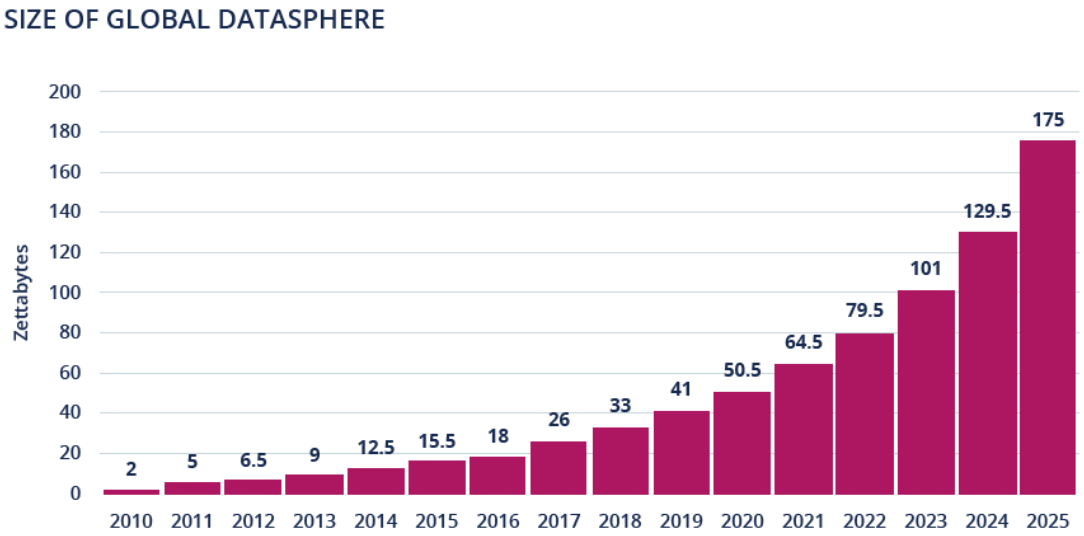
Alternative Data & Sentiment Analyzer

Eshwar Venugopal

What is data?

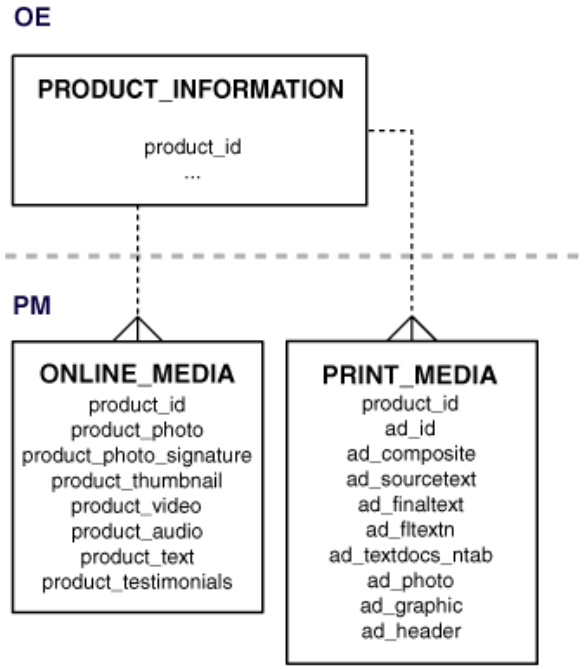
- **Data**: samples of reality recorded as measurements and stored as values.
  - Captures quantitative and qualitative information about units of interest.
  - Stored as structured or unstructured data.
- Data is the new oil. It allows us to
  - Measure the economy
  - Identify new markets
  - Create better credit scores
  - Track our steps, etc.
- Helps us understand our world better
  - Using ground water for irrigation was viewed negatively. A 2022 Science article showed that in Bangladesh pumping water in the summer allowed aquifers to get recharged in rainy season saving 20 trillion gallons of water over 30 years.
  - Finding made possible by missing piece till 2021 – good data!



Source: International Data Corporation

Types of data

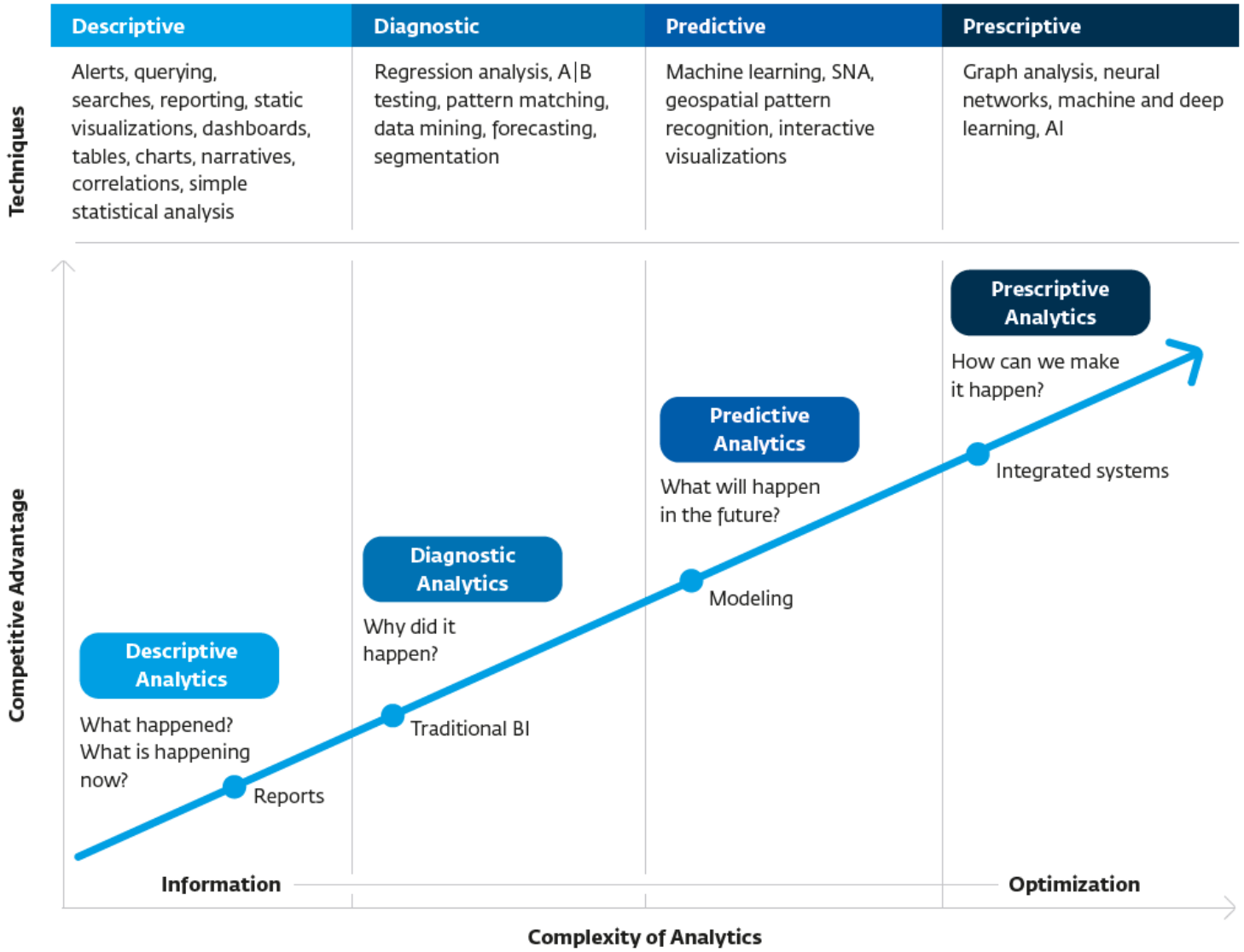
- **Structured data** have a set of attributes and relationships defined at *database* design stage.
  - Have a predetermined organization – schema
  - All elements in the database have the same number of attributes
  - Easy to query and analyze
  - E.g., trades, prices, geo-location, vital stats – age, weight, etc.
- **Unstructured data** are NOT organized by defined schema
  - Flexible, grow in form and shape
  - New attributes may show-up and existing ones missing
  - Could be difficult to query and analyze
  - E.g., emails, social media posts, customer reviews, etc.
- **Semi-structured data** is a combination of the above two
  - Twitter – 280 characters of whatever + images.
  - SEC 10-K filings – have defined sections but each section could be made of text, tables, images
  - Patents – mandatory inventor information and then text/images



Sample scheme

Source: https://docs.oracle.com/database/121/COMSC/diagrams.htm#COMSC00016

Analysis methods

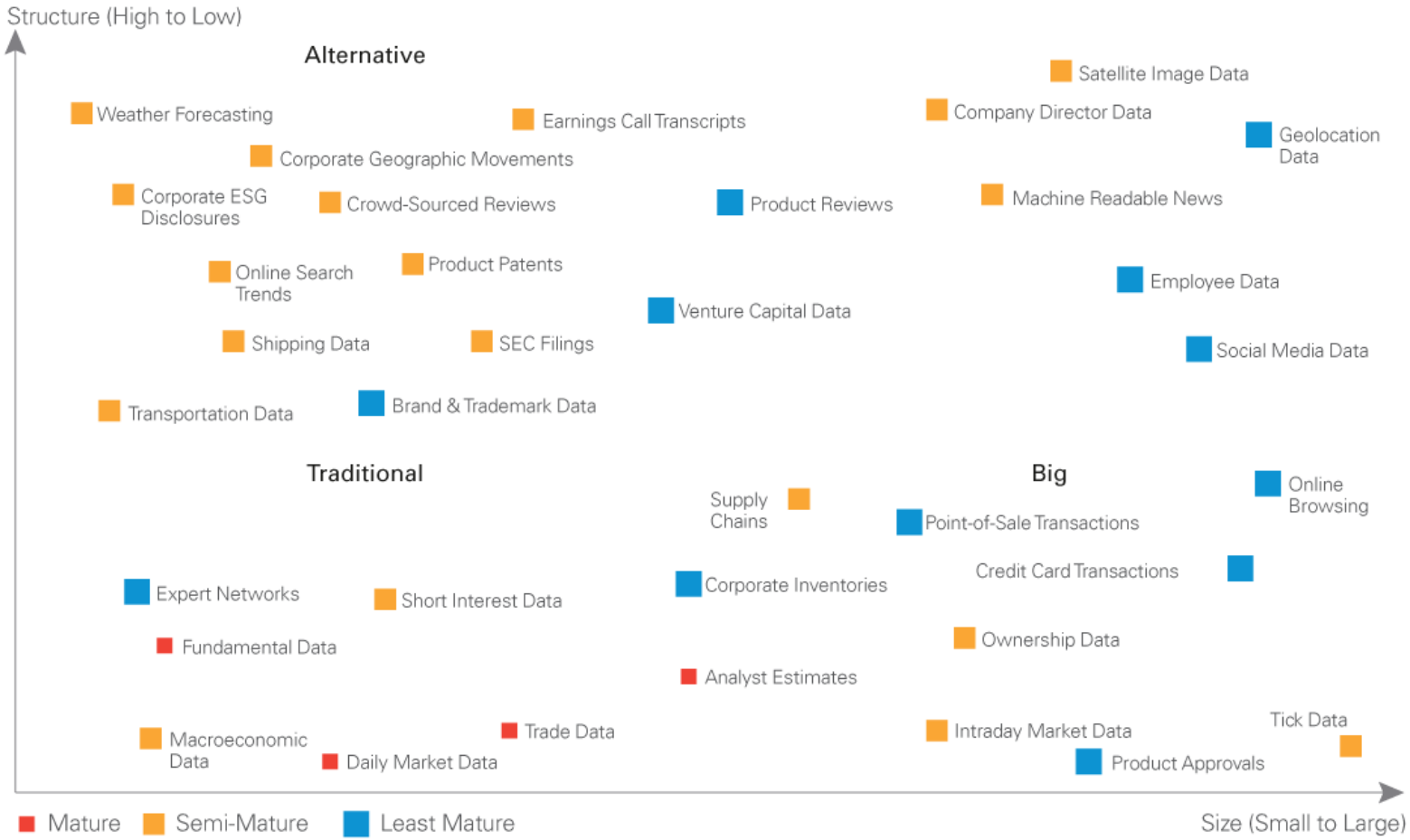


Source: International Finance Corporation, World Bank

Types of data sources

- **Traditional sources**: internal sources that are core to a business or industry
  - *Consumer retail:* products, prices, sales volume, # stores, loyalty points, reviews, etc.
  - *Financial markets:* prices, trading volume, yields, beta, PE, analyst recommendation, filings, third party rating agencies, etc.
  - *Banking:* deposits, clients, transactions, branches, employees, loans, interest rates, third party credit scores, call center recordings, biometric security information, etc.
  - *Telecommunication:* # users, towers, network speeds, complaints, coverage maps, traffic types, etc.
- **Alternative data/non-traditional sources:** structures, semi-structures, unstructured data collected by the firm that typically *was not used* or is sourced from third parties
  - *Consumer retail:* social media engagement, Neilson scanner data, drone survey of parking lots
  - *Financial markets:* news sentiment scores, reddit threads, textual analysis of filings, cloud cover
  - *Banking:* invoices, bills and payment history, geo-spatial survey of customers, social media usage, small business website hits
  - *Telecommunication:* geo-spatial information on users, voice analysis, mobile wallet usage, air-time transfer habits

Types of data sources



As of 30 April 2019. For illustrative purposes only. Data sets represented by red points are examples of the most mature items available, typically first accessible over 40 years ago and ubiquitous in their usage. Data sets represented in orange are less mature, with more recent availability (typically first gathered between 10 and 20 years ago) and requiring a degree of specialized processing. Data sets represented in blue are the least mature in availability, representing the latest developments in data capture and often requiring novel techniques for effective processing.
Source: https://www.lazardassetmanagement.com/uk/en_uk/research-insights/perspectives/the_ubiquity_of_data

Using alternative data

- Blended approach:
  - Creditors blend alternative data with traditional data to create a new score
  - E.g., Lending Club, Prosper, Upstart, etc.

- Second tier approach:
  - Creditors use alternative data only when FICO score is not available or is not satisfactory to obtain credit.
  - E.g., Second Look programs, Sunrise Banks NA

Big data

The new and alternative types of data are increasingly less digestible and voluminous – *big data*.

The V's of big data *[See the Book of Alternative Data by Denev and Amen]*

- Volume (increasing) lots of data
- Variety (increasing) not just numerical data, can be text, image, video etc.
- Velocity (increasing) speed that data is being generated
- Variability (increasing) inconsistencies in the data
- Veracity (decreasing) difficult to tell if accurate (e.g., social media)
- Value (increasing) business value of the data

Advantages of alternative data

- The value alternative data is less likely to erode quickly since not everyone has access to it
- Depending on the quality of analytics, may have higher predictive value
- Possible to extract insights on the economic unit of interest along dimensions that are not captured by traditional sources
  - Hedge funds use retail store parking lot images to formulate trading strategies: https://newsroom.haas.berkeley.edu/how-hedge-funds-use-satellite-images-to-beat-wall-street-and-main-street/
  - Investors used CarMax and Carvana inventory data from Thinknum (platform that tracks public/private companies) to gain insights on the two companies and potential demand for large car manufacturers

Disadvantages of alternative data

- Alternative data is only one piece of the puzzle and has to be combined with other sources.
- Could be extremely difficult to make it analysis ready
  - Unstructured, missing fields
  - May need high level of computing power and technical expertise (e.g., Natural language processing)
- Matching to standard sources or companies may be difficult (e.g., iPhone to Apple)
- Could be expensive (easily upwards of 100K annually)
- Could be difficult to identify the data sources
- Potential legal and privacy issues

Examples: Lenddo

- Philippines does not have a robust credit bureau or national identification system.
- In 2014, less than 10% of the population used traditional banking services
- Loans typically took the form of salary advances.
- Philippines has a large well educated and tech-savvy population.
- Lenddo asked loan applicants permission to access their mobile phone data

Across All Five Social Networks:	7,900+ Total Message Communications:
250+ first-degree connections	250+ first-degree connections
800+ second-degree connections	5,200+ Facebook messages, 1,100+ Facebook likes
2,700+ third-degree connections	400+ Facebook status updates, 600+ Facebook comments
372 photos, 18 videos, 13 groups, 27 interests, 88 links, 18 tweets	250+ emails

- 12,000 data points per user on social network, activity, group memberships, interests
- A new identity verification and scoring system was created – LenddoScore
- Analog identity verification that took 11 days gets completed in 15 seconds.

Examples: Airtel Money

- Airtel Uganda launched Airtel Money is 2012.
- Take-up was only 12.5% among its 7.5 million users.
- Diligent data collection combined with machine learning on 6 months of activity
  - Identified 250,000 ‘high probability’ new clients
  - Showed that 60% of all transfers happened within a 19 km radius of Kampala
  - Airtel Money launched promotions for short-distance P2P transfers around Kampala and expanded campaign to other cities

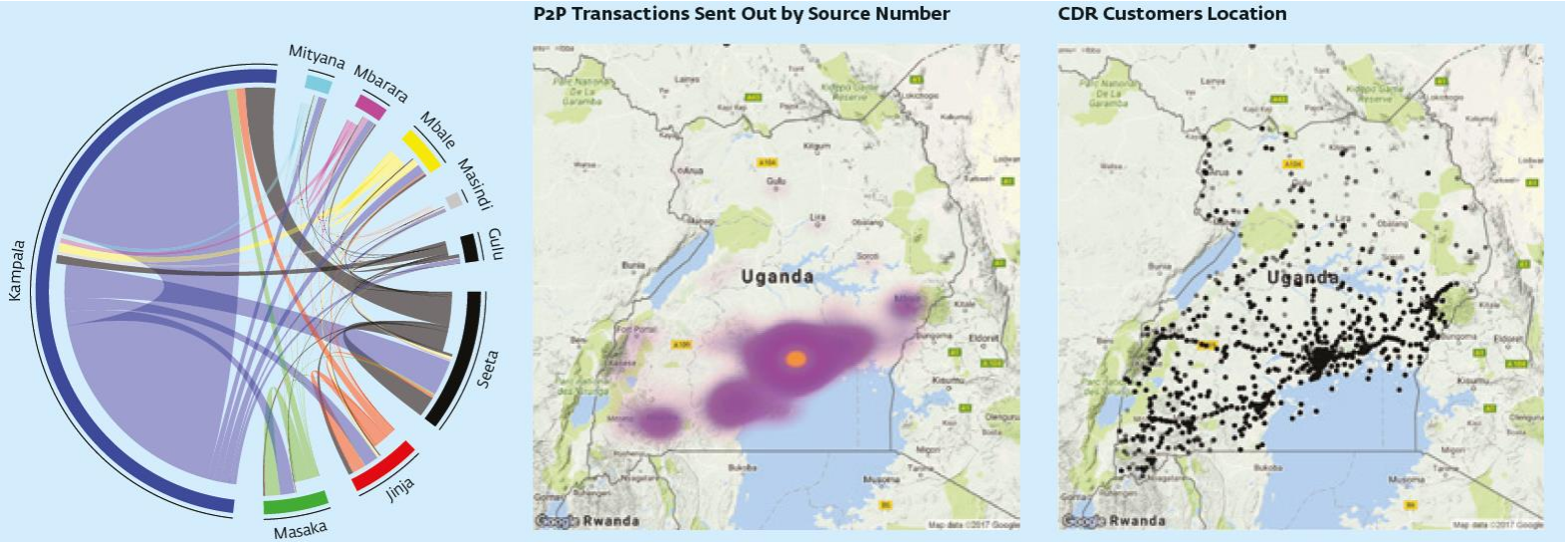


Figure 13: Network analysis (left) of P2P flows between cities and robustness of channel. Also pictured, geospatial density of Airtel Money P2P transactions (center), compared with GSM use distribution (right). Data as of 2014.

Source: International Finance Corporation, World Bank

Measuring innovation using text

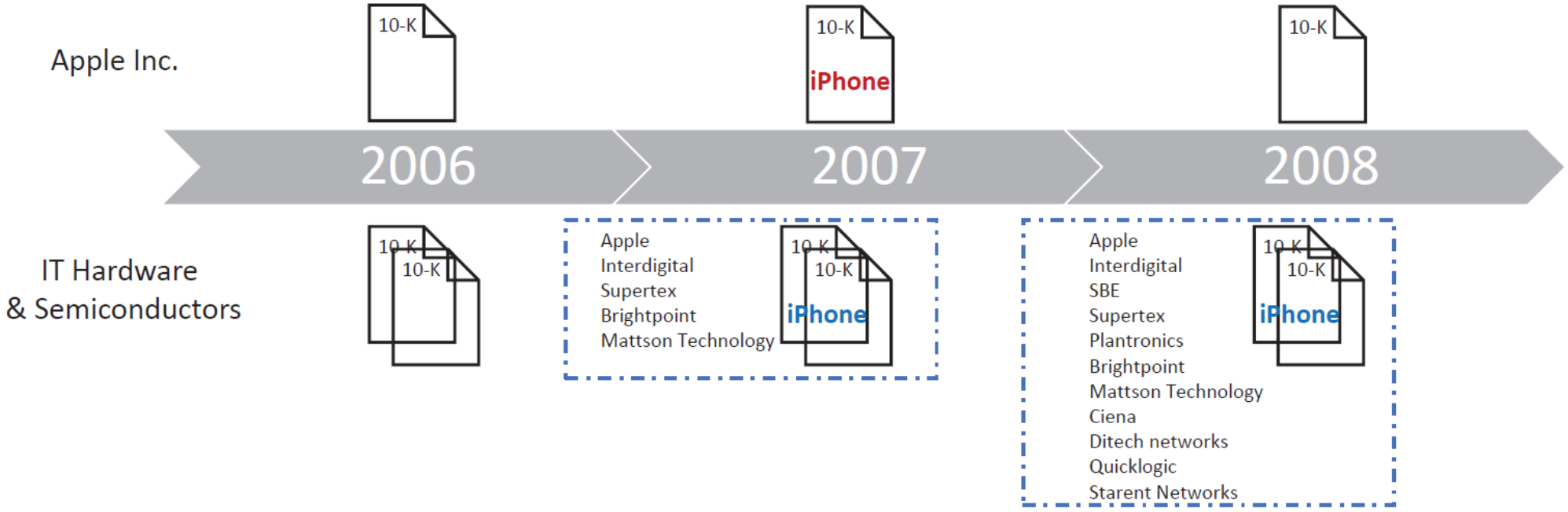
- Corporate innovation is usually measured using patents, R&D expenses.
- But not every firm is able to or willing to patent
  - Finance companies rarely apply for patents.
  - Trade secrets are common (UTSA laws)
  - Transmission Control Protocol/Internet Protocol (TCP/IP) was given away for free by Vincent Cerf and Robert Kahn
- Some ideas cannot be patented

- How to measure innovation across all firms?
  - Common in all innovation is the *creation of something new*.
  - *New words (neologisms) appear in the vocabulary when new innovations appear*
  - E.g., drive-through, Internet, mobile phone

Measuring innovation using text

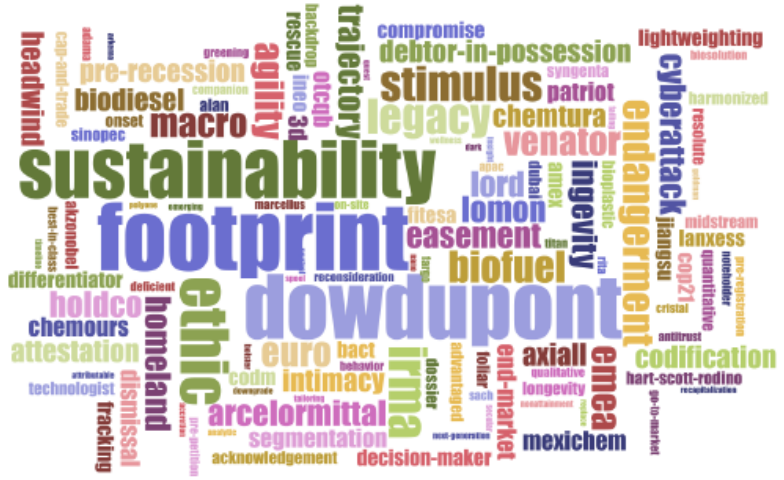
A language-based measure of innovation by Gatchev, Pirinsky & Venugopal (2022) in Research Policy

- Use innovation in language we measure corporate innovation and product creation.
  - Categorize innovations into Product, Process, Market, Supply, and Organization categories.
  - Every firm has an innovation score



Measuring innovation

Panel A: Radical Innovations in IT Software & Services



Panel B: Radical Innovations in Materials

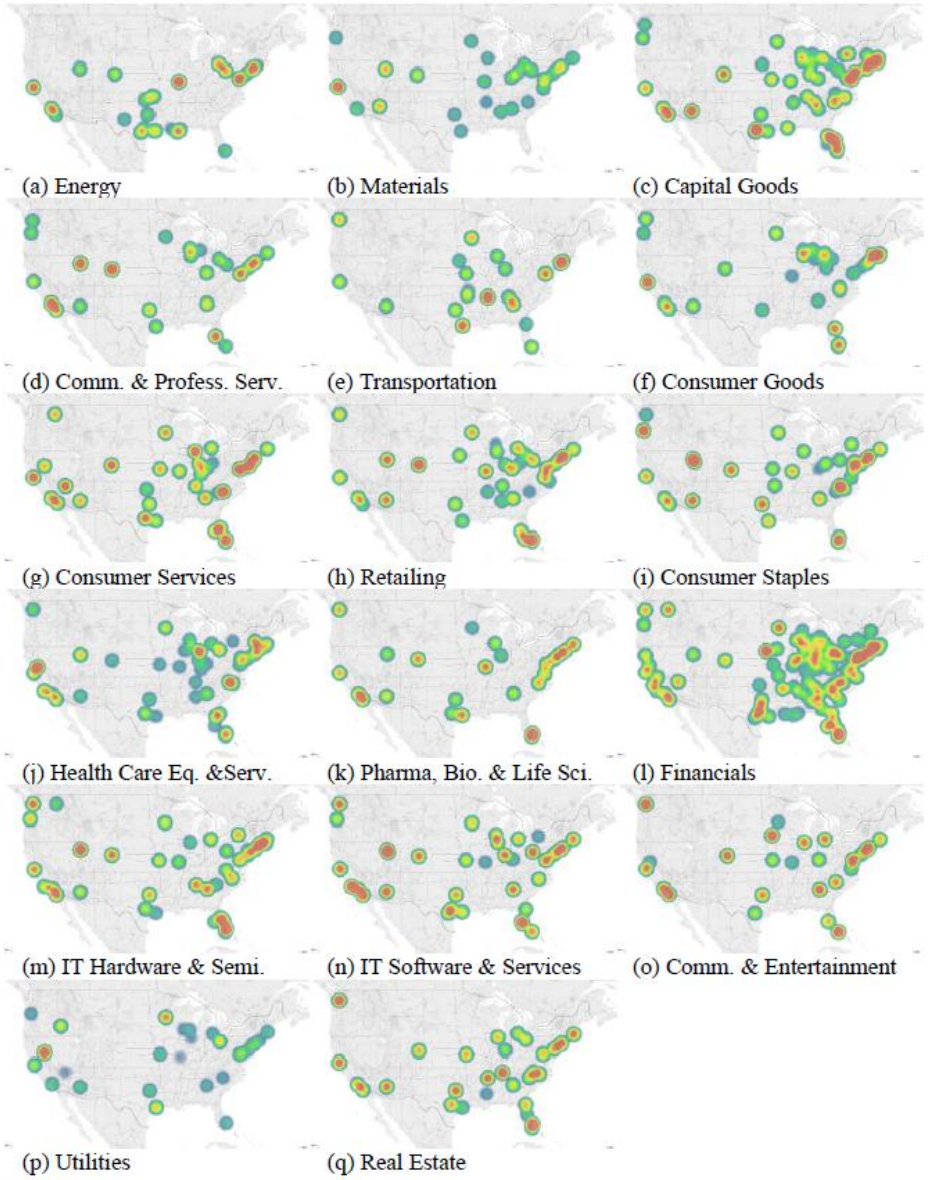


Figure B2. : Geographic Distribution of Innovation by Industry Group.

Source: Gatchev, Pirinsky, Venugopal (2022), Research Policy

Common tools used for collecting alternative data

- Text, images, voice and videos are increasingly used.
- Text is relatively easier to work with.
- Python has a range of text mining tools
  - *Selenium/Requests* are web browser emulator
  - *BeautifulSoup* extracts text from webpages, stripping unnecessary tags
  - *scrapy* is a web scraping crawler
  - Twython is a Python wrapper for Twitter's API to read tweets
  - *search-tweets-python* is a Python wrapper for enterprise Twitter
  - *tabula-py* is a Python wrapper for Tabula (Java), to extract tables from PDF
  - *PDFMiner.six* allows you to extract text from PDF
  - *newspaper* extracts newspaper articles from web

Common tools used for collecting alternative data

- Google Trends - https://trends.google.com/trends/?geo=US
- Common methods used:
  - Word embeddings/bag of words
  - n-gram analysis. See https://books.google.com/ngrams/
  - Word2vec
  - Bidirectional Encoder Representations from Transformers (BERT) and its variations
  - Topic modelling (Latent Dirchlet Allocation – LDA and its variations)

Natural Language Processing (NLP) tools

- Google Natural Language AI - https://cloud.google.com/natural-language
- NLTK - most well-known NLP library for Python
- spaCy - many NLP tasks like extracting entities from text
- textblob - easy to use wrapper for NLTK
- gensim - topic modelling (includes LDA & word2vec)
- Stanford - OpenNLP natural language library
- BERT - TensorFlow code and pre trained models
- Thomson Reuters, Refinitiv, Bloomberg, and many other data vendors provide commercial solutions.

Regulations around alternative data

- Fair Credit Reporting Act (FCRA) and Equal Credit Opportunity Act (ECOA) are among the primary laws that govern the use of data in the credit decision making process.
  - Consumer Finance Protection Bureau (CFPB) and Federal Trade Commission (FTC) and litigation enforce existing laws.
- Under ECOA it is illegal to discriminate against applicants
  - On the basis or race, color, religion, national origin, sex or marital status, or age
  - Because all or part of the applicant's income derives from any public assistance program
  - Because the applicant has in good faith exercised any right under the Consumer Credit Protection Act.
- Regulation B covers creditor activities before, during, or after extension of credit.

- The Federal reserve, Congress, and other regulators in *multiple memos* recognize that alternative data may improve accuracy of credit decisions, reduce information asymmetry and increase inclusion. They also acknowledge the risks alternative data could pose if alternative data is irresponsible used or handled.
- Alternative data, use of AI & ML has led to a "Black box algorithms" situation in which decision process is unclear.

Regulatory concerns with alternative data use

- **Discrimination and Digital Redlining:** The most common concern voiced across government and civil society entities that focus on alternative credit data and AI driven lending was that such programs will continue existing structural roadblocks to credit for marginalized groups, and result in discriminatory outcomes (either intentional or unintentional) that may violate fair lending laws. The overriding concern is that entities will not properly use, calibrate, monitor, or adjust any data sources or algorithms used in new lending platforms to properly control for statistical discrimination (especially that causing a disparate impact on marginalized groups, even through the use of a facially neutral system).
- **Unfair Data Inclusion:** FCRA prohibits certain data elements from being included in consumer reports, and also places requirements on credit reporting agencies (CRAs) to ensure a certain level of accuracy in the data used to create consumer reports. Several commentors and government agencies stated concerns that alternative data may be collected and used outside of FCRA's requirements. For example, social media data (e.g., friend groups and educational institutions) as a proxy for credit worthiness was seen as less reliable and fair compared to the inclusion of rent or utility payments into a credit decision.
- **Improper Additional Data Uses:** A variety of sources noted for the potential misuse of alternative credit data outside of the underwriting context. Because alternative credit data can include a range of data points, including Internet activity, social media, and other data typically associated with digital advertising and other uses of such information, some entities are worried that consumers may grant access to information to receive credit but that the same information will be repurposed for other uses without the consumer's knowledge.
- **Deceptive Terms and Conditions:** The sources we reviewed noted that, because consumers may not be aware of the types of data or activity that has bearing on their credit decisions when a company uses alternative data and AI to make such decisions, they may be unable to address or challenge adverse terms or changes in a credit offer.

(Source: 2022 Venable LLP.)

What do you think about Transunion's use of Zip codes to judge applicants? [recollect last lecture]

Regulatory actions

- Innovation in data and computing have improved credit access but because of the opaqueness of algorithms it is unclear what are the magnitudes of benefits and harm.
- Regulators have taken multiple actions in the recent past
  - CFPB's 2017 Request for Information regarding use of alternative data and modeling techniques in the credit process
  - CFPB's 2020 Adverse action tech sprint
  - DOJ Task Force combatting redlining initiative – American Trustmark National Bank settlement (Oct 2021): Consent order creates $3.85M loan subsidy program for majority Black and Hispanic neighborhoods in Memphis, open new lending office, $5M civil penalty
  - CFPB published Consumer Financial Protection Circular 2022-2023 and reiterated that companies have to explain the specific reason for rejecting applications or other adverse action even if creditors reoky on models using complex algorithms.

Exercise

Suppose an algorithm flags the following variables as indicators with high loan default predictability. Do you think any of them run afoul of federal laws on data use for credit decision making?

- Zip codes
- Time spent on social media
- Typos and grammatical errors

Sentiment Analysis – Part 2

Sentiment Analysis

THE JOURNAL OF FINANCE • VOL. LXII, NO. 3 • JUNE 2007

Giving Content to Investor Sentiment: The Role of Media in the Stock Market

PAUL C. TETLOCK*

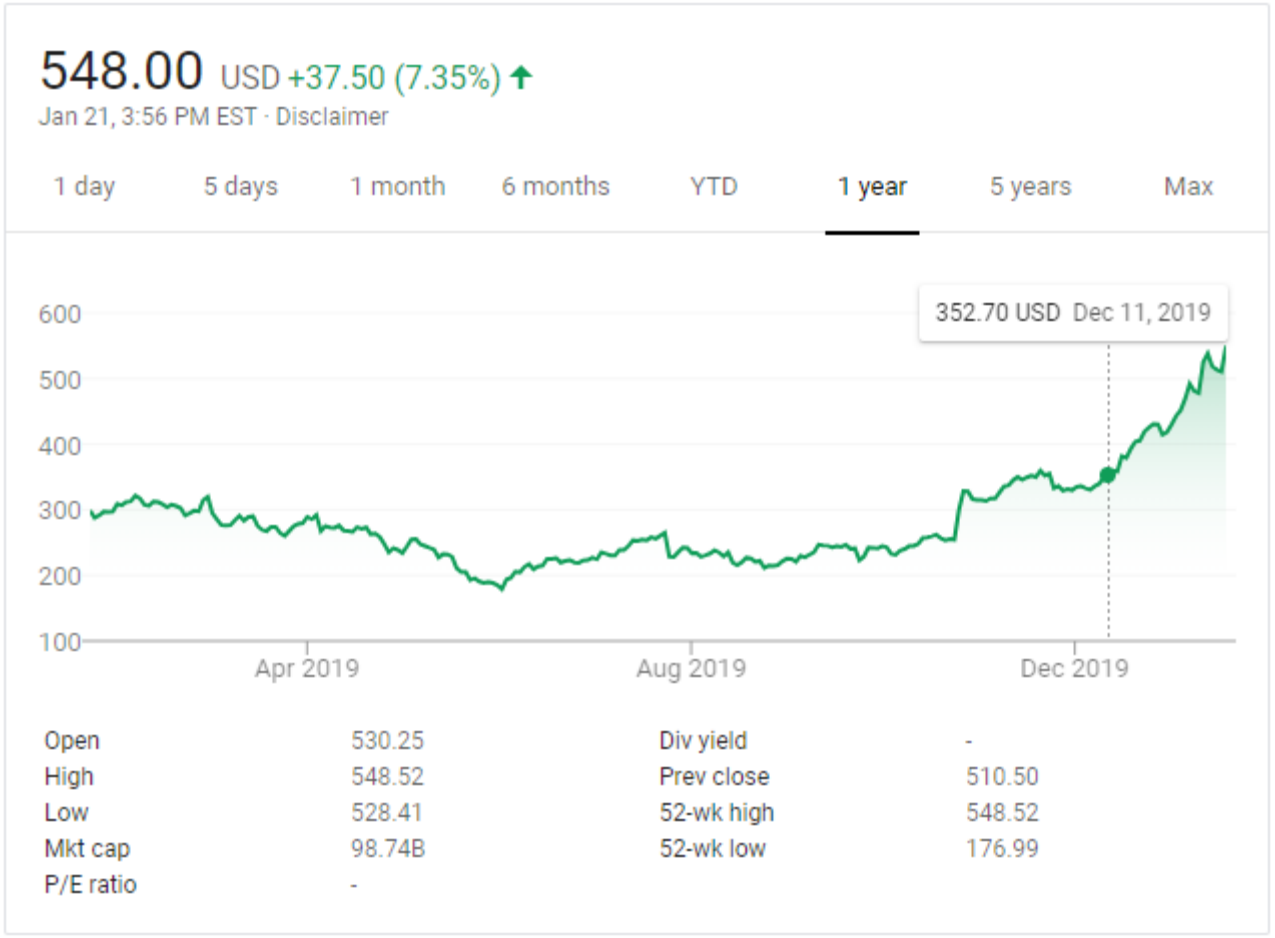
ABSTRACT

I quantitatively measure the interactions between the media and the stock market using daily content from a popular *Wall Street Journal* column. I find that high media pessimism predicts downward pressure on market prices followed by a reversion to fundamentals, and unusually high or low pessimism predicts high market trading volume. These and similar results are consistent with theoretical models of noise and liquidity traders, and are inconsistent with theories of media content as a proxy for new information about fundamental asset values, as a proxy for market volatility, or as a sideshow with no relationship to asset markets.

Sentiment Analysis: Jim Cramer & Tesla

Watch the full video at https://www.youtube.com/watch?v=PXwilboQWvs&t=317s&ab_channel=CNBCTelevision

Sentiment Analysis: Jim Cramer & Tesla



Sentiment Analysis: Social Media

.7AM

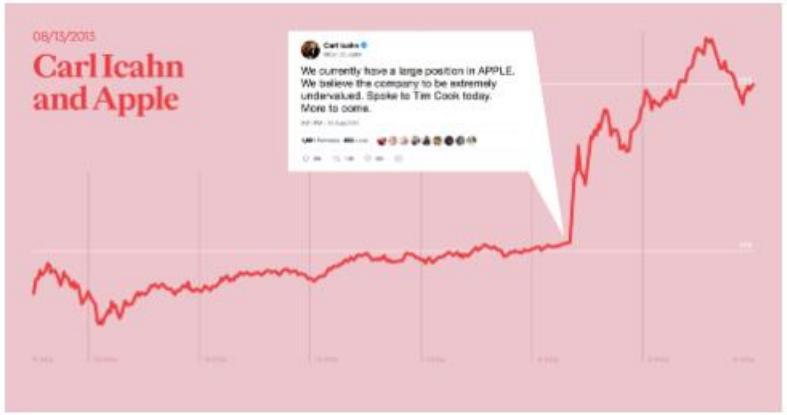
Stefan Nann is co-founder and CEO of data analytics company StockPulse, a data analytics company specializing in mining Emotional Data Intelligence.

"Facts only account for 10% of the reactions on the stock market; everything else is psychology." André Kostolany, a stock market investor who made most of his fortune during the reconstruction of Europe after World War II, made this observation. Renowned for his shrewd and astute mixture of psychology and his sensible knowledge of stocks and markets, Kostolany became one of the most successful investors of the 20th century.

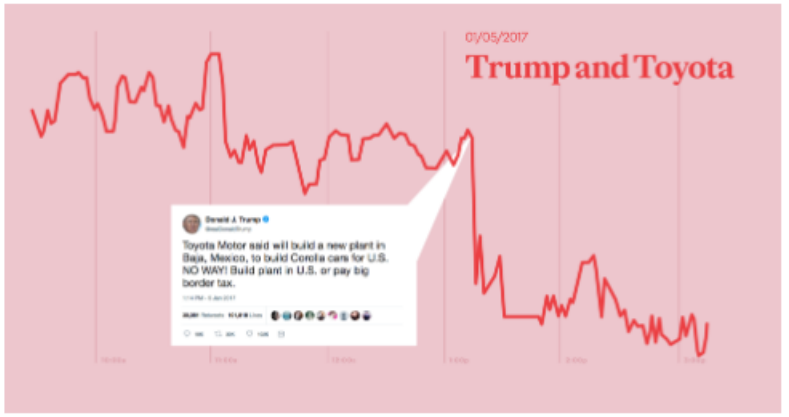
Source: https://www.nasdaq.com/articles/how-does-social-media-influence-financial-markets-2019-10-14

Sentiment Analysis: Social Media

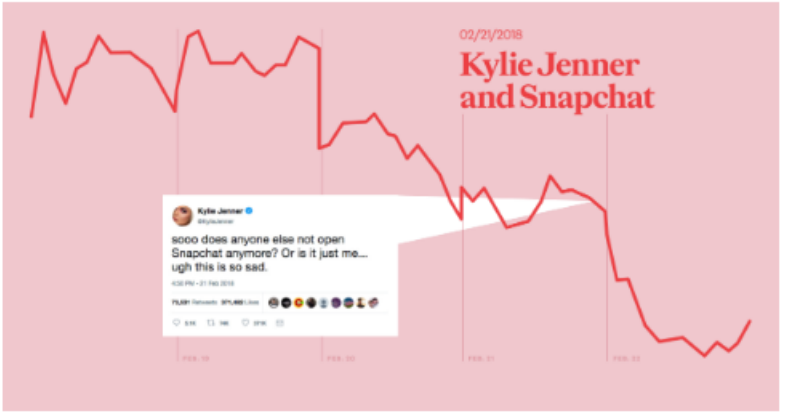
4 - Billionaire Carl Icahn, of Icahn Enterprises L.P. ($IEP), shared his position on Apple ($AAPL) on August 13, 2013, via Twitter, saying that the company is undervalued. Seconds later, Apple's stock spiked. Minutes later Apple gained $17 billion in market cap.



9 - About a month later, in January of 2017, Trump then went after Toyota ($TM), which led to a $1.2 billion decrease in the carmaker's value.



3 - Kylie Jenner made headlines in early 2018 by firing off one single tweet that set off a downward spiral for Snap's stock ($SNAP). A headline from CNN Money read: "Snapchat stock loses $1.3 billion after Kylie Jenner tweet."



Sentiment Analysis: Exercise

- Key people can have an impact on a stock or the economy.
- For example, European Central Bank President's speeches, FOMC meetings, etc. have a palpable impact.
- The 45th U.S. President influenced Regeneron's (and many other) stock prices.

- See: *Loughran & McDonald (2011), When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks, Journal of Finance, 66:1, 35-65.*
- Word lists are available in WebCourses.
- Figure out the following:
  - ***Number of positive and negative*** words.
  - Create positive and negative sentiment scores
  - Create a final sentiment score