Using Computational Linguistics (NLP) and AI in Economics and Finance

Gordon M. Phillips
Professor of Finance
Tuck School of Business, Dartmouth College & NBER

Banque de France/OECD Innovation LAB

New Ways to Gather and Analyze Textual Data!

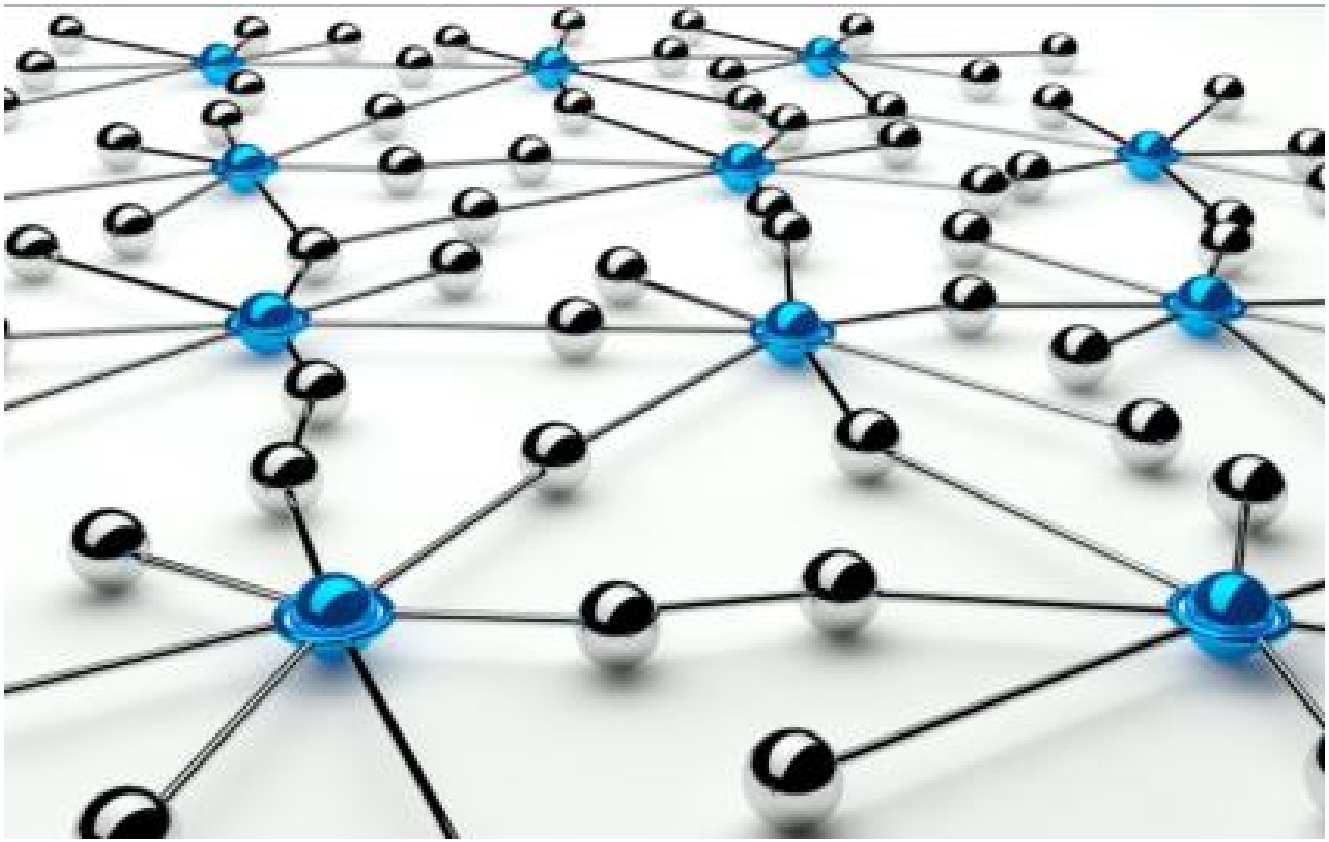
Use scaleable NLP to create a generalized concept of industry, identify competitors, and measure competition. Useful in many finance and economics questions.

Economic Questions Key to Using Big Data

Begin with economic questions and then move to getting and processing the data.

- Most corporate finance studies of interactions of competition and corporate finance decisions have historically focused only on most basic issues of market structure:
  - concentration with preset industry peers according to SIC or NAICS codes.
- Yet real-life competition is rich: market structure is first order in M&A, asset pricing, firm investments, profits and anti-trust.
- Beyond that, these over-simplified industries are mainly used as controls (fixed effects) with little empirical concern for other aspects of market structure.

New data & methods enable *Network representation* of Firms / Individuals



* Modeling which firms (individuals) are in the same industries (group) is essentially “network design”. We can measure localized competition and find related firms/individuals that are important in corporate finance and asset pricing.

Large Language Models: Recent Developments

Recent explosion of language models

2014

doc2vec

New Textual NLP Research in Finance

- Over 1000 papers use TNIC data from these original papers.
  - Hoberg and Phillips (2010 RFS): "Product Market Synergies and Competition in Mergers and Acquisitions: A Text-Based Analysis."
  - Hoberg and Phillips (2016 JPE): "Text-Based Network Industries and Endogenous Product Differentiation."
- Many new "Big Data and NLP" papers in finance and economics.
  - New NLP methods (Doc2Vec) applied to understand firm scope and industry concentration. Hoberg and Phillips (2024, forthcoming *Journal of Finance*: "Scope, Scale and Concentration: The 21st Century Firm"
  - New research using Longformer LLM of patent text from over 600,000 patents to understand innovation competition. Acikalin, Caskurlu, Hoberg and Phillips (2024, wp): **"Intellectual Property Protection Lost and Competition: An Examination Using Large Language Models."**

→ Use Longformer LLM. Being presented in the keynote address by me next Tuesday, the 18th, at 10:50 am, Universite Paris Dauphine-PSL in a conference: Tech for Finance: AI and Blockchain

Large Language Models (LLMs): The Basics

What are language models?

They are functions (parametrized as neural networks) that map free text into high dimensional numerical vectors.

Input text

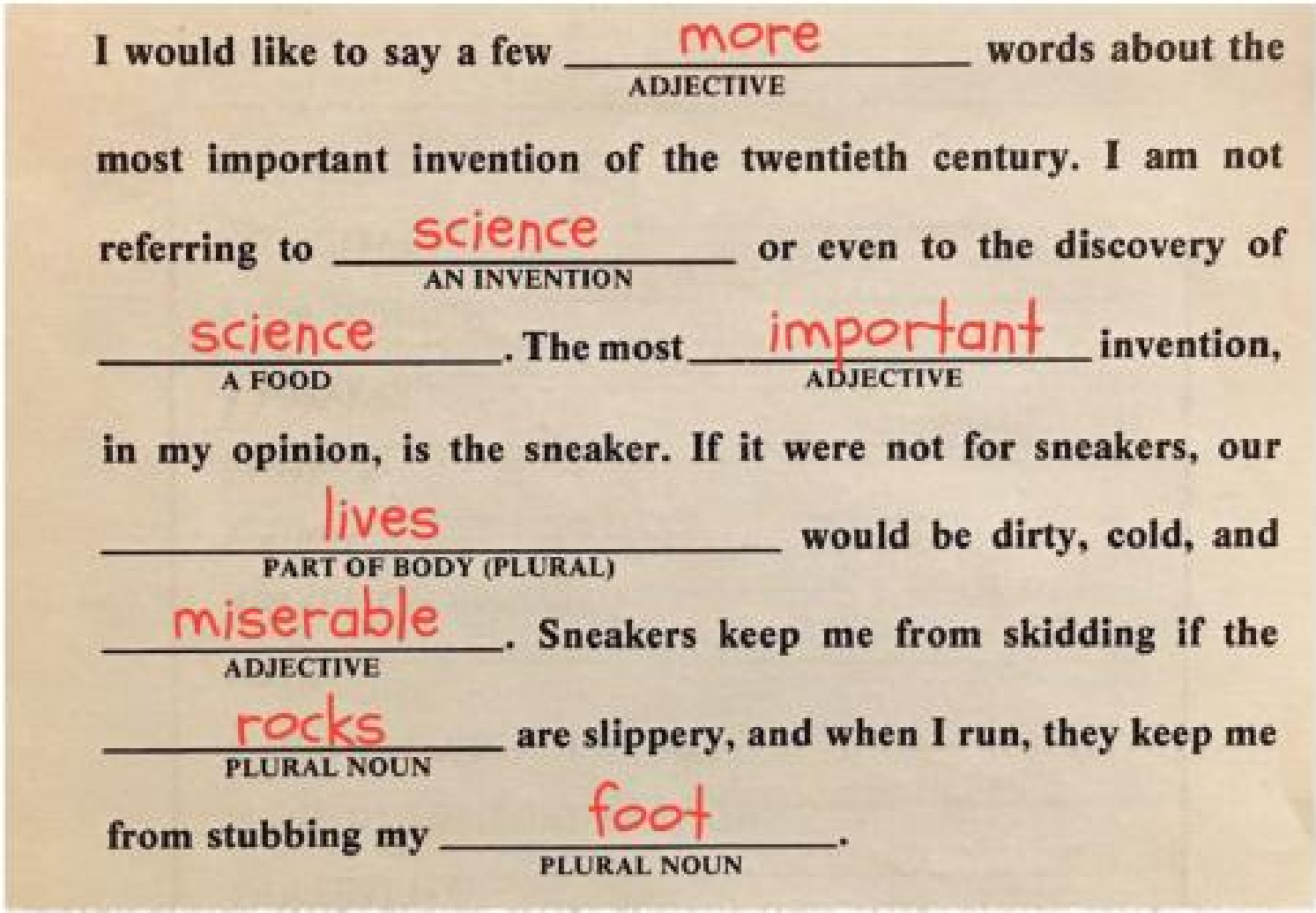
Vectors (aka embeddings)

	A	B	C	D	E	F
1	the	0.056	0.043	0.051	0.08	0.006
2	cat	0.072	0.076	0.1	0.085	0.055
3	dog	0.088	0.099	0.028	0.059	0.06
4	nurse	0.03	0.018	0.058	0.074	0.055
5	doctor	0.097	0.093	0.035	0.057	0.044
6	king	0.013	0.059	0.024	0.032	0.038
7	queen	0.087	0.072	0.029	0.042	0.05
8	bird	0.047	0.044	0.006	0.003	0.003

Can then use these vectors for downstream classification tasks

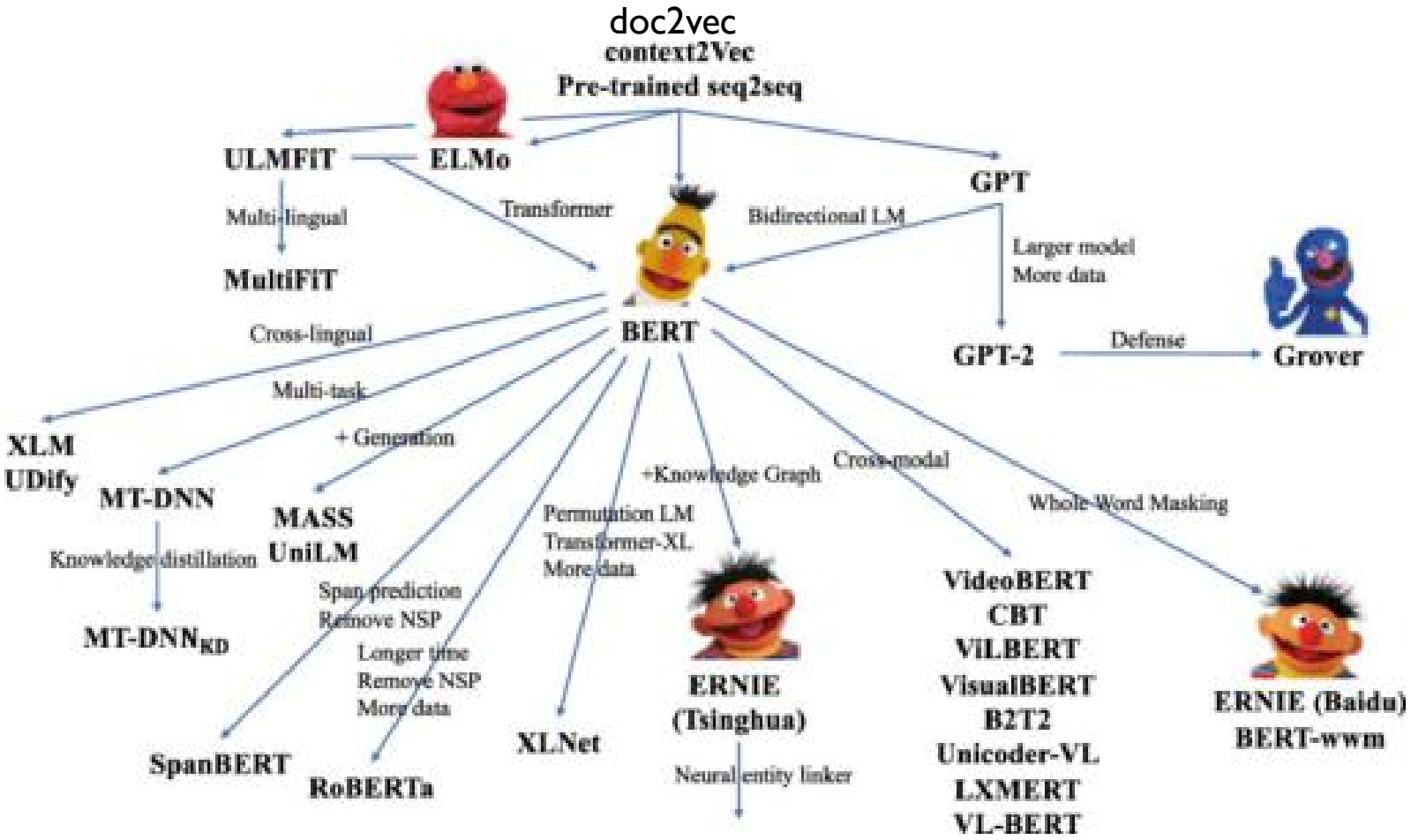
Large Language Models

Language models are trained via mad libs



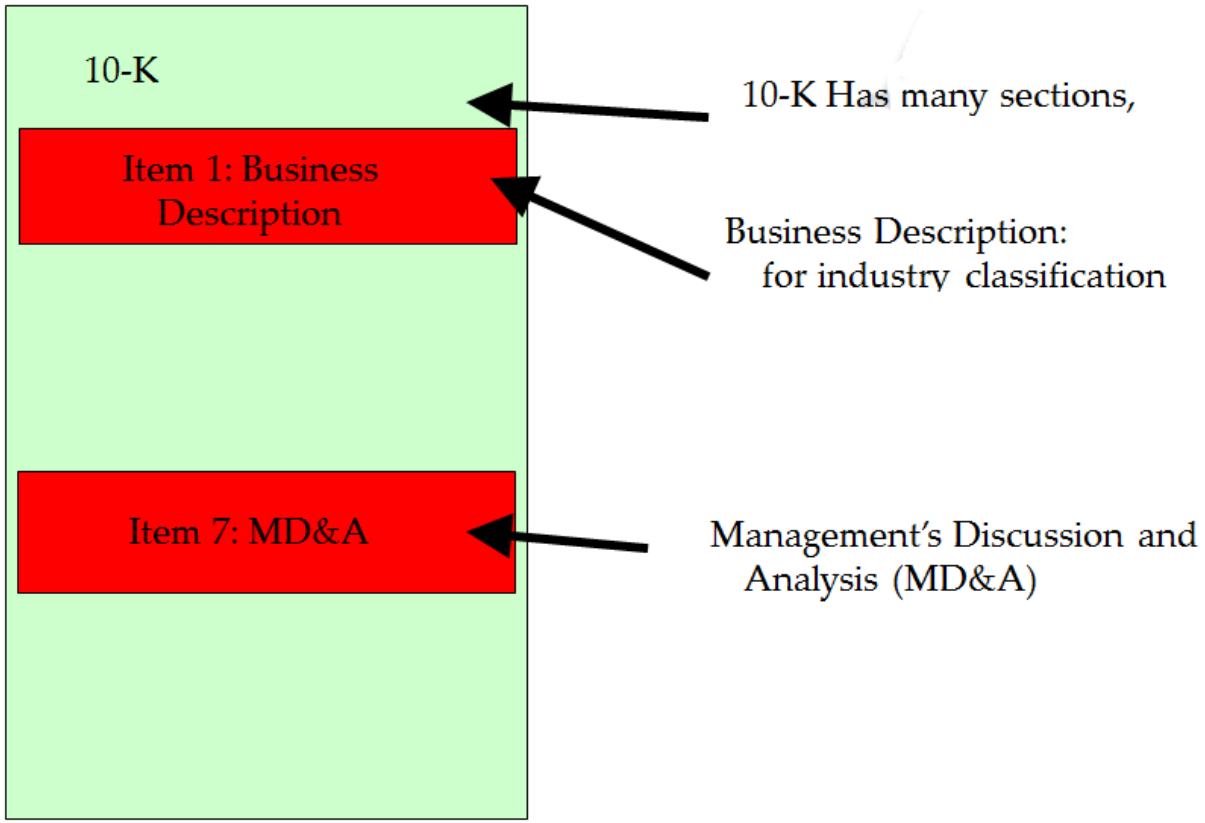
Large Language Models: Doc2Vec in 2014

Recent explosion of language models



Basic Building Blocks: Natural Language Processing (NLP)

We began this analysis in 2006 with PERL on a personal computer. Current work on a high-performance networked cluster with 18TB of stored webpages.



Simply extract Item 1 (Business Description) from every 10-K in each year.

Measuring Product Environment Using Text: Apple

- Apple Computer
  - Apple Product description 1999 10K
    - Power Macintosh, Powerbook, iMac, iBook
  - Apple product description 2009 10K
    - Macbook, MacBook Pro, iMac,
    - iPhone, iPod, iPod Classic, iPodTouch, iTunes, Apple TV

1999 10-K Apple Computer

Apple product description: Dec 1999 10K

Apple Macintosh personal computers were first introduced in 1984, and are characterized by their intuitive ease of use, innovative industrial designs and applications base, and built-in networking, graphics, and multimedia capabilities. The Company offers a wide range of personal computing products, including personal computers, related peripherals, software, and networking and connectivity products. All of the Company's Macintosh products employ PowerPC-Registered Trademark- RISC-based microprocessors.

POWER MACINTOSH-Registered Trademark-

The Power Macintosh line of high-performance personal computers is targeted at business and professional users and is designed to meet the speed, expansion and networking needs of the most demanding Macintosh user. The Company's current line of Power Macintosh systems was introduced in August 1999 and is equipped with PowerPC G4 processors. With the addition of Apple networking software,Power Macintosh systems can be used as workgroup servers.

POWERBOOK-Registered Trademark- G3

The PowerBook G3 family of portable computer products is specifically designed to meet the mobile computing needs of professionals and advanced personal users. Incorporating powerful PowerPC G3 processors, large active-matrix displays, long battery lives, and software designed to enhance mobile computing, the Company's PowerBook G3 family is intended to provide professional desktop performance in a notebook computer.

iMAC-Registered Trademark-

Originally announced in May 1998, the iMac computer is targeted at the education and consumer markets. With an innovative industrial design, easy Internet access, and a powerful PowerPC G3 processor, iMac is suitable for a wide range of education and consumer applications. A completely redesigned iMac was introduced in October 1999 and is available in three models: iMac, iMac DV-Registered Trademark- (Digital Video), and iMac DV Special Edition. Both DV models feature Firewire ports, DVD drives, and the Company's simple-to-use iMovie-TM- digital video editing software.

2 more pages for a total of 874 Words

2009 Apple Inc. 10-K

Apple product description: Dec 2009 10K

Products

The Company offers a range of personal computing products, mobile communication devices, and portable digital music and video players, as well as a variety of related software, services, peripherals, networking solutions and various third-party hardware and software products. The Company designs, develops, and markets to Mac and Windows users its iPhone mobile communication devices and its family of iPod digital music and video players, along with related accessories and services, including the online distribution of third-party digital content and applications through the Company's iTunes Store. In addition, the Company offers its own software products, including Mac OS X, the Company's proprietary operating system software for the Mac; server software and related solutions; professional application software; and consumer, education, and business oriented application software. The Company's primary products are discussed below.

Mac Hardware Products

The Company offers a range of personal computing products including desktop and notebook computers, servers, related devices and peripherals, and various third-party hardware products. The Company's Mac desktop and portable systems feature Intel microprocessors, the Company's Mac OS X Version 10.6 Snow Leopard® ("Mac OS X Snow Leopard") operating system and iLife® suite of software for creation and management of digital photography, music, movies, DVDs and websites.

....

iPhone®

iPhone combines a mobile phone, a widescreen iPod with touch controls, and an Internet communications device in a single handheld product. Based on the Company's Multi Touch user interface, iPhone features desktop-class email, web browsing, searching, and maps and is compatible with both Macs and Windows-based computers. iPhone automatically syncs content from users' iTunes libraries, as well as contacts, bookmarks, and email accounts. iPhone allows users to wirelessly access the iTunes Store to purchase and/or download audio and video content as well as thousands of applications. In July 2008, the Company launched the App Store that allows a user to browse, search for, or purchase third-party applications through either a Mac or Windows-based computer or by wirelessly downloading directly to an iPhone or iPod touch.

2009 Apple Inc. 10-K, part 2

Apple product description: Dec 2009 10K

Products

Music Products and Services

The Company offers its iPod line of portable digital music and video players and related accessories to Mac and Windows users. All iPods work with the Company's iTunes digital music management software ("iTunes") available for both Mac and Windows-based computers. The Company also provides an online service to distribute third-party music, audio books, music videos, short films, television shows, movies, podcasts, and applications through its iTunes Store. In addition to the Company's own iPod accessories, third-party iPod compatible products are available, either through the Company's online and retail stores or from third parties, including portable and desktop speaker systems, headphones, car radio solutions, voice recorders, cables and docks, power supplies and chargers, and carrying cases and armbands.

....

iTunes® 9

iTunes is an application for playing, downloading, and organizing digital audio and video files and is available for both Mac and Windows-based computers. iTunes is integrated with the iTunes Store, a service that allows customers to find, purchase, rent, and download third-party digital music, audio books, music videos, short films, television shows, movies, games, and other applications. Originally introduced in the U.S. in April 2003, the iTunes Store now serves customers in 23 countries. In September 2009, the Company announced iTunes 9, which includes Genius Mixes, a new feature in Genius technology, Home Sharing, which allows users to transfer music, movies and TV shows among up to five authorized computers, and improved syncing functionality that allows users to organize their iPhone applications in iTunes, to sync music by artist and genre, and to sync photos by Events and Faces. In January 2009, the Company announced it would offer all songs in the iTunes catalog without digital rights management software and that iTunes songs would be available at three standard price points, beginning in April 2009 in most countries.

....

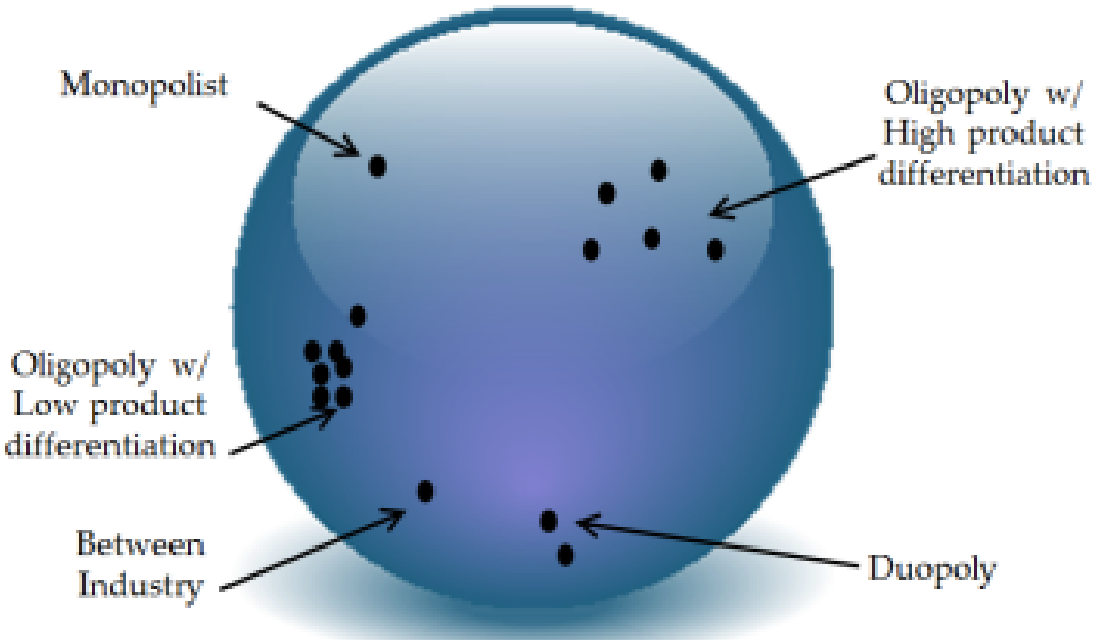
4 more pages for a total of 3,574 Words

10-K Words ==> Over 2 million

- Take all words used in the universe of 10-Ks in product description each year (87,385 in 1997).
- Exclude words (3027 of them in 1997) appearing in more than 5% of all 10-Ks.
- Form boolean vectors for each firm in each year (1=word used, 0=not used).
- Normalize to unit length. Dot products => pairwise product similarity.

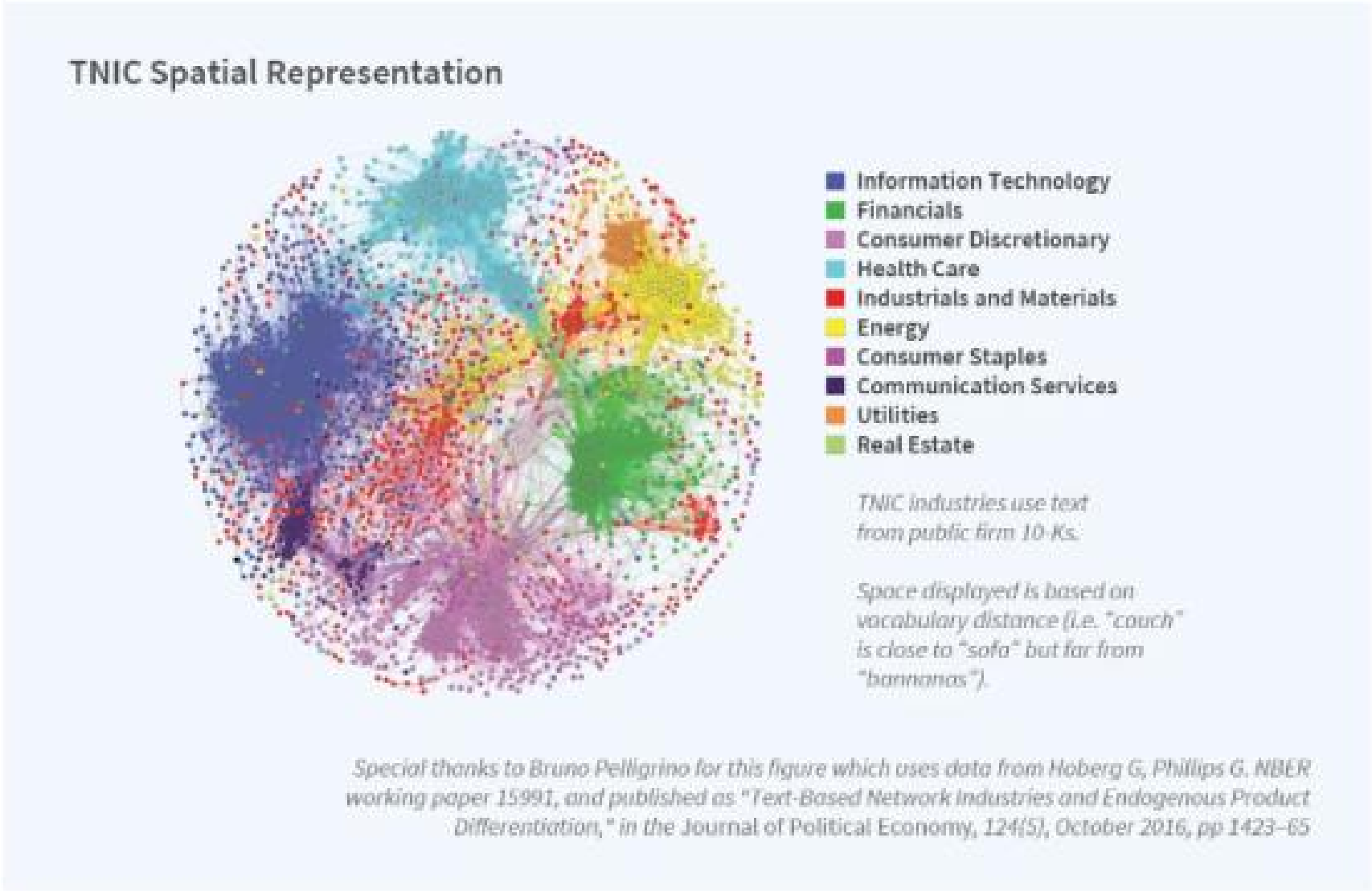
Product Market as a Sphere

Product Market Space is a high dimensional Unit Sphere



Every firm has a location on this sphere. So it has 5000+ public firms and is 80,000 dimensional. Firms residing in dense areas face high levels of competition. Firms in isolated parts of the space are differentiated and effective monopolies.

Hoberg-Phillips TNIC Industries in 2006



Validation is Strong (JPE 2016 paper)! Signal superior to past industries

TNIC Industry Classifications

Adjusted RSQ: various characteristics

Dependent Variable	SIC3 Fixed Eff.	SIC3 Ind.-Yr Avg	NAICS 4 Fixed Eff.	10-K based Fixed Eff.	10-K FIC Ind-Yr Avg	10-K based TNIC
Operating Income/Sales	28.4%	31.2%	28.7%	32.7%	37.2%	45.8%
Adver./Sales	4.1%	8.4%	6.1%	7.1%	16.9%	27.2%
Market Beta	9.6%	15.3%	9.7%	10.4%	15.7%	24.5%

Conclude: Reduces error in variables problem with industry controls

1) Benefits most substantial when complete intransitive network is used

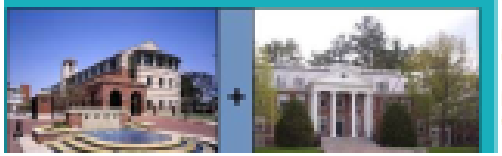
Not only is TNIC more general in modeling of market structure.
It is 50% more informative than SIC! Validation is critical.

TNIC data available on the web: 1989-2021

The Hoberg and Phillips Text Based Industry Classifications have a spatial representation. All firms have a location in a product market space shaped as a unit sphere. Competitive product markets are areas of the sphere where many firms are located. Concentrated areas are sparsely populated.

Some regions of the product space have no firms residing there, as some text descriptions of products would describe products with no demand, such as the word combination "eggs", "paint" and "gardening".

The best way to tap the full research power of this product market grid is to use the Text-based Network Industry Classifications (TNIC), which is a network way of identifying competitors to each firm. Competitors are firms residing in close proximity in product space to each firm based on a continuous measure of similarity. Another key benefit of TNIC industries is that industry composition is updated annually, and our own research indicates that the product market space itself thus dynamically changes over time. As a result, static fixed-location FIC classifications miss out on much of the picture.



Welcome to the Hoberg-Phillips Data Library

<< NEW: Data extended to 2021 (overall coverage now 1989 to 2021)! >>

Data provided by Gerard Hoberg (USC)
and Gordon Phillips (Dartmouth)

Text-based Network Industry Classifications (TNIC) data **[click here]**

* These new industry classifications are based on firm pairwise similarity scores from text analysis of firm 10K product descriptions. Competitors are firm centric with each fi
of competitors - analogous to networks or a "Facebook" circle of friends. These new industry classifications are updated annually and offer more research flexibility, and are
FIC (fixed industry) classifications such as SIC, NAICS, and the 10-K based FIC classifications below. Our research shows they sharply improve upon SIC and NAICS codes
firm specific decisions, including firm profitability, Tobins Q and dividends. These benefits are outlined in Hoberg and Phillips (2010, 2016), with references available by clicki

Industry Concentration and Total Similarity Data **[click here]**

* HHI Concentration metrics and Total Similarity data is available based on TNIC Industries.

Product Market Fluidity Data **[click here]**

* Product Market Fluidity data assesses the degree of competitive threat and product market change surrounding a firm, and is based on Hoberg, Phillips and Prabhala (201

Vertical TNIC Data (VTNIC) **[click here]**

* Vertical TNIC data is comprised of two key databases, and is based on Fresard, Hoberg, and Phillips (2019 forthcoming). The first is a firm-year panel indicating the extent
integrated. The second is a firm-pair-year database indicating the potential for vertical relatedness for every pair of firms in every year.

Papers that develop and use these data

Scope, Scale and Competition. The 21st Century Firm. [Download Paper]
Gerard Hoberg and Gordon Phillips, working paper.

Text-Based Network Industries and Endogenous Product Differentiation. [Download Paper]
Gerard Hoberg and Gordon Phillips, 2016, *Journal of Political Economy* 124 (5), 1423-1465.

Product Market Synergies and Competition in Mergers and Acquisitions: A Text-Based Analysis. [Download Paper]
Gerard Hoberg and Gordon Phillips, 2010, *Review of Financial Studies* 23 (10), 3773-3811.

Real and Financial Industry Booms and Busts. [Download Paper]
Gerard Hoberg and Gordon Phillips, 2010, *Journal of Finance* 65 (1), 45-86.

Now at: http://hobergphillips.tuck.dartmouth.edu/

New Textual NLP Research in Finance

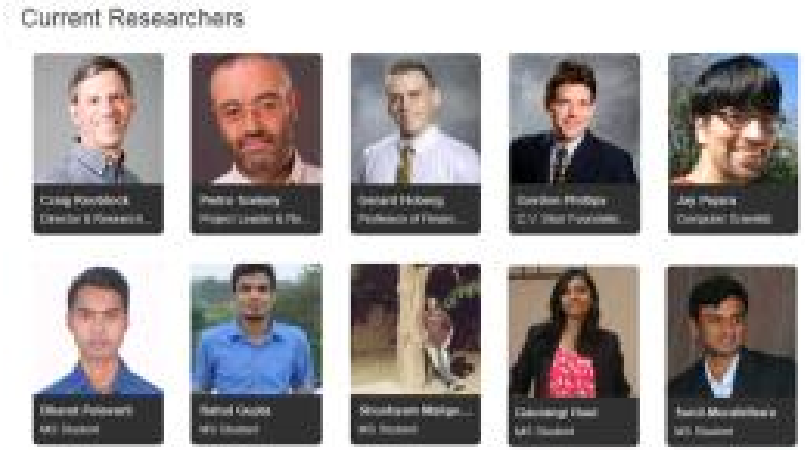
- Over 1000 papers use TNIC data from these original papers.
  - Hoberg and Phillips (2010 RFS): “Product Market Synergies and Competition in Mergers and Acquisitions: A Text-Based Analysis.”
  - Hoberg and Phillips (2016 JPE): “Text-Based Network Industries and Endogenous Product Differentiation.”
- Many new "Big Data and NLP" papers in finance and economics. We will cover These 2 new recent papers.
  - New NLP methods (Doc2Vec) applied to understand firm scope and industry concentration. Hoberg and Phillips (2024, forthcoming *Journal of Finance*: “Scope, Scale and Concentration: The 21st Century Firm”
  - New research using Longformer LLM of patent text from over 600,000 patents to understand innovation competition. Acikalin, Caskurlu, Hoberg and Phillips (2024, wp): “Intellectual Property Protection Lost and Competition: An Examination Using Large Language Models.”

Recent NLP Advances

- Semantic relatedness can improve power beyond cosine similarities.
- Embedding technologies showing strong promise in a pilot study.
- Latent Dirichlet Allocation is not recommended for economic reasons.

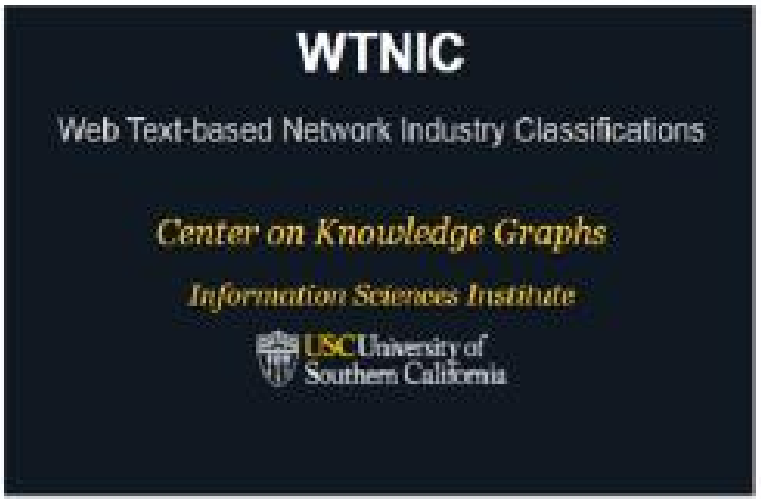
* New WTNIC (Web Text-based Network Industry Classifications) models coming soon using almost 1 million Webpages for over 20 years.

Major Expansion to Private Firms



2 PI's from B-School
2 PI's from Comp Sci (Viterbi)
1 Newly added Viterbi PhD

5 Masters Students in Comp Sci
1 Undergraduate RA



We thank the National Science Foundation!
$500,000 Grant
Joint B-School & USC's Viterbi School (ISI)

Joint NSF-funded research b/t B-School and Comp Sci Researchers.

Finding Competitors Using Web-text

Major concerns about declining competition and innovation in the US.
Getting data on private firms from websites for almost 1 million private firms.

- Over many many websites, we often see:
  1. Employment section.
  2. Social media section
  3. ESG/corporate social value section
  4. Blogs
  5. Press release section
  6. Financials and SEC filings
  7. Specific information about the company's products!
- Our project seeks to focus on 7.

Good news: items (1) to (6) have a strong factor structure, and we can use that to purge this content!

Why do all of this?

- Major concerns about declining competition and innovation in the US.
- Hard to find data on early-stage firms.
- new Economy involves firms with multi-dimensional products. Can measure with text!
- Identifying frictions –> change IP rules or provide “right” incentives –> eventually the right policy choices for next 100 years!
- The counterfactual is weakly established conclusions and a policy of “trial and error” with lackluster economic performance.

* Textual analysis in research has the potential to pick out complex interactions.

Conclusions - Computational Linguistics in Finance

Computational Linguistics (NLP) work began on a PC back in 2006 with simple methods.

- Stage 1, modeling U.S. public firms, was completed in late 2008.
- Much validating research has already been done following stage 1.

New methods include "doc2vec" (paper by me and Jerry Hoberg, forthcoming *Journal of Finance*, and machine learning transformer models (LLMs).

- Stage 2, expansion to U.S. private firms with new transformer language models to be completed soon.

==> Many applications and economic tests can be examined with greater flexibility!