Big Data in Accounting Research

Maclean Gaulin
University of Utah

Overview

I. What accounting research has done so far.
   - History of data and techniques
   - Current practices

II. Where can accounting research go from here?
   - Adapting external techniques to Accounting domain
   - Machine learning for accounting questions

Little ~~Big~~ Data in Accounting Research

Ball and Brown, 1968

- Earnings are related to stock returns
- Data: 1946 – 1966
- 261 firms

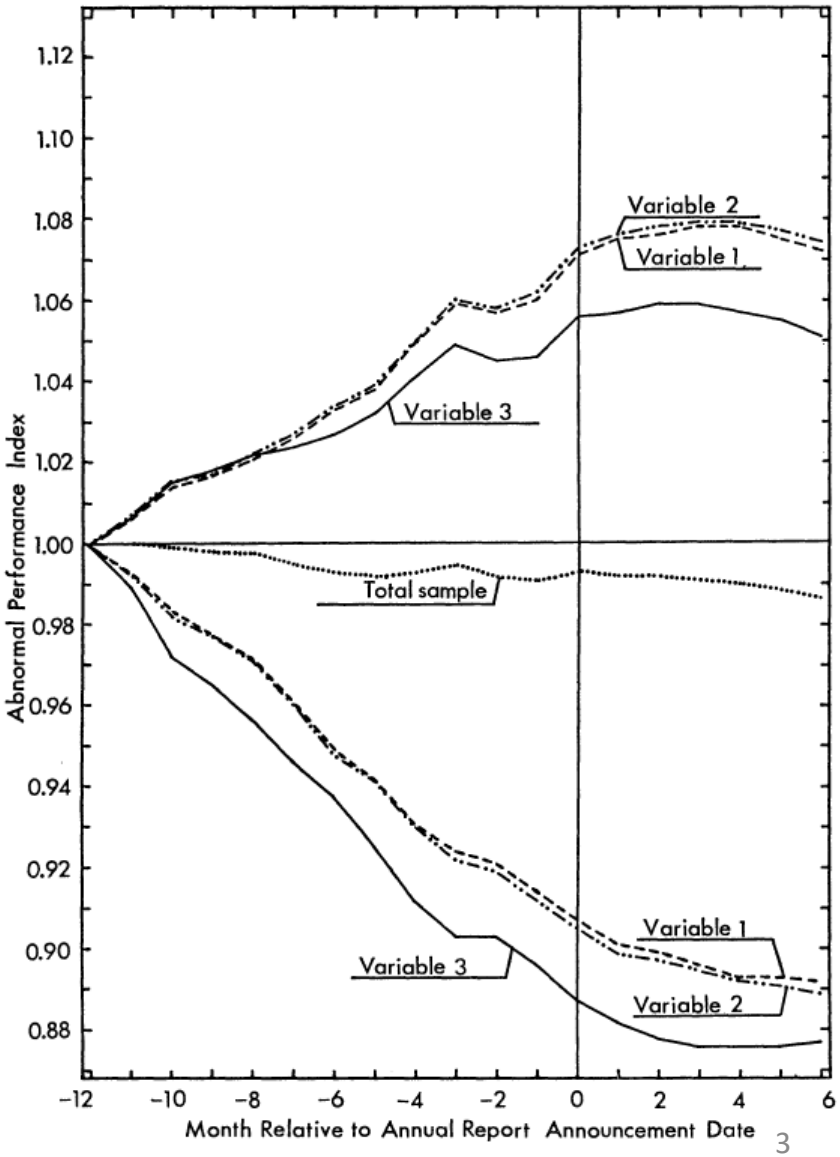


FIG. 1 Abnormal Performance Indexes for Various Portfolios

2008: Bigger Data

Feng Li analyzes textual data from 55,719 annual reports

- FOG Score = (words / sentence) + % complex words
- Length = word count
- Bag of Word (BOW) frequencies for: self-referential words, exclusive words, causation words, positive emotion words, and future tense verbs

Poor performing firms' annual reports are longer and harder to read

Big Data and Machine Learning

Big Data

- K >> N

Features Observations

- Computational complexity
- Data growth

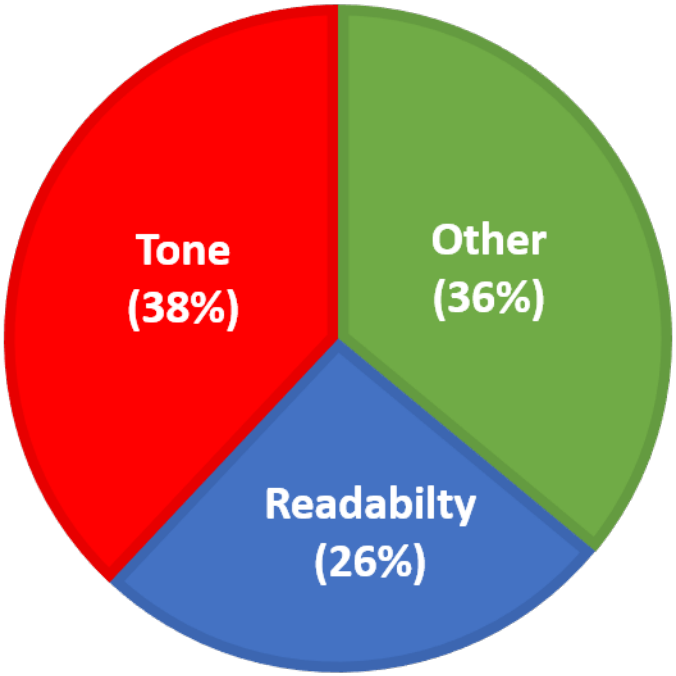
Machine Learning

- Complex modeling
- Myriad model inputs (K, above)
- Supervised / Unsupervised

What is big data in Accounting?

Textual analysis

- Focused on: length, tone, readability (FOG)
- Typically using: word lists



El-Haj, Rayson, Walker, Young, and Simaki, 2018, *Working Paper*

Measures: Tone

Harvard linguistic dictionary negative words

- Example: tax, cost, board, liability, depreciation, gross

Ignores negation

- failed to achieve, overestimated earnings

Does not take context into account

- mitigating bad news with positive tone

Measures: Readability

Readability measured by 'FOG score'

= 0.4 x [(words / sentence) + % complex words]

- corresponds to expected 'grade level'

- Examples:
  - Green Eggs and Ham: 3.24
  - Lord of the Flies: 5.6
  - Harry Potter: 7.2
  - War and Peace: 9.4

  - Average 10-K: 19.4
    - Complex words: competition, accounting, financing, equity, etc.
  - Feng Li's introduction: 19.6

Adapting to Accounting domain

Off the shelf FOG/Tone not necessarily well suited to Business

Targeting data and techniques to specific question

- Can we gain insight by studying disclosures *as conveyed?*
  - What information is being disclosed?
  - How is this information disclosed?

- Setting: Risk factor disclosures from US firms

Risk Factor setting

Unique disclosures: mandatory, no content guidance, purely negative information, non-narrative, no mitigating discussion

Regulation: discussion of the **most significant factors** that make the offering speculative or risky. This discussion must be concise and organized logically. […] Explain how the risk affects the issuer or the securities being offered. *Set forth each risk factor under a subcaption that adequately describes the risk.*

AMD Risk Factor

Intel Corporation's dominance of the microprocessor market and its aggressive business practices may limit our ability to compete effectively.

Intel Corporation has been the market share leader for microprocessors for many years. Intel's market share, margins and significant financial resources enable it to market its products aggressively, to target our customers and our channel partners with special incentives and to influence customers who do business with us.

Intel Risk Factor

We face significant competition.

The industry in which we operate is highly competitive and subject to rapid technological and market developments, changes in industry standards, changes in customer needs, and frequent product introductions and improvements. If we do not anticipate and respond to these developments, our competitive position may weaken, and our products or technologies might be uncompetitive or obsolete.

Measure Risk Factors as disclosed

Individual risk factor as the unit of measure

- Disaggregate disclosure into individual risk factors rather than focus on aggregate risk disclosure

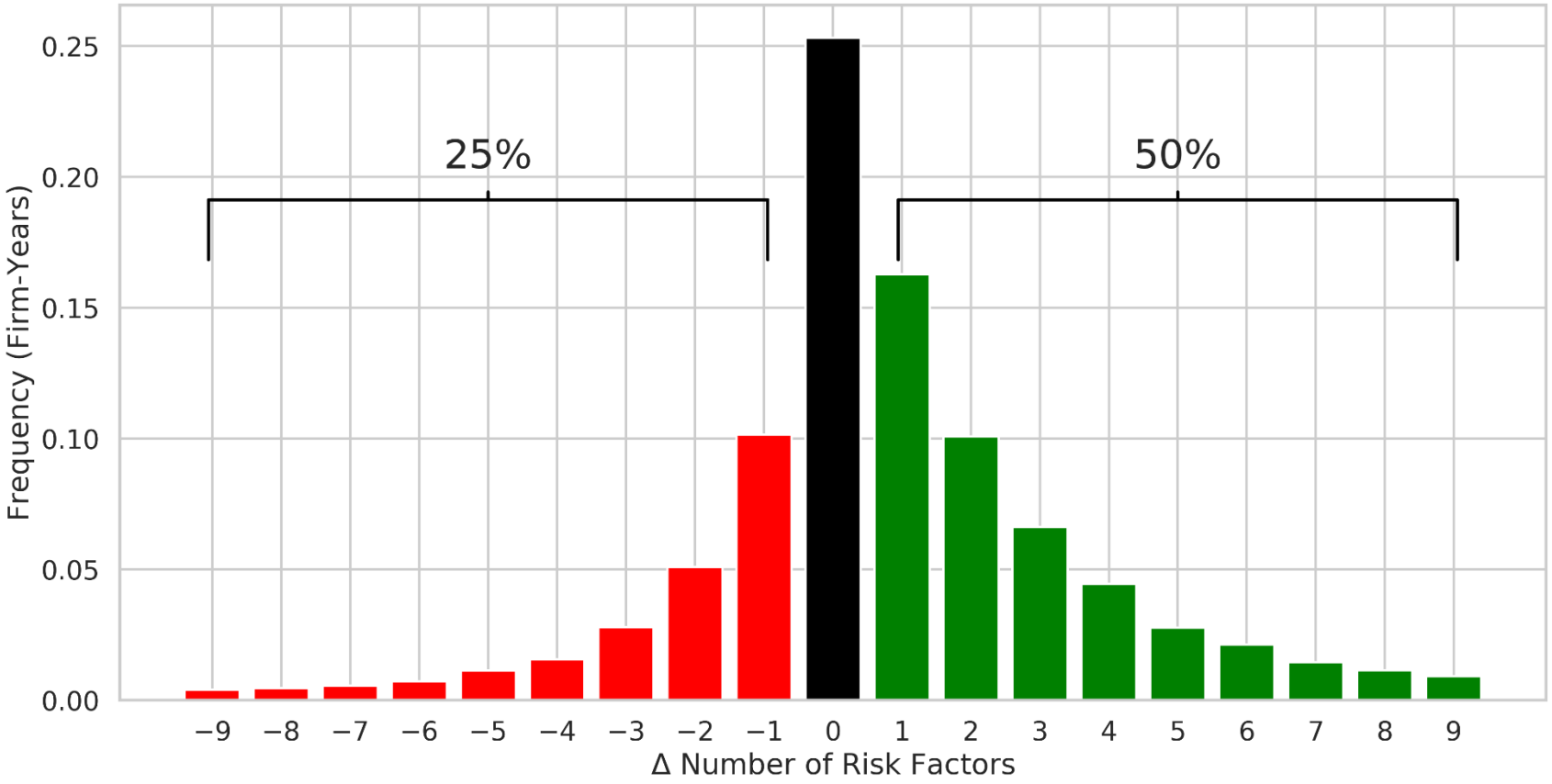
Measures dimensionality of risks as disclosed

- Level of aggregation not chosen by researcher
- Maintains original set of disclosure cost-benefit tradeoffs

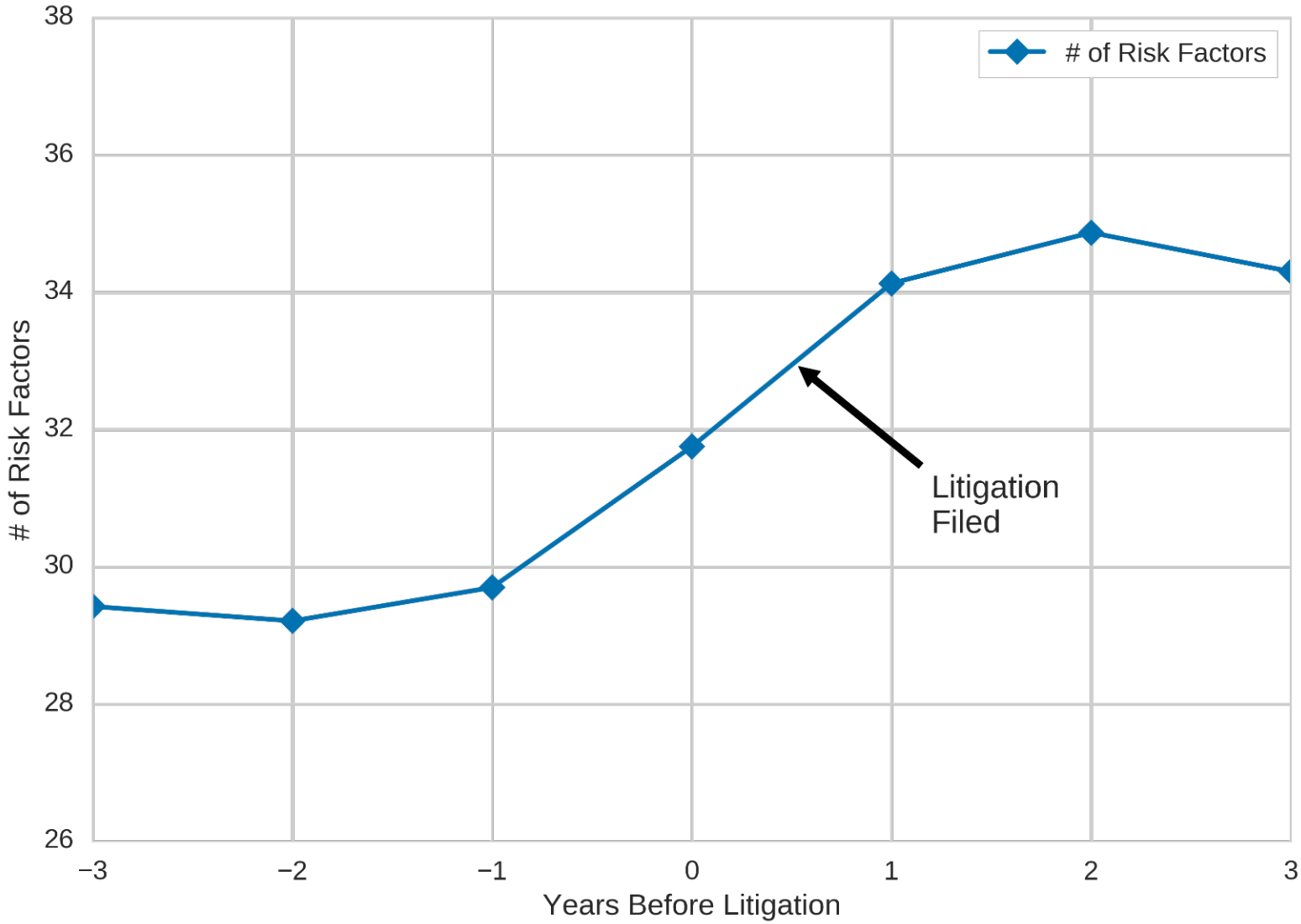
Allows observation over time

- Observe decisions to add/remove risk factors as the underlying conditions evolve

Risk Factor changes



Risk Factors before litigation



Why do managers disclose Risk Factors?

- Managers warn of specific adverse outcomes
  - Net loss, operating loss, sales declines, general lawsuits, securities litigation
  - Managers *remove* stale risk factors when adverse outcomes less likely
  - Disclose specific risks in advance of specific outcomes

- Respond to investors with wider set of risk factors
- Respond to regulators with more definitive disclosures

Why big data?

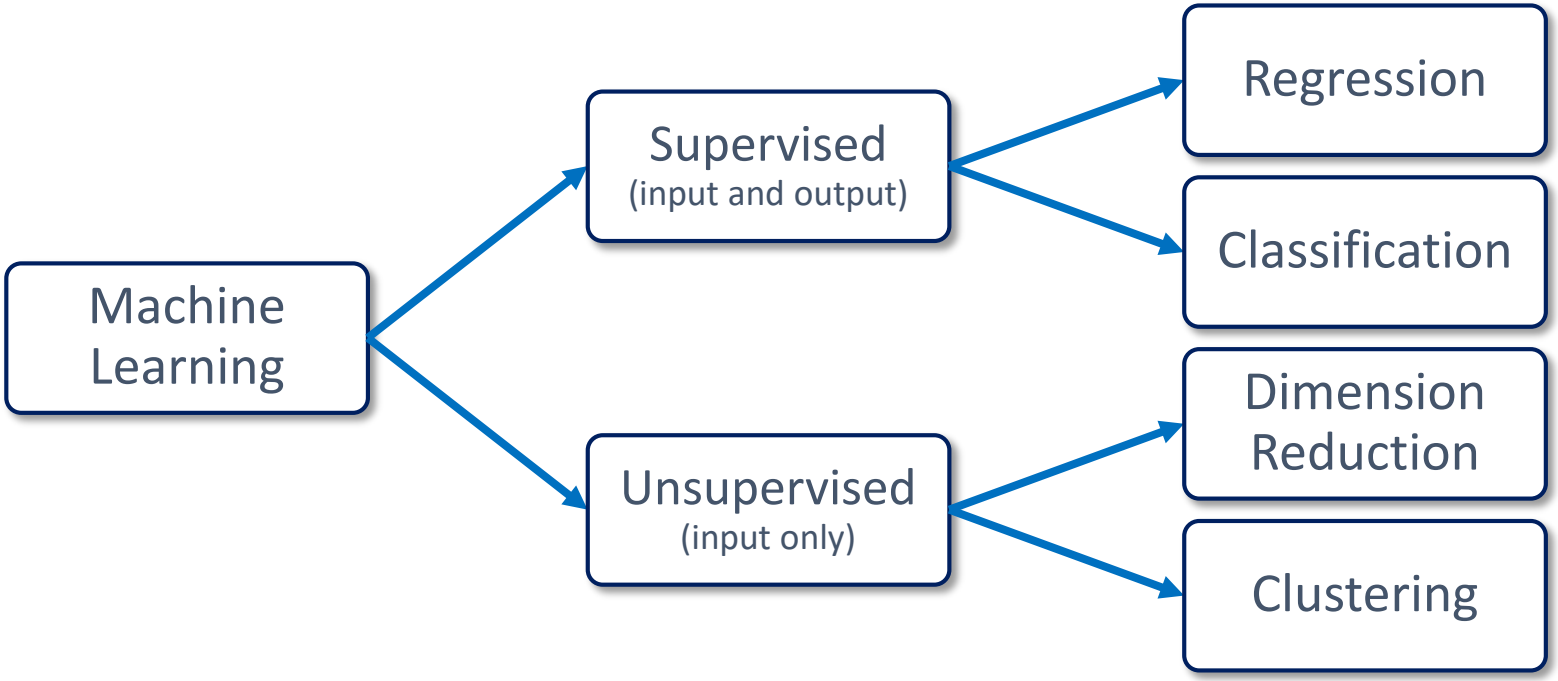
What unique insight does big data facilitate?

- Full picture of information disclosed
- Access to unstructured data – text data
- Access to new data sources

What does machine learning allow?

- Models with inputs from myriad sources
- Discovering what 'matters' – feature selection
- Complex interactions within and between firms

Machine learning about disclosure

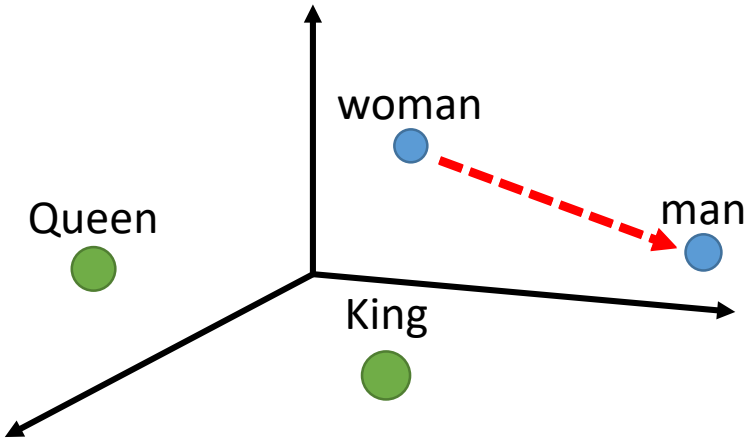


- Supervised (labeled) vs Unsupervised (unlabeled)
- Most accounting data is unlabeled
  - Traditional statistical approaches don't apply
- Networks and cluster analysis

Example: word embeddings

GLOVE/Word2Vec/SkipThought

- 2 layer neural network to predict next word in sentence
- Learns language 'structure' and relationships



King – man + woman = Queen

Example: word embeddings

GLOVE/Word2Vec/SkipThought

- 2 layer neural network to predict next word in sentence
- Learns language 'structure' and relationships
- Pre-trained models available

- Examples:
  - King – man + woman = Queen
  - Motorcycle – engine = bicycle
  - Canada – Loonie + Dollar = United States
  - United States + Tim Hortons = Canada
  - Donut + Donut = Donuts

Example: word embeddings

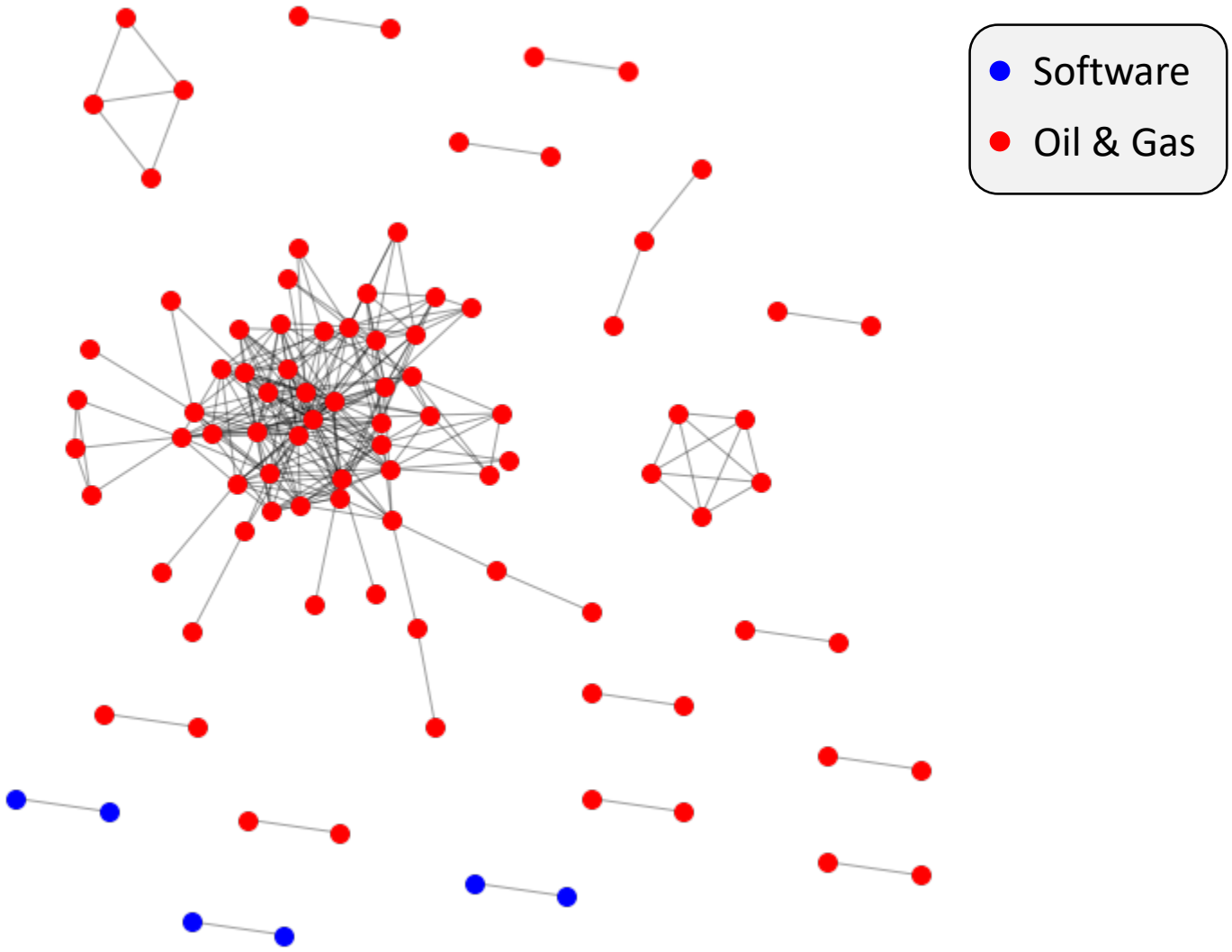
GLOVE/Word2Vec/SkipThought

- 2 layer neural network to predict next word in sentence
- Learns language 'structure' and relationships
- Pre-trained models available

Comparison to model trained on 10-Ks:

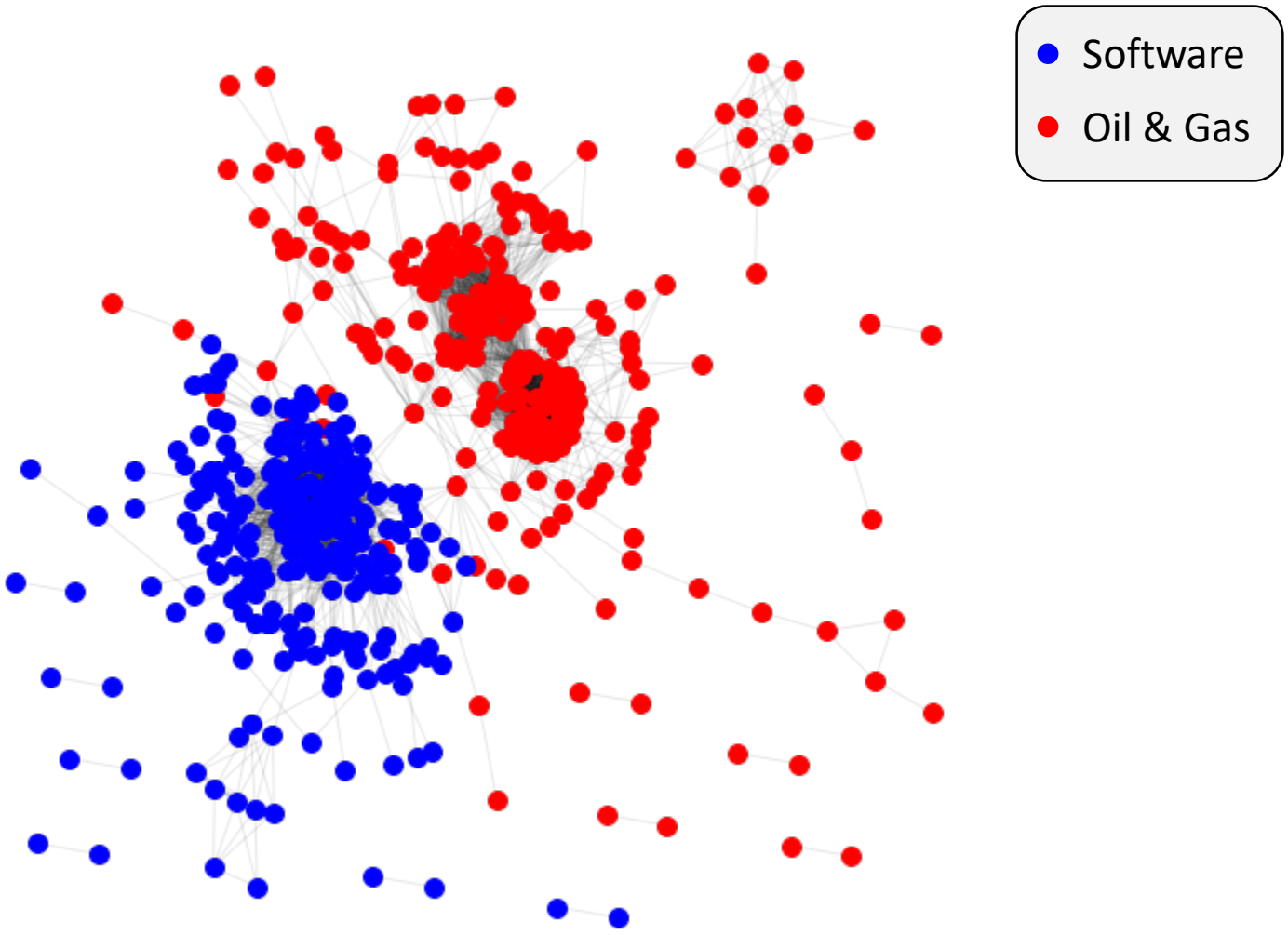
Input	Twitter trained	10-K trained
Android – Google + Apple	Galaxy	iPhone
EPS – splits	Smallville	Diluted
EPS – target	Battlestar	Earnings Loss
Pro + forma	Football reference	Non-recurring, unaudited
GAAP	followinn	Principles, measure, accordance
Netflix	Netflix Streaming	Hulu, YouTube, Amazon

Disclosure 'Network'



Using exact language similarity.

Disclosure 'Network'



Using approximate similarity.

Where to next?

- Adopting machine learning advances to Accounting
  - Unstructured analysis to learn what 'normal' is to detect aberrations in disclosures/audits
  - Network analysis of disclosures to infer future disclosures from peers
- Learning from machine learning
  - Hubel and Wiesel (1959) developed visual neural networks by studying cat brains
  - Ehsani, et al. (2018) program neural networks to mimic dogs and understand how they think
  - Deep Neural Networks outperform models based on expert knowledge

Subasi, 2006. Jiang et al., 2018.

Thank you!