Deliverable 1.4

MSCA-ITN Training for Big Data in Financial Research and Risk Management "BigDataFinance"

Grant Agreement: 675044

This is Deliverable 1.4 of the Work Package 1 (WP1) in "Training for Big Data in Financial Research and Risk Management" (BigDataFinance) Innovative Training Network Marie Sklodowska-Curie project 2015-2019.

Name of the deliverable: "A report and software on a verified and validated knowledge extraction prototype with different data sources"

Description

This report describes efforts to provide a unified platform for analysing and understanding a variety of financially relevant datasets, ranging from textual to semi-structured and time-series data. We include a software library containing supporting classes for the knowledge extraction and analysis processes described. We also present a research paper in the domain of empirical asset pricing, which exemplifies the types of analysis that can be achieved with the infrastructure developed as part of this deliverable.

Date, place: March 26th, 2018, Ljubljana, Slovenia

Name, position: James Hodson, Marie Sklodowska Curie Fellow, Artificial Intelligence Laboratory, Jozef Stefan Institute.

Table of Contents

Abstract

Finance deals with the question of optimal asset allocation. Over the past years, there has been tremendous growth in the availability of large data sets to help explore all different aspects of this question, from highly granular Limit Order Books[1] to high frequency news announcements[2]. We develop a series of tools and empirical explorations in order to understand how Machine Learning can play a role in helping to incorporate some of this information into financial decision making.

As these large 'alternative' data sets become more ubiquitous, there has been a race to identify ways to extract value from them for the investment process[3]. Often, these data sets will be large, noisy, and unstructured, making them difficult to work with from the perspective of traditional econometric techniques. **Large data** can pose practical challenges from the perspective of computational resources, as well as model estimation. **Noisiness** creates a perceived sparsity in the data which can prevent the identification of important patterns--not because they do not exist, but because they are obfuscated by missing or incorrect data points. FInally, a **lack of structure** generally makes it difficult to parameterize the data in a consistent manner for analysis, and makes robust hypothesis testing more tricky.

This report focuses on methods that have been developed in order to cope with these three sources of difficulty across relevant financial data sources. Primarily, unstructured data will be explored through the lens of textual data, since this provides a natural framework for thinking about the problem.

[1] See for example: Kanniainen, Juho, Siikanen, Milla, and Valli, Jaako, "Limit Order Books and Liquidity around Scheduled and Non-Scheduled Announcements: Empirical Evidence from NASDAQ Nordic", Finance Research Letters (Forthcoming), 2018.

[2] See, for example: Fedyk, Anastassia, "News Consumption: From Information to Returns", Retrieved from SSRN, Sunday, March 18th, 2018: News Consumption: From Information to Returns by Anastassia Fedyk ...

[3] Many events have started to focus on Machine Learning and Alternative Data for Finance: e.g. 5th Annual RavenPack Research Symposium: The Big Data & Machine Learning Revolution.

Introduction

Large, unstructured, and noisy textual data sources have become readily available as the web has grown and become a household commodity. Millions of product reviews, resumes, legal and corporate filings, blogs, news articles, governmental releases, (and more) may be freely downloaded and offer tremendous opportunities for unlocking new business value, enabling more effective communities, and supporting more accurate decision making.

Unfortunately, large, unstructured, and noisy textual data sources have three primary disadvantages that have made them difficult to work with, and helped them avoid wide-spread adoption:

They are LARGE,

they are UNSTRUCTURED,

and they are NOISY.

Certainly, with the advent of wide-spread and relatively affordable "elastic cloud" computing grids in the past decade, the problem of scale is not impossible to overcome, although some types of task, especially those that cannot be easily parallelized and require large amounts of fast local memory (e.g. graph building, complex dependent structures), can still pose challenges.

Our main focus is on the latter two problems--lack of structure, and noisiness. The former stems from the inherent infinite property of language, able to generate an endless set of variations when referring to substantially equivalent concrete concepts. The latter stems from the real world consideration that most information (especially text) is still generated by fundamentally flawed and error-prone machines--human beings. People lack the coordination ability to decide on a single referent for each concept or relation, and thus generate noisy data, sometimes missing key pieces of the puzzle, making mistakes, or purposefully misleading the end consumer.

Since these data sources are becoming more and more commonplace, the need for a consistent and reliable set of methodologies and recommendations has also grown. Industries from publishing to finance, recruiting to sales and marketing, and even sophisticated governments are unable to make the most of the vast amounts of unstructured and semi-structured textual data they collect on a daily basis.

In the world of Finance, in particular, there are incredible opportunities to better understand the prospects of firms and viability of different potential investments through the analysis of new alternative data sources. For instance, being able to jointly consider the entire publicly available information set about a firm through news content could allow investors to identify missing links or insights that the rest of the investor community has overlooked. This becomes especially true and salient as the amount of data grows, but also as the complexity of data sets increases.

This is where modern Machine Learning can help to bridge the gap between traditional economic methodology and the large alternative data sets that hold the key to understanding investor behaviour in the wild. We have developed techniques that allow for unstructured data sources to be suitably clustered, taxonomized, corrected, and estimated. These transformations allow unwieldy data sets from disparate sources to be analyzed, understood, modeled, and exploited alongside

traditional econometric data sources like time series, industry classifications, and accounting variables.

We ground our research in the realm of recruiting and the economics of the global labour force (commercial finance). We are interested in understanding the internal workings of individual firms, how people transition from role to role, the skills that they acquire, and how intra-firm workforce movements can affect the financial performance of firms, innovation, industry-wide dynamics, and the economy more broadly. We also would like to better understand when and why people are likely to seek new employment, what the barriers are, and which skills are most in-demand. In the past, such explorations have depended largely on panel data from government surveys (usually restricted to particular industries or companies)[4], or internal Human Resources data from a single large firm[5].

We take a different approach, and instead use hundreds of millions of people's employment histories from resumes and web-based employment profiles. However, each resume is formatted differently, uses different conventions, and may be in any of dozens of languages, and even the semi-structured employment profiles we have access to were manually created, and contain very noisy data, with the potential for many missing or erroneous entries. For instance, every individual is at liberty to write their company name, title, department, location, and skills in whatever way they like best. We are left with a vast number of context-less short strings.

As if to add insult to injury, the concepts that we are interested in working with: companies, roles, skills, etc., do not have readily available reference taxonomies that we can link to, and contain structures that make the problems quite different from entity resolution problems in the literature: when someone lists their role as "Sr. EA to the VP of Marketing (N.A.)", we want to understand that their role is Executive Assistant, with a seniority marker indicating a more senior role than a simple Executive Assistant, and that their boss is a Vice President, in the firm's Marketing department, with responsibilities for North American operations. That's a lot of information, and represents one of tens of thousands of different representation choices.

Ultimately, to study the dynamics of movements in the workforce at the level of an entire economy, we need to move from individual job records to the massive intertwined graph of job transitions, and we need to use the large occurrence of job transitions to infer the correct internal structure of each firm. In order to build such a graph, we need clean disambiguated data. Conversely, in order to get clean, disambiguated data, we could really make use of distributions of roles and employment over the entire graph. This suggests an opportunity to iteratively refine both sides of the equation, and jointly learn how to solve these two problems optimally.

As part of this exploration, we have developed a series of software libraries and tools for dealing with the cleaning, preparation, and modeling of such data sets. The software is described in the section “Software Library Overview”, and is available as a set of Python API’s for academic purposes through the Jozef Stefan International Postgraduate School.

[4] Such as: Gibbons, Robert, and Lawrence F. Katz. "Layoffs and Lemons." *Journal of Labor Economics* 9.4 (1991): 351-380.
[5] See: Baker, George, Michael Gibbs, and Bengt Holmstrom. "The internal economics of the firm: evidence from personnel data." *The Quarterly Journal of Economics* 109.4 (1994): 881-919.

In the remainder of this report we provide an overview of the data sources involved, an overview of the software libraries developed, and summaries of the research and dissemination activities enabled by this work to date. We include a working paper that exemplifies the research methodology, “Trading on Talent: Human Capital and Firm Performance”. This paper was enabled by the libraries and data described in this report, and has won multiple prizes and awards for its contribution to empirical asset pricing (notably, the Q-Group “Jack Treynor” Prize, and the PanAgora Asset Management “Crowell” Prize).

Data Protection and Privacy Notice

Various data sources have been licensed and collected for the purposes of exemplifying the methodology, developing statistical models, and elaborating analytical criteria. These data sources remain the property of the respective entities, and it may not be possible for us to provide access directly. Briefly, the data used in these projects spans large scale employment data (resumes, electronic profiles), news articles, SEC filings and regulatory announcements, global exchange listings data, company websites, wikipedia articles, pricing data from the Center for Research in Security Prices (CRSP), and accounting/corporate events data from Compustat.

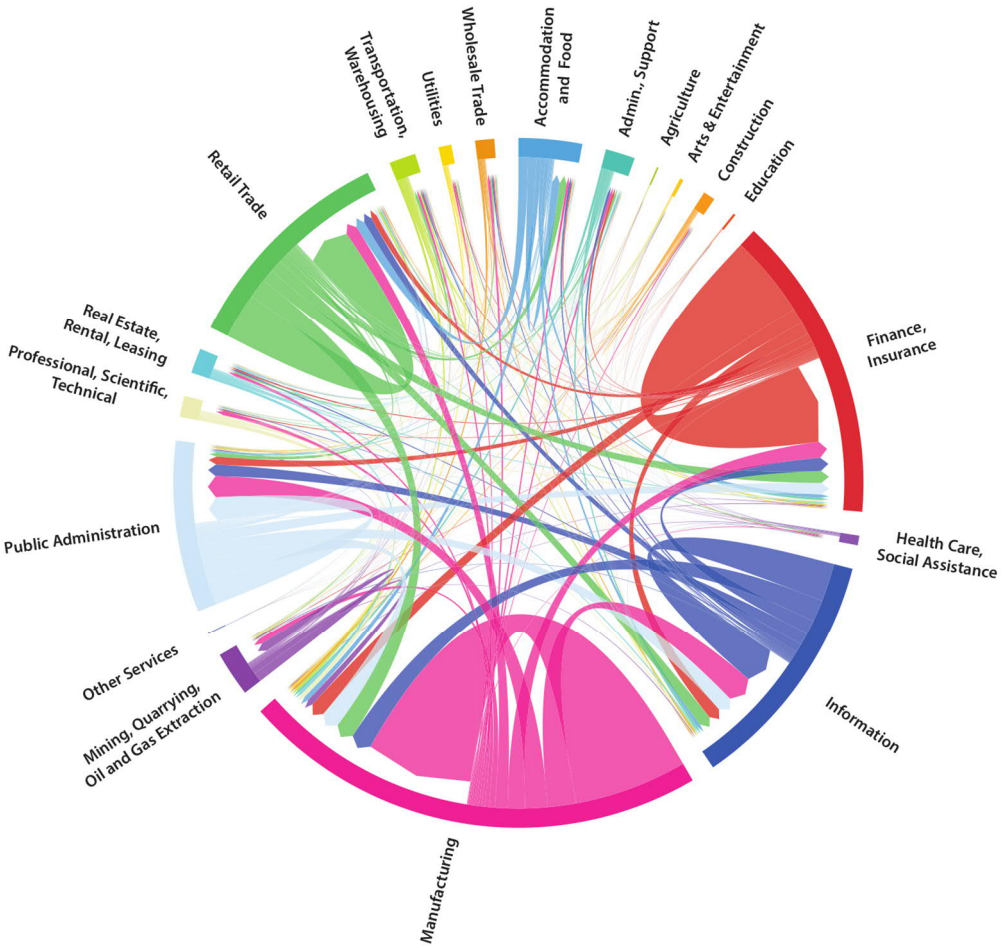
Any information pertaining to the career movements of individuals, or other private individual level data, has only been used for specific research projects where the individuals concerned have specifically provided their consent (for academic research related to career paths and the dynamics of the labour market/economy) to the firm or firms involved. At no point have we directly accessed or downloaded individually identifiable data. As such, all projects have been carried out in compliance with EU legislation on Data Protection and Privacy (GDPR). Additionally, this research received a “not human subject research” determination from the Institutional Review Board at Harvard University (protocol #IRB17-0443).

Financial Data Overview

Employment Data

Employment data for the exploration of how human capital impacts the financial performance of firms, and the dynamics of the labour market, was provided by Cognism Ltd. Cognism has business relationships with human resources departments, recruiting agencies, client relationship management system vendors, and online employment websites, in order to help these firms leverage their data to recommend services, enhance team collaboration, and identify career opportunities. When firms are onboarded with Cognism, and on a regular basis thereafter, specific data uses are opted-in through updated consent agreements with the end employees and individuals concerned. This allows our research projects to be adequately opted-in on a regular basis as part of a robust process of engagement. Of the employees of US publicly listed firms since 1990, 37 million individuals provided their consent to our research project through this framework. This represents approximately 96% of the total pool of available individuals.

Figure 1: Within and Across Industry Transitions, US Public Companies, 2010-2017



Employee resumes and electronic resumes are parsed into basic field types through a set of structured document understanding systems (string, date, integer, float, etc.), and visually clustered. Due to the inherent regularity of records on a resume, a handful of regular rules is able to capture much of the variance in fields. We are provided with rough assignments into likely fields (position, company name, date range, location, degree, etc.). However, these are often mis-assigned, and there is very little regularity in how two individuals refer to the same position, company, location, or even time range.

Our software system provides classification of each entity according to a reference set, instance priors, and available feature aggregates. The model can be learned independently, or jointly within an active learning loop (we generally used Amazon Mechanical Turk supervision by recruited and trained experts). For US public companies we achieve a coverage of approximately 45% of employees, and practically all employer firms.

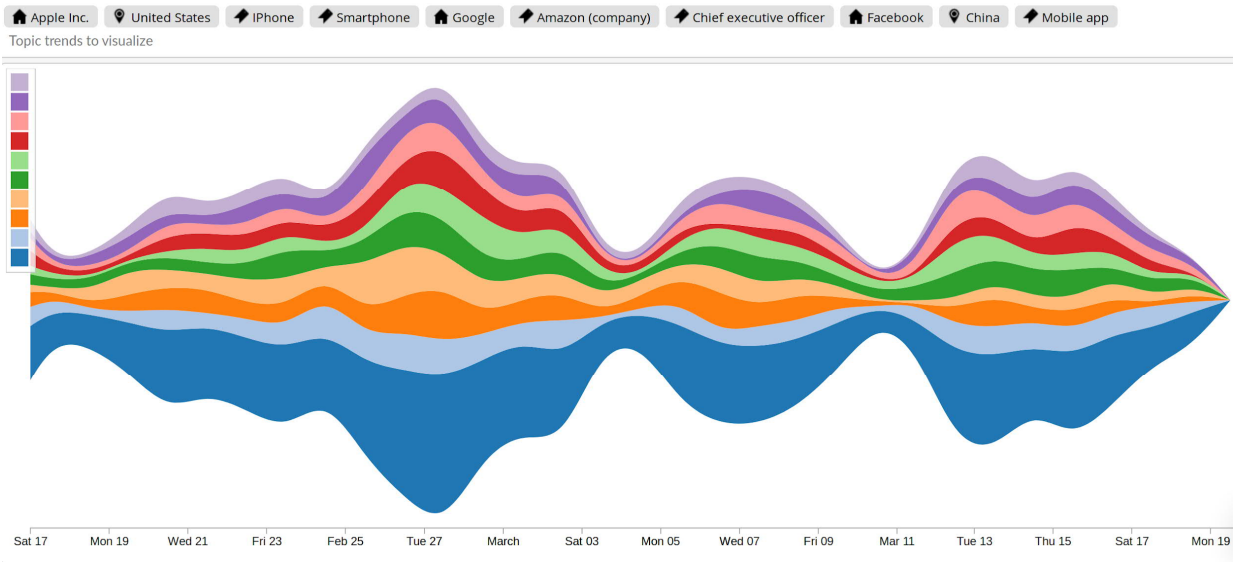
With this data, we are able to show some interesting insights into the US economy on an industry by industry level. Figure 1 shows proportions of transitions within and across industries between 2010 and 2017.

News Articles

The software and methodologies developed as a part of the Big Data Finance project, and explained in this report, are actively integrated into the EventRegistry platform, a project on Global Media Monitoring from the Jozef Stefan Institute[6]. Figure 2 shows a visualization of news about Apple (a corporate US public entity), with relation to the topic trends over the one month period ending Monday, March 19th, 2018.

Figure 2: Tracking Apple Events Through the EventRegistry System

[6] See http://www.eventregistry.org/



EventRegistry operates on top of NewsCrawler, a web based News crawling system that monitors over 100k news sources across 100+ languages globally. Each day, NewsCrawler ingests several hundred thousand news stories, extracts relevant content, and assigns topics, organizations, people, and locations to the text[7]. The EventRegistry system is able to cluster news articles in order to identify "events", or clusters of news articles referring to the same underlying real world events. Information about firms and individuals can be enriched with the help of the EventRegistry system, in order to aggregate pertinent details.

SEC Filings

The Securities and Exchange Commission (SEC) is the central regulatory authority for US financial markets. Publicly traded firms are required to regularly file certain information relating to their health and activities in order to keep the investing public informed. SEC filing requirements include quarterly and yearly accounting disclosures, insider trading disclosures, holdings and divestiture information, material events reporting, press release registration, and a variety of other regular forms.

Most forms required by the SEC are filed as unstructured text (perhaps in a structured set of fields mandated by the form), and firms have latitude with respect to language, presentation, and timing of release. As such, SEC filings have been the focus of much literature on information processing in financial markets, and strategic corporate behaviour. However, since SEC filings are mostly text, it has been difficult for finance/economics researchers to adequately parameterize information except through labor intensive manual annotation, or superficial measures (such as "sentiment").

As part of this project, we collected 6 years of complete SEC filing data through the SEC's web based public interface. Starting with 8K forms (material event disclosure), we have explored how to categorize, structure, and understand the information available in these critical filings.

[7] See http://enrycher.ijs.si/

Global Exchange Listings Data

Global Exchange Listings data is obtained directly from exchanges (e.g. NASDAQ, London Stock Exchange, SIX Swiss Exchange). Our web scraping adapters check exchange websites each month to gather information about newly listed/delisted companies, their industry, aggregate trading volume, website, and summary data. Unfortunately, it is difficult to find ground truth data sources for global historical listing/delisting and corporate information, so this must be done manually on an ongoing basis.

Company Websites Data

Company websites are yet another massive source of unstructured text and media data that is important for financial professionals to process. However, due to the inherent differences between companies, web technologies, choice of information to provide, (and often the size of the website) it is difficult to keep track of changes and understand the relevant information.

Our infrastructure collects website data for all US public companies, and additional global coverage. For each website, we also collect relational web metadata, so that we can understand the links between companies, industries, and technologies. This corpus is ripe to be exploited through the methodologies we discuss and elaborate in this report.

Wikipedia

As a global, multilingual source of validated knowledge about millions of entities, Wikipedia provides a wealth of reference data and “clues” for advanced analysis. We leverage monthly wikipedia article dumps in order to aggregate information on public companies (those that are represented), and to leverage the available structured information (such as tickers and exchange data) in order to better disambiguate company names, and better model the concepts involved.

CRSP and Compustat

We use basic accounting variables, and exchange data obtained through CRSP and Compustat in order to enhance the financial analysis of the companies in our data and answer questions relating to firm performance and investment strategies.

Software Library Overview

We have developed a software library for dealing with large, noisy, unstructured data in the financial domain. Our library provides modular functionality for training a range of classification, clustering, and latent factor models in order to better understand, model, and exploit large heterogeneous data sources. The software library was developed with particular applications in mind, but many of the modules are generic in nature.

Overview of Library

LNUTextLib

Input

A set of classes related to the ingestion of data from diverse data sources. This includes input from json records (semi-structured employment profiles), company data, reference data, etc.

Feature

Classes related to the extraction, ranking, and selection of features for textual document classification. Works for both document and text level annotations, and for supervised and unsupervised problem domains. The classes support arbitrary ranking functions and selection criteria.

Disambiguation

Classes to associate named entity instances with a reference set, given features of interest. The disambiguation classes implement string matching, fuzzy matching, templates, and feature based statistical matching. Can be augmented through active supervision.

Latent

Classes for inferring latent relationships among categorical variables. This includes topic modeling, hierarchical inference, and clustering approaches to grouping in unsupervised settings. In particular, these classes are targeted towards skills and department inference from career data.

Graph

Classes for reconstructing and inferring hierarchical structures under the assumption of missing and spurious edges. Implements a graph pruning measure based on log-likelihoods across graph sub-structures. Primarily used for reconstructing corporate hierarchies.

Output

A set of classes for formatting and outputting enriched data in structured formats that can be leveraged within economic analysis and visualization settings.

Enabled Research

Trading on Talent: Human Capital and Firm Performance

Does human capital impact firm performance? By directly observing the employment and education trajectories of a significant proportion of US public company employees from 1990 to the present, we explore the relationship between performance and two aspects of human capital: turnover and skills. First, we find that firms with higher employee turnover experience significantly worse future returns. A long-short strategy based on employee turnover with a three-month lag generates an excess return of 1.12% per month. Second, firms with a larger emphasis on sales-oriented skills show better subsequent performance, whereas firms with more focus on administrative skills underperform. The effects of skills are heterogeneous across industries, with a larger premium on web development in Information, a higher premium on insurance in Manufacturing, and no benefit from sales-oriented skills in Finance.

Asymmetric Peer Group Learning

We describe a distributed learning paradigm over dynamic network structures that is inspired by human behavior. Our approach exploits node-level heterogeneous attachment communities that learn and act only based on locally available information. We show that when local peer groups are sufficiently heterogeneous and overlapping, we approach globally optimal learning outcomes. This paradigm benefits from a high level of parallelization, converges quickly, and fits a variety of real-world settings. We motivate this method through an application to tracking skilled workforce transitions in the United States between 2010 and 2015, showing that the network can quickly learn to predict which firms will have higher (lower) employee turnover in the future.

Career Consequences of Corporate Scandals

Do corporate events leave black marks on individual employees? The Lehman Brothers bankruptcy offers a setting where individual rank-order employees are affected by a large-scale negative event for which they are not personally responsible. Using rich employment profile data, we match each employee of Lehman Brothers in 2008 to the most similar employees at Goldman Sachs, Morgan Stanley, UBS, and Deutsche Bank based on the following characteristics: age, gender, job position, educational background, and skillsets. By 2017, the former Lehman Brothers employees are 20% more likely to have had at least a year-long break from reported employment and 25% more likely to have left the financial services industry, but show similar career growth conditional on remaining in the industry. On the flip side, the former employees of Lehman Brothers are also approximately 60% more likely to found their own businesses. The negative effects (breaks from employment and industry switches) are larger for more senior individuals such as vice presidents and managing directors, and absent for the junior analysts and associates. By contrast, the increased spillover to entrepreneurship is consistent across the hierarchy, and especially noticeable for the relatively young associates.

Discovering Hierarchical Structure from Bags of Paths

A firm's hierarchy and structure reveal a lot about culture, management, and ability to compete. As a result, companies treat this information as highly confidential. Past studies of corporate structure have relied primarily on panel evidence from single firms. We propose a robust network based method that is able to infer accurate corporate hierarchies and multi-dimensional models of corporate structure dynamically over a multi-year period. Using highly granular employment data from US public firms over a 10 year period covering 2007-2017, we construct large firm-centric networks from individual career paths, and fit a hierarchy through an expectation maximization framework at each period. We show that these hierarchies can provide useful insights into a variety of firm-level, industry-level, and economy-wide dynamics.

Dissemination Activities

5th Annual RavenPack Research Symposium: The Big Data & Machine Learning Revolution

September 19, 2017, New York City
Hodson, J., (Jozef Stefan Institute), "People, Firms, Economies: The new age of data-driven finance".

Financial Data Science Association Workshop

September 3-7, 2017, Dubrovnik, Croatia
Fedyk, A., (Harvard University), "Leveraging Large Scale Employment Data in FInance".

Western Decision Sciences Institute Annual Meeting

April 4-8, 2018, Kauai, United States
Fedyk, A., and Hodson, J., "Trading on Talent: Human Capital and Firm Performance".

PanAgora Crowell Prize

"The Crowell Prize provides a forum for new and cutting-edge research that connects theory and practice."

Fedyk, A., and Hodson, J., "Trading on Talent: Human Capital and Firm Performance".

Q Group Jack Treynor Prize

"The *Q Group's* annual *Jack Treynor* Prize recognizes superior academic working papers with potential applications in the fields of investment management and financial markets."

Fedyk, A., and Hodson, J., "Trading on Talent: Human Capital and Firm Performance".

BlackRock, Goldman Sachs Asset Management, Two Sigma, Blue Mountain Capital

Invited Presentations, 2017-2018.
Hodson, J., "Incorporating large-scale labor market data in investment decision making".

Harvard University

Contracts, Seminar Series, May, 2017.
Fedyk, A., and Hodson, J., "Career Consequences of Corporate Scandals".

Harvard Business School

Entrepreneurship, Seminar Series, July, 2017.

Fedyk, A., and Hodson, J., "Career Consequences of Corporate Scandals".

University of Zurich DOLFINS/Big Data Finance Conference

Winter School Presentations, January 8-13, 2017.
Hodson, J., (Jozef Stefan Institute), "Leveraging Large Scale Employment Data in FInance".

Cubist Systematic Strategies LLC

Invited Presentation to Quantitative Strategies Group, New York City.
Fedyk, A., and Hodson, J., "Trading on Talent: Human Capital and Firm Performance".

Exemplary Research Paper

Research paper titled: “Trading on Talent: Human Capital and Firm Performance” is available at: https://scholar.harvard.edu/files/fedyk/files/humancapital_firmperformance_nov2017.pdf

The paper is the winner of the Jack Treynor Prize from the Institute for Quantitative Research in Finance and received second place in the 2017 PanAgora Asset Management Dr. Richard A. Crowell Memorial Prize.