UCLAAnderson
SCHOOL of MANAGEMENT

MGMTMFE 431:

Data Analytics and Machine Learning

Topic 8:
Textual Analysis and Trading Strategies

Spring 2019

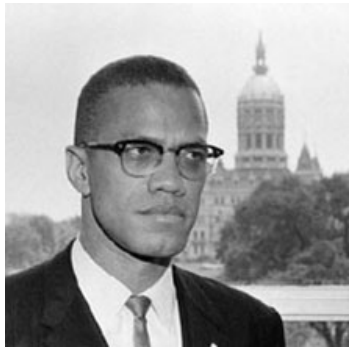
Professor Lars A. Lochstoer

First: A Review of Research on News and the Stock Market

UCLAAnderson
SCHOOL of MANAGEMENT

Motivation

- "The media's the most powerful entity on earth."
  - Malcolm X, American human rights activist



- "I fear three newspapers more than a hundred thousand bayonets."
  - Napolean Bonaparte, first emperor of France

- "Whoever controls the media controls the mind."
  - Jim Morrison, lead singer of The Doors

UCLAAnderson
SCHOOL of MANAGEMENT

Importance of Media in Finance

- Old view: Financial media doesn't matter
  - Capital and product markets are highly efficient
  - Prices and quantities already reflect all information

- Modern view: Media affects and reflects behavior
  - Investors, managers, and consumers are human
  - Humans have limited information processing abilities
    - The way information is transmitted matters
  - Judgmental errors can affect market outcomes
    - Limits to arbitrage in asset markets; frictions in other markets

UCLAAnderson
SCHOOL of MANAGEMENT

Roles of Media in Finance

- Attracts attention
  - To important current events
- Conveys information
  - About the macroeconomy, industries, and firms
  - About politics, laws, and regulations
- Influences beliefs
  - Provides compelling and memorable stories

UCLAAnderson
SCHOOL of MANAGEMENT

News Selection and Promotion

- People notice and remember only a few events
  - We have finite attention and imperfect memories
  - Millions of events occur around the world every day

- Media focuses attention and aids memory by exploiting cognitive heuristics
  - We attend to salient stimuli that stand out
  - We recall memories that are easily available
    - Journalists try to find or construct **dramatic stories**

UCLAAnderson
SCHOOL of MANAGEMENT

Anatomy of a Headline

- Salience
  - Big and bright
  - Evocative language
    - Strips, churn, squirm
- Availability
  - Story-telling
    - Last-minute standoff
    - Wild ride
  - Drama
    - Unprecedented

VOL. CCLVIII NO. 31 ..... WEEKEND ★★★★ $2.00

DOWJONES A NEWS CORPORATION COMPANY — SATURDAY/SUNDAY, AUGUST 6 - 7, 2011 — WSJ.com

S&P Strips U.S. of Top Credit Rating

Unprecedented Downgrade Comes After Last-Minute Standoff; Treasury Says Decision Is 'Flawed by a $2 Trillion Error'

By Damian Paletta And Matt Phillips

A cornerstone of the global financial system was shaken Friday when officials at ratings firm Standard & Poor's said U.S. Treasury debt no longer deserved to be considered among the safest investments in the world.

S&P removed for the first time the triple-A rating the U.S. has held for 70 years, saying the budget deal recently brokered in Washington didn't do enough to address the gloomy outlook for America's finances. It downgraded long-term U.S. debt to AA+, a score that ranks below more than a dozen countries, including Liechtenstein, and on par with Belgium and New Zealand. S&P also put the new grade on "negative outlook," meaning the U.S. has little chance of regaining the top rating in the near term.

FRIDAY (9:30 a.m. to 4 p.m.) DOW JONES INDUSTRIAL AVERAGE — OPEN 11497.96 CLOSE 11444.61

CHANGE IN S&P 500 MARKET CAP THIS WEEK — DOWN $843.79 BILLION

S&P 500 STOCKS UP THIS WEEK — 9

THE DJIA CHANGED DIRECTION — 130 TIMES

CHANGE IN NONFARM PAYROLLS IN THOUSANDS — 235, 194, 217, 117, 68, 53, 45

The unprecedented move came after several hours of high-stakes drama. It began in the morning, when word leaked that a downgrade was imminent and stocks tumbled. Around 1:30 p.m., S&P officials notified the Treasury Department that they planned to downgrade U.S. debt and presented the government with their findings. Treasury officials noticed a $2 trillion error in S&P's math that delayed an announcement for several hours. S&P officials decided to move ahead, and after 8 p.m. they made their downgrade official.

S&P said the downgrade "reflects our opinion that the fiscal consolidation plan that Congress and the administration recently agreed to falls short of what, in our view, would be necessary to stabilize the government's medium-term debt dynamics." It also blamed the weakened "effectiveness, stability, and predictability" of U.S. policy making and political institutions at a time when challenges are mounting.

"A judgment flawed by a $2 trillion error speaks for itself," a Treasury representative said.

The downgrade will force traders and investors to reconsider what has been an elemen-

Please turn to page A4

What's News

★ ★ ★

World-Wide

■ **Syrians poured into the streets of most major towns.** Protesters braved a military crackdown on the first Friday of Ramadan, as security forces fired on marchers. **A11**
Protesters were stymied in the city of Hama, amid a continuing military siege.

■ **A jury found** several New Orleans police officers guilty of covering up their roles in shootings after Katrina. **A3**

■ **Texans are bracing** for more heat as drought strains power and water supplies. **A3**

■ **Japan arrested** two Chinese fishing captains, accusing them of entering Japanese waters illegally. **A11**

■ **Thailand's Parliament** formally elected Yingluck Shinawatra as the country's next prime minister. **A11**

■ **Blasts of solar energy** that could disrupt electrical systems blew by the Earth. ■ **NASA launched** a robotic explorer named Juno on a five-year journey to Jupiter. ■ **Died: Joseph Brady,** 89, "astrochimp" trainer. **A2**

★ ★ ★

Business & Finance

■ **S&P downgraded** the U.S. government's triple-A credit rating, saying the recent budget deal didn't do enough to address the long-term picture for America's finances. **A1, A4**

Markets Go On Wild Ride

By Tom Lauricella And Conor Dougherty

Financial markets seesawed Friday, driven by fast-moving events in Europe and a jobs report that soothed immediate concerns about the U.S. economy but did little to ease longer-term worries.

It was an extraordinary end to a week that saw the Dow Jones Industrial Average collapse nearly 700 points as stocks were flung up and down by skittish traders. On Friday, the Dow gained 60.93 points, or 0.54%, to 11444.61. But that small change masked wicked swings. Within minutes of the opening bell, the Dow was up 245 points. But by midday the Dow had fallen 171 points from Thursday's close, only to soar back to nearly session highs at midafternoon. Bond and currency trading was also volatile.

Even with Friday's gains, the Dow finished the week down nearly 700 points, its largest point decline since the heart of the financial crisis in October 2008. The selloff left the Dow down 10.7% from its high in April of this year. It's in negative territory for 2011, down 1.2%.

While markets closed higher Friday, there was a fresh jolt four hours later: Credit-ratings firm Standard & Poor's downgraded the U.S. government's triple-A debt rating for the first time ever.

The move left investors facing a weekend of uncertainty about how markets will react Monday.

On Friday morning, it seemed that stocks had dodged a bullet. U.S. employers added a better-than-expected 117,000 jobs and the unemployment rate ticked down to 9.1% from 9.2%. In the minutes after the employment report, investors sold safe-haven assets like U.S. Treasurys.

But economists quickly poked holes in the report's positive veneer. The unemployment rate fell not because of stepped-up hiring, but because more discouraged workers gave up their hunt for jobs and dropped out of the work force. At the same time, the pace of job creation, which included 154,000 new private sector jobs, is barely enough to absorb the growth in population.

"One nice number isn't enough to change sentiment at the moment," said Ted Weisberg, president of Seaport Securities. "Folks are scared."

Another reason investors didn't take much heart from the jobs report was that the data was collected in the middle of the month, before the debate intensified over lifting the debt

Please turn to page A4

Markets in Turmoil

More on A4-A9, B1 and B5

As the Financial World Churns, Traders Squirm

BNP Paribas 4:23pm

T3 12:56pm

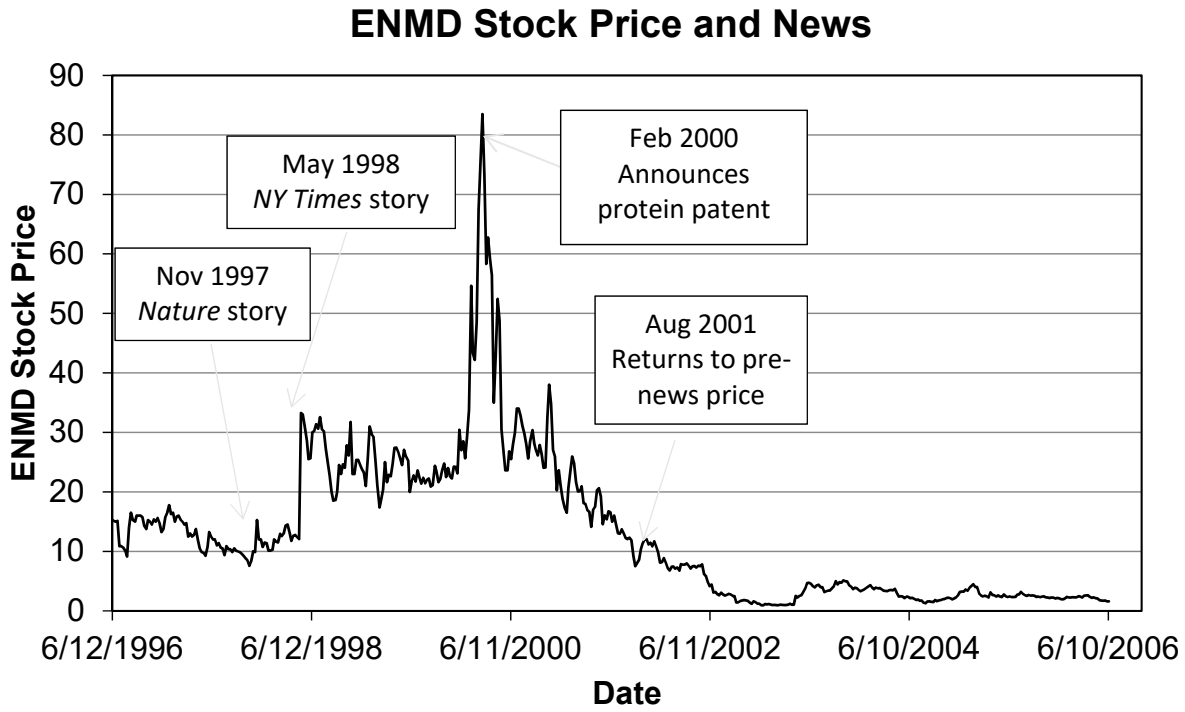
Brencourt Advisors 9:37am

Clockwise from top: John Nunziata, head of U.S. stock trading at BNP Paribas; Michael Dershewitz, research director at Brencourt Advisors; and Scott Redler, chief strategic officer at T3. See page A4.

UCLAAnderson
SCHOOL of MANAGEMENT

The Impact of Media Attention

- Huberman and Regev (2001, *JF*) study EntreMed
  - Possible cure for cancer elicits *many* reactions
  - EntreMed investors seem to ignore past reactions



UCLAAnderson
SCHOOL of MANAGEMENT

Investor Overreaction

- Attention promotes overreaction
  - "Nothing is as important as you think it is while you're thinking about it." - Daniel Kahneman, Nobel laureate

- EntreMed investors focused on salient good news from newspapers and ads
  - They ignored subtle statistical and economic info
    - EntreMed: 90% of new drugs don't receive FDA approval
  - But it's difficult to know whether prices were inefficient

UCLAAnderson
SCHOOL of MANAGEMENT

Stocks Prices and Information

- Prices should reflect info about firm cash flows
  - Stock prices (p) depend on expected cash flows (x)
  - m denotes the stochastic discount factor

$$p_t = E_t[m_{t+1}x_{t+1}|I_t]$$

- Very hard to test even if m takes the CAPM form
  - p and realizations of x (and m) are measurable
  - But expectations of x (and m) are very hard to measure
    - Changes in expectations occur at high frequencies
    - Expectations are conditional on investors' information (I)

How Do We Test Efficiency?

- Null: Measurable info known to investors (*I*) should not predict risk-adjusted returns (*mx*/*p*)
  - Strength of the test is predicated on the quality of info measurement
    - Examples: Past returns, volume, firm characteristics

- If *m* takes the CAPM form, there is no joint hypothesis problem
  - Much AP research looks for the right model of risk
  - Less effort spent on measuring info and expectations

UCLAAnderson
SCHOOL of MANAGEMENT

Reaction to News and Non-news

- News can help us measure expectations of *x*
  - Expectations change when information (news) arrives

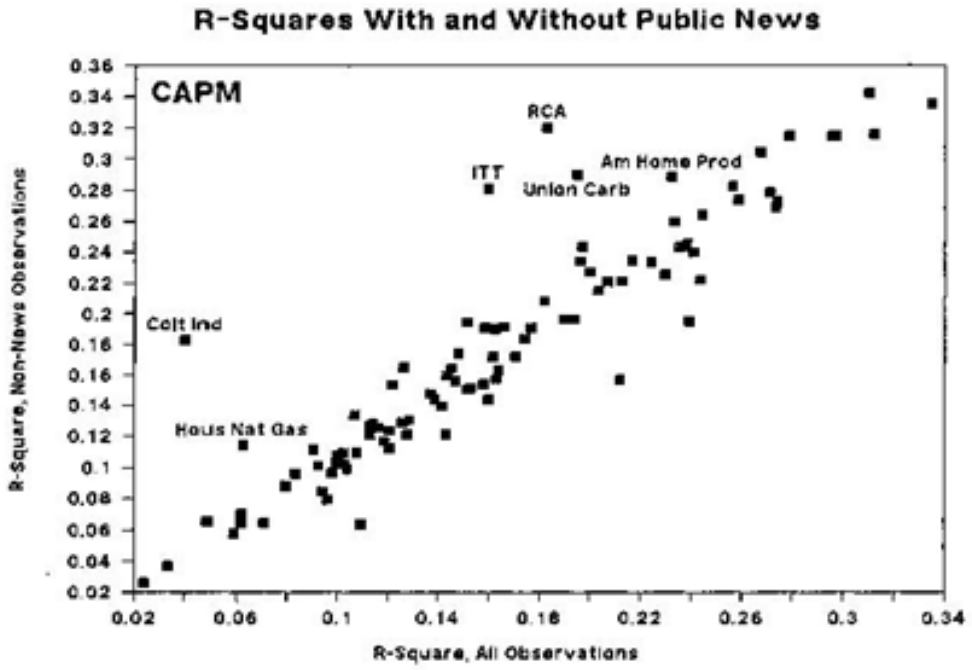
- In efficient markets (EMH), stock prices should:
  1) React accurately to news about fundamentals ($\Delta E(x)$)
  2) Not react to irrelevant facts/rumors (unrelated to *x*)

- EMH's testable implications for news data
  1) Large price moves should occur on major news days
  2) Prices shouldn't move much on non-news days

UCLAAnderson
SCHOOL of MANAGEMENT

Influential Early Test

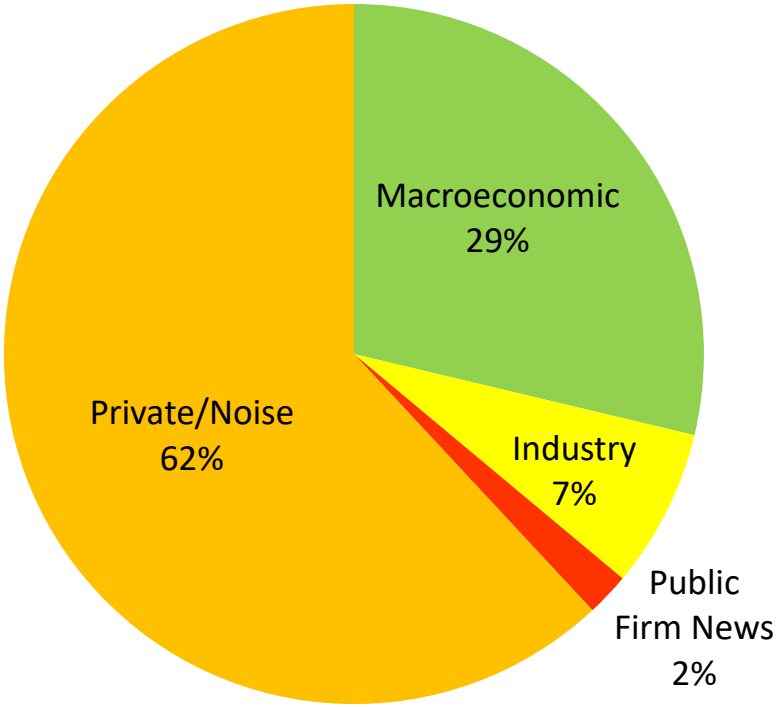
- Roll (1988, *JF*) finds that firm-specific return variance is similar on news and non-news days
  - I.e., R^2s from models of firm returns are similar



UCLAAnderson
SCHOOL of MANAGEMENT

Leading Interpretation of Roll (1988)

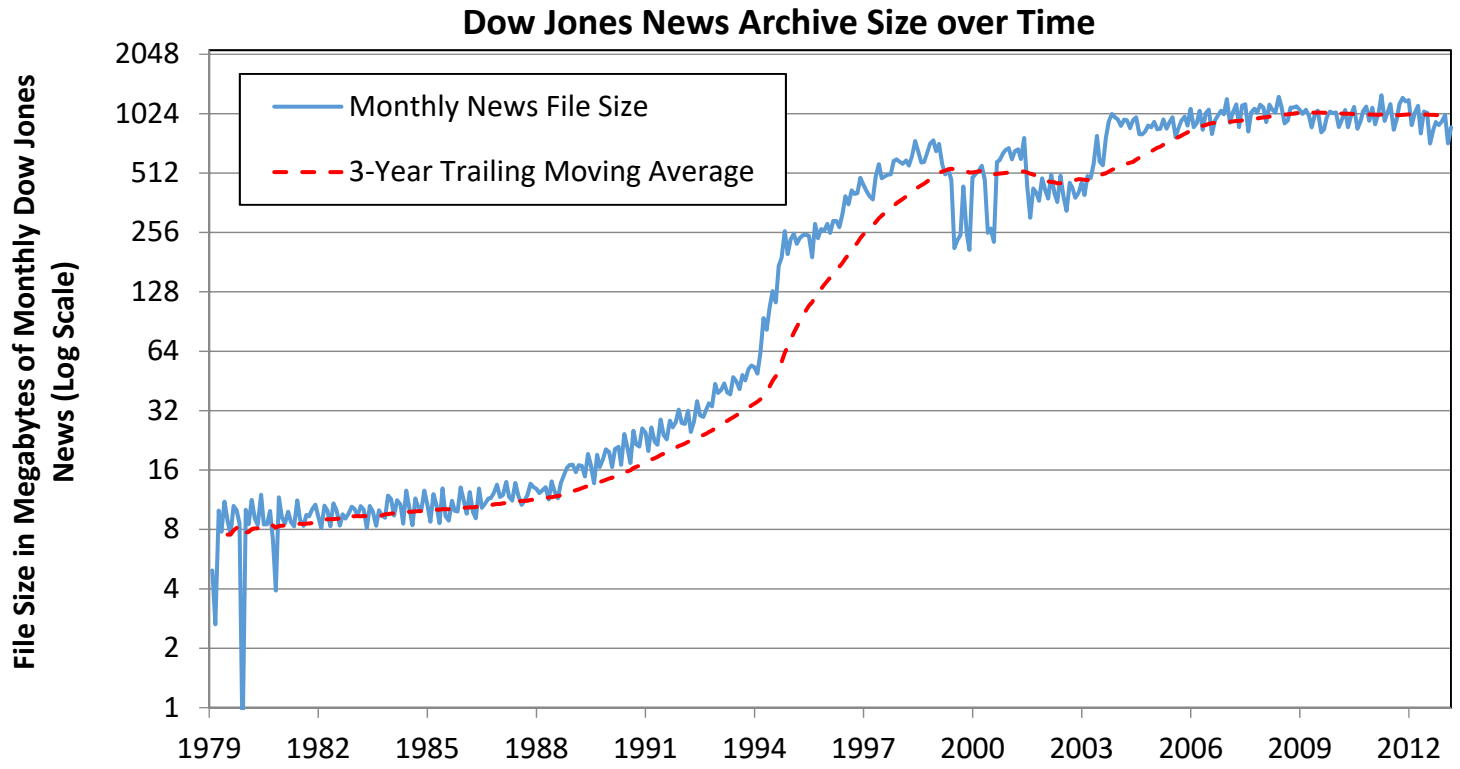
- Public news explains little return variation
  - Most variation comes from private information or noise trading (e.g., liquidity- or sentiment-driven)



UCLAAnderson
SCHOOL of MANAGEMENT

Alternative View

- News measure is a poor reflection of public info
  - Amount of news and ability to process it have increased



UCLAAnderson
SCHOOL of MANAGEMENT

From News Occurrence to Content

- Non-news days are now rare, even for firms
  - Large firms have news every day
    - Internet content continuously evolves
  - Market news occurs every day

- Interpreting content could help explain returns
  - Market and firm-level returns are related to content
    - Tetlock (2014, *ARFE*) surveys the role of media in finance
  - Content comes in linguistic, numeric, and other forms
    - Loughran and McDonald (2015) survey textual analysis

UCLAAnderson
SCHOOL of MANAGEMENT

Interpreting Recent Major News

News Event Description	Categorization	Reaction
Tsunami in Japan (3/11/2011)	Natural Disaster	-17% (2 days)
EU Aid to Spain (6/11/2012)	Central Bank Policy	+2% to +6%
Taper Tantrum (5/22/2013)	Central Bank Policy	-2% to -4%
Flash Crash (5/6/2010)	Trader News	-9% (intraday)
Fake Tweet (4/23/2013)	Conflict	-1% (intraday)

UCLAAnderson
SCHOOL of MANAGEMENT

Interpreting Headlines: 1941 to 1987

Five Important News Events with the Biggest Market Reactions

Rank	Date	Reaction	NY Times Headline News
1	9/26/1955	-6.62%	Eisenhower suffers heart attack
2	6/26/1950	-5.38%	North Korea invades South Korea
3	11/3/1948	-4.61%	Truman defeats Dewey
4	12/8/1941	-4.37%	Japanese bomb Pearl Harbor
5	11/26/1963	3.98%	Orderly transfer of power to Johnson (from Kennedy)

Five Important News Events with the Smallest Market Reactions

Rank	Date	Reaction	NY Times Headline News
1	10/24/1983	0.02%	US Marines killed in Lebanon
2	11/4/1964	-0.05%	Johnson defeats Goldwater
3	5/9/1960	0.09%	U-2 plane shot down; US admits spying
4	12/26/1979	0.11%	Soviet Union invades Afghanistan
5	11/8/1944	-0.15%	Roosevelt defeats Dewey

Source: CRSP Index; Cutler, Poterba, and Summers (J. of Port. Mgmt., 1989)

UCLAAnderson
SCHOOL of MANAGEMENT

Largest Price Moves: 1941 to 1987

Rank	Date	Return	NY Times Explanation
1	10/19/1987	-20.47%	Worry over dollar decline and trade deficit; fear of US not supporting dollar
2	10/21/1987	9.10%	Interest rates continue to fall; deficit talks in Washington; bargain hunting
3	10/26/1987	-8.28%	Fear of budget deficits; margin calls; reaction to falling foreign stocks.
4	9/3/1946	-6.73%	"No basic reason for the assault on prices."
5	5/28/1962	-6.68%	Kennedy forces rollback of steel price hike.
6	9/26/1955	-6.62%	Eisenhower suffers heart attack.
7	6/26/1950	-5.38%	Outbreak of Korean War.
8	10/20/1987	5.33%	Investors looking for "quality stocks"
9	9/9/1946	-5.24%	Labor unrest in maritime and trucking industries.
10	10/16/1987	-5.16%	Fear of trade deficit; fear of higher interest rates; tension with Iran

Source: CRSP Index; Cutler, Poterba, and Summers (J. of Port. Mgmt., 1989)

UCLAAnderson
SCHOOL of MANAGEMENT

Largest Price Moves: 1988 to 2012

Rank	Date	Return	News Explanation
1	10/13/2008	11.49%	Governments throughout the world announce moves to support troubled banks.
2	10/28/2008	9.53%	Late rally on Wall Street as rebound in stocks defies latest economic news.
3	10/15/2008	-8.98%	Falling retail sales and rising wholesale prices spike fears of recession and erase Monday's record rally.
9	8/8/2011	-6.87%	Fearful investors reacted to the US losing its coveted AAA credit rating.
20	1/8/1988	-5.54%	Stocks were hammered by worries about the economy and possibly stricter market regulation.
24	7/29/2002	5.32%	Investors, growing less worried about earnings and corporate bookkeeping, are ready to buy again.
25	1/3/2001	5.29%	Citing signs of economic slowdown, the Fed unexpectedly cut its target to 6%.
31	8/4/2001	-5.03%	Lukewarm jobs report hints that economic deterioration may be stagnation; triggers Fed worries.

Source: CRSP Index; Cornell (J. of Port. Mgmt., 2013)

UCLAAnderson
SCHOOL of MANAGEMENT

Staying "Abreast of the Market"

- Many journalists (and traders) claim to know why the market moved—at least, in hindsight

Fed policy

Housing market

Oil supply

Exchange rates

Innovation

War

UCLAAnderson
SCHOOL of MANAGEMENT

A Simple Content Measure

- Tetlock (2007, *JF*) measures the frequency of positive and negative words in a daily column
  - WSJ – "Abreast of the Market" column
  - Negative words in psychosocial dictionaries include:
    - "fear," "worry," "disappoint," "collapse," "flaw," and "ruin"
  - Compute the relative frequency of each category
    - E.g., **negativity** = negative words / total words
- Stock prices react more to negative words
- Does the market respond appropriately?

UCLAAnderson
SCHOOL of MANAGEMENT

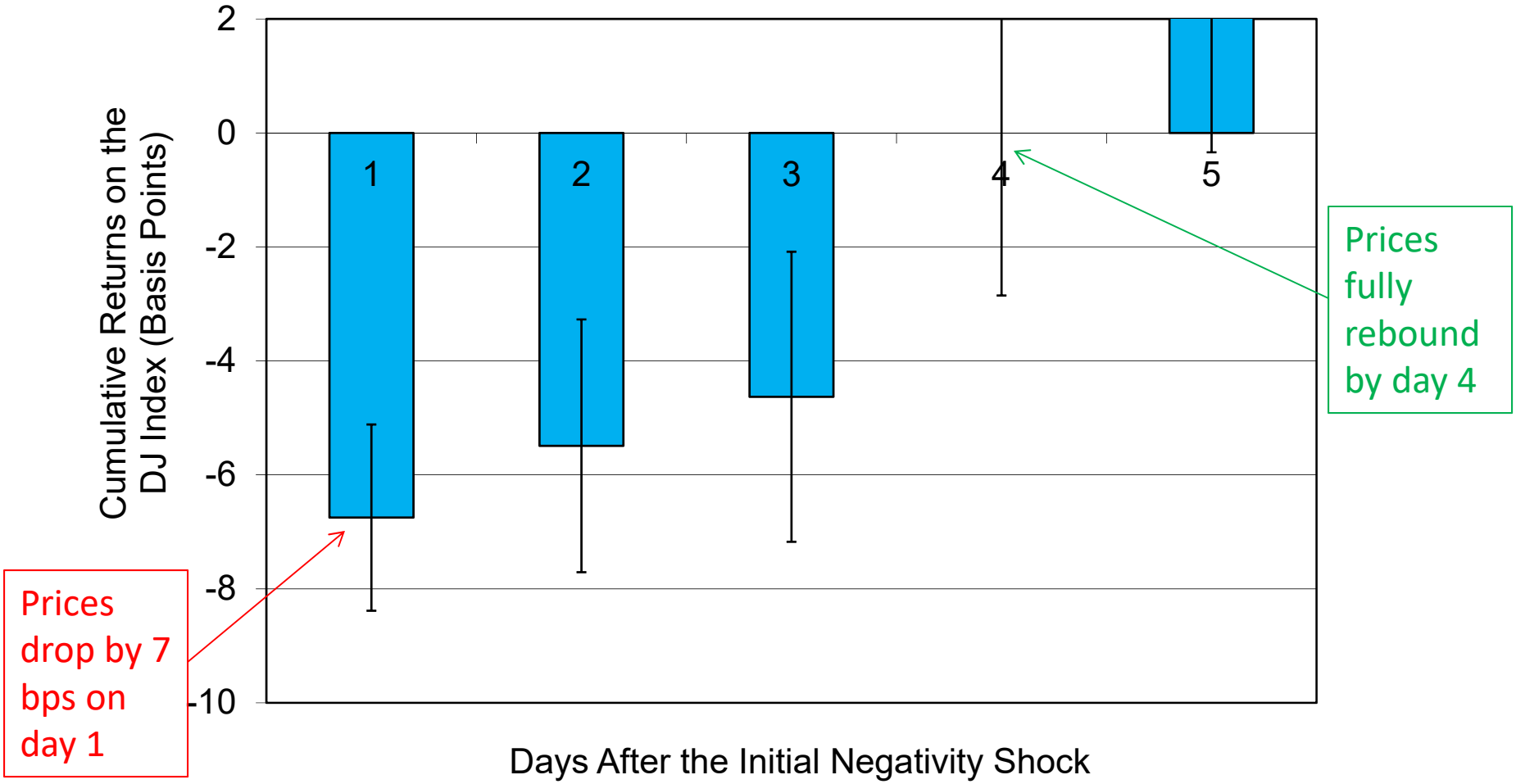
Example: Quantifying Content

- WSJ "Abreast of the Market" column on Feb 17, 2009
  - Headline: *Market's `Hope Balloon' **Loses** Air; Tepid Upturns Haven't Stopped the Slide*
  - Financial markets are supposedly driven by two **competing** forces: **fear** and greed. **Fear** just made another **grab** for the steering wheel.
  - **Disappointment** with the government's planned credit-market bailout and **concerns** that the $787 billion stimulus plan won't jolt the economy fast enough snuffed out the budding stock-market rally. Now investors are **worried** that stocks could fall back to their November **lows** -- and possibly even farther.

- Method: Compute negativity in each day's column
  - E.g., 9 negative / 82 words = 11.0% -- much higher than usual

UCLAAnderson
SCHOOL of MANAGEMENT

Negativity Predicts DJ Index



UCLAAnderson
SCHOOL of MANAGEMENT

Why Do Investors Overreact?

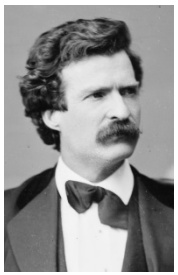
- Journalists' powerful techniques influence beliefs
  - Use evocative imagery
  - Use emotional language
  - Focus on people

- Study of *WSJ* "Abreast of the Market" (1970-2007)
  - Different journalists write the column each day
  - Journalists differ in their writing styles (e.g., optimism)
  - Stock prices increase after days with an optimistic author
  - See Dougal, Engelberg, Garcia, and Parsons (2012, *RFS*)

UCLAAnderson
SCHOOL of MANAGEMENT

Which News Is Informative?

- "If you don't read the newspaper, you are uninformed; if you do … , you are misinformed."
  - Mark Twain, writer



- "It's amazing that the amount of news that happens in the world every day just exactly fits the newspaper."
  - Jerry Seinfeld, comedian

UCLAAnderson
SCHOOL of MANAGEMENT

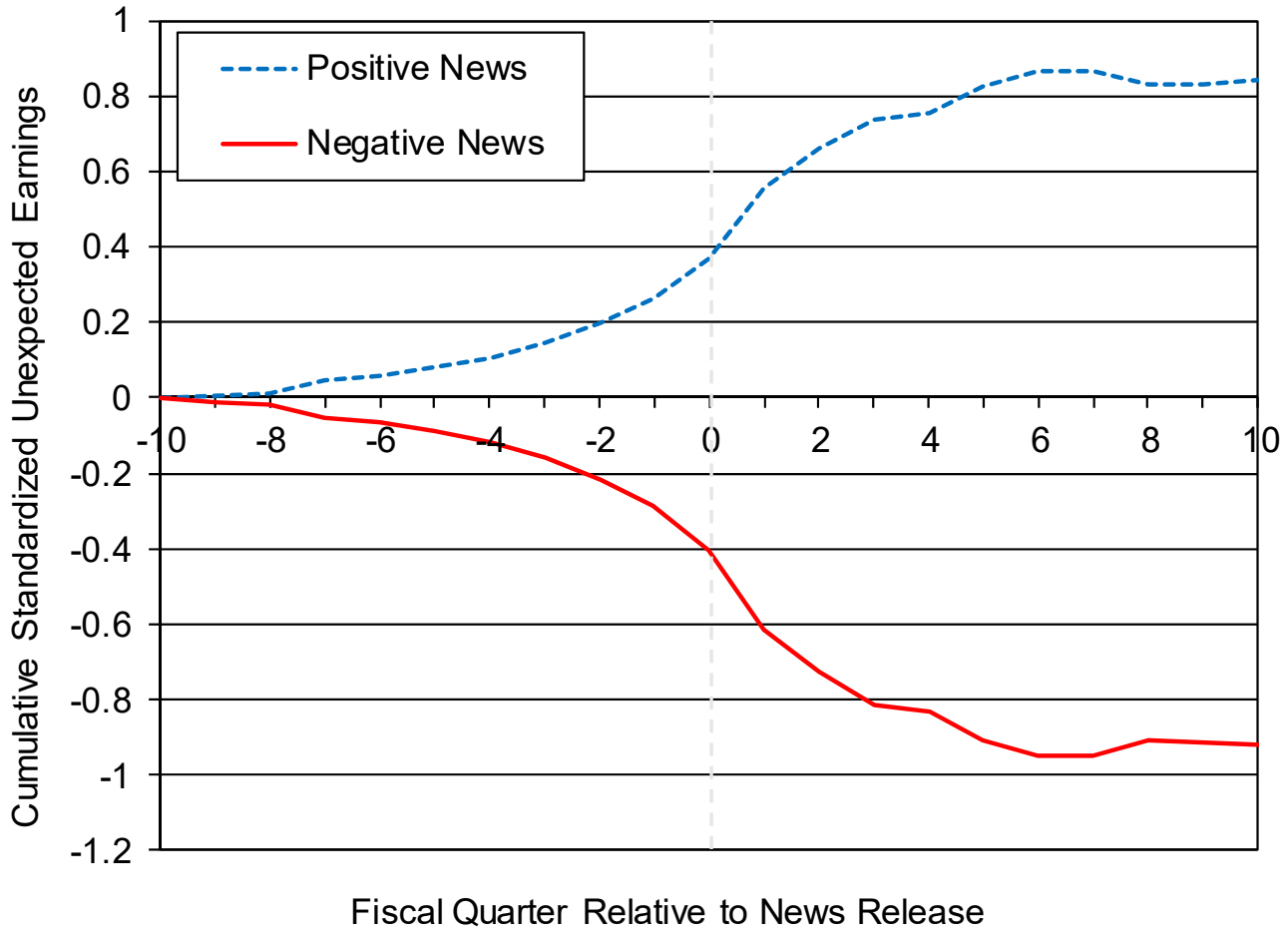
How Informative Is Firm News?

- Much financial news is genuinely informative
  - May be related to firms' fundamental values (*x*)
  - Most news about firms doesn't make the front page

- Tetlock et al. (2008, *JF*) analyze firm news
  - *DJ* newswire and *WSJ* stories about S&P 500 firms
  - Compute daily negativity scores for these stories
  - Examine outcomes before and after negative stories
    - Firms' earnings
    - Firms' stock prices

UCLAAnderson
SCHOOL of MANAGEMENT

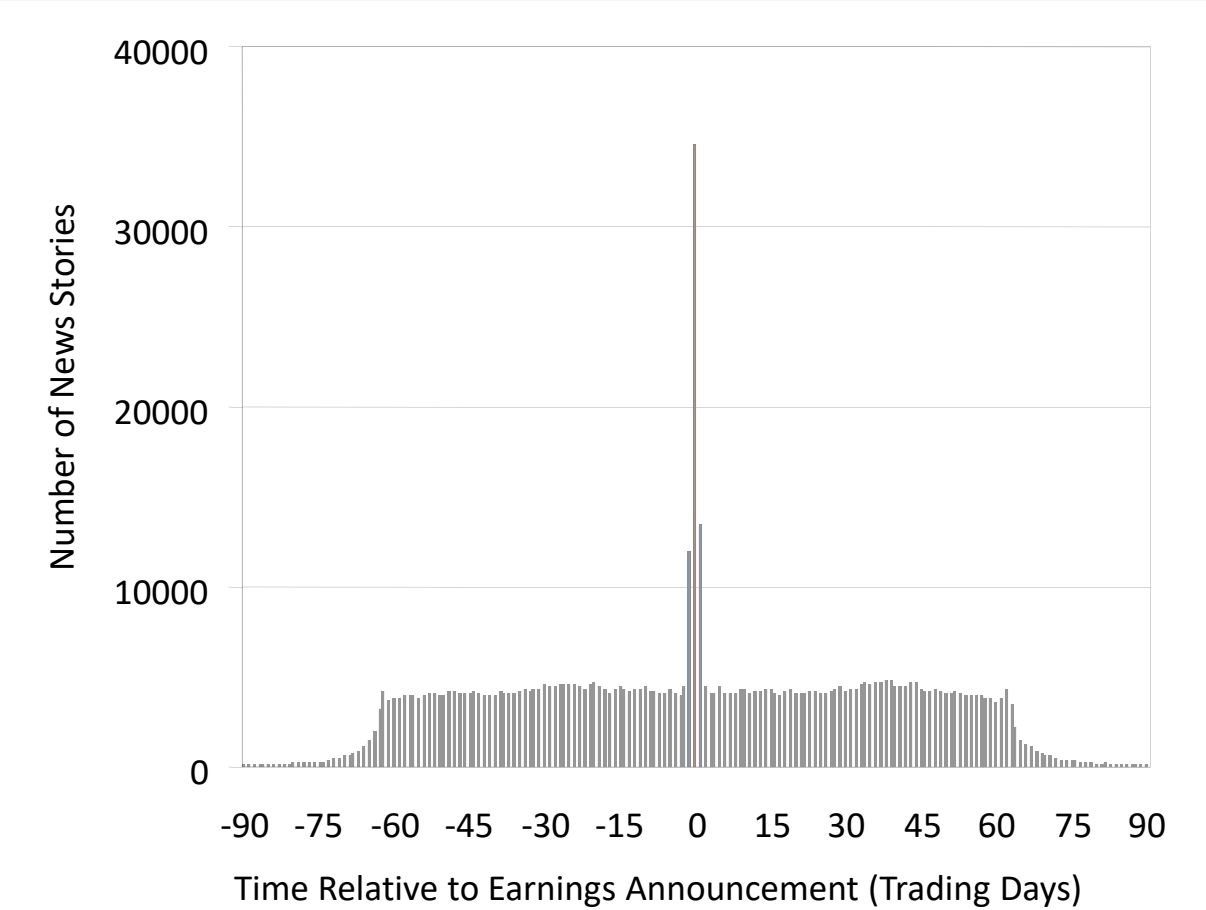
News Content Predicts Firm Earnings



UCLAAnderson
SCHOOL of MANAGEMENT

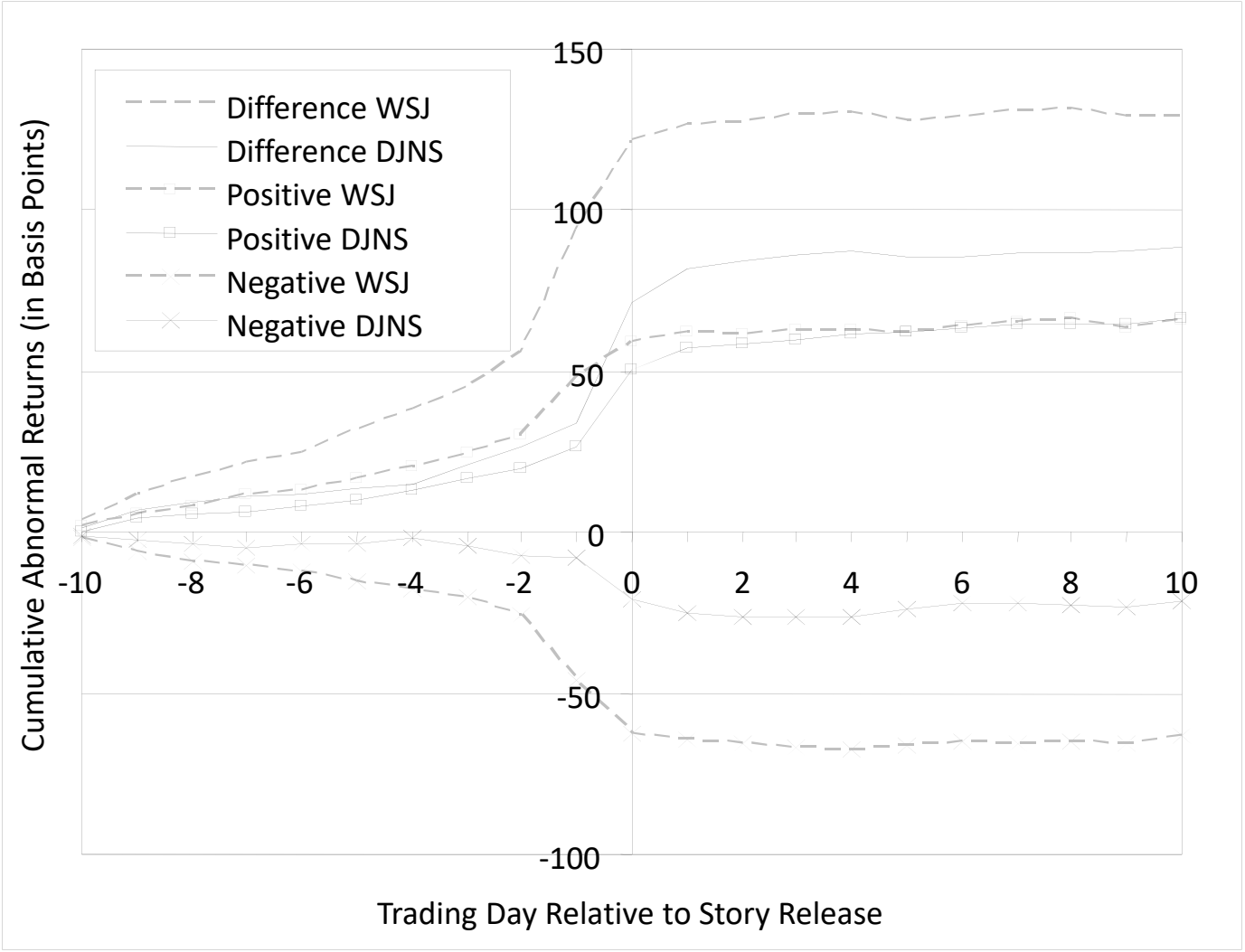
Fundamental Content Is Important

- Coverage patterns suggest earnings news is important



UCLAAnderson
SCHOOL of MANAGEMENT

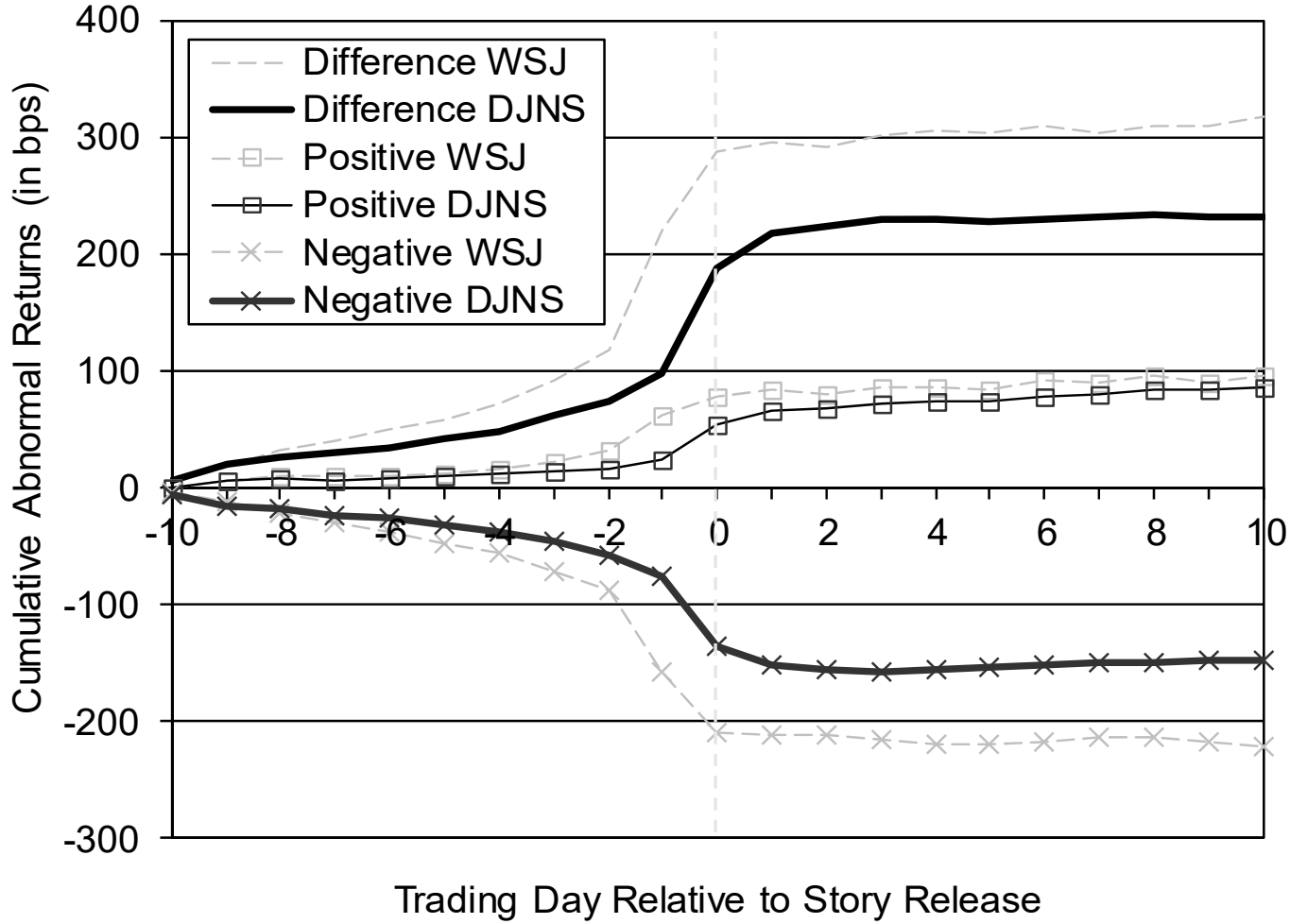
Brief Underreaction to Information



UCLAAnderson
SCHOOL of MANAGEMENT

Underreaction to Earnings News

- Focus on stories that mention "earnings"



UCLAAnderson
SCHOOL of MANAGEMENT

Interpreting Firm News

- Investors can't attend to all relevant news
  - Underreact to relevant news that's not featured
  - Underreaction increases with news relevance
    - Stories about earnings are especially relevant for firm value

- Lack of attention can cause additional biases
  - Failure to distinguish new news from "stale" news
    - Market prices should react more to new news
    - A.J. Liebling – "People everywhere confuse what they read in the newspapers with news."

UCLAAnderson
SCHOOL of MANAGEMENT

Recognizing Stale News

- Tetlock (2011, *RFS*) study of stale news
  - Data: DJ news archive from 1996 to 2008
  - Staleness = similarity of a story to previous stories
    - E.g., 90 words overlap / 150 words = 60% staleness

- Key findings
  - Stock prices react less to stale stories
    - Presumably, stale stories are less informative
  - Still, prices overreact to stale stories
    - Price reactions to stale stories tend to reverse

UCLAAnderson
SCHOOL of MANAGEMENT

Extreme Case of Stale News

- Consider market activity in United Airlines' stock
  - United Airlines filed for bankruptcy in 2002
    - Two published studies of the market reaction(s) to this event
- 2002 United bankruptcy story was new
  - ~100% stock price decline; no rebound
- The firm exited bankruptcy in 2006
  - On Sept. 7, 2008, United's stock market cap is $1.6B

UCLAAnderson
SCHOOL of MANAGEMENT

United Stock on Sept 8, 2008

- *Google News* posts a 6-year-old *Chicago Tribune* story about United's 2002 bankruptcy
  - United's stock falls 76% within minutes
    - United rebounds, but remains down 11% on the day

UCLAAnderson
SCHOOL of MANAGEMENT

Key Lessons from Research

- Trading activity and price movements are related to news, but it's hard to link them

- Market prices reflect both news and noise
  - Overreact to non-information
    - Sensationalist news that grabs investor attention
    - False or stale news when investors aren't paying attention
  - Underreact to genuine information
    - Substantive news—e.g., news about earnings
    - News that's not featured—e.g., firm news in the back pages

UCLAAnderson
SCHOOL of MANAGEMENT

Rest of today

a. Sentiment

b. Using text in regression-based forecasting models

c. It's all about the idea: A text-based trading strategy based on investor *inattention* and "lazy prices"

d. Big data text analysis: Automated downloads from the EDGAR database

e. Group projects

UCLAAnderson
SCHOOL of MANAGEMENT

a. Sentiment

Investor sentiment

- Baker and Wurgler (Journal of Finance, 2006) construct an index intended to capture overall "investor sentiment"

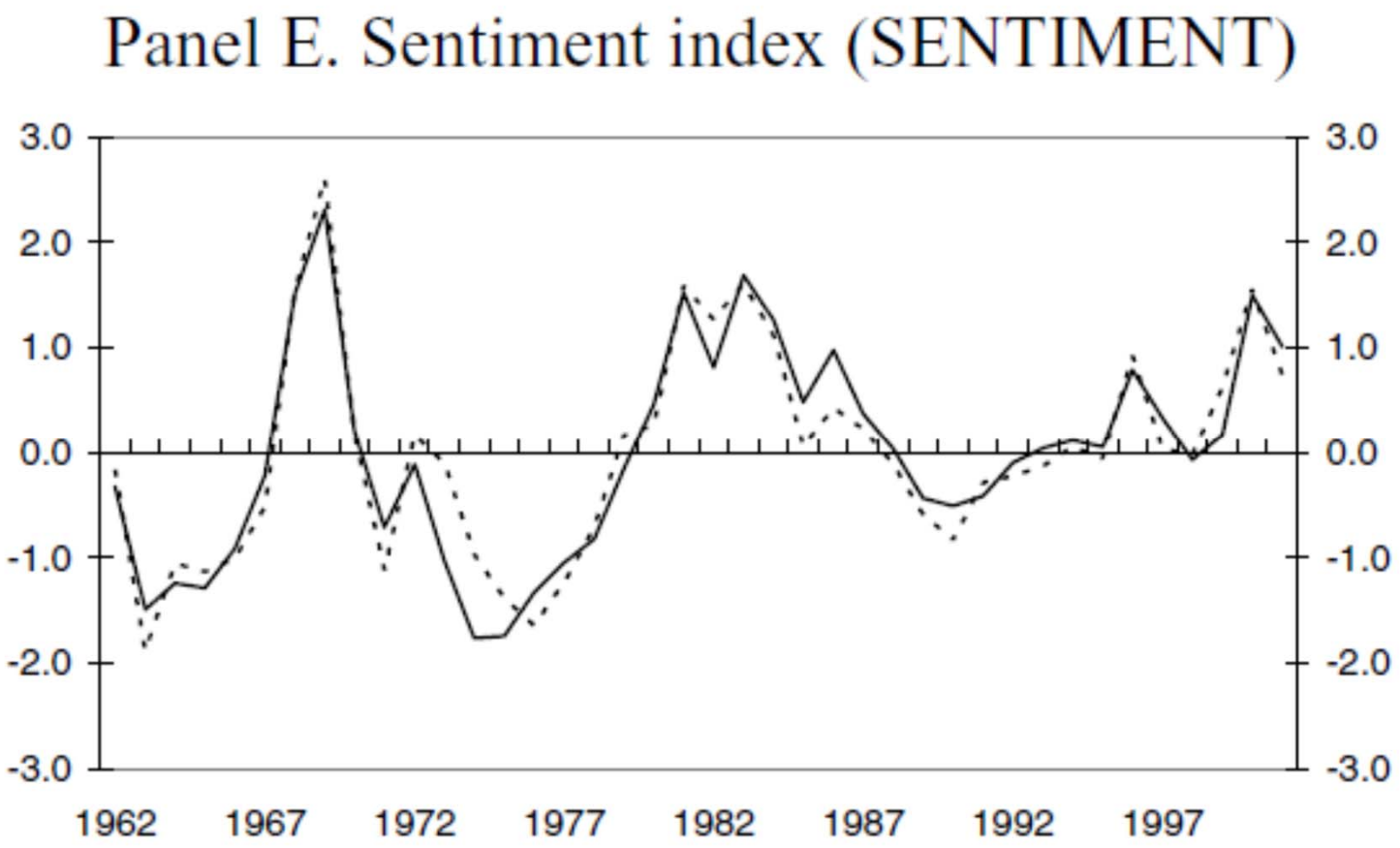
- Roughly speaking, are investors optimistic or pessimistic about business activity and growth?
  - When sentiment is high, they are too optimistic (irrational exuberance)

- Index is a composite of:
  - The closed-end fund discount, NYSE share turnover, the number and average first-day returns on IPOs, the equity share in new issues, and the dividend premium (see paper, pages 1655-1656, for more information)

- The authors find that when sentiment is low, subsequent returns on small stocks, young stocks, high volatility stocks, unprofitable stocks, non-dividend-paying stocks, extreme growth stocks, and distressed stocks are high.
  - High returns are relative to the return on stocks with 'opposite' characteristics; e.g. large stocks, old stocks, low vol, etc.

UCLAAnderson
SCHOOL of MANAGEMENT

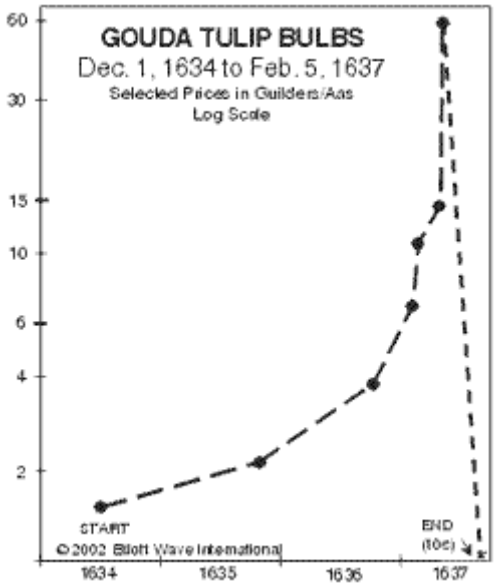
a. B&W Sentiment Index



UCLAAnderson
SCHOOL of MANAGEMENT

a. Sentiment as a general concept

- The idea of investor sentiment affecting valuations (and thus subsequent returns) is an old one, including mentions by Keynes (1936)
  - The Tulip Mania (1636 to 1937 things went truly nuts) is an early example of bubble driven by investor expectations
- Many investors now seek to identify investor sentiment at higher frequencies than B&W's index using alternative data sources
  - We mentioned Ravenpack in the last lecture as a provider of such sentiment indexes.
- In the following, we will create a sentiment indicator based on Dow Jones Newswire headlines.
  - This is Problem Set 6
  - Effectively, we are trying to create an index that helps predicting returns using Supervised Learning (with our usual regression-based forecasting methods)



UCLAAnderson
SCHOOL of MANAGEMENT

a. News Data

- Many papers have used news data to capture investor mood and information
  - Presumably (a) investors get information in part from the news, and (b) presumably the news reflect in part what investors are discussing as current topics

- I have downloaded DJIA Headline News from www.kaggle.com
  - Kaggle.com is a really fun web-site that has cool data and machine learning and analytics code to analyze. If you haven't yet, I encourage you to check it out.

- The data is in DJIA_Headline_News.csv on CCLE under Week 7.
  - Data is from 8/8/2008 until 7/1/2016
  - There are 25 headlines each day
  - There is also an indicator of whether the Dow Jones go up or down that day
  - Next slide shows what data looks like

UCLAAnderson
SCHOOL of MANAGEMENT

a. News Data

- An example of a headline:
  - "*Georgia 'downs two Russian warplanes' as countries move to brink of war*"

- Data example: (Top1 is top headline, Top2 is second to top headline, etc., until Top25)

Date	Label	Top1	Top2	Top3	Top4	Top5
8/8/2008	0	b"Georgia	b'BREAKIN	b'Russia T	b'Russian	b"Afghan
8/11/2008	1	b'Why wo	b'Bush pu	b"Jewish (	b'Georgia	b"Olympi
8/12/2008	0	b'Rememl	b"Russia '	b'"If we h	b"Al-Qa'e	b'Ceasefir
8/13/2008	0	b' U.S. ref	b"When tl	b' Israel cl	b'Britain\'	b'Body of
8/14/2008	1	b'All the e	b'War in S	b'Swedish	b'Russia e	b'Missile

- Label is 0 if Dow Jones Index goes down during the same day as the headline, 1 otherwise

- Ultimate goal: can we construct a model that uses text (headline) data to predict Dow Jones next day returns?
  - Sentiment-related idea: If people become more optimistic (pessimistic), they buy (sell), and prices will go up (down).

UCLAAnderson
SCHOOL of MANAGEMENT

a. News Data: Download and Cleaning

```
# Load data
data <- read.csv("DJIA_Headline_News.csv", stringsAsFactors = FALSE)

# Run pre-processing code from CleanUpScript_PS5.R Make 'Date' column a Date
# object to make train/test splitting easier
data$Date <- as.Date(data$Date)

# Combine headlines into one text blob for each day and add sentence separation token

data$all <- paste(data$Top1, data$Top2, data$Top3, data$Top4, data$Top5, data$Top6,
data$Top7, data$Top8, data$Top9, data$Top10, data$Top11, data$Top12, data$Top13,
data$Top14, data$Top15, data$Top16, data$Top17, data$Top18, data$Top19,
data$Top20, data$Top21, data$Top22, data$Top23, data$Top24, data$Top25, sep = " <s> ")

## Get rid of those pesky b's and backslashes you see if you inspect the raw data
data$all <- gsub("b\"|b'|\\\\|\\\"", "", data$all)

## Get rid of all punctuation except headline separators, alternative to cleaning done in tm-package
data$all <- gsub("([<>])|[[:punct:]]", "\\1", data$all)

## Reduce to only the three columns we need.
data <- data[, c("Date", "Label", "all")]
```

UCLAAnderson
SCHOOL of MANAGEMENT

a. News Data: Create Corpus and Clean

- Use *VectorSource(data)* to read in Corpus such that each line in data is treated as its own document. This is important as we need to have each entry in the Corpus correspond to a particular date
  - Compare to use of *DirSource* in Topic 5, where each text file became an entry in the corpus.

```
Corpus <- Corpus(VectorSource(data$all))

# Cleaning procedure
Corpus = tm_map(Corpus, removePunctuation)
Corpus = tm_map(Corpus, content_transformer(gsub), pattern = "\t", replacement = " ")

# Use 'Regular Expressions' to only keep letters and numbers
Corpus = tm_map(Corpus, content_transformer(gsub), pattern = "[^a-zA-Z0-9 ]", replacement = " ")


Corpus <- tm_map(Corpus, removeNumbers)
Corpus <- tm_map(Corpus, tolower)
Corpus <- tm_map(Corpus, removeWords, c(stopwords(kind = "SMART"), "<s>"))
Corpus <- tm_map(Corpus, stripWhitespace)
```

UCLAAnderson
SCHOOL of MANAGEMENT

a. News Data: Create DTM

- Next, create DocumentTermMatrix to do analysis on terms used in each document.
  - Note: TermDocumentMatrix is the transpose of DocumentTermMatrix

```
dtm <- DocumentTermMatrix(Corpus)
inspect(dtm[5:10, 801:810])
```

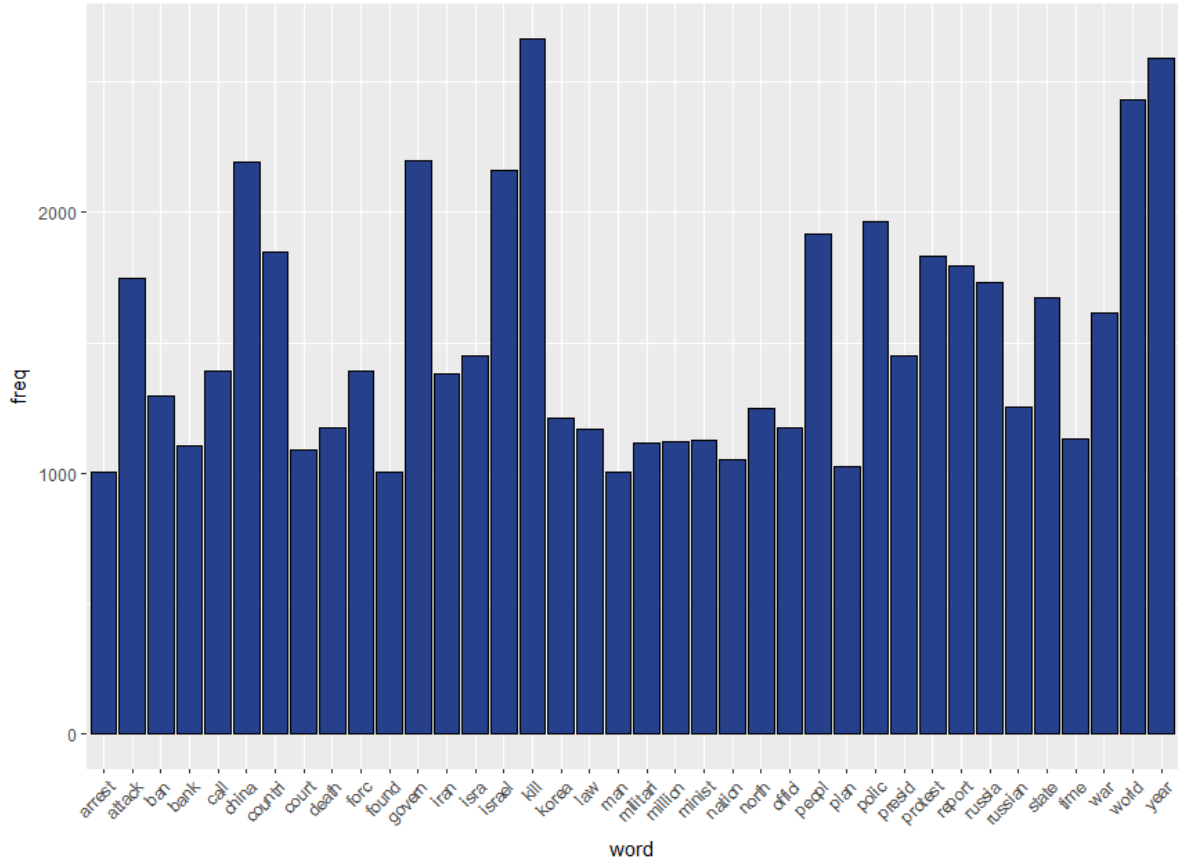
```
Docs biodiesel bring brother bushmccain carrier council crash current cut dead
  10         0     1       0          0       0       0     0       0   0    0
  5          0     0       0          0       0       0     0       0   0    0
  6          0     0       0          0       0       0     0       0   0    0
  7          0     0       0          0       0       0     0       0   0    0
  8          0     0       0          0       0       0     0       0   0    0
  9          1     1       1          1       1       1     1       1   1    1
```

```
# Organize words by frequency
freq <- colSums(as.matrix(dtm))
ord_corpus <- order(freq)

# See most common words
freq[tail(ord_corpus)]
## israel china govern world  year   kill
##  2161 2195   2197   2432 2594 2666
```

UCLAAnderson
SCHOOL of MANAGEMENT

a. Get a sense of data using word frequency

```
# Identify words that appear frequently
word_freq <- data.table(word = names(freq), freq = freq)

# Plot most frequent words along with frequency
ggplot(word_freq[freq > 1000], aes(word, freq)) + geom_bar(fill = "royalblue4", color = "black",
stat = "identity") + theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



UCLAAnderson
SCHOOL of MANAGEMENT

a. Create a WordCloud to get sense of data

```
# Plot 100 most frequent words, and add some color
wordcloud(names(freq), freq, max.words = 100, rot.per = 0.3, colors = brewer.pal(6,"Dark2"))

# note that WordCloud plots are random in terms of word location
```

Pretty clear that typical DJIA headline is about political events, such as wars, foreign affairs, legal issues, etc. Could potentially be related to stock returns and sentiment, but not obviously

UCLAAnderson
SCHOOL of MANAGEMENT

b. Create a Sentiment Indicator

- Baker and Wurgler created an index based on variables they ex ante thought to be related to investor sentiment. We can take a similar approach by pre-defining a set of 'sentiment'-related words.
  - Note: words are expressed in their stemmed form as we have stemmed the document

```
sentiment_words <- c("invest", "growth", "grow", "high", "strong", "lead", "bankrupt",
"good", "bull", "bear", "interest", "market", "hous", "rate", "oil", "loss", "weak", "low",
"fear", "poor", "risk", "stock", "debt", "financi", "fiscal", "reserv", "crash", "war", "recess")
dtm_sentiment <- dtm[, sentiment_words]
x_data_sent <- as.matrix(dtm_sentiment)


# next, compare frequencies of these words with those plotted two slides ago
setkey(word_freq, freq)
word_freq[word %in% sentiment_words]
```

UCLAAnderson
SCHOOL of MANAGEMENT

b. Create a Sentiment Indicator

- Note: the only word (see right column) that has more than 1,000 occurrences (the criterion for showing up on the word frequency plot two slides ago) is 'war'
  - Thus, again, text data is dominated by noise.
  - You are, generally speaking, going to be a lot more successful if you have a good idea how to extract the words/phrases that are most informative for your prediction or classification problem

- To reduce the dimensionality of our problem, we will use this sentiment subset of words (that is, the words given on the right) when we fit models to build our sentiment indicator

```
## word freq
## 1: bull 12
## 2: fiscal 12
## 3: bankrupt 23
## 4: weak 30
## 5: recess 78
## 6: growth 88
## 7: bear 93
## 8: stock 111
## 9: loss 113
## 10: strong 113
## 11: low 118
## 12: reserv 120
## 13: invest 128
## 14: interest 193
## 15: poor 205
## 16: debt 213
## 17: risk 232
## 18: good 238
## 19: crash 252
## 20: financi 254
## 21: market 257
## 22: rate 298
## 23: lead 351
## 24: grow 353
## 25: hous 366
## 26: fear 423
## 27: high 436
## 28: oil 857
## 29: war 1613
```

UCLAAnderson
SCHOOL of MANAGEMENT

b. A regression-based text model

- The outcome-variable (label) is binary, so a logistic regression is natural.

```
# Create full data set
y_data <- as.factor(data$Label)
x_data <- as.matrix(dtm)


# Run logistic regression
glm.fit <- glm(y_data ~ x_data_sent, family='binomial')
summary(glm.fit)
preds_logit <- as.numeric(predict(glm.fit, type='response'))
```

- Note that DTM gives word frequency of each word, each period. Thus x_data are integers with typical values 0, 1, and 2, though higher number of occurrences may happen.
- Output is on next slide. Note that only coefficients on *low* and *stock* have t-statistics higher than 2. In addition, *oil* is marginally significant

UCLAAnderson
SCHOOL of MANAGEMENT

b. Logistic regression results

```
Call:
glm(formula = y_data ~ x_data_sent, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6573  -1.2365   0.9617   1.0925   1.8276

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          0.229378   0.089433   2.565 0.010323 *
x_data_sentinvest   -0.180426   0.172342  -1.047 0.295143
x_data_sentgrowth    0.256280   0.211696   1.211 0.226046
x_data_sentgrow     -0.032557   0.109809  -0.296 0.766855
x_data_senthigh      0.099427   0.097311   1.022 0.306905
x_data_sentstrong   -0.240696   0.190047  -1.267 0.205331
x_data_sentlead      0.022544   0.105989   0.213 0.831561
x_data_sentbankrupt -0.561593   0.437678  -1.283 0.199451
x_data_sentgood     -0.023833   0.125785  -0.189 0.849724
x_data_sentbull     -0.172888   0.588551  -0.294 0.768948
x_data_sentbear      0.301502   0.197119   1.530 0.126129
x_data_sentinterest -0.223533   0.146046  -1.531 0.125877
x_data_sentmarket    0.026550   0.129133   0.206 0.837100
x_data_senthous     -0.026278   0.103165  -0.255 0.798942
x_data_sentrate      0.112261   0.109119   1.029 0.303577
x_data_sentoil      -0.105986   0.057302  -1.850 0.064372 .
x_data_sentloss     -0.036842   0.184135  -0.200 0.841418
x_data_sentweak     -0.093859   0.364760  -0.257 0.796935
x_data_sentlow      -0.633616   0.190324  -3.329 0.000871 ***
x_data_sentfear      0.142049   0.099881   1.422 0.154973
x_data_sentpoor     -0.066665   0.135019  -0.494 0.621488
x_data_sentrisk      0.007246   0.128298   0.056 0.954963
x_data_sentstock    -0.467565   0.185209  -2.525 0.011585 *
x_data_sentdebt      0.195661   0.125995   1.553 0.120442
x_data_sentfinanci   0.040153   0.121906   0.329 0.741871
x_data_sentfiscal    0.055839   0.552284   0.101 0.919467
x_data_sentreserv   -0.124275   0.175667  -0.707 0.479290
x_data_sentcrash    -0.095836   0.109599  -0.874 0.381885
x_data_sentwar      -0.011991   0.043675  -0.275 0.783659
x_data_sentrecess    0.042055   0.217861   0.193 0.846932
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

UCLAAnderson
SCHOOL of MANAGEMENT

b. The ROC Curve for the Logistic Reg.

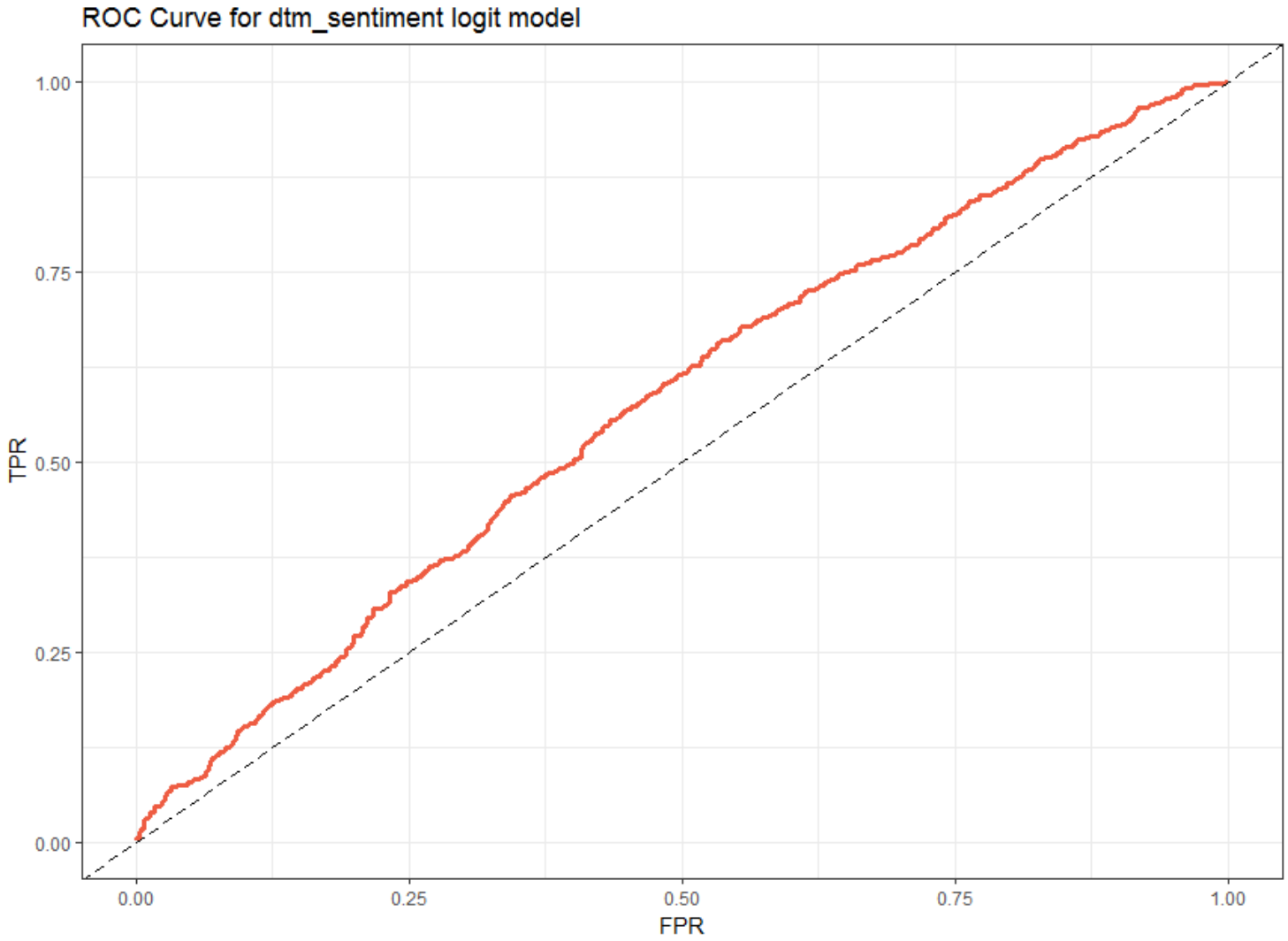
- The outcome-variable (label) is binary, so a logistic regression is natural.

```
# Compute a ROC curve
simple_roc <- function(labels, scores) {
labels <- labels[order(scores, decreasing = TRUE)]
data.frame(TPR = cumsum(labels)/sum(labels), FPR =
cumsum(!labels)/sum(!labels),labels) }

glm_roc <- simple_roc(y_data == "1", preds_logit)
TPR1 <- glm_roc$TPR
FPR1 <- glm_roc$FPR
data1 <- data.table(TPR = TPR1, FPR = FPR1)

# Plot the corresponding ROC curve
ggplot(data1, aes(x = FPR, y = TPR)) + geom_line(color = "tomato2", size = 1.2) +
ggtitle("ROC Curve for dtm_sentiment logit model") + geom_abline(slope = 1,
intercept = 0, linetype = "longdash") + theme_bw()
```

- Output is on next slide. Model is uniformly better than random.

UCLAAnderson
SCHOOL of MANAGEMENT

b. The ROC Curve for the Logistic Reg.

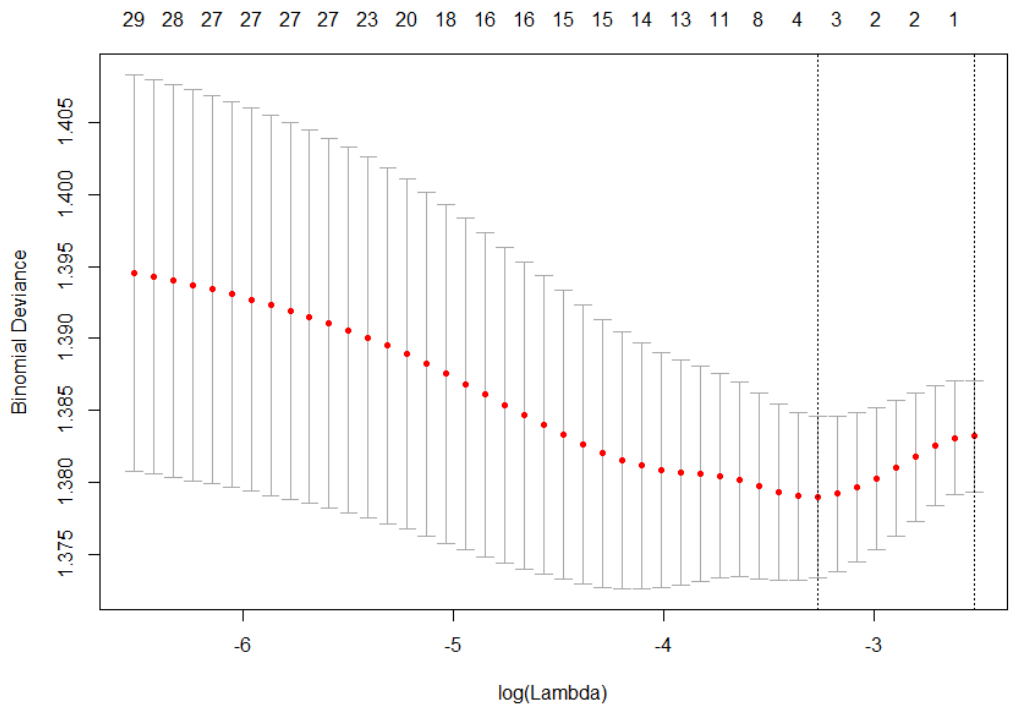


UCLAAnderson
SCHOOL of MANAGEMENT

b. A regularized regression-based text model

- Next, let's consider adding regularization (or, equivalently, a prior) to the logistic regression.
  - In particular, let's use elastic net with alpha = 0.5
  - We will choose the lambda parameter using cross-validation (Default number of folds for the cross-validation exercise is 10).

```
glmnet.fit <- cv.glmnet(x = x_data_sent, y = y_data, family = "binomial", alpha = 0.5)
plot.cv.glmnet(glmnet.fit)
```

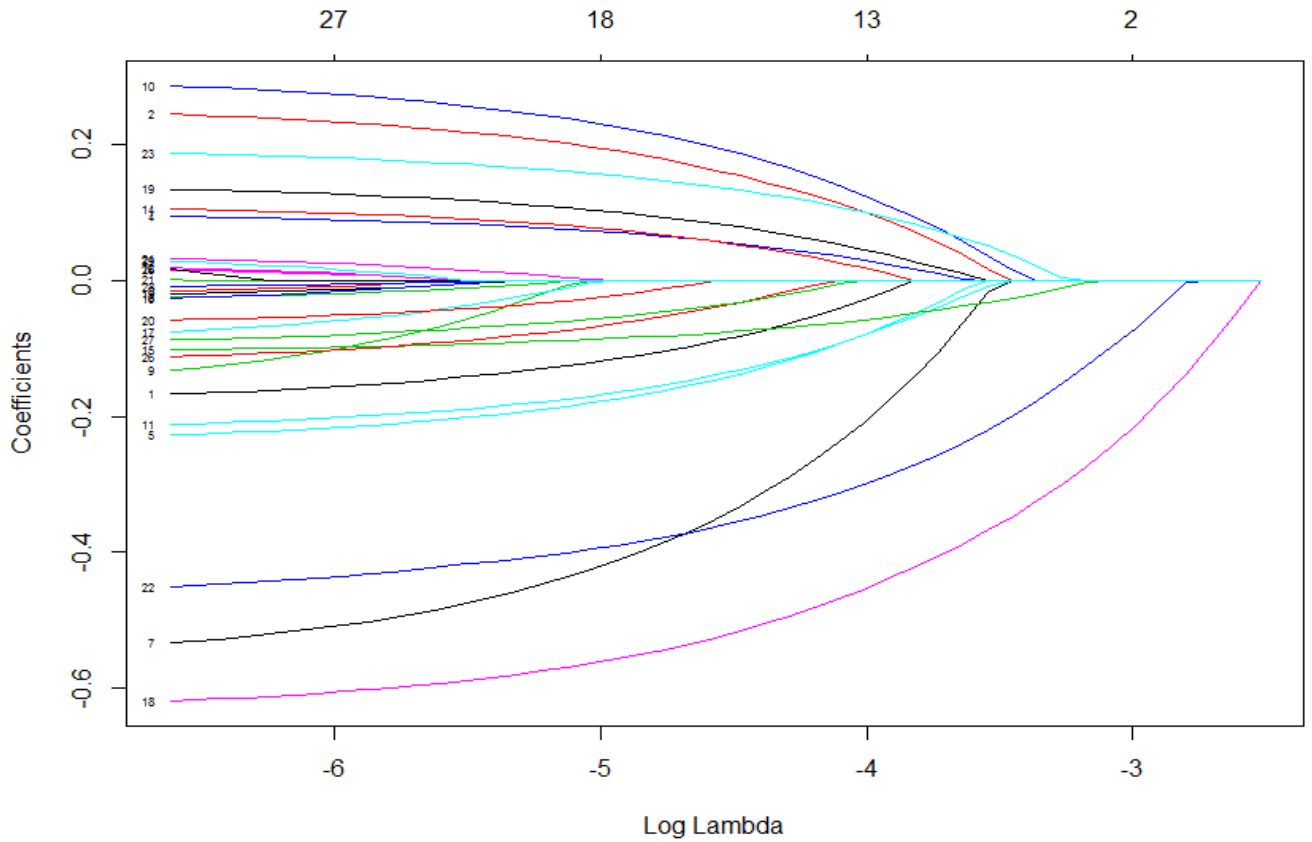


UCLAAnderson
SCHOOL of MANAGEMENT

b. A regularized regression-based text model

```
plot.glmnet(glmnet.fit$glmnet.fit, "lambda", label = TRUE)
```

 # two variables chosen: *low* (18) and *bankrupt* (7)



UCLAAnderson
SCHOOL of MANAGEMENT

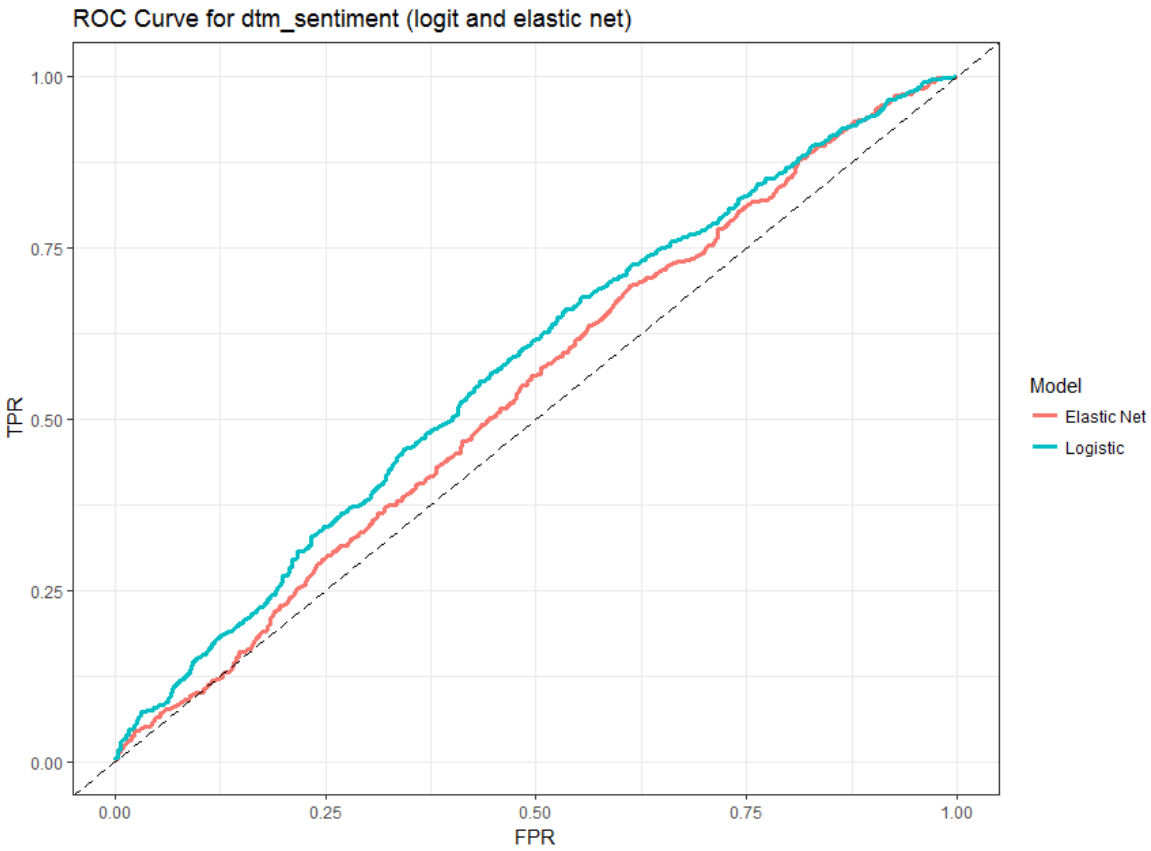
b. The ROC Curve for Regularized Logistic Reg.

```
# Prepare data
preds_enet <- as.numeric(predict(glmnet.fit, newx = x_data_sent, type = "response", s = "lambda.min"))
glmnet_roc <- simple_roc(y_data == "1", preds_enet)
TPR2 <- glmnet_roc$TPR
FPR2 <- glmnet_roc$FPR
data2 <- data.table(TPR = TPR2, FPR = FPR2)
data1[, `:=`(Model, "Logistic")]
data2[, `:=`(Model, "Elastic Net")]
data12 <- rbind(data1, data2)

# Plot the corresponding ROC curve
ggplot(data12, aes(x = FPR, y = TPR, color = Model)) + geom_line(size = 1.2) +
ggtitle("ROC Curve for dtm_sentiment (logit and elastic net)") + geom_abline(slope = 1,
intercept = 0, linetype = "longdash") + theme_bw()
```

UCLAAnderson
SCHOOL of MANAGEMENT

b. The ROC Curve for Regularized Logistic Reg.

- Note that Elastic Net performs worse
- Makes sense as this is all in the same sample
  - Regularizing hurts in-sample performance
  - Really, we want to check out-of-sample performance on a true out-of-sample period
    - Next!
- Note also that both are better than random



UCLAAnderson
SCHOOL of MANAGEMENT

b. Proper out-of-sample testing

- Split data into training data-set and proper out-of-sample (not cross-validation) data set. Let training data be data up until 2014-12-31)

```
# Create training data set
split_index <- data$Date <= as.Date("2014-12-31")
data_train <- data_full[split_index, ]
y_train <- as.factor(data$Label[split_index])

# Create out of sample test data set
data_test <- data_full[!split_index, ]
y_test <- as.factor(data$Label[!split_index])
x_train <- x_data_sent[1:length(y_train), ]
x_test <- x_data_sent[(length(y_train) + 1):length(data_full$Date), ]

# Estimate and test logistic model (no regularization)
logit.fit_train <- glm(y_train ~ x_train, family = "binomial")
preds_logit_test <- predict(logit.fit_train, type = "response", new = data.frame(x_test))
preds_logit_test <- preds_logit_test[1:length(y_test)]
```

UCLAAnderson
SCHOOL of MANAGEMENT

b. Proper out-of-sample testing

- Calculate ROC curve and see how often the model predicts correctly out-of-sample

```
## ROC curve
logit_roc_test <- simple_roc(y_test == "1", preds_logit_test)
TPR1 <- logit_roc_test$TPR
FPR1 <- logit_roc_test$FPR
data1 <- data.table(TPR = TPR1, FPR = FPR1)
prop_correct_logit <- sum(round(preds_logit_test) == y_test)/length(y_test)) # out-of-sample how often correct
## [1] 0.5291005
```

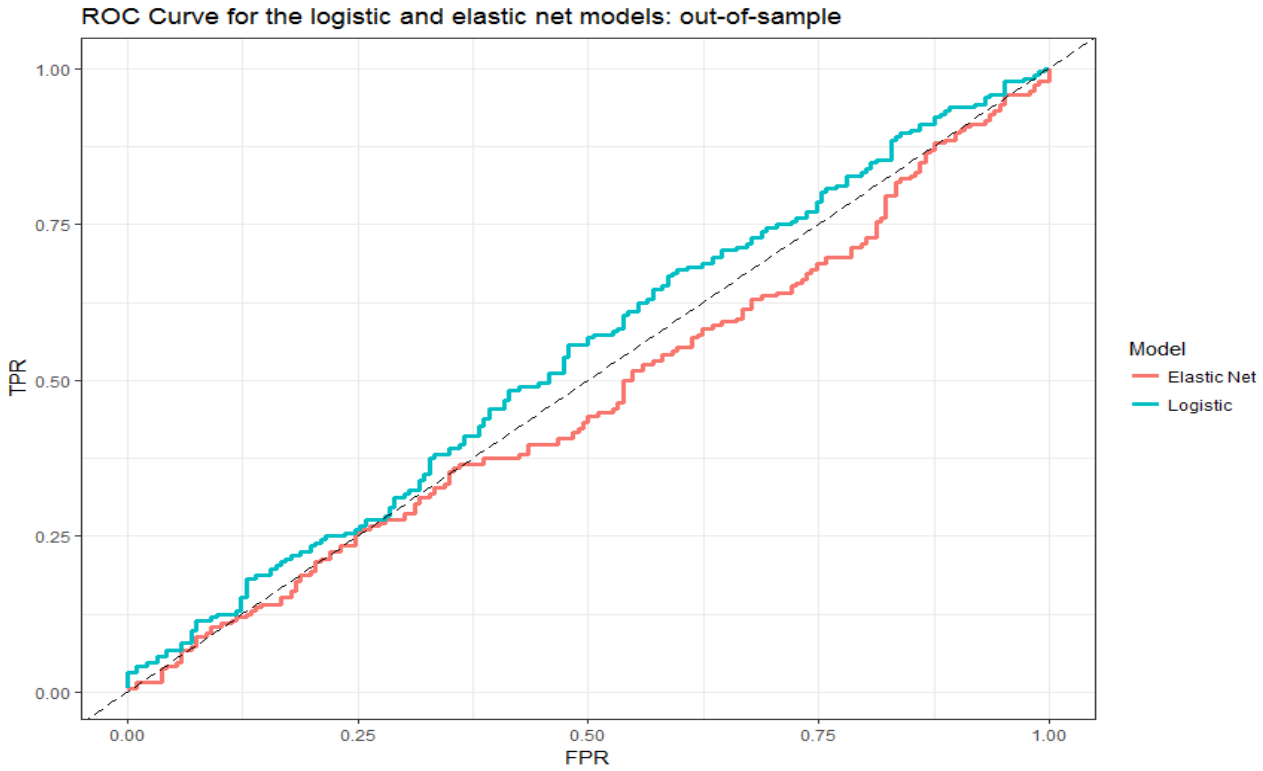
- Next, do Elastic Net model in the same way.

```
# Test elastic net model
enet.fit_train <- cv.glmnet(x = x_train, y = y_train, family = "binomial", alpha = 0.5)
preds_enet_test <- predict(enet.fit_train, newx = x_test, type = "response",
s = "lambda.1se")
## ROC curve
enet_roc_test <- simple_roc(y_test == "1", preds_enet_test)
TPR2 <- enet_roc_test$TPR
FPR2 <- enet_roc_test$FPR
data2 <- data.table(TPR = TPR2, FPR = FPR2)
(prop_correct_enet <- sum(round(preds_enet_test) == y_test)/length(y_test)) # AUC of out-of-sample elastic net model
## [1] 0.5079365
```

UCLAAnderson
SCHOOL of MANAGEMENT

b. Out-of-sample ROC Curves

- Note that unregularized logistic actually did better
  - Regularization no guarantee. Depends on stability of parameters out-of-sample.
    - I.e., some of the ones important in actual out-of-sample was 'regularized away'
    - Closer inspection (display *preds_enet_test*) reveals the prediction never changes, it's always 0.54. In other words, bankrupt and low did not appear as a headline word in the post-2014 period..!
    - This makes the ROC curve for elastic net ill-defined and the below red line in the graph should be ignored. This illustrates the importance of always looking at the data, not just some overall statistics.



UCLAAnderson
SCHOOL of MANAGEMENT

b. Lower frequency co-movement

- This model is estimated daily, but may be a good description of more low frequent stock returns

- A natural way to go to, say, the quarterly frequency is to sum up the predictions over the quarter and related to stock returns over the quarter
  - Note: we are here not trying to actual forecast quarterly stock returns, but instead see if there is comovement between our sentiment indicator and stock returns at the quarterly frequency
  - Of course, you can yourself see if there is a statistically significant forecasting relation as well (probably not, though; textual sentiment data is usually short-lived).

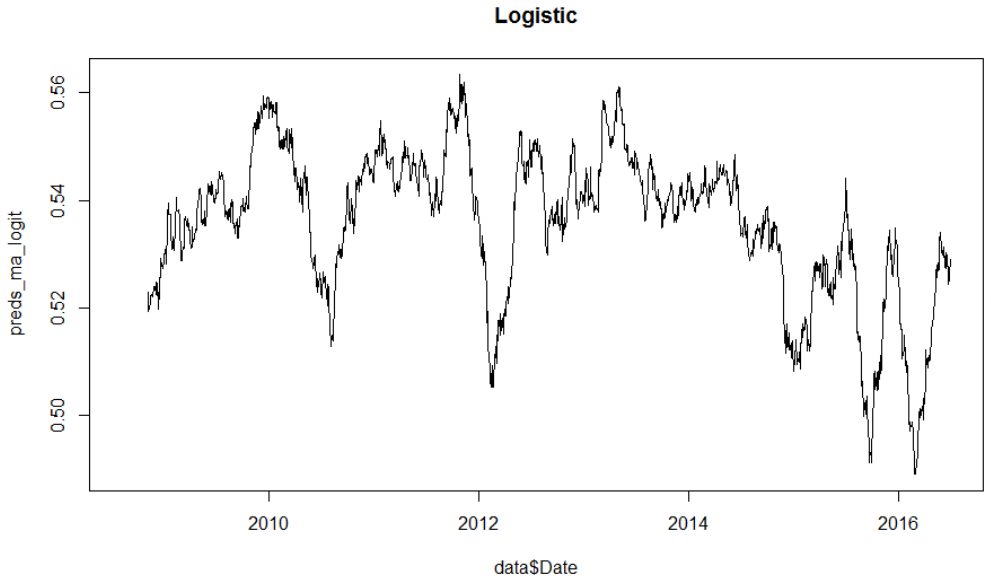
- Create 63-day (3 months) moving average of returns are model predictions, overlapping at the daily frequency

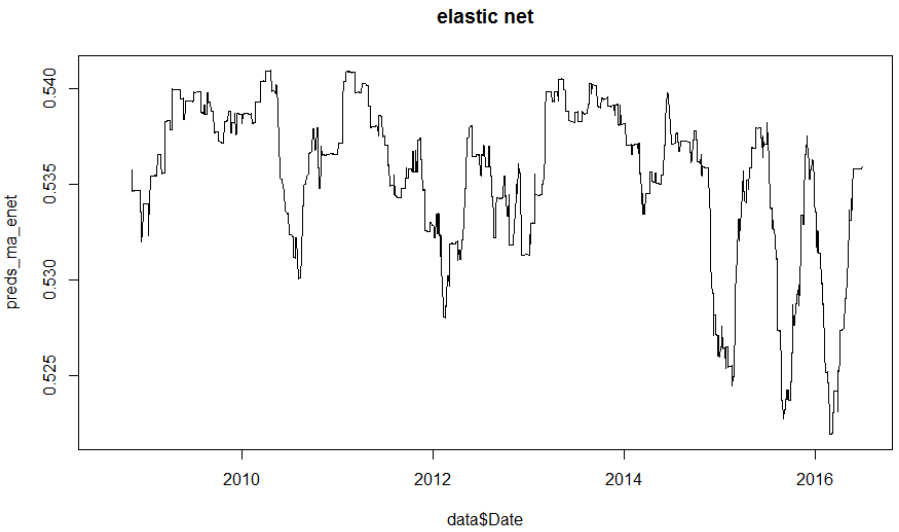
```
# Prepare data and plot
f63 <- rep(1/63, 63)
preds_ma_logit <- stats::filter(preds_logit, f63, sides = 1)
preds_ma_enet <- stats::filter(preds_enet, f63, sides = 1)
y_ma <- stats::filter(y_data, f63, sides = 1)
```

UCLAAnderson
SCHOOL of MANAGEMENT

b. Lower frequency co-movement

- Plot low frequency versions of sentiment measures
  - Looks more like Baker and Wurgler series.
  - Lower frequency plots like this are visually more intuitive and easier to relate to particular historical events
  - Notice much higher volatility of unregularized regression predictions





UCLAAnderson
SCHOOL of MANAGEMENT

b. Lower frequency co-movement

```
# Define Newey-West lag
NW_lag = 90
# Comparison for logistic
reg1 <- lm(as.numeric(y_data[-(1:62)]) ~ preds_logit[-(1:62)])
reg2 <- lm(as.numeric(y_ma) ~ preds_ma_logit)
stargazer(reg1, reg2, coeftest(reg2, NeweyWest(reg2, lag = NW_lag)), type = "text",
column.labels = c("OLS", "MA", "MA with NW SEs"), dep.var.labels.include = F)
```

```
## ================================================================
##                                      Dependent variable:
##                                 --------------------------------
##                                    OLS       OLS      coefficient
##                                                           test
##                                    OLS        MA     MA with NW SEs
##                                    (1)       (2)          (3)
## ----------------------------------------------------------------
## preds_logit[-(1:62)]             1.008***
##                                   (0.155)
##
## preds_ma_logit                              1.487***     1.487**
##                                             (0.094)      (0.710)
##
## Constant                         0.998***   0.740***      0.740*
##                                   (0.084)   (0.051)      (0.378)
##
## ----------------------------------------------------------------
## Observations                       1,927     1,927
## R2                                 0.021     0.114
## Adjusted R2                        0.021     0.114
```

UCLAAnderson
SCHOOL of MANAGEMENT

b. Lower frequency co-movement

```
# Comparison for elastic net
reg3 <- lm(as.numeric(y_data[-(1:62)]) ~ preds_enet[-(1:62)])
reg4 <- lm(as.numeric(y_ma) ~ preds_ma_enet)
stargazer(reg3, reg4, coeftest(reg4, NeweyWest(reg4, lag = NW_lag)), type = "text",
column.labels = c("OLS", "MA", "MA with NW SEs"), dep.var.labels.include = F)
```

```
## ================================================================
##                                      Dependent variable:
##                                 ---------------------------------
##                                    OLS       OLS      coefficient
##                                                          test
##                                    OLS        MA     MA with NW SEs
##                                    (1)       (2)          (3)
## ----------------------------------------------------------------
## preds_enet[-(1:62)]              2.136***
##                                  (0.457)
##
## preds_ma_enet                              4.121***      4.121*
##                                            (0.245)      (2.492)
##
## Constant                          0.394   -0.670***      -0.670
##                                  (0.245)   (0.131)      (1.334)
##
## ----------------------------------------------------------------
## Observations                      1,927     1,927
## R2                                0.011     0.128
## Adjusted R2                       0.011     0.128
## Residual Std. Error (df = 1925)   0.496     0.058
## F Statistic (df = 1; 1925)      21.880*** 283.237***
## ================================================================
## Note:                                  *p<0.1; **p<0.05; ***p<0.01
```

- Note how taking into account overlap (autocorrelation) is critical for correct standard errors.
- 13% of the variation in quarterly stock returns are reflected in DJIA headlines over this sample
- Casuality is likely mainly from stock returns to news, rather than news to stock returns, as evidence by the relatively weak forecasting results

UCLAAnderson
SCHOOL of MANAGEMENT

c. It's all about the idea: Text Similarity and Investor Inattention

- Linking text to trading strategies typically starts with a good idea (as opposed to blind data-mining)
- Investor inattention is a well-established behavioral bias
  - Intuitive: investors do not have mental capacity to keep track of all information and markets and so markets are not fully informationally efficient
- Example of inattention:
  - Predictable changes in health care sector revenues due to aging population not fully priced in stock returns, which lead to subsequent positive "alpha" for such stocks (Dellavigna and Pollet)
  - Anytime publicly available information is not impounded in prices as investors simply have not discovered the link

c. Lazy Prices

- Cohen, Malloy, and Ngyun (2015) argue that investors are inattentive to certain information in quarterly/annual reports
- In particular, firms typically repeat almost all information in their reports from their last report (copy-paste)
- Using data from EDGAR, 1994-2014, they show that active changes to the wording is associated with negative alpha of up to 22% p.a.
- The reporting changes are concentrated in the MD&A section (management discussion).
- Changes in language referring to executive team (CEO and CFO), or regarding litigation, are especially informative.
- Authors use textual analysis tools to execute their study

UCLAAnderson
SCHOOL of MANAGEMENT

c. Lazy Prices: Motivating Example

This table shows the first few paragraphs that are taken from Item 7, "Management's Discussion and Analysis of Financial Condition and Results of Operations", for Schweitzer-Mauduit International's (NYSE:SWM) 2004 and 2005 10-K reports. The new discussion in the 2005 10-K is highlighted.

10-K 2005

Outlook

Consistent with recent historical trends, worldwide cigarette consumption is expected to increase at a rate of approximately one-half to one percent per year. The anticipated decline in the production of cigarettes in developed countries is expected to be more than offset by increased cigarette production in developing countries that currently represent approximately 70 percent of worldwide cigarette production. Age demographics and expected increases in disposable income are expected to support the increased consumption of cigarettes in developing countries. In addition, the litigation environment is different in most foreign countries compared with the United States, having less of an impact on the pricing of cigarettes, which, in turn, affects cigarette consumption. Cigarette production in the United States is expected to continue to decline as a result of a decline in domestic cigarette consumption **caused by increased cigarette prices, health concerns and public perceptions. As well, cigarette consumption has declined in France and Germany following recent tax increases on cigarette sales in those countries.**

We are experiencing weakness in our tobacco-related paper sales in western Europe caused by reduced cigarette consumption in several large European markets and new cigarette paper manufacturing capacity that was added in western Europe in mid-2004. This is expected to result in increased cigarette paper machine downtime in France in 2005.

In developing countries, there is a trend toward consumption of more sophisticated cigarettes, which utilize higher quality tobacco-related papers, such as those we produce, and reconstituted tobacco leaf. This trend toward more sophisticated cigarettes reflects increased governmental regulations concerning tar delivery levels and increased competition from multinational cigarette manufacturers.

Based on these trends, we expect worldwide demand for our products to continue to increase, with a shift from developed countries to developing countries. As a result, we are increasing some of our production capacity in developing countries such as Brazil, Indonesia and the Philippines.

The new RTL production line added at our Spay, France mill, which started up in the fourth quarter of 2003, is expected to continue to contribute positively to sales volumes and operating profit in 2005.

10-K 2004

Outlook

The markets for the Company's products are expected to remain relatively stable during 2004. Trends of improvement are expected to continue in tobacco-related paper sales in several key markets. Cigarette production in the United States continues to decline as a result of declines in domestic cigarette consumption and exports of cigarettes manufactured in the United States. The anticipated decline in the production of cigarettes in developed countries is expected to be more than offset by increased cigarette production in developing countries.

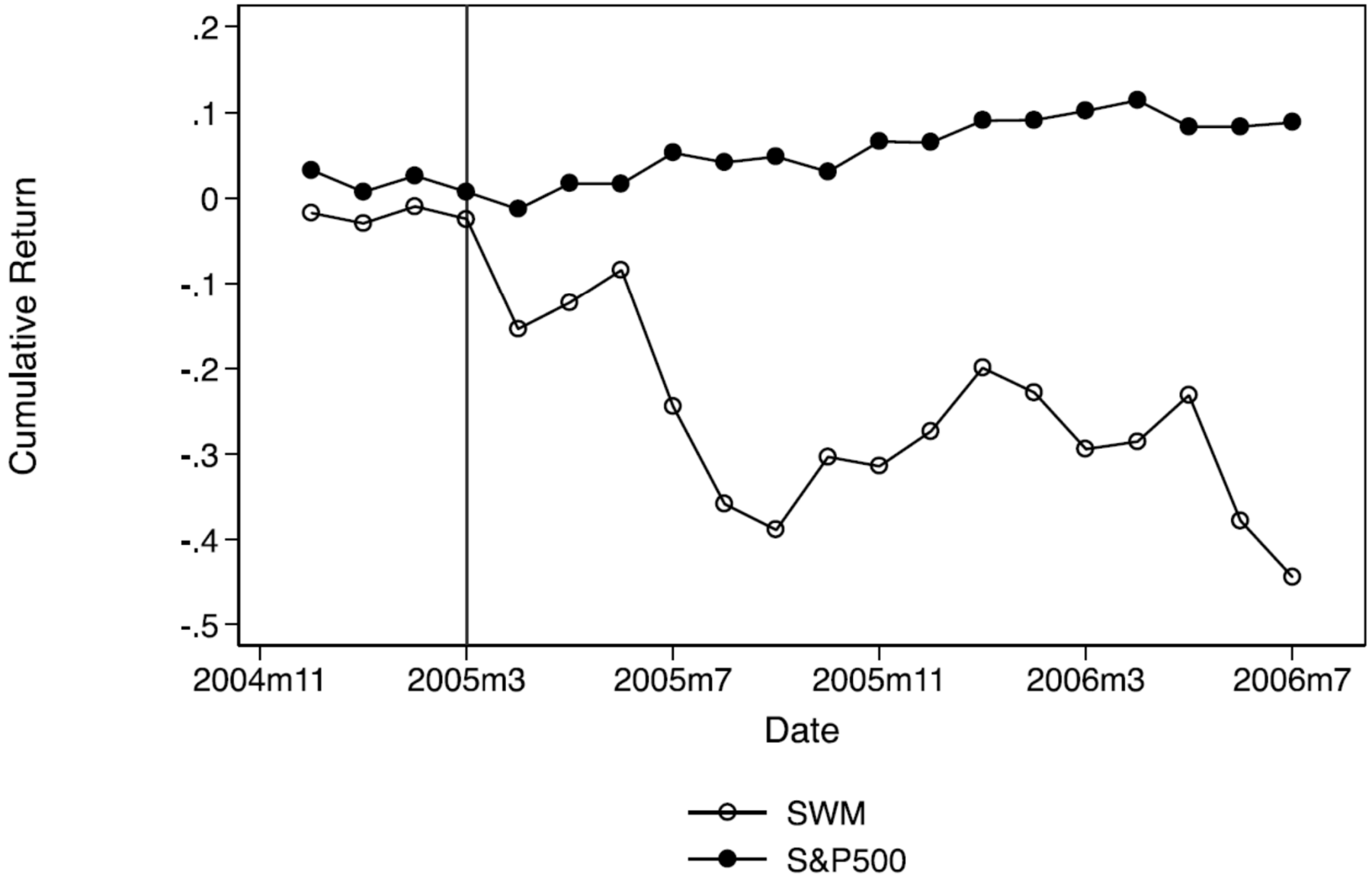
The new RTL production line added at the Company's Spay, France mill, which started up in the fourth quarter of 2003, is expected to be a major contributor to increased operating profit in 2004 compared with 2003. The new RTL production line is expected to achieve end of curve production rates by the end of the second quarter of 2004. The acquisition of a tobacco-related papers manufacturer in Indonesia that was completed in February 2004 is also expected to have a favorable impact on operating profit in 2004.

The Company did not have significant production or sale of banded or print banded cigarette papers during 2003. The Company continues to work with its customers in their development of papers for reduced ignition propensity cigarettes. In December 2003, the State of New York announced the adoption of final regulations for reduced ignition propensity cigarettes. The cigarette fire safety standard requires that all cigarettes sold in the State of New York as of June 28, 2004 have reduced ignition propensity properties. The regulations do contain a provision that allows wholesalers and retailers to transition their existing inventories. As a result of the new fire safety standards in the State of New York, the Company expects increased sales of reduced ignition propensity cigarette papers during 2004. These reduced ignition propensity papers sell for a higher price than the conventional cigarette papers they replace and are expected to have a positive impact on the Company's financial results. Since the State of New York only represents approximately ten percent of U.S. cigarette consumption and the regulations will only be in effect for one-half of 2004, the favorable impact on the Company's financial results is not expected to be significant in 2004.

Selling prices for the Company's tobacco-related products are expected to remain relatively stable during 2004. The recent weakening of the U.S. dollar versus the euro and certain other foreign currencies and higher wood pulp costs could enable the Company to implement selective selling price increases.

UCLAAnderson
SCHOOL of MANAGEMENT

c. Lazy Prices: Motivating Example



c. Text Similarity

- Authors argue this is a systemic pattern
- Use four different text similarity measures, comparing quarter-on-quarter reports
  1. cosine similarity
  2. Jaccard similarity
  3. minimum edit distance
  4. simple similarity

c. Cosine Similarity

- Let D_{S1} and D_{S2} be the set of terms in documents D_1 and D_2, respectively
- Define T as the union of words in D_{S1} and D_{S2} and let t_i be the i'th element of T
- Define the *term frequency vectors*

$$\begin{aligned} D_1^{TF} &= \left[nD_1\left(t_1\right), nD_1\left(t_2\right), ..., nD_1\left(t_N\right)\right], \\ D_2^{TF} &= \left[nD_2\left(t_1\right), nD_2\left(t_2\right), ..., nD_2\left(t_N\right)\right] \end{aligned}$$

  where $nD_j\left(t_i\right)$ is the number of occurences of term t_i in document D_j
- Cosine similarity is defined as

$$\text{Sim\_Cosine} = \frac{\left(D_1^{TF}\right)' D_2^{TF}}{\left\|D_1^{TF}\right\| \times \left\|D_2^{TF}\right\|}$$

c. Cosine Similarity (cont'd)

- To see how this works, consider the following (short) documents

$$\begin{aligned} D_A &: \text{We expect demand to increase} \\ D_B &: \text{We expect worldwide demand to increase} \\ D_C &: \text{We expect weakness in sales} \end{aligned}$$

- Clearly, D_A and D_B are more similar than, say, D_A and D_C
- Let's calculate the cosine measure between A and B:

$$T(D_A, D_B) = [\text{we, expect, worldwide, demand, to, increase}]$$

UCLAAnderson
SCHOOL of MANAGEMENT

c. Cosine Similarity (cont'd)

- From last slide

$$T\left(D_A, D_B\right) = [\text{we, expect, worldwide, demand, to, increase}]$$

- Then

$$\begin{aligned} D_1^{TF} &= [1,1,0,1,1,1] \\ D_2^{TF} &= [1,1,1,1,1,1] \end{aligned}$$

- So cosine similarity is:

$$\frac{\left(D_1^{TF}\right)' D_2^{TF}}{\left\|D_1^{TF}\right\| \times \left\|D_2^{TF}\right\|}$$

$$\begin{aligned} &= \frac{1\times 1+1\times 1+0\times 1+1\times 1+1\times 1+1\times 1}{\sqrt{1^2+1^2+0^2+1^2+1^2+1^2}\times\sqrt{1^2+1^2+1^2+1^2+1^2+1^2}} \\ &= 0.91 \end{aligned}$$

c. Cosine Similarity (cont'd)

- Next, let's do D_A and D_C

$$T(D_A, D_C) = [\text{we, expect, demand, to, increase, weakness, in, sales}]$$

- Then

$$\begin{aligned} D_1^{TF} &= [1,1,1,1,1,0,0,0] \\ D_2^{TF} &= [1,1,0,0,0,1,1,1] \end{aligned}$$

- And so:

$$\frac{\left(D_1^{TF}\right)' D_2^{TF}}{\|D_1^{TF}\| \times \|D_2^{TF}\|}$$

$$\begin{aligned} &= \frac{1\times1+1\times1+1\times0+1\times0+1\times0+1\times1+1\times1+1\times1}{\sqrt{1^2+1^2+1^2+1^2+1^2}\times\sqrt{1^2+1^2+1^2+1^2+1^2}} \\ &= 0.40 \end{aligned}$$

UCLAAnderson
SCHOOL of MANAGEMENT

c. Jaccard Similarity

- Definition:

$$\text{Sim\_Jaccard} = \left| D_1^{TF} \cap D_2^{TF} \right| / \left| D_1^{TF} \cup D_2^{TF} \right|$$

Interaction divided by size of union

- Then

$$\begin{aligned} \text{Sim\_Jaccard}(D_A, D_B) &= |\{\text{we,expect,demand,to,increase}\}| / \\ &\quad |(\text{we,expect,worldwide,demand,to,increase})| \\ &= 5/6 \end{aligned}$$

- And

$$\begin{aligned} \text{Sim\_Jaccard}(D_A, D_C) &= |\{\text{we,expect}\}| / \\ &\quad |(\text{we,expect,demand,to,increase,weakness,in,s} \\ &= 2/8 \end{aligned}$$

c. Min-Edit Similarity

- Minimum number of changes to make the to documents the same
- So, for D_A vs. D_B need only to add "worldwide" to document A (1 change)
- For D_A vs D_C need to delete "demand", "to", "increase" from A and add "weakness", "in", "sales" (6 changes)

UCLAAnderson
SCHOOL of MANAGEMENT

c. Simple Similarity

- Simple similarity uses "track changes" in word to identify "changes", "additions", and "deletions" while comparing the old to new document
- Count the number of words in these categories, sum, and divide by count of words in first document (that is being change to the second)
- Use all available 10-Q's and 10-K's (and some other more rare reports)

UCLAAnderson
SCHOOL of MANAGEMENT

c. Results: FMB regressions 1994 - 2014

Table V: Main Results – Fama MacBeth Regression

This Table reports the Fama-MacBeth cross-sectional regressions of individual firm-level stock returns on our 4 similarity measures and a host of known return predictors. Size is log of market value of equity, log(BM) is log book value of equity over market value of equity, Ret(-1,0) is previous month's return, and Ret(-12, -1) is the cumulative return from month -12 to month -1. SUE is the standardized unexpected earning and computed as actual earnings per share minus average analyst forecast earnings per share, divided by the standard deviation of the forecasts.

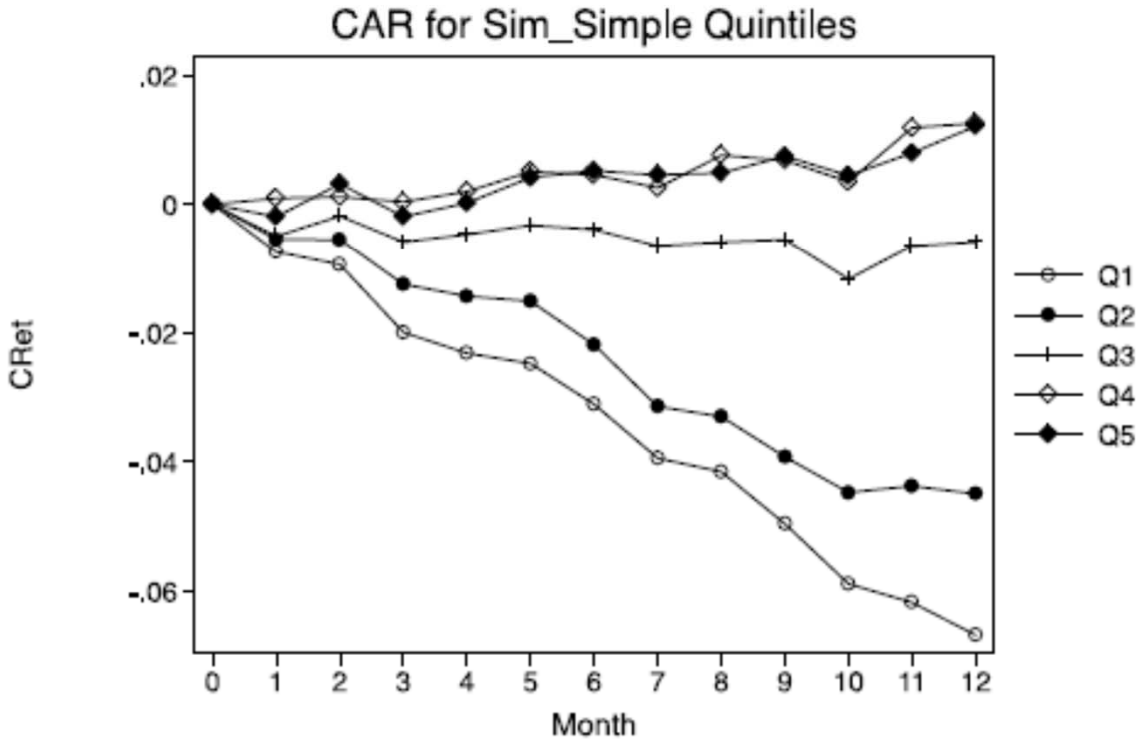
	(1) Ret	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Sim_Cosine	0.0045***	0.0031**	0.0037**									
	(2.6469)	(2.5103)	(2.1751)									
Sim_Jaccard				0.0082***	0.0066***	0.0059***						
				(3.2607)	(3.8197)	(3.4063)						
Sim_MinEdit							0.0054**	0.0041***	0.0029**			
							(2.5398)	(2.7795)	(1.9970)			
Sim_Simple										0.0404**	0.0302**	0.0292**
										(2.1031)	(2.2484)	(2.1099)
Size		0.0000	0.0000		0.0001	0.0001		0.0001	0.0001		0.0001	0.0000
		(0.1111)	(0.0507)		(0.2496)	(0.1133)		(0.2558)	(0.0980)		(0.2385)	(0.0485)
log(BM)		0.0017*	0.0016*		0.0017*	0.0016*		0.0017*	0.0016*		0.0017*	0.0016*
		(1.8936)	(1.7142)		(1.8797)	(1.7047)		(1.8955)	(1.7163)		(1.8740)	(1.6957)
Ret(-1,0)		-0.0260***	-0.0243***		-0.0263***	-0.0244***		-0.0263***	-0.0244***		-0.0263***	-0.0245***
		(-3.9281)	(-3.6827)		(-3.9704)	(-3.7026)		(-3.9731)	(-3.6930)		(-3.9852)	(-3.7105)
Ret(-12,-1)		0.0064**	0.0036		0.0064**	0.0036		0.0064**	0.0036		0.0064**	0.0037
		(2.3394)	(1.2457)		(2.3407)	(1.2502)		(2.3357)	(1.2438)		(2.3469)	(1.2934)
SUE			0.0007***			0.0007***			0.0007***			0.0007***
			(6.5591)			(6.5442)			(6.5584)			(6.4993)
Cons	0.0058	0.0058	0.0067	0.0064	0.0046	0.0069	0.0076**	0.0057	0.0084	-0.0238	-0.0176	-0.0142
	(1.4516)	(0.6721)	(0.5684)	(1.6348)	(0.5171)	(0.5814)	(1.9765)	(0.6369)	(0.7057)	(-1.3069)	(-1.0217)	(-0.7060)
R-Squared	0.0006	0.0427	0.0485	0.0017	0.0432	0.0489	0.0017	0.0432	0.0488	0.0019	0.0435	0.0492
N	713451	713451	496084	713451	713451	496084	713451	713451	496084	713680	713680	495931

UCLAAnderson
SCHOOL of MANAGEMENT

c. Results: Quintile portfolio returns

- No reversal
- Effect is that on average changes are bad news

This figure shows the average cumulative abnormal return for each quintile portfolio sorted based on firms' similarity score, for 1 month to 12 months after portfolio formation.



UCLAAnderson
SCHOOL of MANAGEMENT

c. Which sections are important?

Table VIII: Mechanism – In which sections do changes matter most?

This Table reports the calendar-time portfolio returns. Similarity measures for each item are computed using only the textual portion in that item. For each of the four similarity measures, we compute quintiles based on the prior year's distribution of similarity scores across all stocks. Stocks then enter the quintile portfolio in the month after the public release of one of their 10-K or 10-Q reports. Firms are held in the portfolio for 3 months. We report Excess Return (return minus risk free rate), Fama-French 3-factor alphas (market, size, and value), and 5-factor alphas (market, size, value, momentum, and liquidity) of the top minus bottom quintile portfolio (Q5 – Q1). Panel A reports equal-weight portfolio returns. Panel B reports value-weight portfolio returns.

Panel B: Value Weighted

	Sim_Cosine			Sim_Jaccard		
	Excess Return	3-Factor Alpha	5-Factor Alpha	Excess Return	3-Factor Alpha	5-Factor Alpha
Management's Discussion and Analysis	0.0027*	0.0028*	0.0022	0.0047***	0.0043***	0.0033**
	(1.8009)	(1.8471)	(1.4237)	(2.8834)	(2.6347)	(2.0151)
Legal Proceedings	0.0035*	0.0032	0.0032	0.0018	0.0010	0.0005
	(1.6643)	(1.5347)	(1.4722)	(0.8050)	(0.4609)	(0.2127)
Quantitative and Qualitative Disclosures About Market Risk	0.0039	0.0044	0.0045	0.0047***	0.0042***	0.0038**
	(1.3980)	(1.5716)	(1.6159)	(2.8918)	(2.6005)	(2.3723)
Risk Factors	0.0144*	0.0150**	0.0156**	0.0118*	0.0165***	0.0156**
	(1.9625)	(2.0069)	(2.0470)	(1.8999)	(2.7450)	(2.5669)
Other Information	0.0073**	0.0075**	0.0080**	0.0054	0.0049	0.0043
	(2.1343)	(2.2083)	(2.3014)	(1.5574)	(1.4249)	(1.2049)

	Sim_MinEdit			Sim_Simple		
	Excess Return	3-Factor Alpha	5-Factor Alpha	Excess Return	3-Factor Alpha	5-Factor Alpha
Management's Discussion and Analysis	0.0047***	0.0044***	0.0033*	0.0038**	0.0037**	0.0025
	(2.6718)	(2.6389)	(1.9706)	(2.0562)	(2.1179)	(1.4231)
Legal Proceedings	0.0014	0.0005	0.0007	0.0030	0.0024	0.0027
	(0.6083)	(0.2467)	(0.2985)	(1.2640)	(1.0351)	(1.1573)
Quantitative and Qualitative Disclosures About Market Risk	0.0000	0.0014	0.0012	0.0013	0.0011	0.0007
	(0.0149)	(0.6396)	(0.6135)	(0.1581)	(0.1319)	(0.0801)
Risk Factors	0.0095	0.0151**	0.0105*	0.0125	0.0133	0.0085
	(1.1777)	(2.2874)	(1.6658)	(1.5388)	(1.6108)	(1.0385)
Other Information	0.0022	0.0011	0.0009	0.0013	0.0002	0.0000
	(0.6272)	(0.3286)	(0.2515)	(0.3783)	(0.0678)	(0.0146)

UCLAAnderson
SCHOOL of MANAGEMENT

c. What type of language changes are most important?

This Table reports robustness checks of the types of textual changes that matter most. We split on median reference to a number of different attributes of the text change itself: Sentiment, Uncertainty, and the Litigiousness of the change.

		Sim Cosine						Sim Jaccard					
		Q1	Q2	Q3	Q4	Q5	Q5 - Q1	Q1	Q2	Q3	Q4	Q5	Q5 - Q1
	Low	-0.0009	-0.0049**	-0.0011	0.0001	0.0018	0.0026	-0.0045***	-0.0044***	-0.0024	0.0023	0.0009	0.0054**
Sentiment		(-0.7123)	(-2.4323)	(-0.8359)	(0.0655)	(1.5807)	(1.4798)	(-2.7913)	(-3.1639)	(-1.2370)	-1.6184	-0.6911	-2.4101
	High	0.0017	-0.0022	0.0004	0.0013	0.0021	0.0006	0.0008	0.0004	0.0013	0.0022	0.0015	0.0011
		(1.2713)	(-1.4511)	(0.2767)	(0.9940)	(1.5911)	(0.3044)	-0.6297	-0.266	-0.7833	-1.5338	-1.2704	-0.6093
	Low	-0.0003	-0.0024	0.0012	0.0014	0.0018	0.0021	-0.0023*	-0.0034**	0.002	0.0025*	0.002	0.0044**
Uncertainty		(-0.2047)	(-1.5217)	(0.8707)	(1.0239)	(1.3515)	(1.0751)	(-1.6548)	(-2.0413)	-1.2431	-1.8589	-1.4689	-2.4187
	High	-0.0022*	-0.0007	0.0006	0.0007	0.0005	0.0032*	-0.0054***	-0.001	0	0.0008	0.0013	0.0072***
		(-1.7899)	(-0.4183)	(0.4222)	(0.4518)	(0.4417)	(1.8134)	(-3.1124)	(-0.7230)	(-0.0218)	-0.5928	-1.1628	-3.5092
	Low	-0.0010	-0.0032**	0.0015	0.0018	0.0004	0.0014	-0.0029**	-0.0042***	0.0013	0.0011	0.0016	0.0047**
Litigious		(-0.7701)	(-2.0781)	(1.0152)	(1.2306)	(0.3863)	(0.8268)	(-1.9848)	(-2.6452)	-0.774	-0.8267	-1.0496	-2.1829
	High	-0.0023*	-0.0007	0.0010	0.0024*	0.0012	0.0040**	-0.0048***	-0.0011	0.0006	0.0024**	0.002	0.0071***
		(-1.8054)	(-0.4501)	(0.7448)	(1.8381)	(1.0190)	(2.2466)	(-2.7580)	(-0.7463)	-0.3233	-2.0542	-1.5655	-3.2909

		Sim MinEdit						Sim Simple					
		Q1	Q2	Q3	Q4	Q5	Q5 - Q1	Q1	Q2	Q3	Q4	Q5	Q5 - Q1
	Low	-0.0036**	-0.0022	0.0016	-0.0008	0.0013	0.0048**	-0.0047***	-0.0024	-0.0001	0.0027**	0.0010	0.0057***
Sentiment		(-2.3516)	(-1.5372)	(1.1200)	(-0.6059)	(0.9551)	(2.1460)	(-3.3643)	(-1.5296)	(-0.1041)	(2.0023)	(0.7035)	(2.6567)
	High	-0.0002	-0.0002	0.0006	0.0004	0.0026*	0.0032	0.0011	0.0006	0.0008	0.0009	0.0020	0.0012
		(-0.1464)	(-0.1844)	(0.4199)	(0.2755)	(1.6932)	(1.5618)	(0.8134)	(0.6002)	(0.5391)	(0.5091)	(1.1541)	(0.5032)
	Low	-0.0033**	0.0004	-0.0015	0.0014	-0.0003	0.0033*	-0.0017	-0.0013	-0.0001	0.0017	0.0022	0.0038*
Uncertainty		(-2.0092)	(0.2767)	(-1.1442)	(0.8347)	(-0.1981)	(1.6723)	(-1.1747)	(-1.0097)	(-0.0768)	(1.3819)	(1.4079)	(1.8473)
	High	-0.0014	-0.0021	0.0012	0.0017	0.0026*	0.0041**	-0.0041**	-0.0008	0.0030***	0.0012	0.0007	0.0051**
		(-1.0799)	(-1.5031)	(0.9572)	(1.2670)	(1.7718)	(2.0624)	(-2.2905)	(-0.6771)	(2.6108)	(0.6432)	(0.3959)	(2.1409)
	Low	-0.0005	-0.0022	-0.0005	-0.0008	0.0032**	0.0038*	-0.0023	-0.0030**	0.0019	-0.0007	0.0016	0.0039*
Litigious		(-0.4520)	(-1.3860)	(-0.3590)	(-0.5422)	(2.0016)	(1.9562)	(-1.6448)	(-2.2771)	(1.6493)	(-0.5575)	(1.0031)	(1.8726)
	High	-0.0032*	0.0001	-0.0004	0.0027**	0.0016	0.0051**	-0.0035**	-0.0001	0.0028**	0.0030**	0.0010	0.0049**
		(-1.9640)	(0.0807)	(-0.3698)	(1.9978)	(0.9775)	(2.2169)	(-2.0759)	(-0.1127)	(2.4679)	(2.1654)	(0.6788)	(2.0119)

UCLAAnderson
SCHOOL of MANAGEMENT

d. Big Data and EDGAR Web Crawling

- An important skill is the ability to obtain new (non-standard) data efficiently and execute your idea
- In an optional homework, we will code up a web-crawler that automates downloading annual reports from EDGAR. The idea is to enable you to do large-scale textual analysis.
- We will then construct the similarity measures (Cosine and Jaccard) on a subset of the annual report data