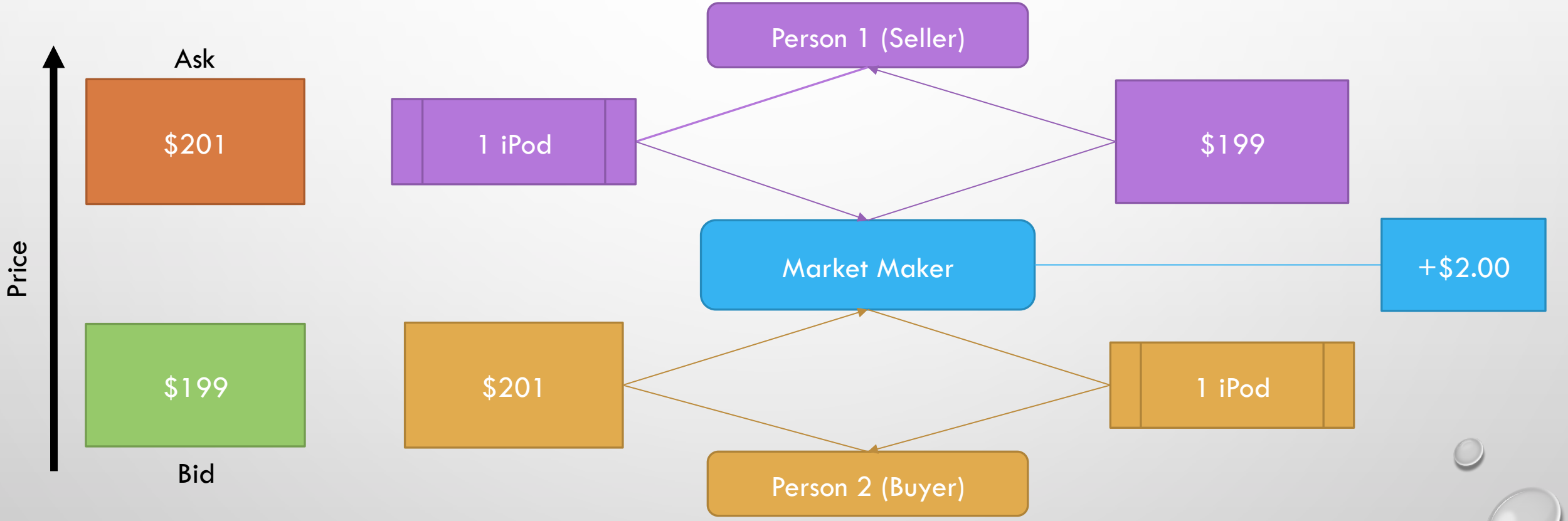
MARKET MAKING WITH MACHINE LEARNING

KEVIN CHEN, KAPIL KANAGAL, YU WU

MARKET MAKING EXPLAINED

- MARKET MAKERS PROVIDE LIQUIDITY IN AN EXCHANGE BY QUOTING BOTH A BID AND ASK PRICE IN A FINANCIAL INSTRUMENT
- MAKE PROFIT BY TAKING ADVANTAGE OF THE BID-ASK SPREAD

MARKET MAKING EXAMPLE

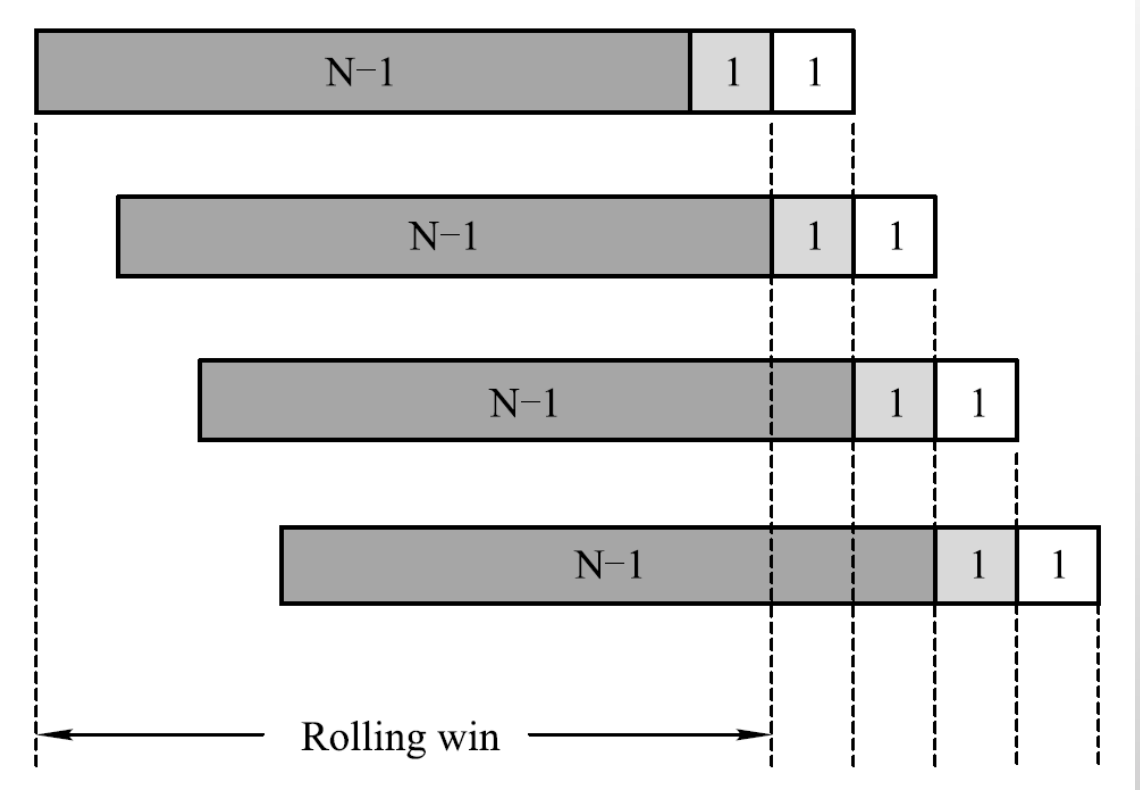


STRATEGY OUTLINE

- LOOK AT MARKET BOOK DATA AND PERIODICALLY PREDICT EVOLUTION OF STOCK PRICE WITH MACHINE LEARNING MODEL
- POST BID AND ASK ORDERS ADJUSTED FOR THIS EVOLUTION
- IF ONLY ONE SIDE ONLY IS HIT AFTER A WAITING PERIOD, CANCEL ORDER AND POST NEW ADJUSTED QUOTES

PREAMBLE STRATEGY TRAINING

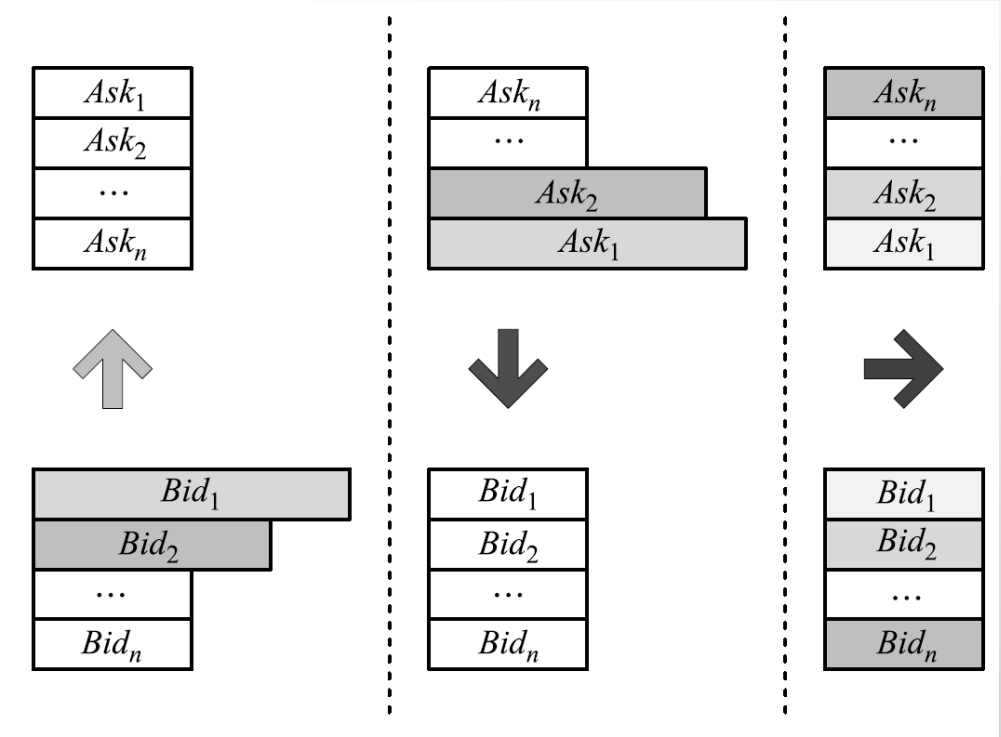
- TRAINING FROM HISTORICAL DATA
- SAMPLING OF MARKET BOOK DATA
- STOP LOSS FUNCTION



Li, Deng, Zhu, Wang, Xie; *An intelligent market making strategy in algorithmic trading*

DATA

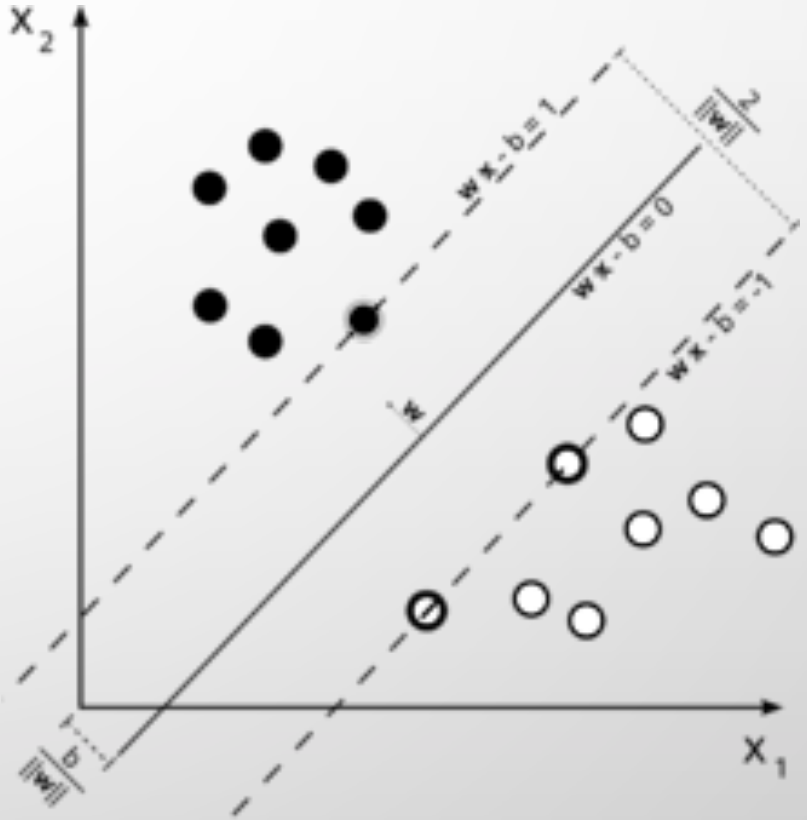
- MARKET BOOK DATA CONTAINING PRICE AND QUEUE SIZE OF THE FIRST FEW LEVELS OF BIDS AND ASKS.
- COMBINED TO MEASURE RELATIVE SIZE OF ASK AND BID SIDES OVER DIFFERENT TIME PERIODS
- CHOICE OF THRESHOLD
- CONSERVATIVE VS LIBERAL ESTIMATES



Li, Deng, Zhu, Wang, Xie; *An intelligent market making strategy in algorithmic trading*

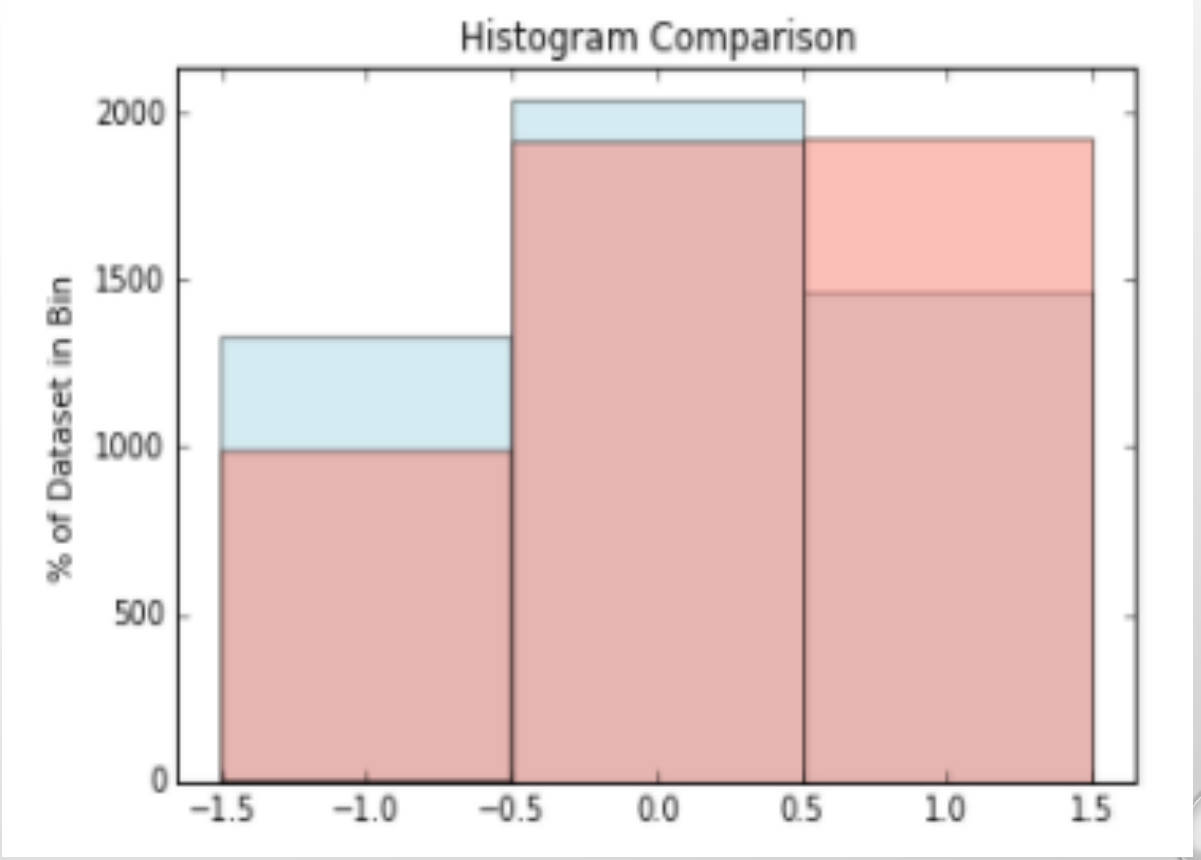
PREDICTING WITH SVM

- Basic idea is to find a hyperplane that separates the two dataset the best (largest margin)
- This optimization can be formulated using kernels
- To find the best parameters, three thing are considered:
    1. Accuracy of prediction
    2. Probability that up is predicted as down or down is predicted as up
    3. Conditional probability of actual going up if predicted going up



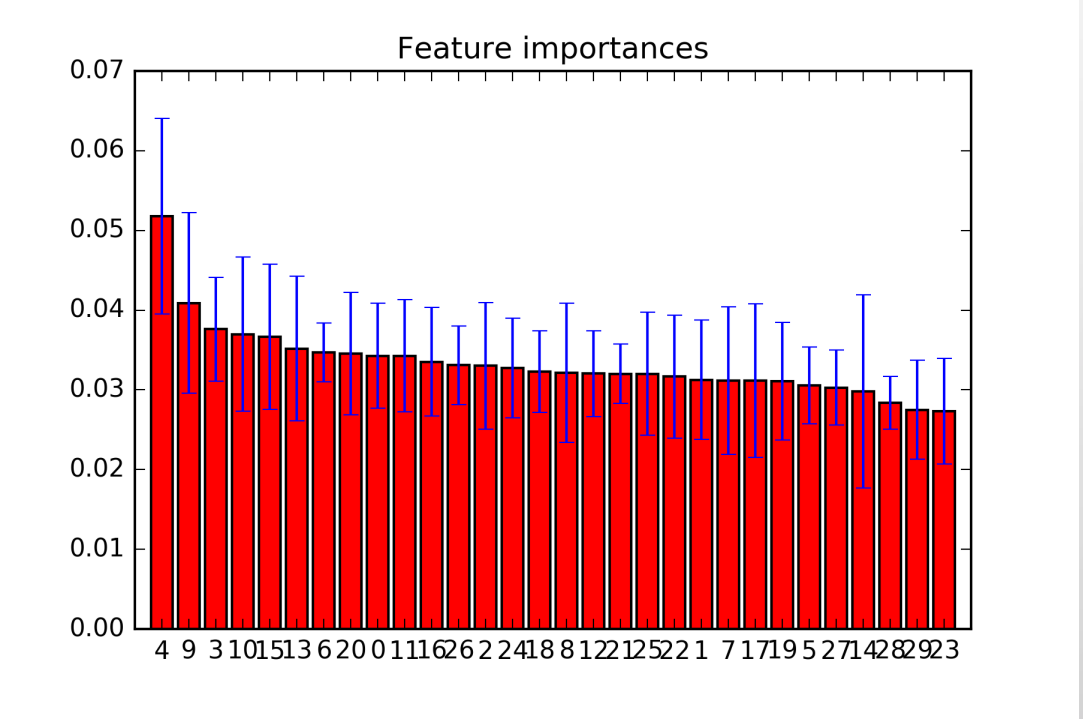
PREDICTING WITH SVM

	pred - 1	pred 0	pred 1
actual -1	398	493	435
actual 0	338	939	759
actual 1	251	483	734



PREDICTING WITH RANDOM FOREST

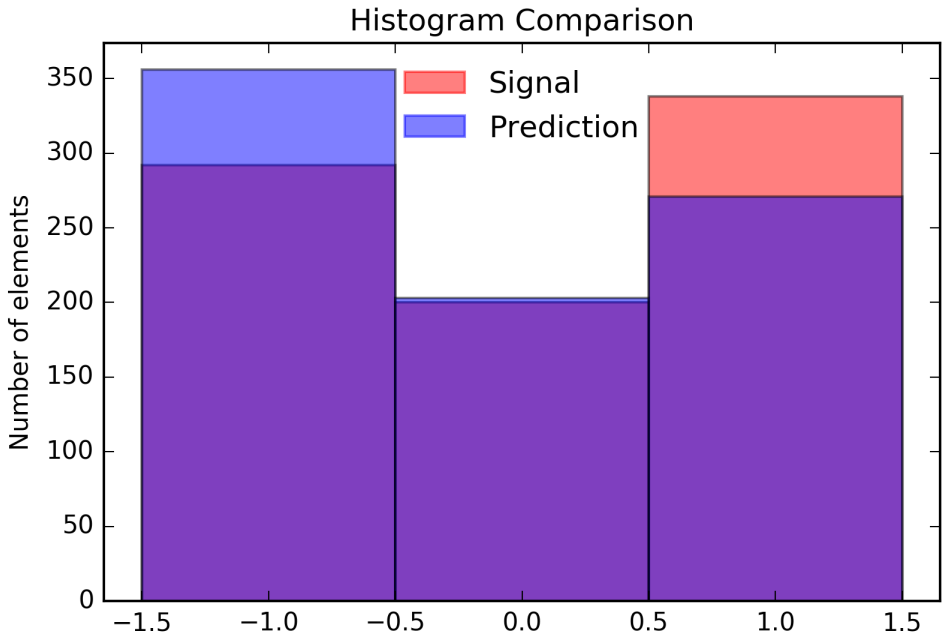
- DECISION TREE
- AVERAGE DECISION TREES TO OBTAIN RANDOM FOREST WITH LESS VARIANCE



PREDICTING WITH RANDOM FOREST

```
Pr(pred = sig) = 0.52

Pr(sig =  1 | pred =  1) = 0.58
Pr(sig =  0 | pred =  0) = 0.37
Pr(sig = -1 | pred = -1) = 0.55
```



PREDICTION WITH LINEAR MODELS - SGD

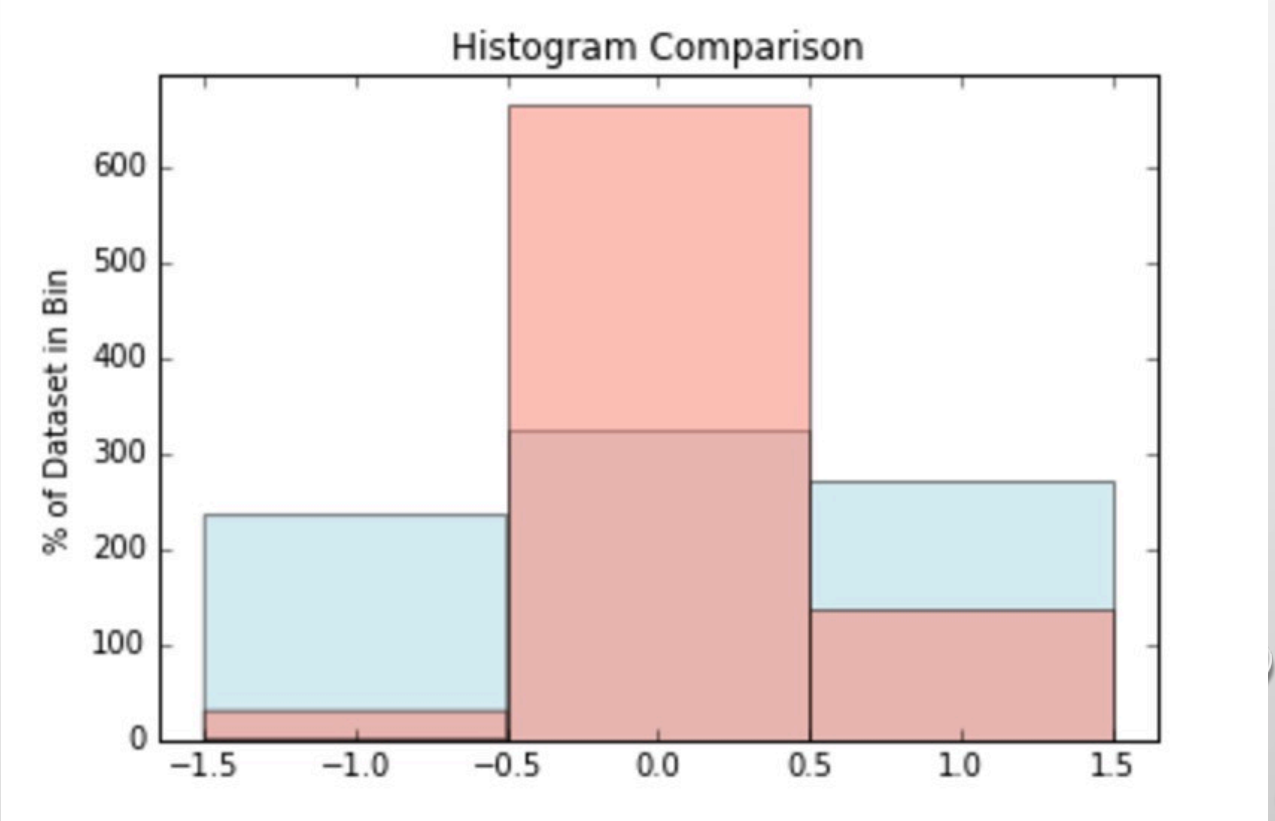
- TRAINING OUR LINEAR MODEL
  - ITERATIVELY FITS 100 LINEAR MODELS ON 4830 DATA POINTS IN OUR 15-DAY ROLLING WINDOW
  - FIRST 4000 POINTS ARE TRAINING SET // ALL OTHER POINTS ARE IN VALIDATION SET
- SGD LINEAR MODEL FROM SKLEARN IS FIT TO TRAINING DATA STOCHASTIC GRADIENT DESCENT
  - W = VECTOR OF ZEROES OF SIZE N
  - N = LENGTH OF TRAINING SET (4000)
  - THIS PROCESS IS REPEATED 500 TIMES TO FIND MIN
  - $Q_I(W)$ IS OBJECTIVE (LOG LOSS) FUNCTION
  - η IS THE LEARNING RATE (INTERNALLY OPTIMIZED)

- Choose an initial vector of parameters w and learning rate η.
- Repeat until an approximate minimum is obtained:
  - Randomly shuffle examples in the training set.
  - For $i = 1, 2, \ldots, n$, do:
    - $w := w - \eta \nabla Q_i(w)$.

PREDICTION WITH LINEAR MODELS - SGD

RESULTS

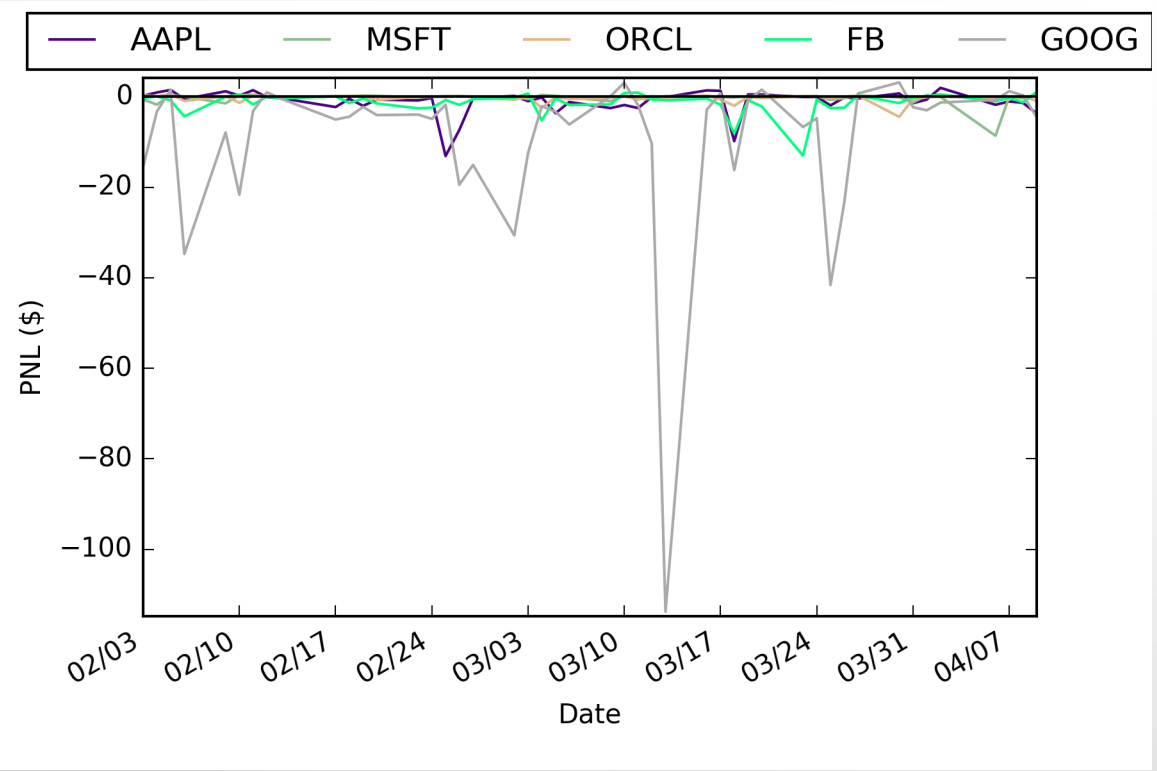
- % OF TOTAL CORRECT PREDICTIONS: 38.733
- % OF TOTAL POSITIVE PREDICTIONS: 14.033
- % OF TOTAL NEUTRAL PREDICTIONS: 82.486
- % OF TOTAL NEGATIVE PREDICTIONS: 7.183
- OVERPREDICTING THE NUMBER OF NEUTRAL SIGNALS → MISSING SOME CORRECT AND MANY NEGATIVE SIGNALS

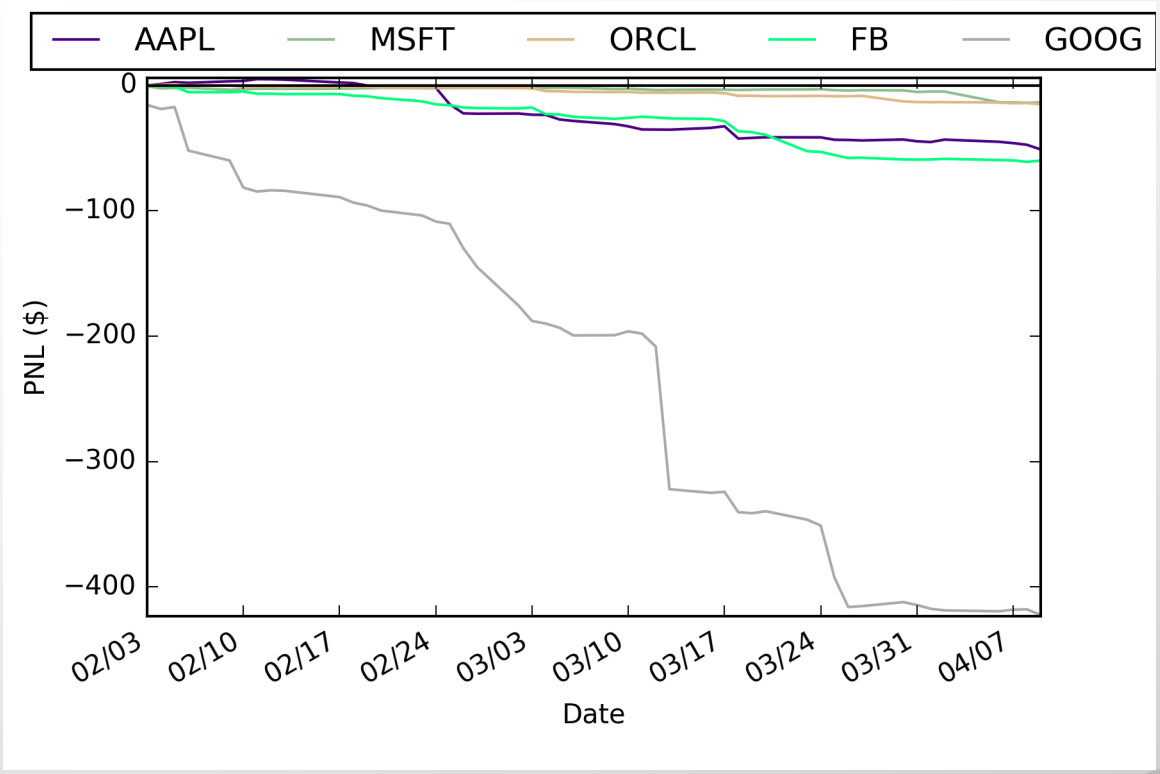


RISK MANAGEMENT

- ADVERSE SCENARIOS
  - EXTRA INVENTORY → NEED TO UNWIND POSITION AT END OF DAY AND DO NOT WANT TO LOSE MONEY DOING SO
  - WANT TO PREDICT SIGNALS TO ENSURE YOU ARE UTILIZING DIRECTIONALITY OF MARKET MOVEMENT IN BID-ASK POSTING
- STOP LOSS
  - REDUCE INVENTORY IF PROFITABLE
  - REDUCE INVENTORY IF UNREALIZED LOSS EXCEEDED A CERTAIN THRESHOLD

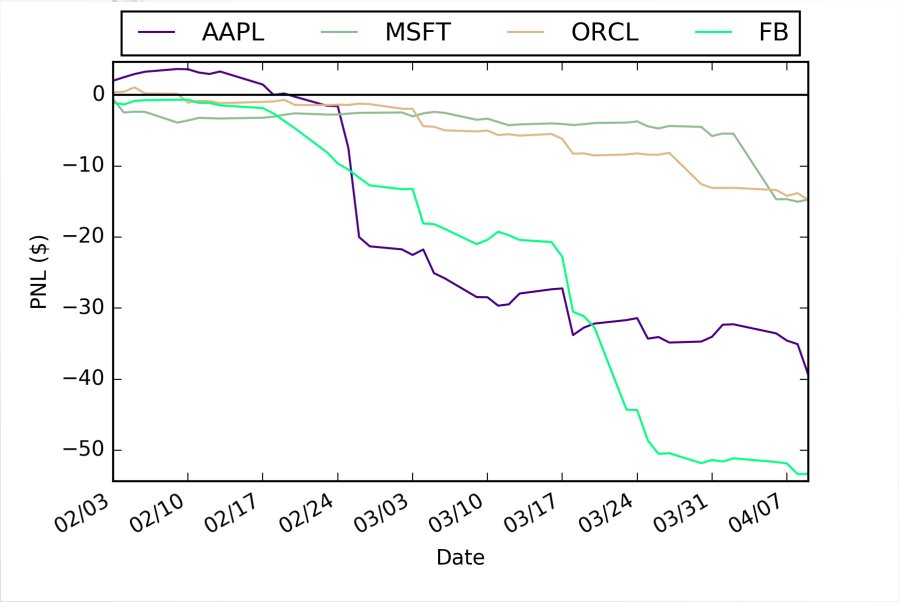
RESULTS: GOOGLE





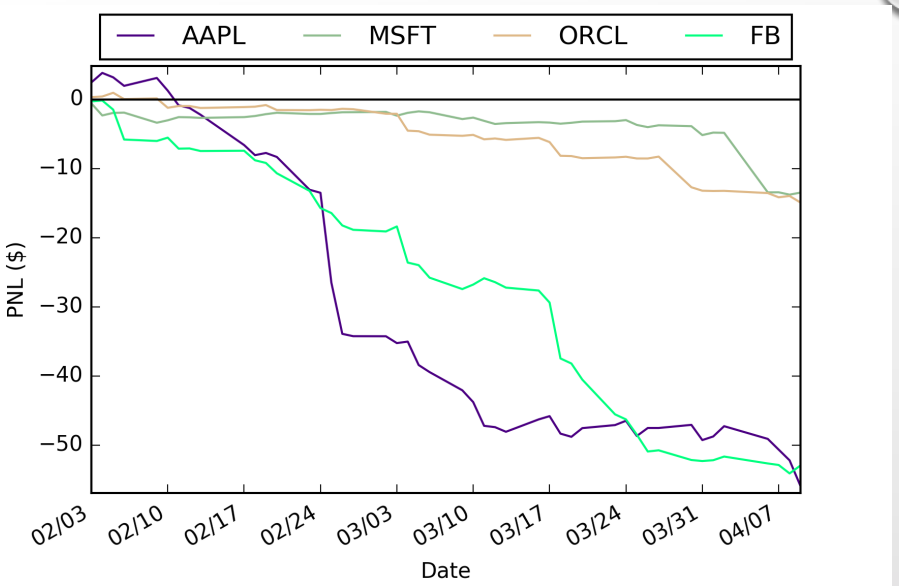
Daily and cumulative PNL with SVM

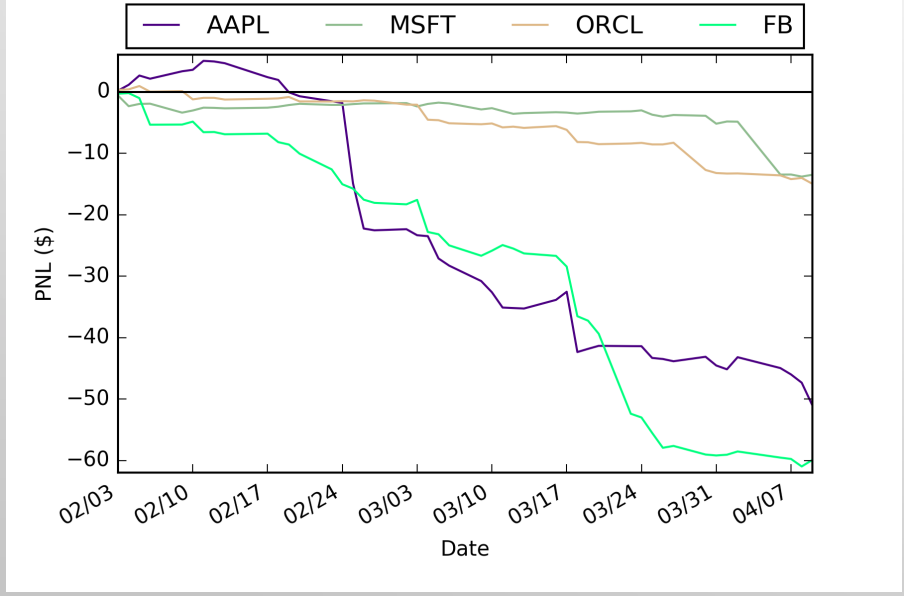
RESULTS: COMPARISON OF CUMULATIVE PNL



RF

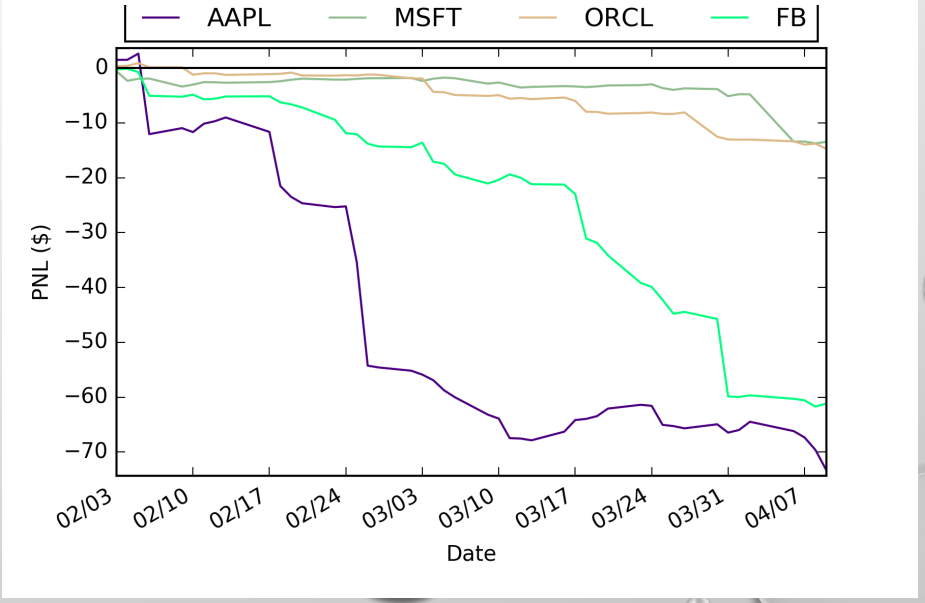
Default





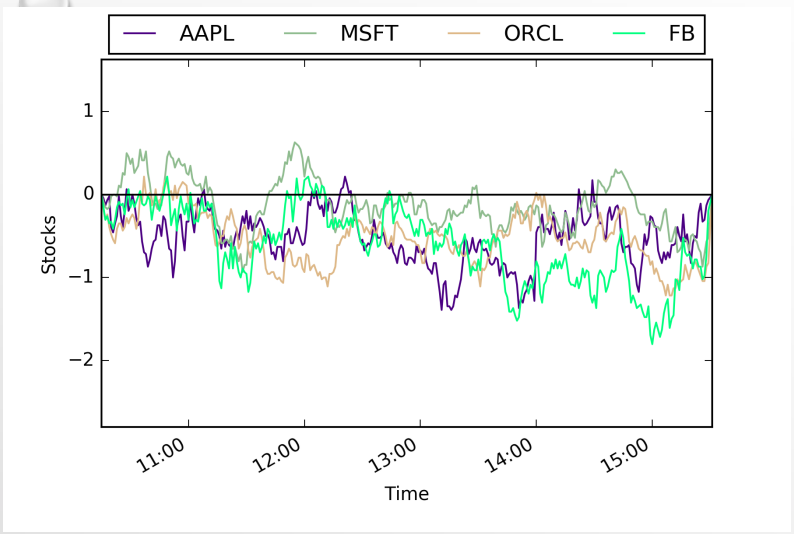
SVM

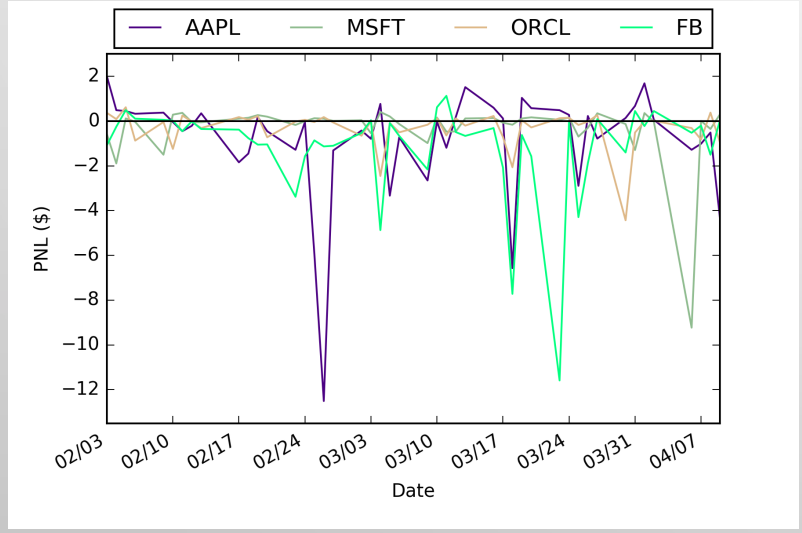
SGD

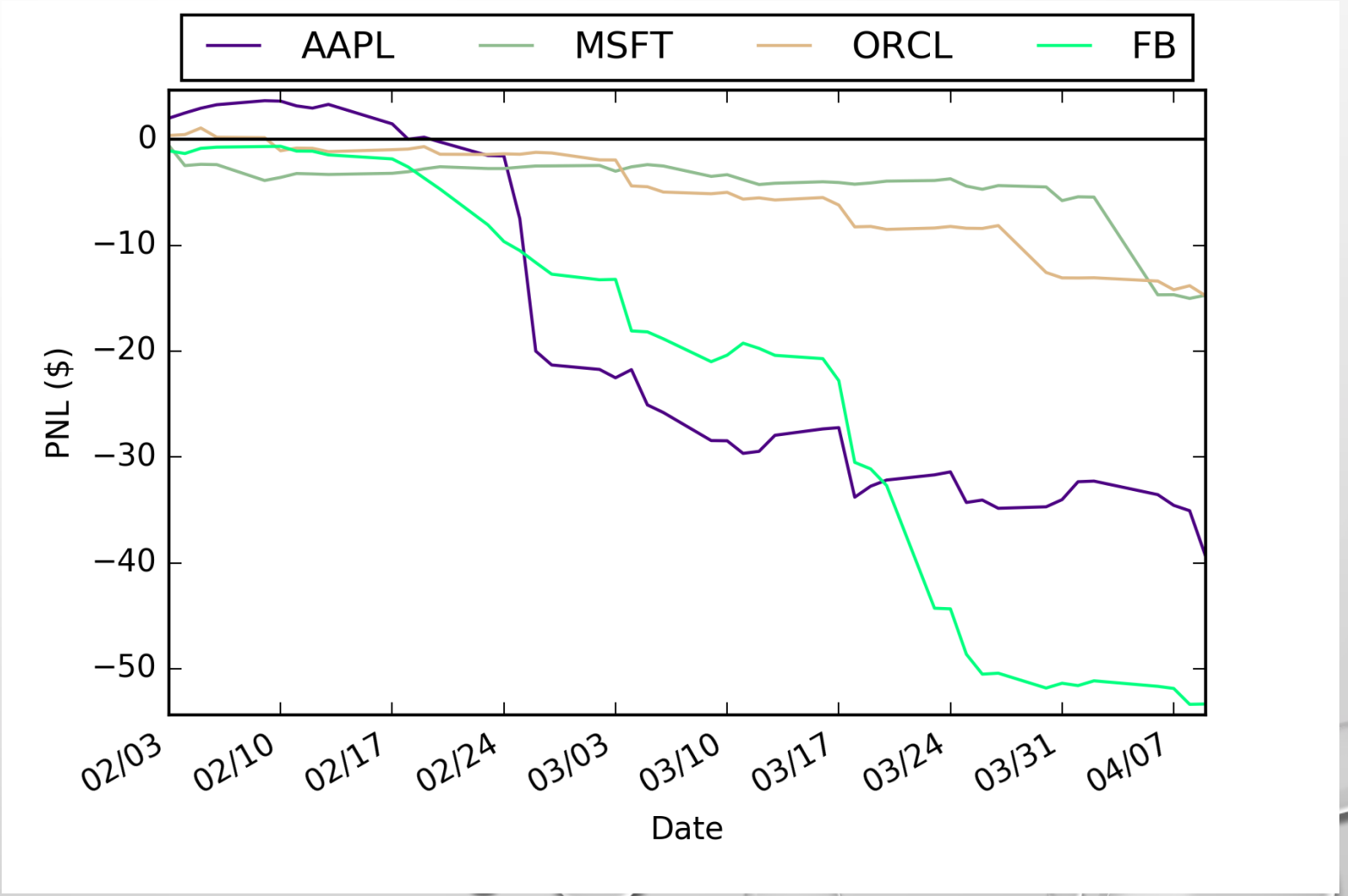


RESULTS: RF

Inventory, daily and cumulative PNL with RF







WHAT WE LEARNED

- STOP LOSS IS REALLY IMPORTANT
- DON'T ALWAYS POST AT BEST BID/ASK
- WHAT WORKS ON ONE MARKET (NIKKEI) IS NOT ALWAYS SUCCESSFUL ON ANOTHER
- THESYS IS UNRELIABLE

THANK YOU!!

- QUESTIONS?
- OUR GITHUB: HTTPS://GITHUB.COM/KKANAGAL/MSE448PROJECT
- HAVE A GREAT SUMMER