This is an Accepted Manuscript of an article published by Taylor & Francis Group in *European Journal of Finance* on 17 Sep 2019, available online: http://www.tandfonline.com/10.1080/1351847X.2019.1662822.

Financial Data Science:

The birth of a new financial research paradigm complementing econometrics?

Chris Brooks[a], Andreas G. F. Hoepner[bc], David McMillan[d], Andrew Vivian[e], Chardin Wese Simen[a]

[a] ICMA Centre, Henley Business School, University of Reading, Whiteknights, Reading RG6 6BA, UK
[b] Michael Smurfit Graduate Business School & UCD Lochlann Quinn School of Business, University College Dublin, Carysfort Avenue, Blackrock, Co. Dublin, Ireland, EU
[c] Technical Expert Group on Sustainable Finance, DG FISMA, European Union
[d] Department of Accounting and Finance, University of Stirling, UK
[e] School of Business and Economics, Loughborough University, UK

Abstract:

In this paper, we compare and contrast financial data science with econometrics and conclude that the former is inevitably interdisciplinary due to the numerous skillsets needed within a competitive research team. The latter, in contrast, is firmly rooted in economics. Both areas are highly complementary, as they share an equivalent process with the former's intellectual point of departure being statistical inference and the latter's being the data sets themselves. Two challenges arise, however, from the age of big data. First, the ever increasing computational power allows researchers to experiment with an extremely large number of generated test subjects and leads to the challenge of *p-hacking*. Second, the extremely large number of observations available in big data sets provide levels of statistical power at which common statistical significance levels are barely a challenge. We argue that the former challenge can be mitigated through adjustments for multiple hypothesis testing where appropriate. However, it can only truly be addressed via a strong focus on the integrity of the research process and the researchers themselves, with pre-registration and actual out-of-sample periods being the best technical though in themselves potentially insufficient tools. The latter challenge can be addressed in two ways. First, researchers can simply use more stringent statistical significance levels such as 0.1%, 0.5% and 1% instead of 1%, 5% and 10%, respectively. Second, and more importantly, researchers can use additional criteria such as economic significance, economic relevance and statistical relevance to assess the robustness of statistically significant coefficients. Especially statistical relevance seems crucial in the age of big data, as it appears not impossible for an individual coefficient to be considered statistically significant when its actual statistical relevance (i.e. incremental explanatory power) is extremely small.

Acknowledgments: We are very grateful to Chris Adcock for providing us the opportunity to empower and develop research on financial data science and econometrics. Our thinking has benefited from discussions with Alexander Arimond, Damian Borth, John Cotter, James Hodson, Yanan Lin, Markus Koch, Valerio Poti and Pei-Shan Yu. Authors are alphabetically listed. All remaining errors are our sole responsibility. *Corresponding author email: andreas.hoepner@ucd.ie

Introduction

"Good with numbers? Fascinated by data? The sound you hear is opportunity knocking." These were the words of the New York Times when it announced "the Age of Big Data" on February 11th 2012.[1] According to Version 6.0 of the Data Never Sleeps report, nowadays it takes less than three minutes for a million tweets to be published, less than 20 seconds for a million Google searches to be conducted and less than four seconds for The Weather Channel to receive 1 million forecast requests.[2]

How do econometricians react to this newly found abundance in data? Some celebrate "the triumph of the empiricists" and announce "the birth of financial data science" (Simonian and Fabozzi, 2019, p. 10), while others warn of p-hacking – the process of arriving at superficially attractive and selective p-values through multiple hypothesis testing, whereby multiple may well mean millions or more (Chordia, Goyal and Saretto, 2017). While such data mining has probably always occurred in academic and professional finance research and similarly always found its critics, it has become much more attractive, more rewarding, and likewise much more dangerous in the age of big data. Even Harry Markowitz himself recently commented on the issue of data mining in the age of big data, stating with his co-authors Arnott and Harvey (AHM in the following):

"We are all data miners, even if only by living through a particular history that shapes our beliefs"

(Arnott, Harvey and Markowitz, 2019, p. 64)

Viewing data mining as an inevitable aspect of being an empirical financial researcher in line with AHM appears pragmatic and sensible. Yet, it implicitly cries out for much more academic research into the analytical measurement opportunities, statistical methods, and new financial products

[1] https://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html
[2] https://web-assets.domo.com/blog/wp-content/uploads/2018/06/18_domo_data-never-sleeps-6verticals.pdf

arising in and from the age of big data as well as the research process and researcher integrity needed to achieve robust conclusions. Furthermore, research on big data in finance needs to be connected with all the knowledge already in place, most notably in the area of financial econometrics. Consequently, in this paper we explore the degree to which the newly emerging financial version of the scientific enquiry into big data is complementary to econometrics and we discuss the opportunities and challenges that arise from the birth of this new research paradigm, which we call 'Financial Data Science'.[3]

Defining Financial Data Science

"[E]conometrics is measurement in economics … [and] financial econometrics will [consequently] be defined as the application of statistical techniques to problems in finance" (Brooks, 2002, p. 1). While the definition of econometrics is well established, embodying the process of measurement and "model-based statistical inference" (Campbell, Lo and MacKinlay, 1997, p. 3), formal definitions of financial data science are yet to emerge. To provide such a formal definition of financial data science, we contrast it with econometrics before subsequently discussing the complementary nature of the two fields.

Financial data science differs from econometrics in its intellectual point of departure, its process and its ambitions. While econometrics' intellectual point of departure is statistical

[3] We would like to note that we used the term 'Financial Data Science' as part of the 'Econometrics and Financial Data Science' workshop at the ICMA Centre of Henley Business School on November 2nd 2017, before the equivalently titled paper by Giudici (2018) or Simonian and Fabbozzi (2019: 10) announcement of "the birth of financial data science" in their first issue of The Journal of Financial Data Science published by IPR journals. Our use of the term 'Financial Data Science' indeed dates back to November 13th 2014, when one of our co-authors used it in a Henley Business School faculty viewpoint (Hoepner, 2014). He had been inspired by discussions he had earlier in the year with Damian Borth, who was a postdoc at the International Computer Science Institute (ICSI) of UC Berkeley at the time. https://www.henley.ac.uk/news/2014/financial-data-science-vs-financial-economics.

inference (i.e., the process), financial data scientists share a common interest in the data sets whose exploration and explanation can advance financial decision making (i.e., the ingredients). Despite the unstoppable move to big data, the availability of high quality (i.e., trustworthy) data sets remains the key practical constraint for the empirical researcher. Consequently, the desire to explain human behaviour through the exploration and critical assessment of new data sets is the common intellectual desire which unites financial data scientists whose statistical techniques can vary from the probabilistic regression models of financial economists to the neural network-based classification models of computer scientists.

To provide researchers with the best odds of explaining human behaviour with ever increasing data sets, financial data science is inevitably inter-disciplinary. In other words, the expertise and skills needed to insightfully extract information from unstructured data, to efficiently process several big datasets, and to design and execute effective statistical analysis, are so plentiful that it normally requires a financial econometrician, a computer scientist and an individual with deep knowledge about financial markets to design a competitive financial data science process. While two researchers maybe able to cover these three required skillsets, it is extremely rare that a given individual truly possesses all three. Consequently, financial data science is inevitably teamwork. To maintain a good interdisciplinary team spirit, it is paramount that no member of single discipline insists on the idiosyncratic attributes of their discipline (i.e., theoretical assumptions) being more worthy or truthful than another discipline's idiosyncratic attributes. Therefore, financial data scientists "minimize … [their] use of assumptions … [and] make every effort to empirically test these. In other words, while .. [others] tend to look at the world from their theoretical angle, financial data scientists .. undertake a deep investigation of

all available data to then arrive at conceptual explanations of what happens in the real world" (Hoepner, 2016, p. 2).

Similarly originating from the inevitably interdisciplinary process, the ambitions of financial data scientists stretch beyond the realms of economics to keep all members of the team highly integrated and motivated. The ReFine Principles of Financial Data Science launched in 2016, [4] for instance, include a clear opposition to any form of discrimination and endorsements of the sharing economy and an open source culture. Financial data scientists adhering to these principles also display a general support for science and aim to enlighten society by "leveraging financial and computer science for the broader good" (Financial Data Science Association, 2016, p. 1). In other words, while financial data scientists focus their work on data-driven research whose conclusions may have the possibility to advance financial decision making, their ambitions as a team are less focused on individual rent seeking and more on societal impact to sustain a strong team spirit.[5]

Consequently, we define financial data science as an interdisciplinary process of scientific enquiry, which is rigorously and repeatedly exploring and explaining the variance in all relevant data sets to advance financial decision making and thereby enlightening not only the interdisciplinary team of researchers but also society as a whole. In line with (Simonian and Fabozzi, 2019, p. 12), we argue "that financial data science is a discipline in its own right, and not merely the application of data science methods to finance", since the self-reinforcing yet mean-reverting nature of many financial markets produces distributions alien to classic data

[4] https://fdsaglobal.org/initiatives/refine-principles-of-financial-data-science/
[5] See also the Asilomar AI Principles which contain a similar team spirit focused ethos https://futureoflife.org/ai-principles/?cn-reloaded=1

scientists and hence require a distinct, interdisciplinary field of enquiry. Nevertheless, the emerging field of financial data science is inevitably complimentary to the intersection of data science and other disciplines (e.g., evidence-based medicine). We argue in the following section that it experiences a similar yin yang style complementary relationship with econometrics.

Yin Yang of Econometrics and Financial Data Science

While econometrics and financial data science differ in their intellectual point of departure (i.e. statistical techniques and data sets, respectively) and exhibit some further divergences largely due to the inevitably interdisciplinary nature of financial data science, the two fields have many more aspects in common than divide them. Both use econometric concepts and techniques, both fields develop their hypotheses informed by some form of economic theorising. Similarly, both are likely to make use of the wealth of newer and bigger data sets resulting from digitalisation and an increased willingness of commercial organisations to share their growing number of proprietary datasets with academics. And finally, both fields are likely to experience an increased practical relevance due to their analysis of bigger and more often proprietary datasets.

Hence, whereas econometrics has more emphasis on statistical inference and financial data science has more emphasis on big data processing, both fields share both concepts. Similarly, neural networks have been described for decades in advanced econometrics textbooks and the concept of explanatory power simultaneously represents the fit of the econometricians' model as well as the degree to which a financial data scientist understands the variation in the respective dependent variable. In other words, econometrics and financial data science represent two complementary perspectives on the same process. Hence, we argue that they enjoy a yin yang type relationship as displayed in Figure 1.

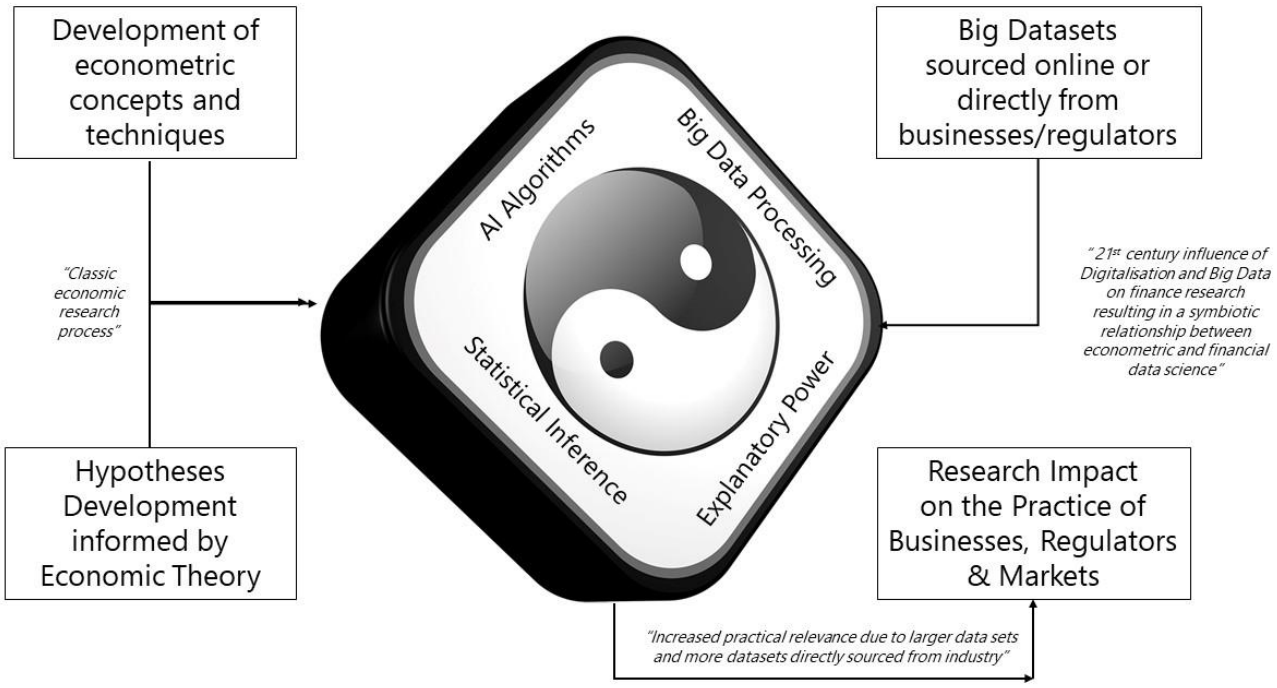

Figure 1: The Yin Yang (i.e. complementary relationship) of Econometrics and Financial Data Science

Notes: The Yin Yang symbol in the middle captures the most crucial aspects of econometrics and financial data science. The text in the boxes provides a formal description of the processes contributing to the complementary relationship between econometrics and financial data science. The text in quotation marks and italics provides practical commentary.

We continue by jointly exploring the implications of the ever increasing amount of human data for the fields of econometrics and financial data science while simultaneously introducing the contributions of this special issue. We commence by discussing new analytical measurement opportunities and new financial products arising from the age of big data. Subsequently, we discuss the challenges that big data impose on the financial economics research process and the resulting need for new research methods and processes to address these. We propose to extend the researchers' focus on statistical significance and economic relevance to also include statistical relevance and economic significance. We conclude with a discussion of the urgent avenues for future research in the fields of econometrics and financial data science at the advent of the age of big data.

Research Opportunities in the Age of Big Data

The most obvious implication of the age of big data is new datasets. A stunning example of such work in progress is (Fedyk and Hodson, 2018), who extract monthly information on career progress from over thirty million curricula vitaes of employees of US firms to investigate the impact of turnover and the skill level of human capital on firm performance. They observe that higher turnover hurts returns, which is intuitive but would previously have been studied only on much smaller and hence less generalisable datasets. Based on a similar text-extraction approach, (Goloshchapova *et al.*, 2019, p. 2) used "a battery of Python code … and … the latest R algorithm" to extract the topics discussed in over 5,000 corporate social responsibility (CSR) reports of more than thousand firms from 15 European countries between 1999 and 2016. They observe topic clustering at the sectoral level with, for instance, industrial firms displaying a bigger concern for employee safety and consumer firms being more engaged in topics such as 'food packaging'. While such results may seem intuitive to the reader, it is the relation of these 'big data statistics' to economic outcomes that presents the deeper appeal of financial data science.

Thng (2019) represents such a paper, which relates text-extracted information to abnormal returns. She first extracts the tone of 647 Initial Public Offerings (IPO) of US firms employing four separate sentiment measurement approaches.[6] She finds that VC-backed IPO have a less optimistic tone and explains this with concerns around litigation risk. Nevertheless, this defensive language does not appear to hamper performance. Much the opposite, Thng (2019) observes VC-backed IPOs to significantly outperform non VC-backed IPOs over longer horizons. A less expected but by no means less interesting application of novel datasets in financial data science is offered by Kumar *et al.*

[6] See Table 1 in her paper for a comparison of these techniques.

(2018). The authors obtained access to a rather unique proprietary dataset: over five million bank accounts with 250 million transactions belonging to clients over the age of 70. Such elderly clients may be the victims of fraud and hence protecting them reduces the operational risk of the major financial institution that provided the data on the condition of anonymity. Employing both logistic regression and classification techniques (support vector machines), the authors develop a new alert model that significantly advances beyond the practical status quo in terms of accuracy. Apart from being a relevant academic contribution and in practical terms representing a significant reduction in operational risk, this paper also received a very positive review from the Wall Street Journal.[7] In short, a truly innovative and successful financial data science research project that also displayed a positive impact on society.

Besides the exploration of new datasets, the ever increasing amounts of information in the age of big data also allow for a deeper exploration of previous overlooked research questions, either through a very large collection of various individual data sets or through a much deeper dive into previously less transparent subjects of analysis. A famous example of a very large collection of data sets is Moskowitz, Hua and Heje's (2012) study of time series momentum in 58 liquid security types. Similarly, Cotter and Suurlaht's (2019) study risk across various asset classes. They include credit risk, equity market risk, interest rate risk, interbank liquidity risk, and real estate market risk, and they find that spillovers between these are led by the equity and real estate markets, which supports the view that these have a special role in terms of financial stability.

Asimakopoulos, Asimakopoulos and Fernandes (2018) focus on a previously less transparent subject of analysis: unlisted firms. More specifically, they compare unlisted firms' cash holdings with those of listed firms and expect that unlisted firms will be more inspired by the precautionary

[7] https://www.wsj.com/articles/banks-monitor-older-customers-for-cognitive-decline-1542730606

principle, therefore holding more cash. The authors confirm their expectation based on a sample of more than hundred thousand Euro-area manufacturing firms. Another category of previously less studied subjects of analysis are those products that only exist as a consequence of bigger data sets and much faster computational processing capabilities. One class of these products is Exchange Traded Funds (ETFs). Studying a specific version of these ETFs – leveraged ETFs based on commodities – Del Brio, Mora-Valencia and Perote (2018) show that semi-nonparametric approaches to risk assessment can perform better than Gaussian approaches in backtests of expected shortfall. In fact, nonparametric approaches are themselves likely to experience a resurgence in popularity due to the classification focused nature of many machine learning approaches. Consequently, Jackson's (2019) theoretical contribution to this special issue is rather timely since the development of new tests and techniques further extends the toolbox available for data scientists to conduct analysis. In particular, non-parametric tests (such as Jackson, 2019) do not require distributional assumptions about the underlying data, a major advantage when there is still much debate over the generating process.

Research Challenges originating from the Age of Big Data

Based on increasing computational power, researchers such as Mclean and Pontiff (2016) or Jacobs and Müller (2019) conduct *'all-in'* studies of any relevant cross-sectional predictor of stock returns and the effect of academic publication on the very predictability of these factors. While McLean and Pontiff (2016) study "only" 97 predictors, Jacobs and Müller (2019) increase this number to 241. Conceptually, there is no theoretical limit to the number of cross-sectional predictors and or time series trading strategies than can be studied, and investment practice is a very willing audience for such kind of academic research. In fact, the increasing literacy of academic scholars

from various disciplines in programming languages such as Python and R is likely to make the occurrence of such *'all-in'* studies a regular sight in both academic and professional seminar series.

Increasing in the numbers game, Psaradellis *et al.* (2018) apply 7,846 technical trading rules to daily data of crude oil futures and the US Oil fund to a sample period of almost 10 years that provides them with significant statistical power. Employing controls for multiple hypothesis testing proposed by Romano and Wolf (2007) and Bajgrowicz and Scaillet (2012), they cannot find systematic, persistent abnormal returns to any of the technical trading rules. Taking the numbers game to the extreme, work in progress by Chordia, Goyal and Saretto (2017) generates 2.1 million trading strategies to evaluate the severity of p-hacking in finance research. They find that most rejections of the null hypothesis under single hypothesis testing disappear using a multiple hypothesis testing framework that accounts for cross-correlations within signals. They conclude that p-hacking is a serious concern for finance research, whose severity is substantially increased by the advent of the age of big data.

Consequently, researchers face the challenge that, due to increasingly large numbers of observations available, the conventional protocols for hypothesis testing are disrupted by shocks to the statistical power of the test datasets (i.e. extremely large number of observations) and shocks to the computing power of the researchers themselves (i.e., extremely large numbers of generated test subjects). While the computing power challenges are theoretically infinite following Moore's law, the statistical power challenge can be precisely illustrated based on the t-statistics that a correlation coefficient would have in a controlled laboratory setting depending on the number of observations. As shown in Table 1, theoretically true correlation coefficients of up to 2% would have t-statistics far below the critical values for conventional significance levels in case of sample sizes of 100 or even 1,000 observations. However, the same theoretically true correlation

coefficients would be declared statistically significant from 0.1% upwards for 10 million observations. This striking difference is neither caused nor helped by the fact that the critical values themselves decrease incrementally with the number of observations. In any case, it highlights how severe the statistical power challenge is in the age of big data. More worryingly, from 10 million observations onwards, regression coefficients are in many cases more likely to be declared highly significant at conventional significance levels than to be considered insignificant or mildly significant, even if they are in fact entirely inconsequential.

True correlation coefficient	t-statistics of correlation coefficients depending on number of observations					
	10 million	1 million	100,000	10,000	1,000	100
0.1%	3.1623	1.0000	0.3162	0.1000	0.0316	0.0099
0.2%	6.3246	2.0000	0.6325	0.2000	0.0632	0.0198
0.3%	9.4869	3.0000	0.9487	0.3000	0.0948	0.0297
0.4%	12.6492	4.0000	1.2649	0.4000	0.1264	0.0396
0.5%	15.8116	5.0001	1.5811	0.5000	0.1580	0.0495
0.6%	18.9740	6.0001	1.8974	0.6000	0.1896	0.0594
0.7%	22.1365	7.0002	2.2136	0.6999	0.2211	0.0693
0.8%	25.2990	8.0002	2.5299	0.7999	0.2527	0.0792
0.9%	28.4616	9.0004	2.8461	0.8999	0.2843	0.0891
1.0%	31.6244	10.0005	3.1624	0.9999	0.3159	0.0990
1.1%	34.7872	11.0007	3.4787	1.1000	0.3475	0.1089
1.2%	37.9501	12.0009	3.7950	1.2000	0.3791	0.1188
1.3%	41.1131	13.0011	4.1113	1.3000	0.4107	0.1287
1.4%	44.2762	14.0014	4.4276	1.4000	0.4423	0.1386
1.5%	47.4395	15.0017	4.7439	1.5000	0.4739	0.1485
1.6%	50.6029	16.0020	5.0602	1.6000	0.5055	0.1584
1.7%	53.7665	17.0024	5.3766	1.7001	0.5371	0.1683
1.8%	56.9302	18.0029	5.6930	1.8001	0.5687	0.1782
1.9%	60.0941	19.0034	6.0094	1.9002	0.6003	0.1881
2.0%	63.2582	20.0040	6.3258	2.0002	0.6319	0.1980

Table 1: Statistical power challenge as illustrated for the simple example of t-statistics of a Pearson correlation coefficient.
Notes: The t-stats have been computed as the true correlation coefficient multiplied by the square root of the degrees of freedom (i.e. observations minus two) scaled by the square root of the difference between one and the squared true correlation coefficient (see Weiss, 2012, pp. 696–697).

These statistical and computing power challenges require new thinking about research protocols and practices to allow researchers to explore the opportunities offered by ever faster computing and exponentially increasing amounts of data being produced, while simultaneously ensuring that the profession maintains its integrity.

New Research Practices to address Research Challenges in the Age of Big Data?

To address the issue of p-hacking resulting from the increasing ability of researchers to generate an extremely large number of usually interrelated test portfolios, Arnott, Harvey and Markowitz (2019) develop a "research protocol for investment strategy backtesting" including 22 questions in 7 sections. While some questions are - as one would expect from AHM – technical such as "[i]s the model resilient to structural change" or "[h]ave the researchers taken steps to produce the simplest practical model specification" (p.73), the vast majority of their questions are procedural if not philosophical and focus on the integrity of the research process.

For instance, AHM ask "[d]id the researchers take steps to ensure the integrity of the data?" Similarly, they question whether "the research culture reward[s] the quality of the science rather than the finding of a winning strategy". These questions about the integrity of the research process are crucial as adjustments for multiple hypotheses testing only work if researchers are transparent about each and every test they conducted. AHM even go beyond the integrity questions and suggest assessing the level of education of the researchers and their managers by asking if "the

researchers [are] aware that true out-of-sample tests are only possible in live trading" and if "researchers and management understand that most tests will fail" (p.73).

While there is clearly a very strong need to assess the integrity of researchers and their research processes in the age of big data, we do not fully understand how AHM aim to practically assess these integrity questions without relying on a researcher self-assessment format that may itself suffer from financial conflicts of interest. AHM recommend determining any relevant research design decision ex-ante before the formal research process has started but this sadly does not prevent researchers from conducting informal explorations to determine ex-ante research design set ups which are supportive of their subsequent formal research process. Further conceptual development and perhaps inspiration from other scientific disciplines such as medicine seem needed to address these significant challenges resulting from researchers' newfound abilities to create extremely large numbers of test portfolios.

Furthermore, the challenge of millions of test portfolios is reasonably specific to a selection of research questions such as the performance of investment strategies, while the challenge of an extremely large number of observations is likely to impact virtually any research question. In our view, an obvious response to this – apart from multiple hypothesis testing where applicable – is to sharpen the conventional significance levels required from 10%, 5% and 1% to 1%, 0.5% and 0.1%, respectively. Such a simple adjustment of expectations regarding statistical significance could be applied across research questions and would simply recognise that one can expect more robustness in conclusions from modern researchers who can see much further and/or in much more detail than previous generations.

However, we argue that an increased focus on concepts beyond statistical significance allows researchers to utilise the benefits of the age of big data while protecting themselves from the pitfalls. These concepts are economic significance, economic relevance and especially statistical relevance as outlined in Table 2. Crucially, while statistical power is vastly increased by the use of big data and hence the difficulty of achieving conventional significance levels (i.e. 5%) has dropped substantially, the remaining three concepts are not negatively affected by the advent of the age of big data.

Economic significance (i.e., the effect size itself) remains unaffected, while more data sets allow for a more seamless comparison of an effect size with other economic indicators. Similarly, economic relevance remains to be assessed by the relationship between the effect size and distributional properties such as the mean and standard deviation of the dependent variable. It probably also slightly benefits from bigger datasets, as these imply that the distributions of the dependent variable are estimated with incrementally increasing accuracy.

Since each individual coefficient's probability of being statistically significant increases with the substantially larger statistical power resulting from the use of big data, the statistical relevance of each coefficient is likely to become a more important assessment criterion for research quality. Statistical relevance can be measured as an incremental explanatory power (e.g. Adjusted R-squared, Shapley R squared) of adding the respective variable to the otherwise identical model. In the age of big data it appears not impossible for an individual coefficient to be considered statistically significant when its actual statistical relevance is negligible or even slightly negative. Consequently, we propose that researchers should discuss not only the statistical significance of the coefficients to key independent variables on which they build their narratives but also measure

and discuss statistical relevance as well as commenting on the economic significance and relevance of key coefficients.

Discipline (horizontal) | Concept (vertical)	Statistical	Economic
Significance	Conventional statistical significance levels of 1, 5% and 10% may need to strengthen to 0.1%, 0.5% and 1% given the vastly increasing statistical power of big data. A multiple hypotheses testing framework may need to be applied where relevant.	The effect size of estimated coefficients can be compared to a wider array of economic alternatives to determine their substance given increasing data availability
Relevance	Since individual coefficients' probabilities of being statistically significant increases with statistical power, their statistical relevance becomes crucial, which can be measured as the incremental explanatory power (e.g. Adjusted R squared) of adding the respective variable to an otherwise identical model	The effect size of the estimated coefficient can be seamlessly compared to the mean, standard deviation and skewness of the dependent variable distribution in the context of bigger data

Table 2: Implication of big data for the significance and relevance of empirical research results in statistical and economic terms

Concluding Thoughts

With millions of tweets being published in less than 10 minutes and millions of google searches being requested in less than one minute, we are living through the advent of the age of big data. Such a shock to the amount of available information appears to result in the emergence of a new research paradigm: financial data science. Acknowledging that we are currently experiencing the advent of the age of big data with ever increasing amounts of data produced on a daily basis, this brings into being exciting new opportunities for academic research which itself will evolve, if not

suddenly then at least gradually, in response to this new environment. We conclude on four aspects.

First and maybe most obviously, the strongly increasing computational power and the seamlessness of open source programming languages such as Python and R are likely leading to significant challenges for their commercial competitors. This process democratises access to statistical packages and economises on limited scholarly research funding, and is therefore beneficial even if it implies that many of us will gradually have to adapt our textbooks and taught courses.

Second, econometrics and financial data science are clearly complimentary fields and we are likely to see an increasing number of studies using innovative fact extraction-based datasets such as Fedyk and Hodson (2018) or Thng (2019) as well as many more *'all in'* studies of any relevant effect such as Mclean and Pontiff (2016) or Jacobs and Müller (2019). Studies such as Kumar *et al.*(2018), who use financial data science techniques to directly advance societal goals such as the protection of the elderly from fraud, are very welcome and hopefully also a more common sight at seminars going forward.

Third, the econometrics and financial data science research community may receive inspiration from medical research and collaborate to establish an institution such as the Cochrane Reviews for jointly synthesizing research results.[8] While the integrity of each individual researcher is hard to ensure, a community of researchers acting jointly should be able to keep itself accountable and thereby maintain its integrity. Pre-registration of research

[8] To the best of our knowledge, such an institution like the Cochrane Library (https://www.cochranelibrary.com/about/about-cochrane-reviews) solely dedicated to synthesizing research only exists in the medical discipline, where communicating the most likely best treatment of a given sympton to general practitioners in the light of conflicting results from empirical studies can potentially be a matter of life or death.

as practiced in much of medicine and psychology,[9] and as suggested by López de Prado (2019), for financial data science research, may also support integrity. Furthermore, we support the idea of actual out-of-sample periods of at least one year occurring past the pre-registration in addition to our proposal of an assessment of the statistical and economic significance and relevance of each key coefficient.

Finally, we clearly need much more engagement with performance management standards. AHM's (2019) protocol for backtesting hypothetical investment strategies is clearly a step in the right direction, but we need further thinking on how to address a new version of the old joint hypothesis problem. While pre-registering the research design and actual out-of-sample periods would certainly help, we might need to develop a research stream on the performance models themselves to avoid researchers registering weak performance models as often as some investment managers cherry pick custom benchmarks. But it is not only financial return models that we are concerned about. We are maybe even more concerned about our models of risk which often are actual models of deviation to both sides (e.g., variance and tracking error are used more often than semi-variance and trailing error, respectively). If risk is measured including upside and downside deviations from the mean (i.e. variance) or an index return (i.e. tracking error), then the researcher assumes that each investor "considers extremely high and extremely low returns equally undesirable" (Markowitz, 1959, p. 194). Since this assumption is incorrect for any profit maximizing investor and (Markowitz, 1959, p. 193) practical caveat that computing based on co-variance instead of variance requires "roughly two to four times as much computing time" does not apply anymore given 2019 computational power, the accurate measurement of risk or the implications of inaccurate measurement of risk appear fruitful avenues for further research in econometrics and financial data science.

[9] https://www.sciencemag.org/news/2018/09/more-and-more-scientists-are-preregistering-their-studies-should-you

References

Arnott, R. D., Harvey, C. R. and Markowitz, H. (2019) 'A Backtesting Protocol in the Era of Machine Learning', *The Journal of Financial Data Science*, 1(1), pp. 64–74. doi: 10.3905/jfds.2019.1.064.

Asimakopoulos, P., Asimakopoulos, S. and Fernandes, F. (2018) 'Cash holdings of listed and unlisted firms: New evidence from the euro area', *The European Journal of Finance, forthcoming*.

Bajgrowicz, P. and Scaillet, O. (2012) 'Technical trading revisited: False discoveries, persistence tests, and transaction costs', *Journal of Financial Economics*, 106(3), pp. 473–491. doi: 10.1016/j.jfineco.2012.06.001.

Del Brio, E. B., Mora-Valencia, A. and Perote, J. (2018) 'Expected shortfall assessment in commodity (L)ETF portfolios with semi-nonparametric specifications', *The European Journal of Finance, forthcoming*. doi: 10.1080/1351847X.2018.1559213.

Brooks, C. (2002) *Introductory Econometrics for Finance*. 1st edn. Cambridge University Press.

Campbell, J. Y., Lo, A. W. and MacKinlay, A. C. (1997) *The Econometrics of Financial Markets*. Princeton, New Jersey: Princeton University Press. doi: 10.2307/j.ctt7skm5.

Chordia, T., Goyal, A. and Saretto, A. (2017) *p-hacking: Evidence from two million trading strategies*, *Swiss Finance Institute Research Paper*.

Cotter, J. and Suurlaht, A. (2019) 'Spillovers in risk of financial institutions', *The European Journal of Finance*, pp. 1–28. doi: 10.1080/1351847X.2019.1635897.

Fedyk, A. and Hodson, J. (2018) 'Trading on Talent: Human Capital and Firm Performance', *SSRN Electronic Journal*. doi: 10.2139/ssrn.3017559.

Financial Data Science Association (2016) *ReFine Principles of Financial Data Science – FDSA*, *December 14 2016*. Available at: https://fdsaglobal.org/initiatives/refine-principles-of-financial-data-science/ (Accessed: 29 July 2019).

Goloshchapova, I. *et al.* (2019) 'Corporate social responsibility reports: topic analysis and big data approach', *The European Journal of Finance*, pp. 1–18. doi: 10.1080/1351847X.2019.1572637.

Hoepner, A. G. F. (2016) *Financial Data Science for Responsible Investors. Forthcoming 10-4-10 Anniversary Book, Environmental Agency Pension Fund, Bristol, UK*.

Jackson, R. H. G. (2019) 'Sub-sequence incidence analysis within series of Bernoulli trials: application in characterisation of time series dynamics', *The European Journal of Finance, forthcoming*. doi: 10.1080/1351847X.2019.1583117.

Jacobs, H. and Müller, S. (2019) 'Anomalies across the globe: Once public, no longer existent?', *Journal of Financial Economics, Forthcoming*. doi: 10.1016/j.jfineco.2019.06.004.

Kumar, G. *et al.* (2018) 'Can alert models for fraud protect the elderly clients of a financial institution?', *The European Journal of Finance, forthcoming*. doi: 10.1080/1351847X.2018.1552603.

López de Prado, M. M. (2019) *Advances in financial machine learning*. 1st edn. Hoboken, New Jersey: John Wiley and Sons, Inc.

Markowitz, H. (1959) *Portfolio Selection: Efficient Diversification of Investments*. 1st edn. New York: John Wiley and Sons.

Mclean, R. D. and Pontiff, J. (2016) 'Does Academic Research Destroy Stock Return Predictability?', *Journal of Finance*, 71(1), pp. 5–32. doi: 10.1111/jofi.12365.

Moskowitz, T. J., Hua, Y. and Heje, L. (2012) 'Time series momentum', *Journal of Financial Economics*. Elsevier, 104(2), pp. 228–250. doi: 10.1016/j.jfineco.2011.11.003.

Psaradellis, I. *et al.* (2018) 'Performance of technical trading rules: evidence from the crude oil market', *The European Journal of Finance, forthcoming*. doi: 10.1080/1351847X.2018.1552172.

Romano, J. P. and Wolf, M. (2007) 'Control of generalized error rates in multiple testing', *Annals of Statistics*, 35(4), pp. 1378–1408. doi: 10.1214/009053606000001622.

Simonian, J. and Fabozzi, F. J. (2019) 'Triumph of the Empiricists: The Birth of Financial Data Science', *The Journal of Financial Data Science*, 1(1), pp. 10–13.

Thng, T. (2019) 'Do VC-backed IPOs manage tone?', *The European Journal of Finance, forthcoming*. doi: 10.1080/1351847X.2018.1561482.

Weiss, N. A. (2012) *Introductory Statistics.* 9th edn. Boston: Addison Wesley/Pearson Higher Ed.