Alternative Data Sources for Measuring Market Sentiment and Events (Using R)

Joe Rothermich, CFA

Natural Selection Financial, Inc.

R/Finance 2011
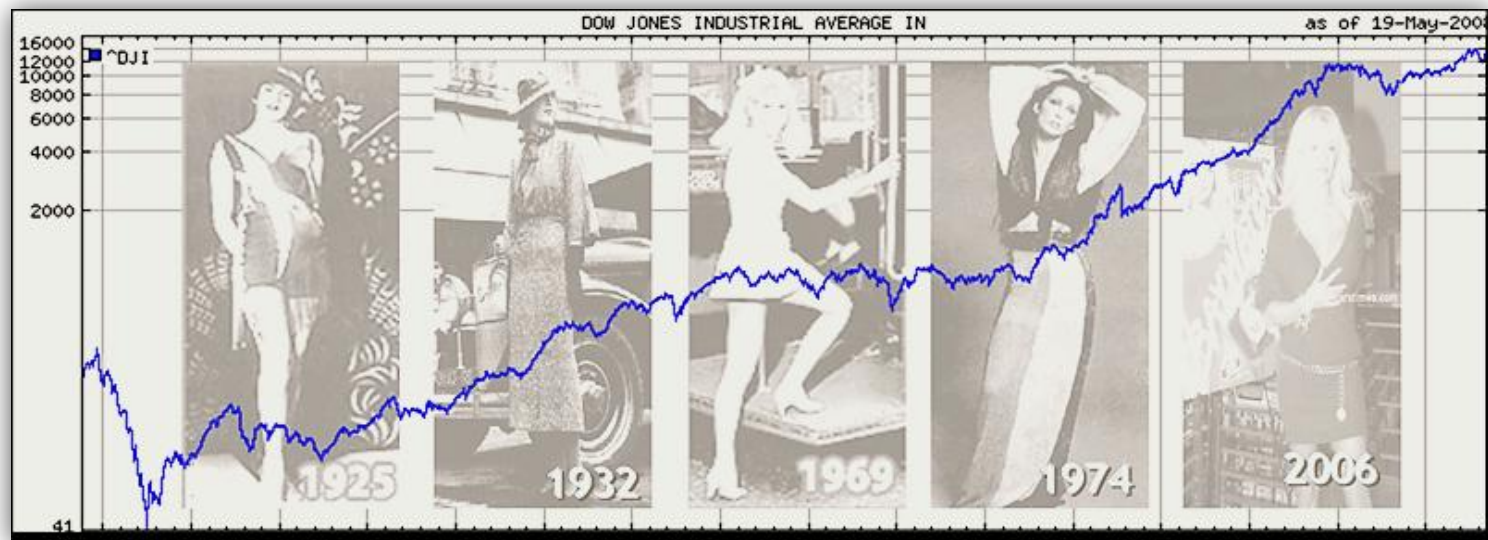
Apr 29, 2011

Chicago, IL

MUSIC HACKDAY

The Hemline Index

Wikipedia: "The theory suggests that hemlines on women's dresses rise along with stock prices. In good economies, we get such results as miniskirts (as seen in the 1960s), or in poor economic times, as shown by the 1929 Wall Street Crash, hems can drop almost overnight."

DOW JONES INDUSTRIAL AVERAGE IN · as of 19-May-200 · ^DJI · 16000 12000 10000 8000 6000 4000 2000 · 1925 1932 1969 1974 2006 · 41

The Danceability Index

Theory: Dancing without moderation suggests a peak of irrational exuberance. When dancing reaches a maximal level, the market may be overvalued.
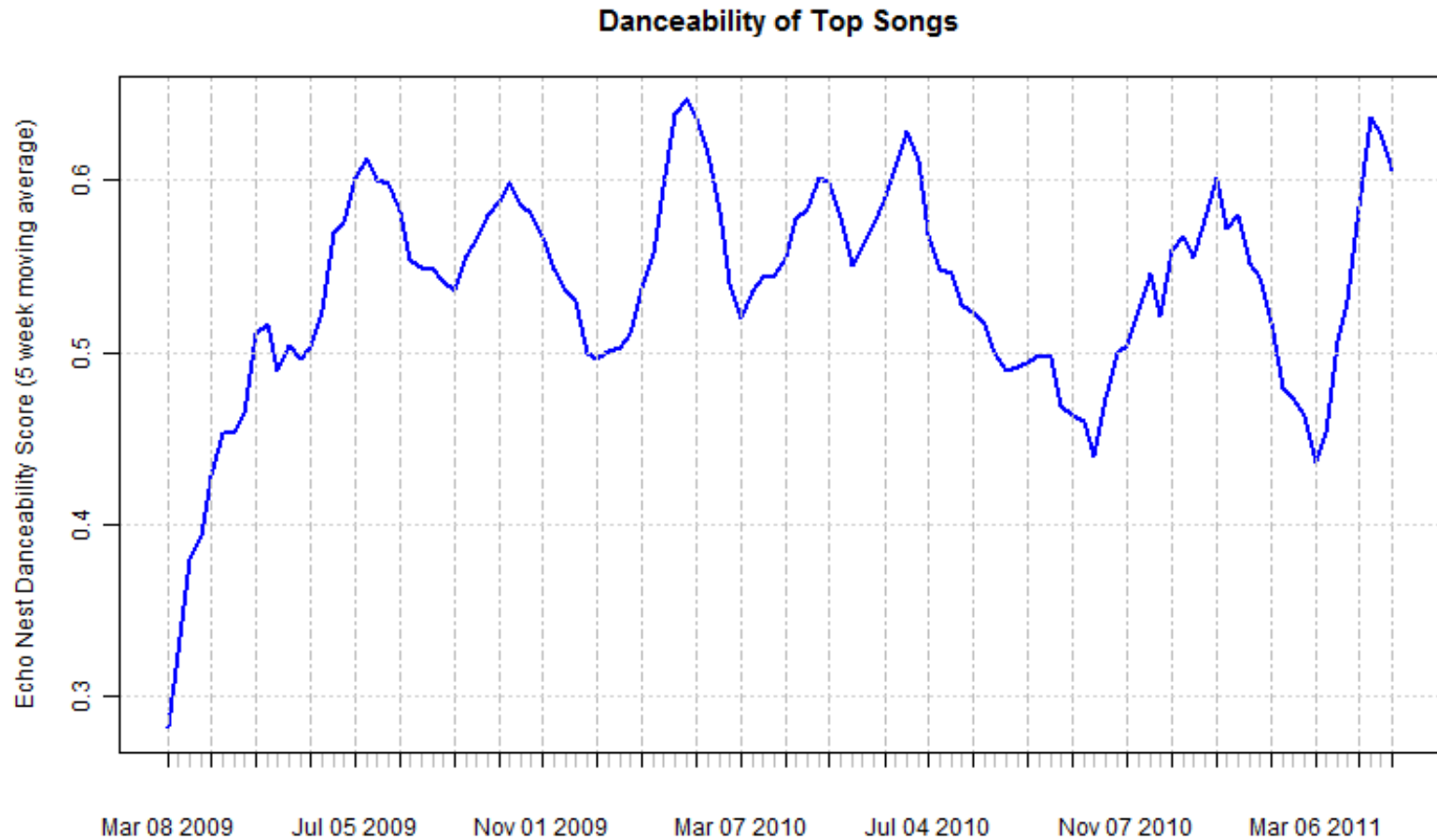

Image source: http://alliedow.wordpress.com/2009/11/17/dancing-in-your-nuddy-pants/

Data Sources
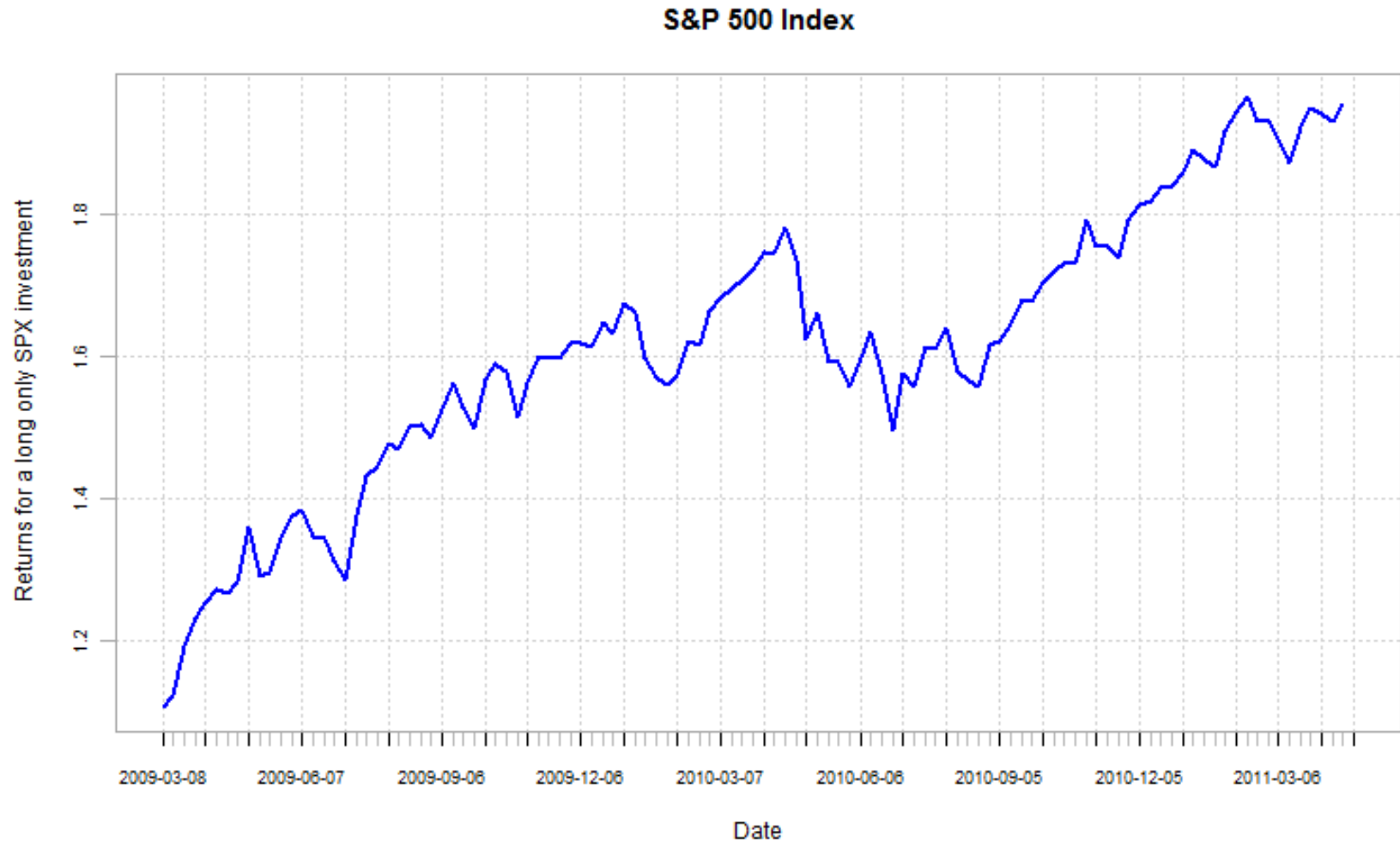
Weekly Historical Top 5 Songs in New York City

Danceability Score of Top Songs

Analysis – Danceability Index

Analysis – S&P 500 Returns

S&P 500 Index

Analysis – Returns vs. Danceability

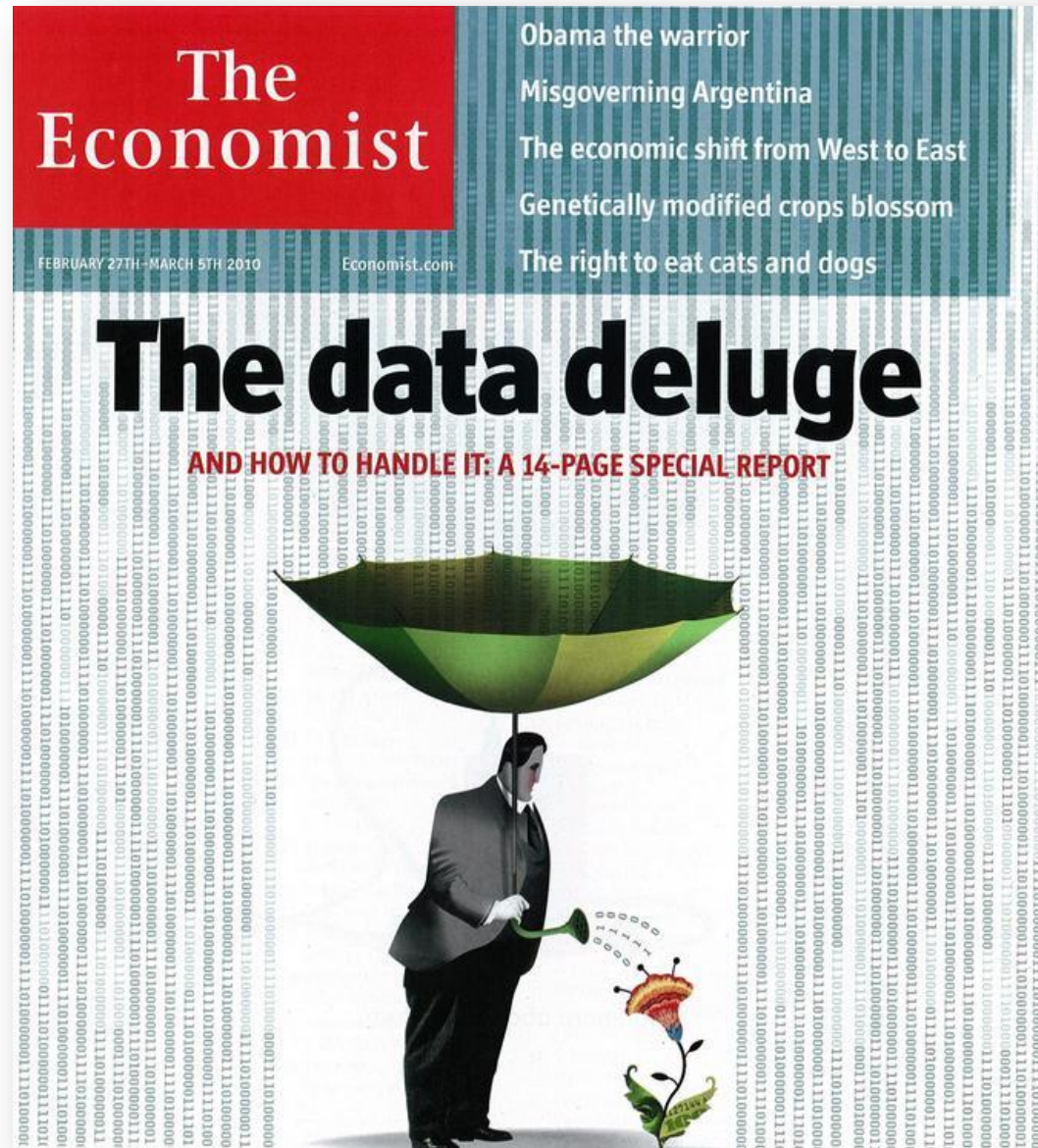Returns by Danceability Intervals
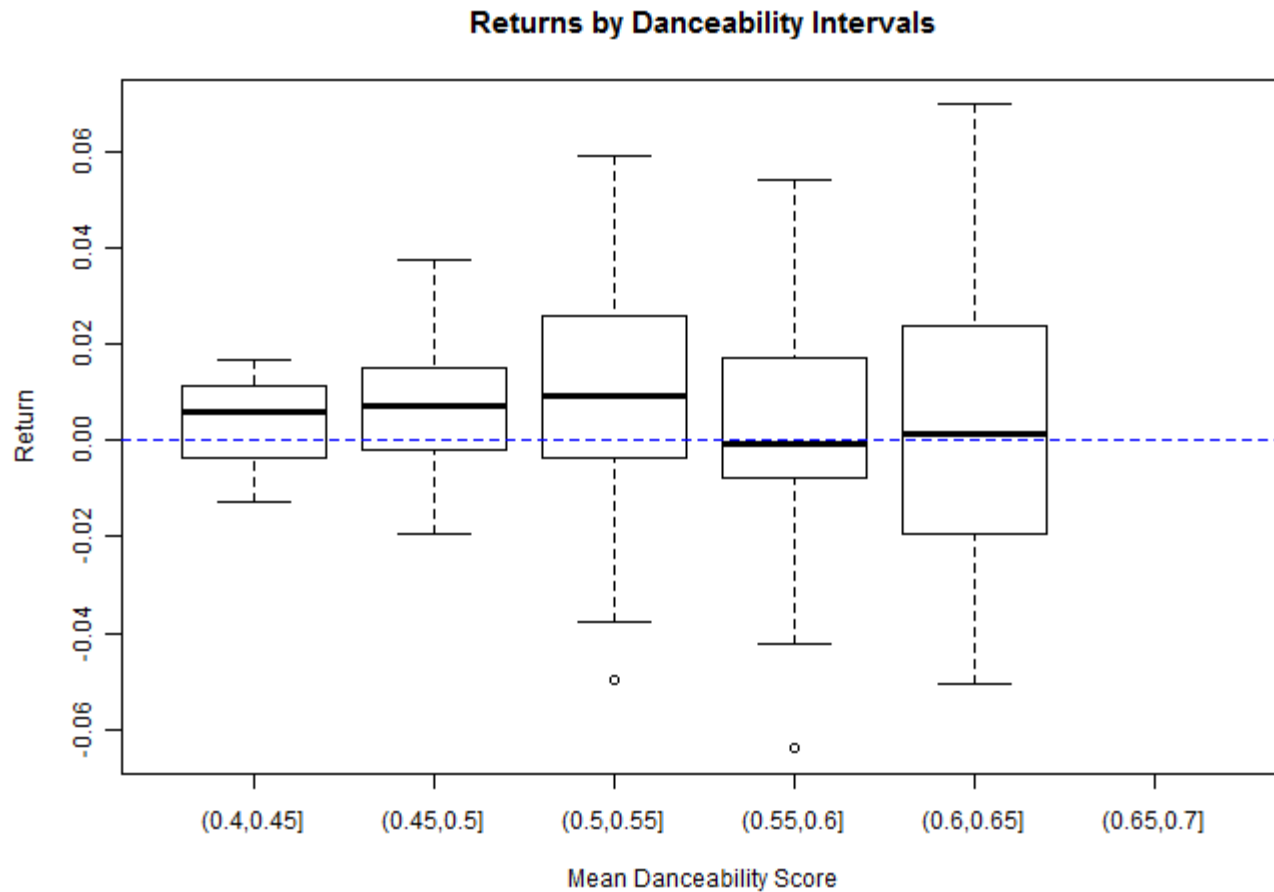
Sentiment Analysis and Events

PART 2: OTHER SOURCES

Data Trends

- Big
- Open
- Structured
- Accessible

R Packages

- XML
- RCurl
- twitteR
- RGoogleTrends
- Infochimps
- tm
- etc.

Sentiment Analysis

- Insider trading
- News
- Social media

Context and volume are important

$$\sqrt{\heartsuit} = ?$$

$$\cos \heartsuit = ?$$

$$\frac{d}{dx} \heartsuit = ?$$

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \heartsuit = ?$$

$$F\{\heartsuit\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t)\, e^{it\heartsuit}\, dt = ?$$

My normal approach is useless here.

Twitter Mood

Soon to be a hedge fund…

Twitter mood predicts the stock market.

Johan Bollen[1,*], Huina Mao[1,*], Xiao-Jun Zeng[2].

*: authors made equal contributions.

Abstract—Behavioral economics tells us that emotions can profoundly affect individual behavior and decision-making. Does this also apply to societies at large, i.e. can societies experience mood states that affect their collective decision making? By extension is the public mood correlated or even predictive of economic indicators? Here we investigate whether measurements of collective mood states derived from large-scale Twitter feeds are correlated to the value of the Dow Jones Industrial Average (DJIA) over time. We analyze the text content of daily Twitter feeds by two mood tracking tools, namely OpinionFinder that measures positive vs. negative mood and Google-Profile of Mood States (GPOMS) that measures mood in terms of 6 dimensions (Calm, Alert, Sure, Vital, Kind, and Happy). We cross-validate the resulting mood time series by comparing their ability to detect the public's response to the presidential election and Thanksgiving day in 2008. A Granger causality analysis and a Self-Organizing Fuzzy Neural Network are then used to investigate the hypothesis that public mood states, as measured by the OpinionFinder and GPOMS mood time series, are predictive of changes in DJIA closing values. Our results indicate that the accuracy of DJIA predictions can be significantly improved by the inclusion of specific public mood dimensions but not others. We find an accuracy of 87.6% in predicting the daily up and down changes in the closing values of the DJIA and a reduction of the Mean Average Percentage Error by more than 6%.

Index Terms—stock market prediction — twitter — mood analysis.

I. INTRODUCTION

STOCK market prediction has attracted much attention from academia as well as business. But can the stock market really be predicted? Early research on stock market prediction [1], [2], [3] was based on random walk theory and the Efficient Market Hypothesis (EMH) [4]. According to the EMH stock market prices are largely driven by *new* information, i.e. news, rather than present and past prices. Since news is unpredictable, stock market prices will follow a random walk pattern and cannot be predicted with more than 50 percent accuracy [5].

There are two problems with EMH. First, numerous studies show that stock market prices do not follow a random walk and can indeed to some degree be predicted [5], [6], [7], [8] thereby calling into question EMH's basic assumptions. Second, recent research suggests that news may be unpredictable but that very early indicators can be extracted from online social media (blogs, Twitter feeds, etc) to predict changes in various economic and commercial indicators. This may conceivably also be the case for the stock market. For example, [11] shows how online chat activity predicts book sales. [12] uses assessments of blog sentiment to predict movie sales. [15] predict future product sales using a Probabilistic Latent Semantic Analysis (PLSA) model to extract indicators of

sentiment from blogs. In addition, Google search queries have been shown to provide early indicators of disease infection rates and consumer spending [14]. [9] investigates the relations between breaking financial news and stock price changes. Most recently [13] provide a ground-breaking demonstration of how public sentiment related to movies, as expressed on Twitter, can actually predict box office receipts.

Although news most certainly influences stock market prices, public mood states or sentiment may play an equally important role. We know from psychological research that emotions, in addition to information, play an significant role in human decision-making [16], [18], [39]. Behavioral finance has provided further proof that financial decisions are significantly driven by emotion and mood [19]. It is therefore reasonable to assume that the public mood and sentiment can drive stock market values as much as news. This is supported by recent research by [10] who extract an indicator of public anxiety from LiveJournal posts and investigate whether its variations can predict S&P500 values.
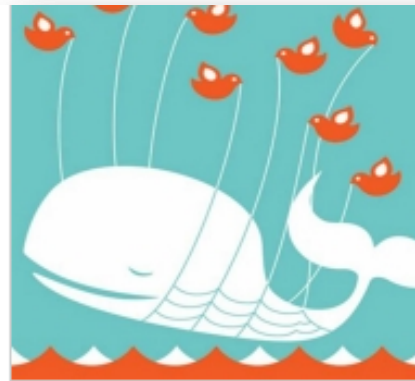
However, if it is our goal to study how public mood influences the stock markets, we need reliable, scalable and early assessments of the public mood at a time-scale and resolution appropriate for practical stock market prediction. Large surveys of public mood over representative samples of the population are generally expensive and time-consuming to conduct, cf. Gallup's opinion polls and various consumer and well-being indices. Some have therefore proposed indirect assessment of public mood or sentiment from the results of soccer games [20] and from weather conditions [21]. The accuracy of these methods is however limited by the low degree to which the chosen indicators are expected to be correlated with public mood.

Over the past 5 years significant progress has been made in sentiment tracking techniques that extract indicators of public mood directly from social media content such as blog content [10], [12], [15], [17] and in particular large-scale Twitter feeds [22]. Although each so-called *tweet*, i.e. an individual user post, is limited to only 140 characters, the aggregate of millions of tweets submitted to Twitter at any given time may provide an accurate representation of public mood and sentiment. This has led to the development of real-time sentiment-tracking indicators such as [17] and "Pulse of Nation"[1].

In this paper we investigate whether public sentiment, as expressed in large-scale collections of daily Twitter posts, can be used to predict the stock market. We use two tools to measure variations in the public mood from tweets submitted

[1] http://www.ccs.neu.edu/home/amislove/twittermood/

…and it's been getting a lot of attention

Investors Slobber Over Cayman Island Hedge Fund That Predicts Stock Market Based on Twitter

4/7/11 at 2:40 PM | 6 Comments

What about when this happens?

It looks like extreme interest from investors over the i of using Twitter to predict stock-market trends is going delay the launch of Derwent Absolute Return Fund. Brothers Paul and Simon Hawtin, wh

THE STRATEGY SESSION
THE 'TWITTER' HEDGE FUND

S&P 500
-4.42   -0.34%
1,286.42

CNBC CNBC CNBC CNBC CNBC CNBC
FIRST IN BUSINESS WORLDWIDE
CNBC

FINalternatives
HEDGE FUND & PRIVATE EQUITY NEWS

Saturday, 9 April 201
Last updated 18 hour

Homepage   About Us   Directory   Events   Jobs   Library

Hedge Funds   Private Equity   People Moves   Regulation   Halls of Jus

Better-Than-Expected Fundraising Delays Twitter Fund Debut

Apr 8 2011 | 12:09pm ET

Here's a problem most hedge fund managers would love to have: a fund that plans to use Twitter feeds has had to delay its debut because it's raised too much money.

Buzz, News Volume, and Headlines

Can signal a change in trends

June 2005

September 2010

twitteR (twitter client for R)

R Package by
Jeff Gantry

Search for tweets with the phrase "Stock Market", then plot frequency of tweets per hour:

```r
library(twitteR)

tweets <- searchTwitter("Stock Market", n=1500)

times <- sapply(tweets, function(x) format(x@created, "%b %d %H:00"))

users <- sapply(tweets, function(x) x@screenName)

times <- times[!duplicated(users)] # removing duplicate users to avoid
spammers and news

counts <- table(times)

bp <- barplot(counts, main="Counts of 'Stock Market' Tweets by Hour",
col="lightblue", border=NA, ylim=c(0,200))

lines(spline(counts ~ bp), lwd=3, lty="dashed", col="darkblue")
```
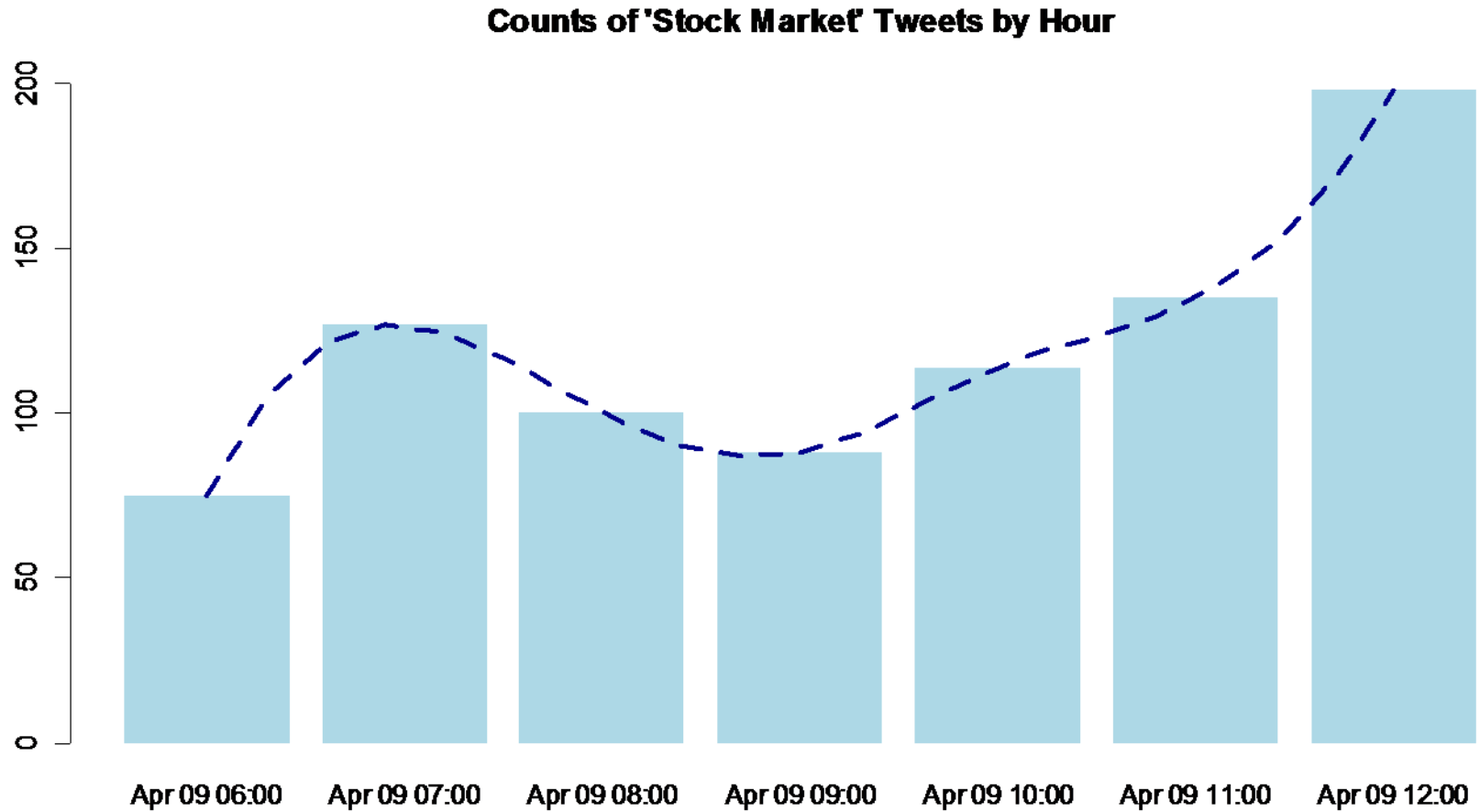
Trend of "Stock Market" tweets

Counts of 'Stock Market' Tweets by Hour

The Hathaway Effect*

*"How Anne gives
Warren Buffett a Rise"*

Oct. 3, 2008 - *Rachel Getting Married* opens:
BRK.A up .44%

Jan. 5, 2009 - *Bride Wars* opens:
BRK.A up 2.61%

Feb. 8, 2010 - *Valentine's Day* opens:
BRK.A up 1.01%

March 5, 2010 - *Alice in Wonderland* opens:
BRK.A up .74%

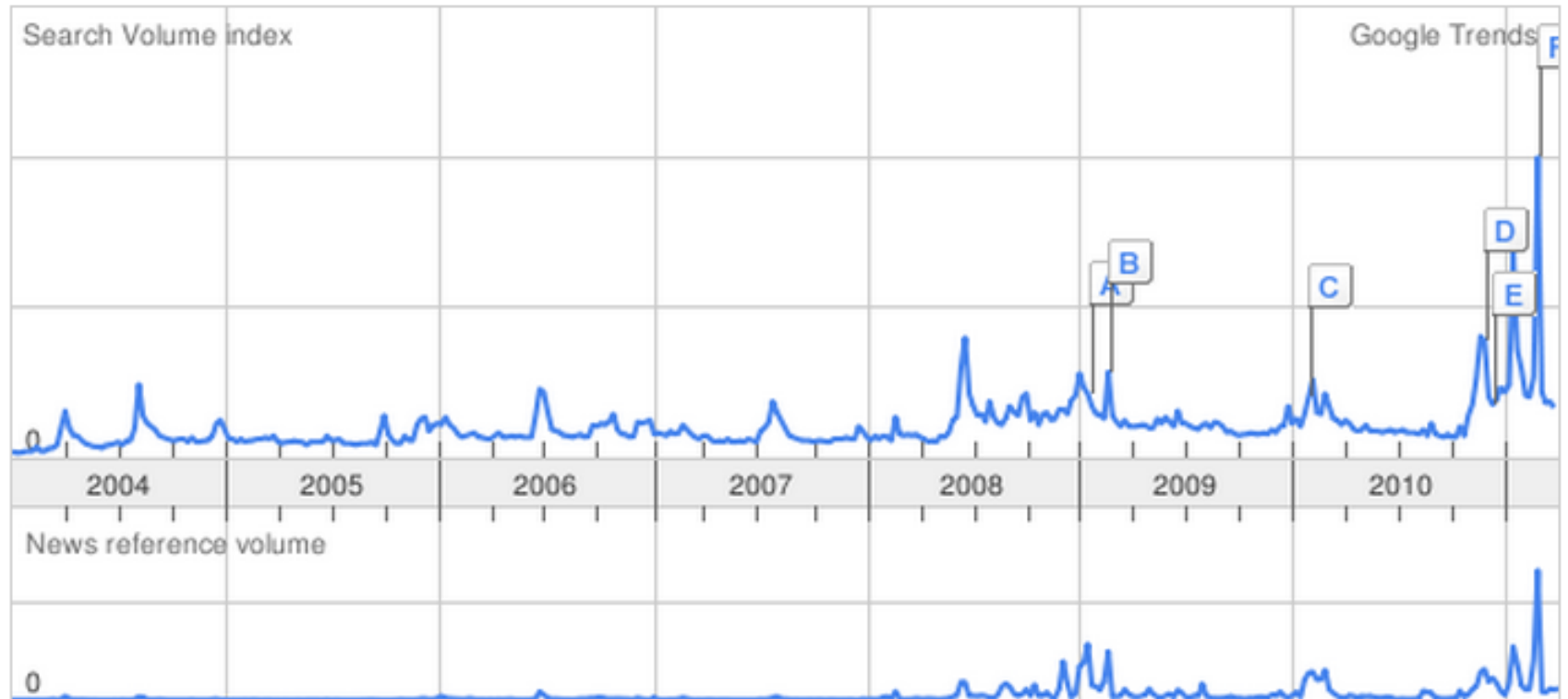Nov. 24, 2010 - *Love and Other Drugs* opens:
BRK.A up 1.62%

Nov. 29, 2010 - Anne announced as co-host of
the Oscars:
BRK.A up .25%

*I don't really believe this is true
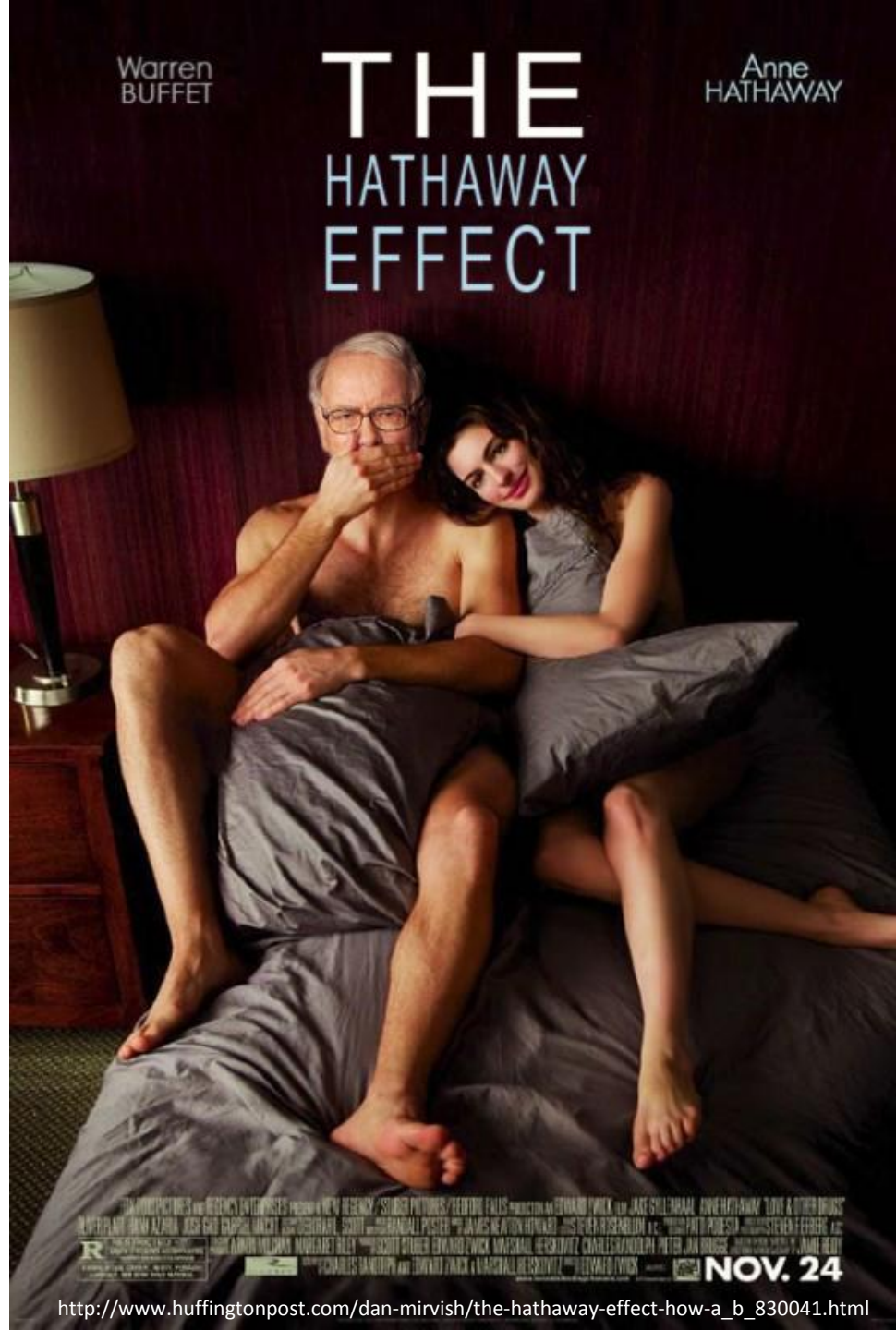
Search Volume with Google Trends

Search volume seems to be a close proximity to news volume

RGoogleTrends

R Package by
Duncan Temple Long

```r
library(RGoogleTrends)
library(quantmod)

ans = getGTrends("Anne Hathaway")
trend <- xts( ans$Week$anne.hathaway, order.by=as.Date(ans$Week$Week, "%b %d %Y"))
brk <- getSymbols("BRK-A", auto.assign=F, from = "2004-01-01" )
x <- na.locf(merge(trend, Cl(brk)))[index(trend)]

#plot returns and search volume
par(mfrow=c(2,1))
plot(x$trend, main="Google Trends: Anne Hathaway", col="blue")
plot(x$BRK.A.Close, main="Berkshire Hathaway Share Price", col="red", cex=.7)

#evaluate returns by search volume
x$Return <- Delt(x$BRK.A.Close)
breaks <- cut(x$trend, seq(0,10, 1))
boxplot(as.numeric(x$Return) ~ breaks, ylab="Weekly Return", xlab="Search Volume",
        col="lightblue", border="darkblue" )
abline(h=0, col="blue")
```
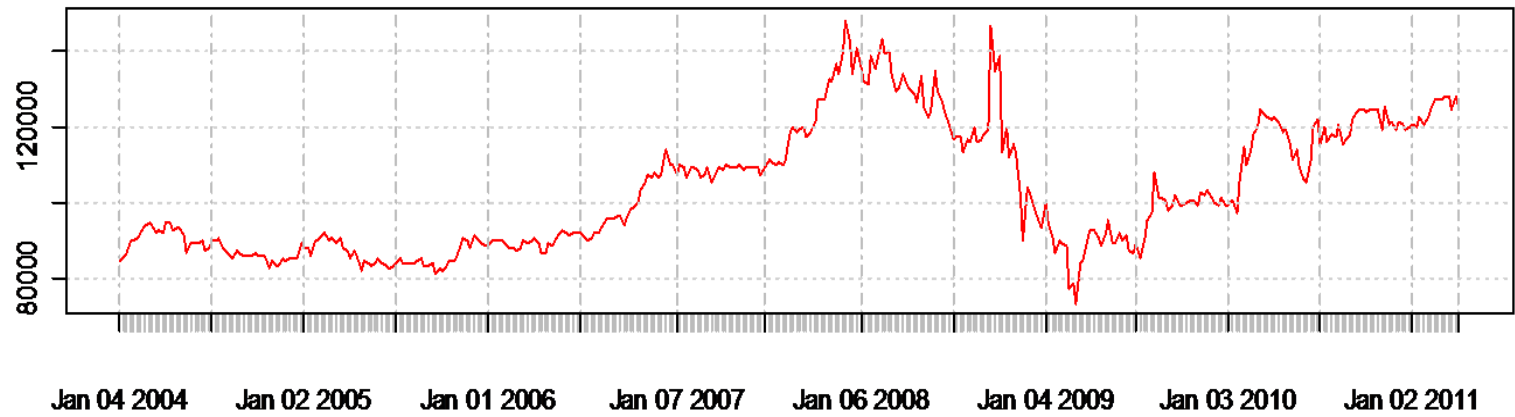
Google Trends: Anne Hathaway
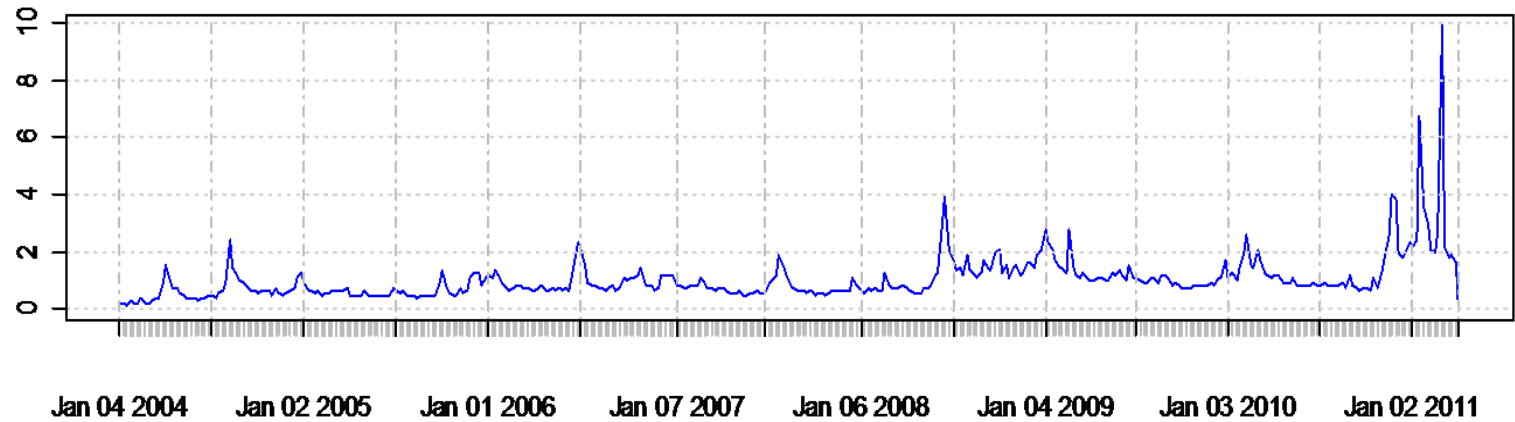
Berkshire Hathaway Share Price

Returns Analysis

Doesn't hold up with search volume

(at least on a weekly basis, and you wouldn't expect short term spikes to last)
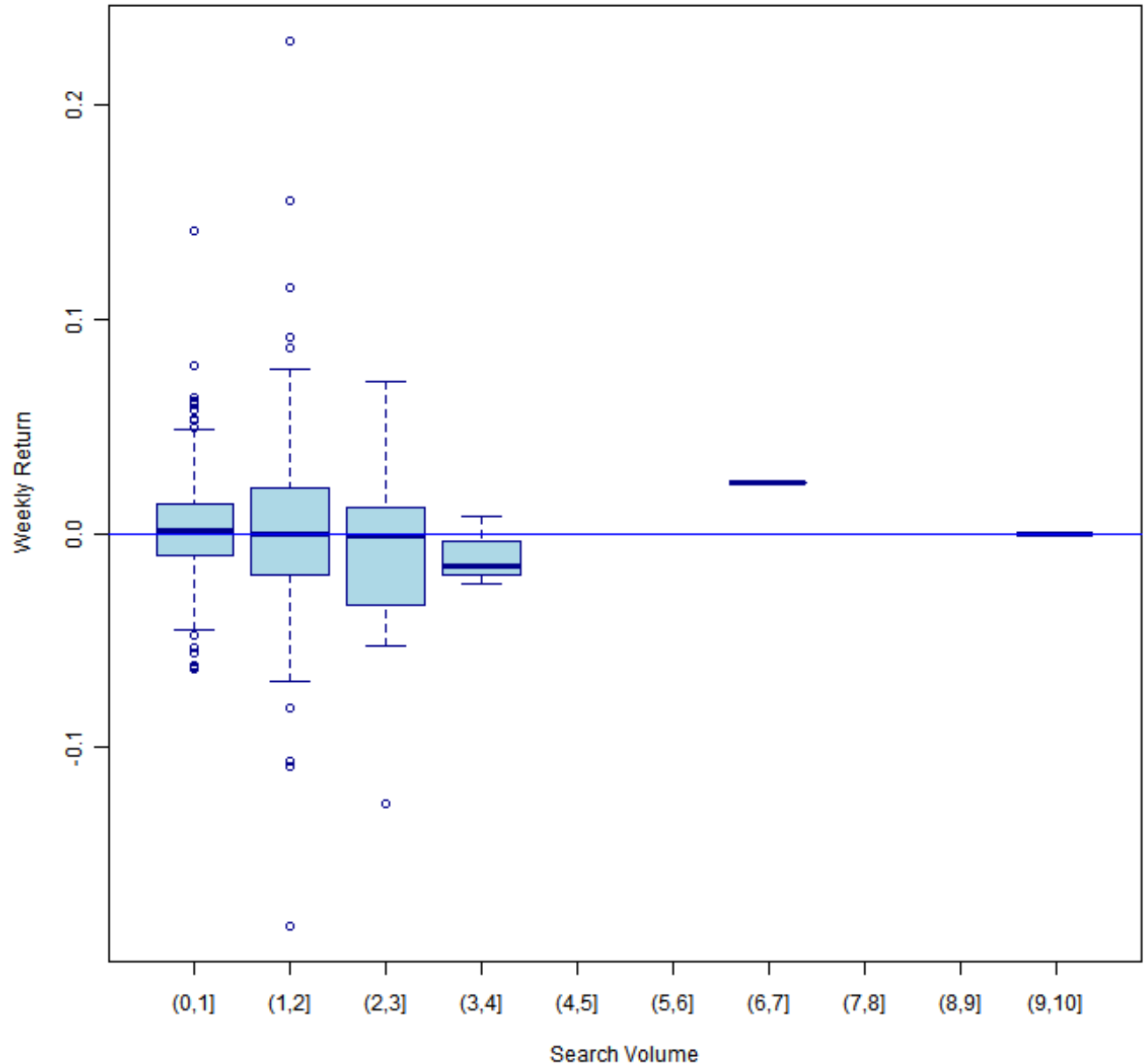
infochimps

R Package by
Drew Conway

- Word frequencies from the British National Corpus
- Twitter influence
- Wikipedia abstracts

- Yahoo stock prices
- Wikipedia abstracts
- etc.

Thank you

joe.rothermich@gmail.com