# Collecting and Analyzing Digital Data

Harm H. Schütt

Tilburg School of Economics and Management

VHB Workshop: “Einfluss der Digitalisierung auf Forschungsmethoden”
Feb 27, 2019

Why Digital Data?

Data Sources

Data Collection

Data Handling

Measurement Basics

NLP Basics

Examples: Word Lists

Examples: Cosine Similarity

Examples: Other

Outlook

# What do these items have in common?

- Sentiment
- Info processing cost
- Deceptive speech
- Macroeconomic uncertainty
- Firm-level political risk
- News audience characteristics
- Product market competition
- Financial constraints
- Product market synergies
- Ex-ante litigation risk
- News staleness
- Fund investing style
- Emerging systemic bank risks
- Ex-ante M&A integration difficulty
- Managerial affective states
- Reporting complexity

All are hard to measure constructs that were successfully measured using digital data

# Constructive constructs

- Accounting deals with lots of latent concepts
- Large measurement issues
- Digital data provides new avenues for better measurement

# Construct → Data → Measurement (Not the other way around)

1. First session is about Data
   - Examples of Construct → Data link
   - Description of common sources of data
   - Illustration of data extraction / scraping techniques
2. Second about Measurement
   - Measurement Theory and NLP Basics
   - Reasons for why certain approaches dominate
   - Critical discussion of most common approaches
   - A look into the future

Purpose: Impart some knowledge but also stimulate ideas

*Some material represents very condensed outtakes from our TiSEM Ph.D. course materials

Why Digital Data?

Data Sources

Data Collection

Data Handling

Measurement Basics

NLP Basics

Examples: Word Lists

Examples: Cosine Similarity

Examples: Other

Outlook

# Sources

- Legal Text
- Company Disclosures
- Websites
- Conference Calls (Transcripts, Video, Audio)
- Server Logs
- XBRL
- TV Interviews
- News Articles
- Social Media
- Logistics Data
- Internal Company Data (e.g., Uber, Facebook, Market Centers, Mint)
- Sensor Data

Construct $\rightarrow$ Data link?

# "Classic" examples

| Source | Study |
|---|---|
| DEF 14A Proxy Statement | *Incentives of compensation consultants and CEO pay* *(Cadman et al., 2010)* CD&A section requires numerous disclosures by the Compensation Committee |
| 8-K filing | *Inducement Grants, Hiring Announcements and Adverse Selection for New CEOs* *(Cadman et al., 2018)* New CEO appointments + new contract features require ad-hoc note |
| XBRL tags | *Measuring Accounting Reporting Complexity with XBRL* *(Hoitash and Hoitash, 2017)* Preparation and disclosure of more items complicated because requires greater knowledge of authoritative accounting standards |

# "Exotic" examples

| Source | Study |
|---|---|
| Hospital admission data | *Worrying about the Stock Market: Evidence from Hospital Admissions* (Engelberg and Parsons, 2016) |
| FAA online airmen inquiry website | *Pilot CEOs and corporate innovation* (Sunder et al., 2017)<br>Pilot credentials captures sensation seeking and associated with better innovation |
| Google Patents (entire history of U.S. patent documents: 7.8mn patents) | *Technological Innovation, Resource Allocation, and Growth* (Kogan et al., 2017)<br>Stock returns around announcement of patents as new measure of economic importance |

Why Digital Data?

Data Sources

Data Collection

Data Handling

Measurement Basics

NLP Basics

Examples: Word Lists

Examples: Cosine Similarity

Examples: Other

Outlook

# How do you get the data?

## Current Developments:

- Firms + government agencies around the world have increasingly digitized records + collect more data
- Willingness to work with academics is slowly increasing
- Increasing number of data vendors

## How to get the Data:

- Company collaborations
- Scraping or APIs
- Online surveys / apps
- Administrative data increasingly publicly available

# Getting data: APIs

## Introduction

The Twitter API platform offers three tiers of search APIs:

**Standard** This search API searches against a sampling of recent Tweets published in the past 7 days. Part of the 'public' set of APIs.

**Premium** Free and paid access to either the last 30 days of Tweets or access to Tweets from as early as 2006. Built on the reliability and full-fidelity of our enterprise data APIs, provides the opportunity to upgrade your access as your app and business grow.

**Enterprise** Paid (and managed) access to either the last 30 days of Tweets or access to Tweets from as early as 2006. Provides full-fidelity data, direct account management support, and dedicated technical support to help with integration strategy.

## Feature summary

| Category | Product name | Supported history | Query capability | Counts endpoint | Data fidelity |
|---|---|---|---|---|---|
| Standard | Standard Search API | 7 days | Standard operators | Not available | Incomplete |
| Premium | Search Tweets: 30-day endpoint | 30 days | Premium operators | Available | Full |
| Premium | Search Tweets: Full-archive endpoint | Tweets from as early as 2006 | Premium operators | Available | Full |
| Enterprise | 30-day Search API | 30 days | Premium operators | Included | Full |
| Enterprise | Full-archive Search API | Tweets from as early as 2006 | Premium operators | Included | Full |

# Getting data: APIs

```
import twitter
authentication = twitter.OAuth(Access_Token, Access_Token_Secret,
                               Consumer_Key, Consumer_Secret)
twitter_api = twitter.Twitter(auth=authentication)
tweets = twitter_api.search.tweets(q="#BigData", count=5)
```

# Getting data: APIs

```
tweets["statuses"][0]
```


```
{'created_at': 'Tue Jan 22 19:19:23 +0000 2019',
'id': 1087791857126518784,
'id_str': '1087791857126518784',
'text': 'RT @reinforcelabtwt: #Triple #Bonanza #Offer - Flat 15% Off + 20% #Cashback + 1 Self-Paced #Course #Free\n\
'truncated': False,
'entities': {'hashtags': [{'text': 'Triple', 'indices': [21, 28]},
{'text': 'Bonanza', 'indices': [29, 37]},
{'text': 'Offer', 'indices': [38, 44]},
{'text': 'Cashback', 'indices': [66, 75]},
{'text': 'Course', 'indices': [91, 98]},
{'text': 'Free', 'indices': [99, 104]},
{'text': 'AI', 'indices': [131, 134]}],
'symbols': [],
'user_mentions': [{'screen_name': 'reinforcelabtwt',
'name': 'Reinforce Lab',
'id': 742544340430389248,
'id_str': '742544340430389248',
'indices': [3, 19]}],
'urls': [{'url': 'https://t.co/xI4T0shH6J',
'expanded_url': 'https://bit.ly/2W8VjGi',
'display_url': 'bit.ly/2W8VjGi',
'indices': [106, 129]}]},
'metadata': {'iso_language_code': 'en', 'result_type': 'recent'},
'source': '<a href="https://www.gretsky.host/" rel="nofollow">Media_Poster_V_1_0</a>',
'in_reply_to_status_id': None,
'in_reply_to_status_id_str': None,
'in_reply_to_user_id': None,
'in_reply_to_user_id_str': None,
'in_reply_to_screen_name': None,
'user': {'id': 2736397896,
'id_str': '2736397896',
'name': 'Ham Gretsky',
'screen_name': 'ham_gretsky',
'location': 'Manhattan, NY',
'description': "I'm not always self aware, but when I am...Wait What? #AI #BigData #Programming #WebDevelopment #IoT
'url': 'https://t.co/p7dovp3d8c',
'entities': {'url': {'urls': [{'url': 'https://t.co/p7dovp3d8c',
```

# Examples of open (REST) APIs

| | |
|---|---|
| TrueFace.AI | Advanced facial recognition in images |
| Zillow | Estimates on real estate listings |
| BigDataCloud | IPv4 address geolocation lookup. |
| Charity Search | Non-profit charity data |
| Clearbit | Search for company logos |
| markerapi | Trademark Search |
| Whitepages Pro | Global identity verification with phone, address, email and IP |
| City Context | Crime, school and transportation data for US cities |
| Zipcodeapi | Zip codes for a city, distance between zip codes, etc. |
| opencorporates | Data collected from legal registers. |

All easily accessible via Python's requests, R's curl / httr, etc.

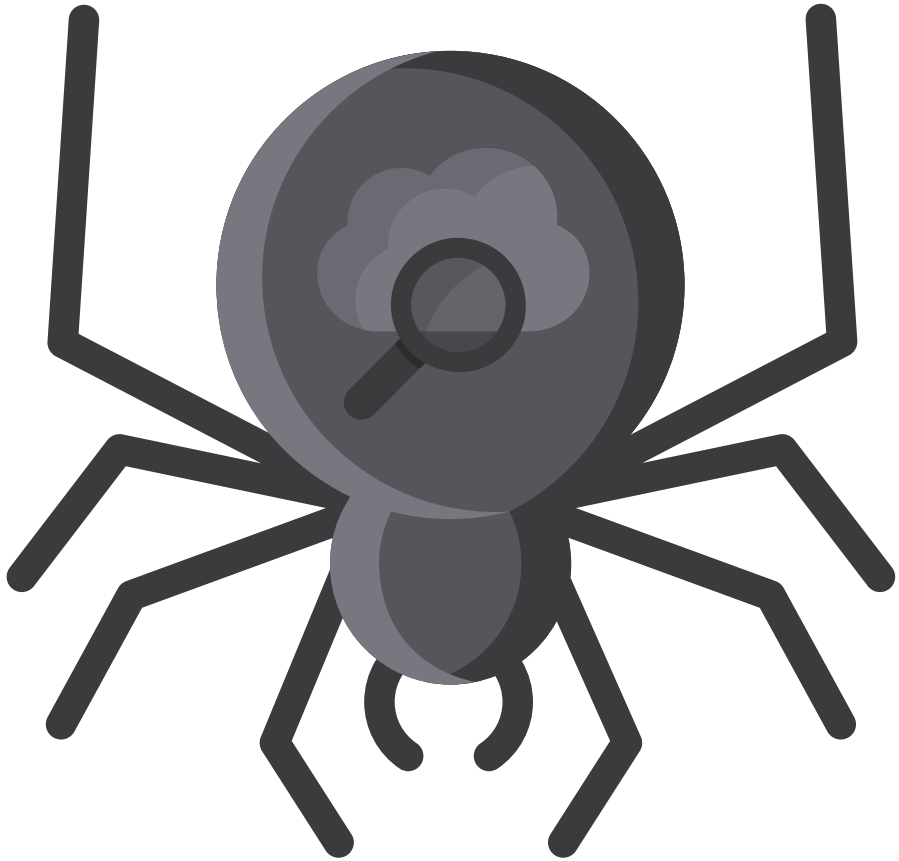
# SEC company filings

```
Description:           Master Index of EDGAR Dissemination Feed by Form Type
Last Data Received:    June 30, 2017
Comments:              webmaster@sec.gov
Anonymous FTP:         ftp://ftp.sec.gov/edgar/
```

| Form Type | Company Name | CIK | Date Filed | File Name |
|---|---|---|---|---|
| 1-A | 1212 Development Corp | 1705510 | 2017-05-10 | edgar/data/1705510/0001705510-17-000002.txt |
| 1-A | 1st stREIT Office Inc. | 1700461 | 2017-06-30 | edgar/data/1700461/0001387131-17-003536.txt |
| 1-A | Advanced Fuel Technologies Ltd | 1706951 | 2017-05-17 | edgar/data/1706951/0001683168-17-001312.txt |
| 1-A | Arcimoto Inc | 1558583 | 2017-06-22 | edgar/data/1558583/0001144204-17-033680.txt |
| 1-A | Campagna Motors USA Inc. | 1688545 | 2017-06-05 | edgar/data/1688545/0001144204-17-030964.txt |

# Scraping

- Everyone scrapes
- With some programming skills, this isn't hard (but hard to do it responsibly)
- Volume of scraping poses significant burden on websites



Not Your Average Web Crawler

# Common tools

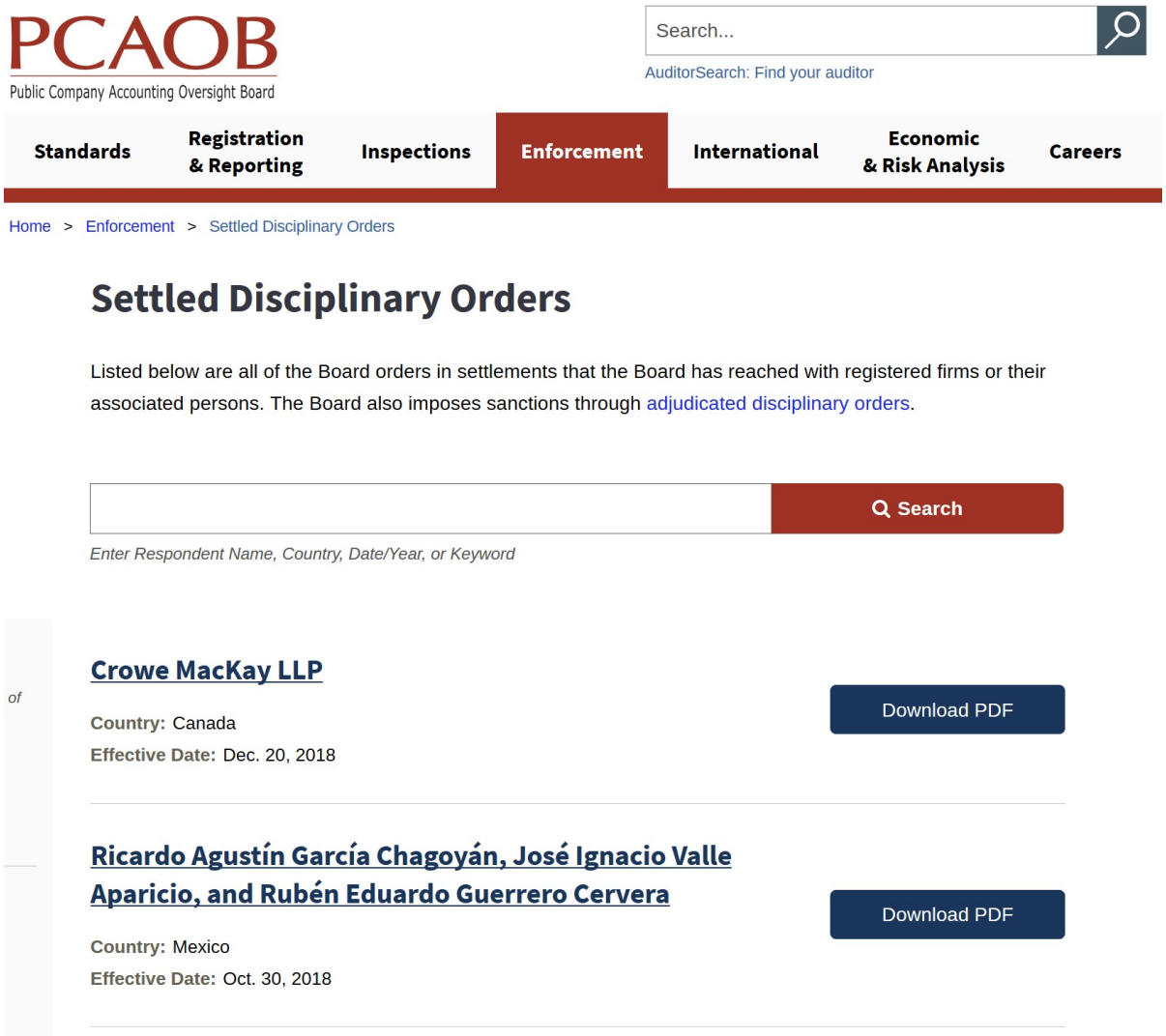
- Programming Languages: Pearl, R, or Python
- Python by far most common now
- Mostly because of powerful open source libraries:
  - requests – sending and getting data to websites or servers
  - beautifulsoup4 – Powerful html scraping functionality
  - selenium bindings – automating web browser activity
  - scrapy – web crawler framework
  - NLTK, spaCy, flair, gensim – powerful NLP libraries
- Alternative: Paid services like scrapinghub.com!

# Responsible scraping

- Negatively affects access to a site
  (people have gotten their university banned from google scholar because of scraping)
- Not always public data! - Ownership of info on private sites a gray area
- Legal consequences  fines, lawsuits
- Be nice to the site
  - Is my webscraper affecting site latency?
  - Minimize the number and rate of requests
  - Is multiprocessing necessary? Can I run it at night?
  - Are multiple pages populated from a single file?
  - If the site had my name and e-mail, would I still scrape?

Heavily inspired by the excellent talk by Graham McDonald (Webscraping Responsibly, November 2018)

# Scraping example



# First Step: Check robots.txt

robots.txt lists the servers policy towards robots

https://pcaobus.org/robots.txt:

```
User-agent: *
Crawl-delay: 1
```

As a comparison:

https://europa.eu/robots.txt

```
# robots.txt for EUROPA httpd-80 production server
# last update on 29/04/2011
# created by Rudi Mosselmans on 8/10/96

User-agent: *          # match any robot name
Disallow: /cgi-bin/    # don't allow robots into cgi-bin
Disallow: /eur-lex/    # don't index old Eurlex - 13/09/2006 Reque
Disallow: /archives/   # don't index the archives
```

# Scraping example

```
</div><div class="ms-webpart-chrome ms-webpart-chrome-fullWidth ">
<div class="ms-webpart-chrome-title" id="WebPartWPQ8_ChromeTitle">
<span title="" id="WebPartTitleWPQ8" class="js-webpart-titleCell"><h2 sty

</div><div WebPartID="00000000-0000-0000-0000-000000000000" HasPers="true
<?xml version="1.0" encoding="utf-8"?>
<p><table><tr>
<td><a href="https://pcaobus.org/Enforcement/Decisions/Documents/
105-2018-025-Crowe-MacKay.pdf">Crowe MacKay LLP</a>
</td><td>https://pcaobus.org/Enforcement/Decisions/Documents/
105-2018-025-Crowe-MacKay.pdf</td><td>Crowe MacKay LLP</td>
<td>Canada</td>
<td>12/20/2018 5:00:00 AM</td>
<td></td><td></td><td></td><td></td><td></td>
</tr><tr>
<td><a href="https://pcaobus.org/Enforcement/Decisions/Documents/
105-2018-021-Chagoyn-Aparicio-Cervera.pdf">Ricardo Agustn Garca Chagoyn,
```

# Simple scraping

```
url = "https://pcaobus.org/Enforcement/Decisions/Pages/default.as
raw = requests.get(url)
soup = BeautifulSoup(raw.content)
tables = soup.find_all(name="table")
decision_table = tables[0]
df = pd.read_html(str(decision_table))[0]
```

df

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| 0 | Crowe MacKay LLP | https://pcaobus.org/Enforcement/Decisions/Docu... | Crowe MacKay LLP | Canada | 12/20/2018 5:00:00 AM | NaN | NaN | NaN |
| 1 | Ricardo Agustín García Chagoyán, José Ignacio ... | https://pcaobus.org/Enforcement/Decisions/Docu... | Ricardo Agustín García Chagoyán, José Ignacio ... | Mexico | 10/30/2018 4:00:00 AM | NaN | NaN | NaN |
| 2 | Deloitte LLP | https://pcaobus.org/Enforcement/Decisions/Docu... | Deloitte LLP | Canada | 10/16/2018 4:00:00 AM | NaN | NaN | NaN |
| 3 | Zhang Hongling CPA, P.C. and Hongling Zhang, CPA | https://pcaobus.org/Enforcement/Decisions/Docu... | Zhang Hongling CPA, P.C. and Hongling Zhang, CPA | United States | 10/2/2018 4:00:00 AM | NaN | NaN | NaN |
| 4 | Breard & Associates, Inc. Certified Public Acc... | https://pcaobus.org/Enforcement/Decisions/Docu... | Breard & Associates, Inc. Certified Public Acc... | United States | 8/9/2018 4:00:00 AM | NaN | NaN | NaN |
| 5 | Brian D. Donahue, CPA | https://pcaobus.org/Enforcement/Decisions/Docu... | Brian D. Donahue, CPA | United States | 7/24/2018 4:00:00 AM | NaN | NaN | NaN |
| 6 | Leigh J Kremer CPA and Leigh J. Kremer, CPA | https://pcaobus.org/Enforcement/Decisions/Docu... | Leigh J Kremer CPA and Leigh J. Kremer, CPA | United States | 7/24/2018 4:00:00 AM | NaN | NaN | NaN |

# More involved scraping

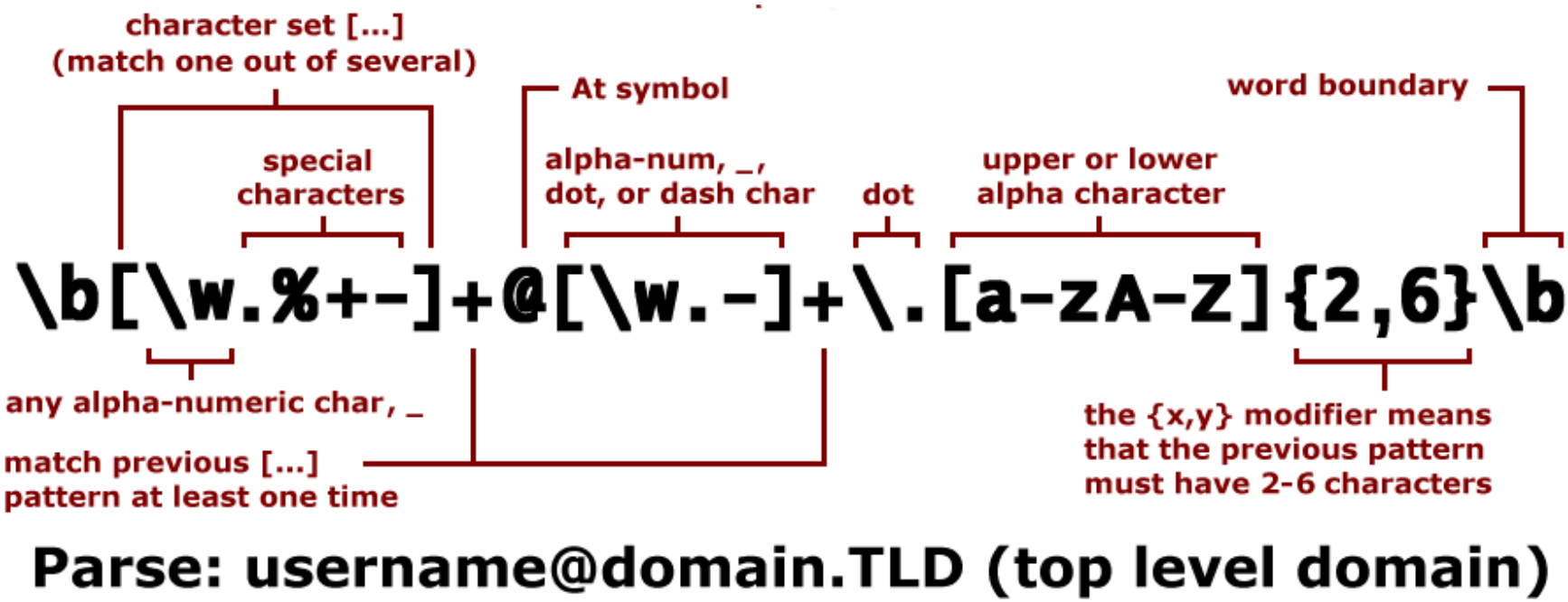
- The PCAOB side is actually full of javascript
- Web content is created dynamically
- Only shows the last 25 cases
- One way to get the full table is to automate your browser

```
from selenium import webdriver
from selenium.webdriver.common.keys import Keys
driver = webdriver.Firefox()
driver.get(url)
python_button = driver.find_element_by_id(f'cccLoadMoreBtnctl00_S
python_button.click()
```

# Parsing

- Find sections in text (e.g., Announcement of new director hiring)
- Find specific entities (e.g., dates, emails, links, headlines, addresses)
- Common combo is html text and well-tested regex rules.

# Parsing with regex




Source: twiki

- String that matches: john@doe.com
- String that doesn't match: john@doe.something (TLD is too long)

# Company collaborations

- Firms have digitized increasing amounts of new and old records
- Nowadays every firm has huge amounts of transaction data
- Probably the best 'new' source for constructs we Accountants care about
- Resurgence in Finance and Economics (e.g, the whole field of HH Finance)
- Not that many efforts yet in Financial Accounting (e.g., Lawrence et al. (2018))

Why Digital Data?

Data Sources

Data Collection

Data Handling

Measurement Basics

NLP Basics

Examples: Word Lists

Examples: Cosine Similarity

Examples: Other

Outlook

# Data issues example I: Firms mentions in texts

```
TRANSCRIPT: 120714a5563679.779

LANGUAGE: ENGLISH

PUBLICATION-TYPE: Transcript

SUBJECT: CONFERENCE CALLS (91%); WEBCASTS (90%); EXECUTIVES (90%); BUSINESS
DEVELOPMENT (90%); TRANSPORTATION SUPPORT SERVICES (90%); INVESTMENT
MANAGEMENT (89%); MANAGERS & SUPERVISORS (78%); BANKING & FINANCE (78%);
INVESTOR RELATIONS (78%); REAL ESTATE (78%); REAL ESTATE INVESTING (78%);
ECONOMIC CRISIS (72%); COMPUTER OPERATING SYSTEMS (65%)

COMPANY: MERRILL LYNCH & CO INC (84%); MORGAN STANLEY (58%); DEUTSCHE BANK AG (58%
CREDIT SUISSE GROUP AG (57%); GLOBAL LOGISTIC PROPERTIES LTD (57%)

TICKER: MS (NYSE) (58%); DEUT (JSE) (58%); DBK (FRA) (58%); DBK (BIT) (58%);
DBETN (JSE) (58%); DB (NYSE) (58%); CSGN (SWX) (57%); CS (NYSE) (57%);
MC0 (SGX) (57%)

COUNTRY: UNITED STATES (95%)

LOAD-DATE: December 12, 2014
```

Transcript: “Global Logistics Properties Transaction Briefing Conference Call and Webcast”

# Data issues example II: Scale

```
START OF THE PROGRAM  --- start time: 2016-07-02 12:51:53.
Preparing collection...
    # documents in collection: 19,512,183
    # documents in collection: 18,572,000
done in 461 min.
Creating aggregation index...
    # documents in collection: 18,572,000
done in 75 min.
Removing duplicates...
    # documents in collection: 18,572,000
performing aggregation... done.
deleting... done.
dropping aggregation index... done.
    # documents in collection: 11,052,250
done in 373 min.
Creating category/date index
    # documents in collection: 11,052,250
done in 33 min.
END OF PROGRAM --- 942 minutes ---
```

# Once you have the data

- Large amounts of text is a pain. Memory hungry and slow
- Do NOT reinvent the wheel, use tested databases
  - SQL if you query frequently
  - NOSQL, elasticsearch, mongoDB otherwise
  - Long term storage, Write speed, querying the data
- AWS, distributed computing, etc. All have pros- and cons
- Cleaning /data errors frequent and time consuming (e.g. server times, ticker in Lexis Nexis)
- Engage with your institution early about long term backup storage
- Prepare for the growing demands on data transparency and accessibility

Why Digital Data?

Data Sources

Data Collection

Data Handling

Measurement Basics

NLP Basics

Examples: Word Lists

Examples: Cosine Similarity

Examples: Other

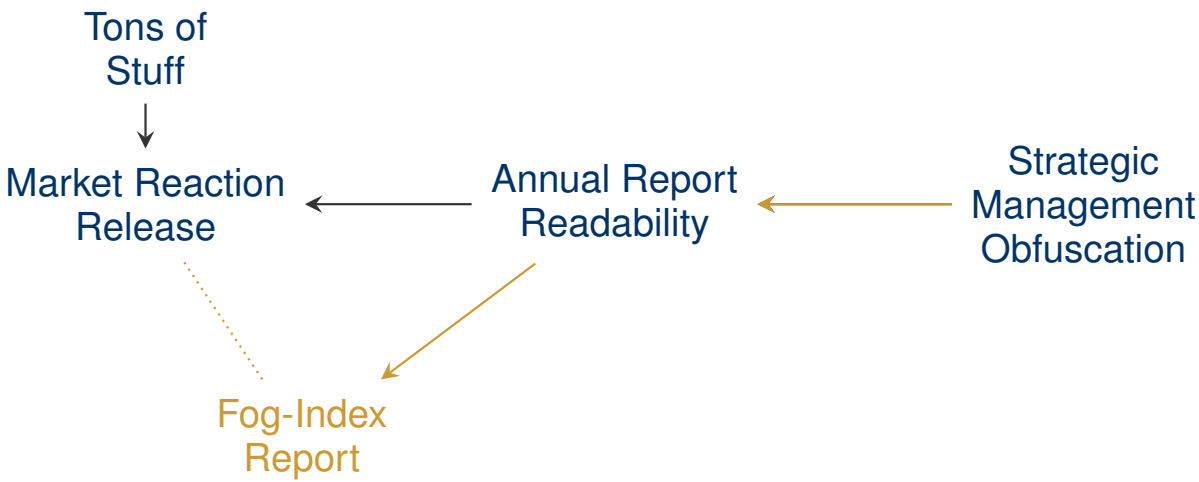
Outlook

# Measuring constructs

- Assume we are interested in
  $MarketReaction = \beta_0 + \beta_1 Obfuscation + u$
- Measure: $MarketReaction = \hat{\beta}_0 + \hat{\beta}_1 FogIndex + u$

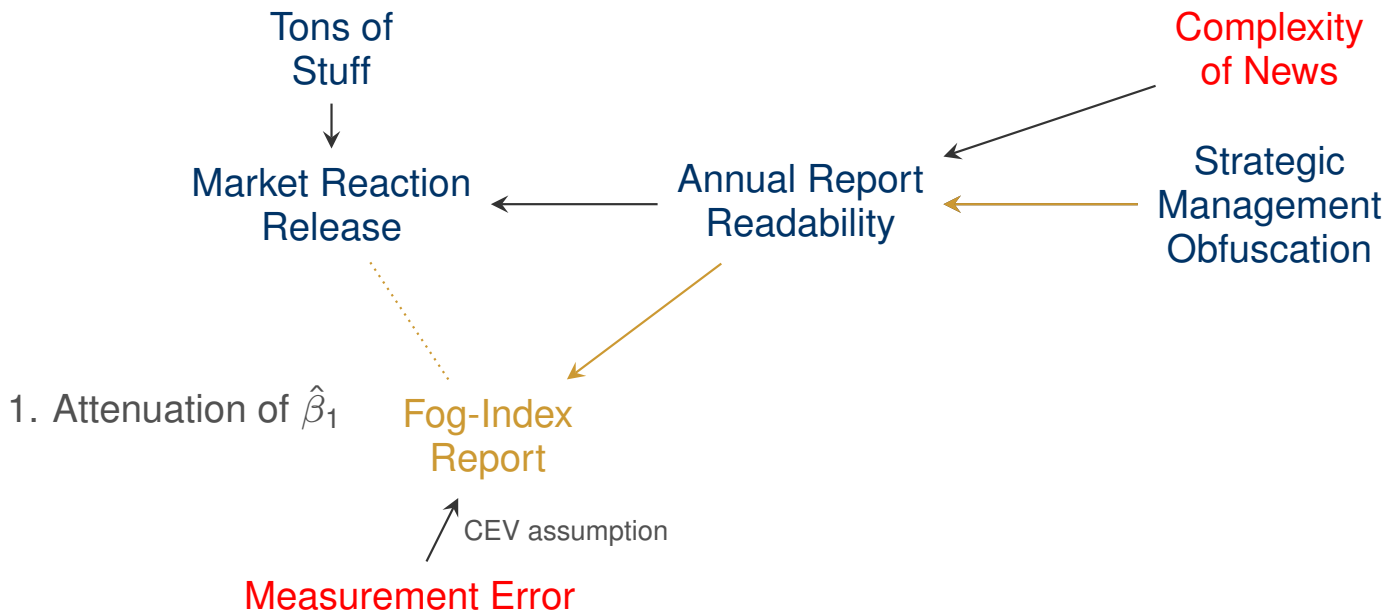
# Measuring constructs

- Assume we are interested in
  $MarketReaction = \beta_0 + \beta_1 Obfuscation + u$
- Measure: $MarketReaction = \hat{\beta}_0 + \hat{\beta}_1 FogIndex + u$



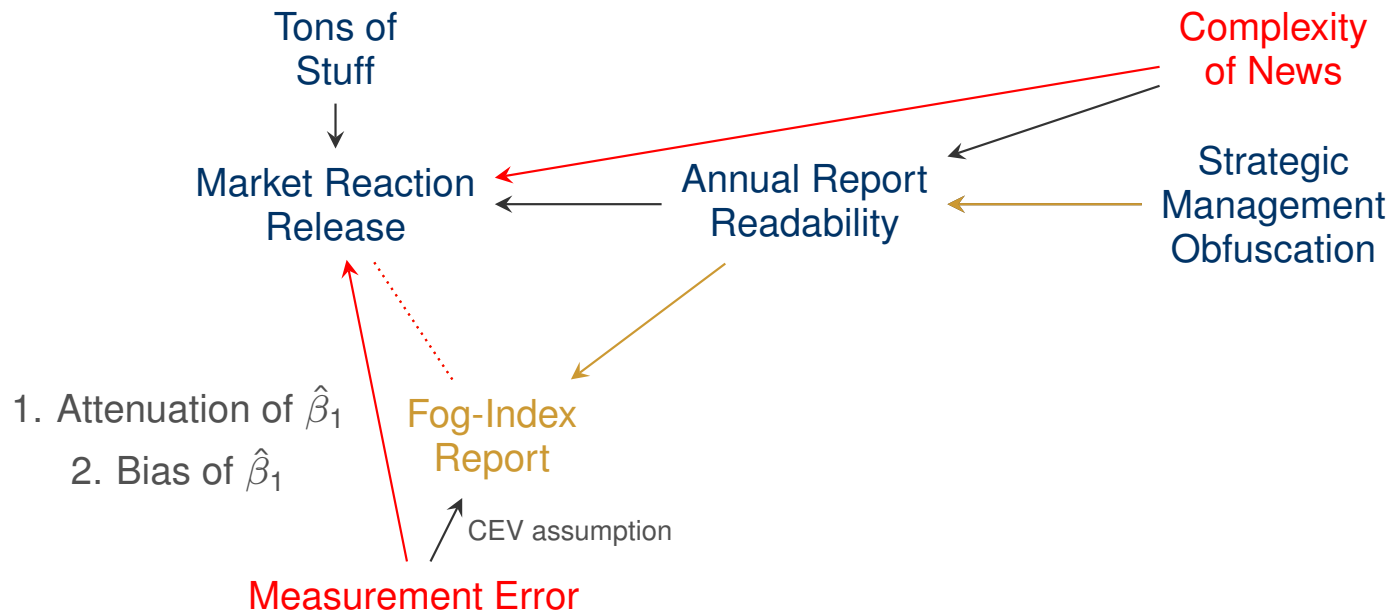
# Measuring constructs

- Assume we are interested in $\textit{MarketReaction} = \beta_0 + \beta_1 \textit{Obfuscation} + u$
- Measure: $\textit{MarketReaction} = \hat{\beta}_0 + \hat{\beta}_1 \textit{FogIndex} + u$



# Measuring constructs

- Assume we are interested in
  $MarketReaction = \beta_0 + \beta_1 Obfuscation + u$
- Measure: $MarketReaction = \hat{\beta}_0 + \hat{\beta}_1 FogIndex + u$



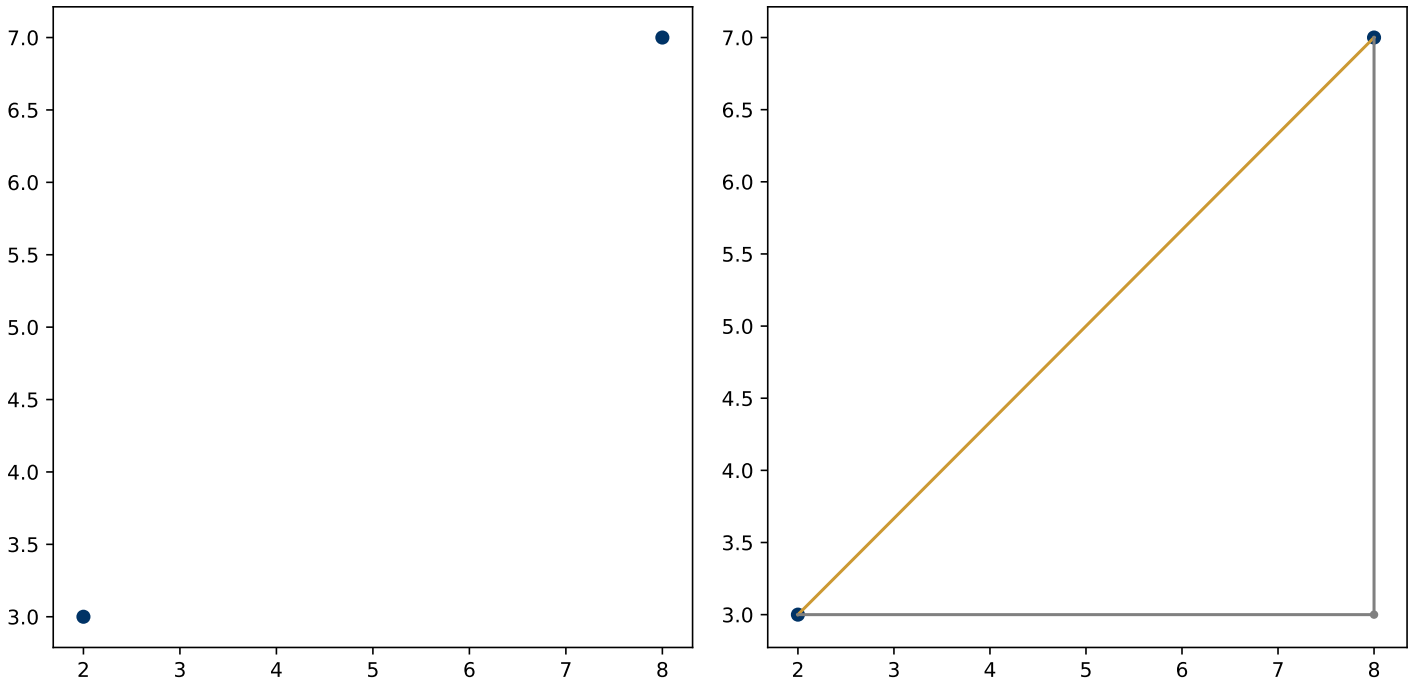
# Measuring constructs using text

*the frequency of word occurrence in an article furnishes a useful measurement of word significance (Luhn 1958, p. 160)*

- Measurement error particular problem with digital data
- Signal to noise ratio
- Often high dimensional (curse of dimensionality)
- Often hard to reason about source of measurement error

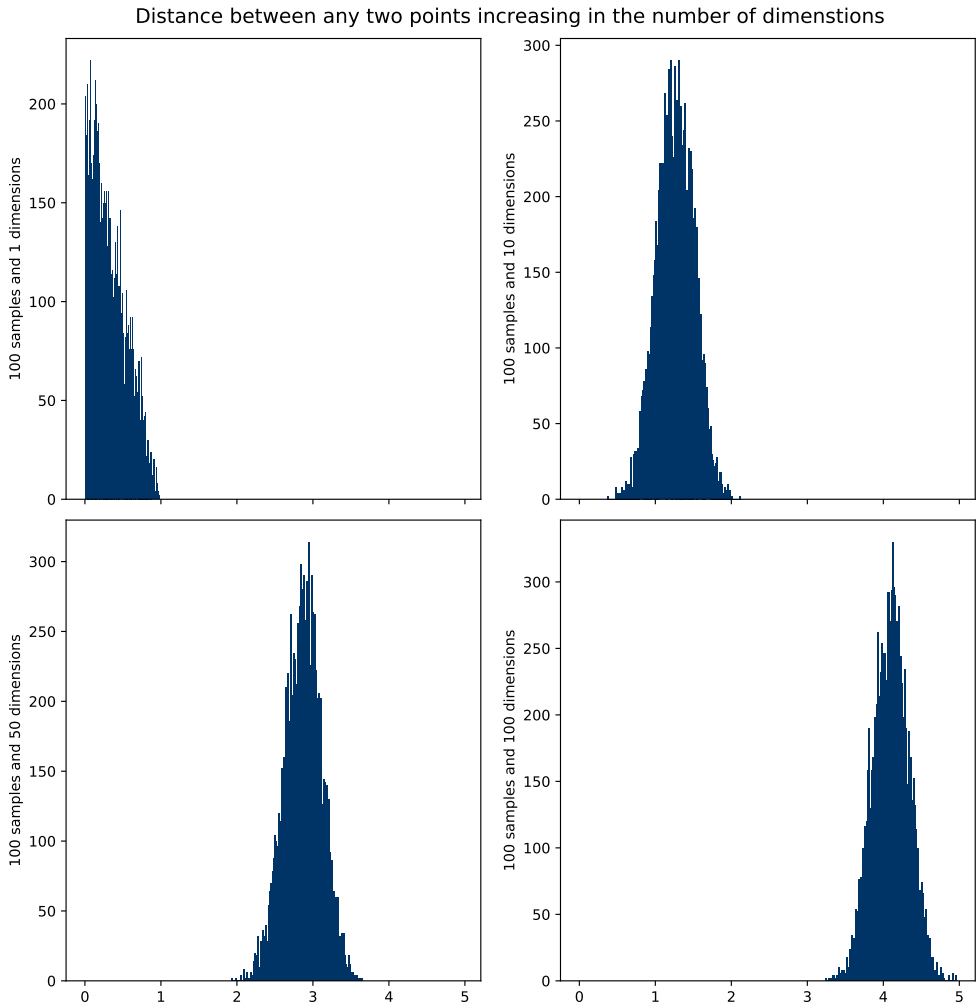
# Problems in high dimensions

- What's distance? Think Pythagoras ($c = \sqrt{a^2 + b^2}$)
- p-dimensions
- $d(q, r) = d(r, q) = \sqrt{(q_1 - r_1)^2 + (q_1 - r_1)^2 + ... + (q_p - r_p)^2}$



# Problems in high dimensions

- 100 random $U(0,1)^p$ points
- Points move further away AND closer to each other
- Measurement error becomes more important
- Text features a classic case



Why Digital Data?

Data Sources

Data Collection

Data Handling

Measurement Basics

NLP Basics

Examples: Word Lists

Examples: Cosine Similarity

Examples: Other

Outlook

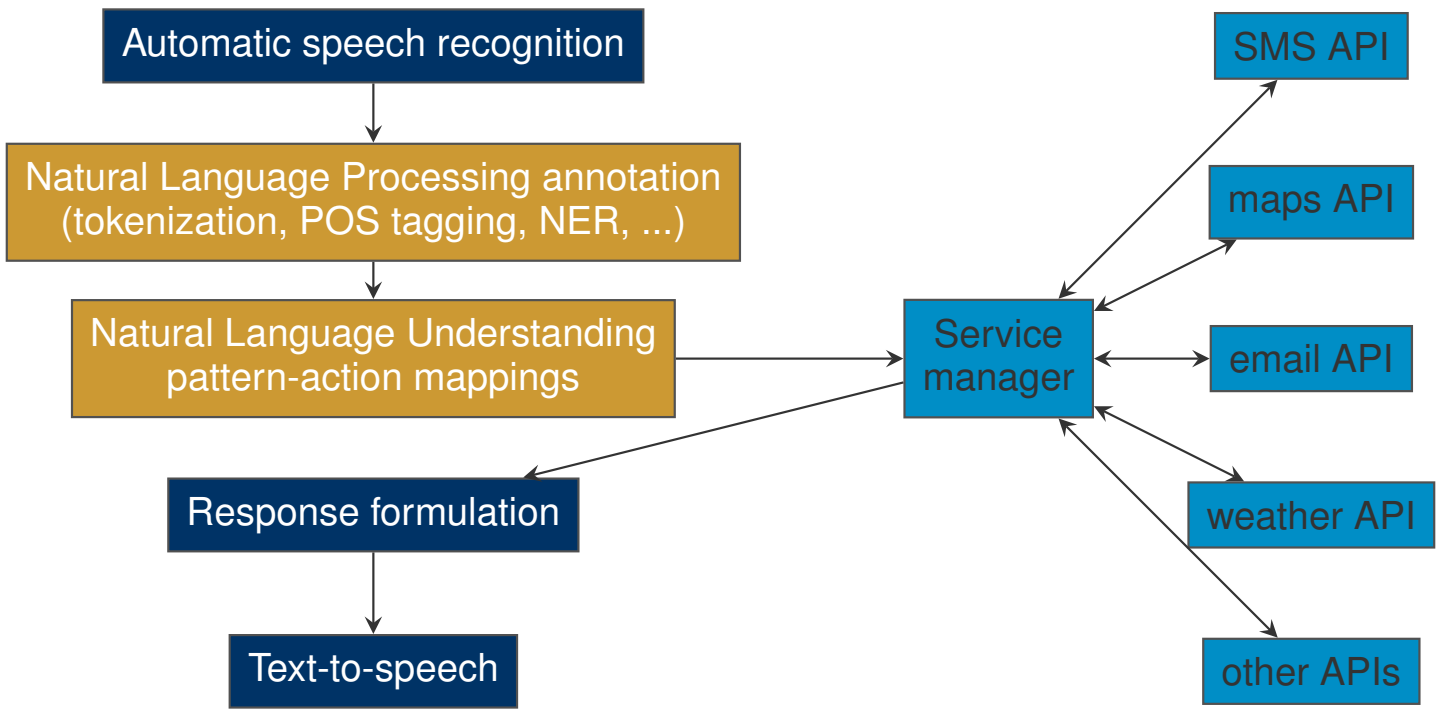
# Analyzing digital data

Main use goal is theory testing: many theories involve very hard to measure constructs of interest

- Exploding number of use cases for digital data.
- Use text analysis as running example
- Experience so far: simple is often enough/better:
  - Simpler approaches are like broadswords – Don't use them when you need a scalpel
  - Depending on use case: broadswords suffice
  - E.g. sentiment word lists work well on standardized texts (e.g., 10-Ks), but reach their limits with online product reviews.
  - Simpler approaches easier to interpret / check for errors
- But, recent trend towards modelling heterogeneity requires more sophisticated statistics

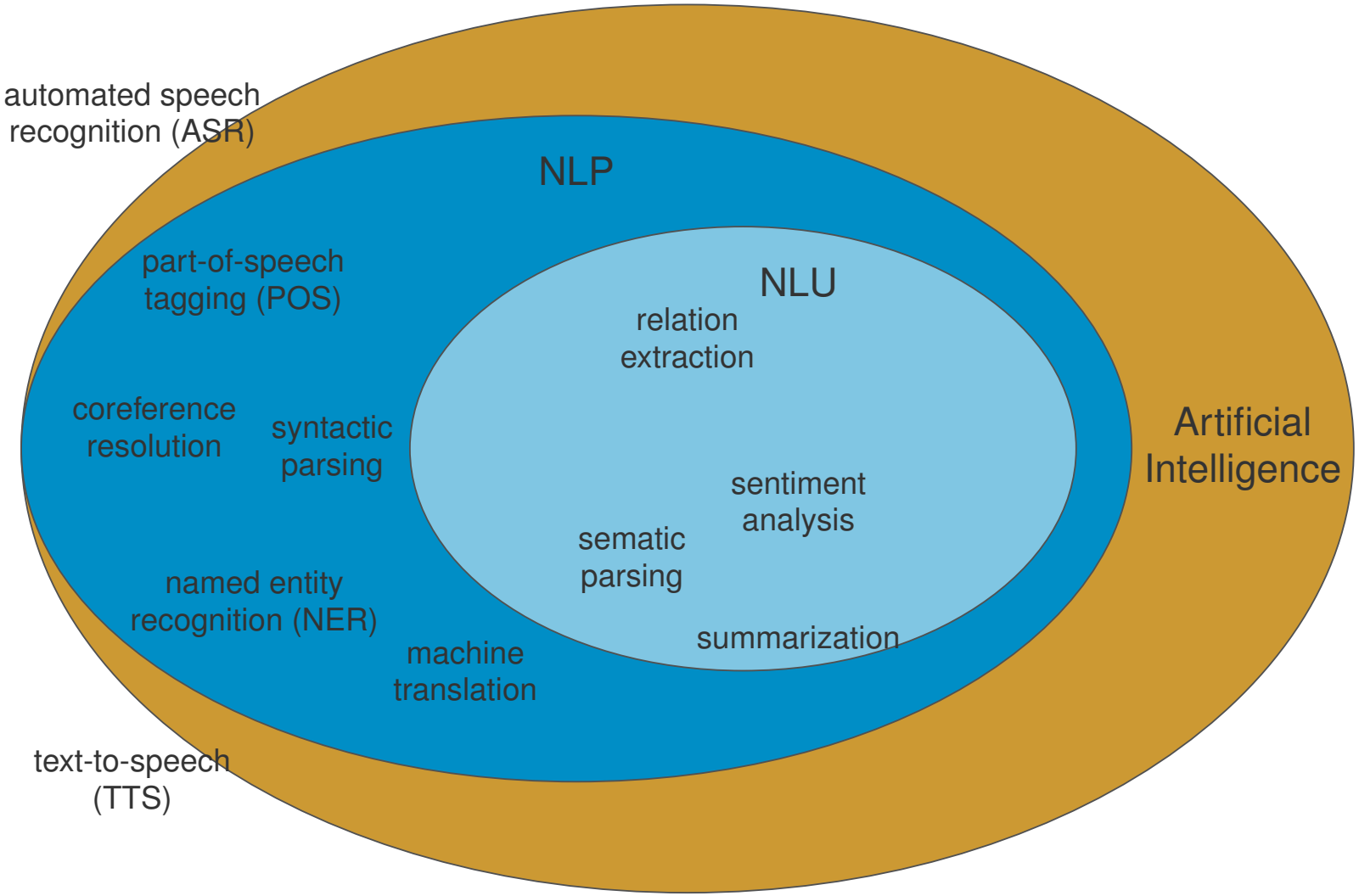
# What is natural language processing?

- NLP: teaching computers how to understand/generate human language
- Good NLP a necessary precursor to textual analysis or similar applications.
- Example: Siri




Source: How does Siri work?: Adapted from Bill MacCartney and Christopher Potts [Stanford CS224U (2015)]

# The field of natural language processing




Source: Adapted from Bill MacCartney (2014)

# Common NLP building blocks

Let's illustrate common NLP processing useful for later analyses. Take the sentence:

> "First quarter sales at luxury conglomerate LVMH proved the point."

FIRST QUARTER SALES AT LUXURY CONGLOMERATE LVMH PROVED THE POINT **Tokenizing**

Source of quote: Ft.com article "LVMH/luxury stocks: wealth of nations"

# Common NLP building blocks

Let's illustrate common NLP processing useful for later analyses. Take the sentence:

> "First quarter sales at luxury conglomerate LVMH proved the point."

| FIRST | QUARTER | SALES | AT | LUXURY | CONGLOMERATE | LVMH | PROVED | THE | POINT | **Tokenizing** |
|---|---|---|---|---|---|---|---|---|---|---|
| **ADJ** | **NOUN** | **NOUN** | **ADP** | **NOUN** | **NOUN** | **PROPN** | **VERB** | **DET** | **NOUN** | |

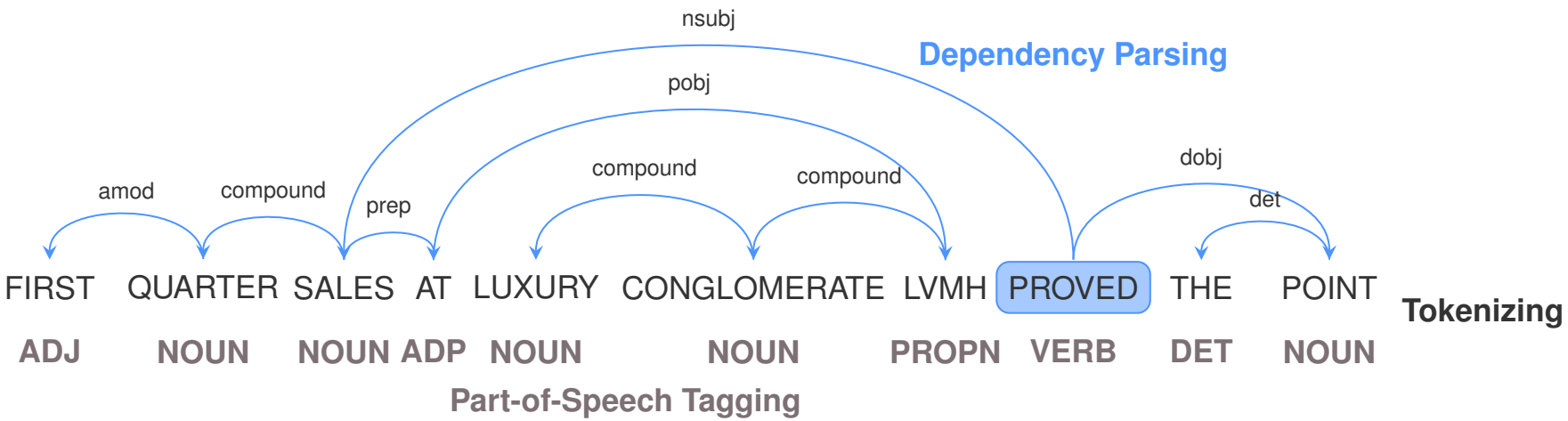
**Part-of-Speech Tagging**

Source of quote: Ft.com article "LVMH/luxury stocks: wealth of nations"

# Common NLP building blocks

Let's illustrate common NLP processing useful for later analyses. Take the sentence:

> "First quarter sales at luxury conglomerate LVMH proved the point."

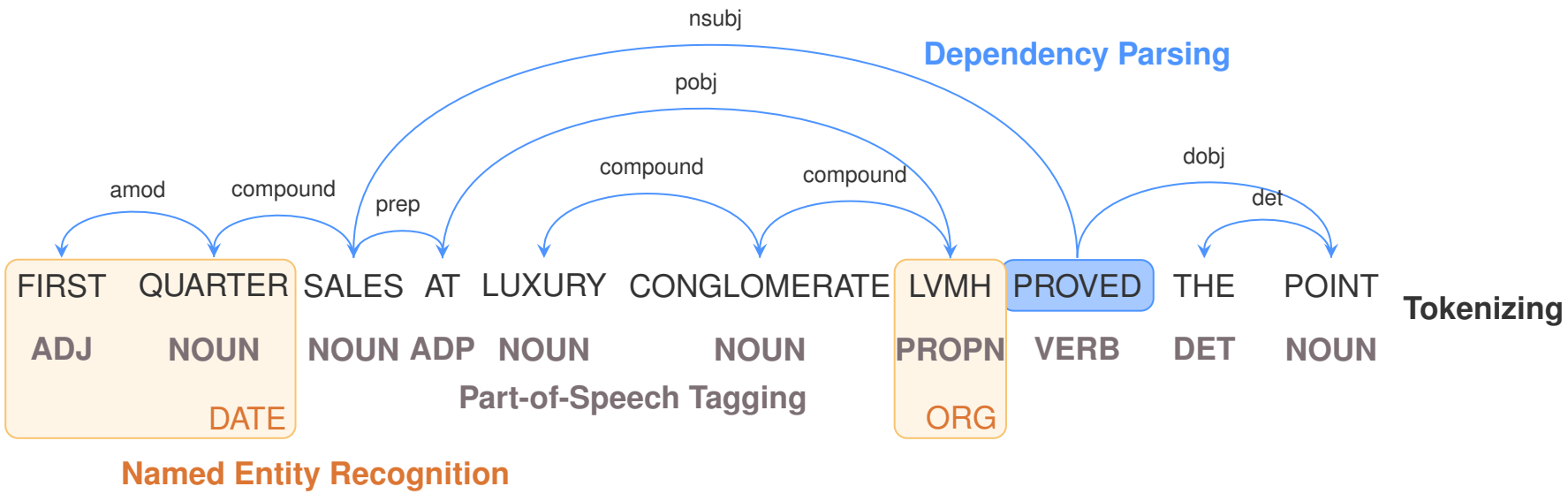



Source of quote: Ft.com article "LVMH/luxury stocks: wealth of nations"

# Common NLP building blocks

Let's illustrate common NLP processing useful for later analyses. Take the sentence:

> "First quarter sales at luxury conglomerate LVMH proved the point."




Source of quote: Ft.com article "LVMH/luxury stocks: wealth of nations"

# NLP Technology

## NLP is part of AI but heavily utilizes machine learning

Before talking about applications, let's briefly discuss how NLP systems work

- Early 60s, mostly rule based approaches
- At some point, became apparent that it's impossible to put language into a rule set
- These days: mostly neural nets trained on tons of pre-labeled data
- Many different models and training data out there: performance for use case depends on both
- For common use cases, probably no need to fit your own NLP models (unless you have very special texts)

# NLP Technology

## SpaCy's EN model

**en_core_web_lg**

RELEASE DETAILS

Latest: `2.0.0`

English multi-task CNN trained on OntoNotes, with GloVe vectors trained on Common Crawl. Assigns word vectors, context-specific token vectors, POS tags, dependency parse and named entities.

| | |
|---|---|
| LANGUAGE | EN English |
| TYPE | CORE Vocabulary, syntax, entities, vectors |
| GENRE | WEB written text (blogs, news, comments) |
| SIZE | LG 812 MB |
| PIPELINE | `tagger, parser, ner` |
| VECTORS | 685k keys, 685k unique vectors (300 dimensions) |
| SOURCES | OntoNotes 5, Common Crawl |
| AUTHOR | Explosion AI |
| LICENSE | CC BY-SA 3.0 |
| COMPAT | spaCy version |

| SYNTAX ACCURACY | | NER ACCURACY | |
|---|---|---|---|
| UAS | 91.89 | NER F | 85.85 |
| LAS | 90.07 | NER P | 85.54 |
| POS | 97.20 | NER R | 86.16 |

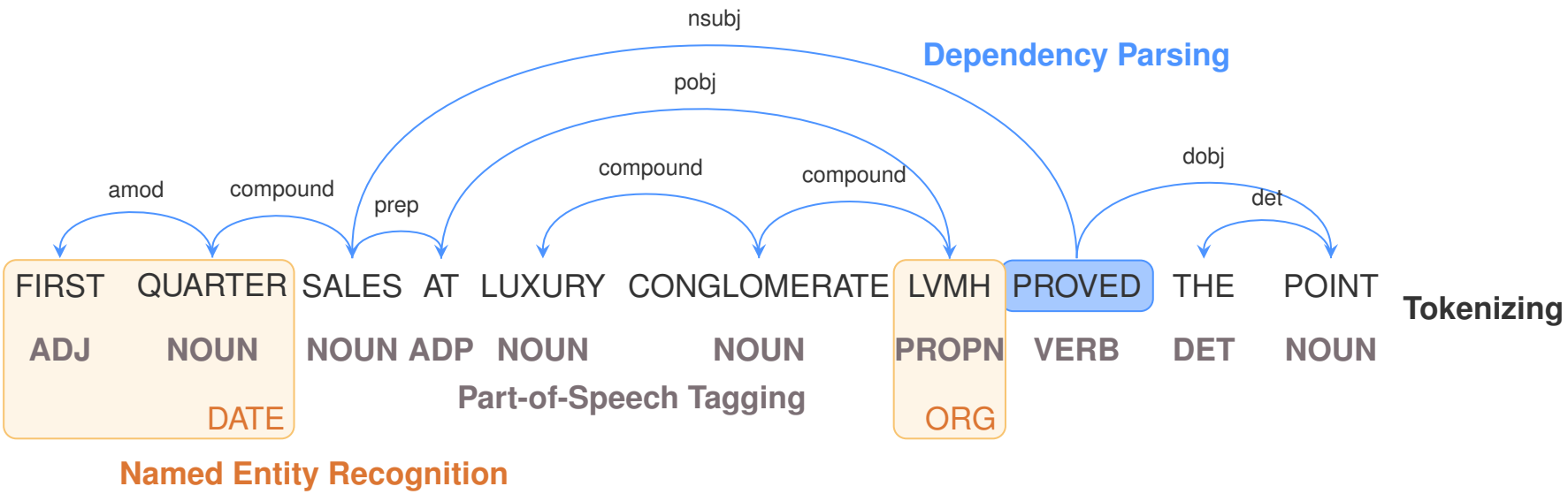
Source: https://spacy.io/models/ and https://catalog.ldc.upenn.edu/ldc2013t19

## OntoNotes Corpus

"The goal of the project was to annotate a large corpus comprising various genres of text (news, conversational telephone speech, weblogs, usenet newsgroups, broadcast, talk shows) in three languages (...) with structural information (syntax and predicate argument structure) and shallow semantics (word sense linked to an ontology and coreference)."

```
---------------------------------------------------------------------------------------------

Plain sentence:
---------------
    I ground the rye on number 6 click-LRB-out of 8-RRB-in my Champion Juicer grinder.

Treebanked sentence:
--------------------
    I ground the rye on number 6 click -LRB- out of 8 -RRB- in my Champion Juicer grinder .

Tree:
-----
    (TOP (S (NP-SBJ (PRP I))
            (VP (VBD ground)
                (NP (DT the)
                    (NN rye))
                (PP-MNR (IN on)
                        (NP (NP (NML (NN number)
                                     (CD 6))
                                (NN click))
                            (-LRB- -LRB-)
                            (PP (IN out)
                                (PP (IN of)
                                    (NP (CD 8))))
                            (-RRB- -RRB-)))
                (PP-LOC (IN in)
                        (NP (PRP$ my)
                            (NML (NNP Champion)
                                 (NNP Juicer))
                            (NN grinder))))
            (. .)))

Leaves:
-------
    0   I
    1   ground
        sense: ground-v.6
    2   the
    3   rye
    4   on
    5   number
```

# Applying NLP




- Different processing steps depending on use case
- NLP (only) creates inputs to the actual measure (e.g., sentiment)
- Best illustrated by examples

Why Digital Data?

Data Sources

Data Collection

Data Handling

Measurement Basics

NLP Basics

Examples: Word Lists

Examples: Cosine Similarity

Examples: Other

Outlook

# Word list use case: Sentiment Analysis I

- **Traditional approach: count sentiment words:**

  "Omni Consumer Products (OCP) delivered an outstanding quarter yet again."

- **But, word list must be chosen carefully:**

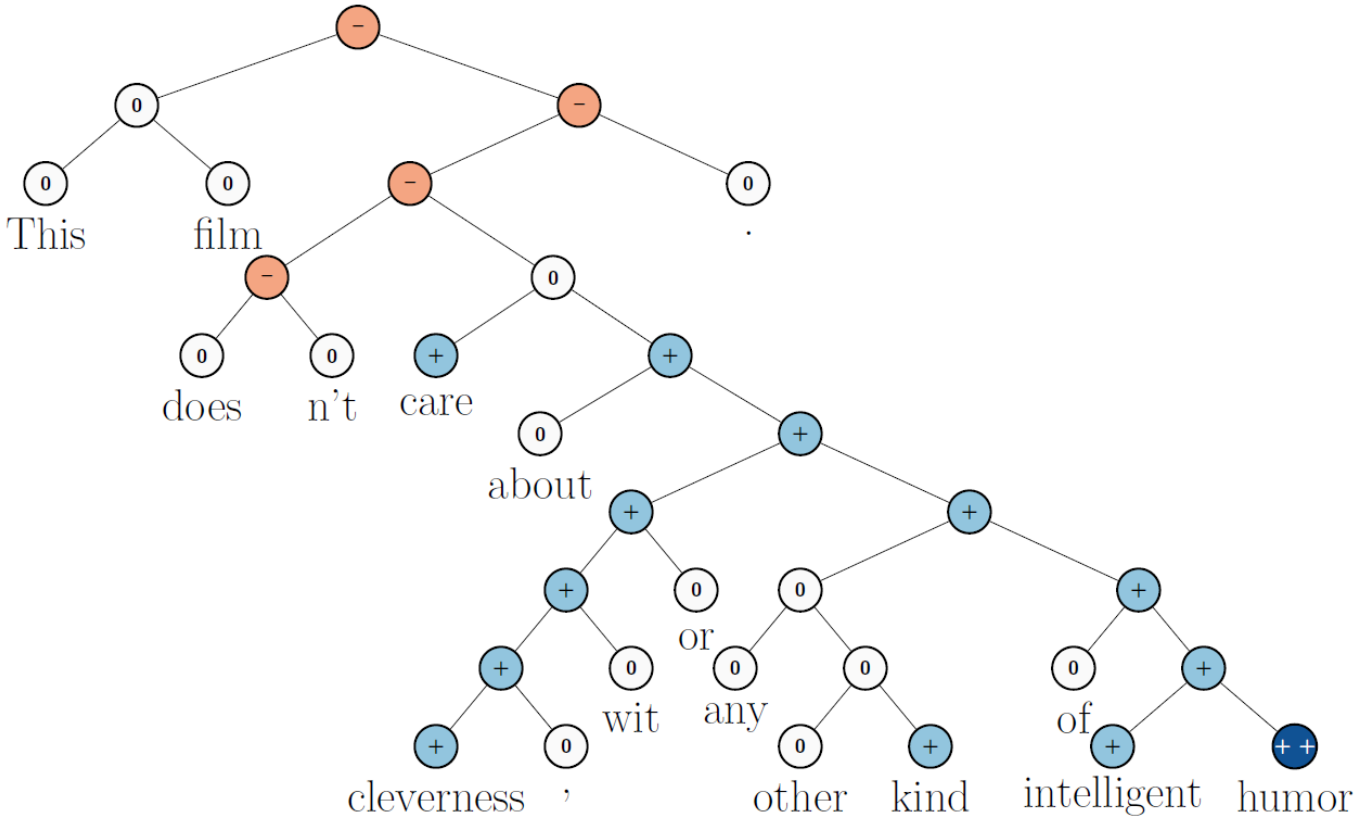
  "Existing Shell Shareholders and former BG Shareholders will own approximately 81% and 19% respectively of the outstanding Shell Shares"

- **Problematic if semantic composition becomes complex:**

  "Motivated, engaged, positive, but ultimately not up to our standards"

How would more sophisticated NLP help in the last case?

# Word list use case: Sentiment Analysis II




”we introduce a Sentiment Treebank. It includes fine grained sentiment labels for 215,154 phrases in the parse trees of 11,855 sentences and presents new challenges for sentiment compositionality. To address them, we introduce the Recursive Neural Tensor Network.”

Source: Socher, Richard, et al. ”Recursive deep models for semantic compositionality over a sentiment treebank.” Proceedings of the 2013 conference on empirical methods in natural language processing. 2013

# Word list use case: Sentiment Analysis III

Optimistic Diction words common in financial texts

- outstanding
- respect
- determined
- power
- trust
- security
- authority

Examples of these words in non-optimistic contexts

- “common shares outstanding”
- “ with respect to ”
- “discount rate is determined by ”
- “electric power generation”
- “Contractual Trust Arrangement”
- “asset-backed security”
- “fiscal authority”

83% of the most common optimistic Diction words do not appear in Loughran McDonalds (2011 JoF) list of positive words custom-made for annual reports.
Depending on use case, sentiment is often measured using negative words only.

# Term-document matrices I

## LVMH/luxury stocks: wealth of nations

Fans of upmarket bling are as conspicuous as a Louis Vuitton monogram

APRIL 10, 2018 0

The best things in life are free. The second best things are very, very expensive. That sentiment — attributed to style icon Coco Chanel — has been taken to heart by brand-conscious fans. Global enthusiasm for upmarket bling is as conspicuous as a Louis Vuitton monogram.

## Netflix and executives sued over bonus scheme

Shareholder complaint alleges 'rigging' and 'misleading' statements from streaming group


© Getty

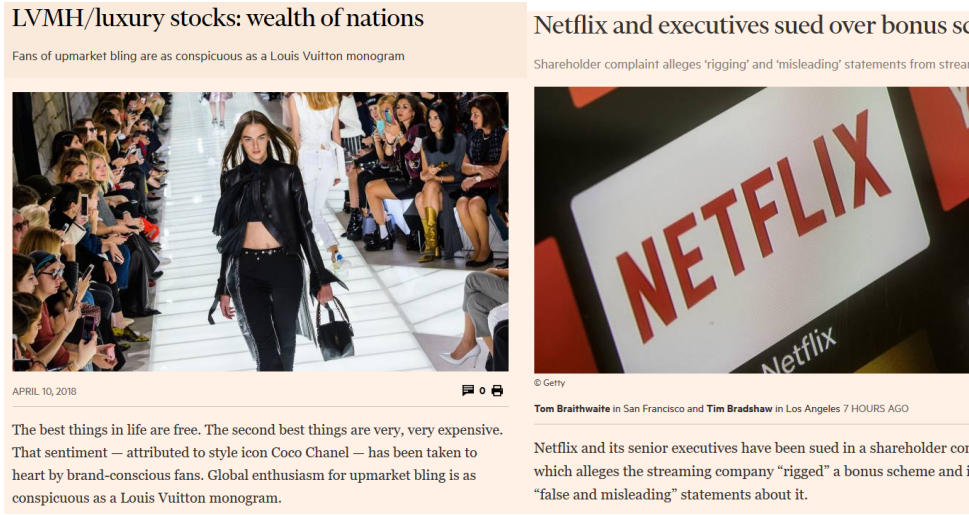
**Tom Braithwaite** in San Francisco and **Tim Bradshaw** in Los Angeles 7 HOURS AGO 1

Netflix and its senior executives have been sued in a shareholder complaint, which alleges the streaming company "rigged" a bonus scheme and issued "false and misleading" statements about it.

- Represent each document as a vector of word (token) counts
- Various NLP pre-processing steps to create + filter tokens
- Often normalize word counts

# Term-document matrices II

| Term | Doc 1 | Doc 2 |
|---|---|---|
| accuracy | | 1 |
| alabama | | 1 |
| analysts | 1 | |
| annual | | 1 |
| appropriate | | 1 |
| article | | 1 |
| asset | 1 | |
| average | 1 | |
| ... | ... | ... |
| chief | | 3 |
| china | 4 | |
| chinese | 3 | |
| ... | ... | ... |
| company | 1 | 5 |
| compensation | | 2 |
| complaint | | 3 |
| ... | ... | ... |
| value | 1 | |
| virgil | 1 | |
| vuitton | 2 | |
| watch | 2 | |
| way | | 1 |
| wednesday | | 1 |
| worth | | 1 |



LVMH/luxury stocks: wealth of nations

Fans of upmarket bling are as conspicuous as a Louis Vuitton monogram

APRIL 10, 2018

The best things in life are free. The second best things are very, very expensive. That sentiment — attributed to style icon Coco Chanel — has been taken to heart by brand-conscious fans. Global enthusiasm for upmarket bling is as conspicuous as a Louis Vuitton monogram.

Netflix and executives sued over bonus sc

Shareholder complaint alleges 'rigging' and 'misleading' statements from strea

© Getty

Tom Braithwaite in San Francisco and Tim Bradshaw in Los Angeles 7 HOURS AGO

Netflix and its senior executives have been sued in a shareholder co which alleges the streaming company "rigged" a bonus scheme and i "false and misleading" statements about it.

# Use cases: Simple word lists

**Market sentiment** Counting occurrences of negative words in news (e.g., Tetlock, 2007)

**Deceptive executive speech** Counting occurrences of certain word lists (e.g., Larcker and Zakolyukina, 2012)

**Macroeconomic uncertainty** Count of newspaper articles that contain terms of three word lists measuring uncertainty, the economy, and policy. (Baker et al., 2016)

**Firm-level political risk** Word counts in 10-Ks pertaining to political risk (Hassan et al., 2017)

**Forward-looking statements** Label sentences in an earnings announcement as forward-looking sentences if they include a ”forward-looking” term (Bozanic et al., 2018)

Why Digital Data?

Data Sources

Data Collection

Data Handling

Measurement Basics

NLP Basics

Examples: Word Lists

Examples: Cosine Similarity

Examples: Other

Outlook

# Approach: Cosine similarity of text documents

Three documents:

1. “He races home”
2. “He was racing home”
3. “He has race issues at home”

NLP Pipeline:

1. Tokenize
2. Lemmatize
3. POS
4. Identify collocations
5. Remove stopwords

Use output to create a term-document-frequency matrix and compute the ”angle” between each document vector in that matrix

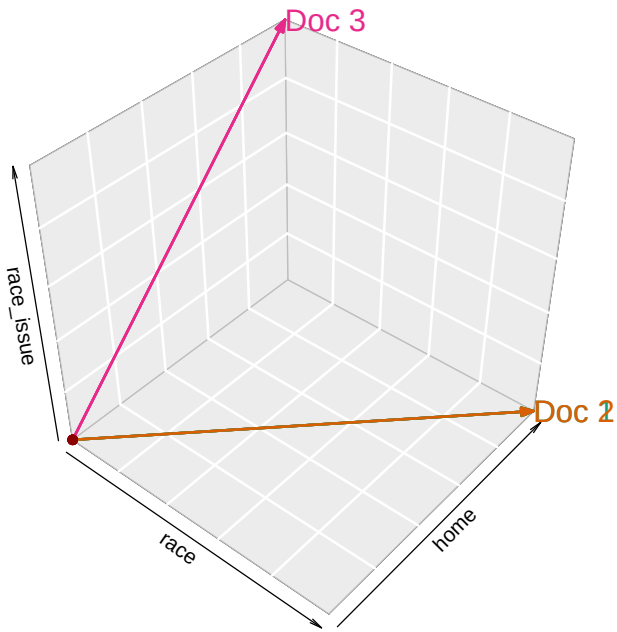
# Approach: Cosine similarity of text documents

1. Tokenize:
   - `(He, races, home)`
   - `(He, was, racing, home)`
   - `(He, has, race, issues, at, home)`
2. Lemmatize
   - `(He, race, home)`
   - `(He, be, race, home)`
   - `(He, be, race, issues, at, home)`
3. POS
   - `(He[PRON], race[VERB], home[ADV])`
   - `(He[PRON], be[VERB], race[VERB], home[ADV])`
   - `(He[PRON], be[VERB], race[NOUN], issues[NOUN], at[ADP], home[NOUN])`
4. Identifying collocations
   - `(He[PRON], race[VERB], home[ADV])`
   - `(He[PRON], be[VERB], race[VERB], home[ADV])`
   - `(He[PRON], be[VERB], race_issues[NOUN], at[ADP], home[NOUN])`

# Approach: Cosine similarity of text documents

5. Remove stopwords:
   - `(race[VERB], home[ADV])`
   - `(race[VERB], home[ADV])`
   - `(race_issues[NOUN], home[NOUN])`
6. Create term-document-frequency matrix

| Term | Doc 1 | Doc 2 | Doc 3 |
|---|---|---|---|
| race | 1 | 1 | 0 |
| home | 1 | 1 | 1 |
| race_issues | 0 | 0 | 1 |




$$cos(v_1, v_2) = \frac{v_1 \cdot v_2}{||v_1|| \cdot ||v_2||}$$

**Finally, straight forward to compute angle between two doc vectors**

# Example: Similarity of business descriptions

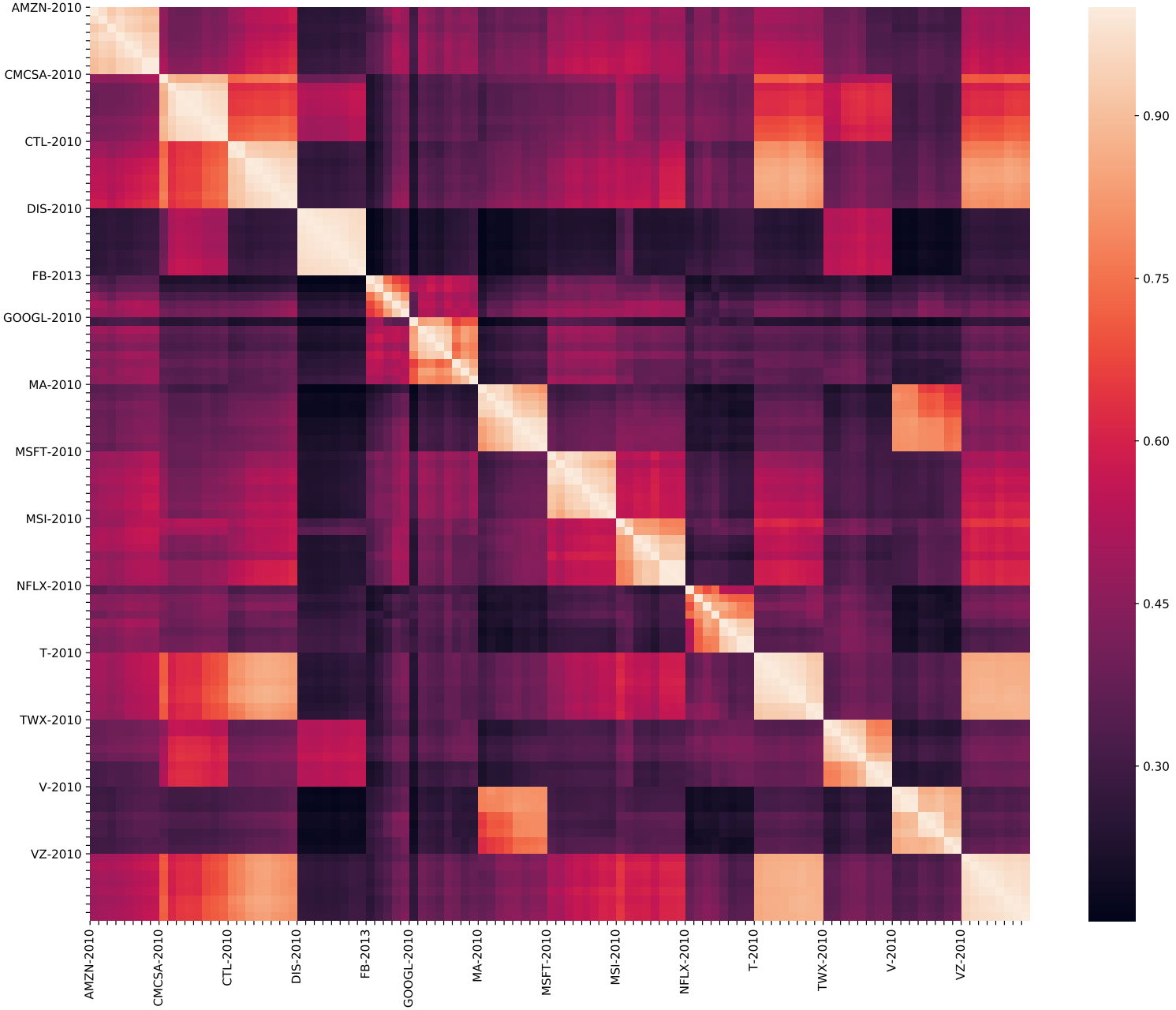
- Item 1 in US 10-Ks: Audited description of a firms business
- Construct the cosine similarity of the following firms and years 2010-2017

| Company | Ticker |
|---|---|
| Amazon | AMZN |
| Comcast | CMCSA |
| CenturyLink | CTL |
| Disney | DIS |
| Facebook | FB |
| Alphabet | GOOGL |
| Mastercard | MA |
| Microsoft | MSFT |
| Motorola | MSI |
| Netflix | NFLX |
| AT&T | T |
| TimeWarner | TWX |
| Visa | V |
| Verizon | VZ |

# Preprocessing

- Lemmatize
- No stopwords or punctuation
- Only nouns, adjectives, and proper nouns
- No digit inside the token
- Drop tokens that occur in less than 1% of the documents
- Compute pairwise similarity based on token count vectors

# Result


AMZN-2010
CMCSA-2010
CTL-2010
DIS-2010
FB-2013
GOOGL-2010
MA-2010
MSFT-2010
MSI-2010
NFLX-2010
T-2010
TWX-2010
V-2010
VZ-2010
0.90
0.75
0.60
0.45
0.30

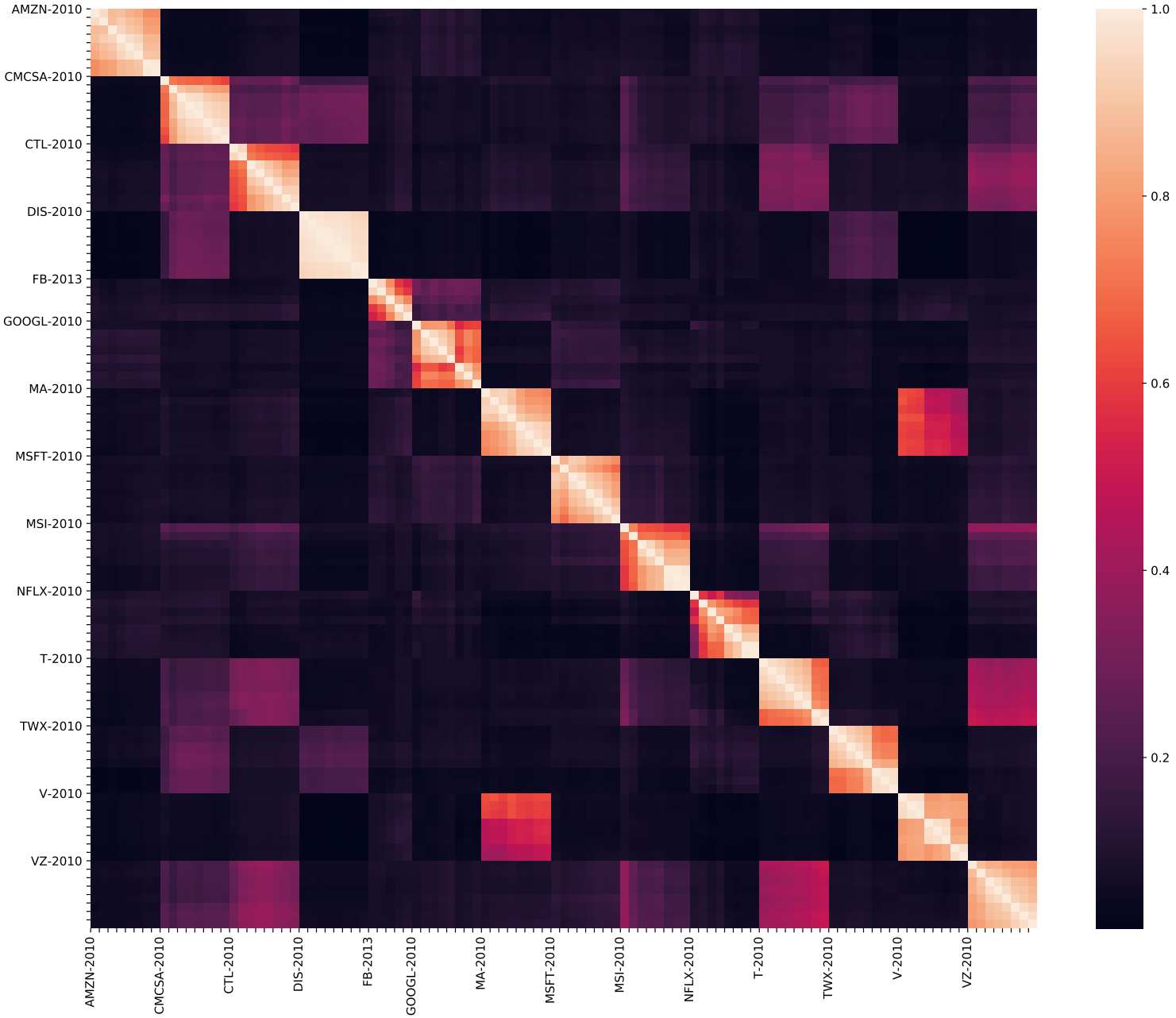
# Alternative preprocessing

- Lemmatize
- No stopwords or punctuation
- Only nouns, adjectives, and proper nouns
- Drop tokens that occur in less than 1% of the documents
- Compute pairwise similarity based on token count vectors

- Lemmatize
- No stopwords or punctuation
- Only nouns, adjectives, and proper nouns
- Drop tokens that occur in less than 5% of the documents
- Drop tokens that occur in more than 95% of the documents
- Compute Bi-grams (e.g ”balance sheet”)
- Weight by *iDF*
- Compute pairwise similarity based on token count vectors

$$idf(t) = log\left(\frac{1 + n_d}{df(d,t)}\right) + 1$$

# Alternative result: Signal to noise is key



# Example use cases: Similarity of text documents

**Financial constraints** Similarity of 10-K capital & liquidity section to firms known to delay their investments due to liquidity issues (Hoberg and Maksimovic, 2015)

**Product market synergies and M&A success** Similarity of 10-K product description (Hoberg and Phillips, 2010) *(Hoberg and Phillips 2010)*

**Product market threats (product market fluidity)** Changes in rivals' 10-K product description that overlaps with own 10-K product description vocabulary (Hoberg and Phillips, 2014)

**News staleness** Similarity between early and later news (Tetlock, 2011)

**Ex-ante litigation risk** IPO's initial prospectus similarity to past sued IPOs (Hanley and Hoberg, 2012)

**Disclosure strategy (intensity)** Revision intensity of IPO prospectuses (amendments) after the initial filing. Time-series of similarity measures (Hanley and Hoberg, 2010)

Why Digital Data?

Data Sources

Data Collection

Data Handling

Measurement Basics

NLP Basics

Examples: Word Lists

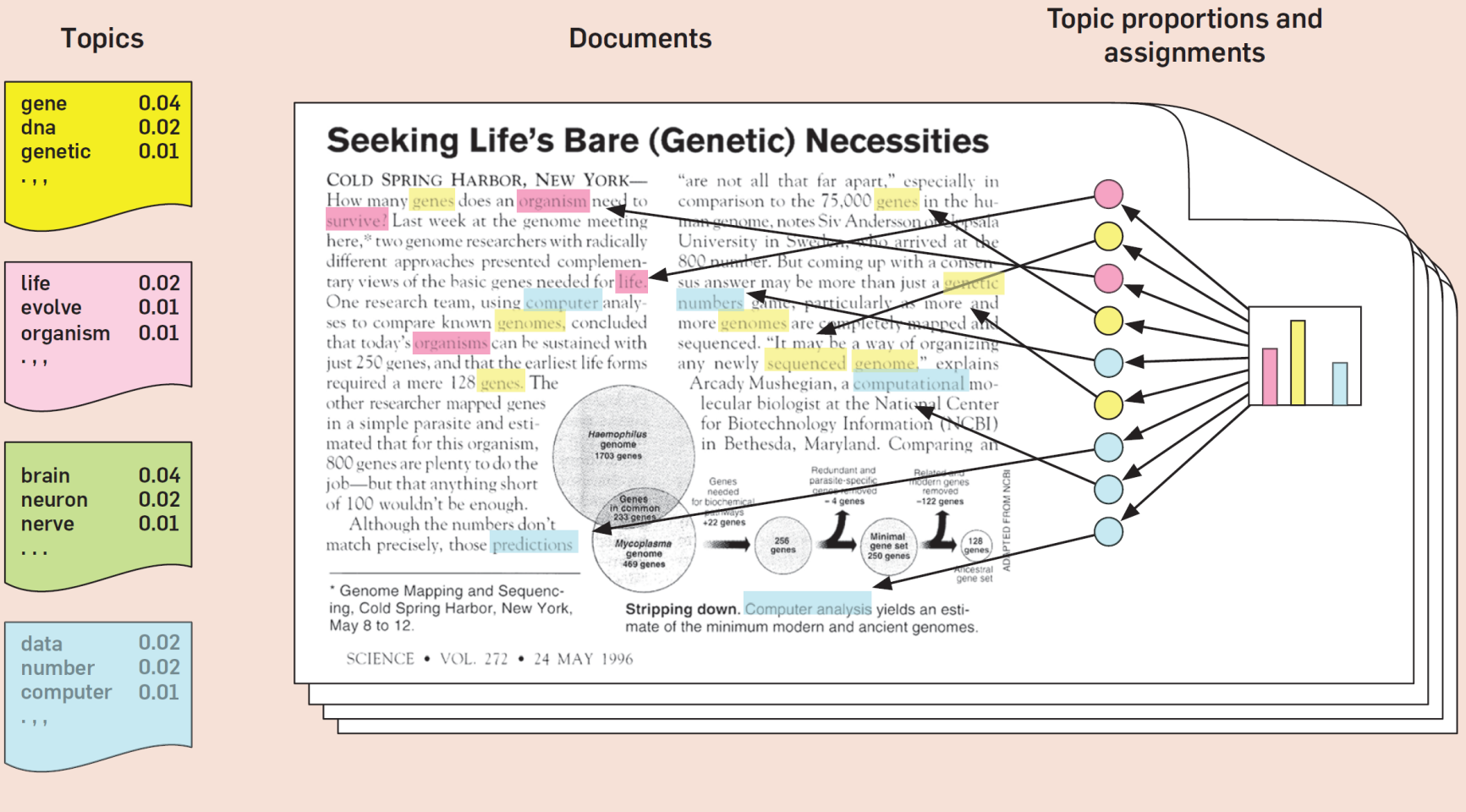
Examples: Cosine Similarity

Examples: Other

Outlook

# Approach: Topic Analysis

**Figure 1. The intuitions behind latent Dirichlet allocation. We assume that some number of "topics," which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.**




Source: Blei, David M. "Probabilistic topic models." Communications of the ACM 55.4 (2012): 77-84.

# Example use cases: Topic Analysis

**Emerging systemic bank risks** Use Latent Dirichlet Allocation (LDA) to produce a unique set of risk topics from 10-K texts of banks in each year (Hanley and Hoberg, 2017)

**Evolution of annual report disclosures** Identify common topics across firms via LDA and finds that 3 of 150 topics – fair value, internal controls, and risk factor disclosures – account for virtually all of disclosure growth over time. (Dyer et al., 2017)

**Analyst information discovery and interpretation** Overlap between topics in firm's earnings announcement topics and analyst report topics. Using LDA (Huang and Thakor, 2013)

# Example use cases: Other measures

**Firm ex-ante integration difficulty** word (non-)occurrence and position in the 10K business descriptions (Hoberg and Phillips, 2017)

**Firm ex-post integration difficulty** word and phrase lists in 10-K MD&A sections (Hoberg and Phillips, 2017)

**Vertical relatedness of firms** Link product vocabularies from the Bureau of Economic Analysis (BEA) Input-Output tables to firms 10-K product descriptions filed with the Securities and Exchange Commission (Frésard et al., 2018)

**Strategic obfuscation vs. complexity** tons of paper using readability measures

# Summary

- Textual analysis: extracting useful data from texts/narratives
- NLP provides preprocessing
- Tools for transforming, filtering, etc. is highly use case specific
- Preprocessing critical for decent signal-to-noise ratio
- Garbage in garbage out: Preprocessing can make or break the application

Why Digital Data?

Data Sources

Data Collection

Data Handling

Measurement Basics

NLP Basics

Examples: Word Lists

Examples: Cosine Similarity

Examples: Other

Outlook

# Current developments

- Constant improvement in NLP preprocessing
- Towards heterogeneous effects models
- Approaches combating measurement error (regularization)
- Penalized Likelihood (e.g., Lasso) or Bayesian adaptive priors

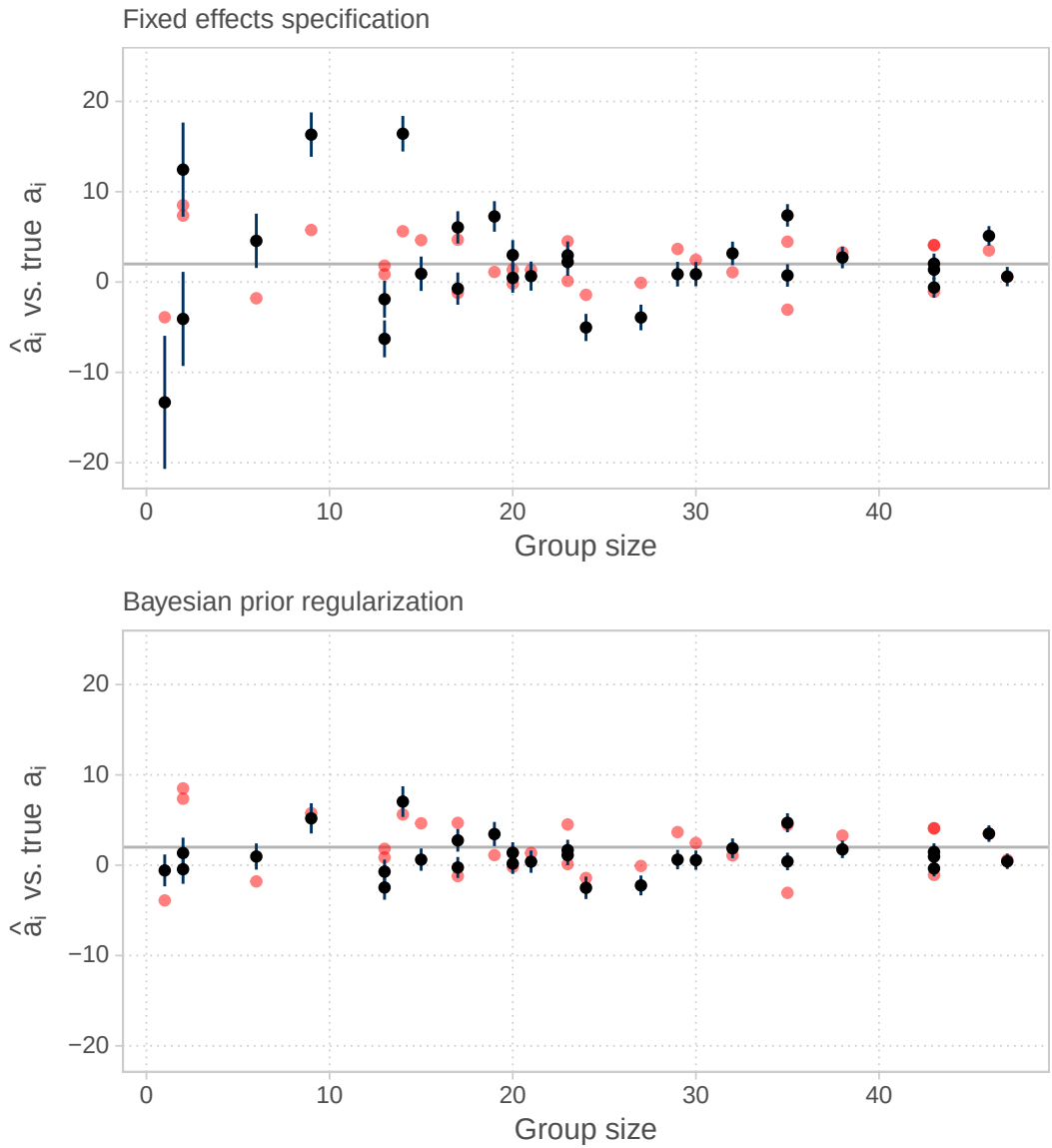
# Heterogeneous effects

$$\begin{aligned}\text{NrNegWords}_{Website,EA} &\sim Binom(N_{words}, p_{neg,WS,EA}) \\ p_{neg,WS,EA} &= \text{RoomInterpretation}_{EA} \times \text{Audience}_{WS} - \text{EconNews}_{EA} \\ &= a_{EA} \times \theta_{WS} - b_{EA}\end{aligned}$$

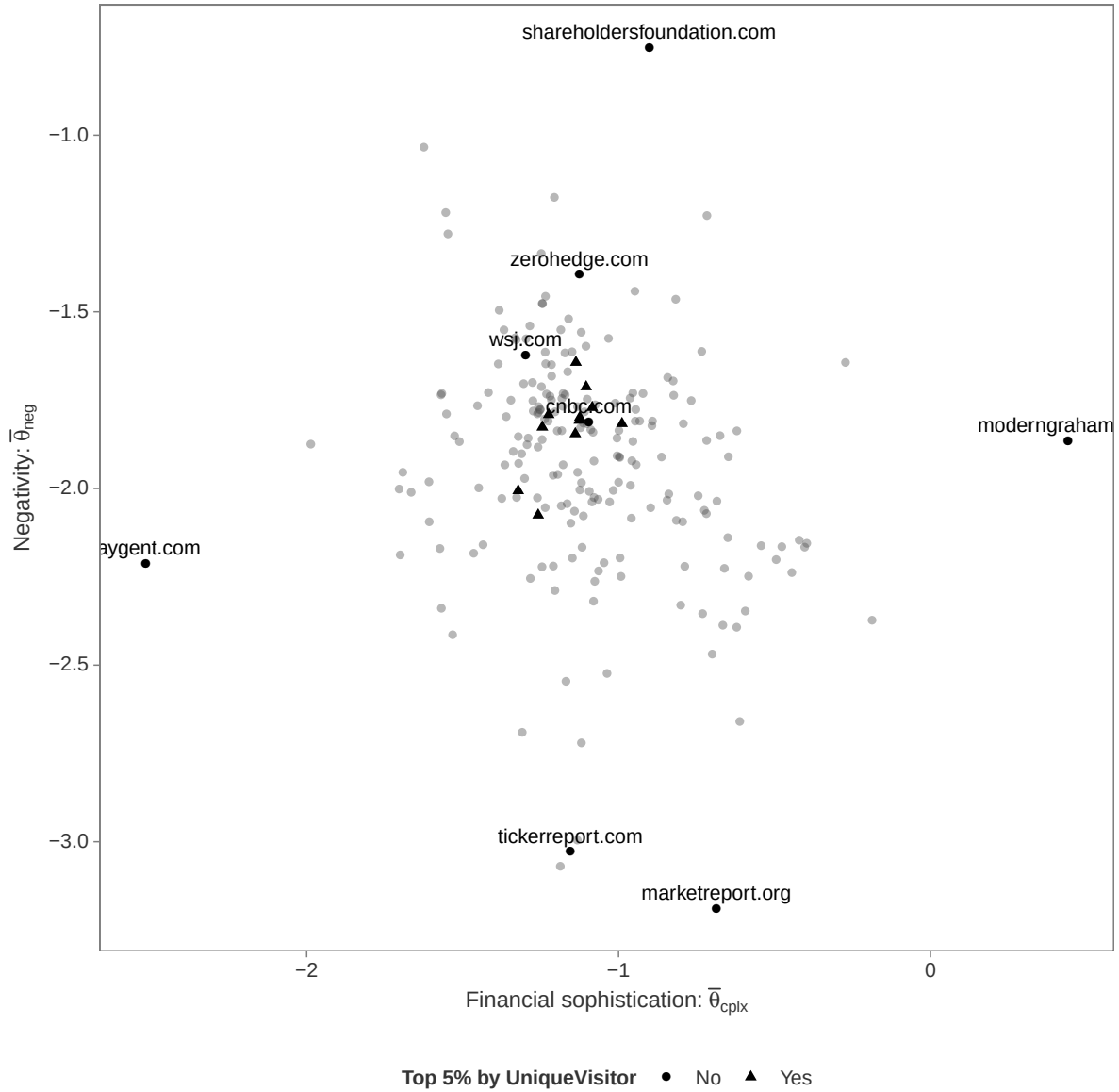
- Measuring news segmentation (Schütt, 2018)
- 212 (2000) Websites, 14,000 Earnings Announcements.
- 212 (2000) + 2 * 14,000 parameters
- Sparse data. Need to regularize in order to not overstate heterogeneity
- Bayesian adaptive regularization to identify and estimate effects

# Heterogeneous effects

- Sparse data. Need to regularize in order to not overstate heterogeneity
- Bayesian adaptive regularization to identify and estimate effects



# Heterogeneous audiences of news site



# Regularizing Example: Has Speech Become More Polarized?

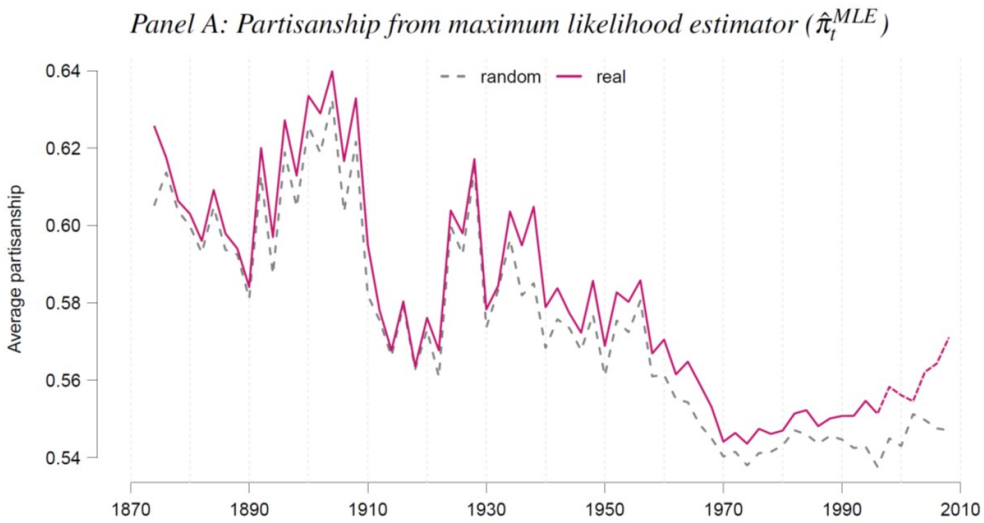
- Gentzkow et al. (2016): Measuring polarization in high-dimensional data: Method and application to congressional speech

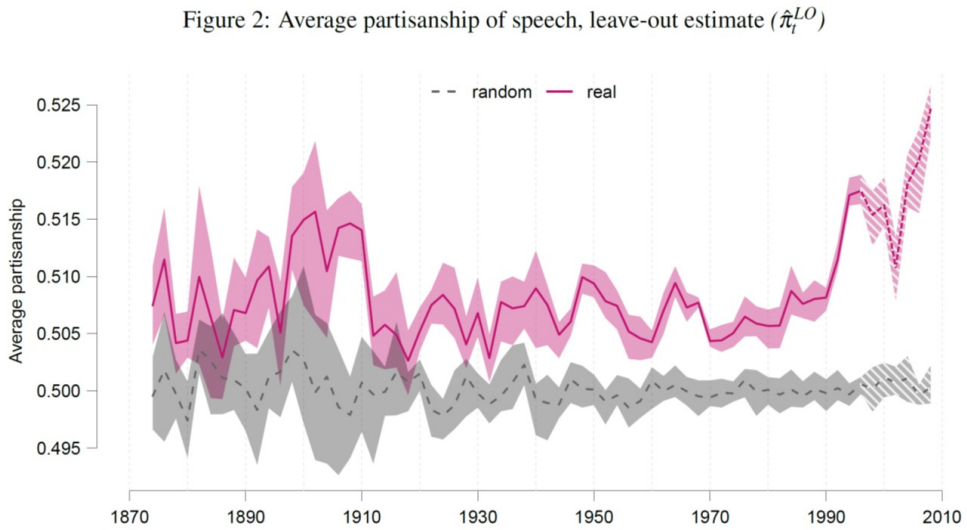
# Regularizing Example: Has Speech Become More Polarized?

- Gentzkow et al. (2016): Measuring polarization in high-dimensional data: Method and application to congressional speech
- “Bias arises because the number of phrases a speaker could choose is large relative to the total amount of speech we observe, meaning many phrases are said mostly by one party or the other purely by chance” (p.3)

# Regularizing Example: Has Speech Become More Polarized?

- Gentzkow et al. (2016): Measuring polarization in high-dimensional data: Method and application to congressional speech
- “Bias arises because the number of phrases a speaker could choose is large relative to the total amount of speech we observe, meaning many phrases are said mostly by one party or the other purely by chance” (p.3)
- Fig. 1 and 2:



# Summary

- Large amounts of data have become much more accessible
- Often text
- Quantifying text poses familiar and new problems (It's all about that Signal-to-Noise ratio)
- Literature amassed a sizable toolkit since 2008
- Increasing awareness of the benefits of regularization

Thank you for your attention

# References I


BAKER, S. R., N. BLOOM, AND S. J. DAVIS (2016): "Measuring economic policy uncertainty," *The Quarterly Journal of Economics*, 131, 1593–1636.

BOZANIC, Z., D. T. ROULSTONE, AND A. VAN BUSKIRK (2018): "Management earnings forecasts and other forward-looking statements," *Journal of Accounting and Economics*, 65, 1–20.

CADMAN, B., M. E. CARTER, AND S. HILLEGEIST (2010): "The incentives of compensation consultants and CEO pay," *Journal of Accounting and Economics*, 49, 263–280.

CADMAN, B. D., R. CARRIZOSA, AND X. PENG (2018): "Inducement Grants, Hiring Announcements and Adverse Selection for New CEOs," .

DYER, T., M. LANG, AND L. STICE-LAWRENCE (2017): "The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation," *Journal of Accounting and Economics*, 64, 221–245.

# References II

ENGELBERG, J. AND C. A. PARSONS (2016): "Worrying about the stock market: Evidence from hospital admissions," *The Journal of Finance*, 71, 1227–1250.

FRÉSARD, L., G. HOBERG, AND G. PHILLIPS (2018): "Innovation and the Incentives for Vertical Acquisitions and Integration," .

GENTZKOW, M., J. M. SHAPIRO, AND M. TADDY (2016): "Measuring polarization in high-dimensional data: Method and application to congressional speech," Tech. rep., National Bureau of Economic Research.

HANLEY, K. W. AND G. HOBERG (2010): "The information content of IPO prospectuses," *Review of Financial Studies*, 23, 2821–2864.

——— (2012): "Litigation risk, strategic disclosure and the underpricing of initial public offerings," *Journal of Financial Economics*, 103, 235–254.

——— (2017): "Dynamic Interpretation of Emerging Risks in the Financial Sector," .

# References III

HASSAN, T. A., S. HOLLANDER, L. VAN LENT, AND A. TAHOUN (2017): "Firm-level political risk: Measurement and effects," Tech. rep., National Bureau of Economic Research.

HOBERG, G. AND V. MAKSIMOVIC (2015): "Redefining Financial Constraints: A Text-Based Analysis," *Review of Financial Studies*, 28, 1312–1352.

HOBERG, G. AND G. PHILLIPS (2010): "Product Market Synergies and Competition in Mergers and Acquisitions: A Text-Based Analysis," *Review of Financial Studies*, 23, 3773–3811.

——— (2014): "Text-Based Industry Momentum," *SSRN Working Paper*.

HOBERG, G. AND G. M. PHILLIPS (2017): "Product Integration and Merger Success," *SSRN Electronic Journal*.

HOITASH, R. AND U. HOITASH (2017): "Measuring accounting reporting complexity with XBRL," *The Accounting Review*, 93, 259–287.

# References IV

HUANG, S. AND A. V. THAKOR (2013): "Investor heterogeneity, investor-management disagreement and share repurchases," *The Review of Financial Studies*, 26, 2453–2491.

KOGAN, L., D. PAPANIKOLAOU, A. SERU, AND N. STOFFMAN (2017): "Technological innovation, resource allocation, and growth," *The Quarterly Journal of Economics*, 132, 665–712.

LARCKER, D. F. AND A. A. ZAKOLYUKINA (2012): "Detecting deceptive discussions in conference calls," *Journal of Accounting Research*, 50, 495–540.

LAWRENCE, A., J. RYANS, E. SUN, AND N. LAPTEV (2018): "Earnings announcement promotions: A Yahoo Finance field experiment," *Journal of Accounting and Economics*, 66, 399–414.

SCHÜTT, H. H. (2018): "Measuring Segmentation in the Financial News Market," Available at SSRN:, available at SSRN:.

SUNDER, J., S. V. SUNDER, AND J. ZHANG (2017): "Pilot CEOs and corporate innovation," *Journal of Financial Economics*, 123, 209–224.

# References V

TETLOCK, P. C. (2007): "Giving content to investor sentiment: The role of media in the stock market," *The Journal of Finance*, 62, 1139–1168.

——— (2011): "All the news that's fit to reprint: Do investors react to stale information?" *Review of Financial Studies*, 24, 1481–1512.