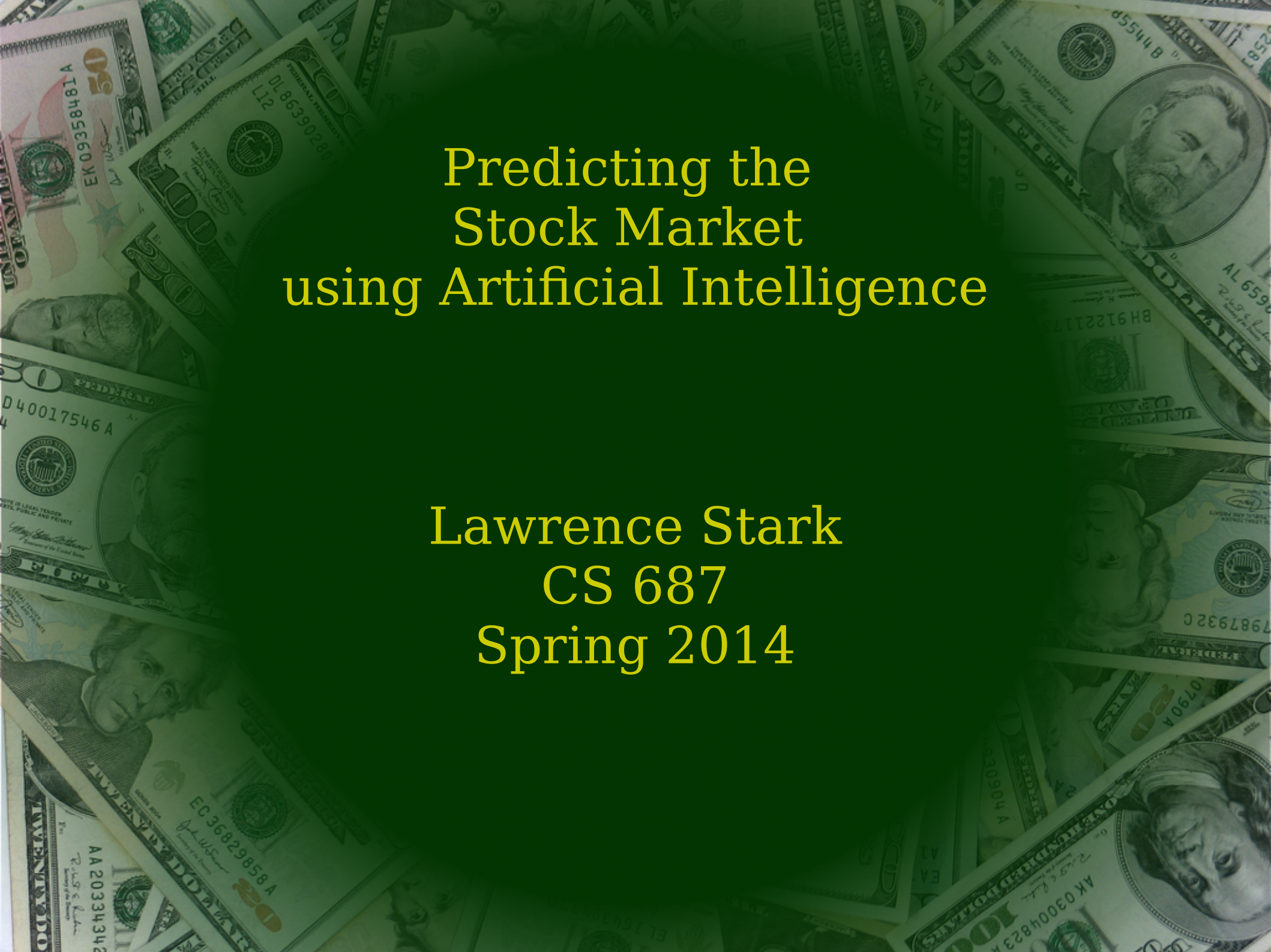


Predicting the Stock Market using Artificial Intelligence

Lawrence Stark

CS 687

Spring 2014

Topic

- Using historical data (3 days), predict whether tomorrow's stock market will close UP or DOWN
- Predict stock market volatility using historical VIX data (16 & 44 days)
- Automated prediction based on model developed from individual stock market data.

Utility

- Get Rich the Quick and Easy Way!
- Personal Finance
  - e.g. Self-managed 401k
- Complex Signal Analysis (Data Mining):
  - Find patterns given unknown distribution
  - Predict future behavior for irrational agents

Method

- Candlestick Pattern
  - Munehisa Homma: Japanese Rich Trader from 1700's
  - Steve Nison: Applied Homma's candlesticks to contemporary investment (stocks)
- Model Market Behavior
  - Use 500 stocks to learn individual stock movement
  - Use model to predict market value for next day

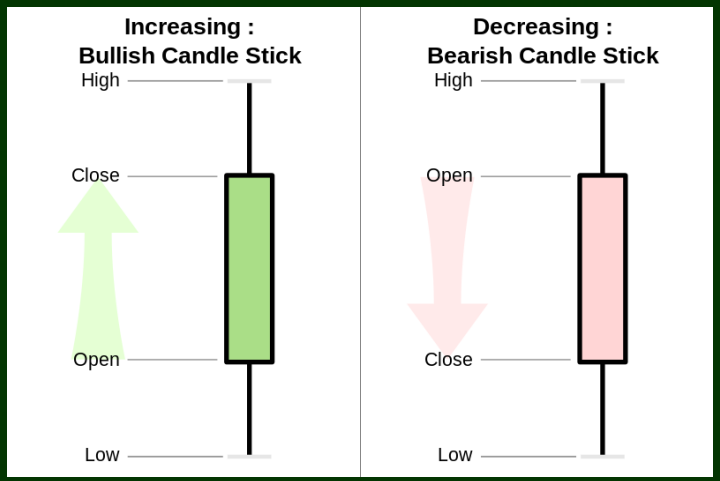
Background

- JPM: Days of loss in 2013 = 0
- Virtu: Days of loss 2009-2013 = 1
- Support Vector Machines
- Neural Networks
- Twitter
- Autoregressive Integrated Moving Average (ARIMA)
- Echostate Networks

Data Source

- Tradestation: www.tradestation.com
- Stocks: S&P 500 + SPDR
- 3 Day Sliding Window (Day 4 = Label)
  - Train/Test : approximately 2.2 million samples
  - Validate: approximately 5,200 samples
- VIX: CBOE
  - Approximately 5,200 samples
  - Same 20 year span as S&P 500 data

Data



- Features:
  - Open, High, Low, Close
  - For each of Day 1 to 3
  - Delta Close Day1/2 and Day 2/3
  - Label: related to line slope: Up, Down, Peak, Trough

Example:
10.97,11.05,10.82,10.97
11.01,11.05,10.56,10.67
10.60,10.67,10.57,10.60
-0.30,-0.07,DOWN

Feature Extraction

- So Far: 3 Day candlestick patterns
  - Only 15 attributes
  - Manually reduced from 24
  - PCA suggests only 3: $\Delta C12$, $\Delta C23$, D_3Vol
- VIX:
  - 16 and 44 Day
  - 80 and 220 attributes respectively

AI Methods

- Baseline: random buy and sell
- Classification:
  - Bayesian Inference
  - Radial Basis Functions
- Regression:
  - Linear Regression
  - Support Vector Machine Regression
  - Radial Basis Function Regression
- Clustering – K-Means

Software Platforms

- WEKA Version 3.7
  - Used only standard algorithms – no plug-ins.
- Java
  - Custom program written to preprocess the data and produce N-Day sliding windows (3, 16, and 44)

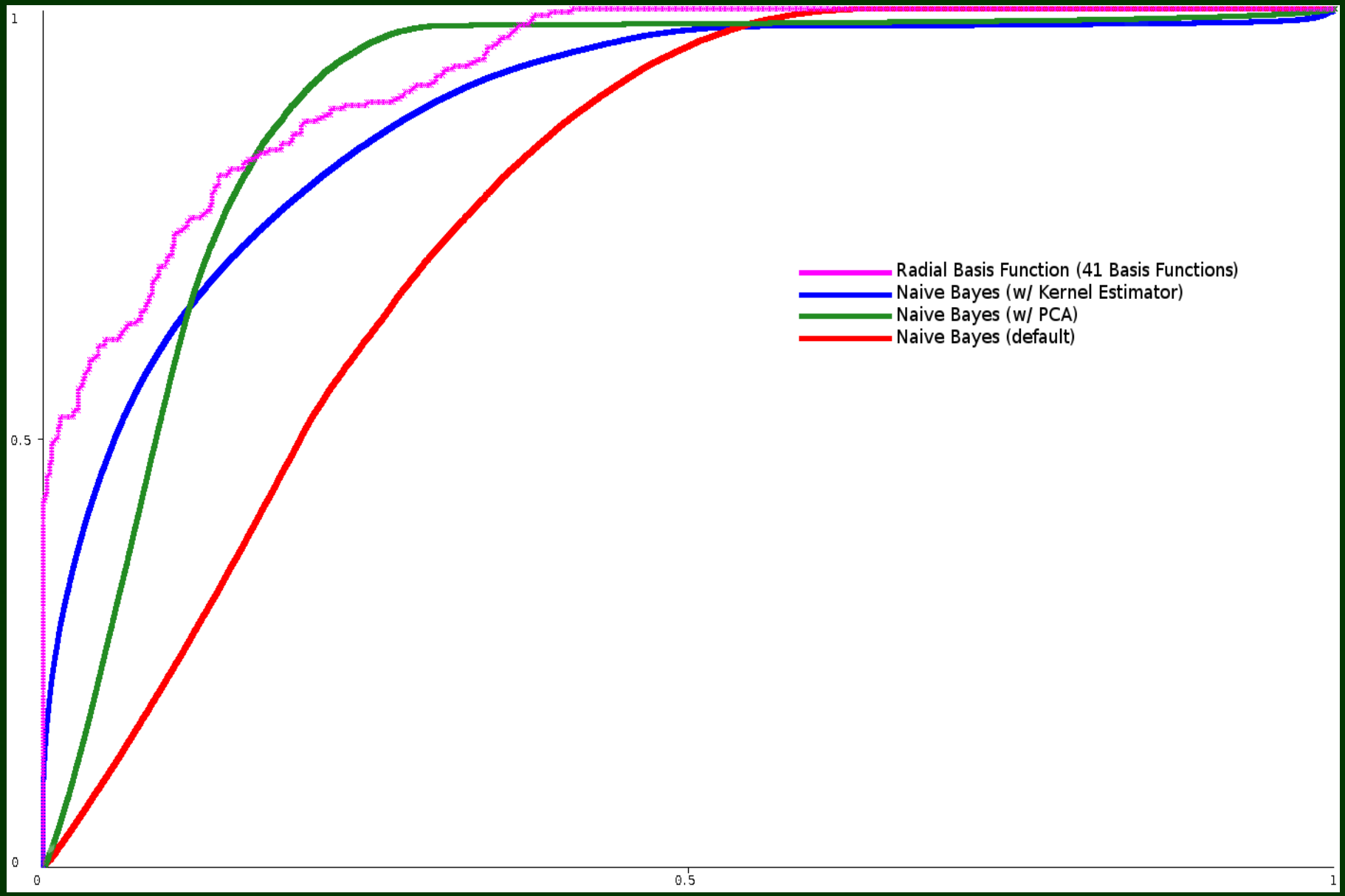
Performance Evaluation

- SPDR (spider)
  - Mimics entire S&P 500
  - Standard for performance evaluation
- Error: $\sqrt{(Z(t+1)-SPDR(t+1))^2}$
- Metrics:
  - Accuracy: predicted market status vs. SPDR
  - ROI: the amount of money gained from trades
  - Market Days: days money is used for trading

Cross Validation

- Training Set
  - 50% of S&P 500 (1.1 million)
- Test Set
  - Remaining 50% of S&P 500 (1.1 million)
- Validation Set
  - 100% of SPDR (5235)
- Validation set deliberately not mixed with train/test sets to mimic real world.

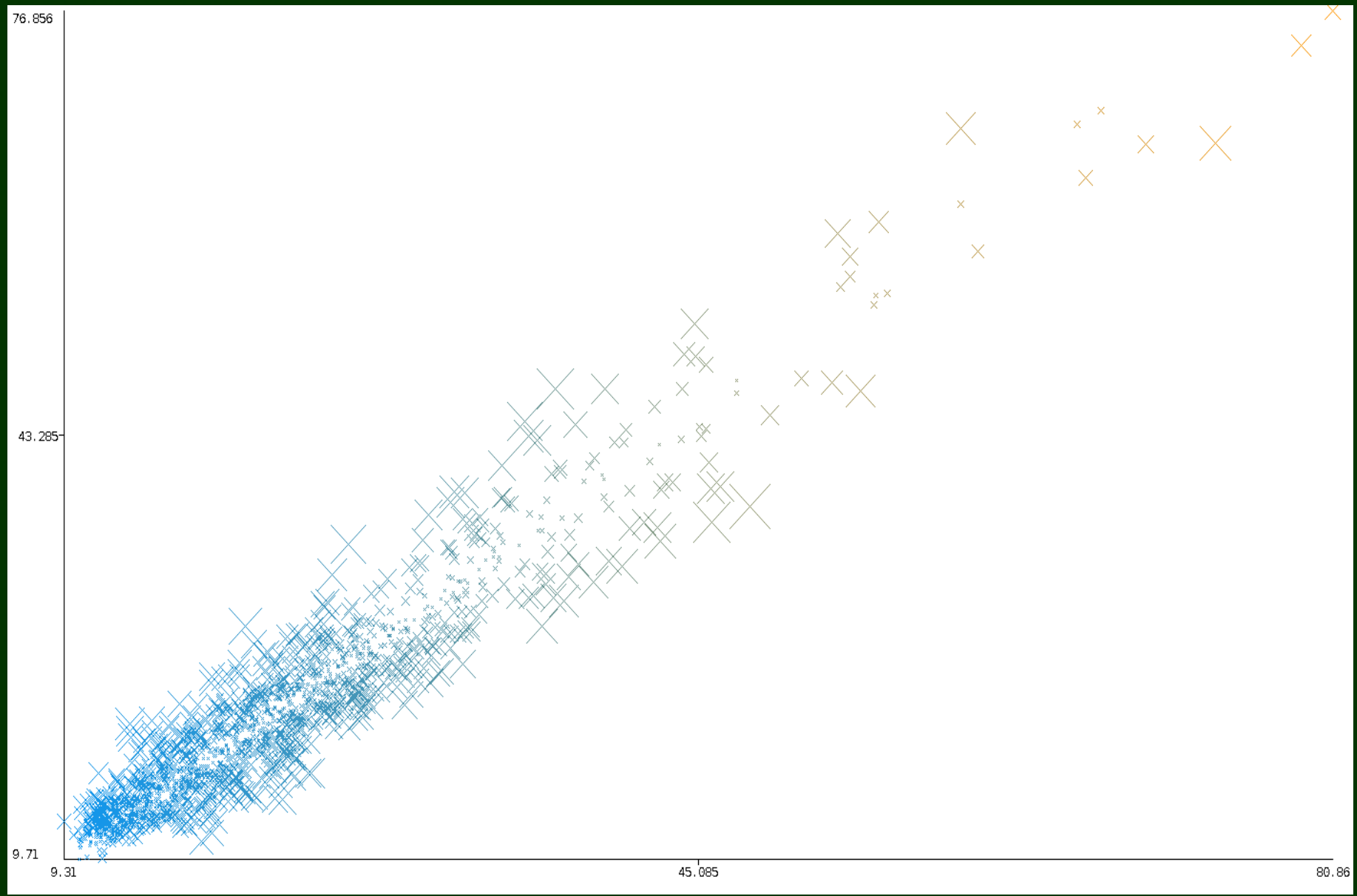
Data Visualization



Final Results

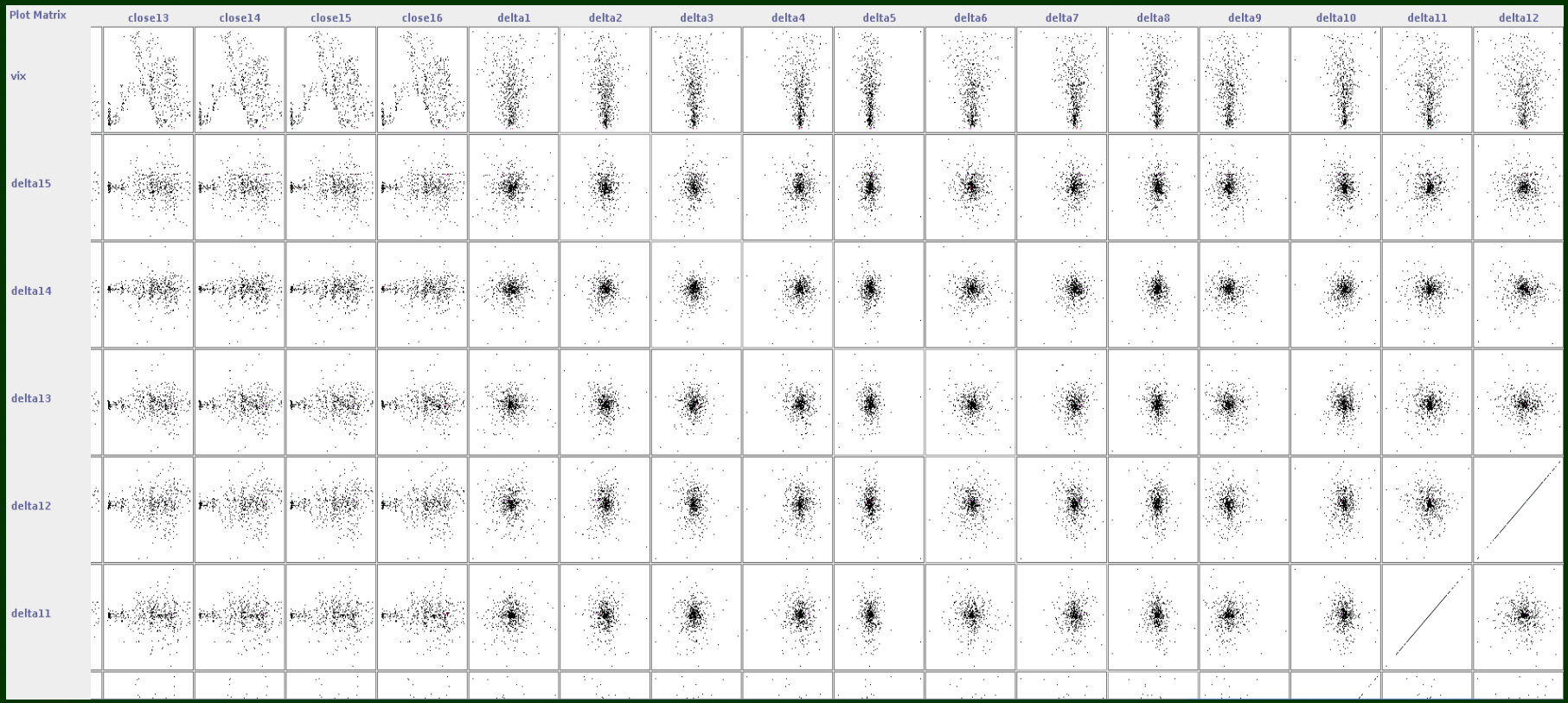
Trial	Accuracy	Market Days	ROI
Random	51%	2618	-31.69%
Naive Bayes3 w/ PCA	55.16%	1201	268.46%
Radial Basis Function Net	80.92%	488	432.10%
Radial Basis Regression	70.49%	N/A	N/A

Visualization of RBF Errors



Results From Clustering

Visualization of K-Means Clusters:



Conclusion

- Accounting for volatility makes a big difference!
- Achieved success as 2 separate models:
  - Classification (discrete categories)
  - Regression
- Next step: combine models
  - Expectation is greater ROI (not accuracy)
  - Predictive ability is maximized with current models
  - Include other factors for greater accuracy