# 95-865
# Unstructured Data Analytics

Mid-mini Quiz Review

Spring 2018

Emaad Ahmed Manzoor

# Things to Know

- **Reading files**
  - numpy `loadtxt`, Python I/O
  - spaCy tokenization
- **Descriptive plots**
  - Scatter, line
  - Bar/histogram
  - Legends, axes ticks/labels
- **Concepts**
  - PMI, PCA, t-SNE
  - Matrix factorization
  - Going from math to code
- **Interpretation skills**
  - Choosing a method/metric

**Look at all the demos!**

- **spaCy:** https://gist.github.com/georgehc/93dc0b8ff4f5d5f56e35b41647122ad3
- **PCA:** https://gist.github.com/georgehc/4100f748246872b4eddeff9b82212d89
- **MDS:** https://gist.github.com/georgehc/6446875f272f2c76e910704f20d7154a
- **t-SNE:** https://gist.github.com/georgehc/83982a65bcec47d5be7bf994946aa043
- **Clustering:** https://gist.github.com/georgehc/6d0f476ac282e51694849e8aec225ac3
- **LDA:** https://gist.github.com/georgehc/d2353feef7e09b4b53fc087d44f75954