# Clustering Class

## Intro to clustering

http://www.mmds.org/mmds/v2.1/ch07-clustering.pdf http://www.mmds.org/mmds/v2.1/ch07-clustering.pptx

### Topics:

- Kmeans
- Agglomerative (Single-link; Complete-link)

### Extra Entity Resolution (query driven)

Query-Driven Sampling for Collective Entity Resolution https://www.dropbox.com/s/xvub1d2f450ek2v/cgrant-iri-2016.key?dl=0

## In-class activity

We are going to cluster the movie reviews. Split up in groups and complete the following activity.

0) Download the sentiment review data set

http://ai.stanford.edu/~amaas/data/sentiment/aclImdb_v1.tar.gz

Use the review data in to pos/neg folder. Cluster the training set first.

1) Normalize and vectorize the data

Create a vector from each term so we cannoit

2) Cluster the data set

Use one of the following algorithms:

- Kmeans
- Agglomerative

3) Visualize the data set

Feel free to use any library.

- Matplotlib
- Seaborn
- ggplot

4) Share your progress with the class.