Clustering Chemical Datasets
- Clustering chemical datasets is a fundamental procedure in library design and selection
- Clustering can also be used as a data-partitioning method to focus on smaller subsets
- Clustering can be used as a classification technique

Why is Fast SVD a Better Way?
- The original matrix is randomly sampled
- The sub-matrix is then decomposed
- Though the fast SVD algorithm utilizes the SVD, the matrix being decomposed can be significantly smaller

Datasets & Descriptors
- AMES Test: 4337 compounds
- Categorical
- 166 bit MACCS fingerprints
- Constitutional, geometric, and topological descriptors
- Calculations and analysis performed with MOE and R
- Aqueous Solubility: 1256 compounds
- Real valued

Methodology
- Evaluate:
  - Intra-cluster similarity
  - Class enrichment
- Visualize:
  - Singular values
  - PAM clustering
- Cluster the rows of the projection

How Fast is the Fast SVD?
- Benchmarked on a 4337 x 4337 matrix
- Each case was run 10 times
- No significant error until less than 10% of the rows are sampled
- Simple SVD = 368s

Comparison of Timings
- PAM is a more robust version of k-means
- Claar is an extension of PAM but is more sensitive to data structure
- However, the fast SVD is significantly faster than all other partitioning methods

Applications of Spectral Clustering
To Chemical Datasets
Rajarsht Guha and Peter C. Jurs
Department of Chemistry
The Pennsylvania State University

Summary
- Fast SVD based clustering gives nearly identical results compared to slow SVD based clustering at significantly higher speed
- The average intra-cluster similarities are comparable to PAM and k-means
- SVD based clustering appears to emphasize specific clusters over others
- The algorithm appears to handle correlated and information-poor descriptors well

Further Work
- Use feature selection when considering real-valued descriptors
- Combine the asymmetric affinity matrix with the fast SVD algorithm
- Investigate methods to avoid the distance matrix calculation

Acknowledgements
- Dr. Petros Drineas for clarifications about the fast SVD algorithm
- Dr. Igor Fischer for providing code to test the asymmetric affinity approach
- Chemical Computing Group for providing MOE

SVD Clustering
- Based on the Singular Value Decomposition
- Projects the original matrix onto a k-D subspace
- Clustering is performed in the reduced subspace
- It has been shown that the SVD itself represents a clustering