# Machine Learning

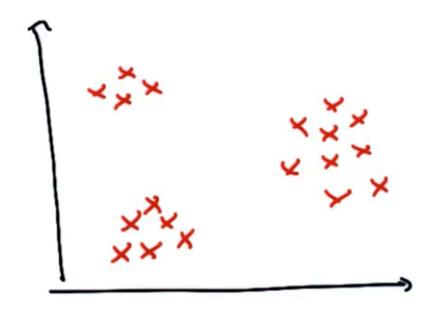
Aaron

## **Bonus Target And Features**

- Regressing Bonus Against LTI
- Salary Vs. LTI For Predicting Bonus
- Outliers Break Regressions

- Slope and Score of Regression with Outliers
- Slope and Score after Cleaning
- Enron Outliers
- Any More Outliers?
- Identifying Two More Outliers. Remove These Outliers?

# Unsupervised Learning



http://www.naftaliharris.com/blog/visualizing-k-means-clustering/

- What features will your clustering algorithm use? (*k\_means/k\_means\_cluster.py*)
- Deploying Clustering
- Clustering With 3 Features
   ('salary', 'exercised\_stock\_options', 'total\_payments')
- Stock Option Range
- Salary Range

- What Kind Of Scaling
- Computing Rescaled Features (a "salary" value of \$200,000 and an "exercised\_stock\_options" value of \$1 million)
- Computing Rescaled Features
- When To Deploy Feature Scaling

- Warming Up With ParseOutText()
  (run "tool/parse\_out\_email\_text.py")
- Deploying Stemming
- Clean Away "Signature Words"
- TfIdf It
- Accessing TfIdf Features