Imperial College
Business School

Imperial means
Intelligent Business

Big Data in Finance

Asset Management - II

Tarun Ramadorai

Today's Agenda

- Cross-sectional approaches.
  - A brief introduction.
- A review of methods:
  - Fama-MacBeth.
  - Event Studies.
- Bringing unstructured text into the equation: leveraging "Big Data".

Market Efficiency

- As we saw before, Fama (1970) defines a market as efficient if "prices fully reflect all available information". In practice this means:

$$R_{i,t+1} = \Theta_{it} + U_{i,t+1}, \tag{1}$$

  where Θ_{it} is the rationally expected return on asset i, and $U_{i,t+1}$ has zero expectation with respect to the information set at t.

- This will only have content if we can restrict Θ_{it} using an economic model.

- Thus market efficiency is not testable except in combination with a model of expected returns.
  - This is called the joint hypothesis problem.

Cross-Sectional vs. Time-Series Efficiency

- Time-series efficiency. (What we have just looked at).
  - Fix i, model returns over t.

Cross-Sectional vs. Time-Series Efficiency

- Time-series efficiency. (What we have just looked at).
  - Fix i, model returns over t.

- Cross-sectional efficiency.
  - Take average returns over t and consider various partitions of i.
  - Economic model for Θ_{it} in (1) is a cross-sectional asset pricing model like the CAPM.
  - Can only conduct joint tests of the CAPM and market efficiency.

Cross-Sectional vs. Time-Series Efficiency

- Time-series efficiency. (What we have just looked at).
  - Fix i, model returns over t.

- Cross-sectional efficiency.
  - Take average returns over t and consider various partitions of i.
  - Economic model for Θ_{it} in (1) is a cross-sectional asset pricing model like the CAPM.
  - Can only conduct joint tests of the CAPM and market efficiency.

- "Event" efficiency.
  - Does the market respond "correctly" to changes in relevant information?
  - Combination of cross-section and time-series.
  - Important target for trading strategies.

Today's Agenda

- Anomalies.
  - Relevant for Practitioners: How do we detect anomalies?
- Methods.
  - Fama MacBeth regressions.
  - Event studies and FFJR.

Detecting Anomalies

- What is the right method to detect a mispricing (or anomaly) from our preferred asset pricing model?

- Standard Methodology:
  - Sort all firms by the trait under study at the end of every period (month, quarter, year etc.).
  - Form the top quantile (usually decile) and the bottom quantile into portfolios.
  - Track their returns post-formation for some holding period, and then rebalance the portfolios.
  - Report the average difference in excess returns between top and bottom decile portfolios – or use the Patton-Timmerman (2010) MR test.

- There are different ways to do a study like this, and they can produce different results.

Issues

- Should returns be equal- or value-weighted?
- What is the correct way to handle delistings and mergers?
- What breakpoints should be used to form portfolios?
- Why use a sorting approach rather than using a continuous variable?
- Robustness often requires trying different approaches and showing that the answers are invariant.
  - Write flexible code!
- An important issue: Are performance differentials *statistically significant*?
  - Gets to the heart of why finance is skeptical of *data-mining*.

Defining 'Unexpected Returns'

- Benchmarking is critical.
  - Average *realized* return differences should be compared against *expected* return differences.

- In terms of intellectual history: CAPM was widely regarded as the correct benchmark.
  - In the 1980s, there were virtually no competing theories to the CAPM.

- Portfolio unexpected returns were computed in two steps:
  - First, estimate each portfolio's market beta.
  - Second, subtract beta times realized return on the market.
  - What is the best way to estimate betas? Portfolios or assets? What about the effect of illiquidity?
  - How often should we re-estimate betas?
  - How should the "market portfolio" be defined?

Implications

- Why is the correct choice of the benchmark model critical?
- Implicitly any test of an anomaly is also about *testing the benchmark model*.
  - This is called the joint hypothesis problem.
- Tests in the 1980's rejected the prevalent theory at the time, i.e., the CAPM.
  - Anomalies detected relative to CAPM predictions about expected returns.
  - Situation can be summarized for example as: "The CAPM fails to price small stocks."
- Also important (esp. for practitioners): Do we expect the return premium to small, high book-value stocks to continue?

Econometric Issue: Data Mining

- In empirical asset pricing, strong concerns about *data mining* naturally arise when thinking about anomalies.
- How many anomalies would we expect to discover in any large enough set of returns just by chance?
- Data mining refers to the process of selectively reporting tests based on advance knowledge of which ones work.
- Researchers almost always run a lot of tests that *don't* work before reporting the ones that *do*.
  - But this *by itself* invalidates classical statistical inference.

Jegadeesh and Titman (1993)

		Panel A				Panel B			
J		$K=$ 3	6	9	12	$K=$ 3	6	9	12
3	Sell	0.0108	0.0091	0.0092	0.0087	0.0083	0.0079	0.0084	0.0083
		(2.16)	(1.87)	(1.92)	(1.87)	(1.67)	(1.64)	(1.77)	(1.79)
3	Buy	0.0140	0.0149	0.0152	.0156	0.0156	0.0158	0.0158	0.0160
		(3.57)	(3.78)	(3.83)	(3.89)	(3.95)	(3.98)	(3.96)	(3.98)
3	Buy-sell	0.0032	0.0058	0.0061	0.0069	0.0073	0.0078	0.0074	0.0077
		(1.10)	(2.29)	(2.69)	(3.53)	(2.61)	(3.16)	(3.36)	(4.00)
6	Sell	0.0087	0.0079	0.0072	0.0080	0.0066	0.0068	0.0067	0.0076
		(1.67)	(1.56)	(1.48)	(1.66)	(1.28)	(1.35)	(1.38)	(1.58)
6	Buy	0.0171	0.0174	0.0174	0.0166	0.0179	0.0178	0.0175	0.0166
		(4.28)	(4.33)	(4.31)	(4.13)	(4.47)	(4.41)	(4.32)	(4.13)
6	Buy-sell	0.0084	0.0095	0.0102	0.0086	0.0114	0.0110	0.0108	0.0090
		(2.44)	(3.07)	(3.76)	(3.36)	(3.37)	(3.61)	(4.01)	(3.54)
9	Sell	0.0077	0.0065	0.0071	0.0082	0.0058	0.0058	0.0066	0.0078
		(1.47)	(1.29)	(1.43)	(1.66)	(1.13)	(1.15)	(1.34)	(1.59)
9	Buy	0.0186	0.0186	0.0176	0.0164	0.0193	0.0188	0.0176	0.0164
		(4.56)	(4.53)	(4.30)	(4.03)	(4.72)	(4.56)	(4.30)	(4.04)
9	Buy-sell	0.0109	0.0121	0.0105	0.0082	0.0135	0.0130	0.0109	0.0085
		(3.03)	(3.78)	(3.47)	(2.89)	(3.85)	(4.09)	(3.67)	(3.04)
12	Sell	0.0060	0.0065	0.0075	0.0087	0.0048	0.0058	0.0070	0.0085
		(1.17)	(1.29)	(1.48)	(1.74)	(0.93)	(1.15)	(1.40)	(1.71)
12	Buy	0.0192	0.0179	0.0168	0.0155	0.0196	0.0179	0.0167	0.0154
		(4.63)	(4.36)	(4.10)	(3.81)	(4.73)	(4.36)	(4.09)	(3.79)
12	Buy-sell	0.0131	0.0114	0.0093	0.0068	0.0149	0.0121	0.0096	0.0069
		(3.74)	(3.40)	(2.95)	(2.25)	(4.28)	(3.65)	(3.09)	(2.31)

Data Mining
An Illustration from J&T (2001)

- In their 1993 paper, the $J = 6, K = 6$ strategy earned abnormal returns of 95 bp per month from 1965 to 1989 with a t-statistic of 3.07.

- Suppose researchers collectively tested N independent trading strategies over the 1965 to 1989 period.

- Assume that the momentum strategy yielded the highest test statistic among those N strategies.

Data Mining

- Cumulative distribution of the *maximum* of N draws is approximately ϕ^N where ϕ is the standard normal cdf.
  - Assuming sample is large enough to use normcdf instead of tcdf.

- If the J&T test is viewed as one draw, done in isolation, then $N = 1$ and the probability of observing a t-statistic this large (the p-value) is 0.11%.
  - That is, $1 - \phi(3.07) = 1 - 0.9989 = 0.0011$, or in Matlab, $1 - normcdf(3.07)$.

- But if $N = 100$, then the probability that the largest test statistic is 3.07 is only about 10%.

Data Mining

- Why? If you did two tests, then the probability of seeing a t-statistic as high as 3.07 drops to $1 - 0.9989\hat{}2 = 0.0021$
  - The probability of seeing one such t-statistic equals one minus the probability of not seeing even one such t-statistic (if the tests are independent).

- For $N = 100$ (100 tests), take $1 - normcdf(3.07)\hat{}100 = 0.1016$ as the appropriate p-value.
  - The appropriate critical value for 5% significance for a two-sided test is 3.477 in this case.

- If $N = 650$, then the p-value based on the test statistic drops below 50% [!]
  - The appropriate critical value for 5% significance for a two-sided test is 3.951 in this case.

Data Mining

- This is a well-known problem. Nevertheless, it is a very difficult one to tackle.
- (At least) two possible solutions:

1. Model statistical selection processes, to correct inferential tests.
   - 1.1 This is hard. (see White (2000), Sullivan, Timmermann and White (2001), and Patton and Ramadorai (2010) for an application).
2. Get new data samples that haven't been mined. (Think cross-validation!).
   - 2.1 J&T (2001) look at 1989 to 1999 to show that momentum is not simply 'mined'.
   - 2.2 There is evidence of momentum in this later period, and it is just as strong. This is very persuasive.

Out of Sample Evidence

Jegadeesh and Titman (2001)

Momentum Portfolio Returns

	1965-89	1990-1997	1965-1997
P1 (Past Winners)	1.70	1.55	1.67
P2	1.50	1.25	1.44
P3	1.40	1.18	1.35
P4	1.32	1.17	1.29
P5	1.31	1.13	1.27
P6	1.25	1.11	1.22
P7	1.22	1.19	1.21
P8	1.16	1.11	1.15
P9	1.07	1.13	1.08
P10 (Past Losers)	0.59	0.54	0.58
P1-P10	1.11	1.01	1.09
t-statistic	4.61	4.10	5.65

Note: This table reports the monthly returns for momentum portfolios formed with stocks traded on the NYSE/AMEX. All stocks priced less than $5 at the beginning of the holding period are excluded from the sample. The momentum deciles are formed based on 6-month lagged returns and held for six months. P1 is the equal-weighted portfolio of ten percent of the stocks with the highest six-month lagged returns, P2 is the equal-weighted portfolio of the ten percent of the stocks with the next highest returns and so on.

Plan

- We've just seen another important reason to care about cross-validation, especially in empirical asset pricing.
- We're now going to move towards a rapidly growing area of "big data" in finance, which is the use of unstructured text as an input used to generate alpha in trading strategies.
- To understand progress made in that literature, we first need to quickly review some simple "alpha evaluation techniques" in cross-sectional asset pricing, and in the analysis of event time.
  - Fama MacBeth.
  - Event studies.
- We will return to recent progress made in "text-mining" immediately thereafter.

Fama-MacBeth (1973)

- **Step 1:** Estimate β_i by running separate (time-series) regressions of each asset/portfolio's return on the market portfolio:

$$R^e_{it} = k_i + \beta_i R^e_{mt} + u_{it}, \text{ for each } i = 1, ..., N \tag{2}$$

  - These regressions can be run on the full sample of data for each asset.
  - They can also be run on rolling windows (re-doing the regression each time using a moving window of data).
  - Or they can be run on separate subsamples to the rest of the analysis (below).
  - In the latter two cases, we would get a time-series of $\hat{\beta}_i$'s.

- **Step 2:** Run cross-sectional regressions of excess returns on the estimated $\hat{\beta}_i$.
  - First stack the $\hat{\beta}_i$ into a vector, $\hat{\beta} = \left[\hat{\beta}_1 \ \hat{\beta}_2 \ ... \ \hat{\beta}_N \right]$.

- Then run one cross-sectional regression in each *time period* t, i.e.:

$$R_t^e = \alpha_t + \hat{\beta}' \lambda_t + \varepsilon_t \tag{3}$$

  Note that $R_t^e = \left[R_{1t}^e \ R_{2t}^e \ ... \ R_{Nt}^e \right]'$.

- Then, Fama and MacBeth suggest that we estimate λ and α as the average of the cross-sectional regression estimates:

$$\hat{\lambda} = \frac{1}{T} \sum_{t=1}^{T} \hat{\lambda}_t$$

$$\hat{\alpha} = \frac{1}{T} \sum_{t=1}^{T} \hat{\alpha}_t$$

- We can also compute the variance of the two estimates using the time series:

$$\hat{\sigma}^2(\hat{\lambda}) = \frac{1}{T(T-1)} \sum_{t=1}^{T} \left(\hat{\lambda}_t - \hat{\lambda}\right)^2$$

$$\hat{\sigma}^2(\hat{\alpha}) = \frac{1}{T(T-1)} \sum_{t=1}^{T} \left(\hat{\alpha}_t - \hat{\alpha}\right)^2$$

- Note that we have $T(T-1)$ in the denominator, because it is the *variance of the mean* rather than the variance of the estimates themselves that we are interested in.

  - These formulas assume uncorrelatedness of the $\hat{\lambda}_t$'s and $\hat{\alpha}_t$'s.

  - $\hat{\lambda}$ is the 'market risk-premium'. If the CAPM is right, it should be 'priced', i.e. $\lambda > 0$.

  - The $\hat{\alpha}$ in this case is the 'pricing error'. If the CAPM is right, it should be zero.

- **Step 3:** Test whether $\hat{\lambda}$ is statistically positive and whether $\hat{\alpha}$ is statistically zero.

  - But this is now easy: we have the means and standard deviations of each from Step 2 above.

  - So we can do a standard t-test, with $T-1$ degrees of freedom:

$$t(\hat{\lambda}) = \frac{\hat{\lambda}}{\sqrt{\hat{\sigma}^2(\hat{\lambda})}}$$

$$t(\hat{\alpha}) = \frac{\hat{\alpha}}{\sqrt{\hat{\sigma}^2(\hat{\alpha})}}$$

- **Step 4:** Realize that there is a 'generated regressors' problem.
  - Essentially, you used the data to estimate the betas in Step 1, and then used those same betas as regressors in Step 2.
  - This means you have to correct the estimates and the test, since the statistics are not asymptotically valid any more.
  - Shanken (1992) presents a correction: multiply $\hat{\sigma}^2(\hat{\lambda})$ and $\hat{\sigma}^2(\hat{\alpha})$ by the factor:

$$shanken = \left(1 + \frac{(\hat{\mu}_m - \hat{\alpha})^2}{\hat{\sigma}_m^2}\right)$$

Fama, Fisher, Jensen, and Roll (1969)

- FFJR is first "event study"
  - Became a standard tool for testing semi-strong EMH
- Event study has 3-step methodology
- Step 1. Pick a model of expected returns and calculate "abnormal returns" (AR) for each security, for each day around the event:

$$AR_{it} = R_{it} - \mathrm{E}[R_{it}],$$

where t is the date relative to the event. (FFJR use a CAPM-like methodology to estimate $\mathrm{E}[R_{it}]$)

- - Generally, event window excluded in estimation of market model to avoid influencing normal parameter estimates.
  - FFJR use later as well as prior data.

- Step 2. Take the average abnormal return across securities ($N = 940$ for FFJR), each day, to get an average abnormal return in "event time".

- Note that this is the cross-sectional mean in each time period:

$$AR_t = \frac{1}{N} \sum_{i=1}^{N} AR_{it}$$

- Step 3. Construct the cumulative abnormal return (CAR) from time k through l $(k < l)$.

- If the market reacted properly to the event, and the $\mathrm{E}[R_{it}]$ model is right, then CAR should be zero for any interval starting at $k > 0$.

$$CAR_{k,l} = \sum_{t=k}^{l} AR_l$$

FFJR(1969)
Figure 2a, 2b

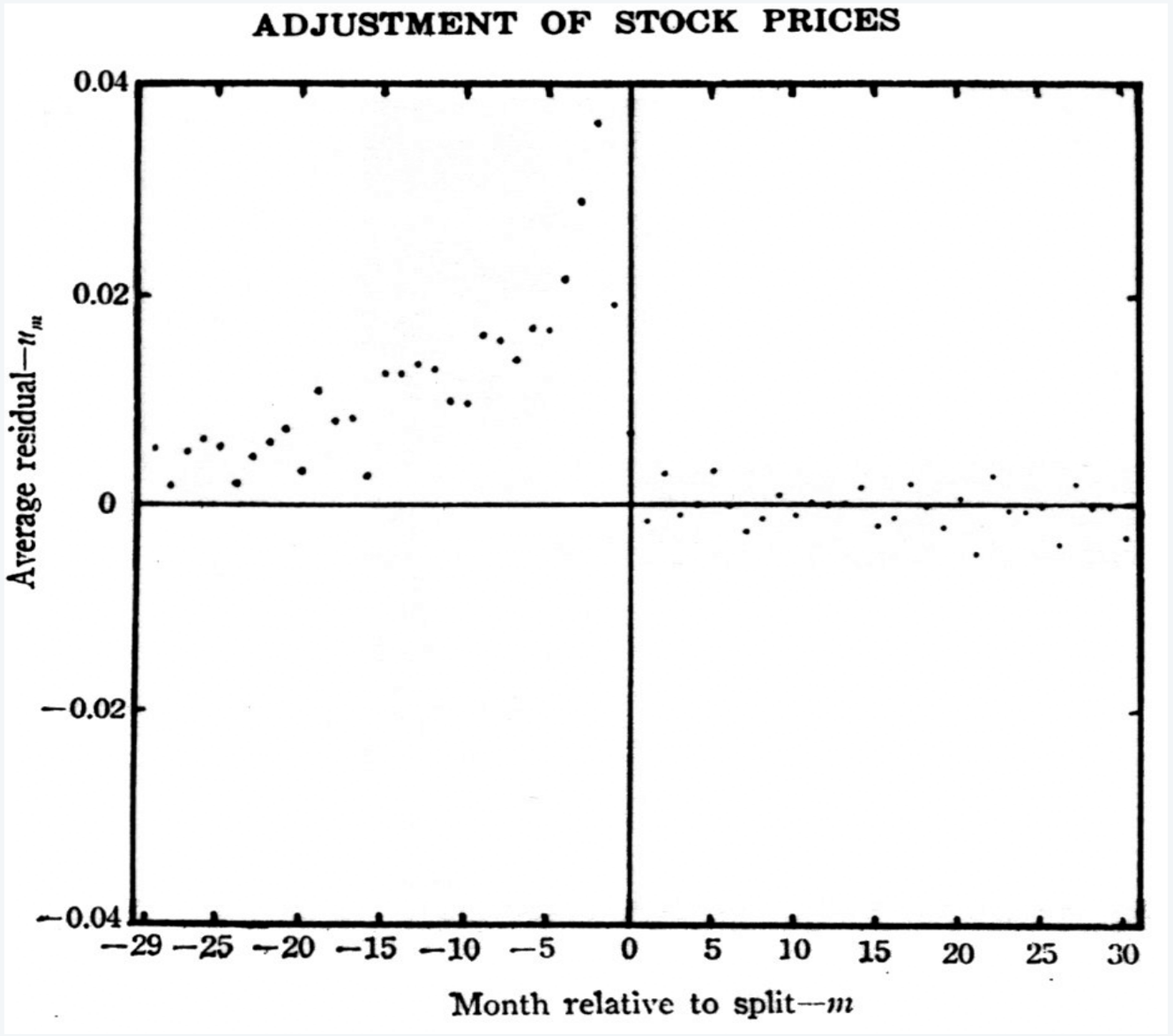
- On average, investors don't seem to under- or over-react to stock splits.

- A restatement: cannot use a stock split, once it becomes public information, to generate abnormal returns.

- Some relationship to dividend signaling, behaviour is consistent with rationality.

- This is consistent with semi-strong EMH.

FFJR(1969)
Figure 2a, 2b



FFJR(1969)
Figure 2a, 2b

FIGURE 2a

AVERAGE RESIDUALS—ALL SPLITS

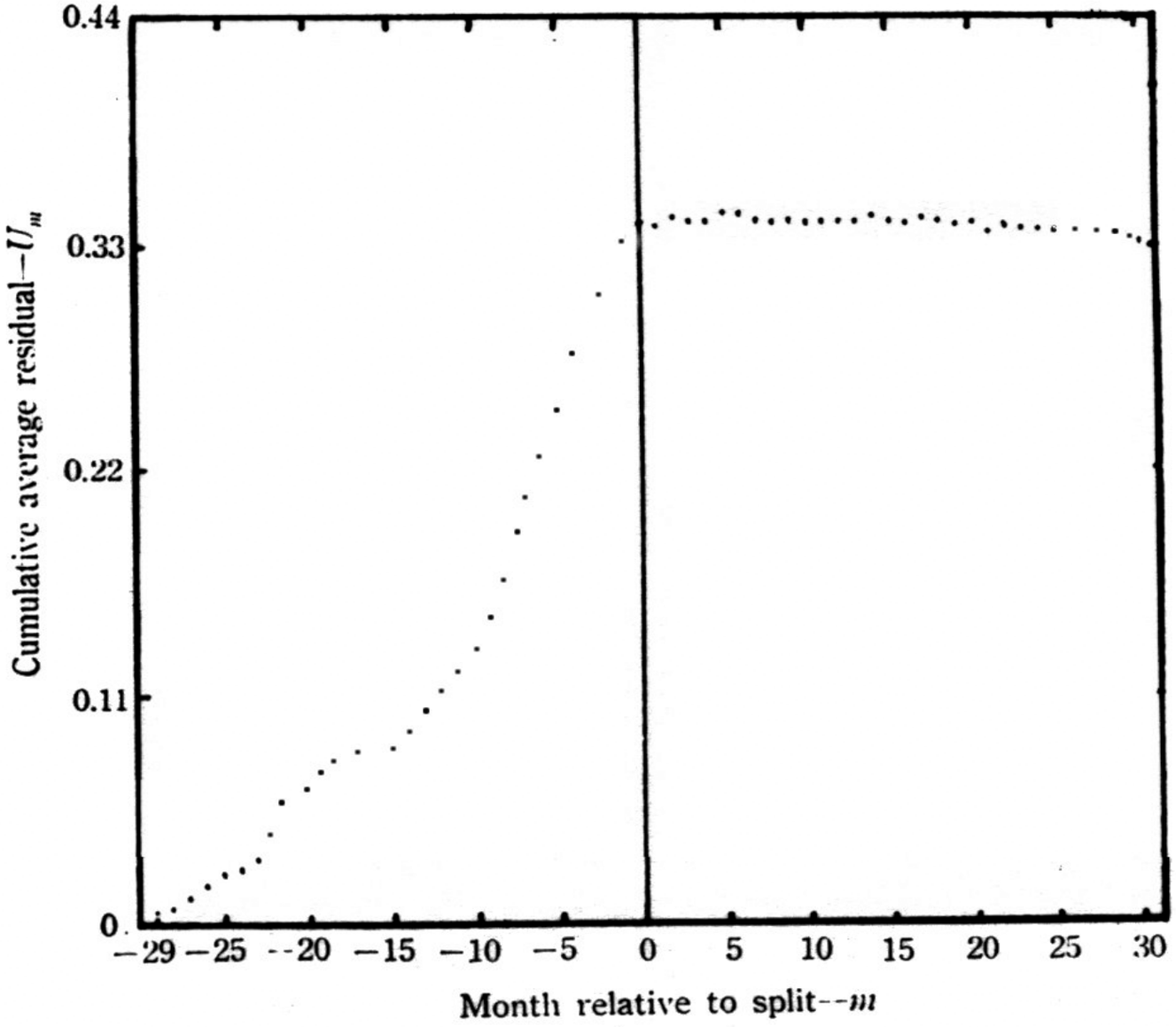


FIGURE 2b

CUMULATIVE AVERAGE RESIDUALS—ALL SPLITS

Unstructured Text

- FFJR was written in 1969, but the underlying questions are not incredibly different to this day.

- If you are a hedge fund manager, you are constantly attempting to discover new sources of alpha.

- One important way in which to do this is to find sources of information that you:
  - Acquire faster than others, who will process it in the same way as you once they get it.
  - Process more efficiently than others, who may process it incorrectly, or even better, process sloppily even though they get it at the same time as you.
  - Process faster than others, who may not immediately see the implications of the information you have acquired.

- Of course, a big risk in implementing such strategies is *noise trader risk*.

Textual Analysis

- As Li (2011) puts it, *The fundamental problem of understanding textual disclosure is data reduction. The goal of data reduction is to aggregate the information contained in a large amount of text into manageable numerical variables for further analysis.*

- Obviously, manual approaches are the most accurate way to process language. However, this is a seriously costly approach.

- The question, therefore, is how best to automate natural language processing. There are two broad classes of approaches.
  - Rule-based or "dictionary" approaches.
  - Statistical approaches.

Textual Analysis

- Dictionary approaches:

- These generally use a mapping schema, which classifies each word or phrase into a particular class using a set of rules or a dictionary.

- This is applied to entire documents alongside some rules to translate classified word frequencies into numerical output.

Textual Analysis

- Statistical approaches:

- These are a more traditional machine-learning approach. The general idea is to use a training dataset (usually manually classified) to train a classification algorithm.

- The classification algorithm generally relates a set of key words (or word frequencies) to a set of smaller classes, in an attempt to capture the essential meaning of the document.

Some Recent Examples

- We will consider two recent and interesting examples of the use of both types of strategies applied to unstructured data to help with predicting stock returns and accounting earnings.

- Note: there are many possible examples, which use:
  - Financial news (Tetlock et al., 2008)
  - Management discussion and analysis in annual reports (Li, 2010, Cohen, Malloy, Nguyen, 2016)
  - Social media. (Chen, De, Hu, Hwang, 2015)
  - Conference call tone. (Mayew and Venkatachalam, 2012)
  - Real-time sales forecasts. (Froot, Kang, Ozik, Sadka, 2016)
  - and others...

Financial News
Tetlock et al., 2008

- These authors attempt to quantify the language used in financial news stories to predict stock price movements and accounting earnings.

- They use data from the Wall Street Journal and Dow Jones News service between 1980 and 2004 to show this strategy works, in an early example which serves as a "proof of concept".

- Basic point they make is that if analyst forecasts and accounting fundamentals aren’t capturing all of the information, there may be more to be learned from textual analysis.
  - Of course, there is still a question here about whether this is cash-flows or discount factors.

Financial News
Tetlock et al., 2008

- The authors use the dictionary-based approach to classify each news story.
  - Use the Harvard IV psychosocial dictionary.

- The approach uses what is called the "Bag of Words" scheme.
  - In this scheme, all documents are represented as a document-term matrix, i.e., if each row is a document, each column of this matrix represents the frequency of representation of a particular term (word or phrase) in the document.

- To simplify the analysis, they just consider a binary classification scheme - the frequency of positive and negative words in each document.
  - They just use fraction of negative words to total in most analysis.

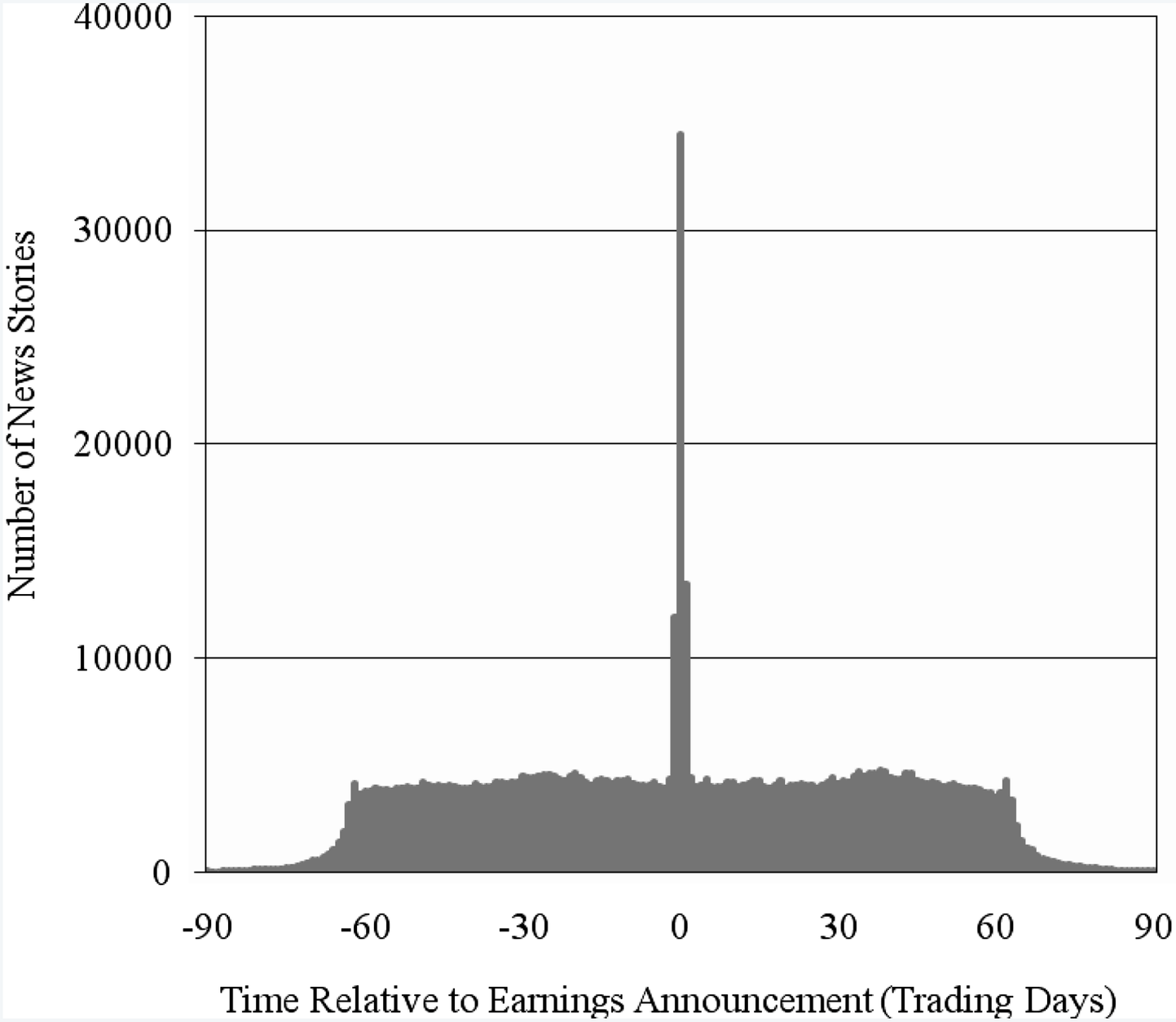
Data

- In total, they retrieve ~350,000 qualifying news stories from DJNS and WSJ—that contain over 100 MM words.

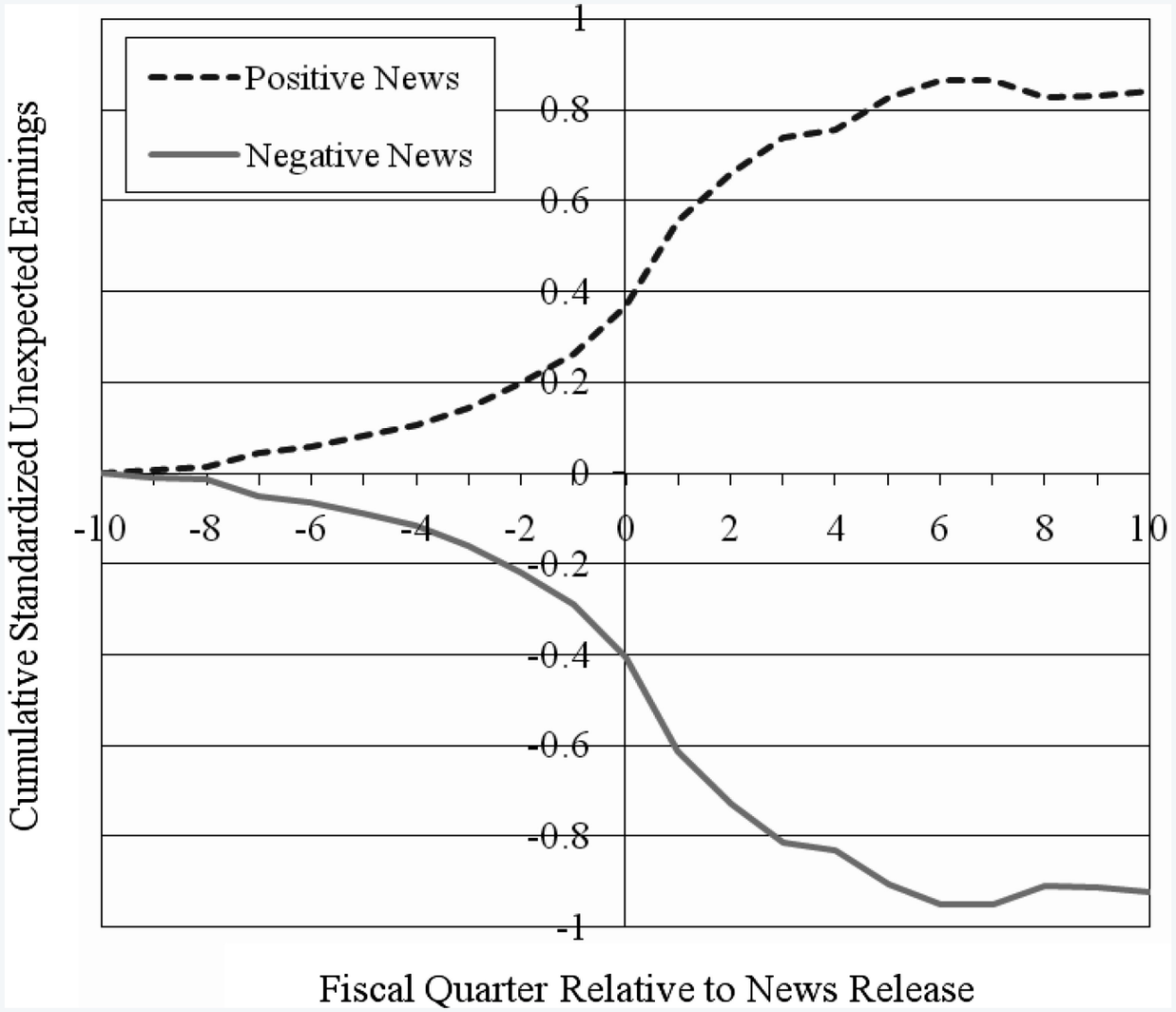
- They find at least one story for 1,063 of 1,110 (95.8%) of the firms in the S&P 500 from 1980 to 2004.

- They require that each story mentions the firm's name at least once within the first 25 words; that each story contains at least 50 words in total; and at least 5 words are either "Positive" or "Negative," where at least 3 of the 5 must be unique.
  - Such filters help eliminate "non-stories," or the influence of outliers.

News Timing

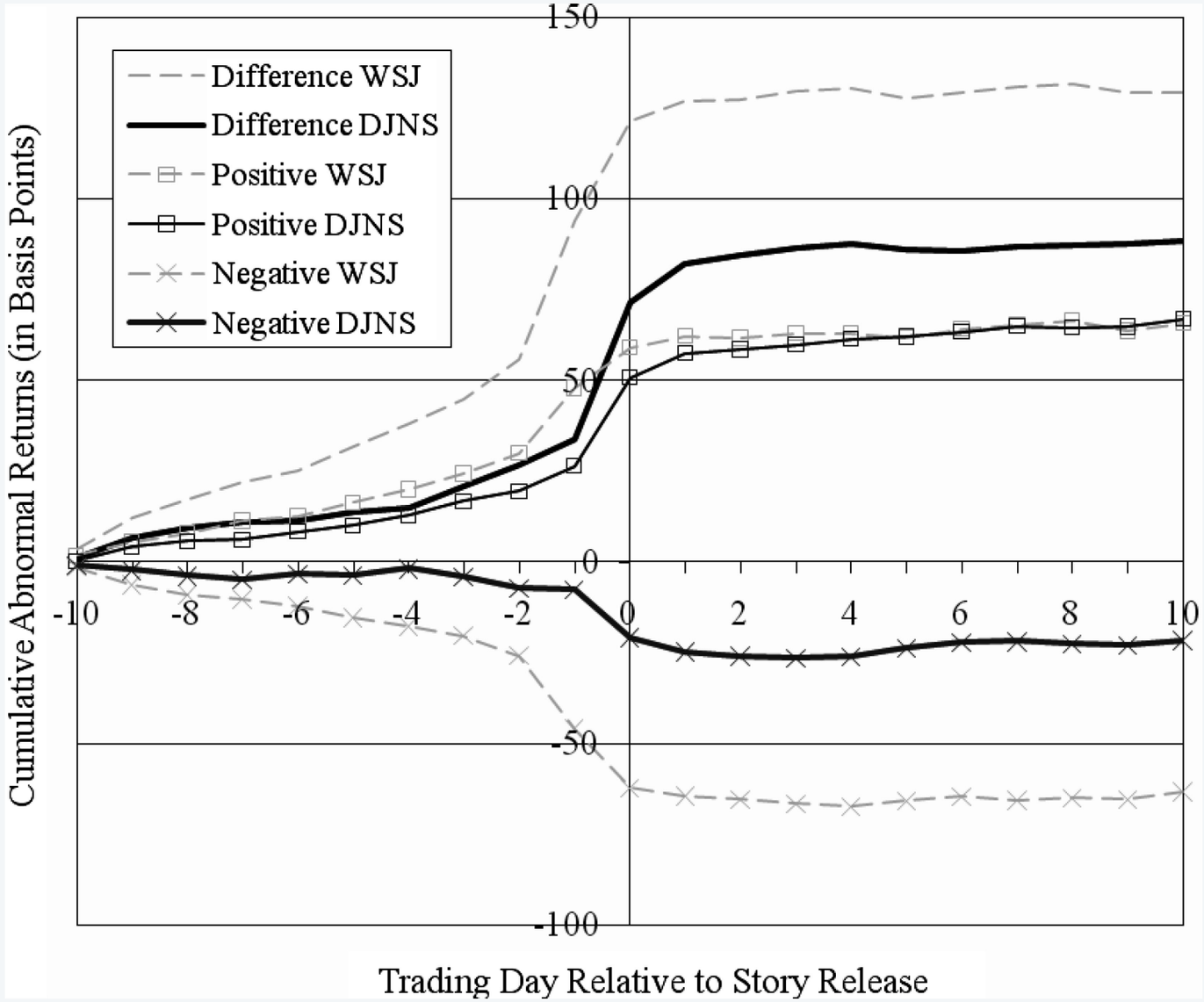

40000
30000
20000
10000
0
Number of News Stories
-90
-60
-30
0
30
60
90
Time Relative to Earnings Announcement (Trading Days)

Earnings Prediction



Note also: Table 1, columns 4 and 6 show that negative words work roughly half as well as CARs, and provide predictive power simultaneously.

Stock Return Prediction



Alpha

	1980–1994	1995–2004	1980–2004	1980–1994	1995–2004	1980–2004
Alpha	0.0919	0.1175	0.1031	0.0952	0.1131	0.1013
	(2.83)	(3.93)	(4.55)	(2.81)	(3.78)	(4.38)
Market	−0.0994	−0.1087	−0.0983	−0.0831	−0.1001	−0.0999
	(−0.93)	(−1.99)	(−1.86)	(−0.75)	(−1.87)	(−1.87)
SMB	−0.0767	0.0475	−0.0081	−0.0647	0.0341	−0.0128
	(−0.35)	(0.70)	(−0.08)	(−0.29)	(0.49)	(−0.12)
HML	−0.1869	−0.2590	−0.2372	−0.1819	−0.2500	−0.2365
	(−1.24)	(−2.81)	(−2.94)	(−1.20)	(−2.75)	(−2.93)
UMD				−0.0911	0.0930	0.0444
				(−0.74)	(2.01)	(0.90)
Trading Days	3398	2497	5895	3398	2497	5895
Adjusted R^2	0.0003	0.0081	0.0026	0.0004	0.0106	0.0027

Trading Costs
Is the strategy truly profitable?

Table IV
Sensitivity of News-Based Trading Returns to Trading Cost Assumptions

This table shows estimates of the impact of transaction costs on the news-based trading strategy's profitability (see the text or Table III for strategy details). We recalculate the trading strategy returns for 11 alternative assumptions about a trader's round-trip transaction costs: 0, 1, 2, 3 . . . or 10 basis points (bps) per round-trip trade. The abnormal and raw annualized cumulative news-based strategy returns for each assumption appear below. The risk-adjustment is based on the full-sample Fama-French three-factor loadings of the news-based portfolio shown in Table III.

Trading Costs (bps)	Abnormal Annualized Returns (%)	Raw Annualized Returns (%)
0	23.17	21.07
1	20.30	18.25
2	17.50	15.49
3	14.76	12.80
4	12.09	10.17
5	9.47	7.60
6	6.92	5.09
7	4.43	2.64
8	1.99	0.25
9	−0.39	−2.09
10	−2.71	−4.37

Statistical Approaches

- What we've just seen is a dictionary approach. But statistical approaches have also been used effectively.

- Let's quickly discuss a classification algorithm - naïve Bayes, which has found use in this literature.

- Recall the Bayes optimal classifier - where we sought two key inputs.
  - The first was π_k, or $P(Y = k)$, the prior class probability.
  - The second was $g_k(x)$ or $P(X = x|Y = k)$.

- As in LDA and QDA, naïve Bayes estimates π_k using the fraction of observations in the training data set that are in the kth class.

Naïve Bayes

- As in LDA and QDA, naïve Bayes estimates π_k using the fraction of observations in the training data set that are in the kth class.

- However, the assumption that is important here is on $g_k(x)$. In this classifier, we assume that:

$$P(X = x|Y = k) = P(x_1|Y)P(x_2|Y)..P(x_d|Y)$$

- That is, we assume that the distribution of the features is conditionally independent, i.e., independent within classes.

- This is a big assumption. Nevertheless if this does hold (unlikely), this will be a Bayes optimal classifier.

Naïve Bayes

- For discretely-valued data, we can estimate the probability $P(X = x|Y = k)$ simply using fractions of each variable/feature that takes each discrete value in the training set.

- For continuous data, can assume that data are normally distributed, for example.
  - Then just estimate the mean and variance for each feature in each class.
  - And use the normal pdf to tell you about the probability of any new $X = x$ given the prior class label $Y = k$.

- How does this help with processing unstructured data?

MD&A
Li, 2010

- Now let's consider an interesting case: how can we use management's own statements to improve earnings and return forecasts?

- In 1980, the Securities and Exchange Commission (SEC) mandated that public companies include in their annual reports a section for Management's Discussion and Analysis of Financial Condition and Results of Operations (MD&A).

- The MD&A is intended to provide management's view on the firm's liquidity, capital resources, and operations.

- There are several forward-looking statements (FLS) in the MD&A section of 10-K and 10-Q filings. These are what Li (2010) analyzes.

MD&A
Li, 2010

- To assess the content and tone of FLS in MD&As, Li uses a naïve Bayes learning algorithm.

- First, he manually classifies 30,000 sentences of randomly selected FLS extracted from the MD&A section of 10-Q filings.
  - Classification occurs along two dimensions: tone (i.e., positive versus negative) and content (e.g., profitability operations, liquidity, etc.).
  - He uses these sentences as a training dataset for the learning algorithm.

- The goal is for the algorithm to be used to classify the tone and content of other FLS in 10-Q and 10-K filings.

Training Dataset

TABLE 1

Percentage Distributions of MD&A FLS Tone and Content

Positive tone	19.59	Cetegory 1: Revenues	15.06
Neutral tone	39.97	Category 2: Cost	10.45
Negative tone	17.82	Category 3: Profits	8.72
Uncertain tone	22.55	Category 4: Operations	28.58
		Sum of 1–4	*62.81*
		Category 5: Liquidity	11.57
		Category 6: Investing	10.79
		Category 7: Financing	16.45
		Sum of 5–7	*38.81*
		Category 8: Litigation	2.14
		Category 9: Employees	1.41
		Category 10: Regulation	4.05
		Category 11: Accounting	2.78
		Category 12: Other	3.32
		Sum of 8–12	*13.70*

This table shows the percentage distributions of the 30,000 sentences (i.e., the training data) that are manually coded into different tone and content categories. The 30,000 forward-looking sentences are extracted randomly from the MD&As. Fifteen research assistants manually categorize them along two dimensions: tone and content. The details of the procedures are presented in appendices C and D.

Applying Naïve Bayes

- The algorithm estimates:

$$
\begin{aligned}
cat^* &= \arg\max_{cats} P(cat|words) \\
&= \arg\max_{cats} P(words|cat)P(cat) \\
&= \arg\max_{cats} P(w_1|cat)P(w_2|cat)..P(w_n|cat)P(cat)
\end{aligned}
$$

- In other words, we pick the category which maximizes the product of word frequencies times the prior probability.

Cross-Validation

TABLE 2

N-fold Cross-Validation Tests

	Tone		Content	
N	4 Categories	3 Categories	12 Categories	3 Categories
Bayesian learning				
3	59.15	66.95	62.52	82.31
5	59.30	67.00	62.76	82.37
10	59.31	67.02	62.91	82.42
25	59.27	66.99	62.88	82.40
50	59.37	67.11	63.02	82.46
Informed guessing				
3	32.44	40.47	15.21	44.64
5	32.05	40.17	15.54	44.50
10	32.25	40.19	15.47	44.51
25	32.22	40.26	15.92	44.37
50	31.77	39.80	15.32	44.25

Prediction

- After developing the training data, Li (2010) uses the Bayesian learning algorithm to categorize the tone and content of:
  - 13 million FLS from more than 140,000 10-Q and 10-K filings.
  - Sample period between 1994 and 2007.
  - Over 145,000 firm-quarters.

- Plenty of work in the paper about how *TONE* correlates with a set of other company-related variables. But how well does it predict earnings?

Predicting Earnings

TABLE 6

Future Earnings and MD&A Tone

COEFFICIENT	(1) EARN($t+1$)	(2) EARN($t+2$)	(3) EARN($t+3$)	(4) EARN($t+4$)
TONE	0.006***	0.005***	0.004**	0.003
	(4.63)	(3.25)	(2.20)	(1.34)
EARN	0.679***	0.616***	0.618***	0.626***
	(24.27)	(21.11)	(21.52)	(22.56)
RET	0.011***	0.008***	0.006***	0.005**
	(6.59)	(4.05)	(4.64)	(2.54)
ACC	−0.240***	−0.243***	−0.257***	−0.272***
	(−10.43)	(−10.88)	(−12.32)	(−9.89)
SIZE	0.002***	0.002***	0.002***	0.002***
	(7.40)	(7.09)	(6.74)	(5.96)
MTB	−0.002***	−0.003***	−0.003***	−0.005***
	(−3.95)	(−4.55)	(−5.95)	(−6.03)
RETVOL	−0.048***	−0.055***	−0.053***	−0.046***
	(−8.22)	(−6.36)	(−5.47)	(−5.52)
EARNVOL	−0.016***	−0.017***	−0.016**	−0.006*
	(−3.06)	(−2.94)	(−2.55)	(−1.83)
FOG	−0.001***	−0.001***	−0.001***	−0.001***
	(−7.92)	(−6.15)	(−6.85)	(−5.27)
NITEMS	0.000	0.000	0.000	0.000
	(0.01)	(0.24)	(0.15)	(0.75)
NBSEG	0.000	0.000	0.000	−0.000
	(0.91)	(0.15)	(0.17)	(−0.69)
NGSEG	0.002***	0.002***	0.003***	0.003***
	(3.90)	(4.78)	(4.58)	(4.86)
FIRMAGE	0.000***	0.000***	0.000***	0.000***
	(5.14)	(4.48)	(4.56)	(3.36)

Predicting Earnings

TABLE 7

Future Earnings Changes and MD&A Tone

COEFFICIENT	(1) DEARN($t+1$)	(2) DEARN($t+2$)	(3) DEARN($t+3$)	(4) DEARN($t+4$)
TONE	0.006***	0.004***	0.003**	0.002
	(4.74)	(2.93)	(1.98)	(1.25)
EARN	−0.324***	−0.370***	−0.357***	−0.362***
	(−12.37)	(−12.29)	(−14.71)	(−12.14)
RET	0.011***	0.008***	0.006***	0.005**
	(6.69)	(4.10)	(4.51)	(2.51)
ACC	−0.245***	−0.244***	−0.259***	−0.273***
	(−11.41)	(−10.63)	(−11.69)	(−9.91)
SIZE	0.002***	0.002***	0.002***	0.002***
	(7.89)	(7.01)	(6.75)	(6.02)
MTB	−0.002***	−0.003***	−0.003***	−0.005***
	(−4.06)	(−4.23)	(−6.22)	(−6.02)
RETVOL	−0.049***	−0.052***	−0.050***	−0.045***
	(−8.28)	(−5.71)	(−5.53)	(−5.30)
EARNVOL	−0.015***	−0.018***	−0.016**	−0.006*
	(−3.06)	(−2.86)	(−2.48)	(−1.80)
FOG	−0.001***	−0.001***	−0.001***	−0.001***
	(−8.16)	(−5.49)	(−6.45)	(−5.13)
NITEMS	−0.000	0.000	0.000	0.000
	(−0.00)	(0.28)	(0.16)	(0.74)
NBSEG	0.000	0.000	0.000	−0.000
	(0.97)	(0.27)	(0.30)	(−0.68)
NGSEG	0.002***	0.002***	0.002***	0.003***
	(3.90)	(4.72)	(4.64)	(4.92)
FIRMAGE	0.000***	0.000***	0.000***	0.000***
	(5.40)	(3.91)	(4.34)	(3.24)
MA	0.001	−0.000	−0.001	−0.004***
	(1.14)	(−0.20)	(−0.99)	(−2.61)
SEO	0.002*	0.000	0.002	0.003*
	(1.88)	(0.15)	(1.58)	(1.71)
SI	−0.429***	−0.406***	−0.404***	−0.363***
	(−12.05)	(−10.26)	(−10.29)	(−6.99)
DLW	−0.002***	−0.002***	−0.002***	−0.003**
	(−2.98)	(−3.33)	(−2.99)	(−2.58)
Q2	−0.004***	−0.011***	0.005***	−0.002**
	(−5.92)	(−7.93)	(3.58)	(−2.30)

Conclusion

- We have seen work on how to use textual analysis to extract unstructured information that can be used to predict earnings and stock returns.

- An extremely active area of investigation in many fields and throughout the private sector.

- Important differences between dictionary-based and statistical approaches, with the latter gaining popularity.

Imperial College
Business School

Imperial means
Intelligent Business