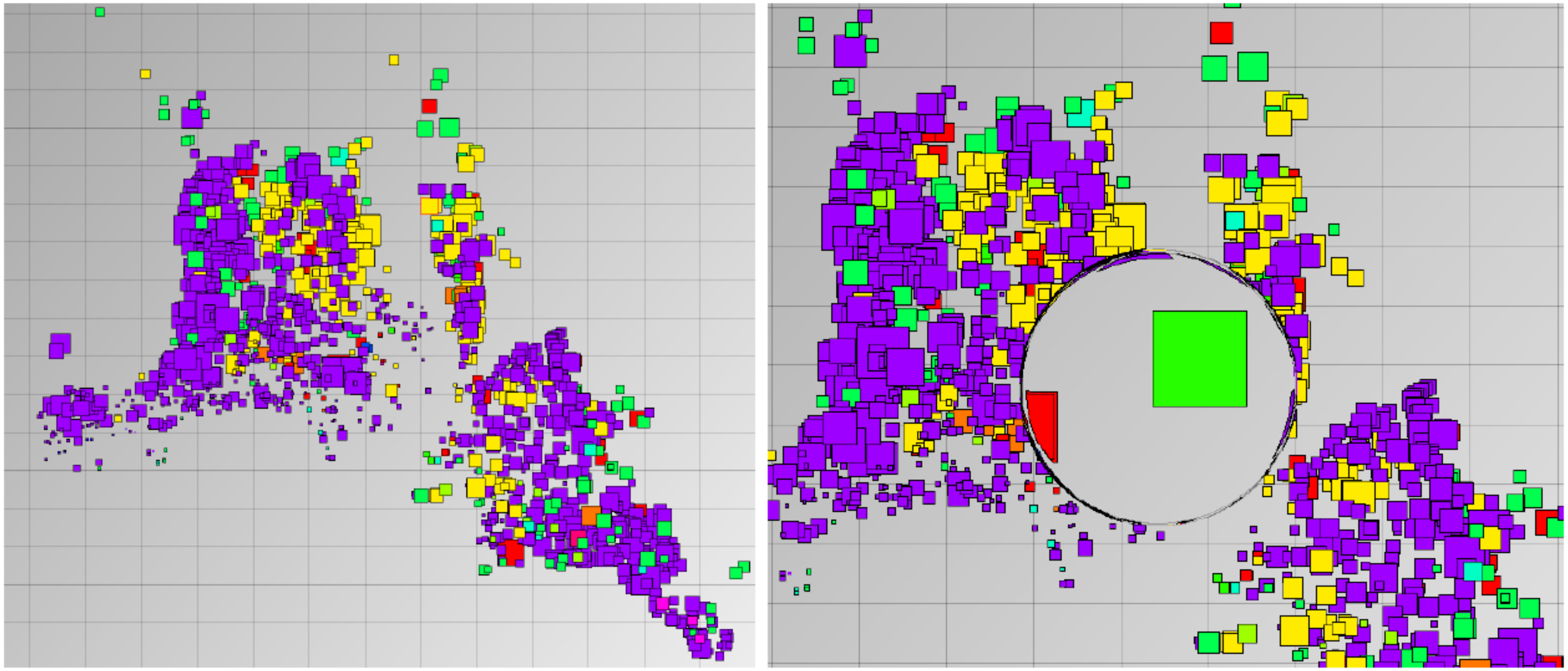
Multi-Source Visual Analytics

Jieping Ye
Computer Science and Engineering
The Biodesign Institute
Arizona State University

Co-PIs: Anshuman Razdan, Peter Wonka

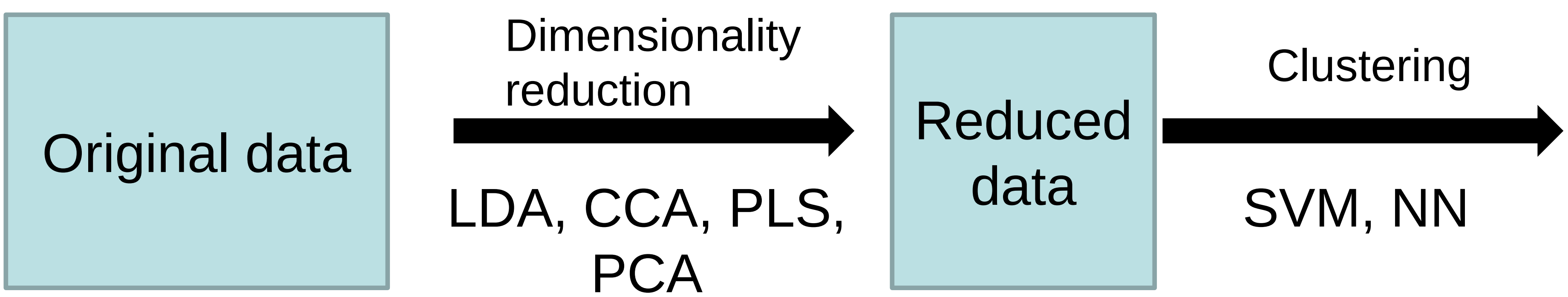
Dimensionality Reduction for Data Visualization



Left: Visualizing data points as rectangles. Right: A magnifying lens

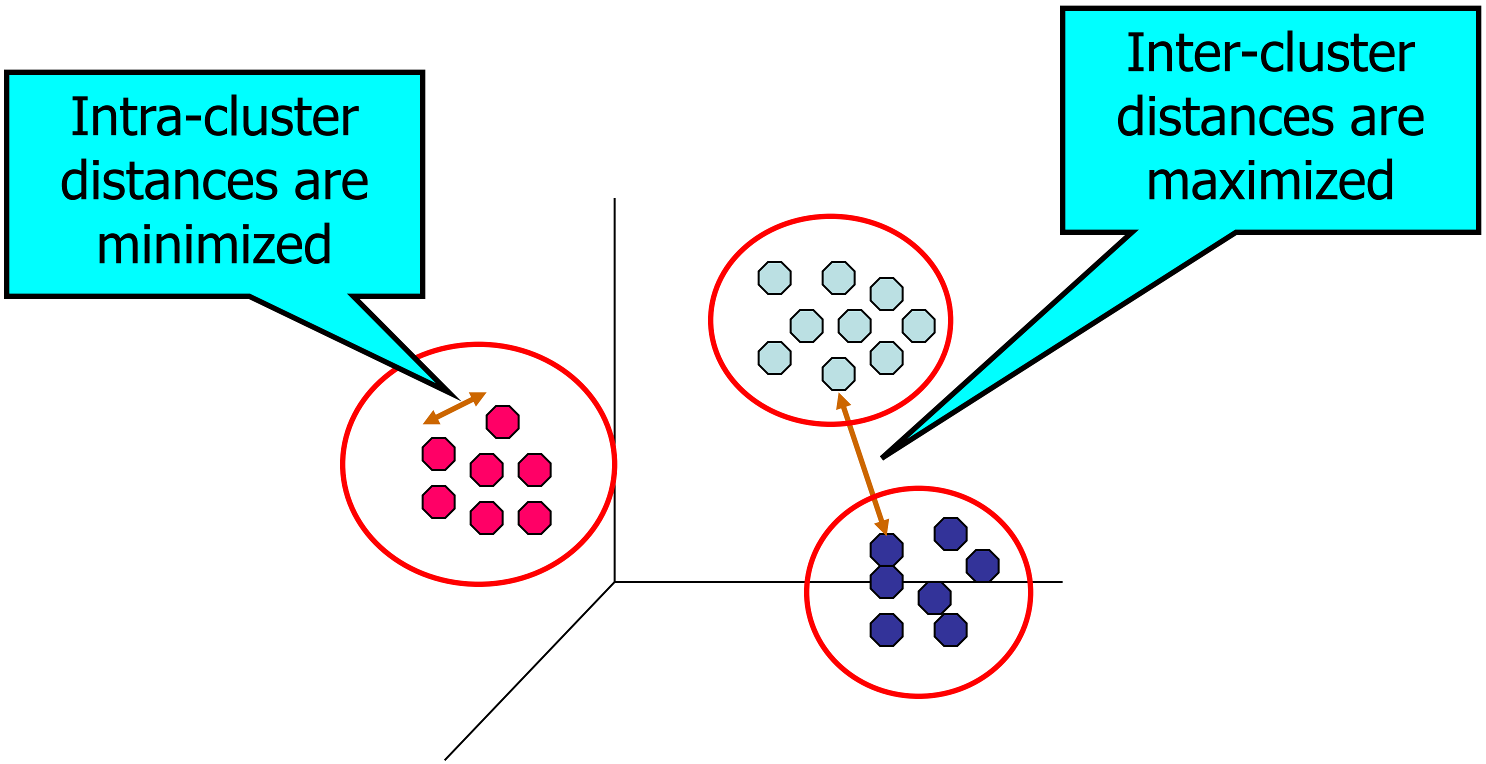
Dimensionality Reduction Algorithms

- Supervised:
  - Linear discriminant analysis (LDA)
  - Canonical correlation analysis (CCA)
  - Partial least squares (PLS)
- Unsupervised:
  - Principal component analysis (PCA)
  - Manifold learning (Isomap, LLE, Laplacian Eigenmap)

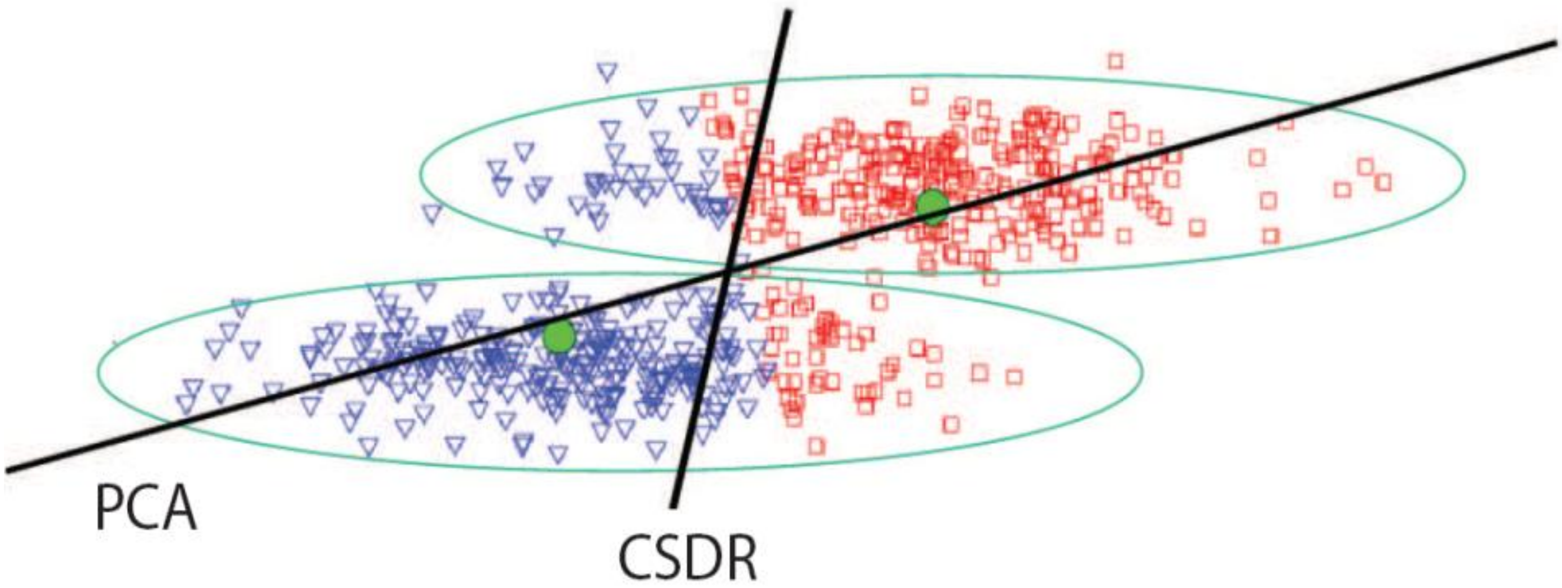


Clustering and Dimensionality Reduction (1)

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



Clustering and Dimensionality Reduction (2)



Standard PCA fails to detect these two natural clusters, whereas the proposed cluster sensitive dimensionality reduction (CSDR) does a much better job of separating the data.

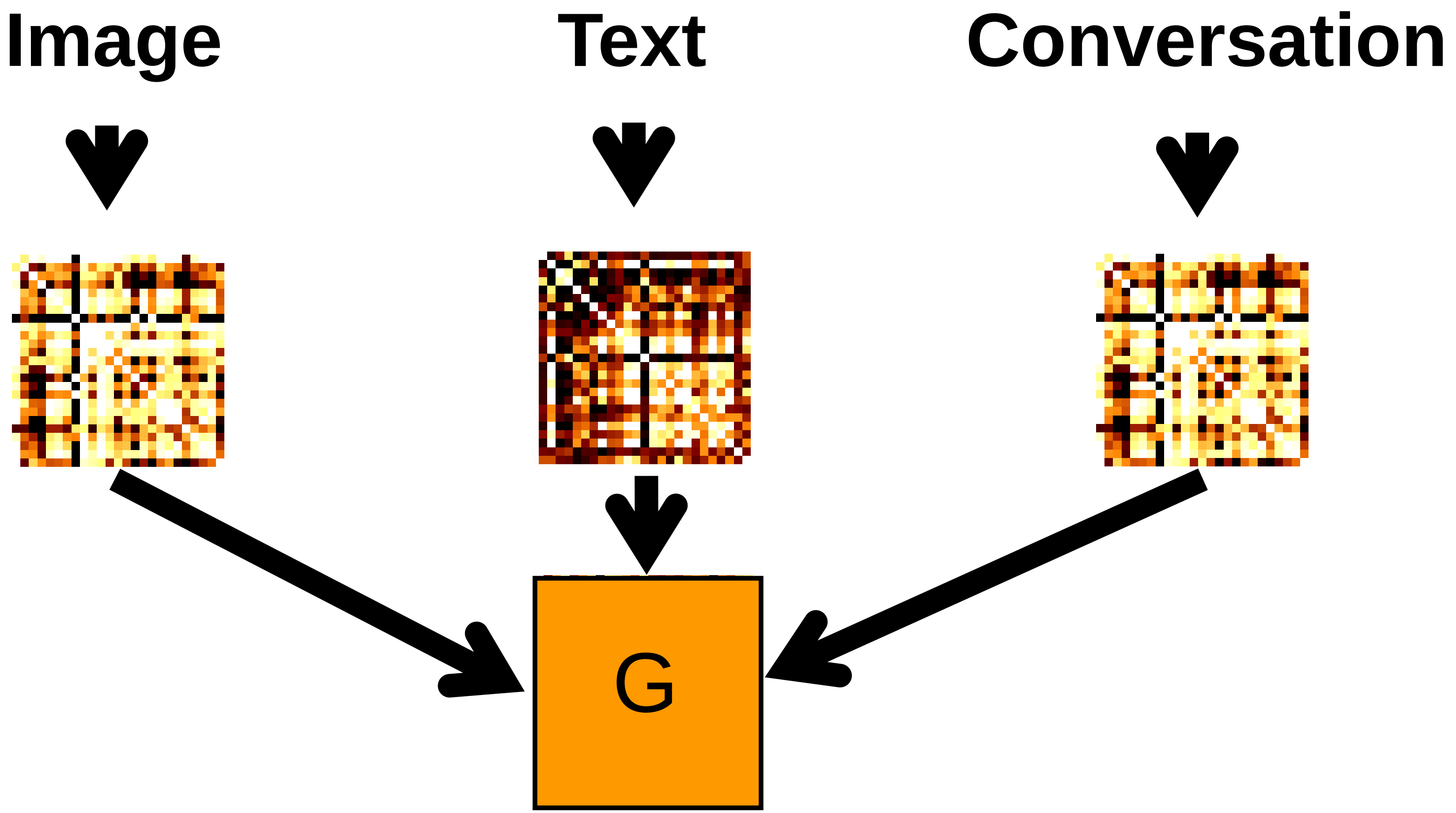
How can we combine clustering and dimensionality reduction to improve visual analytics tasks?

Multi-source Data Transformations

- Processing heterogenous data is a significant challenge in visual analytics.
  - For example, an analyst may want to analyze data from multiple sources like images, text (emails), and telephone conversations.

- We propose to investigate techniques to transform entities that come from different sources.

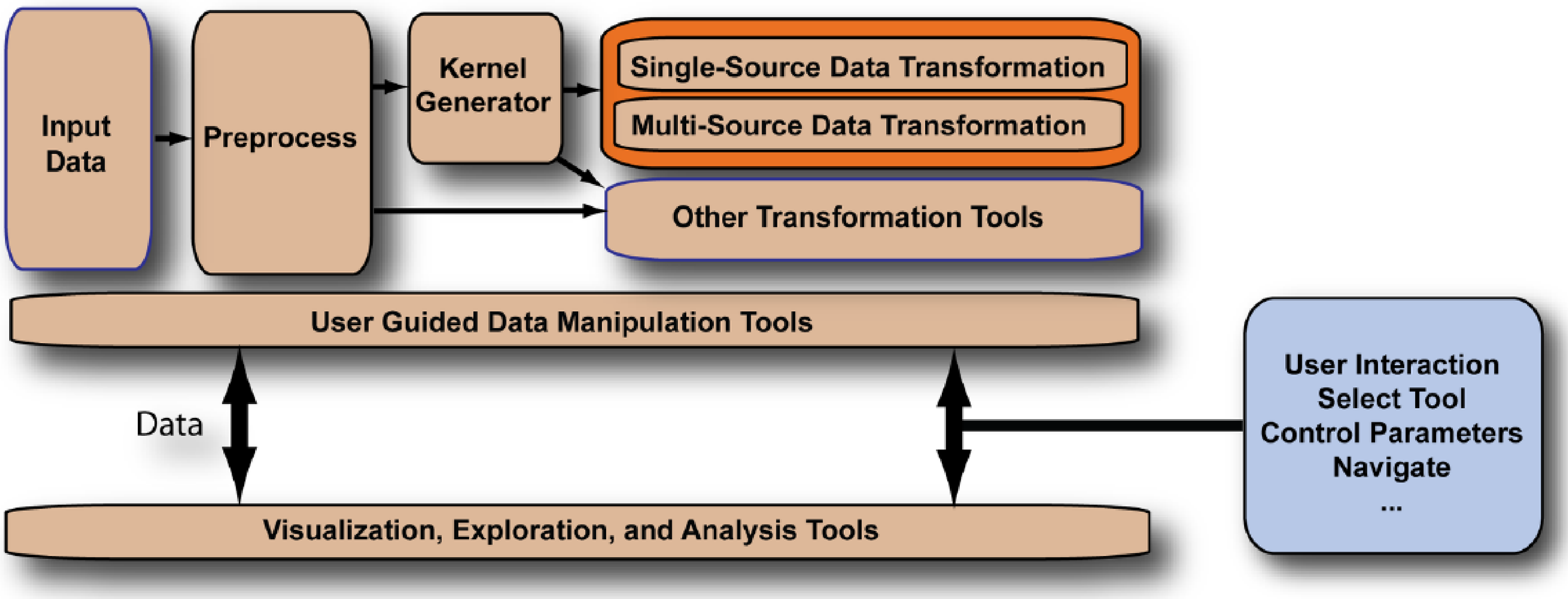
Multiple Kernel Learning for Data Fusion



Research Aims

- Clustering and dimensionality reduction
  - Single source data transformation
- Clustering and dimensionality reduction
  - Multi-source data transformation
- MSVA a novel Visual Analytics Framework

The Proposed MSVA framework



A user can draw from a number of data transformation and visual analysis tools.
A typical sequence of data processing is shown by the arrows.
The user can interactively provide feedback and update the transformation.

Preliminary Work: Problem Setup

Given $\{x_1, x_2, \cdots, x_n\} \in \Re^m$

Let $X = [x_1, x_2, \cdots, x_n]$ be the data matrix

Linear projection $W \in \Re^{m \times l} : x_i \in \Re^m \Rightarrow \hat{x}_i = W^T x_i \in \Re^l$

Clustering $C_1, C_2, \cdots, C_k$

- It has been shown that for most high-dimensional data sets, almost all low dimensional projections are nearly normal.
  - Diaconis and Freedman. *Annals of Statistics*, 1984.
  - Hall and Li. *Annals of Statistics*, 1993.

Mahalanobis Distance

$$\{\hat{x}_1, \hat{x}_2, \cdots, \hat{x}_n\} \in \Re^l$$ nearly normal for large m

Mahalanobis distance $d_M(\hat{x}_i, \hat{x}_j) = \sqrt{(\hat{x}_i - \hat{x}_j)^T \hat{S}^{-1} (\hat{x}_i - \hat{x}_j)}$

where $$\hat{S} = \frac{1}{n}\sum_{i=1}^{n}(\hat{x}_i - \hat{\mu})(\hat{x}_i - \hat{\mu})^T = W^T S W$$

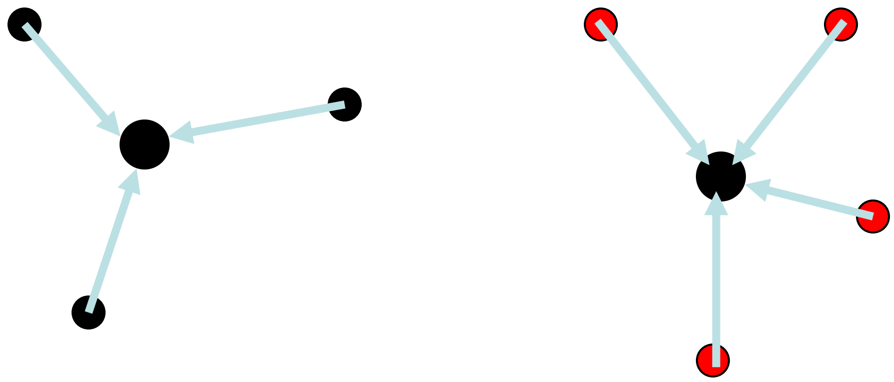
$$S = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T$$

regularization $$S = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T + \lambda I_m$$

Sum of Squared Error

Under this new distance measure, K-means clustering assigns the data into k disjoint clusters, which minimizes the Sum of Squared Error (SSE):

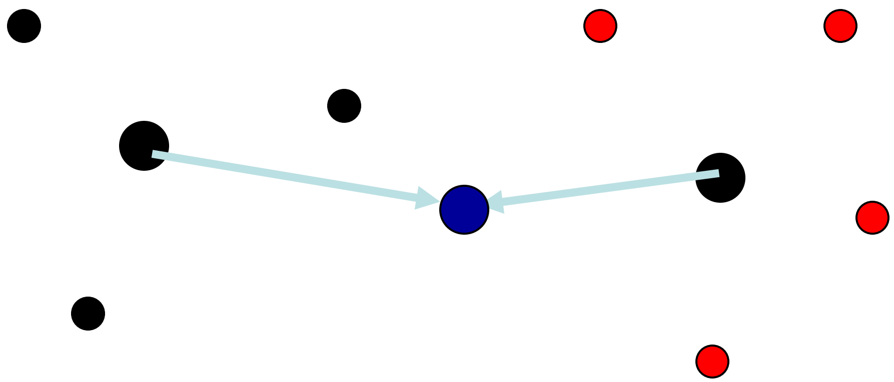
$$\mathrm{SSE}\left(\{C_j\}_{j=1}^{k}\right)=\sum_{j=1}^{k}\sum_{\hat{x}_i\in C_j} d_M\left(\hat{x}_i,\mu_j\right)^2.$$



Sum of Squared Inter-Cluster Error

As the summation of all pair-wise distances is a constant for a fixed W, the minimization of SSE is equivalent to the maximization of Sum of Squared Inter-Cluster Error (SSIE):

$$\mathrm{SSIE}\left(\{C_j\}_{j=1}^{k}\right) = \sum_{j=1}^{k} n_j d_M\left(\hat{\mu}_j, \hat{\mu}\right)^2.$$



Compact Matrix Formulation

Sum of Squared Intra-Cluster Error (SSIE) can be expressed in a compact matrix form as follows:

$$\mathrm{SSIE}\left(\{C_j\}_{j=1}^{k}\right) = \mathrm{trace}\left(\mathrm{L}^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}\mathrm{W}^{\mathrm{T}}\left(\mathrm{W}^{\mathrm{T}} SW\right)^{-1}\mathrm{WXL}\right)$$

L is the weighted cluster indicator matrix, whose *i*-th row is

$$L_i = \frac{1}{\sqrt{n_i}}\left(0,\cdots,0,\overbrace{1,\cdots,1}^{n_i},0,\cdots,0\right)^T$$

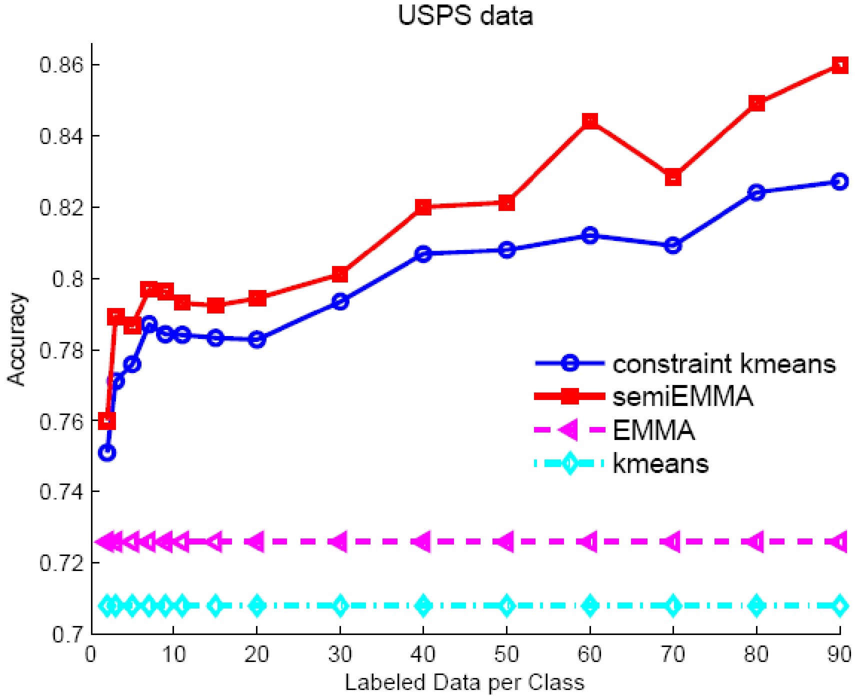
Joint dimensionality reduction and clustering formulation:

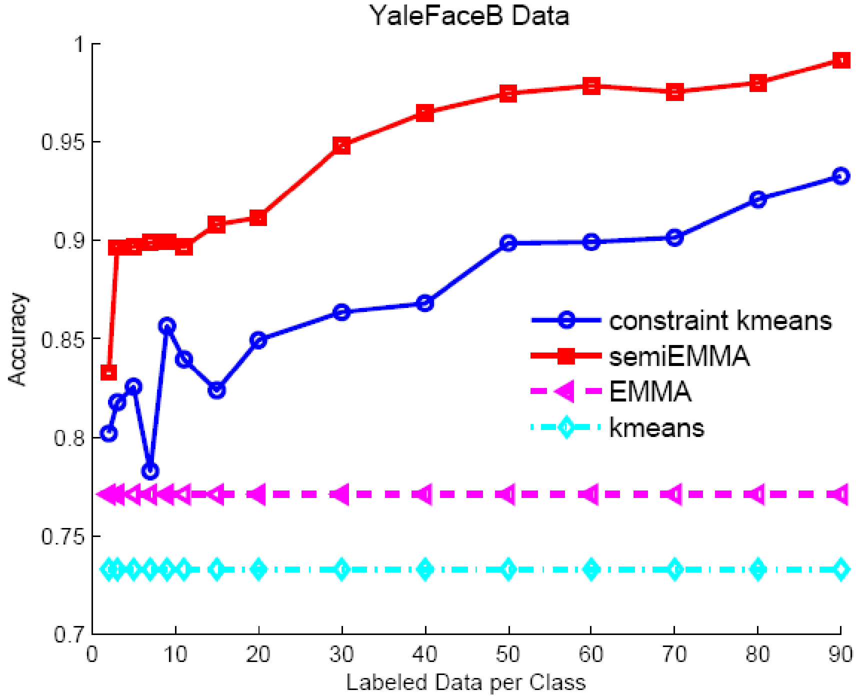
$$\max_{\mathrm{W,L}}\ \mathrm{trace}\left(\mathrm{L}^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}\mathrm{W}^{\mathrm{T}}\left(\mathrm{W}^{\mathrm{T}} SW\right)^{-1}\mathrm{WXL}\right)$$

Preliminary Study

	Proposed		PCA	LLE
GCM	**0.583**	0.568	0.569	0.571
	-	0	0	0
Soybean	**0.725**	0.671	0.668	0.705
	-	0	0	0.002
Segment	**0.644**	0.552	0.551	0.533
	-	0	0	0
Letter (a-d)	**0.662**	0.606	0.606	0.647
	-	0	0	0.003
USPS	**0.726**	0.708	0.709	0.655
	-	0	0.001	0
YaleFaceB	**0.771**	0.733	0.733	0.746
	-	0	0	0.002
Average	0.685	0.64	0.639	0.643

Semi-supervised Setting





Domain knowledge

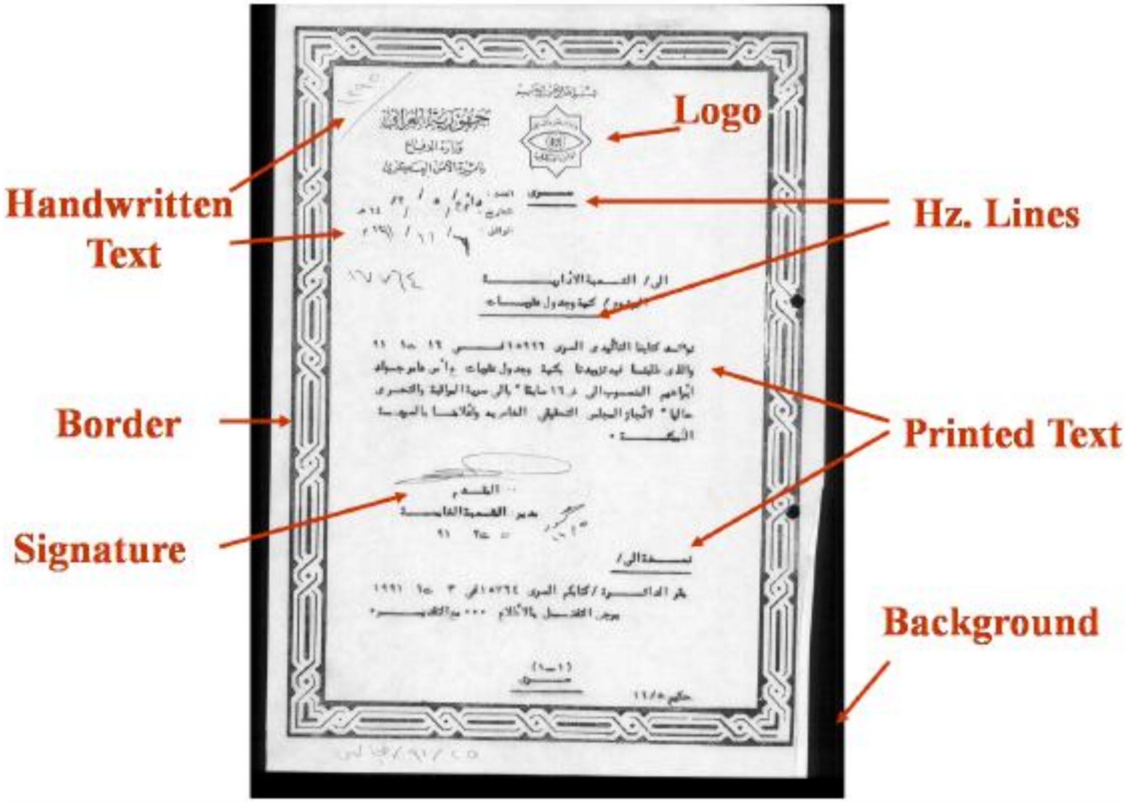
Use feedback

Proposed research

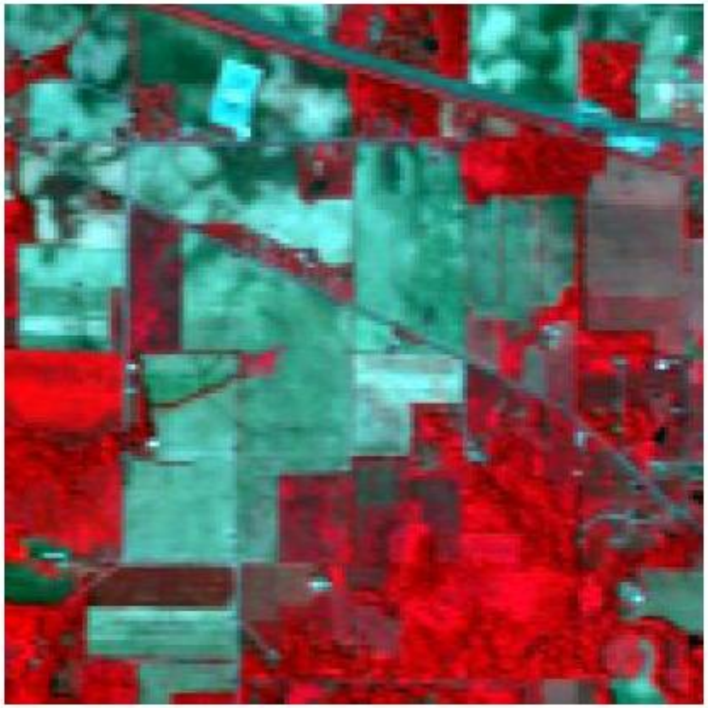
- Single source data transformation
- Multi-source data transformation
- Sparse data transformation
- Applications
  - Visual document analysis
  - Geo-spatial analysis
  - Health information analysis

Application I: Visual Document Analysis

- The capability to quickly process, tag/annotate, triage and classify volumes of information is key to enabling effective and useful information analysis.



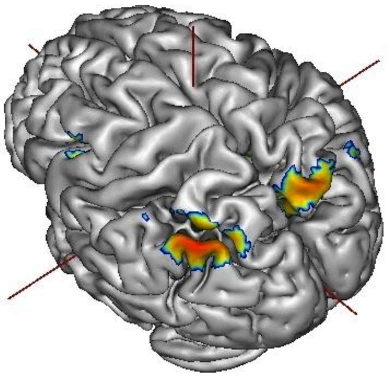
Application II: Geo-spatial Analysis

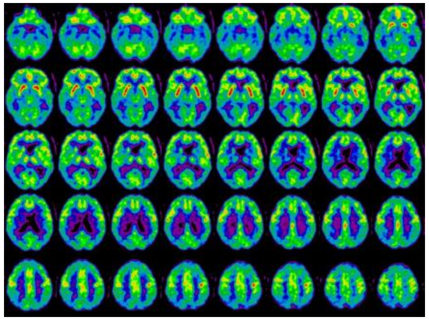


The hyperspectral image is shown on the left, and a thematic map of land cover classes is on the right.

Application III: Health Information Analysis

ADNI ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE





Demographic, genetic, cognitive measures

Questions!