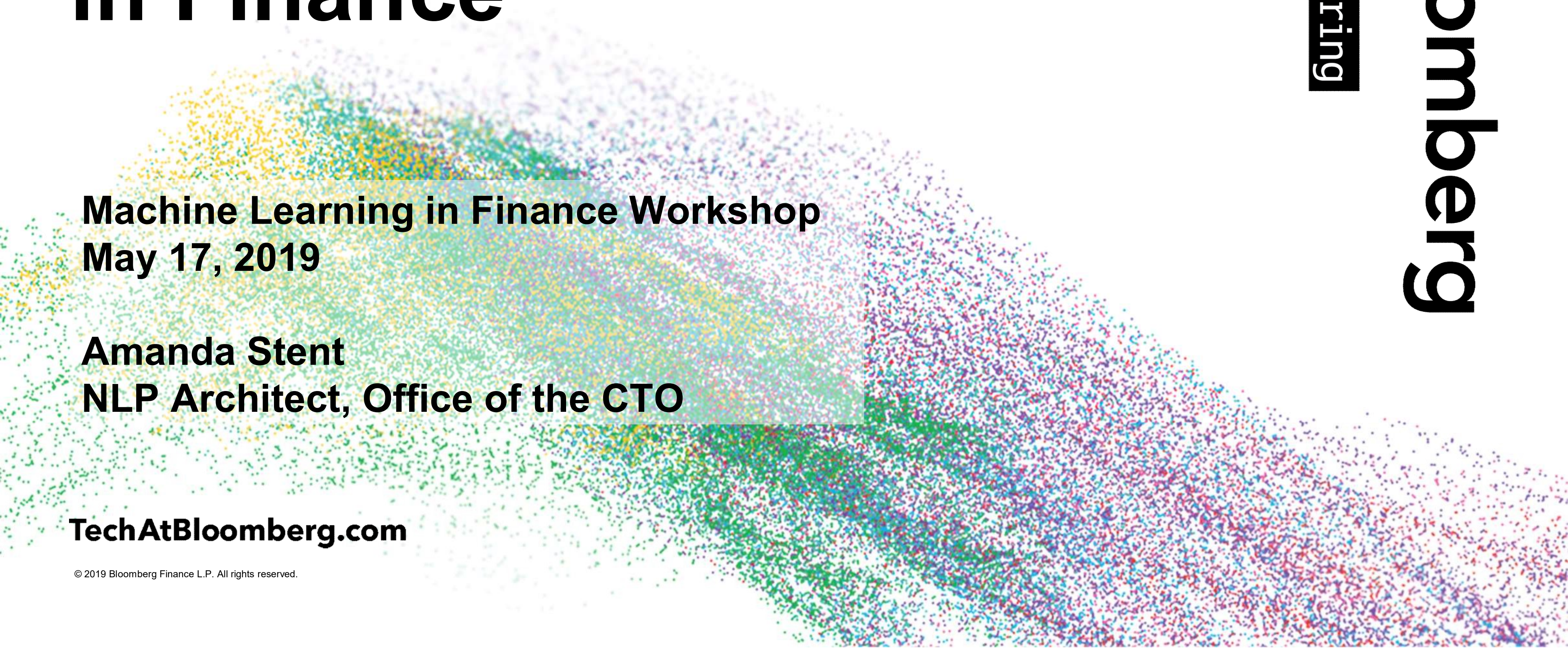
AI for Unstructured Data in Finance



Machine Learning in Finance Workshop
May 17, 2019

Amanda Stent
NLP Architect, Office of the CTO

TechAtBloomberg.com

We Are in an "AI Revolution"

BUSINESS | JOURNAL REPORTS: LEADERSHIP

The Optimistic Promise of Artificial Intelligence

Andrew Ng and Tong Zhang on how AI is going to be like electricity, transforming every industry

THE AI REVOLUTION IS ON

Microsoft researchers achieve speech recognition milestone

China's AI Agenda Advances

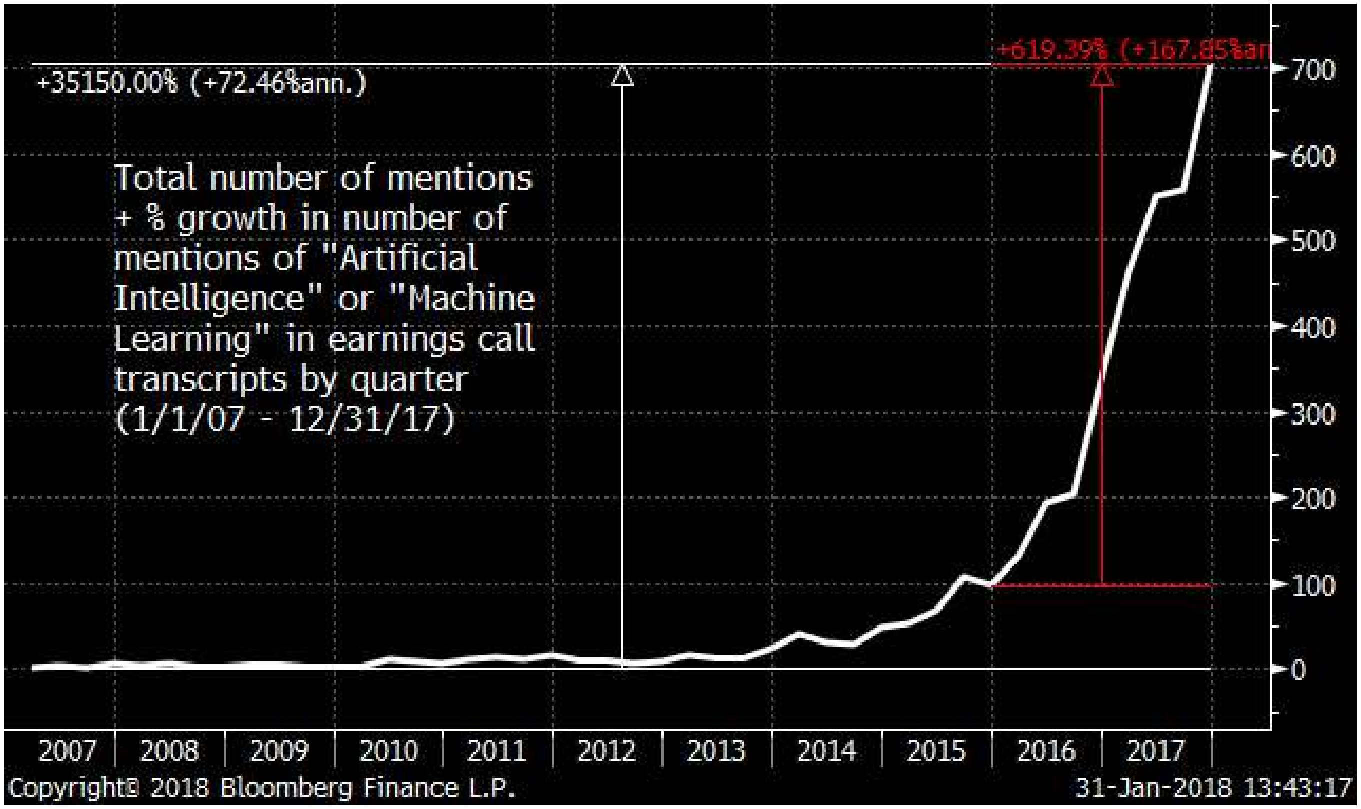
Google Translate Has Reached Human-Like Accuracy Thanks To Neural Machine Translation Engine

Alibaba's AI Outguns Humans in Reading Test

The Sublime and Scary Future of Cameras With A.I. Brains

Stanford team creates computer vision algorithm that can describe photos

The “AI Revolution” is Changing the World



Total number of mentions + % growth in number of mentions of "Artificial Intelligence" or "Machine Learning" in earnings call transcripts by quarter (1/1/07 - 12/31/17)

+35150.00% (+72.46%ann.)

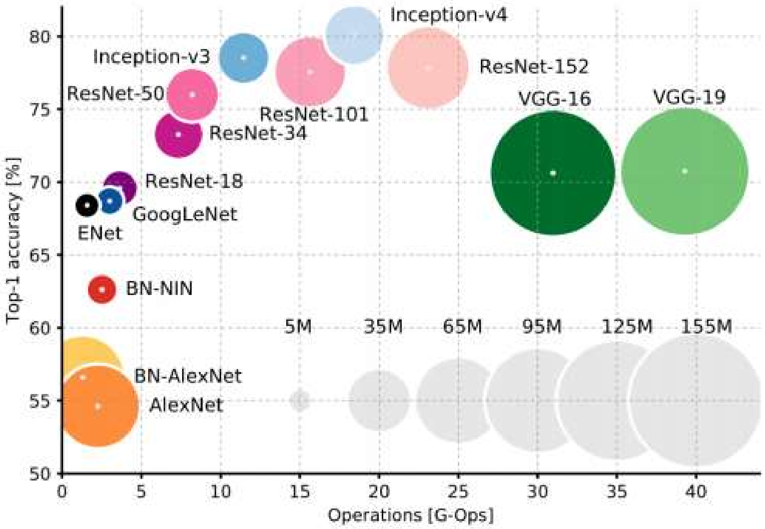
+619.39% (+167.85%an

Copyright© 2018 Bloomberg Finance L.P.

31-Jan-2018 13:43:17

The "AI Revolution" is Data Driven

- Deep-learning based AI depends on:
  — Lots of specialized (tailored, expensive) compute
  — Lots of labeled data (supplied by you and me!)

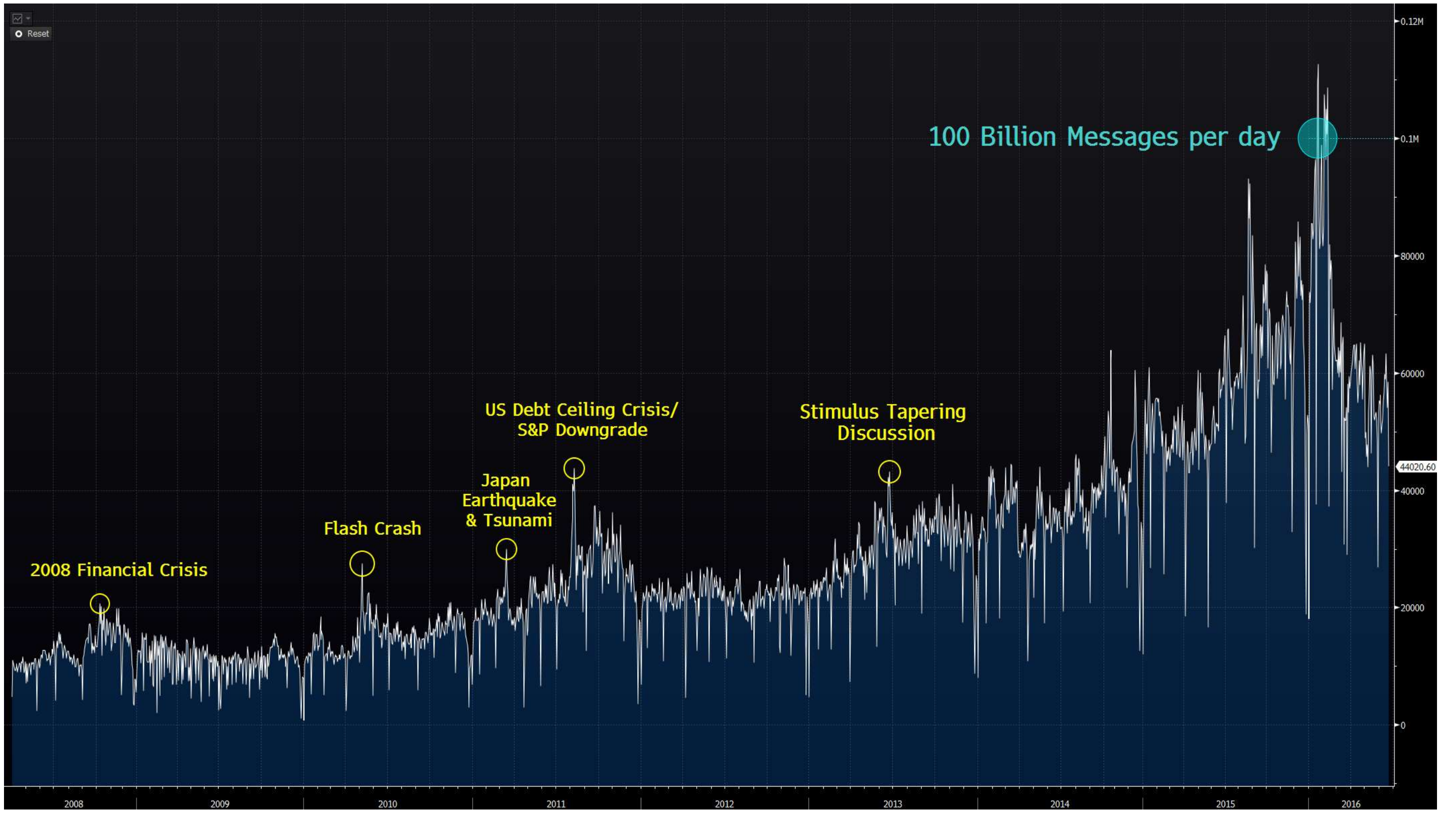


Source: Michael Felsberg

TechAtBloomberg.com

Bloomberg
Engineering

Bloomberg is the Home of Market Data

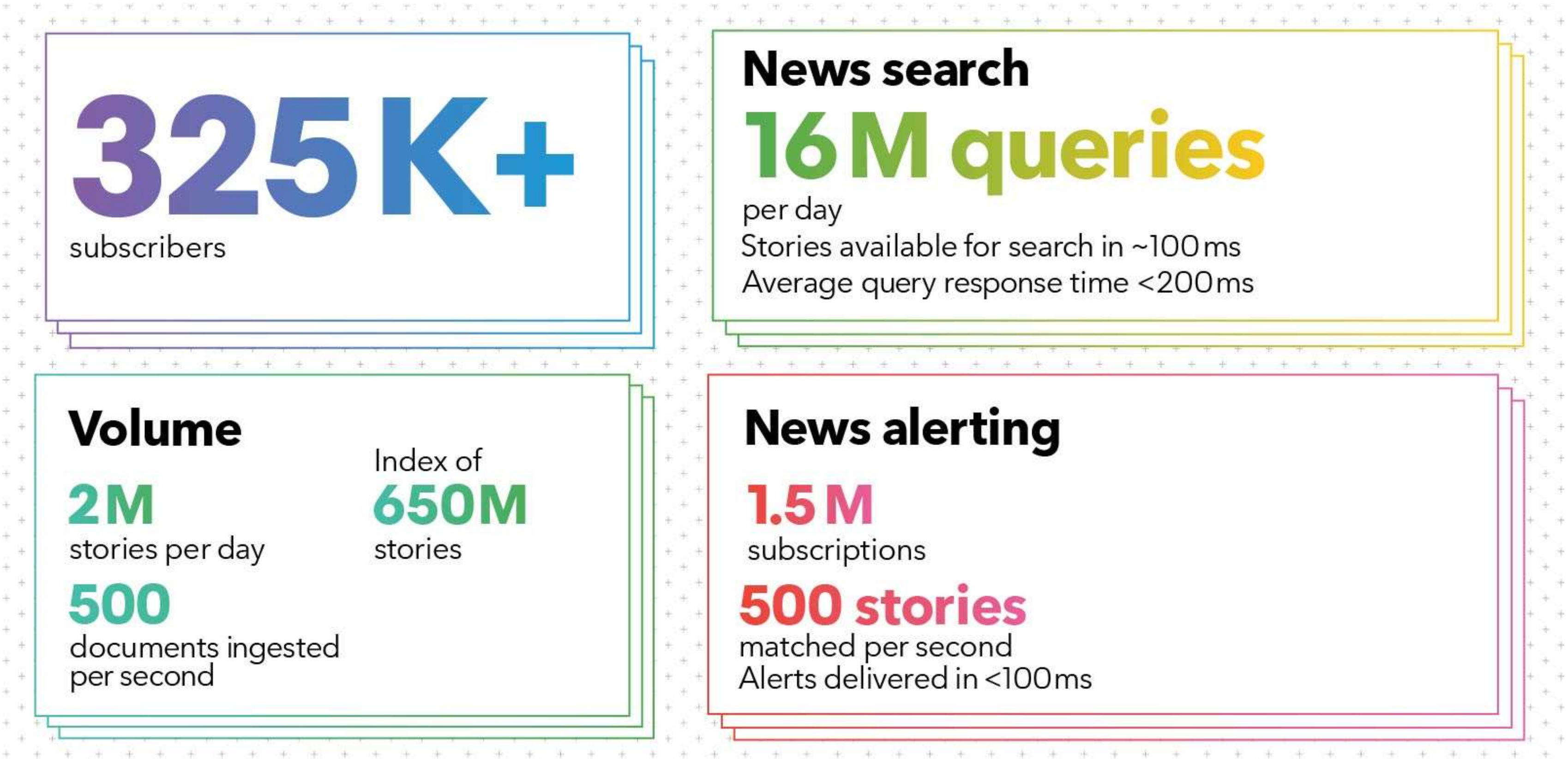


TechAtBloomberg.com

Bloomberg

Engineering

Bloomberg Is the Home of Finance News



325K+
subscribers

News search

16M queries
per day
Stories available for search in ~100ms
Average query response time <200ms

Volume

2M
stories per day

Index of
650M
stories

500
documents ingested per second

News alerting

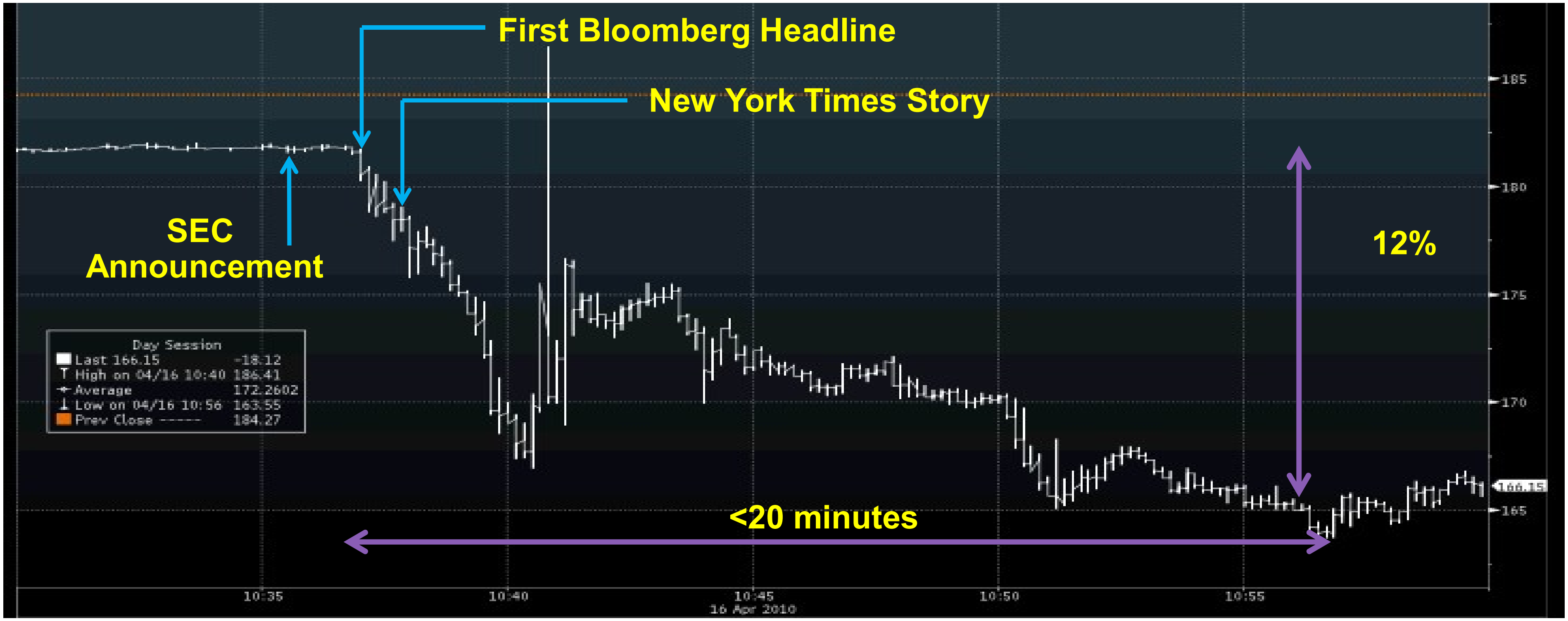
1.5M
subscriptions

500 stories
matched per second
Alerts delivered in <100ms

TechAtBloomberg.com

Bloomberg
Engineering

News Moves Markets (Fast)



First Bloomberg Headline

New York Times Story

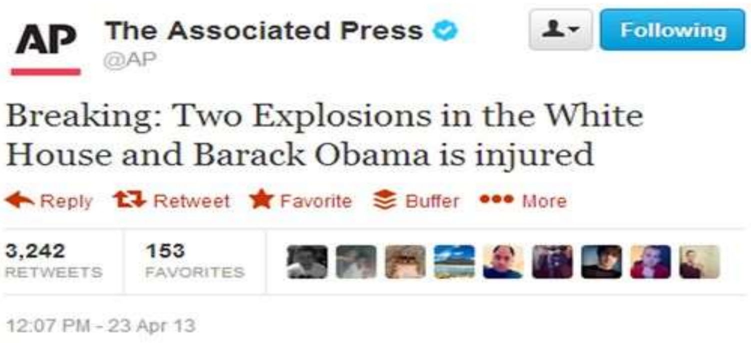
SEC Announcement

12%

<20 minutes

TechAtBloomberg.com

Social Media Moves Markets



AP The Associated Press @AP

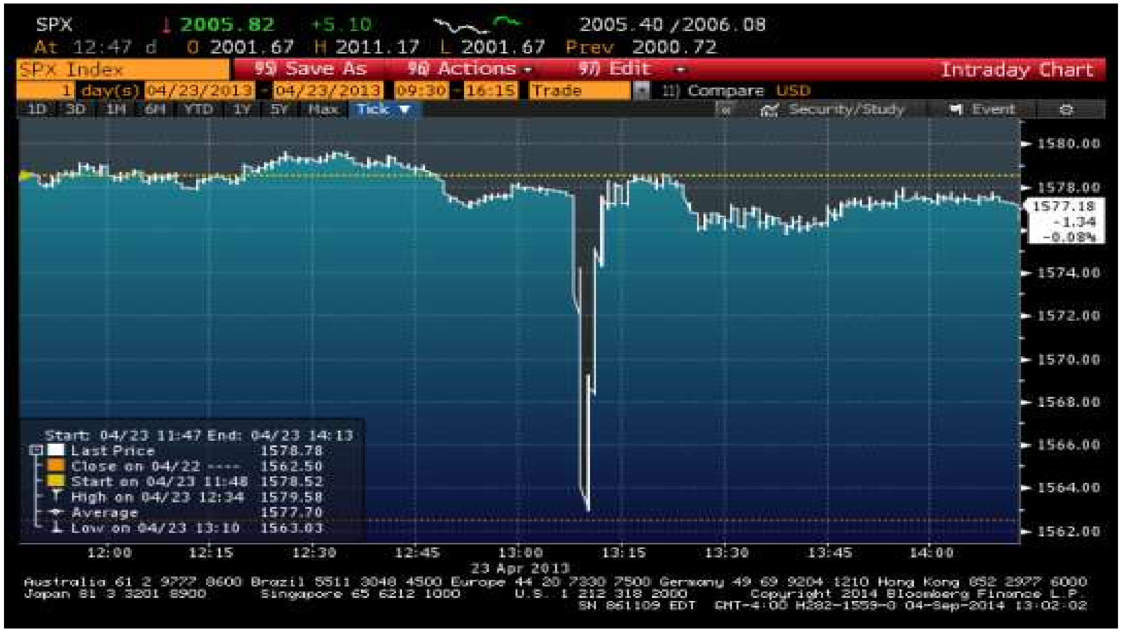
Breaking: Two Explosions in the White House and Barack Obama is injured

3,242 RETWEETS

153 FAVORITES

12:07 PM - 23 Apr 13

Over $136 billion was wiped out in minutes



SPX 2005.82 +5.10 2005.40 /2006.08
At 12:47 d O 2001.67 H 2011.17 L 2001.67 Prev 2000.72

SPX Index

1 day(s) 04/23/2013 - 04/23/2013 09:30 - 16:15 Trade

Compare USD

Start: 04/23 11:47 End: 04/23 14:13

Last Price	1578.78
Close on 04/22 ----	1562.50
Start on 04/23 11:48	1578.52
High on 04/23 12:34	1579.58
Average	1577.70
Low on 04/23 13:10	1563.03

Australia 61 2 9777 8600 Brazil 5511 3048 4500 Europe 44 20 7330 7500 Germany 49 69 9204 1210 Hong Kong 852 2977 6000
Japan 81 3 3201 8900 Singapore 65 6212 1000 U.S. 1 212 318 2000 Copyright 2014 Bloomberg Finance L.P.
SN 861109 EDT GMT-4:00 H282-1559-0 04-Sep-2014 13:02:02

Our Key Challenges

- The need for speed
- The need for precision
- The need for human in the loop

Text Analytics vs. NLP

Natural Language Processing (computational linguistics)

“the scientific and engineering discipline concerned with understanding written and spoken language from a computational perspective, and building artifacts that usefully process and produce language” (*Stanford Encyclopedia of Philosophy*)

Text Analytics

“the process of deriving information from text sources” (*Gartner IT Glossary*)

Why Do Text Analytics?

Most companies that use text analytics have one goal

- Assess customer feedback via voice, SMS, email… (e.g., to assess NPS)
- NLP tasks: entity recognition, sentiment analysis, key terms, visualization

A few have the goal of ranking/serving the text

- Treat text as a “first class object” (i.e., text is both input and output)
- NLP tasks: entity recognition and linking, topic detection, sentiment, clustering

A few have a different kind of goal

- Assess market signals (financial reports, news articles, tweets…) in order to predict some (financial) outcome (e.g., movie grosses, stock prices)
- NLP tasks: entity recognition and linking, sentiment analysis, relation/event extraction, signal extraction, classification

How Text Analytics Is Done

Generation 0: write a bunch of rules (“templates”, “grammars”)

- High-precision
- Slow, manual, difficult to maintain or update

Generation 1: train a statistical classifier

- For sequence tagging: conditional random fields; for document classification: linear/logistic regression, SVMs, decision trees/random forests
- Need labeled data

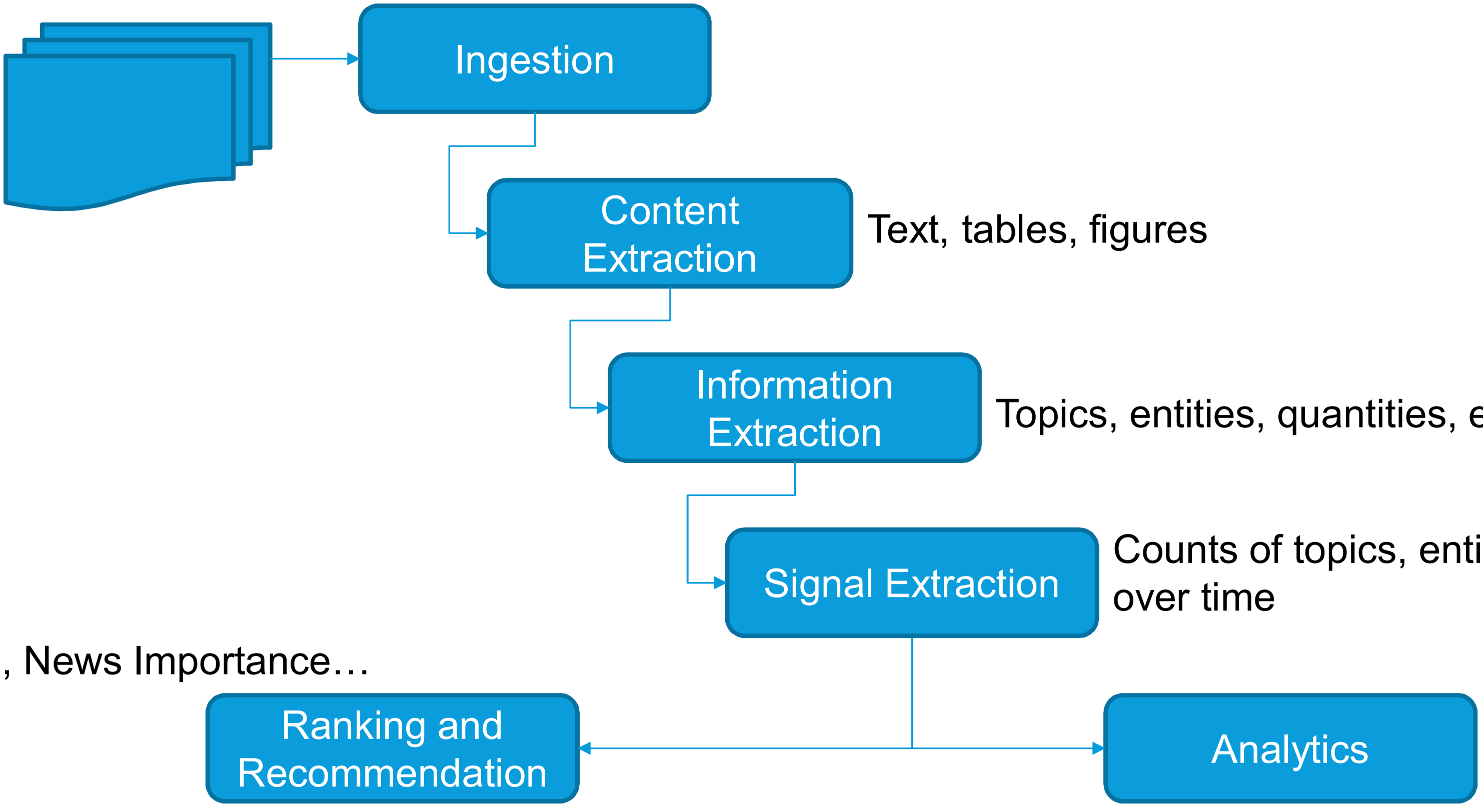
Generation 2: deep learning and human in the loop

- Need a lot of labeled data, or distant supervision
- May be slower

We still need separate “models” for each language/genre pair

English news, English social, English generic; Chinese news, Chinese social…

The Life of a Document



Ingestion

Content Extraction

Text, tables, figures

Information Extraction

Topics, entities, quantities, events, sentiment

Signal Extraction

Counts of topics, entities, events… over time

Key Insights, News Importance…

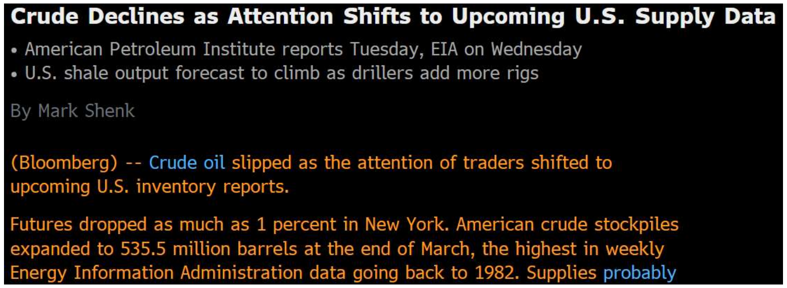
Ranking and Recommendation

Analytics

News Trends, Market Moving News…

Topics are Central to AI for Finance

- Human curated in-house taxonomy of topic codes
- Exposed as an important metadata of text documents



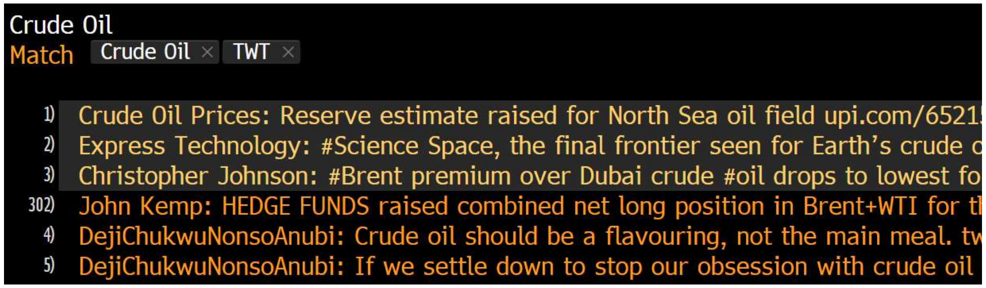
Crude Declines as Attention Shifts to Upcoming U.S. Supply Data

- American Petroleum Institute reports Tuesday, EIA on Wednesday
- U.S. shale output forecast to climb as drillers add more rigs

By Mark Shenk

(Bloomberg) -- Crude oil slipped as the attention of traders shifted to upcoming U.S. inventory reports.

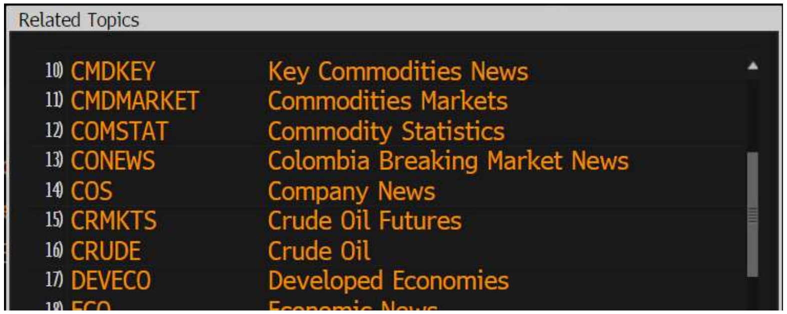
Futures dropped as much as 1 percent in New York. American crude stockpiles expanded to 535.5 million barrels at the end of March, the highest in weekly Energy Information Administration data going back to 1982. Supplies probably



Crude Oil

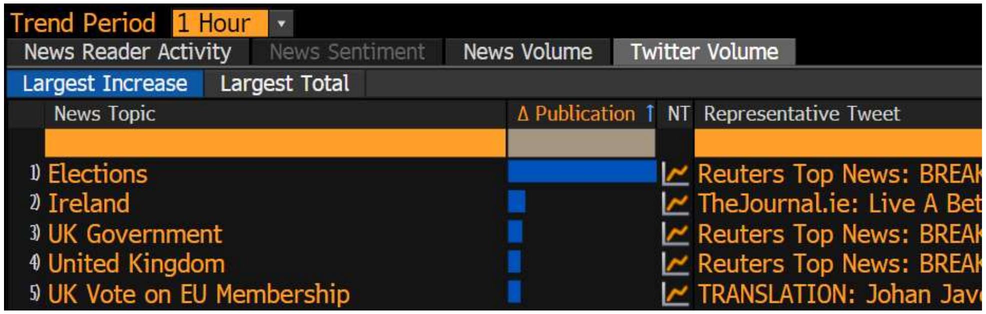
Match Crude Oil × TWT ×

1) Crude Oil Prices: Reserve estimate raised for North Sea oil field upi.com/6521
2) Express Technology: #Science Space, the final frontier seen for Earth's crude o
3) Christopher Johnson: #Brent premium over Dubai crude #oil drops to lowest fo
302) John Kemp: HEDGE FUNDS raised combined net long position in Brent+WTI for t
4) DejiChukwuNonsoAnubi: Crude oil should be a flavouring, not the main meal. tw
5) DejiChukwuNonsoAnubi: If we settle down to stop our obsession with crude oil



Related Topics

10) CMDKEY	Key Commodities News
11) CMDMARKET	Commodities Markets
12) COMSTAT	Commodity Statistics
13) CONEWS	Colombia Breaking Market News
14) COS	Company News
15) CRMKTS	Crude Oil Futures
16) CRUDE	Crude Oil
17) DEVECO	Developed Economies
18) ECO	Economic News



Trend Period 1 Hour ▾

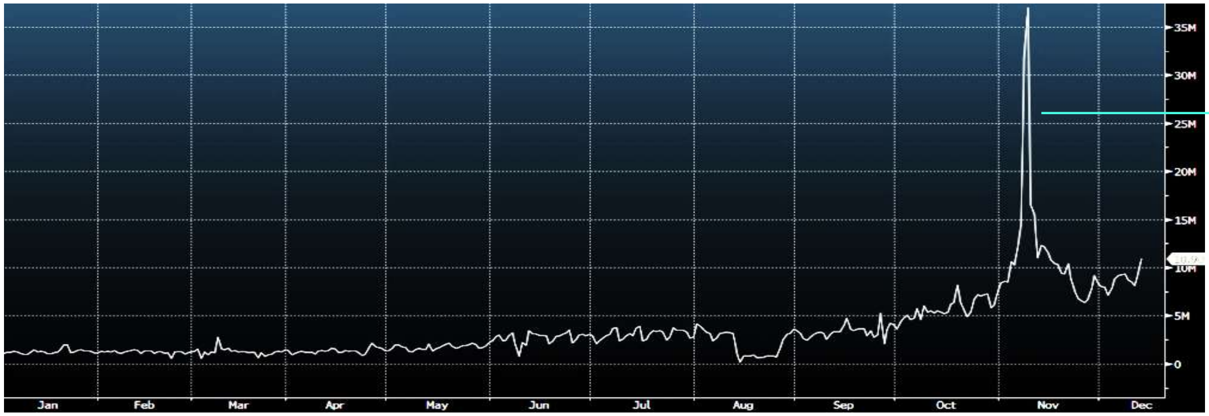
News Topic	Δ Publication ↑	NT	Representative Tweet
1) Elections			Reuters Top News: BREAK
2) Ireland			TheJournal.ie: Live A Bet
3) UK Government			Reuters Top News: BREAK
4) United Kingdom			Reuters Top News: BREAK
5) UK Vote on EU Membership			TRANSLATION: Johan Jav

TechAtBloomberg.com

Bloomberg
Engineering

Topics & Entities Convert Unstructured Content to Structured Data

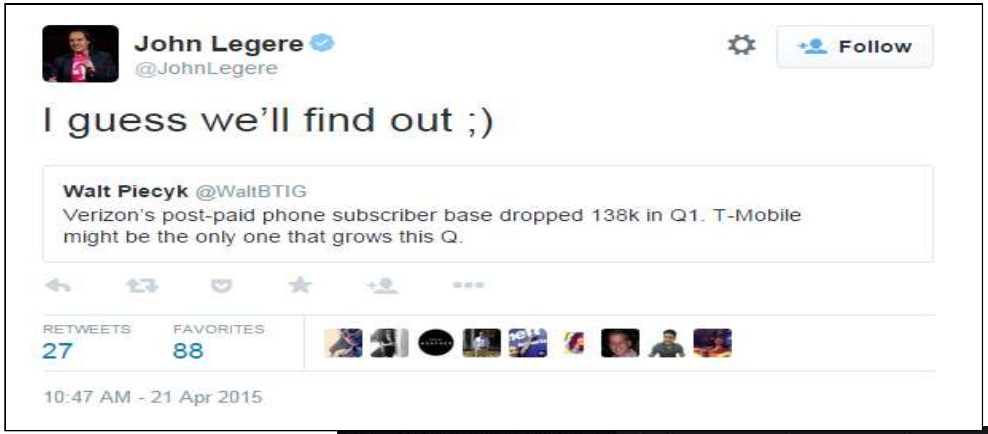
- Over 1B tweets processed in the past 12 months
- Over 700 topic models deployed (still growing)
- Over 1M tweets labelled in-house for model training
- Peak-rate capacity of the real-time classification system



2016 U.S. Presidential Election

- **~100M** tweets in 4 days
- peak rate at **4K+** tweets/sec

Topics & Entities Link Unstructured Content to the World



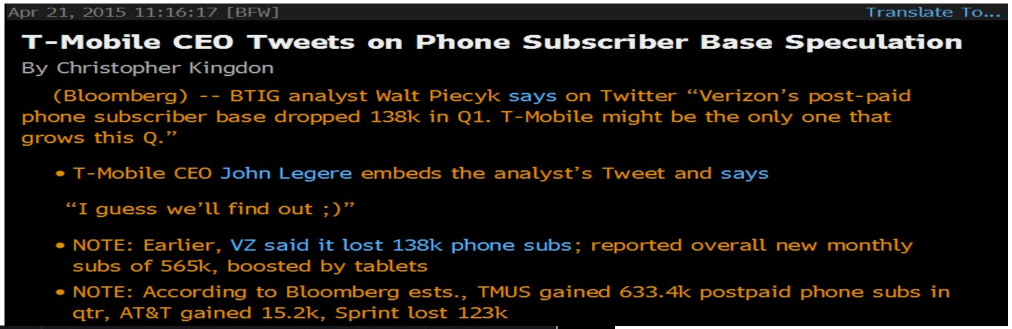
John Legere @JohnLegere

I guess we'll find out ;)

> **Walt Piecyk** @WaltBTIG
> Verizon's post-paid phone subscriber base dropped 138k in Q1. T-Mobile might be the only one that grows this Q.

RETWEETS 27 FAVORITES 88

10:47 AM - 21 Apr 2015



Apr 21, 2015 11:16:17 [BFW]

T-Mobile CEO Tweets on Phone Subscriber Base Speculation

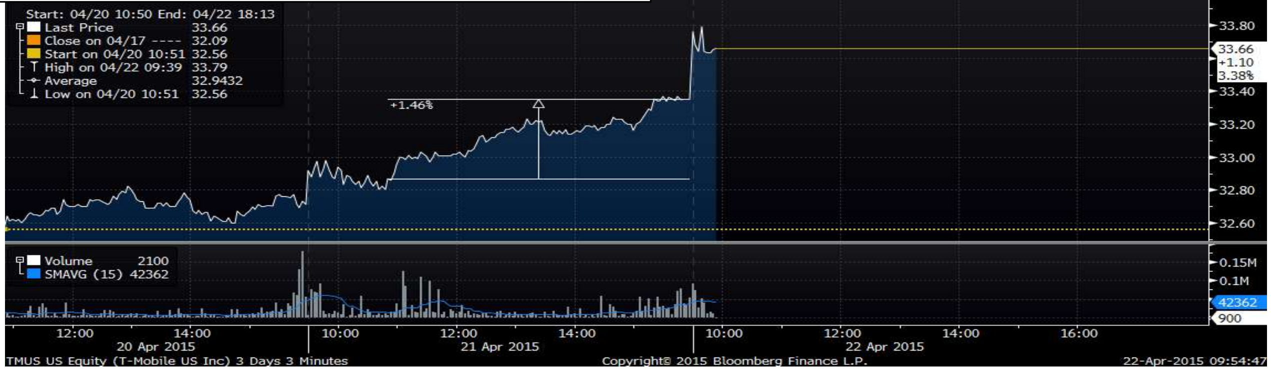
By Christopher Kingdon

(Bloomberg) -- BTIG analyst Walt Piecyk says on Twitter "Verizon's post-paid phone subscriber base dropped 138k in Q1. T-Mobile might be the only one that grows this Q."

- T-Mobile CEO John Legere embeds the analyst's Tweet and says

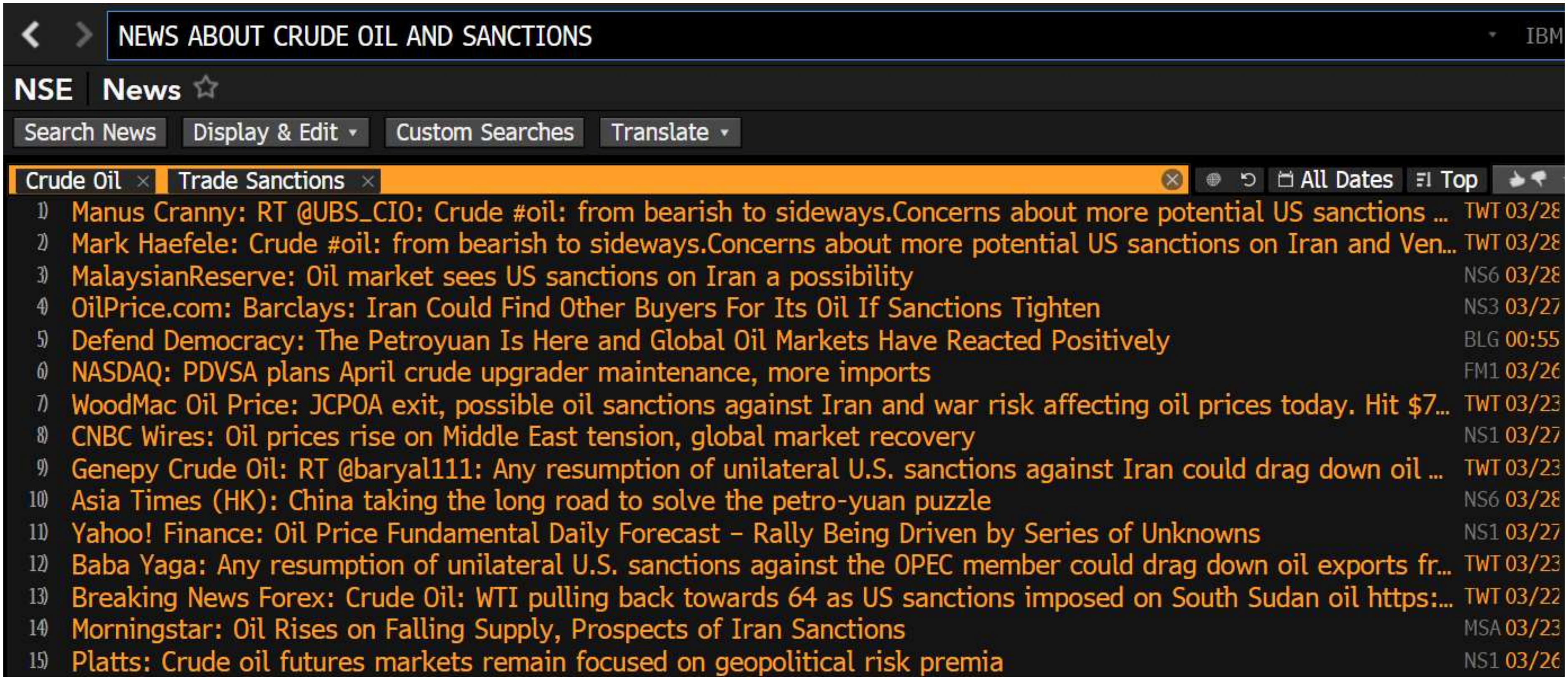
  "I guess we'll find out ;)"

- NOTE: Earlier, VZ said it lost 138k phone subs; reported overall new monthly subs of 565k, boosted by tablets
- NOTE: According to Bloomberg ests., TMUS gained 633.4k postpaid phone subs in qtr, AT&T gained 15.2k, Sprint lost 123k



TechAtBloomberg.com

Topics + Entities + Interaction = Insight



NEWS ABOUT CRUDE OIL AND SANCTIONS

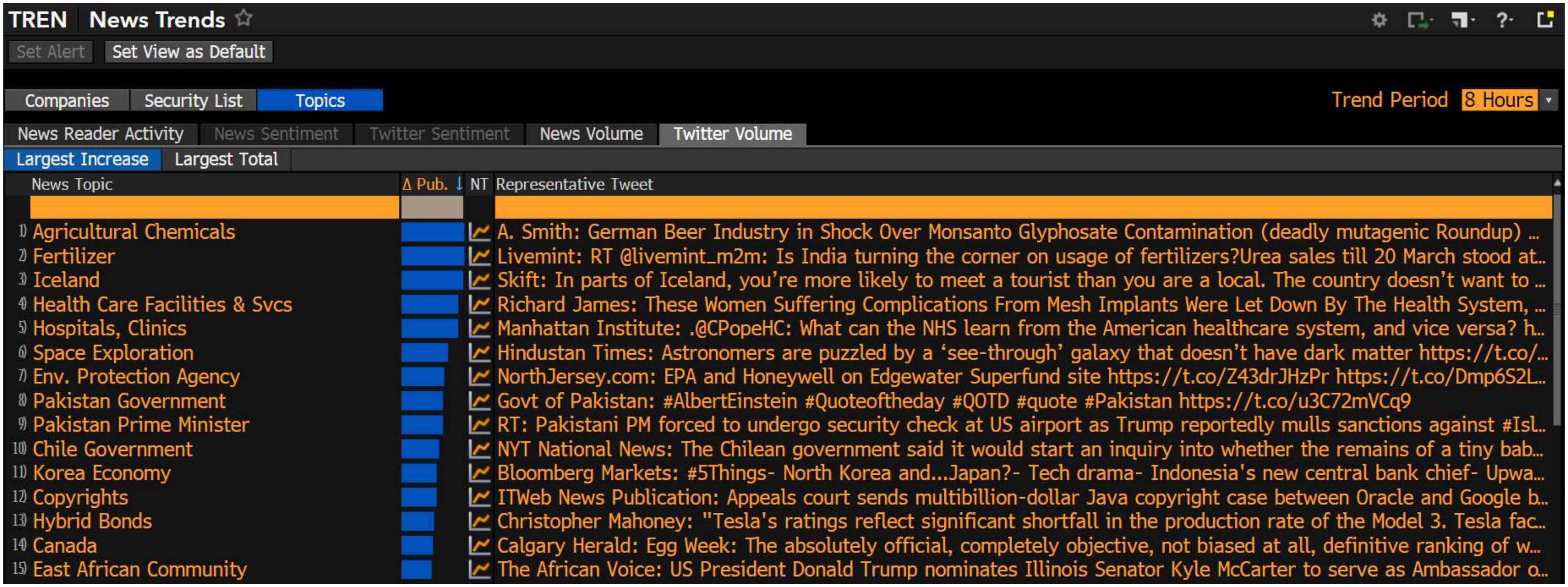
Crude Oil × Trade Sanctions ×

1) Manus Cranny: RT @UBS_CIO: Crude #oil: from bearish to sideways.Concerns about more potential US sanctions ... TWT 03/28
2) Mark Haefele: Crude #oil: from bearish to sideways.Concerns about more potential US sanctions on Iran and Ven... TWT 03/28
3) MalaysianReserve: Oil market sees US sanctions on Iran a possibility NS6 03/28
4) OilPrice.com: Barclays: Iran Could Find Other Buyers For Its Oil If Sanctions Tighten NS3 03/27
5) Defend Democracy: The Petroyuan Is Here and Global Oil Markets Have Reacted Positively BLG 00:55
6) NASDAQ: PDVSA plans April crude upgrader maintenance, more imports FM1 03/26
7) WoodMac Oil Price: JCPOA exit, possible oil sanctions against Iran and war risk affecting oil prices today. Hit $7... TWT 03/23
8) CNBC Wires: Oil prices rise on Middle East tension, global market recovery NS1 03/27
9) Genepy Crude Oil: RT @baryal111: Any resumption of unilateral U.S. sanctions against Iran could drag down oil ... TWT 03/23
10) Asia Times (HK): China taking the long road to solve the petro-yuan puzzle NS6 03/28
11) Yahoo! Finance: Oil Price Fundamental Daily Forecast – Rally Being Driven by Series of Unknowns NS1 03/27
12) Baba Yaga: Any resumption of unilateral U.S. sanctions against the OPEC member could drag down oil exports fr... TWT 03/23
13) Breaking News Forex: Crude Oil: WTI pulling back towards 64 as US sanctions imposed on South Sudan oil https:... TWT 03/22
14) Morningstar: Oil Rises on Falling Supply, Prospects of Iran Sanctions MSA 03/23
15) Platts: Crude oil futures markets remain focused on geopolitical risk premia NS1 03/26

TechAtBloomberg.com

Bloomberg
Engineering

(Topics + Entities + Market Fundamentals) / Time = Insight



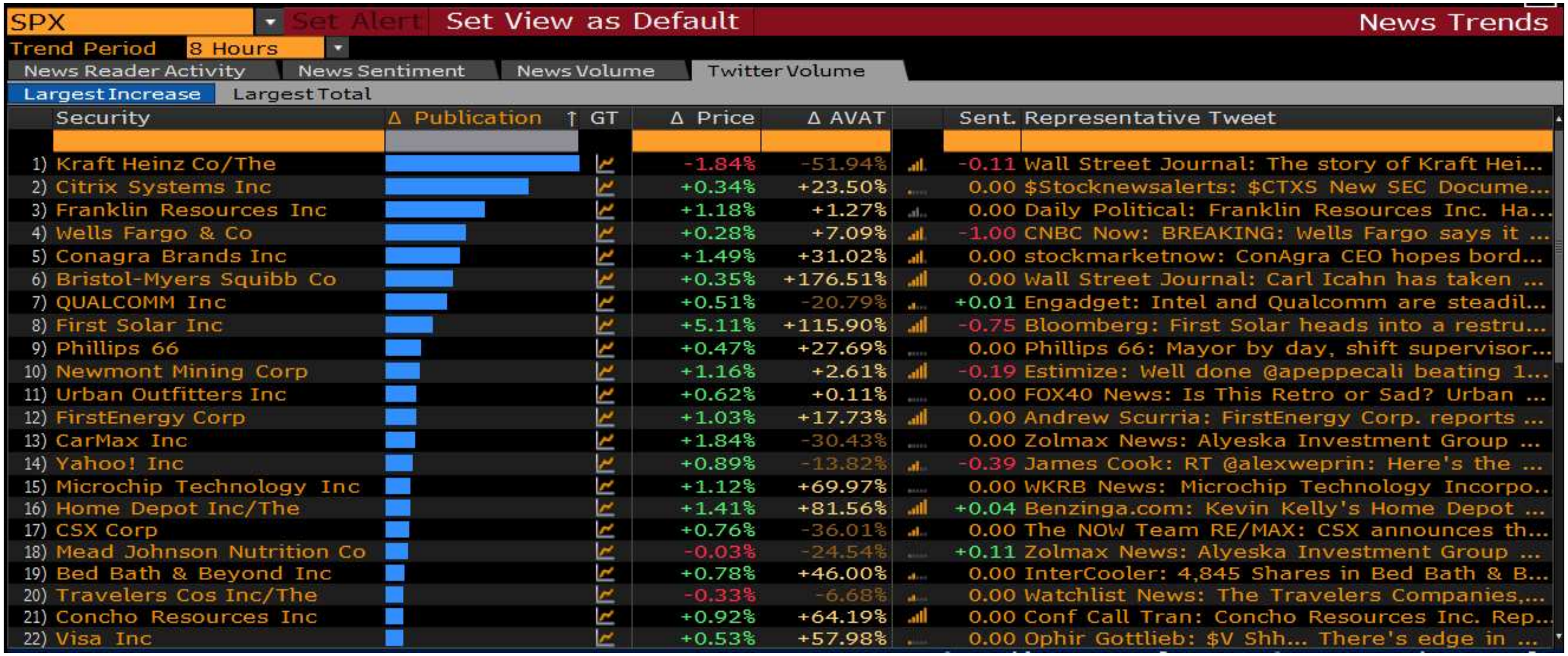
	News Topic	Δ Pub. ↓	NT	Representative Tweet
1)	Agricultural Chemicals			A. Smith: German Beer Industry in Shock Over Monsanto Glyphosate Contamination (deadly mutagenic Roundup) ...
2)	Fertilizer			Livemint: RT @livemint_m2m: Is India turning the corner on usage of fertilizers?Urea sales till 20 March stood at...
3)	Iceland			Skift: In parts of Iceland, you’re more likely to meet a tourist than you are a local. The country doesn’t want to ...
4)	Health Care Facilities & Svcs			Richard James: These Women Suffering Complications From Mesh Implants Were Let Down By The Health System, ...
5)	Hospitals, Clinics			Manhattan Institute: .@CPopeHC: What can the NHS learn from the American healthcare system, and vice versa? h...
6)	Space Exploration			Hindustan Times: Astronomers are puzzled by a ‘see-through’ galaxy that doesn’t have dark matter https://t.co/...
7)	Env. Protection Agency			NorthJersey.com: EPA and Honeywell on Edgewater Superfund site https://t.co/Z43drJHzPr https://t.co/Dmp6S2L...
8)	Pakistan Government			Govt of Pakistan: #AlbertEinstein #Quoteoftheday #QOTD #quote #Pakistan https://t.co/u3C72mVCq9
9)	Pakistan Prime Minister			RT: Pakistani PM forced to undergo security check at US airport as Trump reportedly mulls sanctions against #Isl...
10)	Chile Government			NYT National News: The Chilean government said it would start an inquiry into whether the remains of a tiny bab...
11)	Korea Economy			Bloomberg Markets: #5Things- North Korea and...Japan?- Tech drama- Indonesia's new central bank chief- Upwa...
12)	Copyrights			ITWeb News Publication: Appeals court sends multibillion-dollar Java copyright case between Oracle and Google b...
13)	Hybrid Bonds			Christopher Mahoney: "Tesla's ratings reflect significant shortfall in the production rate of the Model 3. Tesla fac...
14)	Canada			Calgary Herald: Egg Week: The absolutely official, completely objective, not biased at all, definitive ranking of w...
15)	East African Community			The African Voice: US President Donald Trump nominates Illinois Senator Kyle McCarter to serve as Ambassador o...

TechAtBloomberg.com

Bloomberg

Engineering

(Extracted Signals + Market Fundamentals) / Time = Insight



SPX ▾ Set Alert Set View as Default News Trends

Trend Period 8 Hours ▾

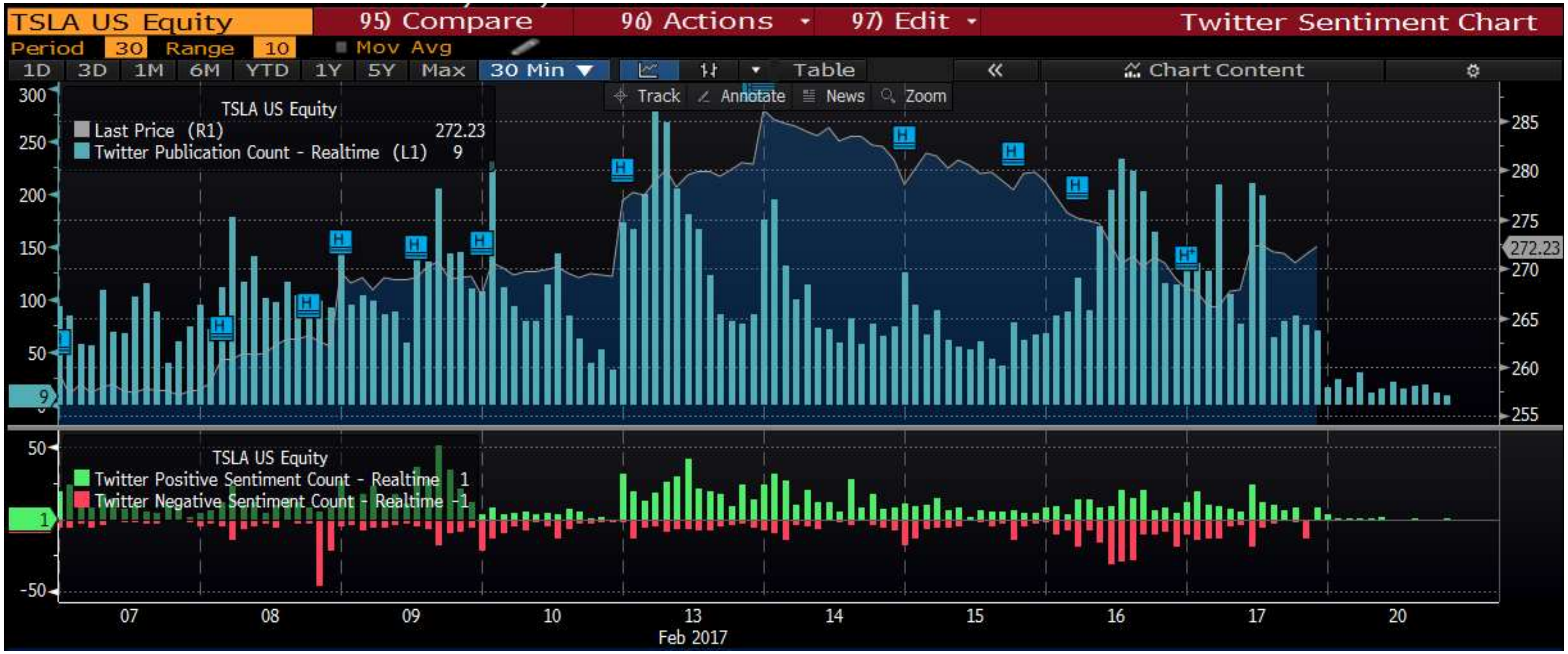
News Reader Activity | News Sentiment | News Volume | Twitter Volume

Largest Increase | Largest Total

	Security	Δ Publication ↑	GT	Δ Price	Δ AVAT	Sent.	Representative Tweet
1)	Kraft Heinz Co/The			-1.84%	-51.94%	-0.11	Wall Street Journal: The story of Kraft Hei...
2)	Citrix Systems Inc			+0.34%	+23.50%	0.00	$Stocknewsalerts: $CTXS New SEC Docume...
3)	Franklin Resources Inc			+1.18%	+1.27%	0.00	Daily Political: Franklin Resources Inc. Ha...
4)	Wells Fargo & Co			+0.28%	+7.09%	-1.00	CNBC Now: BREAKING: Wells Fargo says it ...
5)	Conagra Brands Inc			+1.49%	+31.02%	0.00	stockmarketnow: ConAgra CEO hopes bord...
6)	Bristol-Myers Squibb Co			+0.35%	+176.51%	0.00	Wall Street Journal: Carl Icahn has taken ...
7)	QUALCOMM Inc			+0.51%	-20.79%	+0.01	Engadget: Intel and Qualcomm are steadil...
8)	First Solar Inc			+5.11%	+115.90%	-0.75	Bloomberg: First Solar heads into a restru...
9)	Phillips 66			+0.47%	+27.69%	0.00	Phillips 66: Mayor by day, shift supervisor...
10)	Newmont Mining Corp			+1.16%	+2.61%	-0.19	Estimize: Well done @apeppecali beating 1...
11)	Urban Outfitters Inc			+0.62%	+0.11%	0.00	FOX40 News: Is This Retro or Sad? Urban ...
12)	FirstEnergy Corp			+1.03%	+17.73%	0.00	Andrew Scurria: FirstEnergy Corp. reports ...
13)	CarMax Inc			+1.84%	-30.43%	0.00	Zolmax News: Alyeska Investment Group ...
14)	Yahoo! Inc			+0.89%	-13.82%	-0.39	James Cook: RT @alexweprin: Here's the ...
15)	Microchip Technology Inc			+1.12%	+69.97%	0.00	WKRB News: Microchip Technology Incorpo..
16)	Home Depot Inc/The			+1.41%	+81.56%	+0.04	Benzinga.com: Kevin Kelly's Home Depot ...
17)	CSX Corp			+0.76%	-36.01%	0.00	The NOW Team RE/MAX: CSX announces th...
18)	Mead Johnson Nutrition Co			-0.03%	-24.54%	+0.11	Zolmax News: Alyeska Investment Group ...
19)	Bed Bath & Beyond Inc			+0.78%	+46.00%	0.00	InterCooler: 4,845 Shares in Bed Bath & B...
20)	Travelers Cos Inc/The			-0.33%	-6.68%	0.00	Watchlist News: The Travelers Companies,...
21)	Concho Resources Inc			+0.92%	+64.19%	0.00	Conf Call Tran: Concho Resources Inc. Rep...
22)	Visa Inc			+0.53%	+57.98%	0.00	Ophir Gottlieb: $V Shh... There's edge in ...

TechAtBloomberg.com

(Extracted Signals + Market Fundamentals) / Time = Insight



TSLA US Equity | 95) Compare | 96) Actions | 97) Edit | Twitter Sentiment Chart

Period 30 Range 10 ■ Mov Avg

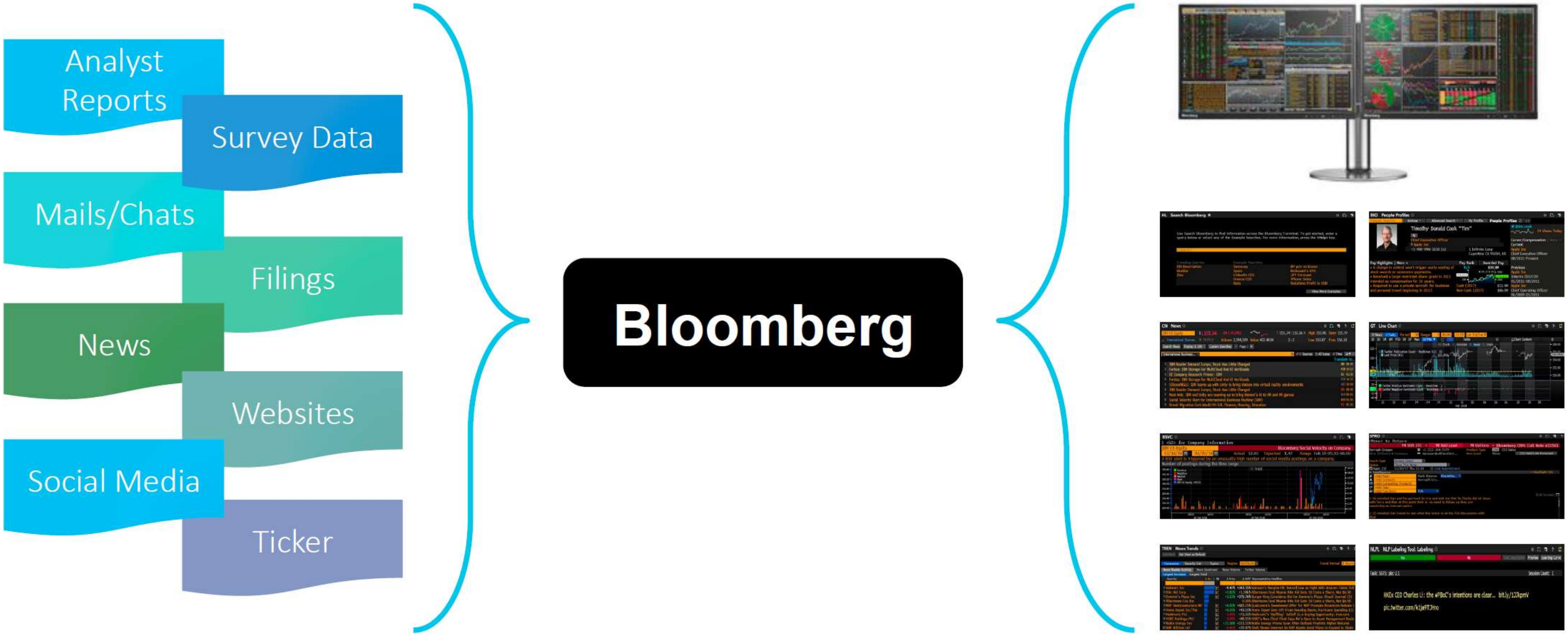
TSLA US Equity
- Last Price (R1) 272.23
- Twitter Publication Count - Realtime (L1) 9

TSLA US Equity
- Twitter Positive Sentiment Count - Realtime 1
- Twitter Negative Sentiment Count - Realtime -1

TechAtBloomberg.com

Bloomberg
Engineering

Bloomberg in a Nutshell



Analyst Reports

Survey Data

Mails/Chats

Filings

News

Websites

Social Media

Ticker

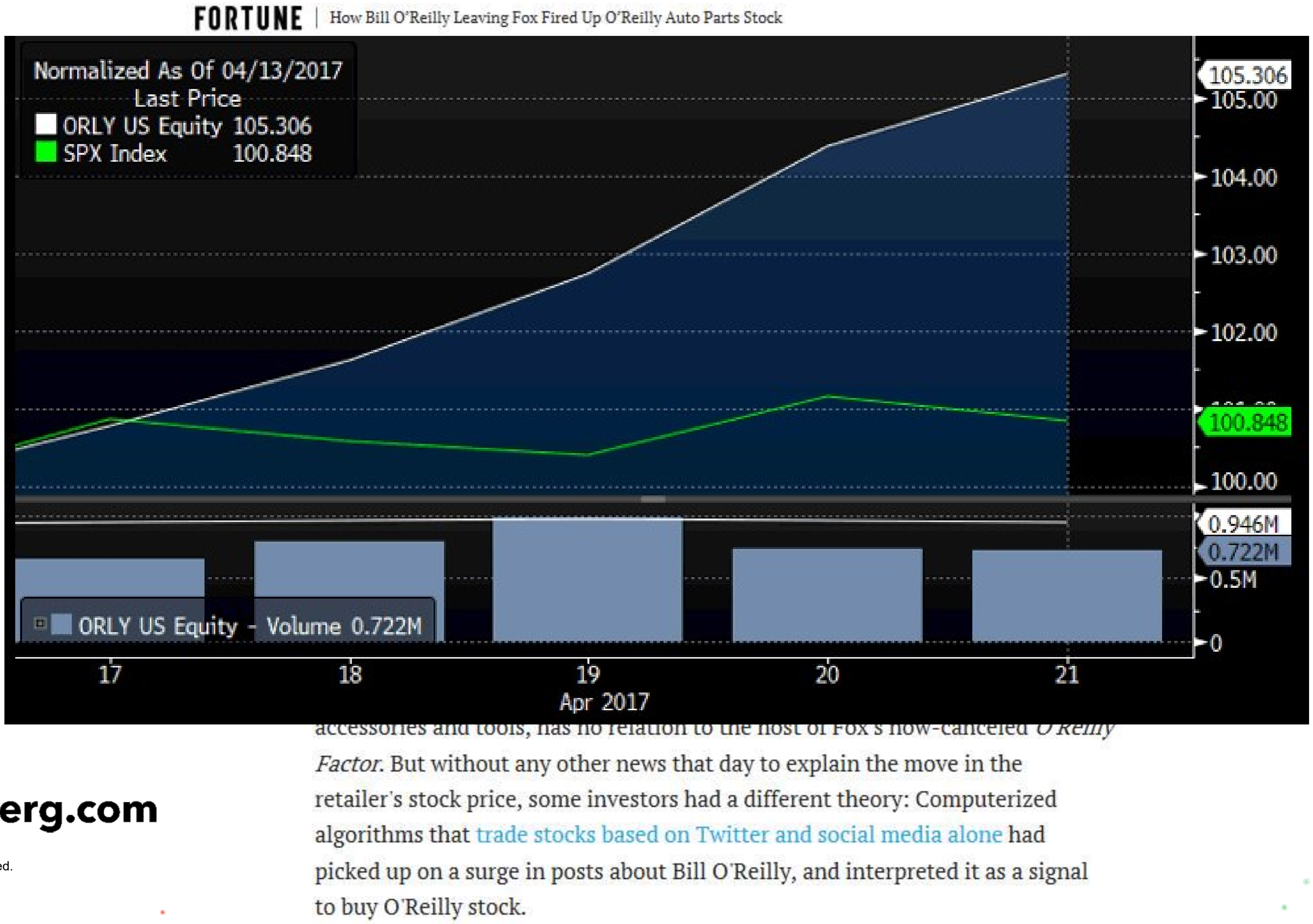
Bloomberg

Ingestion + Creation

Analyses

Discovery

The Need for Precision



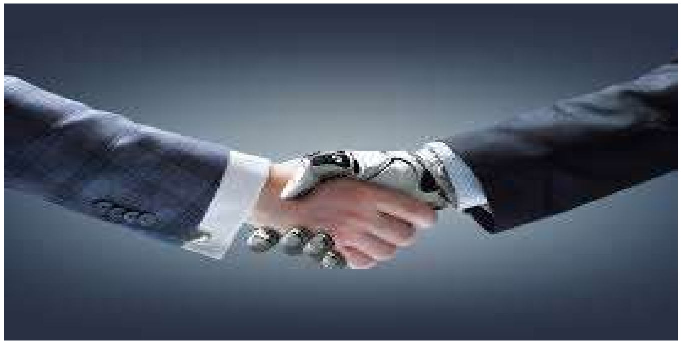
FORTUNE | How Bill O'Reilly Leaving Fox Fired Up O'Reilly Auto Parts Stock

accessories and tools, has no relation to the host of Fox's now-canceled *O'Reilly Factor*. But without any other news that day to explain the move in the retailer's stock price, some investors had a different theory: Computerized algorithms that trade stocks based on Twitter and social media alone had picked up on a surge in posts about Bill O'Reilly, and interpreted it as a signal to buy O'Reilly stock.

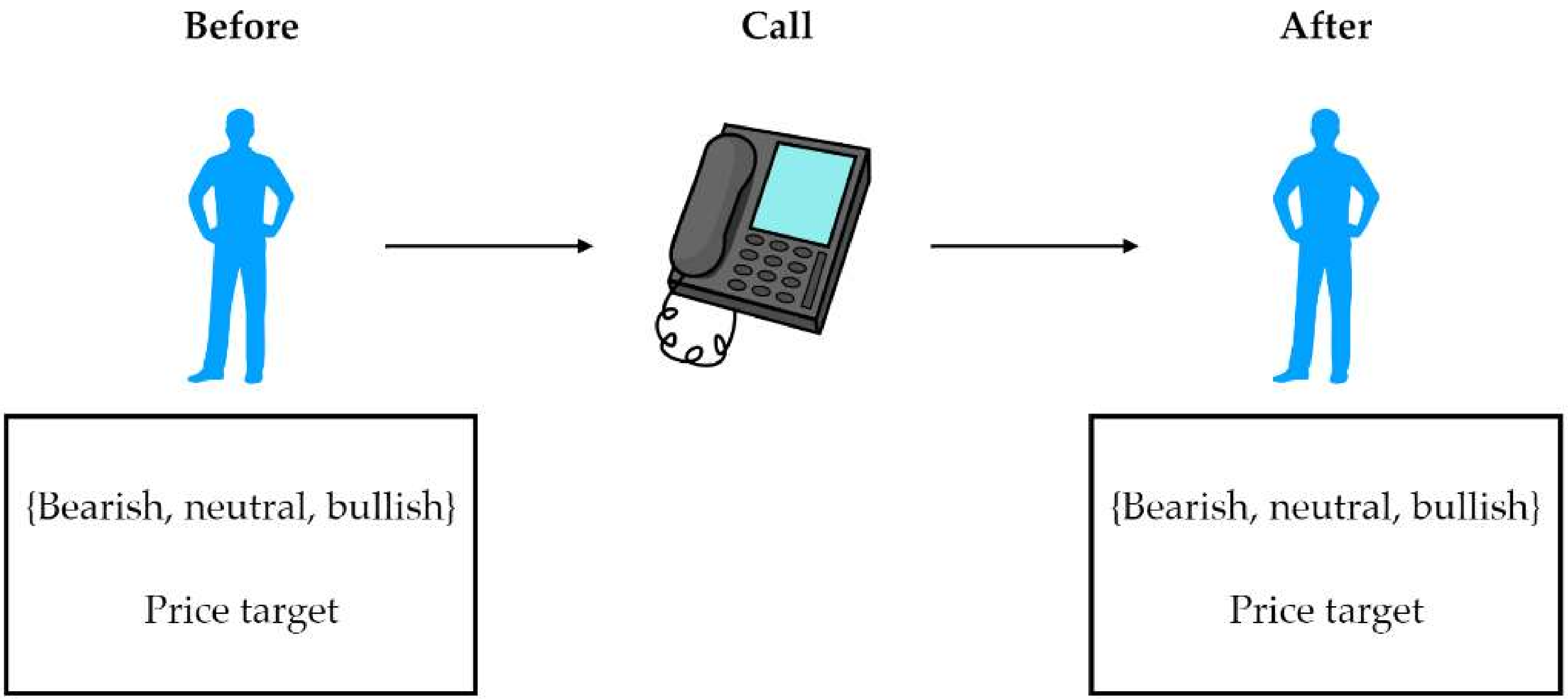
TechAtBloomberg.com

Human in the Loop

- In industry, no model is static
  - New entities
  - New contexts
  - New events and relationships
- Humans in the loop can:
  - Ameliorate lack of recall, precision
  - Provide important training data
- To enable hybrid workflow, more work is needed on:
  - Models "knowing what they know"
  - Active learning
  - Tools for humans to work with ML algorithms



Going Beyond Topics, Entities and Sentiment



Before

Call

After

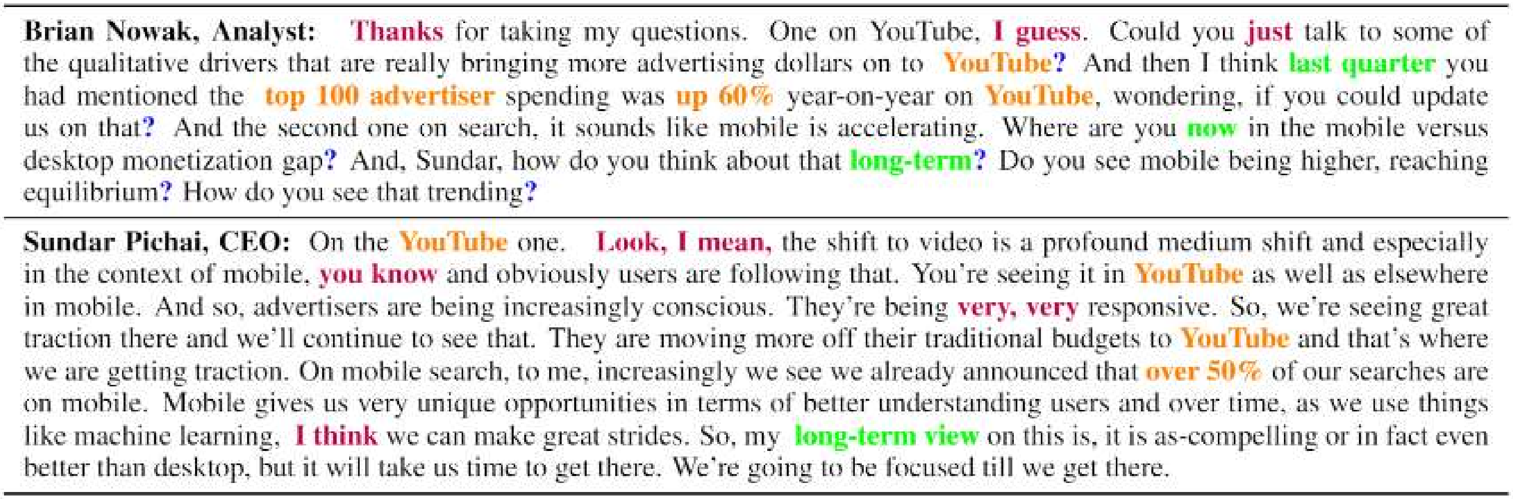
{Bearish, neutral, bullish}

Price target

{Bearish, neutral, bullish}

Price target

Going Beyond Topics, Entities and Sentiment



Brian Nowak, Analyst: **Thanks** for taking my questions. One on YouTube, **I guess**. Could you **just** talk to some of the qualitative drivers that are really bringing more advertising dollars on to **YouTube?** And then I think **last quarter** you had mentioned the **top 100 advertiser** spending was **up 60%** year-on-year on **YouTube**, wondering, if you could update us on that**?** And the second one on search, it sounds like mobile is accelerating. Where are you **now** in the mobile versus desktop monetization gap**?** And, Sundar, how do you think about that **long-term?** Do you see mobile being higher, reaching equilibrium**?** How do you see that trending**?**	
Sundar Pichai, CEO: On the **YouTube** one. **Look, I mean,** the shift to video is a profound medium shift and especially in the context of mobile, **you know** and obviously users are following that. You're seeing it in **YouTube** as well as elsewhere in mobile. And so, advertisers are being increasingly conscious. They're being **very, very** responsive. So, we're seeing great traction there and we'll continue to see that. They are moving more off their traditional budgets to **YouTube** and that's where we are getting traction. On mobile search, to me, increasingly we see we already announced that **over 50%** of our searches are on mobile. Mobile gives us very unique opportunities in terms of better understanding users and over time, as we use things like machine learning, **I think** we can make great strides. So, my **long-term view** on this is, it is as-compelling or in fact even better than desktop, but it will take us time to get there. We're going to be focused till we get there.	

Table 1: Eearnings calls are extremely complex examples of naturally-occurring discourse. In this example question-answer pair from a Google earnings call on October 27, 2016, the analyst asks **six distinct questions** in a single turn. Because the interaction originates as speech, there are **discourse markers and hedging.** The analyst and executive discuss **concrete entities and performance statistics** and **past, present and future** performance.

Going Beyond Topics, Entities and Sentiment

- Bullish analysts are called on earlier in earnings calls
- Bullish analysts ask more positive questions; neutral/bearish analysts ask more negative questions
- Bullish analysts ask more concrete questions
- Neutral/bearish analysts ask about the past more
- Earnings calls are moderately predictive of changes in analysts’ forecasts

Some Recent Work

- Meerkamp, P. & Zhou, Z. (2017) “Information extraction with character-level neural networks and free noisy supervision.” arXiv 1612.04118.

- Yang, Y., Irsoy, O., & Rahman, S. (2018) “Collective entity disambiguation with structured gradient tree boosting. In *Proc. NAACL*.

- Keith, K. & Stent, A. (2019) “Modeling financial analysts’ decision making via the pragmatics and semantics of earnings calls.” In *Proc. ACL*.

Data Science at Bloomberg

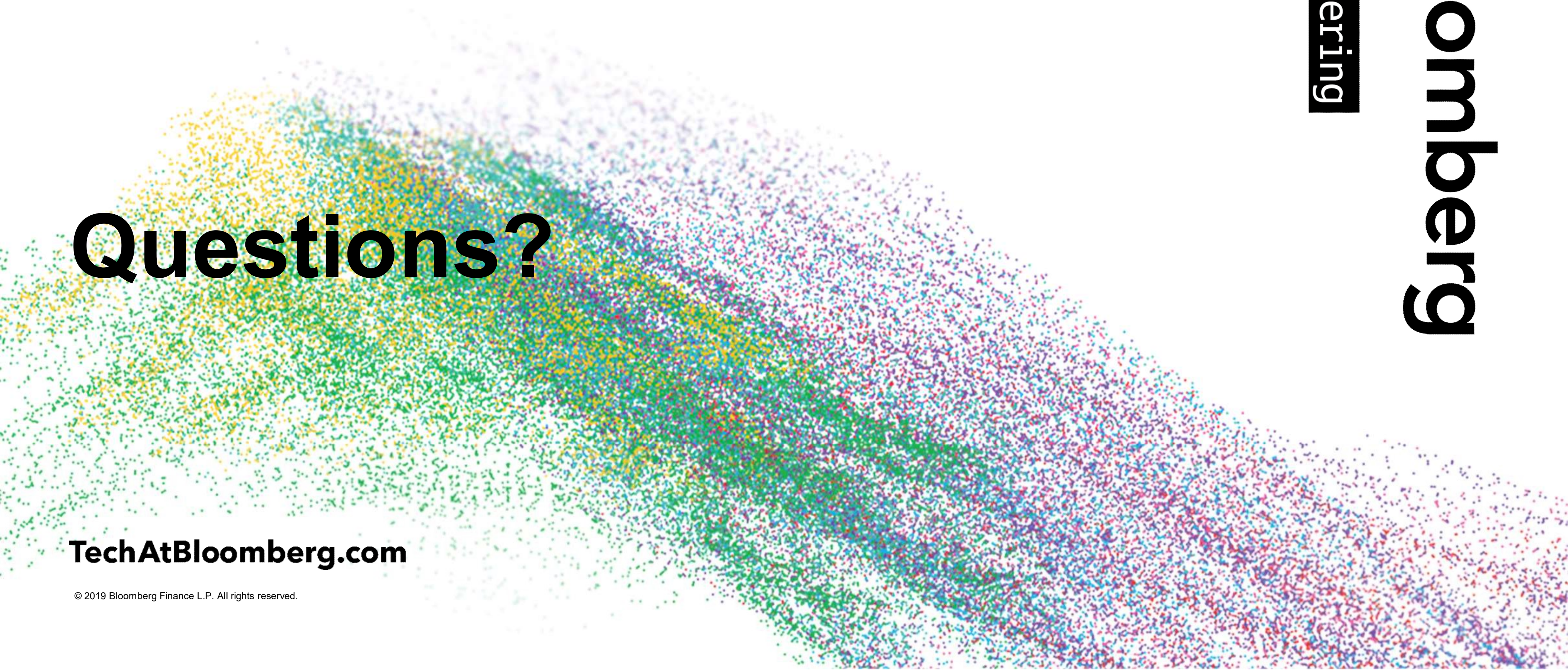
- **Our data:** We have the most text labeled with financial (*and legal!*) entities, instruments and topics anywhere
- **Our databases:** We have the most information about financial entities, instruments and topics anywhere
- **Our communications:** We connect the most extraordinary individuals across global finance
- **Our challenges:** Our business involves *every* AI challenge

Where to Learn More

- TechAtBloomberg.com
  - Papers published in 2018 at EMNLP, SIGIR, NAACL HLT, WWW, and more…
  - Data for Good Exchange, Machines + Media, and other exciting events
  - Our work with JupyterHub, Solr, and other open source frameworks
  - Foundations of Machine Learning (FOML), David Rosenberg's ML class

- GitHub.com/Bloomberg
  - bqplot: fast interactive plotting
  - sgtb: structured gradient tree boosting
  - foml: Foundations of Machine Learning
  - cnn-rnf: Convolutional Neural Networks with Recurrent Neural Filters
  - scatteract: extraction of labeled quantities from scatter plots

Thank You!



Questions?

Bloomberg Engineering

TechAtBloomberg.com