Paper to be presented at DRUID18
Copenhagen Business School, Copenhagen, Denmark
June 11-13, 2018

The Potentials of Machine Learning and Big Data in Entrepreneurship Research - The Liaison of Econometrics and Data Science

Daniel S. Hain
Aalborg University
Department of Business and Management
dsh@business.aau.dk

Roman Jurowetzki
Aalborg University
Business and Management
roman@business.aau.dk

Abstract

While the potential data landscape in econometric research is undergo- ing dramatic changes, terminologies such as ?Big Data? and the associated discipline of ?Data Science? and ?Machine Learning? techniques have so far received little attention among econometricians. In this chapter, we illustrate the potentials for quantitative entrepreneurship research to benefit from the availability unprecedentedly rich datasets and non-traditional data sources such as text, audio, or image data. However, we also highlight that such datasets are in need of new approaches. Both computational and epistemological. We proceed with introducing the Data Science approach to quantitative analysis which is geared towards optimizing the predictive performance, contrast it with standard practices in econometrics which focus on producing good parameter estimates. This chapter also introduces to machine learning techniques such as out of out-of-sample model verification, variable selection, and generalization procedures, and finally the popular model classes of regression trees, artificial neural networks, and vector space models. We provide guidance on how to apply these techniques for quantitative research in entrepreneurship and point towards promising avenues of future research which could be enabled by the use of new data sources and estimation techniques.

The Potentials of Machine Learning and Big Data in Entrepreneurship Research*

– The Liaison of Econometrics and Data Science –

1 Introduction

While the potential data landscape in econometric research is undergoing dramatic changes, terminologies such as "Big Data" and the associated discipline of "Data Science" and "Machine Learning" techniques have so far received little attention among econometricians. In this chapter, we illustrate the potentials for quantitative entrepreneurship research to benefit from the availability unprecedentedly rich datasets and non-traditional data sources such as text, audio, or image data.

However, we also highlight that such datasets are in need of new approaches, both computational and epistemological. While "Big Data" has become somewhat of a buzzword since 2013, we argue that it is particularly the variety of accessible data, coupled with increasing volume, which can provide inputs for quantitative research of entrepreneurial phenomena that were hardly approachable by quantitative techniques in the past. Yet, the extraction and processing of such data, from online sources or the emerging Internet of things requires the understanding of Data Science approaches, and the analysis at least to some extend machine learning techniques.

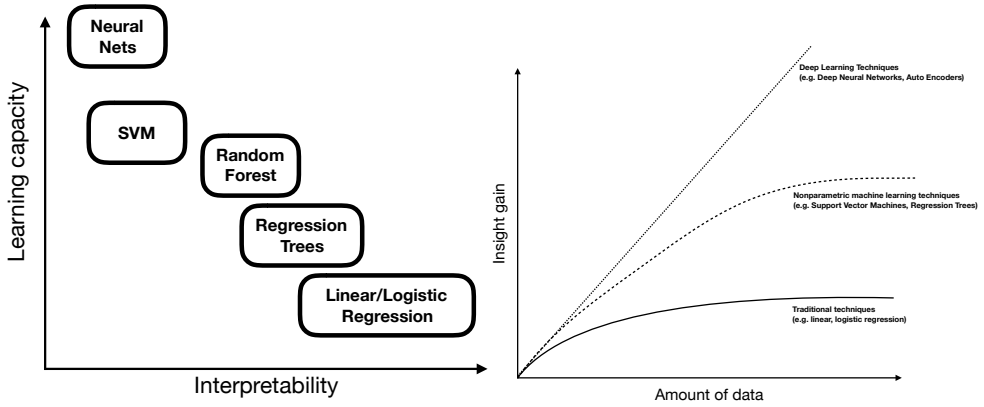
We proceed with introducing the Data Science approach to quantitative analysis which is geared towards optimizing the predictive performance, contrasting it with standard practices in econometrics which focus on producing good parameter estimates. We discuss the potential synergies between the two fields against the backdrop of this at first glans "target-incompatibility".

This chapter also introduces to machine learning techniques such as out of out-of-sample model verification, variable selection, and generalization procedures, and finally the popular model classes of regression trees, artificial neural networks, and vector

*Manuscript prepared for the DRUID18 conference, Copenhagen Business School, Denmark

space models. We provide guidance on how to apply these techniques for quantitative research in entrepreneurship and point towards promising avenues of future research which could be enabled by the use of new data sources and estimation techniques.

This chapter deals also with some central trade-off problems that arise when thinking about model selection. When it comes to comparing the performance of different model classes that are available to social science scholars, we can find many dimensions on which the features of techniques may be accessed. In figure 1 we depict two trade-offs that we find relevant to consider in a paradigmatic discussion of data science and econometric approaches. On the one hand, and as presented in Figure 1a, there is a general trade-off between the learning capacity of model classes and their interpretability. The relationships between inputs and outputs captured by a linear regression model are easy to understand and interpret. As we move up and to the left in this chart, the learning capacity of the models increases. Considering the extreme case of deep neural networks, we find models that can capture interactions and nonlinear relations across large datasets, fitting in their complex functions between in- and outputs across the different layers with their multiple nodes. However, for the most part, it is fairly difficult if not impossible to understand the fitted functional relationship. This is not necessarily a problem for predictive modeling but of much use in cases where the aim is to find causal relationships between in- and outputs. Below we discuss these challenges in detail and outline potential strategies for choosing and combining techniques.



(a) Learning capacity vs. interpretability of selected quantitative methods

(b) Gain of insight vs. Data amount for different model classes

Figure 1: Learning capacity, amount of data and interpretability for different modeling techniques

The other central relationship that should be considered by social science scholars concerns the amount of data. Having more data and utilizing sophisticated econometric techniques has sometimes been equated with doing better research. We argue that the

utility of greater data amounts not only depends on the relationship that is studied but also on the applied techniques. While the assumption of the more, the better is overall valid for newer machine learning models such as deep neural networks and to some extent for non-parametric approaches such as support vector machines, more traditional models only benefit from more data to a certain degree, from which on the marginal gains from having more data flatten. This discussion is taken up and explored in detail in Section 3.

Section 4 identifies potential avenues to combining techniques from econometrics and machine learning and presents examples of recent empirical work that has done that. This section, together with the conclusion, outline various technical and institutional challenges that have to be overcome to benefit from the potentials.

2 New Data Sources

"*There was five exabytes of information created between the dawn of civilization through 2003, but that much information is now created every two days, and the pace is increasing.*"
–Eric Schmidt, former CEO of Google, 2010

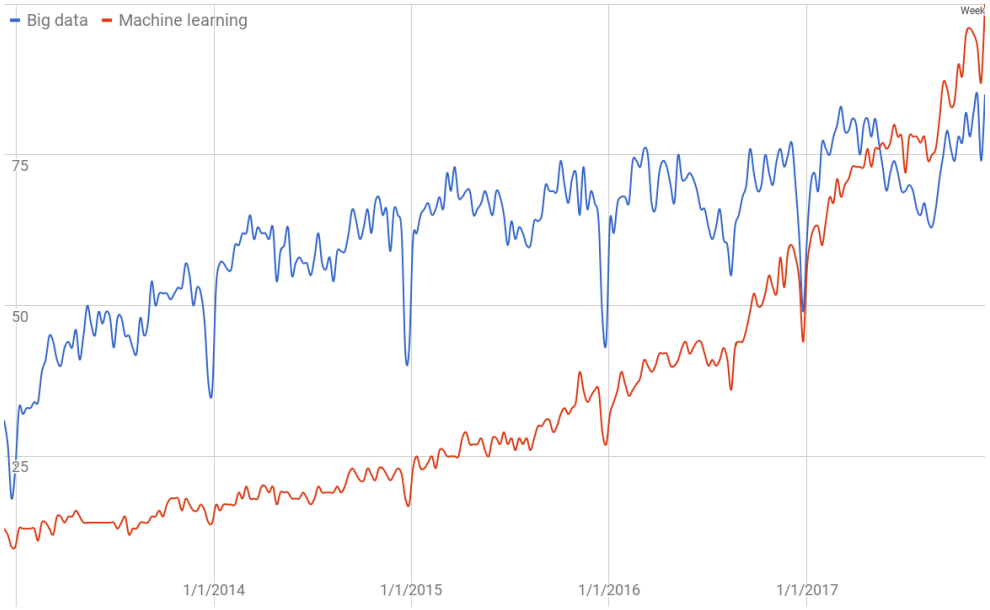
Even twenty or thirty years ago, data on economic activity was relatively scarce. In just a short period this has changed dramatically. Within a very short time, the amount of data one can submit to analysis became big and rich, in quantity and quality alike. Not only have technological developments led to the structured documentation and storage of information about existing economic activity that was not available before, but we have also witnessed the evolution of new digital business models (Chen et al., 2012; McAfee et al., 2012) – think sharing economy ventures – that rely on and generate highly granular data on new forms of economic, often entrepreneurial, activity. In the following, we briefly discuss sources of big data, its implications for statistical analysis, and point towards promising areas where it can be applied in quantitative entrepreneurial research.

2.1 On Big Data

The discussion on "Big Data" has been growing a lot around 2013, and often the 3 Vs – Volume, Variety, Velocity – have been recited like a mantra when talking about this supposedly new kind of data. The first V, Volume, has received arguably most attention with giga-, tera-, and petabytes being used to talk about the "how much" of big data. The simple answer: Data on a scale so large that we cannot manage, analyze and understand it with techniques we developed to make sense of smaller data in the past. This also seems a more appropriate definition given that the recent machine learning discussion (2015 and on) is more interested in the analysis aspects rather

than storage and mapping architecture questions. Figure 2 shows the development of Worldwide Google searches for *Big Data* and *Machine Learning*. The level of interest in *Big Data* has been strongly growing throughout 2013 but slowed down in the following years. *Machine learning* as a search term experienced a stronger and steadier growth surpassing *Big Data* for the first time in 2017. An additional insight from this data is also that for both topics, the majority of search requests come from China.

Figure 2: Worldwide Google searches for the topics *Big Data* and *Machine Learning* 2013 - 2017, Google Trends



When it comes to analytics to much data can become a problem technically with some model classes being better suited for volume than others. Artificial neural networks have been seen as well suited to process volume as they are trained sequentially and do not require the whole dataset to be loaded into memory. Recent developments in memory hardware – potentially soon eroding the difference between memory and disk storage – may change these current technical restrictions in the foreseeable future. The question is therefore not: How much data can this model take? But rather: Which model is better suited to produce sensible results given a certain amount of data.
In an analysis context the second V, Velocity, has to be seen against the backdrop of prediction from trained models and sequential or transfer learning. However, in the research environment, this aspect is not of particular interest.
More interesting is the third V, Variety. Modern forms of communication, commerce, and information consumption lead to a constant generation of data in various ways. Logs of all kinds of interactions produce rich network data: Who has been in contact with whom? How much? About what? When? This relational data adds a new

dimension to data and can be relatively easily obtained. The majority of the data produced today is unstructured data, such as natural language or images, adding to the variety. Natural language processing allows increasingly to make use of information contained in text while deep learning powered computer vision lets us access image data. The question here is therefore: Which models or combinations of approaches are best suited to make sense of this variety of data on different dimensions?

2.2 Sources and Applications of Big Data

2.2.1 Large Scale Register Data

Large-scale administrative data sets and proprietary private sector data can considerably improve the way we measure, track, and describe economic activity. They can also enable novel research designs that allow researchers to trace the consequences of different events or policies (Einav and Levin, 2014a). While in the past mostly Scandinavian researchers enjoyed access to population-wide and fine-grained firm and even personal data on educational and employment history, health data (and greatly capitalized on that in management, entrepreneurship, labor- and health-economics research), the access to comparable datasets becomes increasingly common in many developed and also developing countries Einav and Levin (2014b).

2.3 Online data generally

The data presented in Figure 2 is a simple example of (big) online data. Here Google openly provides anonymised regionally aggregated data of search requests for arbitrary terms through its service *Google Trends*. The company offers two other interesting data sources for social science researchers: *Google Correlate* and *Google Consumer Surveys*. The former one allows identifying country-specific search requests that are correlated with a term or an arbitrary time-series that one can upload. Correlate allows for lags and different time frames. The latter is a tool, created for market research, but has also been used by social science scholars - mainly in economics an political science. Stephens-Davidowitz and Varian (2015) provide a hands-on guide to using these data sources in social science contexts.

Social media provides a rich data source. Some platforms, for instance, Twitter are considerably open – reflecting the type of communication they offer – while others, e.g., Facebook, are more restrictive when it comes to data access. Twitter allows access to user-timelines of open profiles, as well as followship information. Thus, one has access to up to 3200 recent tweets, user meta-data as well as the option to construct user-networks from followship or mentions patterns.[1] In a recent project, the authors,

[1]The Twitter API documentation can be found at `https://developer.twitter.com/en/docs/api-reference-index`.

used such data to identify the structure of the entrepreneurial ecosystem that supports the digital startup scene in Nairobi, Kenya. These type of data can be very useful in contexts of emergence or data scarcity, or both as in the mentioned case.
LinkedIn can be used as a source of professional information on individual or company levels. Detailed but sometimes sparse data describing individuals (e.g. educational and career tracks, endorsed skills) is complemented by a rich network structure.
More specific social networks can provide insight into interaction in particular groups. A less known but interesting example is https://nomadlist.com, a social network for remote workers that lists cities, co-working spaces, and individuals with their travel itineraries. Not only does the data allow to identify movement patterns of individual remote workers but also their "friendship "networks between them, indicating a certain level of interaction. An interesting feature of this database is that the usernames are twitter handles, meaning that the data can be easilly aggregated with twitter data on an individual level. Exploring such datasets may help to identify location independent entrepreneurship models, as well as individuals forming such specific samples.

Another specialized network is GitHub, a platform for collaborative software development with extensive interactive functions. Entrepreneurship scholars may use data extracted from this platform to spot trends in digital entrepreneurship, collaboration patterns and other developments related to innovation in software and data-analytics.

Crowdfunding platforms such as *kickstarter* provide another fascinating source of data. The rich descriptions of proposed ventures, the entrepreneurial team and the detailed information on the funding process allow studying finance processes at an unprecedented level of granularity. Access to these data and other crowdfunding data can be gained using their APIs, scraping approaches, as well as through the Crowd-Berkeley database https://crowdfunding.haas.berkeley.edu/wp/ a project that conveniently aggregated crowdfunding data 2005-2016.

Finally, one should also mention organizations, commonly associated with the sharing economy, such as Airbnb. Activity on and through these platforms may be understood as a form of entrepreneurship, and given the way how transactions are moderated through apps and online platforms, a lot of data documenting these transactions becomes available. For the example of Airbnb, the page http://insideairbnb.com provides access to large publicly available datasets scraped from the Airbnb page with the intention to document arguably negative effects from the renting activity for neighborhoods and communities. Often, pages such as *insideairbnb* provide best-practice guides to scraping and curation of online data. Generally, supported by packages and

wrappers in high-level programming languages, web-scraping and -crawling turns out to be easier than most might imagine.[2]

Many of the above-mentioned platforms provide a comfortable way to access the data using application programming interfaces (APIs) that commonly return responses in XML or JSON format. The latter format is the currently most popular format for transfer of hierarchical data between devices. APIs are usually developed for the interaction of applications with online servers, and thus data retrieval requires some understanding into the functioning of HTTP requests. Also, some of the necessary skills include insights into storage and parsing of these datatypes. Document-oriented "NoSQL" (standing for not-only-SQL) databases such as the open source MongoDB, using a local installation, have proven to be useful solutions to storage and structured retrieval of online data.[3]

At this point, we should mention that researchers which intend to use crawling and scraping techniques to extract online data should make themselves familiar with the particular legal issues that may arise. These may relate to different aspects such as copyright, personal data protection or terms of service.

2.4 Non-traditional data formats

The main share in the recent growth of available data can be found within unstructured data – text, images, audio, video. Devices from the evolving Internet of Things (IoT) are also a new source of big data, often including temporal and geospatial dimensions. While the latter type of data can usually be used for more traditional quantitative analysis after some traditional preprocessing, and given that one is familiar with spatiotemporal approaches, using unstructured data is not straightforward.

When it comes to text, the recent decade has seen a renewed interest in Natural Language Processing (NLP). The central aim of NLP is to make machines analyze natural language using various approaches. The approaches vary widely in terms of their sophistication, ranging from simple word counts, over rule- and grammar-based approaches, to vector space models (which we discuss in Section X), and more recently word embeddings such as Word2Vec (Mikolov et al., 2013) or GloVe (Pennington et al., 2014). Vectorization of, at the onset of the approach, words and later sentences, and paragraphs crystallized out as the dominant paradigm, outperforming previous approaches at traditional NLP tasks such as dependency parsing, sentiment analysis and named entity recognition. In addition, word vectors showed – somewhat surprisingly

[2] Common packages for web-scraping are `rvest` for R, and `scrapy` for Python. To process the "harvest" of such activities (cleaning) of HTML data ect.), packages such as `BeautifulSoup` become handy.

[3] `https://www.mongodb.com`. `Pymongo` for Python and `Mongolite` for R allow to connect to the database

– to carry latent semantic features. Algebraic operations performed on these vectors allow today for many interesting applications, but most importantly text is represented as an n-dimensional vector.

The representation of unstructured data in the form of vectors seems to become the dominant paradigm also for other unstructured data types. For images, convolution has become the standard for transforming 2-dimensional pixel matrices into n-dimensional vectors that can be passed on into neural networks Krizhevsky et al. (2012). Recent examples where image data has been used to estimate demographic and economic outcomes can be found in Glaeser et al. (2016) and Gebru et al. (2017). The later contribution estimated socioeconomic attributes such as income, race, education, and voting patterns from cars detected in Google Street View images for 200 cities in the US on street and neighborhood-level. Such estimates can then be used as extremely granular inputs for more traditional modeling.

The field of machine learning has developed potent approaches to structuring of unstructured data and used these representations to predict some classes or values for individual observations. Quantitative entrepreneurship research that aims at causal inference may benefit from these developments by including inputs resulting from such predictions. Thereby unstructured data can be used to study phenomena at a level of detail where data was yet not available or apply quantitative methods to contexts where this was not an option at all.[4] However, we should also acknowledge that despite the availability of new data sources and methods to making use of them, it will require innovation in the way how entrepreneurship and management science researchers are socialized in doctoral training to benefit from these innovations (George et al., 2016).

The Promises of Big Data for Entrepreneurship Research The high dimensionality and general richness of Big Data, combined with the ability the fit flexible and complex functional forms also allows us to generate variables previously not, or only to a limited extent, available for an econometrician. We here distinguish between three kinds of data sparsity: (i.) physical environments, in which traditionally used dependent and independent variables are no available; (ii.) conceptual or physical systems for which their structure and dynamics are not or only insufficiently captured by traditional data sources, (ii.) theoretical environments, where appropriate measures for theoretical constructs are not available.

To start with, traditional sources, such as public register data, are often not freely accessible, when then only in highly aggregated form, or simply not gathered at all.

[4]For the interested reader, the exhaustive survey on big data sources for social and economic analysis by Blazquez and Domenech (2017) can be recommended.

While this is true across many western countries, the lack of official data on basic economic activity is particularly prevalent in many developing countries. Here, the use of satellite data, where petabytes of high-resolution images with global coverage have become available, gain popularity among economists. Given the knowledge and infrastructure to process and analyze such data, it allows collecting fine-grained panel data at low marginal costs, which can be used to construct proxies for a wide range of hard-to-measure characteristics. For instance, Jean et al. (2016) illustrate how a convolutional neural network can be trained to identify satellite image features that can explain up to 75% of the variation in local-level economic outcomes while Henderson et al. (2012) use a similar procedure to produce a granular map of growth on regional level in a selection of countries in sub-Saharan Africa, Lobell (2013) predict future crop yield with satellite imagery, and Ernst and Jurowetzki (2016) use night images to identify black markets at the North Korean border and analyze their impact on human trafficking. Even more granular, Dong et al. (2017) among others use high-resolution satellite imagery to estimate growth at the firm level. Zooming further in, the use of Google StreetView data has recently been utilized to measure city characteristics and outcome variables such as income at higher collection frequencies and more granular geographic scales than ever before (e.g., Gebru et al., 2017; Glaeser et al., 2016). Another source of non-traditional granular data comes from mobile phone usage, which can be used to make inference on socioeconomic status, and assess the wealth distribution from the country to the household scale (Blumenstock et al., 2015; Blumenstock, 2016)

However, the combination of non-traditional and high dimensional data sources with machine learning methods has not only proven useful in data sparse environments lacking in comprehensive public records, but also in areas where the quantitative data we traditionally use appears to be rather limited to assess and predict quality indicators related to the unit of observation. For instance, one could be interested if a certain regional policy attracts potentially high performing firms, or if some educational measures lead to more high-quality entrepreneurial activity. Here, public register data on the number of firms founded in a region tells us little about their quality, and to assess the quality in measures such as firm growth or patent applications, we would have to wait for quite some time. Here, the high predictive power of many machine learning techniques fed with a lot of data becomes helpful. Such predictions, as exhaustively discussed, tend to be non-causal in nature. Yet, if these estimates turn out to be of high accuracy, we see no obvious reason why such predicted values could not be utilized as inputs for models aiming at causal inference, either as independent or dependent variables. Examples for well performing non-causal predictions are provided by Droll et al. (2017), who apply various forms of web-scraping techniques combined

with NLP to predict firm-level growth potential,[5] and the deep neural network-based prediction models of firm performance by Lee et al. (2017). de la Paz-Marín et al. (2012) employ a Product-Unit Neural Network models both trained by evolutionary algorithms to predict research and development performance in 25 European countries, and Hajek and Henriques (2017) a self-organizing map to visualize and predict innovation performance on regional level. Such combinations of prediction and causal inference techniques offer the potential for granular and timely analysis of phenomena which currently cannot, or only to a limited extent, be addressed using traditional techniques and data sources. The "Startup Cartography Project" at the MIT (Andrews et al., 2017; Fazio et al., 2016; Guzman and Stern, 2015, 2017) provides a good example of such efforts. Coining it "nowcasting" and "placecasting", the project uses large amounts of business registration records and predictive analytics to estimate entrepreneurial quality for a substantial portion of registered firms in the US (about 80%) over 27 years. When inspecting the most relevant predictors (sector controls, firm has patents, firm has trademarks *et cetera*) it pretty soon becomes clear that they are non-causal, yet the model predicts very accurately out-of-sample. Such "predictions in the service of estimation" (Mullainathan and Spiess, 2017), it is not hard to see how these predictions of start-up quality might serve as dependent or independent variables in many interesting hypothesis-testing settings. Further, such predictors might also ease the construction of variables attempting to capture complex socio-techno-economic phenomena, for which the selection of a single hard-coded measurement appears challenging or impossible. For example, Kwon et al. (2017) use a latent semantic analysis (LSA) to measure and visualize the social impact of emerging technologies, and Akgün et al. (2010) a rough set data analysis (RSDA) to identify the critical factors that determine the embeddedness level of rural entrepreneurs.

Furthermore, big and granular data, particularly on interaction pattern, enables us to map– and eventually analyze– the structure and dynamics of complex socio-techno-economic constructs such as entrepreneurial ecosystems or technological systems, which are hard to grasp relying solely on traditional sources of data. In the case of Kenya, Park et al. (2017) use a combination of CrunchBase, Linkedin, and Twitter data to identify key actors and institutions in Nairobi's digital economy, and map the interaction between them. Wang et al. (2017) use a similar methodology to depict entrepreneurial networks in the United States. Kim et al. (2016) uses LDA and fuzzy cognitive maps (FCMs) to create narrative future scenarios on the development of certain technologies, and (Jurowetzki and Hain, 2014) use a socially enhanced web

[5]Data generated from such web-scraping techniques currently start to get recognized as potential valid proxies for specific variables obtained from more classical methods such as surveys Beaudry et al. (2016); Li et al. (2016); Te and Cvijikj (2017).

search, NLP entity extraction, and evolutionary dynamic analysis to map the evolution of technology.

3 New Methods

While there is quite some common ground in the econometrics and machine learning approach to data analysis, there are also substantial differences in the logic, methods, process, and terminology, making the transition between the two approaches less smooth than it might be. In the following, we discuss the fundamental – if not epistemological – differences, their origins, as well as if and how they matter. We proceed with introducing some popular and illustrative supervised machine learning techniques econometricians might already be acquainted with. Finally, we introduce some currently promising machine learning techniques that may be less familiar to econometricians, namely artificial neural networks and vector space modeling approaches. Since techniques and approaches in machine learning are numerous (cf. figure 6 in the appendix for a first glance) this is not an attempt to provide an exhaustive overview, which is done excellently elsewhere (Blazquez and Domenech, 2017) anyhow. Our aim is rather to illustrate the general logic, and point towards some new approaches econometricians might also find useful.

3.1 The Econometric vs. the Data Science approach

As applied econometricians –in entrepreneurship research and elsewhere– we are for the most part interested in producing good *parameter estimates*. We construct models with unbiased estimates for some parameter β, capturing the relationship between a variable of interest x and an outcome y.[6] Such models are supposed to be "structural", where we not merely aim to reveal correlations between x and y, but rather a causal effect of directionality $x \rightarrow y$, robust across a variety of observed as well as up to now unobserved settings. Therefore, we carefully draw from existing theories and empirical findings and apply logical reasoning to formulate hypotheses which articulate the expected direction of such causal effects. Typically, we do so by studying one or more bivariate relationships under *cetris paribus* conditions (everything else equal). We implement this by building a regression model, which we "hand-curate" with a set of supposedly causal variables of interest while controlling for further variables known to influence the level of our outcome y. The primary concern here is to minimize the standard errors ϵ of our β estimates, the difference between our predicted $hat(y)$ and

[6] For the sake of illustration, we here portray the "archetypal" econometrician and data scientist. Since we are well aware of in-group heterogeneity as well as increasing efforts at the intercept of disciplines, we upfront apologize for offending everyone who not neatly fits in the provided categories.

the observed y, conditional to a certain level of x. We are less interested in the overall predictive power of our model (Usually measured by the models R^2), as long as it is in a tolerable range.[7] However, we are usually worried about the various type of endogeneity issues inherent to social data which could bias our estimates of β. For instance, when our independent variable x can be suspected to have a bidirectional causal relationship with the outcome y, drawing a causal inference of our interpretation of β is obviously limited. Are entrepreneurs more successful because of their extensive social network, or do they have an extensive social network because they are successful entrepreneurs? Hard to say, probably a bit of both. To produce unbiased parameter estimates of arguably causal effects, we are indeed willing to sacrifice a fair share of our models' explanatory power.

A good example here is the popular use of instrumental variables in two-stage least square regressions (2SLS) to tackle endogeneity issues in our structural model. Here, we first regress $x = \gamma' z + \delta$, where we predict the supposedly endogenous x by some carefully chosen instrument z, which can be expected to have predictive power over x but is otherwise exogenous to the model. In a second step, we use the predicted $\hat{x}$ to regress $y = \beta' x + \epsilon$. If our selected instrument z is indeed uncorrelated to ϵ and y while also doing a reasonable job in predicting x, our β' indeed expresses the strength of a unidirectional relationship $x \rightarrow y$. Here, it is obvious that a model using the original values of x would do a better job in prediction y.[8] Anyhow, our estimated β captures a structural effect which can be, depending on the representativeness of our sample, generalized to a larger population.

A data science approach to statistical modeling is, however, fundamentally different. To a large extent driven by the needs of the private sector, data analysis here concentrates on producing trustworthy predictions of outcomes. Familiar examples are the recommender systems employed by companies such as Amazon and Netflix, which predict with "surprising" accuracy the types of books or movies one might find interesting. Likewise, insurance companies or credit providers use such predictive models to calculate individual "risk scores", indicating the likelihood that a particular person has an accident, turns sick, or defaults on their credit. Instances of such applications are numerous, but what most of them have in common is that: (i.) they rely on a lot of data, in terms of the number of observations as well as possible predictors, and (ii.) they are not overly concerned with the properties of parameter estimates, but very rigorous in optimizing the overall prediction accuracy. The underlying socio-psychological forces which make their consumers enjoy a specific book are presumably only of minor

[7] At the point where our R^2 exceeds a threshold somewhere around 0.1, we commonly stop worrying about it.

[8] This approach is popular in cross-sectional settings. When having the convenience of panel data, we often do not bother with such procedures and circumvent endogeneity issues just by using the lagged value x_{t-1}. However, the argument also holds.

interest for Amazon, as long as their recommender system suggests them books they ultimately buy.

Until recently, the community of applied econometricians – and more broadly, quantitative researchers in social science – was not overly eager to embrace and apply the methodological toolbox and procedural routines developed within the discipline of data science. An apparent reason is given by inter-disciplinary boundaries and intra-disciplinary methodological "comfort zones" (Aguinis et al., 2009) as well as by path-dependencies, reinforced through the way how researchers are socialized during doctoral training (George et al., 2016). However, as sketched before, there also seems to be inherent – if not epistemological – tension between the econometrics and the data science approach to data analysis, and how both could benefit from each other's insights is not obvious on first glance. Indeed, could an apparatus fully geared towards prediction be of use for deductive or even inductive theory testing and development, the main method we apply on our quest to unveil the underlying structural forces governing economies and societies?

We argue it does, and when using this toolbox in the right way, there are quite some "tricks" an econometrician might find extremely useful (Varian, 2014). We expect such methods to broadly diffuse within quantitative social science research, and suggest the upcoming liaison of econometrics and data science to shake up our current routines. However, before discussing the potentials as well of challenges for work at the disciplinary intercept, we consider a brief introduction to the main principles, approaches, and methods in data science to be necessary.

While the broader discipline of data science covers the whole value chain of collecting, storing, processing analyzing and visualizing data, we here focus on data analysis and the associated sub-discipline of machine learning. Broadly, we can divide machine learning methods into two categories, supervised and unsupervised machine learning. Plainly speaking, in supervised machine learning, we have an observed outcome y and fit a model that predicts this outcome well. In contrast, for unsupervised machine learning tasks, there exists no observed and true outcome y on which we can fit a model. Typical applications are clustering, pattern recognition, and dimensionality reduction. Since most prediction exercises similar to the "bread-and-butter" work of an econometrician fall into the category of supervised machine learning, we naturally start our introduction here.

3.2 Supervised Machine Learning 101

3.2.1 General idea

At its very core, in supervised machine learning, we seek for models and functions that do the best possible job in predicting some output variable y. This is done

by considering some loss function $L(\hat{y}, y)$, such as the popular root-mean-square error (RMSE),[9] and then searching for a function $\hat{f}$ that minimizes our predicted loss $E_{y,x}[L(\hat{f}(x), y)]$. To warm up, lets consider the simplest case, an ordinary least squares (OLS) regression, with a given functional form of $f(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$. If we do not have to care about asymptotic properties, causality and the p values of our β, minimizing the RMSE is simply a task of selecting a set of variables $x_1 - x_n$ which do the best job in predicting y. We could do so until $n = k - 1$, when our degrees of freedom are fully exploited. In data science settings, we typically have a large set of potential variables x_i, where $i \leq k$; and if not, we could just add the best performing transformations of our xs' ($ln(x)$, $\sqrt[2]{x}, x^2$ and the like), interaction terms between them, or panel settings lagged values of $xt - n$. However, one can expect such a model to be prone to over-specification, where the model would perform very well for within-sample prediction, but does a bad job of predicting data it was not fitted for. This is obviously not very helpful, neither for academic econometrics nor professional machine learning tasks and therefore such models are in need of techniques that prevent them from over-specification or overfitting, to use the term commonly used in the data science community.

3.2.2 Out-of-Sample validation

Again, as econometricians, we focus on parameter estimates, and we implicitly take their out-of-sample performance for granted. Once we set up a proper identification strategy that delivers unbiased estimates of a causal relationship between x and y, Depending on the characteristics of the sample, this effect supposedly can be generalized on a larger population. Such an exercise is *per se* less prone to over-specification since the introduction of further variables with low predictive power or correlation with x tends to "water down" our effects of interest. Following a machine learning approach geared towards boosting the prediction accuracy of the model, the best way to test how a model predicts is to run it on data it was not fitted for. Here, we randomly divide our data in a *training sample* on which we fit the data, and a *test sample* on which we run the final model. Consequently, we aim at minimizing the *out-of-sample* instead of the *within sample* loss function. Since such a procedure is sensitive to potential outliers in the training or test sample, it is good practice to not validate your model on one single test-sample, but instead perform a *k-fold cross-validation*, where the loss function is computed as the average loss of k (commonly 5 or 10)) separate test samples.[10]

[9] As the name already suggest, this simply expresses by how much our prediction is on average off: $RMSD = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$.

[10] Such k-fold cross-validations can be conveniently done in R with the `caret`, and in Python with the `scikit-learn` package.

While out-of-sample performance is a standard model validation procedure in machine learning, it has yet not gained popularity among econometricians.[11] As a discipline originating from a comparably "small data" universe, it appears counterintuitive for most cases to "throw away" a big chunk of data. However, the size of data-sources available for mainstream economic analysis, such as register data, has increased to a level, where sample size cannot be taken anymore as an excuse for not considering such a goodness-of-fit test, which delivers much more realistic measures of a model's explanatory power. What econometricians like to do to minimize unobserved heterogeneity and thereby improve parameter estimates is to include a battery of categorical control variables (or in panel models, fixed effects) for individuals, sectors, countries, *et cetera*. It is needless to say that this indeed improves parameter estimates in the presence of omitted variables but typically leads to terrible out-of-sample prediction.

3.2.3 Variable selection

Turning back to our problem of out-of-sample prediction, now that we have a good way of measuring it, the question remains how to optimize it. As a general rule, the higher the complexity of a model, the better it tends to perform within-sample, but also to loose predictive power when performing out-of-sample prediction. Since finding the right level of complexity is a crucial, researchers in machine learning have put a lot of effort in developing "regularization" techniques which penalize model complexity.

To stay for a moment in the convenient linear world of OLS models, when assuming the functional form of $f(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$ as given, then minimizing the out-of-sample loss function $L(\hat{y}, y)$ becomes a question of (i.) how many variables, and (ii.) which variables to include. Such problems of *variable selection* are well known to econometricians, which use them mainly for the selection of control variables, including stepwise regressions (one-by-one adding control variables with the highest impact on our $\bar{R}^2$), partial least squares (PLS), different information criterion (e.g., Aikon: AIC, Bayesian: BIC), to only name a few.[12] One class of estimators for penalized linear regression models that lately also became popular among econometricians are the *elastic nets*. Here, we first standardize all our variables $\mu = 0, \sigma = 0$. The parameter estimates for our β are here as usual computed by minimizing the sum of squared residuals (SSR), but includes another term (λ) that penalizes the coefficient by its contribution to the RSME, which has the form of:

[11] However, one instantly recognizes the similarity to the nowadays common practice among econometricians to bootstrap standard errors by computing them over different subsets of data. The difference here is that we commonly use this procedure, (i.) to get more robust parameter estimates instead of evaluating the model's overall goodness-of-fit, and (ii.) we compute them on subsets of the same data the model as fitted on.

[12] For an exhaustive overview on model and variable selection algorithms consider Castle et al. (2009).

$$\lambda \sum_{p=1}^{P} [1 - \alpha|\beta_p| + \alpha|\beta_p|^2]. \quad (1)$$

Of this general formulation, we know two popular cases. When $\alpha = 1$, we are left with the quadratic term, leading to a *ridge regression*. If $\alpha = 0$, we are left with $|\beta_i|$, turning it to a lately among econometricians very popular "Least Absolute Shrinkage and Selection Operator" (LASSO) regression.[13] While maintaining the asymptotic properties under normality, it has proven to be a useful tool for interference on high-dimensional data. [14] Especially among macro-economists, such "tricks" developed in machine learning became popular, since they nowadays face a similar data situation, meaning high n (worldwide country level panels, 50 years+) and large p (uncountable national account, trade, and composite indicators to choose from).[15] In entrepreneurship, however, this situation is usually different, where quantitative research often is based on "relatively" small-scale cross-sectional ($n < 5.000$) surveys. Since large-scale administrative data is increasingly getting available across countries, and alternative data on the actions and interactions of entrepreneurs can now be extracted, as briefly presented earlier, in creative ways, we expect this situation to change. Consequently, familiarizing oneself with new methods of variable selection and dimensionality reduction will become increasingly important for research in the field.

3.2.4 Non-parametric approaches in machine learning

While for OLS and similar regression techniques, the problem of avoiding over-specification and optimizing out-of-sample fit mainly boils down to a variable selection problem, it gets quite a bit more complicated when we do not take the functional form as given. While the mostly followed parametric approach (again, we usually aim for parameter estimates) requires an *a priori* specification of the functional form,[16] methods in machine learning are to a large extend non-parametric, where the functional form is determined within the model. In such models, the balance between within- and out-of-sample prediction is subject to more than one *tuning parameter*, which comes with some challenges.

[13] For an exhaustive discussion on the use of LASSO, consider Belloni et al. (2014). LASSOs are integrated, among others, in the R package `Glmnet`, and Python's`scikit-learn`.

[14] However, the usefulness of LASSOs is limited to data with $n > p$ (what is usually the case in the kind of cross-sectional and panel data we use. If $n < p$, the LASSO will saturate when n variables are selected.

[15] A good example is provided by Hendry and Krolzig (2004) in their paper "We ran one regression", where they use a general unrestricted model (GUM) to eliminate the myriads of alternative growth regression setups. However, increasing computational power also makes room for the more "brute-force" full model search approaches, as Hanck (2016), who ran two trillion growth regressions instead.

[16] An exception here are kernel regressions, which particularly in microeconomics (cf. e.g, Blundell and Duncan, 1998) enjoy some popularity these days.

Let us start with a classical example, where our task is to predict a dichotomous outcome variable. In the machine learning jargon, this is the simplest form of a *classification problem*, where the available classes are 0=no and 1=yes. As econometricians, probably our intuition would lead us to apply a linear probability (LPM) or some form of a logistic regression model. Again, while such models are indeed very useful to deliver parameter estimates, if our goal is pure prediction, there exist much richer model classes.

The data-science toolbox for such problems is rather rich and diverse, but our point here will be a class of models that have proven to be rather powerful but still to some extend interpretable by the human mind, *classification and regression trees* (CART, in business application also known as *decision trees*).[17] The idea behind this approach is to step-wise identify *feature* values (where *feature* is the data science jargon for what we call a *variable*) explaining the highest variance of outcomes. This can be done in various ways, but in principle you aim to at every step use some criterion to identify the most influential feature X of the model (e.g., the lowest p value), and then another criterion (e.g., lowest χ^2 value) to determine a cutoff value of this feature. Then, the sample is split according to this cutoff. This is repeated for every subsample, leading to a tree-like decision structure, which eventually ends at a terminal node (a *leaf*).

If one lets this tree grow unconstrained, it would further split, until every observation ends up in its' own leave, leading to a perfect in-sample fit, and a 100% accuracy of prediction. Needless to say, that such a model would be highly overfitted, and produce poor out-of-sample predictions. As a consequence, like in linear regression models, there is a need to restrict the complexity by some tuning parameter to achieve good out-of-sample prediction performance. The most common way to do so is to "prune" a tree, taking the number of leaves as a tuning parameter. Typically, one aims to strike a balance between the improved *learning rate* (improvement of prediction accuracy) and higher complexity by deciding directly or indirectly on the depth (number of decision layers) in the tree. In practice, that is done by a mix of expert intuition and experience, but also a set of supporting graphical analyses and techniques.

3.2.5 Regularization and model tuning

As already discussed, when only without mechanisms to restrict the model's complexity, such methods are prone to overfitting, particularly when we do not rely on assumptions regarding the asymptotic properties of our sample. What ensures how well our model performs on new data is given by the results out-of-sample testing, where

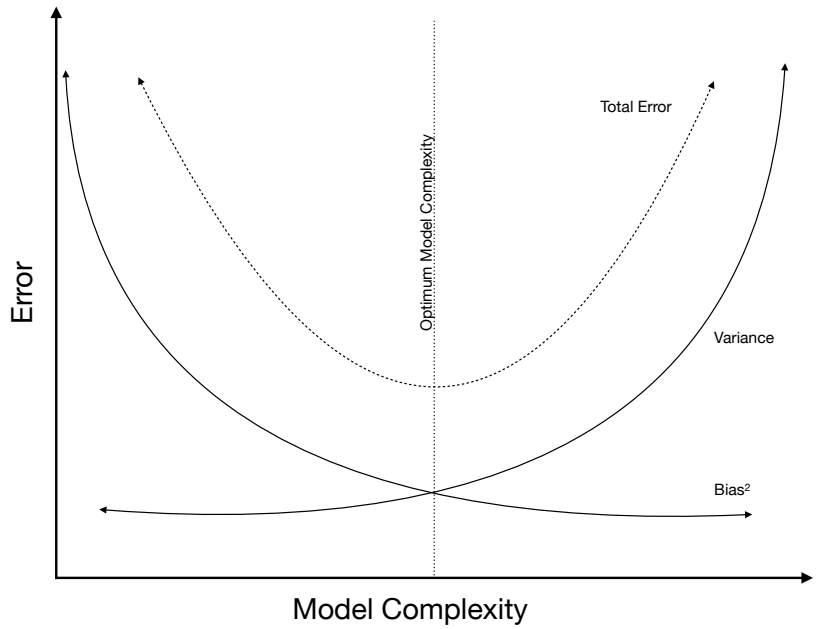
[17] There are quite many packages dealing with different implementations of regression trees in common data science environments, such as `caret, rpart, tree, party` for R, and again the machine learning allrounder `scikit-learn` in Python. For a more exhaustive introduction to CART models, consider Strobl et al. (2009)

prediction quality ($\hat{y} \rightarrow y$) is observable and provides empirical validation. Once we have obtained our out-of-sample results, there is no way back to further model tuning, since then we would again run in danger of overfitting our model on the test sample.

While regression trees are still relatively straightforward to interpret, we already see that model tuning becomes somewhat more intricate and involves more choices and options than the to econometricians well-known problem of variable selection. Flexible functional forms are able to fit models with complex non-linear relationships, which becomes particularly appealing when having high dimensional data with a large number of observations at hand.[18]

For approaches with even more flexible functional forms, for instance, the varieties of neural networks that we review later, it does not get easier. Consequently, the search for optimal tuning parameters (in machine learning jargon called *regularization*)[19] is at the heart of machine learning research efforts, somewhat its "secret sauce". The idea in it's most basic form can be described by the following equation, as expressed by (Mullainathan and Spiess, 2017):

Figure 3: In- vs. out-of-sample loss relationship



[18] Indeed, regression trees appear to show their benefits over traditional logistic regression approaches mostly in settings where we have a large sample size (Perlich et al., 2003), and where the underlying relationships are really non-linear (Friedman and Popescu, 2008).

[19] For exhaustive surveys on regularization approaches in machine learning particularly focused on high-dimensional data, consider Pillonetto et al. (2014); Wainwright (2014)

$$minimize \underbrace{\sum_{i=1}^{n} L(f(x_i), y_i)}_{in-sample\ loss},\ over\ \overbrace{f \in F}^{function\ class}\ \ subject\ to\ \ \underbrace{R(f) \leq c.}_{complexity\ restriction} \tag{2}$$

Basically, we here aim at minimizing the in-sample loss of a prediction algorithm of some functional class subject to some complexity restriction, with the final aim to minimize the expected out-of-sample loss. Depending on the technique applied, this can be done by either selecting the functions features x_i (as we discussed before in "variable selection"), the functional form and class f, or the complexity restrictions c. This process of regularization in practice often is a mixture of internal estimation from the training data, expert intuition, and best practice, as well as trial-and-error. Depending on the complexity of the problem, this can be a quite tedious and lengthy process.footnoteAs an analogy the backward propagation training procedure in neural networks via gradient descent, this process is sometimes humorously called the "graduate-student-descent", where a motivated Ph.D. student spends some time on fiddling around with model parameters until it produces the best results.

The type of regularizations and model tuning techniques one might apply varies, depending on the properties of the sample, the functional form, and the properties of the desired output. Again, the primary aim of regularization is to improve out-of-sample classifier performance by minimizing the loss function. For parametric approaches such as OLS and logistic regressions, regularization is primarily centered around feature selection and parameter weighting, where the above discussed elastic nets are a classic example of such. Many model tuning techniques are iterative, such as model *boosting*, a linear combination of prediction of residuals, where initially misclassified observations are given increased weight in the next iteration. *Bootstrapping*, the repeated estimation of random subsamples, is another technique most applied econometricians are well-acquainted with, yes used for a slightly different purpose. In econometrics, bootstrapping represents a powerful way to circumvent problems arising out of selection bias and other sampling issues, where the regression on several subsamples is used to adjust the standard errors of the estimates. In machine learning, bootstrapping is used primarily to adjust the parameter estimates by weighting them across subsamples (which is then called *bagging*). Similarly, *ensemble* techniques use the weighted combination of different features or even functional forms to determine the final classification of the model.

While it is, for instance, possible to determine which variables across all trees of a random forest on average contributed most to the prediction accuracy, they can in no way be interpreted as causal or global marginal effects. This holds true for most

machine learning approaches and represents a danger for econometricians using them blindly. Again, while an adequately tuned machine learning model may deliver very accurate estimates, it is misleading to believe that a model designed and optimized for predicting $\hat{y}$ *perse* also produces β's with the statistical properties we usually associate with them in econometric models.

3.2.6 Neural Networks and Deep Learning

Regression trees might still be familiar to some econometricians. Now we would like to introduce to another class of models which due to current breakthroughs which delivered unprecedented prediction performance on large high-dimensional data enjoys a lot of popularity: Neural networks. Connecting to the former narrative, neural networks represent *regression trees on steroids*, which are flexible enough to – given enough data – fit every functional form and thereby theoretically can produce optimal predictions to every well-defined problem.

While early ideas about artificial neural networks (ANNs) were already developed in the 1950s and 60s by among others Frank Rosenblatt 1958 and the formal logic of neural calculation described by McCulloch and Pitts (1943), it took several decades for this type of biology-inspired models to see a renaissance in the recent few years.[20]

This revival can be attributed to three reasons: (i) New training techniques, (ii) the availability of large training datasets, and (iii) hardware development, particularly the identification of graphical processing units (GPUs) – normally used, as the name suggests, for complex graphics rendering tasks in PCs and video game consoles – as extremely well suited for modeling neural networks (LeCun et al., 2015).

To understand the neural network approach to modeling, it is essential to get a basic grasp of two main concepts. First, the logic behind the functioning of single neurons[21], and second the architecture and sequential processes happening within ANNs.

A single neuron receives the inputs x_1,x_2,...,x_n with weights w_1,w_2,...,w_n that are passed to it through synapses from previous layer neurons (i.e. the input layer). Given these inputs the neuron will "fire" and produce an output y passing it on to the next layer, which can be a hidden layer or the output layer. In praxis, the inputs can be equated to standardized or normalized independent variables in a regression function. The weights play a crucial role, as they decide about the strength with which signals

[20] It has to be stressed that even though neural networks are indeed inspired by the most basic concept of how a brain works, they are by no means mysterious artificial brains. The analogy goes as far as the abstraction that a couple of neurons that are interconnected in some architecture. The neuron is represented as some sigmoid function (somewhat like a logistic regression) which decides based on the inputs received if it should get activated and send a signal to connected neurons, which might again trigger their activation. Having that said, calling a neural network an artificial brain is somewhat like calling a paper-plane an artificial bird.

[21] for the sake of simplicity here we will not distinguish between the simple perceptron model, sigmoid neurons or the recently more commonly used rectified linear neurons (Glorot et al., 2011)

are passed along in the network. As the network learns, the initially randomly assigned weights are continuously adjusted. As the neuron receives the inputs, it first calculates a weighted sum of $w_i x_i$ and then applies an activation function ϕ.

$$\phi(\sum_{i=1}^{m} w_i x_i) \tag{3}$$

Depending on the activation function the signal is passed on or not.

Figure 4: Illustration of a neuron

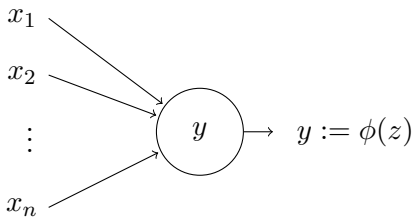
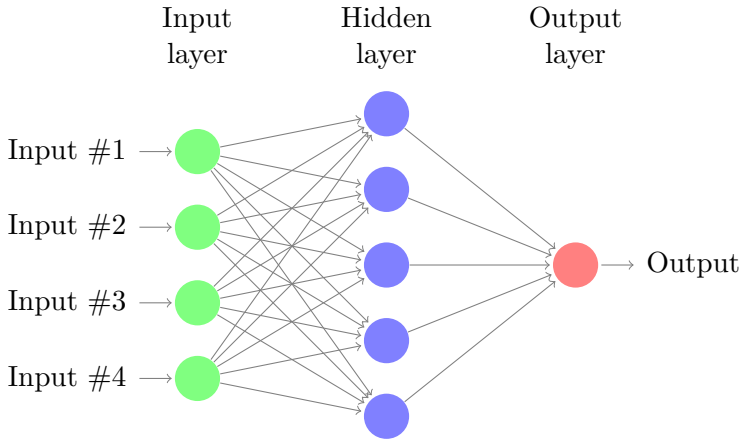


Figure 5 represents an artificial neural network with three layers: One input layer with four neurons, one fully connected hidden layer with five neurons and one output layer with a single neuron. As the model is trained for each observation inputs are passed on from the input layer into the neurons of the hidden layer and processed as described above. This step is repeated and an output value $\hat{y}$ is calculated. This process is called forward propagation. Comparing this value with the actual value y (i.e. our dependent variable for the particular observation) allows to calculate a *cost function* e.g. $C = \frac{1}{2}(\hat{y} - y)^2$. From here on *backpropagation*[22] is used to update the weights. The network is trained as these processes are repeated for all observations in the dataset.

Figure 5: Illustration of a neural network



[22]This complex algorithm adjusts simultaneously all weight in the network, considering the individual contribution of the neuron to the error.

Artificial neural networks have many interesting properties that let them stand out from more traditional models and make them appealing when approaching complex pattern discovery tasks, confronting nonlinearity but most importantly dealing with large amounts of data in terms of the number of observations and the number of inputs. These properties, coupled with the recent developments in hardware and data availability, led to a rapid spread and development of artificial nets in the 2010s. Today, a variety of architectures has evolved and is used for a large number of complex tasks such as speech recognition (Recurrent neural networks: RNNs and Long Short Term Memory: LSTMs), computer vision (CNNs and Capsule Networks, proposed in late October 2017) and as backbones in artificial intelligence applications. They are used not only because they can approach challenges where other classes of models struggle technically but rather due to their performance. Despite the numerous advantages of artificial neural nets they are yet rarely seen in entrepreneurial and more general social science research. Here CNNs may be so far the most often used type, where its properties were employed to generate estimates from large image datasets Gebru et al. (e.g. 2017). The simplest architecture of a CNN puts several convolutional, and pooling layers in front of an ANN. This allows transforming images, which are technically two-dimensional matrices, into long vectors, while preserving the information that describes the characteristic features of the image.

The predictive performance of neural nets stands in stark contrast to the explainability of these models, meaning that a trained neural net is more or less a black box, which produces great predictions but does not allow to make causal inference. In addition, this leads to asking: What is the reason the model produced this or that prediction. This becomes particularly important when such models are deployed for instance in diagnostics or other fields to support decision making. There are several attempts to address this problem under the heading of "explainable AI" (e.g. Ribeiro et al., 2016).

3.3 Unsupervised Machine Learning at the example of Vector Space Models

In this section, we illustrate unsupervised machine learning at the more specific example of Vector Space Models (VSMs) and their application to natural language data. While dimensionality reduction and clustering approaches would also fall into this category, we find that VSMs are a particularly interesting example of techniques that have a great potential to contribute to entrepreneurship and more generally social science research.

VSMs originate in abstract algebra, were developed in the 1970s for information retrieval and are mainly associated with the field of Natural Language Processing (NLP). This is mainly because VSMs exploit a core feature of language which is that terms that appear together (co-occurrence) tend to be related and *vice-versa*. VSMs are frequently used for topic modeling using natural text, where given a corpus – a set of documents – they can identify latent topics. Despite their close link with NLP they can be applied more generally to arbitrary collections of item groups to detect "topics" based on various co-occurrence measures. The most widely used variations of VSMs are TF-IDF (term frequency-inverse document frequency), Latent Semantic Indexing (LSI) (Deerwester et al., 1990) and Random Indexing. Latent Dirichlet Allocation (LDA) is a popular topic modeling approach but not a VSM.

Let's consider the following strongly simplified example of the functioning of LSI: When applied to text, the training process takes a bag of words (BoW) corpus (a collection of usually preprocessed documents), for instance, descriptions of entrepreneurial crowdfunding campaigns. In the first step, a sparse matrix is created, mapping each document onto the set of terms appearing in all documents. Such a matrix would describe a *n-space* where n is the number of unique terms in the corpus. In the following step, this n-dimensional matrix is collapsed using singular-value decomposition (SVD) and the number of dimensions reduced to the desired amount m where m is also the number of topics that are identified. Such a compression may seem counterintuitive and appear to lead to information loss, but in fact, it reduces noise, emphasizing similarities and maintaining differences. The created document-topic matrix (the topic model) can be used to calculate document similarity between all input documents, allowing to group documents semantically. Furthermore, new documents can be projected into the generated vector space. This class of models is often used for mapping and exploration but can also create inputs for analytical or predictive models. In the crowdfunding case we could, for instance, use document vectors of campaign description as independent inputs in a regression model where the outcome variable would be some campaign performance measure. The results of such an analysis could give some indication of successful semantic features. In a predictive setting, a well-trained model could make a prediction of campaign success based on a description draft.

While VSMs and other topic models are powerful and relatively easy to use (e.g. using *gensim* for Python or *topicmodels* for R), they are increasingly crowded out by various deep learning approaches, such as word embedding e.g. Word2Vec (Mikolov et al., 2013), deep neural nets (mainly RNNs) or some combined architectures.

4 Opportunities and Challenges of Big Data and Machine Learning for Quantitative Entrepreneurship Research

We depicted towards some common approaches of variable selection and dimensionality reduction in econometrics and data science, yet the question remains: Are there any synergies between the two approaches, and do data science methods have the potential to enhance theory building in social science and improve our ability to provide meaningful advice to business and policy?

We argue for a clear yes, and in the following point towards some areas where we see the greatest potentials. These are (i.) simply in improving statistical models particularly with explicit out-of sample evaluation, (ii.) the utilization oh high dimensional and granular data to enhance our understanding of human behavior, (iii.) understanding the nature of rare events, (iv.) the generation of quality indicators to quantify complex and up to now often unmeasurable theoretical concepts, (v.) doing quantitative research in data-sparse environments, (vi.) the mapping of dynamic entrepreneurial ecosystems, and (vii.) to provide real-time policy advice. We, however, also identify a set of challenges to be overcome for such opportunities arise.

4.1 Opportunities

In the following, we therefore outline what we consider the main synergies of big data, econometrics, and machine learning methods

Better estimates The first and obvious way econometrics can, and already to some extent does benefit from data science techniques would be to improve the parameter estimates of models by improving the up to now "implicitly predictive" parts of the model building strategy. Here, econometricians already started to apply such tricks as LASSO and other forms of elastic nets to identify relevant control variables among a large number of candidates (which is typically the case in macro studies), which can be computed reasonably efficient while boosting the model's predictive power.

In the same vein, machine learning techniques can be used for choosing instruments in 2SLS and similar regression settings. There we face a variable selection problem, where the aim is to find an instrument z which optimally does a good job in predicting x but also has further characteristics such as being uncorrelated with ϵ. In cases where we do not have the luxury of an undoubtedly exogenous instrument which also delivers a reasonable $\hat{x}$ (which, sadly, is more often not the case than it is), adaption regularization techniques such as LASSO and other elastic nets (eg. Belloni et al., 2012; Carrasco, 2012; Hansen and Kozbur, 2014), and even neural network approaches (cf. Hartford et al., 2016, who develop "Deep Instrumental Variables Networks") have

proven practical. Likewise, techniques such as *genetic matching* (Diamond and Sekhon, 2013), a method of multivariate matching that uses an evolutionary search algorithm following the concepts introduced earlier, and similar machine learning approaches (e.g. Chernozhukov et al., 2016), represent a powerful and convenient alternative to create propensity scores for all kind of matching exercises.[23] Furthermore, machine learning techniques are useful for improving the accuracy of all kind of predictive "side-tasks" we usually discuss only in the footnotes, such as imputation of missing values,[24] , or name disambiguation and matching.[25]

Lastly, the practice of out-of-sample testing might help to improve the external validity of our models by explicitly testing how good our model performs in terms of parameter estimates and overall prediction (Athey and Imbens, 2017).

Understanding human behavior Furthermore, as discussed earlier, new non-traditional data sources tied to individuals or companies are starting to become publicly available on a large scale. While in the past particularly Scandinavian researchers have been "spoiled" by their access to population-wide data on for example individuals education and employment history, similar large-scale public datasets are starting to become available all around the globe (Einav and Levin, 2014b). Besides just increasing our sample size and assuring a p-value associated with three asterisks in our regression analysis, such data has proven valuable for nuanced investigations of subsamples, and to measure concepts which are usually operationalized binary instead in a continuum. For example, Østergaard et al. (2011) use Danish data from the matched employer-employee database (IDA) to measure the effect of employee diversity on firm-level innovation activities, where the variance and amount of available data allowed them to measure firm level employee diversity as a continuous index. Taking this approach one step further, Coad and Timmermans (2014) utilize a non-parametric approach to shed light on the composition, structure, and performance of entrepreneurial pairs, where the structure was operationalized as a matrix of all possible combinations of individual characteristics (age, education, gender and so forth).

Additionally, private companies such as Amazon, Facebook, Linkedin, and Google hold an enormous amount of data. But also less known providers of all kinds of mobile applications (language learning, mobility, health ect.), digital payment and more

[23] Creating propensity scores is a process that typically involves a research assistant running regressions for a weekend or so. Nowadays, packages such as `matching` in R are good examples how machine learning makes repetitive labor tasks redundant.

[24] Indeed, many of the readers will find themselves to be users of machine learning approaches already, since the newer versions of the `mice` package in R for "Multivariate Imputation by Chained Equations" deploys a neural network in the background.

[25] machine learning approaches such as the use of VSMs for author name disambiguation(e.g. Arif et al., 2014) have proven as quite powerful, and making the "name game" (Raffo and Lhuillery, 2009) while working with patents, publications, and similar data sources, an order of magnitude more convenient.

broadly fintech companies,as well as and manufacturers producing consumer goods that include sensors of some sort e.g. everything within the "Internet of things" or that use use sensors to measure internal processes (industry 4.0, advanced manufacturing) have piled up considerable data repositories which can be utilized to put different behavioral theories to a test.[26] While such data is not always easy to obtain, a fair share of companies displays general willingness to cooperate with academic research institutions and share their data. Undoubtedly, such cooperation may make intensive upfront coordination, sophisticated data access and storage processes, as well as sensible confidentiality agreements necessary. But the promises of such data sources are likely to justify such efforts.

As such, the creative use of granular and high dimensional (big) data and machine learning techniques has emerged as a strong fit with behavioral economics (Mahmoodi et al., 2017; Taylor et al., 2014), and for instance been used to evaluate behavioral models of choice under risk and ambiguity Peysakhovich and Naecker (2017). Furthermore, current realizations that the combination of machine learning and high dimensional granular data from a persons *digital footprint* – the trace left in social media, the bloogosphere, online shopping activities, cellphone data *et cetera* – enables surprisingly accurate prediction on psychological characteristics and other personal traits (e.g., Kosinski et al., 2013, 2016) as well as preferences and interests (e.g., Raghuram et al., 2016, demonstrate how to predict individuals interest by their twitter activity). It is not hard to see that such techniques could spur renewed interest in research on the entrepreneurial personality, emotions, and behavior. Recently, a set of publications mainly applying natural language processing techniques to data obtained from twitter represent first interesting demonstrations of such approaches and call for further work (Obschonka and Fisch, 2017; Obschonka et al., 2017; Tata et al., 2017).

Rare event prediction One task our traditional econometrics toolkit has performed particularly bad, is the explanation but also prediction of extremely *rare events*. However, being able to explain impactful low probability events (also coined as "black swans, cf." Taleb, 2010) such as which start-up is going to be the next gazelle, which technology our patent is going to be the futures "next big thing", when does the next financial crisis hit or firm defaults (cf. e.g. van der Vegt et al., 2015), and so forth, would certainly be of enormous interest for research, policy, and business alike. One obvious challenge is, as the name implies that such events are not observed often, leading to insufficient samples to analyze. While in the area of big data, and with the progress of time generally, the number of rare events observed tends to increase,

[26] Such private data naturally tends to carry various types of selection biases. Yet, if that *per se* disqualifies them in comparison to experiments with the typical sort of "representative" agents (undergraduate students), then a stronger argument against their use has to be found.

that will likely not solve the problems econometricians face in predicting or explaining them. The bigger problem here are the models we would use for such a task, most probably some kind of logistic regression or related duration models. If the task is to fit a regression to predict something happening in, let's say, one in a million cases, the best model we would produce would predict only failures, which would be highly accurate since it only would predict wrong in one in a million cases. Here, the flexible functional form and ability to handle high dimensional data inherent to many machine learning model classes such as deep neural networks have proven to do a way better job for such tasks. While such models are, as discussed earlier, prone to overfitting, they for sure only can be taken seriously after performing well in an out-of-sample test. Promising approaches include the work of Cheng et al. (2014) on the prediction of online *cascades* (topics, posts, videos etc. which go viral, associated with rapid growth in mentions or shares), which could potentially be adapted to predict extremely high firm growth.

4.2 Challenges

In this chapter, we made the case that a "liaison of econometrics and data science" indeed creates numerous opportunities to enhance quantitative research in the field of entrepreneurship. But we also see potential challenges and barriers that may hamper the progress along this avenue of research. Some are more practical and related to inter-disciplinary boundaries, yet others are more epistemic and related to intra-disciplinary methodological comfort zones.

To begin with, the growth of data useful for quantitative research in the field of economics, in volume and variety alike, is projected to continue at an exponential rate, making traditionally used data manipulation tools, statistical software, and techniques for analysis increasingly inadequate to handle its sheer complexity and exploit its potentials. Machine learning as a discipline historically emerged in a data-rich environment and has developed many useful techniques to conducts analyses that appreciate and utilize this richness to obtain the highest information gain. Further, necessity as well as the disciplinary closeness to computer science enabled the machine learning community to develop, design and apply methods and workflows to gather, store, and process such enormous amounts of data. And again, with vast we do not refer to the econometrics understanding of a "few million" datapoints, but rather many billions, or only a few with millions of variables (as it is the case for example when analyzing gene samples), or some ten terabytes of image data.

For exploratory analysis, one might want to work on small sub-samples. When stored in a relational database, and accessed via a Structured Query Language (SQL), that performs reasonably well for mid-sized data volumes, and can be learned with

reasonable effort. However, with increasing data size and complexity, one soon wants to change to "NoSQL", which offer more flexibility in how data is stored and accessed. Currently, many different database technologies such as MongoDB, As technology progresses fast in the big data ecosystem, a detailed mapping would add limited value, but for an exhaustive overview of current systems and practices, we point towards Grover and Kar (2017). The bottom line is that the efficient use of (really) big data requires knowledge of current IT hardware, infrastructure, systems, and often also the programming languages (e.g., Java, C) to address them. This rarely overlaps with the skillset of most econometricians.[27] We are confident that econometricians can manage mid-sized machine learning projects with reasonable efforts. However, more ambitious projects, like many of the examples we presented in this chapter, might be in need of closer cross-departmental cooperation with trained computer scientists.

When finally submitting the data to analysis, one will further find most of the traditional proprietary statistical software that have been the standard tools of applied econometricians to be limited in its usefulness. The main and obvious reason is that most statistical software was developed in – and optimized for – a paradigm where data comes in a comparably small scale and is neatly organized as numerical values in a spreadsheet. This fundamental logic and architecture (e.g., working with only a single dataset at a time) are hard to adapt to a new paradigm of large, distributed, and unstructured data sources. However, there are further reasons. Since the machine learning approach is more geared towards deriving algorithms with high predictive performance while econometricians focus on deriving desired statistical properties Wu et al. (2008), machine learning methodology is less constrained in its development of new methods and techniques. Here, an algorithms usefulness is not necessarily in need to fulfill specific statistical requirements, including extended mathematical proofs and theorems. When an algorithm predicts well enough out of sample, it is useful. And since prediction is useful in many domains in almost all areas of science and business alike, new and more powerful techniques are developed at a high frequency by a large cross-disciplinary and mostly cooperative community. The result of such developments mainly manifests in the creation of freely available packages and libraries for open-source programming languages. Currently, the by far most popular programming environments for statistics and machine learning alike are R and Python. Both are relatively easy to learn for anybody with experience in statistical programming in proprietary software solutions. Hence, we highly recommend econometricians interested in leveraging the potential of machine learning in their research to do so.

To fit a complex and computationally intense model class such as a deep neural network on a big dataset, the general architecture is mostly created on a dataset of a small subsample, and only the final "production model" is trained on the whole

[27] This is also true for the typical campus IT support, for that matter.

dataset. The obvious reason is that the training of a deep neural network on large amounts of data most likely exceeds the computing power of even a premium laptop,[28] and should better be run on an appropriate high-performance computer, or directly send to the cloud (eg., Amazon Web Services). This is usually done within an ML&AI framework such as Tensorflow, Pytorch, Caffe or Theano, and parallelized with tools such as MapReduce (making some knowledge of λ-calculus necessary), which again requires some expertise beyond causal statistical coding. To wrap up, big data and machine learning workflows are generally accessible for a typical econometrician, yet beyond a certain scale the involvement of computer scientists becomes necessary, at least to set up the general architecture.

5 Conclusion, ways forward, and avenues for future research

After outlining as what we consider the main opportunities but also challenges in the "liaison of econometrics and data science", we now would like to provide some first suggestions on how to overcome these challenges and leverage the opportunities in future research.

During the last two years (status late 2017) we witness a paradigm shift, where editorial boards of highly ranked journals call for methodological pluralism and the exploration of big data and machine learning approaches (e.g., George et al., 2016), and such approaches also get published. Indeed, AI and big data have become the fastest growing topic in social and natural science research alike (Akoka et al., 2017). So far, such an emerging paradigm shift could up to now not be observed in the field of entrepreneurship research.[29] We hope that to change in the near future and encourage efforts in this directions. We believe that particularly the field of entrepreneurship research, dealing with the complex interactions of individuals and the socio-techno-economic they are embedded in, could considerably benefit from such approaches which enables the analysis of rare events and multidimensional concepts that can not necessarily be captured with a neat functional form.

[28] Interestingly enough, recent breakthroughs have shown that deep neural networks are can be trained on the computers graphic processor (GPU) orders of magnitude faster than on the CPU. This might at least in the near future to a situation where researchers in machine learning on conferences easily can be spotted by their "gaming laptops" with blinking keyboards, which tend to have very powerful GPUs (cf. Shi et al., 2016), and are at the moment the first choice for mobile training of deep neural nets. This amusing fact is, however, important to be communicated to the campus IT procurement office.

[29] To be fair, no rule is without exceptions, such as Obschonka and Fisch (2017); Obschonka et al. (2017); Tata et al. (2017), in the recent call (announced just as we are writing this chapter) for papers in *Small Business Economics: An Entrepreneurship Journal* on "Rethinking the entrepreneurial (research) process: Opportunities and challenges of Big Data and Artificial Intelligence for entrepreneurship research".

To further stimulate such a paradigm shift, we are also in need of adjusting our academic training system. In contrast to the last decade, it now is increasingly realistic the get a tenure position as a "heterodox econometrician". However, we also have to incentive and train the new generation of graduate and undergraduate social science students to embrace this new heterodox approach to data analysis and make use of the richness of new data sources and methods while still pursuing research that aims to rigorously identify robust causal effects.

Promising areas of research incorporating big data and machine learning approaches are by far too numerous to list here. Yet, we still make an attempt to point towards a couple of directions we consider particularly promising.

First, extracting the digital footprint of entrepreneurs, particularly from social media and networks, provides access to three type of valuable information, which are up to now seldom used and analyzed on large scale. The first and obvious information contained is the entrepreneurs social and professional network. Former research clearly indicates the embeddedness of entrepreneurs in such networks to affect their behavior, access to potential customers and finance, and general business success. But we still lack large-scale and nuanced quantitative research to verify many of the claims made. Still, some challenges are likely to arise, for example, the identification of entrepreneurs in social networks (e.g.,matching with register data), the disentanglement of social and professional networks (e.g., via classification of links with machine learning techniques), and how to capture the dynamics of such networks (e.g., publicly available Facebook or Linkedin data does not come with a timestamp indicating when a connection was formed). Second, social networks can reveal much of an entrepreneurs private, educational, and professional history (e.g., extracted from Linkedin Hain et al., 2017), which can be utilized to analyze the progress of becoming an entrepreneur, but also to explain things such as access to finance, performance, and general survival. Third, text data produced by the entrepreneurs represents a rich source of information to classify them statically (psychological traits, preferences), or follow their behavioral or emotional change in a dynamic way (e.g., with semantic language analysis such as LDI).

Such data on subsamples or a whole population of entrepreneurs can probably be matched with investment data from for instance venture capital or crowdfunding databases, to examine how networks, characteristics and traits influence their access to particular types of funding, and surely the impact thereof on success and behavior.

When zooming out even further, such data can be used to map the structure and dynamics of entrepreneurial ecosystems on various scales. Combining different forms of data (e.g., social media, satellite imagery, register data, investments), fine-grained pictures of such ecosystems, including the dynamics of beliefs, sentiments, interaction, institutions and so forth can be produced and analyzed.

Finally, all approaches mentioned before rely only to a minimal extent on governmental register data hence can also be applied to populations which tend to be underrepresented in quantitative entrepreneurship research, such as semi-formal entrepreneurs, entrepreneurship in less developed countries, and nascent entrepreneurs. One notoriously underrepresented population could potentially be identified in such an exercise: failed entrepreneurs, a population where we up to now know very little about the reasons for failure, coping strategy, alternative career-paths and so forth.

References

Aguinis, H., Pierce, C. A., Bosco, F. A., and Muslin, I. S. (2009). First decade of organizational research methods: Trends in design, measurement, and data-analysis topics. *Organizational Research Methods*, 12(1):69–112.

Akgün, A. A., Nijkamp, P., Baycan, T., and Brons, M. (2010). Embeddedness of entrepreneurs in rural areas: a comparative rough set data analysis. *Tijdschrift voor economische en sociale geografie*, 101(5):538–553.

Akoka, J., Comyn-Wattiau, I., and Laoufi, N. (2017). Research on big data–a systematic mapping study. *Computer Standards & Interfaces*, 54:105–115.

Andrews, R. J., Fazio, C., Guzman, J., and Stern, S. . (2017). The startup cartography project: A map of entrepreneurial quality and quantity in the united states across time and location. MIT Working Paper.

Arif, T., Ali, R., and Asger, M. (2014). Author name disambiguation using vector space model and hybrid similarity measures. In *Contemporary Computing (IC3), 2014 Seventh International Conference on*, pages 135–140. IEEE.

Athey, S. and Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *The Journal of Economic Perspectives*, 31(2):3–32.

Beaudry, C., Héroux-Vaillancourt, M., and Rietsch, C. (2016). Validation of a web mining technique to measure innovation in high technology canadian industries. OECD Blue Sky Forum on Science and Innovation Indicators, Ghent, Belguim.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650.

Blazquez, D. and Domenech, J. (2017). Big data sources and methods for social and economic analyses. *Technological Forecasting and Social Change*.

Blumenstock, J., Cadamuro, G., and On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264):1073–1076.

Blumenstock, J. E. (2016). Fighting poverty with data. *Science*, 353(6301):753–754.

Blundell, R. and Duncan, A. (1998). Kernel regression in empirical microeconomics. *Journal of Human Resources*, pages 62–87.

Carrasco, M. (2012). A regularization approach to the many instruments problem. *Journal of Econometrics*, 170(2):383–398.

Castle, J. L., Qin, X., Reed, W. R., et al. (2009). How to pick the best regression equation: A review and comparison of model selection algorithms. Working Paper No. 13/2009, Department of Economics and Finance, University of Canterbury.

Chen, H., Chiang, R. H., and Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS quarterly*, 36(4).

Cheng, J., Adamic, L., Dow, P. A., Kleinberg, J. M., and Leskovec, J. (2014). Can cascades be predicted? In *Proceedings of the 23rd international conference on World wide web*, pages 925–936. ACM.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., et al. (2016). Double machine learning for treatment and causal parameters. *arXiv preprint arXiv:1608.00060*.

Coad, A. and Timmermans, B. (2014). Two's company: Composition, structure and performance of entrepreneurial pairs. *European Management Review*, 11(2):117–138.

de la Paz-Marín, M., Campoy-Muñoz, P., and Hervás-Martínez, C. (2012). Non-linear multiclassifier model based on artificial intelligence to predict research and development performance in european countries. *Technological Forecasting and Social Change*, 79(9):1731–1745.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990). Indexing by latent semantic analysis. *JASIS*, 41(6):391–407.

Diamond, A. and Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, 95(3):932–945.

Dong, L., Chen, S., Cheng, Y., Wu, Z., Li, C., and Wu, H. (2017). Measuring economic activity in china with mobile big data. *EPJ Data Science*, 6(1):29.

Droll, A., Khan, S., Ekhlas, E., and Tanev, S. (2017). Using artificial intelligence and web media data to evaluate the growth potential of companies in emerging industry sectors. *Technology Innovation Management Review*, 7(6).

Einav, L. and Levin, J. (2014a). The data revolution and economic analysis. *Innovation Policy and the Economy*, 14(1):1–24.

Einav, L. and Levin, J. (2014b). Economics in the age of big data. *Science*, 346(6210):1243089.

Ernst, M. and Jurowetzki, R. (2016). Satellite data, women defectors and black markets in north korea: A quantitative study of the north korean informal sector using nighttime lights satellite imagery. *North Korean Review*, 12(1).

Fazio, C., Guzman, J., Murray, F., and Stern, S. (2016). A new view of the skew: Quantitative assessment of the quality of american entrepreneurship. MIT Innovation Initiative Paper.

Friedman, J. H. and Popescu, B. E. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics*, pages 916–954.

Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E. L., and Fei-Fei, L. (2017). Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states. *Proceedings of the National Academy of Sciences*, page 201700035.

George, G., Osinga, E. C., Lavie, D., and Scott, B. A. (2016). From the editors: Big data and data science methods for management research. *Academy of Management Journal*, 59(5):1493–1507.

Glaeser, E. L., Kominers, S. D., Luca, M., and Naik, N. (2016). Big data and big cities: The promises and limitations of improved measures of urban life. *Economic Inquiry*.

Glorot, X., Bordes, A., and Bengio, Y. (2011). Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 513–520.

Grover, P. and Kar, A. K. (2017). Big data analytics: A review on theoretical contributions and tools used in literature. *Global Journal of Flexible Systems Management*, pages 1–27.

Guzman, J. and Stern, S. (2015). Where is silicon valley? *Science*, 347(6222):606–609.

Guzman, J. and Stern, S. (2017). Nowcasting and placecasting entrepreneurial quality and performance. In Haltiwanger, J., Hurst, E., Miranda, J., and Schoar, A., editors, *Measuring Entrepreneurial Businesses: Current Knowledge and Challenges*, chapter 2. University of Chicago Press.

Hain, D., Christensen, J., and Jurowetzki (2017). The value of human capital signals for investment decision making under uncertainty – an analysis of cross-border venture capital investments in europe and sub-saharan africa. *Proceedings of the 10^{th} European Commission Conference for Corporate R&D and Innovation (CONCORDi), Seville, Spain*.

Hajek, P. and Henriques, R. (2017). Modelling innovation performance of european regions using multi-output neural networks. *PloS one*, 12(10):e0185755.

Hanck, C. (2016). I just ran two trillion regressions. *Economics Bulletin*, 36(4):2037–2042.

Hansen, C. and Kozbur, D. (2014). Instrumental variables estimation with many weak instruments using regularized jive. *Journal of Econometrics*, 182(2):290–308.

Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. (2016). Counterfactual prediction with deep instrumental variables networks. *arXiv preprint arXiv:1612.09596*.

Henderson, J. V., Storeygard, A., and Weil, D. N. (2012). Measuring economic growth from outer space. *The American Economic Review*, 102(2):994–1028.

Hendry, D. F. and Krolzig, H.-M. (2004). We ran one regression. *Oxford bulletin of Economics and Statistics*, 66(5):799–810.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., and Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794.

Jurowetzki, R. and Hain, D. S. (2014). Mapping the (r-) evolution of technological fields – a semantic network approach. In Aiello, L. M. and McFarland, D., editors, *Social Informatics*, volume 8851 of *Lecture Notes in Computer Science*, pages 359–383. Springer International Publishing.

Kim, J., Han, M., Lee, Y., and Park, Y. (2016). Futuristic data-driven scenario building: Incorporating text mining and fuzzy association rule mining into fuzzy cognitive map. *Expert Systems with Applications*, 57:311–323.

Kosinski, M., Stillwell, D., and Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805.

Kosinski, M., Wang, Y., Lakkaraju, H., and Leskovec, J. (2016). Mining big data to extract patterns and predict real-life outcomes. *Psychological methods*, 21(4):493.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Kwon, H., Kim, J., and Park, Y. (2017). Applying lsa text mining technique in envisioning social impacts of emerging technologies: The case of drone technology. *Technovation*.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436.

Lee, J., Jang, D., and Park, S. (2017). Deep learning-based corporate performance prediction model considering technical capability. *Sustainability*, 9(6):899.

Li, Y., Arora, S., Youtie, J., and Shapira, P. (2016). Using web mining to explore triple helix influences on growth in small and mid-size firms. *Technovation*.

Lobell, D. B. (2013). The use of satellite data for crop yield gap analysis. *Field Crops Research*, 143:56–64.

Mahmoodi, J., Leckelt, M., van Zalk, M. W., Geukes, K., and Back, M. D. (2017). Big data approaches in social and behavioral science: four key trade-offs and a call for integration. *Current Opinion in Behavioral Sciences*, 18:57–62.

McAfee, A., Brynjolfsson, E., Davenport, T. H., et al. (2012). Big data: the management revolution. *Harvard business review*, 90(10):60–68.

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Mullainathan, S. and Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

Obschonka, M. and Fisch, C. (2017). Entrepreneurial personalities in political leadership. *Small Business Economics*, pages 1–19.

Obschonka, M., Fisch, C., and Boyd, R. (2017). Using digital footprints in entrepreneurship research: A twitter-based personality analysis of superstar entrepreneurs and managers. *Journal of Business Venturing Insights*, 8:13–23.

Østergaard, C. R., Timmermans, B., and Kristinsson, K. (2011). Does a different view create something new? the effect of employee diversity on innovation. *Research Policy*, 40(3):500–509.

Park, E., Hain, D. S., and Jurowetzki, R. (2017). Entrepreneurial ecosystem for technology start-ups in nairobi: Empirical analysis of twitter networks of start-ups and support organizations. *Proceedings of the 17th DRUID Summer Conference, New York, USA*.

Pennington, J., Socher, R., and Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Perlich, C., Provost, F., and Simonoff, J. S. (2003). Tree induction vs. logistic regression: A learning-curve analysis. *Journal of Machine Learning Research*, 4(Jun):211–255.

Peysakhovich, A. and Naecker, J. (2017). Using methods from machine learning to evaluate behavioral models of choice under risk and ambiguity. *Journal of Economic Behavior & Organization*, 133:373–384.

Pillonetto, G., Dinuzzo, F., Chen, T., De Nicolao, G., and Ljung, L. (2014). Kernel methods in system identification, machine learning and function estimation: A survey. *Automatica*, 50(3):657–682. cited By 115.

Raffo, J. and Lhuillery, S. (2009). How to play the names game: Patent retrieval comparing different heuristics. *Research Policy*, 38(10):1617–1627.

Raghuram, M., Akshay, K., and Chandrasekaran, K. (2016). Efficient user profiling in twitter social network using traditional classifiers. In *Intelligent Systems Technologies and Applications*, pages 399–411. Springer.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.

Shi, S., Wang, Q., Xu, P., and Chu, X. (2016). Benchmarking state-of-the-art deep learning software tools. *arXiv preprint arXiv:1608.07249*.

Stephens-Davidowitz, S. and Varian, H. (2015). A Hands-on Guide to Google Data.

Strobl, C., Malley, J., and Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods*, 14(4):323.

Taleb, N. (2010). *The black swan: The impact of the highly improbable*. Random House Trade Paperbacks.

Tata, A., Martinez, D. L., Garcia, D., Oesch, A., and Brusoni, S. (2017). The psycholinguistics of entrepreneurship. *Journal of Business Venturing Insights*, 7:38–44.

Taylor, L., Schroeder, R., and Meyer, E. (2014). Emerging practices and perspectives on big data analysis in economics: Bigger and better or more of the same? *Big Data & Society*, 1(2):2053951714536877.

Te, Y.-F. and Cvijikj, I. P. (2017). Design of a small and medium enterprise growth prediction model based on web mining. In *International Conference on Web Engineering*, pages 600–607. Springer.

van der Vegt, G. S., Essens, P., Wahlström, M., and George, G. (2015). Managing risk and resilience. *Academy of Management Journal*, 58(4):971–980.

Varian, H. R. (2014). Big data: New tricks for econometrics. *The Journal of Economic Perspectives*, 28(2):3–27.

Wainwright, M. (2014). Structured regularizers for high-dimensional problems: Statistical and computational issues. *Annual Review of Statistics and Its Application*, 1:233–253. cited By 24.

Wang, F., Mack, E. A., and Maciewjewski, R. (2017). Analyzing entrepreneurial social networks with big data. *Annals of the American Association of Geographers*, 107(1):130–150.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Philip, S. Y., et al. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1):1–37.

Appendix

Figure 6: Map of machine learning classes, techniques, and algorithms