EECS 349: Machine Learning
Prof. Bryan Pardo

Diagnosis of Viral Infection Using Gene Expression Profiles

Harleen Billing harleen@u.northwestern.edu
Clifton McFate mcfateclifton79@gmail.com
Sahil Shah sahil@u.northwestern.edu

NORTHWESTERN UNIVERSITY

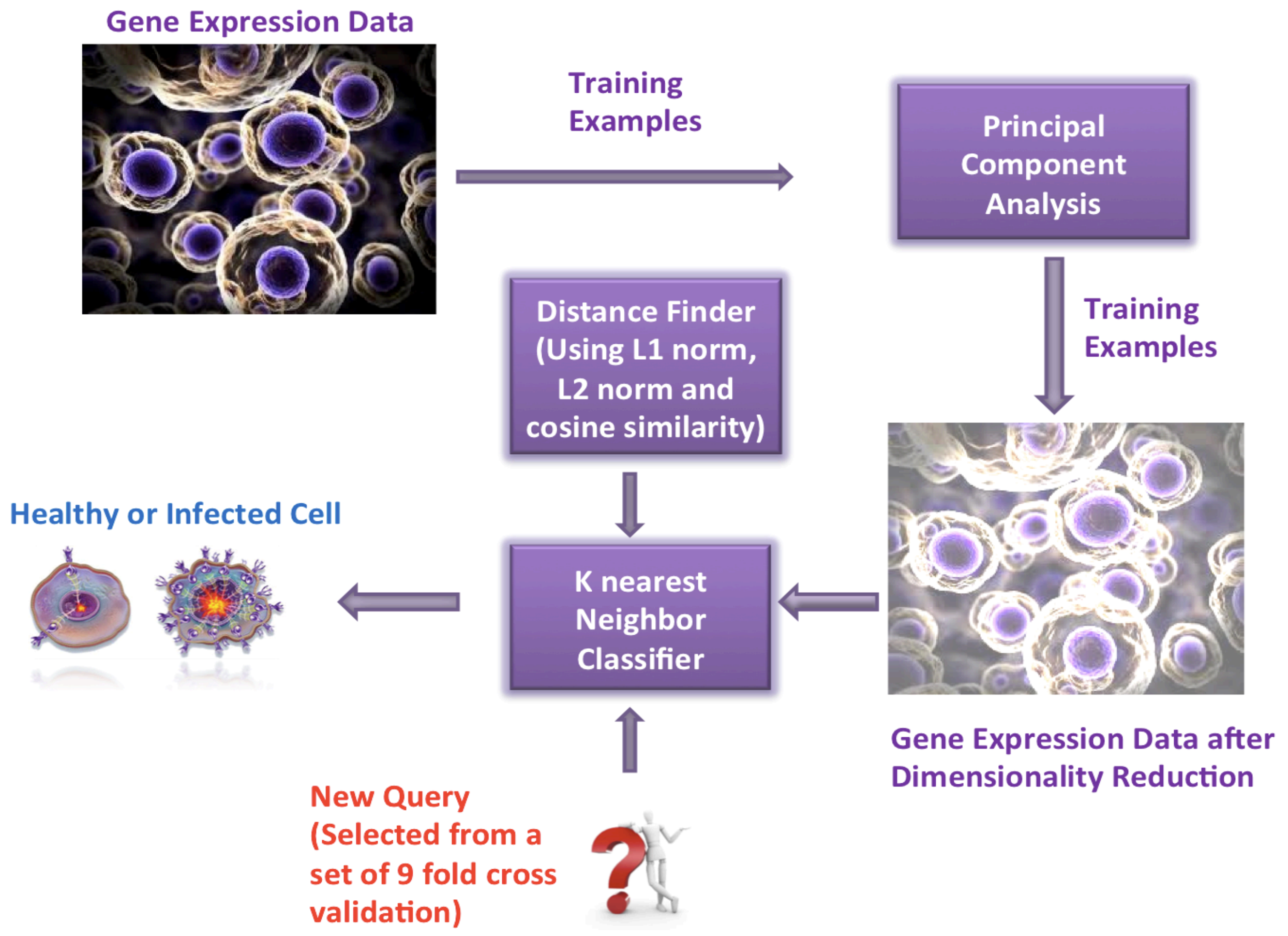
Motivation and Introduction

According to the Centers for Disease Control and Prevention, respiratory syncytial virus (RSV) results in an average 177,000 hospitalizations and 14,000 deaths per year among adults older than 65 and influenza resulted in 64 pediatric deaths during the 2012-2013 flu season. [1],[2] A reliable method for early detection for respiratory infections like these could result in better health outcomes.

The gene expression levels of an infected cell change upon infection and may be a good predictor of illness. These gene expression levels of a cell correspond to the state of the cell and are measured by the abundance of mRNA produced during transcription of each gene.

However, for any cell there are tens of thousands of genes measured and limited patient data available. Because of these constraints, we believe a k-nearest classifier would work well to classify infection and we seek to answer the question of which k values and distance measures perform best.

System Overview



Gene Expression Data
Training Examples
Principal Component Analysis
Training Examples
Distance Finder (Using L1 norm, L2 norm and cosine similarity)
Healthy or Infected Cell
K nearest Neighbor Classifier
Gene Expression Data after Dimensionality Reduction
New Query (Selected from a set of 9 fold cross validation)

- •We first use principal component analysis to reduce the dimensionality of the gene expression profiles.
- •We propose a k-nearest neighbor classifier that uses gene expression profiles of individual patient cells to classify them as infected or uninfected
- •We iterate with different k values and different distance measures to find method which works best for this data.

Dataset and Methodology

We evaluate the ability of a k-nearest neighbor classifier (with different k values and distance measures) to classify infection from gene expression profiles using cross validation on a pre existing dataset.

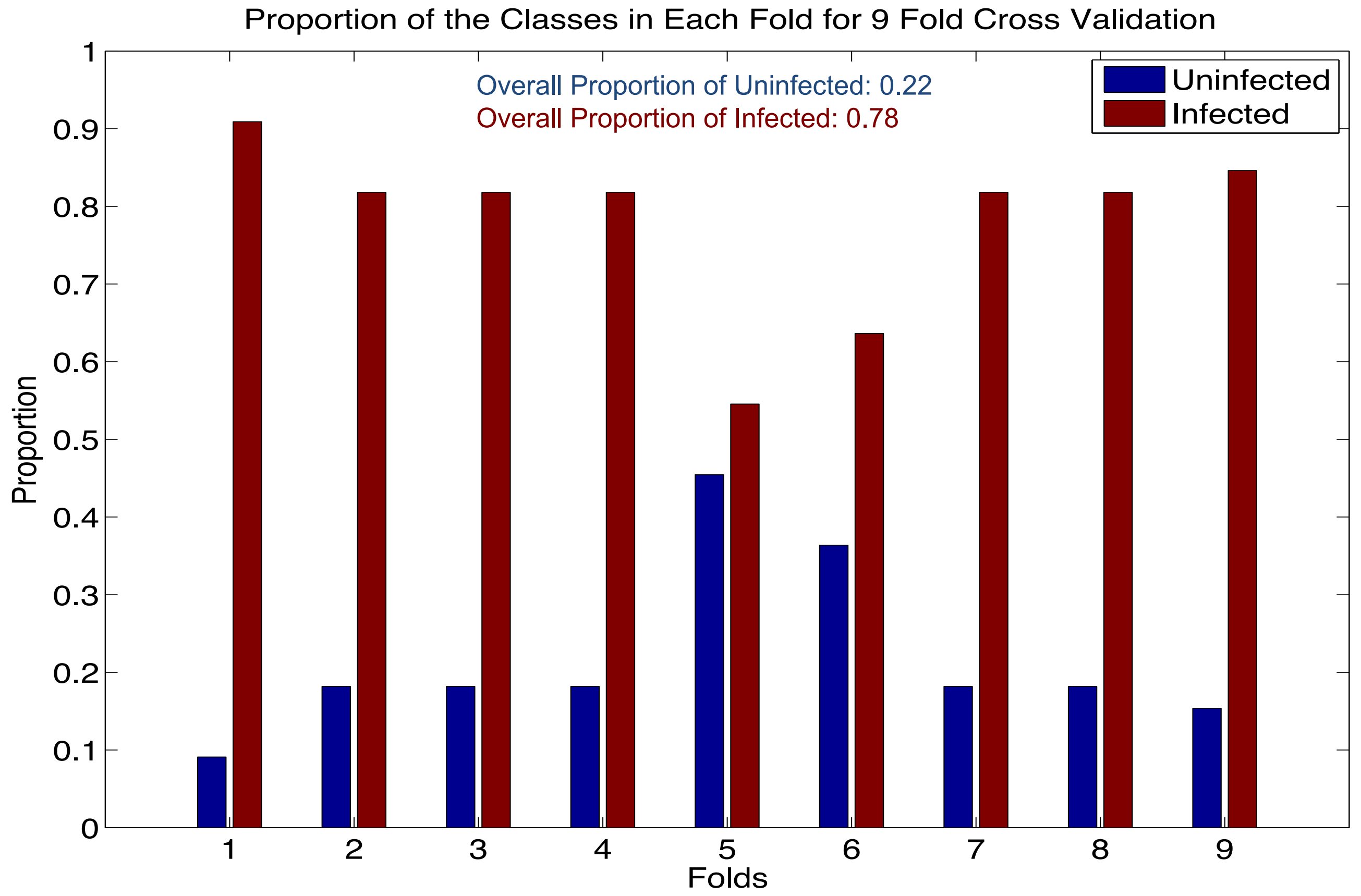
Dataset Summary

We used PCA to reduce the dimensionality

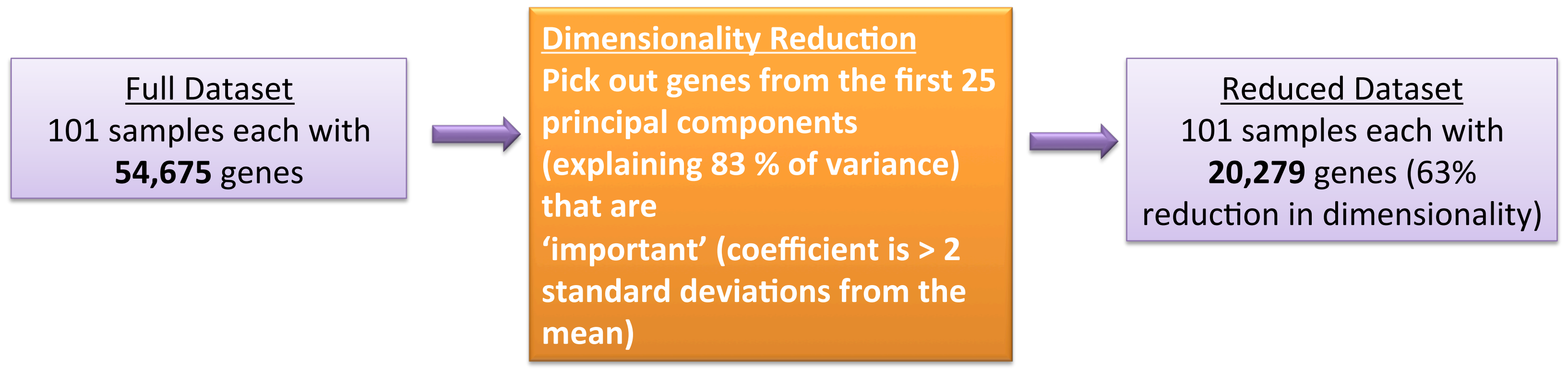
Samples	Gene 1	Gene 2	....	Gene 54,675	Classification
1	1.1628	1.0233	...	2.1372	Infected
2	1.6068	1.0327	...	.6898	Uninfected
...	...	...	...	...	...
101	1.0475	1.890	...	.5679	Uninfected

79 infected samples
22 uninfected samples

For training and evaluation we separated the data into 9 folds and performed cross-validation. We randomized the order of the cells to create an even distribution of infected and uninfected cells per fold



We use principal component analysis to reduce the dimensionality of the gene expression profiles.



Full Dataset
101 samples each with **54,675** genes

Dimensionality Reduction
Pick out genes from the first 25 principal components (explaining 83 % of variance) that are 'important' (coefficient is > 2 standard deviations from the mean)

Reduced Dataset
101 samples each with **20,279** genes (63% reduction in dimensionality)

With the reduced data, we train and test a k-nearest classifier using 9-fold validation across k values from 1 to 3 and using L1 and L2 norms as well as cosine similarity as distance measures.

Analysis of Results

Our top performing method was a k-value of 1 and L1 norm at 92.075% accuracy. We evaluated performance using the F1 measure:

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

The F1 measure was 0.9445. This was an increase of 15.25% in F1 from the lowest performing set (cosine similiarity, k = 5). Against a baseline that always chose infected (Fscore = 0.8779) our method provided a 7.5% increase.

Average F Score

K value	Manhattan	Euclidean	Cosine
1	0.94452	0.93458	0.8301
3	0.92718	0.92718	0.82596
5	0.93974	0.93974	0.81953
7	0.92933	0.94415	0.82512

Furthermore, we evaluated our PCA strategy against a baseline of choosing random loadings. The top performing measure again was k = 1 with L1 norm. With an average F score of 0.944, there was no significant difference between the two methods. There also does not seem to be a difference between k-values across measures.

Conclusion & Future Work

We conclude:

- K-Nearest can effectively differentiate between infected and uninfected cells using gene expression data.
- Simple was better: L1 Norm and K=1 was the best.
- Dimensionality reduction is necessary, but it is not necessary to use PCA.

In the future we plan to:

- Combine different data sets to make up for limited data.
- Apply to time-course data and different diseases.

Acknowledgements

Thanks to Prof. Bryan Pardo, Prof. Rosemary Braun, and Prof. Doug Downey for their guidance in developing this project.

[1] http://www.cdc.gov/rsv/research/us-surveillance.html
[2] http://www.cdc.gov/flu/pastseasons/1213season.htm

For more information, see:
eecs349-infection-classifier.weebly.com/