

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

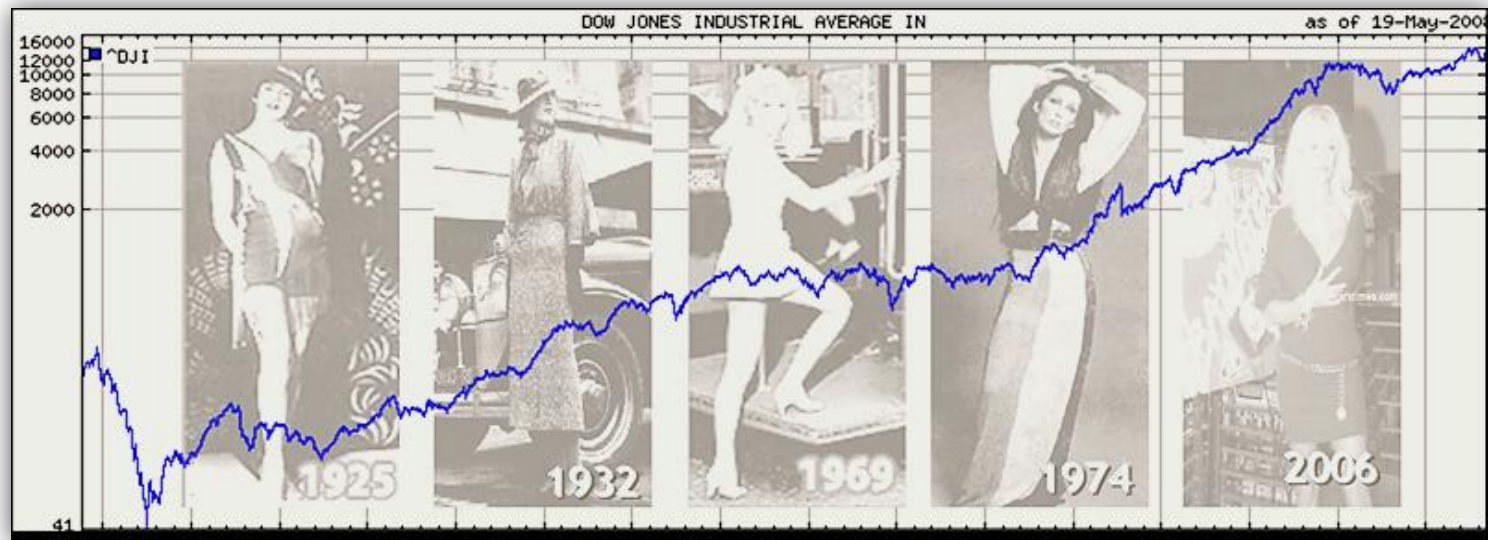
(Introduction)

MUSIC HACKDAY

The image shows a complex music production environment. In the background, there are several racks of modular synthesizers. The left rack has a white faceplate with various knobs and switches, and the right rack is more densely packed with modules. A large, chaotic web of multi-colored patch cables (red, yellow, blue, white) connects the modules across the racks and hangs down. In the foreground, there are three keyboards: one on a white stand to the left, one on a desk in the center-right, and another partially visible on the right. A black office chair is positioned in front of the central keyboard. A desk lamp with a silver shade is on the right side of the desk. The overall scene suggests a hands-on, experimental approach to music technology.

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."



The Danceability Index

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



Data Sources

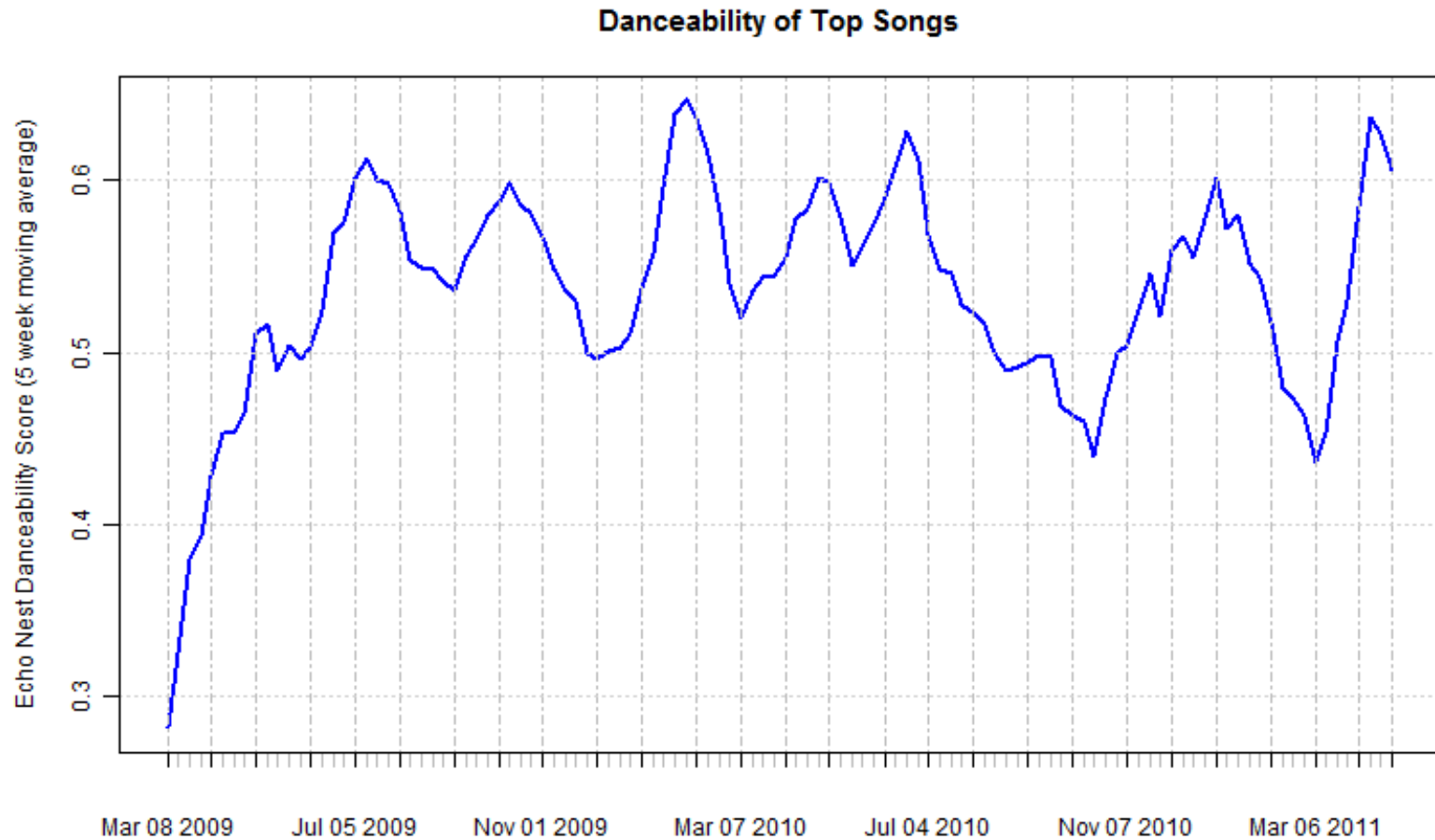


Weekly Historical Top 5
Songs in New York City

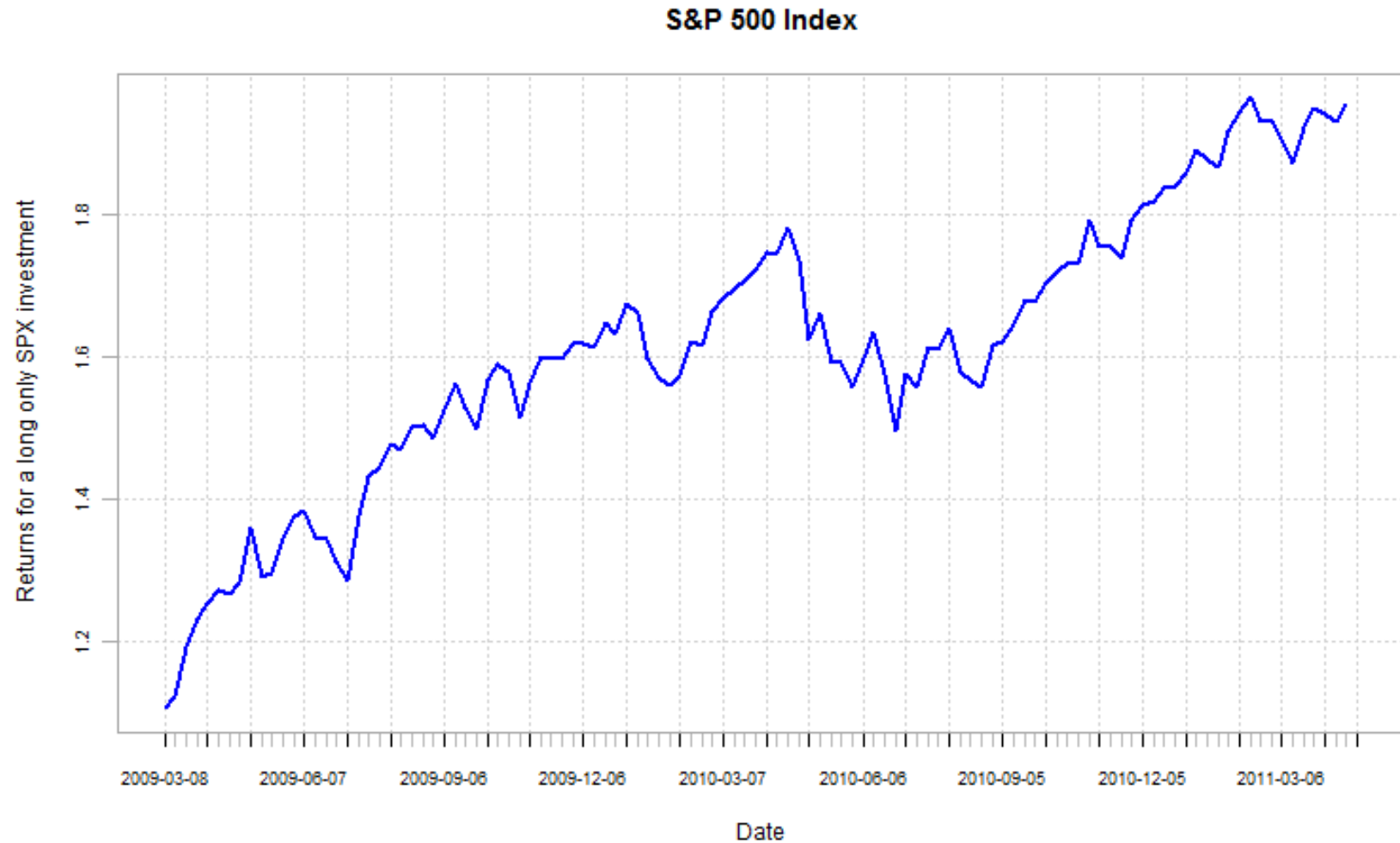


Danceability Score of
Top Songs

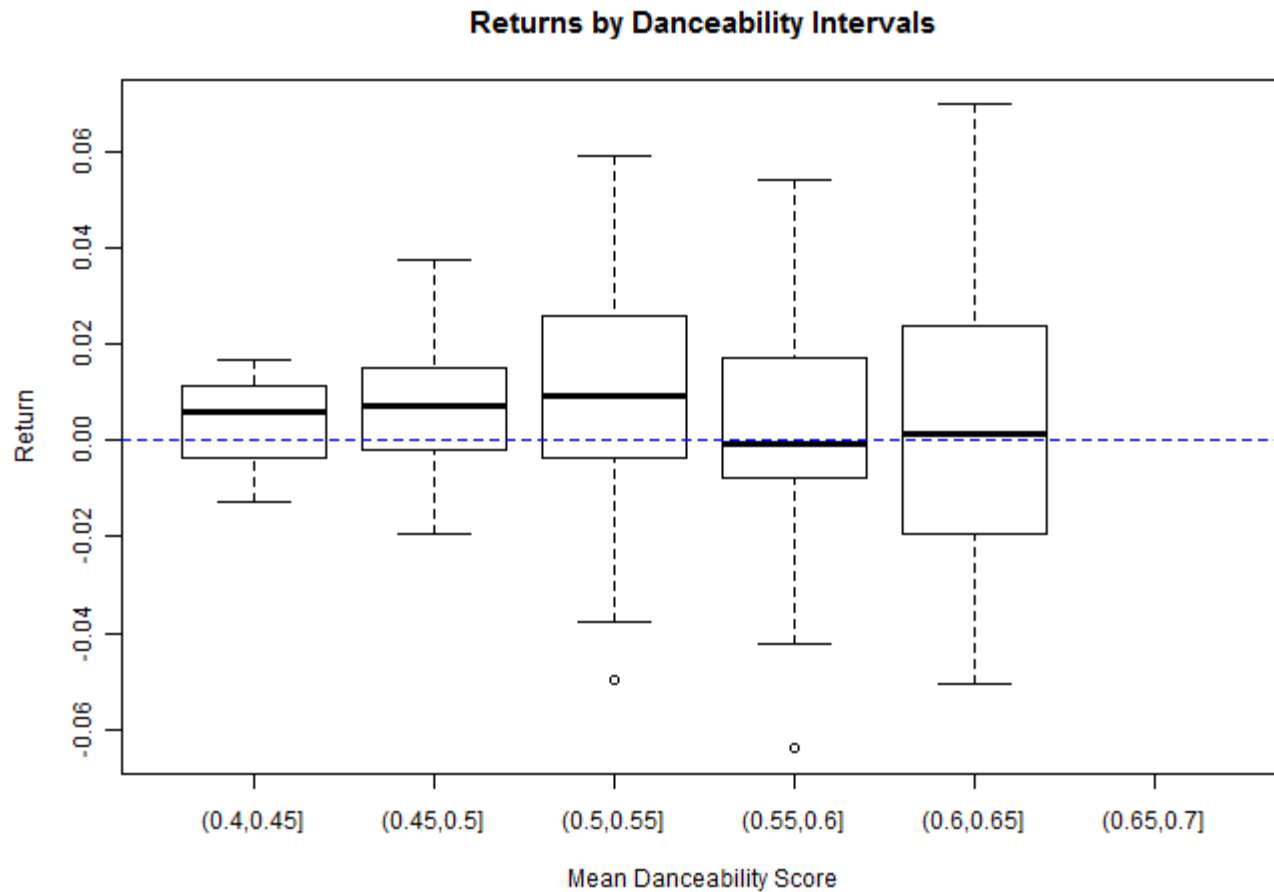
Analysis – Danceability Index



Analysis – S&P 500 Returns



Analysis – Returns vs. Danceability



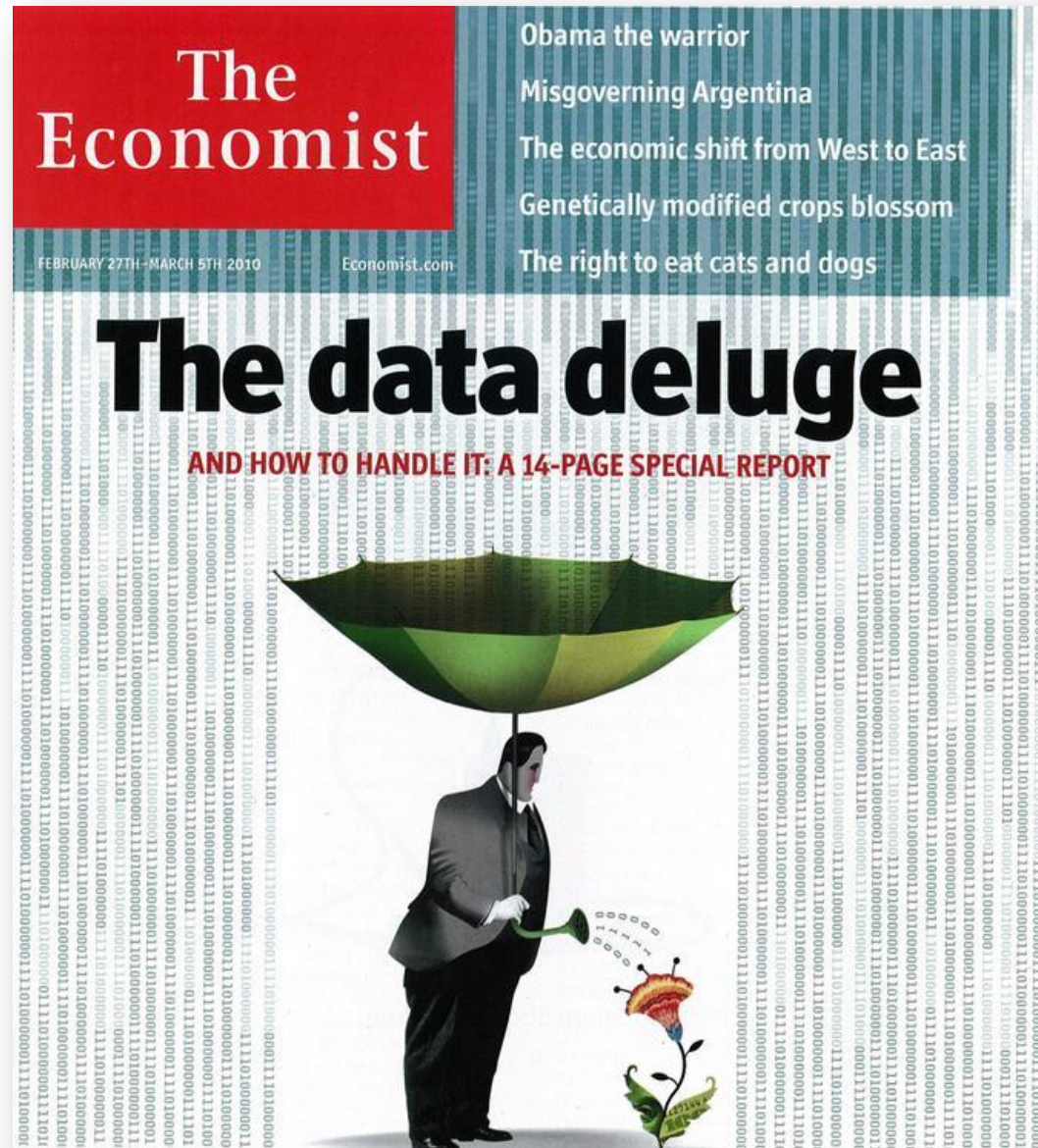
A photograph of a vintage modular synthesizer setup. The image shows several wooden cases filled with various electronic modules, including oscillators, filters, and amplifiers. A dense network of colorful patch cables (red, orange, yellow, green, blue, and white) connects the modules, creating a complex web of connections. The setup is arranged on a desk, with a keyboard visible in the foreground and a lamp on the right. The background is a plain blue wall.

Sentiment Analysis and Events

PART 2: OTHER SOURCES

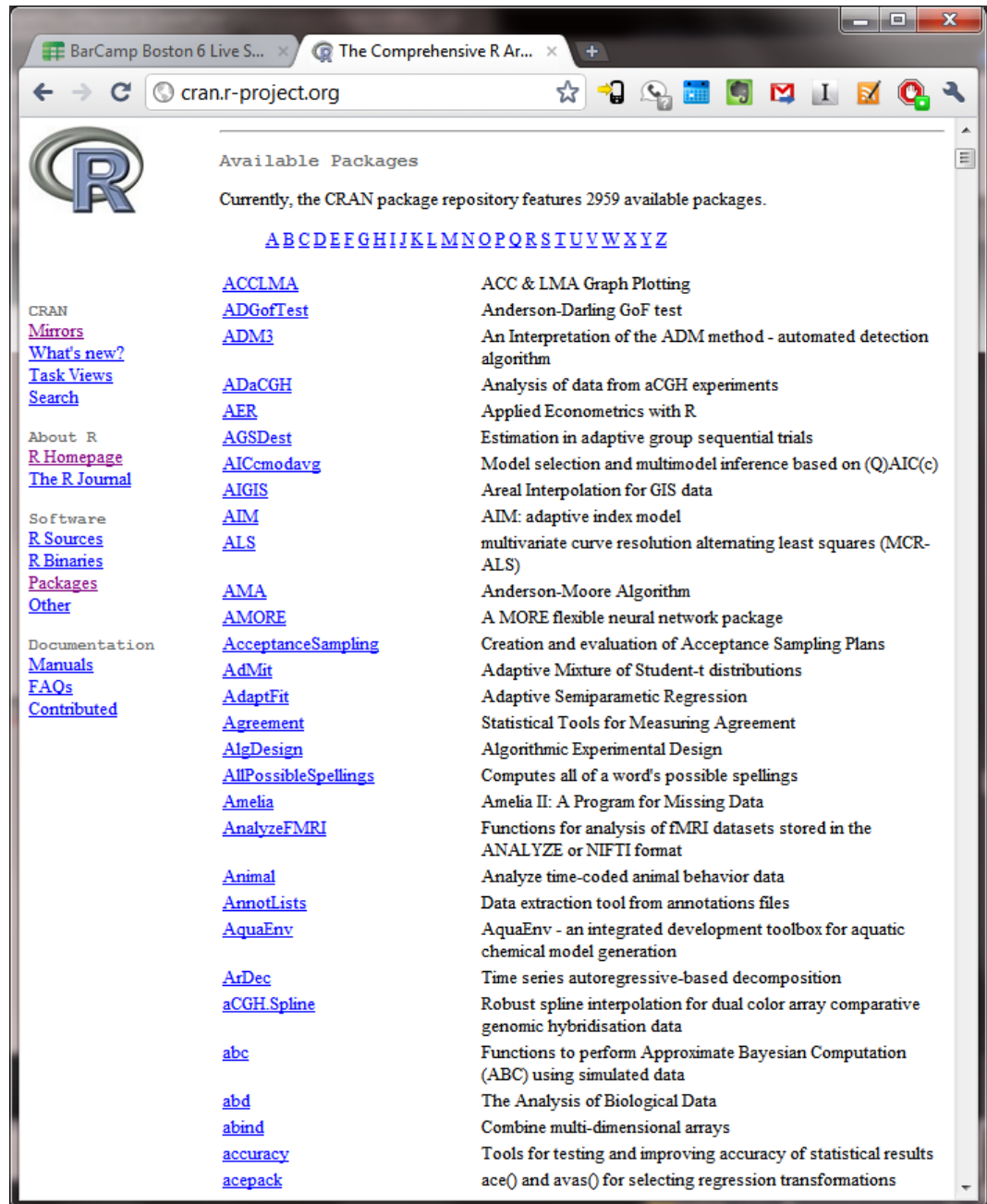
Data Trends

- Big
- Open
- Structured
- Accessible



R Packages

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



The screenshot shows the CRAN (Comprehensive R Archive Network) website. The browser address bar displays "cran.r-project.org". The page features the R logo on the left and a list of available packages on the right. The text "Available Packages" is at the top, followed by "Currently, the CRAN package repository features 2959 available packages." Below this is a navigation bar with letters A through Z. The left sidebar contains links for CRAN (Mirrors, What's new?, Task Views, Search), About R (R Homepage, The R Journal), Software (R Sources, R Binaries, Packages, Other), and Documentation (Manuals, FAQs, Contributed). The main content area lists packages in three columns, including ACC & LMA Graph Plotting, Anderson-Darling GoF test, An Interpretation of the ADM method, Analysis of data from aCGH experiments, Applied Econometrics with R, Estimation in adaptive group sequential trials, Model selection and multimodel inference based on (Q)AIC(c), Areal Interpolation for GIS data, AIM: adaptive index model, multivariate curve resolution alternating least squares (MCR-ALS), Anderson-Moore Algorithm, A MORE flexible neural network package, Creation and evaluation of Acceptance Sampling Plans, Adaptive Mixture of Student-t distributions, Adaptive Semiparametric Regression, Statistical Tools for Measuring Agreement, Algorithmic Experimental Design, Computes all of a word's possible spellings, Amelia II: A Program for Missing Data, Functions for analysis of fMRI datasets stored in the ANALYZE or NIFTI format, Analyze time-coded animal behavior data, Data extraction tool from annotations files, AquaEnv - an integrated development toolbox for aquatic chemical model generation, Time series autoregressive-based decomposition, Robust spline interpolation for dual color array comparative genomic hybridisation data, Functions to perform Approximate Bayesian Computation (ABC) using simulated data, The Analysis of Biological Data, Combine multi-dimensional arrays, Tools for testing and improving accuracy of statistical results, ace() and avas() for selecting regression transformations.

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...

arXiv:1010.3003v1 [cs.CE] 14 Oct 2010

Twitter mood predicts the stock market.

Johan Bollen^{1,*}, Huina Mao^{1,*}, Xiao-Jun Zeng².

*: 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 [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 a 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.

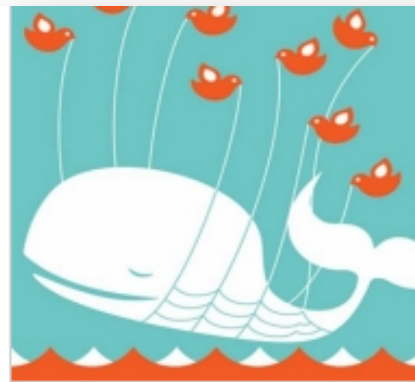
However, if it is our goal to study how public mood influences the stock markets, we need reliable, scalable and early assessments of the public mood at a time-scale and resolution appropriate for practical stock market prediction. Large surveys of public mood over representative samples of the population 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



What about when this happens?

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

4/7/11 at 2:40 PM | [6 Comments](#)

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



FINalternatives
HEDGE FUND & PRIVATE EQUITY NEWS

Saturday, 9 April 2011

Last updated 18 hours ago

[Homepage](#) [About Us](#) [Directory](#) [Events](#) [Jobs](#) [Library](#)

[Hedge Funds](#) [Private Equity](#) [People Moves](#) [Regulation](#) [Halls of Justice](#)

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

twitterR (twitter client for R)

R Package by
Jeff Gentry

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

```
library(twitterR)

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)

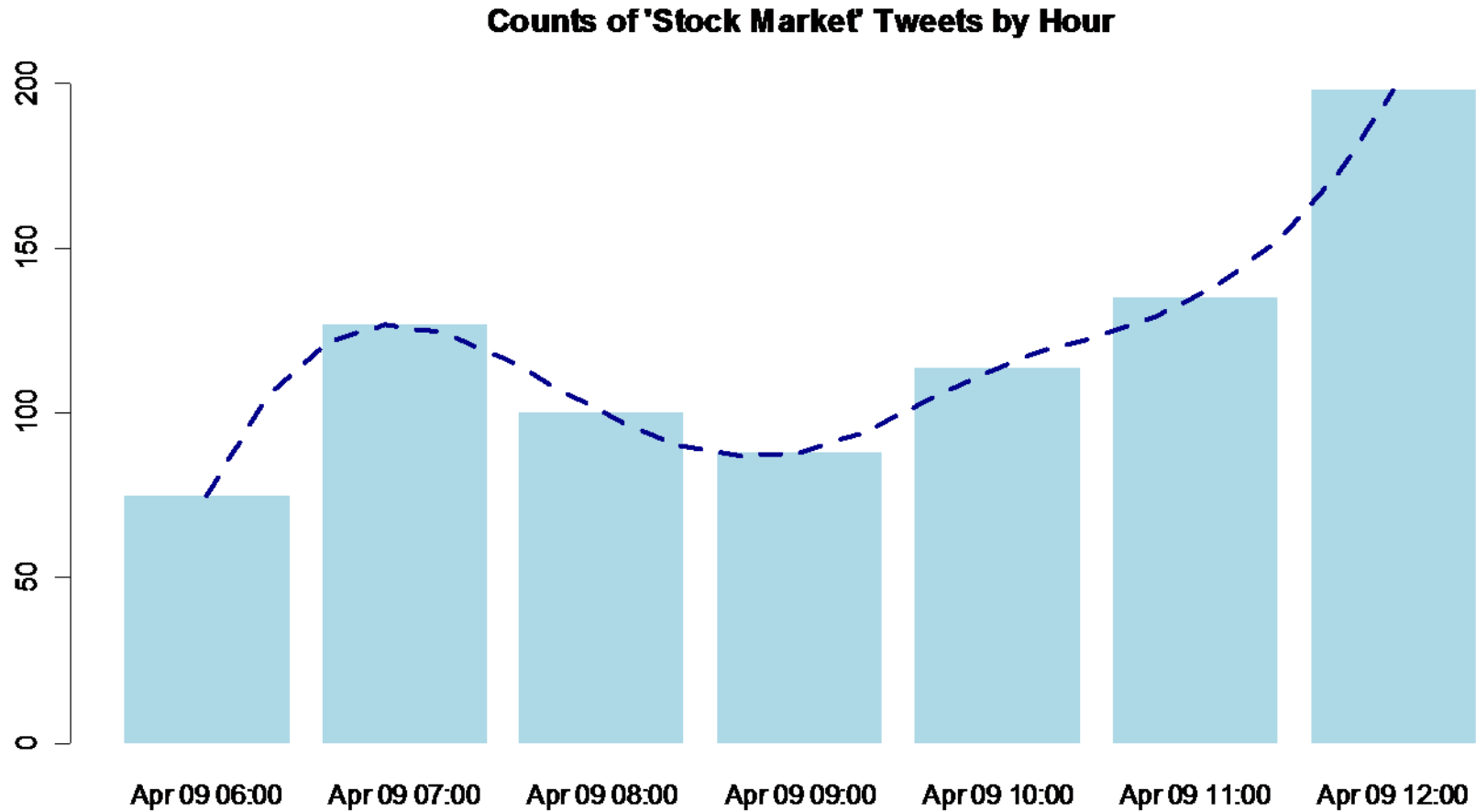
times <- times[!duplicated(users)] # removing duplicate users to avoid
spammers and news

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



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%

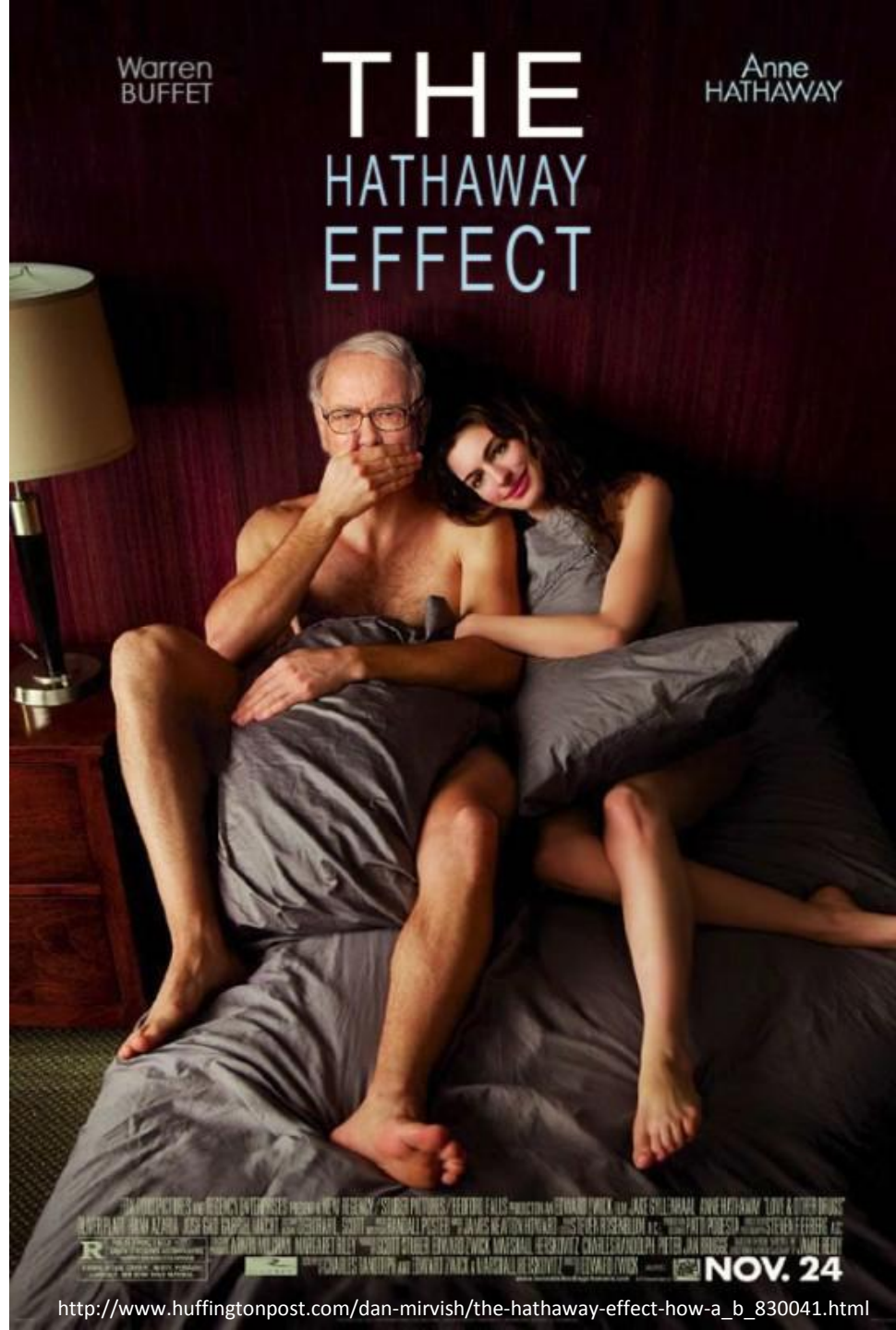
Feb. 8, 2010 - *Valentine's Day* opens:
BRK.A up 1.01%

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

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

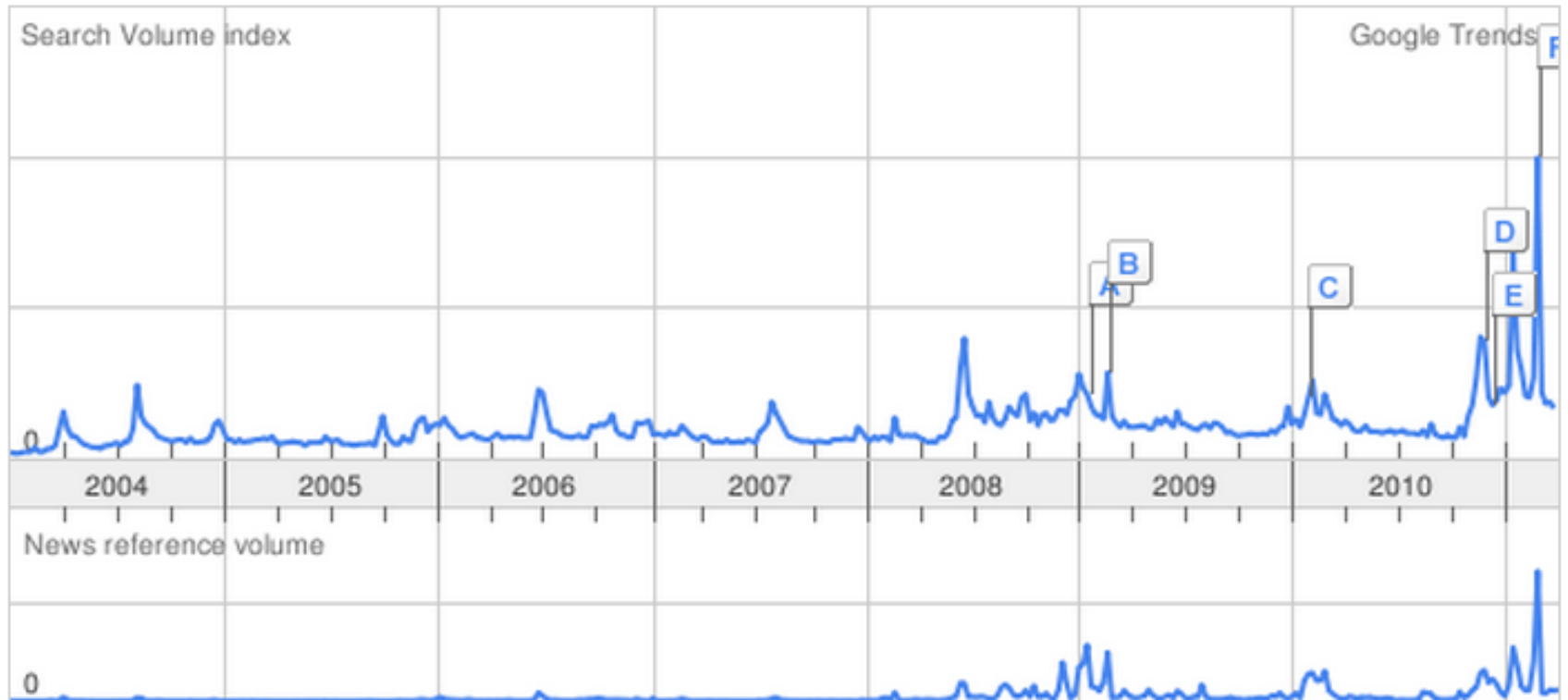
Nov. 29, 2010 - Anne announced as co-host of
the Oscars:
BRK.A up .25%

*I don't really believe this is true



Search Volume with Google Trends

● anne hathaway



Search volume seems to be a close proximity to news volume

RGoogleTrends

R Package by
Duncan Temple Long

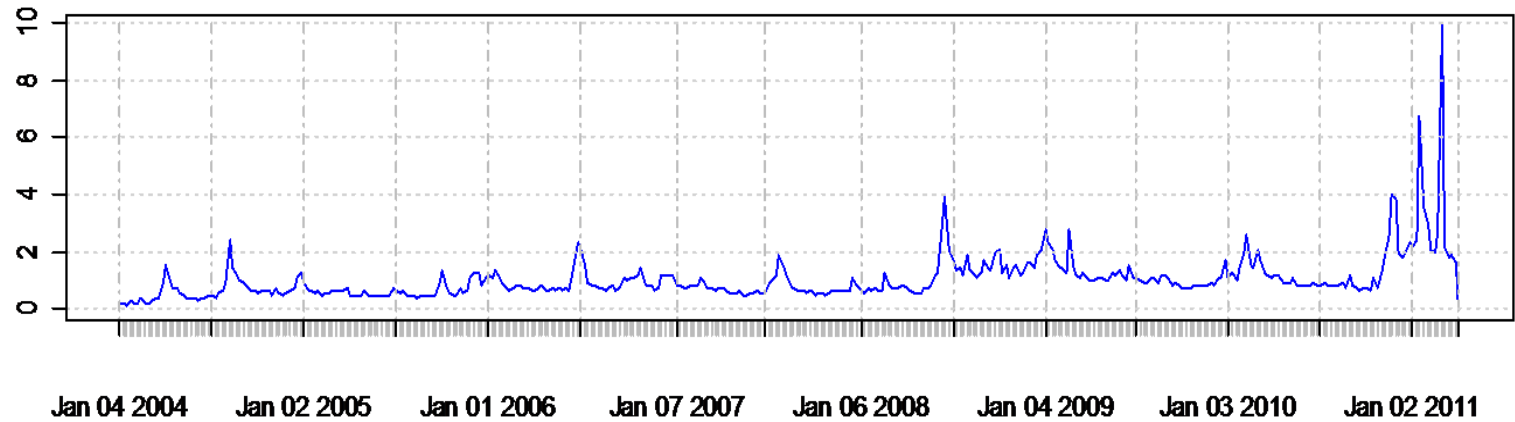
```
library(RGoogleTrends)
library(quantmod)

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)]

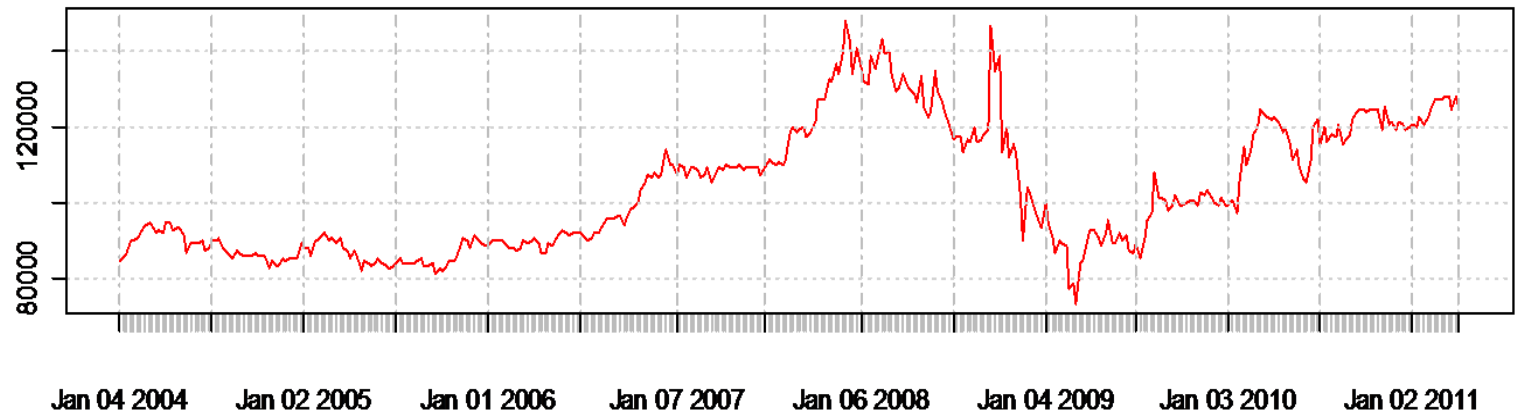
#plot returns and search volume
par(mfrow=c(2,1))
plot(x$trend, main="Google Trends: Anne Hathaway", col="blue")
plot(x$BRK.A.Close, main="Berkshire Hathaway Share Price", col="red", cex=.7)

#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



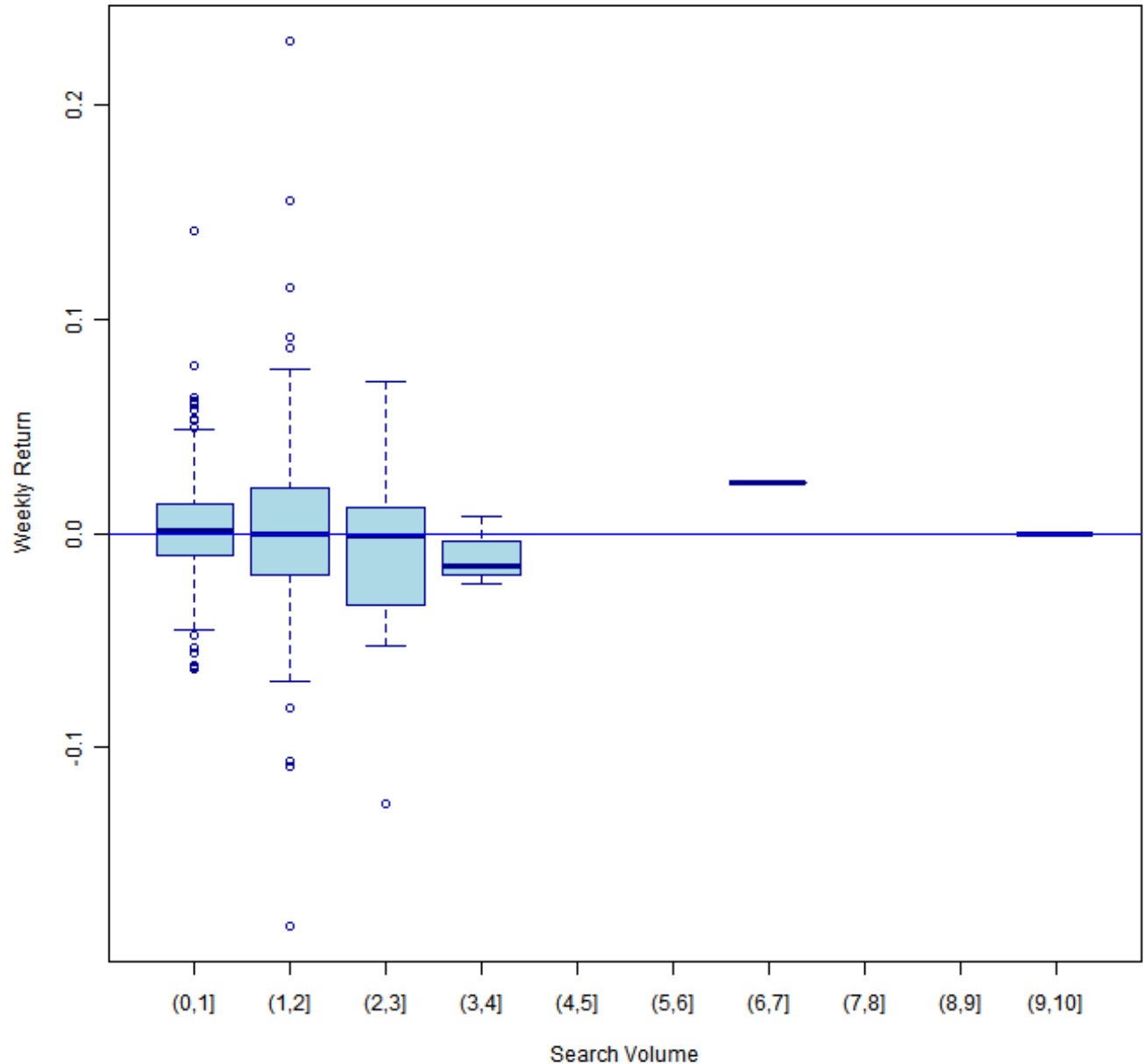
Berkshire Hathaway Share Price



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

The screenshot shows the infochimps website interface. At the top, there's a dark navigation bar with the 'infochimps' logo, a welcome message, and links for 'Sign Up' or 'Log In'. To the right are icons for 'Home', 'Find Data' (highlighted), 'APIs', 'Publish Data', and 'FAQ'. Below the navigation bar is a large search area with the text 'Find data' and a search input field containing 'Find data...'. To the right of the input field, it says '13,524 results'. Below the search bar is a filter section with icons for list and grid views, and three dropdown menus: 'Show free and paid data', 'Show all types of datasets', and 'Sort by best match'. To the right of these are pagination controls showing '1', '2', '3', '...', and '423'. Below the filter section, a message says: 'Please enter a search term above, and we'll find some data for you! Or, browse our [data sets](#) or [tags](#).' Below this message are two columns of links. The left column is titled 'Popular Data Sets (Browse All Data sets)' and lists various datasets like 'Word List - 350,000+ Simple English Words', 'Word List - 100,000+ official crossword words', 'GeoNames.org Postal Code files', 'Crime Rates by State', 'National Center for Educational Statistics (NCES): Tables and Figures', 'Retrosheet: Transactions in Major League Baseball', 'Daily 1970-2010 Open, Close, Hi, Low and Volume (NYSE exchange)', 'Social Security--Beneficiaries, Annual Payments, and Average Monthly Benefit and by State and Other', and '60,000+ Documented UFO Sightings With Text Descriptions And Metadata'. The right column is titled 'Popular Tags (Browse All Tags)' and lists various tags like 'Government', 'Social', 'Music', 'Locations', 'Economics', 'Chemistry', 'Zipcode', 'Pollution', 'Health', 'Law', 'Sports', 'Statistics', 'Survey', 'Language', 'Spending', 'Income', 'Word', 'Age', 'Football', 'Death', 'Demographics', 'Character', 'Geonames', 'Longitude', 'Literature', 'Maps', 'Science', 'Census', 'National', 'Housing', 'Employment', 'Corpora', 'Population', 'Commodities', 'Twitter', and 'Size-large'.

infochimps Welcome. Please [Sign Up](#) or [Log In](#)

Home Find Data APIs Publish Data FAQ

Find data Find data... 13,524 results

Show free and paid data Show all types of datasets Sort by best match

Please enter a search term above, and we'll find some data for you! Or, browse our [data sets](#) or [tags](#).

Popular Data Sets (Browse All Data sets)

- [Word List - 350,000+ Simple English Words \(with Definitions, Excel format\)](#)
- [Word List - 100,000+ official crossword words \(Excel readable\)](#)
- [GeoNames.org Postal Code files - US Zip Code Geolocations](#)
- [Crime Rates by State, 2004 and 2005, and by Type, 2005 \(Cleaned up version\)](#)
- [National Center for Educational Statistics \(NCES\): Tables and Figures](#)
- [Retrosheet: Transactions in Major League Baseball \(Trade, Signing, Draft, etc.\)](#)
- [Daily 1970-2010 Open, Close, Hi, Low and Volume \(NYSE exchange\)](#)
- [Social Security--Beneficiaries, Annual Payments, and Average Monthly Benefit and by State and Other](#)
- [60,000+ Documented UFO Sightings With Text Descriptions And Metadata](#)

Popular Tags (Browse All Tags)

- [Government](#) [Social](#) [Music](#) [Locations](#) [Economics](#) [Chemistry](#)
- [Zipcode](#) [Pollution](#) [Health](#) [Law](#) [Sports](#) [Statistics](#) [Survey](#) [Language](#)
- [Spending](#) [Income](#) [Word](#) [Age](#) [Football](#) [Death](#) [Demographics](#)
- [Character](#) [Geonames](#) [Longitude](#) [Literature](#) [Maps](#) [Science](#)
- [Census](#) [National](#) [Housing](#) [Employment](#) [Corpora](#)
- [Population](#) [Commodities](#) [Twitter](#) [Size-large](#)

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

- Yahoo stock prices
- Wikipedia abstracts
- etc.

Thank you

joe.rothermich@gmail.com