20598 – Finance with Big Data

Week 5 Lecture: Asset Pricing with Big Data

Clément Mazet-Sonilhac
clement.mazetsonilhac@unibocconi.it

Department of Finance, Bocconi University

Outline

Outline

What did we learned so far ?

Lecture 1: Markowitz's **Portfolio Theory**

- Theory: mean-variance framework
- Model: $\mathbb{E}[\tilde{r}_i] = ?$

What did we learned so far ?

Lecture 2: The **CAPM**

- Theory: mean-variance framework + Market equilibrium
- Model: $\mathbb{E}[\tilde{r}_i - r_f] = \beta_i \times (r_m - r_f)$

What did we learned so far ?

Lecture 3: More data, more factors

- Theory: No theory !
- Model: $\mathbb{E}[\tilde{r}_i] = \alpha_i + \beta_i^1 \times (r_m - r_f) + \beta_i^2 \times \text{SMB} + \beta_i^3 \times \text{HML} + \epsilon_i$

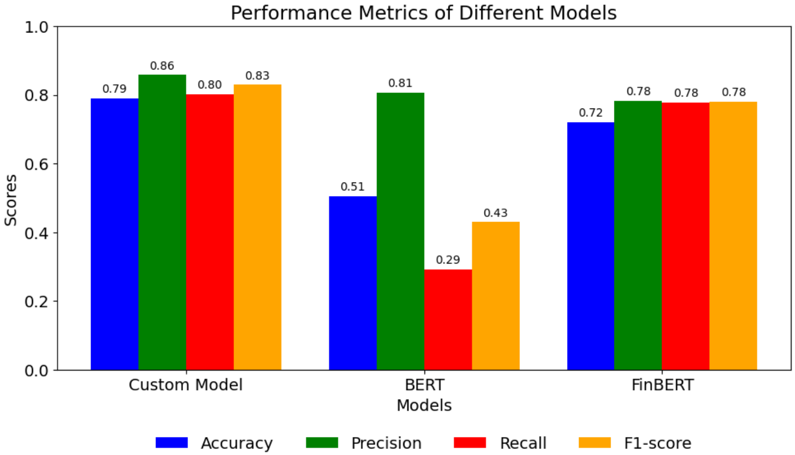
What did we learned so far ?

Lecture 3: More data, **too many** factors

- Theory: No theory !
- Model(s): $\mathbb{E}[\tilde{r}_i] = \alpha_i + \beta_i^1 \times (r_m - r_f) + \sum_k^N \beta_i^k \times \text{Factor}_k + \epsilon_i$

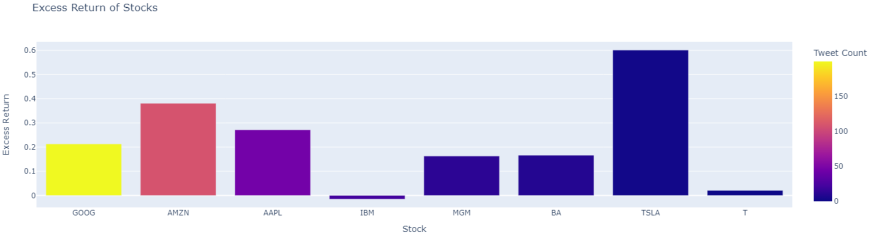
What did we learned so far ?

PC Lab#3: A Media Attention Factor ? – Great job overall



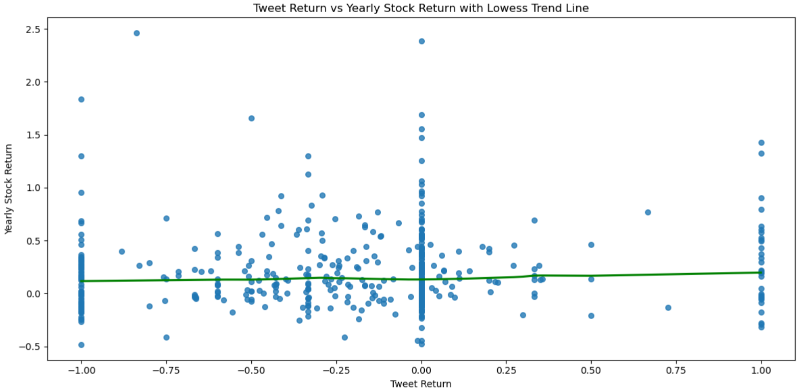
What did we learned so far ?

PC Lab#3: A Media Attention Factor ? – Great job overall



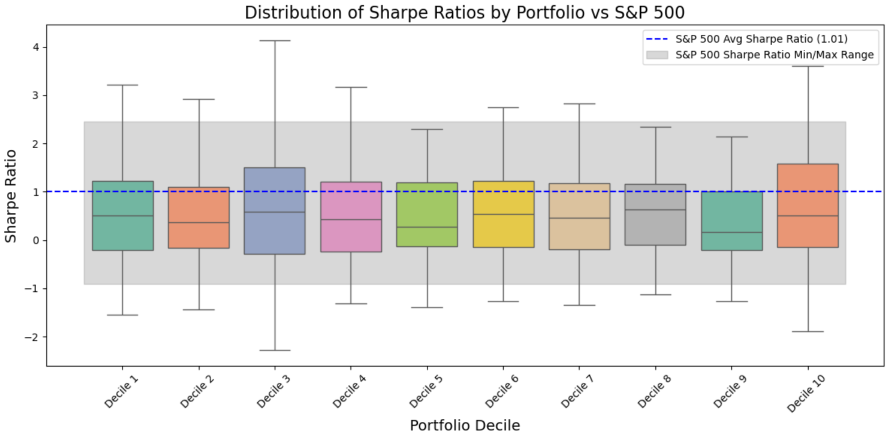
What did we learned so far ?

PC Lab#3: A Media Attention Factor ? – Great job overall



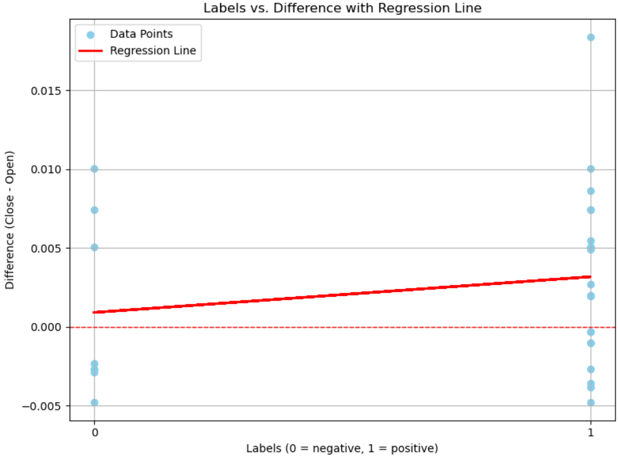
What did we learned so far ?

PC Lab#3: A Media Attention Factor ? – Great job overall



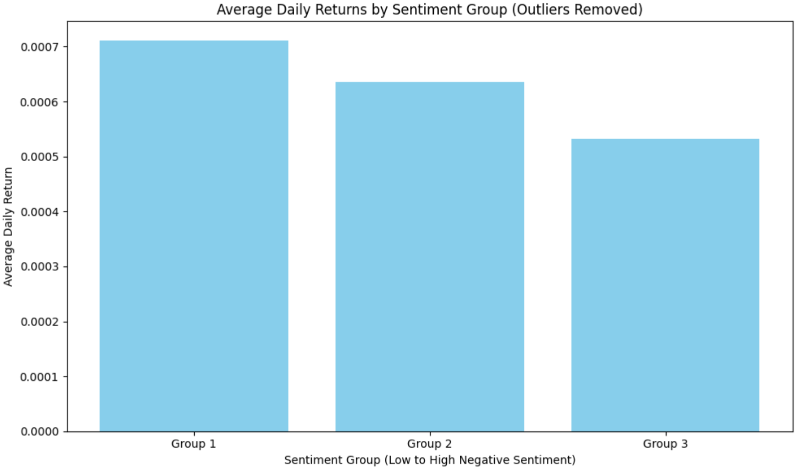
What did we learned so far ?

PC Lab#3: A Media Attention Factor ? – Great job overall



What did we learned so far ?

PC Lab#3: A Media Attention Factor ? – Great job overall



What did we learned so far ?

- Theory: No theory !
- Model(s): $\mathbb{E}[\tilde{r}_i] = \alpha_i + \beta_i^1 \times (r_m - r_f) + \sum_k^N \beta_i^k \times \text{Factor}_k + \epsilon_i$

What do we know and what is missing ?

What do people know?
Mental Models of the Stock Market, 2023

- Very (very) new research paper!
- Research question: Which mental models underlie stock return expectations ?
  - Market efficiency and risk factors (= what we've learned)
  - Mispricing or equilibrium neglect

Mental Models of the Stock Market, 2023

A thought experiment

Scenario 1
Neutral news, four weeks old

Scenario 2
Good news, four weeks old

You invest now, four weeks after the news broke.
In which of the two scenarios can you expect higher returns?

→ Your answer will depend of what is your model of the stock market

Sample

US general population (n ≈ 2,400)
Dynata, broadly representative

US retail investors (n ≈ 400)
Prolific, invested in stocks (median $90k), high income (median $100-150k)

US financial professionals (n ≈ 400)
- Financial advisors / investors / analysts
- CloudResearch, as in Chinco, Hartzmark and Sussman, 2022

Academic experts (n ≈ 100)
- Invited researchers who publ. in top journals, finance topics (2015–2019)

Mental Models of the Stock Market, 2023

Experiment

Please think about the following two hypothetical scenarios.

Scenario 1: Nike maintains supplier partnership
Four weeks ago, on April 29, 2023, Nike Inc. announced the continuation of its partnership with major polyester supplier Toray Industries Inc., in a move aimed at retaining its current supply chain. The continuation of the partnership is expected to maintain the company's current cost structure. Industry experts were not surprised by the announcement, as continuity in supplier relationships is a common practice in the industry.

Scenario 2: Nike secures cost-saving partnership
Four weeks ago, on April 29, 2023, Nike Inc. announced a new strategic partnership with leading recycled polyester supplier Unifi Inc., aimed at reducing raw material costs by 20%. The deal is expected to have a significant impact on Nike's bottom line, making its products more price-competitive. Industry experts were pleasantly surprised by the news and dubbed it an "unexpected success" for the company. They projected the move to significantly enhance Nike's market position in the sports apparel industry.

In both scenarios, the announcement was made four weeks ago and received a lot of attention from stock market traders.

Mental Models of the Stock Market, 2023

Experiment

- **Scenario 1: Nike maintains supplier partnership**
- **Scenario 2: Nike secures cost-saving partnership**

The announcements were made four weeks ago and received a lot of attention.

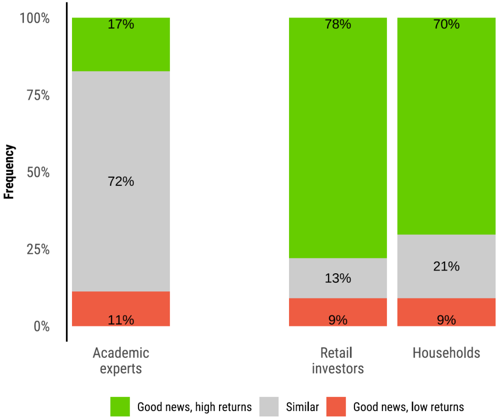
Imagine that you invest $1,000 in Nike stocks today, **four weeks after the announcement was made in the two scenarios**. Imagine that you sell these stocks in twelve months from now.

What would you expect? In which scenario would the return of this investment in Nike stocks be higher?

	The expected return would be ...		
	higher in scenario 1	**similar in both scenarios**	**higher in scenario 2**
Return of investment over the next twelve months Invest $1,000 in Nike stocks today, *four weeks after the announcement*. Sell these stocks one year from now.	○	○	○

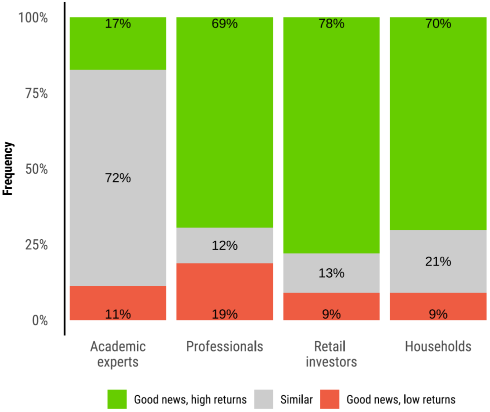
Mental Models of the Stock Market, 2023

Results



Mental Models of the Stock Market, 2023

Results



- Theory: No theory !
- Model(s): $\mathbb{E}[\tilde{r}_i] = \alpha_i + \beta_i^1 \times (r_m - r_f) + \sum_k^N \beta_i^k \times \text{Factor}_k + \epsilon_i$

What do we know and
what is missing ?

What did we learn so far: Limitations

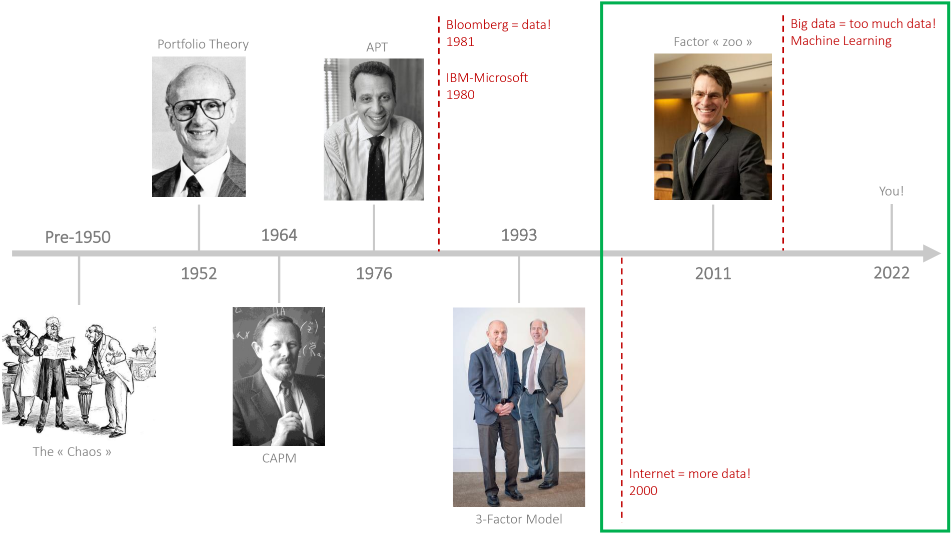
- Only traditional financial data

What did we learn so far: Limitations

- Only traditional financial data
- Explaining the cross section of expected return vs. predicting returns?

What did we learn so far: Limitations

- Only traditional financial data
- Explaining the cross section of expected return vs. predicting returns?
- The asset pricing world is linear



Portfolio Theory
APT
Bloomberg = data!
1981
IBM-Microsoft
1980
Factor « zoo »
Big data = too much data!
Machine Learning
You!
Pre-1950
1964
1993
1952
1976
2011
2022
The « Chaos »
CAPM
3-Factor Model
Internet = more data!
2000

Today: Alternative Data + ML

> Nowadays, analysts sift through non-traditional information such as satellite imagery and credit card data, or use artificial intelligence techniques such as machine learning and natural language processing to glean fresh insights from traditional sources such as economic data and earnings-call transcripts.
>
> —Robin Wigglesworth (2019)

Today : Alternative Data + ML

"Fund industry must use the right type of machine learning"

COMMENTARY | ISSUE 130

Artificial Intelligence

Chances and challenges in quantitative asset management

Finance & economics | Return on AI

Hedge funds embrace machine learning—up to a point

In investing, more artificial intelligence need not mean less of the human kind

"Why hedge fund managers are happy to let the machines take over"

"AI will rewrite the future of fund management"

NEWSLETTERS · EYE ON A.I.

Can an A.I. hedge fund beat the market?

BY JEREMY KAHN

August 25, 2020 at 4:05 PM GMT+2

"Fund managers must embrace AI disruption"

"Money managers seek AI's 'deep learning'"

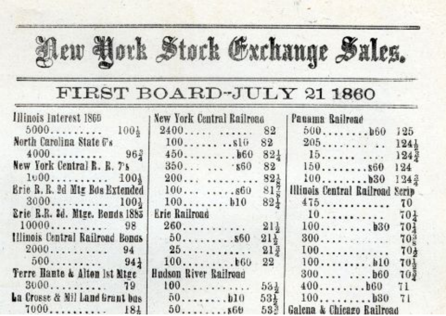
Part I : Alternative Data

A Brief History of Financial Data

- Financial econometrics has been driven the availability of financial data
- In 1800, how many stock markets?

A Brief History of Financial Data

- Financial econometrics has been driven the availability of financial data
- In 1800, how many stock markets? Only 4 in the world : London, Amsterdam, Paris and the US
- From 1815 to 1925, data from 3 major newspapers in the US : The New York Shipping List, The New York Herald and The New York Times (Goetzmann et al., 2001)



New York Stock Exchange Sales.

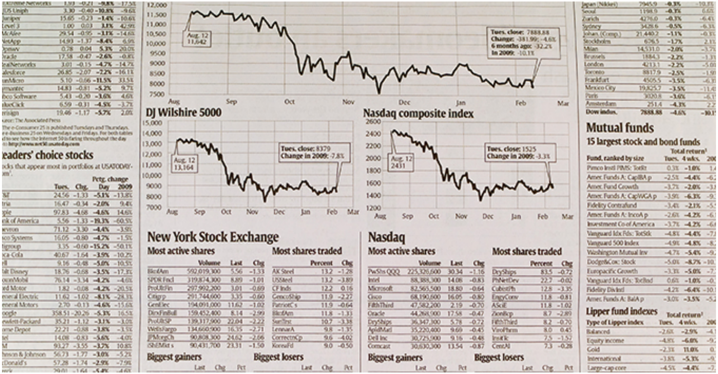
FIRST BOARD--JULY 21 1860

Illinois Interest 1860
5000.......... 100½
North Carolina State 6's
4000.......... 96¾
New York Central R. R. 7's
1000.......... 100½
Erie R. R. 2d Mtg Bds Extended
3000.......... 100½
Erie R.R. 3d. Mtge. Bonds 1883
10000.......... 98
Illinois Central Railroad Bonds
2000.......... 94
500.......... 94¼
Terre Haute & Alton 1st Mtge
3000.......... 79
La Crosse & Mil Land Grant bds
7000.......... 18½

New York Central Railroad
2400.......... 82
100..........s10 82
450..........b60 82¼
350..........s60 82
200.......... 82⅛
100..........s60 81⅞
100..........b10 82¼
Erie Railroad
260.......... 21½
50..........s60 21½
25.......... 21¾
100..........b60 22
Hudson River Railroad
100.......... 53½
50..........b10 53½
50..........s60 53⅜

Panama Railroad
500..........b60 125
205.......... 124½
15.......... 124¾
150..........s60 124
100..........b30 124¾
Illinois Central Railroad Scrip
475.......... 70
10.......... 70¼
100..........b30 70¼
300.......... 70⅜
100.......... 70½
100..........b10 70½
300..........b60 70¾
400..........b60 71
100..........b30 71
Galena & Chicago Railroad

A Brief History of Financial Data



→ Before the 1960s, financial data was scarce and almost impossible to compile

A Brief History of Financial Data

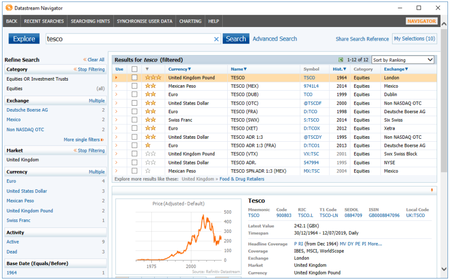
- Digitization: Ultronics Systems (1964), Reuters (1973) and Bloomberg (1981)

A Brief History of Financial Data

- 1960s-2000s, financial research was mainly driven by relatively small data sets
- Mostly price data (end-of-day quotes): low frequency

A Brief History of Financial Data

- From the 2000s, high-speed Internet diffusion allowed a better, faster access to larger datasets



- Prices + balance sheet data + news
- Still small compared to today standards: 1850 assets $\times$ 12 months $\times$ 20 years

A Brief History of Financial Data

- Today, finance professionals have relied on data terminals from companies such as Refinitiv (see Eikon Terminal) or Bloomberg (see Bloomberg Terminal)
- The major breakthrough in data-driven finance is to be seen in the programmatic availability of data via application programming interfaces (APIs)

In [ ]:
```
import eikon as ek
import configparser
```

In [ ]:
```
c = configparser.ConfigParser()
c.read('../../../data/aiif.cfg')  # adjust path
```

In [ ]:
```
ek.set_app_key(c['eikon']['app_id'])
```

A Brief History of Financial Data

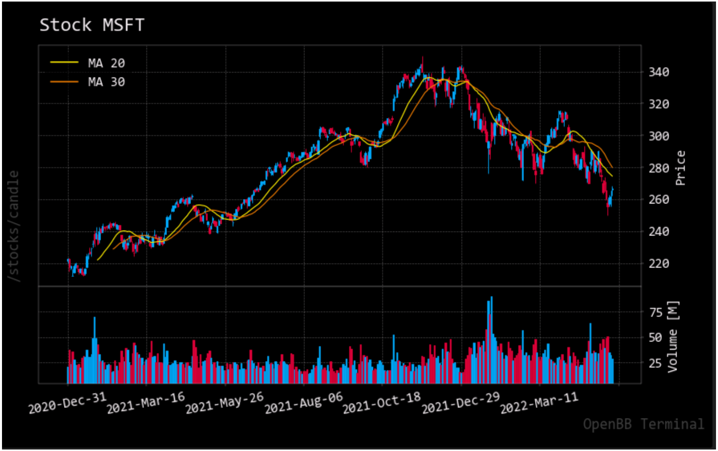
- Today, finance professionals have relied on data terminals from companies such as Refinitiv (see Eikon Terminal) or Bloomberg (see Bloomberg Terminal)
- The major breakthrough in data-driven finance is to be seen in the programmatic availability of data via application programming interfaces (APIs)

```
In [ ]: import eikon as ek
        import configparser

In [ ]: c = configparser.ConfigParser()
        c.read('../../../data/aiif.cfg')  # adjust path

In [ ]: ek.set_app_key(c['eikon']['app_id'])
```

- Still, some limitations: subscription is expensive ($22,000 per year for Eikon, $27,660+ for Bloomberg)

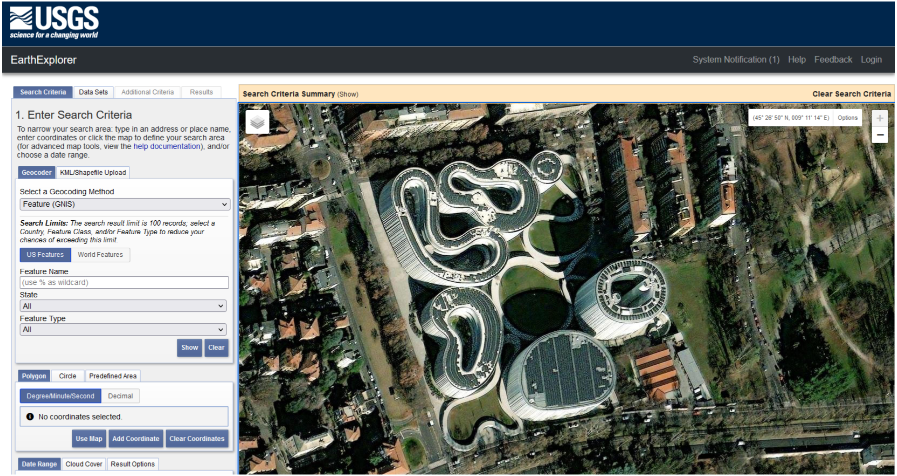
Today : OpenBB [Link]

- OpenBB: Python-based integrated environment for investment research
- Open-source Bloomberg-like



Beyond Financial Data

Satellite Data



- Free access : Sentinel Copernicus (UE), NASA, USGS
- High frequency (every hour)

Text as Data

Eric Trump
@EricTrump
Follow

In my opinion, it's a great time to buy stocks or into your 401K. I would be all in... let's see if I'm right...

11:47 AM - 28 Feb 2020 from Doral, FL

10,874 Retweets 51,901 Likes

Donald J. Trump
@realDonaldTrump

The Dow Jones Industrial just closed above 29,000! You are so lucky to have me as your President😉With Joe Hiden' it would crash😟

Tweet übersetzen

10:05 nachm. · 2. Sep. 2020 · Twitter for iPhone

52.505 Retweets 17.599 Zitierte Tweets 292.950 „Gefällt mir"-Angaben

Elizabeth Warren
@SenWarren
Following

BB&T Corporation has over 2,000 branches across the country and controls $221 billion in assets. It received $3.13 billion in taxpayer bailouts during the 2008 crash. But the #BankLobbyistAct could apply the same rules to @BBT as your neighborhood community bank.

9:45 AM - 13 Mar 2018

FINANCIAL TIMES

TUESDAY 4 FEBRUARY 2020 WORLD BUSINESS NEWSPAPER MIDDLE EAST

Fleet of dreams
An Uber plan to pay for millions of driverless cars — INSIDE FINANCE, PAGE 12

Unlocking Japan Inc
Activist investors force change in a risk-averse world — BIG READ, PAGE 7

Lost in the Valley
Venture capitalists are losing their edge — JOHN THORNHILL, PAGE 9

High stakes
Johnson sets trade stance

Briefing

Goldman nears Amazon loans pact as Main St push gathers pace

Bid to cater for small business ● Plan to use retailer's platform ● Bezos eyes expansion

Datawatch

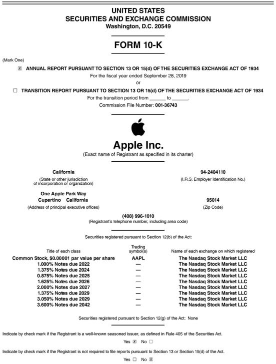
Citi suspends senior bond trader over alleged food theft from office canteen

Brazil faces backlash over shift in dirty money policy

World Markets

You have the power to change the world

A Nikkei Company



UNITED STATES
SECURITIES AND EXCHANGE COMMISSION
Washington, D.C. 20549

FORM 10-K

(Mark One)

☒ ANNUAL REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934

For the fiscal year ended September 28, 2019

or

☐ TRANSITION REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934

For the transition period from ______ to ______.

Commission File Number: **001-36743**

Apple Inc.
(Exact name of Registrant as specified in its charter)

California	**94-2404110**
(State or other jurisdiction of incorporation or organization)	(I.R.S. Employer Identification No.)
One Apple Park Way Cupertino California	**95014**
(Address of principal executive offices)	(Zip Code)

(408) 996-1010
(Registrant's telephone number, including area code)

Securities registered pursuant to Section 12(b) of the Act:

Title of each class	Trading symbol(s)	Name of each exchange on which registered
Common Stock, $0.00001 par value per share	**AAPL**	**The Nasdaq Stock Market LLC**
1.000% Notes due 2022	—	**The Nasdaq Stock Market LLC**
1.375% Notes due 2024	—	**The Nasdaq Stock Market LLC**
0.875% Notes due 2025	—	**The Nasdaq Stock Market LLC**
1.625% Notes due 2026	—	**The Nasdaq Stock Market LLC**
2.000% Notes due 2027	—	**The Nasdaq Stock Market LLC**
1.375% Notes due 2029	—	**The Nasdaq Stock Market LLC**
3.050% Notes due 2029	—	**The Nasdaq Stock Market LLC**
3.600% Notes due 2042	—	**The Nasdaq Stock Market LLC**

Securities registered pursuant to Section 12(g) of the Act: None

Indicate by check mark if the Registrant is a well-known seasoned issuer, as defined in Rule 405 of the Securities Act.

Yes ☒ No ☐

Indicate by check mark if the Registrant is not required to file reports pursuant to Section 13 or Section 15(d) of the Act.

Yes ☐ No ☒

Text as Data

First attempt of textual analysis

CAN STOCK MARKET FORECASTERS FORECAST?

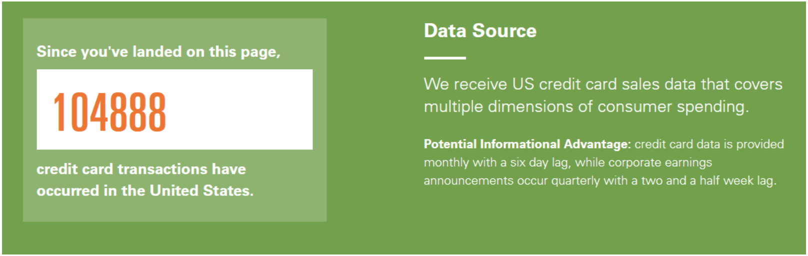
By Alfred Cowles 3rd

A paper read before a joint meeting of the Econometric Society and the American Statistical Association, Cincinnati, Ohio, December 31, 1932

INTRODUCTION

This paper presents results of analyses of the forecasting efforts of 45 professional agencies which have attempted, either to select specific common stocks which should prove superior in investment merit to the general run of equities, or to predict the future movements of the stock market itself. The paper falls into two main parts. The first deals with the attempts of two groups, 20 fire insurance companies and 16 financial services, to foretell which specific securities would prove most profitable. The second part deals with the efforts of 25 financial publications to foretell the future course of the stock market. Various statistical tests of these results are given.

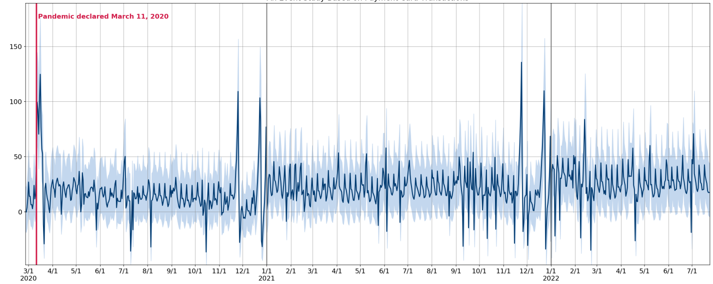
Credit Card Data



- Private data (e.g., see Quantinomics)
- Very high frequency

Spending on Food and Beverage Stores

An Event Study Based on Payment Card Transactions

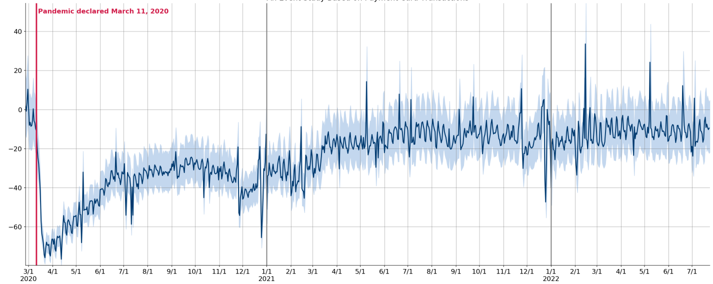


Note: Chart shows the difference from the typical level of spending without COVID-19-related changes in the economy. The typical level corresponds to a value of 0.
The shaded area represents 95 percent confidence interval bands.

U.S. Bureau of Economic Analysis

Spending on Food Services and Drinking Places

An Event Study Based on Payment Card Transactions



Note: Chart shows the difference from the typical level of spending without COVID-19-related changes in the economy. The typical level corresponds to a value of 0.
The shaded area represents 95 percent confidence interval bands.

U.S. Bureau of Economic Analysis

Beyond Financial Data

- Text: newspapers, twitter, companies' reports
- Images: satellite data
- Audio, videos
- Real-time consumption: credit card data
- Much more: search trends, web traffic, supply chain, energy production, etc.

Alternative Data : Recent Applications to Asset Pricing

How Is This Useful in Asset Pricing ?

Example 1 : The Wisdom of Twitter Crowds (Azar and Lo, 2016)

- Research Questions :

    1. Do user messages contain relevant information for asset pricing ?

How Is This Useful in Asset Pricing ?

Example 1 : The Wisdom of Twitter Crowds (Azar and Lo, 2016)

- Research Questions :

    1. Do user messages contain relevant information for asset pricing ?
    2. Can this information be inferred from more traditional sources, or is it truly **new** information ?

How Is This Useful in Asset Pricing ?

Example 1 : The Wisdom of Twitter Crowds (Azar and Lo, 2016)

- Research Questions :

  1. Do user messages contain relevant information for asset pricing ?
  2. Can this information be inferred from more traditional sources, or is it truly **new** information ?
  3. Can social media data help predict future asset returns and shifts in volatility ?

How Is This Useful in Asset Pricing ?

Example 1 : The Wisdom of Twitter Crowds (Azar and Lo, 2016)

- Research Questions :

  1. Do user messages contain relevant information for asset pricing ?
  2. Can this information be inferred from more traditional sources, or is it truly **new** information ?
  3. Can social media data help predict future asset returns and shifts in volatility ?

- Context: analysis of a recurring event that has a high impact on asset prices :

→ Federal Open Markets Committee (FOMC) meetings

The Wisdom of Twitter Crowds (Azar and Lo, 2016)

Context : Federal Open Markets Committee

- Eight times a year, the FOMC meets to determine monetary policy

- Decisions made by the FOMC are highly watched by all market participants + significantly affect asset prices

The Wisdom of Twitter Crowds (Azar and Lo, 2016)

Tweets as Data (1)

- New dataset of tweets that cite the Federal Reserve
  - Period : 2007–2014
  - Tweets that mentioned *FOMC* or *Federal Reserve* (Topsy API)
  - Tweets that mentioned “Bernanke” or “Yellen” (Why ?)
- Using NLP methods, they assign a polarity score to each tweet

The Wisdom of Twitter Crowds (Azar and Lo, 2016)

Tweets as Data (2)

- Tweet sentiment use Python package called `Pattern` (De Smedt et al., 2012)
- Each tweet is associated with a polarity score between −1 and +1

Tweet	Polarity
@GStuedler this was caused by the worst regulation of all time, the Federal Reserve Act of 1913. Decoupled money from reality #topprog #tcot	−1.00
@SenJohnMcCain Maybe instead of partisan bickering, you should all come together and go after the real bad guys... The Federal Reserve	−1.00
Bernanke Says Biggest Worry is That Politicians Abandon Banks: (CEP News) - U.S. Federal Reserve Chairman Ben Be.. http://tinyurl.com/cosb9m	−0.75
Treasurys rise as Fed meets: Treasury prices rose Tuesday amid speculation that the Federal Reserve will begin b.. http://tinyurl.com/c42qjb	0.60
RT @jordangunderson: is proud that Jason Chaffetz is 1 of 28 Congressmen cosponsoring Ron Paul's Federal Reserve Transparency Act (HR 1207).	0.80
Very impressed w/Federal Reserve Chair Ben Bernanke!	1.00

The Wisdom of Twitter Crowds (Azar and Lo, 2016)

Tweets as Data (3)

- Tweets are associated with the number of followers of their user (used as weight)

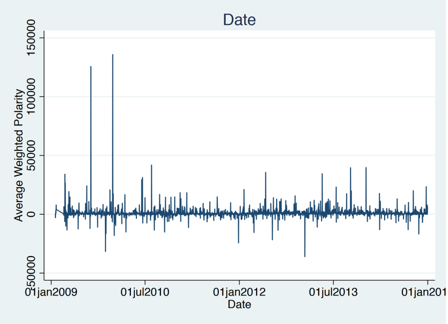


Figure 1 – Polarity

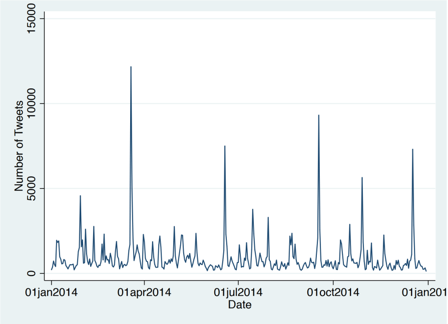


Figure 2 – Number of tweets

The Wisdom of Twitter Crowds (Azar and Lo, 2016)

Empirical Analysis

- Baseline equation estimated by OLS :

$$r_t - r_f = \alpha + \beta_1 \mathbb{I}_{\mathsf{FOMC},t} + \beta_2 \mathsf{TweetPolarity}_{t-1} + \beta_3 \mathbb{I}_{\mathsf{FOMC},t} \times \mathsf{TweetPolarity}_{t-1} \\ + \gamma_1 \mathsf{HML}_t + \gamma_2 \mathsf{SMB}_t + \gamma_3 \mathsf{UMD}_t + \gamma_4 \mathsf{VIX}_t + \gamma_5 (r_{t-1} - r_f) + \epsilon_t$$

- Coefficient of interest = β_3

The Wisdom of Twitter Crowds (Azar and Lo, 2016)

Empirical Analysis

- Baseline equation estimated by OLS :

$$r_t - r_f = \alpha + \beta_1 \, \mathbb{I}_{\mathsf{FOMC},t} + \beta_2 \text{TweetPolarity}_{t-1} + \beta_3 \, \mathbb{I}_{\mathsf{FOMC},t} \times \text{TweetPolarity}_{t-1} \\ + \gamma_1 \text{HML}_t + \gamma_2 \text{SMB}_t + \gamma_3 \text{UMD}_t + \gamma_4 \text{VIX}_t + \gamma_5 (r_{t-1} - r_f) + \epsilon_t$$

- Coefficient of interest = β_3
- $r_t - r_f$ = Daily Excess Return of the market (Why to use a lag ?)

- Baseline equation estimated by OLS :

$$r_t - r_f = \alpha + \beta_1 \, \mathbb{I}_{\textsf{FOMC},t} + \beta_2 \textsf{TweetPolarity}_{t-1} + \beta_3 \, \mathbb{I}_{\textsf{FOMC},t} \times \textsf{TweetPolarity}_{t-1} \\ + \gamma_1 \textsf{HML}_t + \gamma_2 \textsf{SMB}_t + \gamma_3 \textsf{UMD}_t + \gamma_4 \textsf{VIX}_t + \gamma_5 (r_{t-1} - r_f) + \epsilon_t$$

- Coefficient of interest = β_3
- $r_t - r_f$ = Daily Excess Return of the market (Why to use a lag ?)
- HML, SMB, UMD $\approx$ Fama-French factors = controls (something to say ?)

- Baseline equation estimated by OLS :

$$r_t - r_f = \alpha + \beta_1 \mathbb{I}_{\mathsf{FOMC},t} + \beta_2 \mathsf{TweetPolarity}_{t-1} + \beta_3 \mathbb{I}_{\mathsf{FOMC},t} \times \mathsf{TweetPolarity}_{t-1}$$
$$+ \gamma_1 \mathsf{HML}_t + \gamma_2 \mathsf{SMB}_t + \gamma_3 \mathsf{UMD}_t + \gamma_4 \mathsf{VIX}_t + \gamma_5 (r_{t-1} - r_f) + \epsilon_t$$

- Coefficient of interest = β_3
- $r_t - r_f$ = Daily Excess Return of the market (Why to use a lag ?)
- HML, SMB, UMD $\approx$ Fama-French factors = controls (something to say ?)
- VIX = Volatility Index

The Wisdom of Twitter Crowds (Azar and Lo, 2016)

Results

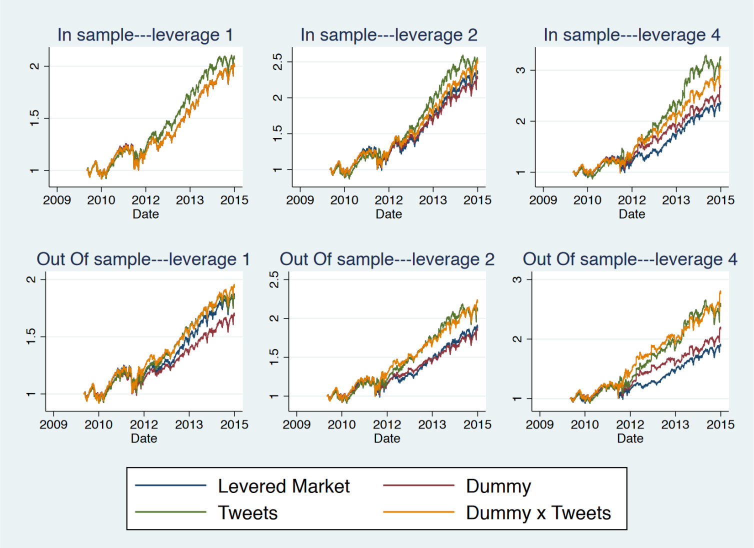
VARIABLES	(1) Return	(2) Return	(3) Return	(4) Return
IndicatorFOMC	0.331*	0.338*	0.398**	0.343**
	(0.187)	(0.187)	(0.177)	(0.140)
TweetPolarity		0.0510**	0.0493*	0.0156
		(0.0249)	(0.0252)	(0.0195)
TweetPolarityFOMC			0.490	0.625**
			(0.529)	(0.296)
hml				0.843***
				(0.0757)
smb				0.857***
				(0.0640)
umd				-0.141**
				(0.0602)
L.Return	-0.0722*	-0.0715*	-0.0716*	-0.0334
	(0.0384)	(0.0384)	(0.0384)	(0.0296)
Constant	0.0630**	0.0628**	0.0628**	0.0516**
	(0.0313)	(0.0312)	(0.0313)	(0.0246)
Observations	1,506	1,506	1,506	1,506

Standard errors in parentheses

*** $p<0.01$, ** $p<0.05$, * $p<0.1$

The Wisdom of Twitter Crowds (Azar and Lo, 2016)

Performance of the Twitter-Enhanced Trading Strategy



A bit more sophisticated :

Ke, Kelly and Xiu, 2019

How Is This Useful in Asset Pricing ?

Predicting Returns with Text Data (Ke, Kelly, Xiu, 2019)

- This paper revisits the most studied text-based research question in finance:

→ Do business news explains and predicts observed asset price variation ?

- Many finance papers manage the text dimensionality challenge by:
  - restricting their analysis to words in pre-existing sentiment dictionaries
  - using ad hoc word-weighting schemes

How Is This Useful in Asset Pricing ?

Predicting Returns with Text Data (Ke, Kelly, Xiu, 2019)

- This paper revisits the most studied text-based research question in finance:

→ Do business news explains and predicts observed asset price variation ?

- Many finance papers manage the text dimensionality challenge by:
  - restricting their analysis to words in pre-existing sentiment dictionaries
  - using ad hoc word-weighting schemes

- This paper: machine learning techniques can be used to understand the sentimental structure of a text corpus without relying on pre-existing dictionaries

Predicting Returns with Text Data (Ke, Kelly, Xiu, 2019)

- Key aspect: learning the sentiment scoring model from the joint behavior of article text and stock returns – rather than taking sentiment scores off the article

Predicting Returns with Text Data (Ke, Kelly, Xiu, 2019)

- Key aspect: learning the sentiment scoring model from the joint behavior of article text and stock returns – rather than taking sentiment scores off the article
- Idea: find individual terms – positive or negative – that most frequently coincide with returns of the same sign
- Three steps :
  1. isolate a list of sentiment terms via predictive screening
  2. assign sentiment weights to these words
  3. aggregate terms into an article-level sentiment score via penalized likelihood

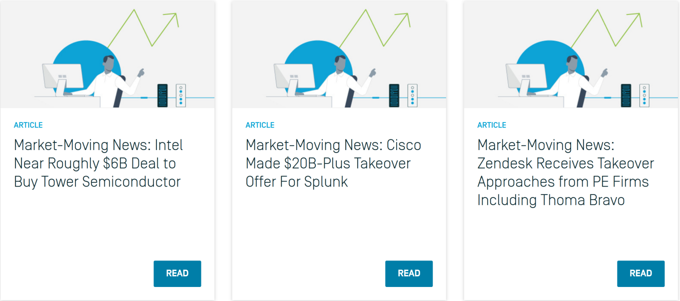
Predicting Returns with Text Data (Ke, Kelly, Xiu, 2019)

- Key aspect: learning the sentiment scoring model from the joint behavior of article text and stock returns – rather than taking sentiment scores off the article
- Idea: find individual terms – positive or negative – that most frequently coincide with returns of the same sign
- Three steps :
  1. isolate a list of sentiment terms via predictive screening
  2. assign sentiment weights to these words
  3. aggregate terms into an article-level sentiment score via penalized likelihood
- Empirical analysis :
  - text-mining one of the most actively monitored streams of news articles
  - supervised sentiment model excels at extracting return-predictive signals

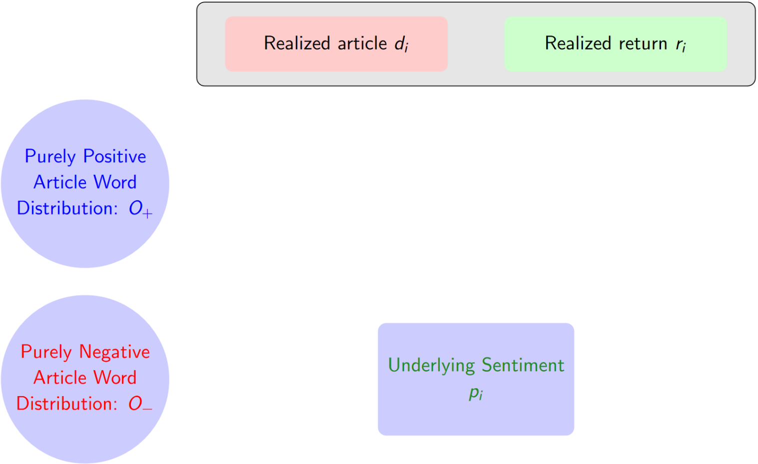
Predicting Returns with Text Data (Ke, Kelly, Xiu, 2019)

Data : Dow Jones Newswires (1.56 billion-article database)

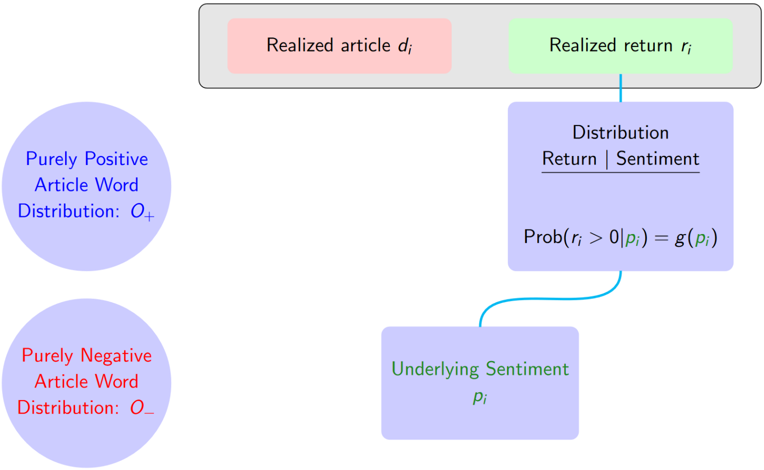
- 38-year of data
- Time-stamped articles + tagged with identifiers of firms involved
- Match articles with stock return data from CRSP



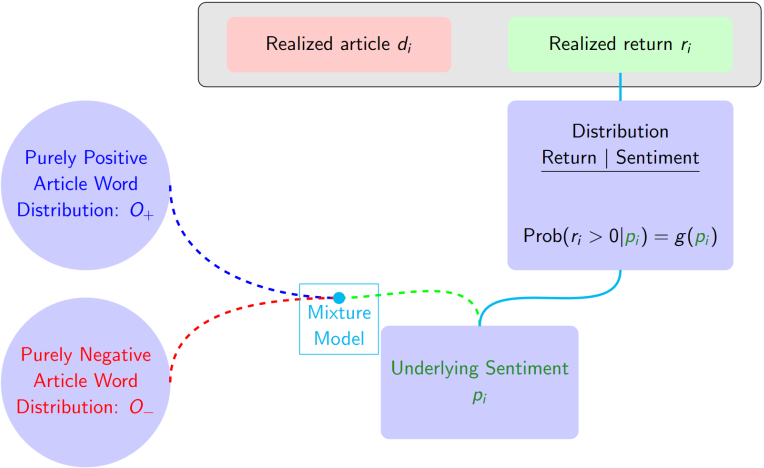
Predicting Returns with Text Data (Ke, Kelly, Xiu, 2019)



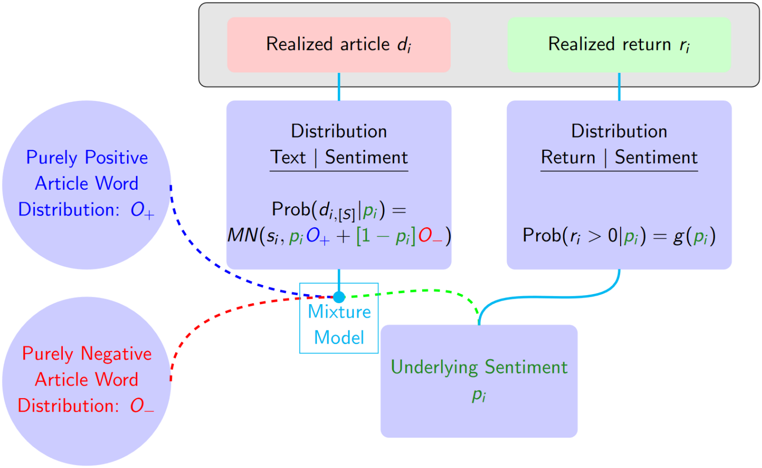
Predicting Returns with Text Data (Ke, Kelly, Xiu, 2019)



Predicting Returns with Text Data (Ke, Kelly, Xiu, 2019)



Predicting Returns with Text Data (Ke, Kelly, Xiu, 2019)



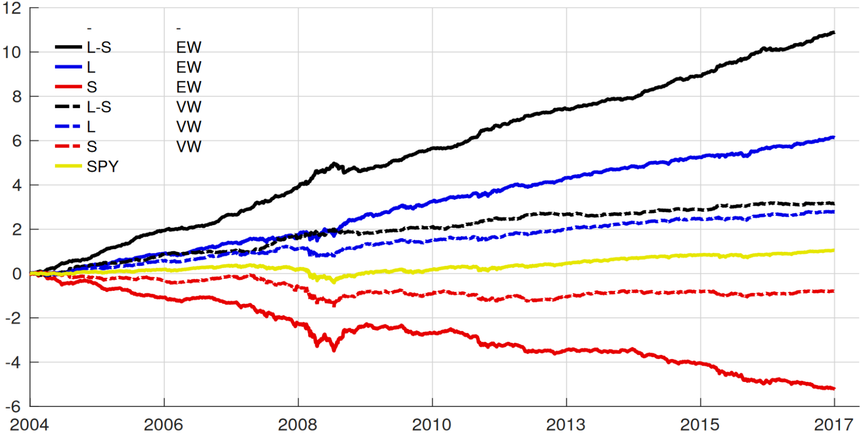
Predicting Returns with Text Data (Ke, Kelly, Xiu, 2019)

Results

- Simple trading strategy that buys assets with positive recent news sentiment and sells assets with negative sentiment
- The portfolio based on the model delivers excellent risk-adjusted out-of-sample returns
- Moreover, it outperforms a similar strategy based on scores from RavenPack (the industry-leading commercial vendor of financial news sentiment scores)

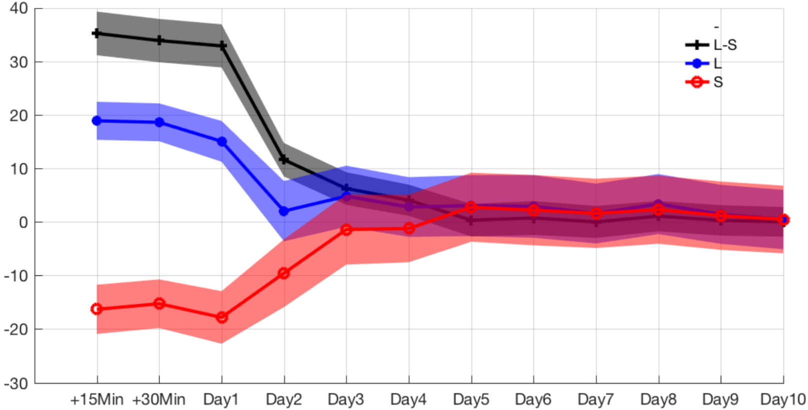
Predicting Returns with Text Data (Ke, Kelly, Xiu, 2019)

Results: Long-term portfolio performance



Predicting Returns with Text Data (Ke, Kelly, Xiu, 2019)

Results: Short-term arbitrage opportunities



So, it's improving markets' efficiency ?

So, it's improving markets' efficiency ?

Hmmmmm, let's see...

On the Capital Market Consequences of Alternative Data: Evidence from Outer Space (Katona et al., 2022)

- Tom Diamond: co-founder of RS-Metrics
- In 2009, visit his brother who worked for DigitalGlobe, a company that sells satellite photos
- Finding: the number of cars in a retailer's parking lots accurately predicts the company's revenues
- Created a company, became extremely rich

RS-Metrics (> $100M)

Example of Data Use : The Case of China Recovery

SIGN IN SUBSCRIBE

THE WALL STREET JOURNAL.

BUSINESS

China Says Growth Is Fine. Private Data Show a Sharper Slowdown.

Beneath China's stable headline numbers, there is a growing belief that the real picture is much worse

On the Capital Market Consequences of Alternative Data: Evidence from Outer Space (Katona et al., 2022)

- Researchers asked the Diamond brothers for their retail-parking-lot data
- Daily car counts conducted from 2011 to 2017 at 67,000 stores representing 44 major U.S. retailers, among them Costco, Nordstrom, Starbucks, Target, Walmart, and Whole Foods.

On the Capital Market Consequences of Alternative Data : Evidence from Outer Space (Katona et al., 2022)

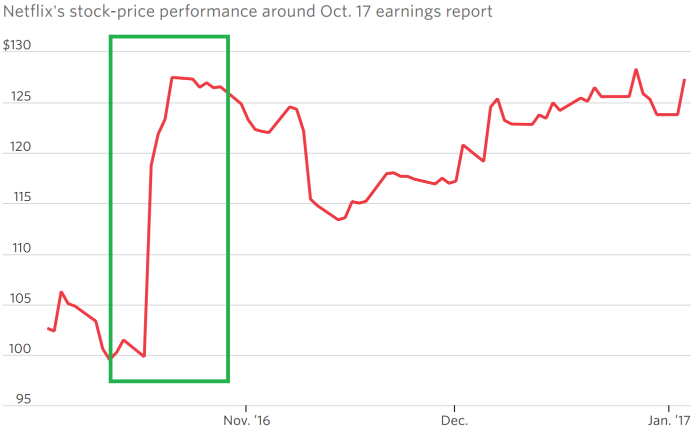
- Sophisticated investors with access to satellite data formulate profitable trading strategies :

> *During the weeks before a retailer reported quarterly earnings, if you had bought its shares when parking-lot traffic increased abnormally, and sold its shares when it declined, you would have earned a return that was 4.7 percent higher than the typical benchmark return [...]*

- Use of satellite imagery creates opportunities for sophisticated investors at the expense of small individual investors

> *[...] Exacerbate capital income inequality between more sophisticated investors, who can afford to acquire costly private information, and less sophisticated investors, who have little access to private information.*

Credit-Card Data Adds Volatility to Markets



- On Oct. 5 2015, credit-card data supplier sent a note to clients saying Netflix's streaming subscriber numbers were below consensus at the end of the 3^{rd} quarter
- On Oct. 18, shares increased by 19% after Netflix beat consensus estimates

Part II : New Methods
ML in Asset Pricing

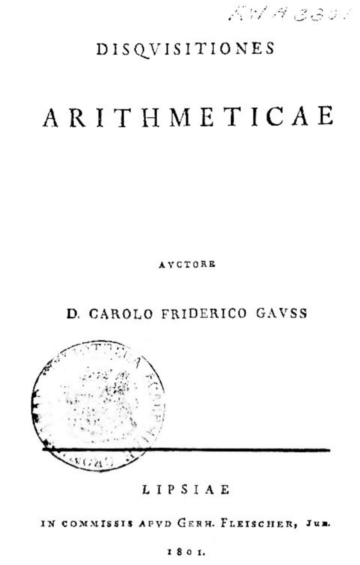
Low-Tech Finance : Linear Regression

- OLS = one of the central tools in financial econometrics

- Many reasons why, can you guess ?

Low-Tech Finance : Linear Regression

- OLS = one of the central tools in financial econometrics
- Many reasons why, can you guess ?
  - Centuries old (18th-century technology, mastered by Gauss in 1794)
  - Simplicity: mathematics of OLS is easy to understand and to implement
  - Scalability: no limit regarding the data size to which OLS can be applied
  - Speed: OLS regression is fast to evaluate, even on larger data sets
  - Availability: efficient implementations many programming languages



Low-Tech Finance : Linear Regression

Limitations

- The application of OLS regression rests on a number of (strong) assumptions :
- Can you guess ?

Low-Tech Finance : Linear Regression

Limitations

- The application of OLS regression rests on a number of (strong) assumptions :
- Can you guess ?
  1. Linearity: The model is linear in its parameters, with regard to both the coefficients and the residuals
  2. Independence: variables are not perfectly (to a high degree) correlated with each other (no multicollinearity)
  3. Number of observations $\gg$ number of parameters
- Many modern asset pricing questions contradict (1), (2) and/or (3)

How to Apply ML in Asset Pricing ?

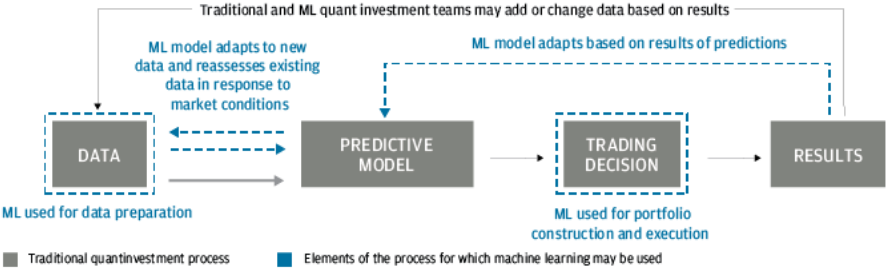
- Dimensionality and Independence: we have seen last week that Penalized Regressions : LASSO, Ridge, Elastic-Net

- Non-linearity: Regressions Trees and Random Forests, RNN, DNN, etc.

How to Apply ML in Asset Pricing ?

- Dimensionality and Independence: we have seen last week that Penalized Regressions : LASSO, Ridge, Elastic-Net

- Non-linearity: Regressions Trees and Random Forests, RNN, DNN, etc.



- What would YOU do ?

Empirical Asset Pricing via Machine Learning (Gu et al., 2020)

- Perform a comparative analysis of machine learning methods for the canonical problem of empirical asset pricing: predicting asset returns (i.e., measuring asset risk premium)
- Researchers demonstrate large economic gains to investors using machine learning forecasts
- In some cases, doubles the performance of leading regression-based strategies from the literature

Empirical Asset Pricing via Machine Learning (Gu et al., 2020)

- Perform a comparative analysis of machine learning methods for the canonical problem of empirical asset pricing: predicting asset returns (i.e., measuring asset risk premium)

- Researchers demonstrate large economic gains to investors using machine learning forecasts

- In some cases, doubles the performance of leading regression-based strategies from the literature

- They identify the best-performing methods (trees and neural networks) and trace their predictive gains to allowing nonlinear predictor interactions missed by other methods

- All methods agree on the same set of dominant predictive signals, a set that includes variations on momentum, liquidity, and volatility

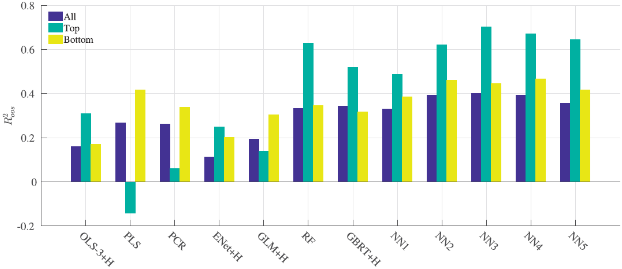
Empirical Asset Pricing via Machine Learning (Gu et al., 2020)

- Large-scale empirical analysis, investigating nearly 30,000 individual stocks over 60 years from 1957 to 2016
- Predictor set includes 94 characteristics for each stock, interactions of each characteristic with 8 aggregate time-series variables, and 74 industry sector dummy variables $\rightarrow$ totaling more than 900 baseline signals
- Divide the 60 years of data into 18 years of training sample (1957–1974), 12 years of validation sample (1975–1986), and the remaining 30 years (1987–2016) for out-of-sample testing
- Many models : OLS with all variables (OLS), OLS using only size, book-to-market, and momentum (OLS-3), PLS, PCR, elastic net (ENet), generalize linear model (GLM), random forest (RF), gradient boosted regression trees (GBRT), and neural networks with 1 to 5 layers (NN1–NN5).

First step : Model error

Monthly out-of-sample stock-level prediction performance (percentage R^2_{oos})

	OLS +H	OLS-3 +H	PLS	PCR	ENet +H	GLM +H	RF	GBRT +H	NN1	NN2	NN3	NN4	NN5
All	−3.46	0.16	0.27	0.26	0.11	0.19	0.33	0.34	0.33	0.39	0.40	0.39	0.36
Top 1,000	−11.28	0.31	−0.14	0.06	0.25	0.14	0.63	0.52	0.49	0.62	0.70	0.67	0.64
Bottom 1,000	−1.30	0.17	0.42	0.34	0.20	0.30	0.35	0.32	0.38	0.46	0.45	0.47	0.42



In this table, we report monthly R^2_{oos} for the entire panel of stocks using OLS with all variables (OLS), OLS using only size, book-to-market, and momentum (OLS-3), PLS, PCR, elastic net (ENet), generalize linear model (GLM), random forest (RF), gradient boosted regression trees (GBRT), and neural networks with 1 to 5 layers (NN1–NN5). "+H" indicates the use of Huber loss instead of the l_2 loss. We also report these R^2_{oos} within subsamples that include only the top-1,000 stocks or bottom-1,000 stocks by market value. The lower panel provides a visual comparison of the R^2_{oos} statistics in the table (omitting OLS because of its large negative values).

Empirical Asset Pricing via Machine Learning

Second step : ML-portfolios performance

- They design a new set of portfolios to directly exploit machine learning forecasts
- At the end of each month, they calculate 1-month-ahead out-of-sample stock return predictions for each method
- Sort stocks into deciles based on each model's forecasts and reconstitute portfolios each month using value weights
- Construct a zero-net-investment portfolio that buys the highest expected return stocks (decile 10) and sells the lowest (decile 1).

Empirical Asset Pricing via Machine Learning

Second step : ML-portfolios performance

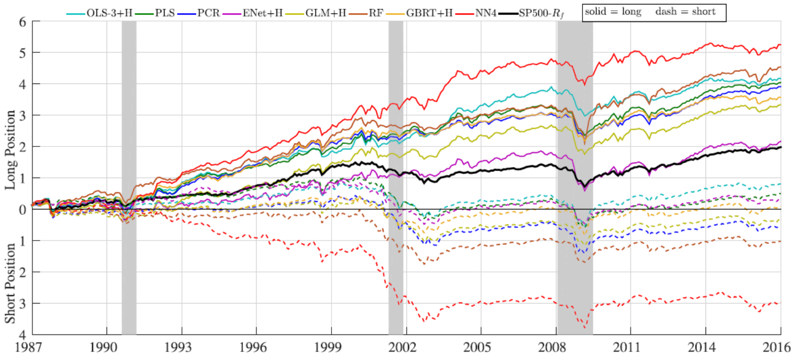


Figure 9
Cumulative return of machine learning portfolios
The figure shows the cumulative log returns of portfolios sorted on out-of-sample machine learning return forecasts. The solid and dashed lines represent long (top decile) and short (bottom decile) positions, respectively. The shaded periods show NBER recession dates. All portfolios are value weighted.

Empirical Asset Pricing via Machine Learning (Gu et al., 2020)

- A portfolio strategy that times the S&P 500 with neural network forecasts enjoys an annualized out-of-sample Sharpe ratio of 0.77 versus the 0.51 Sharpe ratio of a buy-and-hold investor
- A value-weighted long-short decile spread strategy that takes positions based on stock-level neural network forecasts earns an annualized out-of-sample Sharpe ratio of 1.35, more than doubling the performance of a leading regression-based strategy from the literature

To go further : Deep Learning Asset Pricing

Link to the paper

Education Strategies About Us MyAQR

Deep Learning in Asset Pricing

Luyang Chen, Stanford University; Markus Pelger, Stanford University; Jason Zhu, Stanford University

In this paper, the authors use deep neural networks to estimate an asset pricing model for individual stock returns that takes advantage of the vast amount of conditioning information, while keeping a fully flexible form and accounting for time-variation. The key innovations are to use the fundamental no-arbitrage condition as criterion function, to

Download

(Re)-imag(in)ing Price Trends

Link to the paper

- Technical analysis: (old-school) trading method based on the **observation of statistical trends**, such as price movement and volume.

(Re)-imag(in)ing Price Trends

Link to the paper

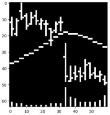
- Technical analysis: (old-school) trading method based on the **observation of statistical trends**, such as price movement and volume.

- This paper: Test **trend-based predictability** using methods that flexibly learn price patterns that are most predictive of future returns

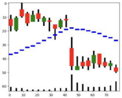
(Re)-imag(in)ing Price Trends

Link to the paper

Ticker	Date	Price	Volume
A	20030101	19.14	2,418,500
A	20030102	19.05	1,875,900
ZTS	20221028	153.28	1,615,515
ZTS	20221031	150.78	1,899,045



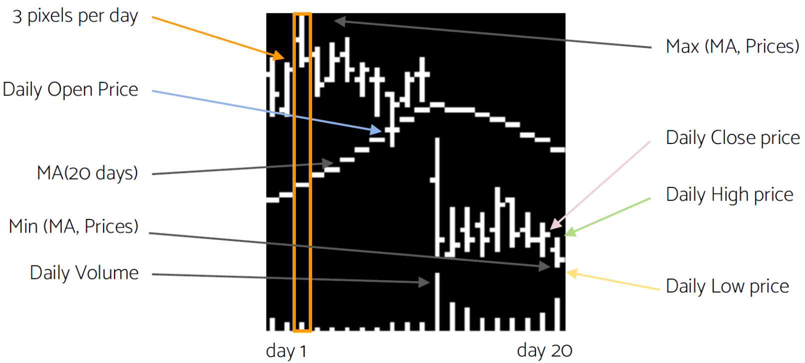
Kelly et al. (2020, *The Journal of Finance*)



Thesis extension

(Re)-imag(in)ing Price Trends

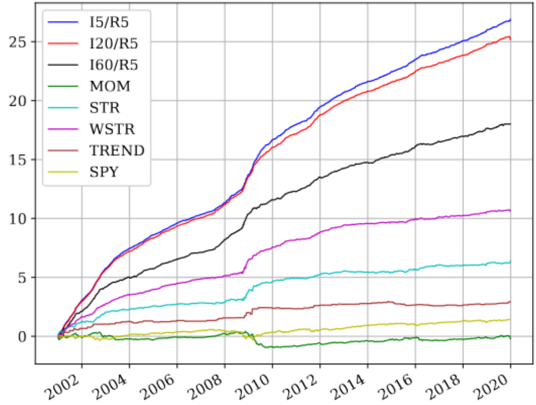
Link to the paper



(Re)-imag(in)ing Price Trends

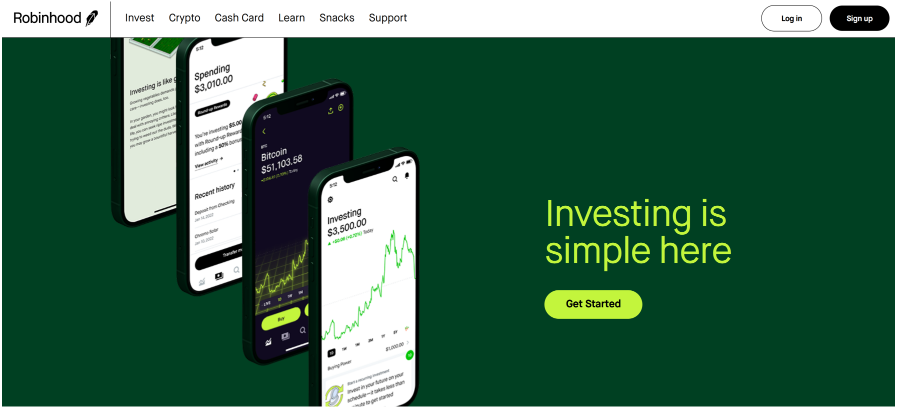
Link to the paper

Figure 5: Cumulative Volatility Adjusted Returns of Equal-Weighted Portfolios



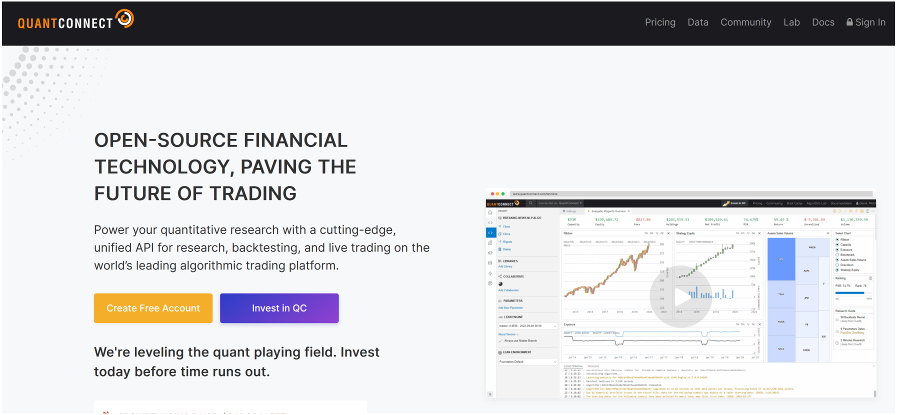

Wealth
Mangement
Disruption?
2:53
S&P 500
4,287.48
AAPL
136.19
+1.41
1D
1W
1M
3M
6M
1Y
2Y
261
249
224
29
Open
High
Low
250.20
261.08
250.01
Vol
P/E
Mkt Cap
6.955M
-
67.03B
52W H
52W L
Avg Vol
More Data from Yahoo Finance
News

Available data + open-source algorithms : RobinHood



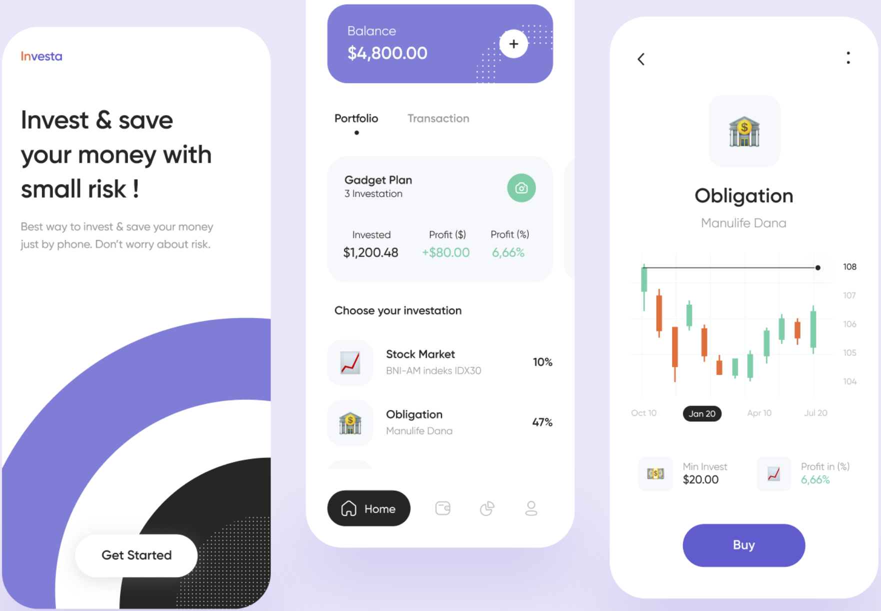
- Online retail brokerage company founded in 2013
- No brokerage fees, which allowed clients to buy and sell single shares of stocks.
- Attracts customers with many tech. innovations, e.g., friendly mobile-first user interface
- As of mid-2020, RH had attracted a clientele of over 13 million investors

Available data + open-source algorithms : QuantConnect



- Open-source platform with two-sides :
  - Algorithm-developer members who developed and tested for free
  - Investor-members pay a subscription to get funds managed by the winning algorithms
- Successful developer-members could get a royalty or commission from investor-members

Available data + open-source algorithms : Many more !




Warning

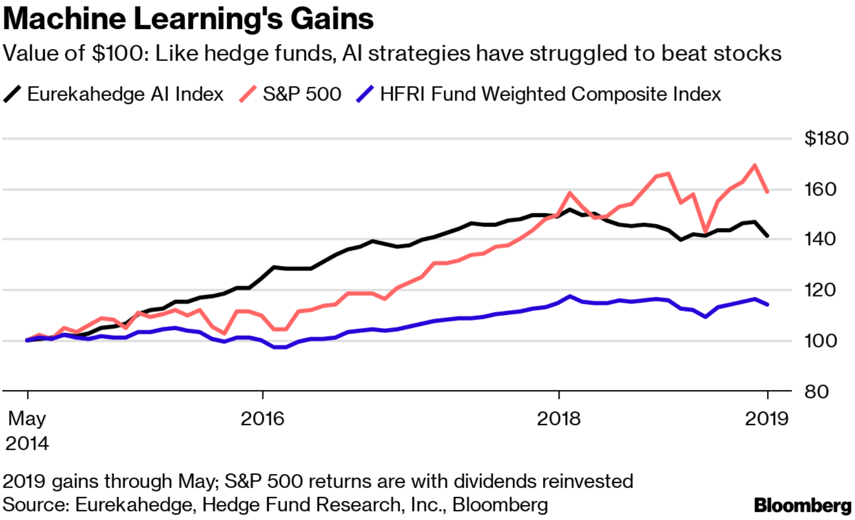
Machine Learning + Alternative Data is **not a Magic Formula**

Don’t believe the hype about AI and fund management

Machine learning can generate marginal improvements but nothing truly transformational

The legendary AI program AlphaGo beat a human champion at Go. Playing the markets is another game entirely. © PA

Machine Learning + Alternative Data is **not a Magic Formula**

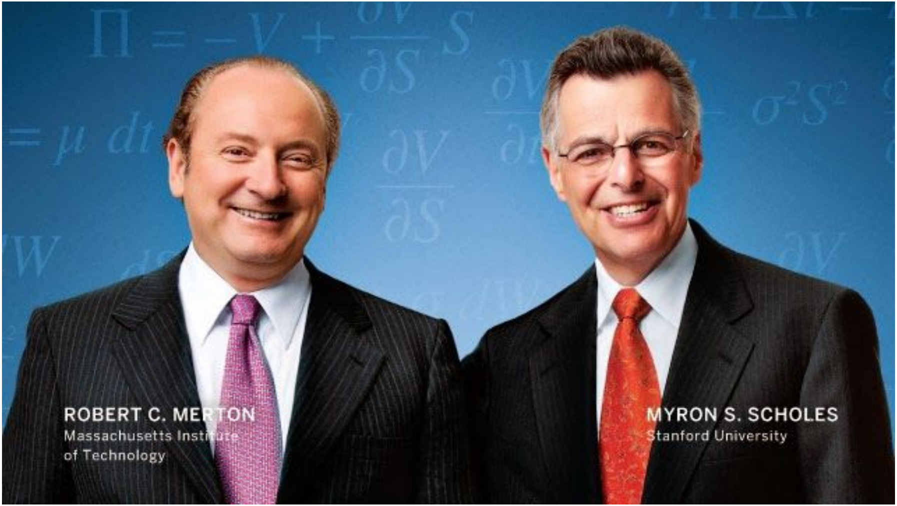


2019 gains through May; S&P 500 returns are with dividends reinvested
Source: Eurekahedge, Hedge Fund Research, Inc., Bloomberg

Long-Term Capital Management: John Meriwether

- Education : Northwestern University + MBA degree from the University of Chicago
- Moved to New York City, where he worked as a bond trader at Salomon Brothers.
- In 1991, he was caught in a Treasury securities trading scandal

Long-Term Capital Management



$\rightarrow$ 1997 Nobel Prize of Economics for the Black-Scholes-Merton formula

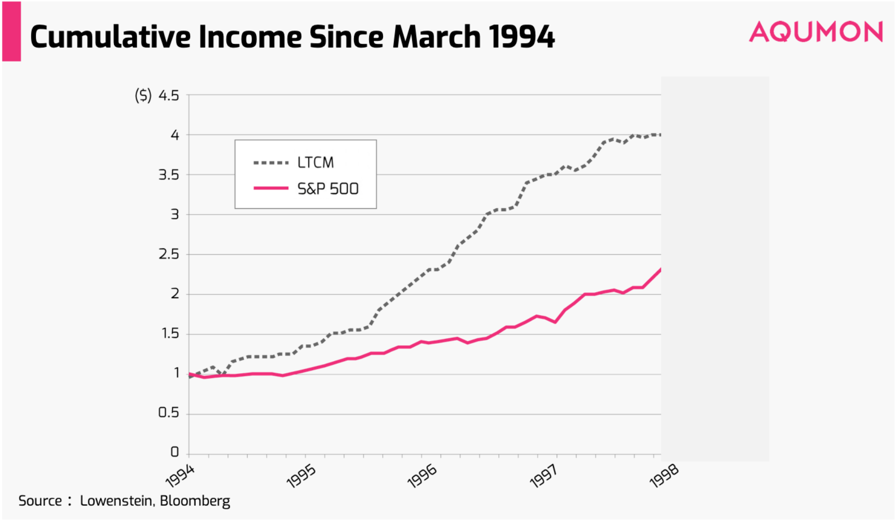
Long-Term Capital Management : $4.7 billion of AUM in 1998

LTCM's Annualized Return on Capital	
YEAR	**RETURN**
1994*	**20%**
1995	**43%**
1996	**41%**
1997	**17%**

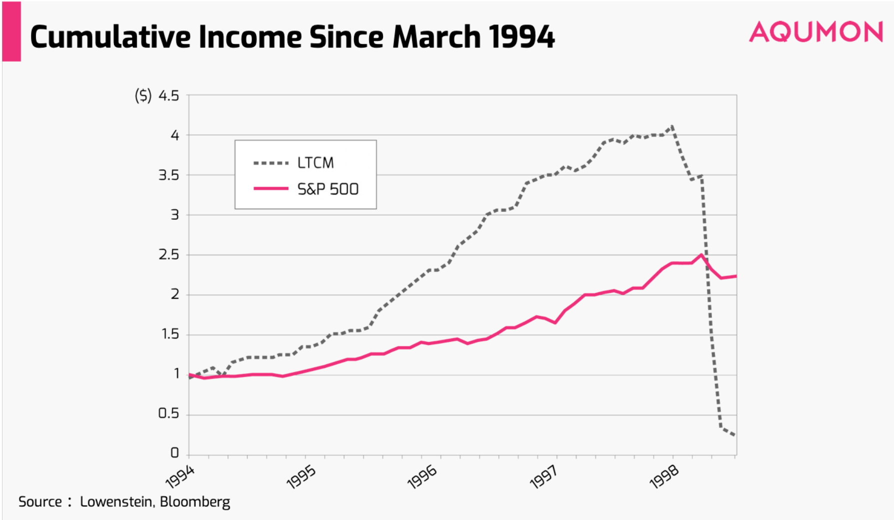
*** last 10 months of 1994**

$\rightarrow$ $1.01 billion in capital in 1994. Leverage : 27/1

Long-Term Capital Management



Long-Term Capital Management : 1998 Russian financial crisis



→ $4.6 billion losses in less than 4 months

Long-Term Capital Management : The Aftermath

- The Federal Reserve Bank of New York organized a bailout of $3.625 billion by the major creditors to avoid a wider collapse in the financial markets
- UBS CEO had to resign

Long-Term Capital Management : The Aftermath

- The Federal Reserve Bank of New York organized a bailout of $3.625 billion by the major creditors to avoid a wider collapse in the financial markets
- UBS CEO had to resign
- R. Merton and M. Scholes

Long-Term Capital Management : The Aftermath

- The Federal Reserve Bank of New York organized a bailout of $3.625 billion by the major creditors to avoid a wider collapse in the financial markets

- UBS CEO had to resign

- R. Merton and M. Scholes won the Nobel Prize. Merton is still Professor at MIT today

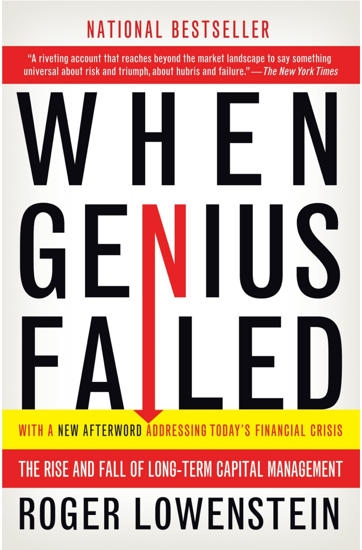
- John Meriwether

Long-Term Capital Management : The Aftermath

- The Federal Reserve Bank of New York organized a bailout of $3.625 billion by the major creditors to avoid a wider collapse in the financial markets
- UBS CEO had to resign
- R. Merton and M. Scholes won the Nobel Prize. Merton is still Professor at MIT today
- John Meriwether started two new hedge funds (JWM Partners and JM Advisors)

Long-Term Capital Management : The Aftermath

- The Federal Reserve Bank of New York organized a bailout of $3.625 billion by the major creditors to avoid a wider collapse in the financial markets
- UBS CEO had to resign
- R. Merton and M. Scholes won the Nobel Prize. Merton is still Professor at MIT today
- John Meriwether started two new hedge funds (JWM Partners and JM Advisors)... and failed twice



What do we learned ?

- Alternative data + machine learning algorithm = huge progress in predicting stock returns
- For you: tons of exciting job opportunities
- Financial markets remain a mystery, investing comes with high risks

What do we learned ?

- Alternative data + machine learning algorithm = huge progress in predicting stock returns
- For you: tons of exciting job opportunities
- Financial markets remain a mystery, investing comes with high risks
- Bigger picture: How is it useful ? Welfare implications ?

What do we learned ?

- Alternative data + machine learning algorithm = huge progress in predicting stock returns
- For you: tons of exciting job opportunities
- Financial markets remain a mystery, investing comes with high risks
- Bigger picture: How is it useful ? Welfare implications ?
  - Responsible investing ? → Next week !

Discussion