Real-Time Visual Analytics for User-Driven Machine Learning

Jaegul Choo

Assistant Professor

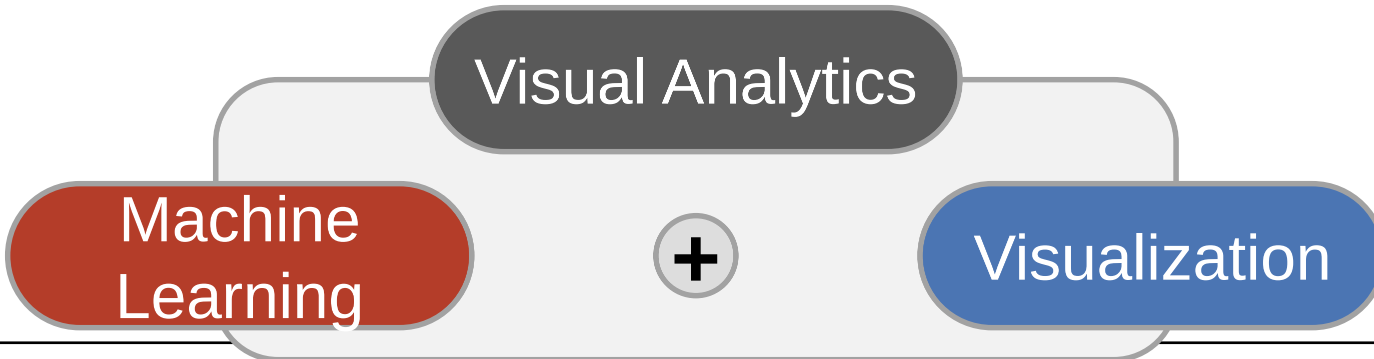
Dept. of Computer Science and Engineering

Korea University

Two Approaches for Data Analysis

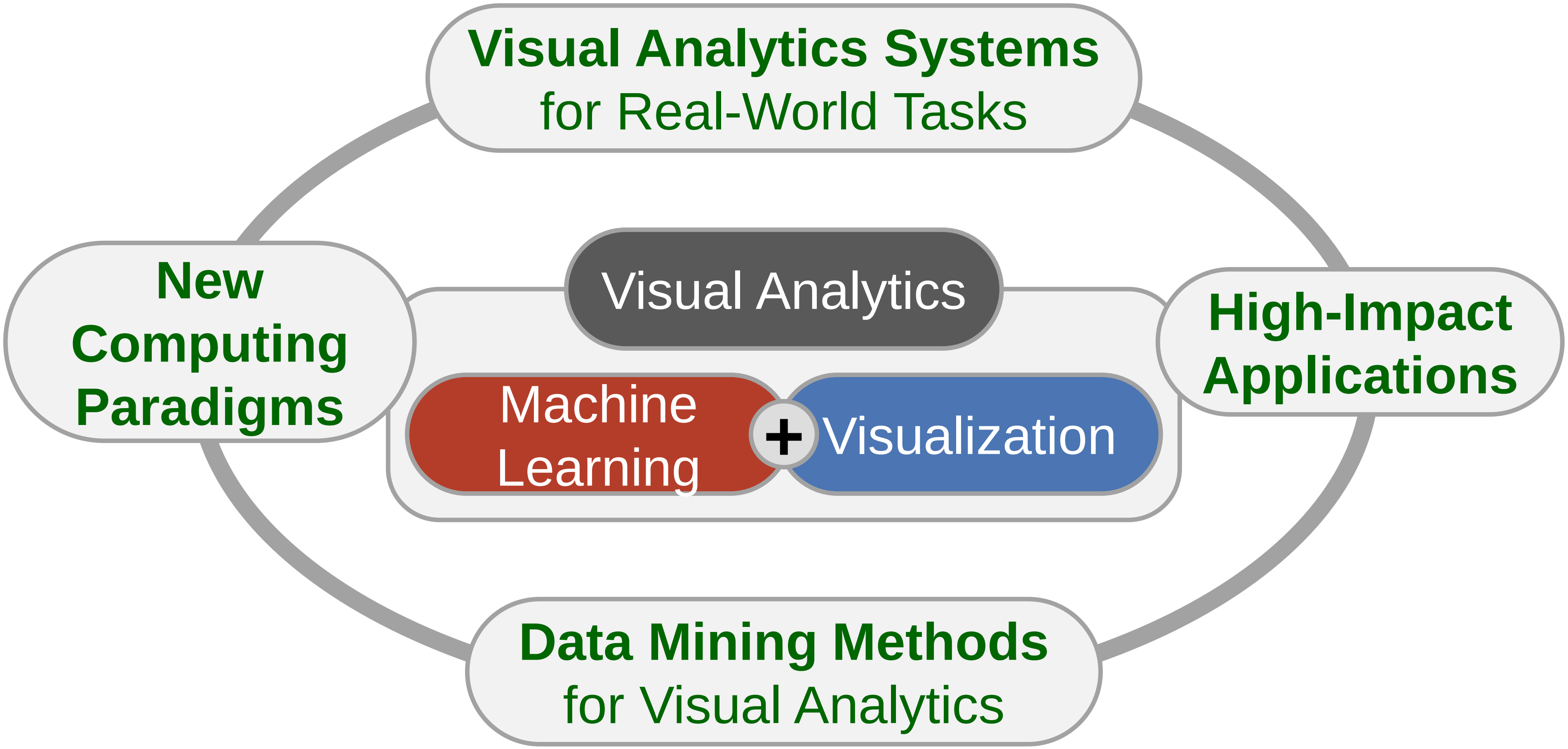
Machine Learning	Visualization
Automated	Interactive (human in the loop)
Clearly defined tasks	Exploratory analysis
Fast computation	Deeper understanding
>Millions of data items	Thousands of data items

Visual Analytics

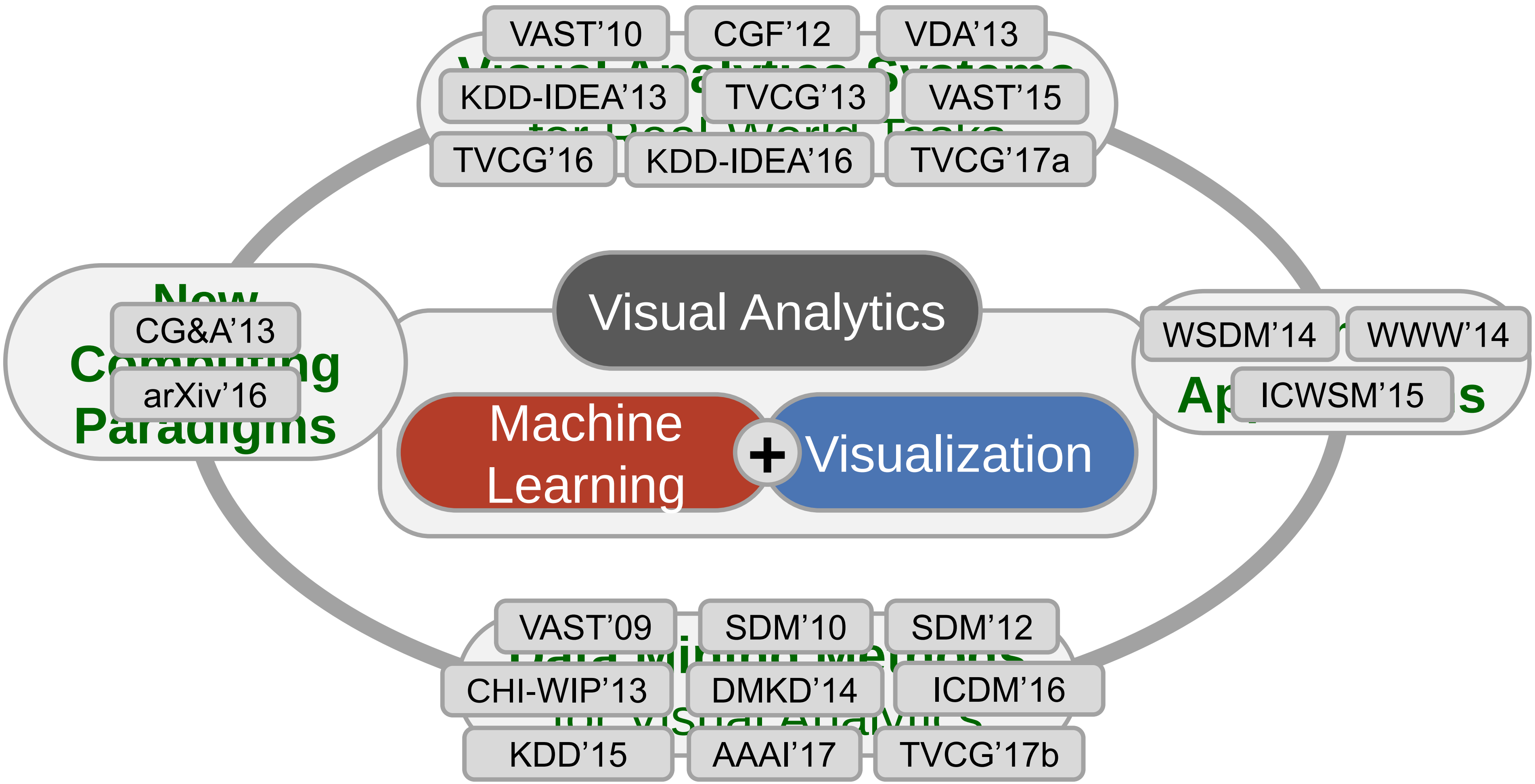


Automated	Interactive (human in the loop)
Clearly defined tasks	Exploratory analysis
Fast computation	Deeper understanding
>Millions of data items	Thousands of data items

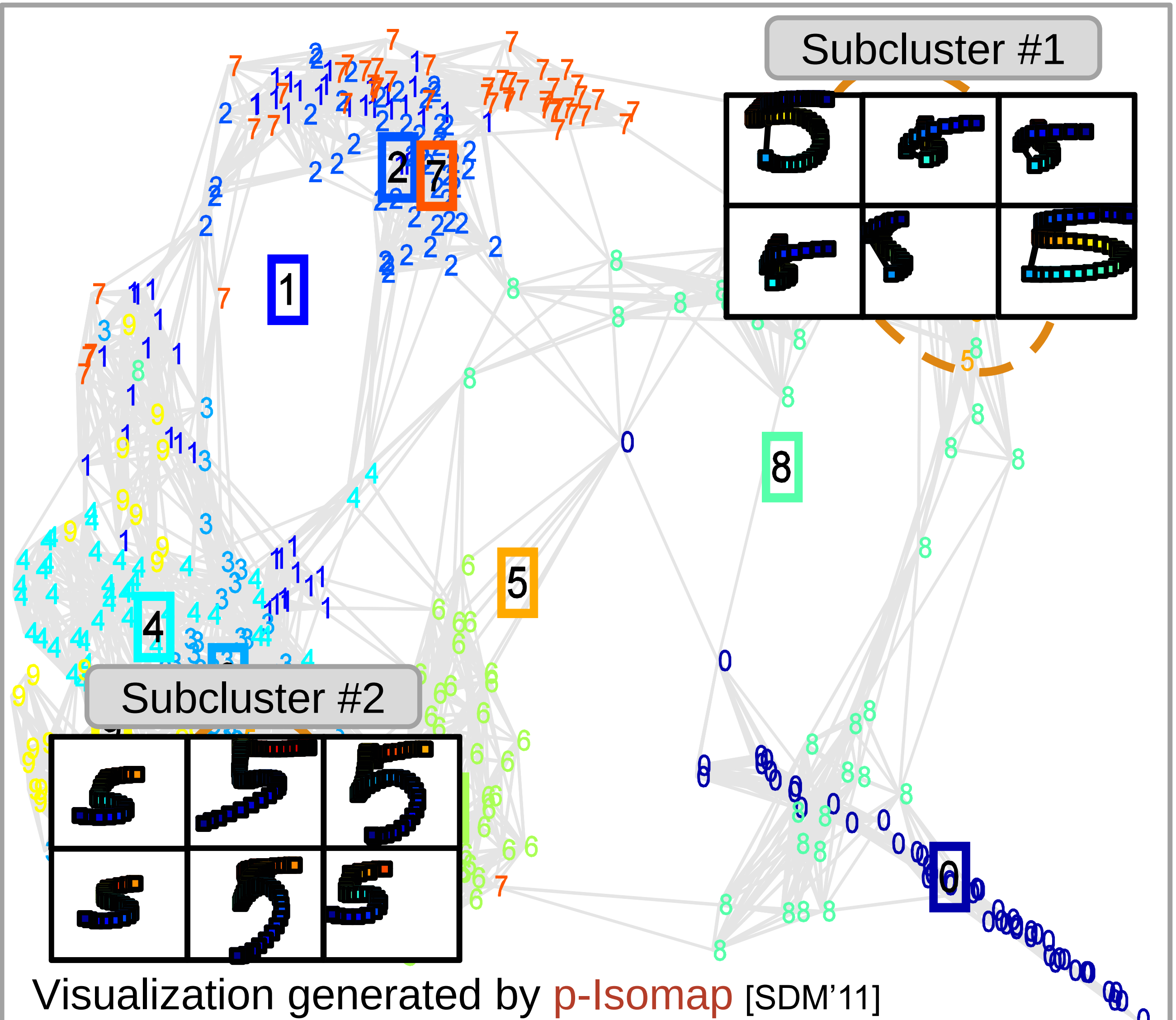
My Research:
True Integration of Both Worlds



My Research:
True Integration of Both Worlds



Visual Insight to Machine Learning
Handwritten Digit Recognition



Visualization generated by p-Isomap [SDM'11]

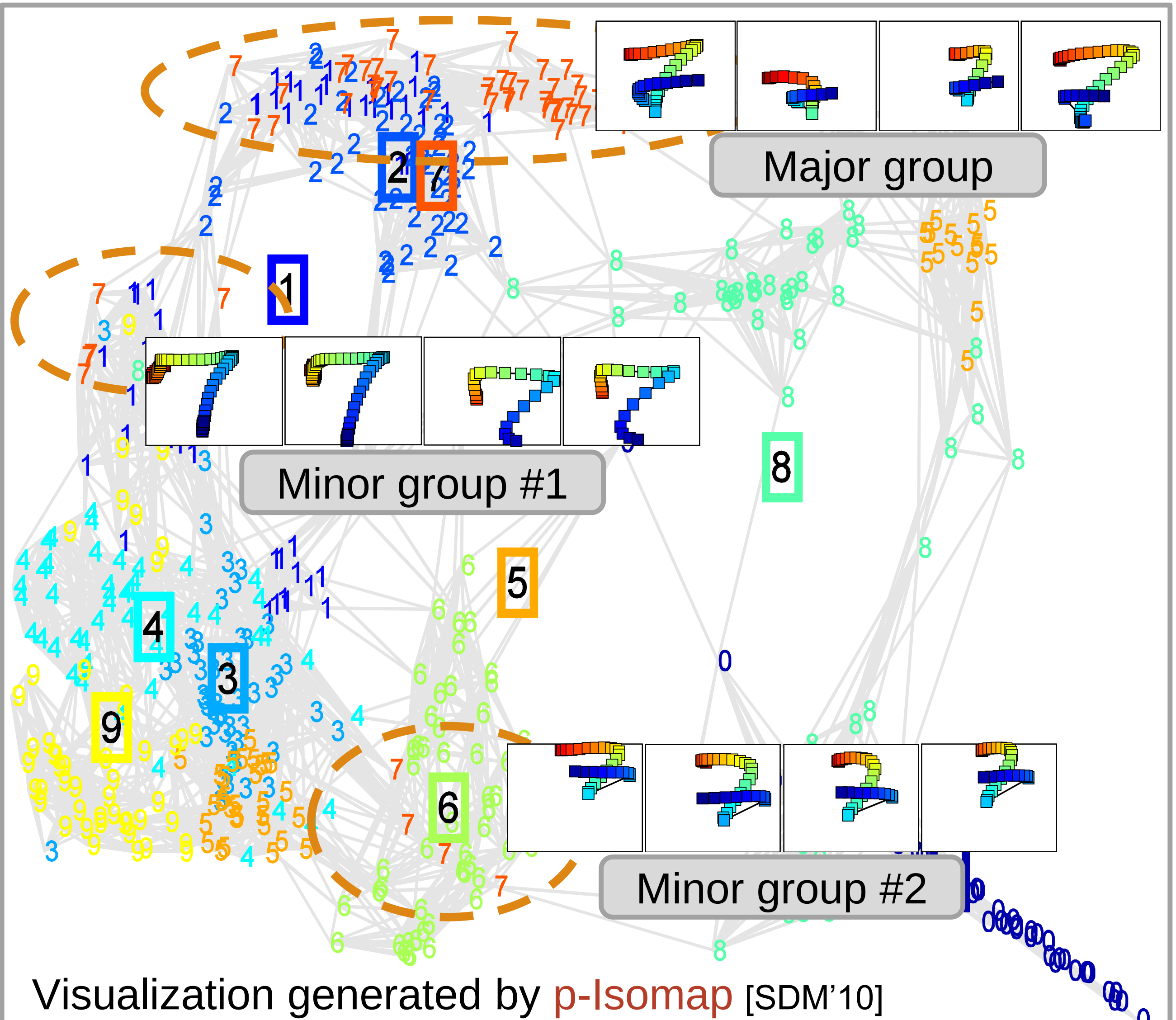
Subclusters in digit '**5**'

Handling them as separate clusters

Better prediction ($89\% \rightarrow 93\%$)

Visual Insight to Machine Learning

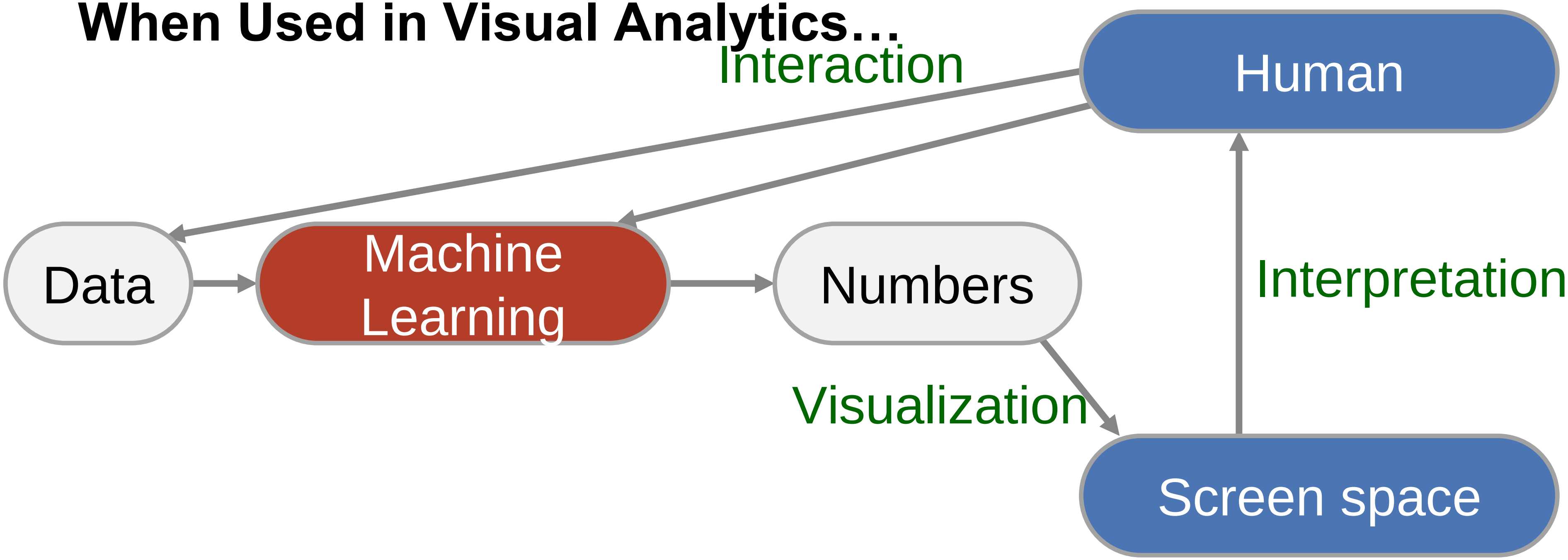
Handwritten Digit Recognition



Visualization generated by p-Isomap [SDM'10]

Challenges in
Machine Learning + Visualization

When Used in Visual Analytics…

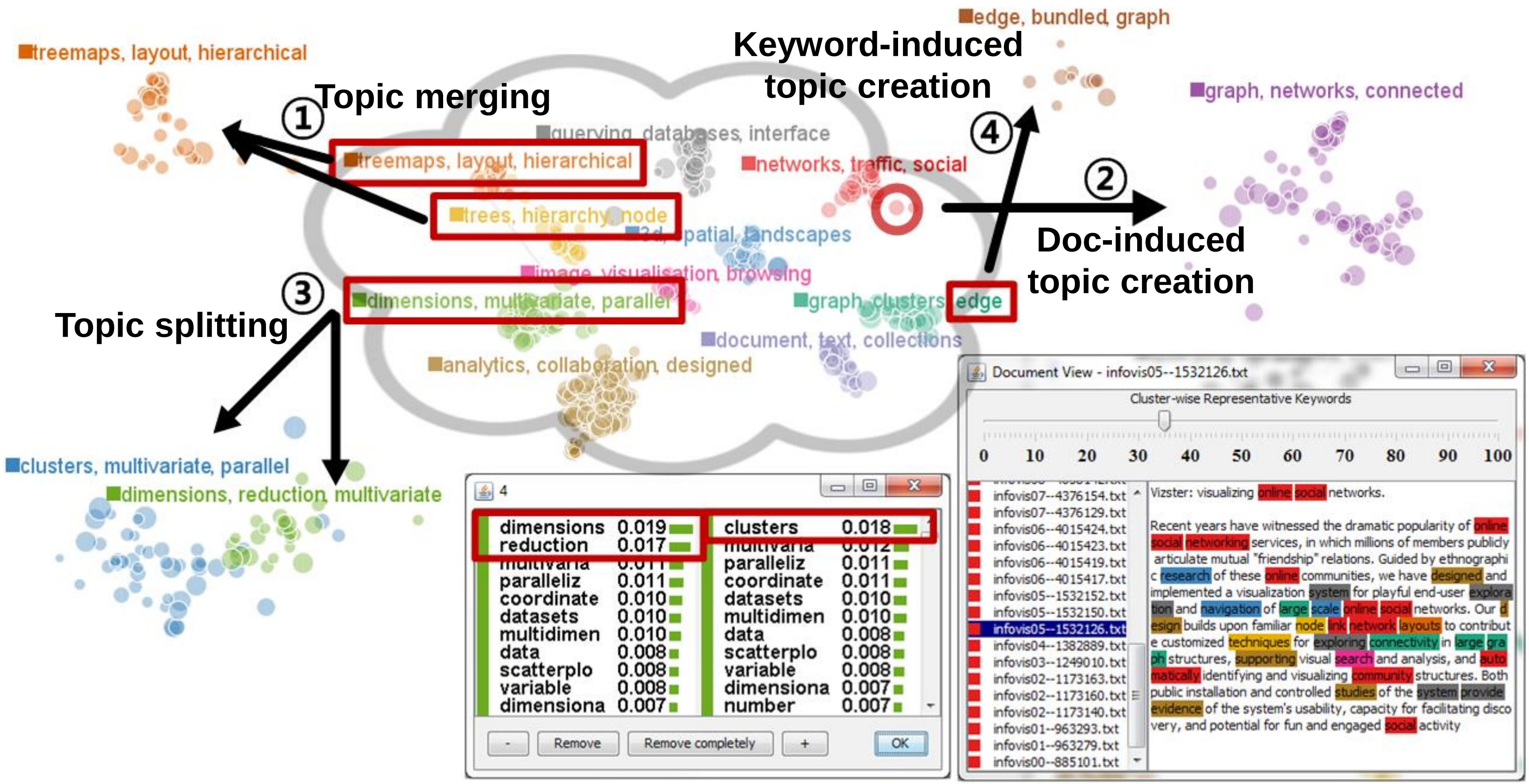


Machine learning methods should be

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

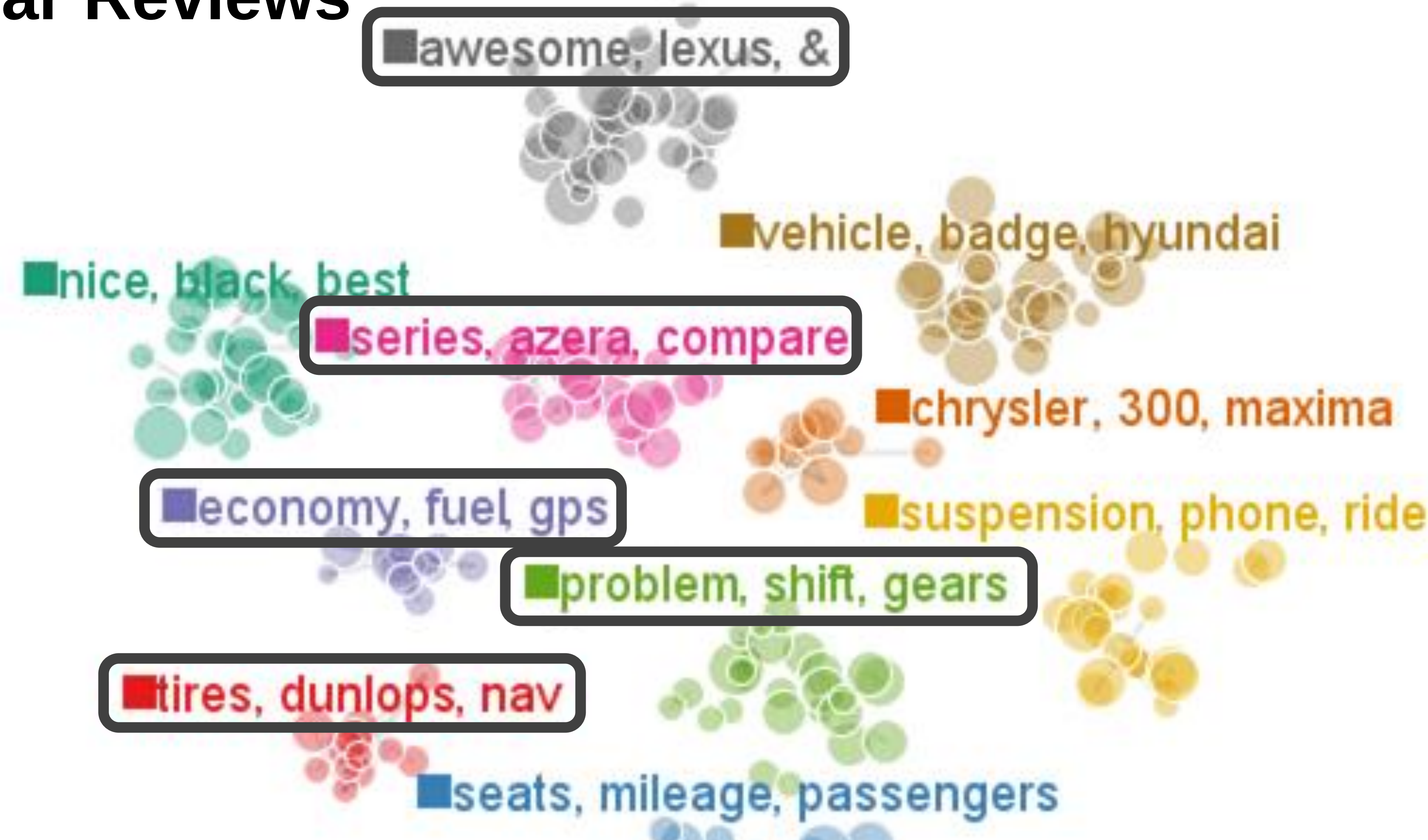
TVCG'13

UTOPIAN: User-Driven Topic Modeling Based on Interactive NMF



Visualization Example:
Car Reviews

TVCG'13



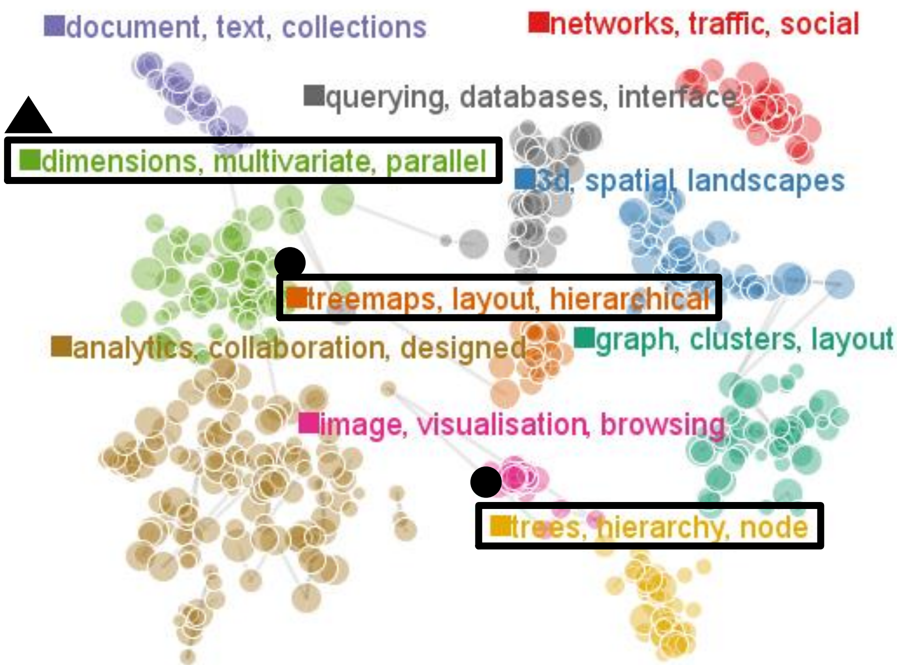
Topic summaries are NOT perfect.
➡ UTOPIAN allows user interactions for improving them.

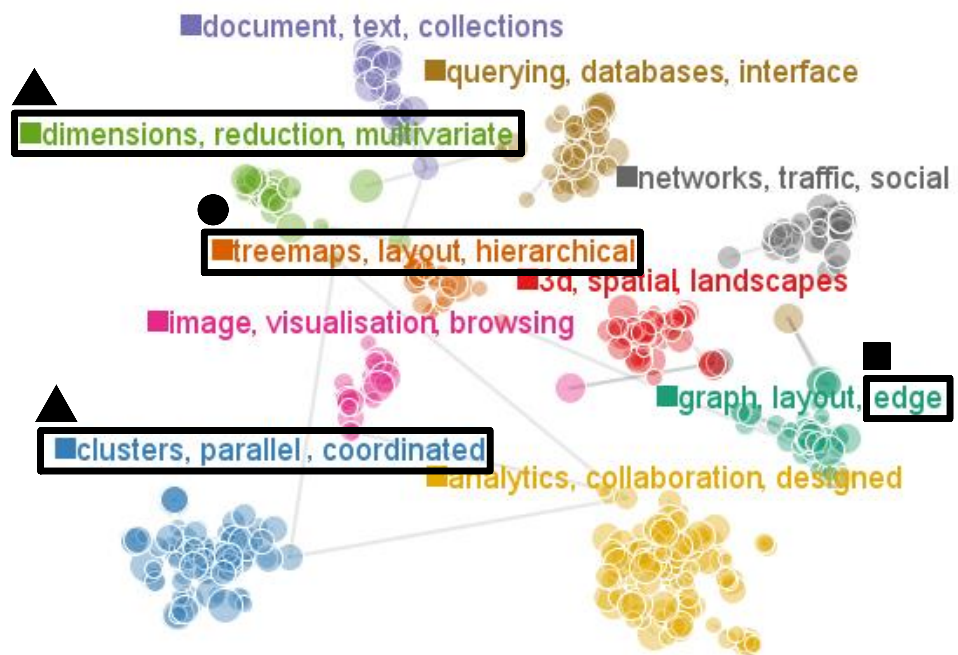
TVCG'13

UTOPIAN Demo

http://tinyurl.com/UTOPIAN2013

InfoVis-VAST Paper Data



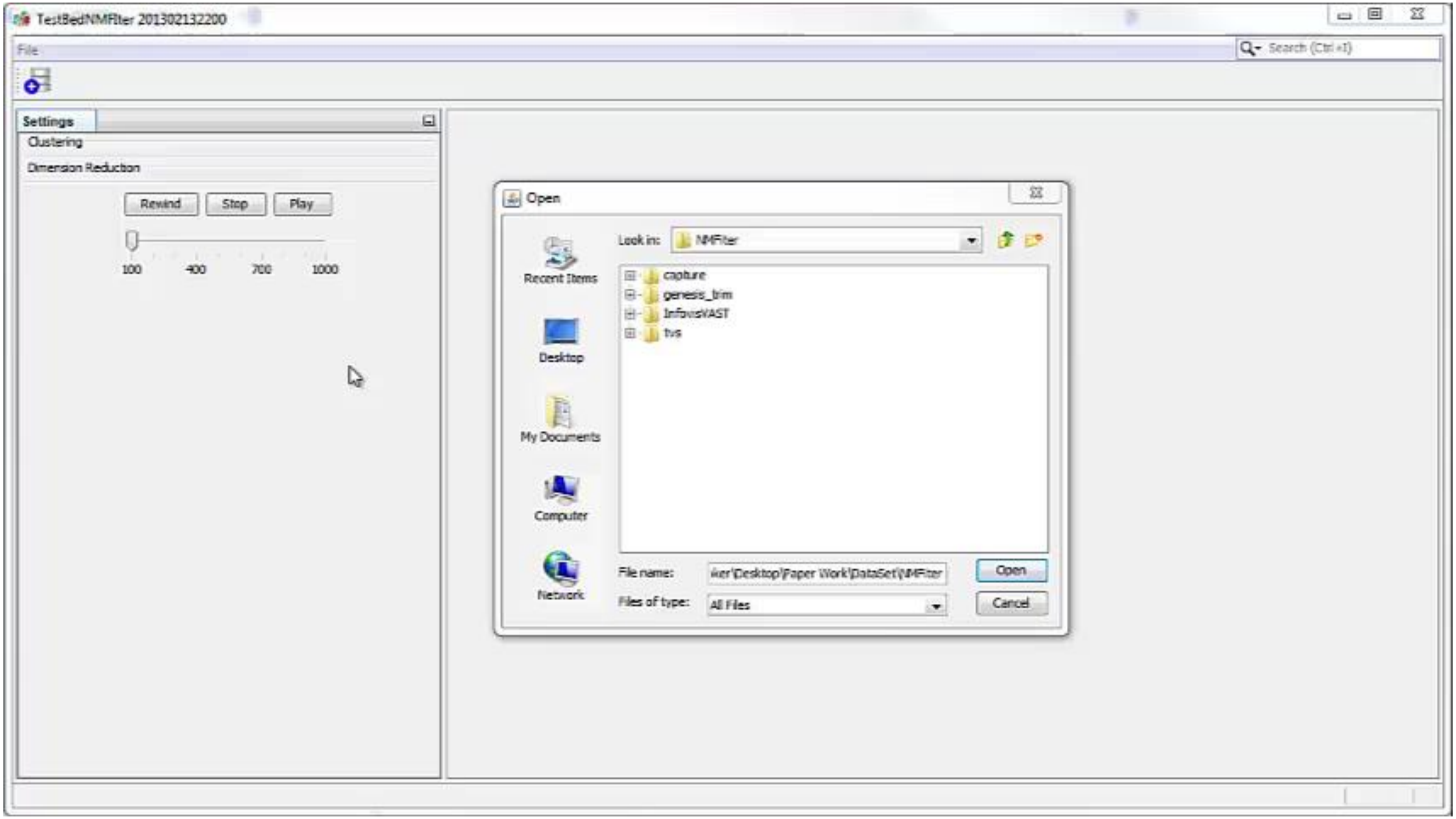


Before interaction

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

UTOPIAN Demo

http://tinyurl.com/UTOPIAN2013

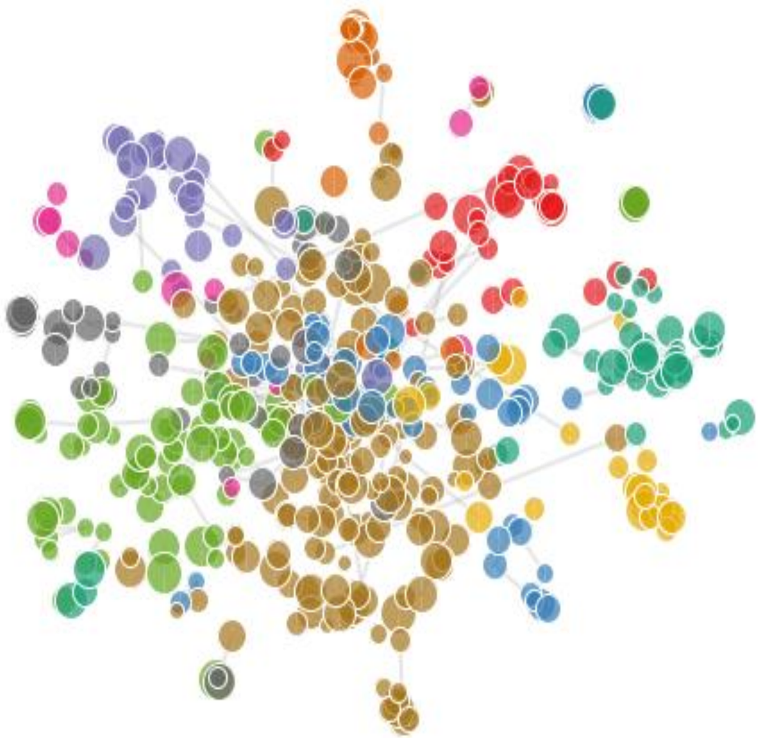


TVCG'13

Supervised t-SNE:
Visualizing documents

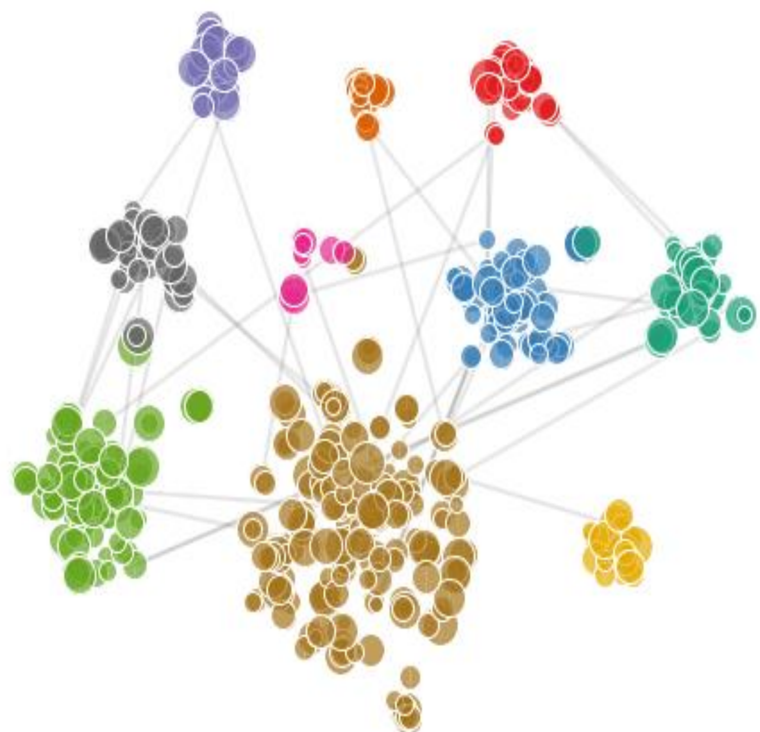
Original t-SNE

- 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

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

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

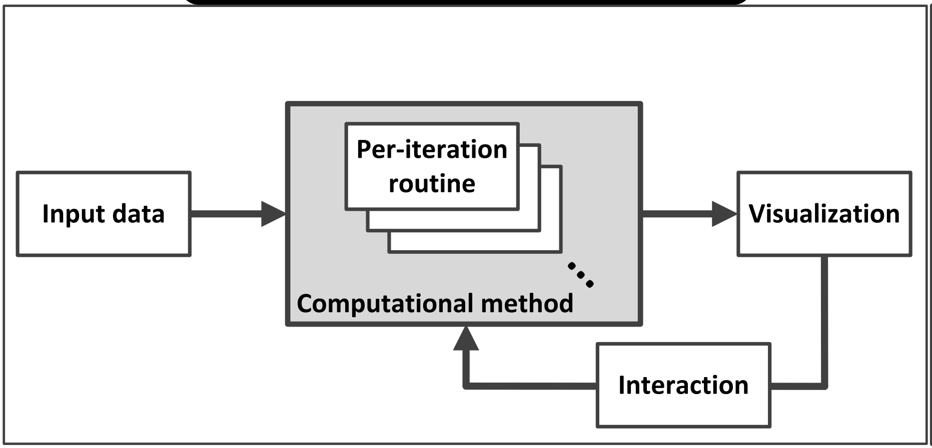
Motivation

- Many algorithms are iterative methods.

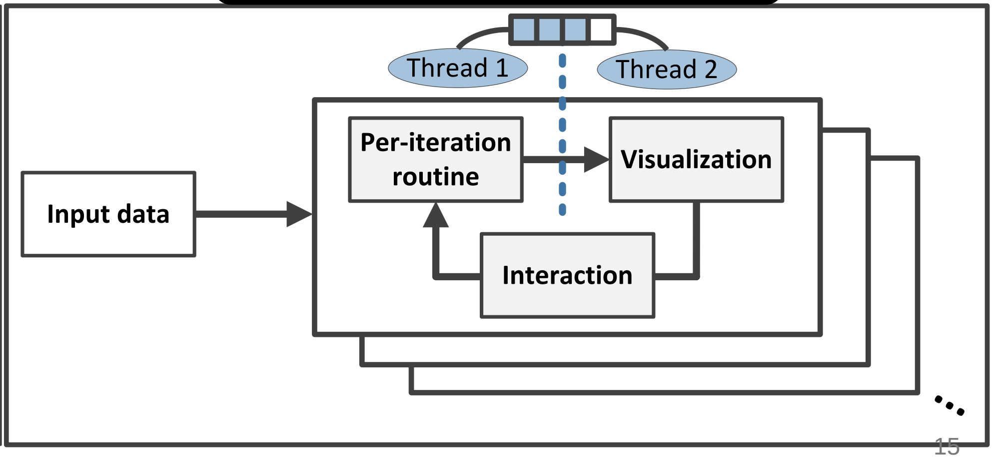
PIVE

- Integration methodology of iterative methods for **Real-Time** interactive visualization

Standard approach



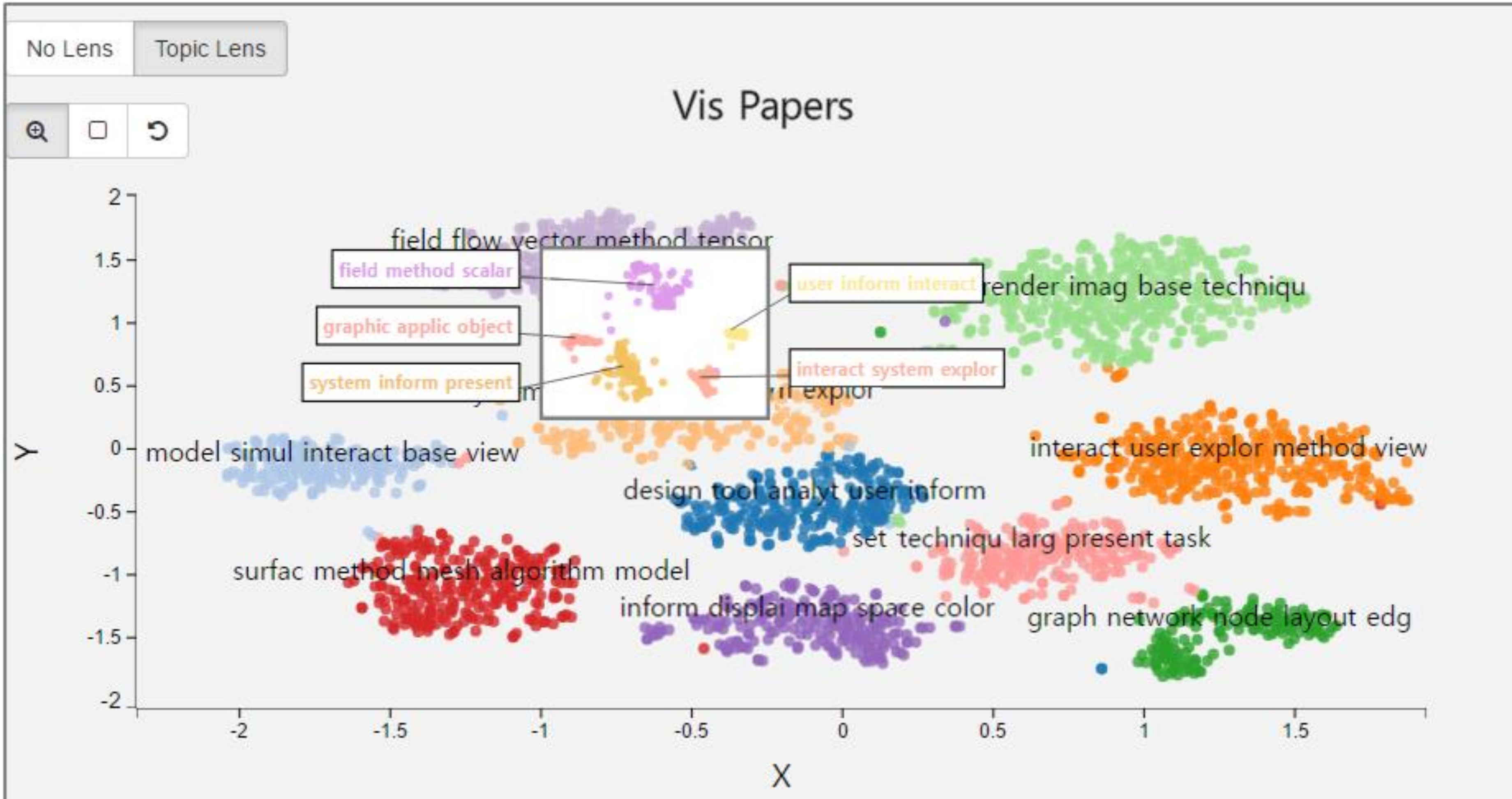
PIVE approach



PIVE Demo

TVCG'17

TopicLens: Efficient Multi-Level Visual Topic Exploration



TVCG'17

TopicLens: Efficient Multi-Level Visual Topic Exploration

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?

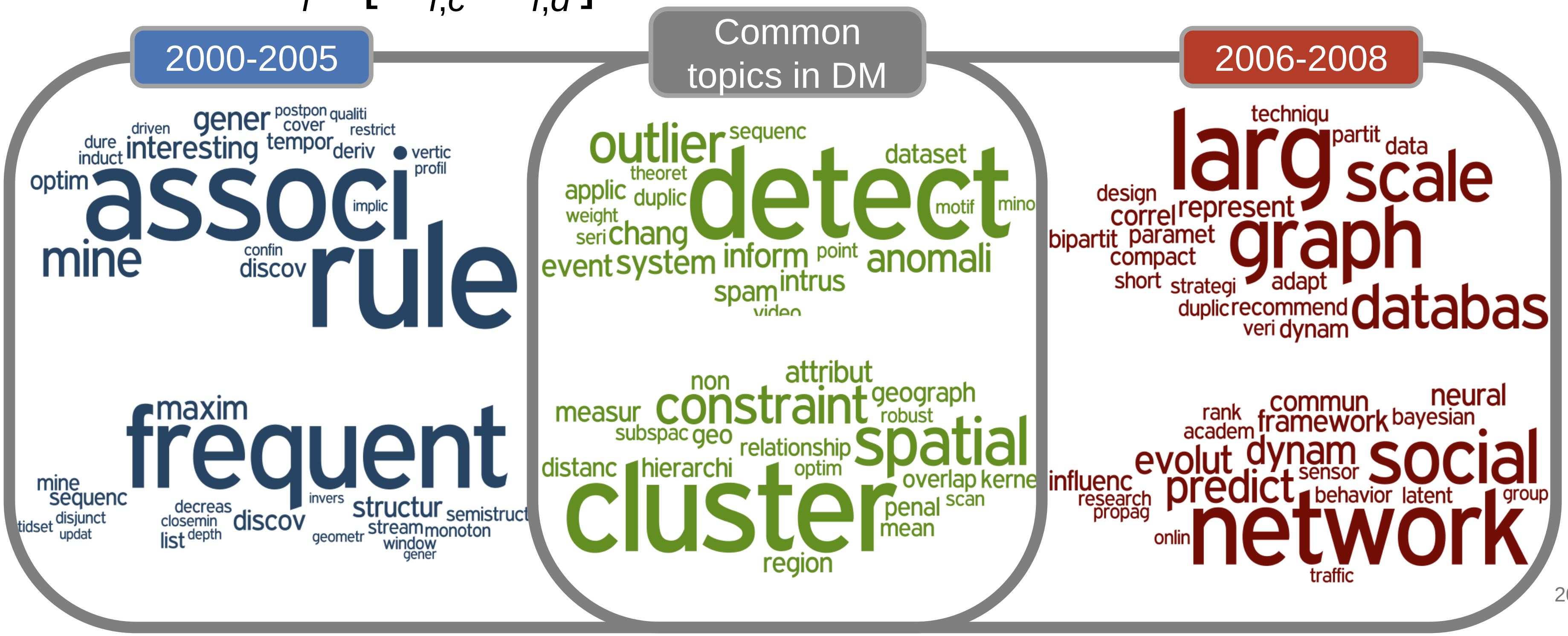
TopicLens Demo

Compare and Contrast: Joint Topic Discovery

Formulation

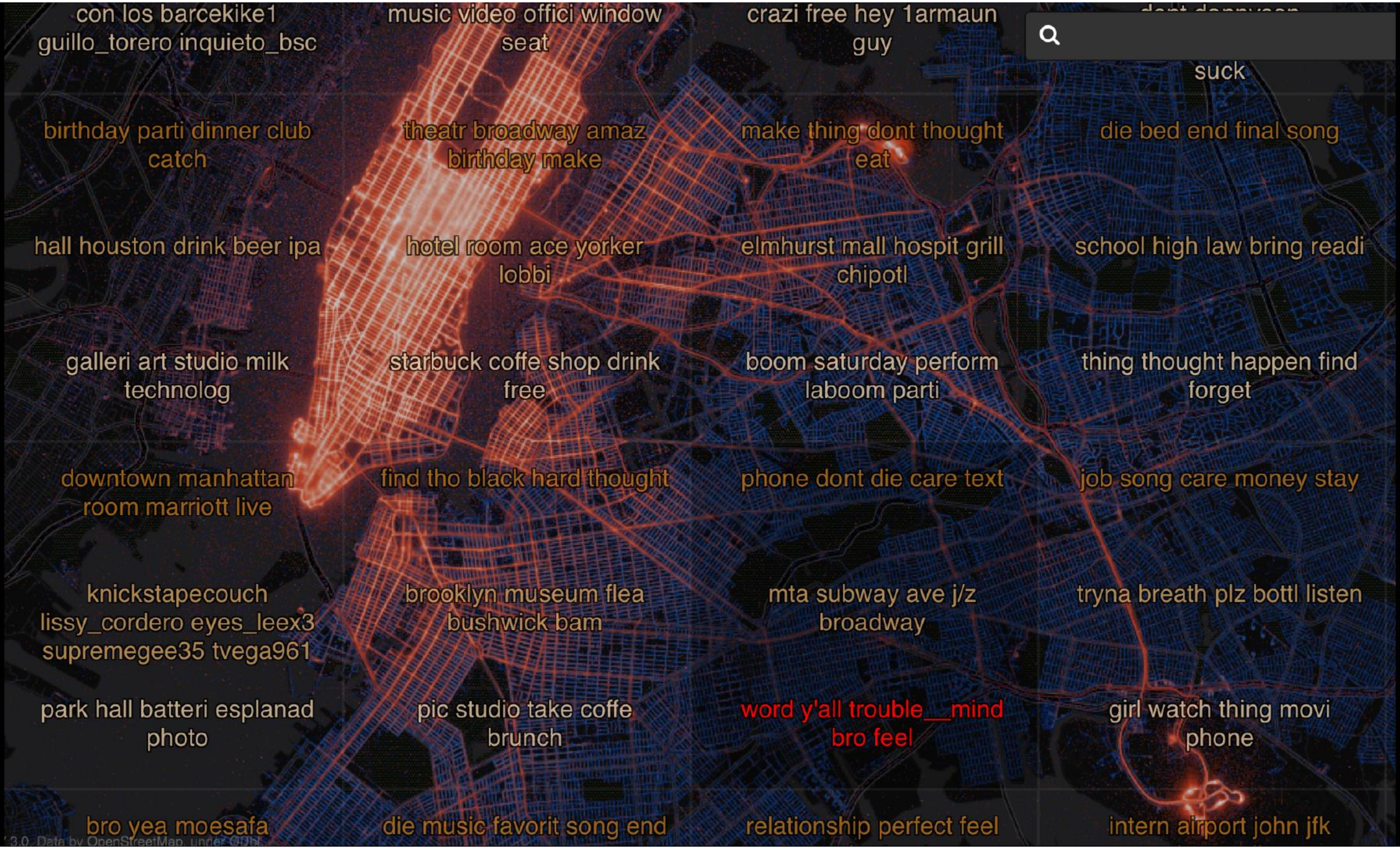
$$\min_{W>=0,\ H>=0} \quad 1/n_1 \| A_1 - W_1 H_1 \|_F^2 + 1/n_2 \| A_2 - W_2 H_2 \|_F^2 + \alpha \| W_{1,c} - W_{2,c} \|_F^2 + \beta \| W^T_{1,d} W_{2,d} \|_F^2$$

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



Geospatio-Temporal Topic Modeling

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



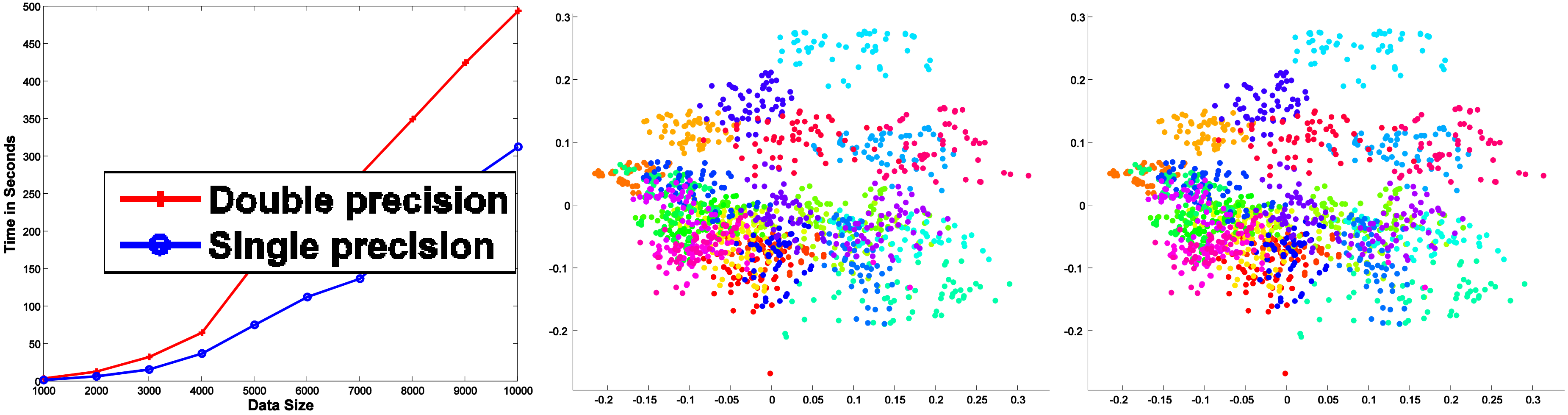
Perception- and Screen Space-Driven Integration Framework

Motivation

- Humans and computer screens do not require high precision.

Approach

- **Approximate computing**



Computing time vs. data size

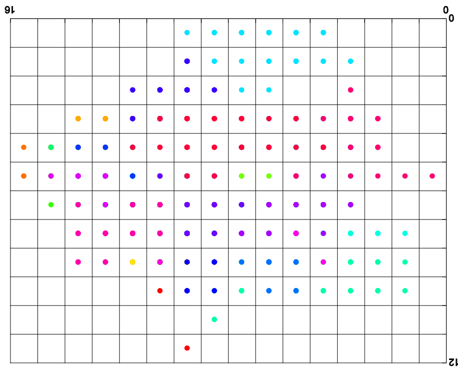
Double-precision PCA

Single-precision PCA

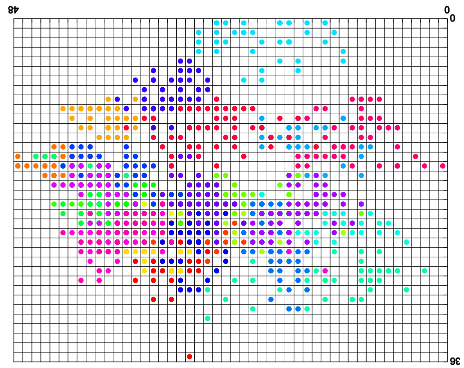
CG&A'13

New Computing Paradigms for Visual Analytics

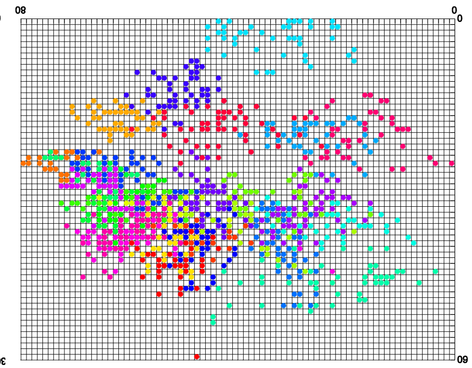
Adaptive hierarchical refinement



16x12



48x36



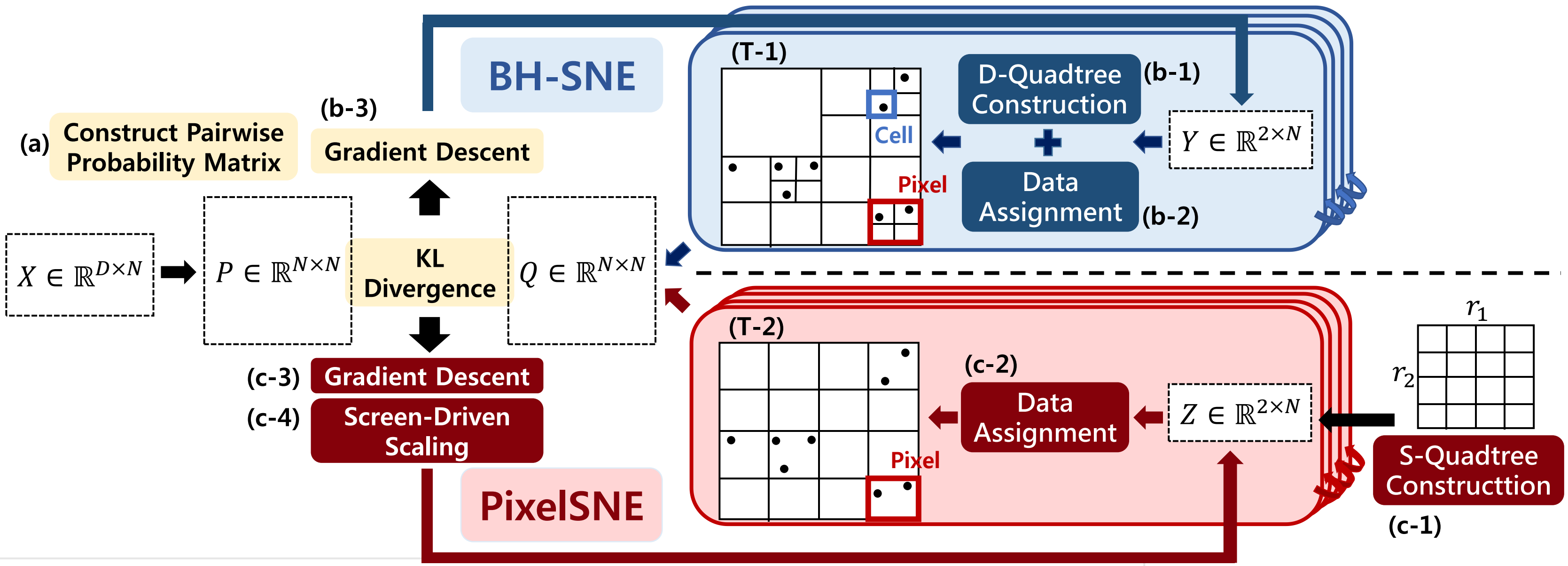
80x60

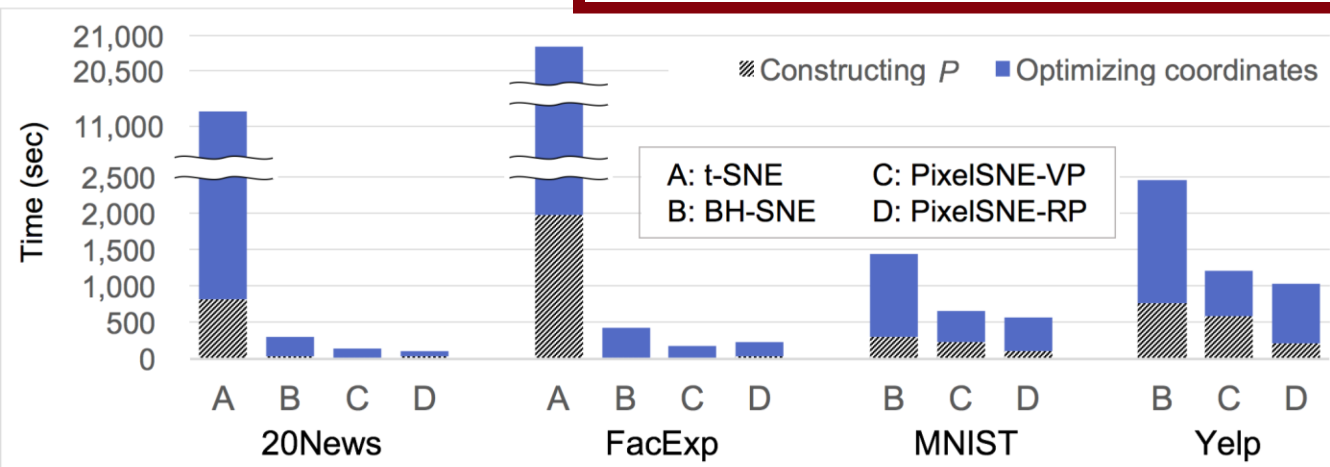
- Leveraging ideas from other literatures, e.g., wavelet

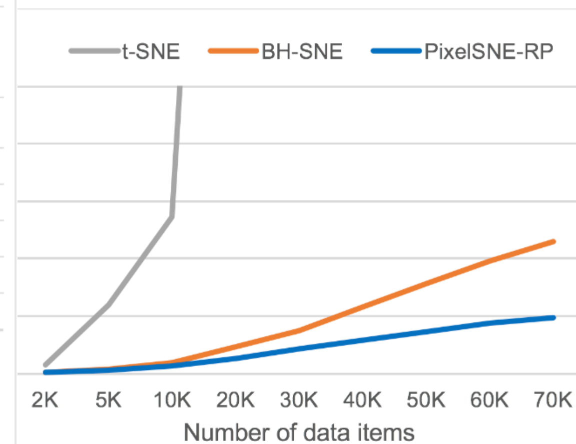
Images src: http://www.cse.lehigh.edu/~spletzer/rip_f06/lectures/lec013_Pyramids.pdf

ArXiv'16

PixelSNE: Pixel-Aligned t-SNE



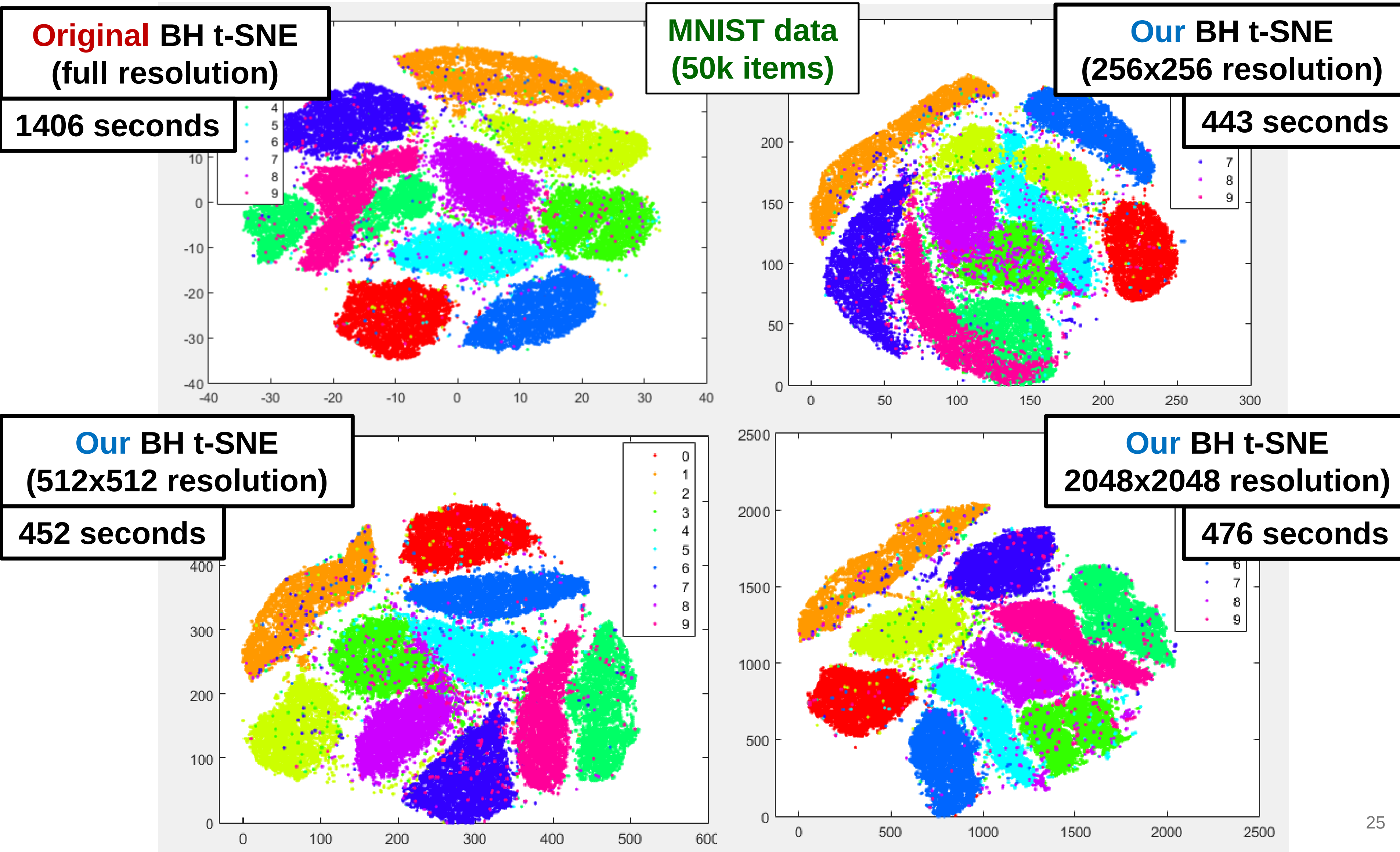




ArXiv'16

PixelSNE: Pixel-Aligned t-SNE

MNIST data (50k items)



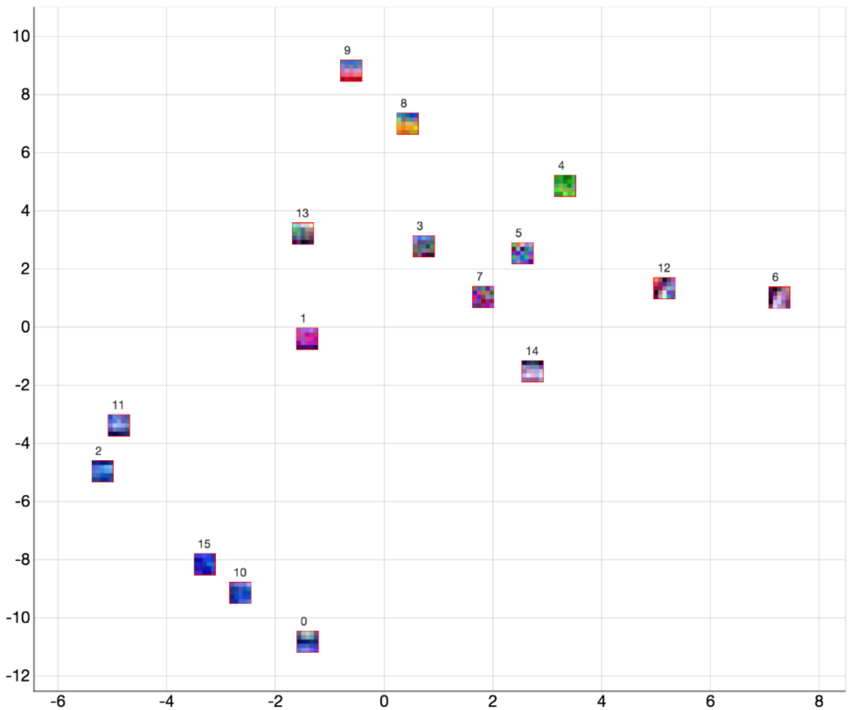
ReVACNN: Real-Time Visual Analytics for CNN

[KDD’16 IDEA Workshop, NIPS’16 FILM Workshop]

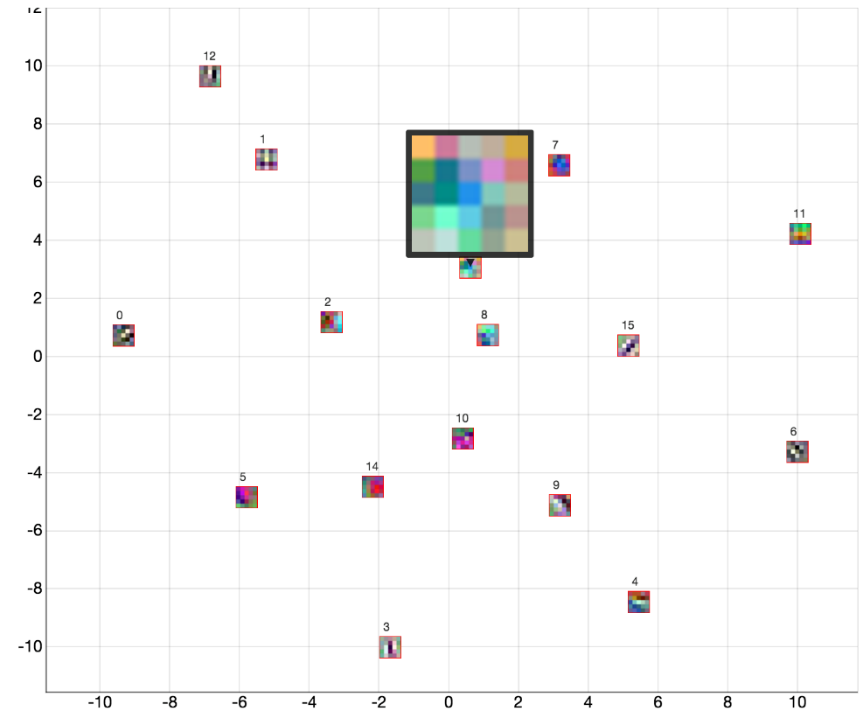
ReVACNN: Real-Time Visual Analytics for CNN

[KDD’16 IDEA Workshop, NIPS’16 FILM Workshop]

2D embedding of first-layer filters



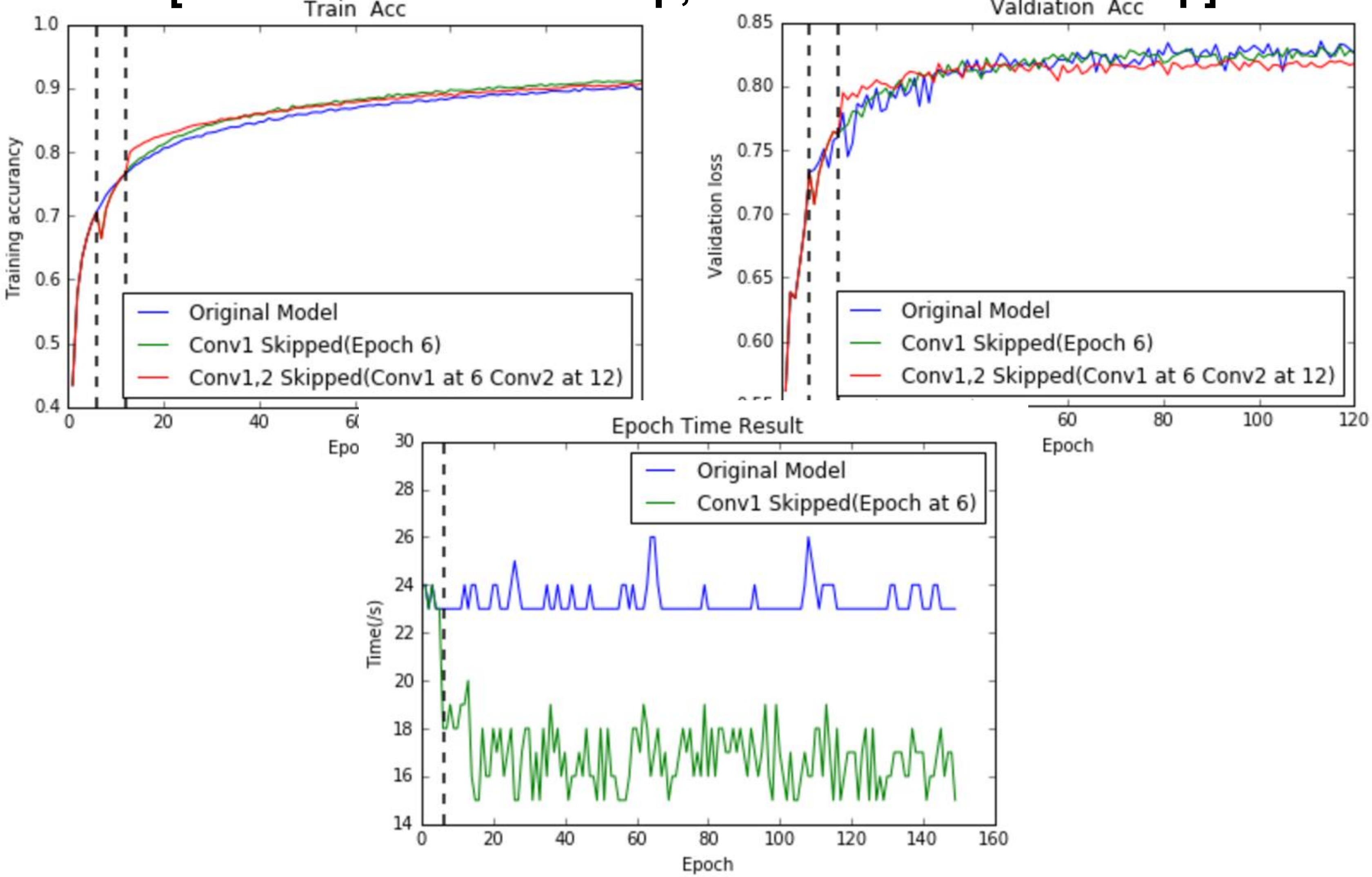
Improperly trained pattern, showing clear clusters



Properly trained pattern, showing no clusters

ReVACNN: Real-Time Visual Analytics for CNN

[KDD'16 IDEA Workshop, NIPS'16 FILM Workshop]



On-Going and Future Work

- Scalable visual analytics for deep networks
  - Tracking activations on residual deep network
- Fast, low-powered deep network on mobile devices
  - Personalized predictive keywords
- End-to-end learning integrated with handcrafted features
  - Automatic debugging on programs
- Semantic word embedding
  - Nonnegative matrix factorization + word embedding
- Direction-agnostic deep networks

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

- PIVE: Per-Iteration Visualization Environment for Real-time Interactive Visualizations, **AAAI**, 2017
- AxiSketcher: Interactive Nonlinear Axis Mapping through Users' Drawing on Visualization, **TVCG**, 2017
- TopicLens: Efficient Multi-Level Visual Topic Exploration of Large-Scale Document Collections, **TVCG**, 2017
- L-EnsNMF: Boosted Local Topic Discovery via Ensemble of Nonnegative Matrix Factorization, ICDM, 2016
- PixelSNE: Visualizing Fast with Just Enough Precision via Pixel-Aligned Stochastic Neighbor Embedding, arXiv, 2016
- 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
- p-ISOMAP: An Efficient Parametric Update for ISOMAP for Visual Analytics, **SDM,** 2010