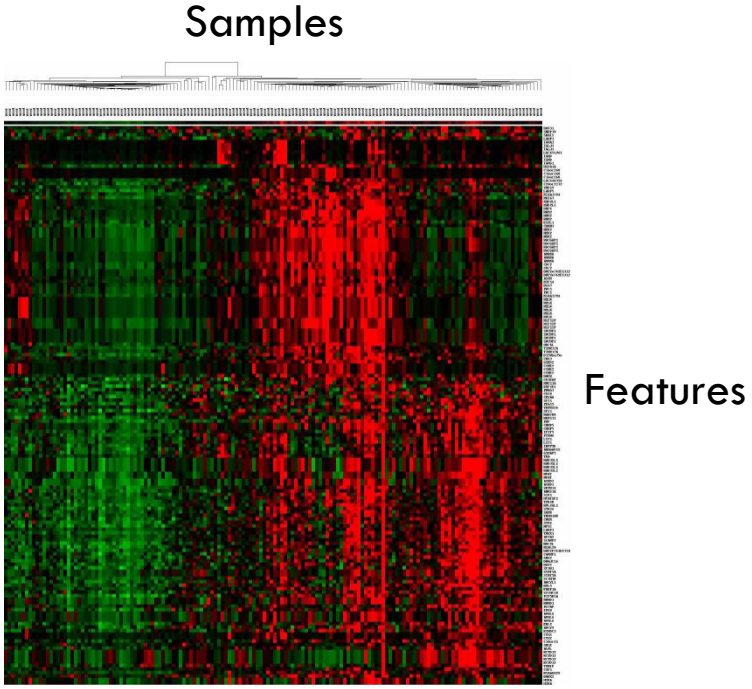
Semi-Supervised Graph Embedding Scheme with Active Learning (SSGEAL): Classifying High Dimensional Biomedical Data

George Lee and Anant Madabhushi

Rutgers, The State University of New Jersey
Dept. of Biomedical Engineering
Laboratory of Computational Imaging and Bioinformatics (LCIB)

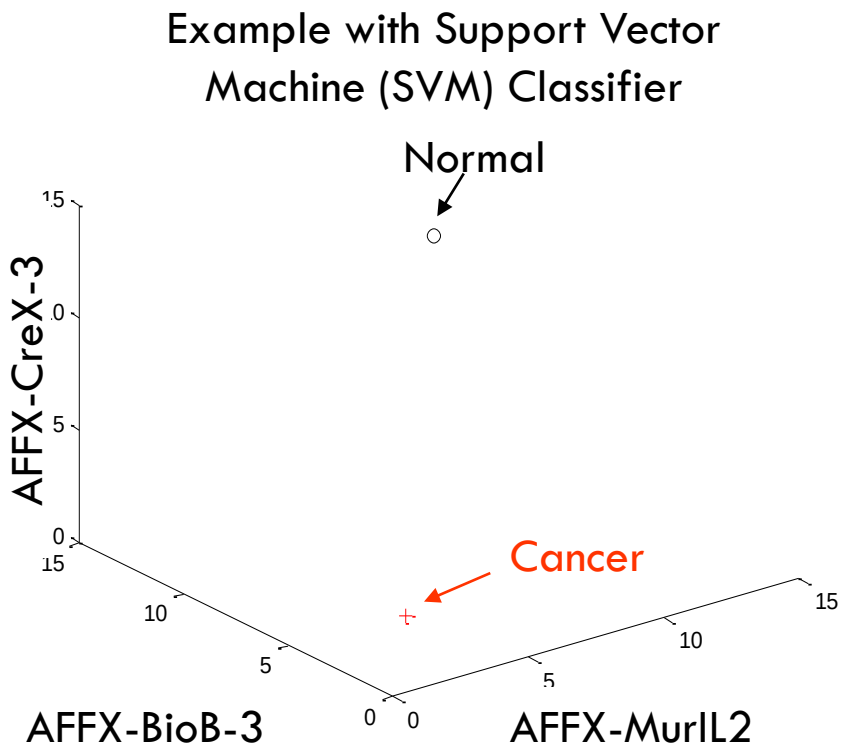
Introduction

- Gene Expression data is difficult to analyze due to high dimensionality



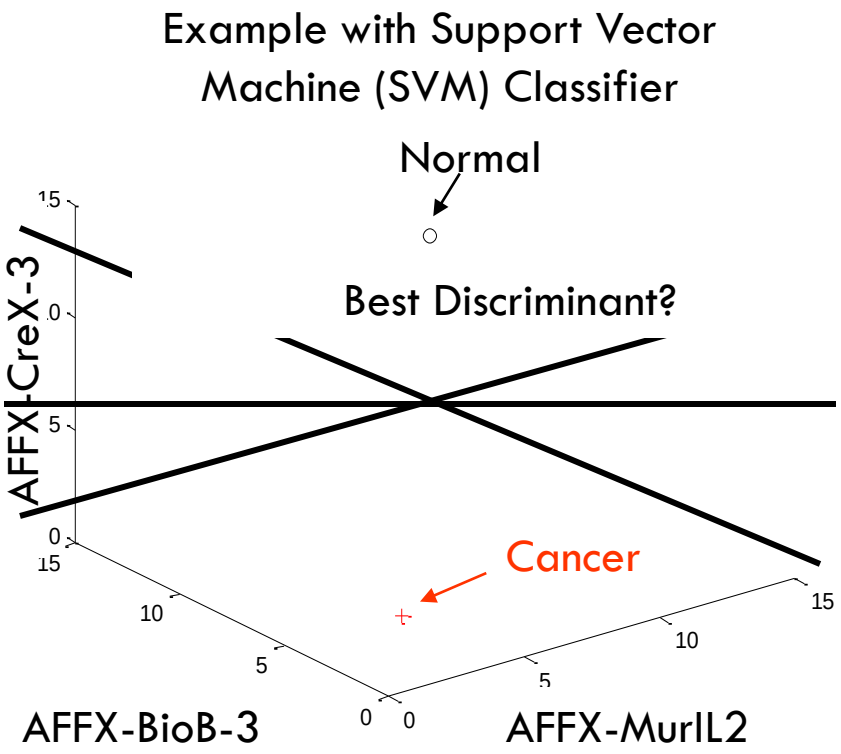
Introduction

- Gene Expression data is difficult to analyze due to high dimensionality
- "Curse of Dimensionality"
  - Too many features (dimensions compared to sample size



Introduction

- Gene Expression data is difficult to analyze due to high dimensionality
- "Curse of Dimensionality"
  - Too many features (dimensions compared to sample size



Introduction

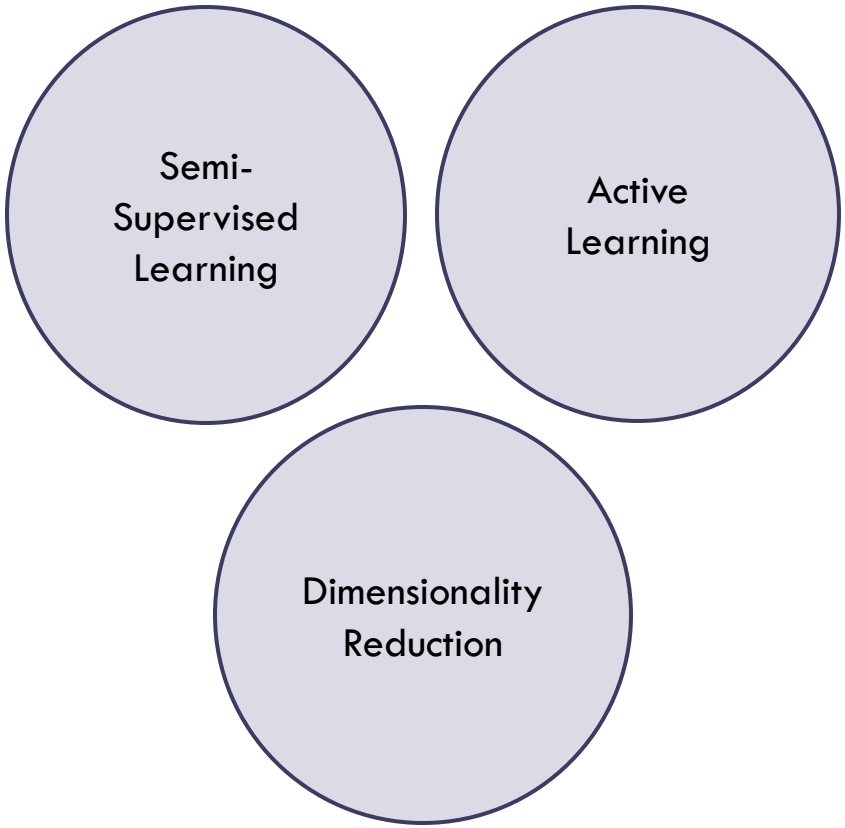
- Gene Expression data is difficult to analyze due to high dimensionality

- "Curse of Dimensionality"
  - Too many features (dimensions compared to sample size

- Goal: to provide a low dimensional feature space such that classification performance improves

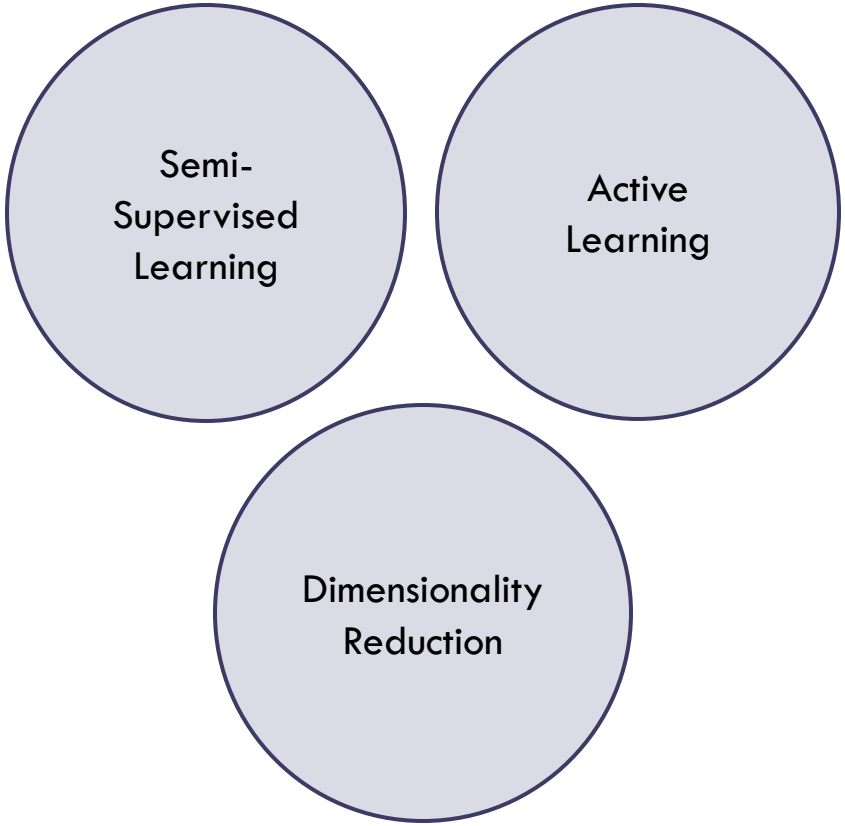
Novel Contribution

- Several strategies exist for improving classification



Novel Contribution

- Our Method (SSGEAL) leverages these strategies to obtain the best possible classification



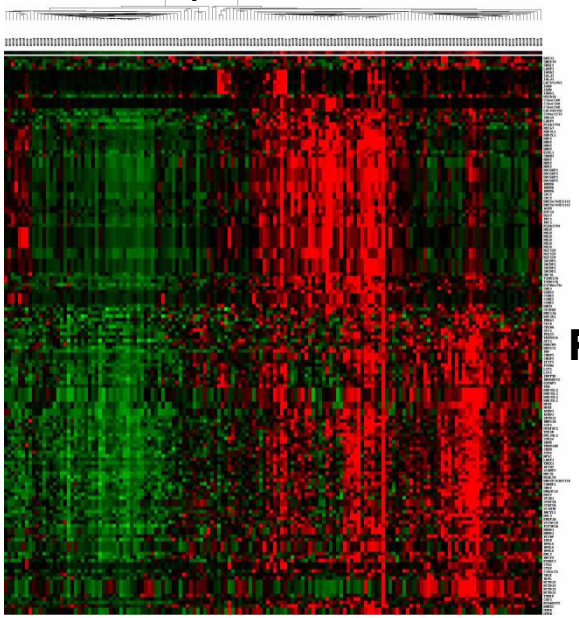
Overview

- Introduction
- Background
  - Dimensionality Reduction (DR)
  - Semi-Supervised Dimensionality Reduction (SSDR)
  - Active Learning (AL)
- Overview of SSGEAL
- Experimental Design
- Results
- Concluding Remarks

Dimensionality Reduction (DR)

- DR assumes a low-dimensional embedding in the high-D space

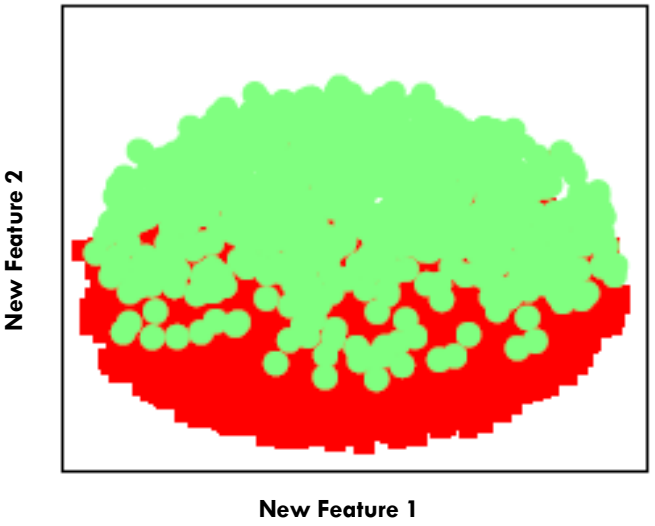
High Dimensional Representation



Patients/ Samples

Genes/ Features

Low Dimensional Embedding

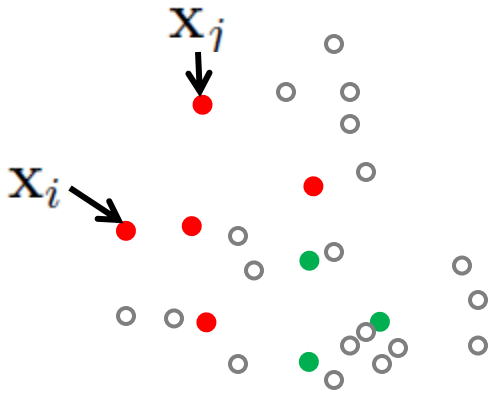


Graph Embedding
Shi et al. (2000)

Semi-Supervised Dimensionality Reduction (SSDR)

- DR typically unsupervised
- Class labels can be used to improve embeddings

Unsupervised DR Embedding



$$W(\mathbf{x}_i, \mathbf{x}_j) = e^{\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|_2}{\sigma}}$$

$$\tilde{W}(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{ii}^{N} W(\mathbf{x}_{ii}, \mathbf{x}_j) \times \sum_{jj}^{N} W(\mathbf{x}_i, \mathbf{x}_{jj})\right)^{-1} W(\mathbf{x}_i, \mathbf{x}_j)$$

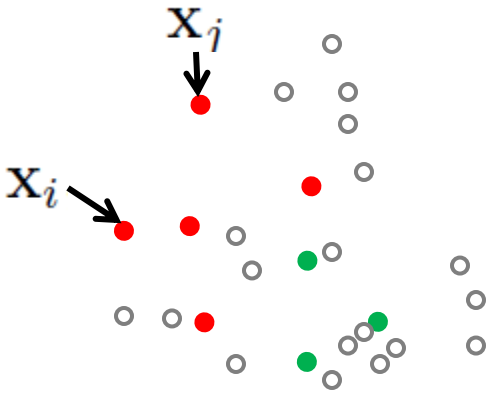
$$(D - \tilde{W})\mathbf{z} = \lambda D\mathbf{z}$$

Solve for embedding vector $\mathbf{z}$

Semi-Supervised Dimensionality Reduction (SSDR)

- DR typically unsupervised
- Class labels can be used to improve embeddings

Unsupervised DR Embedding



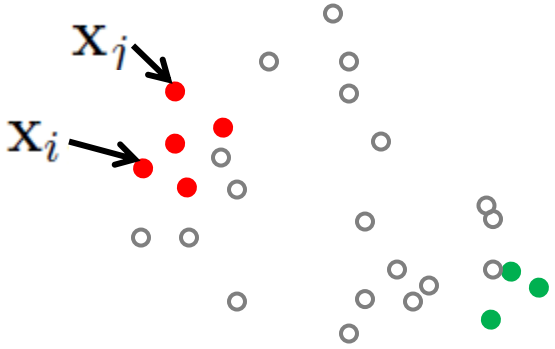
$$W(\mathbf{x}_i, \mathbf{x}_j) = e^{\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|_2}{\sigma}} (1 + e^{\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|_2}{\sigma}})$$

If same class, weight to be more similar

Semi-Supervised Dimensionality Reduction (SSDR)

- DR typically unsupervised
- Class labels can be used to improve embeddings

New SSDR Embedding



Known within-class samples are now mapped closer together

$$W(\mathbf{x}_i, \mathbf{x}_j) = e^{\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|_2}{\sigma}} (1 + e^{\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|_2}{\sigma}})$$

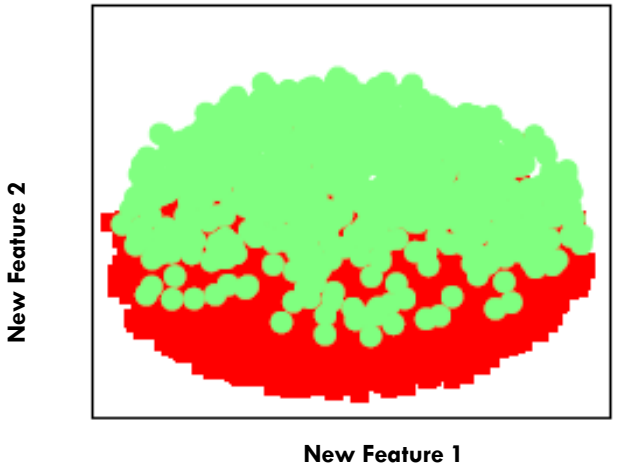
Similarity Matrix

$$\tilde{W}(\mathbf{x}_i, \mathbf{x}_j) = (\sum_{ii}^{N} W(\mathbf{x}_{ii}, \mathbf{x}_j) \times \sum_{jj}^{N} W(\mathbf{x}_i, \mathbf{x}_{jj}))^{-1} W(\mathbf{x}_i, \mathbf{x}_j)$$

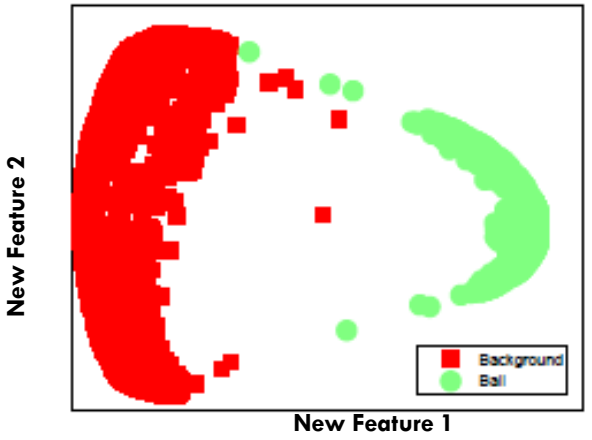
Normalize weight matrix

Semi-Supervised Dimensionality Reduction (SSDR)

Unsupervised DR
(Graph Embedding
(GE) Shi et al. 2000)

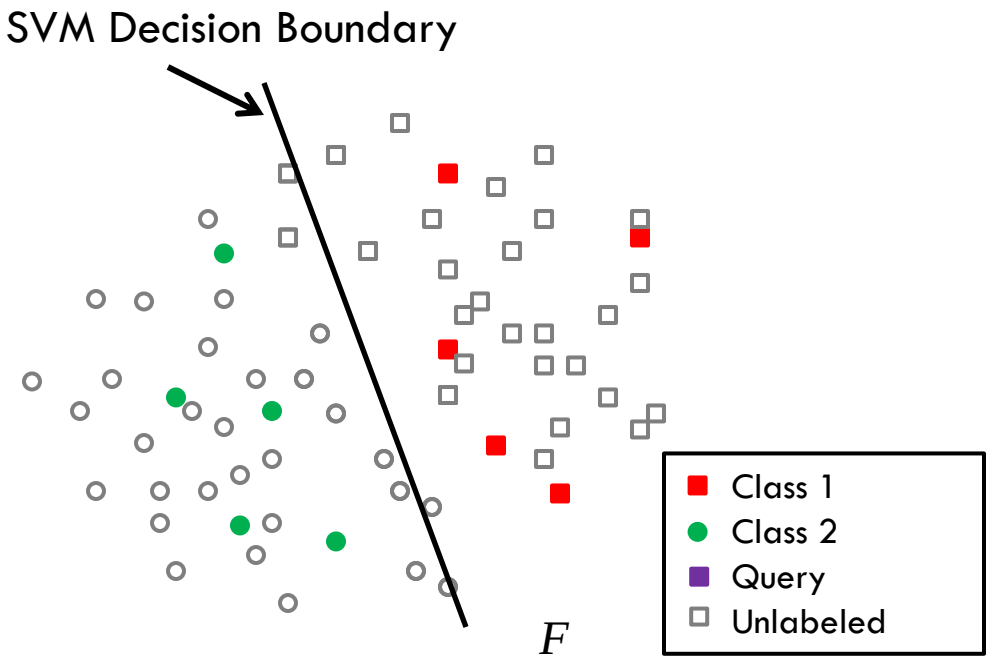


Semi-Supervised DR
(SSAGE
Zhou et al. 2006)



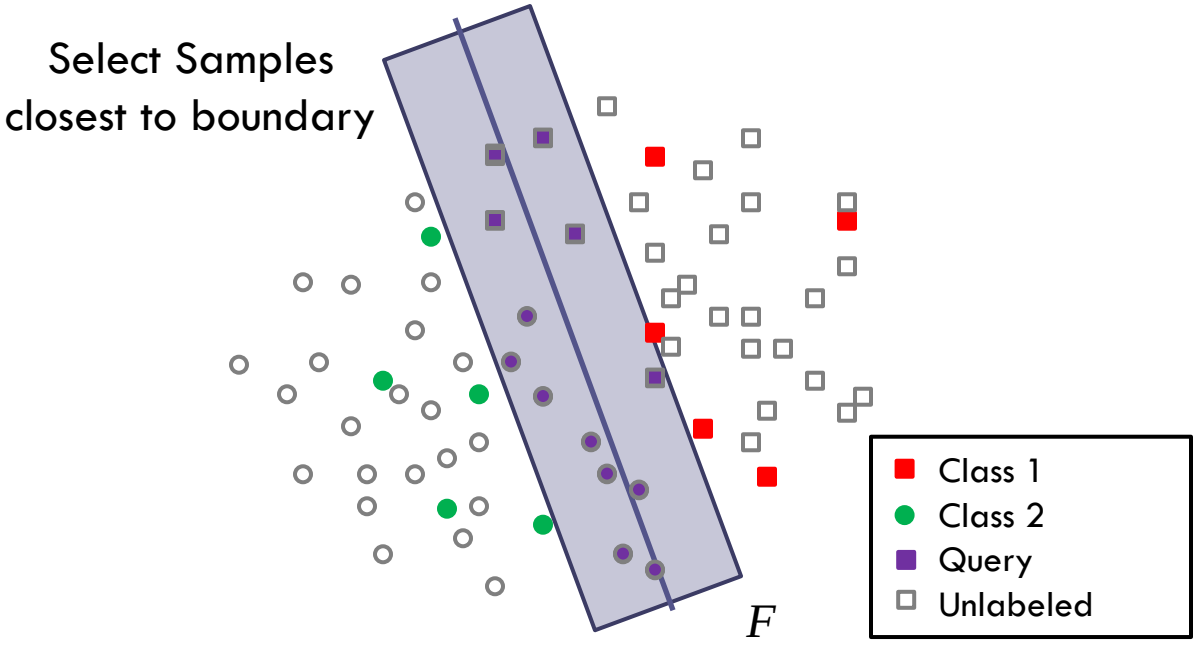
Active Learning (AL)

- Active Learning selects class labels with greatest contribution for improving classification.



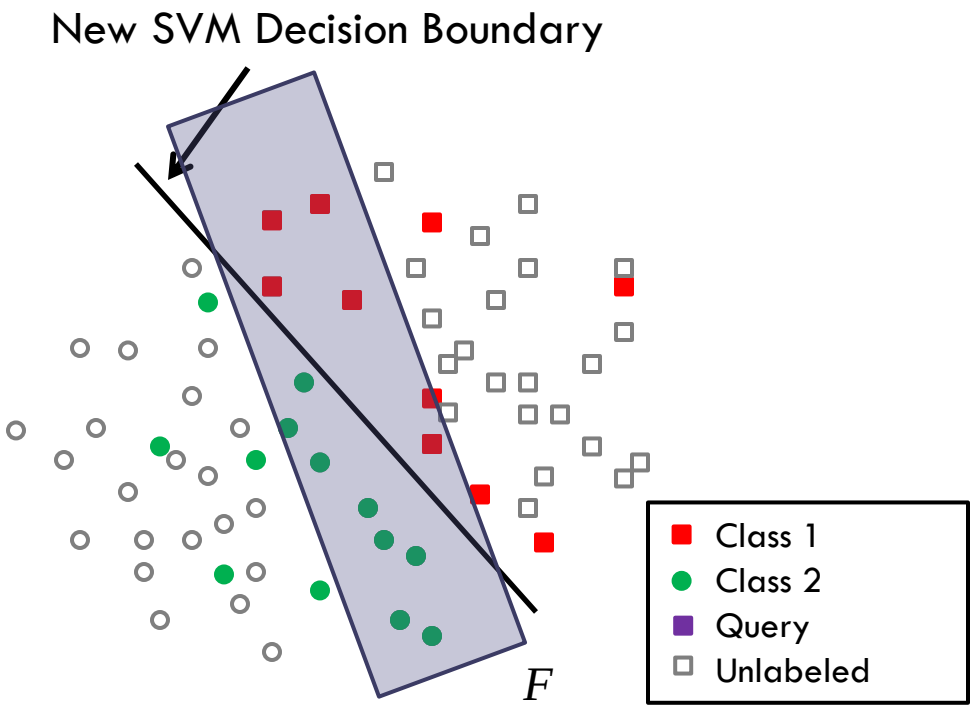
Active Learning (AL)

- Samples closest to the decision boundary are selected, as these are deemed most important for improving classification accuracy



Active Learning (AL)

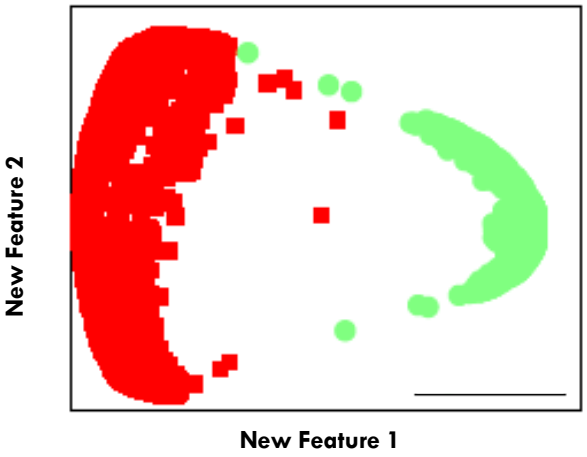
- Samples are used to re-train classifier
- New classifier is a better predictor



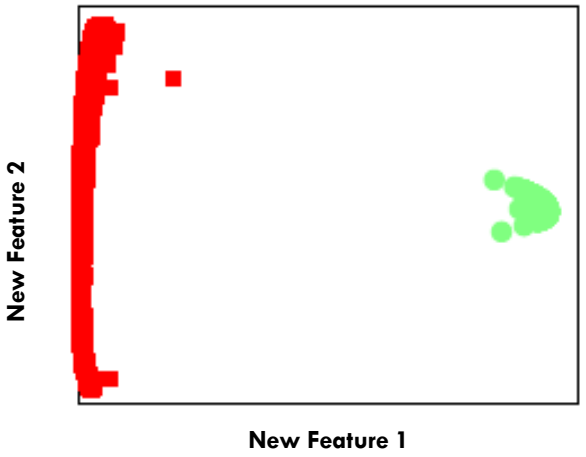
Semi-Supervised Graph Embedding with Active Learning (SSGEAL)

- Active Learning optimizes the SSDR embedding using the most informative samples

Semi-Supervised DR
(SSAGE
Zhou et al. 2006)



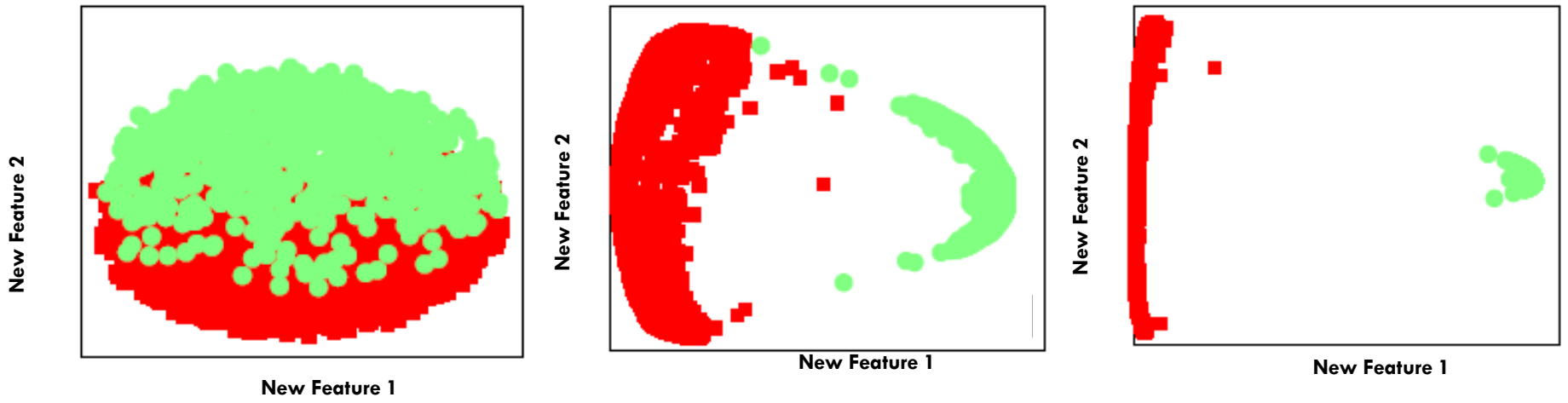
SSDR with Active Learning
(SSGEAL
Lee et al. 2010)



Recap

- SSAGE: Using labeled data > Unsupervised DR
- SSGEAL: Using informative labels > SSDR

Unsupervised DR (Graph Embedding (GE) Shi et al. 2000) → Semi-Supervised DR (SSAGE Zhou et al. 2006) → SSDR with Active Learning (SSGEAL Lee et al. 2010)

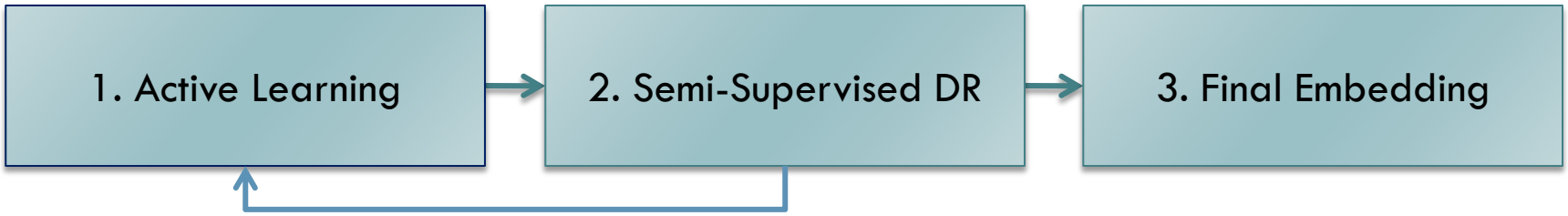


Overview

- Introduction/Background
- Overview of SSGEAL
  1. SVM-based Active Learning
  2. Semi-Supervised Dimensionality Reduction
  3. Stopping Criterion for Final Embedding
- Experimental Design
- Results
- Concluding Remarks

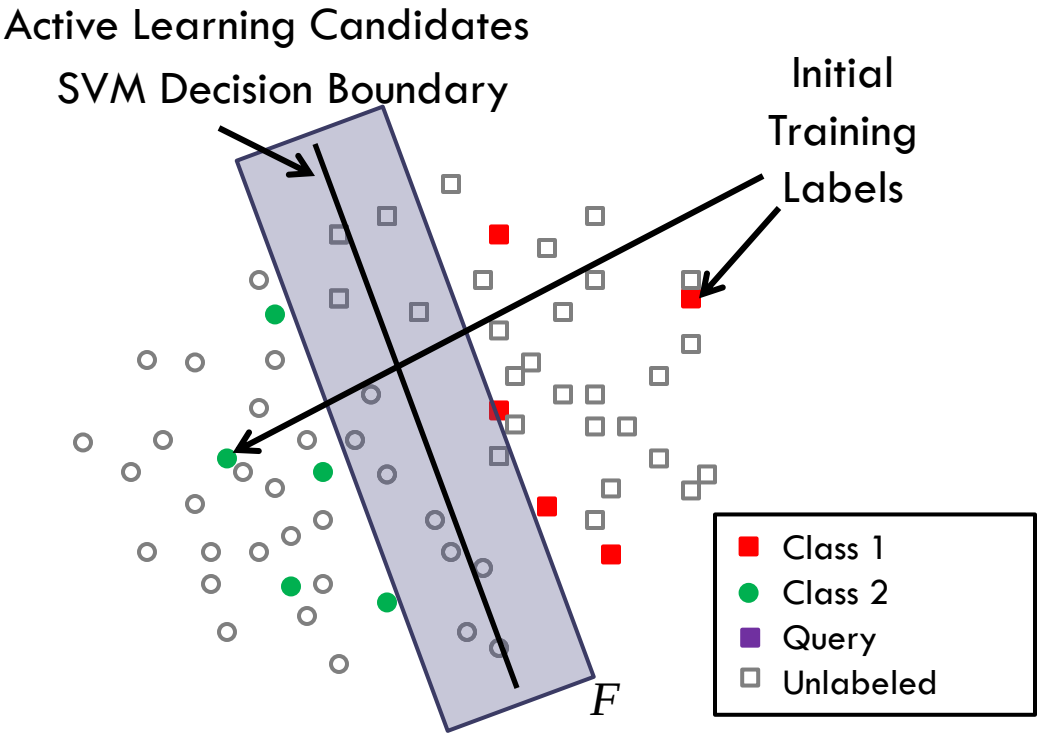
SSGEAL: Overview

- SSGEAL is an iterative approach for leveraging active learning and semi-supervised learning to obtain the best possible embedding.



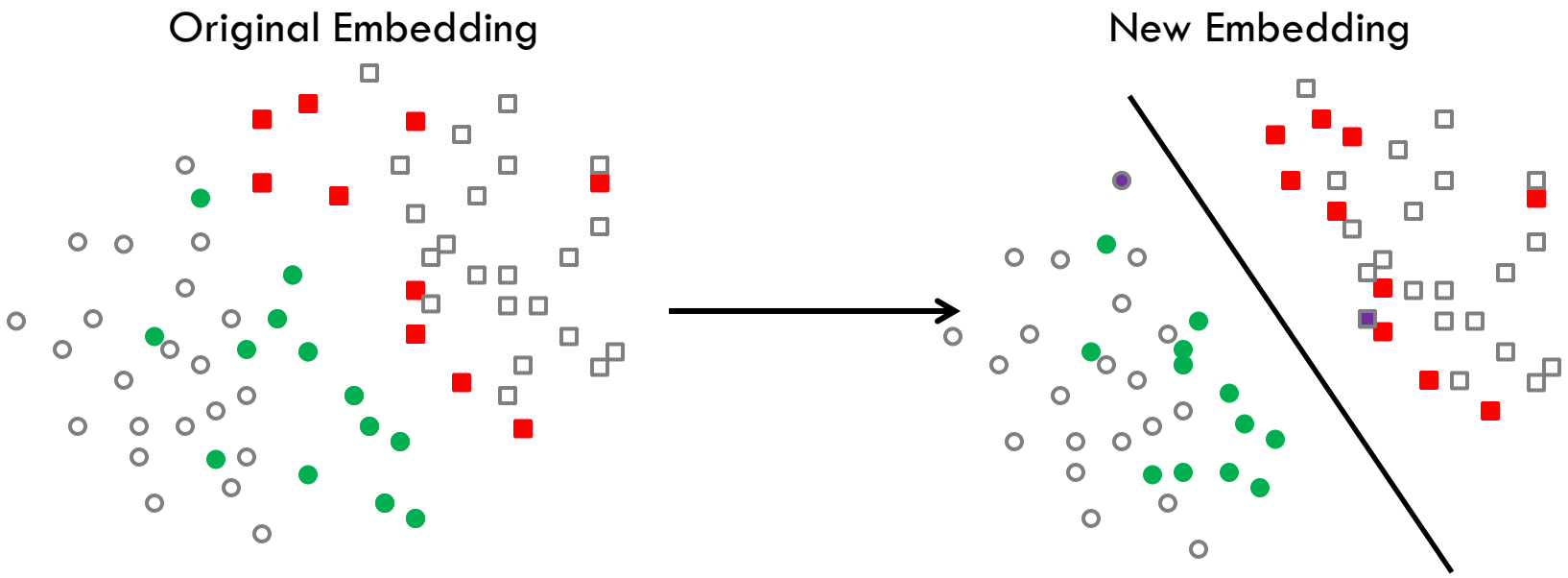
1. Active Learning

- An Initial Training Set $X^{Tr} = \{x_1, \ldots, x_n\}$ is used to kick start the Algorithm
- SVM classification performed in embedding space
- Points close to the boundary are added to X^{Tr}



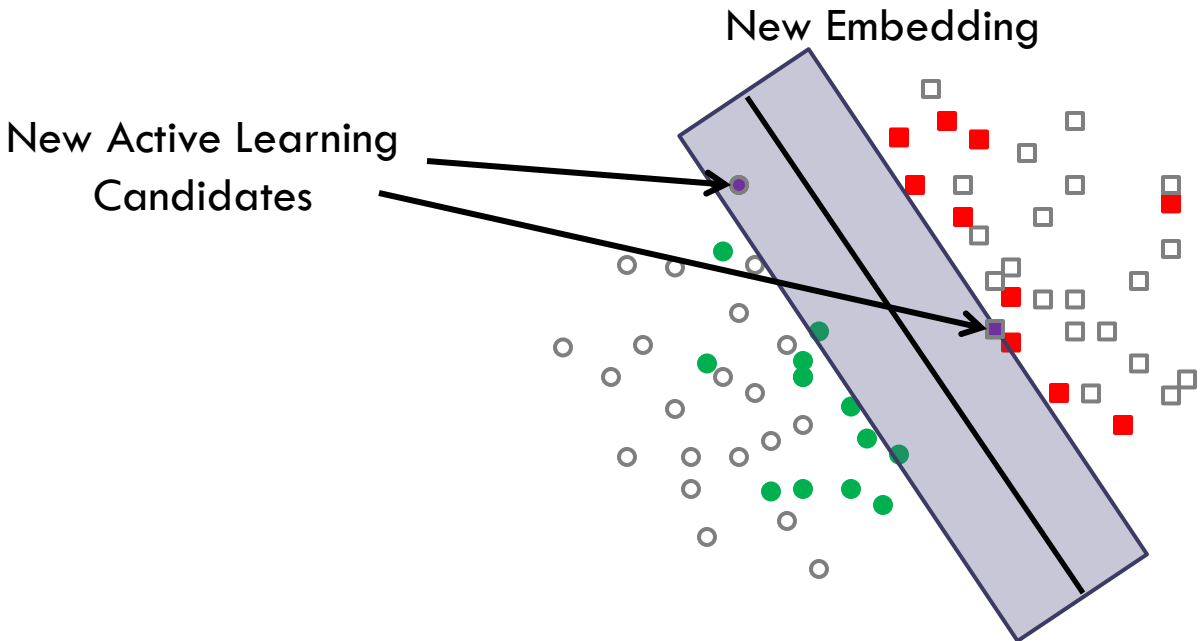
2. Semi-Supervised DR

- To build New Embedding, SSDR uses
  - Similarity Information from Original Embedding
  - Training Set X^{Tr}



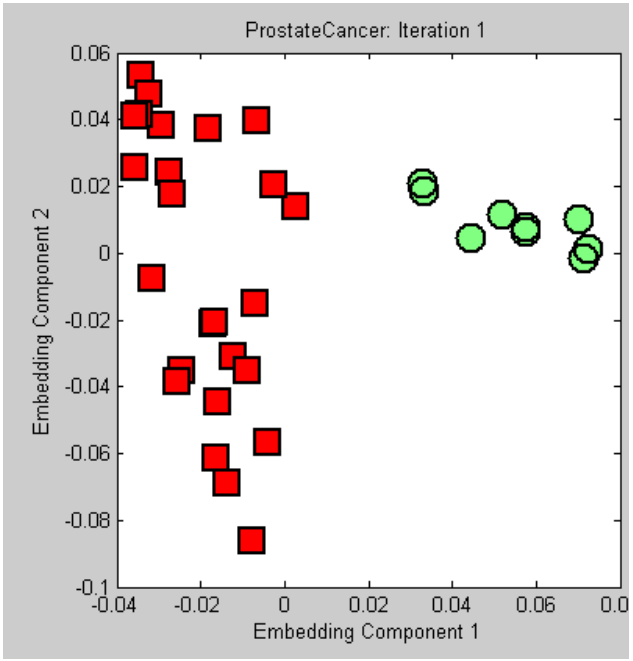
2. Semi-Supervised DR

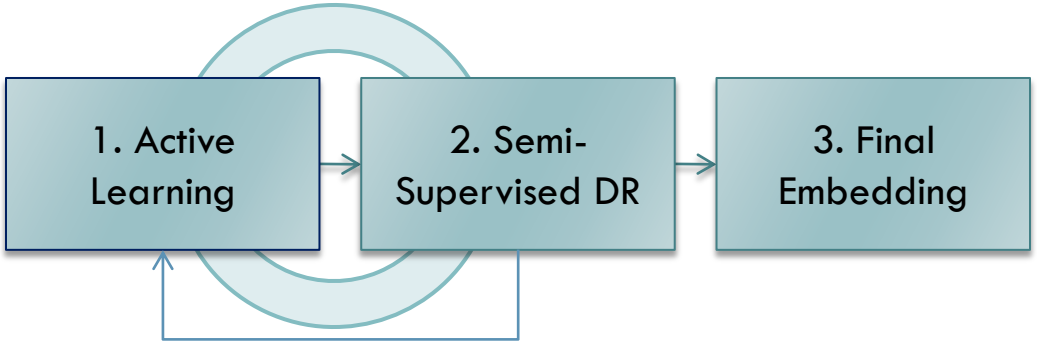
- To build New Embedding, SSDR uses
  - Similarity Information from Original Embedding
  - Training Set X^{Tr}
- New Embedding and Training Set X^{Tr} used for Active Learning

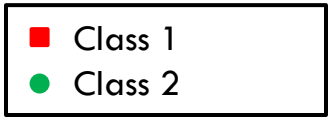


3. Final Embedding

- SSGEAL continues to improve the embedding until a stopping criterion is met
- At this point, the final embedding is achieved

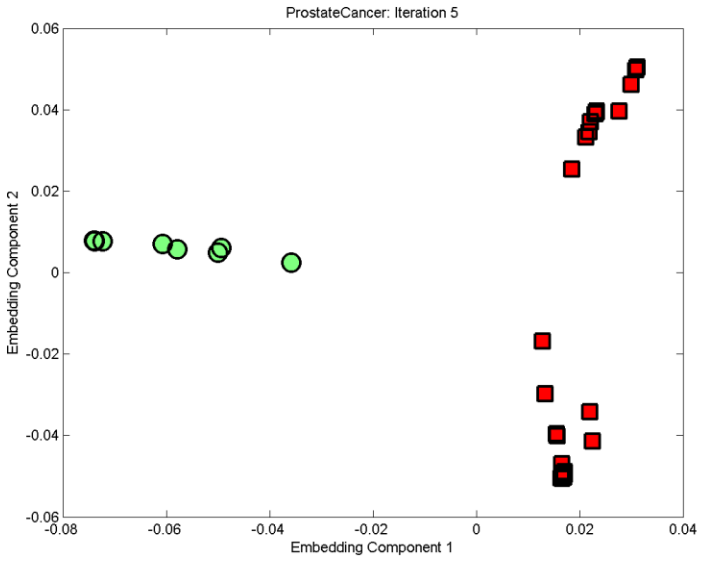


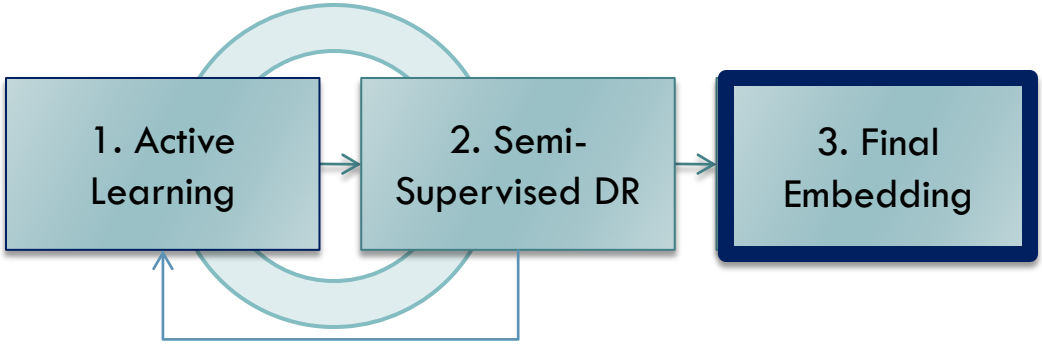




3. Final Embedding

- SSGEAL continues to improve the embedding until a stopping criterion is met
- At this point, the final embedding is achieved





Overview

- Introduction/Background
- Overview of SSGEAL
- Experimental Design
  - Datasets
  - DR Methods Used for Comparison
  - Objectives
- Results
- Concluding Remarks

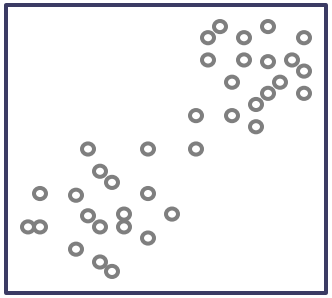
Gene Expression Cancer Datasets

- 7 binary-class gene expression datasets
- Goal is to discriminate between the cancer classes
  - ie. For the Prostate Cancer dataset, Tumor versus Normal classes

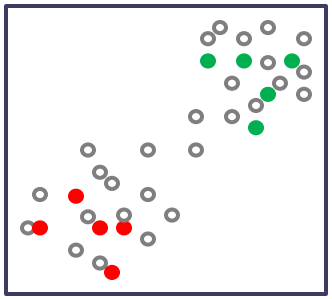
Datasets	Samples	Dimensions	Source
Prostate Cancer	25 Tumor, 9 Normal	12600 genes	Singh et al. 2002
Colon Cancer	22 Tumor, 40 Normal	2000 genes	Alon et al. 1999
Lung Cancer	15 MPM, 134 ADCA	12533 genes	Gordon et al. 2002
ALL/AML	20 ALL, 14 AML	7129 genes	Golub et al. 1999
DLBCL Tumor	58 Tumor, 19 Normal	6817 genes	Shipp et al. 2002
Lung Cancer (Mich)	86 Tumor, 10 Normal	7129 genes	Beer et al. 2002
Breast Cancer	10 Tumor, 20 Normal	54675 genes	Turashvili et al. 2007

Comparison of 3 DR Methods

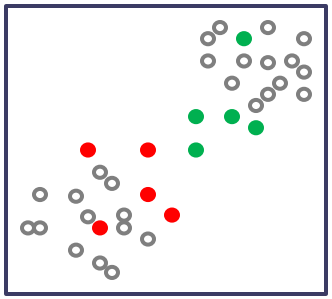
Graph Embedding (GE)
(Shi 2000)



Semi-Supervised Agglomerative Graph Embedding (SSAGE)
(Zhao 2006)



Semi-Supervised Graph Embedding with Active Learning (SSGEAL)
(Lee 2010)



Quantitative Evaluation

- Quantitative Improvement via 2 measures
  - Cluster Overlap using Silhouette Index (SI)

Mean Within-Class Distance

$$A_i = \sum_{j, Y(\mathbf{x}_j) = Y(\mathbf{x}_i)} \|\mathbf{x}_i - \mathbf{x}_j\|_2$$

Mean Inter-cluster Distance

$$B_i = \sum_{j, Y(\mathbf{x}_j) \neq Y(\mathbf{x}_i)} \|\mathbf{x}_i - \mathbf{x}_j\|_2$$

$$\phi^{SI} = \sum_{i}^{N} \frac{B_i - A_i}{\max[A_i, B_i]}$$

Minimize Within-Class Distances A,
Maximize Inter-Class Distances B

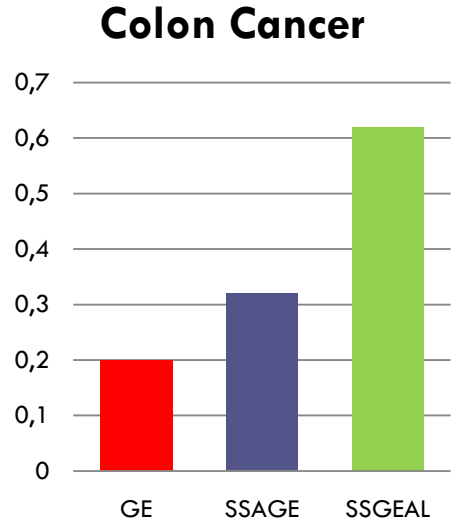
SI ranges between -1 and 1,
where 1 shows highest class
separability

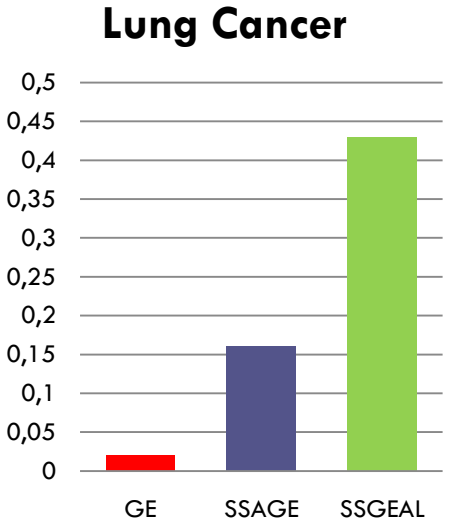
Quantitative Evaluation

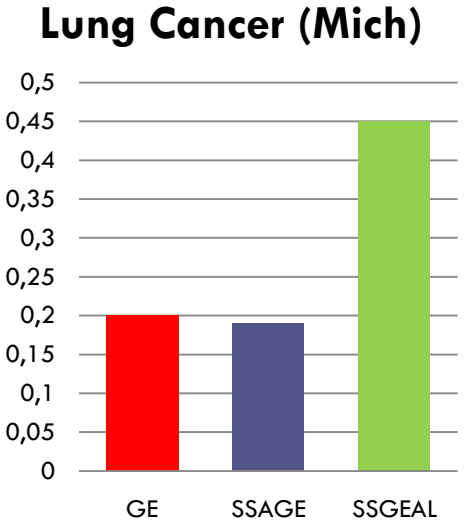
- Quantitative Improvement via 2 measures
  - Cluster Overlap using Silhouette Index (SI)
  - Classification Performance using Random Forest AUC
    - Random Forest: Bagging of 50 Decision Trees
    - AUC: Area Under Curve

Results - Gene Expression Data

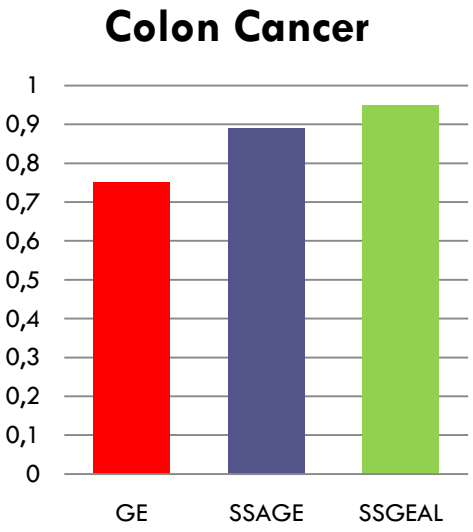
Silhouette Index

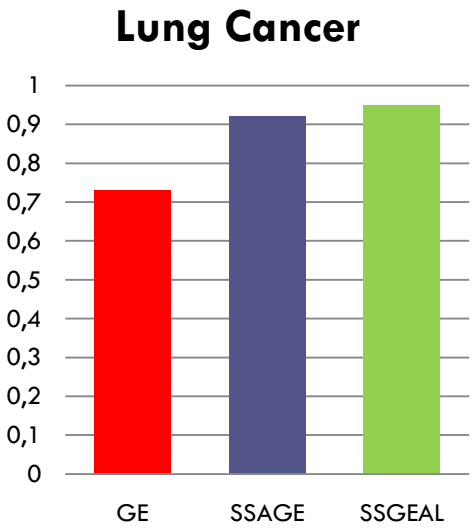


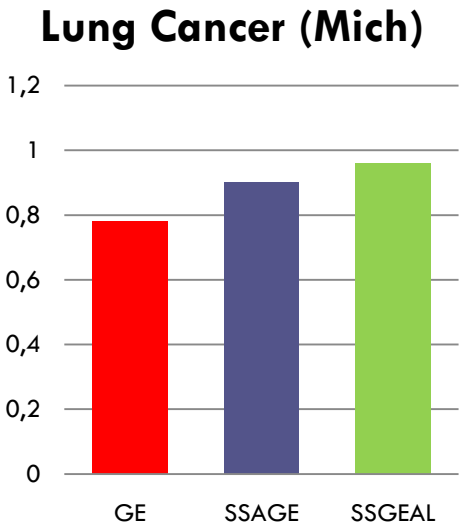




Classification AUC







Concluding Remarks

- Formalized a new DR scheme
- Integration of Active Learning into a SSDR scheme
  - Improves the Embedding
  - Improves the Classification
- SSGEAL shown to outperform SSAGE
  - 7 gene expression datasets
  - 2 Evaluation Measures

Acknowledgements

- Funding for this work made possible by
  - Wallace H. Coulter Foundation
  - New Jersey Commission on Cancer Research
  - National Cancer Institute
    - R01CA140772-01
    - R01CA136535-01
    - R21CA127186-01
    - R03CA128081-01)
  - The Cancer Institute of New Jersey (CINJ)
  - Bioimagene Inc.