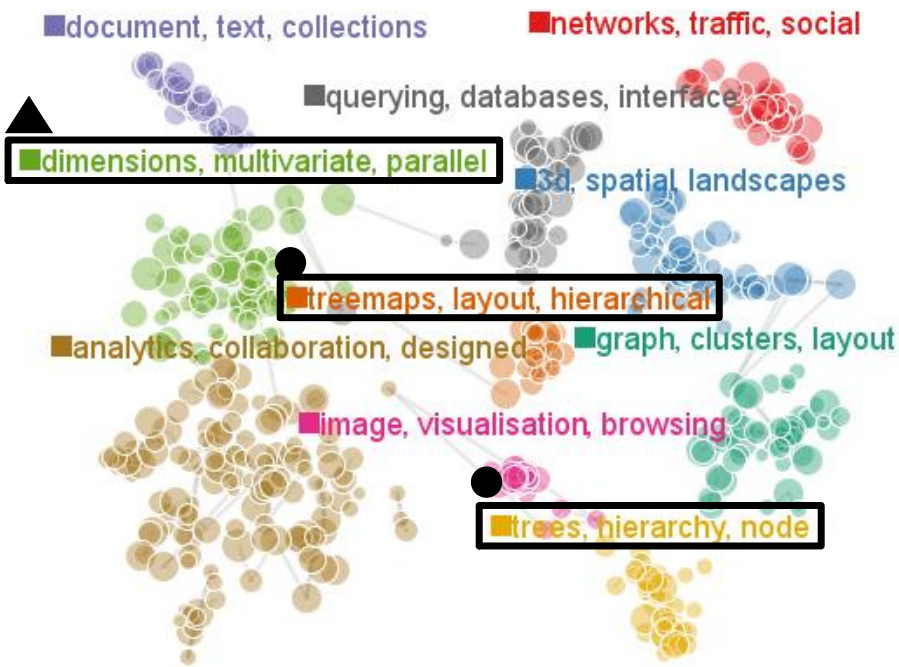
Topic summaries are NOT perfect.

➡ UTOPIAN allows user interactions for improving them.

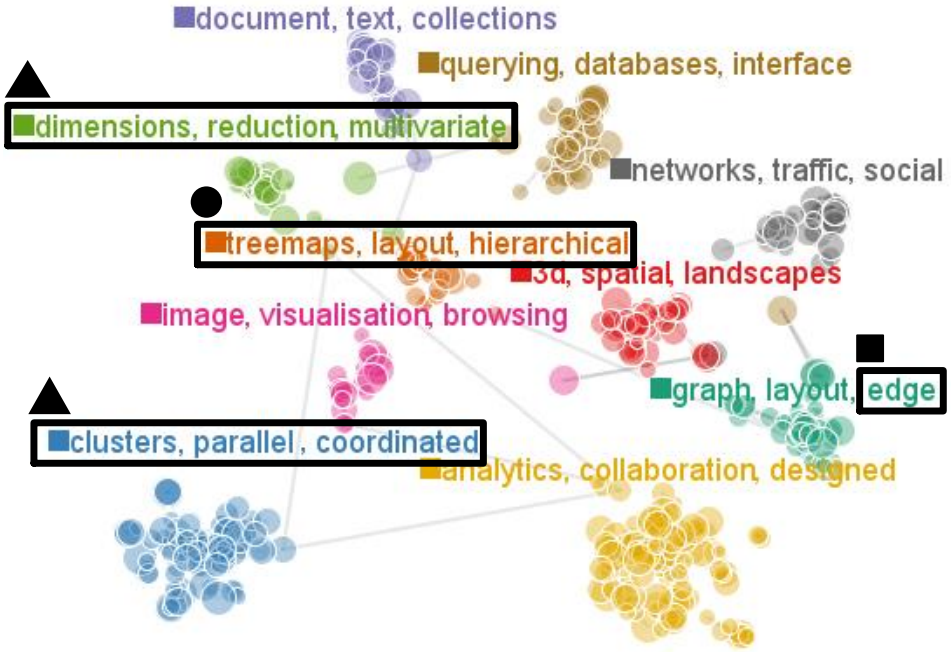
Interaction Demo Video

http://tinyurl.com/UTOPIAN2013

InfoVis-VAST Paper Data



Before interaction



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

UTOPIAN

Interactions and Key Techniques

Topic modeling

- NMF

Visualization

- **Supervised t-SNE**

Interaction

- Refining topic keywords
- Merging topics
- Splitting a topic
- Creating new topics from seed documents/keywords

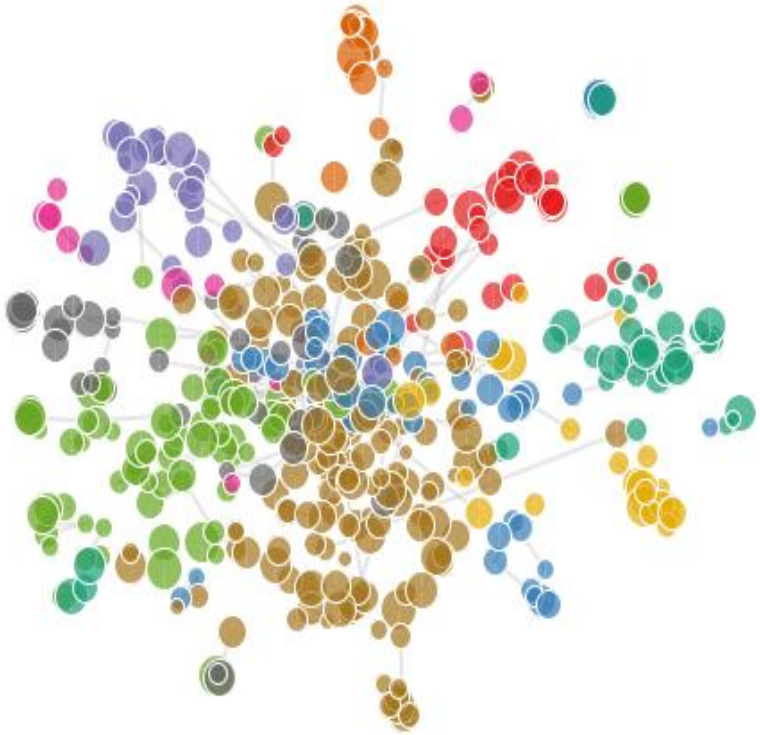
Weakly-supervised NMF

Per-Iteration Visualization Framework

Supervised t-SNE:
Visualizing documents

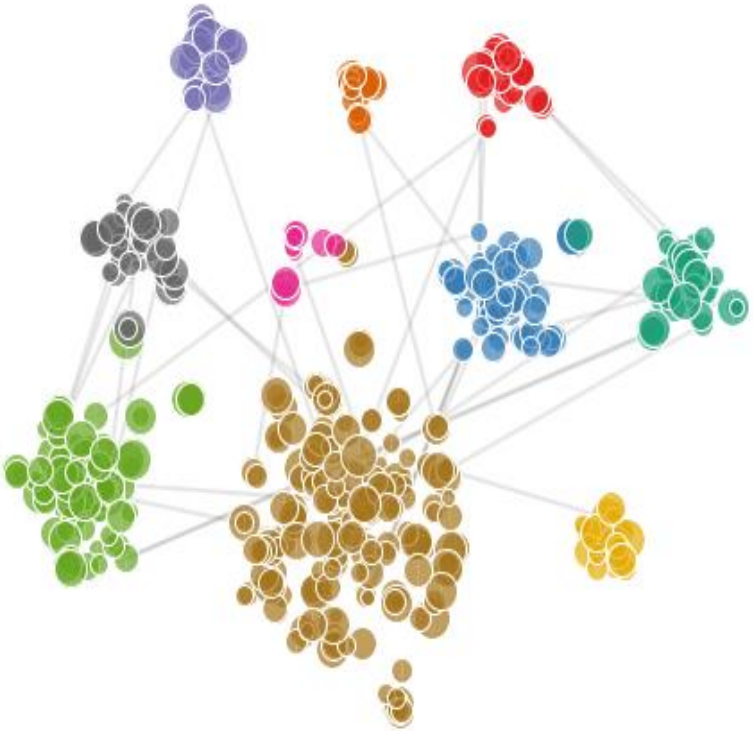
Original t-SNE

- Documents do not have clear topic clusters.



Supervised t-SNE

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



Weakly Supervised NMF:
Supporting user interactions

Weakly supervised NMF [DMKD 2014]

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

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

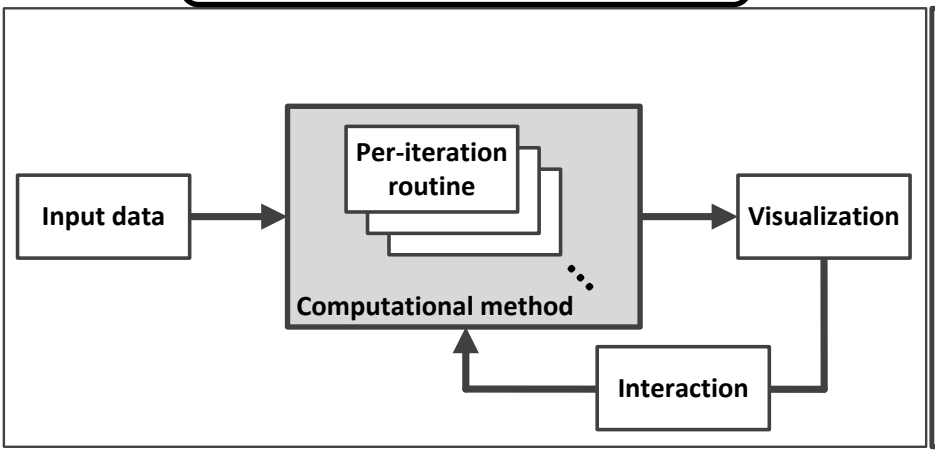
Motivation

- Many algorithms are iterative methods.

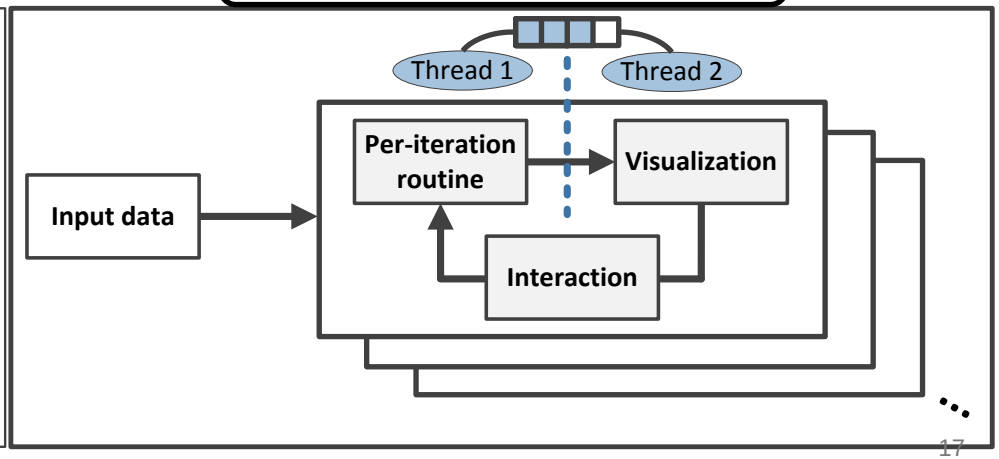
PIVE

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

Standard approach



PIVE approach



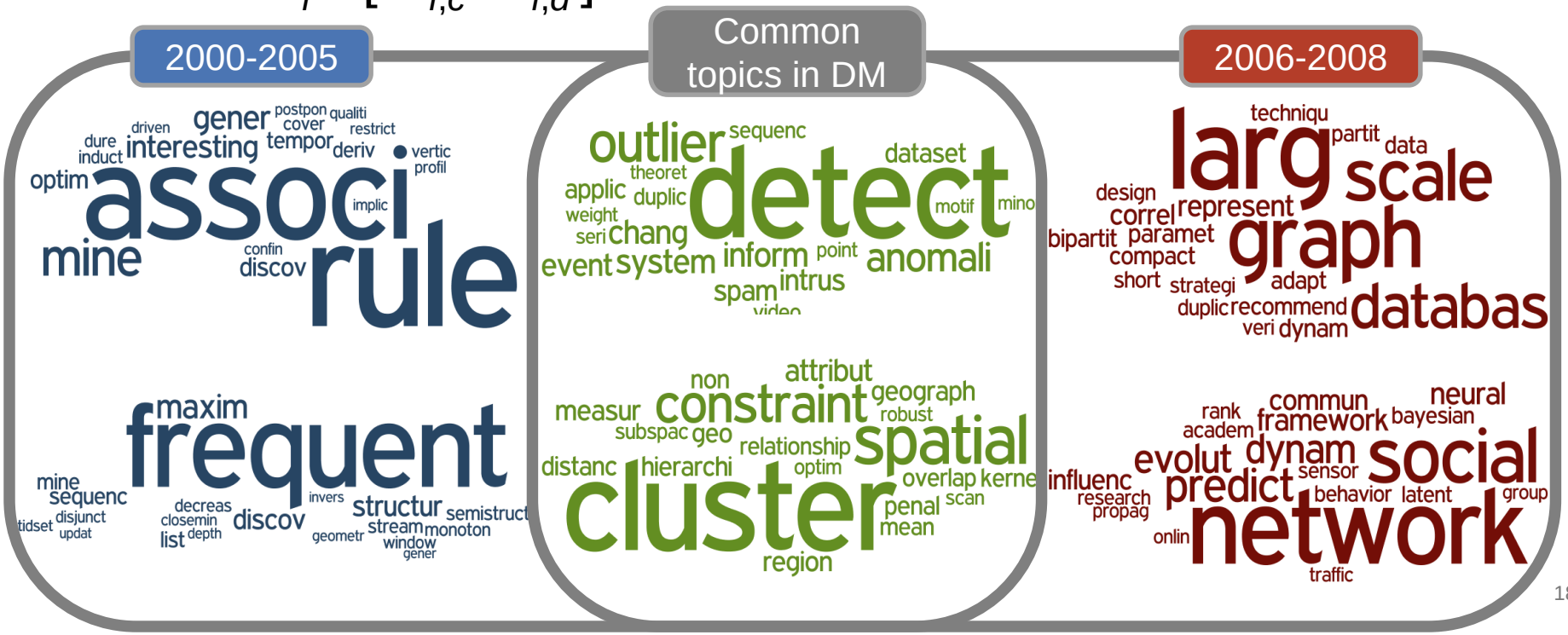
Compare and Contrast: Joint Topic Discovery

[KDD'15]

Formulation

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

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



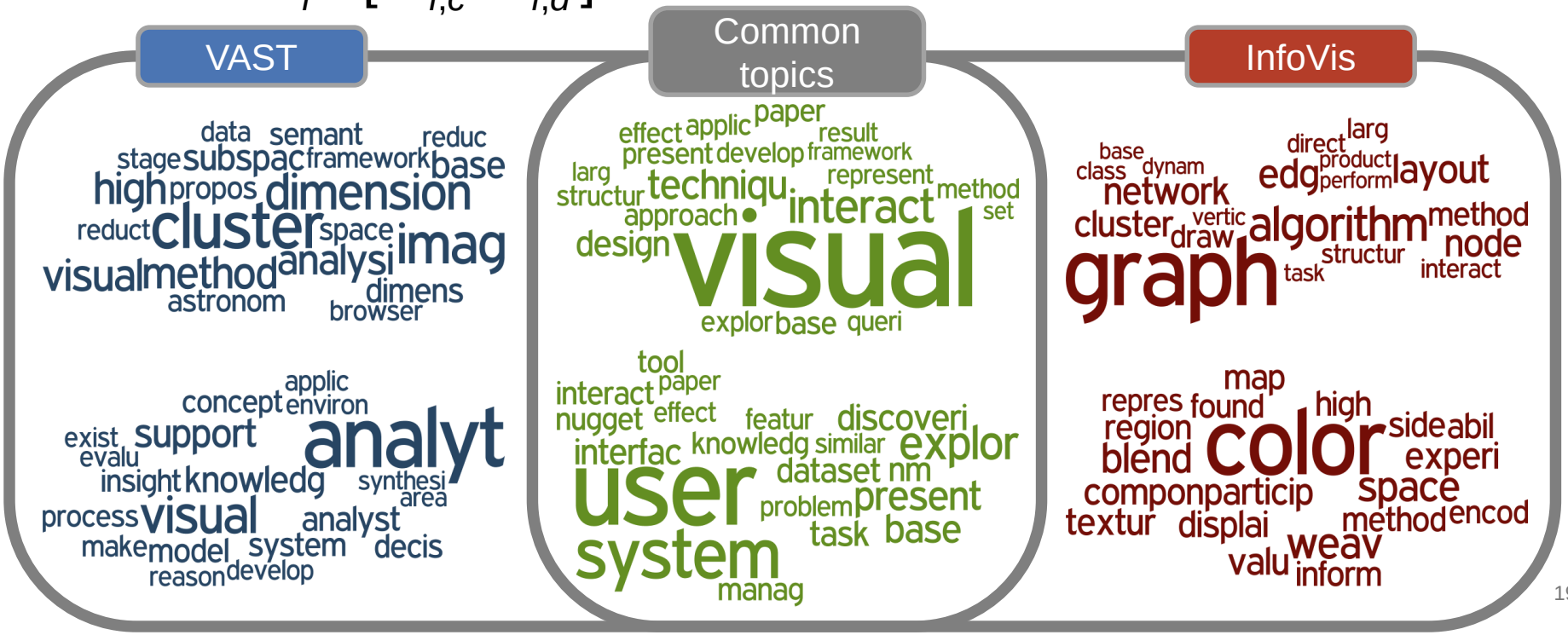
Compare and Contrast: Joint Topic Discovery

[KDD'15]

Formulation

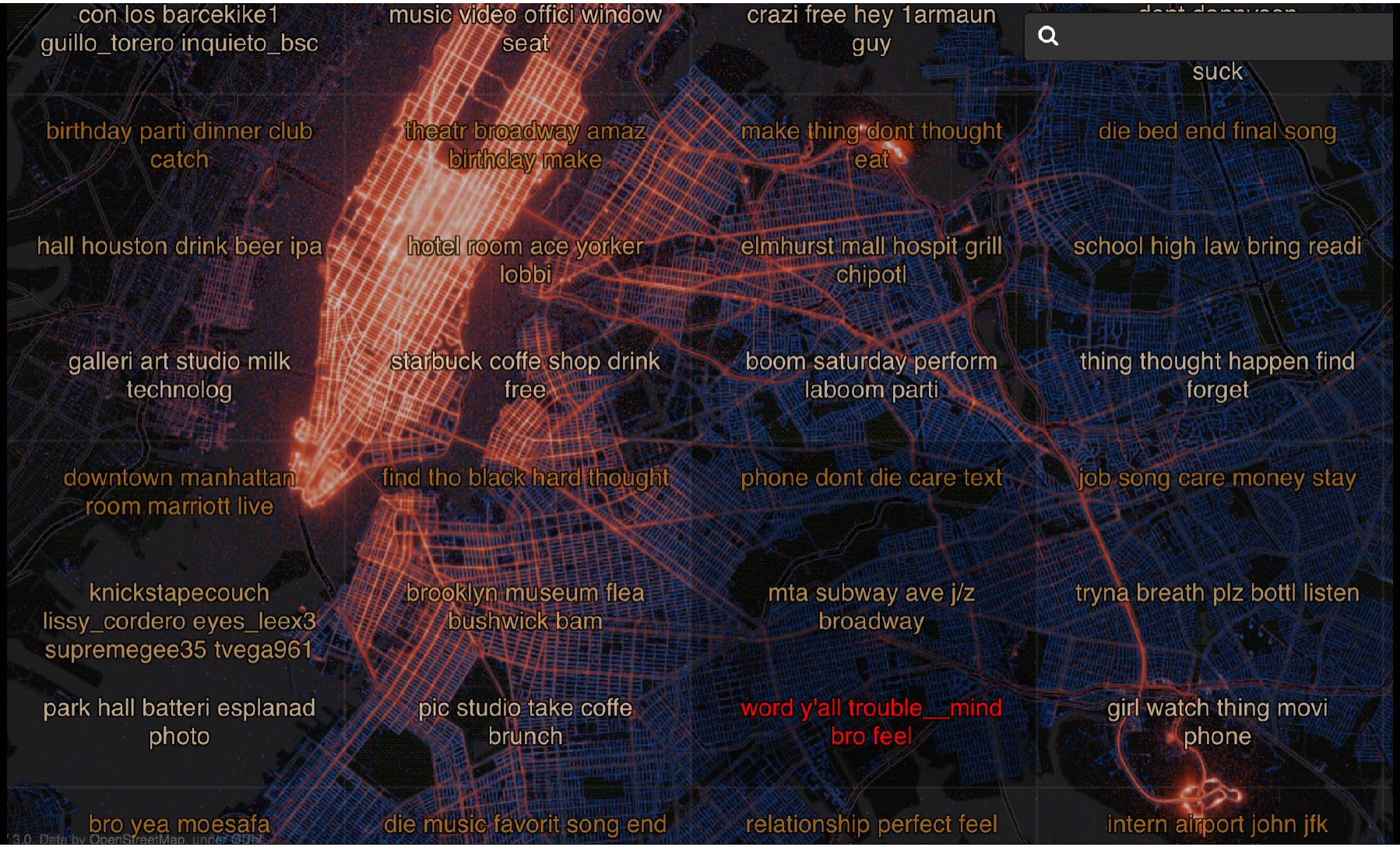
$$\min_{W>=0,\ H>=0} \quad 1/n_1 \| A_1 - W_1 H_1 \|_F^2 + 1/n_2 \| A_2 - W_2 H_2 \|_F^2 + \alpha \| W_{1,c} - W_{2,c} \|_F^2 + \beta \| W^T_{1,d} W_{2,d} \|_F^2$$

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



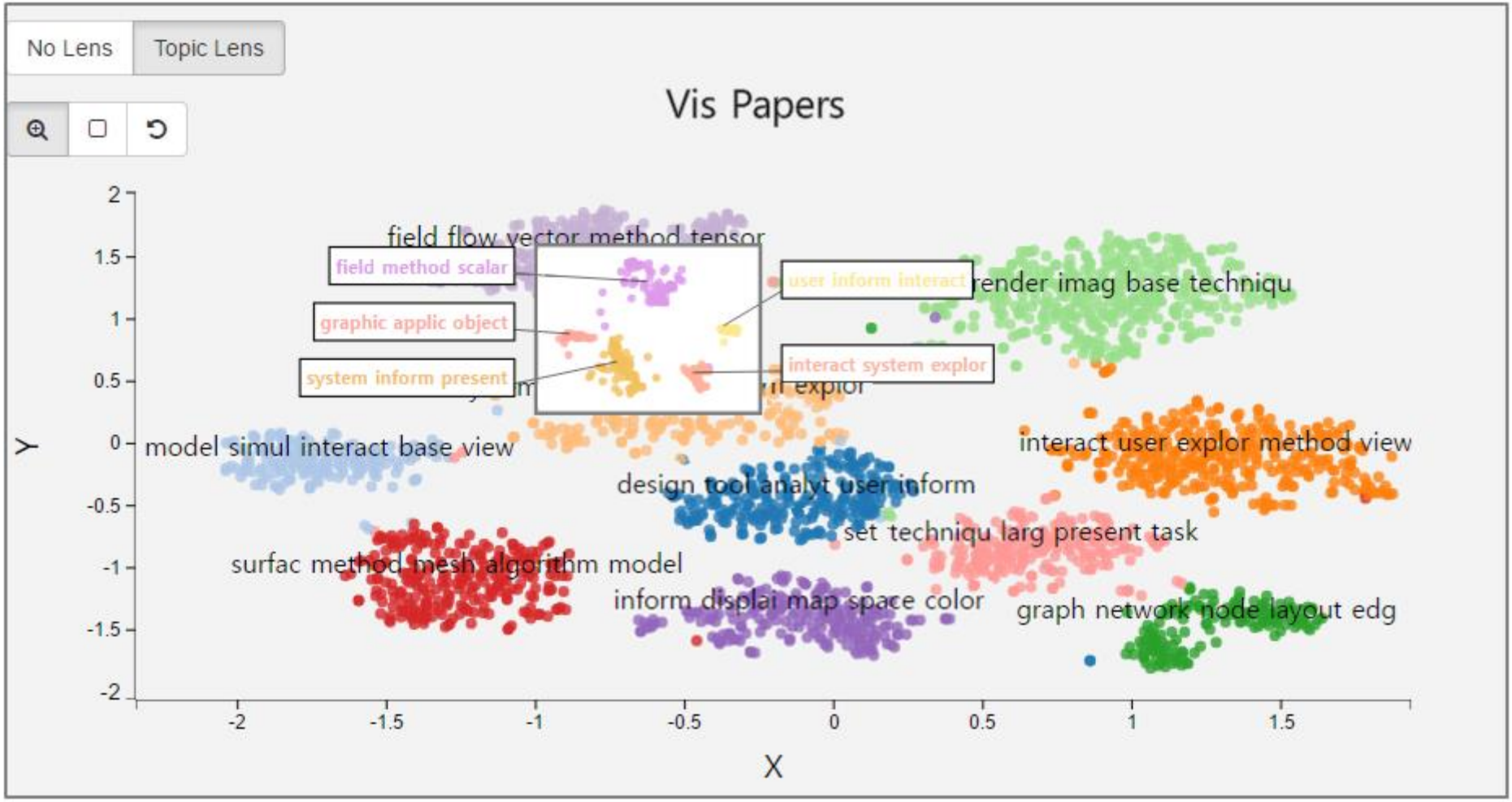
Geospatio-Temporal Topic Modeling

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



TopicLens: Efficient Multi-Level Visual Topic Exploration

[*Under submission*]



TopicLens: Efficient Multi-Level Visual Topic Exploration

[*Under submission*]

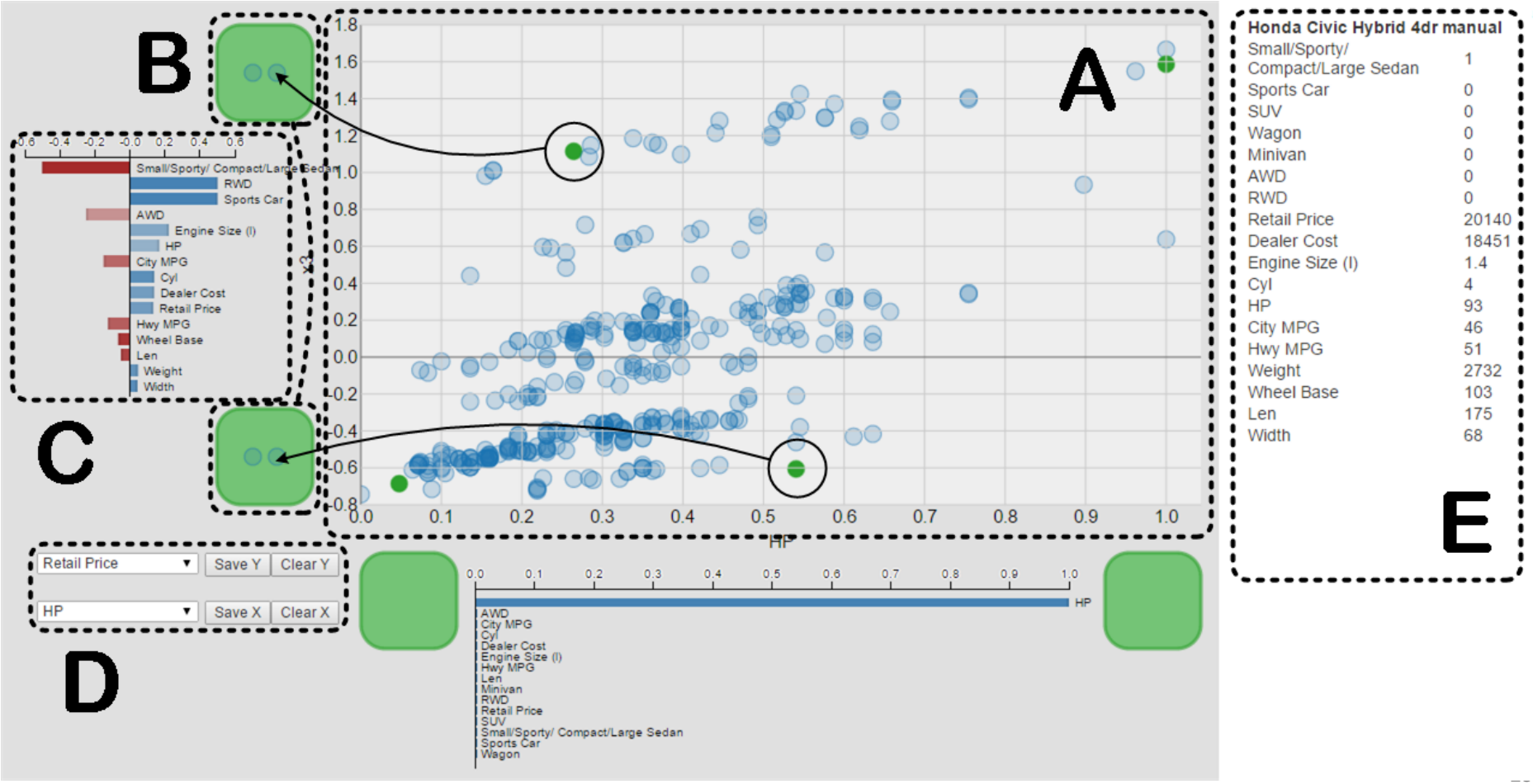
Key aspects of backend topic modeling and dimension reduction methods

- 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 submission*]

http://conceptvector.org/

Concept Name: Immigration-related
Concept Type: Unipolar

Save | Go Back

Positive Words Input

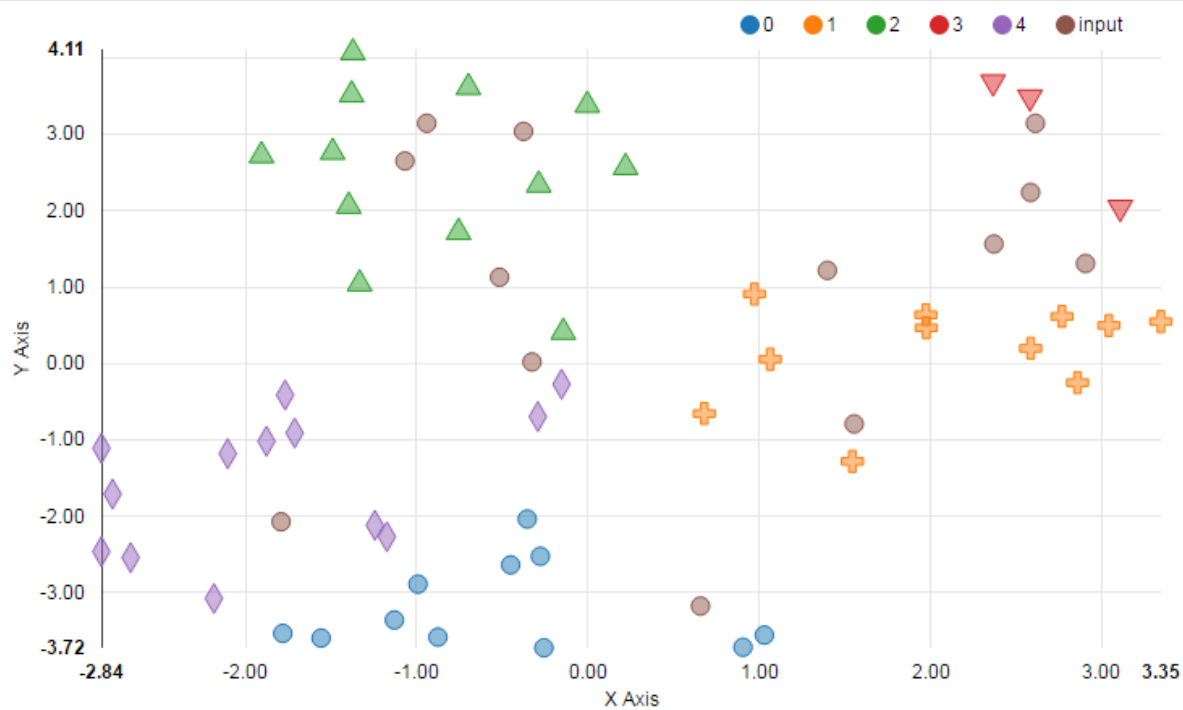
Advanced Settings

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

immigration × citizenship × naturalization × asylum × nationality × deportation × visa × visas × extradition × custody × immigrants × undocumented × migrants × 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



Perception- and Screen Space-Driven Integration Framework

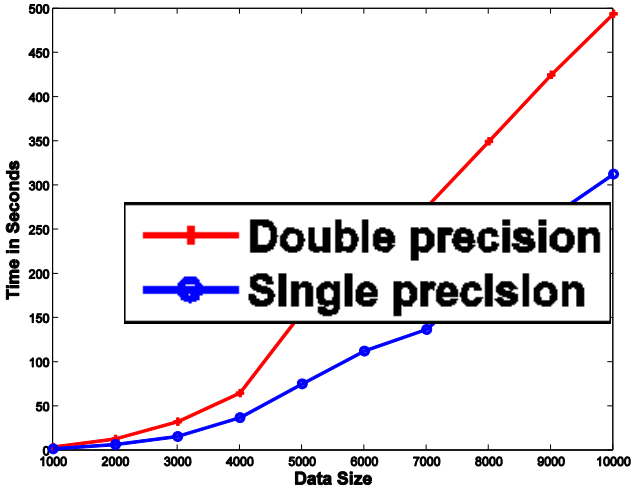
[CG&A, 2013]

Motivation

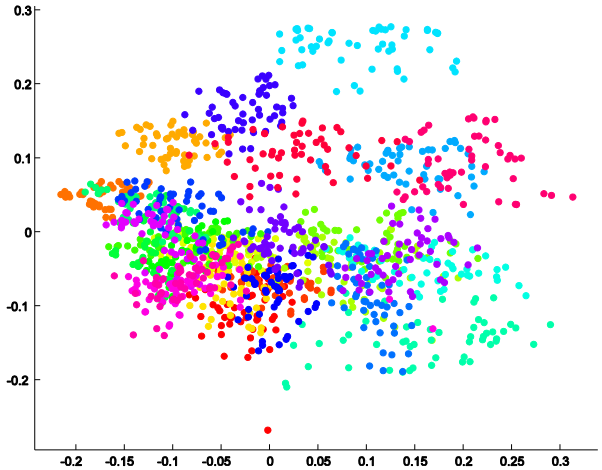
- Humans and computer screens do not require high precision.

Approach

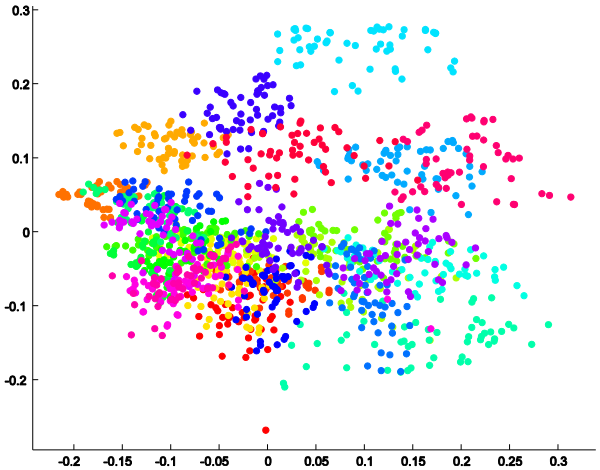
- **Approximate computing**



Computing time vs. data size



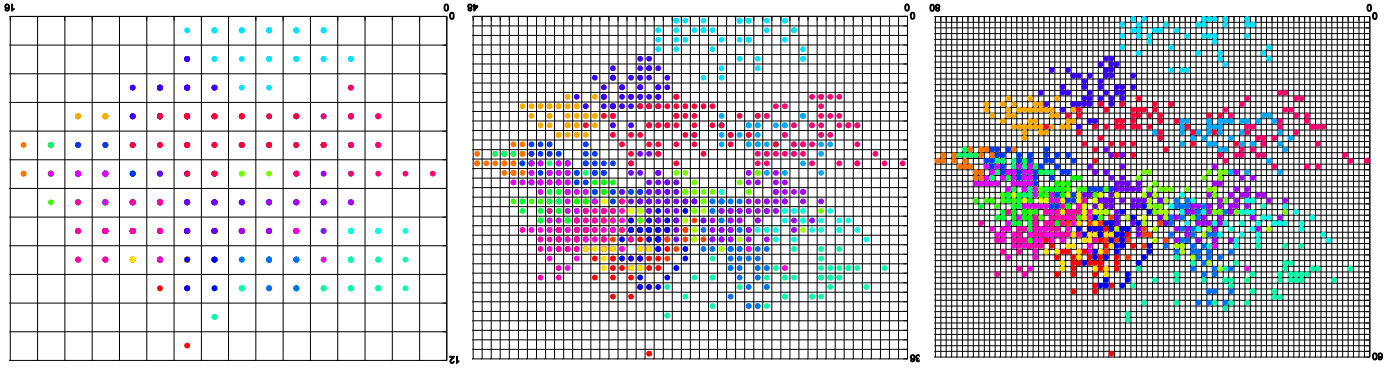
Double-precision PCA



Single-precision PCA

New Computing Paradigms for Visual Analytics

Adaptive hierarchical refinement



16x12 48x36 80x60

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

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

On-going Work

- Real-time visual analytics for deep learning
  - Visualizing the training process in real time
  - Steering the model in a user-driven manner
- Large-scale geospatio-temporal topic modeling
  - Improving NMF capability on tile-based visualization for large-scale topic modeling
- Nonlinear extension of Interaxis
  - Interactive nonlinear dimension reduction
  - Semi-supervised principal curves
- Novel applications
  - Recommendations based on brand-movie-music association

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