

Event-Driven Stock Prediction



By:
Xu Gao,
Scott Edenbaum

Table Of Contents

Overview

- Research
 - Reference Paper
 - Project topic
- Data
 - Web scraping
 - Data Cleaning
 - Database storage
- Model development
- Prediction and trading results
- Conclusions and insights
- Q & A



Preliminary Research

Reference Papers

- Deep Learning for Event-Driven Stock Prediction
- *Xiao Ding, Yue Zhang, Ting Liu, Junwen Duan*
- Built a neural tensor network to transform word embeddings to event embeddings to be used as the input for stock prediction neural network
- Deep Learning in Finance
- *J. B. Heaton , N. G. Polson, J. H. Witte*
- Introduced deep learning techniques used in finance
 - Eg: Shallow Factor Models, Default Probabilities, Event Studies

Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)

Deep Learning for Event-Driven Stock Prediction

Xiao Ding^{†*}, Yue Zhang[‡], Ting Liu[†], Junwen Duan[†]

[†]Research Center for Social Computing and Information Retrieval

Harbin Institute of Technology, China

{xding, tlui, jwduan}@ir.hit.edu.cn

[‡]Singapore University of Technology and Design

yue.zhang@sutd.edu.sg

Abstract

We propose a deep learning method for event-driven stock market prediction. First, events are extracted from news text, and represented as dense vectors, trained using a novel neural tensor network. Second, a deep convolutional neural network is used to model both short-term and long-term influences of events on stock price movements. Experimental results show that our model can achieve nearly 6% improvements on S&P 500 index prediction and individual stock prediction, respectively, compared to state-of-the-art baseline methods. In addition, market simulation results show that our system is more capable of making profits than previously reported systems trained on S&P 500 stock historical data.

1 Introduction

It has been shown that the financial market is “informationally efficient” [Fama, 1965] — stock prices reflect all known information, and the price movement is in response to news or events. As web information grows, recent work has applied Natural Language Processing (NLP) techniques to explore financial news for predicting market volatility.

Pioneering work mainly uses simple features from news documents, such as bags-of-words, noun phrases, and named entities [Kogan et al., 2009; Schumaker and Chen, 2009]. Although useful, these features do not capture structured relations, which limits their potentials. For example, representing the event “Microsoft sues Barnes & Noble.” using term-level features {“Microsoft”, “sues”, “Barnes”, “Noble”} alone, it can be difficult to accurately predict the price movements of *Microsoft Inc.* and *Barnes & Noble Inc.*, respectively, as the unstructured terms cannot differentiate the accuser (“Microsoft”) and defendant (“Barnes & Noble”).

Recent advances in computing power and NLP technology enables more accurate models of events with structures. Using open information extraction (Open IE) to obtain structured events representations, we find that the actor and object

*This work was done while the first author was visiting Singapore University of Technology and Design

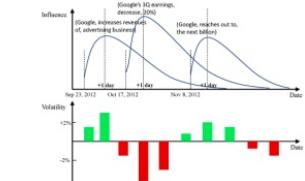


Figure 1: Example news influence of *Google Inc.*

of events can be better captured [Ding et al., 2014]. For example, a structured representation of the event above can be (*Actor* = *Microsoft*, *Action* = *sues*, *Object* = *Barnes & Noble*). They report improvements on stock market prediction using their structured representation instead of words as features.

One disadvantage of structured representations of events is that they lead to increased sparsity, which potentially limits the predictive power. We propose to address this issue by representing structured events using *event embeddings*, which are dense vectors. Embeddings are trained such that similar events, such as (*Actor* = *Nvidia fourth quarter results*, *Action* = *miss*, *Object* = *views*) and (*Actor* = *Delta profit*, *Action* = *didn't reach*, *Object* = *estimates*), have similar vectors, even if they do not share common words. In theory, embeddings are appropriate for achieving good results with a density estimator (e.g. convolutional neural network), which can misbehave in high dimensions [Bengio et al., 2005]. We train event embeddings using a novel neural tensor network (NTN), which can learn the semantic compositionality over event arguments by combining them multiplicatively instead of only implicitly, as with standard neural networks.

For the predictive model, we propose to use deep learning [Bengio, 2009] to capture the influence of news events over a history that is longer than a day. Research shows diminishing effects of reported events on stock market volatility. For example, Xie et al. [2013], Tetlock et al. [2008] and Ding et al. [2014] show that the performance of daily prediction is better than weekly and monthly prediction. As shown in Figure 1, the influences of three actual events for Google

Project Topic

- Event-driven stock prediction
 - What is it?
 - How can we use it?
- Project Motivation
 - Background in finance
 - Interest in financial modeling
 - Lack of research
 - NLP analysis



Data Sources

- Data Sources
 - News events
 - Seeking Alpha
 - Selenium
 - BeautifulSoup
 - Pricing data
 - Bloomberg terminal

Seeking Alpha 

Market News Stock Ideas Dividends Market Outlook Investing Strategy ETFs & Funds Earnings PRO

AAPL \$140.38 ▼ -0.25 (-0.18%) Get Alerts 1M followers

Apple Inc. | NASDAQ 10:58 AM 3/27/17 Bats BZX Real-Time Price

ALWAYS BE FINDING TRADE OPPORTUNITIES

Real-Time Analytics on Active Trader Pro® GET 500 FREE TRADES

Scottrade \$6.95* ONLINE EQUITY *Stocks/ETFs > \$1

1D 5D 1M 6M 1Y 5Y 10Y Advanced Chart

144.00 52wk high: 142.80
142.00 52wk low: 89.47
140.00 EPS: 8.35
138.00 PE (t/m): 16.84
Div Rate (fwd): 2.28
Yield (fwd): 1.62%
Market Cap: \$737.87B
Volume: 8,395,685

The All-New PRIUS PRIME Full efficiency. It makes the planet and your wallet smile.

Let's Go Places LEARN MORE

Ameritrade OPEN AN IRA TODAY! Plus the Satisfaction Guarantee Get the details

options4house \$4.95 Flat Rate Stocks When you may need OPEN AN ACCOUNT

Scottrade Open An Account

TradeStation TRADE NOW

Earn 1.00% APY and a \$200 bonus Start Saving

when you deposit \$10,000 or more

Capital One 360 Money Market

FDIC

LATEST | ANALYSIS | NEWS | EARNINGS | STOCKTALK | KEY DATA | FINANCIALS

All | Dividends | M&A | Press Releases | Other News

Paten case in China involving iPhone 6 ruled in favor of Apple

- Claims by Shenzhen Baili Marketing Services Co. of design infringement regarding its 100C smartphone by Apple (AAPL -0.4%) and local Chinese retailer Zoomlight had previously been in agreement with a patent regulator in Beijing, though have at this point been overturned by the Beijing Intellectual Property Court on grounds the models are distinguishable and infringement absent.
- Apple, though, had applied to deprive Baili of the involved patent, a request that has been denied by the same court.

Today, 10:21 AM | 1 Comment

Apple reported to intensify efforts on augmented reality eyewear

- An area targeted at Facebook [Oculus] (NASDAQ:FB), Microsoft [HoloLens] (NASDAQ:MSFT), Magic Leap [Private:MLEAP] and elsewhere, AR eyewear is sourced as sharpening in focus at Apple (NASDAQ:AAPL) as the technology is in view as prospectively an eventual major hardware evolution beyond the smartphone.
- Previously – Bloomberg: Apple contemplating smart glass move (Nov. 14, 2016) / Apple CEO Tim Cook perceives augmented reality a larger opportunity than virtual (Sept. 14, 2016)

Today, 9:44 AM | 13 Comments

Apple set with \$165 December 2018 target at J.P. Morgan

- Elevating from a \$142 mark prior, firm adds Apple to focus list, citing pent-up

SPONSORED FINANCIAL CONTENT

- Must-Own Stocks as Dow Rallies (Biggest Bull Market in History) Banyan Hill
- Before Applying For A Citi Card, Check If You Pre-Qualify Citi
- Here's The Safe Way To Add Consistent Income For Retirement Profits Run
- [Exclusive] Little-Known Stock Could be the Next Amazon The Motley Fool
- China stoking anger over South Korea

Data hurdles

- Finding news source
 - Content relevance
 - Content consistency
 - Historical data
- Bot & web scraping detection
- Inconsistent structure of news content
- Seemingly random XPath values



Data Scraping

- Selenium
- BeautifulSoup

```
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
import time
import csv
import random
#Import XML
from bs4 import BeautifulSoup
print "importing libraries..."

# Windows users need to specify the path to chrome driver you just downloaded.
# driver = webdriver.Chrome('path\to\where\you\download\the\chromedriver')

def seekalpha(ticker):
    file = ticker
    file_f = '../data/' + file + '.csv'
    driver = webdriver.Chrome()
    url='http://seekingalpha.com/symbol/' + file + "/news"
    print "setting up url..."
    driver.get(url)

    lenOfPage = driver.execute_script("window.scrollTo(0, document.body.scrollHeight);var lenOfPage=document.body.scrollHeight;return lenOfPage;")
    print 'defined lenOfPage'
    match=False
    i=1
    print 'opening : ' + file_f
    with open(file_f, 'w') as f:
        print file_f + ' is open'
        f.write("Date")
        f.write("Headline")
        print 'writing header...'
        while match==False and i<10000:
            soup = BeautifulSoup(driver.page_source, 'lxml')
            print 'Connecting BeautifulSoup...'
            news = soup.findAll("li", {"class": "mc_list_li"})
            print 'Grabbing all news...'
            try:
                print news[-0].contents[1].text.strip('\n')
                print news[-0].contents[3].text.strip('\n').split("\n")[0].strip().split(" \n\n")[0] + "\n" + '\n'
                #print news[-0].contents[3].contents[1].text.strip('\n').strip() + '\n'
                #print 'length of news[-0]... ' + len(news[-0].contents)
            except Exception, e:
                print "error..."
                print e
                continue
            i=i+1
            lastCount = lenOfPage
            time.sleep(1)
            lenOfPage = driver.execute_script("window.scrollTo(0, document.body.scrollHeight);var lenOfPage=document.body.scrollHeight;return lenOfPage;")
            if lastCount==lenOfPage:
                print 'x120' + '\n\n\n'
                print 'FULL LIST' + '\n\n\n'
            #for i in len(news):
            x = 0
            N = len(news)
            while (x < N):
                print news[x].contents[1].text.strip('\n')
                #print news[x].contents[3].contents[1].text.strip('\n').strip() + '\n'
                print news[x].contents[3].text.strip('\n').split("\n")[0].strip() + '\n'
                f.write(news[x].contents[1].text.encode('utf-8').strip('\n') + '\n')
                #f.write(news[x].contents[3].contents[1].text.encode('utf-8').strip('\n').strip() + '\n')
                f.write(news[x].contents[3].text.encode('utf-8').strip('\n').split("\n")[0].strip().split(" \n\n")[0] + "\n")
                x +=1
                if x == N:
                    break
            print 'x120 + '\nComplete!\n'
            print str(N) + " News entries for " + file
        
```

● * 3.3k ts-input.py Python

unix | 76:12 | Top

Data Cleaning

- Extract from data
 - Headline
 - Content
 - Date & time

Date|Headline|Mon, Mar. 13, 10:24 AM|Global miners move higher as China affirms capacity cuts

China's steel and iron ore futures resume their climb overnight, surging a respective 5.9% and 4.3%, after the head of China's Ministry of Information and Technology reaffirmed the government's resolve to cut excess capacity in the steel sector. Today's strength also is attributed to seasonal demand, as Chinese construction activity picks up after a slowdown over winter. Among relevant tickers: SLX +2.4%, BHP +2.1%, RIO +3.5%, VALE +4.1%, CLF +4%, X +3.9%, MT +4.2%, AKS +3%, WOR +2.7%, FCX +2.3%, TECK +2.6%, AA +2.4%, CENX +1.7%, ACH +1.2%, KALU +1%. |Mon, Mar. 6, 8:53 AM|Alcoa upgraded to Buy at Goldman Sachs

Alcoa (NYSE:AA) +1.5% premarket as Goldman Sachs upgrades shares to Buy from Neutral and raises its six-month price target \$52 from \$34, TheFly.com reports. Goldman says China's capacity cuts are improving industry dynamics, and the firm expects AA to generate a 12% free cash yield on average through 2018 amid higher aluminum and alumina prices. |Mon, Mar. 6, 8:14 AM|China proposes capacity cuts in steel, coal

China says it will cut steel capacity by 50M metric tons and coal output by more than 150M metric tons in 2017, according to a report issued at the weekend's National People's Congress. The expected cuts come as the government also reduces its annual economic growth target to ~6.5% vs. last year's 6.5%-7% range. The government also will target cuts in energy consumption per unit of gross domestic product by 3.4% and in carbon intensity by 4% this year. No mention was made of aluminum cuts. Potentially relevant tickers include: SLX, X, MT, AKS, NUE, STLD, BHP, RIO, AA, CENX, ACH |Thu, Mar. 2, 10:06 AM|Alcoa combines business units, names Reyes to run aluminum business

Data Pre-Processing

ReVerb

- Java tool for text processing
 - identifies “similar” words
- Identifies & extracts binary relationships in English sentences
- Designed for use with “Web-scale” information extraction
- Eg: “is” -> “be”, “was” -> “be”
 - Conceptually similar to nltk “stemming”
- Each news item/event is a separate input for ReVerb

ReVerb Results

ReVerb Sample Output

- 1 *China 's steel and iron ore futures resume their climb overnight , surging a respective 5.9 % and 4.3 % , after the head of China's Ministry of Information and Technology reaffirmed the government's resolve to cut excess capacity in the steel sector .* NNP POS NN CC NN NN NNS VBP PRP\$ NN RB , VBG DT JJ CD NN CC CD NN , IN DT NN IN NNP NNP NNP IN NNP CC NNP VBD DT NN VBD NN TO VB JJ NN IN DT NN NN .
B-NP I-NP I-NP I-NP I-NP I-NP B-VP B-NP I-NP B-ADVP O B-VP B-NP I-NP I-NP I-NP O B-NP I-NP O B-PP B-NP I-NP B-VP B-NP I-NP B-VP B-NP B-VP I-VP B-NP I-NP B-PP B-NP I-NP I-NP O china 's steel and iron ore futures resume their
- Not human readable

ReVerb Code

Reverb is slow

- Tweaked code for concurrent data processing

```
from sentiment import parse_1

def reverb():
    ticker, df=parse_1()
    news=df.content
    confidence=[]
    subject=[]
    object_=[]
    verb=[]

    single_news = "single_news_" + ticker + ".txt"
    ticker_output = "output_" + ticker + ".txt"
    for each_news in news:
        news=each_news+'.'
        print news
        f=open(single_news,'w')
        f.write(news)
        f.close()
        javaosarg = "java -Xmx512m -jar reverb-latest.jar " + single_news + " > " + ticker_output
        os.system(javaosarg)
        output=open(ticker_output,'r')
        info=output.read()
        output.close()
        print "reverb finished, next parse"
        confid,subj,obje,verb_=reverb_parse(info)
        confidence.append(confid)
        subject.append(subj)
        object_.append(obje)
        verb.append(verb_)

    osrmarg1 = "rm " + single_news
    osrmarg2 = "rm " + ticker_output
    os.system(osrmarg1)
    os.system(osrmarg2)
    df['confidence']=confidence
    df['subject']=subject
    df['object']=object_
    df['verb']=verb
    file_='../data/reverb/'+ticker+'reverb.csv'
    df.to_csv(file_)

def reverb_parse(info):
    reverb_list=info.split('\n')
    confi=[]
    sub=[]
    obj=[]
    verb=[]
    if len(reverb_list)==0:
        confidence=0
        subject=''
        verb=''
        object_=''
    else:
        for i in range(len(reverb_list)-1):
            buf=reverb_list[i].split('\t')
            confi.append(float(buf[11]))
            sub.append(buf[-3])
            verb.append(buf[-2])
            obj.append(buf[-1])
        if len(confi)==0:
            confidence=0
            subject=''
            verb=''
            object_=''
        else:
            idx=np.argmax(confi)
            confidence=confi[idx]
            subject=sub[idx]
            object_=obj[idx]
            verb_=verb[idx]
            print "finish this news"
    return confidence,subject,object_,verb_
```

ReVerb

- Processing the data can be quite time consuming
 - Solution – concurrent instances to speed up data processing

```
reverb finished, next parse
finish this news
FDA OKs Lilly and Boehringer's extended-release combo pill for T2D

The FDA approves once-daily Jentadueto XR (linagliptin and metformin hydrochloride extended-release) tablets for the treatment of adults with type 2 diabetes (T2D). It is indicated as an adjunct to diet and exercise to improve glycemic control in adult type 2 diabetics where the use of both drugs is appropriate. The product is the seventh new treatment approved in the U.S. from the 2011 diabetes alliance between Eli Lilly (LLY +0.7%) and Boehringer Ingelheim. The FDA approved the original formulation of Jentadueto in January 2012.

Initializing ReVerb extractor...Done.
Initializing confidence function...Done.
Initializing NLP tools...Done.
Starting extraction.
Extracting from single_news_LLY.txt
Done with extraction.
Summary: 4 extractions, 4 sentences, 1 files, 1 seconds
reverb finished, next parse
finish this news
Lilly sees as many as 20 new product launches by 2023

In a presentation to investors, Eli Lilly (NYSE:LLY) says it has the potential to launch 20 new products in the period 2014 - 2023, and an average of two new indications for already-approved products per year during the same period. In R&D, the company invests in areas such as immunotherapy, neurodegeneration and pain. In diabetes, the company targets glucose control, metabolic control and organ protection. In other tumor cells, PF-06438179 (infliximab-Pfizer) in patients with rheumatoid arthritis (RA) met its primary endpoint demonstrating equivalent efficacy with Remicade as measured by the number of participants achieving ACR20 (20% improvement in RA symptoms), as many as 11 by the end of 2018. Immunology key events: Taltz (ixekizumab) recently launched, baricitinib (under review) at Week 14 of treatment. According to ClinicalTrials.gov, the estimated study completion date is May 2017. A replay of the webcast will be available for 90 days on the company's investor website.

Done with extraction.
Summary: 1 extractions, 2 sentences, 1 files, 1 seconds
reverb finished, next parse
finish this news
Alere voluntarily withdraws INRatio and INRatio2 PT/INR Monitoring System from U.S. market

Following a collaborative process with the FDA, point-of-care diagnostic firm Alere (NYSE:ALR) will be initiating a voluntary withdrawal of its Alere INRatio and INRatio2 PT/INR Monitoring System. In December 2014, the company initiated a voluntary correction to inform users of the device that patients with certain conditions should not be tested with the device due to the potential misreporting of results. Over the past two years, Alere updated the software to address the issue. At the end of 2015, the FDA notified the company that it did not wish this news not agree that the problem was fixed and advised the company to submit a plan for the voluntary removal of INRatio from GE Energy Connections' portfolio. Microsoft's Martin as chief digital officer in the market.

Alere is working with the FDA to determine the most appropriate timing for product discontinuation and will provide guidance on transitioning patients to an alternate solution to allow them to continue anti-coagulation monitoring in the least disruptive manner possible. Initializing ReVerb extractor...Done.
Initializing confidence function...Done.
Initializing NLP tools...Done.
Starting extraction.
Extracting from single_news_ABV.txt
Done with extraction.
Summary: 4 extractions, 5 sentences, 1 files, 1 seconds
reverb finished, next parse
FDA approves Abbott's absorbable stent for coronary artery disease

The FDA approves Abbott Vascular's (ABV +0.3%) Absorb GT1 Bioresorbable Vascular Scaffold System, the first full absorbable stent approved for sale in the U.S. to treat coronary artery disease. The GT1 releases everolimus (Novartis' (NVS -1.1%) Afatinib) which limits the growth of scar tissue. It is gradually absorbed by the body in about three years.

Initializing ReVerb extractor...Done.
Initializing confidence function...[]

finish this news
Geron inks license deal with Janssen Pharma valued up to $80M

Geron (GERN +6.4%) heads north on increased volume in response to its announcement of a license deal with Janssen Pharmaceutica. The company has granted Janssen exclusive global rights to develop and commercialize products based on Geron's specialized oligonucleotide backbone chemistry in addition to RNA interference candidates for the treatment of all human disorders except blood and bone marrow cancers and exclusive of telomerase inhibitors. Geron has also granted Janssen a non-exclusive global license to develop and commercialize oligonucleotides based on Geron's patents covering the synthesis of monomers (building blocks of oligonucleotides). Janssen Biotech has non-exclusive rights to the same intellectual property under their November 2014 imetelstat agreement. Under the terms of the deal, Geron will receive an upfront payment of $5M, up to $75M in milestones and low-single-digit royalties on global net sales of each licensed product.

Initializing ReVerb extractor...Done.
Initializing confidence function...Done.
Initializing NLP tools...Done.
Starting extraction.
Extracting from single_news_JNJ.txt
Done with extraction.
Summary: 6 extractions, 4 sentences, 1 files, 2 seconds
reverb finished, next parse
Pfizer's Remicade biosimilar candidate successful in late-stage study

In a presentation to investors, Eli Lilly (NYSE:LLY) says it has the potential to launch 20 new products in the period 2014 - 2023, and an average of two new indications for already-approved products per year during the same period. In R&D, the company invests in areas such as immunotherapy, neurodegeneration and pain. In diabetes, the company targets glucose control, metabolic control and organ protection. In other tumor cells, PF-06438179 (infliximab-Pfizer) in patients with rheumatoid arthritis (RA) met its primary endpoint demonstrating equivalent efficacy with Remicade as measured by the number of participants achieving ACR20 (20% improvement in RA symptoms), as many as 11 by the end of 2018. Immunology key events: Taltz (ixekizumab) recently launched, baricitinib (under review) at Week 14 of treatment. According to ClinicalTrials.gov, the estimated study completion date is May 2017. A replay of the webcast will be available for 90 days on the company's investor website.

Done with extraction.
Summary: 1 extractions, 2 sentences, 1 files, 1 seconds
reverb finished, next parse
finish this news
Novartis' (NYSE:NVS) Sandoz unit has commercialization rights to PF-06438179 in the European Economic Area (28 countries) while Pfizer retains the rights elsewhere.

Initializing ReVerb extractor...[]

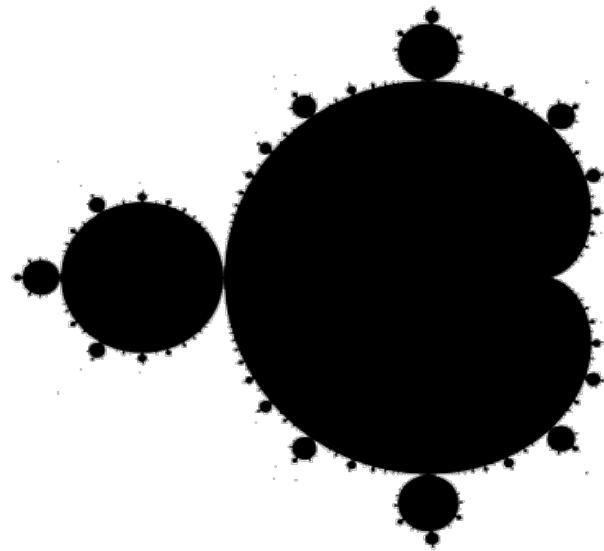
years at GE Aviation, where he served since 2014 as president and CEO of GE Aviation Services; he had been VP and general manager of GE's global sales and marketing since 2008. .
Initializing ReVerb extractor...Done.
Initializing confidence function...Done.
Initializing NLP tools...Done.
Starting extraction.
Extracting from single_news_GE.txt
Done with extraction.
Summary: 3 extractions, 3 sentences, 1 files, 1 seconds
reverb finished, next parse
finish this news
The revival of supersonic transport?

Honeywell (NYSE:HON) has agreed to supply avionics for a proposed supersonic jetliner - intended to carry only premium passengers - that could cruise 10% faster than the now extinct Concorde. The demonstrator vehicle, called Baby Boom, is slated to take to the air in 2017. If development goes as planned, the full-size version, including engines manufactured by General Electric (NYSE:GE), could start carrying passengers as early as next decade.

Initializing ReVerb extractor...Done.
Initialization confidence function...[]
```

Data Transform

- News and text data have many possible transformations to assist in modeling/analysis
 - Sentiment Analysis with TextBlob
 - Word Embedding with Word2Vec
 - Document Embedding with Doc2Vec
 - Event Embedding



TextBlob



 word2vec
Tool for computing continuous distributed representations of words.

Data Storage

Stock Pricing Data

- MySQL DB

```
mysql> describe common_data;
+-----+-----+-----+-----+
| Field | Type      | Null | Key | Default | Extra |
+-----+-----+-----+-----+
| SPX   | decimal(19,10) | YES |   | NULL    |       |
| NYSEVOL | decimal(19,10) | YES |   | NULL    |       |
| CCMP  | decimal(19,10) | YES |   | NULL    |       |
| date   | date        | NO  |   | NULL    |       |
+-----+-----+-----+-----+
4 rows in set (0.00 sec)

mysql> describe stock_data;
+-----+-----+-----+-----+
| Field | Type      | Null | Key | Default | Extra |
+-----+-----+-----+-----+
| ticker | varchar(5) | NO  |   | NULL    |       |
| close  | decimal(19,4) | YES |   | NULL    |       |
| date   | date        | NO  |   | NULL    |       |
+-----+-----+-----+-----+
3 rows in set (0.00 sec)
```

```
#####
# update the db_host, db_user, db_name, db_pass values
# db_get_all only grabs data for stocks included in ../data/Price
# db_get_all outputs a dictionary of dataframes named by the stock ticker (capital)
# eg:
# stocks = db_get_all()
# stocks['AA'] -> dataframe for AA
# stocks['AA']['close'][0] -> 81.2958
# stocks['AA']['date'][0] -> datetime.date(2008, 2, 21)

import MySQLdb as mdb
import pandas as pd
import os
import datetime as dt
from pandas.io import sql

def db_get_ticker(ticker):
    t = ticker
    cols = ['ticker', 'closer', 'date']

    db_host = '216.230.228.88:3306'
    db_user = 'bc8_scottede'#scottedb'
    db_name = 'securities'
    db_pass = '#####'
    con = mdb.connect(host = db_host, user = db_user, passwd = db_pass, db=db_name)
    con = mdb.connect(host = db_host, user = db_user, db=db_name)
    query = ("SELECT * FROM stock_data WHERE ticker = '%s'" % (t))
    df = sql.read_sql(query, con=con)
    print "grabbing stock data for %s" % (t)
    return (df)

def db_get_all():
    stocks = []
    pricedir = '../data/Price/' #directory
    for file_ in set(os.listdir(pricedir)):
        try:
            g = len(file_)
            if g < 10:
                stocks.append(str(file_).split(".")[0])
        except Exception, e:
            print e
            continue
    print "getting all stock data: "
    [stock for stock in stocks]
    stock_data = {stock: db_get_ticker(stock) for stock in stocks}

    return (stock_data)
```

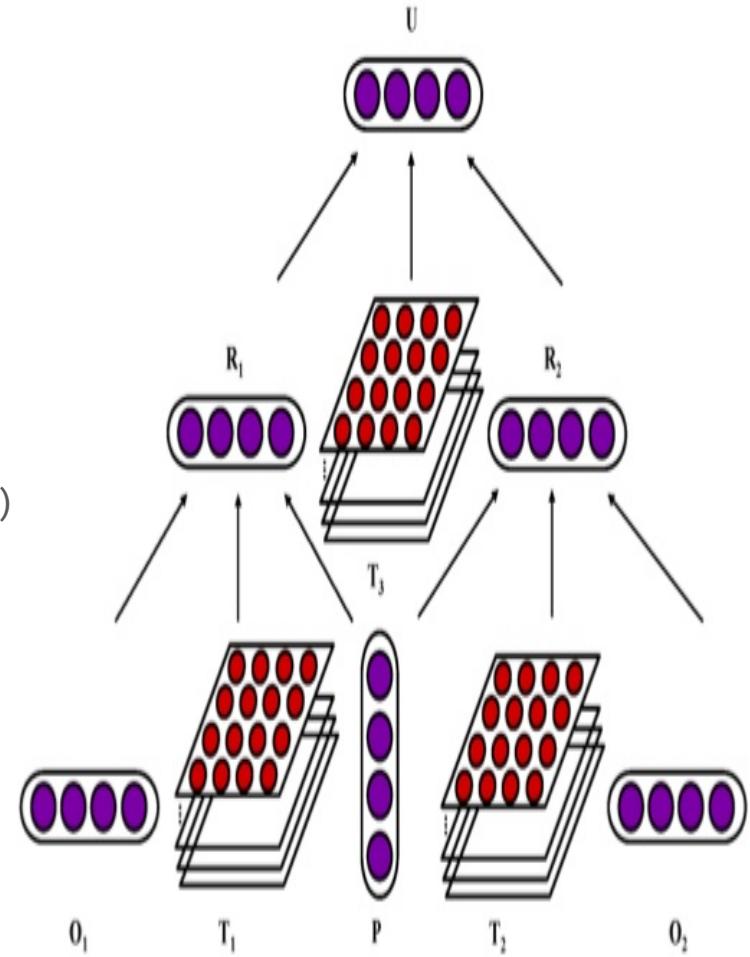
● * 1.5k dbgrab.py Python

unix | 9:55 All

Model Development

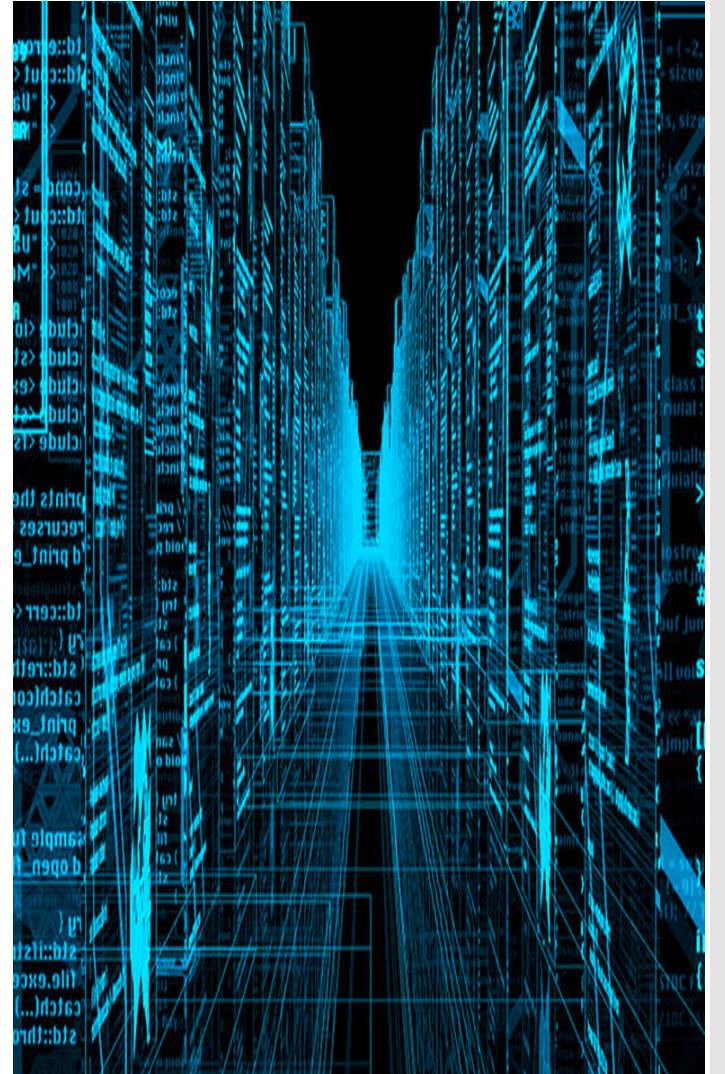
Foundational model

- Neural Network
 - 2 hidden layers
 1. 300 nodes, activation function: $\tanh(x)$
 2. 50 nodes, activation function: $\tanh(x)$
 - Output layer
 - Single node
 - Binary result with Sigmoid activation function



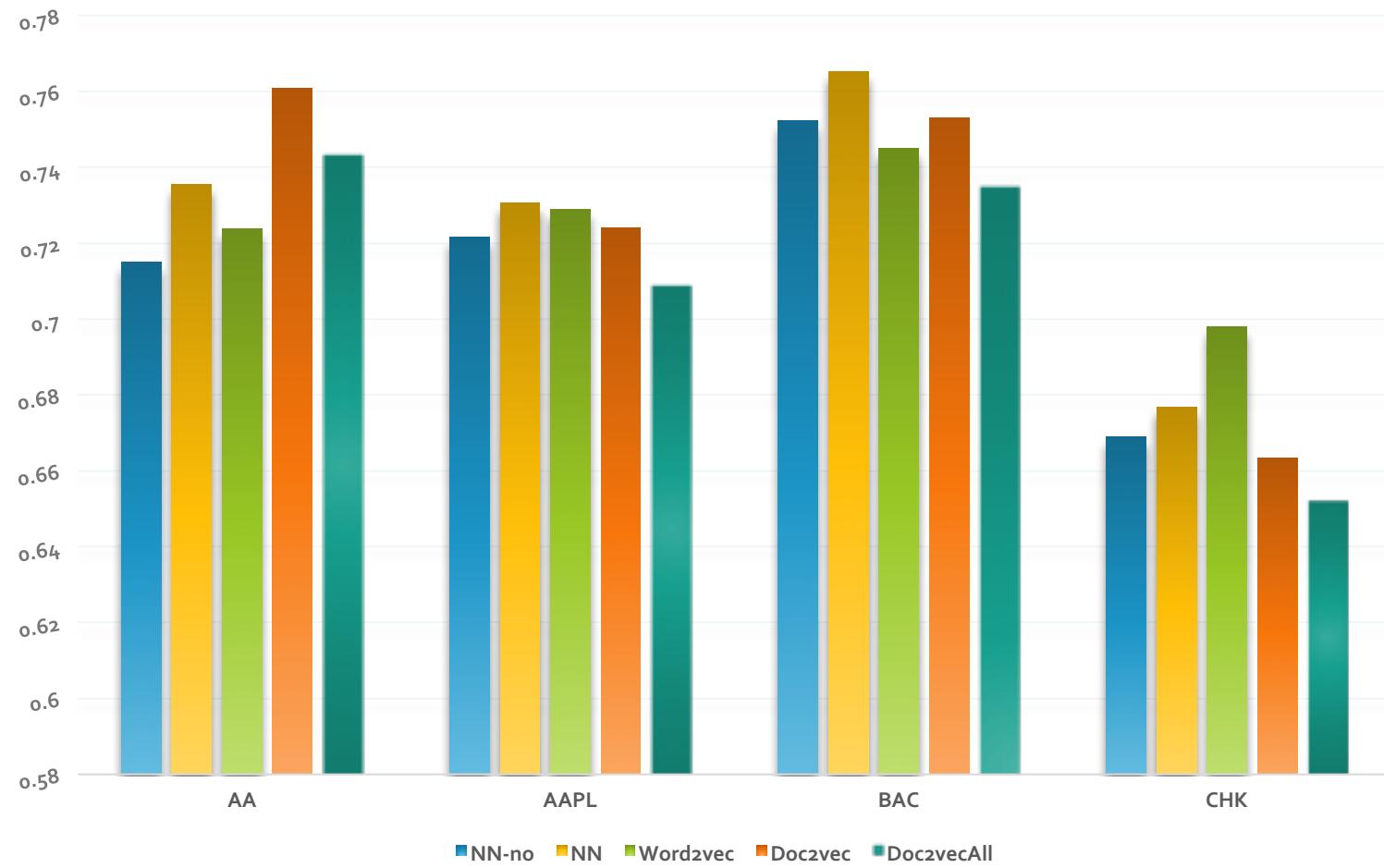
Dataset Development

1. Market Data
 - 5 day window
 - S&P 500, NASDAQ Composite, NYSE Volume
2. Sentiment polarity
 - Generated with TextBlob
3. Word Embeddings
 - Uses output from ReVerb
 - Grabs sentence with highest confidence level
 - Subject, object, and verb 'tuple' forms word embeddings
 - Downsides: - loss of information through selection process
4. Doc2vec
 - Converts paragraph to single vector
 - Use vector as input for Neural Network
 - Generate training dataset for each stock and combines
 - Downsides: - uses all info, does not filter out irrelevant text



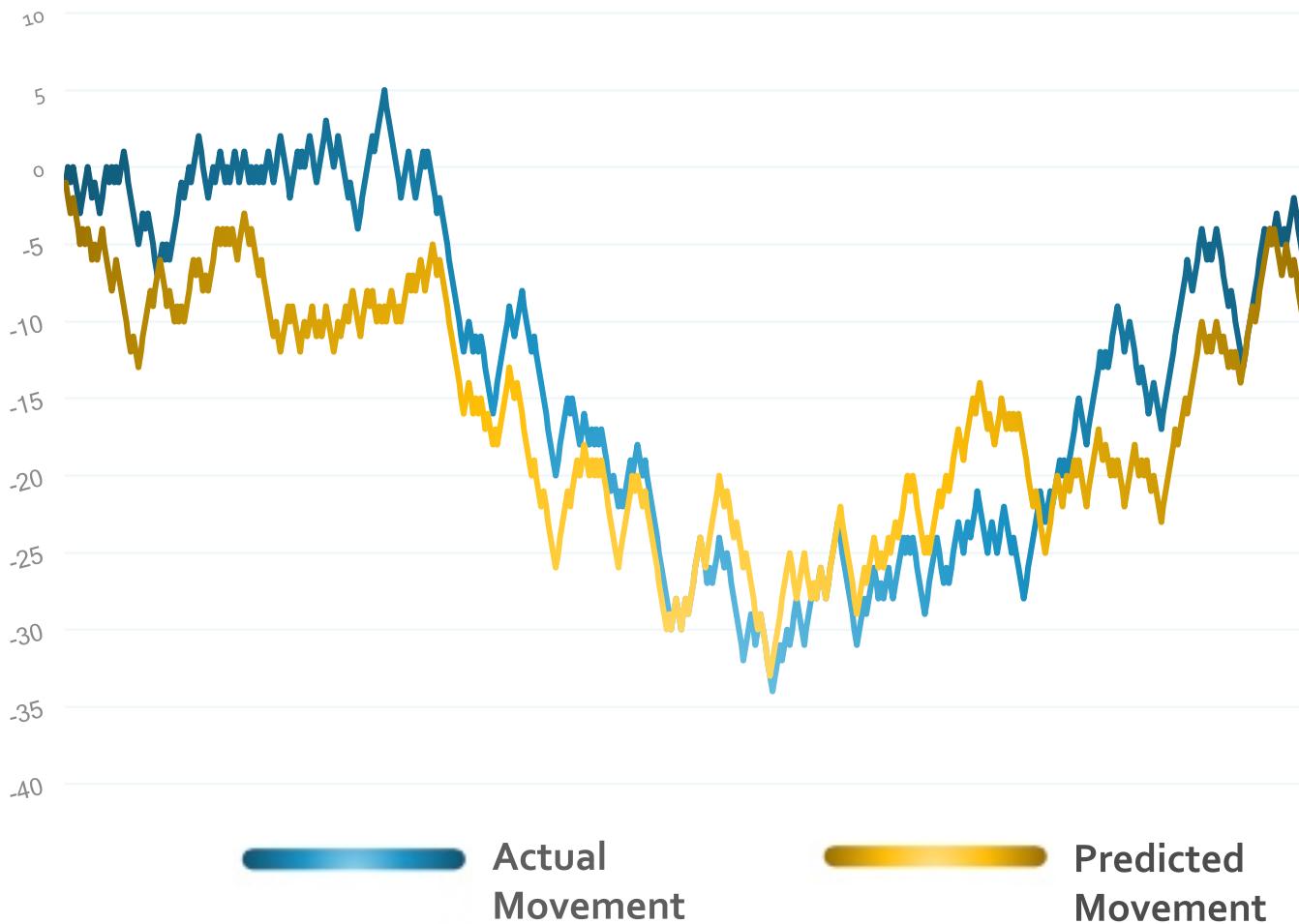
Results

Accuracy With Different Models



Results: AA

AA: Stairs for prediction

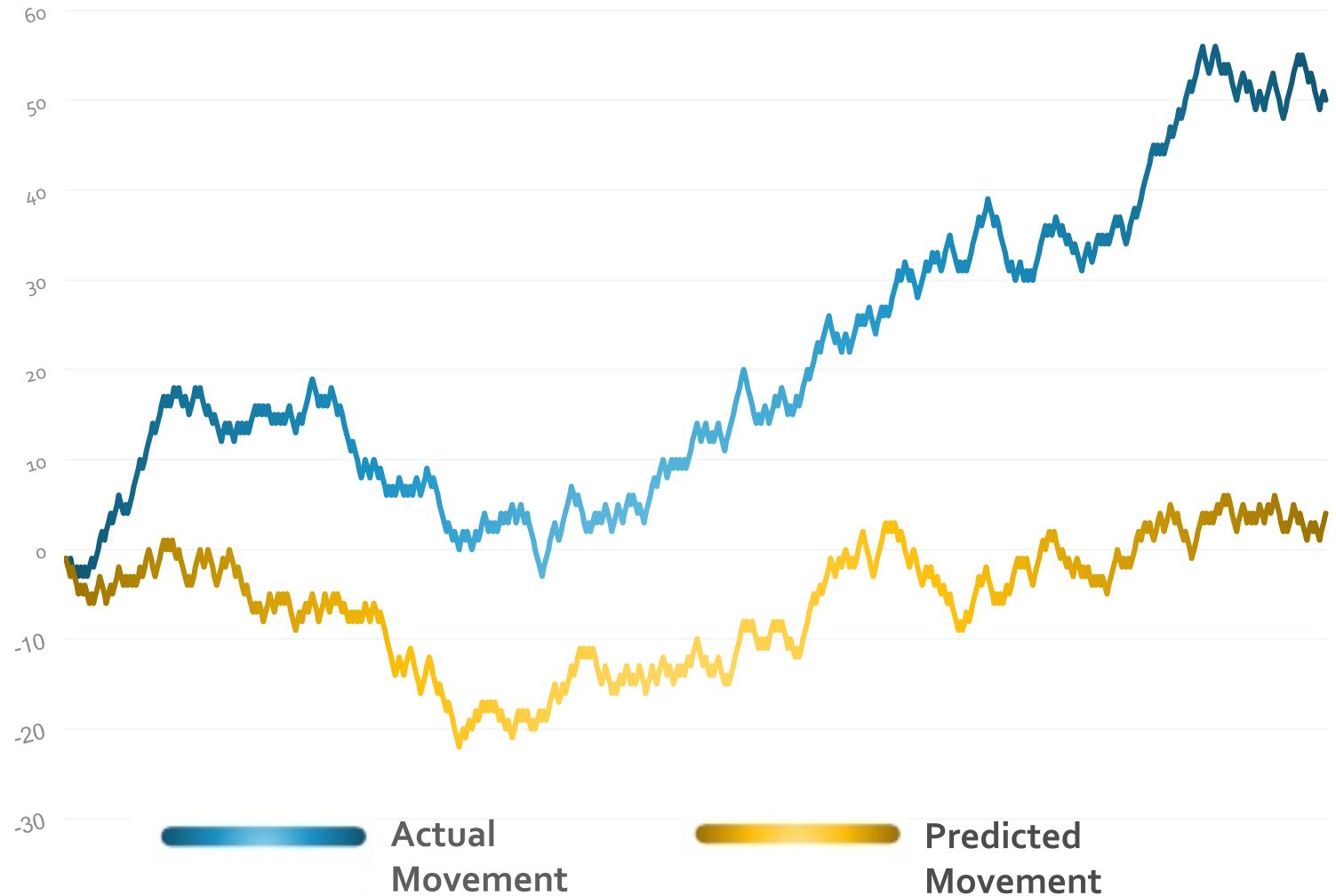


Results:
AA

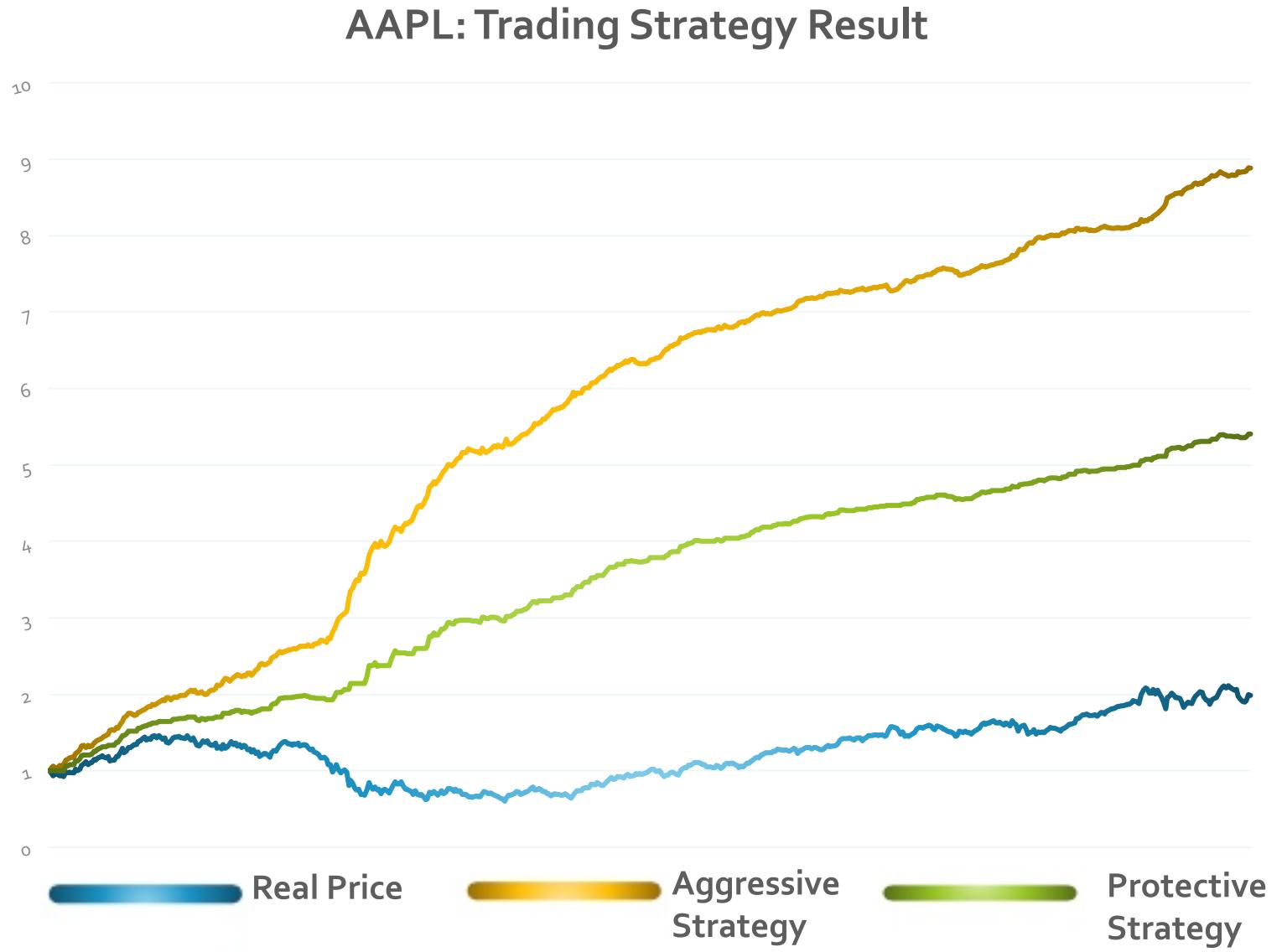


Results: AAPL

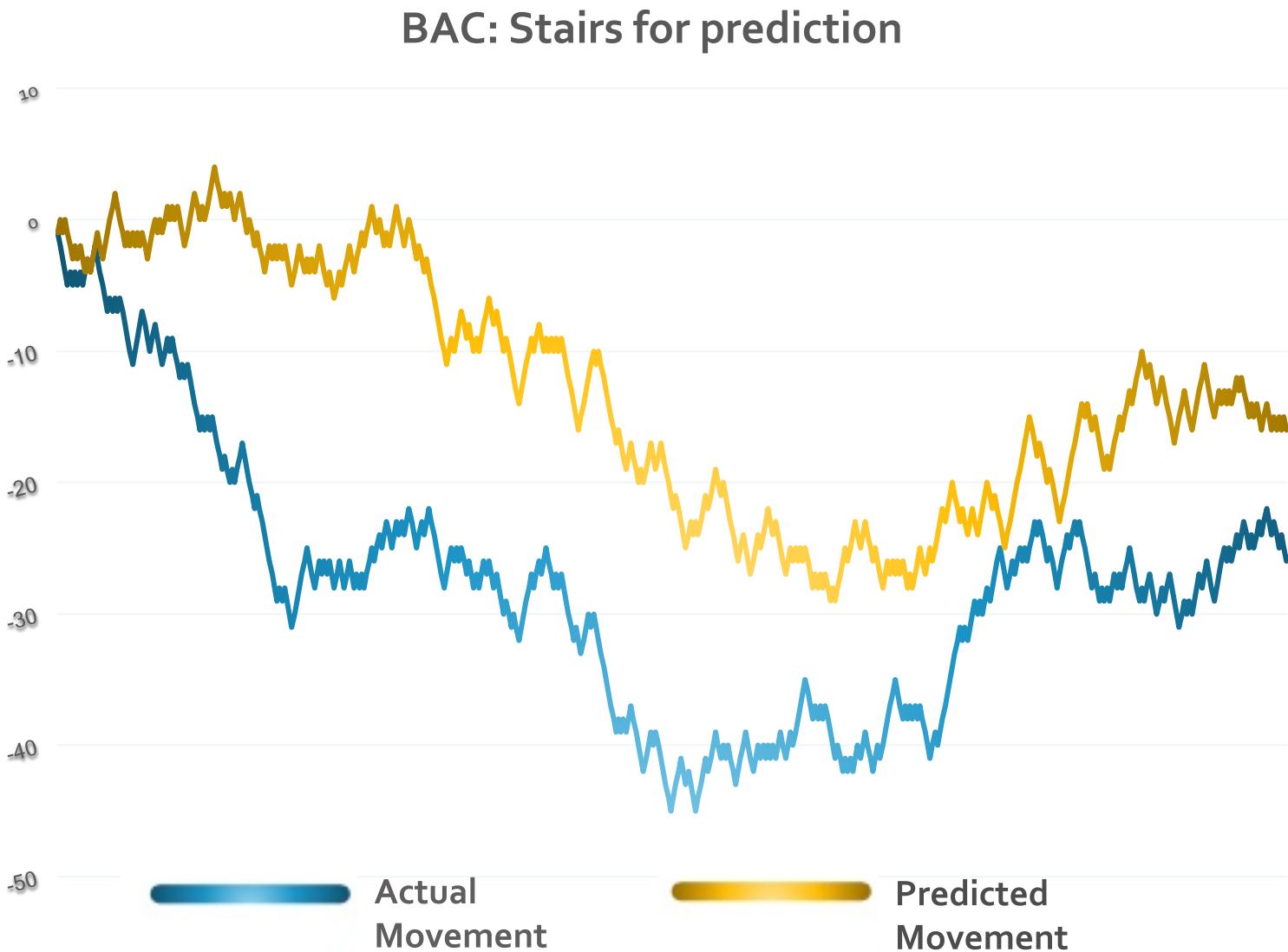
AAPL: Stairs for prediction



Results: AAPL

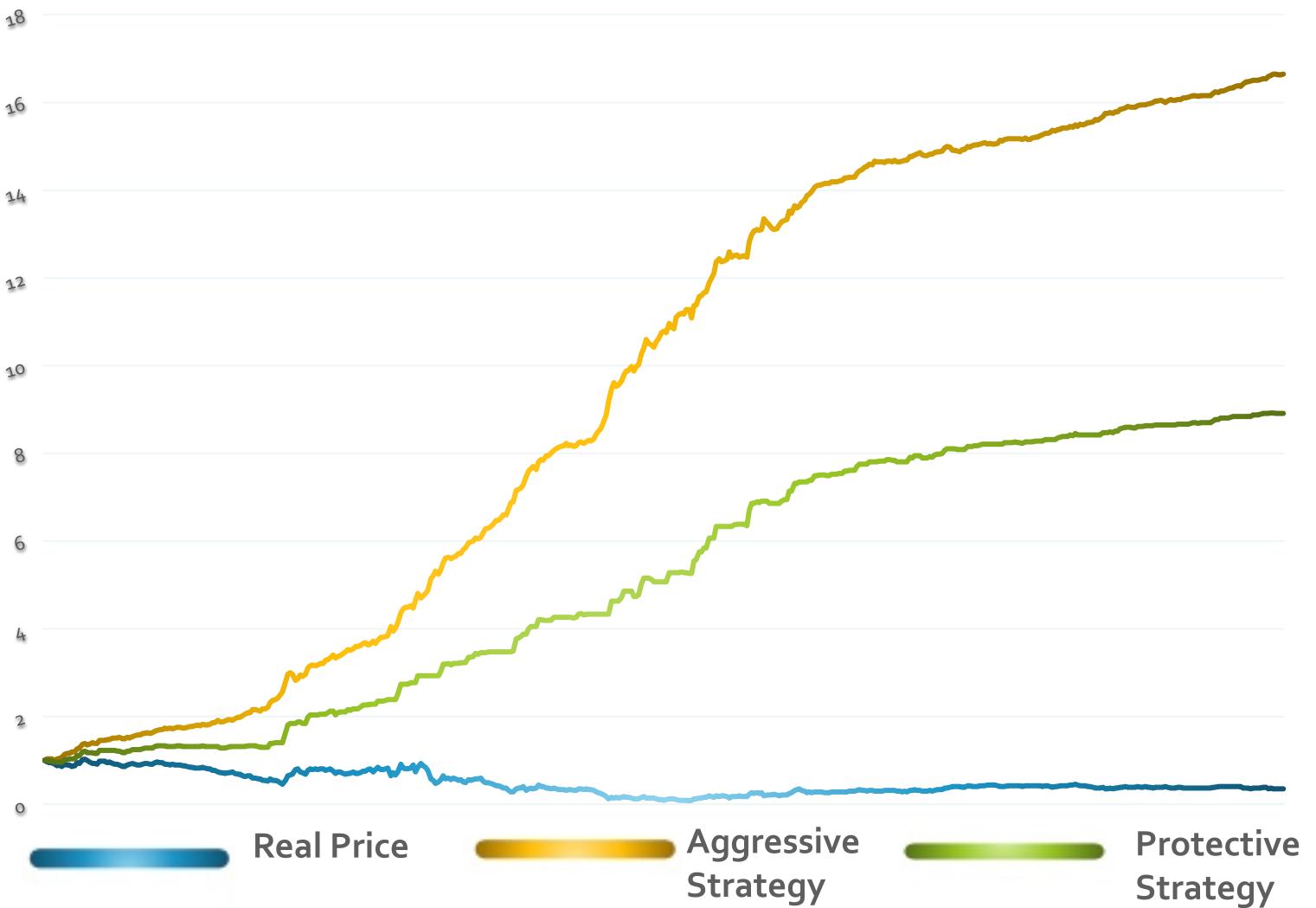


Results: BAC

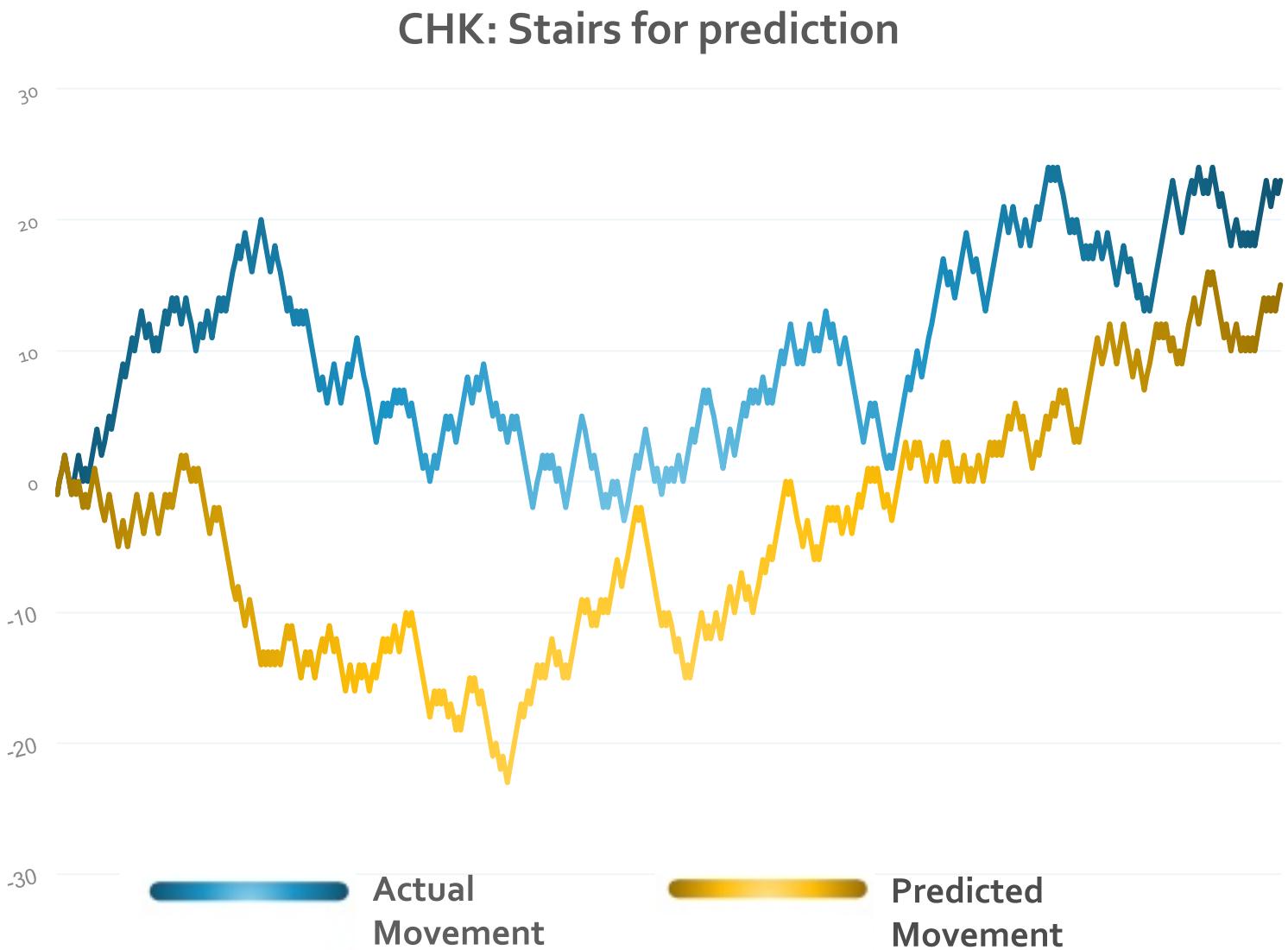


Results: BAC

BAC: Trading Strategy Result



Results: CHK



Results: CHK

CHK: Trading Strategy Result



Conclusions and Insights

- Accurate historical stock news is hard to acquire!
- Each 'model' has strengths & weaknesses
 - Model accuracy is for relevance of text content to sentence subject
 - Does not imply impact of news event
- Our pricing model naively assumes stocks open at previous close value
- Model only predicts direction of price movement, not magnitude
- This is a new approach to financial modeling is bound to grow in popularity as tools for text processing grow and historical data become more available
- A next step would be to categorize events
 - Ie: including a corporate events overlay to better identify scheduled corporate actions

Q&A

Growing Popularity of Event-Driven Strategies

- 19% increase in net demand for Event-Driven strategies, according to 2014 Credit Suisse survey
- Event-Driven strategies provide an alternative investment strategy with low correlations to economic and market trends